MBE Advance Access published December 20, 2006

Comparison of *Pax1/9* locus reveals 500-Million-Year-Old syntenic block and evolutionary conserved non-coding regions

Authors: Wei Wang[*], Jing Zhong[*], Bing Su[†], Yan Zhou[¶] and Yi-Quan Wang[*]

* Key Laboratory of the Ministry of Education for Cell Biology and Tumor Cell Engineering, School of Life Sciences, Xiamen University, Xiamen, 362005, China

† Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology and Kunming Primate Research Center, The Chinese Academy of Sciences, Kunming, 650223, China

¶ Chinese Human Genome Center at Shanghai, Shanghai, 201203, China

Corresponding author: Yi-Quan Wang; Address: School of Life Sciences, Xiamen University, Xiamen, 362005, China; Tel: +86-592-2184427; Fax: +86-592-2181015; E-mail: wangyq@xmu.edu.cn

Key words: amphioxus; conserved synteny; ncECR; duplication; evolution

Running head: Comparison *Pax1/9* locus reveals conserved region

Comparison of *Pax1/9* locus reveals 500-Million-Year-Old syntenic block and evolutionary conserved non-coding regions

Abstract: Identification of conserved genomic regions within and between different genomes is crucial when studying genome evolution. Here, we described regions of strong synteny conservation between vertebrates (tetraopods and teleosts) and invertebrate deuterostomes (amphioxus and sea urchin). The shared gene contents across phylogenetically distant species demonstrate that the conservation of the regions stemmed from an ancestral segment instead of a series of independent convergent events. Comparison of the syntenic regions allows us to postulate the primitive gene organization in the last common ancestor of deuterostomes and the evolutionary events that occurred to the three distinct lineages of sea urchin, amphioxus and vertebrates after their separation. In addition, alignment of the syntenic regions led to the identification of 8 non-coding evolutionary conserved regions shared between amphioxus and vertebrates. To our knowledge, this is the first report of conserved non-coding sequences shared by vertebrates and non-vertebrates. These non-coding sequences have high possibility of being elements that regulate neighboring genes. They are likely to be a factor in the maintenance of conserved synteny over long phylogenetic distance in different deuterostome lineages.

Key words: amphioxus; conserved synteny; ncECR; duplication; evolution

Introduction

The nature of the last common ancestor of vertebrates is of enormous research interest. This area of interest has being rigorously researched in the past and present. Knowledge in this area has increased steadily which, in the not too distant time, we hope to understand the evolution process. It has been recognized that the origin of vertebrates is significantly related to the profusion of gene duplication events (Pennisi 2001). The correlation between the increase of gene number and dramatic jumps in morphological complexity has led to a widely accepted notion that the duplication of existing genes was the source of raw materials for the great bursts of innovation seen in the vertebrate lineage. However the extent and nature of the events (two successive polyploidizations, single polyploidization or limited segmental duplications) remained a subject of vigorous debate (Larhammar, Lundin, and Hallbook 2002; McLysaght, Hokamp, and Wolfe 2002; Friedman and Hughes 2003). Comparisons of vertebrate species to one another and to invertebrate outgroups could bring further knowledge to the debate. Prior studies have shown that cephalochordate amphioxus is the closest living invertebrate relative of vertebrates (Wada and Satoh 1994) and it has evolved into its present form almost instantly preceding the vertebrate gene duplication (Panopoulou et al. 2003). This knowledge of cephalochordate amphioxus offers us a compelling understanding of the ancestral state of preduplicated genome (Panopoulou et al. 2003; Horton and Gibson-Brown 2003; Luke et al. 2003; Castro and Holland 2003; Castro, Furlong, and Holland 2004).

To better comprehend the duplication pattern at the origin of vertebrates, it is important to research the evolutionary history of given chromosomal paralogy regions. Such research will provide evidences for different duplication models, as well as offer new insights into the dynamics of genome evolution and the emergence of new functions. In 2003, Santagati et al. identified a pair of paralogy regions containing five genes from four distinct gene families (*Pax1/9*, *Nkx2*, *Slc*, and *FoxA*) in human chromosome (HSA) 20p and 14q. These regions are hypothesized to have formed either through segmental or whole genome duplication in early vertebrate evolution. The availability of genomic sequences from invertebrate species would greatly increase the reliability of comparative genomic analysis to delineate the ancestry and divergence of the paralogy region. In this study, we have sequenced and annotated the *Pax1/9* containing BAC clone from Chinese amphioxus

Branchiostoma belcheri. Through comparison of syntenic regions from human, amphioxus and a variety of other species, we are able to illustrate the stigmata of conservation among different deuterostome lineages exists in spite of more than 500 million years of divergence and evolution from their last common ancestor. Moreover, cross-species alignment of the contiguous sequences corresponding to the conserved syntenic blocks enables the identification of non-coding evolutionary conserved regions (ncECR, Shin et al. 2005) that are shared among the divergent groups of vertebrate and invertebrates. To our knowledge the conservation over such a great evolutionary distance has not been reported before. These ncECRs represent potential regulatory elements associated with the presence of genes in the conserved syntenic blocks.

Materials and Methods

Sequencing of the *Pax1/9* locus and the surrounding genes in *B. belcheri*

We have documented the isolation of BAC clone 71P5 from *B. belcheri* containing *Pax1/9* gene located on a 140 kb genomic insert in the BAC vector pBACe3.6 in our previous work (Wang et al. 2005). This clone was sonicated and the resulting fragments (2 kb) were subcloned and sequenced to cover the BAC clone with 8x overlap. The sequences were then analyzed and assembled with the PHRED/PHRAP/CONSED package (University of Washington, Seattle, WA, USA; http://www.phred.org). For the annotation, we adopted the gene finding program GENESCAN to predict potential genes (http://bioweb.pasteur.fr/seqanal/interfaces/genscan.html). The entire sequence and the predicted genes were investigated by homology searching against NCBI databases (http://ncbi.nlm.nih.gov/BLAST/) with BLASTN, BLASTX, and TBLASTX algorithms (Altschul et al. 1990).

Synteny Conservation of *Pax1/9* Loci among Deuterostomes

Scaffold bearing amphioxus *Pax1/9* gene was retrieved from JGI v.1.0 *B. floridae* draft genome assembly (http://genome.jgi-psf.org/) using TBlastN search with the correspondent full-length protein sequence (*BfPax1/9*, GenBank accession no. U20167) as query. The adjacent genes of *BfPax1/9* were investigated to determine whether the orthologues could be mapped to the vicinity of *Pax1* or *Pax9* loci in human genome.

The loci conservation of syntenic genes shared between amphioxus and human were determined via

database (JGI, UCSC and Ensembl) searching against the genome assemblages of 15 species including six mammals (chimpanzee *Pan troglodytes*; dog *Canis familiaris*, cow *Bos taurus*, mouse *Mus musculus*, rat *Rattus norvegicus*, opossum *Monodelphis domestica*) one bird (chicken *Gallus gallus*), one amphibian (western clawed frog *Xenopus tropicalis*), three teleost fishes (zebrafish *Danio rerio*, pufferfish *Tetraodon nigroviridis*, and fugu *Takifugu rubripes*), one urochordate (ascidian *Ciona intestinalis*), one echinoderms (sea urchin *Strongylocentrotus purpuratus*), one insect (fruitfly *Drosophila melanogaster*), and one worm (nematode *Caenorhabditis elegans*).

Phylogenetic reconstruction

Phylogenetic reconstruction was employed to access the orthology and paralogy relationships of genes residing in the *Pax1/9* loci. The putative protein sequences were either retrieved by database searching (accession numbers provided in Supplementary Table 1) or predicted through alignment of genomic sequences to known cDNA sequences by Genewise (http://www.ebi.ac.uk/Wise2/). Amino acid sequences were aligned using the CLUSTAL X program (Thompson, Higgins, and Gibson 1994). The alignments were visually verified and corrected where necessary to improve accuracy. The phylogenetic trees were constructed using neighbor-joining (p-distance) method from version 3.1 of MEGA program (Kumar, Tamura, and Nei 2004). Confidence on each node was assessed by 1,000 bootstrap replicates.

Identification of ncECR

Syntenic genomic regions were retrieved from genome assemblies. Each gene and exon in the regions was annotated from start to end. Repetitive sequences were detected and removed by the program RepeatMasker (http://www.repeatmasker.org/) on slow setting. Local multiple alignments of the resulting masked sequences were generated and visualized by using MULAN free web server (Ovcharenko et al. 2005; http://mulan.dcode.org/) with the "increase alignments sensitivity" option. To detect ncECR presence across vertebrates and invertebrates, a high-resolution threshold of 50% identity in a 50 bp window was adopted. All identified ncECRs were tested individually by using BLASTN to exclude their presence in other positions of the genomes. Statistical analyses were conducted with paired t-test (two-tailed) to determine whether ncECRs

were significantly more conserved than their adjacent segments. Potential binding sites for transcription factors were analyzed using Matinspector (http://www.genomatrix.de, Cartharius et al. 2005).

Results

Pax1/9 Gene Environment in Amphioxus

From *B. belcheri* BAC clone 71P1, we obtained 134 kb of sequence with 4 gaps organized into five contigs. Combining GenScan prediction and homology search, a total of five genes are identified. They are dehydrogenase/reductase family member 7 (*Dhrs7*), *Pax1/9*, solute carrier family 25 member 21 (*Slc25A21*), egg laying defective nine homolog 3 (*Egln3*) and hepatocyte nuclear factor 3 (*HNF3*) family member *HNF3-1*. The clone 71P1 contains a complete coding sequence of the former four genes, and a 649 bp segment at the sequence end, which is identical to the 3'UTR of *B. floridae HNF-3-1* mRNA (GenBank accession no. X96519) indicating that this gene is present immediately beside the genomic stretch. Since the nomenclature of the *HNF3* family has been revised to *FoxA* genes (Kaestner 2000), we refer to amphioxus *HNF-3-1* as *FoxA1/2A* hereafter.

To determine if the gene loci are conserved in amphioxus lineage, the genome assembly of another amphioxus species *B. floridae* is examined and the result shows that the corresponding sequence of clone 71P5 is localized in Scaffold 42. Comparison of the orthologous regions between two amphioxus species reveals identical gene contents and spatial organization (Fig 1). In addition, two more genes are found in the downstream region of *B. floridae*. One gene encoding *HNF3-2* (GenBank accession no. Y09236) is present in close tail-to-tail arrangement to *FoxA1/2A*. We refer to it as *FoxA1/2B* hereafter. The second one is mirror-image polydactyly gene 1 (*Mipol1*) residing downstream of *FoxA1/2B.*

Region of Conserved Synteny Maintained throughout 500 Myr of Evolution

As indicated in Fig 1, synteny between the segment in amphioxus and two regions in human genome is found through genomic comparison of *Pax1/9* gene environments. Among seven genes in the amphioxus segment, four (*Pax1/9*, *Slc25A21*, *FoxA1/2A* and *FoxA1/2B*) can be mapped to the pair of human 20p-14q paralogy regions (Santagati et al. 2003), the other three genes (*Mipol1*, *Egln3* and *Dhrs7*) have orthologues on

14q but not 20p. *Egln3* and *Dhrs7* are distantly associated with the physically linked four genes of *Pax9*, *Slc25A21*, *Mipol1* and *FoxA1*.

To better understand the evolution of ancient syntenic region shared between amphioxus and human, the conservation of orthologous gene loci of the available complete genome sequences from diverse species were investigated (Fig 1). Comparison of vertebrate orthologous regions showed that the organization of syntenic gene group *Egln3-Pax9-Slc25A21-Mipol1-FoxA1-Dhrs7* is highly conserved in diverse tetrapod species. However in teleost lineage only the linkage between *Pax9* and *Slc25A21* is observed. The paralogous cluster of *Pax9*-containing syntenic group includes three genes of *Pax1*, *Slc25A5l* and *FoxA2*. The association between *Pax1* and *FoxA2* is conserved throughout vertebrate. The computer predicted pseudogene *Slc25A5l* is only encountered in the syntenic region of human and chimpanzee, suggesting the remnant of ancient duplicated copy had been erased from non-primate vertebrates.

In invertebrate species, the strongest synteny conservation is observed in the sea urchin genome. The close linkage of four genes including *Slc25A21*, *Pax1/9*, *Mipol1* and *FoxA1/2* is identified in scaffold 37. In the genome draft of *C. intestinalis*, a genomic segment of 25 kb long (Scaffold 764) comprising of *Slc25A21* and *Pax1/9* is found, but no *Mipol1* or *FoxA1/2* coding sequence is identified, suggesting these two genes are either discarded or positioned in the gap of the *Ciona* genome assembly. In the genome of *C. elegans*, the genes of *Pax1/9*, *Egln3* and *FoxA1/2* are mapped to the same chromosome albeit the large intervening genomic space between them. In *D. melanogaster*, all identifiable orthologues are not syntenic. This result is not surprising since the lineage of protostomes diverged to a great extent from that of deuterostomes.

Phylogenetic Resconstruction

We constructed neighbour-joining trees to access the phylogenetic relationship of 6 gene families localized in the surveyed region of syntenic conservation (Supplementary Fig 1). Genes used in this analysis are listed in supplementary Table 1. In general, the internal topology of each tree agrees fairly well with the accepted evolutionary relationship of the organisms and is supported by high bootstrap values. The invertebrate genes are placed at the base of the trees. The two gene families of *Pax1/9* and *FoxA1/2* both have two vertebrate members and exhibited similar topology. The two vertebrate genes show greater similarity to each other than

their invertebrate counterparts. In the gene family *FoxA1/2*, the two *B. floridae* genes (*BfFoxA1/2A* and *BfFoxA1/2B*) are clustered together, suggesting gene duplication specific to the amphioxus lineage.

Identification of ncECR

To detect putative regulatory elements conserved across vertebrates and invertebrates, we searched for ncECRs in the syntenic region conserved in diverse deuterostomes species using multiple sequence alignment tool Mulan. A total of 8 conserved elements shared by amphioxus and vertebrates were identified in highly variable sequence background. These elements are 50~124 bp in amphioxus genome and can be divided into two groups. The first group includes 4 elements (A1~A4) shared with vertebrate *Pax1-FoxA2* genomic segments and the second comprises another 4 elements (B1~B4) shared with vertebrate *Pax9-Slc25A21-Mipol1-FoxA1* intervals. The spatial arrangement and alignment are shown in Fig 2 and supplementary Fig. 2. The sequence similarities range from 52.8%~74% between the two amphioxus species, 68.9%~95.2% between the vertebrate species, and 53.2%~73.2% between the two clades of amphioxus and vertebrates. Statistics analysis reveals that with the exception of element A1, the ncECR identified in our study are significantly more conserved than the adjacent segments (paired t-test, two-tailed $P<0.05$). In addition, element A1 does not widely deviate from the significance ($p=0.056$). The result indicates that negative selection has acted on these elements.

The order, position and orientation of those identified elements are strictly conserved within each individual lineage of amphioxus or vertebrate, but highly variant between the two lineages (Fig. 2). One exception is element A4 residing downstream of *FoxA2* (*FoxA1/2B*) with relative conserved orientation in both lineages. Additionally, this element also displayed the highest level of conservation as suggested by wider evolutionary spectrum (including teleosts) and higher overall sequence identity (90% versus 55~80%) than to the other 7 elements. These evidences strongly indicate that A4 is involved in fundamental regulating function of *FoxA*.

We employed the MatInspector program to examine whether any known transcription factor-binding sites were contained in the conserved elements. The results show that a putative *FoxA2* binding site exists in the element A3 of both amphioxus and vertebrates. Previous studies demonstrated that both cross-regulation (by

FoxA2 and *FoxA3*) and autoregulation (by *FoxA1* itself) featured in the regulation of *FoxA1* gene (Kaestner et al. 1999). The binding site of *FoxA2* in element A3 is probably required to mediate the regulation of *Fox* genes.

Discussion

Phylogenetic analysis

Of the 7 genes corresponding to 6 gene families predicted from the amphioxus genomic sequences, three (*Pax1/9*, *FoxA1/2A* and *FoxA1/2B*) were cloned previously and their phylogenetic relationships analysed (Holland, Holland, and Kozmic 1995; Shimeld 1997). Here, we performed phylogenetic analysis on all 6 gene families by incorporating sequence of taxa examined in our study (Supplementary Fig 1 and Supplementary Table 1). The topology of *Pax1/9* is consistent with previous studies (Holland, Holland, and Kozmic 1995) indicating that the vertebrate *Pax1* and *Pax9* originated by duplication after the divergence of amphioxus. The phylogenetic analysis of *FoxA1/2* gene family confirms the earlier hypothesis that the ancient *FoxA1/2* had independently duplicated in the two lineages of amphioxus and vertebrate (Shimeld 1997). Furthermore, the identification of physically linked *FoxA1/2A* and *FoxA1/2B* in *B. floridae* genome revealed that these two genes were formed through tandem duplication. The four newly predicted amphioxus genes (*Dhrs7*, *Egln3*, *Slc25A21* and *Mipol1*) displayed an orthology relationship toward vertebrate sequences.

Evolution of Syntenic Block

Evidences from genomic loci and evolutionary reconstruction strongly suggest that the region surrounding *Pax1/9* genes are evolutionarily related between human and amphioxus. Such homology can be explained by two alternative hypotheses. The first hypothesis is the conservation of an ancestral state, and the other is the convergence of evolutionary events. In order to test the hypotheses, we investigated whether any conservation could be found in the genomes of other species. This investigation reveals strong synteny conservation over wide evolutionary spectrum of deuterostome, supporting the hypothesis of a conserved ancestral state.

Identification of homologous genomic regions is an essential prerequisite to study the evolution of genomes, both within and between organisms. Identifying intergenomic homology allows researchers to assess the impact of rearrangement events, while intragenomic homology gives insights into the duplication history of

a genome (Simillion, Vandepoele, and Van de Peer 2004). Analysis of the conserved synteny between vertebrate paralogy regions and non-vertebrate deuterostome chromosomal segments supports the dynamics that shaped the region hosting *Pax1/9* gene members. We propose that a linked array of genes comprising *Pax1/9, Slc25A21/5l, Egln3, Mipol1* and *FoxA1/2* existed in the genome of early deuterostomes which was then passed to sea urchin, amphioxus and vertebrates. Lineage specific rearrangement of the major deuterostome clades can be inferred through the comparison of positional relationship of orthologous genes. A putative model for the evolution of the syntenic region harboring *Pax1/9* is given in Fig. 3.

In amphioxus, vertebrates and *C. intestinalis*, *Slc25A21* lies downstream of *Pax1/9* (*Pax9*). However, this gene is located upstream of *Pax1/9* in sea urchin. We also noticed that a gap of 36 kb long is present immediately upstream of *Slc25A21* in the sea urchin genome assembly. This observation allows us to speculate that during the evolution of sea urchin, the local rearrangement had translocated the genomic segment encompassing *Slc25A21* and *Egln3* (block1 in Fig. 3) upstream of *Pax1/9* and the *Egln3* gene localized in the gap of the region.

Compare to sea urchin and tetrapods, the amphioxus region features two peculiarities. One is the reversed order and orientation of the block containing *Mipol1* and the neighboring *FoxA1/2B* gene, the second is that two *FoxA* family members (*FoxA1/2A* and *FoxA1/2B*) are closely linked in tail-to-tail fashion. Phylogenetic analysis shows that *FoxA1/2A* and *FoxA1/2B* from amphioxus are co-orthologous to the *FoxA1* and *FoxA2* from human (Shimeld 1997). Therefore we hypothesise that in amphioxus lineage, the genomic region harboring *Mipol1* and *FoxA1/2* (block2 in Fig3) underwent segmental inverse duplication followed by the disposal of one *Mipol1* copy.

In the genome of diverse vertebrate species there are two chromosomal segments showing similar gene contents and organization to the single invertebrate region. Thus, we infer that during the early stages of vertebrate evolution, the chromosomal segment containing a cluster of at least 5 genes were duplicated. The congruent timing of vertebrate gene expansion and the duplication of two paralogous gene pairs of *Pax1/9* and *FoxA1/2* (Holland et al. 1995; Shimeld 1997) implies that the sister segments could have formed from whole genome duplication at the early stage of the vertebrate formation. Subsequently, the two sets ofparalogous genes diverged independently up to the present time, acquiring distinct functions. In the region containing *Pax1*,

the paralogues of *Mipol1* and *Egln3* were discarded, and the *Slc25A5l* was pseudogenized after the genome duplication. In the region harboring *Pax9,* all single copies of ancestral genes were retained albeit an interchromosal translocation that moved *Egln3* to a distant position.

The evolutionary scenario we present here supports the idea that the ancestral segment containing *Pax1/9, Slc25A21 and FoxA1/2* had duplicated in vertebrate progenitors as proposed by Santagati et al (2003). Our finding that the two additional genes of *Egln* and *Mipol1* are also part of the duplicated segment extends further the characterization of the syntenic region. Furthermore, by incorporating genomic information of invertebrate species we are able to deduce that the evolutionary events are specific occurrences to the lineages of cephalochordates (amphioxus) and echinoderms (sea urchin). Another marked difference is that our scenario does not include the tandem pair of *NK2* class genes (*Nkx2-1/4* and *Nkx2-2/9*), which are previously regarded as part of the syntenic region (Santagati et al, 2003). Our genome assembly search reveals that there is a pair of tightly linked *NK2* class genes in the genome of amphioxus and sea urchin, however none resides in the scaffold harbouring *Pax1/9* gene (data not shown). With separate evolution paths, the correlation diminishes to insignificant level between NK class genes and the synetenic regions.

Conserved Non-coding Sequences in the Region

The existence of conserved syntenies between vertebrates and invertebrates has been studied previously. Such studies have focused on two areas: the clustering of families of related genes and the relative positions of genes physically associated in one lineage whose orthologues are linked to the genomes of another lineage. The former is characterised by the clustering of *Hox* (Garcia-Fernandez and Holland 1994), *Parahox* (Brooke, Garcia-Fernandez, and Holland 1998) and *Nkx* genes (Luke et al. 2003), while the later is represented by the MHC (Castro and Holland 2003) and *Insulin-Relaxin* gene families (Olinski, Lundin, and Halbook 2006). The case of highly conserved genomic region encompassing unrelated genes in vertebrates and invertebrates as deciphered in our study has not been addressed before. The degree of conservation in the gene content in the proximity of *Pax1/9* may suggest that the regulatory elements of the neighboring genes are largely conserved across great evolutionary distances. One of the selective forces that keep the genes closely linked may stem from the fact that adjacent genes share common cis-regulatory elements (Peifer, Karch, and Bender 1987).

Dispersing of these genes would deprive one of them of its cis-regulatory elements and lead to deleterious mutation. In the study of Santagati et al (2003), two conserved non-coding sequences that function as tissue-specific enhancers of neighboring genes are identified through the comparison of the syntenic regions spanning *Nkx2-9*, *Pax9*, and *Slc25a21* from human, mouse and fugu. Interestingly, one of the elements residing in the intron of *Slc25A21* was found to be the cis-regulatory sequences of neighboring *Pax9* gene (Santagati et al. 2003), suggesting the fixation of regulatory elements inside the territory of neighboring genes might constitute a functional bond to limit the gene rearrangement and as a result, the linked organization is retained in the phylogenetically distant species.

In this study, we provided a novel expansion to earlier genomic comparison by including species from both the invertebrate and vertebrate with double conserved synteny blocks surrounding *Pax1* and *Pax9* loci. The analysis allows the identification of 8 non-coding regions conserved between amphioxus and vertebrates. The sequence similarity among vertebrate species is higher than that between vertebrates and invertebrates, as can be expected from the evolutionary distance of the compared species. Interestingly, in the five elements (A1, A2, B1, B3 and B4) where the sequences of both amphioxus species are available, the amphioxus sequences exhibited low level of conservation as suggested by nucleotide identity ranging from 52.8%~74%, the degree comparable to that between amphioxus and vertebrates (53.2%~73.2%). The sister taxa *B. floridae* and *B. belcheri* diverged about 112 million years ago (Nohara et al. 2004), while the ancestor they share with vertebrates, existed approximately 500 million years. Taking the variation level and the divergence time into account, the amphioxus elements appeared to have experienced lower selective constraint than their vertebrate counterparts. It is possible that the observed changes are generally neutral and do not alter their regulatory role extensively. It is known that stabilizing selection on transcriptional output allows slightly deleterious mutations to persist, compensated for by adaptive changes elsewhere in the promoter and resulting in continuous binding-site turnover (Ludwig et al. 2000; 2005; Gompel et al. 2005; Balhoff and Wray 2005). The insufficient sensitivity of comparative genomic analysis could leave many of the functional elements undetected. Although phylogenetic footprinting is widely accepted and a powerful approach, we still need to be cautious of its limitation.

To our knowledge, the non-coding sequence similarity between vertebrates and invertebrates has never

been described before. In 2005, Woolfe et al. reported sequences that are highly conserved between human and fugu, but failed to find similarity in the invertebrate sequence databases, including the whole-genome sequences of *C. intestinalis*, *D. melanogaster*, and *C. elegans*. The non-coding sequence homology between amphioxus and vertebrates identified here indicates the presence of conserved regulatory modules between the two lineages. *In vivo* reporter constructs using amphioxus genomic DNA in transgenic mice (Manzanares et al. 2000) and chicken (Wada et al. 2006) support the hypothesis. Therefore we can infer that for the purpose of better understanding the mechanism and evolution of gene regulation, amphioxus is preferred to other invertebrate models.

As a preliminary attempt to exploit the extent of non-coding conservation between vertebrates and invertebrates, we have also inspected 3 individual loci of *Pax2/5/8, Pax3/7* and *Pax6* but no ncECR could be identified. The enrichment of ncECR in the *Pax1/9* locus is congruent with the recent knowledge that highly conserved non-coding sequences are located in and around genes involved in the regulation of transcription and development (Bejerano et al. 2004; Woolfe et al. 2005; Shin et al. 2005; Siepel et al. 2005). In vertebrates and invertebrates there are two developmental regulators (*Pax1/9* and *FoxA1/2)* residing in the regions of synteny being researched in this paper. The expression profiles of *Pax1/9* and *FoxA1/2* further supports the hypothesis that the functional relationship between ncECRs with several associated genes may be one of the selective forces that maintain the associated genes tightly through 500 Myr evolution. In amphioxus, the *Pax1/9*, *FoxA1/2A and FoxA1/2B* expressed in the developing pharyngeal (Holland, Holland, and Kozmik 1995; Shimeld 1997), whereas the vertebrate *FoxA1*, *FoxA2*, *Pax1* and *Pax9* all expressed in thymus (Dooley, Erickson, and Farr 2005), strongly indicating the linked genes might share common regulatory elements.

Based on the estimation that the common ancestors of amphioxus and vertebrates diverged 500 million years ago (Holland et al. 1992), the non-coding sequences identified in this study is conserved over half a billion years of parallel evolution, thus they are valid candidates for functional significance. Although the transient transgenic method developed by Yu et al (2004) for amphioxus embryos is generally the accepted method, the system is still not well established yet due to the limitation of the embryo resource. Other well-developed systems, such as mice or zebra fish, probably offer the experimental evaluation for these cis-regulatory modules. Furthermore, the completion of amphioxus genome project provides a valuable

resource to promote our understanding of the relationships between conserved sequences and the biological functions they confer, and shed light on the evolutionary forces that have shaped chordates genomes.

Acknowledgements

Authors are grateful to two anonymous reviewers for their helpful suggestions and comments and Mr. Bang Yoke Leong for his linguistical help. This work is supported by NSFC (No. 30470938 & No. 30570208), Natural Science Foundation of Fujian Province, China (No. D0510002) and grant from the Science and Technology Bureau of Xiamen (No. 3502Z20042015).

Literature Cited

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403-410.

Balhoff, J. P., and G. A. Wray. 2005. Evolutionary analysis of the well characterized endo16 promoter reveals substantial variation within functional sites. Proc. Natl. Acad. Sci. U S A. 102:8591-8596.

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

Brooke, N. M., J. J. Garcia-Fernandez, and P. W. H. Holland. 1998. The *ParaHox* gene cluster is an evolutionary sister of the *Hox* gene cluster. Nature. 392: 920–922.

Cartharius, K., K. Frech, K. Grote, B. Klocke, M. Haltmeier, A. Klingenhoff, M. Frisch, M. Bayerlein, and T. Werner. 2005. MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics 21:2933-2942.

Castro, L. F., and P. W. Holland. 2003. Chromosomal mapping of ANTP class homeobox genes in amphioxus: piecing together ancestral genomes. Evol. Dev. 5:459-465.

Castro, L. F., R. F. Furlong, and P. W. H. Holland. 2004. An antecedent of the MHC-linked genomic region in amphioxus. Immunogenetics 55:782-784.

Dooley, J, M. Erickson, and A. G. Farr. 2005. An organized medullary epithelial structure in the normal thymus expresses molecules of respiratory epithelium and resembles the epithelial thymic rudiment of nude mice. J. Immunol. 175:4331-4337.

Friedman, R., and A. L. Hughes. 2003. The temporal distribution of gene duplication events in a set of highly conserved human gene families. Mol. Biol. Evol. 20:154-161.

Garcia-Fernandez, J., P. W. Holland. 1994. Archetypal organization of the amphioxus *Hox* gene cluster. Nature. 370:504-505.

Gompel, N., B. Prud'homme, P. J. Wittkopp, V. A. Kassner, S. B. Carroll. 2005. Chance caught on the wing: Cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature. 433: 481–487

Holland, N. D., L. Z. Holland, and Z. Kozmik. 1995. An amphioxus *Pax* gene, *AmphiPax-1,* expressed in embryonic endoderm, but not in mesoderm: implications for the evolution of class I paired box genes. Mol. Mar. Biol. Biotechnol. 4:206-214.

Holland, P. W., L. Z. Holland, N. A. Williams, and N. D. Holland. 1992. An amphioxus *homeobox* gene: sequence conservation, spatial expression during development and insights into vertebrate evolution. Development 116:653-661.

Horton, A. C., and J. J. Gibson-Brown. 2003. Evolution of developmental functions by the Eomesodermin, T-brain1, *Tbx21* subfamily of *T-box* genes: insights from amphioxus. J. Exp. Zool. Mol. Dev. Evol. 294:112-121.

Kaestner, K. H. 2000. The hepatocyte nuclear factor 3 (*HNF3* or *FOXA*) family in metabolism. Trends Endocrinol. Metab. 11:281-285.

Kaestner, K. H., J. Katz, Y. Liu, D. J. Drucker, and G. Schutz. 1999. Inactivation of the winged helix transcription factor HNF3alpha affects glucose homeostasis and islet glucagon gene expression *in vivo*. Genes Dev. 13:495-504.

Kumar, S., K. Tamura, and M. Nei, 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform. 5: 150-163.

Larhammar, D., L. G. Lundin, and F. Hallbook. 2002. The human *Hox*-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. Genome Res. 12:1910-1920.

Ludwig, M. Z., A. Palsson, E. Alekseeva, C. M. Bergman, J. Nathan, and M. Kreitman. 2005. Functional evolution of a cis-regulatory module. PloS. Biol. 3: e93.

Ludwig, M. Z., C. Bergman, N. H. Patel, and M. Kreitman. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403: 564-567.

Luke, G. N., L. F. Castro, K. McLay, C. Bird, A. Coulson, and P. W. Holland. 2003. Dispersal of NK homeobox gene clusters in amphioxus and humans. Proc. Natl. Acad. Sci. USA 100:5292-5295.

Manzanares, M., H. Wada, N. Itasaki, P. A. Trainor, R. Krumlauf, and P. W. Holland. 2000. Conservation and elaboration of *Hox* gene regulation during evolution of the vertebrate head. Nature 408:854-857.

McLysaght, A., K. Hokamp, and K. H. Wolfe. 2002. Extensive genomic duplication during early chordate evolution. Nat. Genet. 31:200-204.

Nohara, M., M. Nishida, V. Manthacitra, and T. Nishikawa. 2004. Ancient phylogenetic separation between Pacific and Atlantic cephalochordates as revealed by mitochondrial genome analysis. Zoolog. Sci. 21: 203-210.

Olinski, R. P., L. G. Lundin, and F. Hallbook. 2006. Conserved synteny between the *Ciona* genome and human paralogons identifies large duplication events in the molecular evolution of the *insulin-relaxin* gene family. Mol. Biol. Evol. 23:10-22.

Ovcharenko, I., G. G. Loots, B. M. Giardine, M. Hou, J. Ma, R. C. Hardison, L. Stubbs, and W. Miller. 2005. Mulan: multiple-sequence local alignment and visualization for studying function and evolution. Genome Res. 15:184-194.

Panopoulou, G., S. Hennig, D. Groth, A. Krause, A. J. Poustka, R. Herwig, M. Vingron, and H. Lehrach. 2003. New evidence for

genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. Genome Res. 13:1056-1066.

Peifer, M., F. Karch, and W. Bender. 1987. The bithorax complex: Control of segmental identity. Genes & Dev. 1: 891–898.

Pennisi, E. 2001. Molecular evolution. Genome duplications: the stuff of evolution? Science 294: 2458-2460.

Santagati, F., K. Abe, V. Schmidt, T. Schmitt-John, M. Suzuki, K. Yamamura, and K. Imai. 2003. Identification of Cis-regulatory elements in the mouse *Pax9/Nkx2-9* genomic region: implication for evolutionary conserved synteny. Genetics 165:235-242.

Shimeld, S. M. 1997. Characterisation of amphioxus *HNF-3* genes: conserved expression in the notochord and floor plate. Dev. Biol. 183:74-85.

Shin, J. T., J. R. Priest, I. Ovcharenko, A. Ronco, R. K. Moore, C. G. Burns, and C. A. MacRae. 2005. Human-zebrafish non-coding conserved elements act in vivo to regulate transcription. Nucleic. Acids. Res. 33:5437-5445.

Siepel, A., G. Bejerano, J. S. Pedersen, et al. (16 co-authors). 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034-1050.

Simillion, C., K. Vandepoele, and Y. Van de Peer. 2004. Recent developments in computational approaches for uncovering genomic homology. Bioessays 26:1225-1235.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673-4680.

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

Wada, H. and N. Satoh. 1994. Details of the evolutionary history from invertebrates to vertebrates, as deduced from the sequences of 18S rDNA. Proc. Natl. Acad. Sci USA 91:1801-1804.

Wang, W., H. L. Xu, L. P. Lin, B. Su, and Y. Q. Wang. 2005. Construction of a BAC library for Chinese amphioxus *Branchiostoma belcheri* and identification of clones containing *Amphi-Pax* genes. Genes Genet. Syst. 80:233-236.

Woolfe, A., M. Goodson, D. K. Goode, et al. (11 co-authors). 2005. Highly conserved non-coding sequences are associated with vertebrate development. PloS. Biol. 3:e7

Yu, J. K., N. D. Holland, and L. Z. Holland. 2004. Tissue-specific expression of *FoxD* reporter constructs in amphioxus embryos. Dev. Biol. 274:452-461.

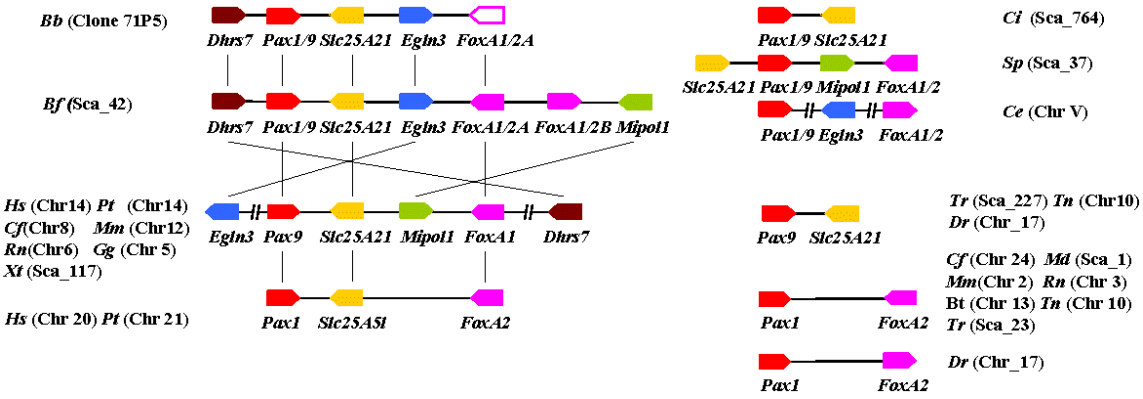

Fig. 1. ─Comparative map of the region encompassing syntenic genes.

Hexagons represent genes. Synteny conservation is highlighted by the same color between various species. Breaks indicate larger genomic distances with a number of intervening genes (not displayed). The gene name is indicated beneath the hexagon. The organism names and chromosome/scaffold numbers (in parentheses) are given beside. *FoxA1/2A* gene in *B. belcheri* is shown as unfilled hexagon because only UTR sequence was identified at the terminal of the BAC clone.

Homo sapiens: *Hs*; *Pan troglodytes*: *Pt*; *Canis familiaris*: *Cf*; *Bos taurus*: *Bt*; *Mus musculus*: *Mm*; *Rattus norvegicus*: *Rn*; *Monodelphis domestica*: *Md*; *Gallus gallus*: *Gg*; *Xenopus tropicalis*: *Xt*; *Danio rerio*: *Dr*; *Tetraodon nigroviridis*: *Tn*; *Takifugu rubripes*: *Tr*; *Ciona intestinalis*: *Ci*; *Strongylocentrotus purpuratus*: *Sp*; *Caenorhabditis elegans*: *Ce*; *Branchiostoma floridae*: *Bf*.

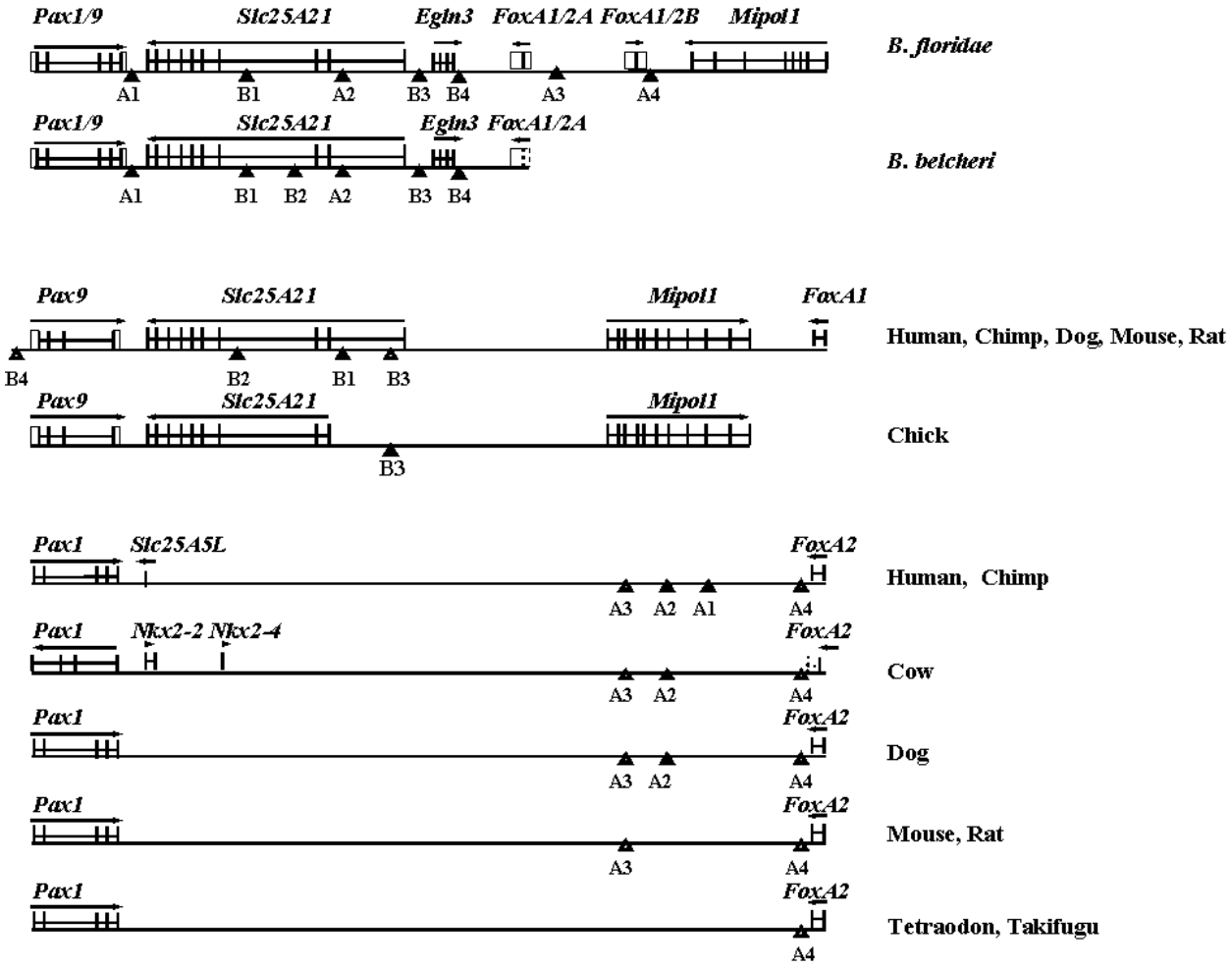

Fig. 2. ─Position of ncECRs in the syntenic region of amphioxus and vertebrates.

The arrowheads suggest the transcriptional direction of genes. Identified ncECRs shared between amphioxus and vertebrates are indicated by triangle. Open triangle: ncECR in the reverse strand respect to amphioxus. Filled triangle: ncECR in the same strand respect to amphioxus. Bold vertical line: exon. Open box: UTR. This scheme is not drawn to scale.

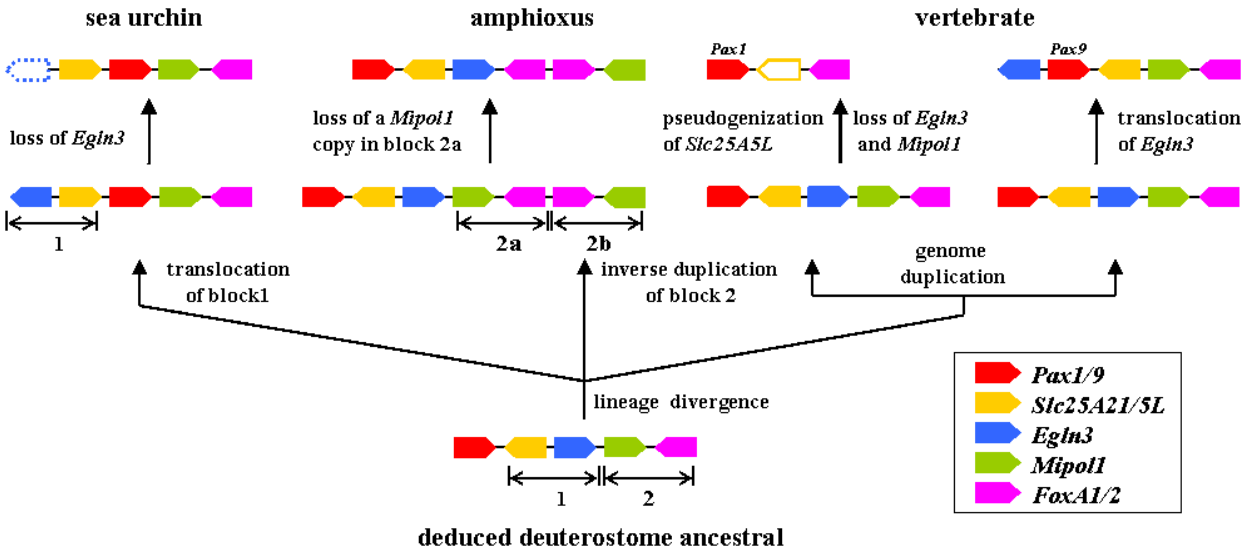

Fig. 3. —Evolution of *Pax1/9* ancient locus.

The cascade starts from the bottom (ancestral) and proceeds upwards to the contemporary blocks in sea urchin, amphioxus and vertebrates. Hexagons denote the genes belonging to the synteny group. The pseudogene *Slc25A5L* is shown as unfilled hexagon. The *Egln3* gene is postulated to present in the gap of present sea urchin genome assembly and indicated by unfilled hexagon of dotted line.