African Journal of Microbiology Research Vol. 6(23), pp. 5005-5014, 21 June, 2012
Available online at http://www.academicjournals.org/AJMR
DOI: 10.5897/AJMR12.300
ISSN 1996-0808 ©2012 Academic Journals

Full Length Research Paper

Isolation and sequencing of the HMG domains of fifteen Sox genes from *Hyla sanchiangensis,* and analysis of the evolutionary behaviors of Sox duplicated copies based on bioinformatics

CHEN Qi-long[1,2]*, QIAO Zi-jun[1,2#], CHEN Jian[2#], HU Sheng[2#], ZHOU Hui[1,2#], LI Zhong-hui[1], LU Shun-Qing[1] and MA Wen-li[2]

[1]College of Life and Environment Science, Huangshan University, Huangshan, 245021, China.
[2]Institute of Genetic Engineering, Southern Medical University, Guangzhou, 510515, China.

Accepted 8 March, 2012

Sox gene is a large gene family which encodes transcription factors and contains a HMG box that is responsible for a variety of developmental processes. In the present study, we obtained fifteen clones representing Sox gene HMG-boxes from male and female *Hyla sanchiangensis*, distributed as Sox1, Sox2, Sox3, Sox4, Sox11, and Sox33. The sequences analysis indicated that Sox1 and Sox4 have two duplicated copies, respectively, Sox2 has three duplications and Sox11 has six different copies. Furthermore, the amino acid of Sox1, Sox2, Sox4 and Sox11 duplicated copies has been exchanged indicating that the gene functional selection might be necessary for Sox gene duplicated process. Phylogenetic analysis was carried out and suggested that HMG domain-encoding sequences are members of the SoxB and SoxC groups. The topologies implied that the duplicated Sox genes might evolve independently in *H. sanchiangensis*. Substitution rate showed that the evolutionary behaviors of Sox duplicated copies are dissymmetry, which would cause two parallel evolutionary patterns at the molecular level. One is the duplicated genes were suffering from a period of relaxed selection and caused the asymmetric evolutionary rate in one copy, then accelerated gene evolution. Another is that the Sox duplicated genes experienced identical selection constraints and had no greater genetic diversity. In this case, we proposed that all Sox genes in *H. sanchiangensis* obtained in this examination are under strong purifying selection, and duplicated Sox genes evolved independently.

Key words: *Hyla sanchiangensis,* Sox gene, gene duplication.

INTRODUCTION

The Sox gene family is a group of transcription factors with a high mobility- group (HMG) box DNA-binding domain that determines sex region in mouse and human Y chromosome (Sry) (Gubbay et al., 1990). Sry gene is necessary and sufficient for proper development of the testes, and to trigger differentiation of genital ridge somatic cells into Sertoli cells (Koopman et al., 1991). The HMG domain of Sox protein contains a 79-amino-acid and forming three-helices (fold to form a “L” shape), which is thought to bind the minor groove of DNA with preference for the consensus motif (A/T)AACAAT and induces a 70 to 85°C bend in the DNA strand (Harley et al., 1994; Wegner, 1999; Werner et al., 1995). This bending enhances the recruitment and binding of other transcription factors adjacent to the HMG-box biding sites

*Corresponding author. E-mail: cql@hsu.edu.cn. Tel: +86-559-2546552. Fax: +86-559-2656630.

#Authors contributed equally.

(Chung et al., 2011), and it has been shown the outside of HMG-box might facilitate interactions and influence the specificity of Sox proteins (Wilson and Koopman, 2002).

The Sox genes are highly conserved and are known to play important roles in embryonic development including roles in gonad, central nervous system, neural crest and skeletal development (Nagai, 2001). At present, over 40 orthologous pairs of Sox genes have been cloned in animal kingdom, including more than 30 in vertebrates and over a dozen from invertebrates (Hagiuda et al., 2003). Based on the sequence similarity of the HMG-box, function, gene structure and chromosome location, the Sox gene family can be further subdivided into groups A-J (Bowles et al., 2000). Sox group A (Sry) is specific to eutherian mammals, groups G-J are restricted to particular lineages and other groups are found in all higher metazoans. In addition, Sox HMG-domain sequence is congruent with relatedness overall gene and protein structure. It has been shown HMG boxes can be used to identify Sox genes without scrutinizing the entire sequence (Bowles et al., 2000; Hett and Ludwig, 2005). Noticeably, most of Sox genes appear to have been duplicated in amphibians and teleost fishes such as *Odorrana schmackeri* (Wang et al., 2009b), *Danio rerio* and *Takifugu rubripes* (Koopman et al., 2004). This might be the results of the partition of ancestral subfunctions and have exhibited some hints of important mechanism leading to the preservation of multiple gene copies (Force et al., 1999).

Amphibians are a class of vertebrate animals which are characterized as non-amniote ectothermic tetrapods. Most amphibians undergo metamorphosis from a juvenile water-breathing form to an adult air-breathing form, from which evolves a large diversity of morphological changes that are different from aquatic vertebrate. Therefore, as a transitional group from aquatic to terrestrial in vertebrate evolution, they play a key role in the analysis of the genetic basis of the morphological and lifestyle transition and the evolution of genes that function well in different animals (Mannaert et al., 2006). *Hyla sanchiangensis* (Pope, 1929) is a species of frog in the Hylidae family, and also an endemic species to China (Zhao and Adler, 1993). In the present study, we describe the isolation and characterization of the Sox genes in *H. sanchiangensis* with the aim of researching the diversity and evolution of this gene family, and it indicates that some of these genes are very significant for understanding of the gene duplication process.

MATERIALS AND METHODS

PCR amplification, cloning, and sequencing

Specimens used in this study were collected in a recovering subtropical forest and adjacent bamboo plantation (29.45°N, 118.15°E) near the Lingnan Nature Reserve, Xiuning County, Anhui Province in China, including two male and two female *H. sanchiangensis*. Total genomic DNA was isolated from muscle tissue using the standard phenol-chloroform method. Sox genes were amplified by genomic polymerase chain reaction (PCR) using a pair of degenerate primers (p1: ATGAAYGCNTTYATGGTNTGG; p2: GGNCGRTAYTTRTARTCNGG). The sequences of degenerate primers were designed using multiple alignments of the HGM-box sequence of SRY, corresponding to the MNAFMVW and PDYKYRP motifs in the HMG box region.

PCR reactions were conducted in a volume of 50 µL PCR mix consisting of 1×buffer with 1.5 mM Mg^{2+}, 0.25 mM dNTPs, 2.5 U Taq DNA polymerase, 0.3 pM each oligonucleotide primer, and approximately 50 ng genomic DNA. The PCR amplification profile included an initial 5 min denaturing period at 94°C. The following PCR conditions by 35 cycles with 40 sec. at 94°C, 40 sec. at 53°C, 1 min at 72°C, and finally 72°C for 10 min were used to complete the final reaction. Resulting PCR products were resolved on 0.8% agarose gels, and purified with AxyPrep DNA gel extraction kit (Axygen). Then the purification products were ligated and cloned using the TA cloning kit (Invitrogen). Positive clones were screened by SSCP (single-strand conformation polymorphism) analysis and sequenced using an ABI 3730 automatic sequencer. To ensure authenticity, all sequences were sequenced in both directions. Sox sequences had been deposited in GenBank database with accession numbers listed in Table 1.

Evolutionary analysis of Sox sequences

The identity of each sequence was determined using BLAST analysis. Then, inferred amino acid sequences of Sox genes HMG domain, using the key signature residues of different Sox genes (Koopman et al., 2004), were used to make further identification of Sox genes in the *H. sanchiangensis*. The evolutionary phylogenetic relationships were investigated by performing phylogenetic analyses of amino acid sequences with Neighbor-joining (NJ) calculation in MEGA5.0 (Tamura et al., 2011), maximum-likelihood (ML) calculation and maximum-parsimony (MP) implemented in PAUP*4.0b10 (Swofford, 2003). Heuristic MP Searches were executed in 1000 random addition replicates with all characters unordered and equally weighted, and using tree bisection reconnection (TBR) branch-swapping. Bootstrap branch proportions (BBP) were calculated with 1000 MP replicates. Based on the Akaike Information Criterion (AIC), the GTR+I+G model was selected for maximum likelihood (ML) analyses by Modeltest 3.7 (Posada and Buckley, 2004). Heuristic Searches were executed in 10 replicates with the GTR+I+G model, and using TBR branch swapping. BBP values were calculated with 10 ML replicates (CAI et al., 2007).

Bayesian inference (BI) was carried out using MrBayes 3.1 (Ronquist and Huelsenbeck, 2003). The Bayesian posterior probabilities (BPP) used models estimated with Modeltest 3.7 under the AIC. Two separate runs were performed with four Markov chains. Each run was conducted with 1,000,000 generations and sampled every 100 generations. When the log-likelihood scores were found to stabilize, a consensus tree was calculated after omitting the first 25% trees as burn-in (Ronquist and Huelsenbeck, 2003).

Based on genealogical relationships, the synonymous and nonsynonymous substitution rates (Ks and Ka) between duplicated Sox genes were calculated using the modified KaKs_Calculator 2.0 version (Wang et al., 2009a). The Z test was also performed using MEGA5.0 to detect deviation from neutrality of those duplicated genes. The following sequences were also included in the phylogenetic analyses as comparison, obtained from GenBank database.

Genetic nomenclature

In reference to the previous nomenclature system (Hett and Ludwig, 2005; Koopman et al., 2004) and the actual conditions of

Table 1. Summary of Sox genes in *H. sanchiangensis*.

Gene	Accession number	Human (%)	Mouse (%)	Variable positions for each duplications									
				15	**18**	**21**	**24**	**139**	**196**	**198**			
HsSox1-1	JQ283978	99	99	C	A	T	T	***G***	***A***	***T***			
HsSox1-2	JQ283979	96	96	T	C	C	C	***A***	***G***	***C***			
				15	18	21	195	196	201	204			
HsSox2-1	JQ283980	97	97	T	C	C	C	***A***	C	G			
HsSox2-2	JQ283981	96	96	C	T	C	T	***G***	T	A			
HsSox2-3	JQ283982	96	96	C	G	T	T	***A***	C	A			
HsSox3	JQ283983	92	92	**No variation observed**									
				19	22	25	33	45	84	97	99	111	117
HsSox4-1	JQ283984	94	94	T	***C***	C	T	G	G	C	C	G	G
HsSox4-2	JQ283985	92	92	C	***T***	A	G	C	C	A	A	C	A
				118	126	135	144	151	153	171	177	186	204
				C	***G***	***C***	T	C	G	C	G	G	G
				A	***T***	***G***	C	A	A	G	A	C	A
				15	18	21	24	31	153	195	196	204	
HsSox11-1	JQ283986	96	96	C	G	T	C	***T***	A	C	***G***	G	
HsSox11-2	JQ283987	96	96	T	T	C	T	***T***	A	C	***A***	A	
HsSox11-3	JQ283988	96	96	T	T	T	T	***T***	A	C	***A***	A	
HsSox11-4	JQ283989	96	96	C	G	T	C	***T***	A	C	***A***	G	
HsSox11-5	JQ283990	96	96	C	G	T	C	***T***	G	C	***A***	A	
HsSox11-6	JQ283991	94	94	T	T	T	C	***C***	G	A	***A***	A	
		X. laevis (%)	*B. maxima* (%)										
HsSox33	JQ283992	97	94	No variation observed									

In last column, numbers indicate variable positions corresponding to positions in the HMG-box sequence; substitutions leading to amino acid changes are indicated in bold.

Sox genes in *H. sanchiangensis*, the duplicated homologous genes were designated with *Arabic numerals* according to their evolutionary relationships.

RESULTS

PCR amplification and clone sequenced

PCR using the degenerate primers with genomic DNA generated a single electrophoretic band about 220-bp in size. Taken together, the sequences of 131 clones isolated from the male and female were sequenced. Finally, we obtained a 216-bp fragment of the HMG boxes for 15 different Sox genes, in which no sex-specific genes were detected between male and female. These genes have been submitted to GenBank under the accession numbers JQ283978-JQ283992.

BLAST analysis revealed high levels of similarity at the amino acid level (92 to 99%) to HMG-boxes of previously published Sox sequences in other vertebrates. Therefore, the sequences were named according to their corresponding Sox genes. Differing sequences that coded the amino acid sequence for same Sox gene were considered duplications of this gene. We found the following 6 Sox genes: Sox1, Sox2, Sox3, Sox4, Sox11 and Sox33. Interestingly, the amino acid sequence of Sox33 had 76% similarity

```
Hs-Sox1-1  KRPMNAFIVW SRGQRRKMAQ ENPKMHNSEI SKRLGAEWKV MSEAEKGPFI DEAKRLRALH MKEHPNYKYR PR
Hs-Sox1-2  .......... .......... .......... .......... ......R... .......... .....D.... ..
Hs-Sox2-1  .......... .......... .......... .........L L.....R... .......... .......... ..
Hs-Sox2-2  .......... .......... .......... .........L L.....R... .......... .....D.... ..
Hs-Sox2-3  .......... .......... .......... .........L L.....R... .......... .......... ..
Hs-Sox3    .......... .......... ........K. ......D..L L.DS..R... ........V. ...Y.TINID LV
Hs-Sox4-1  QATHEC.HR. .QIE...IME QS.D...A.. .....KR..L LKDSD.I... R..E...LK. .ADY...... ..
Hs-Sox4-2  QATHECLYR. .QIE...IME QS.D...A.. .....KR..M LNDS..I... R..E...LK. .ADY...... ..
Hs-Sox11-1 .......... .KIE...IME QS.D...A.. .....KR..M LNDS..I... R..E...LK. .ADY.D.... ..
Hs-Sox11-2 .......... .KIE...IME QS.D...A.. .....KR..M LNDS..I... R..E...LK. .ADY...... ..
Hs-Sox11-3 .......... .KIE...IME QS.D...A.. .....KR..M LNDS..I... R..E...LK. .ADY...... ..
Hs-Sox11-4 .......... .KIE...IME QS.D...A.. .....KR..M LNDS..I... R..E...LK. .ADY...... ..
Hs-Sox11-5 .......... .KIE...IME QS.D...A.. .....KR..M LNDS..I... R..E...LK. .ADY...... ..
Hs-Sox11-6 .......... PKIE...IME QS.D...A.. .....KR..M LNDS..I... R..E...LK. .ADY...... ..
Hs-Sox33   .......... .QIE...LMS LC.N...AD. .RS..QR..L LEDTD.I.YV R..E...LK. .ADY.RLQVS TS
```

Figure 1. Alignment of Sox HMG-box amino acid sequences in *H. sanchiangensis*.

to the Sox4 of human and mouse, 74% similarity to Sox11; however, it had 97% homology to the *X. laevis* Sox-K1 which was isolated from African clawed frog and named xSox33 (Hagiuda et al., 2003), and 94% similarity to BmSox33 of *B. maxima* (Jing et al., 2009), individually. According to the homologous, we named the sequence of this clone as HsSox33.

The number of duplicated copies for each Sox gene varied between 1 (Sox3 and Sox33) and 6 (Sox11). All variable positions were found independently in both male and female. This makes it unlikely that these differences are due to sequencing errors or PCR artefacts. The number of variable positions ranged 7 substitutions in the case of Sox1 and Sox2, individually, 20 substitutions in the case of Sox4, and 9 substitutions in the case of Sox11. The numbers of duplicated copies for each gene, and the variable positions are given in Table 1.

In the case of Sox genes, Sox1 showed 2 different duplications, and 7 substitutions were detected, including 4 synonymous and 3 nonsynonymous substitutions, resulting in two amino acids exchange (G↔R, and N↔D). Three different copies were detected in the case of Sox2, which varied in 7 positions and showing 1 nonsynonymous substitution, resulting in an amino acid exchange (N ↔ D). Within 2 different duplications isolated, Sox4 showed 20 substitutions, four of them were found for nonsynonymous substitutions and the amino acid exchange including F ↔ L, L ↔ M, K ↔ N, and D ↔ E, respectively. Remarkably, Sox11 showed 6 duplicated copies and 9 substitutions were detected, including 2 nonsynonymous substitution (S↔P, and D ↔ N) (Figure. 1).

Phylogenetic analysis of Sox genes

The genes were clustered in phylogenetic trees based on nucleic acid sequences and their inferred amino acid identities. In Figure 2, phylogenetic methods (MP, ML, and BI) resulted in very similar trees, which could conclude that the analysis about these Sox genes is credible. Phylogenetic analysis showed 68 Sox genes were clustered into ten subfamilies (A-J). According to the classification of Bowles et al. (2000), 6 clones from *H. sanchiangensis* (*Hs-Sox1-1, Hs-Sox1-2, Hs-Sox2-1, Hs-Sox2-2, Hs-Sox2-3, Hs-Sox3*) belonged to groups B, and 9 isolations (*Hs-Sox4-1, Hs-Sox4-2, Hs-Sox11-1, Hs-Sox11-2, Hs-Sox11-3, Hs-Sox11-4, Hs-Sox11-5, Hs-Sox11-6, Hs-Sox33*) belonged to groups C.

In order to understand the relationship of Sox genes in *H. sanchiangensis*, phylogenetic trees of *Hs-Sox* genes were reconstructed using Mus-Sox genes, Bom-Sox33 and Xen-SoxK1 as comparisons (Figure 3). The following clusters were supported independently by the phylogenetic methods (NJ and ML): the members of the Sox33 were jointed by a strongly supported bootstrap values of 100/100; Sox2, Sox3 and Sox4, the sequences are also supported by high bootstrap values (93/88, 87/90 and 91/93, respectively). In addition, Sox1 representatives were unified in the neighbor-joining tree (75) and the maximum-likelihood tree (63), and Sox11was found in moderated bootstrap values (77/72). To consider the consequences of the phylogenetic calculations, the members of each subgroup were clustered together with their mouse orthologues.

Evolution of duplicated Sox genes in *H. sanchiangensis*

Similarities including both nucleic acid and amino acid sequences between the Sox genes in *H. sanchiangensis* and their reported orthologs in Human and Mouse were calculated, and listed in Table 1. Generally, amino acid similarities of most of those Sox genes are extremely high compared with their orthologs, and the difference of similarities is tiny among them.

To understand the evolutionary process of Sox genes

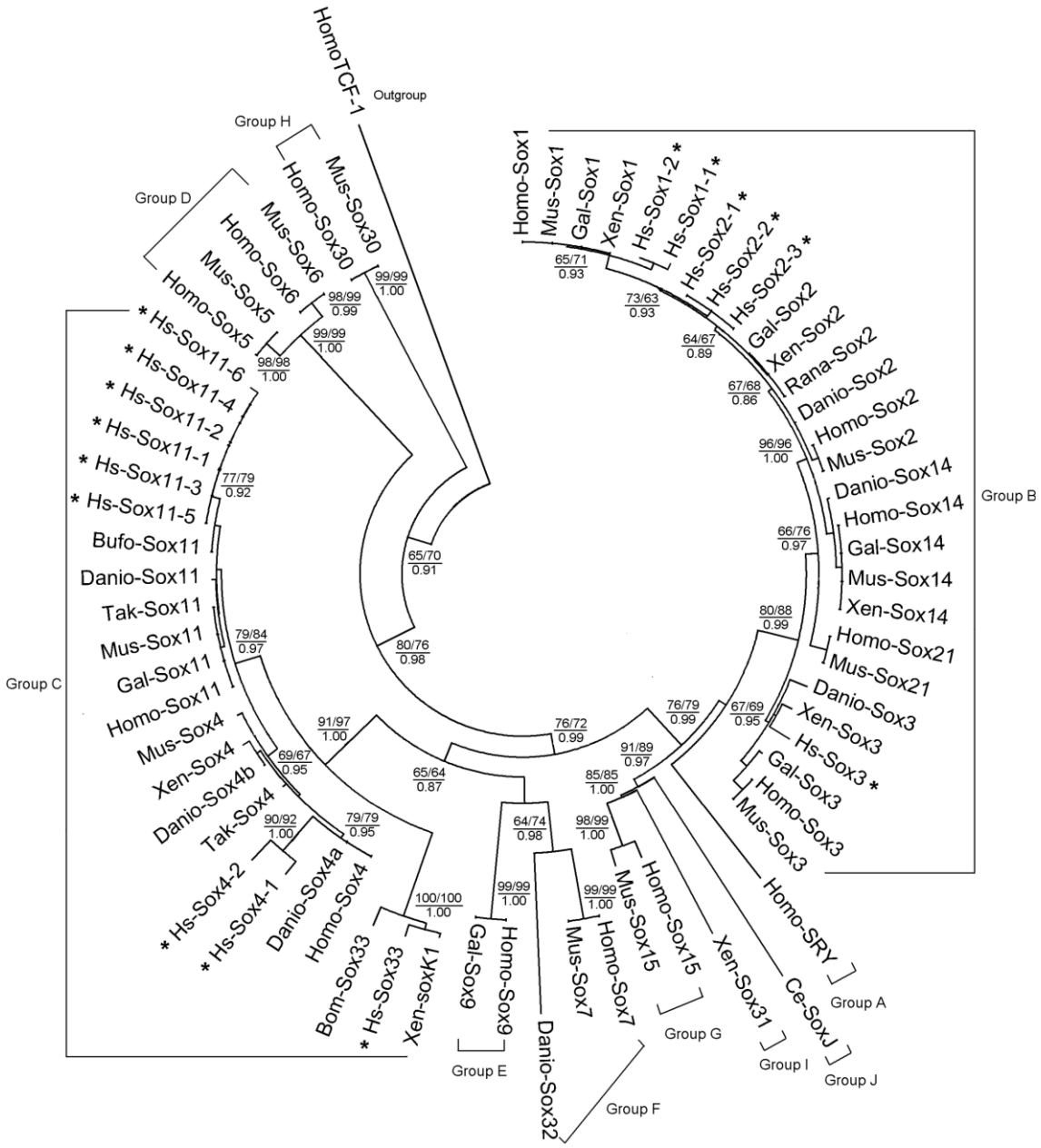

Figure 2. Bayesian inference tree derived from SOX/Sox HMG-box amino acid sequences. Numbers above branches are bootstrap support for MP (1000 replicates) / ML (10 replicates) analyses (> 50 retained), and numbers below branches indicate Bayesian posterior probabilities (>85% retained). Based on the topological structure of phylogenetic trees, 68 Sox genes were defined approximated ten groups (A - J).

in close related duplications, according to the evolutionary phylogenetic relationships of these duplicated Sox gene pairs, we calculated their pairwise divergences using nucleotide sequences and amino acid sequences, respectively (Table 2).

In the case of Sox11 duplications, the value of divergences for nucleic acid is from 0.0047 (Sox11-2 and Sox11-3, Sox11-1 and Sox11-4) to 0.0287 (Sox11-1 and Sox11-2, Sox11-1 and Sox11-6), notable, the value for amino acid is from 0 (Sox11-2, Sox11-3, Sox11-4 and Sox11-5) to 0.0282 (Sox11-1 and Sox11-6). It is revealed that evolution rates of Sox11 duplications were statistically quickly than others. The relative value of nucleotide for Sox2 duplications is lower than Sox11 genes, and the value of amino acid has no difference among them. In addition, the relative values of nucleotide and amino acid for Sox3 duplications are 0.0284 and 0.0140, and Sox4 duplicated copies are 0.0998 and 0.0720, respectively.

To evaluate whether duplicated Sox genes evolved

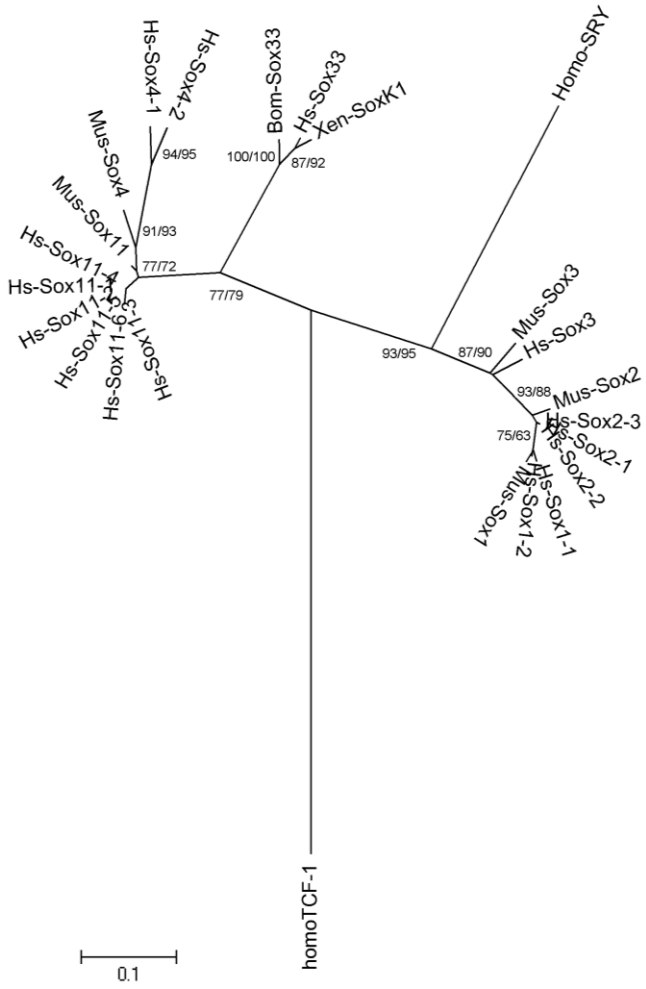

Figure 3. Phylogenetic relationships (NJ/ML) of duplicated Sox gene in *H. sanchiangensis* using amino acid sequence, and HomoTCF-1 as outgroup.

neutrally, substitution rate (Ka/Ks) among these duplicated Sox genes was calculated, as presented in Table 3. Comparison across all duplicated Sox pairs revealed that the ratios of Ka/Ks ranged from 0.0010 to 0.1306. In addition, codon-based Z Test of Neutrality for analysis between sequences, and the probability (P) of rejecting the null hypothesis of strict-neutrality is shown above the diagonal. Values of P less than 0.05 are considered significant at 5% level and are highlighted.

DISCUSSION

Sox gene diversity in *H. sanchiangensis*

Six different Sox genes (Sox1, Sox2, Sox3, Sox4, Sox11, and Sox33) were detected in *H. sanchiangensis*, and these represented groups B and C. Other groups are found in mammals, and not observed in *H. sanchiangensis*. The primer set was designed based on Cremazy et al. (1998), who reported a bias leading to preferential amplification of group B Sox genes. As expected, the distribution of the obtained sequences demonstrates that not all Sox genes are amplified equally by the primers. The veracious number of Sox genes in *H. sanchiangensis* is probably much higher; moreover, is probably not completely represented by the genes detected in our study.

In the case of Sox genes, we found 2 isforms of Sox1 and 3 different copies of Sox2, characterized by differences in their amino acid sequence in the HMG-binding domain. With Sox4, 2 duplicated copies were isolated, 20 substitutions were represented, including 16 synonymous and 4 nonsynonymous sites. There are two exchanges in the amino acid sequences of Sox11, which have 9 substitutions. Clearly, the high number of

Table 2. The pairwise divergences among the duplicated PcSox genes were shown, using nucleic acid Sequences (above diagonal) and amino acid sequences (below diagonal).

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	Hs-Sox1-1		0.0337	0.0842	0.0842	0.0686	0.1567	0.4381	0.4215	0.4401	0.4308	0.4223	0.4313	0.4227	0.4395	0.5722
2	Hs-Sox1-2	0.0282		0.0583	0.0634	0.0685	0.1330	0.4053	0.4053	0.4223	0.4142	0.4227	0.4308	0.4223	0.4227	0.5110
3	Hs-Sox2-1	0.0426	0.0426		0.0239	0.0190	0.1556	0.3292	0.3816	0.3896	0.3819	0.3899	0.3817	0.3896	0.3899	0.4830
4	Hs-Sox2-2	0.0572	0.0282	0.0140		0.0190	0.1440	0.3511	0.4054	0.3896	0.3899	0.3980	0.3977	0.3896	0.3980	0.4924
5	Hs-Sox2-3	0.0426	0.0426	0.0000	0.0140		0.1441	0.3511	0.3816	0.3740	0.3896	0.3817	0.3663	0.3586	0.3817	0.5110
6	Hs-Sox3	0.2513	0.2336	0.1991	0.1991	0.1991		0.4381	0.4054	0.3749	0.3907	0.3825	0.3830	0.3749	0.3825	0.5241
7	Hs-Sox4-1	0.6391	0.6657	0.5878	0.6131	0.5878	0.7213		0.0998	0.1161	0.1053	0.1107	0.1106	0.1106	0.1053	0.2600
8	Hs-Sox4-2	0.6391	0.6657	0.6131	0.6391	0.6131	0.7503	0.0720		0.0533	0.0433	0.0383	0.0483	0.0483	0.0483	0.2803
9	Hs-Sox11-1	0.4480	0.4265	0.4265	0.4055	0.4265	0.5155	0.1991	0.1658		0.0287	0.0238	0.0047	0.0142	0.0287	0.2602
10	Hs-Sox11-2	0.4265	0.4480	0.4055	0.4265	0.4055	0.5155	0.1823	0.1495	0.0140		0.0047	0.0238	0.0238	0.0190	0.2605
11	Hs-Sox11-3	0.4265	0.4480	0.4055	0.4265	0.4055	0.5155	0.1823	0.1495	0.0140	0.0000		0.0190	0.0190	0.0142	0.2673
12	Hs-Sox11-4	0.4265	0.4480	0.4055	0.4265	0.4055	0.5155	0.1823	0.1495	0.0140	0.0000	0.0000		0.0094	0.0238	0.2670
13	Hs-Sox11-5	0.4265	0.4480	0.4055	0.4265	0.4055	0.5155	0.1823	0.1495	0.0140	0.0000	0.0000	0.0000		0.0142	0.2809
14	Hs-Sox11-6	0.4480	0.4700	0.4265	0.4480	0.4265	0.5390	0.1991	0.1658	0.0282	0.0140	0.0140	0.0140	0.0140		0.2743
15	Hs-Sox33	0.7503	0.7503	0.6931	0.6931	0.6931	0.6391	0.4925	0.5631	0.3848	0.3848	0.3848	0.3848	0.3848	0.4055	

Analyses of nucleotide and amino acid sequences were conducted using the Kimura 2-parameter model and the Poisson correction model, individually. All calculations of standard error estimate are based on bootstrap procedure (1000 replicates), respectively.

synonymous and nonsynonymous substitutions observed in the duplicated Sox genes indicates that they are under selective pressure and therefore represent functional genes. In addition, Sox3 and Sox33 only one copy is available in this work. This is probably due to insufficient sequencing and/or biased amplification of certain alleles; but it could also be caused by gene deletion or other unknown reason.

Generally, the number of alleles at a single locus should be a minimal reflection of the ploidy level, and loci with more than 2 alleles in a single individual were considered gene duplications (David et al., 2003; Hett and Ludwig, 2005; Ludwig et al., 2001). For the differential copies of Sox gene, several possibilities might contribute to their existence. One important mechanism for functional innovation during evolution is the duplication of genes and entire genomes. For example, a large number of clusters of Hox genes have been identified in fishes such as zebrafish and Fugu, and gave rise to the genome duplication theory (Amores et al., 1998). Analysis of the triplets reveals accelerated evolution or relaxation of constraint in the peptides of the X. laevis pairs, which supports duplicate genes are retained through a process of subfunctionalization and/or relaxation of constraint on both copies of an ancestral gene (Hellsten et al., 2007). Meyer and Schartl (1999) had proposed that the genome underwent two rounds of duplication leading from a single ancestral deuterostome genome to two after the first duplication, and then to four genomes after the second genome duplication (the 'one-to-two-to-four' rule). Evidence for 1-2-4 rule is that genes from the same gene family are often arranged in linked clusters, and maintain the same gene order on different chromosomes (Pebusque et al., 1998). In fishes, analyses have indicated that duplicated genes are the result of a large scale segmental duplication before the radiation of teleosts, lending support to a 'fish-specific whole-genome duplication' theory and have exhibited some hints of important functional genes evolution in duplicated genome (Guo et al., 2009; Koopman et al., 2004).

Next important model is the duplication-degeneration-complementation (DDC) model, which predicts that the probability of gene conservation will be higher in more complex genes with a larger number of subfunctions, and suggests that the partition of ancestral subfunctions is an important mechanism leading to the preservation of multiple gene copies

Table 3. Substitution rate (Ka/Ks) among the duplicated Sox genes in *H. sanchiangensis* (below diagonal). Z Test of Neutrality for analysis between sequences is shown above the diagonal.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	Hs-Sox1-1		0.0121	0.0000	0.0001	0.0004	0.0000	0.0013	0.0233	0.0158	0.0257	0.0294	0.0158	0.0026	0.0019	0.0688
2	Hs-Sox1-2	0.0994		0.0029	0.0009	0.0010	0.0000	0.0164	0.0960	0.0128	0.0366	0.0241	0.0128	0.0020	0.0026	0.1508
3	Hs-Sox2-1	0.0684	0.1218		0.0325	0.0364	0.0000	0.1131	0.0862	0.0287	0.0429	0.0280	0.0287	0.0024	0.0031	0.5065
4	Hs-Sox2-2	0.0858	0.0625	0.0521		0.0876	0.0000	0.0747	0.0612	0.0226	0.0497	0.0334	0.0226	0.0040	0.0043	0.4117
5	Hs-Sox2-3	0.0842	0.0837	0.0010	0.0896		0.0000	0.0413	0.0770	0.0389	0.0269	0.0317	0.0389	0.0079	0.0040	0.2862
6	Hs-Sox3	0.1904	0.1889	0.1406	0.1363	0.1524		0.0173	0.1594	0.1512	0.1662	0.1833	0.1512	0.0569	0.0528	0.2061
7	Hs-Sox4-1	0.5857	0.6112	0.8845	0.6605	0.7796	0.2804		0.0147	0.0078	0.0004	0.0002	0.0078	0.0069	0.0123	0.0976
8	Hs-Sox4-2	0.5805	0.6063	0.5921	0.5171	0.6126	0.3215	0.0928		0.1068	0.0088	0.0165	0.1068	0.1118	0.1212	0.0939
9	Hs-Sox11-1	0.3079	0.3052	0.3966	0.3291	0.4288	0.3219	0.2544	0.4901		0.0206	0.0346	1.0000	0.1373	0.0178	0.0285
10	Hs-Sox11-2	0.2995	0.3702	0.3853	0.3667	0.3292	0.2564	0.2127	0.4934	0.0498		0.3208	0.0206	0.0242	0.0444	0.0561
11	Hs-Sox11-3	0.3258	0.3412	0.3563	0.3389	0.3563	0.2849	0.2306	0.5681	0.0633	0.0010		0.0346	0.0377	0.0788	0.0372
12	Hs-Sox11-4	0.2995	0.3138	0.3853	0.3388	0.4166	0.3139	0.2413	0.4617	NA	0.0010	0.0010		0.1373	0.0178	0.0285
13	Hs-Sox11-5	0.3252	0.3405	0.3556	0.3661	0.4497	0.3132	0.2409	0.4612	0.1306	0.0010	0.0010	0.0010		0.0726	0.0146
14	Hs-Sox11-6	0.2815	0.3500	0.3661	0.3766	0.3959	0.3211	0.2241	0.4091	0.1001	0.0633	0.0857	0.0498	0.0856		0.0254
15	Hs-Sox33	0.1608	0.3348	0.3316	0.3389	0.2559	0.2650	NA	0.1514	0.1253	0.1725	0.1558	0.1216	0.1211	0.1433	

The variance of the difference was computed using the bootstrap method (1000 replicates). Analyses were conducted using the Nei-Gojobori method.

(Force et al., 1999). This means that subfunctions will be maintained after subsequent rounds of duplication will be reduced, and most of the duplicated gene copies will be lost during the diploidization process (Hett and Ludwig, 2005).

In the present work, we found that the number of differentiated sequences of Sox11 duplication was more than the duplicated copies of Sox1, Sox2 and Sox4 in *H. sanchiangensis*, under the assumption that no random duplication of Sox genes occurred. Interestingly, Sox4 represents 20 substitutions in 2 different duplicated copies. Sox4 and Sox11 belong to SoxC group, together with Sox12. It seems more possible that the duplicated copies of group C might play an important role in this species. Furthermore, we note that duplicated copies of Sox1, Sox2, Sox4 and Sox11 have different amino acid sequences. This phenomenon indicated that these copies are more likely to be necessary for the functional selection of Sox genes in duplicated progress. In this case, it implies that the mechanism of duplicated copies of Sox gene in *H. sanchiangensis* would be lopsided for DDC model, although, this needs further investigation.

Evolutionary behavior of Sox duplicated genes

Phylogenetic analyses of nucleic acid sequences and inferred protein sequences basically confirmed the classification of SOX genes proposed in previous studies(Bowles et al., 2000), and also supported Sox group classification and orthology assignments of Sox genes in *H. sanchiangensis*. It is clear in the phylogenetic tree that those Sox genes of *H. sanchiangensis* fall into two Sox gene groups, SoxB (Sox1, Sox2, and Sox3) and SoxC (Sox4, Sox11, and Sox33).

SoxB proteins play crucial roles in embryo development in vertebrates. Based on the full-length protein sequence alignment and functional roles, SoxB can be more correctly divided into subgroups B1 and B2 (Koopman et al., 2004; Uchikawa et al., 1999). In terms of function, SoxB1 acts as transcriptional activators and SoxB2 plays a role as repressors; interesting, they display overlaps of expression domains in developing tissues (Uchikawa et al., 1999). In addition, Sox group B expanded independently via different trajectories, which is parallel increase in complexity at the molecular level in vertebrates (Zhong et al., 2011).

In order to resolve the evolutionary behavior of Sox gene duplicated copies, one interesting question is how duplicated Sox gene pairs affect each other after duplication. The topologies imply

that the duplicated Sox genes would evolve independently in *H. sanchiangensis* (Figure 2). This is because all duplicated Sox genes are phylogenetically sister to their orthologous counterparts rather than close to their paralogs, which means no interaction among duplicated Sox gene pairs in *H. sanchiangensis*.

Another striking question is whether the evolutionary constraint of duplicated copies of Sox gene in *H. sanchiangensis* is similar or not. Compared with the duplications, the ratio of non-synonymous to synonymous substitution rates (Ka/Ks) for these duplicate Sox genes is presented in Table 3. Among the Sox 1, Sox2, Sox4, and portion of Sox11 copies, the Ka/Ks rate of duplicated pairs is close to zero, which means that both copies of these duplicated genes evolved equally as well as their paralogs and are under the similar selection constraints at the Sox-HMG domain. The high amino acid similarity of the Sox genes also supported the above result, which implies that most of the nucleotide substitutions are silent mutation (synonymous substitutions) and these sequences are under selective pressure, especially for strongly supporting that they might represent functionally important genes at the DNA level. However, the ratios of Ka/Ks of Sox11-5 and Sox11-1, Sox11-6 and Sox11-1 are representing 0.1306 and 0.1001. Interestingly, the ratio of Sox11-4 and Sox11-1 is NA, because the value of Ks is close to zero. Obviously, this result is not coincident with above mechanism, indicating that the evolutionary rate of Sox11 copies might dissymmetry.

For these phenomena, we proposed that there would have two evolutionary patterns of Sox duplicated genes, and the patterns would occur with parallel process in complexity and diversification at the molecular level in *H. sanchiangensis*. One is consistent with the general pattern: duplicated genes suffering from a period of relaxed selection (Lynch and Conery, 2000) and causing the asymmetric evolutionary rate in one copy, then accelerating gene evolution (Brunet et al., 2006; Hellsten et al., 2007). Another pattern is that the alleles of duplicated genes experience identical selection constraints and have no greater genetic diversity in *H. sanchiangensis*, which might be mainly due to strongly functional constraint of the Sox-HMG domain (Guo et al., 2009).

Conclusion

In conclusion, our analysis indicates that all Sox genes in *H. sanchiangensis* obtained in this examination are under strongly purifying selection, and duplicated Sox genes evolved independently. Both copies of duplicated Sox genes would cause two parallel evolutionary patterns.

However, although the mounting evidence that Sox genes are the key players in the development of vertebrates, limited data are unavailable regarding the evolutionary and functions of Sox genes in *H. sanchiangensis* and would require further investigation.

ACKNOWLEDGEMENTS

The studies were supported by the National Science Foundation of China (30870281), the Innovative Foundation of Huangshan University (2011xdkj018) and the Foundation of PhDs of huangshan University (2009xkjq004).

REFERENCES

Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL, Westerfield M, Ekker M, Postlethwait JH (1998). Zebrafish hox clusters and vertebrate genome evolution. Science, 282(5394): 1711-1714.

Bowles J, Schepers G, Koopman P (2000). Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. Dev. Biol., 227(2): 239-255.

Brunet FG, Roest Crollius H, Paris M, Aury JM, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Mol. Biol. Evol., 23(9): 1808-1816.

Cai HX, Che J, Pang JF, Zhao EM, Zhang YP (2007). Paraphyly of Chinese Amolops (Anura, Ranidae) and phylogenetic position of the rare Chinese frog, *Amolops tormotus*. Zootaxa, 1531: 49-55.

Chung MI, Ma AC, Fung TK, Leung AY (2011). Characterization of Sry-related HMG box group F genes in zebrafish hematopoiesis. Exp. Hematol., 39(10): 986-998.

Cremazy F, Soullier S, Berta P, Jay P (1998). Further complexity of the human SOX gene family revealed by the combined use of highly degenerate primers and nested PCR. FEBS Lett., 438(3): 311-314.

David L, Blum S, Feldman MW, Lavi U, Hillel J (2003). Recent duplication of the common carp (*Cyprinus carpio* L.) genome as revealed by analyses of microsatellite loci. Mol. Biol. Evol., 20(9): 1425-1434.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999). Preservation of duplicate genes by complementary, degenerative mutations. Genetics, 151(4): 1531-1545.

Gubbay J, Collignon J, Koopman P, Capel B, Economou A, Munsterberg A, Vivian N, Goodfellow P, Lovell-Badge R (1990). A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. Nature, 346(6281): 245-250.

Guo B, Tong C, He S (2009). Sox genes evolution in closely related young tetraploid cyprinid fishes and their diploid relative. Gene, 439(1-2): 102-112.

Hagiuda J, Hiraoka Y, Hasegawa M, Ogawa M, Aiso S (2003). A novel *Xenopus laevis* SRY-related gene, xSox33. Biochim. Biophys. Acta., 1628(2): 140-145.

Hellsten U, Khokha MK, Grammer TC, Harland RM, Richardson P, Rokhsar DS (2007). Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog *Xenopus laevis*. BMC Biol., 5: 31.

Hett AK, Ludwig A (2005). SRY-related (Sox) genes in the genome of European Atlantic sturgeon (*Acipenser sturio*). Genome, 48(2): 181-186.

Koopman P, Gubbay J, Vivian N, Goodfellow P, Lovell-Badge R (1991). Male development of chromosomally female mice transgenic for Sry. Nature, 351(6322): 117-121.

Koopman P, Schepers G, Brenner S, Venkatesh B (2004). Origin and diversity of the SOX transcription factor gene family: genome-wide analysis in *Fugu rubripes*. Gene, 328: 177-186.

Ludwig A, Belfiore NM, Pitra C, Svirsky V, Jenneckens I (2001). Genome duplication events and functional reduction of ploidy levels in sturgeon (Acipens*er, Huso* and *Scaphirhynchus)*. Genetics, 158(3): 1203-1215.

Lynch M, Conery JS (2000). The evolutionary fate and consequences of duplicate genes. Science, 290(5494): 1151-1155.

Mannaert A, Roelants K, Bossuyt F, Leyns L (2006). A PCR survey for posterior Hox genes in amphibians. Mol. Phylogenet. Evol., 38(2):

449-458.
Meyer A, Schartl M (1999). Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. Curr. Opin. Cell Biol., 11(6): 699-704.
Nagai K (2001). Molecular evolution of Sry and Sox gene. Gene, 270(1-2): 161-169.
Pebusque MJ, Coulier F, Birnbaum D, Pontarotti P (1998). Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. Mol. Biol. Evol., 15(9): 1145-1159.
Posada D, Buckley TR (2004). Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Syst. Biol., 53(5): 793-808.
Ronquist F, Huelsenbeck JP (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics, 19(12): 1572-1574.
Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol., 28(10): 2731-2739.
Uchikawa M, Kamachi Y, Kondoh H (1999). Two distinct subgroups of Group B Sox genes for transcriptional activators and repressors: their expression during embryonic organogenesis of the chicken. Mech. Dev., 84(1-2): 103-120.
Harley VR, R Lovell-Badge, PN Goodfellow (1994). Definition of a consensus DNA binding site for SRY. Nucleic Acids Res, 22(8): 1500–1501.
Wang JJ, Wang N, Nie LW (2009a). Cloning and analysis of the HMG domains of Sox genes from *Bombina maxima* (Amphibia: Anura). Afr. J. Biotechnol., 8(8): 1441-1448.
Wang N, Wang JJ, Jia R, Xu JW, Nie LW (2009b). Isolation and sequencing of the HMG domain of ten Sox genes from *Odorrana schmackeri* (Amphibia: Anura). Zoologia, 26(1): 109-117.
Wegner M (1999). From head to toes: the multiple facets of Sox proteins. Nucleic Acids Res., 27(6): 1409-1420.
Werner MH, Huth JR, Gronenborn AM, Clore GM (1995). Molecular basis of human 46X, Y sex reversal revealed from the three-dimensional solution structure of the human SRY-DNA complex. Cell, 81(5): 705-714.
Wilson M, Koopman P (2002). Matching SOX: partner proteins and co-factors of the SOX family of transcriptional regulators. Curr. Opin. Genet. Dev., 12(4): 441-446.
Zhao EM, Adler K (1993). Herperology of China, The Society for the study of Amphibians and Reptiles. Oxford. p. 130-132.
Zhong L, Wang D, Gan X, Yang T, He S (2011). Parallel expansions of Sox transcription factor group B predating the diversifications of the arthropods and jawed vertebrates. PLoS One, 6(1): e16570.