



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

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

C. Luo, X. Lu, L. Stubbs, J. Kim

February 9, 2006

Genomics

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

Editorial Manager(tm) for Genomics
Manuscript Draft

Manuscript Number: GENO-D-05-00188R1

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

Article Type: Regular Article

Section/Category:

Keywords: retroposition, evolution, zinc finger transcription factor

Corresponding Author: Dr. Joomyeong Kim, PhD

Corresponding Author's Institution: Louisiana State University

First Author: Chunqing Luo, PhD

Order of Authors: Chunqing Luo, PhD; Xiaochen Lu, PhD; Lisa Stubbs, PhD; Joomyeong Kim, PhD

Manuscript Region of Origin:

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-terminal 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 very similar sequence structure as 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 neuronal cells of cerebrum, but at very high levels in testis. The expression levels of YY2 is much lower than 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.

Louisiana State University

202 Life Sciences Building
Baton Rouge, LA 70803

Tel: (225) 578-7692
Fax: (225) 578-2597
E-mail: jkim@lsu.edu



Department of Biological Sciences

November 10, 2005

Dear Editor,

We are pleased to hear that our manuscript entitled “Rapid evolution of a recently retroposed transcription factor YY2 in mammalian genomes” can be published in your journal *Genomics*. We appreciate the thorough comments from a reviewer and the editor, which have been very helpful for correcting mistakes and improving our manuscript. The followings are the changes and responses for the various comments we have received.

Regarding comment#1: providing and commenting on the results of YY1-paralogs searches in non-eutherians. We have repeated a series of searches again using all the available genome sequences, including two fish and frog, and chicken genomes, and we did find evidence suggesting that two fish genome sequences, puffer and zebra fish, have the second copy of YY1. The second copies of these fish appear to have been duplicated through DNA-mediated mechanisms based on their multi-exon structures. Therefore, our speculative conclusion that YY1 might have been duplicated only once in vertebrates turns out to be invalid. Therefore, we changed the first paragraph of the Discussion section from this speculative conclusion to the paragraph summarizing potential YY1 paralogs in vertebrates.

Regarding comment#2: commenting relationship YY1 and ZFP42. We agree with the reviewer that ZFP42 has unusual sequence similarity with YY1 and also that ZFP42 might be another paralog of YY1. This has been mentioned in the first paragraph of the Discussion section.

Regarding comment#3: improving the poor quality of Figure 5 showing the spatial expression patterns of YY1, YY2, and Mbtps2. We have performed very thorough *in situ* hybridization for this study, and in fact the images we have used for this manuscript are part of an entire set of images examining the spatial expression of these genes. Therefore, we are confident about our conclusion: the spatial expression patterns of YY2 are similar to those of Mbtps2. Also, one alternative form of YY2 transcripts shares 5'-exons with Mbtps2, suggesting that some of YY2 transcripts should be regulated in a similar manner as Mbtps2. We have printed out the PDF file of our manuscript from the *Genomics* website, and we realized that the quality of the figure image is very poor. We believe that converting our original image to a PDF file may have resulted in this poor-quality picture. To

respond this criticism, we are sending the original file in the powerpoint file format along with the revised version of our manuscript.

Regarding comment#4: TGIFX (which should be TGIFLX) is located in the intron of another gene. This is an incorrect statement. The proper name for this gene is also TGIFLX as pointed out by the reviewer. The incorrect statement was made based on our over-interpretation of the NCBI mapviewer annotation and thus we are not certain at moments whether this retroposed gene is really located in the intron of another gene. Therefore, we have removed that statement and also corrected the misspelling of the gene name.

Regarding comment#5 and #6: providing and discussing more examples of other retrogenes and discussing the X-chromosomal linkage of YY2 in light of the recent studies by Emerson et al., (Science. 2004 303:537-540). These two comments have been very educational and informative in discussing the unusual location of YY2 in X chromosomes. To incorporate these comments, we have rewritten the final paragraph of the Discussion section. We have mentioned briefly more retrogenes and also added sentences discussing potential evolutionary constraints that might have been a contributing factor to recruiting retroposed copies to X chromosomes.

Regarding comment#7: shortening introduction and discussion sections. We have shortened the first paragraph of the Introduction section as suggested by the reviewer.

Regarding comment#8: contrasting YY2 with SNAIL-like and TGIFLX. This has been discussed along with other retrogenes in the newly added, final paragraph of the Discussion section.

Regarding comment#9: the phrase “adapted into new functions” in our abstract is speculative without any firm evidence. We accept this criticism, and thus removed that phrase in the abstract.

Overall, we have rewritten two paragraphs of the Discussion section 1) to correct our speculative conclusion regarding YY1 duplication in vertebrates and 2) to provide more updated, insightful discussion regarding the X-chromosomal linkage of YY2. These two paragraphs are marked by red color. We have also marked as red color some sections that have been revised. We are sending a set of images corresponding to Fig. 5 as a separate file to the editor. If this set is still unsatisfactory, please let us know. We would be happy to provide more images.

If you have any further concerns, please contact me at the above correspondence.

Thank you for your help.

Sincerely,

Joomyeong Kim, Ph.D.
Associate Professor

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

¹Chunqing Luo, ²Xiaochen Lu, ²Lisa Stubbs, ^{1,3}Joomyeong Kim

¹Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, 70803.

²Genome Biology Division, Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94551.

³Correspondence should be forward to:
jkim@lsu.edu, 225-578-7692(ph), or 225-578-2597(fax)

Running title: Rapid evolution of retroposed gene *YY2*

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-terminal 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 very similar sequence structure as *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 neuronal cells of cerebrum, but at very high levels in testis. The expression levels of *YY2* is much lower than *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.**

Key words: 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 upon the sequence context of *YY1*-binding sites with respect to other regulator elements [Thomas and Seto, 1999]. 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 [Seto et al., 1993; Lee et al., 1993; Chiang et al., 1995; Usheva and Shank, 1994; Austen et al., 1997]. Other studies also indicated that *YY1* recruits histone-modifying enzymes including p300, HDACs and PRMT1 for transcription control [Lee et al., 1995; Yang et al., 1996; Rezai-Zadeh, 2003]. Physiological roles of *YY1* have been demonstrated in mouse by gene knockout experiments, in which homozygous mutant mice show peri-implantation lethality and subset of heterozygous mice show developmental abnormalities, such as exencephaly (or open brain) [Donohoe et al., 1999].

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 [Brown et al., 1998]. Many key transcription factors, including *Sp1* and *E2F*, have evolutionary histories similar to that of *YY1*. These transcription factors are conserved throughout the vertebrate lineage as well as 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 [Kingsley et al., 1992; Hagen et al., 1995; Aravind et al., 2001]. **Genome-wide and segmental duplications, DNA-mediated, are thought to be responsible for this increase of gene number in vertebrates [Friedman and Hughes, 2001]. Occasional retropositions, RNA-mediated, have also contributed to the increase of gene number in vertebrates [Emerson et al., 2004].**

Consistently, another gene sequence with significant similarity to *YY1* has recently been identified in the human genome, and thus 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 similar binding motifs as *YY1* [Nguyen et al., 2004]. 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 non-mammalian 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. As compared to *YY1*, *YY2* shows different expression patterns and also appears to have evolved in a very unusual pace in the mammalian genomes.

Results

1. *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 with 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 comprised of eleven 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* 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 *YY2*-coding sequence in the 5th intron (Fig.2). The single marsupial mammal, opossum, as well as other vertebrates do not have *YY2* sequences in

either the introns or 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 (Fig2). 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 potential to encode a zinc finger protein 371 amino acid in length. The second form represented by BC012905.1 is a fused transcript consisting of the first five exons of *Mbtps2* and *YY2*-coding region. In the second form, the joining of the 5th exon and *YY2*-coding exon occurs at the 6th 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 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 long 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 length from human testis sample [Nguyen et al., 2004]. The third type of transcripts derived from this 50-kb locus is represented by NM_015884. This form splices out the *YY2*-coding region along with the 5th intron and subsequently generates a 1759-bp transcript encoding a 551 amino acid long *Mbtps2* protease without the zinc finger domain of *YY2*. A series of similar searches 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 5th 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.

2. 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 so far are transcribed and maintain a full coding ORF, indicating that *YY2* is a functionally active gene despite its unusual duplication mode from *YY1*. In order 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 families* (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 each other with the exception of dog *YY2* (GenBank accession No. XM_548891). Whereas predicted *YY2* proteins are 372 amino acid long for human (Hs) and chimp (Pt), and 378 amino acid long for mouse (Mm) and rat (Rn), we predict a 451 amino acid protein for dog (Cf). The reason for this difference is currently unknown, but it appears to be due to the expansion of a tandem repeat sequence located within the N-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% Mm vs Cf, and 52.7% between Hs and Mm (Table1). 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 in much closer distances than *YY2* in these trees (Fig3). More detailed analyses with two separate regions of *YY2*, the N-terminal (1-255) and C-terminal (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 lineage of placental mammals (Table 1). In contrast, comparison of *YY1* protein sequences indicates very high levels of conservation in both the N-terminal and 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 ask if 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-terminal 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 as compared to the N-terminal region of *YY1*.

3. 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*. Further, 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* 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 all 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 expression of *YY2* and *Mbtps2* was not detected in this tissue.

The expression of all three genes was similarly observed in the epithelium 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 expression similarity between *YY2* and *Mbtps2* suggests that *YY2* may be subject to similar transcriptional controls as *Mbtps2*, consistent with the possibility that one of the two transcripts involving *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 has occurred after the divergence of placental mammals from other vertebrates based on the presence of *YY2* only in the placental mammals. The N-terminal 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 *YY1*, suggesting that *YY2* and *Mbtps2* share transcriptional control.

Our study indicates that *YY2* has been 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, puffer and zebra fish, contains two copies of *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). 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* paralogs have been generated through DNA-mediated duplication based on the detection of sequence similarity beyond the *YY1*-coding regions as well as the obvious multi-exonic structures observed in these paralogs. As compared to these *YY1* paralogs, *YY2* is thought to have undergone a different evolutionary path due to its unusual retroposition-mediated duplication from *YY1*. This is well reflected on the two different selection pressures

imposed on the N-terminal 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* paralogs 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 very similar sequence structure as the C-terminal region of *YY1*. Consistently, the previous study demonstrated that the C-terminal region of *YY2* encoding four zinc finger units binds to similar binding motifs as *YY1*. 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 [Thomas and Seto, 1999]. However, the N-terminal region of *YY2* is so diverged from *YY1* and also differs significantly between species that it is difficult to identify any conserved domain. The divergent sequence structure within the N-terminal region of *YY2* supports a possibility that the functions of *YY2* protein in placental mammals should differ from *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’ [Emerson et al., 2004]. These retrogenes share several unusual features with *YY2*. First, some of 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 *mUtp14b*, *NUP62*, and *SNAIL-like* (Rohozinski and Bishop, 2004; Wiemann et al., 2005; Locascio et al., 2002). In particular, *mUtp14b* 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 decreasing disadvantageous effects on females [Emerson et al., 2004], 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.

Database search was performed using the BLAST program (<http://www.ncbi.nlm.nih.gov/BLAST>). The 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, from 56.19 Kb to 92.45 Kb region), *Rattus norvegicus* (NW_048042.1, from 2.95 Kb to 30.02 Kb region), *Pan troglodytes* (ChrX, from 22.38 Mb to 22.43 Mb region, version panTro1), *Canis familiaris* (ChrX, from 17.55 Mb to 17.60 Mb region, Jul. 2004 assembly of MIT), *Rhesus monkey* (version rheMac1, SCAFFOLD65289, 1529 bp – 2653 bp 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 *Rattus norvegicus* (NM_173290.1), *YY2* of *Homo sapiens* (AY567472 and AK091850.1), and *YY2* of *Canis 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'-GGTTTTTCGTCACGCTCTCTC-3' and 5'-CCCAGGCTTCAAAGGATCT-3'. The PCR reaction was performed in a Bio-Rad iCycler Thermal Cycler under the following conditions: 95°C, 3 min for initiation; 33 cycles of 95°C, 30 sec, 63°C, 30 sec, 72°C, 30 sec; and 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 as GenBank accession No. DQ107161.

Sequence alignment, phylogeny and mutation rate computation.

Sequence alignment was carried out with the CLUSTALW program [Thompson et al., 1994], 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 the Nei and Gojobori's method [Nei and Gojobori, 1986], modified as recommended by Zhang et al [Zhang et al., 1998]. The gene trees were constructed by the neighbor-joining method implemented by Mega3 [Saitou and Nei, 1987; Kumar et al., 2004].

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 minimum modifications [Kim et al., 2000], sections were dewaxed and rehydrated through 3 changes of Xylene, and 2 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 (DakoCytomation S1699) at 95°C for 20 min and cooled down to room temperature for another 20 min. Slides were treated with methanol containing 3% hydrogen peroxide for 1 hr and then rinsed by PBS. Deproteinization was done by proteinase K for 10 min and fixed with fresh 4% paraformaldehyde for 10 min. Acetylation was carried out with 100 ml of triethanolamine buffer (pH8.0) 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 μ l hybridization solution (DakoCytomation S3304) containing 1 μ g of labeled probes. Hybridization was performed at 42°C inside a humidified chamber for overnight. Strain wash (DakoCytomation S3500) was performed at 45°C for 20 minutes. The Tyramide Signal Amplification (TSA) system kit (PerkinElmer NEL702) was used to amplify signals. DAPI was used as a counter stain.

Acknowledgements

We thank Drs. Mark Batzer and Richard Cordaux for helpful discussions; Jinchuan Xing for drawing phylogenetic trees; Jeong Do Kim for providing technical tips for cloning and sequencing; and Jennifer Thompson for critical reading of 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 with University of California, Lawrence Livermore National Laboratory.

References

- [1] Aravind, L, VM Dixit, and EV Koonin. 2001. Apoptotic molecular machinery: Vastly increased complexity in vertebrates revealed by genome comparisons. *Science* **291**: 1279-1284
- [2] Austen, M, B Luscher and JM Luscher-Firzlaff. 1997. 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**: 1709-1717.
- [3] Brown, JL, D Mucci, M Whiteley, ML Dirksen and JA Kassis. 1998. The Drosophila Polycomb group gene pleiohomeotic encodes a DNA binding protein with homology to the transcription factor YY1. *Mol Cell.* **1**: 1057-1064.
- [4] Chiang, CM. and RG Roeder. 1995. Cloning of an intrinsic human TFIID subunit that interacts with multiple transcriptional activators. *Science.* **267**: 531-536.
- [5] Carroll SB, JK Grenier, and SD Weatherbee. 2001. From DNA to diversity-Molecular genetics and the evolution of animal design. Blackwell Science, Malden, Massachusetts.
- [6] Donohoe, ME, X Zhang, L McGinnis, J Biggers, E Li and Y Shi. 1999. Targeted disruption of mouse Yin Yang 1 transcription factor results in peri-implantation lethality. *Mol Cell Biol.* **19**: 7237-7244.
- [7] Emerson, JJ, HK Kaessmann, E Betran, and M Long. 2004. Extensive gene traffic on the mammalian X chromosome. *Science.* **303**:537-540.
- [8] Friedman, R and AL Hughes. 2001. Pattern and timing of gene duplication in animal genomes. *Genome Res.* **11**: 1842-1847.
- [9] Hagen, G, J Dennig, A Preiss, M Beato and G Suske. 1995. Functional analyses of the transcription factor Sp4 reveal properties distinct from Sp1 and Sp3. *J Biol Chem.* **270**: 24989-24994.
- [10] Kim, J, VN Noskov, X Lu, A Bergmann, X Ren, T Warth, P Richardson, N Kouprina and L Stubbs. 2000. 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**: 1138-1147.

- [11] Kingsley, C and A Winoto. 1992. Cloning of GT box-binding proteins: a novel Sp1 multigene family regulating T-cell receptor gene expression. *Mol Cell Biol.* **12**: 4251-4261.
- [12] Kumar, S, K Tamura and M Nei. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform.* **5**: 150-163.
- [13] Lee, JS, KM Galvin, and Y Shi. 1993. Evidence for Physical Interaction Between the Zinc-Finger Transcription Factors YY1 and Sp1. *Proc Natl Acad Sci U S A.* **90**: 6145-614.
- [14] Lee, JS, KM Galvin, RH See, R Eckner, D Livingston, E Moran, and Y Shi. 1995. Relief of YY1 transcriptional repression by adenovirus E1A is mediated by E1A-associated protein p300. *Genes Dev.* **9**: 1188-1198.
- [15] Locascio, A, S Vega, CA de Frutos, M Manzanares and MA Nieto. 2002. Biological potential of a functional human SNAIL retrogene. *J Biol Chem.* **277**: 38803-38809.
- [16] Miklos, GL and GM Rubin. 1996. The role of the genome project in determining gene function: insights from model organisms. *Cell.* **86**: 521-529.
- [17] Nei, M and T Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol.* **3**: 418-426.
- [18] Nguyen, N, X Zhang, N Olashaw and E Seto. 2004. Molecular cloning and functional characterization of the transcription factor YY2. *J Biol Chem.* **279**: 25927-25934.
- [19] Rezai-Zadeh, N, X Zhang, F Namour, G Fejer, YD Wen, YL Yao, I Gyory, K Wright, and E Seto. 2003. Targeted recruitment of a histone H4-specific methyltransferase by the transcription factor YY1. *Genes Dev.* **17**: 1019-1029.
- [20] Rohozinski, R and CE Bishop. 2004. The mouse juvenile spermatogonial deletion (jsd) phenotype is due to a mutation in the X-derived retrogene, mUtp14b. *Proc. Natl. Acad. Sci. USA.* **101**:11695-11700.
- [21] Saitou, N and M Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* **4**: 406-425.
- [22] Seto, E, B Lewis, and T Shenk. 1993. Interaction between transcription factors Sp1 and YY1. *Nature.* **365**: 462-464.
- [23] Thomas, MJ and E Seto. 1999. Unlocking the mechanisms of transcription factor YY1: are chromatin modifying enzymes the key? *Gene.* **236**: 197-208.
- [24] Thompson, JD, DG Higgins and TJ Gibson. 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**: 4673-4680.
- [25] Usheva, A, and T Shenk. 1994. TATA-binding protein-independent initiation: YY1, TFIIB, and RNA polymerase II direct basal transcription on supercoiled template DNA. *Cell*

76: 1115-1121.

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

[27] Yang, WM, C Inouye, Y Zeng, D Bearss, and E Seto. 1996. Transcriptional repression by YY1 is mediated by interaction with a mammalian homolog of the yeast global regulator RPD3. *Proc Natl Acad Sci U S A.* **93**: 12845-12850.

[28] Zhang, J, HF Rosenberg and M Nei. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci U S A.* **95**: 3708-3713.

Figure legends

Figure 1. **A)** Schematic representation of *YY2* and its surrounding region structure on the 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 bar with red border represents *YY2* and the green bar without red border represents *Mbtps2* exons.

Figure 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 on the top of each branch. The presence of *YY2* in the 5th 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 border represent *YY2* whereas the green bars without red border represent *Mbtps2* exons. The evolutionarily conserved *Mbtps2* exons flanking *YY2* are marked by dots on the top.

Figure 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: *Homo Sapiens* (Hs), *Pan troglodytes* (Pt), *Rhesus monkey*

(Rm), *Mus musculus* (Mm), *Canis families* (Cf), *Rattus norvegicus* (Rn), *Rattus rattus* (Rr), and *Monodelphis domestica* (Md).

Figure 4. Alignment of YY2 protein sequences. The amino acid residues identical to the human YY2 are depicted by dots, and gaps by dashes. The conserved zinc finger region is marked by shade and the zinc finger residues, Cys₂His₂, with red color. The YY2 sequences of different species are represented by the following abbreviations: *Homo Sapiens* (Hs), *Pan troglodytes* (Pt), *Mus musculus* (Mm), *Rattus norvegicus* (Rn), and *Canis families* (Cf).

Figure 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 eight-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 counter stain (blue color). In the reproductive organs, the expression of *YY1* is observed in ovary follicles (A), but no detectible expression of *YY2* and *Mbtps2* (B, C). In testis three genes were all highly expressed in all layers of spermatocytes (D, E, 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, H, 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 expression levels of *YY2* (L) were detected in these two types of cells.

Tables

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

	whole protein		C-terminal		N-terminal		<i>Mbps2</i>		
	<i>YY1</i>	<i>YY2</i>	<i>YY1</i>	<i>YY2</i>	<i>YY1</i>	<i>YY2</i>	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)

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

	C-terminal Region						N-terminal Region					
	<i>YY1</i>			<i>YY2</i>			<i>YY1</i>			<i>YY2</i>		
	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)

Figure 1

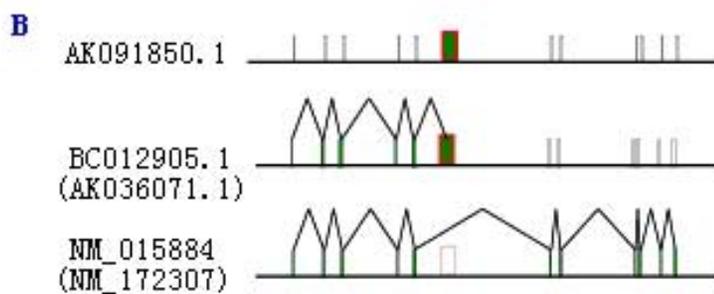
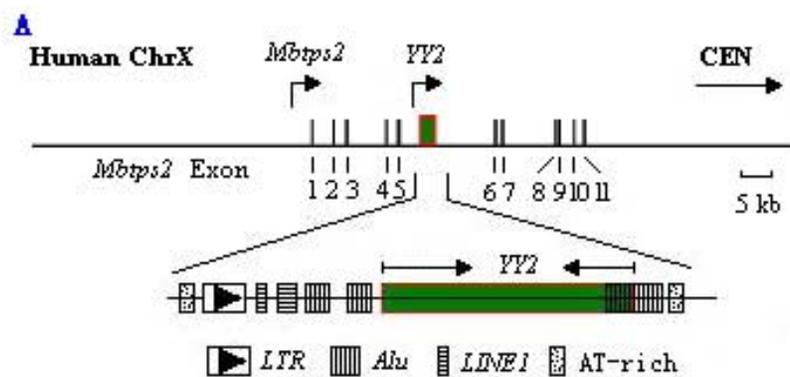


Figure 2

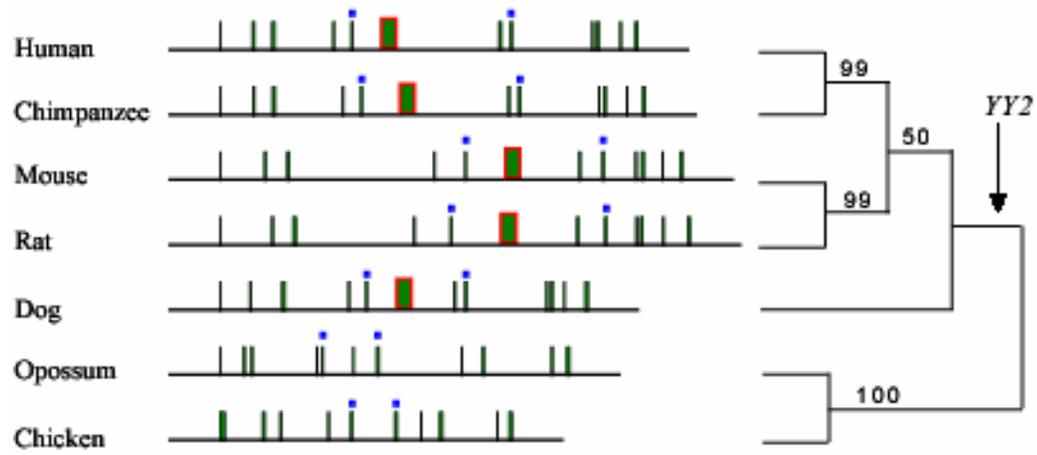
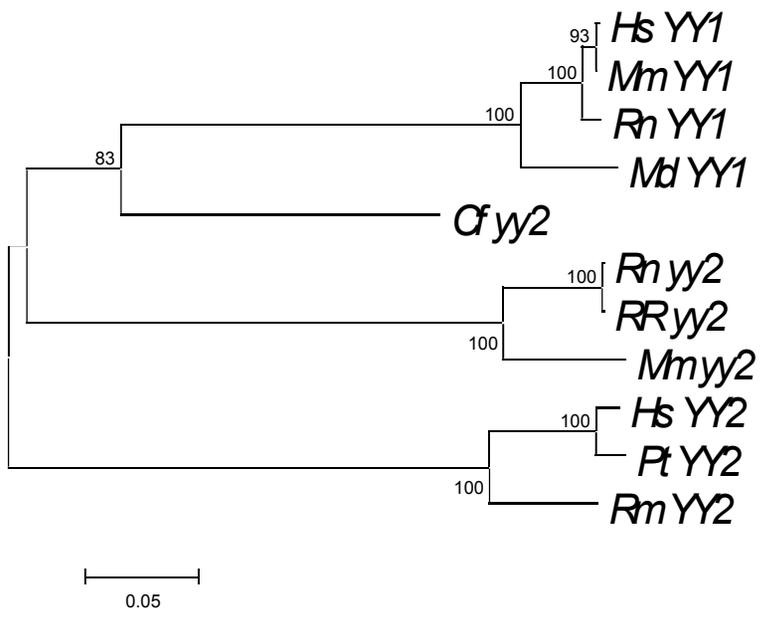


Figure 3

A



B

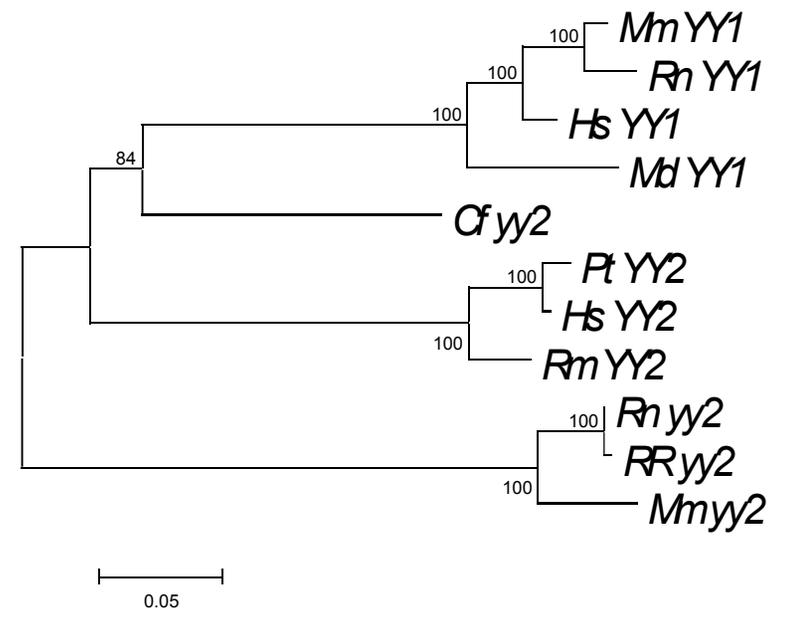


Figure 4

```

      10      20      30      40      50      60      70      80      90     100
Hs MAS-MEDFST--TQDLEIPADIVELHDIN---VEPLPMEDIPTESV-QYEDVDGNQIYGGHMHPPLMVLQPLFTN-TGYGDHDQEMMLLQTQ-EEVWGY
Pt .....I-----S-----M-----S-----
Mm ...ET.KLLCLN.ESA...F...LPPDNIGDI.AVSL.TSUGQTIEV.G..GVD.AH.SQY.S.VIA...VGSLSLR...K.FVV..RE.....
Rn ...DT.KLMCLT.EMA...F...QPLD---EI.TVSL.TNGSQTIEV.G..GVD.AH...Y.S..IA...AGSNLSM.....IIV..R....D.
Cf -SIPQ.SIPQESIPQES..VESIPQES.P---L.SI.V.AMM.TITE..IIS.S.VH...H...IA...V..NPNQ.....LI.V...

      110     120     130     140     150     160     170     180     190     200
Hs CDSIN-QLGNDLEDQLALPDSIEDERFQMTLASLSASAASTSTST-TQSRSKPKPKRPSGKSAATSTEAMPAGSSS ELGTENQEQKQMQQKLEGEFSTVIMQ
Pt .....K.....E.....
Mm Q...LLFSPEFGS.MV...-VM..DYL.P.T..FTGFL.AENGQ---GELSPYEGMLC.LTTFIEAGAEESVNA.D..DKQ.....-IDG.D...PF...
Rn Q...LL..TEF.S.MV...-VM..DYL.P.T.TF.GFM.AENGQ---DELSPYEGMLC.LTTIIEAGAE.VMPD..DKQ....I.IDG.D...PFA...
Cf YE.E.L..AT.MF...MVF..-VD..DG..Q...G...S.STY.RSKK.GGQR.AS.K.NH.AS.DQAGS...KMDCK.....V.I.....L.

      210     220     230     240     250     260     270     280     290     300
Hs SPNDMDQGAUGEGQAEN-PPDQSEYLNKGNLPPGGLPG IDLSDPKQLAETTKVWPKRKSNGEPPKTVPCSYSGC EMMFEDQAAEMKHLHINQPGVHGC AE
Pt .....V.....A.....T.
Mm DDGMEKEDIPIAE.QAG.S-.....MT...F..E.I.....SMR..KP..DF.EPIA..HK.....K.NS.....
Rn ED.NLKEDEPVAE.EAG.S-T.....MT...F..E.I.....SM..KP..DF.EP.A..HK..G..K.NS.....
Cf ..SEKR.H--ET..I.DSA.....MT.....E.I.....RM...-P.EAALR.IA.PHK..V..K.NS.....T.....

      310     320     330     340     350     360     370     380
Hs CGKAFLESKLERHQLAMTGEKPFQCTFEFGCGKRFSLDFMLRHLRHTGDKPFVCPFDGCMKFAQSTMLKHILTHVKTAMMP
Pt .....
Mm ...V...K.....Y.....R.....V.....A.K.....S.....N.DQ
Rn ...V...K.....Y.....R.....V.....A.K.....S.....N.DQ
Cf ...V...K.....C.....V.....R.YI...A.K.....S.....A.N..SQ

```

Figure 5

