Clone and Sequence Analysis of Sox Genes in Rana tientaiensis

YAN ZHANG¹, LIU WANG NIE^{1,*}, LIANG YAN¹, PING PING ZHENG¹ AND WEN CHENG¹

¹Life Science College, Anhui Normal University; The Key Laboratory of Biotic Environment and Ecological Safety in Anhui Province, Wuhu 241000, Anhui, China. *Corresponding author E-mail: lwnie@mail.ahnu.edu.cn

Abstract.- The *Sox* genes of *Rana tientaiensis* were amplified and cloned using highly degenerate primers designed from the conservative motif (HMG-box) of the human SRY gene. The SSCP technique was used to detect different clones. Seven distinct *Sox* gene fragments were obtained from both male and female *R. tientaiensis*; no sexual differences were observed. Seven of these fragments (named *RtSox3a, RtSox3b, RtSox3c, RtSox4, RtSox11, RtSox12,* and *RtSox14*) exhibited 95%, 95%, 95%, 97%, 98%, 97%, and 97% similarity (respectively) to the corresponding homologous human SOX genes. The eighth fragment showed 79% and 77% similarity to the human *SOX21* and *SOX14* genes, as well as varying levels of similarity to other group B Sox genes. The eighth gene, provisionally named Rt*SoxB*14, may be a new member of the Sox gene family or a derivative of an existing *Sox* gene. Phylogenetic analysis further illustrates that the gene *Sox3* found in *R. tientaiensis* are duplicates of those seen in the mammalian *Sox* gene family. Amino acid positions 15–19 are characteristic of each group in the *Sox* family.

Keywords.- Sox genes, SSCP, Rana tientaiensis, subgroup diagnosis.

Introduction

The Y chromosome-linked gene SRY is a dominant inducer of testis development in mammals (Sinclair et al., 1990) and a founding member of a gene family with sequence homology to the High Mobility Group (HMG) domain (Fawcett and Klymkowsky, 2004). Since discovery of the SRY gene, many members of the SOX/Sox (SRY-related HMG box) gene family have been found throughout the vertebrates, showing at least 60% protein similarity to the SRY HMG domain. Genes in each subgroup show over 80% similarity. These SOX/Sox genes have also been found to be involved in physiological processes such as sex determination and the development of the CNS, neural crest and endoderm (Bowles et al., 2000). Sox1, Sox2, Sox3 and Sox11, for instance, are expressed mainly in the developing nervous system (Collignon et al., 1996; Pevny et al., 1998), and Sox4 is essential for heart and lymphocyte development (Schilham et al., 1996). The SOX/Sox genes have been divided between ten subgroups, named A to J (Table 1), not all of which occur in the same taxa (Bowles et al., 2000); groups I and J (containing sox31, sox32 and sox33), for instance, are only found in Zebrafish (Girard et al., 2001; Lunde et al., 2004).

Some amphibians have ZZ/ZW or XX/XY modes of sex determination, but most species do not have heteromorphic chromosomes, making the study of evolution and the mechanism of sex determination in these organisms interesting. *Rana tientaiensis* (2n = 26) is one of these species without identifiable sex chromosomes (Guo et al., 1991). In this paper, we describe the cloning and sequencing of the eight Sox genes in *Rana tientaiensis* with the aim of researching the diversity and evolution of this gene family. Sequence analysis indicates that some of these genes in *R. tientaiensis* are duplicated.

Materials and Methods

Two male and two female *Rana tientaiensis* were captured from Taolin and Tingxi Anhui Provinces, China. Genomic DNA was isolated from muscle tissues using routine protocols. A pair of degenerate primers (sn1: ATGAAYGCNTTYATGGTNTGG; sn2: GGNCGR-TAYTTRTARTCNGG) were designed using multiple alignments of the HGM-box sequence of *SRY*, corresponding to the MNAFMVW and PDYKYRP motifs found in the HMG boxes of a wide range of *Sox* proteins.

PCR reactions were 30 μ l in volume, including 18.75 μ l ddH₂O and approximately 100 ng of genomic DNA, 1.5mM Mg²⁺, 120 μ M dNTP, 0.3 μ M/primer and 1.25 μ l Taq polymerase. PCR cycling conditions were 1 cycle for 5 minutes at 97°C, followed 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 minutes to complete the final reaction.

Clones were genetically sequenced to detect the positive clones with *Sox* DNA insertions. PCR products were detected by 1.5% agarose gels and cloned by pMD 18-T Vector (purchased from TAKARA). Positive clones were screened by SSCP (single-strand conforma-

A	В	С	D	E	F	G	Н	I	J
SRY	Sox1	Sox4	Sox5	Sox8	Sox7	Sox15	Sox30	Sox31	Sox32
	Sox2	Sox11	Sox6	Sox9	Sox17	Sox16			Sox33
	Sox3	Sox24	Sox12	Sox10	Sox18	Sox20			
	Sox14	Sox22	Sox13						
	Sox21		Sox23						
	Sox25								

Table 1. Classification of the Sox gene family.



Figure 1. Amplified *Sox* gene fragments from 1: male *Rana tientaiensis*, 2: female *R. tientaiensis*,3: Human; 4: negative control; M: DL2000 marker (TaKaRa).

tion polymorphism) analysis and sequenced with the universal sequencing primers on an ABI377 autosequencer. DNA sequences were analyzed using the BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) and CLUSTALX programs. Molecular Evolutionary Genetic Analysis (MEGA) software was used to construct the phylogenetic tree.

Results and Discussion

A 203 bp fragment of *Rana tientaiensis* genomic DNA was obtained using the degenerate PCR primers listed above. The fragment was identical to that found in the human genome (Fig. 1), indicating that these fragments belong to homologous genes.

Of the 113 white clones (i.e., those with insertions), 64 were positive for the *Sox* insertion, as confirmed through nucleotide analysis following genomic amplification. Eight distinct positive clones were found in both male and female *Rana tientaiensis*; there were no sexual differences.

Seven of the eight distinct genes were named as follows: *RtSox3a*, *RtSox3b*, *RtSox3c*, *RtSox4*, *RtSox11*, *RtSox12*, and *RtSox14*. The amino acid sequences from these genes had 95%, 95%, 95%, 97%, 98%, 97%, 97% and 97% similarity (respectively) to homologous *SOX* genes in Human. These seven genes belonged to the *SoxB*, *SoxC* and *SoxD* subgroups, all of which lack introns (Bowles et al., 2000). The 9th *Sox* gene was provisionally named *RtSoxB14*; the amino acid sequences from this gene had 79% similarity to the Human *SOX21* gene, 77% similarity to the *Human Sox14* gene, as well as varying levels of similarity to other group B *Sox* genes. Nucleotide and putative amino acid sequences for the eight *Sox* genes are listed (Fig. 2)

The amino acid sequences from the eight clones were compared to 44 published *Sox* gene sequences in GenBank, including sequences from Human (*HomoSRY*, *SOX1, 2, 3, 4, 7, 9, 11, 12, 14, 15, 21, 30*), Mouse (*MusSox1, 2, 3, 4, 7, 9, 11, 12, 14, 15, 21, 30*), Gallus gallus (*GallSox2, 3, 9, 11, 14, 21*), Danio rerio (DaniSox1, 2, 4, 11), Xenopus laevis (XenopusSox2, 4, 11), Takifugu rubripes (TakifuguSox1, 14b) and Eremias breuchleyi (EbSox2, 4, 11, 12, 14, 21) (Table 2). All sequences were analyzed using neighbor-joining (NJ) methods by MEGA 2.0 (Fig. 3).

Amino acid sequences between the *RtSox* genes were highly conserved. Representatives of the *Sox* gene in other species were also highly conserved with much similarity between sequences. Gene duplication has likely caused most of the diversity seen in the HMG box superfamily, for which the *Sox* genes show the highest mutation rate (Laudet et al., 1993). The high similarity seen between the *Rana* and human genes in this study are certainly indicative of gene duplication.

In the case of Sox12, amino acid sequences were nearly identical, although the 10th amino acid was N instead of H in Human and Mouse (Fig. 2). In *Sox 4*, the 48th amino acid was R instead of Q in Mouse, and in *Sox11*, the 48th amino acid was D instead of N in Mouse, Human and several other species. The high degree of similarity amongst these genes suggests that they belong to the same gene family, and may perform similar roles amongst taxa. For example, the three highly conserved genes in group C display overlapping expression patterns, and *Sox4* and *Sox11* display overlapping expression patterns in the mouse embryonic pancreas (Lioubinski et al., 2003).

According to Laudet et al. (1993), *Sox4* is considered to be an early offshoot of the *SRY* gene in the *Sox* family phylogeny. The conservative nature of *Sox4* homologues in non-mammalian amniotes is interesting because in mammals, the *SRY* gene exhibits rapid evolu-

Sequence	Accession number	Sequence	Accession number			
Human sapiens		Mouse musculus				
HomoSOX1	NP_005977	MusSox1	BAC75667			
HomoSOX2	CAA83435	MusSox2	NP_035573			
HomoSOX3	CAA50465	MusSox3	AAH52024			
HomoSOX4	NP_003098	MusSox4	NP_033264			
HomoSOX7	NP_113627	MusSox7	NP_035576			
HomoSOX9	CAA86598	MusSox9	AAH04064			
HomoSOX11	BAA88122	MusSox11	NP_033260			
HomoSOX12	CAB81632	MusSox12	NP_035568			
HomoSOX14	AAI06731	MusSox14	XP 284529			
HomoSOX15	NP_008873	MusSox15	NP_033261			
HomoSOX21	NP_009015	MusSox21	NP_808421			
HomoSOX30	NP_848511	MusSox30	AAF99391			
HomoSRY	AAT37462	Danio rerio				
Gallus gallus		DaniSox1	NP_001032751			
GallSox2	NP_990519	DaniSox2	NP_001002483			
GallSox3	NP_989526	DaniSox4	BC065354			
GallSox9		DaniSox11	CAB87379			
GallSox11	NP_990518	Xenopus laevis				
GallSox14	NP_990092	XenopusSox2	AAB62821			
GallSox21	BAA77266	XenopusSox3	P55863			
Eremias breuchleyi		XenopusSox11	Q91731			
EbSox2	DQ067423	Takifugu rubripes				
EbSox4	DQ067426	TakifuguSox1	AAQ18494			
EbSox11	DQ067427	TakifuguSox14b	AAQ18499			
EbSox12	DQ067428					
EbSox14	DQ067430					
EbSox21	DQ067433					

Table 2. Sox genes sequence in different species

tion, possibly caused by Y-linked inheritance (Tucker and Lundrigan, 1993). The limited diversity of this gene in non-mammalian taxa may be due to the retention of an ancient conserved function.

It is likely that *Sox3* is the closest homologue to the *Sry* gene based on nucleotide sequence data. Furthermore, *Sox3* is located on the mammalian X chromosome, and is highly similar to SRY (Sinclair et al., 1990), suggesting that they arose through duplication of their common ancestral during differentiation of the sex chromosomes (Collignon et al., 1996; Foster and Graves, 1994; Stevanovic et al., 1993). This is significant because the X and the Y chromosomes are thought to have arisen from a common "autosome" ancestor in the lineage that gave rise to mammals (Wright et al., 1993). Further examination of *Sox* genes in lower vertebrates, Prototheria (monotremes) and Metatheria (mar-

supials) will be necessary to establish the evolutionary origins of *Sry*. The three conservative *Sox3* genes (sometimes found within the same species or individual) can be identified by the following variations in amino acid sequence: *RtSox3a* has an F at position 46 and an H at position 58; *RtSox3b* has an F at position 46 and an M at position 58; *RtSox3c* has an I on position 46 and an M on position 58.

The *RtSoxB14* is unique among the Sox genes in having the amino acid sequence VITEH at positions 15–19, a K at position 44 and an S at position 50. In the phylogenetic tree (Fig. 3), although *RtSoxB14* was more closely related to subgroup B than other groups, the origin and classification of this gene is ambiguous.

Previously, genes encoding proteins with more than 60% similarity to the *SRY* HMG domain have been named *Sox* (*SRY* box) genes, and *Sox* genes with at least

RtSox12	ATGAAT	GCGT	TT	ATGGT	ATGGT	ССС	AGA	ACG	AGO	GGG	GG	FAA	GAI	ICA	TGG.	AC	CAG	ΤG	GCCG
RtSox11	ATGAAT	GCTT	TT	ATGGT	ATGGT	CC A	AGA	TCG	AGO	CGG4	٩Ġ	ιÅÅ.	AA1	CA	TGG.	ÅG	CAG	TC	GCCC
RtSox4	ATGAAO	GCGT	TT	ATGGT	TTGGT	CGC	AGA	TCG	AGO	GGG	GGC	CAA	GAI	ICA	TGG.	AG	CAG	TC	GCCC
RtS ox3a	ATGAAT	GCTT	TT	ATGGT	TTGGT	CGC	GGG	GGC	AGO	CGGG	GG	CAA	GAI	ſGG	CTC.	AG	GAA	АA	сссс
RtSox3c	ATGAAT	GCGT	TT	ATGGT	TTGGT	CGC	GGG	GGC	AGO	CGGG	GG	CAA	GAI	ſGG	CTC.	AG	GAA	АA	сссс
RtS ox3b	ATGAAT	GCGT	TT	ATGGT	TTGGT	CGC	GGG	GGC	AGO	CGGC	GG	CAA	GAI	ſGG	CTC.	AG	GAA	АA	сссс
RtSoxB14	ATGAAO	GCGT	TT	ATGGT	CTGGT	CC A	GAG	TAC	AG	AGG/	٩GG	GAA	GG1	ſĠĂ	TTA	CA	GAA	СA	TCCI
RtSOX14	ATGAAT	GCGT	TC /	ATGGT	GTGGT	CC A	GGG	GGC	AG	AGG/	٩GG	GAA	GAI	ſGG	CCC.	AG	GAC	AA	TCCC
	*****	** *	* >	lokokok	*****	*		,	**	**	*	**	*	¢			*		**
RtSox12	GACAT	CACA.	AC (GCTGA	GATCT	CC A	AGC	GCC:	rco	GCC	G	CG	CT	GGC.	AGC	тс	CTG	СА	GGAC
RtSox11	GACAT	CACG.	AC (GCCGA	GATCT	CC A	AGC	GCC:	r G(GC /	A AC	ЮG	GT (3GA	AAA	ΤG	CTG	AA	GGAC

I GCACA AC TCTGA A. I GCACA AC TCTGA A. I GCACA AT TCGGA G.	ATTAGC AAAAAGTT G ATCAGT AAAAGACT T	GGGGGCACAGT GGA	AGAT CETTO COCOAC AGAT CETTO GEGAT
I GCACAAC TCTGAA	ATTAGE AAAAAGTT G	GGGGGCACAGT GGA	AGATCCTTGGCGAT
I OCACAA ICOORO.	AICICCAROCOCCI (000000000000000000000000000000000000000	MOCT OCTOROCORC
rankek kereaak a	ATCTCC & AGCGCCT G	асполольствол	ACCT CCTCA CCCAC
I GCACA AC TOGGAG.	ATCTCC AAGCGCCT G	GGCGCGGGACT GGA	AGET GET GAGEGAE
I GCACA AC TOGGA G.	ATCTCC AAGCGCCT G	GGCGCGGGACT GGA	AGCT GCTGA GCGAC
I GTACA AC GCCGAG.	ATCTCC AAGCGGCT A	AGGCA AACGCT GGA	AGCT GCTCA AGGAC
I GCACGAC GCCGAG.	ATCTCC AAGCGCCT G	GGCA AGCGGT GGA	AAAT GCTGA AGGAC
	I GEAEGAE GEEGAG. I GTAEAAE GEEGAG. I GEAEAAE TEGGAG. I GEAEAAE TEGGAG.	I GEACGAE GEEGA GATETEE AAGEGEET (I GTACAAE GEEGA GATETEE AAGEGGET / I GEACAAE TEGGA GATETEE AAGEGEET (I GEACAAE TEGGA GATETEE AAGEGEET (I GEACGAE GEEGA GATETEE AAGEGEET GGGEA AGEGGT GGA I GTAEA AE GEEGA GATETEE AAGEGGET AGGEA AAEGET GGA I GEACA AE TEGGA GATETEE AAGEGEET GGGEGGEGGAET GGA I GEACA AE TEGGA GATETEE AAGEGEET GGGEGEGGAET GGA

GICON	SAMA O	ROCOTRO										· · · ·	. on	10070	•
CTCCA	GAAAAG	<u>ՆՐՐՐԾՆՐ</u>	ATTGAC	GAAG	ЮСА	AAA	GGT	TGA	GGGI	CTCA	ACA	CAT	'GA I	IGGAI	1
TCAGA	GAAGA A	GCCTTTT	ATAGAC	GAAT	CAA	A A A	GGC	TGA	GAG	CTCA	.GCA	TAI	[GG]	TGA	;
GCGGA	GAAGC G	CCCTATT	ATCGAC	GAGG	ЮСА	AGC	GGC	TCC	GCGI	CCGT	CCA	CAI	(GA)	AGGAJ	ł
GCGGA	GAAGC G	CCCTTTT	ATCGAC	GAGG	ЮСА	AGC	GGC	TCC	GCG	CCGT	CCA	CAI	[GA]	AGGAJ	Ł
GCGGA	GAAGC G	CCCTTTC	ATCGAC	GAGG	ЮСА	AGC	GGC	TCC	GCGI	CCGT	CCA	CAC	GA J	AGGAJ	ł
AGCGA	CAAGAT.	ICCGTIC	ATCCAG	GAGO	ЮGG	AGC	GAC	TGC	GCC	TCAA	.GC A	CAI	GGG	CTGAO	2
AGCGA	GAAGAT(CCCCTTC	ATCCGC	GAGG	ЮCG	AGC	GGC	TGC	GAC.	ГСАА	.GCA	CAI	GGG	CTGAO	2
TCGGA	GAAGAT	CCCCTTT	GTGAAG	GAGO	ЮTG	AGC	GGC	TGC	GAC	TCAA	.GCA	CAI	GGG	CTGAO	2
	TCGGA AGCGA AGCGA GCGGA GCGGA GCGGA TCAGA	TCGGAGAAGAT AGCGAGAAGAT AGCGACAAGAT GCGGAGAAGCG GCGGAGAAGCG GCGGAGAAGCG TCAGAGAAAGAA GTCGAGAAAAG	TCGGAGAAGATCCCCTTT AGCGAGAAGATCCCCTTC AGCGACAAGATTCCGTTC GCGGAGAAGC @CCTTTT GCGGAGAAGC @CCTTTT TCAGAGAAGC @CCTTTT GTCGAGAAAAGACCCTTA	TEGGA GAAGATE CECTT TGTGAAG AGEGA GAAGATE CECTT CATECGE AGEGA CAAGATT CEGTT CATECAG GEGGA GAAGE GE CETTT CATEGAE GEGGA GAAGE GE CETTT TATEGAE GEGGA GAAGE GE CETAT TATEGAE TEAGA GAAGA AGECETA CATTGAE	TEGGA GAAGA TE CECTT TGTGAAG GAGG AGEGA GAAGA TE CECTT CATECGE GAGG GEGGA GAAGA TT CEGTT CATECAG GAGG GEGGA GAAGE Œ CETTT CATEGAE GAGG GEGGA GAAGE Œ CETTT TATEGAE GAGG TE AGA GAAGE Œ CETTT TATEGAE GAG TE AGA GAAGA AGECTTT TATAGAE GAAT GTEGA GAAAA GAECETA CATTGAE GAAC	TEGGA GAAGA TE CEETT TOTGAAG GAGGETG AGEGA GAAGA TE CEETT CATEEGE GAGGEEG AGEGA CAAGA TE CEGTT CATEEGE GAGGEEG GEGGA GAAGE OF CETTT CATEGAE GAGGEEA GEGGA GAAGE OF CETTT TATEGAE GAGGEEA GEGGA GAAGE OF CETAT TATEGAE GAGGEEA TEAGA GAAGA AGEETT TATAGAE GAATEAA GTEGA GAAAA GA EECTT ATTGAE GAAGEEA	TEGGA GAAGA TE CECTT TGTGAAG GAGGETGAGE AGEGA GAAGATE CECTT E ATEEGE GAGGEEGA GE GEGGA GAAGA TT EEGTT E ATEEGAE GAGGEEGA GE GEGGA GAAGE Œ CETTT E ATEGAE GAGGEEAA GE GEGGA GAAGE Œ CETTT TATEGAE GAGGEEAA GE TE AGA GAAGE Œ CETTT TATEGAE GAGGEEAA AA GTEGA GAAAA GAEEET A E ATTGAE GAAGEEAA AA	TEGGA GAAGA TE CECTT TGTGAAG GAGGETGA GEGGE AGEGA GAAGA TE CECTT CATEEGE GAGGEEGA GEGGE GEGGA GAAGA TT CEGTT CATEEGE GAGGEEGA GEGGE GEGGA GAAGE GE CETTT CATEGAE GAGGEEAA GEGGE GEGGA GAAGE GE CETTT TATEGAE GAGGEEAA GEGGE TEAGA GAAGE GE CETTT TATEGAE GAGGEEAA GEGGE GEEGA GAAAG AGEETTT TATAGAE GAATEAAA AAGGE	TEGGA GAAGATE CECTT TGTGAAGGAGGETGAGEGGETGE AGEGA GAAGATE CECTT EATECGE GAGGEEGA GEGGETGE AGEGA GAAGATE CECTT EATECAG GAGGEGGAGEGAETGE GEGGA GAAGE OF CETTT EATEGAE GAGGEEAA GEGGETEE GEGGA GAAGE OF CETTT TATEGAE GAGGEEAA GEGGETEE TEAGA GAAGE OF CETTT TATEGAE GAGGEEAA GEGGETEG TEAGA GAAGA AGEETTT TATAGAE GAAGEEAA AAGGETGA	TEGGA GAAGA TE CECTT TGTGAAG GAGGETGA GEGGE TGEGAE AGEGA GAAGA TE CECTT E ATEEGE GAGGEEGA GEGGE TGEGAE AGEGA CAAGA TE CEGTT E ATEEGA GAGGEEGA GEGGE TGEGGE GEGGA GAAGE GE CETTT E ATEGAE GAGGEEAA GEGGE TEEGEG GEGGA GAAGE GE CETTT TATEGAE GAGGEEAA GEGGE TEEGEG GEGGA GAAGE GE CETAT TATEGAE GAGGEEAA GEGGE TEEGEG TE AGA GAAGA AGECTTT TATAGAE GAATEAAA AAGGE TGAGAG GTEGA GAAAA GAEEETA E ATTGAE GAAGEEAA AAGGT TGAGGG	TEGGA GAAGATE CECTT TGTGAAGGAGGETGAGEGGE TGEGAE TEAA AGEGA GAAGATE CECTT EATECGE GAGGEEGAGEGGE TGEGAET CAA AGEGA GAAGATT CEGTT EATECAG GAGGEEGAGEGGAE TGEGEE TEAA GEGGA GAAGE OF CETTT EATEGAE GAGGEEAA GEGGE TEEGEGEECGT GEGGA GAAGE OF CETTT TATEGAE GAGGEEAA GEGGE TEEGEGECEGT TEAGA GAAGE OF CETTT TATEGAE GAGGEEAA GEGGE TEEGEGECEGT TEAGA GAAGA AGECTTT TATAGAE GAAGEEAA AAGGE TGAGAGETEA GTEGA GAAAA GAEECTT TATAGAE GAAGEEAA AAGGT TGAGGGETEA	TEGGA GAAGA TE CEETT TGTGAAG GAGGETGA GEGGE TGEGAE TEAAGEA AGEGA GAAGATE CEETT E ATEEGE GAGGEEGA GEGGE TGEGAET EAAGEA AGEGA CAAGA TT EEGTT E ATEEGAE GAGGEEGA GEGGE TGEGEE TEAAGEA GEGGA GAAGE GE CETTT E ATEEGAE GAGGEEAA GEGGE TEEGEGE EGTEEA GEGGA GAAGE GE CETTT TATEEGAE GAGGEEAA GEGGE TEEGEGE EGTEEA GEGGA GAAGE GE CETAT TATEGAE GAGGEEAA GEGGE TEEGEGE CEGTEEA TEAGA GAAGE AGECETTT TATEGAE GAAGEEAA AGGE TGAGAGETEAEA	TEGGA GAAGA TE CECTT TGTGAAG GAGGETGA GEGGE TGEGAE TE AAGEA CAT AGEGA GAAGA TE CECTT E ATEEGE GAGGEEGA GEGGE TGEGAET CAAGEA CAT AGEGA GAAGA TT CEGTT CATEEGE GAGGEEGA GEGGE TGEGEE TE AAGEA CAT GEGGA GAAGE GE CETTT E ATEGAE GAGGEEAA GEGGE TE GEGEE CEGEC GTE CA CA GEGGA GAAGE GE CETTT TATEGAE GAGGEE AA GEGGE TE GEGEC GTE CA CA TE AGA GAAGE GE CETTT TATEGAE GAGGEEAA GEGGE TE GEGEC GTE CA CA TE AGA GAAGE GE CETTT TATEGAE GAGGEE AA AGGE TGAGAGETCA CA TAT GEGGA GAAAG AGECTTT TATEGAE GAAGEE AA AAGGE TGAGAGETCA CA TAT	TEGGA GAAGA TE CEETT TETGAAG GAGGETGA GEGGE TGEGAE TEAAGEA CATGGE AGEGA GAAGATE CEETT CATEEGE GAGGEEGA GEGGE TGEGAET CAAGEA CATGGE GEGGA GAAGE GE CETTT CATEEGAE GAGGEEGAE GEGGE TGEGEGE CTEAAGEA CATGGE GEGGA GAAGE GE CETTT TATEGAE GAGGEEAA GEGGE TEEGEGE EGTECA CAEGA/ GEGGA GAAGE GE CETTT TATEGAE GAGGEEAA GEGGE TEEGEGE CGTECA CAEGA/ TEAGA GAAGE GE CETTT TATEGAE GAGGEEAA GEGGE TEEGEGECEGTECA CAEGA/ TEAGA GAAGE GE CETTT TATEGAE GAGGEEAA AGEGE TEEGEGECEGTECA CAEGA/ GEGGA GAAGE AGECETTT TATEGAE GAGGEEAA AGEGE TEEGEGECEGTECA CAEGA/ TEAGA GAAGA AGECTTT TATAGAE GAATEAAA AAGGE TGAGGEETEAACAE ATGGA	TEGGA GAAGA TE CECTT TOTGAAG GAGGETGA GEGGE TGEGAE TEAAGEA CATGGETGAE AGEGA GAAGATE CECTT CATEEGE GAGGEEGA GEGGE TGEGAE TEAAGEA CATGGETGAE AGEGA GAAGATE CECTT CATEEGE GAGGEEGA GEGGE TGEGECETEAAGEA CATGGETGAE GEGGA GAAGE GE CETTT CATEGAE GAGGEEAA GEGGE TEEGEGEEGTEEA CATGA AGGA/ GEGGA GAAGE GE CETTT TATEGAE GAGGEEAA GEGGE TEEGEGEEGTEEA CATGA AGGA/ GEGGA GAAGE GE CETTT TATEGAE GAGGEEAA GEGGE TEEGEGECGTEEA CATGA AGGA/ TEAGA GAAGE GE CETTT TATEGAE GAGGEEAA AEGGE TEEGEGECGTEEA CATGA AGGA/ TEAGA GAAGE AGECTTT TATEGAE GAAGEEAA AAGGE TGAGAGETEACAEATATGATGAE GTEGA GAAAA GA EEETTA CATTGAE GAAGEEAA AAGGE TGAGAGETEACAEATATGATGAE

RtSox12	TACCE TGACT AC AAATA TEGECE
RtSox11	TACCC CGATT AC AAATACCGCCC
RtSox4	TACCE TGACT AC AAATACEGEEE
RtS ox3a	TACCE GGATT AC AAATACEGEEE
RtS ox3c	TACCC CGATT AC AAATACCGCCC
RtS ox3b	TACCC TGACT AT AAATACCGCCC
RtSoxB 14	CATCC CGACT AC AAAT ACCGCCC
RtSOX14	CACCCTGACT AC AAATATCGCCC
	* ** ** ** ** ** **

Figure 2. (A) Alignment of nucleotide sequences (above), (B) Alignment of amino acid sequences (Opposite page, top), (C) Percentage amino acid similarity between *Rana tientaiensis Sox* clones as determined by the sequence identity matrix function in Bioedit (Opposite page, bottom).

RtSox <u>3a</u>	MNAFMVWSRGQRRKMAQENPKMHNSEISKRLGADWKLLSDAEKRPFIDEAKRLRAVHTKEYPDYKYR
RtSox <u>3c</u>	***************************************
RtSox3b	***************************************
RtSox14	*******Y*******D*****D****************
RtSoxB14	********V****VIT*H********K***Q****G*S**K***S*************
RtSox11	******KIE***IMEQS*D**DA****KR**M*K*S**I***R**E***LK*MAD******
RtSox4	******QIE***IMEQS*D*Y*A****KR***K*SD*I***Q**E***LK*MAD******
RtSox12	******QNE***IMDQW*D***A*****RR*Q**Q*S**I**VK**E***LK*MAD******

Sox B Sox C Sox D

	HomoSRY	RtSox3a	RtSox3b	RtSox3c	RtSox14	RtSoxB14	RtSox11	RtSox4	RtSox12
Homo SRY	100	67.6	66.1	67.6	63.2	60.2	54.4	54.4	52.9
RtSox3a		100	97.0	98.5	88.2	76.4	64.7	64.7	64.7
RtSox3b			100	98.5	89.7	76.4	64.7	64.7	64.7
RtSox3c				100	89.7	77.9	66.1	66.1	66.1
RtSox14					100	76.4	61.7	61.7	61.7
RtSoxB14						100	63.2	61.7	61.7
RtSox11							100	91.1	83.8
RtSox4								100	85.2
RtSox12									100

Figure 2 (continued).

80% similarity have been placed in the same subgroup. However, as more and more *Sox* genes are identified, the ability to accurately classify these genes decreases. For example, the genes *sox30* and *Ce-soxj* have only 46% and 48% similarity to *Human SRY* HMG. Bowles (2000) attempted to alternatively diagnose the gene family by possession of the amino acid sequence "RPMNAF", which is highly conserved across genes, however, this sequence was also found in the taxonomically ubiquitous gene *cic*, so the sequence "RPMNAFMVW" was provided as a replacement (Bowles et al., 2000).

Now that a sequence identifying the Sox gene family has been identified, can sequences be found to characterize the Sox subgroups? Following analysis of the available sequences (Figure 4), it appears that positions 15-19 may have be useful for this purpose. The sequence "MAQE(D)N" may work for group B (except for HomoSOX3 and MusSox3), "IMEQS" for group C, "IMEQW" (for Sox12) or "ILQAF" (for Sox5, Sox6 and Sox13) for group D, "LADQY" for group E, "LAVQN" (for Sox7) or "LAQQN" (for Sox17 and Sox18) for group F, "MAQQN" for group G and "LAKAN" for group H. RtSoxB14 can be separated from the remaining Sox genes by the unique sequence VITEH, although in mammals (HomoSOX3 and MusSox3) it is changed to "MALEN", like the sequence for SRY. This further supports a close relationship between the SOX3 and SRY genes.

Several Sox genes appear to have been duplicated in Rana tientaiensis: RTSox3a, RTSox3b and RTSox3c. Similar duplications in amphibians are uncommon, but they are more frequently encountered in teleosts: Sox1, 4, 9 and 14 has been duplicated in the sea bass (Malyka et al., 2003) and Sox21 has been duplicated in the Zebrafish (Argenton et al., 2004). The "duplicationdegeneration-complementation" model developed by Force et al. (1999) suggests that the partition of ancestral subfunctions is an important mechanism leading to the preservation of multiple gene copies; this model predicts that the probability of gene conservation will be higher in more complex genes with a larger number of subfunctions (Force et al. 1999). Most duplicate genes in Rana tientaiensis have silent mutations (except RtSox3a, which has an encoded amino acid mutation), but it would appear that the sequences are under selective pressure and may indeed perform separate subfunctions. Future studies investigating Sox genes in Rana tientaiensis will likely provide much insight into duplicate genes.

Acknowledgments

We are grateful to the reviewers for numerous valuable suggestions on the manuscript. This research was supported by National Natural Science Foundation of China (No. 30640048, No. 30770296), the Natural Science Foundation of Anhui Education Department (KJ2007A022).



2008



Figure 3. Phylogenetic analysis of Sox/SOX gene family

mus sox30	M NAFMV WARIHRP ALAKANP AANNAEIS VQLGLEWNKLS EEQKKPY YDEAQ KIKEKHREE F PGW VYQ P
hom o s o x 30	M NAFMV WA RIHRP ALAKANP AANNAEIS VQLGLEWNKLS EEQKKPY YDEAQ KIKEKHREE F PGW Y Q P
MusSox4	M NAFMV WSQIERR NIMEQSE DMHNAEIS KRLGKRWKLLKDSDKIPF IQEAERLRLKHM AD Y PDYKYRP
RTS ox4	M NAFMV WSQIERR KIMEQSP DMYNAEIS KRLGKRWKLLKDSDKIPF IQEAE RLRLKHM AD Y PDYKYR P
Hom oSox4	M NAFMV WSQIERRRIMEQSP DMHNAEIS KRLGKRWKLLKDSDKIPF IREAE RLRLKHM AD Y PDYKYRP
DanioSox4	M NAFMV WSQIERR KIMEQSP DMHNAEIS KRLGKRWKLLKDSDKI PFIREAE RLRLKHM AD Y PDYKYR P
EbS ox4	M NAFIV WS RIERR KIMEQSP DMHNAEIS KRLGKRWKLLKDSDKIPF IQEAE RLRLKHM AD Y PNYKYR P
Hom oSox11	M NAFMV WS KIERR KIMEQS [®] DMHNAEIS KRLGKRWKMLKDSEKIPF IREAERLRLKHM AD Y PDYKYRP
MusSox11	M NAFMV WS KIERR KIMEQSP DMHNAEIS KRLGKRWKMLKDSEKIPF IREAE RLRLKHM AD Y PDYKYR P
GallusSox11	M NAFMV WS KIERREIMEQSP DMHNAEIS KRLGKRWKMLEDSEKIPF IREAE RLRLEHM AD YPDYKYRP
XenopusSox11	M NAFMVWS KIERRKIMEQSP DMHNAEIS KRLGKRWKMLKDSEKIPF IREAE RLRLKHMAD Y PDYKYRP
RTS ox11	M NAFMV WS KIERR KIMEQSE DMHDAEIS KRLGKRWKMLKDSEKIPF IREAERLRLKHM AD Y PDYKYRP
EbSox11	M NAFIVWS KIERRKIMEQSP DMHNAEIS KRLGKRWKMLKDSEKIPF IREAE RLRLKHMAD Y PDYKYRP
DanioSox11	M NAFMV WS KIERRKIMEQSP DMHNAEIS KRLGKRWKMLKDSEKIPF IREAE RLRLQHM AD Y PDYKYRP
Hom oSox12	M NAFMV WSQHERR KIMDQWP DMHNAEIS KRLGRR WQLLQDSEKIPF VREAERLRLKHM AD Y PDY KYRP
MusSox12	M NAFMV WSQHERRKIMDQWP DMHNAEIS KRLGRRWQLLQDSEKIPF VREAE RLRLKHM AD Y PDYKYRP
Hom o s ox 22	M NAFMV WSQHERR NIMDOWE DMHNAEIS KRLGR RWQLLQ DSEKIPF VREAE RLRLKHM AD Y PDYKYR P
RTS ox12	M NAFMV WSQNERR KIMDQWP DMHNAEIS KRIGR RWQLIQ DSEKIPF V KEAE RIRIKHM AD Y PDY KYR P
EbS ox12	M NAFI V WS QNERR NI MDQWE DMHNAEI SKRLGRR WQLLQ DSEKI PF V KEAERLRLKHM AD YFDY KYR P
MusSox7	M NAFMV WA KDERKELAVQNP DLHNAELS KMLGKSWKALT LSQKRPY VDEAE RLRLQHMQD Y FNYKYR P
Hom oSox7	M NAFMV WA KDERK FLAVQNP DLHNAELS KMLGKSWKALT LSQKRPY VDEAE RLRLQHMQD Y FNYKYR P
Hom o s ox7	M NAFMV WA KDERK HLAVQNP DLHNAELS KMLGKSWKALT LSQKRPY VDEAE RLRLQHMQD Y FNYKYR P
Homosox17	M NAFMV WA KDERK RLAQQNP DLHNAELS KMLGKSWKALT LAEKRPF VEEAE RLRVQHMQD HF NYKYR P
Hom o sox18	M NAFMV WA KDERK FLAQONP DLHNAVLS KMLGK AW KELN AA EKR PF VEEAE RLRVOHLRD HF NY KYR P
Hom oSox9	M NAFMV WAQAARR KLADQYP HLHNAELS KTLGKLWRLLNESEKRPF VEEAE RLRVQHKKD HFDYKYQP
MusSox9	MNAFMV WAQAARR MLADOYP HLHNAELS KTLGKLWRLLNESEKRPF VEEAE RLRVQHKKD HFDYKYOP
GallusSox9	MNAFMV WAQAARR KLADOYP HLHNAELS KTLGKLWRLLNE SEKRPF VEEAERLRVOHKKD HEDYKYOP
Homosox10	M NAFMV WAQAARR KLADQYP HI HNAELS KTLOKI WRLLNESDKRPF TEEAE RI RMQHKKD HEDYKYQP
MusSox3	M NAFMV WS RGORR MMALENP KMHNSEIS KRLGADWKLLT DAEKRPF IDEAKRLRAVHM KE Y FDYKYRP
Hom oSox3	M NAFMV WS RGORR MMALENP KMHNSETS KRLGA DWKLLTD AFKRPF TDFAK RLRAVHM KE Y PDY KYRP
GallusSox3	M NAFMV WS RGORR KMADENP KMHNSETS KRI (34 DWKT I S DAEKRPF TDEAKRI RAVHM KE Y PDY KYRP
RTS or 3c	M NAFMY WS ROORR IMAGENE KMHNSETS KELGA DWKLLS DARKEPET DE AKELEAVHM KE Y PDY KYRP
RTS ox3a	M NAFMV WS RGORE KMADENP KMHNSETS KELGA DWKLLS DAEKEPF TDEAKELBAVHT KE V PDVKVEP
RTS ox3b	M NAFMV WS RGORE KMADENP KMHNSETS KELGA DWKLLS DAEKEPT TDEAKELEAVHM KE Y PDY KYRP
XenonusSor3	M NAFMY WS ROORER MADENE KINNESETS KELOADWILLS DEDKRET TERAKTI RAVIM KELYDDIR THE THE
HomoSov1	M NAFMANS ROORE IMAGENE KIMINGETS KRI GAFWENNE FAFEREFT DEAKRI RALIMEET POTITI
MurSov1	M NARMAWS ROBERTMADEND FUMINERTS FRI GARWEIMES RAREERET DRAFFI RALIMMER MEDVEVER
DanioSov1	M NAFMV WS RCORE EN ADENE KINDELES KELCA FWKVMS FAFKREFT DEAKELENED WER HED YEVEF
Taki fuguSov1	M NAFMV WS RGORE KNADENP KMENSETS KELGA FWKVMT FAFKREFT DEAKEL RAMEM KEHPDY KYRF
GollueSor?	MARKIN STORIETANINGEN AND DESTRUCTION AND DESTRUCTION AND TO DESTRUCT AND
VenonusSox2	M NAFM V WS ROORE KMAOENP KMHNSETS KRIJGA FWKLIS EARKREFT DEAKRIJAALMM KE HED V KYRF
EbS ox2	M NAFT VWS RGORR MMADENP KMHNSETS KRLGA RWKLLS RAEKRPFTDEAKRLRALHM KEH PNYKYRF
HomoSox2	M NAFMV WS RGORE KMADENP KMENSETS KELGA FWKLLS ETEKEPFTDEAKELEALEM KEHPDYKYRF
MurSov2	M NAFMAN'S BOORD IM AOFIND FININGETS FRI GAFWELLS FTEERPETDEAFRI RALIM FEMPINEUR F
DanioSov2	M NAFMV WSRCORRENADENFEMINISETS KRIGA FWKI I SESEKREFTIDEAKRI RALIMISETH DIKITI
MusSov21	M NAFMU WS RAORE WAADENF KINDISETS KREGAFWKTET FESKREFT DEAKRERAMIN KENPEYKYRF
HonoSov21	M NARMAWS RAORRINA OF MENDERS KRI CARWALL TECHNIKA TO DAARA BODANA AT DIATAT
FbS og 21	MARTIN'S DAORD MAADENE KUNDEETE KELOA DWELLTEETE DEAKELTEETE AKDEL DAMAN KE MEDITEETE
CollineSon21	IN MALEY IS INVESTIGATION OF THE AND T
Var Sould	III WATHIY IS RAQUE HIMAQENE NII USETS KUSAEN KUSA
Hun Sould	III MATIII Y IIS KOORKAAN MARENT KUUMSELS KALONEN KLES ENEKALI TEENKALANGUU KE HI DI KI KI H MARUUNS DOODD MAADIN VUUMSETE VDI OA DUVI I O DAVDDV TERANDI DAOUH VE UDEVIVO D
musbox14	MARTMY NEROSORDAN AGENE AND AND AND AND AND A COMPACT AND
Gallusbox14	M NAPHY AS NOW RETAINING TO AND AND AND AND A COMPANY A COMPANY AND A COMPANY AND A COMPANY AND A COMPANY AND A CO
EDD 0X14	IN MALINA NO DOWN AND AND AND AND AND A CARACTER AND A CARACTER AND A COMPANY AND A COMPANY AND A CARACTER AND A CA
A150X14	M NACHAN NO KONDRAN KANANGEN KANANGEN KANANGA BAKALAS BARKELI IDEAK KIKANAN KE HEDYKIRP
lakifuguSox145	M NAPROVECTIVE DIVERTIME AND ADDRESS OF A COMPANY ADDRESS TO A COMPANY ADDRESS TO A COMPANY ADDRESS AND ADDRESS AND ADDRESS ADDRE
ALS 0XB14	M NALMA A DE VARK MATTERE KWUNDELE KUTA MAKTIGIZEKKEL IDEZK KI KAMA A HADAKABA
HomoSox15	MNAFMYNSSAURRUMAUUNF KMENSEISKELGAUWKLIDEDEKRPFVEEAKRIRARHIRD Y PDYKYRP
MusSox15	M NAFMY WS SVYKKYMAJQNP KMHNSEIS KRLGAQWKLLGDEEKRPF VEEAKRLRARHLRD Y PDYKYRP
HomoSry	M NAF 1 Y N5 KUQKK KMALENF KMKNSE 15 KQLGI QWKMLT BAEKWPF FQEAQ KLQAMHREK Y PNYKYR P
Hom o sox5	M NAFMV WA KUERR KILQAFP DMHNSNIS KILGS RWKAMT NLEKQPY YEEQA RLSKQHLEK Y PDY KYKP
Homosox13	M NAFMV WA KDERR KILLQAFP DMHNSSIS KILGS RWKSMT NQEKQPY YEEQA RLSRQHLEK Y PDYKYKP
Hom o s o x 6	M NAFMV WAKUERRH <mark>ilgafp</mark> DMHNSNIS KILGS RWKSMS NQEKQPY YEEQARLSKIHLEK YP NYKYKP
HomoTcf-1	MYKETVYSAFNLLMHYPPPSGAGQHPQPQPPLHKANQPPHGVPQLSLYEHFNSPHPTPAPADISQKQV

Figure 4. Characteristic Sox amino acid sequences.

Literature Cited

- Bowles, J., G. Schepers, and P. Koopman. 2000. Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. Developmental Biology 227: 239– 255.
- Collignon, J., S. Sockanathan, A. Hacker, M. Cohen-Tannoudji, and D. Norris. 1996. A comparison of the properties of Sox3, with SRY and two related genes, Sox1 and Sox2. Development 122: 509–520.

- deMartino, S., Y. L. Yan, and T. Jowett. 2000. Expression of SozII gene duplicates in zebrafish suggests the reciprocal loss of ancestral gene expression patterns in development. Developmental Dynamics 217(3): 279–292.
- Fawcett, S. R. and M. W. Klymkowsky 2004. Embryonic expression of Xenopus laevis SOX7. Gene Expression Patterns 4(1): 29–33.
- Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531–1545.
- Foster, J. W. and J. A. Graves. 1994. An SRY-related sequence on the marsupial X chromosome: Implications for the evolution of the mammalian testis-determining gene. Proceedings of the National Academy of Sciences. USA 91: 1927– 1931.
- Girard, F., F. Gremazy and P. Berta. 2001. Expression pattern of the Sox 31 gene during zebrafish embryonic development. Mechanisms of Development 100(1): 71–73.
- Guo, C. W., Y. W. Dong and S. H. Zhang.1991. Studies on the karyotype and ag-NoRs of *Rana tientaiensis* and *Microhyla*. Hereditas 13(2): 6–8.
- Hagiuda, J., Y. Hiraoka, M. Hasegawa, M. Ogawa and S. Aiso. 2003. A novel Xenopus laevis SRY-related gene, xSox33. Biochem. Biophys. Acta 1628(2): 140–145.
- Laudet, V., D. Stehelin and H. Clevers. 1993. Ancestry and diversity of the HMG box superfamily. Nucleic Acids Research 21(10): 2493–2501.
- Lioubinski, O., M. Muller, M. Wegner and M. Sander. 2003. Expression of Sox transcription factors in the developing mouse pancreas. Developmental Dynamics 227: 402–408.
- Lunde, K., H.-G. Belting and W. Driever. 2004. Zebrafish pou5f1/pou2, homolog of mammalian Oct4, functions in the endoderm specification cascade. Current Biology 14(1): 48–55.
- Ohno, S. 1970. Evolution by gene and genome duplication. Springer Verlag, Berlin.
- Pevny, L. H., S. Sockanathan, M. Placzek, and R.

Lovell-Badge. 1998. A role for SOX1 in neural determination. Development 125(10): 1967–78.

- Rimini, R., M. Beltrame, F. Argenton, F. Cotelli and M. E. Bianchi. 1999. Expression patterns of zebrafish SOZ11A, SOS11B and So.z21. Mechanisms of Development 89: 167–171.
- Schilham, M. W., M. A.Oosterwegel, P. Moerer, J. Ya, P. A. deBoer, M. Van der Wetering, S. Verbeek, W. H. Lamers, A. M. Kruisbeek, A. Cumano and H. Clevers.1996. Defects in cardiac outflow tract formation and pro-B-lymphocyte expansion in mice lacking Sox-4. Nature 380(6576): 711–714.
- Sinclair, A. H., P. Berta and M. S. Palmer. 1990. A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. Molecular Phylogenetics and Evolution 3(1): 1–9.
- Sinclair, A. H., P. Berta, M. S. Palmer, J. R. Hawkins, B. L. Griyths, M. J. Smith, J. W. Foster, A. M. Frischauf, R. Lovell-Badge, and P. N. Goodfellow. 1990. A gene from the sex-determining region encodes a protein with homology to a conserved DNA-binding motif. Nature 346: 240–244.
- Stevanovic, M., R. Lovell-Badge, J. Collignon, and P. N. Goodfellow. 1993. SOX3 is an X-linked gene related to SRY. Human Molecular Genetics 2: 2013– 2018.
- Tucker, P. K. and B. L. Lundrigan. 1993. Rapid evolution of the sex determining locus in Old World mice and rats. Nature 364(6439): 715-717.
- Wright, E. M., B. Snopek. and P. Koopman. 1993. Seven new members of the Sox gene family expressed during mouse development. Nucleic Acids Research 21: 744.