Louisiana State University
LSU Digital Commons

Faculty Publications | Department of Biological Sciences

3-1-2006

Rapid evolution of a recently retroposed transcription factor YY2 in mammalian genomes

Chunqing Luo
Louisiana State University

Xiaochen Lu
Lawrence Livermore National Laboratory

Lisa Stubbs
Lawrence Livermore National Laboratory

Joomyeong Kim
Louisiana State University

Follow this and additional works at: https://digitalcommons.lsu.edu/biosci_pubs

Recommended Citation

Luo, C., Lu, X., Stubbs, L., & Kim, J. (2006). Rapid evolution of a recently retroposed transcription factor YY2 in mammalian genomes. *Genomics, 87* (3), 348-355. https://doi.org/10.1016/j.ygeno.2005.11.001

This Article is brought to you for free and open access by the Department of Biological Sciences at LSU Digital Commons. It has been accepted for inclusion in Faculty Publications by an authorized administrator of LSU Digital Commons. For more information, please contact ir@lsu.edu.

Available online at www.sciencedirect.com
SCIENCE DIRECT
Genomics 87 (2006) 348 – 355
GENOMICS
www.elsevier.com/locate/ygeno

Rapid evolution of a recently retroposed transcription factor *YY2* in mammalian genomes☆,☆☆

Chunqing Luo [a], Xiaochen Lu [b], Lisa Stubbs [b], Joomyeong Kim [a,*]

[a] *Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA*
[b] *Genome Biology Division, Lawrence Livermore National Laboratory, Livermore, CA 94551, USA*

Received 30 June 2005; accepted 10 November 2005
Available online 27 December 2005

Abstract

YY2 was originally identified due to its unusual similarity to the evolutionarily well-conserved zinc finger gene *YY1*. In this study, we have determined the evolutionary origin and conservation of *YY2* using comparative genomic approaches. Our results indicate that *YY2* is a retroposed copy of *YY1* that has been inserted into another gene locus named *Mbtps2* (membrane-bound transcription factor protease site 2). This retroposition is estimated to have occurred after the divergence of placental mammals from other vertebrates based on the detection of *YY2* only in the placental mammals. The N- and C-terminal regions of *YY2* have evolved under different selection pressures. The N-terminal region has evolved at a very fast pace with very limited functional constraints, whereas the DNA-binding, C-terminal region still maintains a sequence structure very similar to that of *YY1* and is also well conserved among placental mammals. *In situ* hybridizations using different adult mouse tissues indicate that mouse *YY2* is expressed at relatively low levels in Purkinje and granular cells of cerebellum and in neuronal cells of cerebrum, but at very high levels in testis. The expression levels of *YY2* are much lower than those of *YY1*, but the overall spatial expression patterns are similar to those of *Mbtps2*, suggesting a possible shared transcriptional control between *YY2* and *Mbtps2*. Taken together, the formation and evolution of *YY2* represent a very unusual case where a transcription factor was first retroposed into another gene locus encoding a protease and survived with different selection schemes and expression patterns.
© 2005 Elsevier Inc. All rights reserved.

Keywords: Retroposition; Evolution; Zinc finger transcription factor

Introduction

The transcription factor *YY1* is a *Gli-Kruppel* type zinc finger protein and controls the transcription of a large number of viral and cellular genes. *YY1* can function as a repressor, activator, or transcriptional initiator depending on the sequence context of YY1-binding sites with respect to other regulator elements [21]. The protein has a DNA-binding domain at the C-terminus and other modulating domains at the N-terminus displaying repression, activation, and protein-protein interaction activities. *YY1* interacts with several key transcription factors, including TBP, TAFs, TFIIB, and Sp1 [2,4,12,20,23]. Other studies also indicated that *YY1* recruits histone-modifying enzymes including p300, HDACs, and PRMT1 for transcription control [13,17,25]. Physiological roles for *YY1*have been demonstrated in mouse by gene knockout experiments, in which homozygous mutant mice show peri-implantation lethality and a subset of heterozygous mice show developmental abnormalities, such as excenphaly (or open brain) [5].

YY1 is evolutionarily well conserved throughout all vertebrate lineages although no systematic and comprehensive studies to date have addressed the evolutionary history of this gene. At least two genes similar to vertebrate *YY1* are found even in fly, and one of them is known to be involved in a heritable silencing mechanism as a component of the Polycomb complex [3]. Many key transcription factors, including *Sp1* and *E2F*, have evolutionary histories similar to that of *YY1*. These

☆ Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under Accession No. DQ107161.

☆☆ The U.S. Government's right to retain a nonexclusive royalty-free license in and to the copyright covering this paper, for governmental purposes, is acknowledged.

* Corresponding author. Fax: +1 225 578 2597.
E-mail address: jkim@lsu.edu (J. Kim).

0888-7543/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2005.11.001

transcription factors are conserved throughout the vertebrate lineage as well as in some invertebrates. In most cases, the gene copy number of these transcription factors has increased with the increase of physiological complexity of vertebrate animals and they exist as multigene families in the available genome sequences of vertebrates [1,8,10]. Genome-wide and segmental duplications, DNA-mediated, are thought to be responsible for this increase of gene number in vertebrates [7]. Occasional retropositions, RNA-mediated, have also contributed to the increase in gene number in vertebrates [6].

Consistently, another gene sequence with significant similarity to *YY1* has been identified in the human genome and thus was named *YY2*. The human *YY2* located in the X chromosome shows unusual similarity to *YY1* at the amino acid and nucleotide sequence levels and also encodes for a zinc finger protein that recognizes binding motifs similar to those recognized by *YY1*[16]. In this study, we sought to determine the evolutionary origin and conservation of *YY2* using comparative genomic approaches. We have identified YY2 homologues from the genomes of various mammals by database searching and sequencing. Our studies show that *YY2* is placental mammal-specific and is not present in marsupial and nonmammalian vertebrate species. Its intronless genomic structure and the character of surrounding regions suggest that *YY2* is a duplication product from *YY1* that has been generated through retroposition. Compared to *YY1*, *YY2* shows different expression patterns and also appears to have evolved in a very unusual pace in the mammalian genomes.

Results

YY2 is a retroposed copy of YY1 in placental mammals

We analyzed in detail the deposited cDNA sequence of human *YY2* (GenBank Accession No. AK091850) and its genomic locus to determine the genomic structure of *YY2*. Alignment of the YY2 cDNA with the human genome sequences indicated that they are in co-linearity without any interruption (Fig. 1A). This intronless structure is different from the exon structure of the available vertebrates' *YY1* sequences: the similar coding region of *YY1* is divided into five exons. Despite the sequence similarity between YY1- and YY2-coding regions, the immediate surrounding genomic regions of *YY2* lack any sequence similarity to those of *YY1*, suggesting an unusual duplication mode that has generated these two similar genes. Further analyses of the 50-kb genomic region flanking human *YY2* indicated that this genomic interval contains another gene named *Mbtps2* (membrane-bound transcription factor protease site 2). *Mbtps2* is composed of 11 exons distributed over the entire 50-kb genomic interval (Fig. 1A) and *YY2* turns out to be located in the middle of Mbtps2 intron 5. Therefore, this locus bears an unusual "gene-within-another-gene" structure.

To investigate the origin of this unusual genomic structure of *Mbtps2/YY2*, we first searched all of the available genomes with the sequence of human *Mbtps2* (GenBank Accession No. NM_015884). We successfully identified orthologous *Mbtps2*

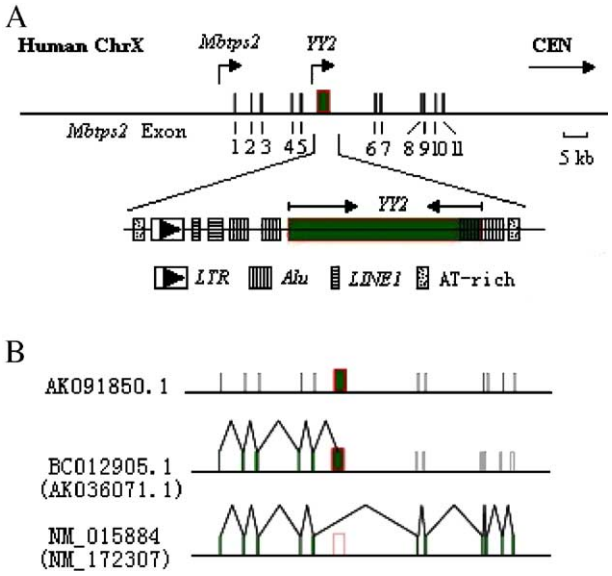

Fig. 1. (A) Schematic representation of the structure of *YY2* and its surrounding region on human chromosome X. (B) Three different forms of transcripts derived from the *Mbtps2*/*YY2* locus in human and mouse. The mouse transcripts are shown within parentheses. The green bars with red borders represent *YY2* and the green bar without a red border represents Mbtps2 exons.

sequences from sequenced vertebrate genomes, including fish, frog, chicken, marsupial, and several placental mammals with available genomic sequences. The identified Mbtps2 sequences show high levels of conservation among the vertebrates in terms of exon structure as well as coding sequences (Fig. 2). The exon structures of the identified *Mbtps2* were further examined to confirm the presence of *YY2* in the introns. Among the vertebrate sequences we examined, only the placental mammals have a YY2-coding sequence in the fifth intron (Fig. 2). The single marsupial mammal, opossum, and other vertebrates do not have YY2 sequences in either the introns or the flanking regions of *Mbtps2*. Since only placental mammals harbor the *YY2* gene in the *Mbtps2* genomic locus, *YY2* is most likely not part of the ancestral *Mbtps2*. Instead, we surmise that *YY2* was inserted into the *Mbtps2* locus after the divergence of placental mammals from the other vertebrates (Fig. 2). Furthermore, the intronless structure of *YY2* and no significant sequence similarity between *YY1* and *YY2* beyond the coding region suggest that *YY2* was duplicated from *YY1* through an RNA-mediated retroposition event.

We then searched all the available cDNA sequences derived from the 50-kb Mbtps2/YY2 locus in placental mammals. Three different forms of transcripts were detected to arise from this locus (Fig. 1B). The first form represented by mRNA sequence AK091850.1 corresponds to the intronless YY2 structure, containing an open reading frame (ORF) with the potential to encode a zinc finger protein 371 amino acids in length. The second form represented by BC012905.1 is a fused transcript consisting of the first five exons of *Mbtps2* and the YY2-coding region. In the second form, the joining of the fifth exon and YY2-coding exon occurs at the sixth amino acid of the open reading frame of the first transcript form (AK091850.1), indicating that the second form of YY2 transcripts may utilize an alternative start codon located in

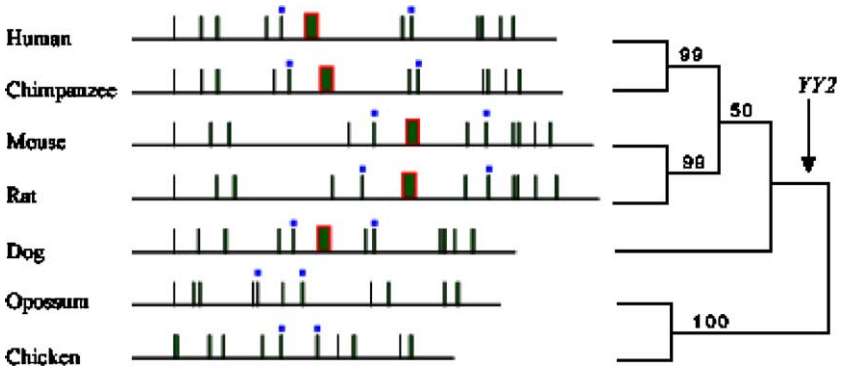

Fig. 2. Genomic structures of *Mbtps2* and *YY2* derived from seven vertebrates. The phylogenetic distance tree is based on Mbtps2 protein sequences using the neighbor-joining method. The bootstrap values derived from 1000 replicates are indicated above each branch. The presence of *YY2* in the fifth intron of *Mbtps2* is detected in all the placental mammals, but not in other vertebrates, including opossum, chicken, frog, and fish. An arrow indicates the estimated evolutionary time point of *YY2* insertion into the *Mbtps2* locus. The green bars with red borders represent *YY2*, whereas the green bars without red borders represent Mbtps2 exons. The evolutionarily conserved Mbtps2 exons flanking *YY2* are marked by blue squares.

one of the five upstream exons of *Mbtps2*. In fact, the start codon of *Mbtps2* is in-frame with the zinc finger exon of *YY2*, making a potential 587-amino-acid ORF with a fusion protein structure of *Mbtps2* and *YY2*. The functional significance of this predicted protein needs to be confirmed, but it is noteworthy that a previous study did indeed detect two YY2 proteins of different lengths from the human testis sample [16]. The third type of transcript derived from this 50-kb locus is represented by NM_015884. This form splices out the YY2-coding region along with the fifth intron and subsequently generates a 1759-bp transcript encoding a 551-amino-acid Mbtps2 protease without the zinc finger domain of *YY2*. A series of similar searches that focused on the mouse genomic interval also identified three different forms of transcripts isolated from various tissues, indicating the evolutionary conservation of the three different forms of Mbtps2/YY2 transcripts. Overall, the Mbtps2/YY2 locus produces three different forms of transcripts and their transcription starts at two different regions, one located in the fifth intron and the other immediately upstream of the first exon of *Mbtps2*, suggesting that at least two different promoter regions are involved in the transcriptional control of alternative transcripts produced by this 50-kb locus.

Rapid evolution of YY2 proteins

According to our analyses described above, *YY2* is a retroposed copy of *YY1* unique to placental mammals. Yet all the *YY2* sequences identified thus far are transcribed and maintain a full coding ORF, indicating that *YY2* is a functionally active gene despite its unusual duplication mode from *YY1*. To understand potential functional constraints that have shaped *YY2* during mammalian evolution, we performed a series of comparative analyses using seven YY2 and four YY1 sequences derived from eight different mammals, including *Homo Sapiens* (Hs), *Pan troglodytes*(Pt), *Mus musculus*(Mm), *Canis familiaris*(Cf), *Rhesus monkey*(Rm), *Rattus rattus*(Rr), *Rattus norvegicus*(Rn), and *Monodelphis domestica* (Md) (Fig. 3 and Table 1).

The predicted sizes of YY2 protein are similar to one another with the exception of dog *YY2*(GenBank Accession No. XM_548891). Whereas predicted YY2 proteins are 372 amino acids long for human (Hs) and chimpanzee (Pt), and 378 amino acids long for mouse (Mm) and rat (Rn), we predict a 451-amino-acid protein for dog (Cf). The reason for this difference is unknown, but it appears to be due to the expansion of a tandem repeat sequence located within the N-

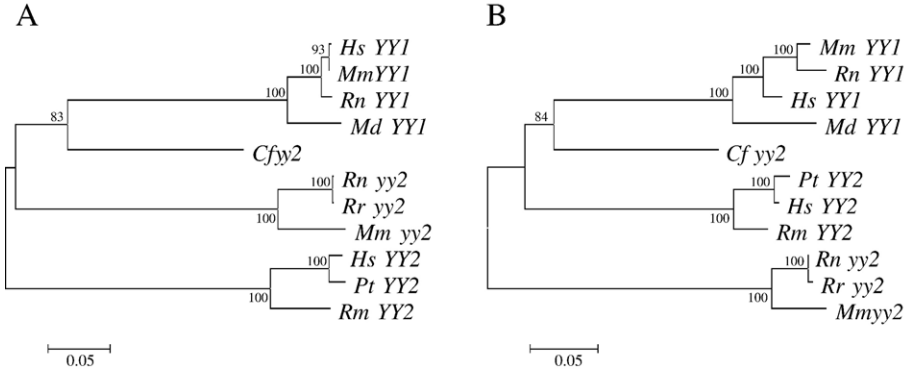

Fig. 3. Gene trees connecting YY1 and YY2 sequences. The trees were constructed with the neighbor-joining method using the Mega3 program. (A) Protein and (B) DNA sequences of *YY1* and *YY2* were used for this analysis. In each case, the bootstrap values calculated from 1000 replicates are indicated above each branch. Different species' *YY1* and *YY2* are indicated with the following abbreviations: Hs, *Homo sapiens*; Pt, *Pan troglodytes*; Rm, *Rhesus monkey*; Mm, *Mus musculus*; Cf, *Canis familiaris*; Rn, *Rattus norvegicus*; Rr, *Rattus rattus*; Md, *Monodelphis domestica*.

Table 1
Protein similarity of *YY1*, *YY2* and *Mbtps2*

	Whole protein		C-terminal		N-terminal		Mbtps2		
	YY1	YY2	YY1	YY2	YY1	YY2	Exon1-11	Exon1-5	Exon6-11
Hs vs. Pt	1	0.975	1	0.990	1	0.960	0.988	0.991	0.986
Mm vs. Rn	0.980	0.871	1	0.990	0.973	0.798	0.984	0.968	0.993
Mm vs. Cf	0.737	0.465	1	0.910	0.658	**0.302**	0.945	0.901	0.976
Hs vs. Mm	0.985	0.527	1	0.881	0.979	**0.291**	0.967	0.946	0.979
Hs vs. Cf	0.740	0.508	1	0.836	0.662	**0.405**	0.967	0.935	0.996
Md vs. Hs	0.921	–	1	–	0.891	–	0.703	0.678	0.718
Md vs. Mm	0.914	–	1	–	0.881	–	0.704	0.683	0.718
Md vs. Cf	0.700	–	1	–	0.610	–	0.694	0.656	0.722

Homo Sapiens(Hs), *Pan troglodytes*(Pt), *Mus musculus*(Mm), *Canis families*(Cf), *Rattus norvegicus*(Rn), *Monodelphis domestica*(Md).

terminal part of the dog *YY2* sequence. This repeat region was excluded from our comparative analyses. Initial comparison of these YY2 protein sequences showed relatively low levels of conservation among placental mammals: 50.8% between Hs and Cf, 46.5% between Mm and Cf, and 52.7% between Hs and Mm (Table 1). In contrast, *YY1* shows much higher levels of conservation, which is more evident in phylogenetic trees generated using YY1 and YY2 sequences: YY1 sequences are clustered together at much closer distances than YY2 in these trees (Fig. 3). More detailed analyses with two separate regions of *YY2*, the N-terminus (1–255) and C-terminus (256–365), revealed that the two regions have very different sequence conservation levels (Table 1 and Fig. 4). The C-terminal region encoding the DNA-binding, zinc finger domain still shows high levels of conservation among placental mammals, averaging 90% sequence identity. However, the N-terminal region has only 30% similarity among different lineages of placental mammals (Table 1). In contrast, the comparison of YY1 protein sequences indicates very high levels of conservation in both the N-terminal and the C-terminal regions among different mammals. In particular, the N-terminal region of *YY1* still shows high levels of conservation among placental mammals as well as among other vertebrates, ranging from 61 to 100%. This is quite different from the conservation levels observed from YY2 protein sequences. The protein sequences of *Mbtps2* also show high levels of sequence conservation among vertebrates, ranging from 69% (Md vs. Cf) to 96% (Hs vs. Cf) (Table 1). This rules out the possibility that the low levels of sequence conservation detected in *YY2* might be related to overall divergence rates at the inserted location. Instead, this analysis indicates that *YY2* has evolved under a selection scheme that is different from that of *YY1*.

Since the N-terminal region of *YY2* appears to have evolved at an unusually fast pace, we performed additional analyses to

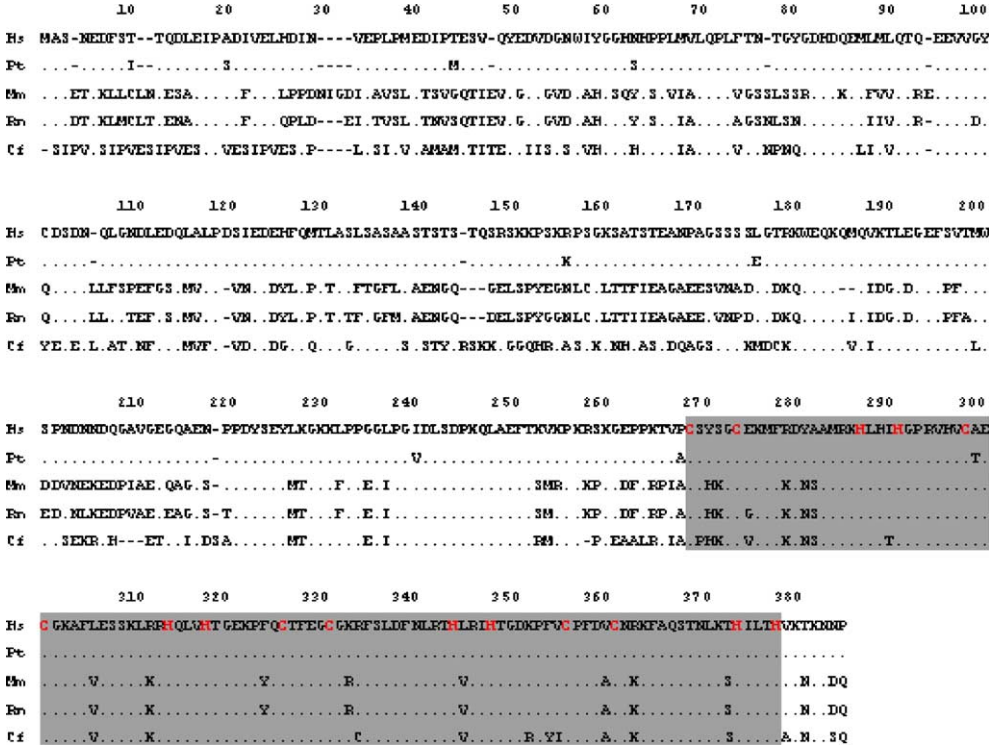

Fig. 4. Alignment of YY2 protein sequences. The amino acid residues identical to human *YY2* are depicted by dots and gaps are indicated by dashes. The conserved zinc finger region is shaded gray and the zinc finger residues, Cys_2His_2, are shown in red. The YY2 sequences of different species are represented by the following abbreviations: Hs, *Homo sapiens*; Pt, *Pan troglodytes*; Mm, *Mus musculus*; Rn, *Rattus norvegicus*; Cf, *Canis familiaris*.

Table 2
d_N and d_S values of two different regions of *YY1* and *YY2*

	C-terminal region						N-terminal region					
	YY1			YY2			YY1			YY2		
	d_N	d_S	d_N/d_S	d_N	d_S	d_N/d_S	d_N	d_S	d_N/d_S	d_N	d_S	d_N/d_S
Hs vs. Pt	0.000	0.000	–	0.000	0.014	**0.000**	0.000	0.000	–	0.015	0.024	**0.625**
Mm vs. Rn	0.000	0.028	**0.000**	0.004	0.208	**0.019**	0.014	0.167	**0.084**	0.074	0.095	**0.747**
Mm vs. Cf	0.000	0.235	**0.000**	0.038	1.311	**0.029**	–	–	–	–	–	–
Hs vs. Mm	0.000	0.238	**0.000**	0.063	1.588	**0.040**	–	–	–	–	–	–
Hs vs. Cf	0.000	0.217	**0.000**	0.082	0.988	**0.083**	–	–	–	–	–	–

Homo Sapiens(Hs), *Pan troglodytes*(Pt), *Mus musculus*(Mm), *Canis families*(Cf), *Rattus norvegicus*(Rn).

determine whether the N-terminal region of *YY2* has evolved under different evolutionary selection pressures (Table 2). The numbers of synonymous (d_S) and nonsynonymous (d_N) nucleotide substitutions per site were calculated using *YY1* and *YY2* sequences of five placental mammals, as summarized in Table 2. The N- and C-terminal regions of *YY1* have evolved under strong negative selection, with the d_N/d_S ratio being almost zero. The C-terminal region of *YY2* has also been under a similar level of negative selection with the d_N/d_S ratio ranging from 0.0 to 0.1. The values derived from *YY1* and the C-terminal region of *YY2*, indicating strong negative selection pressure, are consistent with the high levels of sequence conservation observed from the comparison analyses described above (Table 1). However, the N-terminal region of *YY2* shows relatively higher values of the d_N/d_S ratio, ranging from 0.6 to 0.7, indicating that this region has evolved in recent evolutionary times under slightly negative selection. This supports the idea that the selection pressure on the N-terminal region of *YY2* has been very minimal compared to the N-terminal region of *YY1*.

Comparison of spatial expression patterns of YY1, YY2, and Mbtps2

Since *YY2*was duplicated from *YY1* through retroposition, a process that does not duplicate regulatory regions for transcription, it is likely that *YY2* is subject to transcriptional control that is different from that of *YY1*. Furthermore, since *YY2* is located inside the *Mbtps2* locus, it is possible that the *YY2* gene is influenced by transcriptional regulators controlling expression of the host gene. To compare the expression patterns of *YY2* with those of *YY1* and *Mbtps2*, we conducted a series of *in situ* RNA hybridization experiments using sectioned adult mouse tissues (Fig. 5). Two unique regions of *YY2* and *Mbtps2* were selected and used for preparing *in situ* RNA probes to differentiate the expression patterns of *YY2* from *Mbtps2*.

In the nervous system, *YY1* and *Mbtps2* are highly expressed in both neuronal and glial cells of the cerebral cortex, whereas very low expression levels of *YY2* were detected in these two types of cells. In the cerebellum, the expression of all three genes was detected in Purkinje cells, but only *YY2*and *Mbtps2* were detected in the granular layers of cerebellum. In reproductive organs, all three genes are highly expressed in all layers of spermatocytes, but the expression of *YY2* was not detected in sperm cells. The expression of *YY1* was observed in ovary follicles, but the expression of *YY2* and *Mbtps2* was not detected in this tissue. The expression of all three genes was similarly observed in the epithelial cells of the uterus as well as in intestine (data not shown). The overall expression levels of *YY1* are much higher than those of *YY2* and *Mbtps2* in all the tissues examined, except for adult testis, where all genes are highly expressed. In terms of spatial expression patterns, *YY2* is similar to *Mbtps2*, but these two genes also show some distinctive differences. In particular, *YY2* is not expressed in sperm cells, whereas *Mbtps2* is highly expressed. The overall similarity in expression between *YY2* and *Mbtps2* suggests that *YY2* may be subject to similar transcriptional controls as *Mbtps2* is subject to, consistent with the possibility that one of the two transcripts involving a YY2-coding exon shares a promoter with *Mbtps2* (Fig. 1B).

Discussion

In the current study, we have analyzed the evolutionary origin and conservation of *YY2* using comparative genomic approaches. According to our results, *YY2* is a retroposed sequence derived from an evolutionarily well-conserved zinc finger gene, *YY1*, and this retroposition event occurred after the divergence of placental mammals from other vertebrates based on the presence of *YY2* in only the placental mammals. The N- and C-terminal regions of *YY2* have evolved under quite different selection pressures. The N-terminal region has evolved at a very fast pace with very limited functional constraints. The spatial expression pattern of *YY2* is similar to that of *Mbtps2* but different from that of *YY1*, suggesting that *YY2* and *Mbtps2* share transcriptional control.

Our study indicates that *YY2* was derived from *YY1* by retroposition and yet that *YY2* is conserved among all the placental mammals as an active gene. Our separate searches with YY1 and YY2 sequences against vertebrate genomes independently revealed that each of two published fish genome sequences, pufferfish and zebrafish, contains two copies of the *YY1* gene sequence and also that mammalian genome sequences contain another gene sequence, named *ZFP42* or *Rex-1*, showing sequence similarity to *YY1* (Kim et al., unpublished results). These results suggest that *YY1* has also increased its copy number during vertebrate evolution as seen in other conserved transcription factors, such as *Sp1* and *E2F* families. It is also likely that all of these *YY1* paralogues have been generated through DNA-mediated duplication based on

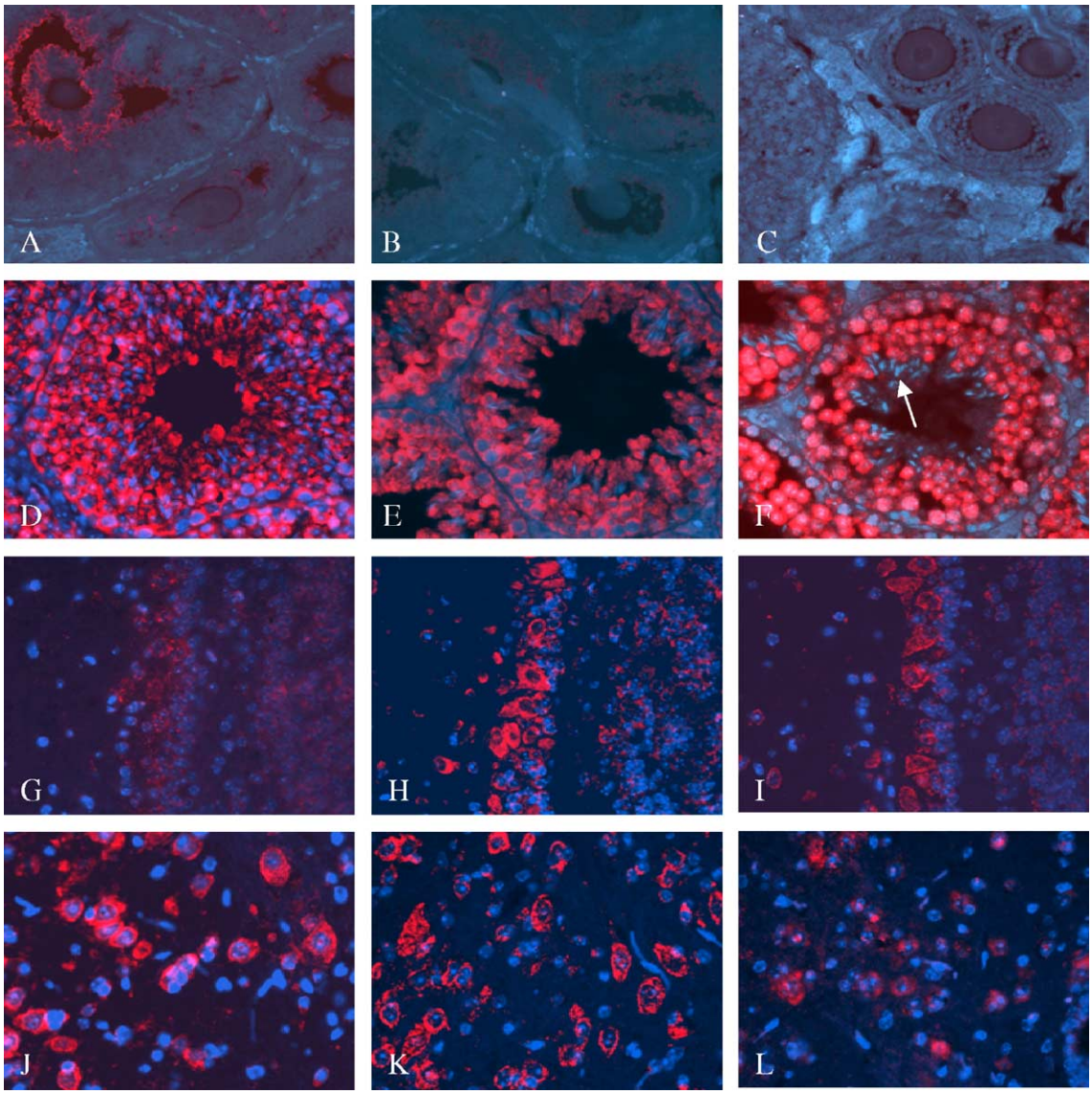

Fig. 5. Spatial expression patterns of *YY1* (A, D, G, J), *Mbtps2* (B, E, H, K), and *YY2* (C, F, I, L). Paraformaldehyde-fixed, sectioned tissues derived from 8-week-old C57BL/6 female and male mice were hybridized by DIG-labeled RNA probes and the signals (red color) were amplified with TSA tetramethylrhodamine. DAPI was used as a counterstain (blue color). In the reproductive organs, the expression of *YY1* is observed in ovary follicles (A), but there is no detectible expression of *YY2* and *Mbtps2* (B, C). In testis, the three genes were all highly expressed in all layers of spermatocytes (D–F), but the expression of *YY2* is not detected in sperm cells (arrow in F). In the nervous system, the expression of all three genes (G–I) is detected in the Purkinje cells of the cerebellum, but only *Mbtps2* (H) and *YY2* (I) were detected in the granular layers of the cerebellum. In the cerebral cortex, *YY1* (J) and *Mbtps2* (K) are highly expressed in both neuronal and glial cells, whereas very low levels of expression of *YY2* (L) were detected in these two types of cells.

the detection of sequence similarity beyond the YY1-coding regions as well as the obvious multiexonic structures observed in these paralogues. Compared to these *YY1* paralogues, *YY2* is thought to have undergone a different evolutionary path due to its unusual retroposition-mediated duplication from *YY1*. This is well reflected in the two different selection pressures imposed on the N- and C-terminal regions of *YY2* and the hybrid exon structure of *YY2* with its host gene, *Mbtps2*. It will be interesting to investigate in the future what different selection schemes have driven the evolution of these *YY1* paralogues in each lineage of vertebrates.

The two regions of YY2 protein have evolved under different selection pressures. The C-terminal region of *YY2* has evolved under strong purifying selection and still shows a sequence structure very similar to that of the C-terminal region of *YY1*. Consistently, the previous study demonstrated that the C-terminal region of *YY*2, encoding four zinc finger units, and *YY1* bind to similar binding motifs. In contrast, the N-terminal region of *YY2* has evolved at a very fast pace with very minimum constraints, which is very different from the N-terminal region of *YY1*, showing high levels of conservation among all the vertebrates. According to previous studies, the N-terminal region of *YY1* can be further divided into several domains based on different functional contributions provided by each domain, including two acidic activation domains, a spacer domain, and other domains responsible for protein-protein interactions [21]. However, because the N-terminal region of *YY2* is so diverged from *YY1* and also

because it differs significantly between species, it is difficult to identify any conserved domain. The divergent sequence structure within the N-terminal region of *YY2* supports the possibility that the functions of YY2 protein in placental mammals should differ from those of *YY1*, mainly based on the difference observed between the N-terminal regions of *YY1* and *YY2*.

According to recent genome-wide surveys, mammalian genomes contain several hundred copies of retroposed sequences and some of these are functional as "retrogenes" [6]. These retrogenes share several unusual features with *YY2*. First, some retrogenes are also located in the introns of another host genes, resulting in a similar hybrid genomic structure, as seen in *Mbtps2/YY2*. These include rodent-specific *Utp14b*, *NUP62*, and *SNAIL-like* [14,18,24]. In particular, *Utp14b* and *SNAIL-like* are transcribed as a fused transcript between host and inserted genes. The expression patterns of these retrogenes are also somewhat similar to those of the host genes. Second, the localization of *YY2* in X chromosomes is consistent with frequent retroposition-mediated gene movements between X chromosomes and autosomes in mammalian genomes. Many retrogenes exported from X to autosomes tend to show male germline-specific expression, whereas many retrogenes recruited from autosomes to X chromosomes show another unusual pattern, the paucity of female-specific tissue expression among these retrogenes. Interestingly, a similar pattern is also observed in *YY2*: no expression of *YY2* in ovary despite the fact that the parental gene, *YY1*, is expressed in both male and female germ cells (Fig. 5). It remains to be tested whether avoiding female tissue expression among X-linked retrogenes is caused by natural selection reducing disadvantageous effects on females [6], but this unusual pattern provides a potential clue regarding the X-chromosomal linkage and subsequent functional impacts of *YY2*on mammalian genomes.

Materials and methods

Database search and gene prediction

A database search was performed using the BLAST program (http://www.ncbi.nlm.nih.gov/BLAST). Gene prediction of various mammals' *Mbtps2* and *YY2* was carried out using the known human or mouse homologous protein sequences as references and further confirmed by EST evidence. The genomic regions containing *Mbtps2* and *YY2* are as follows: *Monodelphis domestica* (AAFR0102815, region from 56.19 to 92.45 kb), *Rattus norvegicus* (NW_048042.1, region from 2.95 to 30.02 kb), *Pan troglodytes* (chromosome X, region from 22.38 to 22.43 Mb, version panTro1), *Canis familiaris* (chromosome X, region from 17.55 to 17.60 Mb, July 2004 assembly of MIT), rhesus monkey (version rheMac1, SCAFFOLD65289, bp 1529–2653 of *YY2*). GenBank accession numbers used for this study are *Mbtps2* of *Mus musculus* (NM_172307), *Mbtps2* of *Homo sapiens* (NM_015884), *YY1* of *Homo sapiens* (NM_003403.3), *YY1* of *Mus musculus* (NM_009537), *YY1* of *R. norvegicus* (NM_173290.1), *YY2* of *H. sapiens* (AY567472 and AK091850.1), and *YY2* of *C.s familiaris*(XM_548891).

Genomic DNA amplification and sequencing

The YY2-coding region of *Rattus rattus* was amplified from genomic DNA using the following two primers: 5′-GGTTTTCGTCACGCTCTCTC-3′ and 5′-CCCAGGCTTCAAAAGGATCT-3′. The PCR was performed in a Bio-Rad iCycler Thermal Cycler under the following conditions: 95°C for 3 min for initiation; 33 cycles of 95°C for 30 s, 63°C for 30 s, 72°C for 30 s; followed by terminal elongation for 7 min at 72°C. The products were subcloned into the Topo TA Cloning system and sequenced with an ABI Prizm 3100 sequencer. Four independent clones were sequenced in both directions and the final sequence has been deposited with GenBank under Accession No. DQ107161.

Sequence alignment, phylogeny, and mutation rate computation

Sequence alignment was carried out with the CLUSTALW program [22] and then manually adjusted using the BioEdit sequence alignment editor (Tom Hall, Department of Microbiology, North Carolina State University, North Carolina, USA; http://www.mbio.ncsu.edu/RNaseP/home.html). Numbers of synonymous (d_S) and nonsynonymous (d_N) nucleotide substitutions per site were estimated by Nei and Gojobori's method [15], modified as recommended by Zhang et al. [26]. The gene trees were constructed by the neighbor-joining method implemented by Mega3 [11,19].

In situ hybridization analysis of YY2, YY1, and Mbtps2

The following three regions of mouse were used for generating *in situ* RNA probes: *YY1* (nt 1358–1794 of GenBank Accession No. NM_009537), *YY2* (nt 2272–2476 of GenBank Accession No. NM_178266), and *Mbtps2* (nt 1039–1208 of GenBank Accession No. NM_172307). Following the published method with minimal modifications [9], sections were dewaxed and rehydrated through three changes of Xylene and two changes of 100, 90, 80, 70% ethanol and water with each washing step for 5 min. Sections were treated with heat using the Target Retrieval Solution (Dako Cytomation S1699) at 95°C for 20 min and cooled to room temperature for another 20 min. Slides were treated with methanol containing 3% hydrogen peroxide for 1 h and then rinsed with PBS. Deproteinzation was performed using proteinase K for 10 min and fixed with fresh 4% paraformaldehye for 10 min. Acetylation was carried out with 100 ml of triethanolamine buffer (pH 8.0) containing 0.25 ml of acetic anhydride for 15 min. The slides were dehydrated through two rounds of a gradient series of 70, 90, 100% ethanol washes and finally air-dried. Each slide was hybridized with 100 μl hybridization solution (Dako Cytomation S3304) containing 1 μg of labeled probes. Hybridization was performed at 42°C inside a humidified chamber overnight. Strain wash (Dako Cytomation S3500) was performed at 45°C for 20 min. The Tyramide Signal Amplification (TSA) system kit (Perkin-Elmer NEL702) was used to amplify signals. DAPI was used as a counterstain.

Acknowledgments

We thank Drs. Mark Batzer and Richard Cordaux for helpful discussions; Jinchuan Xing for drawing phylogenetic trees; Jeong Do Kim for providing technical tips about cloning and sequencing; and Jennifer Thompson for critically reading the manuscript. We also thank an anonymous reviewer for providing many helpful comments. This study was supported by NIH Grant GM66225 (to J.K.) and the U.S. Department of Energy under Contract No. W-7405-ENG-48.

References

[1] L. Aravind, V.M. Dixit, E.V. Koonin, Apoptotic molecular machinery: vastly increased complexity in vertebrates revealed by genome comparisons, Science 291 (2001) 1279–1284.

[2] M. Austen, B. Luscher, J.M. Luscher-Firzlaff, Characterization of the transcriptional regulator YY1. The bipartite transactivation domain is independent of interaction with the TATA box-binding protein, transcription factor IIB, TAFII55, or cAMP-responsive element-binding protein (CPB)-binding protein, J. Biol. Chem. 272 (1997) 1709–1717.

[3] J.L. Brown, D. Mucci, M. Whiteley, M.L. Dirksen, J.A. Kassis, The

Drosophila Polycomb group gene pleiohomeotic encodes a DNA binding protein with homology to the transcription factor YY1, Mol. Cell. 1 (1998) 1057–1064.

[4] C.M. Chiang, R.G. Roeder, Cloning of an intrinsic human TFIID subunit that interacts with multiple transcriptional activators, Science 267 (1995) 531–536.

[5] M.E. Donohoe, X. Zhang, L. McGinnis, J. Biggers, E. Li, Y. Shi, Targeted disruption of mouse Yin Yang 1 transcription factor results in peri-implantation lethality, Mol. Cell. Biol. 19 (1999) 7237–7244.

[6] J.J. Emerson, H.K. Kaessmann, E. Betran, M. Long, Extensive gene traffic on the mammalian X chromosome, Science 303 (2004) 537–540.

[7] R. Friedman, A.L. Hughes, Pattern and timing of gene duplication in animal genomes, Genome Res. 11 (2001) 1842–1847.

[8] G. Hagen, J. Dennig, A. Preiss, M. Beato, G. Suske, Functional analyses of the transcription factor Sp4 reveal properties distinct from Sp1 and Sp3, J. Biol. Chem. 270 (1995) 24989–24994.

[9] J. Kim, V.N. Noskov, X. Lu, A. Bergmann, X. Ren, T. Warth, P. Richardson, N. Kouprina, L. Stubbs, Discovery of a novel, paternally expressed ubiquitin-specific processing protease gene through comparative analysis of an imprinted region of mouse chromosome 7 and human chromosome 19q13.4, Genome Res. 10 (2000) 1138–1147.

[10] C. Kingsley, A. Winoto, Cloning of GT box-binding proteins: a novel Sp1 multigene family regulating T-cell receptor gene expression, Mol. Cell. Biol. 12 (1992) 4251–4261.

[11] S. Kumar, K. Tamura, M. Nei, MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment, Brief Bioinform. 5 (2004) 150–163.

[12] J.S. Lee, K.M. Galvin, Y. Shi, Evidence for Physical Interaction Between the Zinc-Finger Transcription Factors YY1 and Sp1, Proc. Natl. Acad. Sci. USA 90 (1993) 6145–6614.

[13] J.S. Lee, K.M. Galvin, R.H. See, R. Eckner, D. Livingston, E. Moran, Y. Shi, Relief of YY1 transcriptional repression by adenovirus E1A is mediated by E1A-associated protein p300, Genes Dev. 9 (1995) 1188–1198.

[14] A. Locascio, S. Vega, C.A. de Frutos, M. Manzanares, M.A. Nieto, Biological potential of a functional human SNAIL retrogene, J. Biol. Chem. 277 (2002) 38803–38809.

[15] M. Nei, T. Gojobori, Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions, Mol. Biol. Evol. 3 (1986) 418–426.

[16] N. Nguyen, X. Zhang, N. Olashaw, E. Seto, Molecular cloning and functional characterization of the transcription factor YY2, J. Biol. Chem. 279 (2004) 25927–25934.

[17] N. Rezai-Zadeh, X. Zhang, F. Namour, G. Fejer, Y.D. Wen, Y.L. Yao, I. Gyory, K. Wright, E. Seto, Targeted recruitment of a histone H4-specific methyltransferase by the transcription factor YY1, Genes Dev. 17 (2003) 1019–1029.

[18] R. Rohozinski, C.E. Bishop, The mouse juvenile spermatogonial deletion (jsd) phenotype is due to a mutation in the X-derived retrogene, mUtp14b, Proc. Natl. Acad. Sci. USA 101 (2004) 11695–11700.

[19] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Mol. Biol. Evol. 4 (1987) 406–425.

[20] E. Seto, B. Lewis, T. Shenk, Interaction between transcription factors Sp1 and YY1, Nature 365 (1993) 462–464.

[21] M.J. Thomas, E. Seto, Unlocking the mechanisms of transcription factor YY1: are chromatin modifying enzymes the key? Gene 236 (1999) 197–208.

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

[23] A. Usheva, T. Shenk, TATA-binding protein-independent initiation: YY1, TFIIB, and RNA polymerase II direct basal transcription on supercoiled template DNA, Cell 76 (1994) 1115–1121.

[24] S. Wiemann, A. Kolb-Kokocinski, A. Poustka, Alternative pre-mRNA processing regulates cell-type specific expression of the IL411 and NUP62 genes, BMC Biol. 3 (2005) 16–27.

[25] W.M. Yang, C. Inouye, Y. Zeng, D. Bearss, E. Seto, Transcriptional repression by YY1 is mediated by interaction with a mammalian homolog of the yeast global regulator RPD3, Proc. Natl. Acad. Sci. USA 93 (1996) 12845–12850.

[26] J. Zhang, H.F. Rosenberg, M. Nei, Positive Darwinian selection after gene duplication in primate ribonuclease genes, Proc. Natl. Acad. Sci. USA 95 (1998) 3708–3713.