Comparative genomic analysis reveals the evolutionary conservation of *Pax* gene family

Wei Wang^{1,2}, Jing Zhong¹ and Yi-Quan Wang^{1*}

¹Key Laboratory of Ministry of Education for Coastal and Wetland Ecosystems,
School of Life Sciences, Xiamen University, Xiamen, 361005, China

²Key Laboratory of Marine Biogenetic Resources, Third Institute of Oceanography,
State Oceanic Administration, Xiamen 361005, China

(Received 16 February 2010, accepted 19 May 2010)

The *Pax* gene family encodes a group of transcription factors whose evolution has accompanied the major morphological and functional innovations of vertebrate species. The evolutionary conservation throughout diverse lineages of metazoan and the functional importance in development rendered *Pax* family an ideal system to address the relationship inside Chordata phylum. In the present study, we sequenced and annotated four genomic regions containing Chinese amphioxus (*Branchiostoma belcheri*) *Pax* genes, and retrieved homologous sequences from public database. In comparison with vertebrate homologues, the predicted amphioxus *Pax* proteins display high sequence conservation. Evidences from the molecular phylogenetic studies and gene organization analyses supports cephalochordates have a much closer relationship to vertebrates than that between tunicates and vertebrates, contrasting to urochordate relatives hypothesis proposed by several latest studies. Analysis of phylogenetic topology derived from concatenated subfamily datasets uncovered a potential statistical bias of supermatrix approach. Furthermore, we deduced an evolutionary scenario of *Pax* gene family. This scenario provided a plausible explanation for the origin and dynamics of the *Pax* gene members.

Key words: cephalochordate, comparative genome, evolution, *Pax* gene family

INTRODUCTION

Sharing chordate characters, vertebrates are appended with evolutionary novelties, including the acquisition and diversification of neural crest, placodes, and endoskeletal elements, as well as the reorganization and elaboration of the brain (Shimeld and Holland, 2000). It is now accepted that gene duplication that enables functional diversity of existing developmental genes has a causal relationship to the major evolutionary transition featured by dramatic increases in morphological complexity (Holland and Garcia-Fernandez, 1996; Dover, 2000). In this context, studies on families of developmental control genes are of particular interest, as they provide clues to the gene duplication events, as well as the functional diversity associated with the evolution of vertebrate innovations. One group of those genes that have drawn extensive attention is *Pax* gene family (Mazet et al., 2003;

Holland et al., 1995; Kozmik et al., 2003; Glardon et al., 1997; Hadrys et al., 2005; Chalepakidis et al., 1993).

Transcriptional regulators of *Pax* gene family are robust markers of several key structures that define the vertebrate clade, including the cranial and peripheral ganglia, pharyngeal arch, skeletomusculature, and definitive neural crest (Peters et al., 1998; Herbrand et al., 1998). The genes encode proteins characterized by a well-conserved DNA-binding domain (paired domain, PD) of 128 amino acids (aa), and are highly involved in animal embryonic patterning and organogenesis. Nine family members (denoted *Pax1* to *Pax9*) were identified in vertebrates encode proteins exhibiting highly conserved structure, genomic organization, expression patterns and biological functions (Underhill, 2000; Mansouri et al., 1996). *Pax* genes can be further classified into four subgroups: subgroup I (*Pax1*, 9), II (*Pax2*, 5, 8), III (*Pax3*, 7), and IV (*Pax4*, 6). Genes belonging to same subgroup display higher similarities in both expression pattern and function, which is suggestive of similar or partly overlapping roles in development. Experimental evidences show mutations cause profound developmental defects in organisms as diverse as zebrafish, mice and human

Edited by Fumio Tajima

* Corresponding author. E-mail: wangyq@xmu.edu.cn

Note: Nucleotide sequence data reported are available in the GenBank databases under the accession numbers DQ991501–DQ991504.

(Mackereth et al., 2005; Eccles et al., 2002). Further investigation of the gene expression pattern of chordates indicates the evolution of this gene family is highly relevant to the function and major morphological innovations along the basal chordate and vertebrate lineage. The genomic information contained in amphioxus *Pax* genes can be of key importance to address the question concerning how the expansion of regulatory gene family had led to the morphological innovations during the invertebrate-to-vertebrate transition.

In addition, evidences also support that the origination of *Pax* gene occurred shortly after the emergence of metazoans about one billion years ago (Breitling and Gerber, 2000). Homologues have been cloned from some primitive organisms including coral *Acropora millepora* (Cnidaria) (Miller et al., 2000), tunicates *Ciona intestinalis* (Urochordata) (Wada et al., 2003) and amphioxus *Branchiostoma floridae* (Cephalochordata) (Holland et al., 1995; Holland et al., 1999). It is generally agreed that the complexity of *Pax* subfamilies in vertebrates has arisen by gene duplication after the cephalochordate-vertebrate split, and each of the subfamily is represented by a single *Amphi-Pax* gene (Holland et al., 1995; Kozmik et al., 1999; Glardon et al., 1998; Holland et al., 1999). The functional domains in each class of vertebrate *Pax* genes are conserved between amphioxus and vertebrates, and moreover, they are conserved among all vertebrate duplicates, suggesting strong evolutionary constraints for maintaining the particular domain combinations even after gene duplication (Short and Holland, 2008). Amphioxus possesses the unduplicated *Pax* complements with characteristic but simplified features of vertebrate genome organization, and therefore, offers a suitable bridge between vertebrates and basal organisms like protostomes. The genomic information contained in amphioxus *Pax* genes will help to address the gene family expansion led to the morphological innovations during the invertebrate-to-vertebrate transition and the evolutionary relationship between Cephalochordata and Vertebrata. Although the genomic sequence of one of three frequently studied amphioxus *B. floridae* is recently released in JGI, we sequence four BAC clones containing *Amphi-Pax* genes of Chinese amphioxus *B. belcheri* in the present study for two reasons: Firstly, the database information is of insufficient quality, and secondly, the amphioxus *B. belcheri* as another frequently adopted species distributing in the West Pacific Ocean is distantly related to *B. floridae* in genetics in despite of both species are taxonomically put in same genus (Zhong et al., 2009). So, this investigation will provide additional sequence information of amphioxus genome and more insights into to chordate phylogeny.

MATERIALS AND METHODS

Sequencing and annotation of BAC clones containing *BbPax* genes Four clones (BAC 71P5, 114D9, 80p18 and 89L24), each containing a complete *Pax* gene (*BbPax* gene) of Chinese amphioxus (*B. belcheri*), were selected from the BAC library constructed previously (Wang et al., 2005). Plasmids with insert of amphioxus genomic DNA were isolated from those clones by alkaline lysis/phenol:chloroform extraction, ethanol precipitation, and then, mechanically fragmented to generate random fragments. These fragments were subcloned into appropriate library vectors and sequenced. Sequences were analyzed and assembled with the PHRED/PHRAP/CONSED package (University of Washington, Seattle, WA, USA; <http://www.phred.org>). The initial assembling of the sequences resulted in several contigs and the relative order and orientation were established by standard PCR amplifications using primer pair specific to each end of the contigs in various combinations. Gap filling were performed by amplicons sequencing to yield contiguous DNA sequence covering the full *BbPax* coding region for each BAC clone.

For the purpose of annotation, we adopted program GENESCAN to predict potential genes (<http://bioweb.pasteur.fr/seqanal/interfaces/genscan.html>). The entire sequence or predicted gene was further analyzed by homology search against NCBI databases (<http://ncbi.nlm.nih.gov/BLAST/>) with BLASTN, BLASTX, and TBLASTX algorithms.

Sequences retrieval from public database Sequences of previously reported *Pax* genes were retrieved through an initial PSI-BLAST search of the non-redundant database at NCBI. Genomic structure analysis and additional gene predictions were conducted through database search (JGI, <http://www.jgi.doe.gov>; UCSC <http://genome.cse.ucsc.edu/> and Ensembl, <http://www.ensembl.org/>) against the genome assemblages of 17 species including two primates (human *Homo sapiens* and chimpanzee *Pan troglodytes*), one carnivora (dog *Canis familiaris*), one artiodactyla (cow *Bos taurus*), two rodents (mouse *Mus musculus* and rat *Rattus norvegicus*), one marsupial (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 cephalochordate (amphioxus *Branchiostoma floridae*), one urochordate (ascidian *Ciona intestinalis*), one echinoderms (sea urchin *Strongylocentrotus purpuratus*), one insect (fruit fly *Drosophila melanogaster*), and one worm (nematode *Caenorhabditis elegans*). The two-letter abbreviations, initial letter of genus and species names, are designed to each species. First, we employed TBLASTN to identify

the genomic locations of putative *Pax* genes in a genome with previously reported *Pax* peptide sequences as queries, and then applied the genomic DNA sequence of a putative *Pax* gene and known cDNA sequences to conduct a cDNA-to-genomic sequence alignment using Genewise (<http://www.ebi.ac.uk/Wise2/>), which provided exon/intron structure and full-length protein sequence of a putative gene. Finally, we examined deduced protein sequences and excluded those with a large proportion of missing data or incomplete paired domain for further analysis. It should be noted that some described *Pax* cDNAs did not match exactly with the genome sequences. In such cases, we adopted gene sequences derived from the genome database only.

Phylogenetic analysis and genomic comparison of *Pax* genes

The deduced amino acid sequences were aligned using Clustal W program (Thompson et al., 1994) with default settings, and manually adjusted to improve accuracy of alignment. The phylogenetic trees of *Pax* gene family were constructed using neighbor-joining (p-distance) and maximum parsimony (standard parsimony) method implemented in the program MEGA version 3.1 (Kumar et al., 2004) with *Pseudomonas* transposase sequence as out-group (Breitling and Gerber, 2000). Bootstrapping values were estimated using 1000 replications for the measure of confidence. In order to test the closest relative to vertebrates, the four-cluster likelihood mapping analyses on the whole protein sequence data of each individual subfamily were performed with TREE-PUZZLE program (Schmidt et al., 2002, <http://www.tree-puzzle.de>).

Web site references <http://bioweb.pasteur.fr/seqanal/>

[interfaces/genscan.html](http://www.genscan.html); Genscan—Gene identification program.

<http://www.ensembl.org>; Ensembl genome browser.

<http://www.ebi.ac.uk/Wise2/>; Genewise—Intelligent algorithms for DNA searches (EBI).

<http://www.ncbi.nlm.nih.gov/BLAST>; NCBI BLAST.

<http://www.phred.org>; PHRED/PHRAP/CONSED package—DNA calling, assembling and editing program.

<http://www.jgi.doe.gov>; JGI (Joint Genome Institute) genome browser.

<http://genome.cse.ucsc.edu/> University; UCSC (California at Santa Cruz) genome browser.

<http://www.tree-puzzle.de>; tree-puzzle: Maximum likelihood phylogenetic analysis program.

RESULTS

Sequence and characterization of *BbPax* genes Four BAC clones each encompassing one full coding region of *BbPax* genes were sequenced. After assembling, 132 kb of sequence with 4 gaps organized into five contigs was obtained from the clone 71P1, 111 kb of sequence with 1 gap in two contigs from 114A21, 106 kb of sequence with 7 gaps in eight contigs from 80P18, and no gap in 115 kb of sequence from 89L24 (Fig. 1, GenBank accession numbers DQ991501–DQ991504). Addition to four *Pax* genes, total of 30 genes were also predicted across the four BAC clones using GENESCAN. Although most of those genes are not considered to be true genes due to no match or low similarity to known genes or ESTs (expect value higher than 10^{-10}) in the GenBank database, four of them exhibited significant homology to those known genes: dehydrogenase/reductase family member 7 (*Dhrs7*), egg laying defective nine homolog 3 (*Egln3*), solute carrier family 25

Fig. 1. Structure and gene content of the BAC clones. Hollow arrowhead and filled arrowhead represent the orientation of genes and contigs respectively. The diagram is not drawn to scale.

member 21(*Slc25A21*) and *Actin* genes. The former three located at the adjacent region of *BbPax1/9* in clone 71P1 and the last one is neighboring to *Pax 3/7* in clone 80P18. In addition, similarity search revealed a 649 bp BAC end sequence of clone 71P1 is absolutely identical to 3'UTR of *B. floridae* *AmHNF-3-1* mRNA (hepatocyte nuclear factor-3, X96519), indicating the presence of this gene immediately beside the genomic stretch contained in clone 71P1. Since the nomenclature of *HNF-3* family has been revised to *FoxA* genes (Kaestner et al., 2000), we refer amphioxus *HNF-3-1* gene as *FoxA1A* hereafter.

Comparison between genomic sequences and four cDNAs of *B. floridae* *Pax* genes (*BfPax*) allowed us to deduce the coding sequence, the intron/exon boundaries of *BbPax* genes and, furthermore, to compare them with the homologues in other species. The deduced proteins of *BbPax1/9*, *BbPax2/5/8*, *BbPax3/7* and *BbPax6* are 363, 444, 470 and 461 aa in length separately, comparing to 363, 445, 464 and 463 aa of their *BfPax* counterparts, all of which possess a diagnostic domain of corresponding *Pax* subfamily. Amino acid identities of the four genes are 97.2%, 98.9%, 70.1% and 97.0% between the two species of amphioxus. Actually, the lower similarity between two *Pax3/7* proteins is owing to relative high divergence at C-terminal part of the proteins. The similarity value goes up to 89% after truncating this portion. The *BbPax1/9*, *BbPax2/5/8*, *BbPax3/7* and *BbPax6* span 15.9 kb, 68.1 kb, 13.1 kb, 38.2 kb in the genome and compose of 5, 9, 5 and 11 exons respectively. The intron sizes range from 357 bp (*BbPax6*, intron 5) to 20,935 bp (*BbPax2/5/8*, intron 1), and all intron/exon organization of four genes is conserved between two species of amphioxus despite little variance in length among the genes.

***Pax* genes identified from public database** To examine the phylogenetic distribution of *Pax* gene family, we conducted PSI-blast search on GenBank and TBLASTN search against the genomic assemblies of variety species (from *C. elegans* to mammals, see MATERIALS AND METHODS). The identified genes (Table 1) were assigned to subfamilies based on BLAST similarity score, and then annotated according to the previously characterized genes. We cannot exclude the possibility that those genomes contain additional genes in gaps in the initial assemblies.

The vertebrate *Pax* genes are subdivided into four classes based on the presence or absence of motifs in addition to paired domain (Robson et al., 2006). The class I contains *Pax1* and *Pax9* that encode an octapeptide but lack a homeodomain. The class II *Pax* genes, *Pax2*, *Pax5*, and *Pax8*, encodes the octapeptide and a partial homeodomain. The class III genes, *Pax3* and *Pax7*, encoding the octapeptide and a full homeodomain, and the class IV *Pax* genes, *Pax4* and *Pax6*, encoding the full

homeodomain but missing the octapeptide. Most of the invertebrates *Pax* genes fall into the four classes based on the comparisons of domain structure and sequences.

In addition to these four classes that include orthologs from a wide range of animals, many "orphan" *Pax* genes are found in invertebrates. Similarity searching against genomic assemblies identified two genes encoding uncanonical paired domains. One is from sea urchin *S. purpuratus*, the other from *B. floridae*. These two predicted genes probably do not contain other functional domains since no sequence encoding octapeptide or homeodomain could be identified in the vicinity of the fragments. BLASTP search revealed that the two genes hit the *Poxneuro* gene from hemichordate species *Saccoglossus kowalevskii* (ABD97270) best. Considering of deduced protein sequence of our newly identified gene from *S. purpuratus* is same to that of *PaxA* protein segment (AAB69869) of *Paracentrotus lividus* (Echinoderms), we referred this gene as *SpPaxA* and its orthologue of *B. floridae* as *BfPaxA*.

In addition to *PaxA*, *Pax* genes in many other invertebrates displayed atypical functional domain combination. Some *Pax* genes miss one of functional domains, e.g., the octapeptide of *Pax1/9* subfamily members in *C. elegans* (*CePax19*, U53336), *D. melanogaster* (*DmPoxmeso*, NM_079552), and that of *Pax3/7* subfamily members in *D. melanogaster* (*DmPrd*, NM_078832) and *C. intestinalis* (*CiPax3/7*, AB210632); the homeodomain of *Pax2/5/8* subfamily members in *C. elegans* (*CePax2/5/8*, U50072; *CeEgl38* NP_501836), *D. melanogaster* (*DmSparkling*, AF010256) and *C. intestinalis* (*CiPax2/5/8B*). There are also two genes of *Pax2/5/8* subfamily miss both homeodomain and octapeptide. One is the gene in *C. intestinalis* (*CiPax2/5/8A*, NM_001032480), the other in *S. purpuratus* (*SpPaxB*, AF016886).

A total of 12 lineage specific gene duplications are identified, eight of which occurred exclusively in teleost fish (*Pax2A/B*, *Pax3A/B* and *Pax6A/B* in pufferfish, *Pax1A/B*, *Pax5A/B* and *Pax6A1/2* and *Pax6B1/2* in zebrafish), two in *D. melanogaster* (*GsbP/GsbD* and *Eyeless/Toy*), one in *C. elegans* (*CeEgl38/CePax2/5/8*) and *C. intestinalis* (*Pax2/5/8A/B*). Some of the predicted copies were not mapped to chromosome (e.g. *Pax1B* of zebrafish) yet. To ensure the two gene copies are not resulted from sequence polymorphism, we compared their genomic structure and environment. The results indicate that the paralogues of *D. melanogaster* and *C. elegans* are all located in the same chromosome with varying distance (from 9.8 kb to 5.9 Mb). This is also the case of *Pax5A/B* (353 kb apart) and *Pax6A2/B2* (11.8 Mb apart) in zebrafish. The distantly associated paralogues are probably raised from tandem duplication and separated by local genomic rearrangement.

Genomic organization of *Pax* genes Most of the pre-

Table 1. *Pax* genes from major model species

Gene name	Database	Ensembl gene name	Chrom/ Scaf	PD	OP	Truncated HD	Full HD
<i>CePax1/9</i>	Ensembl	K07C11.1	V	X			
<i>DmPoxmeso</i>	Ensembl	CG8246	2R	X			
<i>SpPax1/9</i>	UCSC	None	Sca_37	X	X		
<i>CiPax1/9</i>	Ensembl	ENSCING00000002978	Sca_145	X	X		
<i>BbPax1/9</i>	Present study	None	Clo_71P1_Con_145	X	X		
<i>BfPax1/9</i>	JGI	None	Sca_42	X	X		
<i>HsPax1</i>	Ensembl	ENSG00000125813	20	X	X		
<i>MmPax1</i>	Ensembl	ENSMUST00000047382	2	X	X		
<i>RnPax1</i>	Ensembl	ENSRNOG00000024882	3	X	X		
<i>PtPax1</i>				X	X		
<i>CfPax1</i>	Ensembl	ENSACFG00000005170	24	X	X		
<i>BtPax1</i>	Ensembl	ENSBTAG00000006742	13	X	X		
<i>MdPax1</i>	Ensembl	ENSMODG00000006198	Sca_1	X	X		
<i>XtPax1</i>	Ensembl	ENSXETG00000023497	Sca_352	X	X		
<i>DrPax1A</i>	UCSC	None	17	X	X		
<i>DrPax1B</i>	UCSC	None	Unknown	X	X		
<i>TnPax1</i>	Ensembl	GSTENG00031343001	10	X	X		
<i>TrPax1</i>	Ensembl	SINFRUG00000125776	Sca_1159	X	X		
<i>HsPax9</i>	Ensembl	ENSG00000198807	14	X	X		
<i>MmPax9</i>	Ensembl	ENSMUSG00000001497	12	X	X		
<i>RnPax9</i>	Ensembl	ENSRNOG00000008826	6	X	X		
<i>CfPax9</i>	Ensembl	ENSACFG00000013690	8	X	X		
<i>BtPax9</i>	Ensembl	ENSBTAG00000009186	Sca_10035	X	X		
<i>MdPax9</i>	Ensembl	ENSMODG00000012681	Sca_73	X	X		
<i>GgPax9</i>	Ensembl	ENSGALG00000010114	5	X	X		
<i>XtPax9</i>	UCSC	None	Sca_117	X	X		
<i>DrPax9</i>	Ensembl	ENSARG00000019354	17	X	X		
<i>TrPax9</i>	UCSC	None	Unknown	X	X		
<i>TnPax9</i>	Ensembl	GSTENP00035716001	10	X	X		
<i>CeEgl38</i>	Ensembl	C04G2.7	IV	X	X		
<i>CePax2/5/8</i>	Ensembl	K06B9.5	IV	X	X		
<i>DmSparkling</i>	Ensembl	CG11049	4	X	X		
<i>SpPax2/5/8</i>	UCSC	None	Sca_75861	X	X	X	
continued	UCSC	None	Sca_1593				
<i>SpPaxB</i>	UCSC	None	Sca_773	X			
<i>CiPax2/5/8-A</i>	Ensembl	ENSCING00000006915	Sca_368	X			
<i>CiPax2/5/8-B</i>	UCSC	None	Sca_140	X	X		
<i>BbPax2/5/8</i>	Present study	None	Clo_114A21_Con_5	X	X	X	
<i>BfPax2/5/8</i>	JGI	None	Sca_131	X	X	X	
<i>HsPax2</i>	Ensembl	ENSG00000075891	10	X	X	X	
<i>PtPax2</i>	UCSC	None	8	X	X	X	
<i>BtPax2</i>	Ensembl	ENSBTAG00000021566	Sca_936	X	X	X	
<i>MmPax2</i>	Ensembl	ENSMUSG00000004231	19	X	X	X	
<i>RnPax2</i>	Ensembl	ENSRNOG00000014253	1	X	X	X	
<i>XtPax2</i>	Ensembl	ENSXETG00000014206	Sca_186	X	X	X	
<i>GgPax2</i>	Ensembl	ENSGALG00000005689	6	X	X	X	
<i>TrPax2A</i>	UCSC	None	Unknown	X	X	X	
<i>TnPax2A</i>	Ensembl	GSTENG00016448001	17	X	X	X	
<i>MdPax2</i>	UCSC	None	Sca_19214	X	X	X	
<i>CfPax2</i>	UCSC	None	28	X	X	X	

Continued

Table 1. Continued

<i>TrPax2B</i>	Ensembl	SINFRUG00000134443	Sca_37	X	X	X
<i>TnPax2B</i>	UCSC	None	2	X	X	X
<i>DrPax2</i>	Ensembl	ENSDARG00000032578	Sca_Zv5_NA10828	X	X	X
<i>HsPax5</i>	Ensembl	ENSG00000196092	9	X	X	X
<i>PtPax5</i>	UCSC	None	11	X	X	X
<i>CfPax5</i>	Ensembl	ENSCAFG00000002357	11	X	X	X
<i>BtPax5</i>	Ensembl	ENSBTAG00000012498	Sca_10220	X	X	X
<i>MmPax5</i>	Ensembl	ENSMUSG00000014030	4	X	X	X
<i>MdPax5</i>	Ensembl	ENSMODG00000002720	Sca_19	X	X	X
<i>GgPax5</i>	UCSC	None	2	X	X	X
<i>XtPax5</i>				X	X	X
<i>DrPax5A</i>	UCSC	None	1	X	X	X
<i>DrPax5B</i>	UCSC	None	1	X	X	X
<i>TrPax5</i>	Ensembl	SINFRUG00000147084	Sca_1125	X	X	X
<i>TnPax5</i>	UCSC	None	18	X	X	X
<i>HsPax8</i>	Ensembl	ENSG00000125618	2	X	X	X
<i>CfPax8</i>	Ensembl	ENSCAFG00000007309	17	X	X	X
<i>BtPax8</i>	Ensembl	ENSBTAG00000019354	11	X	X	X
<i>MmPax8</i>	Ensembl	ENSMUSG00000026976	2	X	X	X
<i>RnPax8</i>	Ensembl	ENSRNOG00000026203	3	X	X	X
<i>MdPax8</i>	Ensembl	ENSMODG00000013770	Sca_63	X	X	X
<i>XtPax8</i>	Ensembl	ENSXETG00000001272	Sca_217	X	X	X
<i>TrPax8</i>	Ensembl	SINFRUG00000123885	Sca_27	X	X	X
<i>TnPax8</i>	UCSC	None	4	X	X	X
<i>DrPax8</i>	Ensembl	ENSDARG00000015879	1	X	X	X
<i>CePax3/7</i>	UCSC	None	2	X	X	X
<i>DmGsbP</i>	Ensembl	CG3388	2R	X	X	X
<i>DmGsbD</i>	Ensembl	CG2692	2R	X	X	X
<i>DmPrd</i>	Ensembl	CG6716	2L	X		X
<i>CiPax3/7</i>	Ensembl	ENSCING00000005904	Sca_134	X		X
<i>BbPax3/7</i>	Present study	None	Clo_80P18_Con_18	X	X	X
<i>BfPax3/7</i>	JGI	None	Sca_259	X	X	X
<i>HsPax3</i>	Ensembl	ENSG00000135903	2	X	X	X
<i>PtPax3</i>	Ensembl	ENSPTRG00000012969	13	X	X	X
<i>BtPax3</i>	Ensembl	ENSBTAG00000009950	sca_628	X	X	X
<i>MmPax3</i>	Ensembl	ENSMUSG00000004872	1	X	X	X
<i>RnPax3</i>	Ensembl	ENSRNOG00000013670	9	X	X	X
<i>MdPax3</i>	Ensembl	ENSMODG000000021190	Sca_28	X	X	X
<i>GgPax3</i>	UCSC	None	Unknown	X	X	X
continued	UCSC	None	9			
<i>XtPax3</i>	Ensembl	ENSXETG00000007640	Sca_209	X	X	X
<i>CfPax3</i>	Ensembl	ENSCAFG00000016041	37	X	X	X
<i>TrPax3A</i>	Ensembl	SINFRUG00000161042	Sca_123	X	X	X
<i>TrPax3B</i>	Ensembl	SINFRUG00000132501	Sca_73	X	X	X
<i>TnPax3A</i>	Ensembl	GSTENG00026741001	16	X	X	X
<i>TnPax3B</i>	Ensembl	GSTENG00028437001	15	X	X	X
<i>DrPax3</i>	Ensembl	ENSDARG00000010192	2	X	X	X
<i>CfPax7</i>	UCSC	None	2	X	X	X
<i>BtPax7</i>	Ensembl	ENSBTAG00000014967	2	X	X	X
<i>MmPax7</i>	Ensembl	ENSMUSG00000028736	4	X	X	X
<i>RnPax7</i>	Ensembl	ENSRNOG00000018739	5	X	X	X
<i>MdPax7</i>	Ensembl	ENSMODG00000002540	Sca_109	X	X	X

Continued

Table 1. Continued

continued	Ensembl	ENSMODG00000020885	Sca_276			
<i>GgPax7</i>	UCSC	None	Unknown	X	X	X
continued	UCSC	None	21			
<i>XtPax7</i>	Ensembl	ENSXETG00000018527	Sca_930	X	X	X
<i>DrPax7</i>	Ensembl	ENSDARG00000005173	23	X	X	X
<i>TrPax7</i>	ensembl	NEWSINFRUG00000129602	Sca_56	X	X	X
<i>TnPax7</i>	Ensembl	GSTENG00027574001	11	X	X	X
<i>HsPax7</i>	Ensembl	ENSG00000009709	1	X	X	X
<i>PtPax7</i>	Ensembl	ENSPTRG00000000253	1	X	X	X
<i>CePax6</i>	Ensembl	F14F3.1	X	X		X
<i>DmEyeless</i>	Ensembl	CG1464	4	X		X
<i>DmToy</i>	Ensembl	CG11186	4	X		X
<i>SpPax6</i>	UCSC	None	Sca_160	X		X
<i>CiPax6</i>	Ensembl	ENSCING00000006503	Sca_7	X		X
<i>BbPax6</i>	Present study	None	Clo_89L24_Con_28	X		X
<i>BfPax6</i>	JGI	None	Sca_23	X		X
<i>HsPax4</i>	Ensembl	ENSG00000106331	7	X		X
<i>PtPax4</i>	UCSC	None	6	X		X
<i>CfPax4</i>	Ensembl	ENSCAFG00000001719	14	X		X
<i>BtPax4</i>	Ensembl	ENSBTAG00000000410	Sca_200442	X		X
<i>MmPax4</i>	Ensembl	ENSMUSG00000029706	6	X		X
<i>RnPax4</i>	Ensembl	ENSRNOG00000008020	4	X		X
<i>TrPax4</i>	UCSC	None	Unknown	X		X
<i>TnPax4</i>	UCSC	None	19	X		X
<i>HsPax6</i>	Ensembl	ENSG00000007372	11	X		X
<i>BtPax6</i>	Ensembl	ENSBTAG00000004561	Sca_124	X		X
<i>CfPax6</i>	Ensembl	ENSCAFG00000007491	18	X		X
<i>MmPax6</i>	Ensembl	ENSMUSG000000027168	2	X		X
<i>RnPax6</i>	Ensembl	ENSRNOG00000004410	3	X		X
<i>MdPax6</i>	Ensembl	ENSMODG00000009611	Sca_10	X		X
<i>GgPax6</i>	Ensembl	ENSGALG00000012123	3	X		X
<i>XtPax6</i>	Ensembl	ENSXETG00000008175	Sca_624	X		X
<i>TrPax6A</i>	Ensembl	SINFRUG00000121553	Sca_227	X		X
<i>TrPax6B</i>	Ensembl	SINFRUP00000160172	Sca_430	X		X
<i>TnPax6A</i>	Ensembl	GSTENG00020822001	5	X		X
<i>TnPax6B</i>	UCSC	None	Unknown	X		X
<i>DrPax6B1</i>	Ensembl	ENSDARG00000016640	7	X		X
<i>DrPax6B2</i>	Ensembl	ENSDARG00000045936	25	X		X
<i>DrPax6A1</i>	Ensembl	ENSDARG00000045045	12	X		X
<i>DrPax6A2</i>	Ensembl	ENSDARG00000045746	25	X		X
<i>DmPoxneuro</i>	Ensembl	CG8246	2R	X		
<i>SpPaxA</i>	UCSC	None	Sca_487	X		
<i>BfPaxA</i>	JGI	None	Sca_404	X		

Note: Due to incomplete genome sequences of some species and lack of the linkage between *Pax* bearing region and chromosome number, Scaffold numbers are displayed instead in the table. In the cases of *Pax* gene spanning two scaffold (e.g. *MdPax7*, *SpPax2/5/8*), the gene name is marked with "continued" in the second portion. Structural domains include paired domain (PD), octapeptide motif (OP) and homeodomain (HD). *Pax2/5/8* subfamily contained truncated homeodomain, whereas *Pax3/7* and *Pax4/6* subfamilies have a full homeodomain.

viously reported deuterostome *Pax* genes have single short exon (ranging from 4–85 bp) immediately before the paired domain and 2–5 exons at downstream region where other characteristic domains locate. In each subfamily, position of intron is highly conserved among

vertebrate orthologues and comparable to that of non-vertebrate deuterostomes, but largely variant to genes from protostomes.

The *Pax1/9* subfamily, consisting of 4 (vertebrate *Pax9*) or 5 (protochordates *Pax1/9* and vertebrate *Pax1*) exons

with two characteristic functional domains (paired domain and the octapeptide) residing in the same exon, is the least complex among the four subfamilies. The *Pax1/9* in *D. melanogaster* (*DmPoxmeso*) and *C. elegans* (*CePax1/9*) lack the octapeptide.

The *Pax2/5/8* subfamily genes encode proteins containing a paired domain, an octapeptide and a partial homeodomain. Generally, the paired domain is encoded by three exons. Furthermore, the C-terminal portion of the paired domain, the octapeptide and the homeodomain of vertebrate *Pax2/5/8* are located in three separate exons (exon 4, 5 and 7). However, the last portion of paired domain and octapeptide of invertebrate *Pax2/5/8* is encoded by a single exon (exon 4/5), suggesting that vertebrate exon 4 and 5 originated from ancestral *Pax2/5/8* gene via inserting an intron between the two exons. The last two functional domains in *Pax2/5/8* subfamily are absent in six invertebrate homologous genes. Among them four genes (*CePax2/5/8* and *CeEgl38*, *DmSparkling* and *CiPax2/5/8B*) lack of the homeodomain and the other two (*CiPax2/5/8A* and *SpPaxB*) miss both homeodomain and octapeptide. There are three introns in *Pax2/5/8* genes (*CeEgl38* and *CePax2/5/8*) of *C. elegans*, but two introns within the paired domain in all other members of the subfamily.

The *Pax3/7* is the only subfamily possessing a complete set of functional domains among *Pax* genes. In vertebrates, exon 2, 3 and 4 encode paired domain and a portion of exon 4 also encode octapeptide. This scheme differs from that of all invertebrate *Pax3/7* genes. In amphioxus, single exon (exon 1/2/3/4) encodes the complete paired domain and octapeptide. The homeodomain encoded by single exon in four protostome genes (*CePax3/7* and *DmGsbP*, *DmGsbD* and *DmPrd*) is split into two separate exons (exon 5 and 6) in chordate genes. Two of invertebrate *Pax3/7* genes (*CiPax3/7* and *DmPrd*) lack of octapeptide. Two introns locating at residue 74/75 and 116/117 within the paired domain of vertebrates are absent in all invertebrate *Pax3/7* genes, whereas the intron between residues 46/47 in the homeodomain is conserved throughout all chordate *Pax3/7* genes.

Pax4/6 subfamily has two functional domains, the paired domain and the complete homeodomain, both of which are split by two introns in almost all referred species. *Pax6* genes of some vertebrates (human, mouse, chicken, and zebrafish) have an alternatively spliced form with a 14 amino acid encoded by one additional exon (exon3) inserted in the paired domain. Inspection on the genomic sequences revealed that this exon was present in most vertebrate *Pax6* genes but missing from all invertebrate genes, suggesting the alternative form was vertebrate-specific and generated after their divergence from other chordates. Among the four introns distributed in the functional domains, three have very conserved positions in both vertebrate and invertebrate lineages, the

other one in the paired domain have no consistent placement between the two lineages.

In addition, three genes retrieved from public database that cannot be clearly assigned into the four canonical *Pax* subfamilies. They are *DmPoxneuro* from *D. melanogaster*, *SpPaxA* from *S. purpuratus* and *BfPaxA* from *B. floridae*. These genes display highly homologous to each other and exhibited similar genomic characteristics, encoding a single paired domain which possessing a common intron position between residues 74/75, suggesting they are probably derived from a common ancestor.

Phylogenetic analysis of the *Pax* gene family The paired domain of *Pax* genes is highly conserved throughout all examined lineages. In the present study, we construct a NJ tree and a MP tree based on the amino acid sequence of the paired domain using *Pseudomonas* transposase as out-group. Since two trees display a similar topological character, we show the MP tree here as a representative for further discussion (Fig. 2). In general, the members of *Pax* gene family are divided into four major clades, corresponding four subfamilies. The gene from a coral *A. millepora*, namely *AmPaxD*, takes the most basal position in the tree, probably representing a common ancestor of all *Pax* genes.

Among the four major clades, the first one branches out at base of the tree, composing of nematode *C. elegans* (*CePax3/7*), tunicate *C. intestinalis* (*CiPax3/7*) and amphioxus (*BbPax3/7* and *BfPax3/7*), fruit fly *D. melanogaster* (*DmGsbD*, *DmGsbP* and *DmPrd*) and vertebrate (*Pax3* and *Pax7*) genes. *CePax3/7* first branches out at the clade base, representing the most ancient form in the *Pax3/7* subfamily. Whereas all vertebrate *Pax3* and *Pax7* genes gather together at the top of the clade forming a monophyly with *BbPax3/7* and *BfPax3/7*. In vertebrate lineage, the genes are further divided into two subgroups, *Pax3* and *Pax7* clusters. The genes of tunicate and fruit fly compose another cluster out of amphioxus-vertebrate cluster.

The second major clade composes of *Pax1/9* subfamily genes, including genes of invertebrate species *D. melanogaster* (*DmMeso*), *C. elegans* (*CePax1/9*), sea urchin *S. purpuratus* (*SpPax1/9*), *C. intestinalis* (*CiPax1/9*), amphioxus (*BbPax1/9* and *BfPax1/9*), and two vertebrate genes (*Pax1* and *Pax9*). Two protostome (fruit fly and nematode) genes branch out at the base against all other genes in this clade, and protochordates (amphioxus and tunicate) genes constitute a sister group to the lineage of vertebrates *Pax1* and *Pax9* genes. Those genes of vertebrates also subsequently form two separate clusters.

The third major clade is unexceptionally constituted of *Pax4/6* subfamily genes, containing all *Pax4* and *Pax6* genes of both vertebrates and invertebrates. The topol-

Fig. 2. Maximum parsimony tree based on pairwise comparison of the paired domains of Pax proteins. A pseudomonas transposase sequence (AF169828) serves as outgroup (Breitling and Gerber, 2000). The four major clades are colored for easy comparison and the color corresponding to the subgroups are annotated above the tree.

ogy of this clade is particular since it neither conforms the pattern observed in *Pax3/7* and *Pax1/9* clade nor agrees with traditional interpretation of taxonomic phylogeny. Two *Pax4* genes branches, one composing of teleost genes and the other of tetrapod genes, are placed at the base of the major clade. The third branch containing only all *Pax6* genes of both vertebrates and invertebrates is further divided into two subgroups, one including the *Pax6* genes from *C. elegans* (*CePax6*), *C. intestinalis* (*CiPax6*), *S. purpuratus* (*SpPax6*), amphioxus (*BfPax6* and *BbPax6*) as well as two species of puffer fish (*TnPax6*, *TrPax6*), and other composing of *Pax6* genes from most vertebrates. Interestingly, instead of gathering into the first subgroup mainly constituted by invertebrates, two *Pax* genes of *D. melanogaster*, *DmEyeless* and *DmToy*, clustered together with vertebrates.

The last major clade divided into two subgroups. The first subgroup at the base of this clade is composed of two

sister clusters, one containing the genes *DmPoxneuro* (*D. melanogaster*), *SkPoxneuro* (*S. kowalevskii*), *BfPaxA* (*B. floridae*) and *SpPaxA* (*S. purpuratus*), the other consisting of the genes of *HfPaxA* (hydra *Hydra littoralis*), *CePaxA* (jellyfish *Cladonema californicum*) and *AmPaxA*, *AmPaxC* (coral *A. millepora*). In the second subgroup, *EfPax2/5/8* (sponge *Ephydatia fluviatilis*) and *SpPaxB* (*S. purpuratus*) genes branch out independently at the basal place, whereas all members of vertebrate *Pax2/5/8* genes gathering closely occupy the top of the tree. Those vertebrate genes are again divided into three branches, respectively representing *Pax2*, *Pax5* and *Pax8* genes of vertebrates. The *Pax2/5/8* genes from most invertebrates form four clusters locating between the sponge and vertebrate.

To resolve the phylogenetic relationship among three subphyla of chordate, we employed the Four-Cluster likelihood mapping approach on the analysis of the whole pro-

Fig. 3. Four-cluster likelihood-mapping of Pax proteins. A: Hypotheses of the phylogenetic relationship among four predefined groups of phylum. B: likelihood-mapping results, represented as a barycentric triangle showing the likelihood support for the three alternative topologies indicated by the numbers outside corners of the triangle corresponding to the topologies shown in Fig. 3A. Values in the corners indicate the percentage of fully resolved quartet topologies, numbers in the rectangular sections give the percentage of partially resolved topologies, and the value at the center of the triangle represents the percentage of unresolved topologies. The cumulatively high percentage from the corner values indicates the data sets are phylogenetically informative.

tein sequence of each subfamily. Since phylogenetic analysis did not support *Pax4* genes descended directly from protochordata *Pax6* gene like ancestor, only vertebrate *Pax6* genes are tested with invertebrate *Pax6* in the analysis. The results showed that 74.0%, 96.8%, 80.6% and 96% of quartets had resolved phylogenies for *Pax1/9*, *Pax2/5/8*, *Pax3/7* and *Pax6*, and only 9.8%, 1.2%, 14.2% and 2.8% of all quartet points were in the star-tree region, indicating that all data sets contain reasonably high degree of phylogenetic information (Fig. 3). The all mapping placed significant majority of the resolved quartets (64.3%, 98.9%, 89.5% and 100%, respectively) in the region of tree 1, suggesting a branching pattern groups amphioxus and vertebrates versus *Ciona* and protostomes. The topology with *Ciona* as more ancestral split followed by the divergence of amphioxus and vertebrate is favored. Therefore, likelihood mapping gives out extra support to the hypothesis that amphioxus is the closest relative of vertebrates.

DISCUSSION

Data set selection in phylogeny study The comparisons of morphological characters and embryogenesis between cephalochordate and vertebrate have demonstrated that amphioxus is the closest living relative to the vertebrate. The scenario was widely accepted by evolutionary biologists in the past decades, and the hypothesis was further proved by a series of molecular data (Wada and Satoh, 1994; Swalla et al., 2000; Cameron et al., 2000; Winchell et al., 2002). Nonetheless, controversy about the relationship among three subphyla of chordate arises again in recent years. Some phylogenetic analyses based on sequences of multiple nuclear genes/proteins from deuterostomes provide compelling evidences claiming that tunicates, but not cephalochordates, represent the closest living relatives of vertebrates (Holland et al., 2008; Putnam et al., 2008; Delsuc et al., 2006, 2008). These researchers argued that a single gene or a small number of genes would induce the impact of systematic

biases on phylogenomic studies, thus more genes were adopted for their phylogenetic analysis. Nevertheless, those supermatrix data still cannot avoid the biases because even several hundreds of selected genes are only representatives of less one percentage to whole genome of any metazoa. On the other hand, current computer system cannot perform phylogenetic analysis for multi-species of whole genome data yet, and in addition, whole genome data without evolutionary weightiness consideration will introduce much more noise to the analysis. The situation is similar to that of high-lever phylogeny tree reconstruction on the data of morphological characters. Biologists selected several evolutionary important characters from huge numbers of data including anatomical, embryonic and fossil traits, whereas those characters specialized in definite lineages are not under consideration when they were rebuilding the relationship among distantly related organisms because the specialized characters will obstruct us from approaching to the real history. Therefore, to avoid evolutionary noise, it is also necessary to select data of functional importance in the evolutionary history carefully but not randomly before the phylogeny tree reconstruction. Considering of the critical importance of *Pax* gene family relating to formation of body plan in animals, especially relating to some important organogenesis in the vertebrate, as well as the gene appearance along with the metazoa from the very beginning to current vertebrates, we investigated sequence of *Pax* genes in amphioxus and characterized their genomic organization to address the issue of relationship of subphyla of Chordata.

Last common ancestor of vertebrates As presented in the results, the overview of phylogenetic tree based on the PD of *Pax* gene family indicated that cephalochordates were much more frequently gathered together with vertebrates than urochordates did. In order to examine the relationship of three subphyla of Chordata, we further constructed phylogenetic trees respectively based on the complete amino acid sequences of each gene subfamily. The trees derived from three subfamilies (*Pax1/9*, *Pax2/5/8* and *Pax3/7*) show us that amphioxus cluster first with vertebrates, but tunicates form a sister group to cephalochordate-vertebrates lineage consistent with the four-Cluster likelihood analysis (Fig. 3). The tree based on the sequences of subfamilies *Pax4/6* also generates a same outcome with vertebrate *Pax 4* genes forming an independent clade. For further testing, we concatenated subfamilies datasets and reconstructed phylogenetic trees again. Interestingly, the trees continue to supporting that cephalochordates closely related to vertebrates but not tunicates, contrasting to urochordate relatives hypothesis (Holland et al., 2008; Putnam et al., 2008; Delsuc et al., 2006, 2008), and the bootstrap values also arose evidently with the increasing of data till up to 100 per-

centage. These observations lead us presume that accumulation of statistical biases in random data selection might also produce a higher bootstrap value but an incorrect clustering relationship.

In addition to information draw from the sequences, comparison of the functional domains could provide more clues to reveal the relationship of *Pax* proteins. Among five *Ciona* *Pax* proteins, three have atypical functional domain combination, e.g., CiPax3/7 lacks the octapeptide (OC), CiPax2/5/8-B misses homeodomain (HD), and CiPax2/5/8-A loses both OC and HD. In contrast to this, *Pax* proteins (excluding PaxA) in amphioxus unexceptionally exhibited canonical domain combinations, which are diagnostic to the each subfamily of vertebrates. The genome structure of sea urchin *Pax2/5/8* is same to those of cephalochordates and vertebrates indicating that *Pax2/5/8* of these three lineages originated from a common ancestral gene and *Ciona Pax2/5/8* genes was undergone a independent evolution after divergence between ascidian and major lineage of chordates. Therefore, organization of functional domain of *Pax* genes demonstrates that cephalochordates and vertebrates have a much closer relationship than that between tunicates and vertebrates.

Evolution of *Pax* gene family In cnidarians, four classes of *Pax* genes had been identified (Miller et al., 2000). Based on the overall structure, Catmull et al. (1998) assumed that the *Pax* gene family was originated from *PaxA* precursor. Alternatively, the fact that in the Porifera and Placozoa only *PaxB* genes have been found led some authors to suggest that the first divergence of this family was *PaxB* gene (Miller et al., 2000). Taken together with the phylogenetic analyses of paired domain sequences and the structure feature, we presume that cnidarian *PaxD* is the most 'primitive' representative of *Pax* gene family. Breitling and Gerber (2000) proposed that the precursor of all *Pax* genes arose via the fusion between the DNA binding domain of an ancestral transposase and a homeodomain followed by acquisition of an octapeptide-encoding motif. The *PaxD* gene, resembling the intermediate state in possessing a paired domain and homeodomain, was considered to represent the precursor genes between the two capture events. However, since no *PaxD*-like gene has been found in more basal metazoans like Porifera and Placozoa, we cannot exclude the fact that these species may also have a *PaxD* gene that has not yet been found.

To provide more perspectives to the temporal order of events in *Pax* gene evolution, we deduced a plausible model that is compatible with the available data from phylogenetic affinity and schematized structure of *Pax* proteins. We speculate that the acquiring of the octapeptide as well as the intron IV and VI in paired domain occurred in the last common ancestor (LCA) of the sub-

families, and followed by subsequent divergence of *Pax3/7*, *Pax1/9*, *Pax4/6* and *Pax2/5/8* in order (Fig. 4).

Each subfamily had independently undergone loss/gain of functional domains and splice sites. In *Pax3/7* subfamily the intron IV and VI had lost in all but vertebrate gene members. Before the separation of protostomes and chordates the intron V had emerged and retained in the present *DmGsbP* from *D. melanogaster* and *CiPax3/7* from *C. intestinalis*. In *Pax1/9* subfamily the homeodomain, intron IV and VI lost before the divergence of protostomes. The gene organization of the subfamily members presented a very conservative property without any modification throughout deuterostomes. Before the divergence of ancestral *Pax4/6* and *Pax2/5/8* genes, the intron III inserted into the paired domain, and the capture of intron VIII and the lost of octapeptide occurred in the early evolution of *Pax4/6* subfamily. At the onset of vertebrate evolution, the intron III was lost and the intron II specific to vertebrate lineage was formed, indicating the duplication of *Pax4* and *Pax6* happened in the early evolution.

The subfamily of *Pax2/5/8* is homologous to *PaxA*, *PaxB* and *PaxC* genes. The intron VI was lost before the separation between two ancestors of *PaxA/C* and *PaxB/2/5/8* lineages. During the evolution of *PaxA/C*, the intron III, octapeptide and homeodomain had lost in succession. The failure to detect any vertebrate homologues suggests a case of gene death event had accompanied with extensive gene duplication in *Pax* family.

Pax2/5/8 subfamily is closely related with *PaxB*. The ancestor of *PaxB/2/5/8* lineage obtained an intron VII and lost intron IV. In the genome of *S. purpuratus*, both *Pax2/5/8* and *PaxB* genes existed, indicating a paralogy relationship between the two genes. In *Pax2/5/8* subfamily, the partial losing of homeodomain before the divergence of deuterostomes resulted in a typical gene organization of the subfamily, which organization are identical among all *Pax2/5/8* members and their homologues in Cephalochordata and Echinoderms. In the urochordate, however, the lineage specific duplication occurred after ancestor of *CiPax2/5/8* lost its homeodomain.

Fig. 4. Evolutionary scenario of *Pax* gene family depicting the conservation of structural features. This scenario is speculated according to the tree of Fig. 2. The paired domain is depicted as green box, homeodomain as red box and octapeptide as blue circle. The numbers designated introns shared by different *Pax* genes within paired domain and homeodomain. These introns are denoted by inverted triangle with color represent specific positions between neighboring amino acids within the domain. The intron positions that only occurred in single *Pax* gene are marked with asterisks and black triangle. The assumed episodes of the gain (▲) and losses (▼) of introns and functional domains are itemized. The putative ancestral structures at the nodes are proposed as shown in dotted boxes.

This work was supported by the National Natural Science Foundation of China (No. 30830023 and 30800857), the National High Technology Research and Development Program of China (No. 2008AA092602) and Specialized Research Fund for the Doctoral Program of Higher Education from the Ministry of Education, China (No. 20070384041).

REFERENCES

- Breitling, R., and Gerber, J. K. (2000) Origin of the paired domain. *Dev. Genes Evol.* **210**, 644–650.
- Cameron, C. B., Garey, J. R., and Swalla, B. J. (2000) Evolution of the chordate body plan: new insights from phylogenetic analyses of deuterostome phyla. *Proc. Natl. Acad. Sci. USA* **97**, 4469–4474.
- Catmull, J., Hayward, D. C., McIntyre, N. E., Reece-Hoyes, J. S., Mastro, R., Callaerts, P., Ball, E. E., and Miller, D. J. (1998) *Pax-6* origins—implications from the structure of two coral *Pax* genes. *Dev. Genes Evol.* **208**, 352–356.
- Chalepakakis, G., Stoykova, A., Wijnholds, J., Tremblay, P., and Gruss, P. (1993) *Pax* gene regulators in the developing nervous system. *J. Neurobiol.* **24**, 1367–1384.
- Delsuc, F., Brinkmann, H., Chourrout, D., and Philippe, H. (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439**, 965–968.
- Delsuc, F., Tsagkogeorga, G., Lartillot, N., Philippe, H. (2008) Additional molecular support for the new chordate phylogeny. *Genesis* **46**, 592–604.
- Dover, G. (2000) How genomic and developmental dynamics affect evolutionary processes. *Bioessays* **22**, 1153–1159.
- Eccles, M. R., He, S., Legge, M., Kumar, R., Fox, J., Zhou, C., French, M., and Tsai, R. W. (2002) *PAX* genes in development and disease: the role of *PAX2* in urogenital tract development. *Int. J. Dev. Biol.* **46**, 535–544.
- Gardon, S., Callaerts, P., Halder, G., and Gehring, W. J. (1997) Conservation of *Pax-6* in a lower chordate, the ascidian *Phallusia mammillata*. *Development* **124**, 817–825.
- Gardon, S., Holland, L. Z., Gehring, W. J., and Holland, N. D. (1998) Isolation and developmental expression of the amphioxus *Pax-6* gene (*AmphiPax-6*): insights into eye and photoreceptor evolution. *Development* **125**, 2701–2710.
- Hadrys, T., Desalle, R., Sagasser, S., Fischer, N., and Schierwater, B. (2005) The trichoplax *PaxB* gene: a putative *Proto-PaxA/B/C* gene predating the origin of nerve and sensory cells. *Mol. Biol. Evol.* **22**, 1569–1578.
- Herbrand, H., Guthrie, S., Hadrys, T., Hoffmann, S., Arnold, H. H., Rinkwitz-Brandt, S., and Bober, E. (1998) Two regulatory genes, *cNkx5-1* and *cPax2*, show different responses to local signals during otic placode and vesicle formation in the chick embryo. *Development* **125**, 645–654.
- Holland, P. W., and Garcia-Fernandez, J. (1996) *Hox* genes and chordate evolution. *Dev. Biol.* **173**, 382–395.
- Holland, L. Z., Schubert, M., Kozmik, Z., and Holland, N. D. (1999) *AmphiPax3/7*, an amphioxus paired box gene: insights into chordate myogenesis, neurogenesis, and the possible evolutionary precursor of definitive vertebrate neural crest. *Evol. Dev.* **1**, 153–165.
- Holland, L. Z., Albalat, R., Azumi, K., Benito-Gutiérrez, E., Blow, M. J., Bronner-Fraser, M., Brunet, F., Butts, T., Candiani, S., Dishaw, L. J., et al. (2008) The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res.* **18**, 1100–1111.
- Holland, N.D., Holland, L.Z., and Kozmik, Z. (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.
- Kaestner, K. H., Knochel, W., and Martinez, D. E. (2000) Unified nomenclature for the winged helix/forkhead transcription factors. *Genes Dev.* **14**, 142–146.
- Kozmik, Z., Holland, N. D., Kalousova, A., Paces, J., Schubert, M., and Holland, L. Z. (1999) Characterization of an amphioxus paired box gene, *AmphiPax2/5/8*: developmental expression patterns in optic support cells, nephridium, thyroid-like structures and pharyngeal gill slits, but not in the midbrain-hindbrain boundary region. *Development* **126**, 1295–1304.
- Kozmik, Z., Daube, M., Frei, E., Norman, B., Kos, L., Dishaw, L. J., Noll, M., and Piatigorsky, J. (2003) Role of *Pax* genes in eye evolution: a cnidarian *PaxB* gene uniting *Pax2* and *Pax6* functions. *Dev. Cell* **5**, 773–785.
- Kumar, S., Tamura, K., and Nei, M. (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform.* **5**, 150–163.
- Mackereth, M. D., Kwak, S. J., Fritz, A., and Riley, B. B. (2005) Zebrafish *Pax8* is required for otic placode induction and plays a redundant role with *Pax2* genes in the maintenance of the otic placode. *Development* **132**, 371–382.
- Mansouri, A., Hallonet, M., and Gruss, P. (1996) *Pax* genes and their roles in cell differentiation and development. *Curr. Opin. Cell Biol.* **8**, 851–857.
- Mazet, F., Hutt, J. A., Millard, J., and Shimeld, S. M. (2003) *Pax* gene expression in the developing central nervous system of *Ciona intestinalis*. *Gene Expr. Patterns* **3**, 743–745.
- Miller, D. J., Hayward, D. C., Reece-Hoyes, J. S., Scholten, I., Catmull, J., Gehring, W. J., Callaerts, P., Larsen, J. E., and Ball, E. E. (2000) *Pax* gene diversity in the basal cnidarian *Acropora millepora* (Cnidaria, Anthozoa): implications for the evolution of the *Pax* gene family. *Proc. Natl. Acad. Sci. USA* **97**, 4475–4480.
- Peters, H., Neubuser, A., Kratochwil, K., and Balling, R. (1998) *Pax9*-deficient mice lack pharyngeal pouch derivatives and teeth and exhibit craniofacial and limb abnormalities. *Genes Dev.* **12**, 2735–2747.
- Putnam, N. H., Butts, T., Ferrier, D. E., Furlong, R. F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J. K., et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064–1071.
- Robson, E. J., He, S. J., and Eccles, M. R. (2006) A panorama of *PAX* genes in cancer and development. *Nat. Rev. Cancer* **6**, 52–62.
- Schmidt, H. A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502–504.
- Shimeld, S. M., and Holland, P. W. (2000) Vertebrate innovations. *Proc. Natl. Acad. Sci. USA* **97**, 4449–4452.
- Short, S., and Holland, L. Z. (2008) The evolution of alternative splicing in the *Pax* family: the view from the Basal chordate amphioxus. *J. Mol. Evol.* **66**, 605–620.
- Swalla, B. J., Cameron, C. B., Corley, L. S., and Garey, J. R. (2000) Urochordates are monophyletic within the deuterostomes. *Syst. Biol.* **49**, 52–64.
- Thompson, J. D., Higgins, D. G., and Gibson, T. J. (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.
- Underhill, D. A. (2000) Genetic and biochemical diversity in the

- Pax* gene family. *Biochem. Cell Biol.* **78**, 629–638.
- Wada, H., and Satoh, N. (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.
- Wada, S., Tokuoka, M., Shoguchi, E., Kobayashi, K., Di Gregorio, A., Spagnuolo, A., Branno, M., Kohara, Y., Rokhsar, D., Levine, M., et al. (2003) A genomewide survey of developmentally relevant genes in *Ciona intestinalis*. II. Genes for homeobox transcription factors. *Dev. Genes Evol.* **213**, 222–234.
- Wang, W., Xu, H. L., Lin, L. P., Su, B., and Wang, Y. Q. (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.
- Winchell, C. J., Sullivan, J., Cameron, C. B., Swalla, B. J., and Mallatt, J. (2002) Evaluating hypotheses of deuterostome phylogeny and chordate evolution with new LSU and SSU ribosomal DNA data. *Mol. Biol. Evol.* **19**, 762–776.
- Zhong, J., Zhang, Q. J., Xu, Q. S., Schubert, M., Laudet, V., and Wang, Y. Q. (2009) Complete mitochondrial genomes defining two distinct lancelet species in the West Pacific Ocean. *Mar. Biol. Res.* **5**, 278–285.