Available online at www.sciencedirect.com

SCIENCE DIRECT®

Gene 352 (2005) 82 – 91

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

Coupling and decoupling of evolutionary mode between X- and Y-chromosomal red-green opsin genes in owl monkeys

Kenji Nagao, Naomi Takenaka[1], Momoki Hirai, Shoji Kawamura*

Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, Seimeitou #502, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan

Received 15 December 2004; received in revised form 17 March 2005; accepted 1 April 2005

Received by T. Gojobori

Abstract

We previously discovered Y-chromosomal red-green opsin genes in two types of owl monkeys with different chromosomal characteristics. In one type, the Y-linked opsin gene is a single-copy intact gene and in the other, the genes exist as multiple pseudogenes on a Y/autosome fusion chromosome. In the present study, we first distinguished the two types of monkeys as distinct allopatric species on the basis of karyotypic characteristics: *Aotus lemurinus griseimembra* (Karyotype III, diploid chromosome number [$2n$]=53) and *Aotus azarae boliviensis* (Karyotype VI; male $2n=49$; female $2n=50$), belonging to the northern and southern species groups, respectively, separated by the Amazon River system. Our sequence analysis revealed a common L1-Alu-Alu insertion between the two species in the 3′-flanking region of the X-linked opsin genes. The insertion was absent in the Y-linked opsin genes and in the human red and green opsin genes, indicating that it occurred in the X copy before the split into northern and southern species and after the X to Y duplication, i.e. duplication preceded speciation. We also show that in the northern species, the Y-linked opsin gene has evolved concomitantly with the X-linked copy whereas in the southern species, the Y-autosome fusion possibly led to decoupling evolutionary processes between X- and Y-linked copies and subsequent degeneration and duplications of the Y-linked opsin gene.

Keywords: *Aotus*; New World monkeys; Visual pigments; Gene conversion; Chromosomal fusion

1. Introduction

Photoreceptor cells in the vertebrate retina are distinguished into rods and cones. Rods work for dim light (scotopic) vision while cones work for daylight (photopic) vision. Color vision, an entity of photopic vision, can be achieved only with the presence of at least two spectrally distinct types of cones. Owl monkeys (*Aotus*), a genus of the New World monkeys (platyrrhine primates), are the only nocturnal higher primates (simians). They have features specialized for nocturnal vision. Their orbit size is the largest among simians. This, together with a relative increase in rod density (peak density 325,000/mm^2) compared to that of diurnal simians (~180,000/mm^2), supports their high visual sensitivity and enhanced scotopic acuity (Wikler and Rakic, 1990). In addition, they lack a foveola (a major central peak in cone density in the retina characteristic of diurnal primates). In diurnal simians, peak cone density is ~200,000/mm^2 (Wikler and Rakic, 1990) whereas peak cone density is only ~7000/mm^2 in owl monkeys, comparable to that in bushbaby (~8500/mm^2), a nocturnal prosimian. Owl monkeys have only one type of red-green cone visual pigment maximally sensitive to 539 nm (Hiramatsu et al., 2004); the blue visual pigment has been lost due to disruptive mutations (Jacobs et al., 1996) and the allelic polymorphism of the red-green opsins, a common feature of most New World monkeys, is

Abbreviations: bp, base pair(s); cDNA, complementary DNA; CDR, coding region; kb, kilobase(s); M/LWS, middle-to-long wavelength-sensitive; PAR, pseudoautosomal region.

* Corresponding author. Tel.: +81 4 7136 5422; fax: +81 4 7136 3692.
E-mail address: kawamura@k.u-tokyo.ac.jp (S. Kawamura).
[1] Present address: Department of Biology, Emory University, Atlanta, GA, USA.

0378-1119/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2005.04.008

absent (Jacobs et al., 1993). These features preclude the possibility of color vision in owl monkeys and appear to support only a level of photopic acuity significantly inferior to that of diurnal simians.

Despite these nocturnal features, owl monkeys lack a tapetum lucidum, a reflective retinal structure in back of the photoreceptor layer and characteristic of nocturnal mammals. The rod density, though higher than that in diurnal primates, is relatively low compared to that in the nocturnal bushbaby (~450,000/mm^2) (Wikler and Rakic, 1990). Based on these characteristics, owl monkeys are considered to have a diurnal ancestry (Fleagle, 1999). Although utilizing a variety of auditory and olfactory signals, owl monkeys appear to be highly dependent on vision because their social behaviors (including intra-group calling and playing and inter-group fighting) almost exclusively occur under bright moon light; insect foraging also occurs at dawn, dusk and on moonlit nights (Wright, 1994). Owl monkeys are one of the most successful genera among New World monkeys, broadly distributed across South America from Panama to northern Argentina (Fleagle, 1999). They are adept leapers found in a variety of forest habitats and are primarily frugivorous, supplemented by both foliage and insects.

Phylogenetically, owl monkeys have traditionally been linked with the titi monkeys (*Callicebus*). Molecular studies, however, have classified them in the family Cebidae, in the subfamily Aotinae, together with Callitrichinae (marmosets and tamarins) and Cebinae (capuchins and squirrel monkeys) (Schneider, 2000). Owl monkeys can be divided into northern (gray-necked) and southern (red-necked) groups, widely inhabiting regions north and south of the Amazon River system, respectively. The northern and southern groups are comprised of four and five allopatric species, respectively, and are distinguished by karyotypes, geographic origins and pelage patterns (Hershkovitz, 1983) though details of the intra-group taxonomy are still controversial (Pieczarka et al., 1993). Pre-mating reproductive isolation between the two species groups appears to be established in the natural habitat since hybrids have not been found in river-bend cutoffs where species from the two groups are sympatric, though the two groups can be crossed in captivity with reduced fertility (Pieczarka et al., 1992). The karyotypical diversity is one of the most conspicuous characteristics among owl monkeys, exhibiting many intra- and inter-specific chromosomal variations, both numerical and structural; chromosome diploid numbers ranging from 46 to 58 in 18 karyotypes can be assigned to general karyotypically-defined taxa (Torres et al., 1998). Characteristic of the southern species group, three widely-distributed species (*Aotus nigriceps*, *Aotus azarae*, and *Aotus infulatus*) have been documented to have a Y/autosome fusion chromosome (Pieczarka and Nagamachi, 1988; Ma et al., 1989; Pieczarka et al., 1993).

In addition to the authentic X-chromosomal gene, we have recently identified extra red-green opsin genes on the Y chromosome in captive owl monkeys maintained in an institutional breeding colony (Kawamura et al., 2002). In Kawamura et al. (2002), two genes of red-green opsin were identified in one male monkey (no. 14), only one of which was transmitted to a single daughter. In situ hybridization indicated an XY location of these genes. The Y-linked opsin gene was found to have no structural defect (denoted 14Y). Multiple red-green opsin genes were identified in another male monkey (no. 29); one appeared X-linked and four others appeared Y-linked on the basis of inheritance patterns. These genes were also mapped to the sex chromosomes by in situ hybridization. The four Y-linked genes were pseudogenes (29Ys) and the Y chromosome was found to be fused with an autosome (Kawamura et al., 2002). The two types of monkeys were both classified as a single, widespread polytypic species, *Aotus trivirgatus*, which is now recognized as being comprised of multiple species as described above (Hershkovitz, 1983). Therefore, our previous species identification needs to be reconsidered by more detailed chromosomal characterization.

In the present study, we aimed to determine the species origins of two types of the owl monkeys (types 14 and 29) by re-examining their karyotypes. On the basis of the species relationships, we then sought to determine the translocation origin of the two types of Y-linked opsin genes by extensive DNA sequence analysis in order to gain better insight into the evolution of Y-chromosomal red-green opsin genes.

2. Materials and methods

2.1. Owl monkeys

The two types of owl monkeys (types 14 and 29) were found in a breeding colony in the Primate Research Institute of Kyoto University, Japan, on the basis of genomic organization of the red-green opsin genes in our previous study (Kawamura et al., 2002). Briefly, type 29 has multiple red-green opsin genes (29Y-1, 29Y-2, 29Y-3 and 29Y-4) with premature stop codons on the Y chromosome that is fused to an autosome. Type 14 has a canonical Y chromosome and has an intact red-green opsin gene (14Y). The X-linked opsin genes of type 14 (14X) and type 29 (29X) can be distinguished by their restriction site distribution. Founders of type 29 monkeys have a Bolivian origin; the origin of type 14 has not been recorded. All founder owl monkeys were introduced to the Institute in 1973–1977. Until the work of Hershkovitz (1983), owl monkeys were generally thought to contain a single widespread polytypic species, *A. trivirgatus*, and the founder monkeys were all recorded with this species name.

2.2. Chromosomal typing

Peripheral blood lymphocytes from four male owl monkeys (no. 14 from type 14 and nos. 24, 28 and 44 from type 29; see Kawamura et al., 2002 for their kin

relationships) were cultured in RPMI1640 medium supplemented with fetal bovine serum (15%) and phytohemagglutinin (3%). Both conventional and R-banded chromosome preparations were made as described previously (Hirai et al., 1994).

2.3. Genomic library screening

A genomic library of one type 29 male owl monkey (no. 29) was previously constructed (Kawamura et al., 2002) and re-screened for the red-green opsin gene regions uncloned in the previous study using the full-length 29X cDNA (Hiramatsu et al., 2004) as a probe. The probe was labeled with [α-^{32}P] dCTP using the random primer method. Plaque hybridization was carried out at 65 °C in the solution consisting of 6 × SSC, 5 × Denhardt's solution, 0.5% SDS and 5 μg/ml *E. coli* DNA. The hybridized membranes were washed in 1 × SSC/0.1% SDS at 65 °C four times (20 min each), which allows approximately 20% mismatch. The plasmid subclones were sequenced in both strands by using LI-COR 4200L-1 automated DNA sequencer or ABI PRISM 3100-Avant Genetic Analyzer with M13 forward and reverse primers.

The newly cloned regions in this study are: the upstream region from exon 2 in 29Y-2; the upstream region from exon 3 in 29Y-3; and the upstream region from exon 3 and the downstream region from exon 5 in 29Y-4 (compare Fig. 1B of Kawamura et al., 2002 and Fig. 2 of this article). The newly sequenced regions are: exons 1, 2 and 6 (with their respective flanking regions) in 29Y-1 (Genbank accession nos. AB181207, AB181208 and AB181209, respectively), 29Y-2 (AB181210, AB181211 and AB181212) and 29Y-3 (AB181213, AB181214 and AB181215); exon 6 and its flanking regions in 29Y-4 (AB181216); flanking regions of exons 3, 4 and 5 in 29Y-1 (AB081278, AB081279 and AB081280 updated from the ones containing only exon sequences), 29Y-2 (AB081281, AB081282 and AB081283 updated likewise), 29Y-3 (AB081284, AB081285 and AB081286 updated likewise), and 29Y-4 (AB084907, AB084908 and AB084909 updated likewise); 3′-flanking regions of all X- and Y-linked genes (the *Sac*I-*Sac*I region in 14X and 29X and the *Sac*I–*Bgl*II region in the others [see Fig. 2]) (AB181217 [29X], AB181218 [29Y-1], AB181219 [29Y-2], AB181220 [29Y-3], AB181221 [29Y-4], AB181222 [14X] and AB181223 [14Y]). For readers' convenience sake, Genbank accession nos. for the other gene regions given in our previous study (Kawamura et al., 2002) are given again here: AB081260–AB081265 (exons 1, 2, 3, 4, 5, and 6 with their respective flanking regions in 14X), AB081266–AB081271 (those in 14Y) and AB081272–AB081277 (those in 29X).

2.4. DNA sequence analysis

Repetitive sequence elements were identified using Repeat Masker (http://www.repeatmasker.org). The sequence data of human red and green opsin genes, including introns and flanking regions, were retrieved from human X chromosome sequences (Genbank accession nos. Z68193 and AC092402). For the single-copy marmoset red-green (M/LWS) opsin gene, the 5′-flanking sequence was from Genbank AF155218 and coding regions, introns and 3′-flanking sequences were from AB046561–AB046566. Nucleotide sequences were aligned using CLASTAL W and alignment was refined visually. Subsequent phylogenetic analyses were conducted using the MEGA2 program version 2.1 (Nei and Kumar, 2000). The number of nucleotide substitutions per site (d) for two sequences was estimated from their number of nucleotide differences per site by the method of Tamura and Nei (1993) which takes into account differences in substitution rates between nucleotides, inequality of nucleotide frequencies and compensates for multiple substitutions. In the calculation, gap sites were removed in a pair-wise fashion. Phylogenetic trees were constructed by applying the neighbor-joining

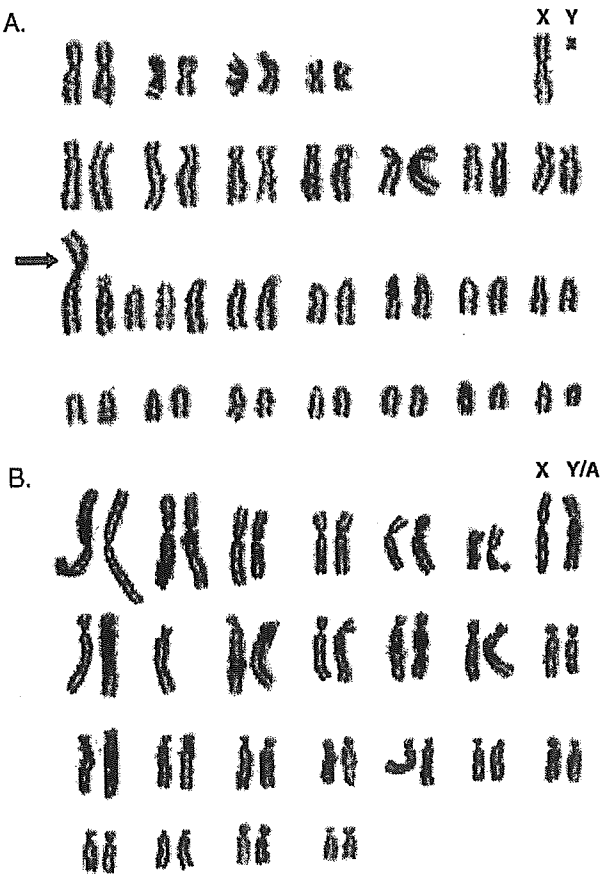

Fig. 1. Two different karyotypes of male owl monkeys detected in the present study. (A) Chromosomes of type 14 owl monkey (no. 14; $2n=53$) with a normal sex chromosome complement (X, Y). The large metacentric chromosome derived from a possible Robertsonian translocation between two medium-sized acrocentric chromosomes is indicated by an arrow. (B) Chromosomes of type 29 owl monkeys (nos. 24, 28 and 44; $2n=49$) with an autosome/Y-chromosome (Y/A) translocation.

method. The reliability of the tree topology was evaluated by bootstrap analysis with 1000 replications.

3. Results

3.1. Species distinction by chromosomal typing

The diploid chromosome number of the type 14 male owl monkey (no. 14) was found to be 53. Each metaphase cell contained a large metacentric chromosome, which may be the derived chromosome from a Robertsonian translocation between two medium-sized acrocentric chromosomes (Fig. 1A, arrow). This karyotype corresponds to karyotype III according to the classification of owl monkey karyotypes by Ma (1981). This also corresponds to karyomorph denomination 2 as defined by Reumer and de Boer (1980). All type 29 male owl monkeys (nos. 24, 28 and 44) showed 49 chromosomes with an autosome/Y-chromosome translocation (Fig. 1B, Y/A). This corresponds to karyotype VI by the Ma's definition (1981) and karyomorph 5 described by Reumer and de Boer (1980). Based on the known geographic distribution of owl monkey species associated with karyotypes, owl monkey type 14 was assigned as *Aotus lemurinus griseimembra*, inhabiting northern Colombia, and type 29 as *A. azarae boliviensis*, inhabiting northern Bolivia (Hershkovitz, 1983; Torres et al., 1998). This is consistent with the institutional record that type 29 monkeys have a Bolivian origin.

3.2. Genomic organization of owl monkey red-green opsin genes

We have cloned the entire gene region for all X- and Y-linked red-green opsin genes, with the exception of 29Y-4, from the two species of owl monkeys in this and previous studies (Kawamura et al., 2002) (Fig. 2). In 29Y-4, the upstream region from exon 3 had a distinct restriction site distribution from those of the other genes and the region was devoid of exons 1 and 2. No other clones containing these exons were isolated from the genomic library after multiple rounds of gene screening. Hence, we consider that the 29Y-4 gene is a product of partial gene duplication involving only the downstream region from exon 3.

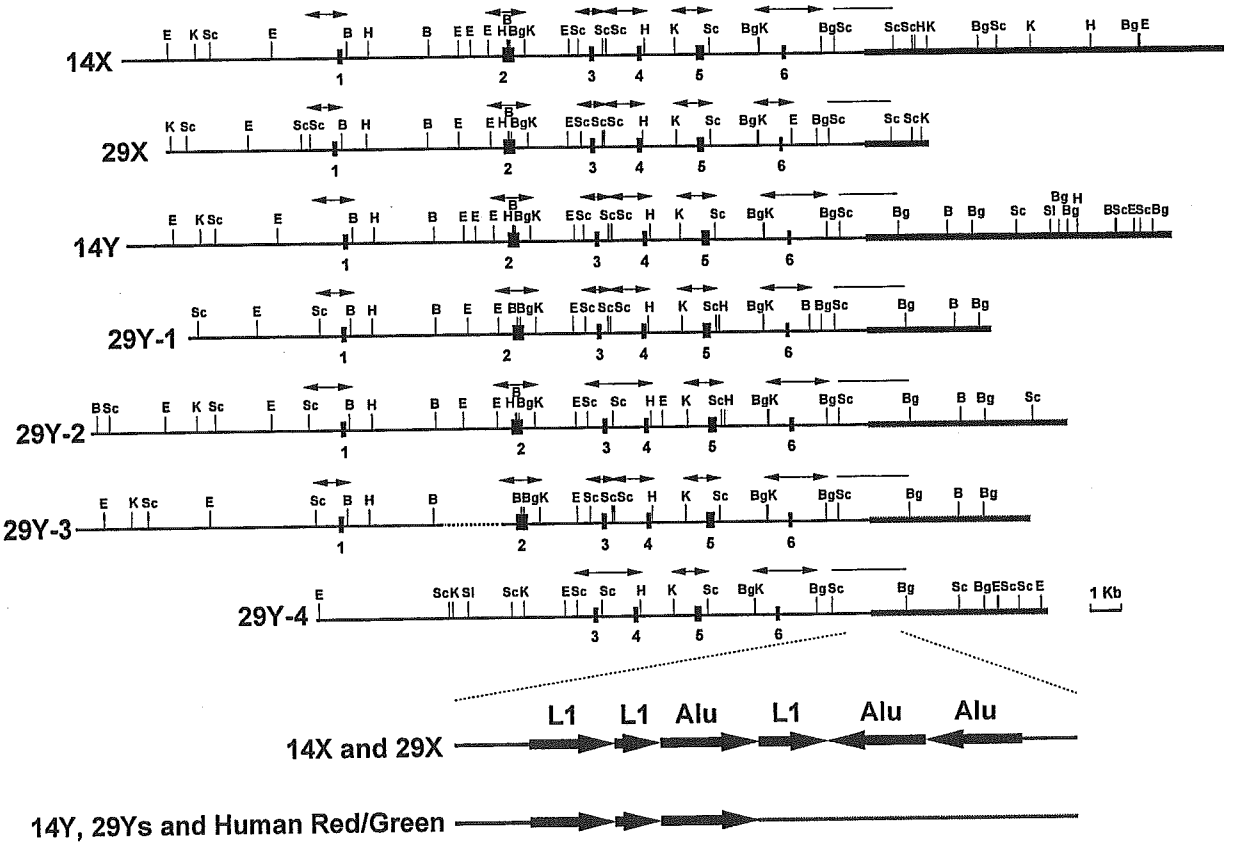

Fig. 2. Genomic organization of red-green opsin genes of owl monkeys. The six exons are depicted with solid boxes. Sequenced regions encompassing exons, determined by this investigation or in our previous study (Kawamura et al., 2002), are indicated by double-headed arrows. The distinct downstream regions between X- and Y-linked genes are indicated with blue and red colors, respectively. The boundaries of the colored regions subjected to DNA sequencing are indicated with horizontal bars. The L1 and Alu repetitive elements identified in the boundary region are depicted in the expanded view of this region with arrows showing their orientations. An intron region in 29Y-3, not determined for restriction site distribution, is depicted with a dotted line. B, *Bam*HI; Bg, *Bgl*II; E, *Eco*RI; H, *Hin*dIII; K, *Kpn*I; Sc, *Sac*I; Sl, *Sal*I.

From these genes, all coding regions (the six exons) and their surrounding non-coding regions (introns and 5′- and 3′-flanking regions) were sequenced (Fig. 2; regions indicated with double-headed arrows). Table 1 lists total lengths of nucleotide sequences determined for the coding (CDR) and non-coding regions (non-CDR). By sequencing only exons 3, 4 and 5, we previously found that 29Y-1, 29Y-2, 29Y-3 and 29Y-4 genes have a premature stop codon in exon 5 and that 29Y-1, 29Y-3 and 29Y-4 genes have an additional stop codon in exon 4 (Kawamura et al., 2002). In this study, we also found one frame-shift nucleotide insertion in exon 2 of 29Y-1, 29Y-2 and 29Y-3.

3.3. Common translocation origin of Y-linked opsin genes revealed by L1-Alu-Alu insertion

We previously noted that Y-linked opsin genes have a similar restriction site distribution with each other in the downstream region of the gene, distinct from the site distribution in the downstream region of X-linked genes (colored regions in Fig. 2; see also Fig. 1B of Kawamura et al., 2002). This implies a common translocation origin of Y-linked opsin genes between the two species. Nucleotide sequencing of the region encompassing the boundary of the common and distinct regions between X- and Y-linked genes (the regions indicated with horizontal bars in Fig. 2; the sequenced lengths listed in Table 1 under "Repeat") revealed characteristic insertions of repetitive elements in this region (Fig. 2 bottom). All genes had two L1 and one Alu repetitive elements in common. The same insertion was also found at the same downstream location in human red and green opsin genes, though they contained an additional Alu element within the most proximal L1 element to the gene.

Importantly, 14X and 29X contained an additional L1 and two Alu elements (blue arrows in the Fig. 2 bottom) adjacent to the common L1-L1-Alu insertion site. The L1-Alu-Alu elements were absent in 14Y and all 29Ys (29Y-1, 29Y-2, 29Y-3 and 29Y-4) as well as in human red and green opsin genes. The nucleotide difference between owl monkey (14Y and 29Ys) and human (red and green opsin genes) in the red-colored region sequenced in Fig. 2 was 9.1–10.0%, comparable to the nucleotide differences between them in the introns and 5′- and 3′-flanking regions (10.5–11.9%). On the other hand, in the 100-kb region encompassing the human red and green opsin genes (Genbank AC092402), we could not find any significant similarity to the blue-colored region of 14X and 29X sequenced in Fig. 2. This implies that an insertion (of the blue-colored region in Fig. 2) rather

Table 1
Lengths of nucleotide sequences (bp) determined for the coding and non-coding regions of the owl monkey red-green opsin genes

	CDR[a]	Non-CDR[b]	Repeat[c]
14X	1095	4793	1951
29X	1095	5264	1941
14Y	1095	4692	1875
29Y-1	1096	5538	2296
29Y-2	1096	6146	2283
29Y-3	1096	5898	2303
29Y-4	686	4407	2301

[a] Coding regions comprising the coding exon sequences.
[b] Combined non-coding regions comprising introns and immediate flanking regions to the start and stop codons.
[c] The *Sac*I–*Sac*I (14X and 29X) or *Sac*I–*Bgl*II (the others) genomic region containing the Alu and L1 repeat sequences downstream from the opsin gene (see Fig. 2).

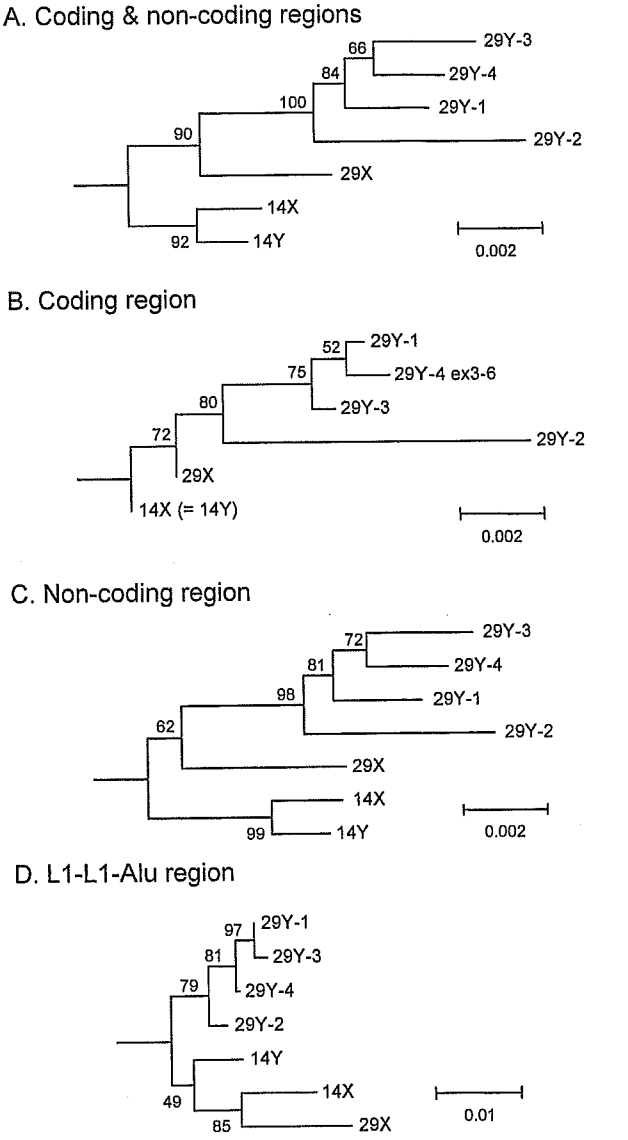

Fig. 3. Phylogenetic trees of owl monkey red-green opsin genes reconstructed using (A) exons 1–6, introns and 5′- and 3′-flanking regions, (B) exons 1–6, (C) introns and 5′- and 3′-flanking regions, and (D) the downstream L1-L1-Alu repetitive element complex. To give a phylogenetic root, the marmoset M/LWS opsin gene was used for (A), (B) and (C), and the human red opsin gene was used for (D). Bootstrap percent probabilities are indicated at each node. Scale bar: number of nucleotide substitutions per site.

Table 2
The number of nucleotide substitutions per 100 sites in the coding (d_C; above diagonal) and non-coding (d_N; below diagonal) regions±their standard errors and the number of compared nucleotides (parenthesized) among the owl monkey and marmoset red-green opsin genes

	14X	29X	14Y	29Y-1	29Y-2	29Y-3	29Y-4	CJA[a]
14X		0.09±0.09 (1095)	0.00±0.00 (1095)	0.55±0.23 (1095)	0.92±0.29 (1095)	0.46±0.21 (1095)	0.59±0.29 (686)	1.67±0.40 (1095)
29X	1.01±0.15 (4792)		0.09±0.09 (1095)	0.46±0.21 (1095)	0.83±0.28 (1095)	0.37±0.18 (1095)	0.44±0.25 (686)	1.76±0.41 (1095)
14Y	0.32±0.08 (4683)	1.03±0.15 (4690)		0.55±0.23 (1095)	0.92±0.29 (1095)	0.46±0.21 (1095)	0.59±0.29 (686)	1.67±0.40 (1095)
29Y-1	1.21±0.16 (4742)	1.01±0.14 (5198)	1.17±0.16 (4641)		0.92±0.29 (1096)	0.27±0.16 (1096)	0.15±0.15 (686)	2.23±0.46 (1095)
29Y-2	1.26±0.16 (4746)	1.24±0.16 (5213)	1.20±0.16 (4645)	0.76±0.12 (5538)		1.01±0.31 (1096)	1.18±0.42 (686)	2.61±0.50 (1095)
29Y-3	1.31±0.17 (4786)	1.13±0.15 (5073)	1.27±0.17 (4685)	0.62±0.11 (5358)	0.83±0.12 (5857)		0.15±0.15 (686)	2.14±0.45 (1095)
29Y-4	1.10±0.18 (3472)	0.87±0.16 (3480)	1.05±0.18 (3371)	0.42±0.11 (3813)	0.82±0.14 (4407)	0.44±0.10 (4301)		2.38±0.60 (686)
CJA[a]	7.78±0.51 (3383)	7.70±0.47 (3273)	7.75±0.51 (3826)	7.78±0.48 (3820)	8.03±0.48 (3934)	7.88±0.49 (3656)	8.65±0.66 (2220)	

[a] *Callithrix jacchus* (common marmoset).

than a deletion (of red-colored region) occurred in the 14X and 29X downstream region, possibly mediated by L1 and Alu repeat elements. Whichever the event, it must have occurred on the X-chromosomal owl monkey opsin gene before the split of the two species of owl monkeys and after the X to Y duplication of the opsin gene. This indicates that the Y-translocation of the opsin gene occurred in the common ancestor of the two species; that is, the common ancestor of the northern and southern species groups of the owl monkeys.

3.4. Concerted evolution between X- and Y-linked opsin genes

To elucidate the evolutionary history of X- and Y-linked owl monkey opsin genes in detail, a phylogenetic tree was reconstructed using entire coding and non-coding regions sequenced (corresponding to the CDR and Non-CDR, respectively, in Table 1). Contrary to the conclusion derived from the L1-Alu-Alu insertion that the Y-linked opsin genes of the two owl monkey species occurred in their common ancestor, topology of the reconstructed tree inferred independent origins of the Y-linked opsin genes in the two species, showing that the X- and Y-linked genes clustered together by species with high bootstrap probabilities (90% in type 29, *A. azarae boliviensis*, and 92% in type 14, *A. lemurinus griseimembra*) (Fig. 3A). Virtually identical trees were obtained when Jukes and Cantor's, Kimura's two-parameter, Tajima and Nei's and Tamura's methods (Nei and Kumar, 2000) were used for estimating d values. When coding and non-coding regions were separated, both trees again showed clustering of X- and Y-linked opsin genes by each species (Fig. 3B and C). When the L1-L1-Alu region was considered, which is located at the 3′ edge of the homologous region among all red-green opsin genes (Fig. 2), 14X and 29X clustered with a high bootstrap probability (85%), though the phylogenetic position of 14Y was not unambiguously determined in the tree (49% bootstrap probability) (Fig. 3D). These results indicate that some process of sequence homogenization, such as gene conversion or homologous recombination, occurred in each species between X- and Y-linked opsin genes in the coding and surrounding non-coding regions but not in the downstream L1-L1-Alu region.

Table 3
Differences between d_N and d_C values (d_N–d_C) among the owl monkey red-green opsin genes

	14X	29X	14Y	29Y-1	29Y-2	29Y-3
29X	0.92±0.17***					
14Y	0.32±0.08***	0.94±0.17***				
29Y-1	0.66±0.28*	0.55±0.25*	0.62±0.28*			
29Y-2	**0.34±0.33**	**0.41±0.32**	**0.28±0.33**	**−0.16±0.31**		
29Y-3	0.85±0.27**	0.76±0.23***	0.81±0.27**	**0.35±0.19**	**−0.18±0.33**	
29Y-4	**0.51±0.34**	**0.43±0.30**	**0.46±0.34**	**0.27±0.19**	**−0.36±0.44**	**0.29±0.18**

Statistical significance of differences between d_N and d_C values listed in Table 2 was evaluated by the two-tail Z test. Gene pairs with no significant difference between d_N and d_C values are indicated with boldface letters.

* Significant at 5% level.

** Significant at 1% level.

*** Significant at 0.1% level.

Table 4
Probabilities that evolutionary rate is the same between two red-green opsin genes of owl monkeys evaluated by the Tajima's relative rate test using the marmoset M/LWS opsin gene as an out-group reference in the coding (above diagonal) and non-coding (below diagonal) regions

	14X	29X	14Y	29Y-1	29Y-2	29Y-3	29Y-4
14X		0.317	N.E.[a]	**0.014**	**0.002**	**0.025**	**0.046**
29X	0.752		N.E.[a]	**0.025**	**0.003**	**0.046**	0.083
14Y	0.405	0.411		N.E.[a]	N.E.[a]	N.E.[a]	N.E.[a]
29Y-1	0.647	0.739	0.330		0.206	0.564	0.317
29Y-2	0.555	0.572	0.206	0.705		0.132	0.157
29Y-3	1.000	1.000	0.639	0.670	0.465		0.317
29Y-4	0.695	0.346	0.827	0.257	0.827	1.000	

Statistically significant rate differences (<0.05) are indicated by boldface letters.
[a] Not evaluated because nucleotide sequences are identical between 14X and 14Y coding regions.

3.5. Subsequent decoupling of evolutionary mode between X- and Y-linked opsin genes in the southern species

Once duplicated genes are homogenized, these genes are closely placed in a reconstructed phylogenetic tree. However, if the homogenization ceases at some evolutionary time point, the two genes thereafter start evolving independently while keeping clustered in the tree. This seems to be the case in the relationship between 29Ys and 29X of the southern species of owl monkeys. In the coding region, evolutionary rates appear to be considerably higher in 29Y-1, 29Y-2, 29Y-3 and 29Y-4 than in the others (Fig. 3B). Table 2 lists the evolutionary distances (d) among the owl monkey opsin genes in both the coding and the non-coding regions (CDR and Non-CDR, respectively, in Table 1). Evolutionary distances in the coding region (d_C) are not significantly different from those in the non-coding region (d_N) between any two of 29Ys, while d_C values are significantly smaller than d_N at 0.1% level between any two of 14X, 29X and 14Y (Table 3).

We then evaluated evolutionary rate differences between the opsin genes in the owl monkeys by applying a relative rate test (Tajima, 1993) with the marmoset M/LWS opsin gene as an out-group reference. In the non-coding region, no significant difference was detected between any two red-green opsin genes of the owl monkeys. In the coding region, a significant difference was detected between 29Ys and the rest (14X, 14Y [sequence identical with 14X] and 29X; $p=0.002-0.046$); i.e., evolutionary rates of 29Ys were significantly faster than those of the rest, except between 29Y-4 and 29X ($p=0.083$) (Table 4). The non-significance in the 29Y-4/29X pair is likely due to the short length of the 29Y-4 sequence although the 0.083 probability is still lower than that for the 14X/29X pair and intra-29Ys pairs (Table 4). In summary, evolutionary rates in the coding regions of 29Ys have increased to the rate in their non-coding regions and are significantly faster than those of 29X, 14X and 14Y, thus showing clear decoupling of evolutionary mode between X- and Y-linked opsin genes in the southern species.

4. Discussion

We previously reported that the Y-linked red-green opsin genes existed in owl monkeys as multiple pseudogenes on the Y/autosome fusion chromosome or a single-copy intact gene on the regular Y chromosome (Kawamura et al., 2002). In the present study, we verified the two types of monkeys being distinct allopatric species on the basis of karyotypic characteristics: the former being *A. azarae boliviensis* and the latter *A. lemurinus griseimembra*, belonging to the southern and northern species groups of owl monkeys, respectively (Fig. 1). The presence of the unique L1-Alu-Alu repetitive elements in the downstream region from the X-linked opsin genes in both species (Fig. 2) indicates that the X to Y duplication of the opsin gene predates the split of the northern and southern species groups. Clustering of the X- and Y-linked opsin genes by species in the reconstructed phylogenetic trees for the coding and adjacent non-coding regions (Fig. 3A–C) elucidates the occurrence of concerted evolution between the X- and Y-linked opsin genes (Fig. 4 top row). The Y-linked opsin genes of the southern species accumulated nonsense mutations, multiplicated and evolved as fast in the coding region as in the non-coding region, an indication that the genes have lost functional constraint and evolved independently from the X-lined copy (Fig. 4 bottom row).

4.1. Gene conversion between X and Y copies of opsin genes

How is concerted evolution between X- and Y-linked genes possible? One possibility is that these genes might be located in a pseudoautosomal region (PAR) in owl monkeys. A second possibility is that the gene conversion has occurred repeatedly between the two chromosome genes. At PAR, X- and Y-linked genes are paired at meiosis and frequently recombined with each other similar to allelic genes on autosomes. The red-green opsin genes are located at the telomeric region of the long arm of the X chromosome in mammals (Xq28 in human) (Nathans et al., 1986; Kawamura et al., 2001). The red-green opsin genes of owl monkeys on the X chromosome and on the ordinary Y chromosome are also mapped to the distal region of the long arm (Kawamura et al., 2002). The human X chromosome has two PARs, PAR1 and PAR2, placed at the short- and long-arm tips, respectively (Graves et al., 1998). In humans, the red and green opsin genes are located close to but not in PAR2, approximately 1.5 Mb toward centromeric direction from the HSPRY3 gene that is located at the most centromeric side of the PAR2 (NCBI Human Genome Resources; http://www.ncbi.nlm.nih.gov/genome/guide/human). Whereas homologous regions to PAR1 are widely observed among eutherian mammals on the X and Y chromosomes, PAR2 appears to be specific to humans and absent in the Y chromosomes of non-human primates (Ciccodicola et al., 2000; Charchar et al., 2003). Therefore, it is unlikely that the opsin genes on the X and Y

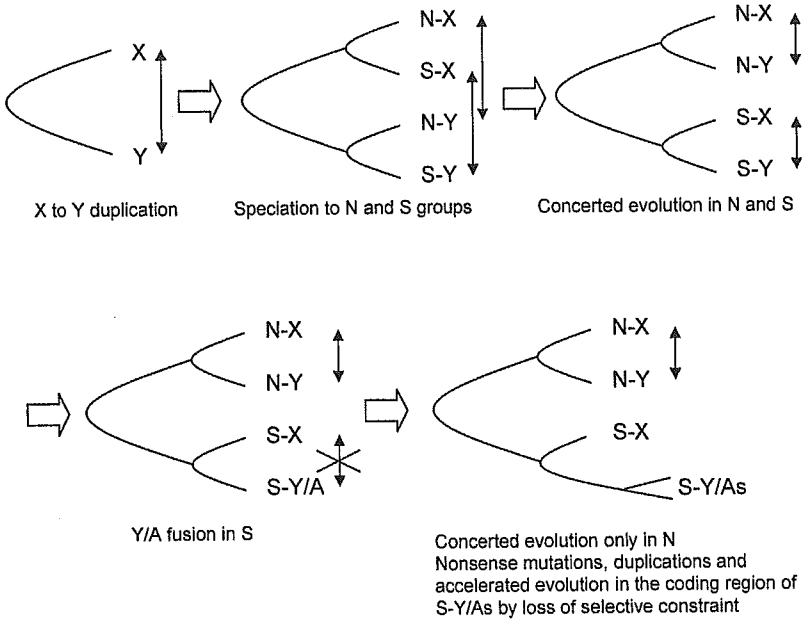

Fig. 4. Evolutionary scenario of owl monkey red-green opsin genes. X and Y represent X- and Y-linked opsin genes, respectively. Y/A indicates the red-green opsin gene on the Y-autosome fusion chromosome. N and S represent the northern and southern species groups, respectively. Arrows indicate DNA sequence exchange between X- and Y-linked opsin genes. Note that sequence exchange could have occurred even before the split of the two species groups.

chromosomes are part of a PAR in owl monkeys. Moreover, it is difficult to explain the difference in the 3′ flanking region of the X and Y opsin genes under PAR hypothesis. On the other hand, gene conversion events are local and could leave some genomic region different between the two loci. Gene conversion between X and Y copies and among Y copies that are not in PAR have been documented (Pecon Slattery et al., 2000; Rozen et al., 2003). Taken together, gene conversion would be the most plausible mediator of the concerted evolution between the two chromosomal copies of opsin genes in owl monkeys.

4.2. Y-autosome fusion and decoupling of evolutionary process between X- and Y-linked opsin genes

What is the cause of the subsequent decoupling of evolution between the X- and Y-linked opsin genes in the southern species? Chromosomal fusion between Y and an autosome has been reported in *A. nigriceps*, *A. infulatus*, *A. azarae boliviensis* and *A. a. azarae*, all of which belong to the southern species group (Pieczarka and Nagamachi, 1988; Ma et al., 1989). In the first three, the Y chromosome fusion occurs with the short arm of a medium-size subtelocentric autosome. G-banding and gene mapping provided evidence that this autosome is the same chromosome among them though all three appear to have experienced different chromosomal rearrangements in the Y-autosomes after the fusion event. In *A. a. azarae*, the Y-fusion is to a small acrocentric autosome, but all the other chromosomes are identical in G- and C-banding patterns to *A. a. boliviensis* chromosomes. Taken together, the Y-autosome fusion is considered to have occurred once in the evolution in their common ancestor (Pieczarka and Nagamachi, 1988; Ma et al., 1989). We speculate that the Y-autosome fusion could have impeded gene conversion between the X and Y copies of red-green opsin genes through some change in the macrostructure or locality of the chromosome in the cell and insulated the Y-linked copy from the genetic information flow from the original X-chromosomal copy. This could allow for the accumulation of nonsense mutations in the Y-linked copy, as in many other Y-chromosomal genes without male-specific function (Graves, 1995) and for the multiplication of the gene, though correlation between gene multiplication and accumulation of nonsense mutations is not clear.

4.3. Functional implication on the Y-linked opsin genes

We previously inferred a potential advantage in parental behaviors for males with extra red-green opsin, which almost exclusively occur in males, by having higher light sensitivity (Kawamura et al., 2002). The intact Y-linked opsin gene (14Y) has an identical DNA sequence with the X-linked counterpart (14X) in the coding region. On the other hand, non-coding regions of 14X and 14Y accumulate differences independently (the $d_N - d_C$ difference is statistically significant at 0.1% level; see Tables 2 and 3). Suppose the rate of the homogenizing event, such as gene conversion, between the X- and Y-linked opsin genes is equal between the coding and non-coding regions, the significant $d_N - d_C$ difference between 14X and 14Y implies that 14Y is under as strict a functional constraint as is 14X.

Expression of 14Y, however, is not currently verified because 14X and 14Y are identical not only in the coding region but also in the putative untranslated regions and are indistinguishable even if expressed. Considering that the Y-linked opsin genes are all non-functional in the southern species and that this is likely due to the insulation of the Y-linked gene from genetic information flow from the X-linked copy, the functional constraint on the intact Y-linked opsin gene in the northern species is possibly an indirect one, with 14Y influencing 14X by gene conversion. This would predict that in other owl monkey species, Y-linked opsin genes on Y-autosome are likewise pseudogenes and those on ordinary Y are similar to X-linked counterparts and remain intact. A survey of the Y-linked opsin gene in other owl monkey species may reveal sequence variation (SNPs) of the Y-linked opsin genes, which would be useful for examination of their expression and provide better insight on their biological significance.

4.4. Divergence between the northern and southern owl monkeys

The origin of owl monkeys can be traced back by fossil records to the early Miocene (~19 million years ago [MYA]): *Aotus*-like *Tremacebus harringtoni* in Argentina and the middle Miocene (~9–14 MYA); *Aotus dindensis* in Colombia (Fleagle, 1999). Recent molecular data also supports the early split of the *Aotus* lineage from Cebinae and Callitrichinae species, dating it at ~23 MYA (Schneider, 2000). Despite the early separation of *Aotus* from the other genera, extant chromosomal variability in *Aotus* has been suggested to originate in Pleistocene (~2–0.01 MYA) as a result of isolation of populations in forest refugia by repeated dramatic climatic changes (Ma, 1981). However, intra-genus genetic distances and taxonomy have not been evaluated by DNA sequence data.

By comparing non-coding regions, the divergence time between the northern and southern species groups can be estimated from the evolutionary distances (d_N) listed in Table 2. The average d_N value between the northern (14X and 14Y) and southern (29X and 29Ys) species was 0.0116±0.0016 (per site); that between 29X and 29Ys was 0.0106±0.0015, and the d_N between 14X and 14Ywas 0.0032±0.0008. When the evolutionary rate of 3.5×10^{-9}/site/year for mammalian intron sequences was applied (Li, 1997), divergence time of 1.7±0.2, 1.5±0.2, and 0.5±0.1 MYA were given for 14X/14Y vs. 29X/29Ys, 29X vs. 29Ys, and 14X vs. 14Y, respectively. Given the effect of the concerted evolution between the X and Y copies, this estimate for 14X/14Y vs. 29X/29Ys could include the divergence time between X and Y genes before the speciation (Fig. 4). The smaller estimate for 14X vs. 14Y than for 29X vs. 29Ys is supposed to reflect the continued effect of concerted evolution between 14X and 14Y. Therefore, divergence time of the two species can be estimated as no later than 1.5 MYA and no earlier than 1.7 MYA, i.e. ~1.6 MYA. This is consistent with the Pleistocene divergence of extant owl monkeys. Based on the pelage and karyotypic characters, the southern species group is considered to be derived from the northern species group (Galbreath, 1983; Hershkovitz, 1983). If this is the case, it can be predicted that the Y-chromosomal opsin gene will be found in all the other southern species and at least some other northern species since X to Y duplication of the opsin gene predates the northern-southern species separation. The date of the X to Y duplication event, however, is impossible to estimate from our present data set because the differences between the X- and Y-linked opsin genes must have been largely erased by concerted evolution, especially in the northern species, and possibly even before the speciation event (Fig. 4). To elucidate the date, it is necessary to establish the intra-genus phylogeny and divergence times of owl monkeys using appropriate molecular markers, such as mitochondrial DNA, and to survey the presence/absence status of Y-linked opsin genes on the phylogeny. A similar behavior concerning the opsin gene has never been reported in other primates nor in other vertebrates. At any rate, because of their recent origin and diversity, the Y-linked opsin genes of owl monkeys should become an excellent model to study the evolution of genes translocated to the Y chromosome.

Acknowledgements

This study was supported by Grant-in-Aid for Scientific Research (B) from Japan Society of the Promotion of Science (12440243) to S.K. and Cooperative Research Program of the Primate Research Institute of Kyoto University through Dr. Osamu Takenaka to S.K. and N.T. This manuscript was proofread by BioMed Proofreading Service.

References

Charchar, F.J., et al., 2003. Complex events in the evolution of the human pseudoautosomal region 2 (PAR2). Genome Res. 13, 281–286.

Ciccodicola, A., et al., 2000. Differentially regulated and evolved genes in the fully sequenced Xq/Yq pseudoautosomal region. Hum. Mol. Genet. 9, 395–401.

Fleagle, J.G., 1999. Primate Adaptation and Evolution, 2 ed. Academic Press, San Diego.

Galbreath, G.J., 1983. Karyotypic evolution in *Aotus*. Am. J. Primatol. 4, 245–251.

Graves, J.A., 1995. The origin and function of the mammalian Y chromosome and Y-borne genes-an evolving understanding. BioEssays 17, 311–320.

Graves, J.A., Wakefield, M.J., Toder, R., 1998. The origin and evolution of the pseudoautosomal regions of human sex chromosomes. Hum. Mol. Genet. 7, 1991–1996.

Hershkovitz, P., 1983. Two new species of night monkeys, genus *Aotus* (Cebidae, Platyrrhini): a preliminary report on *Aotus* taxonomy. Am. J. Primatol. 4, 209–243.

Hirai, M., Suto, Y., Kanoh, M., 1994. A method for simultaneous detection of fluorescent G-bands and in situ hybridization signals. Cytogenet. Cell Genet. 66, 149–151.

Hiramatsu, C., Radlwimmer, F.B., Yokoyama, S., Kawamura, S., 2004. Mutagenesis and reconstitution of middle-to-long-wave-sensitive visual pigments of New World monkeys for testing the tuning effect of residues at sites 229 and 233. Vision Res. 44, 2225–2231.

Jacobs, G.H., Deegan, J.F., Neitz, J., Crognale, M.A., Neitz, M., 1993. Photopigments and color vision in the nocturnal monkey, *Aotus*. Vision Res. 33, 1773–1783.

Jacobs, G.H., Neitz, M., Neitz, J., 1996. Mutations in S-cone pigment genes and the absence of colour vision in two species of nocturnal primate. Proc. R. Soc. Lond., B 263, 705–710.

Kawamura, S., Hirai, M., Takenaka, O., Radlwimmer, F.B., Yokoyama, S., 2001. Genomic and spectral analyses of long to middle wavelength-sensitive visual pigments of common marmoset (*Callithrix jacchus*). Gene 269, 45–51.

Kawamura, S., Takenaka, N., Hiramatsu, C., Hirai, M., Takenaka, O., 2002. Y-chromosomal red-green opsin genes of nocturnal New World monkey. FEBS Lett. 530, 70–72.

Li, W.H., 1997. Molecular Evolution. Sinauer, Sunderland.

Ma, N.S.-F., 1981. Chromosome evolution in the owl monkey, *Aotus*. Am. J. Phys. Anthropol. 54, 293–303.

Ma, N.S.-F., Page, D.C., Harris, T.S., 1989. Molecular evidence of Y-autosomal translocations in owl monkeys. J. Heredity 80, 259–263.

Nathans, J., Piantanida, T.P., Eddy, R.L., Shows, T.B., Hogness, D.S., 1986. Molecular genetics of inherited variation in human color vision. Science 232, 203–210.

Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Pecon Slattery, J., Sanner-Wachter, L., O'Brien, S.J., 2000. Novel gene conversion between X-Y homologues located in the nonrecombining region of the Y chromosome in Felidae (Mammalia). Proc. Natl. Acad. Sci. U. S. A. 97, 5307–5312.

Pieczarka, J.C., Nagamachi, C.Y., 1988. Cytogenetic studies of *Aotus* from eastern Amazonia: Y/autosome rearrangement. Am. J. Primatol. 14, 255–263.

Pieczarka, J.C., De Souza Barros, R.M., Nagamachi, C.Y., Rodrigues, R., Espinel, A., 1992. *Aotus vociferans* × *Aotus nancymai*: sympatry without chromosomal hybridization. Primates 33, 239–245.

Pieczarka, J.C., De Souza Barros, R.M., De Faria Jr., F.M., Nagamachi, C.Y., 1993. Aotus from the southwestern Amazon region is geographically and chromosomally intermediate between *A. azarae boliviensis* and *A. infulatus*. Primates 34, 197–204.

Reumer, J.W.F., de Boer, L.E.M., 1980. Standardization of *Aotus* chromosome nomenclature, with description of the $2n=49-50$ karyotype and that of a new hybrid. J. Hum. Evol. 9, 461–482.

Rozen, S., et al., 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature 423, 873–876.

Schneider, H., 2000. The current status of the New World monkey phylogeny. An. Acad. Bras. Cienc. 72, 165–172.

Tajima, F., 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135, 599–607.

Tamura, K., Nei, M., 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10, 512–526.

Torres, O.M., Enciso, S., Ruiz, F., Silva, E., Yunis, I., 1998. Chromosome diversity of the genus Aotus from Colombia. Am. J. Primatol. 44, 255–275.

Wikler, K.C., Rakic, P., 1990. Distribution of photoreceptor subtypes in the retina of diurnal and nocturnal primates. J. Neurosci. 10, 3390–3401.

Wright, P.C., 1994. The behavior and ecology of the owl monkey. In: Baer, J.F., Weller, R.E., Kakoma, I. (Eds.), *Aotus*: The Owl Monkey. Academic Press, New York, pp. 97–112.

Substitution Rate and Structural Divergence of 5′ UTR Evolution: Comparative Analysis Between Human and Cynomolgus Monkey cDNAs

Naoki Osada,[1] *Makoto Hirata,*[1] *Reiko Tanuma,*[1] *Jun Kusuda,*[1] *Munetomo Hida,*† *Yutaka Suzuki,*† *Sumio Sugano,*† *Takashi Gojobori,*‡ *C.-K. James Shen,*§ *Chung-I Wu,*‖ *and Katsuyuki Hashimoto**

*Division of Genetic Resources, National Institute of Infectious Diseases, Tokyo, Japan; †Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo, Chiba, Japan; ‡Center of Information Biology, National Institute of Genetics, Research Organization of Information and Systems, Shizuoka, Japan; §Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan; and ‖Department of Ecology and Evolution, University of Chicago

The substitution rate and structural divergence in the 5′-untranslated region (UTR) were investigated by using human and cynomolgus monkey cDNA sequences. Due to the weaker functional constraint in the UTR than in the coding sequence, the divergence between humans and macaques would provide a good estimate of the nucleotide substitution rate and structural divergence in the 5′UTR. We found that the substitution rate in the 5′UTR (K_{5UTR}) averaged ≈10%–20% lower than the synonymous substitution rate (K_s). However, both the K_{5UTR} and nonsynonymous substitution rate (K_a) were significantly higher in the testicular cDNAs than in the brain cDNAs, whereas the K_s did not differ. Further, an in silico analysis revealed that 27% (169/622) of macaque testicular cDNAs had an altered exon-intron structure in the 5′UTR compared with the human cDNAs. The fraction of cDNAs with an exon alteration was significantly higher in the testicular cDNAs than in the brain cDNAs. We confirmed by using reverse transcriptase–polymerase chain reaction that about one-third (6/16) of in silico "macaque–specific" exons in the 5′UTR were actually macaque specific in the testis. The results imply that positive selection increased K_{5UTR} and structural alteration rate of a certain fraction of genes as well as K_a. We found that both positive and negative selection can act on the 5′UTR sequences.

Introduction

The most pronounced evolutionary conservation of genomic sequences reflects constraints on protein structure and function, and as a result, protein-coding sequences are much more conserved than noncoding sequences. Because of this, it has long been argued that changes in gene regulation may be more important to phenotypic evolution than changes in protein-coding sequences (King and Wilson 1975; Enard et al. 2002). Although functional constraints in intergenic sequences as promoter or enhancer sequences have been thoroughly studied because of their utility as markers of functional parts of noncoding sequences (e.g., Bejerano et al. 2004; Suzuki et al. 2004), the study of evolution of the untranslated region (UTR) of transcripts has been limited owing to the paucity of transcript sequences in appropriate species. Typical mRNA contains UTR upstream (5′UTR) and downstream (3′UTR) of protein-coding sequence. Watanabe et al. (2004) reported that orthologous genes with high divergences in their 5′UTRs tend to show differences in expression levels between humans and chimpanzees, while no correlation between the nucleotide and expression divergence was found in the 3′UTR. Thus, molecular evolutionary study of the 5′UTR is important to understand how our genome has evolved and become organized. As an initial step, we should measure the nucleotide substitution rate and the magnitude of structural divergence in the 5′UTR to infer what type of natural selection has an influence on the 5′UTR evolution.

Due to the weaker functional constraint in the UTR than in the coding sequence (CDS, e.g., Miyata, Yasunaga, and Nishida 1980; Li 1997; Makalowski and Boguski 1998), there is a limitation to study the UTR sequence evolution by using distantly related species. For example, most UTR sequences between humans and mice are not conserved well enough to be aligned, which would considerably hamper the evolutionary analysis especially when we want to find the signature of positive selection (i.e., accelerated evolution). The divergence between humans and macaques are approximately 5%–7% at the nucleotide level (e.g., Savatier et al. 1987; Kawamura et al. 1991; Osada et al. 2002b; Wang et al. 2003), which allows us to compare the macaque UTR sequence with that of humans. So far, the macaque is the only model organism for which UTRs are readily alignable. We constructed cDNA libraries from cynomolgus monkey (*Macaca fascicularis*) brains and testis by the oligo-capping method for a variety of purposes, such as identification of novel human genes (Osada et al. 2001, 2002*a*, 2002*b*) and evolutionary comparative analysis (Osada et al. 2002*c*; Mesak et al. 2003). The cynomolgus monkey cDNA libraries were used to conduct comparative analysis of the 5′UTR sequences of humans and macaques.

The result of recent studies has suggested that the 5′UTRs in humans may have been under positive selection because of the higher substitution rate and lower polymorphism in the 5′UTR than in the synonymous sites (Hellmann et al. 2003). In this report, we compared macaque cDNA sequences from two different organs (brain and testis) with human orthologous cDNAs and reported the substitution rates for nonsynonymous sites (K_a), synonymous sites (K_s), and 5′UTR sites (K_{5UTR}).

Structural divergence of UTRs, such as gains and/or losses of exons, should be an important factor in 5′UTR evolution, as well as in the evolution related to nucleotide substitution. Several reports have shown species-specific

[1] Present address: Division of Biomedical Resources, National Institute of Biomedical Innovation, Osaka, Japan.

Key words: evolution, substitution rate, 5′UTR, alternative splicing, primates.

E-mail: khashi@nih.go.jp.

Mol. Biol. Evol. 22(10):1976–1982. 2005
doi:10.1093/molbev/msi187
Advance Access publication June 8, 2005

gains and/or losses of protein-coding exons between humans and rodents that have an important role in the creation of proteomic diversity (Modrek and Lee 2003; Nurtdinov et al. 2003; Pan et al. 2005). Similarly, several studies have found that transcripts from the same locus have different 5′UTRs as a result of alternative splicing or use of different promoters (Chew et al. 2003; Bernard, Woodruff, and Plant 2004). In some cases, tissue-specific transcripts are generated starting at a particular transcription start site by using *cis*-regulatory elements different from those used in other tissues (Mao, Chirala, and Wakil 2003; Newton et al. 2003), and many such transcripts have been identified for genes expressed in the testis (P. Mezquita, C. Mezquita, and J. Mezquita 1999; Newton et al. 2003; Sugiura et al. 2003). These variants in the exons for 5′UTRs also have a quantitative effect on the translation of the genes (Wang et al. 1999; Chamas and Sabban 2002; Gellersen et al. 2002; Lammich et al. 2004). In this study, we analyzed the structural divergence in the 5′UTR between human cDNAs registered in public databases and cynomolgus monkey cDNAs.

Materials and Methods

cDNA Library from Cynomolgus Monkey

Two cynomolgus monkeys, a 15-year-old male and a 16-year-old female, were used for tissue collection. The monkeys were cared for and handled according to the guidelines established by the Institutional Animal Care and Use Committee of the National Institute of Infectious Diseases (NIID) of Japan. The tissues were harvested in accordance with all the guidelines in the Laboratory Biosafety Manual of the World Health Organization and were carried out at the P3 facility for monkeys of the Tsukuba Primate Center of NIID. Immediately after removing the organs, the tissue samples were frozen with liquid nitrogen. Oligo-capped cDNA libraries were constructed according to the method described previously (Suzuki et al. 1997). The 5′-ends of the cDNAs were capped with oligonucleotides to preserve the full length of the transcript.

Sequencing of Cynomolgus Monkey cDNA Clones

The 5′-end of the testicular cDNA clones were sequenced with an ABI 3700 sequencer (Applied Biosystems Japan, Cjuo-ku, Tokyo, Japan) and clustered with DYNACLUST (DYNACOM, Mobara, Chiba, Japan). We isolated 10,426 cDNA clones, and sequencing their 5′-ends yielded 4,980 clusters of sequences. To investigate how many macaque cDNA clones have valid human homologous genes, we performed a Blast search of the human RefSeq (Pruitt, Tatusova, and Maglott 2003) database. The 5′-end sequences of 6,151 clones had homology to 2,343 human RefSeq genes at a cutoff value of 1×10^{-60}.

The entire sequences of clones were determined by the primer walking method. Cycle sequencing was performed with an ABI PRISM BigDye Terminator Sequencing kit (Applied Biosystems) according to the manufacturer's instructions. We sequenced approximately 2,200 cDNA clones whose 5′-end sequences had homology to human RefSeq sequences. The CDSs of the macaque cDNA clones were searched in the University of California Santa Cruz (UCSC) human genome database (http://genome.ucsc.edu, verified on May 2004) with a visual inspection. For the further analysis, we selected the macaque cDNA clones of which CDSs covered translation start site and nearly entire CDSs of the homologous human cDNAs and obtained 785 human-macaque cDNA pairs. The homologous regions of human-macaque cDNA sequences were aligned to each other using ClustalW (Thompson, Higgins, and Gibson 1994). After removing the ambiguous alignments shorter than 300 bp (100 aa) in the CDSs and redundant cDNA pairs (we occasionally sequenced more than one macaque cDNA per one RefSeq gene), we compiled 622 one-to-one human-macaque orthologous alignments.

cDNA sequences from the brain were also compiled as a control group for the cDNA sequences from the testis. So far, over 8,000 cDNA sequences from the macaque brain have been accumulated. However, we used only 443 cDNA sequences preliminarily extracted from the initial phase of our macaque cDNA sequencing project. The selected 443 brain cDNAs are supposed to have no assortment bias. The analysis using all the macaque brain cDNA sequences will be presented elsewhere. We identified 302 orthologous gene pairs of human and macaque cDNAs with the same procedure as above.

The name of macaque cDNAs represents the anatomical parts where the clone was derived from (Qtr: temporal lobe, Qfl: frontal lobe, Qnp: parietal lobe, Qcc: cerebellum cortex, Qts: testis; see Supplementary Table 1 [Supplementary Material online] for the accession numbers of the macaque cDNA clones and human orthologs). All 2,331 macaque cDNA sequences analyzed in this study were deposited to the DNA Data Bank of Japan/European Molecular Biology Laboratory/GenBank DNA database (accession numbers AB168131–AB169925, AB178956–AB179491).

Computational Analyses

We calculated K_a and K_s by the method of Li-Pamilo-Bianchi (Li 1993; Pamilo and Bianchi 1993). The K_{5UTR} was estimated by using Kimura's two-parameter method (Kimura 1980). Bootstrap test was performed to estimate the sampling variance of substitution rate for the concatenated 5′UTR sequences. We reconstructed 5′UTR sequences of the testicular and brain genes by randomly choosing nucleotides from the original concatenated sequences and estimated K_{5UTR} 1,000 times. BLAT program against the human genome sequences at the UCSC human genome database (http://genome.ucsc.edu, verified on May 2004) was used to visually inspect whether the macaque 5′UTRs had any structural divergence to the homologous human cDNAs in the database.

Reverse Transcriptase–Polymerase Chain Reaction

The templates of the human total RNA from brain, liver, and testis were purchased from Clontech (Mountain View, Calif.). Total RNA of the cynomolgus monkey brain, liver, and testis was isolated using TRIzol (Invitrogen, Carlsbad, Calif.). One microliter of total mRNA was amplified using One Step RNA PCR Kit (TakaraBio, Otsu, Shiga,

Table 1
The Means and Standard Errors of the Substitution Rate Per 100 Sites Between the Human and Cynomolgus Monkey cDNAs

	N_{CDS} (N_{5UTR})[a]	$K_a \times 100$	$K_s \times 100$	$K_{5UTR} \times 100$
Testis	622 (480)	1.464 ± 0.060	5.726 ± 0.102	5.223 ± 0.178
Brain	302 (254)	0.836 ± 0.065	5.841 ± 0.151	4.611 ± 0.233

[a] Number of gene pairs used to estimate the K_a and K_s (N_{CDS}) and the K_{5UTR} (N_{5UTR}).

Japan). Temperature and time schedule were 40 cycles of 94°C for 30 s, 58°C for 30 s, and 72°C for 1.5 min. The primers were designed to match both human and macaque cDNA sequences, and their sequences are presented in Supplementary Table 2 (Supplementary Material online).

Results

Substitution Rate

Using the full-insert sequences of macaque cDNAs, we obtained 622 and 302 orthologous pairs of human and macaque cDNAs derived from the testis and brain, respectively (see *Materials and Methods* for further information). There was no overlapping of genes between the two data sets. Average length of the alignments was 103.88 bp in the 5′UTR and 1,248.17 bp in the CDS. The K_a, K_s, K_{5UTR}, and length of each alignment are presented in Supplementary Table 1 (Supplementary Material online). Because the 5′UTR sequences of some genes were only several base pairs long, we filtered out the alignments containing 5′UTR shorter than 20 bp to estimate the K_{5UTR}, which yielded 480 and 254 5′UTR alignments for the testicular and brain cDNAs, respectively. After the filtering, average length of the alignments in the 5′UTR was 128.91 bp. The mean values and standard errors of the K_a, K_s, and K_{5UTR} per 100 sites are shown in table 1.

In both the organs, the K_a and K_{5UTR} were significantly lower than the K_s by the Wilcoxon matched-pair signed-rank test (P values ranged from 1×10^{-5} to 1×10^{-15}). Next, we tested whether the substitution rates are different between the testicular and brain cDNAs by the Wilcoxon rank-sum test. The testicular cDNAs showed significantly higher K_a ($P < 1 \times 10^{-12}$) and K_{5UTR} ($P = 0.014$) than the brain cDNAs, whereas the K_s did not significantly differ ($P = 0.359$, table 1 and fig. 1). The cumulative distributions of K_a, K_s, and K_{5UTR} are shown in figure 1. Because we did not use the short alignments (less than 20 bp) in the 5′UTR for the former statistical test, we subsequently concatenated all 5′UTR alignments and estimated the K_{5UTR}. Whether the K_{5UTR} of the testicular cDNAs is significantly greater than the K_{5UTR} of the brain cDNAs was tested by a bootstrap method with 1,000 times iteration. The result was highly significant ($P < 0.001$).

Because the high mutability of CpG sites may increase the substitution rate of each class (Hellmann et al. 2003), we estimated the substitution rate after masking all CG to CA and TG substitutions between human and macaque cDNA sequences, but there was no change in the trend as a result of masking (Supplementary Table 3, Supplementary Material online).

Structural Divergence

When the macaque testicular cDNA sequences were aligned to the sequences of the human ortholog, we frequently found the unaligned blocks of 5′UTR sequences between the two. We used BLAT program (Kent 2002) at the UCSC human genome database (http://genome.ucsc.edu) to visually inspect the loci on the human genome where the macaque testicular cDNAs were mapped and found that the many blocks of macaque 5′UTR sequences, which aligned with the human genome sequences, did not show any homology to human cDNA sequences mapped on the same loci. The unaligned macaque sequences that are possibly due to the extension of macaque transcription start site were removed from the further analysis. The exonlike blocks in the human genome appeared to be "macaque-specific" exons in the 5′UTR. Note that the macaque-specific exon does not always literally refer to specific in macaques (e.g., chimpanzees may transcribe the macaque-specific exon). Throughout this report, however, we shall use the macaque specific in terms of the comparison between humans and cynomolgus monkeys. The number of testicular clones with macaque-specific exons

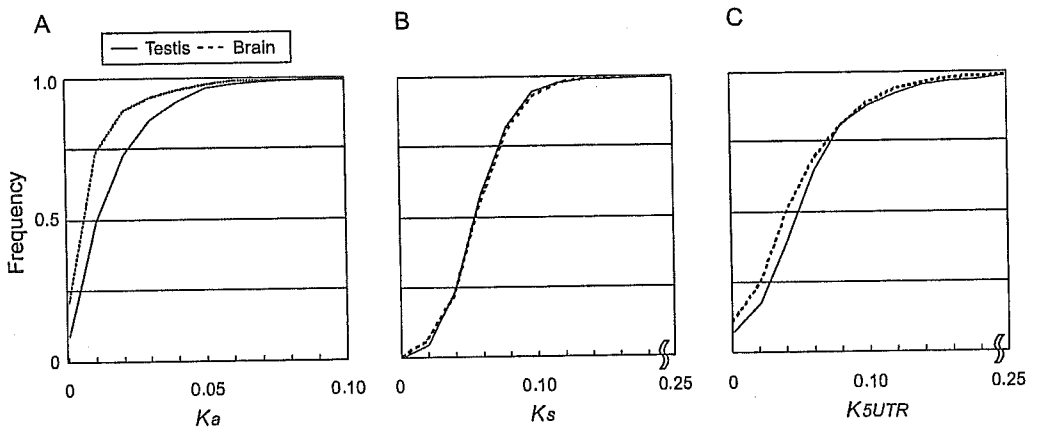

FIG. 1.—The cumulative distribution of the K_a (*A*), K_s (*B*), and K_{5UTR} (*C*) between the human and cynomolgus monkey cDNAs. The solid lines represent the substitution rate of the testicular cDNAs, and dotted lines denote the substitution rate of the brain cDNAs. The testicular cDNAs showed a significantly different distribution from the brain cDNAs in both K_a and K_{5UTR}, while the K_s did not significantly differ.

in the 5′UTR amounted to 169 (27%) of the 622 gene pairs. Supplementary Table 4 (Supplementary Material online) lists the orthologous pairs that were found to carry a macaque-specific exon by in silico analysis.

We investigated whether alteration of the 5′UTR in macaque transcripts occurs more frequently with the clones derived from the testis than the clones from the brain, and the results showed that 41 (14%) of the 302 orthologous pairs in the brain had altered 5′UTRs (data not shown). Thus, the testis-derived transcripts had a significantly higher rate of alteration than the brain-derived transcripts ($P < 1 \times 10^{-5}$: Fisher's exact test).

We classified the macaque-specific exons in the 5′UTR into three categories. Class A consists of macaque cDNA clones whose first exon differs between human and macaque cDNAs. The transcripts in class A have different transcription start sites, and thus may use different promoters and transcription regulatory elements (Supplementary Fig. 1*A*, Supplementary Material online). Class B consists of clones whose first exons start at the same region as human cDNAs but whose intron-exon structure in the 5′UTR differs between humans and macaques (Supplementary Fig. 1*B*, Supplementary Material online). Class C consists of cDNA pairs, part of whose 5′UTR was not found in the human genome (Supplementary Fig. 1*C*, Supplementary Material online).

We randomly selected 18 of the 169 macaque-specific exons found by in silico analysis in the testis and performed the reverse transcriptase–polymerase chain reaction (RT-PCR) to examine whether the 18 macaque-specific exons in silico occurred in vivo. Primers were designed to match both human and macaque sequences and either one of the primer sets on the potential macaque-specific exon. The primer sequences are shown in Supplementary Table 2 (Supplementary Material online). Human and cynomolgus monkey total RNA samples from brain, liver, and testis were used for the expression analysis. Of the 16 pairs of primers that amplified the product at least in one of the macaque samples, six pairs amplified products of the expected size only in the macaque samples. The results are summarized in Supplementary Table 2 (Supplementary Material online). Figure 2*A* (QtsA-12177) and 2*B* (QtsA-17708) shows the examples of the exon-intron structures of transcripts and gel images of the RT-PCR products. The finding that about one-third of macaque-specific exons in silico were the actual macaque-specific exons in vivo indicates that around 10% (169/622 × 6/16) of the testicular clones carry macaque-specific exons in their 5′UTRs.

Discussion

The 622 human-macaque alignments yielded 169 macaque cDNAs carrying macaque-specific exons in the testis in silico. We confirmed that about one-third of the 16 macaque-specific exons in silico are not transcribed in humans but that the rest of them are expressed in human tissues. However, the fact that we did not find any human cDNAs corresponding to the macaque-specific exons in the public databases suggests that these macaque-type transcripts are very rare in humans. This indicates that human cDNA databases require registration of more transcript variants for one locus to represent the whole transcriptome and that transcriptional resources from other species, especially primates, would help to complement human transcriptome databases.

It might be the case that human brain transcripts are overrepresented in the public databases, making it less likely to miss a splice variant than in other tissues. We further surveyed how many transcripts are registered per locus that we used for the study. The average and median of the number of transcripts per locus are presented in Supplementary Table 5 (Supplementary Material online). Indeed, there are more transcripts per locus in the public database for the brain genes than the testicular genes. However, if this assortment bias affected the finding rate of evolutionarily altered exons, the transcripts with macaque-specific exons should have less homologous human transcripts in the public database than the evolutionarily conserved transcripts. As shown in Supplementary Table 5 (Supplementary Material online), we did not find any systematic trend among them. Hence, the assortment bias would not violate our interpretation.

We estimated that around 10% of macaque transcripts contain macaque-specific exons in the testis. We found that all the six experimentally confirmed macaque-specific exons retain the consensus splicing donor-acceptor site (GT-AG) in the human genome, in spite of the fact that the exonlike blocks were not transcribed in the human tissues. Therefore, it is plausible that the exonlike sequences in the human genome were inactivated during the evolution so that they are not expressed or are expressed in only certain tissues. If we assume that the evolutionary exon alteration is mainly due to the exon loss and the evolutionary rate is the same in human and macaque lineages, the number of evolutionarily altered exons in humans and cynomolgus monkeys would be around 20% of the total number of transcripts in the testis.

We found a slower substitution rate in the 5′UTR than in the synonymous sites, suggesting negative selection acting on the 5′UTR evolution at the genome-wide level. However, it is possible that the K_{5UTR} of some genes might have been increased by positive selection (Hellmann et al. 2003; Kohn, Fang, and Wu 2004). Because the length of the 5′UTR sequence of the gene is sometimes very short and the sampling variance of K_{5UTR} is substantially large, the analysis that calculates K_{5UTR}/K_s for each gene cannot be easily applied to our data set. Another source of evidence inferring the type of natural selection would be a different selection pressure acting on different functional classes of genes. Significantly higher K_a and K_{5UTR} in the cDNAs from the testis than those from the brain were observed in this study (table 1 and fig. 1). There is a great deal of evidence that the K_a of genes expressed in reproductive tissues, especially in the testis, has been increased by positive selection (Wyckoff, Wang, and Wu 2000; Swanson and Vacquier 2002). In our data set, more testicular genes showed a signature of positive selection in the CDS ($K_a/K_s > 1$) than the brain genes (25/622 and 5/302 from the testis and brain, respectively). If we apply the same augment to the 5′UTR, higher K_{5UTR} in the testis than in the brain would be driven by positive selection as well. When a mutation yields a gene expression pattern that is spatially

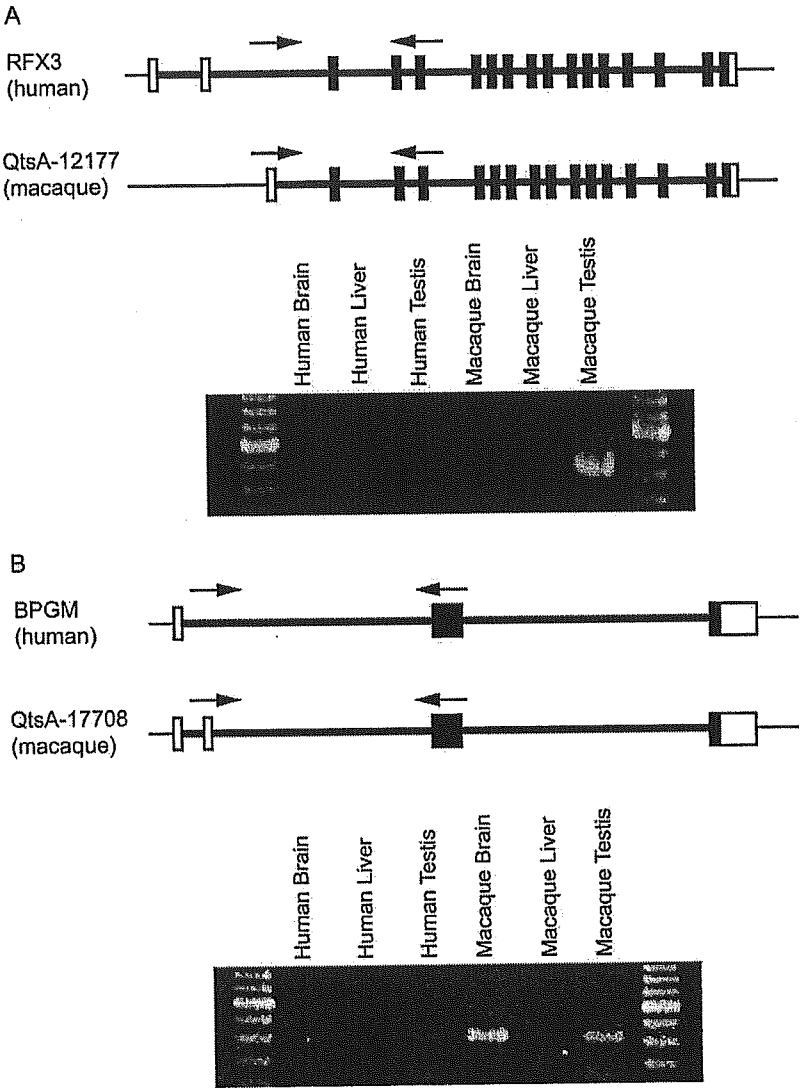

FIG. 2.—Example of a macaque-specific exon confirmed by the RT-PCR experiments. All results and primer sequences are summarized in Supplementary Table 2. Upper panels: open boxes and closed boxes represent the 5′UTR and CDS, respectively. The genes are transcribed from left to right. The primers were designed to match both human and macaque sequences and are indicated by arrows. Lower panels: images of the RT-PCR gels. Tissues from three organs of humans and macaques (brain, liver, and testis) were examined.

and/or temporally beneficial to an organism, the mutation would fix to a population faster than neutral mutations.

We should note that more constraints on brain genes than on genes expressed in other tissues might be the cause of the difference between the substitution rate in the brain and testis (Duret and Mouchroud 2000). Because the brain is a special tissue, especially in primates, whether the evolution of brain genes is fast or slow is still in debate (Dorus et al. 2004). The reference genes expressed in other tissues, such as housekeeping genes, would be useful to access how much of acceleration of evolution in the testis is due to positive selection.

Nucleotide substitution is not the only source of genetic changes in the evolution. We estimated that about 10% of macaque testicular cDNAs have the exons that are not transcribed in humans, and this would have a larger impact on the divergence of gene regulation. The fraction of cDNAs with an exon alteration found by in silico analysis was significantly higher in the testicular cDNAs than in the brain cDNAs. Thus, it is plausible that the 5′UTRs of a certain fraction of genes are under positive selection in terms of both substitution rate and structural alteration.

In this report, we found that both positive and negative selection can act on not only the protein-coding (Fay, Wyckoff, and Wu 2002) and promoter sequences (Kohn, Fang and Wu 2004) but also the 5′UTR sequences. It is worthwhile to study how the 5′UTR divergence affects gene expression and modifies the phenotype of organisms.

Supplementary Material

Supplementary tables 1–5 and supplementary figure 1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This study was supported in part by a Health Science Research grant from the Human Genome Program of the Ministry of Health, Labor and Welfare of Japan. We thank Michael H. Kohn for comments and discussions. We also thank two anonymous reviewers for helpful suggestions.

Literature Cited

Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D. Haussler. 2004. Ultraconserved elements in the human genome. Science **304**:1321–1325.

Bernard, D. J., T. K. Woodruff, and T. M. Plant. 2004. Cloning of a novel inhibin alpha cDNA from rhesus monkey testis. Reprod. Biol. Endocrinol. **2**:71.

Chamas, F., and E. L. Sabban. 2002. Role of the 5′ untranslated region (UTR) in the tissue-specific regulation of rat tryptophan hydroxylase gene expression by stress. J. Neurochem. **82**: 645–654.

Chew, C. H., M. R. Samian, N. Najimudin, and T. S. Tengku Muhammad. 2003. Molecular characterisation of six alternatively spliced variants and a novel promoter in human peroxisome proliferator-activated receptor alpha. Biochem. Biophys. Res. Commun. **305**:235–243.

Dorus, S., E. J. Vallender, P. D. Evans, J. R. Anderson, S. L. Gilbert, M. Mahowald, G. J. Wyckoff, C. M. Malcom, and B. T. Lahn. 2004. Accelerated evolution of nervous system genes in the origin of Homo sapiens. Cell **119**:1027–1040.

Duret, L., and D. Mouchiroud. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. **17**:68–74.

Enard, W., P. Khaitovich, J. Klose et al. (13 co-authors). 2002. Intra- and interspecific variation in primate gene expression patterns. Science **296**:340–343.

Fay, J. C., G. J. Wyckoff, and C. I. Wu. 2002. Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature **415**:1024–1026.

Gellersen, B., R. Kempf, R. Sandhowe, G. F. Weinbauer, and R. Behr. 2002. Novel leader exons of the cyclic adenosine 3′,5′-monophosphate response element modulator (CREM) gene, transcribed from promoters P3 and P4, are highly testis-specific in primates. Mol. Hum. Reprod. **8**:965–976.

Hellmann, I., S. Zollner, W. Enard, I. Ebersberger, B. Nickel, and S. Paabo. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. Genome Res. **13**:831–837.

Kawamura, S., H. Tanabe, Y. Watanabe, K. Kurosaki, N. Saitou, and S. Ueda. 1991. Evolutionary rate of immunoglobulin alpha noncoding region is greater in hominoids than in Old World monkeys. Mol. Biol. Evol. **8**:743–752.

Kent, W. J. 2002. BLAT—the BLAST-like alignment tool. Genome Res. **12**:656–664.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

King, M. C., and A. C. Wilson. 1975. Evolution at two levels in humans and chimpanzees. Science **188**:107–116.

Kohn, M. H., S. Fang, and C. I. Wu. 2004. Inference of positive and negative selection on the 5′ regulatory regions of Drosophila genes. Mol. Biol. Evol. **21**:374–383.

Lammich, S., S. Schobel, A. K. Zimmer, S. F. Lichtenthaler, and C. Haass. 2004. Expression of the Alzheimer protease BACE1 is suppressed via its 5′-untranslated region. EMBO Rep. **5**:620–625.

Li, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36**:96–99.

———. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.

Makalowski, W., and M. S. Boguski. 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc. Natl. Acad. Sci. USA **95**:9407–9412.

Mao, J., S. S. Chirala, and S. J. Wakil. 2003. Human acetyl-CoA carboxylase 1 gene: presence of three promoters and heterogeneity at the 5′-untranslated mRNA region. Proc. Natl. Acad. Sci. USA **100**:7515–7520.

Mesak, F. M., N. Osada, K. Hashimoto, Q. Y. Liu, and C. E. Ng. 2003. Molecular cloning, genomic characterization and overexpression of a novel gene, XRRA1, identified from human colorectal cancer cell HCT116Clone2_XRR and macaque testis. BMC Genomics **4**:32.

Mezquita, P., C. Mezquita, and J. Mezquita. 1999. Novel transcripts of carbonic anhydrase II in mouse and human testis. Mol. Hum. Reprod. **5**:199–205.

Miyata, T., T. Yasunaga, and T. Nishida. 1980. Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc. Natl. Acad. Sci. USA **77**:7328–7332.

Modrek, B., and C. J. Lee. 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat. Genet. **34**: 177–180.

Newton, D. C., S. C. Bevan, S. Choi, G. B. Robb, A. Millar, Y. Wang, and P. A. Marsden. 2003. Translational regulation of human neuronal nitric-oxide synthase by an alternatively spliced 5′-untranslated region leader exon. J. Biol. Chem. **278**:636–644.

Nurtdinov, R. N., I. I. Artamonova, A. A. Mironov, and M. S. Gelfand. 2003. Low conservation of alternative splicing patterns in the human and mouse genomes. Hum. Mol. Genet. **12**:1313–1320.

Osada, N., M. Hida, J. Kusuda, R. Tanuma, M. Hirata, M. Hirai, K. Terao, Y. Suzuki, S. Sugano, and K. Hashimoto. 2002*a*. Prediction of unidentified human genes on the basis of sequence similarity to novel cDNAs from cynomolgus monkey brain. Genome Biol. **3**:RESEARCH0006.

Osada, N., M. Hida, J. Kusuda, R. Tanuma, M. Hirata, Y. Suto, M. Hirai, K. Terao, S. Sugano, and K. Hashimoto. 2002*b*. Cynomolgus monkey testicular cDNAs for discovery of novel human genes in the human genome sequence. BMC Genomics **3**:36.

Osada, N., M. Hida, J. Kusuda et al. (12 co-authors). 2001. Assignment of 118 novel cDNAs of cynomolgus monkey brain to human chromosomes. Gene **275**:31–37.

Osada, N., J. Kusuda, M. Hirata, R. Tanuma, M. Hida, S. Sugano, M. Hirai, and K. Hashimoto. 2002*c*. Search for genes positively selected during primate evolution by 5′-end-sequence screening of cynomolgus monkey cDNAs. Genomics **79**:657–662.

Pamilo, P., and N. O. Bianchi. 1993. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol. Biol. Evol. **10**:271–281.

Pan, Q., M. A. Bakowski, Q. Morris, W. Zhang, B. J. Frey, T. R. Hughes, and B. J. Blencowe. 2005. Alternative splicing of conserved exons is frequently species-specific in human and mouse. Trends Genet. **21**:73–77.

Pruitt, K. D., T. Tatusova, and D. R. Maglott. 2003. NCBI Reference Sequence project: update and current status. Nucleic Acids Res. **31**:34–37.

Savatier, P., G. Trabuchet, Y. Chebloune, C. Faure, G. Verdier, and V. M. Nigon. 1987. Nucleotide sequence of the delta-beta-globin intergenic segment in the macaque: structure and evolutionary rates in higher primates. J. Mol. Evol. **24**:297–308.

Sugiura, S., S. Kashiwabara, S. Iwase, and T. Baba. 2003. Expression of a testis-specific form of TBP-related factor 2 (TRF2) mRNA during mouse spermatogenesis. J. Reprod. Dev. **49**:107–111.

Suzuki, Y., R. Yamashita, M. Shirota, Y. Sakakibara, J. Chiba, J. Mizushima Sugano, K. Nakai, and S. Sugano. 2004. Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions. Genome Res. **14**:1711–1718.

Suzuki, Y., K. Yoshitomo Nakagawa, K. Maruyama, A. Suyama, and S. Sugano. 1997. Construction and characterization of a full length-enriched and a 5′-end-enriched cDNA library. Gene **200**:149–156.

Swanson, W. J., and V. D. Vacquier. 2002. The rapid evolution of reproductive proteins. Nat. Rev. Genet. **3**:137–144.

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

Wang, H. Y., H. Tang, C. K. Shen, and C. I. Wu. 2003. Rapidly evolving genes in human. I. The glycophorins and their possible role in evading malaria parasites. Mol. Biol. Evol. **20**:1795–1804.

Wang, Y., D. C. Newton, G. B. Robb, C. L. Kau, T. L. Miller, A. H. Cheung, A. V. Hall, S. VanDamme, J. N. Wilcox, and P. A. Marsden. 1999. RNA diversity has profound effects on the translation of neuronal nitric oxide synthase. Proc. Natl. Acad. Sci. USA **96**:12150–12155.

Watanabe, H., A. Fujiyama, M. Hattori et al. 2004. DNA sequence and comparative analysis of chimpanzee chromosome 22. Nature **429**:382–388.

Wyckoff, G. J., W. Wang, and C. I. Wu. 2000. Rapid evolution of male reproductive genes in the descent of man. Nature **403**:304–309.

Yeo, G., D. Holste, G. Kreiman, and C. B. Burge. 2004. Variation in alternative splicing across human tissues. Genome Biol. **5**:R74.

Jianzhi Zhang, Associate Editor

Accepted May 30, 2005

D86–D89 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj129

DBTSS: DataBase of Human Transcription Start Sites, progress report 2006

Riu Yamashita, Yutaka Suzuki[1,*], Hiroyuki Wakaguri[1], Katsuki Tsuritani, Kenta Nakai and Sumio Sugano[1]

Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan and [1]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa, Chiba 277-8562, Japan

Received September 15, 2005; Revised and Accepted October 21, 2005

ABSTRACT

DBTSS was first constructed in 2002 based on precise, experimentally determined 5′ end clones. Several major updates and additions have been made since the last report. First, the number of human clones has drastically increased, going from 190 964 to 1 359 000. Second, information about potential alternative promoters is presented because the number of 5′ end clones is now sufficient to determine several promoters for one gene. Namely, we defined putative promoter groups by clustering transcription start sites (TSSs) separated by <500 bases. A total of 8308 human genes and 4276 mouse genes were found to have putative multiple promoters. Third, DBTSS provides detailed sequence comparisons of user-specified TSSs. Finally, we have added TSS information for zebrafish, malaria and schyzon (a red algae model organism). DBTSS is accessible at http://dbtss.hgc.jp.

INTRODUCTION

Recently, a huge amount of comprehensive expression profile data obtained by various experiments, such as microarrays, has been made available. It is a challenging problem to uncover the regulatory networks among the expressed genes from these data. Information about promoters, which contain most of the binding sites of transcription factors, is indispensable for solving this question. To define promoter regions, precise information about transcription start sites (TSSs) is also required. Such data, however, are not easily obtained because the cDNA sequence data in repository sequence databases provide no guarantees regarding the 5′ end of the sequences and because the computational prediction of promoters and TSSs still remains problematic (1). To overcome these difficulties several databases (2), including DBTSS (DataBase of Transcription Start Sites) have been constructed. DBTSS contains TSS information of genes based on specific experiments (3,4). Clones constructed by full-length cDNA methods such as oligo-capping (5,6) or CAP-trapper (7,8) are mapped on to genome sequences to determine TSSs. Each TSS is determined based on the 5′ end of the corresponding clone. DBTSS was first constructed in 2002, and has been improved by several major and minor updates. The original version (version 1) contained only human data (3). Two years later, we reported the addition of mouse TSS information (9) in version 3 (4). Here we introduce the new updates and additions since version 3, the most important one being the addition of putative alternative promoter information.

NEW FEATURES

The current version of DBTSS, version 5, includes some notable improvements since the previous report, in addition to minor updates such as modifications of the interface and the result views.

One major improvement is that the amount of data for human TSSs has been significantly increased: in our report in 2002, we described 190 964 human clones which corresponded to 11 234 NCBI reference sequence cDNAs (RefSeq) (4). Because we added data from a new full-length cDNA project (10), DBTSS now contains 1 359 000 clones corresponding to 19 753 RefSeq cDNAs (Table 1). Since RefSeq cDNAs contain splicing variants as separate entries, we performed clustering of clones' information depending on their coordinate in the genome sequence; if their sequences overlapped, we regard them as the same locus. After clustering, our data correspond to 15 262 genes (Table 1). This is one of the largest collections of human 5′ end cDNA sequences.

To check the quality of our TSS data, we compared DBTSS with the Eukaryote Promoter Database (EPD) (2). In EPD Release 82, there are 1871 promoters collected from the literature. Among them, we could map 1767 promoter

*To whom correspondence should be addressed. Tel: +81 4 7136 3607; Fax: +81 4 7136 3607; Email: ysuzuki@hgc.Jp

sequences to the human genome; 1639 of them mapping within 100 bases of the DBTSS TSSs, indicating that the data in DBTSS are consistent with the data obtained from ordinary methods.

In the next two sections, we will discuss two other major updates: alternative promoters (APs) and promoter comparison.

ALTERNATIVE PROMOTERS

Several genes are known to have multiple promoters which could be regulated in a different manner. These promoters, labeled as APs, could be useful to maximally exploit the relatively limited number of genes in the genome (11). However, no estimation of how many genes might have alternative promotes is available to date. Since DBTSS now has enough 5′ end clones from human and mouse, we performed this estimate. This is the most important addition in version 5. Although the details of our analysis will be reported elsewhere (12), the procedure is summarized below.

To determine APs, we first collected all the TSSs from the same locus. TSSs located inside a RefSeq gene exon, with the exception of the first one, were removed in order to avoid artifacts caused by truncated 5′ ends. We used several intervals to define AP clusters. The distribution of the number of putative alternative promoter containing genes shows a plateau before the interval size reaches 500 bp (12). We, therefore, clustered the clones using a 500 base interval, and defined each cluster as an promoter. We obtained 30 964 promoters, and 26 784 (86.5%) of them are within 500 bp. According to this procedure, 6954 human loci and 9886 mouse loci have only one promoter while 8308 human loci and 4276 mouse loci have two promoters or more. Figure 1A shows the three alternative promoters found in the gene encoding human A kinase

Table 1. Statistics of DBTSS

	No. of genes/ no. of RefSeq	No. of promoters	No. of TSSs	No. of clones
Human	15 262/19 753	30 964	452 117	1 359 000
Mouse	14 162/14 746	19 023	149 876	364 487
Zebrafish	3061/3075	3382	15 198	32 263
Malaria	1527/NA	NA	6908	10 236
Schyzon	3635/NA	NA	14 029	22 923

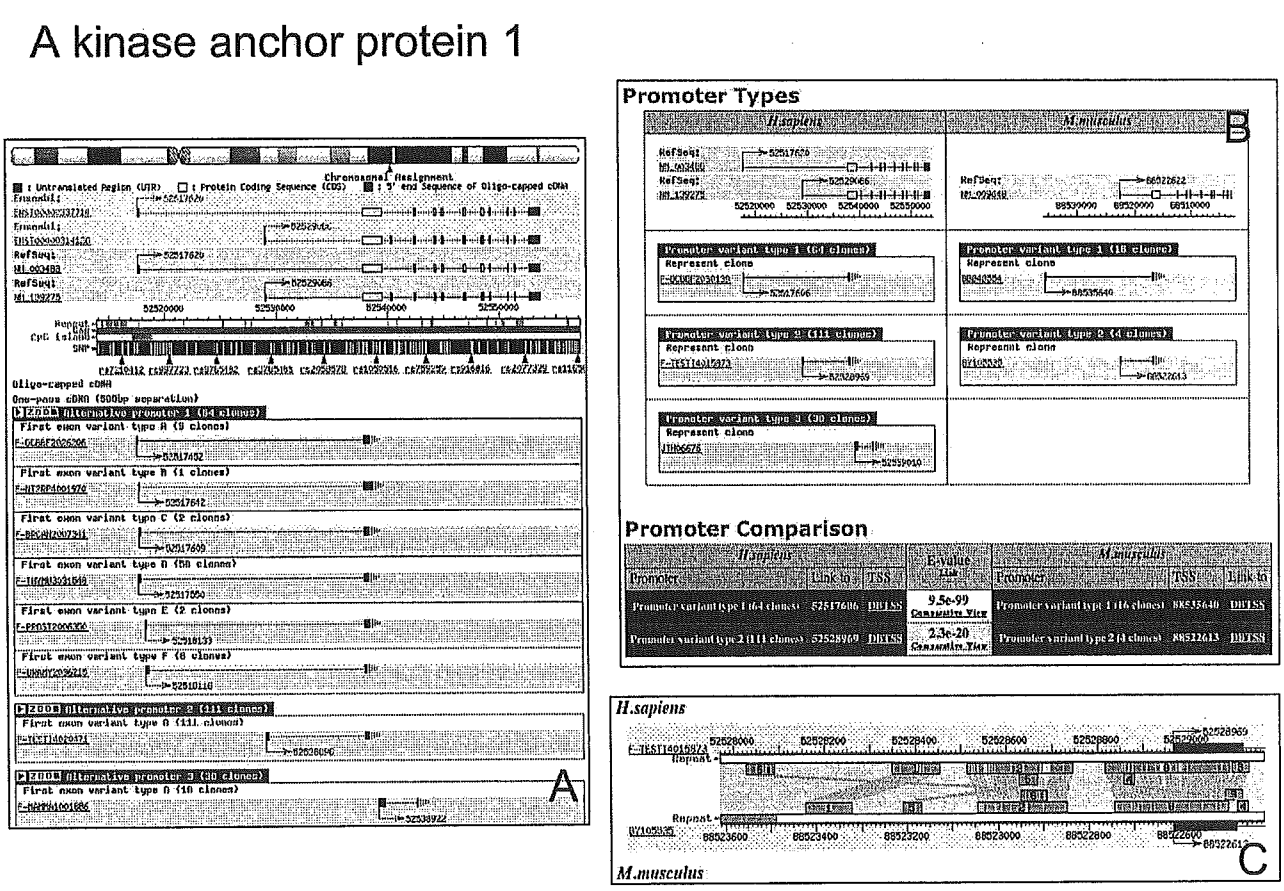

Figure 1. An example of alternative promoter view. Here we show AKAP1 (NM_003488) as an example. (**A**) The putative promoter clusters are given using different colors; therefore, there are three putative promoters in human AKAP1. We observed several patterns of first exon in promoter type 1, so we clustered them and show them as 'First exon variant type A–F'. (**B**) Comparative analysis between human and mouse alternative promoters. There are two putative promoters in mouse. The best match between two promoters is available in 'Promoter Comparison'. (**C**) Clicking the 'Comparative View', the user can obtain the alignment between these promoters.

anchor protein 1 (AKAP1). It is notable that DBTSS also provides comparative information between human and mouse promoters. Figure 1B shows an example of comparative promoter analysis between orthologous genes. Two promoters were identified for the mouse gene for AKAP1. From this view, the representative APs are also available for alignment. By clicking 'Comparative View' in 'Promoter Comparison' in Figure 1B, the LALIGN-based alignment view, shown in Figure 1C is obtained.

COMPARATIVE PROMOTER ANALYSIS

In the previous section, we showed an example of alternative promoter comparison between human and mouse. Before version 5 of DBTSS, these were precomputed, and the user could only obtain alignments between orthologous human and mouse genes. Despite being a useful idea, this sometimes failed to answer the user's need for alignments of arbitrary promoter pairs, for instance, promoters of paralogous genes. We therefore implement a dynamic viewer allowing the alignment of any two TSSs present in DBTSS. Such analyses are necessary to understand how transcriptional regulatory elements were conserved or diverged during gene and exon duplication. For example, in Figure 2A, the clones TST01431 of protamine 1 (PRM1: NM_002761) and TST00906 of protamine 2 (PRM2: NM_002762) are selected for alignment. Both genes are expressed in testis and are paralogous to each other. PRM1 is found in nearly all mammals while PRM2 is observed in relatively few mammals including human and mouse (13). In human, both genes are on chromosome 16, separated by ~5 kb (14). The obtained alignment and the determined conserved regions are shown in Figure 2B. In this case, the blocks '0' and '7' are highly conserved. The details of the alignment of both TSS regions are also available, as shown in Figure 2C. Especially, it is noteworthy that the putative TATA-box is inside block '7' for the PRM1 promoter and outside of it for the PRM2 promoters (15).

FUTURE PERSPECTIVE

As shown in Table 1, we have added data from 32 263 zebrafish (*Danio rerio*) (16), 10 236 malaria (*Plasmodium*

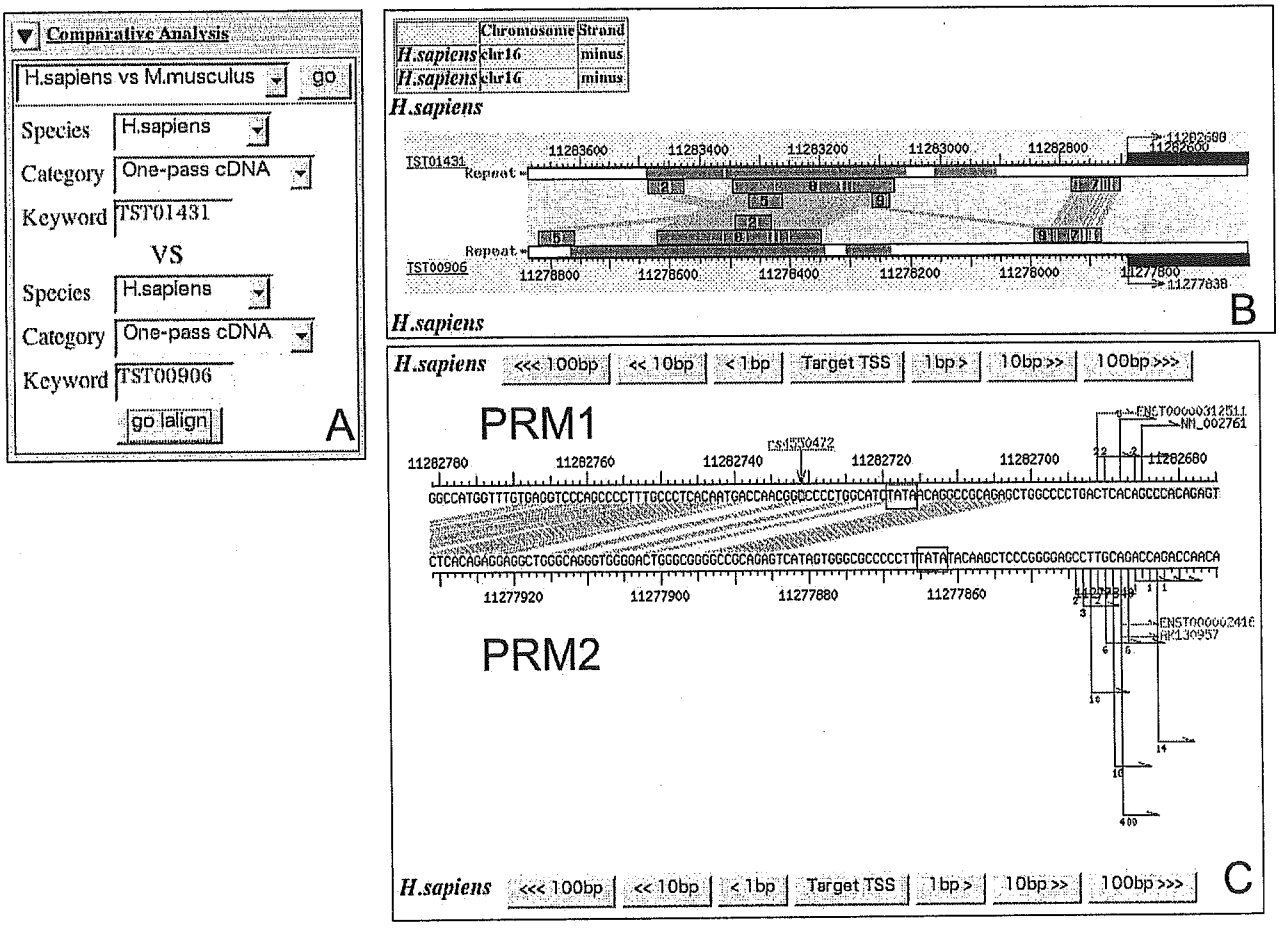

Figure 2. An example of comparative analysis with any pair of TSSs. We show paralogous genes, protamine 1 (PRM1: NM_002761) and protamine 2 (PRM2: NM_002762), as an example. (**A**) By inputting the IDs of clones of PRM1 (TST01431) and PRM2 (TST00906) representative TSSs, users can obtain the results (B and C). (**B**) LALIGN analysis between two sequences. *Note*: smaller numbers indicate more highly conserved blocks. In this figure, the most conserved region between a pair is block 0; however, it includes *Alu* repeats. (**C**) The detail of the alignment of block 7. The putative TATA-boxes are marked with boxes.