The evolutionary analysis of the vertebrate two-round whole genome duplications

Masatoshi Matsunami

DOCTOR OF PHYLOSOPHY

Department of Genetics,

School of Life Science,

The Graduate University for Advanced Studies

(SOKENDAI)

Acknowledgements

My study would not have been possible without help of others. Firstly, I would like to thank my academic supervisor, Naruya Saitou who provide me many helpful discussions, suggestions and comments. I would like to appreciate my committee for helping me to achieve this degree. I thank all the member of Saitou lab, past and present, especially Yosuke Kawai who taught me the basic of programming, Yukuto Sato who provide me helpful discussion and an opportunity to analyze the data of second generation sequencing machine, and Kiyoshi Ezawa who gave me constructive criticisms and advices. I also thank Yoichi Nakatani for giving the paralog data, Takahiko Kawasaki and Yasunori Murakami for lamprey samples, and the member of Fujiyama lab for the help of sequencing.

I appreciate Shigehiro Kuraku and the members of his lab for natural history of the genome, Axel Meyer and the members of his lab at the Universität Konstanz. They allowed me short stay and gave valuable advises for my research.

Lastly, I would also like to express my gratitude to my family and friends for their moral and financial support and warm encouragement.

TABLE OF CONTENTS

List of Figures

List of Tables

Abbreviations

WGD:	whole genome duplication
SSD:	small scale duplication
2R WGD:	two-round whole genome duplications
CNS:	conserved non-coding sequences
TFBS:	transcription factor binding site
TP:	tetra-paralogous
DP :	di-paralogous
RARE:	retinoic acid response element
FCS :	four cluster sequence
UCE:	ultra-conserved element
CNE:	conserved non-coding element
DDC model:	duplication - degeneration – complementation model
NGS:	next-generation sequencing

ABSTRACT

Two-rounds whole genome duplications (2R WGD) occurred in the vertebrate ancestors, and they generated large numbers of duplicated protein-coding genes and their regulatory elements. These events could contribute to the emergence of vertebrate-specific features. However, the evolutionary impact of the 2R WGD is still unclear. To address this issue, I conducted comprehensive studies on both protein-coding and non-coding sequences found in the conserved synteny blocks generated by the 2R WGD. Such conserved synteny blocks are expected to retain duplicated protein-coding and gene regulatory sequences. Consequently, evolutionary changes or some constraints relating to these blocks would have played important roles in the evolution and diversification of vertebrates. On the basis of this view, I focused on evolution of both protein-coding and non-coding sequences of the vertebrate genomes, especially Hox clusters.

Because a part of gene regulatory elements are expected to be conserved according to their functional importance, evolutionarily conserved non-coding sequences (CNSs) might be good candidates of gene regulatory elements. In addition, portion of the paralogous protein-coding genes retained after the 2R WGD show

overlapping expression pattern. Therefore, paralogous genes might share gene expression regulatory mechanisms. Paralogous CNSs have possibility to control overlapping expression patterns of those paralogs. Thus, detecting paralogous CNSs and inferring the relation between paralogous gene and CNSs is important to understand evolution after the 2R WGD.

Four or more paralogous Hox clusters exist in vertebrate genomes because of the 2R WGD. The paralogous genes in the Hox clusters show similar expression patterns, implying shared regulatory mechanisms for expression of these genes. Previous studies partly revealed the expression mechanisms of Hox genes. However, *cis*-regulatory elements that control these paralogous gene expression are still poorly understood. Toward solving this problem, I searched CNSs within vertebrate Hox clusters. I compared orthologous Hox clusters of 19 vertebrate species, and found 208 intergenic conserved regions. I then searched for CNSs that were conserved not only between orthologous clusters but also among the four paralogous Hox clusters. I found three regions that are conserved among the all four clusters and eight regions that are conserved between intergenic regions of two paralogous Hox clusters. In total, 28 CNSs were identified in the paralogous Hox clusters, and nine of them were newly found in this study. One of these novel regions bears a RARE motif. These CNSs are candidates

for gene expression regulatory regions among paralogous Hox clusters. I also compared vertebrate CNSs with amphioxus CNSs within the Hox cluster, and found that two CNSs in the HoxA and HoxB clusters retain homology with amphioxus CNSs through the 2R WGD.

The duplication histories of vertebrate Hox clusters are controversial. Under the assumption of the 2R WGD, phylogenies of Hox gene should show a symmetrical topology. However, some previous studies did not support this symmetrical topology. I thus carried out exhaustive phylogenetic analysis of deuterostome Hox genes. First, to identify outgroup genes of each vertebrate Hox paralog group, I inferred the correct ortholog/paralog relationships among deuterostome posterior Hox genes by comparing available Hox genes. Amphioxus Hox9-11 were generated by amphioxus specific tandem duplications. Because vertebrate Hox10-12, and Hox14-15 genes have no counter parts in amphioxus Hox genes, they were probably lost in the amphioxus lineage. Secondly, the duplication histories of vertebrate Hox genes were inferred by constructing phylogenetic trees and phylogenetic networks. My analysis suggested that the ((A,B), (C,D)) topology is most suitable explanation of Hox cluster duplications.

I then carried out genome-wide identification of paralogous CNSs. A sensitive BLAST search of each synteny block revealed 7,924 orthologous CNSs and 309

paralogous CNSs conserved among 8 high quality vertebrate genomes. I newly detected 194 paralogous CNSs. Their locations are biased nearby the transcription factors coding regions shown expression in brain and neural system. The existence of these paralogous CNSs is difficult to explain by previous duplication models. Because these sequences have same transcription factor binding motifs, they might be backup of paralogous gene expression and/or contribute to the interaction between paralogs.

The 2R WGD occurred after the split of the urochordate ancestors but before the diversification of extant gnathostomes (jawed vertebrates). However, there is no clear evidence whether the timing of the 2R WGD is before or after the split of agnathans (jawless vertebrates including lamprey) and gnathostomes. To clarify this problem is highly important for study of vertebrate evolution and development. The lamprey gene data are also useful for molecular function and developmental studies. Thus, I analyzed the mRNA sequences of Japanese brook lamprey (*Lethenteron reissneri*) and estimated the relative timing of the 2R WGD by combining newly obtained sequence data from Japanese brook lamprey and sea lamprey (*Petromyzon marinus*) data in the database.

The Japanese brook lamprey cDNAs were synthesized from the mRNAs of ammocoetes larva and were sequenced by Roche 454 GS FLX titanium system. After

the assembly of 426,476 sequence reads, I obtained 7,708 contigs with 336 bp length on average. Additionally, I also analyzed the sea lamprey mRNA sequencing data in the SRA database. Including 119,412,170 reads, they were assembled to 78,947 contigs. Based on these lamprey data, I analyzed putative orthologous and paralogous gnathostome sequences corresponding to the lamprey contigs to estimate the relative timing of the 2R WGD. From the homologous gene clustering, phylogenetic trees of 358 gene families are reconstructed. However, if I restrict trees which contain two duplication events and have high statistical supports, only 55 trees were left. The majority (49) of them showed the pattern that two genome duplications both occurred before the lamprey divergence

Recently, the sea lamprey (*Petromyzon marinus*) genome sequences appeared in the public database including 11,429 genes. I also investigated the possibility that gene losses caused misunderstanding of true ortholog/paralog relationships by using these newly released sea lamprey data, as well as with 13 gnathostomes and 6 nonvertebrate species genome data. I reconstructed phylogenetic trees of 545 gene families, and there were 127 trees with one agnathan (A) and two gnathostomes (G) clusters. Although 69 trees showed topology ((A,G),G) suggesting two duplications before the agnathans/gnathostomes divergence, the remaining 58 trees had topology

((G,G),A). I compared the branch lengths connecting the gnathan common ancestor and the agnathan/gnathostomes common ancestor, and found that ((G,G),A)-topology trees had the significantly longer branch than ((A,G),G)-topology trees. This suggests that agnathan genes were lost in the lamprey lineage in ((G,G),A)-topology trees, and the occurrence of duplications erroneously looked like after the agnathans/gnathostomes divergence. I thus conclude that 2R WGD occurred before agnathans/gnathostomes divergence.

CHAPTER 1

General Introduction

1.1 Mode of duplication

Darwin (1859) argued for natural selection as a creative force of new functions in his "*The Origin of Species*". Although the power of natural selection in removing disadvantageous variants was clear, many biologists doubted whether it could build wholly new structures. The canonical work on the subject is Ohno's (1970) "*Evolution by Gene Duplication*", in which he stressed the importance of gene duplication and considered the various types of duplications and their potential for yielding novel functions.

There are two types of duplications. These are whole-genome duplication (WGD) and small-scale duplication (SSD). Both WGD and SSD can produce different kinds of adaptations (Wapinski et al. 2007). Previous studies found a pattern of negative correlation between genes fixed in duplicate after SSD events and those surviving from WGDs (Maere et al. 2005; Wapinski et al. 2007). Duplicates produced by WGD also seem to share more protein interactions after duplication than do genes duplicated by SSD (Guan et al. 2007; Hakes et al. 2007). Moreover, products of WGD are often

highly expressed (Seoighe et al. 1999) and are more likely to show an overexpression phenotype or haploinsufficiency than other duplicates (Wapinski et al. 2007). Strangely, although SSDs tend to be created from genes with smaller than average knockout fitness defects, enzymes that are retained in duplicate after WGD seem to have fitness defects at least as large as those for the genes that are not retained (DeLuna et al. 2008). The WGD can lead to the retention of duplicates of genes whose dosage balance is potentially important (such as transcription factors), whereas this class of gene is rarely duplicated by SSD. This idea implies that WGD events might allow certain evolutionary novelties to appear and be selected for that would have been unlikely to arise otherwise.

1.2 History of genome duplication study

While genome duplications in animals are now well documented, the existence of a polypoid vertebrate that is a salamander *(Ambystoma jeffersonianum*) was accepted in 1960s (Uzzell 1964), much later than studies of plant polyploidization. The first polyploid frogs (*Odontophrynus americanus* and *Ceratophrys ornata*) were described in 1966 (Saez et al. 1966). Although they provided clear figures showing multiple sets of chromosomes and multivalent formation during meiosis, their conclusion did not suggest the existence of WGD event. Bogart (1967) later confirmed that these were both

octoploid species that reproduced bisexually. Earlier research on fish also suggested that polyploidy played a major role in the speciation and the diversification of the Salmonidae (Svärdson 1945) and the genus *Coregonus* (Kupka 1948). However, these were discounted by some researchers.

After these reports, Ohno (1970) addressed the importance of the ancient genome duplications. The possibility that the genome duplication has played an important role in animal evolution has received much attention since the discovery of them (Donoghue et al. 2005; Volff 2005). In contrast to the animals, genome duplications in plants are the focus of modern genomic research not only due to their economic importance, but also due to the much larger than expected genomic signatures of ancient WGD events. A large fraction of plant genomes is generated by duplication, partly because of the frequent occurrence of genomic segmental duplications and polyploidization events in plants. For example, in the Arabidopsis thaliana and rice genomes up to 90% and 62% of loci are duplicated, respectively, and it is estimated that 70–80% of angiosperm species have undergone polyploidization at some point in their evolutionary history (Moore et al. 2005). However, even in plants, we still do not have a complete understanding of the factors that promote the formation and establishment of WGD in the wild, the role ecology plays in polyploid speciation, and whether

polyploidy accelerates diversification rates or is an evolutionary dead end (Levin 2002; Soltis et al. 2010).

1.3 Genome duplications of vertebrates

In the vertebrate evolution, WGDs sometimes occurred (Lewis 1980; Otto and Whitton 2000; Le Comber and Smith 2004; Gregory 2005). However, WGDs are most common in organisms that do not regulate their internal temperature like plants and ectothermic animals (Gregory 2005). Why do some groups are polyploid and others not? Although it is possible that intrinsic mechanisms regulating genome integrity constrain WGD establishment, it may also be possible that ecological factors (living in habitats or conditions that favor polypoidy), in combination with the inherently stochastic nature of establishment of polyploid lineages. Formation in the midst of diploid progenitor (Husband 2000) and producing balanced chromosome sets are some of these inherently stochastic natures.

The WGDs occurred in the vertebrate genomes are divided into ‘ancient’ (i.e. paleopolyploid) WGDs and ‘recent’ WGDs. The ‘recent’ polyploid species usually have twice chromosome number of close relatives. Those ‘recent’ polyploid events are often

occurred in amphibian and fish lineages (Mable et al. 2011). One lineage specific WGD event in mammal is reported (Gallardo et al. 1999). However, this WGD remains unresolved, because to ascertain polyploidy is technically very difficult (Gallardo et al. 2004; Svartman et al. 2005; Gallardo et al. 2006). By contrast, the 'ancient' WGD are known as the two-rounds whole genome duplications (2R WGD) and fish specific genome duplication (FSGD). In this study, I focused on the 2R WGD from these genome duplication events, because these events might generate the vertebrate specific features (Lundin et al. 2003).

1.4 Hox clusters are the hallmarks of the 2R WGD

The Hox genes regulate animal body plans. They were discovered from fruit fly. The mutations of these homeobox (Hox) genes have powerful and interpretable effects on morphology, the most conspicuous being the homeotic transformation in *Drosophila melanogaster* (Lewis 1978; Kaufman et al. 1990). Hox genes are present and expressed in similar patterns in nearly every bilateral animal that has been analyzed, so their roles in morphological diversification probably evolved before the appearance of the first bilateral animal. Indeed, the initial glimpses into the conservation of metazoan developmental control genes came during the study of *D. melanogaster* Hox gene

clusters (McGinnis et al. 1992).

The Hox clusters are also the hallmark of the 2R WGD study. The all deuterostome invertebrates so far studied has only one Hox cluster (Lemons et al. 2006). Major tetrapod species have four Hox clusters in their genome. The identification of Hox quadrupled regions strongly supported the existence of the 2R hypothsis (Lundin 1993; Ruddle et al. 1994). The teleosts have approximately twice number of the Hox clusters, compared with tetrapod species. Additional Hox clusters have been identified in teleost fish occupying different taxonomic positions. The mapping of Hox clusters and many duplicated genes in several fish suggested an extra WGD in ray-finned fish (Amores et al. 1998; Woods et al. 2000; Amores et al. 2004; Naruse et al. 2004). After the finding of duplicated Hox clusters in teleost genomes, genome-wide gene comparison was done (Vandepoele et al 2004; Christoffels et al. 2004). The result indicated a fish-specific large-scale duplication event (called fish specific WGD or 3rd WGD). The definitive proof that a more recent WGD occurred in teleost fish has important consequences for the 2R hypothesis because it indicates that WGD and not segmental duplication was the duplication mechanism responsible for the origin of the additional Hox clusters in this clade. Therefore, people could accept that the Hox clusters are reliable markers of WGDs. However, proofing the existence of the 2R

WGD by genome-wide comparison was difficult at that time, because syntenic outgroup genomes were unavailable.

1.5 Vertebrate genome evolution after the 2R WGD

The 2R hypothesis was proven after the amphioxus genome was sequenced (Putnam et al. 2008). Before the amphioxus genome was reported by Putnam et al. (2008), the 2R hypothesis was extensively debated (e.g., Holland et al. 1994; Gibson and Spring 2000; Hughes et al. 2001; Dehal and Boore 2005). Because gene synteny comparison between amphioxus and tetrapod species shows 1:4 ratio in almost genomic regions, the existence of the 2R WGD is now widely accepted. However, we have unsolved problems about the evolution after the 2R WGD.

First, the duplication history of paralogs derived from the 2R WGD is unclear. If the 2R WGD events occurred, the tree topology of the paralogous genes, say A, B, C, and D, should show a symmetrical ((A,B)(C,D)) topology. However, many gene families show not symmetrical but asymmetrical topology, including Hox genes (Hughes et al. 2001). There is a possibility of homogenization such as recombinations, crossovers and conversions.

Second, the impact of the 2R WGD about the gene expression is unknown.

Genome duplications generated paralogous genes and complex gene regulatory mechanisms in vertebrate evolution. These paralogous genes often share the same expression patterns, but some might have acquired new expression patterns. The changes of gene expression mainly resulted from changes in *cis*-regulatory elements (Carroll 2001). Because gene regulatory elements are expected to be conserved due to their functional importance, searching for evolutionarily conserved non-coding sequences (CNSs) would be an effective strategy for finding candidates of functional elements. Previous studies have already shown that CNSs are abundant in vertebrate genomes (Bejerano et al. 2004; Woolfe et al. 2005). Genome-wide comparative approaches have also reported the existence of paralogous CNSs (Bejerano et al. 2004; Woolfe et al. 2005; McEwen et al. 2006), and most of them are located in paralogous gene clusters that code for transcriptional factors. These results imply that paralogous CNSs contribute to cluster organization and/or their neighboring gene expression patterns. However, paralogous CNSs derived from the 2R WGD are still unclear.

Third, the relative timing of the 2R WGD is not determined yet. Jawless vertebrates (i.e. hagfish and lamprey) branched off at the early timing of the vertebrate evolution. Ohno (1970) speculated that all vertebrate share the 2R WGD events. Force et al. (2002) suggested that at least one duplication of Hox cluster occurred before the

divergence of gnathostome and jawless vertebrates, whereas an independent cluster duplication occurred in the lamprey lineage, after it diverged from the gnathostome lineage. Fried et al. (2003) argued for an independent duplications of these Hox clusters and suggested that the common ancestor of agnathans and gnathostomes had a single Hox cluster. Recently, Kuraku et al. (2008) estimated that gnathostome and jawless vertebrates shared the 2R WGD events by using 55 gene family data. These results are contradictory with each other. We need more reliable genome-wide analysis to clarify the relative timing of the 2R WGD events.

There are some reasons to explain why the relative timing of the 2R WGD is so important. First, the 2R WGD events are deeply related to the acquirement of vertebrate novel structures (listed in Figure1.1), especially neural crest cells. The vertebrate novel structures are derived mainly from the neural crest cells. Hall (2000) considers vertebrates to be not merely usual triploblastic animals, but quadroblastic, with the neural crest constituting a fourth germ layer. Holland and Chen (2001) have even proposed calling vertebrates and their fossilized precursors "cristozoa", the "crest-animals". However, it is unclear that how these unique neural crest cells arose. This is critically important question in evolutionary developmental biology, because it goes to the heart of evolutionary novelty and the origin of vertebrate. The 2R WGD may

contribute to the emergence of this new type cells. Holland et al. (1996) suggested that the origin of the neural crest cells involves the genome duplications. If the 2R WGD events occurred before agnathan divergence and vertebrate share these events, the origin of the neural crest cells is clearly related to the 2R WGD. Otherwise, if vertebrate share only one genome duplication event, the origin of the neural crest cells is related to only the 1st-round WGD. The 2nd-round WGD may generate other vertebrate features, such as jaws, bones and limbs. If vertebrate share no genome duplication events, there is no relation between genome duplications and the origin of neural crest cells. In this case, we should reconsider the origin of neural crest cells. We, thus, can show the genomic change that contributes to the emergence of vertebrate novel structures, if the relative timing of the 2R WGD are identified. Second, developmental biologist use homologous lamprey genes as makers of orhologous structures, in spite of their uncertain orthologies. Because the definition of orthologous structure is difficult, especially evolutionary separated species, these definition sometimes cause misinterpretations. The identification of timing of the 2R WGD will help to show true orthologous structures.

1.6 Questions of this study

Previous studies show the evidence of the 2R WGD events. In this study, I

dissected the unsolved problems of the 2R WGD in different pieces. I focused on especially non-coding region, Hox clusters, and lamprey genome evolution.

In chapter 2, I discussed paralogous conserved gene regulatory elements within the vertebrate Hox clusters. These elements are conserved through the vertebrate evolution and may play important roles in Hox gene expressions. I then concentrate on the Hox gene phylogeny in chapter 3. In this chapter, possible gene duplication history of the vertebrate Hox clusters is reconstructed by using not only phylogenetic tree, but also phylogenetic networks. In chapter 4, I will focus on the genome-wide survey of paralogous non-coding sequences derived from the 2R WGD. These highly conserved sequences are very important when we infer the gene regulatory evolution after the 2R WGD. I will show the challenge of de novo RNA sequencing of Japanese brook lamprey in chapter 5. The next generation sequencers are recent cutting edge techniques. These equipments make it possible to read massive sequence data by low cost. By using these data, the relative timing of the 2R WGD is inferred. I continue to discuss the relative timing of the 2R WGD events in chapter 6. The sea lamprey genome data are recently released, and these data provide us the genome-wide comparison between jawless vertebrates and jawed vertebrates. The timing of the 2R WGD is estimated with a high confidence for the first time. These studies will help a further understanding of

the 2R WGD events.

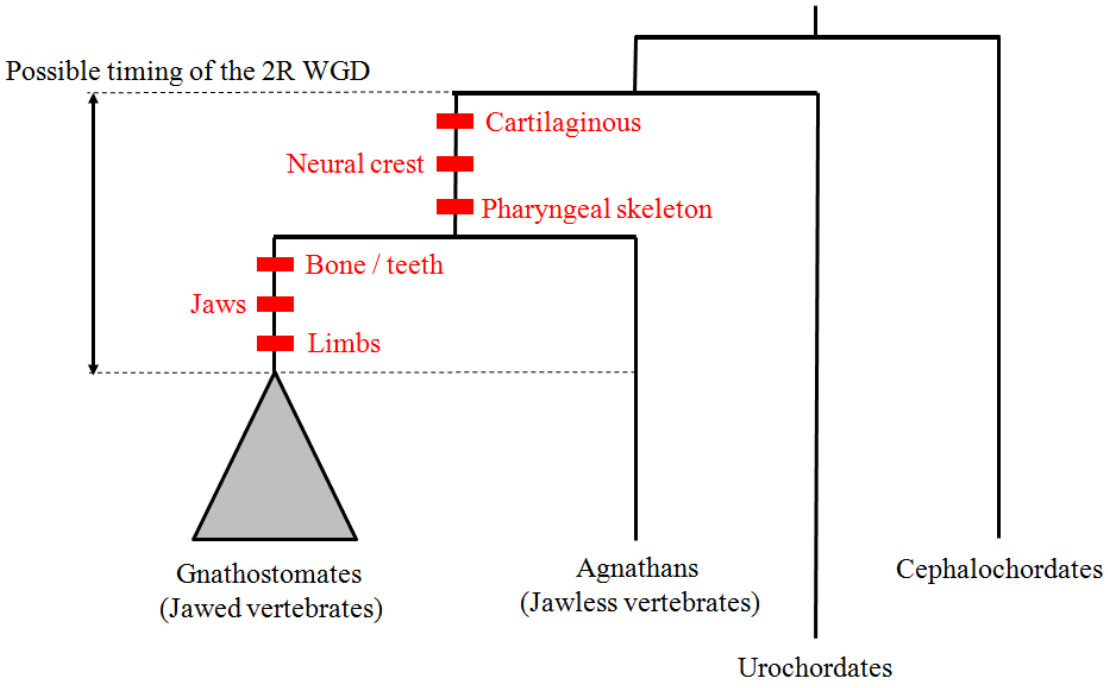


Figure 1.1: Morphological novelty of vertebrate lineages

CHAPTER 2

Evolution of Conserved Non-Coding Sequences Within the Vertebrate Hox Clusters Through the Two-Round Whole Genome Duplications

2.1 Introduction

Vertebrate genomes show evidence of widespread gene duplications compared to invertebrate genomes. Ohno (1970) proposed the existence of two-round whole genome duplications (2R WGD) during the early vertebrate evolution, now known as the 2R hypothesis. Before the amphioxus genome was reported by Putnam et al. (2008), this hypothesis was extensively debated (e.g., Holland et al. 1994; Gibson and Spring 2000; Hughes et al. 2001). Genome duplications generated paralogous genes and complex gene regulatory mechanisms in vertebrate evolution (e.g., Dehal and Boore 2005). These paralogous genes often share the same expression patterns, but some may acquire new expression patterns. The changes of gene expression are mainly resulted from changes in *cis*-regulatory elements (Carroll 2001).

Identifying the *cis*-regulatory sequences that control spatial and temporal gene expression is a challenging issue. Because gene regulatory elements are expected to be

conserved due to their functional importance, searching for evolutionarily conserved non-coding sequences (CNSs) would be an effective strategy for finding candidates of functional elements. We should note that the gene regulatory elements which are not conserved are very rare (Weirauch and Hughes 2010). Previous studies have already shown that CNSs are abundant in vertebrate genomes (Bejerano et al. 2004; Woolfe et al. 2005). Genome-wide comparative approaches have also reported the existence of paralogous CNSs (Bejerano et al. 2004; Woolfe et al. 2005; McEwen et al. 2006), and most of them are located in paralogous gene clusters that code for transcriptional factors. These results imply that paralogous CNSs contribute to cluster organization and/or their neighboring gene expression patterns. I therefore focused on the vertebrate Hox clusters because they contain abundant CNSs.

The Hox genes orchestrate the development of animal body plans. They consist of more than four physically linked clusters in different chromosomes in vertebrate genomes (Pearson et al. 2005; Lemons and McGinnis 2006). Hox genes of each cluster are expressed along the anterior-posterior body axis in the same order as lining up on the chromosome, a feature called “colinearity” (Garcia- Fernàndez 2005). Paralogous genes of the Hox clusters show the similar expression pattern, which suggests that there might be shared gene expression regulatory mechanisms among paralogous Hox

clusters.

The duplication of Hox clusters influences cluster architecture and patterns of non-coding sequence evolution. The duplicated non-coding regions within the Hox clusters are mainly studied for teleost fish (e.g., Chiu et al. 2002; Santini et al. 2003; Prohaska et al. 2004). The third round whole genome duplication occurred after the 2R WGD in the teleost lineage. Chiu et al. (2002) and Prohaska et al. (2004) found massive loss of sequence conservation in teleost HoxA cluster non-coding regions after the 3R WGD. Therefore, teleosts are not suitable for analyzing duplicated Hox cluster non-coding sequences.

In the case of 2R WGD, Kim et al. (2000) described one paralogous CNS within the four Hox clusters. However, analysis of non-coding sequences of the Hox clusters within vertebrates, especially mammalian species, is not sufficient. There are probably two reasons for this. First, the functional paralogous conservation cannot be detected easily. This is because the 2R WGD were very ancient events which occurred approximately half a billion years ago and the non-coding sequences experienced higher evolutionary rates compared to protein coding sequences. This is probably because *cis*-regulatory elements are redundant and may be changed by binding site turnover (Hancock et al. 1999). Secondly, only a few invertebrate sequences that are more

closely related to vertebrates and that still retain cluster structure are available. With the recent abundance in vertebrate genomes sequences, we can now analyze the evolution of non-coding sequences within the Hox clusters after 2R WGD. However, identifying CNSs within Hox clusters before 2R WGD remains a challenge.

Recently, Hox cluster sequences of two different amphioxus species, *Branchiostoma floridae* and *B. lanceolatum* were reported by Amemiya et al. (2008) and Pascual-Anaya et al. (2008), respectively. Because amphioxus is the chordate bearing a syntenic Hox cluster which is most closely-related to vertebrates, these data would be very informative for inferring the evolution of non-coding regions within Hox clusters before 2R WGD.

Detection of the functional turnover of transcription factor binding site (TFBS) is one interesting problem. In the *Drosophila* genome, the TFBS turnover frequently occurred (Ludwig et al. 2005). Ray et al. (2008) developed a program to find the functional turnover motifs by using experimental results as training data. Some *cis*-regulatory regions showed the TFBS turnovers also in vertebrates (Weirauch et al. 2010). But these data are difficult to utilize for finding other functional turnover events for various reasons such as insufficient experimental data, short alignment length, and low mutation rate. Therefore I did not examine the functional turnover of the TFBS in

this study.

In this study, I identified orthologous CNSs within the vertebrate Hox clusters, and found conserved loci among paralogous Hox clusters. I compared these CNSs with amphioxus-human CNSs reported by Pascual-Anaya et al. (2008) by using phylogenetic footprinting to find CNSs that can be dated back to amphioxus. This study identified and mapped vertebrate CNSs within the four vertebrate Hox clusters by using comprehensive genome comparisons.

2.2 Materials and Methods

2.2.1 Identification of vertebrate Hox CNSs

Genomic sequences of Hox clusters were obtained for the following 18 vertebrate species from UCSC Genome Bioinformatics (http://genome.ucsc.edu/): Human (*Homo sapiens*), mouse (*Mus musculus*), chimpanzee (*Pan troglodytes*), orangutan (*Pongo pygmaeus abelii*), rhesus macaque (*Macaca mulatta*), marmoset (*Callithrix jacchus*), rat (*Rattus norvegicus*), guinea pig (*Cavia porcellus*), cat (*Felis catus*), dog (*Canis familiaris*), horse (*Equus caballus*), cow (*Bos taurus*), opossum (*Monodelphis domestica*), platypus (*Ornithorhynchus anatinus*), chicken (*Gallus gallus*), zebra finch (*Taeniopygia guttata*), lizard (*Anolis carolinensis*), and frog (*Xenopus*

tropicalis). Partial sequences of the horn shark (*Heterodontus francisci*) that included Hox clusters (DDBJ/EMBL/GenBank accession numbers are AF224262 and AF224263) were also used for this study. I excluded teleost fishes, which have undergone the additional genome duplication in their lineages. Protein coding regions were filtered based on the RefSeq project (http://www.ncbi.nlm.nih.gov/RefSeq/) annotation. Alternative exons were not considered in this analysis. BLAST homology search (Altschul et al, 1997) was performed on this data set with default parameter setting and cutoff scores of >200.

Orthologous CNSs were systematically named based on their genomic locations and BLAST scores. For example, the CNS that is located at the intergenic region between HoxA7 and HoxA6 with the highest BLAST score was named “A76-1”.

These CNSs were aligned by using CLASTALW (Thompson et al. 1994), and divided into three categories to investigate the depth of conservation: placental mammals, amniotes, and vertebrates. I then searched for conserved sequences that were conserved not only between orthologous clusters but also among paralogous four Hox clusters by using BLAST search with the cutoff score of less than 30. Annotations of TFBS motifs were mainly based on the TRANSFAC database (http://www.biobase-international.com/pages/index.php?id= transfac).

2.2.2 Analysis of paralogous CNSs

To investigate the non-coding transcribed regions of the Hox clusters, transcriptional information of mRNAs and ESTs within the human and mouse Hox clusters were obtained from the UCSC Genome Bioinformatics database and these transcripts were mapped on the region.

Phylogenetic footprinting analysis was carried out for each orthologous CNSs that also have paralogous conservation. Each vertebrate CNS was aligned by using CLASTALW. The substitution number of each aligned site was estimated parsimoniously by using Fitch's (1971) algorithm. The guide phylogenetic tree (**Figure A2.1**) necessary for this analysis was taken from Murphy et al. (2004). In parallel, the likelihood estimation of ancestral sequence of each vertebrate CNS was inferred by using PAML 4 (Yang 2007).

2.2.3 Comparison with amphioxus Hox CNSs

Pascual-Anaya et al. (2008) compared Hox clusters of two different amphioxus species (*Branchistoma floridae* and *B. lanceolatum*) to each human Hox cluster and defined 75 human-amphioxus CNSs (amphiCNS). These amphiCNSs were obtained and

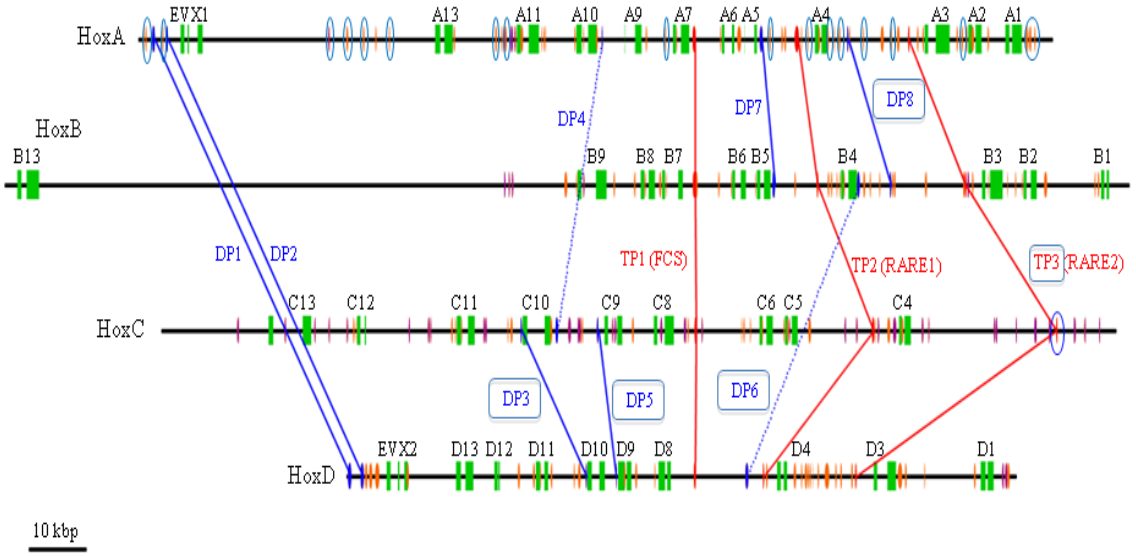


Figure 2.1: The schematic diagram of orthologous CNSs and paralogous CNSs among human Hox clusters

Exons of protein coding genes are represented by light green boxes. The orange ovals are orthologous CNSs. The blue and red ovals indicate locations of paralogous CNSs conserved also among the two clusters and the four clusters, respectively. The blue dotted lines show either microRNA (DP4) or non-syntenic DP (DP6). The paralogous CNSs whose name enclosed by blue rectangle is newly detected. Especially the newly detected HoxC CNS of TP3 is highlighted by blue circle. The light blue circled HoxA CNSs were not identified by Prohaska et al. (2004). Abbreviations are TP; tetra-paralog, DP; di-paralog.

Table 2.1: Conservation depth of each CNS

	HoxA	HoxB	HoxC	HoxD
Placental mammals	4	5	33	2
Above + Marsupials	2	11	0	1
Above + Monotremes	8	6	3	6
Amniotes	17	7	8	7
Tetrapods	17	14	16	9
Vertebrates	16	-	-	16
Total	64 (48)	43 (43)	60 (60)	41 (25)

Note. There is no genomic sequence data for horn shark HoxB and HoxC clusters, so "Vertebrates" depth CNSs are not determined, as shown with hyphens. Because of this, values in parentheses in "Total" are those excluding CNSs shared in all vertebrates.

were sorted by identity. I named amphiCNSs by the order of their identity. The amphiCNSs were compared with vertebrate CNSs to identify significant conserved region among chordates.

2.3 Results

2.3.1 Orthologous CNSs within vertebrate Hox clusters

I defined 208 CNSs in total: 64, 43, 60, and 41 for HoxA, B, C, and D clusters, respectively. Genomic locations of these CNSs are graphically shown in **Figure 2.1**, and detailed information of all these CNSs is shown in **Table A2.1**. Many of these orthologous CNSs overlap microRNAs and *cis*-regulatory elements which are previously described (Mainguy et al. 2003; Yekta et al. 2004). Because sequence information is not complete or homologous sequence is lacking, some CNSs were not found in several species (see **Table A2.2**). As an example of a *cis*-regulatory element, C98-1 corresponds to the HoxC8 early enhancer which is necessary for proper HoxC8 expression (Juan and Ruddle 2003). Other CNSs might bear similar enhancer functions.

Our findings are consistent with previous observations (Prohaska et al. 2004; Chiu et al. 2002), confirming that orthologous CNSs were detected effectively. Moreover, by using our criteria, I also detected 160 new CNSs (see **Figure 2.1** and

Table A2.1).

I detected a larger number of CNSs in Hox5-Hox3 (corresponding to *Drosophila Antp* and *Ubx/abdA*) intergenic sequences than in other intergenic sequences (**Figure 2.1**). This region has abundant alternatively spliced coding RNAs and long non-coding RNAs (Mainguy et al. 2007). This observation suggests that functionally unknown CNSs in this region contribute to these alternative splicing events. In contrast, posterior regions of Hox clusters have poor conservation except for upstream regions of Evx1 and Evx2.

The 208 CNSs were divided into six categories: placental mammals, placental mammals + marsupials, placental mammals + marsupials + monotremes, amniotes, tetrapods, and vertebrates, based on the depth of conservation (**Table 1.1**). The level of conservation of orthologous CNSs varies among the four Hox clusters; HoxA has the highest number (64) of CNSs in total, while HoxD has the smallest number (41) of CNSs due to the small numbers of CNSs conserved among amniotes and tetrapods. The HoxC cluster has the highest number (33) of CNSs conserved only among placental mammals, while the HoxB cluster has the highest number (11) of CNSs in placental mammals + marsupials. This result, however, does not mean that the HoxC cluster is the least conserved (see Discussion).

Table 2.2: Possible functions of Tetra (TP) and Di (DP) paralogous CNSs

Name	ID	Function	Putative TFBS	References
(A) TP CNSs				
TP1	A76-1	Anterior Hox promoter[a]	Homeobox, E-box	Kim et al. (2000),This study
	B76-1	Anterior Hox promoter[a]	Homeobox, E-box	Kim et al. (2000),This study
	C86-1	Anterior Hox promoter[a]	Homeobox, E-box	Kim et al. (2000),This study
	D84-2	Anterior Hox promoter[a]	Homeobox, E-box	Kim et al. (2000),This study
TP2	A54-2	Hox4 Enhancer	RARE	Mainguy et al. (2003)
	B54-3	Hox4 Enhancer	RARE	Mainguy et al. (2003)
	C54-3	Hox4 Enhancer	RARE	Mainguy et al. (2003)
	D84-3	Hox4 Enhancer	RARE	Mainguy et al. (2003)
TP3	A43-7	Hox3 Enhancer	RARE	Mainguy et al. (2003)
	B43-3	Hox3 Enhancer	RARE	Mainguy et al. (2003)
	C4-3	Hox4 Enhancer[a]	RARE	This study
	D43-9	Hox3 Enhancer	RARE	Mainguy et al. (2003)
(B) DP CNSs				
DP1	E1-1	Hox13 Enhancer (distal limb enhancer)	PPAR-α, GATA-1, POU1F1a, Homeobox	Lehoczky et al. (2004)
	E2-1	Hox13 Enhancer	PPAR-α,	Spitz et al. (2001)

		(distal limb enhancer)	GATA-1, POU1F1a, Homeobox	
DP2	E1-2	Hox13 Enhancer (distal limb enhancer)	C-Myb, Homeobox, YY1	Lehoczky et al. (2004)
	E2-3	Hox13 Enhancer (distal limb enhancer)	C-Myb, Homeobox, YY1	Spitz et al. (2001)
DP3	C1110-2	Hox10 enhancer[a]	SF1, CP1, Homeobox	This study
	D1110-3	Hox10 enhancer[a]	SF1, CP1, Homeobox	This study
DP4	A109-2	MicroRNA (mir-196 family)	-	Yekta et al. (2004)
	C109-2	MicroRNA (mir-196 family)	-	Yekta et al. (2004)
DP5	C109-4	Hox9 enhancer[a]	GR, E-box, CAT-box	This study
	D109-1	Hox9 enhancer[a]	GR, E-box, CAT-box	This study
DP6	B43-1	Hox3 enhancer[a]	GR	This study
	D84-1	Hox3 enhancer[a]	GR	This study
DP7	A54-1	Bidirectional promoter[a]	E-box, NF-1, E-box, CAT-box, TATA-box	This study
	B54-1	Bidirectional promoter	E-box, NF-1, E-box, CAT-box, TATA-box	Dinger et al. (2008)
DP8	A43-12	Hox3 enhancer[a]	USF, Homeobox	This study
	B43-5	Hox3 enhancer[a]	USF, Homeobox	This study

[a]Putative function

Figure 2.2: Multiple alignments of three TP CNS sequences

(A) - (C) are results of multiple alignments of paralogous conserved regions derived from each TP CNS. Aligned sequences are ancestral sequences estimated from each CNS using PAML4 program (Yang 2007). Alignments are generated by using CLUSTALW (Thompson et al. 1994). The putative TFBS are highlighted by orange.

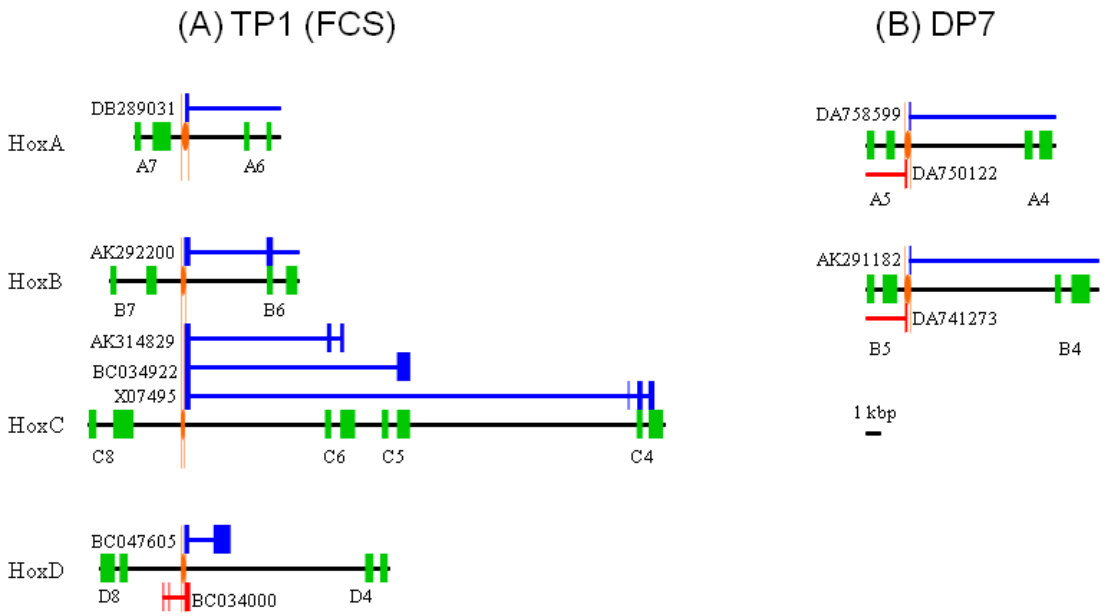


Figure 2.3: The scheme of paralogous conserved bidirectional promoters

We mapped the paralogous CNSs on bidirectional transcript start sites which code alternative splicing RNAs of Hox genes and antisense RNAs. Paralogous CNSs are (A) TP1 and (B) DP7. The blue and red lines are sense RNAs and antisense RNAs, respectively. DDBJ/EMBL/GenBank accession numbers of these RNA sequences are also shown.

2.3.2 Paralogous CNSs among Hox clusters

I found 28 paralogous conserved elements in total (8, 6, 6 and 8 for Hox A, B, C and D clusters, respectively). Three quartets of CNSs are conserved among all four Hox clusters, and I named them TP (tetra-paralogous), as shown in **Figure 2.1**. I carried out the phylogenetic footprinting analysis to infer significantly conserved motifs among these three TPs. I found the highly conserved region in each CNS, and these overlap with paralogous conserved regions (see **Figure A 2.2**). Multiple sequence alignments of three TP CNSs are shown in **Figure 2.2**. It should be noted that these sequences are reconstructed ancestral ones. TP2 and TP3 contain retinoic acid response elements (RAREs). Intergenic regions of upstream or downstream of Hox4 genes are abundant with functional RAREs (Mainguy et al. 2003). Despite this, RAREs located downstream of HoxC4 has not been reported before. I found a new evolutionarily highly conserved sequence containing RARE in this region. These motifs might maintain gene expression pattern of clusters cooperatively.

The remaining TP1 was discovered by Kim et al. (2000), and they named it four cluster sequence (FCS). Though I found conserved motifs in the FCS (**Figure 2.2**), these motifs have no experimental corroboration. Then I mapped transcripts within Hox

clusters. As a result, 136 CNSs overlap with transcribed regions (see **Table A2.1**). FCS corresponds to the bidirectional transcript start sites (TSS) which encode alternative spliced RNAs of Hox genes and antisense non-coding RNAs (**Figure 2.3A**). These CNSs might play important roles in the colinear expression pattern of the Hox cluster. Another paralogous CNS between Hox5 and Hox4 overlapped the region of TSS and alternative exons (**Figure 2.3B**), suggesting that CNSs function as *cis* and *trans* regulatory elements.

Eight pairs of CNSs are conserved between two paralogous Hox clusters, and I named them DP (di-paralogous), as shown in **Figure 2.1**. Results of phylogenetic footprinting analysis and pairwise sequence alignment are shown in **Figure A2.2** and **Figures A2.3**, respectively. The DP6 CNS is not located at syntenic region and the conservation is poor. Other DP CNSs are located at the syntenic region of each cluster and include functional elements (**Table 2.2**). DP1 and DP2 which are located at the upstream of Evx1 and Evx2 have *cis*-regulatory functions (Lehoczky et al. 2004). The region called “distal limb enhancer” in the HoxD cluster is essential for the posterior HoxD gene expression of appendicle (Spitz et al. 2001). The DP4 pair corresponds to microRNAs mir-196b and mir-196a-2. They belong to the mir-196 family. This family is composed of three members, which are mapped between Hox10 and Hox9 of HoxA,

HoxB and HoxC clusters (Yekta et al. 2004). However, another member, mir-196a-1, was difficult to detect because of poor conservation. I thus defined only two microRNA members as CNS.

2.3.3 Comparison between vertebrate CNSs and amphioxus CNSs within Hox clusters

Phylogenetic footprinting can be used to detect significantly conserved sequences between vertebrates and the amphioxus Hox cluster. Because the conservation of non-coding region between amphioxus and vertebrates is poor, Pascual-Anaya et al. (2008) defined CNS in the case of human-amphioxus comparison as approximately 60 % identity and 50 bp length region. They reported 75 amphiCNSs. However, this might include CNSs which are not conserved among all vertebrates, but conserved only between human and amphioxus.

To remove these CNSs and to identify CNSs conserved among all vertebrates, I collected multiple orthologous vertebrate sequences and carried out phylogenetic footprinting analysis. I then identified the highly conserved “core region” of each vertebrate CNS. By comparing amphiCNSs with the vertebrate CNSs, I found that only 16 out of 75 amphiCNSs overlap with the vertebrate CNSs. Eight of them show deep

conservation; they are conserved among all vertebrates used in this study (see vertebrate CNSs information shown in **Table A2.2**). Two of eight amphiCNSs were aligned with the “core region” of the vertebrate CNSs; they are conserved among all chordates used in this study (**Figure 2.4A**). These are located at the HoxA and HoxB anterior regions, and supported a previous observation that the posterior region is more divergent than the anterior region (Ferrier et al. 2000).

The remaining six amphiCNSs did not correspond to the “core region” of the vertebrate CNSs (**Figure 2.4B** and **Figure A2.4**). Interestingly, the “core region” of the vertebrate CNSs is often adjacent to the amphioxus-human conserved regions. At last, only 2 out of 75 amphiCNSs are significantly highly conserved among chordates.

2.4 Discussion

I defined 208 CNSs within the vertebrate Hox clusters. To infer the depth of sequence conservation, I investigated the existence of orthologous CNSs from vertebrate species. The depth of conservation is different with each cluster. The HoxC cluster shows the shallowest conservation. Despite this result, the HoxC cluster retains some paralogous CNSs. Shallow conservation of the HoxC cluster could be an artifact. Because the intergenic sequence data of HoxC cluster is the poorest, I cannot detect the

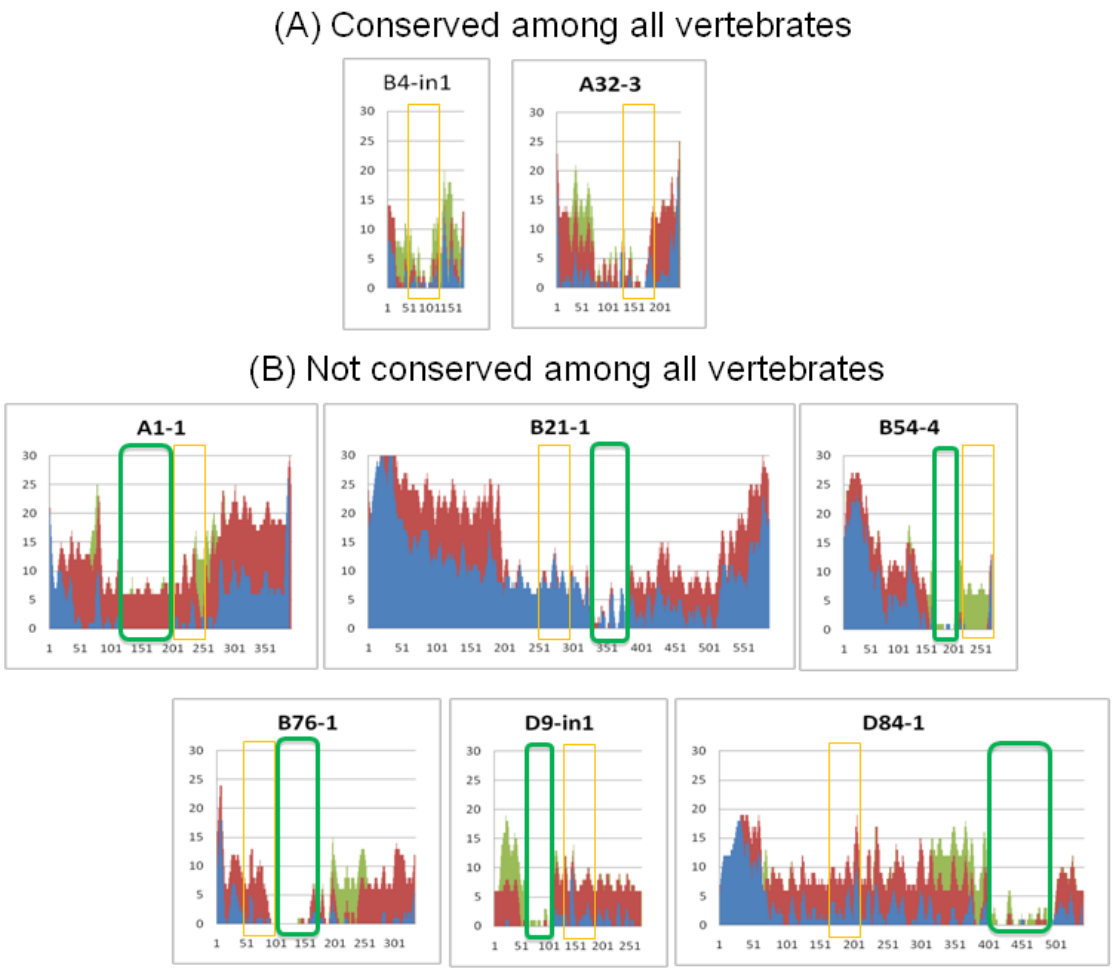


Figure 2.4: The phylogenetic footprinting analysis within chordates

We compared the vertebrate CNSs with amphioxus-human CNSs (amphiCNSs). The results of phylogenetic footprinting are described. Each orange box corresponds to amphiCNS. (A) CNSs conserved among all vertebrates. (B) CNSs not conserved among all vertebrates. Each green box represents highly conserved region among vertebrates identified by phylogenetic footprinting. Each axis and color is the same as Figure A2.2.

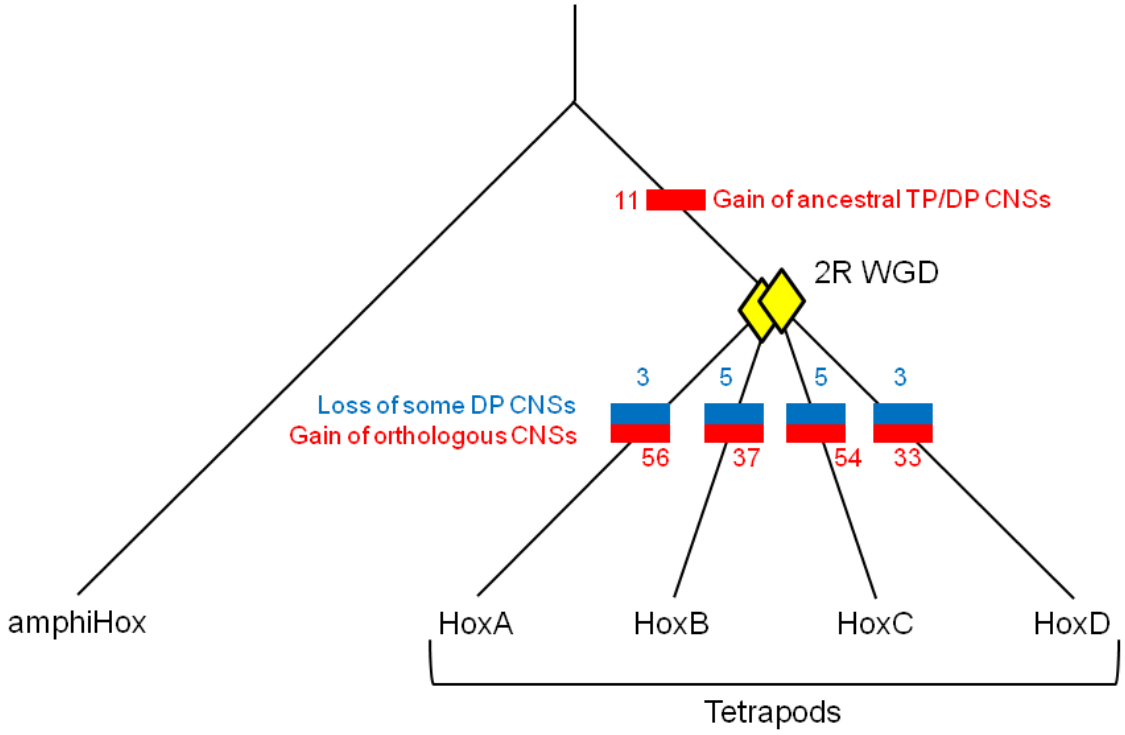


Figure 2.5: The loss and gain of Hox CNSs during the chordate evolution

The numbers of gain and loss of CNSs, shown in red and blue colors, respectively, are apportioned to the known Hox gene tree.

intergenic conservation from several species accurately. If the sequence data of Hox clusters are complete, the abundance of CNS of each Hox cluster may not be so different.

The number of the CNSs located at anterior region is higher than that of CNSs located at posterior region. The divergence of posterior paralogous Hox genes are more rapid compared with other paralogous Hox genes, called “posterior flexibility” (Ferrier et al. 2000). For example, because posterior genes of the HoxD cluster are regulated not only by each gene regulatory element but also by the global control regulatory element located 240 kb upstream of the cluster (Spitz et al. 2003), the intergenic region of the posterior HoxD cluster might have poor conservation. The posterior HoxA genes show similar expression pattern with the posterior HoxD genes. Therefore, this tendency applies to the HoxA cluster. The HoxA cluster also have global control enhancers located at upstream of the cluster (Lehoczky and Innis 2008).

The DP CNSs have many putative TFBSs (**Table 2.2** and **Figure A2.3**). The homeobox binding motifs are especially abundant. This suggests that DP CNSs are important for the auto regulatory mechanism of the four vertebrate Hox clusters. Each Hox protein may bind to *cis*-regulatory regions of other Hox genes and controls the expression patterns. E-box is the motif related to the HLH (helix-loop-helix)

transcription factor. HLH and homeobox proteins mainly regulate the expression pattern of Hox genes. The DP7 CNSs bear the conserved TATA-box. This suggests that the DP7 CNSs have promoter function as I described.

I identified three paralogous regions conserved among the four Hox clusters. One of them, FCS, was previously reported (Kim et al. 2000). Surprisingly, many RNAs are transcribed in this area. Different directional transcripts are started in the HoxA cluster. FCS of the HoxB cluster corresponds with TSS of the HoxB6 gene. In the HoxC cluster, FCS is the TSS of HoxC6, HoxC5 and HoxC4 coding transcripts. In the HoxD cluster, FCS might control different directional transcripts. Not only FCS but also other paralogous CNSs (DP7) between HoxA and HoxB clusters overlap with TSS and alternative exons (**Figure 2.3**). Experimental approach revealed long non-coding antisense RNA started from this HoxB cluster region (Dinger et al. 2008). Because RNA data are insufficient to detect all cluster transcripts, some of these transcripts are partial and were found only in human and/or mouse. It is probable that these paralogous CNSs play important roles in alternative transcription in other tetrapod species.

The other two TP CNSs (TP2 and TP3) include the RARE (Mainguy et al. 2003). Their functions are experimentally confirmed (Morrison et al. 1997). Retinoids are thought to exert their activities at the transcriptional level, acting as ligands to

activate nuclear receptors. These nuclear receptors recognize DNA sequences closely related to 5'-(A/G)G(G/T)TCA-3'. Previous studies suggested that retinoic acids contribute to the expressions of Hox genes (Dubrulle and Pourquié 2004). TP2 and TP3 have type11 and type3 RAREs, respectively. A conserved sequence, TP3, downstream of HoxC4 gene was newly detected in this study. This sequence is located more than 20 kb away of the HoxC4 gene and corresponds to type3 RAREs. Amphioxus also has RARE in this intergenic region (Wada et al. 2006). However, I could not detect this element in this study. Only one motif conservation is difficult to detect by using this method. Other motifs of those paralogous CNSs might function as *cis*-regulatory element that cooperates with RAREs.

It is possible that these TP CNSs are key components of cluster organization. The motifs within them might have already existed in the ancestor of vertebrates who had only one Hox cluster. Because other motifs are not conserved within the orthologous region of invertebrates but conserved in the paralogous region of vertebrates, they were acquired after the emergence of vertebrates.

Pascual-Anaya J et al. (2008) reported 75 amphiCNSs which might include CNSs that are not conserved among all vertebrates but conserved only between human and amphioxus. To remove these CNSs and to increase statistical significance, I

compared multiple orthologous vertebrate sequences. I found that two amphiCNSs are overlapped and conserved in vertebrate CNSs. Ancestral DNA sequences of these CNSs have probably been under strong selective constraint throughout the chordate evolution, though their conservation is detected in only one Hox cluster. Other amphiCNSs might not be conserved among all vertebrates. However, we should deal with this problem carefully, for only two amphioxus genomes were used to detect CNSs conserved among chordates. More information of the Hox cluster from non-vertebrate chordate genome is necessary to obtain the complete picture of chordate CNSs.

The loss and gain of Hox CNSs are shown in the **Figure 2.5**. After the 2R WGD, the massive gains of CNSs were occurred. In contrast, the conservation of non-coding regions in the invertebrate genomes is low. This difference on the Hox clusters may be related with the evolution of various unique features of vertebrates. When vertebrates acquired the more complex morphogenesis, the Hox clusters may become more conservative. To solve why these highly conserved CNS were appeared, we have to consider the relationship between the non-coding functions and evolutionary conservations.

In summary, I efficiently detected orthologous CNSs of vertebrates. I identified three paralogous CNSs, and one of them bears a newly detected RARE motif. These

CNSs are conserved among all paralogous Hox clusters, and might contribute to Hox cluster organization and gene expression patterns.

CHAPTER 3

Phylogenetic Network Analysis of Vertebrate Hox Genes

3.1 Introduction

The Hox genes orchestrate animal body plans in deuterostomes. The Hox genes are originally identified from the mutant of *Drosophila melanogaster* (Lewis 1979; Kaufman et al. 1990), called homeotic transformation. These Hox genes are organized as a cluster on same chromosome in many animal phyla, suggesting that they are generated by series of tandem duplications occurred before their common ancestor of animals. Hox genes of each cluster are expressed along the anterior-posterior body axis in the same order as lining up on the chromosome, called “colinearity” (Pearson et al. 2005; Lemons et al. 2006). However, recent genome sequencing of deuterostomes revealed that Hox genes are not always colinear. They are scattered in different chromosomes (Seo et al. 2004) or translocated (Cameron et al. 2006) in specific deuterostomes. In vertebrate genomes, the two-round whole genome duplications (2R WGD) generated paralogous four Hox clusters (Ruddle et al. 1994). They consist of approximately 40 members that are physically linked on chromosomes and made four clusters. Each paralogous gene of the Hox clusters shows the colinear expression pattern.

There might be shared gene expression regulatory mechanisms among paralogous Hox clusters. However, Hox14 genes are the exception of this colinearity (Kuraku et al. 2008).

The posterior Hox genes are rapidly evolving, and this phenomenon has been termed "posterior flexibility" (Ferrier et al. 2000), so that it is difficult to assign clear ortholog/paralog relationships among deuterostomes. Especially, the orthologies between vertebrate posterior Hox genes (Hox9-Hox14) and amphioxus posterior Hox genes (Hox9-Hox15) are ambiguous. The cephalochordate amphioxus possesses a single Hox cluster, which is regarded as the ancestral state of vertebrate Hox clusters (Amemiya et al. 2008). The clear assignment of 1-to-1 orthologies between amphioxus and vertebrate posterior Hox genes cannot be established without further data (Ferrier 2004; Amemiya et al. 2008; Hueber et al. 2010). For instance, the non-orthology between the amphioxus Hox14 gene and the vertebrate Hox14 genes has been supported by phylogenetic analysis (Kuraku et al. 2008; Feiner et al. 2011) as well as a non-tree-based study (Thomas-Chollier et al. 2010). The identical name of the amphioxus and vertebrate Hox genes is simply derived from the same relative location in the cluster, but does not reflect true orthology. Interestingly, orthology between amphioxus Hox15 and vertebrate Hox13 was previously suggested (Holland et al. 2008;

Thomas-Chollier et al. 2010), despite their non-syntenic location in the cluster. However, the support for this grouping is poor.

Duplication history of Hox cluster is also controversial. Under the assumption of the 2R WGD, phylogenies of Hox gene should show a symmetrical topology, such as ((A,B),(C,D)). However, Lynch et al. (2009) reconstructed phylogeny of paralogous Hox genes that showed (B,(A,(C,D))) topology. This result is contradictory to some of other reports. Kappen et al. (1993) found a single best tree with the topology ((A,B),(C,D)). However, the next best tree with the topology (B,(A,(C,D))) was only a single step away. The (B,(A,(C,D))) topology was also proposed by Zhang et al. (1996) using distance methods, but they could not reject an ((A,B),(C,D)) because of low internal branch support. Recently, it was shown that elephant shark Hox genes support the ((A,B),(C,D)) topology with high statistical significance (Ravi et al. 2009). These results suggest that the phylogeny of Hox clusters is not yet solved. The phylogenetic network study of each paralogous Hox gene family may shed a light on this conundrum.

In this study, I carried out an exhaustive phylogenetic analysis of deuterostome Hox genes. First, to identify outgroup genes of each vertebrate Hox paralog group, I inferred the correct ortholog/paralog relationships between amphioxus and vertebrate posterior Hox genes. Second, the duplication histories of vertebrate Hox genes were

inferred by not only phylogenetic tree, but also phylogenetic network with and without outgroups. My analysis demonstrated that the ((A,B),(C,D)) topology is the most suitable explanation of Hox cluster duplications.

3.2 Materials and Methods

The homeodomain sequences of deuterostome posterior Hox genes were manually downloaded from the GenBank database. Hox amino acid sequences for human (*Homo sapiens*), coelacanth (*Latimeria menadoensis*), horn shark (*Heterodontus francisci*), elephant shark (*Callorhinchus milii*), ascidian (*Ciona intestinalis*), larvacean (*Oikopleura dioica*), two amphioxus (*Branchiostoma floridae*; *Branchiostoma lanceolatum*), two acorn worms (*Ptychodera flava*; *Saccoglossus kowalevskii*), and sea urchin (*Strongylocentrotus purpuratus*) were used to infer the orthologies between vertebrate and other chordates. Vertebrate amino acid sequences for Dlx, Collagen (Col), Hox and ErbB were downloaded from GenBank or identified from BLAST searches of amino acid databases. The sequences of all paralog members for each Hox cluster were aligned with and without outgroups. Amino acid sequences for all genes were aligned by using CLUSTALW (Thompson et al. 1994) and adjusted by visual inspection. Regions with large gaps, ambiguous alignment or repetitive sequences were removed

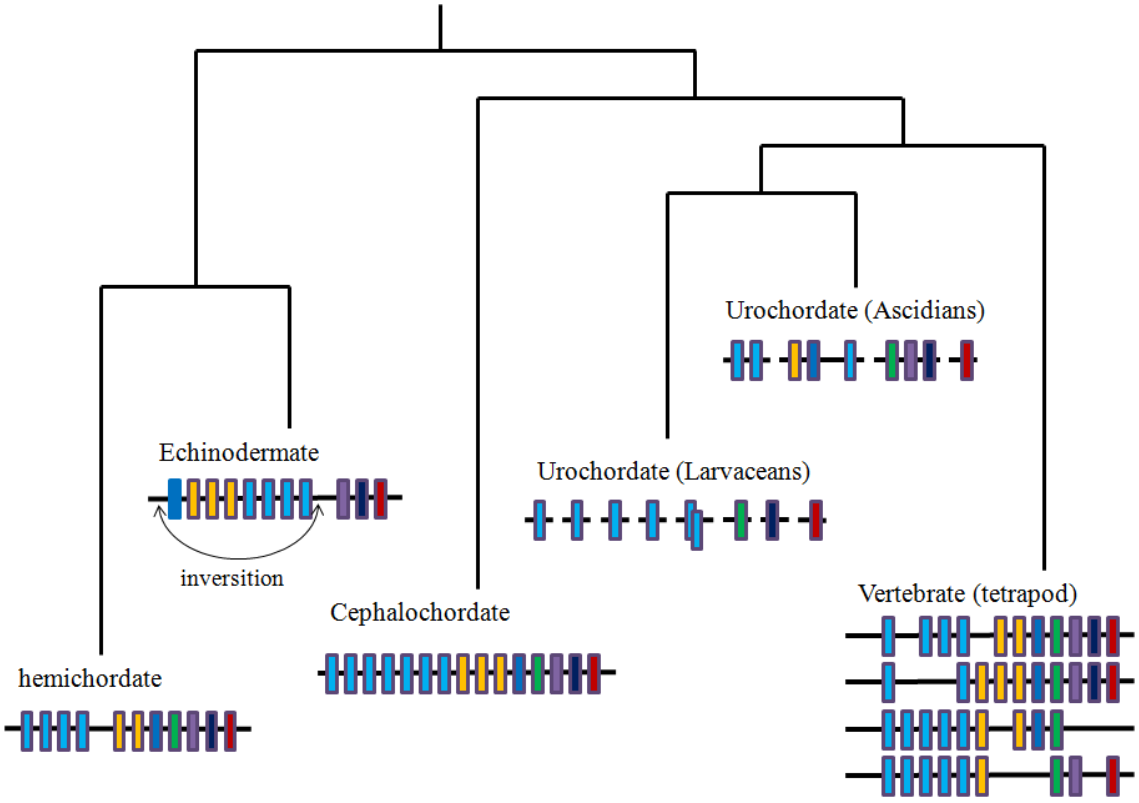


Figure 3.1: The evolution of deuterostome Hox clusters

For each taxon, Hox clusters are illustrated. In vertebrate, HoxA, HoxB, HoxC, and HoxD are shown from top to bottom. Genes are colored to differentiate between Hox family members, and genes that are orthologous between clusters and species are labeled in the same color. In some cases, orthologous relationships are not clear.

Table 3.1: Species used in this study

	phyla	Species name	Common name	Reference
Hs	Vertebrate	*Homo sapiens*	Human	GenBank
Lm	Vertebrate	*Latimeria menadoensis*	Coelacanth	Amemiya et al. 2010
Hf	Vertebrate	*Heterodontus francisci*	Horn shark	Kim et al. 2000
Ci	Chordate	*Ciona intestinalis*	Ascidians	Spagnuolo et al. 2003
Od	Chordate	*Oikopleura dioica*	Larvaceans	Seo et al. 2004
Bf	Chordate	*Branchiostoma floridae*	Amphioxus	Amemiya et al. 2008
Pf	Hemichordate	*Ptychodera flava*	Acorn worm	Peterson 2004
Sk	Hemichordate	*Saccoglossus kowalevskii*	Acorn worm	Aronowicz et al. 2006
Sp	Echinodermata	*Strongylocentrotus purpuratus*	Sea urchin	Cameron et al. 2006

from all genes. Phylogenetic trees were reconstructed using neighbor-joining (NJ) JTT distance, and maximum likelihood (ML) algorithms implemented in the MEGA5 (Tamura et al. 2011) package of programs. Branch support was assessed with 500 and 100 bootstrap resamplings for NJ distance and ML, respectively. A phylogenetic network based on a distance matrix was reconstructed by using the neighbor-net method (Bryant et al. 2004; Huson et al. 2006),

3.3 Results

3.3.1 Ortholog/paralog relation of posterior Hox genes

Because the phylogenetic relationships of deuterostome posterior Hox genes, especially between amphioxus and vertebrates, are still unclear, these relations were inferred by phylogenetic networks. The already known homeodomain sequences of deuterostome posterior Hox genes were collected from the database (**Figure 3.1** and **Table 3.1**) and reconstructed the possible evolutionary history. Although the statistical significance is very low because of the short alignment length (**Figure A3.1**), amphioxus Hox9-11, and Hox15 are clustered with vertebrate Hox9, and Hox13 paralogous groups, respectively (**Figure 3.2**). The vertebrate Hox12 paralog group is not clustered with any other amphioxus Hox genes. This result may imply Hox12

paralog group was generated by vertebrate specific tandem gene duplication or amphioxus ortholog were lost along their lineage.

From the results of analysis, I proposed the possible orthologies of posterior Hox genes among deuterostomates (**Figure 3.3**). From the results, it was suggested that amphioxus Hox9-11 are generated by amphioxus specific tandem duplications. Because vertebrate Hox10-12, and Hox14-15 genes have no counterpart of amphioxus Hox genes, they are lost at amphioxus lineage. Because the posterior Hox genes rapidly evolved (Ferrier et al. 2000), it seems to be usual that tandem duplications happened in this region. Previous studies suggested that vertebrate Hox14 paralog group, that is recently identified from early divergent vertebrates (Ravi et al. 2009; Amemiya et al. 2010; Liang et al. 2011), is born from the tandem duplication (Feiner et al. 2011). Amphioxus Hox15 has a high similarity with vertebrate Hox13 paralog group (Holland et al. 2008; Thomas-Chollier et al., 2010). These report strongly supports our hypothesis. To infer the possible duplication history of vertebrate Hox clusters, these orthologies between vertebrate and amphioxus were used.

3.3.2 Phylogenetic analysis of vertebrate Hox genes

The duplication histories of vertebrate Hox clusters were inferred by

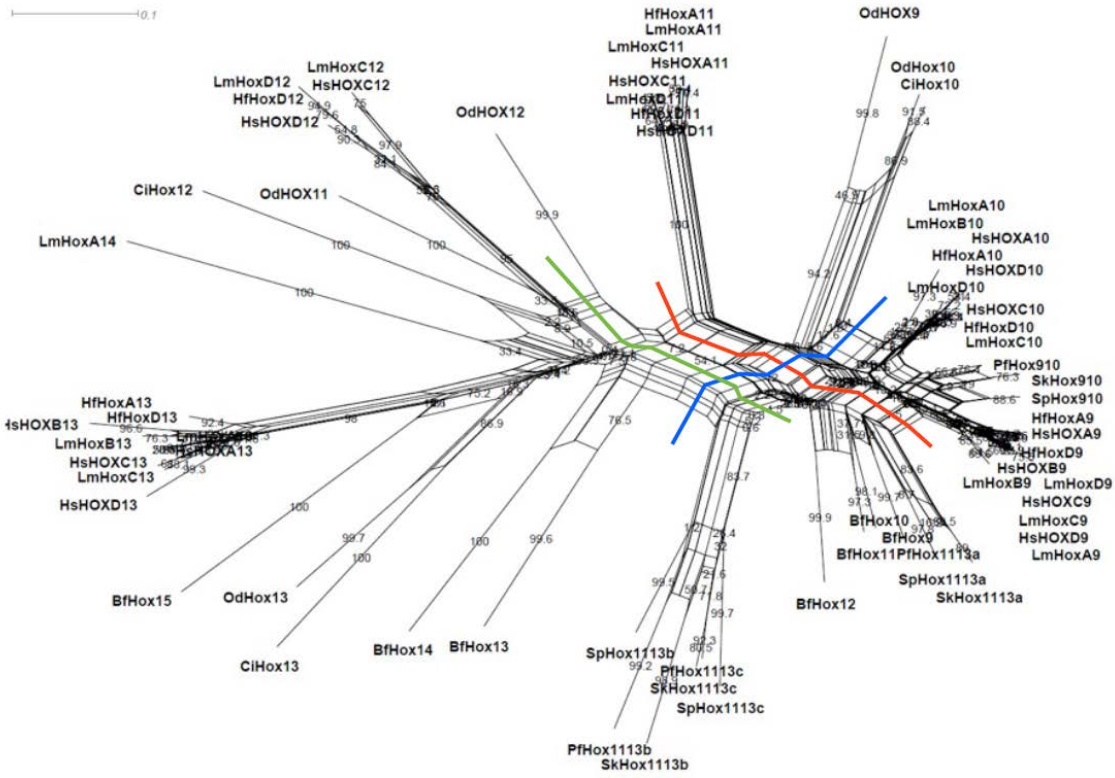


Figure 3.2: The phylogenetic network of deuterostome posterior Hox genes

The phylogenetic network of deuterostome posterior Hox genes are reconstructed by the neighbor net. The homeodomain sequences of vertebrate Hox9-13, amphioxus Hox9-15 and other deuterostome posterior Hox genes are used for alignments. This network is divided into 3 splits, shown in red, blue, and green lines.

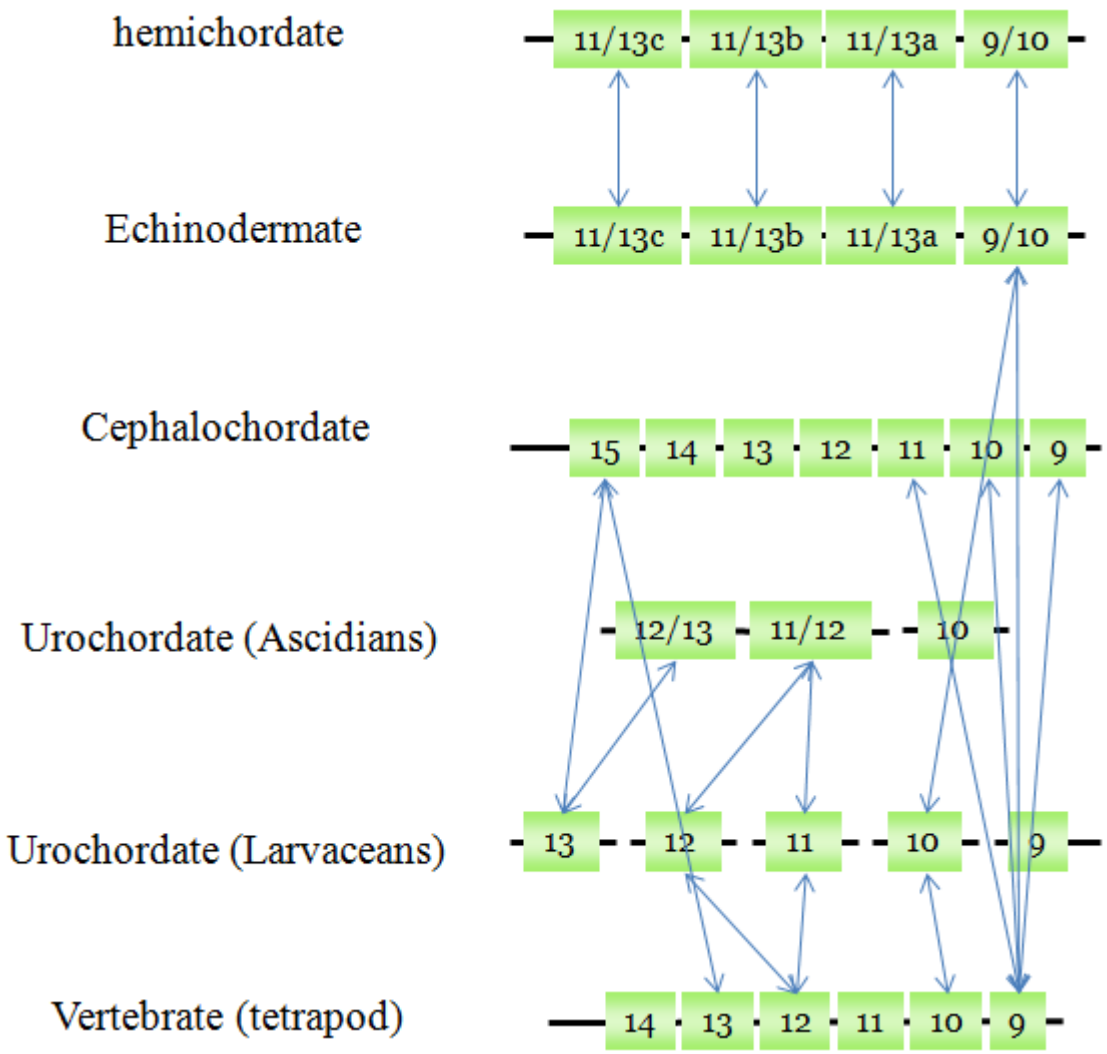


Figure 3.3: Possible orthology of posterior Hox genes

The possible orthologies among deuterostome posterior Hox cluster are estimated from the phylogenetic analysis. The orthologous relationship each other are shown in arrows.

amphioxus and vertebrate Hox genes. Syntenic paralogs linked to Hox cluster were also used for the analysis. Vertebrate Hox clusters are located at highly conserved syntenic regions derived from the 2R WGD (Larhammar et al. 2002). Dlx, Collagen, and ErbB paralogs were reconstructed the phylogenetic history among these syntenic genes. Because they have more than two paralogs and are well-researched previously (Lynch et al. 2009), I decided to use Dlx, Collagen, and ErbB genes for analysis. The possible duplication history of Hox and their syntenic genes are ((A,B),(C,D)), ((A,C),(B,D)) or ((A,D),(B,C)) under the assumption of the 2R WGD. To determine the most suitable topology, NJ and ML phylogenetic trees of each paralog group were reconstructed with and without amphioxus, as an outgroup species. When Hox genes were aligned with outgroup sequences, the aligned regions were short. The orthologies of posterior Hox genes are ambiguous at this time. Therefore, I reconstructed not only rooted trees but also unrooted trees. In the results, majority of paralog groups were shown asymmetrical topologies for both NJ and ML trees with amphioxus (**Table 3.3** and **Table 3.4**). In contrast, all phylogenetic trees without amphioxus showed ((A,B),(C,D)) topologies except for the Hox4 paralog group. Next, concatenated analysis of paralog groups keeping four members was done. Because elephant shark genome retained the largest number of four paralogs, elephant shark Hox genes with and without human Dlx,

Table 3.2: Topology of each paralogous Hox gene NJ tree

Group	Topology without amphioxus	Topology with amphioxus (X)
Hox1	(A,B):86%,(C,D)	{((C,D):85%,B):44%,A),X}
Hox2	-	{((A,D):69%,B),X}
Hox3	(A,B):100%,(C,D)	{(((A,B):95%,D):58%,C),X}
Hox4	(A,C):49%,(B,D)	{(((A,D):40%,C):51%,B),X}
Hox5	(A,B):97%,(C,D)	{(((A,B):88%,C):51%,D),X}
Hox6	-	{((A,B):99%,C),X}
Hox7	-	-
Hox8	-	{((C,D):65%,B),X}
Hox9	(A,B):98%,(C,D)	{(((A,B):40%,D)23%,C),X}
Hox10	(A,B):97%,(C,D)	{((A,B):40%,(C,D):87%),X}
Hox11	-	{((A,C):27%,D),X}
Hox12	-	-
Hox13	(A,B):99%,(C,D)	{(((C,D):91%,A):24%,B),X}
Collagen	(A,D):95%,(B,C)	{(((A,D):85%,B):77%,C),X}
ErbB	(A,B):100%,(C,D)	{((A,B):67%,(C,D):72%),X}
Dlx6/7/1	-	{((A,D):89%,B),X}
Dlx5/3/2	-	{((A,B):52%,D),X}
Hox genes	(A,B):100%,(C,D)	{((A,B):100%,(C,D):57%),X}
Hox/Col/ErbB genes	(A,B):100%,(C,D)	{(((A,B):100%,D):88%,C),X}

Table 3.3: Topology of each paralogous Hox gene ML tree

Group	Topology without amphioxus	Topology with amphioxus (X)
Hox1	(A,B):80%,(C,D)	{(((C,D):82%,B):49%,A),X}
Hox2	-	{((A,D):70%,B),X}
Hox3	(A,B):99%,(C,D)	{(((A,B):53%,(C,D):70%),X}
Hox4	(A,C):49%,(B,D)	{(((B,D):2%,A):47%,C),X}
Hox5	(A,B):84%,(C,D)	{(((A,B):51%,C):55%,D),X}
Hox6	-	{((A,B):94%,C),X}
Hox7	-	-
Hox8	-	{((C,D):46%,B),X}
Hox9	(A,B):98%,(C,D)	{(((C,D):46%,A)34%,B),X}
Hox10	(A,B):100%,(C,D)	{((A,B):69%,(C,D):93%),X}
Hox11	-	{((A,C):19%,D),X}
Hox12	-	-
Hox13	(A,B):100%,(C,D)	{((A,B):30%,(C,D):87%),X}
Collagen	(A,D):89%,(B,C)	{(((A,D):74%,B):76%,C),X}
ErbB	(A,B):100%,(C,D)	{((A,B):48%,(C,D):93%),X}
Dlx6/7/1	-	{((A,D):74%,B),X}
Dlx5/3/2	-	{((A,D):57%,B),X}
Hox genes	(A,B):100%,(C,D)	{(((A,B):100%,D):46%,C),X}
Hox/Col/ErbB genes	(A,B):100%,(C,D)	{(((A,B):100%,D):86%,C),X}

Table 3.4: Topology of each paralogous Hox gene tree based on networks

Group	Topology without amphioxus	Topology with amphioxus (X)
Hox1	A,B,C,D	{((B,C,D):25%,A),X}
Hox2	-	{((A,D):9%,B),X}
Hox3	(A,B):97%	{((A,B):80%,(C,D):92%),X}
Hox4	A,B,C,D	{((A,B,D),C),X}
Hox5	(A,B):73%	{((A,B):67%,D):43%},C,X}
Hox6	-	{((A,B):98%,C),X}
Hox8	-	{(B,C,D),X}
Hox9	(A,B):98%	{((A,B):81%,(C,D):68%,X}
Hox10	(A,B):96%	{((A,B): 89%,(C,D):30%),X}
Hox11	-	{((A,C):78%,D),X}
Hox13	(A,B):15%,	{(C,D):66%,A,B,X}
Collagen	A,B,C,D	{(A,B,D):83%,C),X}
ErbB	(A,B):100%	{((A,B):98%,D):95%,C),X}
Dlx6/7/1	-	{((A,B):11%,D),X}
Dlx5/3/2	-	{((A,B):17%,D),X}
Hox genes	(A,B):100%	{((A,B):100%,(C,D):97%),X}
Hox genes	(A,C):93%	{((A,B):100%,D):92%},C,X}
Hox/Col/ErbB genes	(A,B):100%	{((A,B):100%,D):100%},C,X}
Hox/Col/ErbB genes	~	{((A,B):100%,(C,D):98%),X}

Collagen, and ErbB genes are concatenated and aligned. The phylogenetic trees of concatenated alignments also showed ((A,B),(C,D)) topology. These results suggest that ((A,B)(C,D)) topology is the most plausible duplication history of the vertebrate Hox clusters. The parsimonious gene loss pattern was estimated under the ((A,B),(C,D)) topology (**Figure 3.4**). The gene loss pattern of each Hox cluster supports this ((A,B),(C,D)) topology. HoxA and HoxB clusters share the loss of Hox12 genes, while HoxC and HoxD clusters share the loss of Hox7 genes. These losses probably occurred after the first WGD event and before the second WGD event.

3.4 Discussion

The Hox clusters have played a central role in the genome duplication story, largely because they conform to the 1:4 expectation of the 2R hypothesis and are tightly linked to each other and several non-Hox genes. However, numerous studies of the duplications of the Hox clusters and linked genes have failed to reach a consensus on the mechanisms, number and order of duplications. Many of these studies were hampered by limited sequence data and poor taxon sampling, lack of appropriate out-group data or computational limitations that prevented the use of computationally intensive methods of phylogenetic inference such as maximum likelihood. Given these

difficulties it is not surprising that nearly every study found support for a different duplication order.

In my results, the best duplication history of the vertebrate Hox cluster are ((A,B),(C,D)). However, many asymmetric phylogenetic tree and different symmetric tree, such as ((A,C),(B,D)), were reconstructed. There are two possible reasons to cause these unequal evolutionary histories. First, the tree topologies are violated by homogenization of sequences, such as gene conversion and crossover. Gene conversions frequently occur in yeast genome (Kellis et al. 2004). It also occur in vertebrate species (Ezawa et al. 2010). Lynch et al. (2009) pointed out the possibility of crossover between different chromosomes after the 2R WGD. These homogenization events may produce improper tree topologies. Second, vertebrate Hox clusters may evolve with different evolutionary rate after the 2R WGD. Recently sequenced shark and skate genomes have no HoxC cluster in their genome (Oulion et al. 2010; King et al. 2011). This implies that the Hox clusters did not share the ancestral functions equally. The HoxC cluster might evolve rapidly in specific lineages, because of their less important functions. Different evolution rate make tree topology inaccurate. These facts suggest that the evolution of vertebrate Hox clusters after the 2R WGD is not straight and even. The further studies are required for more clear understanding of the Hox cluster duplication histories.

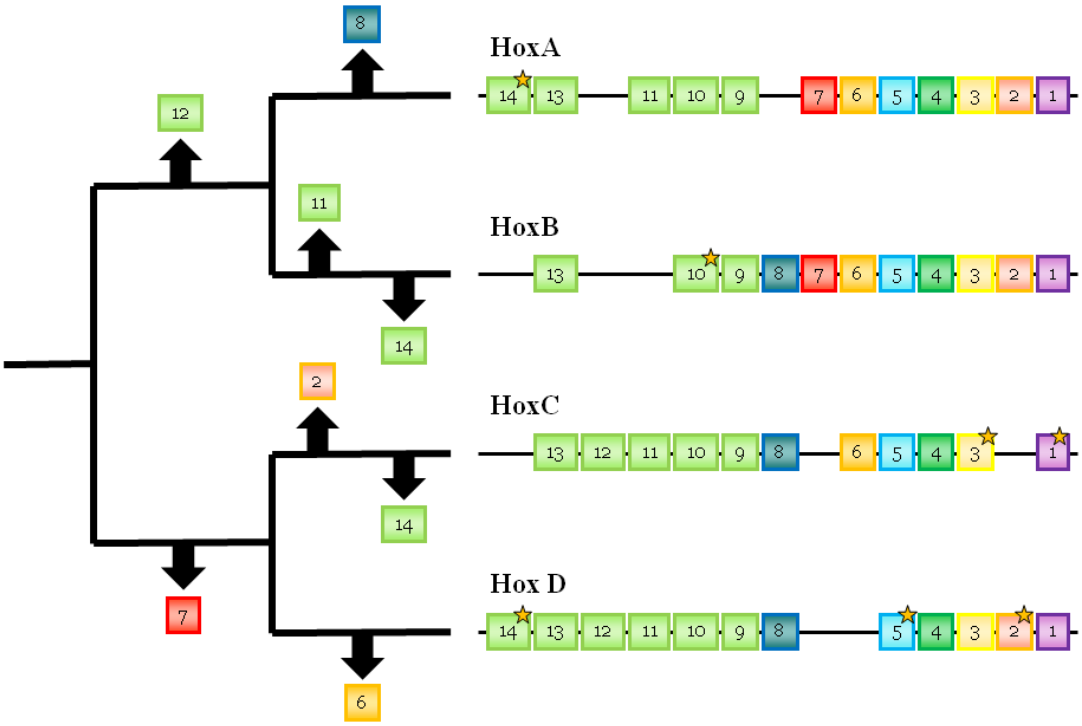


Figure 3.4: Reconstruction of the Hox cluster duplication history

The gene inventory of the Hox clusters in the hypothetical ancestors of major evolutionary lineages are inferred based on parsimony principles. The losses of Hox genes are indicated in boxes along branches. Colored squares indicate Hox genes that have been conserved. The squares with a small star are Hox genes that are lost in specific vertebrate species. A currently accepted phylogenetic tree is shown on the left.

The assumption of ((A,B),(C,D)) topology are supported by gene loss pattern of each Hox cluster. Because HoxA+B clusters, HoxC+D clusters share the loss of Hox12, Hox7, respectively, these losses occurred after the first WGD event and before the second WGD event (**Figure 3.4**). Other combination of Hox clusters never share the gene loss patterns, except Hox14 paralog group. Hox14 paralog group is only retained basal vertebrates and shows non-colinear expression pattern, so that this group is not a typical Hox gene. From the view point of the gene loss pattern within Hox clusters, the ((A,B),(C,D)) topology is reinforced.

CHAPTER 4

Paralogous Conserved Non-coding Sequences in Vertebrates Derived from the Ancient Whole Genome Duplications

4.1 Introduction

Regulation of gene expression in a spatial and temporal manner is crucial during vertebrate development. Such complex transcriptional regulations are thought to be mediated by the coordinated binding of transcription factors. They, known as *cis*-regulatory elements, allow the integration of multiple signals to regulate the expression of specific genes. These elements may not act on the physically closest gene but can act across intervening genes (Spitz et al. 2003). It was shown recently that certain genomic regions contain arrays of conserved non-coding sequences (CNS), which are candidate of *cis*-regulatory elements, clustered around developmental regulatory genes (Bejerano et al. 2004; Woolfe et al. 2005; McEwen et al. 2006; Hufton et al. 2009; Lee et al. 2011). Some of them already tested have been shown to act as enhancers in transgenic reporter assays (Pennacchio et al. 2006; McEwen et al. 2006; Hufton et al. 2009; Lee et al. 2011). These genomic regions also show conserved synteny that is prominent feature of vertebrate genomes. Moreover, these regions are

conserved not only orthologous but also paralogous.

These paralogous synteny are derived from ancient genome duplications. Ohno (1970) proposed the two-round whole genome duplications (2R WGD) happened at the intersection of early vertebrate evolution, known as the 2R hypothesis. Now it is clear that 2R WGDs occurred early in vertebrate evolution from phylogenetic and syntenic analysis of vertebrates and invertebrates (Dehal and Boore 2005; Nakatani et al. 2007; Putnam et al. 2008). These conserved synteny blocks are bearing paralogous genes and CNSs (Kikuta et al. 2007). They are under the strong evolutionary constraint and might have played important roles in the vertebrate evolution. However, the CNSs generated by the 2R WGD and conserved among paralogous synteny blocks are still not completely documented. These sequences have a vertebrate-specific conservation and might be related to vertebrate morphological feature (e.g. jaws and brains). Thus, detecting paralogous CNSs and inferring their characteristic is important in understanding genome evolution after the 2R WGD.

The vertebrate Hox cluster is one of the most famous examples of highly conserved paralogous syntenic regions (Garcia-Fernández 2005). Vertebrates were shown to possess at least four Hox clusters, whose genes are intimately involved in axial patterning and a strict relationship exists between respective genes and their

expression limits in somitic and neural tissues. As a result of their intimate involvement in early development, the change of Hox gene expression often triggers a vertebrate morphological change (Cohn et al. 1999). The paralogous genes of the Hox clusters show the similar expression pattern, which suggests that there might be shared gene regulatory mechanisms among paralogous Hox clusters. In chapter 2, I carried out the search of CNSs within these clusters, not only among orthologous clusters but also among paralogous clusters and identified three paralogous CNSs conserved within all four Hox clusters of vertebrate species experienced no further genome duplication. This work was already published as Matsunami et al. (2010). These CNSs should contribute to Hox cluster organization and gene expression patterns.

I used a region-focused homology search to detect weak paralogous conservations in this chapter. The method of paralogous CNS identification is critical for the result. Previous studies were mainly based on MegaBLAST search (Zhang et al. 2000) of whole genome sequences to detect the paralogous CNSs (Bejerano et al. 2004; McEwen et al. 2006) on whole vertebrate species. This method is faster than conventional BLAST search (Altschul et al. 1997) and is effective to identify the paralogous CNSs, showing prominent high conservation, among vertebrate genomes. However, it is difficult to detect weak conservation of paralogous non-coding regions by

using this method. The orthologous conservation of non-coding region is statistically highly significant and easy to detect. By contrast, the paralogous conservation of non-coding region is usually weaker than orthologous conservation of non-coding regions. To overcome this problem, a region-focused BLAST searches of each synteny block are useful. This improved method allowed me to identify much weaker paralogous conservation derived from the 2R WGD.

In this study, I characterized paralogous synteny blocks derived from the 2R WGD by using vertebrate genome data and identified both orthologous and paralogous CNSs derived from the 2R WGDs. From these data, I found 194 new paralogous CNSs. These paralogous CNSs are frequently located at near the coding regions encoded transcription factors expressed in brain and neural system and have potential to control similar expression patterns of paralogs. These paralogous CNSs are shared among vertebrate lineages and might reflect the interplay between paralogous genes such as similar expression and dosage sensitivity.

4.2 Materials and Methods

4.2.1 Identification of conserved synteny blocks after the 2R WGD

Nakatani et al. (2007) reported 118 conserved vertebrate linkage regions within

the human genome through comparison with medaka fish genome sequences. In these conserved vertebrate linkage regions, 10,618 genes exist. Each of them shows 4, 3, 2 or 1 paralogous gene retention(s). These conserved vertebrate linkages were used as synteny block data. These gene information was downloaded from the Ensembl database using the Ensembl BioMart interface (http://www.ensembl.org/biomart /martview/). To identify paralogous synteny blocks, paralogous genes (4, 3 or 2 gene retention(s)) were used as markers of paralogous conservation. When paralogous genes show same combination of conserved vertebrate linkage groups, these pairs are regarded as paralogous synteny blocks. I then identified pairs of paralogous block sets (di-paralogous sets), trios of paralogous block sets (tri-paralogous sets) and quartets of paralogous block sets (tetra-paralogous sets) (**Table 4.1**). The location of paralogous synteny blocks are shown in **Figure 4.1**.

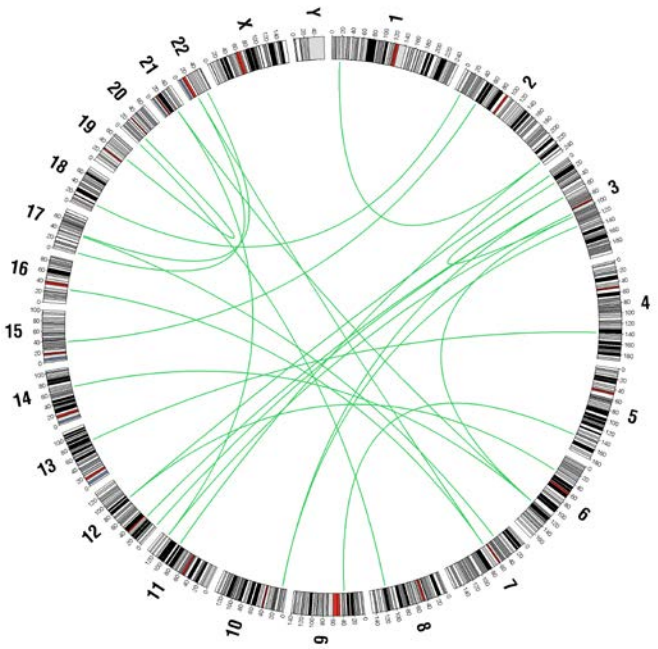
4.2.2 Identification of paralogous CNSs

I identified paralogous CNSs within vertebrate genomes by using the paralogous synteny block data. First, the human genome sequences (*Homo sapiens*; NCBI36) were downloaded from the Ensembl database (http://www.ensembl.org/) and divided into the conserved vertebrate linkage regions. Repeat and coding regions of

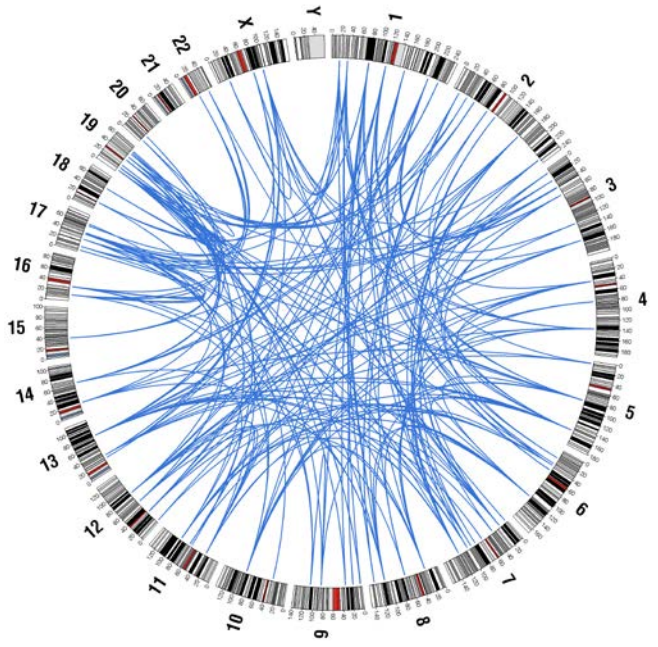
Table 4.1: The number of paralogous CVL

	No. of paralogous group
2 (di-paralogous blocks)	54
3 (tri-paralogous blocks)	118
4 (tetra-paralogous blocks)	38
Sum	**211**

(A)


Y
X
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1

(B)


Y
X
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1

(C)

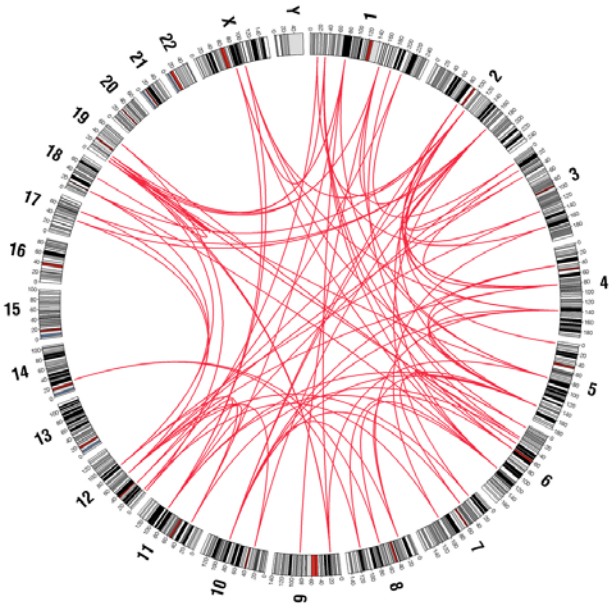


Figure 4.1: Paralogous synteny blocks within human genome

Genomic distribution of paralogous synteny blocks are shown. (A) Di-, (B) Tri- and (C) tetra- paralogous blocks are identified by the gene order and homology.

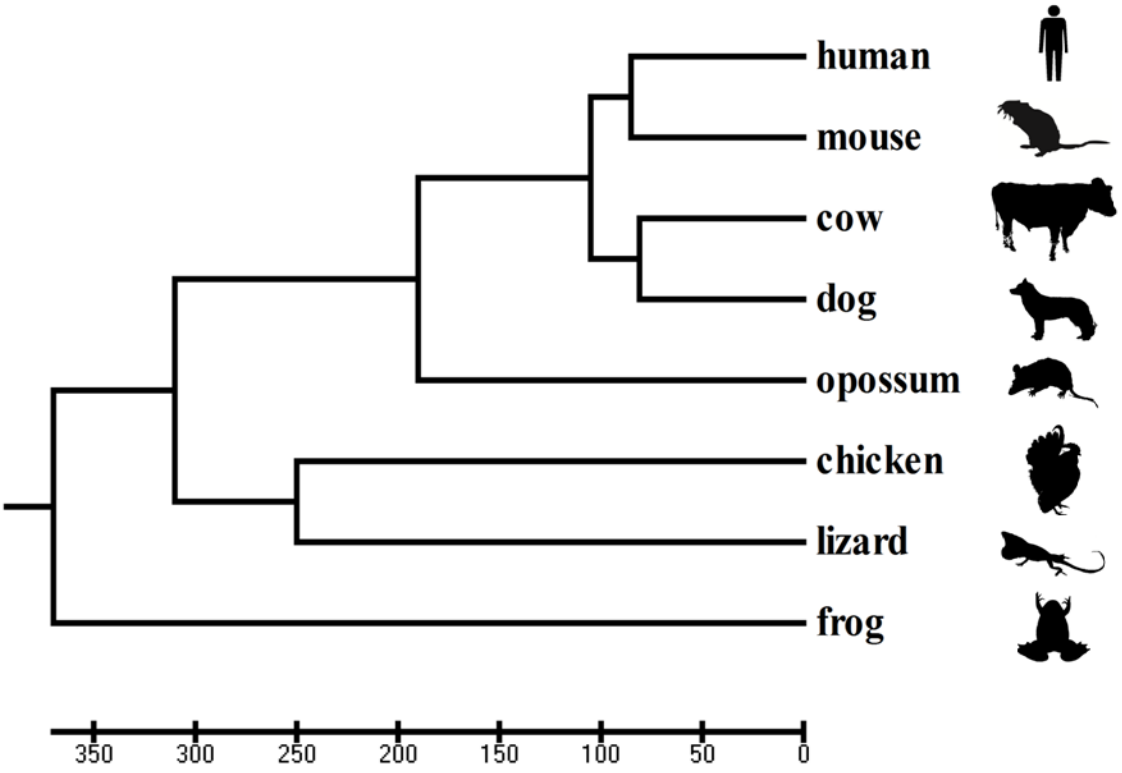


Figure 4.2: Phylogeny of vertebrate species used in this study

The phylogenetic relationships of species used in this study are shown. The scale is one million years ago (MYA).

each block were masked based on the annotations of Ensembl database. BLAST search were carried out between human and mouse (*Mus musculus*; NCBI m37) orthologous blocks to detect orthologous CNSs (Altschul et al. 1997). The default parameter settings of BLAST search were used. The cutoff bit score applied in these comparisons was 200. To evaluate the conservation, human-mouse CNSs were compared with other vertebrate genomes. Species used were dog (*Canis familiaris*; CanFam 2.0), cow (*Bos taurus*; Btau_4.0), opossum (*Monodelphis domestica*; monDom5), chicken (*Gallus gallus*; WASHUC2), lizard (*Anolis carolinensis*; AnoCar1.0), and frog (*Xenopus tropicalis*; JGI 4.1). Teleost were excluded because they underwent an additional genome duplication in their lineages. The cutoff E-value to identify orthologous CNSs was 10^{-5}. From these orthologous CNSs, I determined CNSs conserved among all the 8 vertebrate species included **Figure 4.2**. Those orthologous CNSs were compared with each other with less than 10^{-3} E-value to detect paralogous CNSs.

The detected paralogous CNSs were compared with previously reported sequences that already tested the enhancer activities. The 1,619 human and mouse non-coding elements tested in transgenic mice at 11.5 day spost-coitum (dpc) were downloaded from VISTA Enhancer Browser (http://enhancer.lbl.gov/) on November 2011 (Visel et al.2007). These sequences were compared with my paralogous CNSs

through BLAST search.

4.2.3 Ontology analysis of paralogous CNS-harboring genes

The paralogous CNS-harboring genes were defined and their features were inferred in the following manner. The closest paralogs derived from the 2R WGD, which are conserved among both paralogous loci, were defined as paralogous CNS-harboring genes. I then conducted statistical analysis by using Gene Ontology database (http://www.geneontology.org/) to find significantly enriched paralogous CNS-harboring genes. Analysis of gene function enrichment was performed using Fatigo+ web server (Al-Shahrour et. al. 2007). The paralogous CNS-harboring genes were compared with the entire human genes in the Gene Ontology database to detect the overrepresented paralogous CNS-harboring genes. The expression regions and timings of paralogous CNS-harboring genes were also analyzed. The eGenetics (http://www.nhmrc.gov.au/your_health/egenetics/index.htm) database were used to investigate gene expression of paralogous CNS-harboring genes. Human anatomical system data, which give information about in which organs and when a gene is expressed were obtained from eGenetics database by the Ensembl Biomart (Kelso et al. 2003). I counted how many of genes in total, paralogs derived from the 2R WGD, and

paralogous CNS-harboring genes are expressed in each organ and timing and divided the numbers by the total number of all genes.

4.2.4 Estimation of genes and CNSs loss rate after the 2R WGD

The gene and CNS loss rates were estimated so as to know the tendency of genomic evolution after the 2R WGD. The least-square approach was applied for the calculation of gene and CNS loss rate. I assumed the gene (or CNS) loss rate of the 1st WGD as λ_1, the gene (or CNS) loss rate of the 2nd WGD as λ_2, gene number in ancestor's genome as N and current gene number as M. The following equations can be derived based on these assumptions for number of genes that experienced no loss (M_0), one loss (M_1), two losses (M_2), and three losses (M_3).

$$M_0 = N(1- \lambda_1)^2 (1- \lambda_2)^4 \qquad \text{(Equation 4.1)}$$

$$M_1 = 4N(1- \lambda_1)^2\lambda_2(1- \lambda_2)^3 \qquad \text{(Equation 4.2)}$$

$$M_2 = N\{2\lambda_1^2(1- \lambda_1)\lambda_2^2(1- \lambda_2)^2 + 4(1- \lambda_1)^2\} \qquad \text{(Equation 4.3)}$$

$$M_3 = 8N\lambda_1(1- \lambda_1)\lambda_2^3(1- \lambda_2) \qquad \text{(Equation 4.4)}$$

The minimal values of λ_1 and λ_2 were calculated by the least-square approach, assuming the value of N as 20,000.

Table 4.2: Gene and CNS loss pattern of paralogs derived from the 2R WGD

Conservation level	No. of paralogous gene group (No. of genes)	No. of paralogous CNS group (No. of CNSs)
4	50 (50x4 = 200)	0
3	220 (220x3 = 660)	3 (3x3 = 9)
2	861 (861x2 = 1722)	150 (150x2 = 300)
1	8036 (8036x1 = 8036)	7341 (7341x1 = 7341)
Total	**9167 (10618)**	**7494 (7650)**

4.3 Results

4.3.1 Identification of orthologous CNSs

To identify orthologous CNSs shared among vertebrate species, I carried out comprehensive BLAST searches. The cutoff value of BLAST search directly influences the result of orthologous CNS detection. Because I effectively detected the orthologous CNSs from the Hox gene cluster regions in Chapter 2 (Matsunami et al. 2010), the same cutoff bit score was used for the human-mouse comparison. I discovered 67,052 CNSs from human and mouse genome comparison under the following settings: > 100 bp length, > 78% similarity. Their average length was 318 bp. I compared these human-mouse CNSs with other vertebrate species. The genomes of dog, cow, opossum, chicken, lizard, and frog shared 62611, 65878, 44726, 24549, 19724, and 10664 CNSs with human and mouse, respectively. The number of orthologous CNSs is gradually decreased along the evolutionary distance from human and mouse. Especially, the frog is most distant species from the human. Nevertheless, I identified a large amount of non-coding conservations from the frog genome. Among the 67,052 human-mouse orthologous CNSs, 7,650 CNSs were conserved in all species employed in this study. These CNSs may be recognized among all vertebrate species which did not experience

further genome duplications. The each syntey block has 65 orthologous CNSs per block in average. The blocks are scattered across the whole genome of each species except for the Y chromosome. This conservation might be related to characteristic feature of this sex chromosome. I could not find any correlations between orthologous gene density and orthologous CNS density. Although important development genes such as Hox or Sox have many CNSs, other CNSs are equally distributed in the genome.

4.3.2 Highly conserved synteny blocks

Lundin et al. (2003) reported that the chromosomes bearing the Hox clusters frequently include paralogous genes derived from the 2R WGD and are organized large synteny blocks. These blocks have been considered as a hallmark of the 2R WGD and include not only the Hox clusters but also other important genes such as Dlx, Gbx, Gli and Collagen genes. In my study, Hox linked paralogous synteny blocks also showed prominent paralogous conservation of not only coding regions but also non-coding regions. These Hox linked paralogous synteny blocks were one of highly conserved synteny blocks including abundant paralogous genes. These blocks also have the high numbers of paralogous CNSs, especially di-paralogs CNSs (shown with green lines in the **Figure 4.3**). The paralogous gene-dense regions correspond to the paralogous CNSs

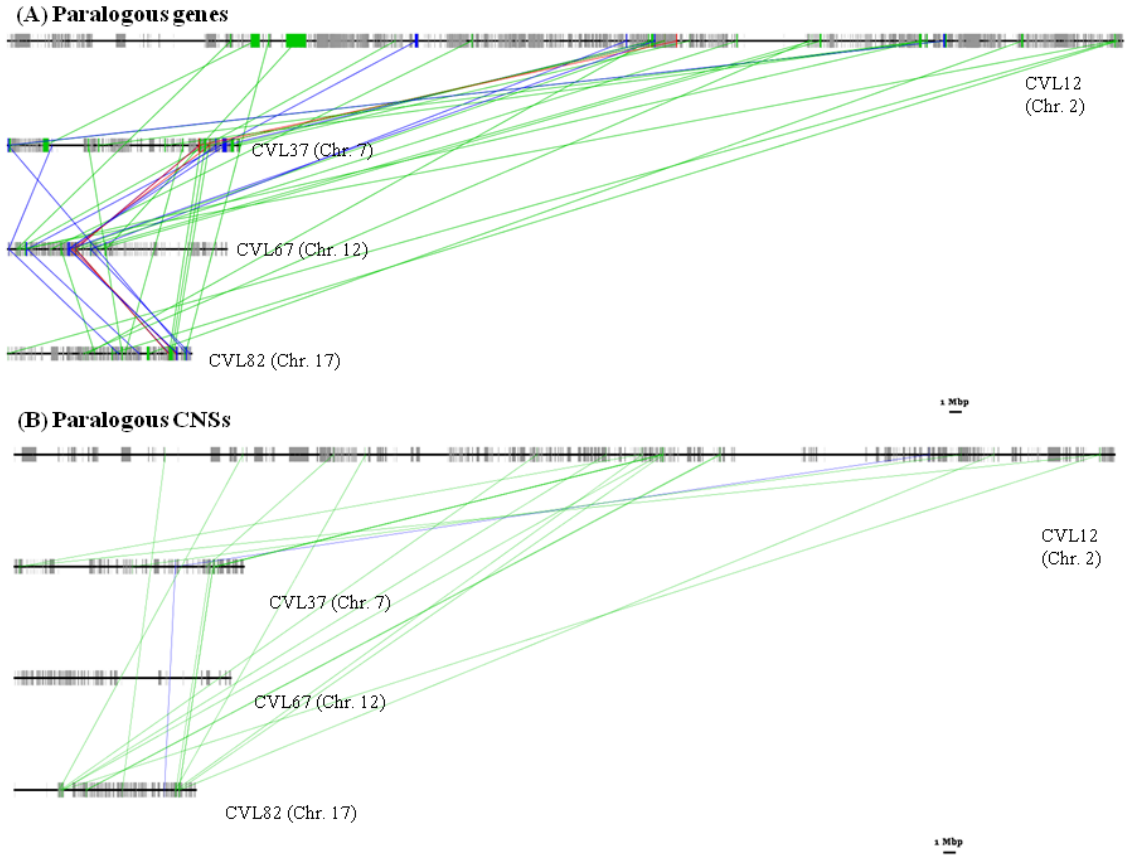


Figure 4.3: Scheme of Hox-linked paralogous block

(A) Paralogous gene conservations and (B) paralogous CNSs conservations were shown. Hox linked paralogous synteny blocks also shows prominent conservation of not only coding regions but also non-coding regions. These tetra-paralogs, tri-paralogs and di-paralogs represents red, blue and green lines, respectively. We could not identify tetra-paralogous CNSs in these regions.

dense regions. Because each paralogous Hox gene shown similar expression pattern controlled vertebrate early development, these CNSs might function as *cis*-regulatory elements such as already known paralogous conserved elements (Lehoczky et al. 2004) and control the similar expression of paralogous Hox or neighboring genes.

4.3.3 Paralogous CNSs

From the vertebrate-specific orthologous CNSs, 309 paralogous CNSs were identified (**Table A4.1**). In the paralogous CNSs, I could find di-paralogs or tri-paralogous CNSs. However, tetra-paralogous CNSs were not detected, because conservation of non-coding regions was lower than that of coding regions, no tetra-paralogous CNSs were conserved. Each paralogous synteny block bears several conservation levels of genes and CNSs, such as trios or pairs. I compared paralogous CNSs I detected with already described CNSs. The results confirmed 115 previously reported paralogous CNSs (McEwen et. al. 2006) and detected 194 new paralogous CNSs. By using enhancer database (Visel et al. 2007), 83 paralogous CNSs were already tested for their enhancer function by the transgenic mice. Out of 83 CNSs, 51 CNSs have positive enhancer functions, when mice were at 11.5 days post-coitum (dpc). Although remaining 22 CNSs have no enhancer activity, they have a possibility of

enhancer functions at other developmental stages. Among CNSs having positive enhancer functions, 42 CNSs show the prominent expression at the developmental brain region. These results, which is a small but significant minority of functional CNSs, suggests that paralogous CNSs may frequently regulate genes which is expressed in brain regions at early developmental stages. Otherwise, I found 196 newly detected CNSs. The functions of these sequences are unknown. The **Figure 4.3** illustrates one of paralogous CNSs located nearby POU3F paralogs. In the alignment, orthologous CNSs are obviously highly conserved. However, the conservation of paralogous CNSs is weaker than that of orthologous CNSs, so that previous studies could not detect these paralogous CNSs. These newly detected paralogous CNSs also have a possibility to work as a distal enhancer.

4.3.4 Location of CNSs and paralogous CNS-harboring genes

I searched paralogous CNS-harboring genes and inferred the functional bias of paralogous CNS-harboring genes based on the Gene Ontology database and the gene expression database. **Table 4.3** shows paralogous CNS-harboring genes, having abundant paralogous CNSs. The majority of paralogous CNSs are located at intron, upstream region or downstream region of the genes encoded transcription factor. The

functions of paralogous CNSs located near well studied transcription factors, such as FoxP1/P2, Sox14/21(McEwen et al. 2006), and Irx cluster (de la Calle-Mustienes et al. 2005), are already known. These paralogous CNSs function as distal enhancers and partially share their gene expression regions between paralogous pairs. Among other paralogous CNSs, some of only one conterpart of paralogous CNS pairs also show experimentally validated distal enhancer functions in the database (**Table A4.1**). This information strongly suggests that other part of paralogous CNS pairs, whose function is not tested yet, have enhancer function, too. Newly detected paralogous CNSs may also have the function of distal enhancer and partially share the gene expression patterns

Table 4.4 is the result of gene ontology analysis. The paralogous CNS-harboring genes were compared with entire human genes in the database. I revealed paralogous CNSs were frequently located genes which function as sequence-specific DNA binding (i.e. transcription factors). These results are consistent with previous studies and suggested that, after whole genome duplication, genes which function as gene regulation are more conservative than genes having other function.

The expression regions and stages of paralogous CNS-harboring genes were investigated by the eGenetics database. I found that paralogous CNS-harboring genes frequently include genes expressed in the brain at early developmental stages (**Figure**

Table 4.3: List of paralogous CNSs harboring genes

Harboring genes	Number of pairs (trios)
FOXP1, FOXP2	6
ZNF503, ZNF703	6
IRX1, IRX3	5
PBX1, PBX3	4
SALL1, SALL3	4
EBF1, EBF3	3
EVX1, EVX2	3
NR2F1, NR2F2	3
POU4F1, POU4F2	3
SOX5, SOX6	3
MEF2A, MEF2C, MEF2D	1 (1)
NFIA, NFIB, NFIX	1 (1)
BCL11A, BCL11B	2
ESRP1, ESRP2	2
FOXB1, FOXB2	2
HOXA5, HOXB5	2
LMO1, LMO3	2
LRBA, NBEA	2
LRP3, LRP12	2
NEUROD1, NEUROD2	2
NRXN1, NRXN3	2
OTX1, OTX2	2
POU3F1, POU3F2	2
POU3F2, POU3F3	2
PRDM16, MECOM	2
SLIT2, SLIT3	2
SOX14, SOX21	2
TCF4, TCF12	2
TFAP2A, TFAP2B	2
TOX, TOX3	2
TSHZ1, TSHZ2	2
VRK1, VRK2	2

GRIA1, GRIA2, GRIA4	(1)

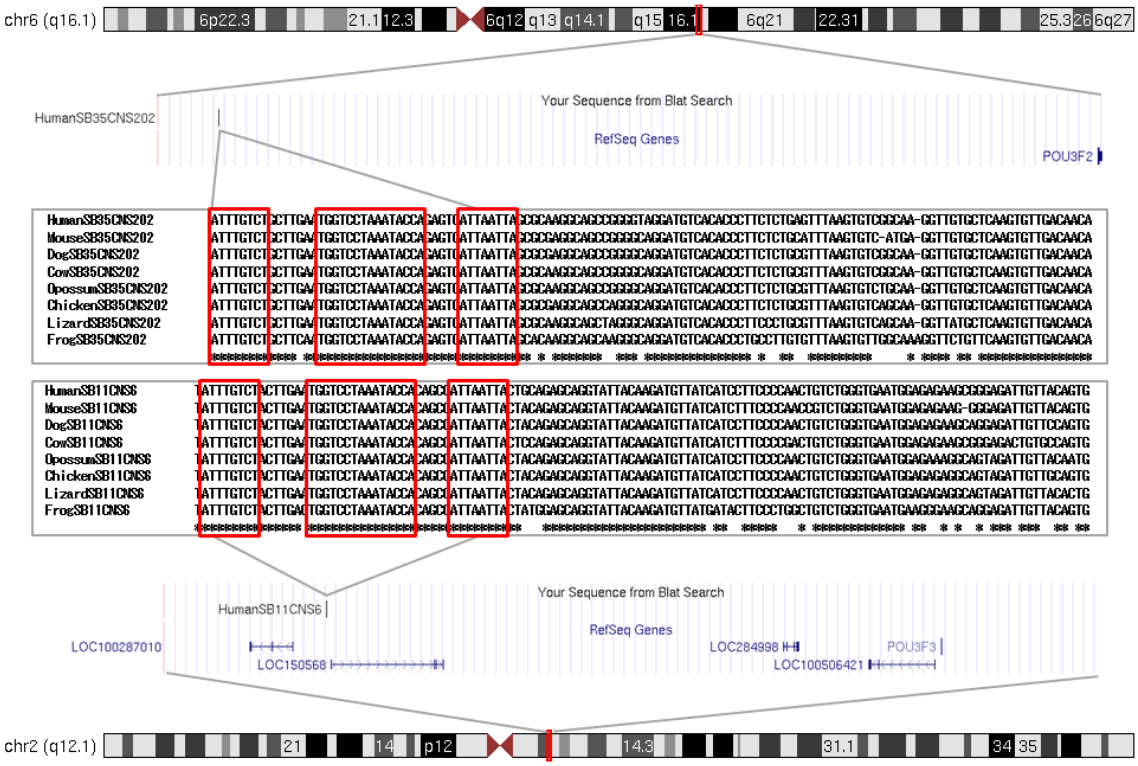


Figure 4.4: Paralogous CNSs shared between POU3F2 and POU3F3 genes

Genomic locations of each orthologous CNS in the human genome and the alignment of paralogous CNS are shown. This paralogous CNS pair is located at nearby POU3 paralogs, POU3F2 (BRN2) and POU3F3 (BRN1), that is derived from the 2R WGDs. These are strong candidates of gene regulatory sequences of these paralogs.

Table 4.4: Overrepresented gene functions of host genes

GO term	P-value
sequence-specific DNA binding (GO:0043565)	3.39E-15
ionotropic glutamate receptor activity (GO:0004970)	7.69E-05
phosphoinositide binding (GO:0035091)	6.05E-05
lipid kinase activity (GO:0001727)	5.33E-04
1-phosphatidylinositol-3-kinase activity (GO:0016303)	8.92E-06
follicle-stimulating hormone receptor activity (GO:0004963)	1.77E-04
low-density lipoprotein receptor activity (GO:0005041)	3.41E-04

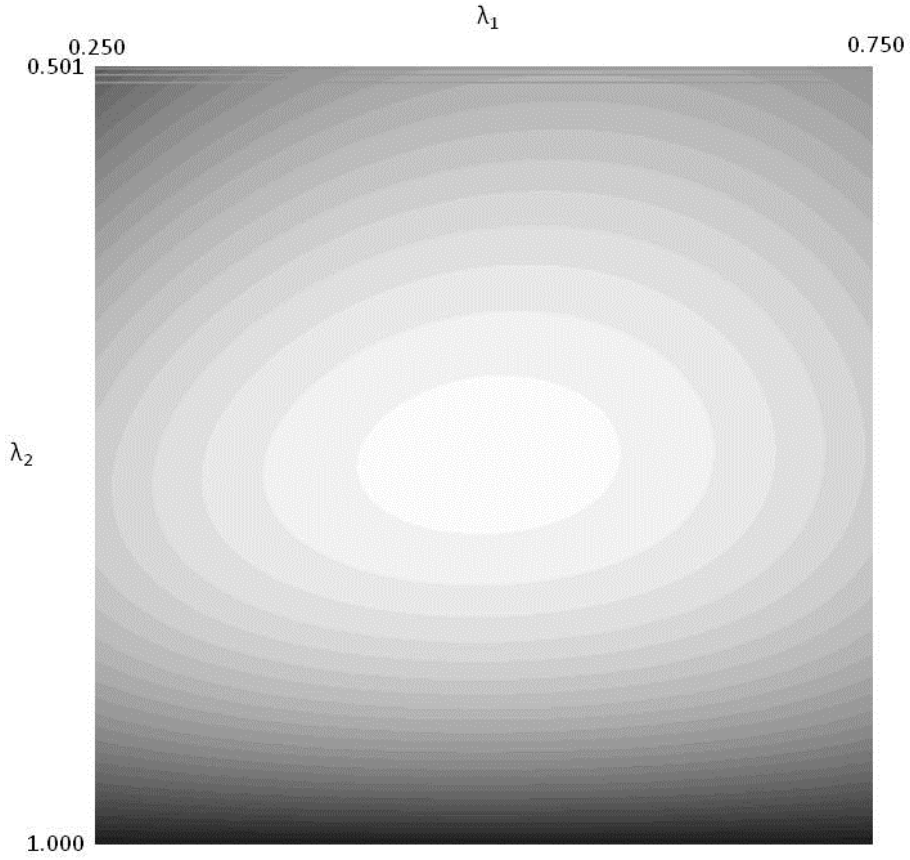
Adjusted P-values are calculated by comparing the distribution of the host genes with that of human genes.

A4.1). Similarly, majority of paralogous CNSs in the enhancer database show expression in brain at early developmental stage (**Table A4.1**). These results imply that existing paralogous CNSs may contribute to vertebrate-specific complex brain morphology at early developmental stages.

4.3.5 Gene loss rate after the 2R WGD

The gene and CNSs loss rates after the 2R WGD were estimated. Recently, the gene loss rate after the fish specific WGD was estimated by Sato et al. (2009). The loss rate after the 2R WGD was unknown, because these events are very ancient. I estimated the loss rate from the existing paralogous gene or CNS combinations such as solos, pairs, trios and quartets (**Table 4.2**). According to the result of least-square approach (**Figure 4.5**), a half of both duplicated genes and CNSs were lost after the first WGD event (gene: $\lambda_1 = 0.504$, CNS: $\lambda_1 = 0.520$). In addition, about three quarters genes and CNSs were lost after the second WGD event (gene: $\lambda_2 = 0.753$, CNS: $\lambda_2 = 0.765$). As we expected, the loss rate of CNSs is higher than that of genes. Although this estimation is very rough, this information is very important to study after the 2R WGD or other lineage specific WGD events.

(A) Gene



(B) CNSs

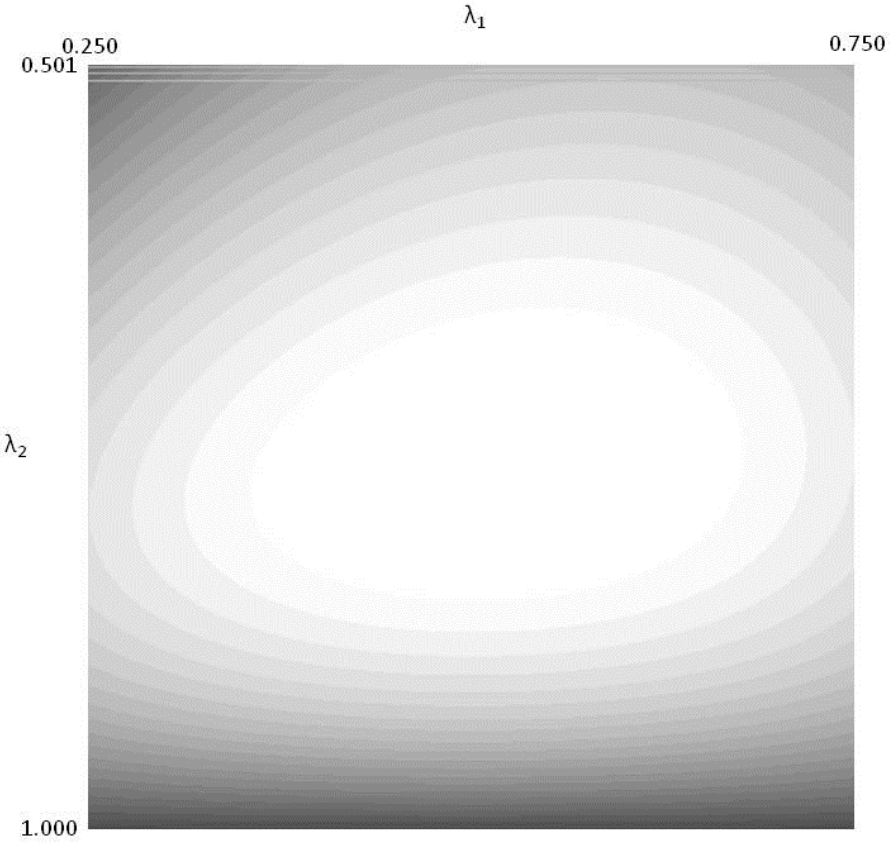


Figure 4.5: Estimation of loss rate after the 2R WGD

(A) gene and (B) CNS loss rate after each WGD event were estimated by the least-square approach. Each gene loss rate was calculated at the interval of 0.001 to estimate smallest value. The minimum values shown in white are (A) $\lambda_1 = 0.504$ and $\lambda_2 = 0.753$, (B) $\lambda_1 = 0.520$ and $\lambda_2 = 0.765$, respectively.

4.4 Discussion

In this chapter, I identified vertebrate specific CNSs. Those may be related to the vertebrate specific features. Previously, conserved non-coding sequences were described by several criteria. Bejerano et al. (2004) defined 483 UCEs. These are including coding regions and ≥200bp length with 100% identity among human, rat and mouse. On the other hand, Woolfe et al. (2007) defined 6957 CNEs. These are ≥40bp length and 65% identity between human and fugu. These genome-wide studies are very rough and often missed the paralogous conservations. It is difficult to compare those results with this study, because used genome sequences are different. My results covered 308 human-rat-mouse UCEs and 3388 human-fugu CNEs. Other conserved elements previously detected are not overlapped with my CNSs. I defined vertebrate-specific CNSs as conserved all 8 vertebrate species in this study. Because my criteria are strict, other conserved elements are difficult to be detected. These missed conserved elements may be overlapped with long gap regions. Although some CNSs are missed because of low genome quality, this study effectively detected the orthologous CNSs.

Why the paralogous synteny block is conserved remains elusive. The genomic regulatory block hypothesis is proposed to explain this enigma (Kikuta et al. 2007;

Becker and Lenhard 2007). This hypothesis suggests that CNSs scattered across each synteny block prevent each block from breakage of the synteny. Under the assumption of this hypothesis, paralogous CNSs maintain paralogous conserved synteny. I inferred the relation between the distribution of paralogous CNSs and the distribution of paralogous genes. As the result, paralogous gene order (synteny) and paralogous CNS conservation are weakly correlated (Pearson's product-moment correlation coefficient = 0.223). This result suggests that paralogous synteny blocks bearing many paralogs also include abundant paralogous CNSs. In other words, highly conserved syntenic regions have more paralogous CNSs. This implies these paralogous CNSs may constrain the synteny blocks from the breakage and play a key role in the genomic regulatory block hypothesis.

The majority of vertebrate CNSs may have been generated before the divergence of the extant vertebrate species. The sequencing of the genome of the cephalochordate, amphioxus (*Branchiostoma floridae*) has uncovered traces of the origins of very small number of vertebrate CNSs (Putnam et al. 2008; Holland et al. 2008). Invertebrate groups have been found to possess their own sets of CNSs (Glazov et al. 2005; Vavouri et al. 2007), and interestingly there is a correlation between the classes of genes around which both vertebrate and invertebrate CNSs cluster. This

suggests parallel evolution of CNS networks (Vavouri et al. 2007). Consequently, whereas the slow evolution of coding sequences can be charted readily across the invertebrate/vertebrate boundary, the CNSs changed very quickly during the vertebrate evolution. Recently, non-coding sequences that are conserved from several basal vertebrates were reported. The elephant shark (*Callorhinchus milii*) is a cartilaginous fish and a basal jawed vertebrate. Its genome contains a few thousand vertebrate CNSs in spite of their low coverage genome information (Lee et al. 2011). However, the number of CNSs retained in the sea lamprey (*Petromyzon marinus*), one of extant jawless vertebrates, is much smaller than that of other vertebrates (McEwen et al. 2009). The lamprey CNSs show remarkably short length and low homology. Whether the jawless vertebrate genomes experience the 2R WGD is still unclear so that the orthologies of sequences are difficult to assign (Kuraku et al. 2008). Nevertheless, these observations suggest that vertebrate CNSs have not constantly evolved. We can interpret that lamprey non-coding sequences are extremely changed such as their coding regions (Qiu et al. 2011), or the evolutionary rate of jawed vertebrate non-coding region has become slower after the jawless vertebrate lineages branched off. To answer these issues, we should analyze the high quality basal vertebrate genomes. In either case, the existence of massive CNSs is not usual situation compare with other invertebrate

species, which are closer to vertebrates. These CNSs might contribute to vertebrate specific features.

The existences of paralogous CNSs detected in this study are difficult to explain by previous duplication models. Classical models predict that the most likely fate of duplicated genes is the degeneration of one of the pair to a pseudogene (or completely lost from the genome) or less frequently the acquisition of novel gene functions as a result of alterations in coding or regulatory sequences in a process known as neo-functionalization. Recently, in the duplication - degeneration - complementation (DDC) model, Force et al. (1999) proposed the possibility of a sub-functionalization in which duplicated genes undergo complementary deleterious mutations in independent subfunctions so that both genes are required to share the functions of the ancestral gene. These models are difficult to explain the existence of paralogous CNSs. The alternative model is gene balance hypothesis proposed by Papp et al. (2003). Although this model has not explicitly been applied to evolutionary fates of non-coding sequences, it postulates that selection against gene dosage imbalances will promote the retention of particular types of genes (Papp et al. 2003). Immediately after a WGD event, genome-wide relative gene dosage is maintained, but subsequent step-wise mutation or deletion of duplicate genes can lead to deleterious dosage imbalances. Genes whose

proteins have many interaction partners may be more sensitive to these dosage changes, possibly leading to an over-retention of highly connected gene functions, such as transcriptional regulators and signaling complexes. Conversely, small-scale genomic duplications immediately disrupt relative dosage, so highly connected genes should avoid this type of duplication during evolution. This different correlation between gene retention after WGD and small-scale duplication is a key distinction between the gene balance hypothesis and the DDC models; DDC should promote the same patterns of gene retention for all types of gene duplication. In support of the gene balance hypothesis, vertebrate genes that function in transcription regulation or signal transduction are over-retained after the 2R WGD events but not after small-scale duplications. I also found that the paralogous CNSs are frequently retained near the transcription factors. The transcription and developmental genes have more complex function than other genes, such as pleiotropic expressions, highly connected protein networks and dosage-sensitive. These characters may allow greater sub-functionalization. They often share gene expression regions and have similar functions among paralogous genes. However, the existence of paralogous CNSs is difficult to explain by the DDC model, because this model does not assume same enhancer functions among paralogous loci. The one possible explanation of the

existence of paralogous CNSs is the gene balance hypothesis. These paralogous CNSs have possibility to control similar expression patterns of paralogs and dosage compensation of paralogs through the highly conserved sequences.

The alternative possible function of paralogoues CNSs is non-coding RNA. Rinn et al. (2007) reported inter-chromosomal interactions between paralogous regions through non-coding RNA. Some enhancers act not only cis but also trans via non-coding RNA transcription (Ørom et al. 2010). Paralogous CNSs may be related to these inter-chromosomal interactions of duplicated genome regions.

CHAPTER 5

De novo transcriptome sequencing of Japanese brook lamprey

5.1 Introduction

Whole transcriptome analysis using next-generation sequencing (NGS) technologies has started to reveal the complex landscape and dynamics of the transcriptome at an unprecedented level of sensitivity and accuracy (Metzker 2010; Ozsolak et al. 2011). Although traditional Sanger EST sequencing only detects abundant transcripts, NGS transcriptome sequencing offers a near-complete snapshot of a transcriptome, including the rare transcripts that have strict regulatory roles with the enormous sequencing depth. In contrast to alternative high-throughput technologies, such as microarrays, RNA-seq achieves base-pair-level resolution and a much higher dynamic range of expression levels, and it is also capable of de novo annotation (Metzker 2010; Ozsolak et al. 2011).

The Roche 454 sequencer, which is massively parallel pyrosequencing machine (Margulies et al. 2005), is most suitable for de novo transcriptome sequencing, because the length of each read is the longest (average: 400 bp) among second generation sequencers. Pyrosequencing had been restricted to model organisms (Bainbridge et al.

2006; Weber et al. 2007; Torres et al. 2008) or closely related species (Toth et al 2007) due to their low coverage data. However, the technical advance increases the depth of coverage of 454 sequencers. Recently, 454 technology has been applied to transcriptome of non-model organisms (Kumar et al. 2010). These studies illustrate the potential of 454 pyrosequencing to rapid characterize expression genes that can be used to address important biological questions.

Because phylogenetic positions of lamprey species are basal of vertebrates, they are an interesting biological group in terms of vertebrate genome evolution. They consist of one clade of jawless vertebrates (Kuraku et al. 2006). Although phylogenetic analysis of these lampreys were already conducted by using a couple dozen gene families (Escriva et al. 2002; Kuraku et al. 2009), it is still an unsolved problem whether jawless vertebrates share the 2R WGD events with other jawed vertebrates or not.

To dissect this problem, Japanese brook lamprey (*Lethenteron reissneri*) was picked up as a sample of this study from several lamprey species (**Figure 5.1**). Because they are living around the Japan sea, we can easily collect them. All lamprey species breed in fresh water, where they spend several years as suspension or detritus feeders (Hubbs et al. 1971). This stage is known as the ammocoetes larval stage. After metamorphosis, while some species parasitize fish and other animals, other species do

not feed after metamorphosis and breed within several months. This Japanese brook lamprey belongs to non-parasitic species (Yamazaki et al. 1998; Yamazaki et al. 2006).

In this study, genome-wide transcriptome sequencing of far east Japanese brook lamprey was carried out by using Roche 454 sequencer. From ammocoetes larval, RNA were extracted and sequenced by Roche 454 sequencer. I also used sea lamprey data in the database. By using massive sequencing data derived from both our experiment and public database, I achieved two aims: to estimate the relative timing of the 2R WGD and to make the pipeline of non-model organism transcriptome sequencing and phylogenetic analysis

5.2 Materials and Methods

5.2.1 Sample preparation

To collect a huge variety of transcripts, I chose a single immature far east brook lamprey (*Lethenteron reissneri*) as a sample (**Figure 5.1A**). The sample was collected by Dr. Kawasaki Tatsuhiko of National Institute of Genetics and Dr. Yasunori Murakami of Ehime University at the Niigata Prefecture. The sample was flash frozen in liquid nitrogen and shipped on dry ice at -80°C until total RNA isolation. After a sample was frozen with liquid nitrogen, it was broken into small pieces. Total RNA was isolated

(A)

(B)

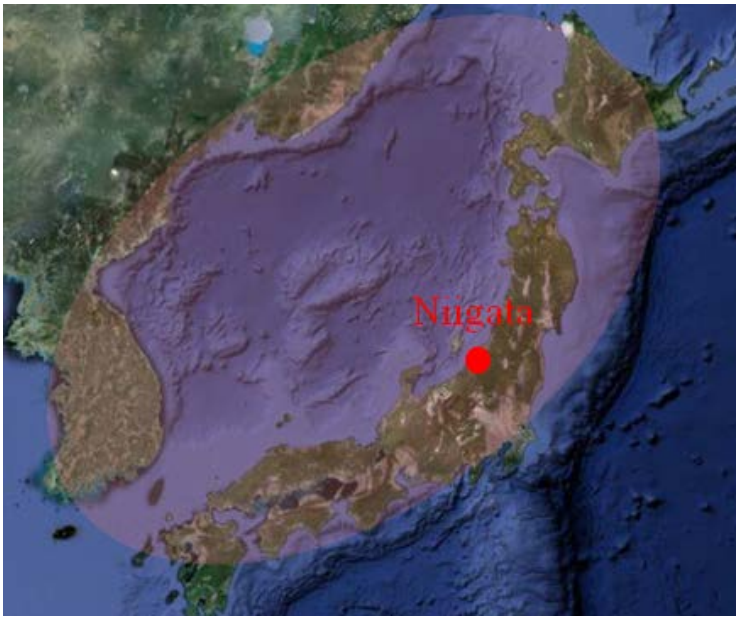


Figure 5.1: The profile of the Japanese brook lamprey (*Lethenteron reissneri*)

(A) The picture of far east brook lamprey at the ammocoetes larval stage. The size is about 10 cm. (B) The habitant of far east brook lamprey. they are living around the Japan sea that is highlighted by light red. The sample used in this study was collected at Niigata prefecture highlighted by red circle in map.

using TRIzol™ (Invitrogen) from whole body homogenization of larval under the manufacture's protocols. The preparation of the sample was done by Dr. Sato Yukuto of National Institute of Genetics.

Because the efficiency of cDNA synthesis is lower in lampreys due to its higher GC-content in general, we synthesized first-strand cDNA directly from the total RNA. Genomic DNA was digested by DNaseI. First and second strand cDNA were synthesized from mRNA using the Takara® cDNA Synthesis Kit (M-MLV version). Double-stranded DNA was fragmented with the Covaris sonicator (Covaris, US). The condition of the sonication is Duty cycle = 10%, Intensity = 5, Cycle/Burst = 200. The fragmented cDNA were cleaned up with AMpure (Beckman coulter genomics, US). Fragmented cDNA were modified to the blunt-end by T4 DNA polymerase (Takara) and adenylated.

Because the poly(A)-priming method reduces accuracy and quality of the 454 sequence reads, we employed a newly developed method for amplification of mRNA-derived cDNA using partially-annealed adaptors. Titanium adaptors (Roche/454 Life Sciences, CT) were put the each end of the cDNA by adding 5 μl sample, 6 μl Ligation solution II, 12 μl Ligation solution I, 2μl (10 μM) Titanium adapter RL-MID1-A/B mix (RL-MID1-A; 5'-CATCTCATCCCTGCGTGTCTCCGACGA

CTACACGACGACT-3', RL-MID1-B; 5'-GTCGTIGTGTIGTCGICGTCTCTCAAGGCA CACAGGGGATAGG-3') and incubated the ligation reaction at 16 °C for 30 minutes. The reaction was cleaned up using AMpure (Beckman coulter genomics, US). To gain enough cDNA molecules from small amount of purified cDNA, PCR amplification was done. Fifteen μl of cDNA was used as template for amplification in 25 μl PCR reactions containing, 2.5 μl 10 Ex Taq buffer (Takara), 0.5 μl (10 pmol/μl) emPCR primer A/B (emPCR primer A; 5'-CCATCTCATCCCTGCGTGTC-3', emPCR primer B; 5'-CCTATCCCCTGTGTGCCTTG-3'), 2.0 μl dNTP mix and Takara Ex Taq (Takara). The PCR conditions were as follows: 94°C for 1 minute, followed by 30 cycles with 94°C for 30 seconds, 60°C for 30 seconds and 72°C for 30 seconds, with a final extension of 72°C for 3 minute. Because adapters were put after the sonication, we could avoid the amplification of poly(A) sequences.

5.2.2 Sequencing and assembly

To estimate more abundant gene families' histories, both the newly sequenced sample and data in the database were used. The samples we prepared were sequenced by Roche 454 GS FLX and assembled by using Newbler 2.3 program at Professor Fujiyama Asao’s Laboratory at National Institute of Genetics. The sea lamprey

(*Petromyzon marinus*) RNA sequences in the sequence read archive (SRA) database were also collected (**Table 5.1**). The data were sampled from liver (ID: SRS117161) and brain (ID: SRS117159). These were sequenced by illumina GA IIX. These data were downloaded from database and trimmed the low quality reads by SolexaQA program (Cox et al. 2010). Trimmed reads were assembled by using velvet program (Zerbino et al. 2008).

5.2.3 Phylogenetic analysis

Putative orthologs were searched from several representative taxa. To detect orthologs, BLASTX homology searches were done between lamprey (as query) and human (*Homo sapiens*), chicken (*Gallus gallus*), and amphioxus (*Branchistoma floridae*) (Altschul et al. 1997). The multiple hit lamprey contigs against these 3 species database were extracted. These putative orthologous groups show overlapping hit region for all 4 species, using same codon frame and not including stop codon in sequences. They were aligned by MAFFT program (Katoh et al. 2002). To confirm the significance of results, the evolutionary distances of each group were calculated by PAML program (Yang 1997). From the evolutionary distances, unrooted phylogenetic trees of each group were reconstructed.

To infer the timing of the 2R WGD, putative orthologous and paralogous relationships of each lamprey contig were estimated by the phylogenetic reconciliation method. Putative orthologous groups and human and chicken paralogs derived from the 2R WGD defined Nakatani et al (2007) were clustered by Perl script written by myself and they were aligned by using MAFFT. The NJ trees (Saitou and Nei 1987) of each gene family were reconstructed with 1,000 bootstraps. The duplication nodes of each tree were determined by using Notung program (Durand et al. 2006). The duplications before lamprey divergence were defined as pre-duplication. The duplications after lamprey divergence were defined as post-duplication.

5.3 Results

5.3.1 Contigs

Each read generated by NGS machines was assembled to contigs. Each of contigs was assumed one putative transcript. The two different assemblers were used for two different data set, because amount and length of reads were different (**Table 5.2**). The data of Japanese brook lamprey and sea lamprey was assembled by using Newbler 2.3 and velvet program, respectively. These programs are specialized in each NGS machine. The total RNA size of whole lamprey genome was calculated assuming that

lamprey genome has 30,000 RNA sequences with average 1,500 bp. From this rough estimation, expected total RNA size is 45 Mbp. This value was used to calculate the expected coverage of sequence data.

When the sea lamprey data was assembled, 78,947 contigs (average = 181.83 bp) were obtained. From the database, about 6.0 Gbp RNA sequences were downloaded. The length of each read was 50 bp. The coverage of data was very high at this step. However, after the trimming of low quality data, the coverage became small. This is because the quality of data was poor. Especially, the data from brain was very poor. The high quality sequences included abundant poly(A) sequences that was no use to analysis. The BLASTX homology search was carried out by using these contigs against human whole protein sequences. Although many hits (6,623 contigs) were detected, each contig length was too short to estimate the molecular phylogeny of each gene. This result was not enough to estimate duplication timings.

When the Japanese brook lamprey data was used, each read we obtained (average = 196.04 bp) was shorter than usual read of Roche 454 sequencer, which is about 400 bp on average (**Figure 5.2**). The factor caused this problem was unknown. The lamprey genome specific feature such as high GC content may be related. The coverage of this data is small because of character of sequencer. After these reads were

assembled by Newbler, 7,708 contigs were obtained. The average of contig length of Japanese brook lamprey is longer than that of sea lamprey. However, the number of contig is small, only one-tenth of sea lamprey. The hit number of BLASTX search against the human whole proteins is also small. The hit length of each contig was not long enough to estimate the duplication history of each gene family with high statistical significance.

In summary, the data amount of all lamprey contigs was not enough. When I focused on the BLAST hit more than 100 amino acid sequences that are suitable for the molecular evolutionary analysis, the number of amino acid sequences was similar between sea lamprey and Japanese brook lamprey. Although the sequencing coverage of Roche 454 was smaller than that of illumina GA IIx, Roche 454 generated similar amount of long contigs. Although the data amount was not enough, the read of 454 Roche sequencer is better suited for de novo transcriptome sequencing than illumina GA IIx.

5.3.2 Orthologous gene clustering

The orthologous genes of each lamprey contig were identified for inferring the ortholog/paralog relationship of each gene family. Human and chicken were chosen as

Table 5.1: Status of Sea lamprey data in the SRA database

Sample ID	SRS117159	SRS117161
Sample	RNA (Brain)	RNA (Liver)
Sequencer	Illumina GA 2x	Illumina GA 2x
Length of read (bp)	50	50
Total Length (Gbp)	2.8	3.1

Table 5.2: Summary of read assembles

	Petromyzon marinus **(Sea lamprey)**	***Lethenteron reissneri*** **(far east brook lamprey)**
Sequencer	Illumina GAIIx	Roche 454 GS FLX
Total Reads	119,412,170	426,476
Average read length /after trimming (bp)	50 / 25.68	196.04 / -
Coverage* /after trimming (bp)	x106 / x68	x2/ -
Assembler	Velvet	Newbler
Total contigs	78,947	7,708
Average contig length (bp)	181.83	335.77
No. of BLASTX hits **	6,623	1,923
No. of More than 100aa contig	704 contigs	423 contigs

*Coverage is calculated as lamprey whole RNA size assuming 45 Mbp

** The cutoff value is E-value < 0.1 against human protein.

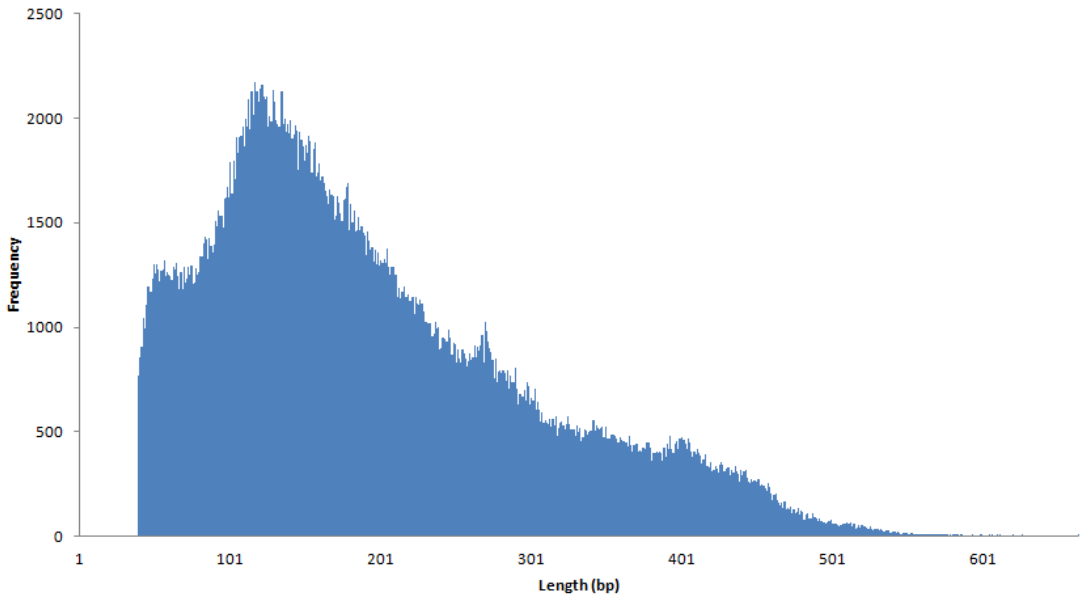


Figure 5.2: Distribution of 454 read length

This is the distribution of 454 read length of far east brook lamprey. The reads less than 40 bp are eliminated. The average of read length (196.04 bp) is clearly shorter than usual 454 read length (400-600 bp).

representative vertebrates. Amphioxus was chosen as the outgroup. The ortholog candidates were searched by BLASTX program against these taxa. The lamprey contigs which hit to all other taxa were extracted (**Table 5.3**). Because these protein sequences were aligned in the next step, each hit of BLASTX search was filtered by following criteria. These are 1) homologous genes shared alignable region for all 4 species, 2) they used same codon frames and 3) they did not include stop codons in their aligned regions. The sea lamprey and Japanese brook lamprey has 1,738 and 656 contigs that are conserved among all four species, respectively. The homologous sequences between two lamprey species were also searched by BLASTX. However, the number of homologous sequence pair was small, only 78 contigs (**Figure 5.3**). These sequences were dealt with independent gene groups. Eventually, 2,394 contigs were remained for next step.

The evolutionary distances between sequences within each group were calculated, and the unrooted phylogenetic trees of each group were reconstructed by using codeml program in the PAML4 package. In the **Figure 5.4**, the four possible topologies of unrooted trees are shown. **Figure 5.4A** is the putative orthologous tree, and **Figure 5.4B** and **5.5C** are putative paralogous trees. The remaining one has no internal branch (**Figure 5.4D**). If the results of clustering are reliable, orthologous tree

Table 5.3: The results of BLASTX homology search

(A) Sea lamprey

Query	Database	No. of hits
Sea lamprey	Human	6,623
Sea lamprey	Chicken	3,252
Sea lamprey	Amphioxus	3,213
Conserved among all 4 species		1,738

(B) far east brook lamprey

Query	Database	No. of hit
Brook lamprey	Human	1,923
Brook lamprey	Chicken	1,336
Brook lamprey	Amphioxus	2,075
Conserved among all 4 species		656

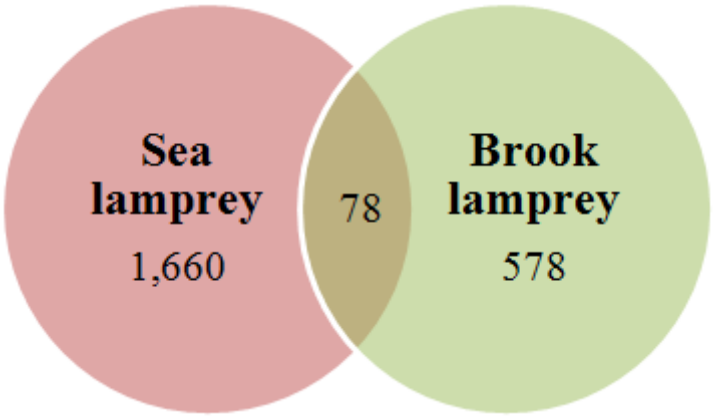
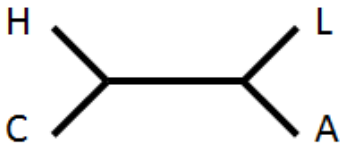


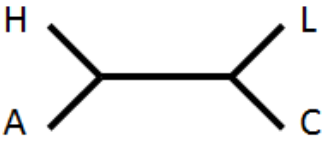
Figure 5.3: Venn diagram of BLASTX hit of each lamprey against human sequences

A small number (78 contigs) of BLASTX hits is overlapped with same human sequences, suggesting that the contigs from far east brook lamprey include newly sequenced RNA.

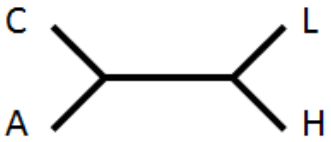
(A)



(B)



(C)



(D)

Figure 5.4: Possible topology of unrooted phylogenetic tree among human, chicken, lamprey and amphioxus

The topology (A) is most ideal orthologous unrooted tree. The topology (B) and (C) are possible paralogous trees with gene losses in each lineage. Because internal branch length is 0, the topology (D) is difficult to assign the ortholog/paralog relationship. H; Human, C; Chicken, L; lamprey, A; Amphioxus

Table 5.4: Summary of orthlogous gene tree topology

	HC-LA	HA-LC	HL-CA	HLCA
Brook lamprey	385	117	112	42
Sea lamprey	746	328	331	333
Total	1131	445	443	375

Note: H; human, C; chicken, L; lamprey, A; amphioxus

may be the majority. In fact, putative orthologous trees were majority (**Table 5.4**). The orthologous genes were effectively detected by this analysis.

5.3.3 Phylogenetic reconciliation

The relative duplication timing of each gene family was estimated to utilize the result of orthologous gene clustering. The best way of this estimation is to use paralogous lamprey sequences. Because paralogous lamprey genes derived from the 2R WGD were not detected in this study, the human and chicken paralogous genes derived from the 2R WGD defined by Nakatani et al. (2007) were used for analysis to estimate the duplication timings of each gene tree. These paralogous genes were clustered with each orthologous gene group. They were aligned and NJ trees were reconstructed. The gene duplication timings of each phylogenetic tree were estimated by using the phylogenetic reconciliation method implementing in the Notung program.

The possible timings of the 2R WGD were estimated by using the results of the phylogenetic reconciliation. Because only 4 species were used in this study, these pre and post duplications corresponded to amphioxus-lamprey split and lamprey-chicken slit, respectively (**Figure 5.5**). The relative duplication timings of each phylogenetic tree were counted with bootstrap probability of duplication nodes. The results were divided

into 4 groups (only pre-duplication, only post-duplication, both pre- and post-duplication and unknown) and shown in **Table 5.5**. The duplication nodes of each gene tree were counted by several bootstrap cutoff. When the cutoff values are more stringent, the number of trees which support each hypothesis is decreasing. The phylogenetic trees show a huge variety of topologies. Because some trees did not have duplication events or were difficult to assign the duplication node as pre or post duplication, these trees were added into the "unknown" group. Among these trees, one phylogenetic tree clearly shows pre-pre-duplication topology (**Figure 5.6**). By contrast, another one shows post-post-duplication topology (**Figure 5.7**).

From the phylogenetic reconciliations, phylogenetic trees for 385 gene families were initially estimated to be used for the relative duplication timings. However, when I chose trees which contained two duplication events and had high statistical branching pattern supports, only 55 trees were left. The majority (49) of them showed the pattern that two genome duplications both occurred before the agnathans/gnathostomes divergence. In addition, when I chose trees which contained only one duplication event with high bootstrap values, trees shown pre-duplication were also majority (195). The original tree topologies of these pre-duplication trees were pre-pre-duplication trees or pre-post-duplication trees under the assumption of the 2R WGD. However, the number

of pre-post-duplication trees with high bootstrap values is very small (6). Most pre-duplication trees, thus, might be originally from pre-pre-duplication trees. However, those are preliminary results. The data amount of this study is not enough. Moreover, the tree topology easily change by taxon sampling or alignment length. To get more reliable result, further study will be needed.

5.4 Discussion

The sequence reads generated in this study were shorter than usual Roche 454 sequencer read. These short reads caused the difficulty of phylogenetic analysis. In general, the lamprey genome, especially protein coding regions, has very specialized features, such as high GC content (Kuraku et al. 2006) and strange codon usages (Qiu et al 2011). These features make sequencing lamprey cDNA difficult, when traditional Sanger method were used. In this study, a new cDNA synthesized method is used to avoid reducing accuracy and quality of the 454 sequence reads because of ploy(A) priming. All together, these factors have possibility that caused unusual short read length in this study. However, it is difficult to identify the main factor of short read length at this time.

The velvet program is applied to assemble illumina NGS reads in this study.

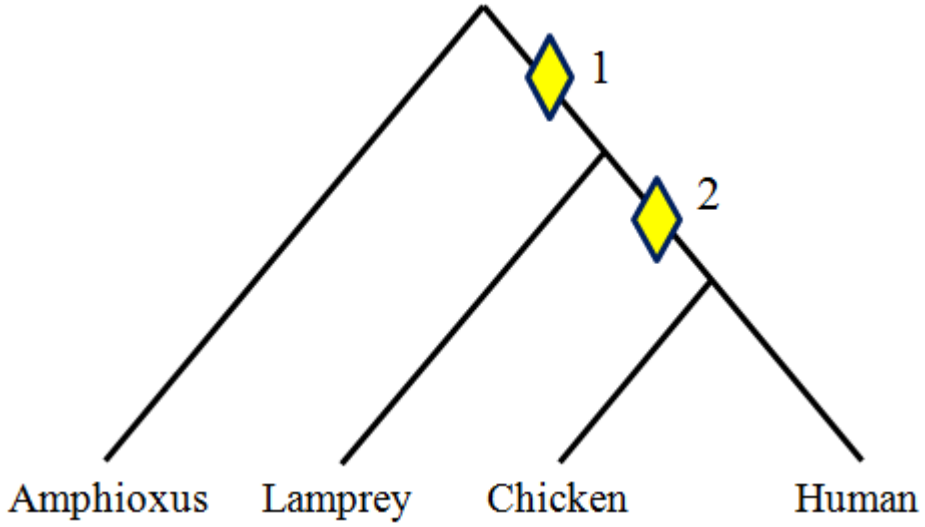


Figure 5.5: The phylogenetic relationship of species used in this study with the possible timing of the WGD event

The species tree are shown. Amphioxus is defined as outgroup of vertebrates. Possible WGD time is shown in yellow diamond. The relative timing of the 2R WGD is still in debate. However, there are two possible candidate timing. These are 1: Pre-duplication (duplication before lamprey divergence) or 2: Post-duplication (duplication after lamprey divergence).

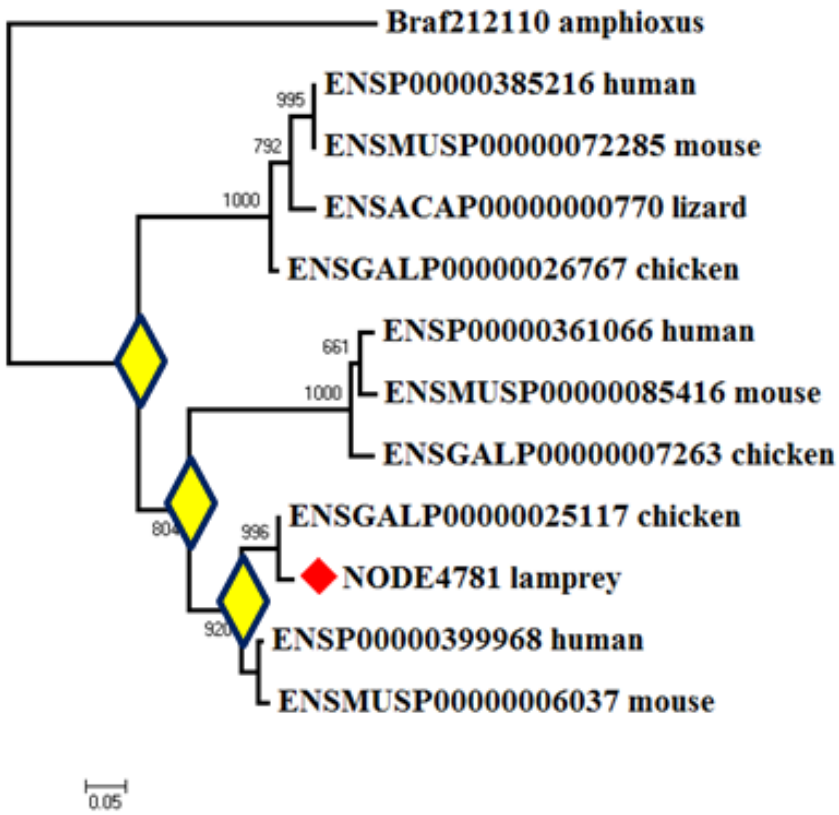


Figure 5.6: Representative Pre-Pre duplicated gene family phylogenetic tree

The phylogenetic tree of each gene family was reconstructed. The values of each node show the bootstrap probability from 1000 trials. The leaf node show common species name and database ID of each gene except for lamprey. The lamprey sequence is highlighted by red diamond with contig ID. The yellow diamonds means duplications. This gene family experienced twice genome duplications before lamprey divergence with some losses.

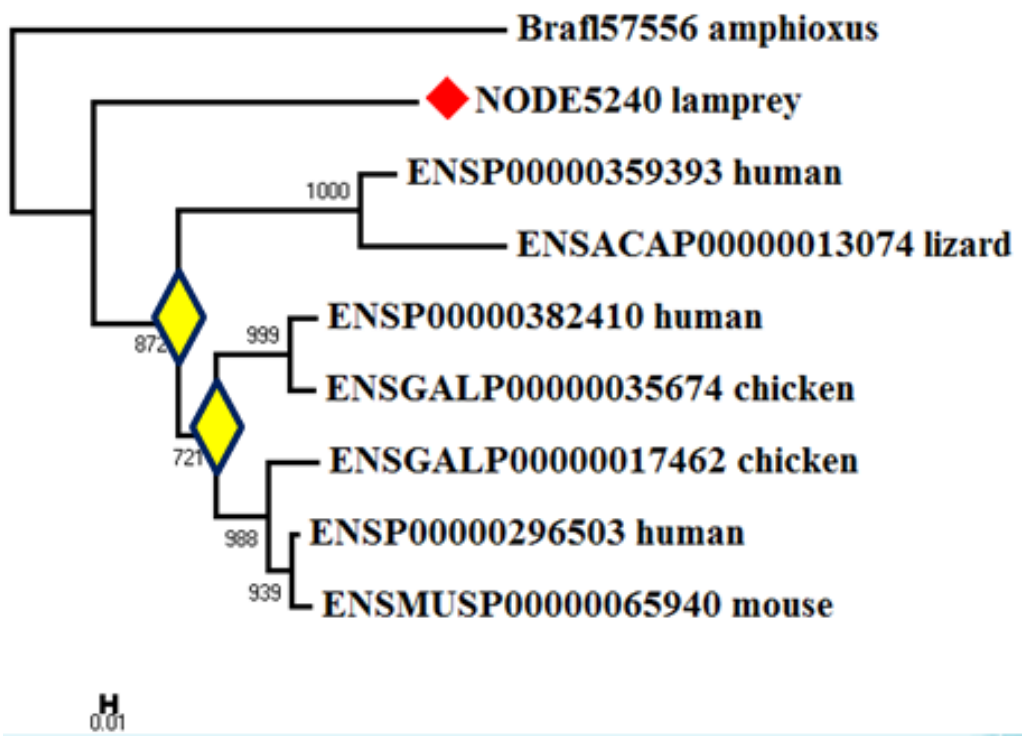


Figure 5.7: Representative Post-Post duplicated gene family phylogenetic tree

The phylogenetic tree of each gene family was reconstructed. The values of each node show the bootstrap probability from 1000 trials. The leaf node show common species name and database ID of each gene except for lamprey. The lamprey sequence is highlighted by red diamond with contig ID. The yellow diamonds means duplications. This gene family experienced twice genome duplications after lamprey divergence with some losses.

Table 5.5: The results of phylogenetic reconciliation analysis

Pre - duplication	Post - duplication	No cutoff	Bootstrap ≥70 %	Bootstrap ≥ 90 %
Only pre-duplication				
4	0	7	0	0
3	0	14	5	0
2	0	76	68	49
1	0	123	163	195
Only post-duplication				
0	4	1	0	0
0	3	4	0	0
0	2	21	4	0
0	1	28	22	14
Both pre- and post-duplication				
3	1	3	0	0
2	2	1	0	0
2	1	9	3	0
1	3	1	0	0
1	2	3	0	0
1	1	13	11	6
Unknown				
0	0	54	82	94

Recently, many assemble programs used for de novo transcriptome sequencing are available (Kumar et al. 2010). Because the result of assemble is very sensitive and easily changed by sequence contents or quality, we should carefully choose the assemble programs, considered with several factors.

The choice of taxa to use evolutionary genomic analysis is important. Because this study focused on vertebrate evolution, we should equally use representative species of each vertebrate clade with similar evolutionary distances. Human, chicken and amphioxus genomes were used in this study. Because this was preliminarily analysis, only four species were used for comparison. However, these four species are representative of each clade. Although the number of species are small, it is possible to estimate the correct duplication timing by these species.

When the relative timings of the 2R WGD are inferred, paralogs of jawless vertebrates are used in previous studies. However, I used already annotated human paralogs derived from the 2R WGD instead of lamprey paralogs to overcome the limitation of analysis caused by small amount lamprey gene data. The rapid progress of NGS technology make reading huge amount of transcripts more easy and will shed a light on this fundamental problem of vertebrate genome evolution. The preliminary result suggests that lamprey and jawed vertebrate genomes share twice WGDs. To

produce more reliable results, we need more massive data of jawless vertebrate transcripts.

CHAPTER6

Inferring the timing of the 2R WGD from lamprey genome data

6.1 Introduction

The relative timings of the 2R WGD are still unclear. There are three possibilities (**Figure 6.1**). Ohno (1970) addressed that the timing of the 2R WGD occurred both before the all vertebrates diverged (**Figure 6.1C**). The alternative hypothesis that jawless vertebrates and jawed vertebrates shared only one genome duplication (**Figure 6.1B**) was supported by Escriva et al (2003) who used data for 33 gene families and by Fried et al. (2004) who used data for Hox genes. More recently, Kuraku et al. (2009) supported Ohno's hypothesis based on data for 55 gene families. Because the divergence time of jawless vertebrates and jawed vertebrates was very ancient, their phylogenetic relationship is not easy to determine. Another problem is a high bias in the codon usage of jawless vertebrate (Qiu et al. 2011). These difficulties make the problem more complex.

Recently, the sea lamprey (*Petromyzon marinus*) genome sequences appeared in the public database. In this chapter, I investigated the possibility that gene losses

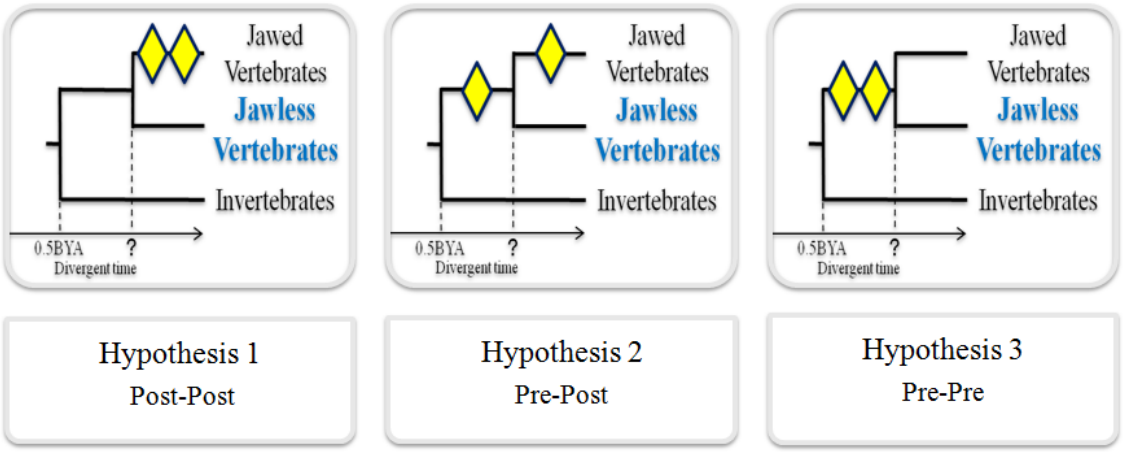


Figure 6.1: Three possible scenarios for timings of 2R genome duplications

Expected tree topologies (hypotheses 1-3) for gene phylogeny are illustrated for an imaginary gene family comprising one invertebrate out-group and jawed vertebrate genes with jawless vertebrate genes.

caused misunderstanding of true ortholog/paralog relationships by using newly released sea lamprey genome data.

6.2 Materials and Methods

6.2.1 Homologous gene clustering

Sequences are clustered into homologous gene families to reconstruct phylogenetic histories. All amino acid sequences were collected from 14 vertebrate species and 6 outgroup species from the Ensembl database (http://www.ensembl.org/index.html). Vertebrate species are human, mouse, rat, dog, cow, opossum, platypus, chicken, lizard, frog, medaka, tetraodon, zebra fish, and sea lamprey. Out-group species are sea squirt, amphioxus, sea urchin, fly, nematode and sea anemone (**Table 6.1**). Sequences were clustered into homologous gene families to reconstruct phylogenetic histories. These genes were clustered to the 14,299 homologous gene clusters by using BLAST search. The default BLAST parameters were used; the cutoffs were the E-value < 0.001, the identity (%) > 30 and the detected homologous length of gene / the full length of gene > 0.75. Each group was aligned by using MAFFT. The NJ trees were reconstructed, only when gene families were at least with one outgroup sequence.

Table 6.1: Sequences used in this study

Common Name	Species Name	Protein Sequence
Human	*Homo sapiens*	21,164
Mouse	*Mus musculus*	23,228
Rat	*Rattus norvegicus*	22,490
Cow	*Bos taurus*	19,030
Dog	*Canis familiaris*	19,292
Opossum	*Monodelphis domestica*	19,453
Platypus	*Ornithorhynchus anatinus*	13,401
Chicken	*Gallus gallus*	16,736
Lizard	*Anolis carolinensis*	17,660
Frog	*Xenopus tropicalis*	18,023
Medaka	*Oryzias latipes*	19,686
Tetraodon	*Tetraodon nigroviridi*	19,589
Zebra Fish	*Danio rerio*	24,147
Sea Lamprey	*Petromyzon marinus*	11,429
Sea Squirt*	*Ciona intestinalis*	14,180
Amphioxus*	*Branchiostoma floridae*	50,817
Sea Urchin*	*Strongylocentrotus purpuratus*	42,420
Fly*	*Drosophila melanogaster*	14,128
Nematode*	*Caenorhabditis elegans*	20,178
Sea Anemone*	*Nematostella vectensis*	27,273

*; used as out-groups

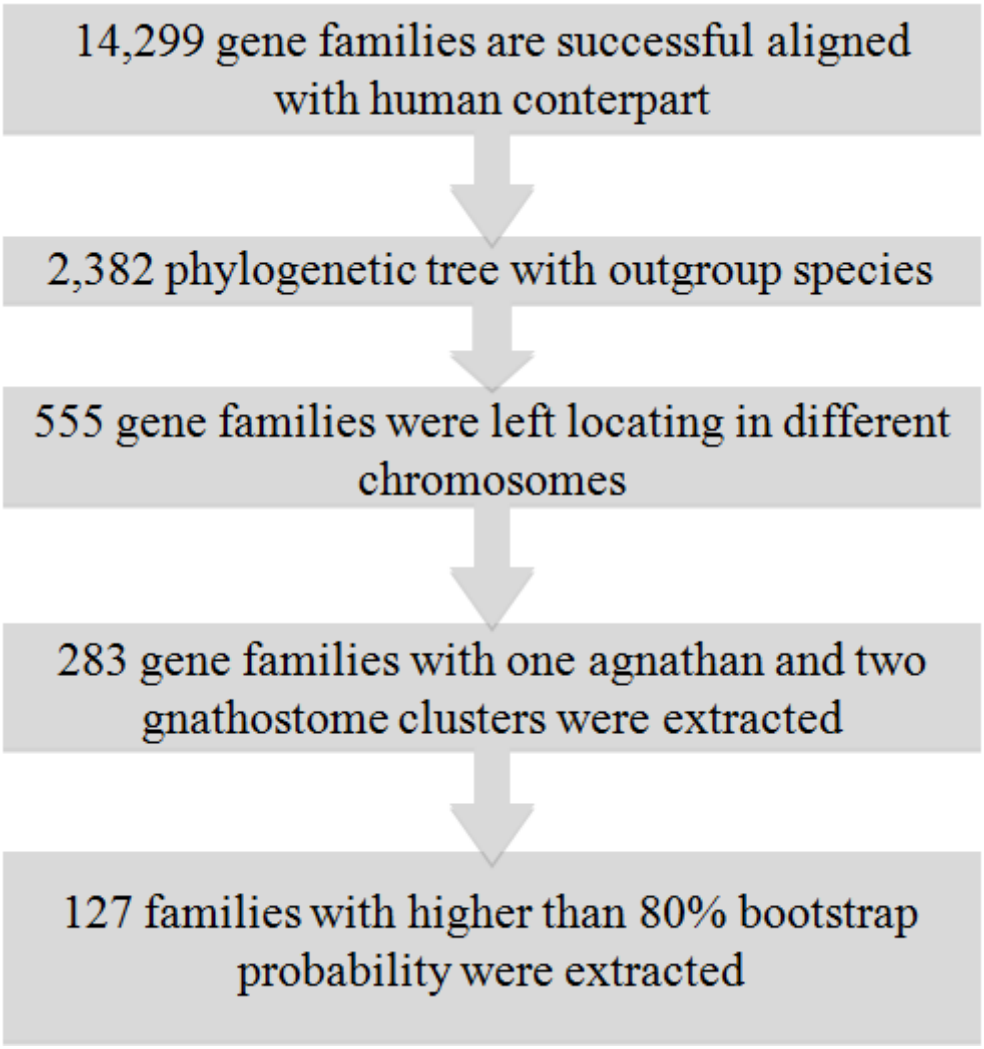


Figure 6.2: Pipeline of analysis

The pipeline of this study is shown. Sequences were grouped into 14,299 gene families.

After the series of filtering, 127 families were remained and analyzed.

To confirm whether paralogs derived from WGD or tandem duplication, synteny information was used. Each human gene location was downloaded from the Ensembl database. The human genes of homologous gene families were mapped to chromosomes. When more than 2 paralog pairs were located at same chromosomes, genes were defined as syntenic paralogs. These procedure are shown in **Figure 6.2**.

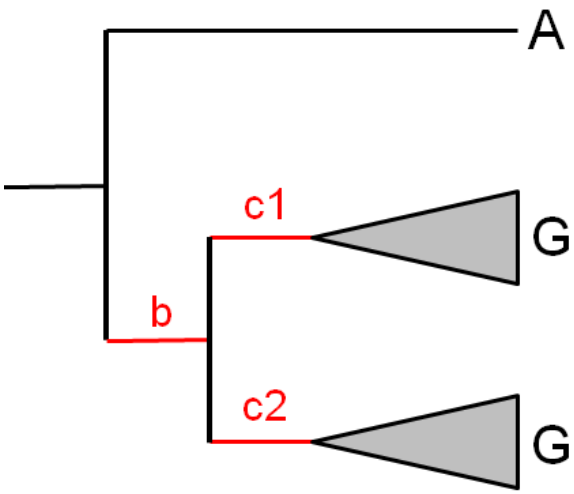
6.2.2 Calculation of branch length

The homologous gene families were divided according to the number of gnathostome clusters and the number of agnathan sequences (**Table 6.2**). The gene families which include one gnathostome cluster and two gnathan sequences were selected for analysis, because these groups were retained the largest number of gene families. The topologies of these gene families were divided into two different groups (**Figure 6.3**). Internal branch lengths of each phylogenetic tree were measured and compared each other.

6.3 Results

I reconstructed phylogenetic trees of 545 syntenic gene families including trees with low bootstrap values (**Table 6.2**). The expected number of paralogs in each gene

(A)



(B)

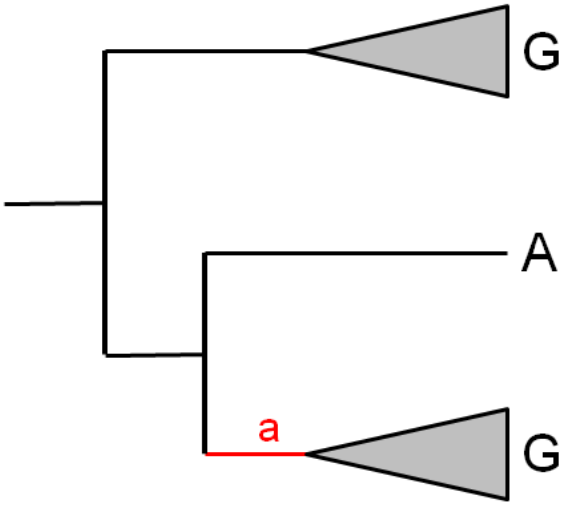


Figure 6.3: Possible topologies of one agnathan and two gnathostome phylogenetic tree

(A) topology share shared gene duplication only among agnathan and (B) topology shared gene duplication shared among all vertebrates are shown. The red internal branch represents distance from agnathan speciation to gnathostome divergence.

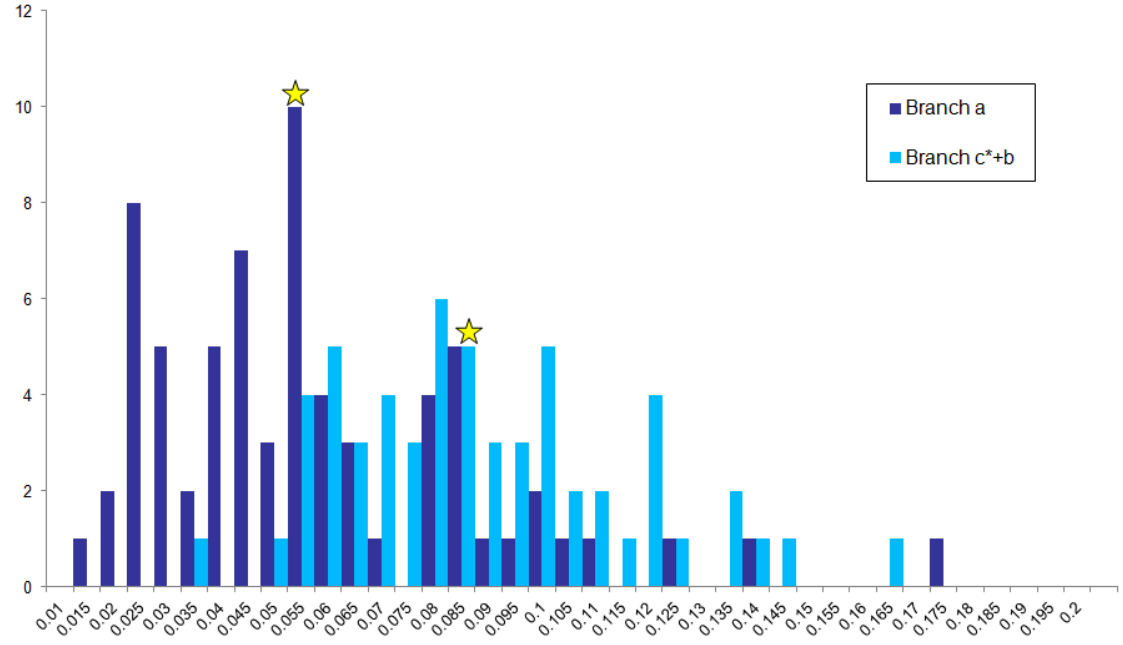


Figure 6.4: Distribution of internal branch length

The frequency of internal branch lengths a and c* + b, corresponding to red branch in Figure 6.3 are shown. The value c* is equal to (c1 + c2) / 2. The average values are highlighted by stars.

Table 6.2: The results of gene clustering

No. of gnathostome cluster	No. of agnathan sequences	No. of phylogenetic tree
4	4	1
4	3	1
4	2	3
4	1	7
3	4	4
3	3	11
3	2	23
3	1	53
2	4	16
2	3	40
2	2	113
2	1	283

family was four under the assumption of the 2R WGD and no gene losses. However, gene losses occurred after the duplication. If three paralogs were lost in gene families, these gene families were difficult to assign the timing of duplication. I thus selected the gene families that retained 2-4 gnathostome clusters and 1-4 agnathan sequences. In the result of classification, groups retained 4 or 3 gnathostome clusters and agnathan sequences was small number. In contrast, the number of groups retained 2 or 1 gnathostome clusters and agnathan sequences was large. Because the number of trees with one agnathan (A) and two gnathostome (G) clusters was largest, these trees were used for further analysis. There were 127 trees with one agnathan (A) and two gnathostome (G) clusters.

The gene loss sometimes covers true gene duplication histories. Under the assumption of no gene loss, two different topologies, ((G,G),A) and ((A,G),G), indicate duplication after the agnathan divergence (**Figure 6.3A**) and before the agnathan divergence (**Figure 6.3B**), respectively. However, the ((G,G),A) topology is compatible with the case in which a genome duplication occurred before the agnathan/gnathostome divergence, if we assume that the two gnathan genes are lost after the duplication. To examine the possibility of these hidden gene duplication sharing, the internal branch lengths of each tree were inferred. If genome duplications are shared among all

vertebrates, internal branch from the agnathan speciation to the gnathostome divergences will be equal between two different gene topologies (a = b + c* in **Figure 6.3**). On the contrary, if it is because of lamprey gene missing, ((A,A),G) topology shows longer internal branch lengths than topologies (a < b + c* in **Figure 6.3**). In the results of analysis, I found that average internal branch length (0.086) of the topology ((A,A),G) is longer than internal branch length (0.055) of the topology ((A,A),G) ($P < 0.001$, t-test) (**Figure 6.4**). These results suggested that tree topologies that do not shared duplication among all vertebrate may cause by agnathan specific gene losses.

6.4 Discussion

The gene loss cause false orthology in the evolution context. Because the sequence data of agnathan are poor previously, it is difficult to identify agnathan specific gene loss. These difficulty lead to misunderstand the ortholog/paralog relation between agnathan and other vertebrates. However, lamprey genome data were released so that we can access this problem correctly.

In my result of clustering, the number of groups that retained 4 gnathostome clusters and 4 agnathan sequences (i.e. no gene losses after the 2R WGD) was very small. This may be caused by small number of lamprey genes or characteristic feature

of the lamprey genome. Although each vertebrate genome usually have approximately 20,000 genes, the gene number of lamprey in the database is only 11,429, suggesting that the half genes in the lamprey genome are not yet annotated. The lamprey gene has characteristic feature such as high GC contents. These features may hamper the gene clustering that facilitate usual vertebrate amino acid sequences.

In this study, I investigated the possibility of improper prediction of gene duplication timing caused by gene loss along the lamprey lineage to compare each internal branch length. The result is preliminary, but suggesting that gene loss violate the true gene duplication history.

CHAPTER 7

General Discussions and Conclusions

The 2R hypothesis has been debated for about forty years. Because genome sequences of many vertebrate and invertebrate were sequenced, we can address this problem by comparative genomic approach. In this thesis, the Hox cluster, that is the symbol of the 2R WGD, is analyzed in detail. Then, I found that these clusters keep many CNSs within non-coding regions (**Chapter 2**) and show ((A,B)(C,D)) symmetry tree topology (**Chapter 3**). In addition, paralogous CNSs derived from the 2R WGD were identified (**Chapter 4**). The proper relative timing of the 2R WGD was estimated by using lamprey gene sets (**Chapter 5** and **6**).

The duplication histories of vertebrate Hox clusters have been discussed since the discovery of them. However, several different hypothesis are proposed, such as ((A,B),(C,D)) or (B,(A,(C,D))), because of short alignment length, taxon sampling and different method of phylogenetic tree reconstruction. Recently, early divergent vertebrate Hox clusters were sequenced (Ravi et al. 2009; Amemiya et al. 2010; Bernard-Samain et al. 2010). These improvement of data makes it possible to estimate precious duplication histories of vertebrate Hox clusters. By using these data, the

((A,B),(C,D)) tree topology was estimated in this study (**Chapter 3**). The distribution of paralogous CNSs within the Hox clusters also suggests this duplication history (**Chapter 2**). The total number of both HoxA/HoxB pair and HoxC/HoxD pair is four out of eight. These clusters share more number of DP CNSs than other pair of clusters. From now, sequence data will be improved more and more, so that we can estimate more precious duplication timing of this clusters.

Our genome-wide survey showed paralogous CNSs derived from the 2R WGD (**Chapter 4**). These conserved sequences may function as enhancers. However, why paralogous CNSs are highly conserved between paralogous loci is unknown. Because paralogous CNSs are frequently located near the coding region encoded transcription factors expressed in brain and/or neural system, they have other unknown functions related to these expression patterns. Recently, the 3D conformation of genome, especially Hox cluster, in the nucleus of the cell are of particular interest (Lanctôt et al. 2007; Ferraiuolo et al. 2010; Noordermeer et al. 2011). The functions of paralogous CNSs may be derived from these 3D conformational information. Otherwise, they may encode non-coding RNA. The further analysis is required to solve this problem.

Whether jawless vertebrate share the 2R WGD with jawed vertebrate is most important question after the proof of the 2R WGD. To dissect this question, lamprey

data generated by second generation sequencers and in public database were used. Although these data do not include the complete gene set of lamprey genome, the results of phylogenetic analysis supports that all vertebrate share the 2R WGD events (**Chapter 5** and **6**). Especially chapter 6, newly released gene data were used. These genes were not used in previous studies (**Figure 7.1**). Because the tree topologies shared only one genome duplication sometimes showed the low bootstrap values, these may be misunderstanding. The cutoff bootstrap value is 50% in Escriva et al. (2002). Otherwise, these topologies may be caused by gene losses. These results suggest that the 2R WGD events are deeply linked to vertebrate specific features, such as limb formation, brain complexity and jaw morphology. Next, we should show clear relation between genome evolution and morphological evolution.

Because the 2R WGD are very ancient events, phylogenetic signals are very weak. This feature causes the difficulty of analysis. In contrast, recent lineage specific WGD events are easy to analyze. The lineage specific WGD are reported in vertebrates. Amphibian and teleost fish genomes show frequent lineage specific WGD event, especially. These lineage specific WGD events are good models for the study of genome evolution after the 2R WGD. Some researchers already start this kinds of analysis, by comparing *Xenopus tropicalis* with *X. laevis* (Sémon et al. 2008). Massive genome data

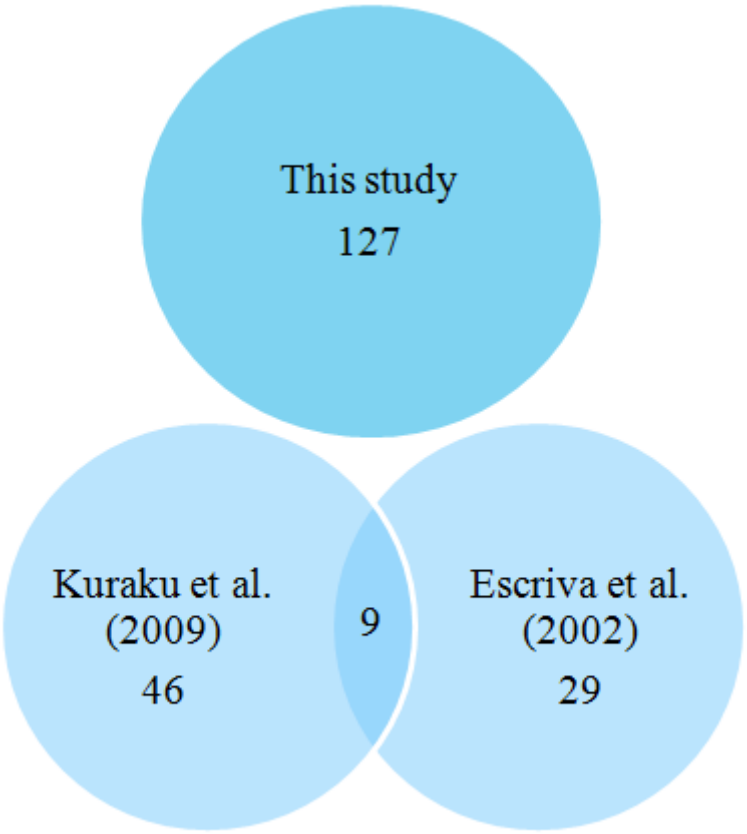


Figure 7.1: Comparison of gene families used for the estimation of relative timing of the 2R WGD

Gene families used Kuraku et al. (2009), Escriva et al. (2002), and this study are compared. Kuraku et al. (2009) shared 9 gene families with Escriva et al. (2002). However, this study did not share gene families with other researches.

of vertebrate experienced lineage specific WGD will be generated by next generation sequencers. We expect that these data show the relation between phenotypic feature and genome evolution after the WGD event more clear.

It is said that there is no correlation between genome conservation and functions in vertebrate (Pennacchio et al. 2010). For example, highly conserved non-coding region showed no enhancer function (Ahituv et al. 2007; Visel et al. 2008) and majority of transcription factor binding region validated by ChIP-seq experiment showed no conservation (Schmidt et al. 2010). Because ChIP-seq experiment can tell the functional regions of only specific timing and regions, it is not sure that other non-coding regions have no enhancer activity. However, particular highly conserved region have no enhancer function, so that they may have other unknown functions. These unknown functions are difficult to detect available through method of experiment or computational analysis. We should consider these unknown functions.

In this thesis, I inferred the possible roles of the 2R WGD in the vertebrate evolution by using available genome data. I detected massive paralogous CNSs and reconstructed duplication histories of each gene family. These studies will help a further understanding of the 2R WGD events.

References

Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM. 2007. Deletion of ultraconserved elements yields viable mice. *PLoS Biol.* **5**:e234.

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

Al-Shahrour F, Minguez P, Tárraga J, Medina I, Alloza E, Montaner D, Dopazo J. 2007. FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Res.* **35**:W91-W96

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

Amemiya CT, Powers TP, Prohaska SJ, Grimwood J, Schmutz J, Dickson M, Miyake T, Schoenborn MA, Myers RM, Ruddle FH, Stadler PF. 2010. Complete HOX cluster characterization of the coelacanth provides further evidence for slow evolution of its genome. *Proc Natl Acad Sci U S A.* **107**:3622-7.

Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL, Westerfield M, Ekker M, Postlethwait JH. 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* **282**:1711-1714

Amores A, Suzuki T, Yan YL, Pomeroy J, Singer A, Amemiya C, Postlethwait JH. 2004. Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. *Genome Res.* **14**:1-10

Aronowicz J, Lowe CJ. 2006. Hox gene expression in the hemichordate Saccoglossus kowalevskii and the evolution of deuterostome nervous systems. *Integr Comp Biol.* **46**:890-901.

Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, Delaney A, Griffith

M, Hickenbotham M, Magrini V, Mardis ER, Sadar MD, Siddiqui AS, Marra MA, Jones SJ. 2006. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7**:246.

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

Becker TS, Lenhard B. 2007. The random versus fragile breakage models of chromosome evolution: a matter of resolution. *Mol Genet Genomics* **278**:487-491

Bogart JP. 1967. Chromosomes of the South American amphibian family Ceratophidae with a reconsideration of taxonomic status of Odontophrynus americanus. *Can. J. Genet. Cytol.* **9**:531-542

Bryant D, Moulton V. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol.* **21**:255-65.

Cameron RA, Rowen L, Nesbitt R, Bloom S, Rast JP, Berney K, Arenas-Mena C, Martinez P, Lucas S, Richardson PM, Davidson EH, Peterson KJ, Hood L. 2006. Unusual gene order and organization of the sea urchin hox cluster. *J Exp Zool B Mol Dev Evol.* **306**:45-58.

Carroll SB. 2001. Chance and necessity: the evolution of morphological complexity and diversity. *Nature* **409**:1102-1109

Chiu CH, Amemiya C, Dewar K, Kim CB, Ruddle F, Wagner GP. 2002. Molecular evolution of the HoxA cluster in three major gnathostome lineages. *Proc. Natl. Acad. Sci. U S A* **99**:5492–5497

Christoffels A, Koh EG, Chia JM, Brenner S, Aparicio S, Venkatesh B. 2004. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol. Biol. Evol.* **21**:1146-1151

Cohn MJ, Tickle C. 1999. Developmental basis of limblessness and axial patterning in snakes. *Nature.* **399**:474-9.

Cox MP, Peterson DA, Biggs PJ. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* **11**:485.

Darwin C. 1859. The Origin of Species by Means of Natural Selection. *John Murry, London*

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3**:1700-1708

de la Calle-Mustienes E, Feijóo CG, Manzanares M, Tena JJ, Rodríguez-Seguel E, Letizia A, Allende ML, Gómez-Skarmeta JL. 2005. A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res.* **15**:1061-72

DeLuna A, Vetsigian K, Shoresh N, Hegreness M, Colón-González M, Chao S, Kishony R. 2008 Exposing the fitness contribution of duplicated genes. *Nature Genet.* **40**: 676–681

Dinger ME, Amaral PP, Mercer TR, Pang KC, Bruce SJ, Gardiner BB, Askarian-Amiri ME, Ru K, Soldà G, Simons C, Sunkin SM, Crowe ML, Grimmond SM, Perkins AC, Mattick JS. 2008. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome. Res.* **18**:1433-1445

Donoghue PCJ, Purnell MA. 2005. Genome duplication, extinction and vertebrate evolution. *Trends. Ecol. Evol.* **20**:312-319

Dubrulle J and Pourquié O. 2004. Coupling segmentation to axis formation. *Development* **131**:5783-5793

Durand D, Halldórsson BV, Vernot B. 2006. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol.* **13**:320-35.

Escriva H, Manzon L, Youson J, Laudet V. 2002. Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. *Mol Biol Evol.* **19**:1440-1450

Ezawa K, Ikeo K, Gojobori T, Saitou N. 2010. Evolutionary pattern of gene homogenization between primate-specific paralogs after human and macaque speciation using the 4-2-4 method. *Mol Biol Evol.* **27**:2152-71.

Feiner N, Ericsson R, Meyer A, Kuraku S. 2011. Revisiting the origin of the vertebrate

Hox14 by including its relict sarcopterygian members. *J Exp Zool B Mol Dev Evol.* **316**:515-25.

Ferraiuolo MA, Rousseau M, Miyamoto C, Shenker S, Wang XQ, Nadler M, Blanchette M, Dostie J. 2010. The three-dimensional architecture of Hox cluster silencing. *Nucleic Acids Res.* **38**:7472-84.

Ferrier DE, Minguillon C, Holland PW, Garcia-Fernàndez J. 2000. The amphioxus Hox cluster: Deuterostome posterior flexibility and Hox14. *Evol. Dev.* **2**:284–293

Ferrier DE. 2004. Hox genes: Did the vertebrate ancestor have a Hox14? *Curr Biol.* **14**:R210-1.

Fitch WM. 1971. Towards defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.* **20**:406-416

Force A, Lynch M, Pickett FB, Amores A, Yan YL. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**:1531-1545

Force A, Amores A, Postlethwait JH. 2002. Hox cluster organization in the jawless vertebrate Petromyzon marinus. *J Exp Zool.* **294**:30-46.

Fried C, Prohaska SJ, Stadler PF. 2003. Independent Hox-cluster duplications in lampreys. *J Exp Zool B Mol Dev Evol.* **299**:18-25.

Gallardo MH, Bickham JW, Honeycutt RL, Ojeda RA, Köhler N. 1999. Discovery of tetraploidy in a mammal. *Nature* **401**:341

Gallardo MH, Kausel G, Jimenez A, Bacquet C, Gonzalez C, Figueroa J, Köhler N, Ojeda R. 2004. Whole-genome duplications in South American desert rodents (Octodontidae). *Biol. J. Linn. Soc.* **82**:443-451

Gallardo MH, Gonzalez CA, Cebrian I. 2006. Molecular cytogenetic and allotetraploidy in the red vizcacha rat, Tympanoctomys barrerae (Rodentia, Octodontidae). *Genomics* **88**: 214-221

Garcia-Fernández J. 2005. The genesis and evolution of homeobox gene clusters. *Nat Rev Genet.* **6**:881–892

Gibson TJ, Spring J. 2000. Evidence in favour of ancient octaploidy in the vertebrate genomes. *Biochem. Soc. Trans.* **28**:259-264

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

Gregory TR. 2005. The evolution of the genome. *Burlington: Elsevier Academic Press*

Guan Y, Dunham MJ, Troyanskaya OG. 2007. Functional analysis of gene duplications in *Saccharomyces cerevisiae*. *Genetics* **175**:933–943

Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol.* **8**: R209

Hall BK. 2000. The neural crest as a fourth germ layer and vertebrates as quadroblastic not triploblastic. Evol Dev. 2:3-5.

Hancock JM, Shaw PJ, Bonneton F, Dover GA. 1999. High sequence turnover in the regulatory regions of the developmental gene hunchback in insects**.** *Mol Biol Evol.* **284**:1083-1094

Holland LZ, Albalat R, Azumi K, Benito-Gutiérrez E, Blow MJ, Bronner-Fraser M, Brunet F, Butts T, Candiani S, Dishaw LJ, Ferrier DE, Garcia-Fernàndez J, Gibson-Brown JJ, Gissi C, Godzik A, Hallböök F, Hirose D, Hosomichi K, Ikuta T, Inoko H, Kasahara M, Kasamatsu J, Kawashima T, Kimura A, Kobayashi M, Kozmik Z, Kubokawa K, Laudet V, Litman GW, McHardy AC, Meulemans D, Nonaka M, Olinski RP, Pancer Z, Pennacchio LA, Pestarino M, Rast JP, Rigoutsos I, Robinson-Rechavi M, Roch G, Saiga H, Sasakura Y, Satake M, Satou Y, Schubert M, Sherwood N, Shiina T, Takatori N, Tello J, Vopalensky P, Wada S, Xu A, Ye Y, Yoshida K, Yoshizaki F, Yu JK, Zhang Q, Zmasek CM, de Jong PJ, Osoegawa K, Putnam NH, Rokhsar DS, Satoh N, Holland PW. 2008. The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res.* **18**:1100-11.

Holland ND, Chen J. 2001. Origin and early evolution of the vertebrates: new insights from advances in molecular biology, anatomy, and palaeontology. Bioessays. 23:142-51.

Holland PWH. 1996. Molecular biology of lancelets: insights into development and evolution. Israel Journal of Zoology 42:S247-S272

Holland PW, Garcia- Fernàndez J, Williams NA, Sidow A. 1994. Gene duplications and the origins of vertebrate development. *Dev Suppl.* **120** (SUPPL.):125-133

Hubbs CI, Potter IC. 1971. Distribution, phylogeny and taxonomy. In: Hardisty MW, Potter IC (Eds.), The Biology of Lampreys Vol. I. *Academic Press, London*

Hueber SD, Weiller GF, Djordjevic MA, Frickey T. 2010. Improving Hox protein classification across the major model organisms. *PLoS One.* **5**:e10820.

Hufton AL, Mathia S, Braun H, Georgi U, Lehrach H, Vingron M, Poustka AJ, Panopoulou G. 2009. Deeply conserved chordate noncoding sequences preserve genome synteny but not drive gene duplication retention. *Genome Res.* **19**:2036-51

Hughes AL, da Silva J, Friedman R. 2001. Ancient genome duplications did not structure the human Hox-bearing chromosomes. *Genome Res*. **11**:771-780

Husband B. 2000. Constraints on polyploid evolution: a test of the minority cytotype exclusion principle. *Proc. Roy. Soc. Lond. Ser. B.* **267**:217-223

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol.* **23**:254-67.

Juan AH, Ruddle FH. 2003. Enhancer timing of Hox gene expression: deletion of the endogenous Hoxc8 early enhancer. *Development* **130**:4823-4834

Kappen C, Ruddle FH. 1993. Evolution of a regulatory gene family: HOM/HOX genes. *Curr Opin Genet Dev.* **3**:931-8.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**:3059-3066.

Kaufman TC, Seeger MA, Olsen G. 1990. Molecular and genetic organization of the antennapedia gene complex of Drosophila melanogaster. *Adv. Genet.* **27**:309–362

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient

genome duplication in the yeast Saccharomyces cerevisiae. *Nature.* **428**:617-24.

Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, Smedley D, Otgaar D, Greyling G, Jongeneel CV, McCarthy MI, Hide T, Hide W. 2003. eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res.* **13**:1222-30.

Kikuta H, Fredman D, Rinkwitz S, Lenhard B, Becker TS. 2007. Retroviral enhancer detection insertions in zebrafish combined with comparative genomics reveal genomic regulatory blocks - a fundamental feature of vertebrate genomes. *Genome Biol.* **8**(Suppl 1):S4

Kim CB, Amemiya C, Bailey W, Kawasaki K, Mezey J, Miller W, Minoshima S, Shimizu N, Wagner GP, Ruddle F 2000. Hox cluster genomics in the horn shark, Heterodontus francisci. *Proc Natl Acad Sci U S A.* **97**:1655–1660

King BL, Gillis JA, Carlisle HR, Dahn RD. 2011. A natural deletion of the HoxC cluster in elasmobranch fishes. *Science.* **334**:1517.

Kumar S, Blaxter ML. 2010. Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics* **11**:571.

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

Kuraku S. 2008. Insights into cyclostome phylogenomics: pre-2R or post-2R? *Zool Sci.* **25**:960-968

Kuraku S, Takio Y, Tamura K, Aono H, Meyer A, Kuratani S. 2008. Noncanonical role of Hox14 revealed by its expression patterns in lamprey and shark. *Proc Natl Acad Sci U S A.* **105**:6679-83.

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

Kupka E. 1948. Chromosomale versheidenheiten bei schweizerischen Coregonen. *Rev. Suisse Zool.* **55**:285-293

Lanctôt C, Kaspar C, Cremer T. 2007. Positioning of the mouse Hox gene clusters in the nuclei of developing embryos and differentiating embryoid bodies. *Exp Cell Res.* **313**:1449-59.

Larhammar D, Lundin LG, Hallböök F. 2002. The human Hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. *Genome Res.* **12**:1910-20.

Le Comber SC, Smith C. 2004. Polyploidy in fishes: patterns and process. *Biol. J. Linn. Soc.* **82**:431-442

Lee AP, Kerk SY, Tan YY, Brenner S, Venkatesh B. 2011. Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. *Mol Biol Evol.* **28**:1205-1215

Lehoczky JA, Williams ME, Innis JW. 2004. Conserved expression domains for genes upstream and within the HoxA and HoxD clusters suggests a long-range enhancer existed before cluster duplication. *Evol. Dev.* **6**:423-430

Lehoczky JA, Innis JW. 2008. BAC transgenic analysis reveals enhancers sufficient for Hoxa13 and neighborhood gene expression in mouse embryonic distal limbs and genital bud. *Evol. Dev.* **10**:421-423

Lemons D, McGinnis W. 2006. Genomic evolution of Hox gene clusters. *Science* **313**:1918-1922

Levin DA. 2002. The role of chromosomal change in plant evolution. *Oxford: Oxford University Press*

Lewis EB. 1978 A gene complex controlling segmentation in Drosophila. *Nature* **276**:565–570

Lewis WH. 1980. Polyploidy: biological relevance. *New York: Plenum Press*

Liang D, Wu R, Geng J, Wang C, Zhang P. 2011. A general scenario of Hox gene inventory variation among major sarcopterygian lineages. *BMC Evol Biol.* **11**:25.

Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M. 2005. Functional evolution of a cis-regulatory module. *PLoS Biol.* **3**:588-598

Lundin LG. 1993. Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics* **16**:1-19

Lundin LG, Larhammar D, Hallböök F. 2003. Numerous groups of chromosomal regional paralogies strongly indicate two genome doublings at the root of the vertebrates. *J Struct Funct Genomics.* **3**: 53-63.

Lynch VJ, Wagner GP. 2009. Multiple chromosomal rearrangements structured the ancestral vertebrate Hox-bearing protochromosomes. *PLoS Genet.* **5**:e1000349.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proc. Natl Acad. Sci. USA* **102**:5454–5459

Mable BK, Alexandrou MA, Taylor MI. 2011. Genome duplication in amphibians and fish: an extended synthesis. *J. Zool.* **284**:151-182

Mainguy G, In der Rieden PM, Berezikov E, Woltering JM, Plasterk RH, Durston AJ. 2003. A position-dependent organization of retinoid response elements is conserved in the vertebrate *Hox* clusters. *Trends Genet.* **19**:476-479

Mainguy G, Koster J, Woltering J, Jansen H, Durston A. 2007. Extensive polycistronism and antisense transcription in the mammalian Hox clusters. *PLoS One* **2**:e356

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**:376-380

Matsunami M, Sumiyama K, Saitou N. 2010. Evolution of conserved non-coding sequences within the vertbrate hox clusters through the two-round whole duplications revealed by phylogenetic footprinting analysis. *J. Mol. Evol.* **71**:427-436

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

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

McGinnis W, Krumlauf R. 1992. Homeobox genes and axial patterning. *Cell* **68**:283–302

Metzker ML. 2010. Sequencing technologies - the next generation. *Nat Rev Genet.* **11**:31-46

Moore RC, Purugganan MD. 2005. The evolutionary dynamics of plant duplicate genes. *Curr Opin Plant Biol.* **8**:122-8.

Morrison A, Ariza-McNaughton L, Gould A, Featherstone M, Krumlauf R. 1997. HOXD4 and regulation of the group 4 paralog genes. *Development* **124**:3135-3146

Murphy WJ, Pevzner PA, O'Brien SJ. 2004. Mammalian phylogenomics comes of age. *Trends Genet.* **20**:631-639

Nakatani Y, Takeda H, Kohara Y, Morishita S. 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.* **17**:1254-1265

Naruse K, Tanaka M, Mita K, Shima A, Postlethwait J, Mitani H. 2004. A medaka gene map: the trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping *Genome Res.* **14**:820-828

Noordermeer D, Leleu M, Splinter E, Rougemont J, De Laat W, Duboule D. 2011. The dynamic architecture of Hox gene clusters. *Science.* **334**:222-5.

Ohno S. 1970. Evolution by gene duplication. *New York: Springer Verlag.*

Ørom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q, Guigo R, Shiekhattar R. 2010. Long noncoding RNAs with enhancer-like function in human cells. *Cell* **43**:46-58

Otto SP, Whiton J. 2000. Polyploidy: incidence and evolution. *J. Anim. Ecol.* **76**:1053-1061

Oulion S, Debiais-Thibaud M, d'Aubenton-Carafa Y, Thermes C, Da Silva C, Bernard-Samain S, Gavory F, Wincker P, Mazan S, Casane D. 2010. Evolution of Hox gene clusters in gnathostomes: insights from a survey of a shark (Scyliorhinus canicula) transcriptome. *Mol Biol Evol.* **27**:2829-38.

Ozsolak F, Milos PM. 2011. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet.* **12**:87-98

Papp B, Pál C, Hurst LD. 2003 Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**:194-197

Pascual-Anaya J, D'Aniello S, Garcia-Fernàndez J. 2008. Unexpected number of conserved noncoding regions within the ancestral chordate Hox cluster. *Dev. Genes Evo.* **218**:591-597

Pearson JC, Lemons D, McGinnis W. 2005. Modulating Hox gene functions during animal body patterning. *Nat. Rev. Genet.* **6**:893-904

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

Peterson KJ. 2004. Isolation of Hox and Parahox genes in the hemichordate Ptychodera flava and the evolution of deuterostome Hox genes. *Mol Phylogenet Evol.* **31**:1208-15.

Pennacchio LA, Visel A. 2010. Limits of sequence and functional conservation. *Nat Genet.* **42**:557-8.

Prohaska SJ, Fried C, Flamm C, Wagner GP, Stadler PF. 2004. Surveying phylogenetic footprints in large gene clusters: application to Hox cluster duplications. *Mol. Phylogenet. Evol.* **31**:581–604

Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK, Benito-Gutie´rrez E, Dubchak I,

Garcia-Fernàndez J, Grigoriev IV, Horton AC, de Jong PJ, Jurka J, Kapitonov V, Kohara Y, Kuroki Y, Lindquist E, Lucas S, Osoegawa K, Pennacchio LA, Salamov AA, Satou Y, Sauka-Spengler T, Schmutz J, Shin-I T, Toyoda A, Gibson-Brown JJ, Bronner-Fraser M, Fujiyama A, Holland LZ, Holland PWH, Satoh N, Rokhsar DS. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**:1064-1071

Qiu H, Hildebrand F, Kuraku S, Meyer A. 2011. Unresolved orthology and peculiar coding sequence properties of lamprey genes: the KCNA gene family as test case. *BMC Genomics.* **12**:325

Ravi V, Lam K, Tay BH, Tay A, Brenner S, Venkatesh B. 2009. Elephant shark (Callorhinchus milii) provides insights into the evolution of Hox gene clusters in gnathostomes. *Proc Natl Acad Sci U S A.* **106**:16327-32.

Ray P, Shringarpure S, Kolar M, Xing EP. 2008. CSMET: Comparative genomic motif detection via multi-resolution phylogenetic shadowing. *PLoS Comput. Biol.* **4**:e1000090

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*. **129**:1311-23.

Ruddle FH, Bartels JL, Bentley KL, Kappen C, Murtha MT, Pendleton JW. 1994. Evolution of Hox genes. *Annu Rev Genet.* **28**:423-42.

Ruddle FH, Bentley KL, Murtha MT, Risch N. 1994. Gene loss and gain in the evolution of the vertebrates. *Dev.* (Suppl.) 155-161

Ruse M. 2003. Is evolution a secular religion? *Science* **299**:1523–1524

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* **4**:406-25

Santini S, Boore JL, Meyer A. 2003. Evolutionary conservation of regulatory elements in vertebrate Hox gene clusters. *Genome Res.* **13**:1111-1122

Sato Y, Hashiguchi Y, Nishida M. 2009. Temporal pattern of loss/persistence of duplicate genes involved in signal transduction and metabolic pathways after

teleost-specific genome duplication. *BMC Evol Biol.* **9**:127

Saez FA, Brum-Zorrilla N. 1966. Karyotype variation in some species of the genus Odontophrynus (Amphibia-Anura). *Caryologia* **19**:7-24

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, Talianidis I, Flicek P, Odom DT. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*. **328**:1036-40.

Sémon M, Wolfe KH. 2008. Preferential subfunctionalization of slow-evolving genes after allopolyploidization in Xenopus laevis. *Proc Natl Acad Sci U S A.* **105**:8333-8.

Seo HC, Edvardsen RB, Maeland AD, Bjordal M, Jensen MF, Hansen A, Flaat M, Weissenbach J, Lehrach H, Wincker P, Reinhardt R, Chourrout D. 2004. Hox cluster disintegration with persistent anteroposterior order of expression in Oikopleura dioica. *Nature* **431**:67-71.

Seoighe C, Wolfe KH. 1999. Yeast genome evolution in the post-genome era. *Curr. Opin. Microbiol.* **2**:548–554

Soltis DE, Buggs RJ, Doyle JJ, Soltis PS. 2010. What we still don't know about polyploidy. *Taxon* **2**:1-17

Spagnuolo A, Ristoratore F, Di Gregorio A, Aniello F, Branno M, Di Lauro R. 2003. Unusual number and genomic organization of Hox genes in the tunicate Ciona intestinalis. *Gene.* **309**:71-9.

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

Spitz F, Gonzalez F, Peichel C, Vogt TF, Duboule D, Zákány J. 2001. Large scale transgenic and cluster deletion analysis of the HoxD complex separate an ancestral regulatory module from evolutionary innovations. *Genes Dev* **15**:2209-2214

Svärdson G. 1945. Chromosome studies on Salmonidae. *Rep. Swed. State. Inst. Fresh. Fish. Res.* **23**:1-151

Svartman M, Stone G, Stanyon R. 2005. Molecular cytogenetics discards polyploidy in

mammals. *Genomics* **85**:425-430

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* **28**:2731-9.

Thomas-Chollier M, Ledent V, Leyns L, Vervoort M. 2010. A non-tree-based comprehensive study of metazoan Hox and ParaHox genes prompts new insights into their origin and evolution. *BMC Evol Biol.* **10**:73.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**:4673-4680

Torres TT, Metta M, Ottenwälder B, Schlötterer C. 2008. Gene expression profiling by massively parallel sequencing. *Genome Res.* **18**:172-177

Toth AL, Varala K, Newman TC, Miguez FE, Hutchison SK, Willoughby DA, Simons JF, Egholm M, Hunt JH, Hudson ME, Robinson GE. 2007. Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. *Science* **318**:441-444

Uzzell TM Jr. 1964. Relations of the diploid and triploid species of the Ambystoma jeffersonianum complex (Amphibia, Caudata). *Copeia* **1964**:257-300

Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y. 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc. Natl Acad. Sci. U S A* **101**:1638–5443

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

Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**:D88-92.

Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I,

Afzal V, Rubin EM, Pennacchio LA. 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet.* **40**:158-60.

Volff JN. 2005. Genome evolution and biodiversity in teleost fish. *Heredity* **94**:280-294

Wada H Escriva H, Zhang S, Laudet V. 2006. Conserved RARE localization in amphioxus Hox clusters and implications for Hox code evolution in the vertebrate neural crest. *Dev. Dyn.* **235**:1522-1531

Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449**:54–61

Weber AP, Weber KL, Carr K, Wilkerson C, Ohlrogge JB. 2007. Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiol.* **144**:32-42

Werauch MT, Hughes TR. 2010. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet.* **26**:66-74

Woods IG, Kelly PD, Chu F, Ngo-Hazelett P, Yan YL, Huang H, Postlethwait JH, Talbot WS. 2000. A comparative map of the zebrafish genome. *Genome Res.* **10**:1903-1914

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards YJ, Cooke JE, Elgar G. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**:116-130

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

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci.* **13**:555-6

Yang Z. 2007. PAML 4: phylogenic analysis by maximum likelihood. *Mol Biol Evol* **24**:1586-1591

Yamazaki Y, Goto A. 1998. Genetic structure and differentiation of four Lethenteron

taxa from the Far East, deduced from allozyme analysis. *Env. Biol. Fish.* **52**: 149-161

Yamazaki Y, Yokoyama R, Nishida M, Goto A. 2006. Taxonomy and molecular phylogeny of Lethenteron lampreys in eastern Eurasia. *J. Fish Biol.* **68**:251-269.

Yekta S, Shih IH, Bartel DP. 2004. MicroRNA-directed cleavage of HOXB8 mRNA. *Science* **304**:594-596

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**:821-829

Zhang J, Nei M. 1996. Evolution of Antennapedia-class homeobox genes. *Genetics.* **142**:295-303.

Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7**:203-214

Appendices

Figures A2.1-4 and **Tables A2.1-2** are avairable on line (http://www.springerlink.com/content/r631411563668m66/239_2010_Article_9396_ESM.html). Figures A2.1-4 nad Tables A2.1-2 correspond to Supplementary Figures 1-4 and Supplementary Tables 1-2 of Matsunami et al. (2010), respectively.

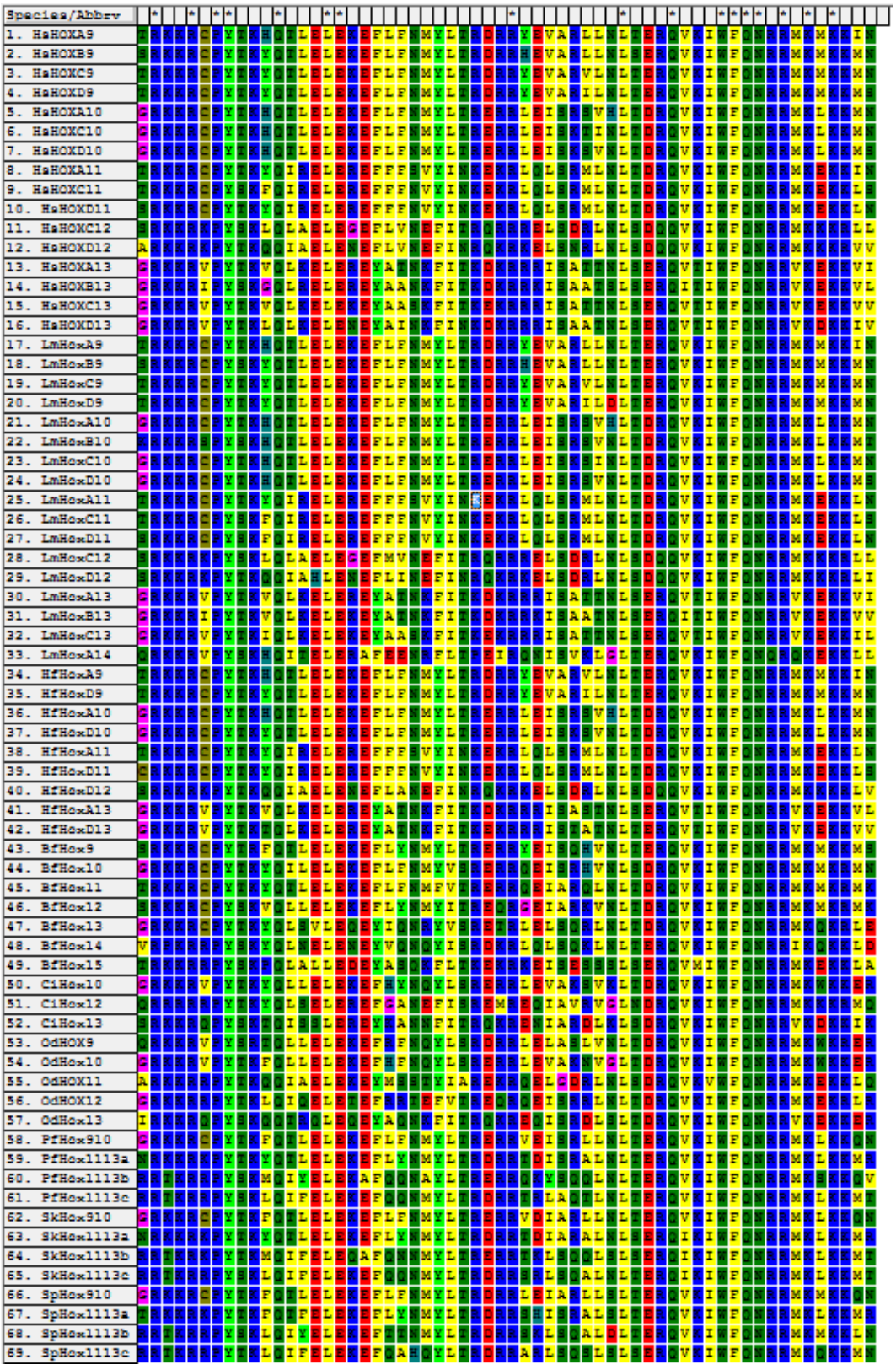


Figure A3.1: Multiple alignments of deuterostomate posterior Hox genes

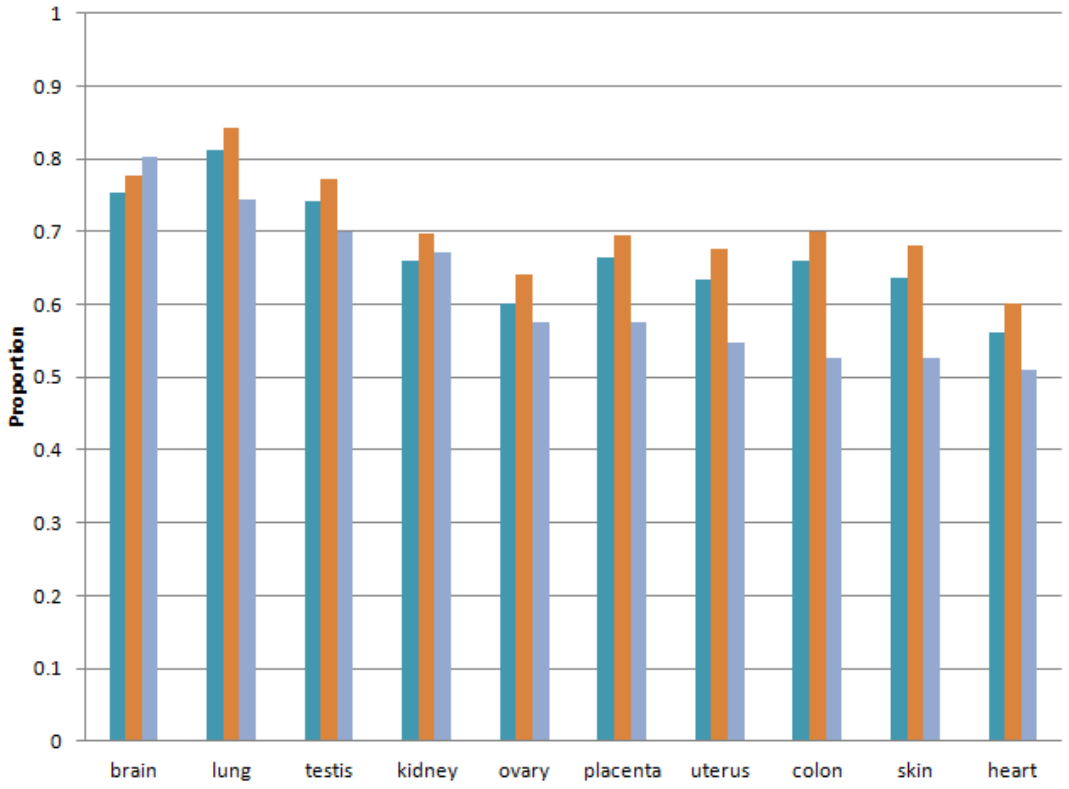


Figure A4.1: Multiple alignments of each amphiCNS

The relative proportion of genes in each anatomical site was calculated. The portion is the gene number expressed each site / all gene number. All human genes (cyan), all genes derived from the 2R WGD (orange) and paralogous CNSs harboring genes (light blue) were compared.

Table A4.1: Feature of each paralogous CNS

Name	Chr	Start	End	Location	Harboring Gene	Relative Location	ParaCNS
SB0CNS30	1	3057640	3058117	Intron	PRDM16	Intron	SB0SB22DP1
SB22CNS8	3	169193752	169194457	Intergenic	MECOM	Intron	SB0SB22DP1
SB0CNS138	1	3086383	3086621	Intron	PRDM16	Intron	SB0SB22DP2
SB22CNS71	3	169152977	169153473	Intergenic	MECOM	Intron	SB0SB22DP2
SB0CNS170	1	1327882	1328160	Intergenic	CCNL2	Intron	SB0SB22DP3
SB22CNS399	3	156869421	156869694	Intergenic	CCNL1	Intron	SB0SB22DP3
SB22CNS1480	3	181410474	181410795	Intergenic	GNB4	Upstream	SB0SB22SB43DP2
SB43CNS15	7	101892998	101893150	3UTR	GNB2	Downstream	SB0SB22SB43DP2
SB113CNS242	11	122017197	122017327	Intergenic	BACE1	Upstream	SB102SB113DP1
SB102CNS53	21	17912029	17912312	Intergenic	BACE2	Upstream	SB102SB113DP1
SB11CNS29	2	104648840	104649405	Intergenic	SEPT10	Downstream	SB11SB108DP4
SB108CNS392	X	128085221	128085647	Intergenic	SEPT6	Upstream	SB11SB108DP4
SB11CNS4	2	105254515	105255294	Intergenic	SH3RF3	Downstream	SB11SB25DP2
SB25CNS214	4	145277467	145277870	Intergenic	SORBS2	Downstream	SB11SB25DP2
SB11CNS163	2	104935832	104936116	Intergenic	SLC9A2	Downstream	SB11SB26DP1
SB26CNS48	5	4014165	4014410	Intergenic	SLC9A3	Downstream	SB11SB26DP1
SB11CNS6	2	105047984	105048689	Intergenic	POU3F3	Upstream	SB11SB35DP1
SB35CNS202	6	98566292	98566728	Intergenic	POU3F2	Upstream	SB11SB35DP1
SB11CNS11	2	104061618	104062163	Intergenic	POU3F3	Upstream	SB11SB35DP4
SB35CNS137	6	133916374	133916844	Intergenic	POU3F2	Downstream	SB11SB35DP4
SB12CNS228	2	182243019	182243526	Intergenic	SLC4A10	Downstream	SB12SB24DP6
SB24CNS58	4	84868501	84868836	Intergenic	SLC4A4	Downstream	SB12SB24DP6
SB12CNS1127	2	201724959	201725201	Intron	CLK1	Intron	SB12SB28DP4
SB28CNS163	5	178044436	178044824	Intron	CLK4	Intron	SB12SB28DP4
SB12CNS2094	2	177013650	177013834	Intergenic	HOXA4	Upstream	SB12SB37DP14
SB37CNS27	7	27172953	27173699	Intergenic	HOXD4	Upstream	SB12SB37DP14
SB12CNS31	2	145257544	145258406	Intron	ZEB2	Intron	SB12SB37DP2
SB37CNS19	7	26497947	26498556	Intergenic	KIAA0087	Upstream	SB12SB37DP2
SB12CNS59	2	176938088	176938725	Intergenic	EVX2	Downstream	SB12SB37DP3
SB37CNS31	7	27290645	27291137	Intergenic	EVX1	Downstream	SB12SB37DP3

SB12CNS115	2	176718728	176719289	Intergenic	EVX2	Downstream	SB12SB37DP5
SB37CNS7	7	27591407	27592224	Intron	EVX1	Downstream	SB12SB37DP5
SB37CNS81	7	27288322	27288670	Intergenic	EVX1	Downstream	SB12SB37DP7
SB12CNS161	2	176940364	176940783	Intergenic	EVX2	Downstream	SB12SB37DP7
SB12CNS3004	2	176979102	176979337	Intergenic	RALB	Downstream	SB12SB38DP14
SB38CNS24	7	39537702	39538234	Intergenic	RALA	Upstream	SB12SB38DP14
SB12CNS592	2	207907044	207907421	Intergenic	IHH	Downstream	SB12SB45DP9
SB45CNS17	7	156539742	156540107	Intron	SHH	Downstream	SB12SB45DP9
SB12CNS493	2	146057604	146057977	Intergenic	LASS6	Upstream	SB12SB76DP26
SB76CNS362	15	96093996	96094252	Intergenic	LASS3	Downstream	SB12SB76DP26
SB12CNS959	2	160546183	160546855	Intergenic	CLK1	Downstream	SB12SB76DP5
SB76CNS443	15	60060702	60061346	Intergenic	CLK3	Upstream	SB12SB76DP5
SB12CNS391	2	182551181	182551750	Intergenic	NEUROD1	Upstream	SB12SB82DP1
SB82CNS67	17	37774594	37774901	Intergenic	NEUROD2	Upstream	SB12SB82DP1
SB12CNS29	2	164870539	164871449	Intergenic	ARL5A	Upstream	SB12SB82DP11
SB82CNS54	17	35236960	35237326	Intergenic	ARL5C	Downstream	SB12SB82DP11
SB12CNS1000	2	182418924	182419244	Intron	NEUROD1	Downstream	SB12SB82DP5
SB82CNS112	17	37719178	37719412	Intergenic	NEUROD2	Downstream	SB12SB82DP5
SB13CNS18	2	236963549	236963919	Intergenic	INPP5D	Upstream	SB13SB108DP6
SB108CNS150	X	148019022	148019450	Intergenic	SH2D1A	Downstream	SB13SB108DP6
SB18CNS13	3	71277494	71278110	Intergenic	FOXP1	Intron	SB18SB44DP1
SB44CNS44	7	114058138	114058595	Intergenic	FOXP2	Intron	SB18SB44DP1
SB18CNS25	3	71052587	71053014	Intron	FOXP1	Intron	SB18SB44DP2
SB44CNS9	7	114288766	114289631	Intron	FOXP2	Intron	SB18SB44DP2
SB18CNS48	3	71254138	71254558	Intergenic	FOXP1	Intron	SB18SB44DP4
SB44CNS38	7	114065464	114066000	Intergenic	FOXP2	Intron	SB18SB44DP4
SB18CNS61	3	71009052	71009511	Intron	FOXP1	Intron	SB18SB44DP5
SB44CNS22	7	114328428	114329032	Intron	FOXP2	Intron	SB18SB44DP5
SB18CNS76	3	71154381	71154780	Intron	FOXP1	Intron	SB18SB44DP6
SB44CNS11	7	114209215	114210152	Intergenic	FOXP2	Intron	SB18SB44DP6
SB18CNS92	3	71290627	71291076	Intergenic	FOXP1	Intron	SB18SB44DP7
SB44CNS49	7	114057426	114057837	Intergenic	FOXP2	Intron	SB18SB44DP7
SB1CNS2	1	39875900	39876965	Intron	MACF1	Intron	SB1SB34DP1
SB34CNS7	6	56405775	56406409	Intron	DST	Intron	SB1SB34DP1
SB1CNS26	1	38792406	38792852	Intergenic	POU3F1	Upstream	SB1SB35DP1
SB35CNS4	6	98492276	98493041	Intergenic	POU3F2	Upstream	SB1SB35DP1

SB1CNS43	1	38494889	38495256	Intergenic	POU3F1	Downstream	SB1SB35DP9
SB35CNS484	6	95488673	95489069	Intergenic	POU3F2	Upstream	SB1SB35DP9
SB22CNS1003	3	181328010	181328191	Intergenic	SOX2	Upstream	SB22SB108DP15
SB108CNS319	X	139620668	139621146	Intron	SOX3	Upstream	SB22SB108DP15
SB22CNS141	3	155705336	155705775	Intergenic	PLS1	Downstream	SB22SB108DP21
SB108CNS1399	X	116097171	116097487	Intergenic	PLS3	Downstream	SB22SB108DP21
SB22CNS107	3	180957015	180957518	Intergenic	ACTL6A	Downstream	SB22SB43DP1
SB43CNS24	7	101894082	101894242	3UTR	ACTL6B	Upstream	SB22SB43DP1
SB22CNS51	3	180773807	180774369	Intergenic	SOX2	Upstream	SB22SB72DP3
SB72CNS328	13	112037543	112037917	Intergenic	SOX1	Upstream	SB22SB72DP3
SB22CNS74	3	137412620	137413009	Intergenic	SOX14	Upstream	SB22SB72DP4
SB72CNS409	13	95403021	95403559	Intergenic	SOX21	Upstream	SB22SB72DP4
SB22CNS15	3	136983463	136984034	Intergenic	SOX14	Upstream	SB22SB72DP1
SB72CNS6	13	95618617	95619477	Intergenic	SOX21	Upstream	SB22SB72DP1
SB22CNS1341	3	152164375	152164491	Intron	MBNL1	Intron	SB22SB72DP8
SB72CNS167	13	98008688	98009048	Intron	MBNL2	Intron	SB22SB72DP8
SB23CNS146	4	20529436	20530028	Intron	SLIT2	Intron	SB23SB28DP2
SB28CNS583	5	168195143	168195387	Intron	SLIT3	Intron	SB23SB28DP2
SB23CNS200	4	20481846	20482160	Intron	SLIT2	Intron	SB23SB28DP3
SB28CNS1314	5	168271924	168272063	Intron	SLIT3	Intron	SB23SB28DP3
SB23CNS43	4	17886376	17886851	Intergenic	SLIT2	Upstream	SB23SB59DP24
SB59CNS1488	10	98714243	98714491	Intron	SLIT1	Downstream	SB23SB59DP24
SB24CNS2	4	84700473	84701152	Intergenic	ODZ3	Upstream	SB24SB28DP7
SB28CNS286	5	164336571	164337000	Intergenic	ODZ2	Upstream	SB24SB28DP7
SB24CNS6	4	80826864	80827747	Intergenic	BMP3	Upstream	SB24SB59DP3
SB59CNS74	10	77990349	77990893	Intron	GDF10	Upstream	SB24SB59DP3
SB25CNS808	4	95117176	95117665	Intergenic	NPNT	Upstream	SB25SB107DP2
SB107CNS221	X	36229132	36229741	Intergenic	EGFL6	Downstream	SB25SB107DP2
SB25CNS1	4	103717357	103718307	Intergenic	GPM6A	Downstream	SB25SB108DP1
SB108CNS1777	X	88164024	88164346	Intergenic	PLP1	Upstream	SB25SB108DP1
SB25CNS1583	4	158281297	158281458	Intron	GRIA2	Intron	SB25SB108DP18
SB108CNS34	X	122598965	122599523	Intron	GRIA3	Intron	SB25SB108DP18
SB25CNS1431	4	153535074	153535459	Intergenic	ARSJ	Upstream	SB25SB27DP14
SB27CNS2443	5	115898234	115898583	Intergenic	ARSB	Upstream	SB25SB27DP14
SB25CNS567	4	158282487	158282688	Intron	GRIA2	Intron	SB25SB28SB63TriP1
SB28CNS643	5	153174707	153175025	Intron	GRIA1	Intron	SB25SB28SB63TriP1

SB63CNS52	11	105842387	105842639	Intron	GRIA4	Intron	SB25SB28SB63TriP1
SB25CNS560	4	124467394	124467672	Intergenic	SEC24D	Upstream	SB25SB59DP1
SB59CNS321	10	77697498	77697864	Intron	SEC24C	Downstream	SB25SB59DP1
SB25CNS531	4	148347125	148347508	Intron	SGMS2	Downstream	SB25SB59DP22
SB59CNS1577	10	54148866	54149269	Intergenic	SGMS1	Upstream	SB25SB59DP22
SB25CNS40	4	151453264	151453881	Intron	LRBA	Intron	SB25SB71DP1
SB71CNS3	13	36088920	36089444	Intron	NBEA	Intron	SB25SB71DP1
SB25CNS862	4	151416823	151417013	Intron	LRBA	Intron	SB25SB71DP2
SB71CNS13	13	36105736	36106124	Intron	NBEA	Intron	SB25SB71DP2
SB25CNS87	4	145766758	145768024	Intergenic	POU4F2	Upstream	SB25SB72DP2
SB72CNS1958	13	79424519	79424889	Intergenic	POU4F1	Upstream	SB25SB72DP2
SB25CNS978	4	147364853	147365155	Intron	POU4F2	Upstream	SB25SB72DP25
SB72CNS59	13	79348073	79348631	Intergenic	POU4F1	Upstream	SB25SB72DP25
SB25CNS94	4	182405565	182406068	Intergenic	ING2	Upstream	SB25SB72DP39
SB72CNS227	13	55188976	55189388	Intergenic	ING1	Upstream	SB25SB72DP39
SB25CNS195	4	131943406	131943850	Intergenic	POU4F2	Upstream	SB25SB72DP5
SB72CNS166	13	79168758	79169227	Intergenic	POU4F1	Upstream	SB25SB72DP5
SB26CNS90	5	2829427	2829653	Intergenic	SLC6A18	Downstream	SB26SB68DP1
SB68CNS74	12	102874161	102874522	5UTR	SLC6A15	Upstream	SB26SB68DP1
SB26CNS1	5	3512400	3513166	Intergenic	IRX1	Upstream	SB26SB78DP2
SB78CNS3	16	55223277	55224019	Intergenic	IRX3	Upstream	SB26SB78DP2
SB26CNS2	5	3186909	3187703	Intergenic	IRX1	Upstream	SB26SB78DP3
SB78CNS4	16	54576592	54577367	Intergenic	IRX3	Upstream	SB26SB78DP3
SB26CNS3	5	2112461	2113335	Intergenic	IRX1	Upstream	SB26SB78DP4
SB78CNS73	16	54323658	54324060	Intergenic	IRX3	Upstream	SB26SB78DP4
SB26CNS4	5	3182352	3183124	Intergenic	IRX1	Upstream	SB26SB78DP5
SB78CNS1260	16	54579810	54579978	Intergenic	IRX3	Upstream	SB26SB78DP5
SB26CNS5	5	3197955	3198809	Intergenic	IRX1	Upstream	SB26SB78DP6
SB78CNS259	16	54540560	54540900	Intergenic	IRX3	Upstream	SB26SB78DP6
SB26CNS69	5	2643228	2643590	Intergenic	IRX2	Downstream	SB26SB78DP7
SB78CNS1230	16	54973802	54973940	Intergenic	IRX5	Downstream	SB26SB78DP7
SB27CNS1453	5	68364721	68365021	Intergenic	ARSB	Downstream	SB27SB28DP222
SB28CNS311	5	173988694	173989254	Intergenic	ARSI	Upstream	SB27SB28DP222
SB27CNS9	5	91037493	91038314	Intergenic	MCTP1	Downstream	SB27SB76DP1
SB76CNS393	15	95336270	95336796	Intergenic	MCTP2	Downstream	SB27SB76DP1
SB27CNS1811	5	87899199	87899491	Intergenic	MEF2C	Downstream	SB27SB76DP15

SB76CNS980	15	45508248	45508599	Intergenic	MEF2A	Upstream	SB27SB76DP15
SB27CNS225	5	93230283	93230774	Intron	NR2F1	Downstream	SB27SB76DP20
SB76CNS77	15	97436395	97436985	Intergenic	NR2F2	Downstream	SB27SB76DP20
SB27CNS702	5	92903113	92903513	Intergenic	NR2F1	Upstream	SB27SB76DP31
SB76CNS823	15	96860239	96860432	Intergenic	NR2F2	Upstream	SB27SB76DP31
SB27CNS416	5	93222693	93222959	Intron	NR2F1	Downstream	SB27SB76DP4
SB76CNS348	15	97347915	97348364	Intergenic	NR2F2	Downstream	SB27SB76DP4
SB27CNS1132	5	88672587	88672894	Intergenic	SLC12A2	Downstream	SB27SB78DP29
SB78CNS20	16	51934604	51935279	Intergenic	SLC12A3	Downstream	SB27SB78DP29
SB27CNS842	5	102365073	102365428	3UTR	ST8SIA4	Upstream	SB27SB86DP7
SB86CNS47	18	53750453	53750835	Intergenic	ST8SIA3	Upstream	SB27SB86DP7
SB28CNS355	5	175953716	175954049	3UTR	RNF44	UTR	SB28SB53DP1
SB53CNS23	9	36336378	36336690	3UTR	RNF38	UTR	SB28SB53DP1
SB28CNS49	5	158341632	158342205	Intergenic	EBF1	Intron	SB28SB59DP1
SB59CNS33	10	131691195	131691924	Intergenic	EBF3	Intron	SB28SB59DP1
SB28CNS563	5	158301808	158302038	Intergenic	EBF1	Intron	SB28SB59DP10
SB59CNS176	10	131685582	131686130	Intergenic	EBF3	Intron	SB28SB59DP10
SB28CNS639	5	158356453	158356709	Intergenic	EBF1	Intron	SB28SB59DP11
SB59CNS195	10	131694031	131694422	Intergenic	EBF3	Intron	SB28SB59DP11
SB28CNS438	5	122165508	122165835	Intergenic	P4HA2	Downstream	SB28SB59DP28
SB59CNS168	10	50572125	50572708	3UTR	P4HA1	Downstream	SB28SB59DP28
SB29CNS18	6	9634909	9635426	Intergenic	TFAP2A	Downstream	SB29SB101DP1
SB101CNS49	20	51339717	51340156	Intergenic	TFAP2C	Downstream	SB29SB101DP1
SB29CNS15	6	8838546	8839172	Intergenic	TFAP2A	Downstream	SB29SB33DP1
SB33CNS219	6	51875258	51875394	Intron	TFAP2B	Downstream	SB29SB33DP1
SB29CNS260	6	10397908	10398127	3UTR	TFAP2A	UTR	SB29SB33DP4
SB33CNS101	6	50815056	50815324	3UTR	TFAP2B	UTR	SB29SB33DP4
SB2CNS1101	1	78239605	78240214	Intron	PDE4B	Downstream	SB2SB27DP12
SB27CNS861	5	61127528	61128110	Intergenic	PDE4D	Upstream	SB2SB27DP12
SB2CNS102	1	88184033	88184495	Intergenic	KANK4	Upstream	SB2SB50DP2
SB50CNS3	9	969049	969541	3UTR	KANK1	Downstream	SB2SB50DP2
SB2CNS2568	1	61330941	61331041	Intergenic	NFIA	Upstream	SB2SB51DP21
SB51CNS541	9	14532022	14532211	Intergenic	NFIB	Upstream	SB2SB51DP21
SB2CNS91	1	61918558	61919114	Intron	NFIA	Intron	SB2SB51SB89TriP1
SB51CNS37	9	14096549	14097003	Intergenic	NFIB	Intron	SB2SB51SB89TriP1
SB89CNS1	19	13202256	13202570	Intron	NFIX	Intron	SB2SB51SB89TriP1

SB2CNS354	1	70696395	70696746	Intron	PTCH2	Upstream	SB2SB54DP7
SB54CNS91	9	98222210	98222557	Intron	PTCH1	Intron	SB2SB54DP7
SB31CNS3	6	41491553	41491967	Intergenic	FOXP4	Upstream	SB31SB44DP1
SB44CNS69	7	114052554	114052975	Intergenic	FOXP2	Intron	SB31SB44DP1
SB33CNS28	6	50516711	50517099	Intron	RHAG	Upstream	SB33SB76DP7
SB76CNS713	15	95718003	95718520	Intergenic	RHCG	Upstream	SB33SB76DP7
SB34CNS0	6	62389140	62390236	3UTR	KHDRBS2	UTR	SB34SB48DP9
SB48CNS1205	8	94507454	94507689	Intergenic	KHDRBS3	Downstream	SB34SB48DP9
SB35CNS158	6	108797659	108798066	Intron	ENPP3	Upstream	SB35SB48DP2
SB48CNS53	8	93623427	93623993	Intergenic	ENPP2	Downstream	SB35SB48DP2
SB35CNS260	6	107813567	107813928	Intron	FOXO3	Upstream	SB35SB72DP3
SB72CNS653	13	100309355	100309705	Intron	FOXO1	Upstream	SB35SB72DP3
SB35CNS1462	6	113902732	113903036	Intergenic	HSF2	Upstream	SB35SB78DP21
SB78CNS1119	16	61089599	61089785	Intergenic	HSF4	Upstream	SB35SB78DP21
SB37CNS6	7	27183227	27184221	5UTR	HOXA5	UTR	SB37SB82DP1
SB82CNS439	17	46671202	46671332	Intergenic	HOXB5	Upstream	SB37SB82DP1
SB37CNS39	7	27179806	27180206	Intergenic	HOXA5	Upstream	SB37SB82DP2
SB82CNS84	17	46667532	46667877	Intergenic	HOXB5	Upstream	SB37SB82DP2
SB43CNS16	7	101723569	101723827	Intron	CUX1	Intron	SB43SB69DP1
SB69CNS67	12	111725243	111725479	Intron	CUX2	Intron	SB43SB69DP1
SB43CNS18	7	101901303	101901535	3UTR	GPC2	Upstream	SB43SB72DP1
SB72CNS155	13	54768429	54768875	Intergenic	GPC6	Upstream	SB43SB72DP1
SB44CNS2	7	114295212	114296219	Intron	CPA1	Upstream	SB44SB72DP10
SB72CNS192	13	79042675	79043190	Intergenic	CPA2	Upstream	SB44SB72DP10
SB46CNS5	8	37238226	37238725	Intergenic	ZNF703	Upstream	SB46SB59DP1
SB59CNS42	10	77406123	77406837	Intergenic	ZNF503	Upstream	SB46SB59DP1
SB46CNS127	8	36957904	36958139	Intergenic	ZNF703	Upstream	SB46SB59DP10
SB59CNS71	10	77726895	77727502	Intron	ZNF503	Upstream	SB46SB59DP10
SB46CNS398	8	23064603	23064906	Intergenic	R3HCC1	Upstream	SB46SB59DP24
SB59CNS1335	10	104332491	104332967	Intron	c10orf28	Downstream	SB46SB59DP24
SB46CNS6	8	37310224	37310744	Intergenic	ZNF703	Upstream	SB46SB59DP2
SB59CNS191	10	77383381	77383857	Intergenic	ZNF503	Upstream	SB46SB59DP2
SB46CNS41	8	37277993	37278318	Intergenic	ZNF703	Upstream	SB46SB59DP4
SB59CNS221	10	77389384	77389846	Intergenic	ZNF503	Upstream	SB46SB59DP4
SB46CNS62	8	37532824	37533223	Intergenic	ZNF703	Upstream	SB46SB59DP5
SB59CNS435	10	77164918	77165215	Intergenic	ZNF503	Upstream	SB46SB59DP5

SB46CNS73	8	37378365	37378666	Intergenic	ZNF703	Upstream	SB46SB59DP7
SB59CNS66	10	77357417	77358116	Intergenic	ZNF503	Upstream	SB46SB59DP7
SB47CNS16	8	61761724	61762604	Intron	SULF1	Upstream	SB47SB68DP1
SB68CNS23	12	85702741	85703365	Intergenic	GNS	Upstream	SB47SB68DP1
SB47CNS50	8	71021993	71022406	Intergenic	TOX	Upstream	SB47SB78DP1
SB78CNS9	16	52792379	52793337	Intron	TOX3	Upstream	SB47SB78DP1
SB47CNS32	8	77647766	77648378	Intron	ZFHX4	Intron	SB47SB78DP3
SB78CNS273	16	51148015	51148300	Intergenic	ZFHX3	Downstream	SB47SB78DP3
SB47CNS6	8	53086472	53087314	Intron	TOX	Downstream	SB47SB78DP5
SB78CNS400	16	80204017	80204371	Intron	TOX3	Upstream	SB47SB78DP5
SB48CNS4	8	144613660	144614440	Intergenic	KCNK9	Upstream	SB48SB101DP2
SB101CNS99	20	50179359	50179613	Intergenic	KCNK15	Downstream	SB48SB101DP2
SB48CNS13	8	100648614	100649227	Intron	ESRP1	Downstream	SB48SB78DP2
SB78CNS31	16	51789433	51789992	Intergenic	ESRP2	Downstream	SB48SB78DP2
SB48CNS81	8	93090734	93091154	Intergenic	ESRP1	Downstream	SB48SB78DP4
SB78CNS126	16	59227451	59228007	Intergenic	ESRP2	Downstream	SB48SB78DP4
SB48CNS172	8	93935357	93935834	Intergenic	LRP12	Downstream	SB48SB91DP3
SB91CNS96	19	32084609	32084956	Intergenic	LRP3	Upstream	SB48SB91DP3
SB48CNS262	8	106743026	106743281	Intron	LRP12	Upstream	SB48SB91DP4
SB91CNS202	19	30602060	30602263	Intergenic	LRP3	Upstream	SB48SB91DP4
SB27CNS91	5	87962585	87963591	Intergenic	MEF2C	Downstream	SB4SB27SB76TriP1
SB4CNS87	1	156390129	156390271	Intron	MEF2D	Downstream	SB4SB27SB76TriP1
SB76CNS971	15	89911098	89911353	Intergenic	MEF2A	Upstream	SB4SB27SB76TriP1
SB54CNS19	9	79627872	79628232	Intergenic	FOXB2	Upstream	SB54SB76DP1
SB76CNS267	15	60285830	60286124	Intergenic	FOXB1	Upstream	SB54SB76DP1
SB54CNS26	9	79628340	79628783	Intergenic	FOXB2	Upstream	SB54SB76DP2
SB76CNS116	15	60286655	60287104	Intergenic	FOXB1	Upstream	SB54SB76DP2
SB5CNS9	1	198215373	198215959	Intron	PBX1	Downstream	SB5SB56DP1
SB56CNS4	9	128521405	128522382	Intergenic	PBX3	Intron	SB5SB56DP1
SB5CNS868	1	179937821	179938144	Intergenic	QSOX1	Upstream	SB5SB56DP11
SB56CNS210	9	127957567	127958064	Intergenic	QSOX2	Downstream	SB5SB56DP11
SB5CNS35	1	164700584	164701100	Intergenic	PBX1	Intron	SB5SB56DP2
SB56CNS6	9	128645901	128646742	Intergenic	PBX3	Intron	SB5SB56DP2
SB5CNS80	1	164325972	164326536	Intergenic	PBX1	Upstream	SB5SB56DP3
SB56CNS832	9	124282209	124282656	Intergenic	PBX3	Upstream	SB5SB56DP3
SB5CNS224	1	164637895	164638102	Intergenic	PBX1	Intron	SB5SB56DP5

SB56CNS514	9	128584081	128584254	Intergenic	PBX3	Intron	SB5SB56DP5
SB5CNS250	1	172113496	172113783	Intron	DNM3	Intron	SB5SB56DP6
SB56CNS771	9	131006995	131007149	Intron	DNM1	Intron	SB5SB56DP6
SB60CNS12	11	16424381	16425029	5UTR	SOX6	UTR	SB60SB66DP1
SB66CNS10	12	24168071	24168699	Intergenic	SOX5	Intron	SB60SB66DP1
SB60CNS21	11	16461607	16462219	Intergenic	SOX6	Intron	SB60SB66DP2
SB66CNS11	12	24291747	24292328	Intergenic	SOX5	Intron	SB60SB66DP2
SB60CNS26	11	8304629	8305146	Intron	LMO1	Upstream	SB60SB66DP3
SB66CNS27	12	16941504	16941911	Intron	LMO3	Upstream	SB60SB66DP3
SB60CNS51	11	16426363	16426911	Intergenic	SOX6	Intron	SB60SB66DP5
SB66CNS34	12	24173340	24173943	Intergenic	SOX5	Intron	SB60SB66DP5
SB60CNS381	11	8290324	8290567	Intron	LMO1	Upstream	SB60SB66DP7
SB66CNS45	12	16941128	16941437	Intron	LMO3	Upstream	SB60SB66DP7
SB6CNS3	1	206702290	206703042	Intron	PIK3C2B	Upstream	SB6SB60DP3
SB60CNS338	11	29747368	29748117	Intron	PIK3C2A	Upstream	SB6SB60DP3
SB71CNS14	13	31035193	31035456	3UTR	HMGB1	UTR	SB71SB108DP1
SB108CNS1510	X	97677844	97678108	Intergenic	HMGB3	Upstream	SB71SB108DP1
SB72CNS1176	13	100376305	100376539	Intron	ZIC2	Upstream	SB72SB108DP13
SB108CNS1833	X	136316342	136316562	Intergenic	ZIC3	Upstream	SB72SB108DP13
SB72CNS1820	13	92002789	92002953	Intergenic	GPC5	Upstream	SB72SB108DP21
SB108CNS1565	X	133304065	133304309	Intergenic	GPC3	Upstream	SB72SB108DP21
SB76CNS8	15	96031795	96032737	Intergenic	SLC12A1	Downstream	SB76SB78DP18
SB78CNS399	16	52436473	52436791	Intergenic	SLC12A3	Upstream	SB76SB78DP18
SB76CNS25	15	68040022	68040567	Intron	SMAD3	Downstream	SB76SB86DP1
SB86CNS56	18	45213210	45213528	Intergenic	SMAD2	Downstream	SB76SB86DP1
SB76CNS42	15	57426167	57426671	Intron	TCF12	Intron	SB76SB86DP2
SB86CNS41	18	53089869	53090198	Intron	TCF4	Intron	SB76SB86DP2
SB76CNS55	15	57425317	57425821	Intron	TCF12	Intron	SB76SB86DP3
SB86CNS141	18	53090854	53091090	Intron	TCF4	Intron	SB76SB86DP3
SB78CNS364	16	51185299	51185548	Intergenic	SALL1	Upstream	SB78SB101DP2
SB101CNS133	20	50418988	50419262	Intergenic	SALL4	UTR	SB78SB101DP2
SB78CNS1	16	49735417	49736423	Intron	ZNF423	Intron	SB78SB85DP1
SB85CNS10	18	22865042	22865614	Intron	ZNF521	Intron	SB78SB85DP1
SB87CNS9	18	75831759	75832470	Intron	SALL3	Upstream	SB78SB87DP7
SB78CNS21	16	51786401	51787036	Intergenic	SALL1	Upstream	SB78SB87DP7
SB78CNS237	16	51732745	51733050	Intergenic	SALL1	Upstream	SB78SB87DP3

SB87CNS19	18	75957049	75957625	Intron	SALL3	Upstream	SB78SB87DP3
SB78CNS784	16	51231620	51231961	Intergenic	SALL1	Upstream	SB78SB87DP5
SB87CNS56	18	76705047	76705441	Intron	SALL3	Upstream	SB78SB87DP5
SB78CNS1100	16	51492664	51492856	Intergenic	SALL1	Upstream	SB78SB87DP6
SB87CNS26	18	76461748	76462227	Intron	SALL3	Upstream	SB78SB87DP6
SB87CNS80	18	72925601	72925883	Intergenic	TSHZ1	Intron	SB87SB101DP3
SB101CNS14	20	51619958	51620394	Intergenic	TSHZ2	Intron	SB87SB101DP3
SB87CNS174	18	73370811	73370962	Intergenic	TSHZ1	Downstream	SB87SB101DP6
SB101CNS34	20	51962511	51962864	Intergenic	TSHZ2	Intron	SB87SB101DP6
SB9CNS308	2	50755790	50756221	Intergenic	NRXN1	Intron	SB9SB74DP13
SB74CNS2246	14	79276541	79276656	Intron	NRXN3	Intron	SB9SB74DP13
SB9CNS576	2	50848388	50848561	Intergenic	NRXN1	Intron	SB9SB74DP19
SB74CNS1814	14	79115552	79115677	Intergenic	NRXN3	Intron	SB9SB74DP19
SB9CNS3	2	63193931	63194666	Intergenic	OTX1	Upstream	SB9SB74DP2
SB74CNS117	14	57476174	57476711	Intergenic	OTX2	Upstream	SB9SB74DP2
SB9CNS86	2	63195795	63196425	Intergenic	OTX1	Upstream	SB9SB74DP26
SB74CNS228	14	57475279	57475844	Intergenic	OTX2	Upstream	SB9SB74DP26
SB9CNS100	2	58860364	58860805	Intergenic	VRK2	Downstream	SB9SB74DP27
SB74CNS769	14	97490229	97490490	Intergenic	VRK1	Downstream	SB9SB74DP27
SB9CNS20	2	60441493	60442175	Intergenic	BCL11A	Downstream	SB9SB74DP4
SB74CNS962	14	99466536	99466727	Intergenic	BCL11B	Downstream	SB9SB74DP4
SB9CNS64	2	60297806	60298245	Intergenic	BCL11A	Downstream	SB9SB74DP7
SB74CNS466	14	99322718	99323056	Intergenic	BCL11B	Downstream	SB9SB74DP7
SB9CNS125	2	58858016	58858435	Intergenic	VRK2	Downstream	SB9SB74DP9
SB74CNS29	14	97430975	97431761	Intergenic	VRK1	Downstream	SB9SB74DP9
SB9CNS69	2	63268471	63269049	Intergenic	MEIS1	Upstream	SB9SB75DP2
SB75CNS89	15	36985668	36986077	Intron	MEIS2	Downstream	SB9SB75DP2