

Rapid evolution of primate *ESX1*, an X-linked placenta- and testis-expressed homeobox gene

Xiaoxia Wang and Jianzhi Zhang*

Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA

Received April 13, 2007; Revised June 1, 2007; Accepted June 14, 2007 DNA sequences reported in this paper have been submitted to GenBank (accession nos EF650070–EF650084 and EF695414–EF695445)

Homeobox genes encode transcription factors that play important roles in various developmental processes and are usually evolutionarily conserved. Here we report a case of rapid evolution of a homeobox gene in humans and non-human primates. *ESX1* is an X-linked homeobox gene primarily expressed in the placenta and testis, with physiological functions in placenta/fetus development and spermatogenesis. *ESX1* is paternally imprinted in mice, but is not imprinted in humans. We provide evidence for a significantly higher non-synonymous substitution rate than synonymous rate in *ESX1* between humans and chimps as well as among a total of 15 primate species. Population genetic data also show signals of recent selective sweeps within humans. Positive selection appears to be concentrated in the C-terminal non-homeodomain region, which has been implicated in regulating human male germ cell division by prohibiting the degradation of cyclins. In contrast, mouse *Esx1* has a substantively different C-terminal region subject to strong purifying selection. These and other results suggest that even the fundamental process of spermatogenesis has been targeted by positive selection in primate and human evolution and that mouse may not be a suitable model for studying human reproduction.

INTRODUCTION

Homeobox genes are characterized by the presence of a sequence motif known as the homeobox, which encodes the ~60-amino-acid homeodomain, a helix-turn-helix DNA binding domain (1). In humans, there are about 230 homeobox genes (2), encoding a large family of transcription factors that play key roles in various developmental processes such as body-plan specification, pattern formation and cell-fate determination (1). Due to their functional importance, most homeodomain proteins are evolutionarily highly conserved in sequence (1,3,4). Hence, the identification of non-conserved homeobox genes would be particularly interesting, because such homeobox genes may regulate important developmental processes that vary among relatively closely related species. Three such rapidly evolving homeobox genes are known, from fruit flies (*OdsH*), rodents (*Rhox5*) and primates (*TGIFLX*), respectively. *OdsH* is an X-linked gene involved in spermatogenesis and it is partly responsible for the hybrid male sterility between *Drosophila simulans* and *D. mauritiana* (5). Mouse *Rhox5* (also known as *Pem*) is expressed in both male and female reproductive tissues (6). Targeted disruption

of *Rhox5* increases male germ cell apoptosis and reduces sperm production, sperm motility and fertility (7). In fact, *Rhox5* is just one member of a recently expanded homeobox gene cluster known as the *Rhox* cluster on the mouse X chromosome (7–10). Several other members of the cluster are also expressed in reproductive tissues (7) and evolve rapidly (9,11). *TGIFLX* is a retroduplicate formed in the common ancestor of primates and rodents by retroposition of the autosomal gene *TGIF2* to the X chromosome, and is specifically expressed in the germ cells of adult testis (12). Interestingly, each of the three cases involves a homeobox gene that is X-linked and testis-expressed. Here we report yet another case of rapid evolution of an X-linked testis-expressed homeobox gene, *ESX1*.

Human *ESX1*, also known as *ESX1L* and *ESXR1*, is a paired-like homeobox gene located on Xq22.1 (13). *ESX1* protein contains two functional domains, the homeodomain and the proline-rich domain (Fig. 1A) (13). *Esx1*, the mouse ortholog, has an extra domain known as the PN/PF motif, located at the C-terminus (Fig. 1B) (14). In humans, *ESX1* is specifically expressed in placenta from 5 weeks of gestation until term (15) and in adult testis (13). A recent study shows

*To whom correspondence should be addressed at: Department of Ecology and Evolutionary Biology, University of Michigan, 1075 Natural Science Building, 830 North University Avenue, Ann Arbor, MI 48109, USA. Tel: +1 7347630527; Fax: +1 7347630544; Email: jianzhi@umich.edu

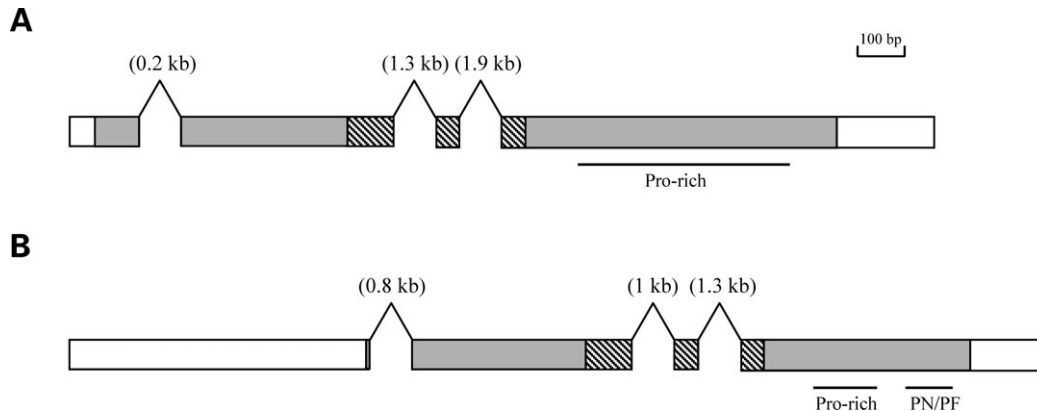


Figure 1. Structures of the orthologous (A) human *ESX1* and (B) mouse *Esx1* genes, adapted from Fohn and Behringer (13) and Li *et al.* (17). Exons are boxed, with coding regions shown in grey and homeobox shown by hatches. The approximate length of each intron is given in parentheses. Pro-rich and PN/PF motifs are indicated underneath the gene structure.

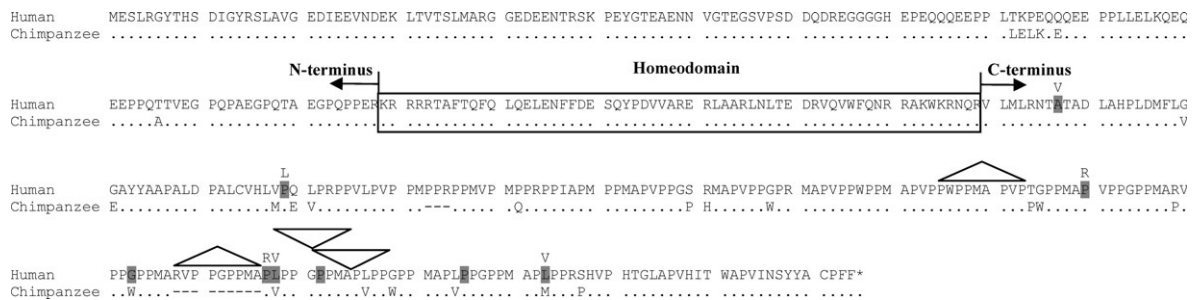


Figure 2. Alignment of human and chimpanzee *ESX1* protein sequences. The human sequence is from GenBank (accession number AY114148). The homeodomain is boxed. SNPs detected in humans are shaded. For each non-synonymous SNP, the alternative amino acid is shown above the human sequence. For each synonymous SNP, no alternative amino acid is shown. Triangles indicate indel polymorphisms observed in humans, with deletions shown by triangles pointing upwards and insertions shown by triangles pointing downwards. The width of the triangle shows the size of the indel. '.' indicates identity to the human sequence and '-' indicates a gap.

decreased *ESX1* expression in human pre-term idiopathic fetal growth restriction, a clinically significant pregnancy disorder in which the fetus fails to achieve its full growth potential *in utero* (16). In mice, *Esx1* is also expressed in placenta and testis (17,18). More specifically, during embryogenesis, it is expressed in the extraembryonic tissues, including the endoderm of the visceral yolk sac, the ectoderm of the chorion and subsequently the labyrinthine trophoblast of the chorio-allantoic placenta (17). In adults, *Esx1* is expressed in male germ cells only, particularly the spermatogonia/preleptotene spermatocytes and round spermatids of spermatogenic stages IV–VII (17,18). These restricted temporal and spatial expression patterns suggest that *ESX1/Esx1* is involved in placental development and spermatogenesis. Mouse *Esx1* is paternally imprinted in the placenta, with only the maternally derived allele expressed (19). Heterozygous female mice inheriting a null *Esx1* allele from their mother are born 20% smaller than normal, suggesting that *Esx1* is required for placental development and fetal growth in mice (19). In contrast, biparental expression of *ESX1* is found in human placenta (20).

Our preliminary comparison between human *ESX1* and mouse *Esx1* proteins showed an unexpectedly high level of sequence divergence (34%), suggesting that the gene might be evolving rapidly in primates and/or rodents as a result of positive Darwinian selection (12). Below we first describe

the evolutionary pattern of *ESX1* in primates and then compare it to the evolutionary pattern in rodents. We show that positive selection has acted on *ESX1* within humans, between humans and chimpanzees, and among a large array of primate species, whereas purifying selection has dominated *Esx1* evolution in rodents. We discuss these evolutionary patterns in light of the structure and function of the gene.

RESULTS

Comparison of *ESX1* sequences between humans and chimps and within humans

We obtained the *ESX1* gene sequence from the chimpanzee genome sequence (<http://genome.wustl.edu/>) and compared it with the human *ESX1* sequence available in GenBank (AY114148). The alignment shows a high level of sequence divergence. Of the aligned 406 amino acid sites, there are 25 amino acid replacements, in addition to two gaps totaling 12 amino acids (Fig. 2). A comparison of synonymous (d_S) and non-synonymous (d_N) nucleotide distances between gene sequences can inform us about the nature and strength of selection acting on a gene. A higher d_N than d_S indicates positive selection, whereas a lower d_N than d_S indicates negative or purifying selection. The vast majority of genes in the

human genome are under negative selection, with a genomic average d_N/d_S ratio of 0.26 (21). In *ESX1*, however, d_N (0.031) is significantly greater than d_S (0.009) [$P = 0.028$; Fisher's exact test (22)]. Because the d_S value of *ESX1* is not significantly different from the genomic average d_S of 0.012 (23), the above observation strongly suggests that positive selection has promoted non-synonymous substitutions in *ESX1* during the divergence between humans and chimps.

To identify the regions where positive selection has been operating, we divided the *ESX1* protein sequence into three segments, the N-terminus, homeodomain and C-terminus (Fig. 2). The homeodomain is completely identical in amino acid sequence between human and chimp and thus has not been targeted by positive selection ($d_N = 0$, $d_S = 0.019$, $P = 0.29$; Fisher's exact test). In the N-terminus, d_N (0.024) and d_S (0) are not significantly different ($P = 0.1$) and hence neutrality cannot be rejected. In the C-terminus, however, d_N (0.047) is significantly greater than d_S (0.012) ($P = 0.035$). Thus, positive selection has been concentrated in the C-terminus. As aforementioned, the C-terminus is mainly composed of proline-rich repeats. The two alignment gaps between human and chimp also occur in the C-terminus (Fig. 2). Compared to the human sequence, the chimp sequence lost a complete nine-amino-acid repeat and part of another repeat.

To further examine whether the positive selection might have happened in the recent history of human evolution, we sequenced exon 4 of *ESX1* in 32 unrelated male humans of diverse geographic origins (4 Pygmy Africans, 6 African Americans, 12 Caucasians, 3 Southeast Asians, 2 Chinese, 2 Pacific Islanders, and 3 Andes Indians). Exon 4 encodes 14 amino acids of the homeodomain, corresponding roughly to the third helix of the homeodomain, and the complete C-terminus of *ESX1*, the likely target of positive selection (Figs 1 and 2). From the 32 alleles, we observed four insertion/deletion (indel) polymorphisms and nine single nucleotide polymorphisms (SNPs). All of these polymorphisms occur in the C-terminus (non-homeodomain) region and none of them disrupt the open reading frame (Fig. 2). Supplementary Material, Table S1 lists the polymorphisms and their associated allele frequencies. The polymorphic data allow us to compute the level of DNA polymorphism in exon 4. Nucleotide diversity per sequence (π) is 0.897 and Watterson's polymorphism per sequence (θ) is 1.27. A comparison between expected and observed distributions of allelic frequencies can tell us whether a genomic region is likely to have been subject to recent selective sweeps, which render π lower than θ and high-frequency alleles enriched, generating negative values of Tajima's D (24) and Fay and Wu's H (25). Combining the information from D and H , Zeng and colleagues recently invented a new test known as the DH test of positive selection (26). This test is superior to the individual D and H tests because it is more powerful and is insensitive to common confounding factors such as background selection, population growth and population subdivision (26). We found that the DH test rejects the neutral hypothesis for the exon 4 sequences of 32 humans ($P < 0.039$). For samples with African, Caucasian and Asian origins, the tail probability of the DH test is 0.067, 0.31 and 0.26, respectively. Thus, selective sweeps might have occurred among Africans. Consistent with this result, the H test also yields a significant result for

the African samples ($P = 0.044$), and this P -value is lower than 128 of the 132 genes that were recently surveyed in Africans (27). In other words, *ESX1* is among the bottom 3% of human genes for H value in Africans. These results support the hypothesis of recent selective sweeps at human *ESX1* or linked genomic regions. We also sequenced exons 1, 2 and 3 of *ESX1* in eight male humans with diverse geographic origins (one Pygmy African, three African Americans, one Caucasian, two Chinese and one Pacific Islander), but observed no polymorphisms.

Positive selection at the C-terminus of *ESX1* in many primates

To examine whether *ESX1* has also been under positive selection in other primates, we obtained the rhesus monkey *ESX1* gene sequence by searching its recently completed draft genome sequence (<http://www.ncbi.nlm.nih.gov/>). We then sequenced exon 4 of *ESX1* in 12 additional primate species, including three hominoids, four Old World monkeys and five New World monkeys (see Materials and Methods). Together with the three known sequences from human, chimp and rhesus, a total of 15 primate sequences of exon 4 were conceptually translated and aligned by Clustal W with manual adjustment. DNA sequences were subsequently aligned by following the protein alignment (Fig. 3A). A gene tree of the 15 sequences was reconstructed using the neighbor-joining method (28). The tree topology is consistent with the known species tree, suggesting that the sequences analyzed are orthologous to each other. We found that the length of exon 4 is highly variable among species. The shortest proline-rich repeat region is found in marmoset and tamarin, whereas the longest is observed in orangutan. The nine-amino-acid repeat unit has variable sequences among the primate species, although proline is always the most frequent amino acid.

To examine the potential action of positive selection in exon 4 of primate *ESX1*, we computed pairwise d_N and d_S among the 15 sequences. Excluding alignment gaps, we analyzed a total of 312 nucleotide sites. Higher d_N than d_S is observed in 79 (75.2%) of 105 pairwise comparisons (Fig. 4A). There is no apparent difference in this pattern among hominoids, Old World monkeys and New World monkeys. When only the (non-homeodomain) C-terminus is analyzed, 86 (81.9%) comparisons showed $d_N > d_S$ (Fig. 4B). In our dataset, the average d_S is 0.12 between hominoids and Old World monkeys, 0.18 between hominoids and New World monkeys and 0.16 between Old World monkeys and New World monkeys. All three numbers are greater than the corresponding values (0.08, 0.12 and 0.15, respectively) previously estimated from multiple different intron and non-coding sequences of the same species pairs (29). Thus, the synonymous substitution rate of *ESX1* is not reduced and the overall higher d_N over d_S suggests positive selection.

Due to the occurrence of many indels in the proline-rich region of primate *ESX1*, the sequence alignment may not be reliable. Because closely related species are more likely to share the same repeat sequence, which facilitates alignment, we made separate alignments for hominoids, Old World monkeys and New World monkeys, respectively (Supplementary Material, Figure S1). The d_N and d_S values were

A

| | | Homeodomain | | | | | | | | | | | | | | | | | | | |
|-----------------|---------------|-------------------------|-------------------|---------------------|------------------------|---------------------|-------------------|----------------|------------------|-------------------|-------------------|-----------------|---------------------|----------------------|----------------|-----------------------|-----------------|-----------------|--------------|--------------|--|
| OW monkeys | Human | VWFQNRRAKW | KRNQRVLMRL | NTATADLAHP | LDMFLGGAYY | AAPALDPALC | VHLVPQLPRP | PVLVPVPPMP | RPPMVPMPFR | PPIAPMPF-- | -MAPVPPGSR | MAPVPPGPRM | APVPPVPPMA | | | | | | | | |
| | Chimpanzee | | | |VE..... | |M.EV..... | |Q..... | |PH..... | | | | | | | | | | |
| | Gorilla | | |A..... |S.....N..... | | |A..... |Q..... | |P..... | | |G..... | | | | | | | |
| | Orangutan | | |I.A.T.R.S..... |N..... | | |L..... |Q..... |M..... |GP..... |P..... | |G.R.V..... | | | | | | | |
| | Gibbon | | |I.A.A..... |E.T.....P.D..... |F..... | |M.EI..... |S..... |G..... |Q.G..... |M..... |GP..... |P.C..... |M..... |G..V..... | | | | | |
| | Rhesus monkey | | |I.A.A.R.S..... |A.EV.....P.N..... |T.S..... | | | |Q..... |M..... |GP..... |P..... |PP..... |R.P..... |V.MQP..... | | | | | |
| | Baboon | | |I.A.A.R.S..... |T.EV.....P.N..... |T.S..... | | | |Q..... |M..... | | |VPP..... |R.P..... |M.V.MQP..... | | | | | |
| | Green monkey | | |I.A.A.R.S..... |T.EV.....P.N..... |T.S..... | | | |Q..... |M..... |GP..... |P..... |PRP.MA.VPP..... |R.P..... |M.V.MQP..... | | | | | |
| | Talapoia | | |I.A.A.R.S..... |A.EV.....P.N..... |T.S..... | | | |Q..... |M..... |R..... | |P.MA.VPP..... |R.P..... |V.MQP..... | | | | | |
| | Langur | | |I.A.AV.P..... |A.EV.....P.N..... |T.S..... | |M..... |T..... | | | | |G..... |M..... |G--PSMV.MPP..... |R.P..... |V.MQP..... | | | |
| | Marmoset | |R.....M..... |V.ALA.PA |VE.I..AP.D |V.V.....W..... |A.RP..... |R.VA..... |V.HTA..... |AG..... |M..... |VG..... | |P.MA.MPP..... |G.P..... |V..... | | | | | |
| | Tamarin | |R.....M..... |V.ALA.PA |VE.I..APHD |V.V.....W..... |A.RP..... |R.VA..... |V.HA..... |AG..... |M..... |VG..... | |P.MA.MPP..... |G.P..... |V..... | | | | | |
| Owl monkey | |T.....R.....M..... |M.DDA.PPA |VEVI.DMP.D |V.V.....W..... |A.PLG..... |G.VV..... |M..... |A.AQA..... |M.G..... |AGP..... |PVV.M..... |AP..... |M..... |M..... | | | | | | |
| Squirrel monkey | |R.....M..... |M.A.P.VP |VEVI..AP.D |V.V.....W..... |N.A..P..... |R.V..... |MQ..... |A.QA..... |M.G..... |ARP..... |P.P.M..... |EQP..... |M..... | | | | | | | |
| Woolly monkey | |R.....M..... |L.A.A.PA |VEVI..AP.D |V.V.....W..... | |A..... | | | | |RQ..... |VL.A..QAG..... |G..... |AGP..... |P..... |M..... |PP..... |M..... |PA..... | |
| | | Human | EVPFPPMAP | VPTGPPMAFV | PPGPPMARVP | PGPPMARVPP | GPPMAPLPPG | PPMAPLPP-- | -GPPMAPLPP | GPPMAPLPPR | SHVP-HTGLA | PVHITWAPVI | NSYYACPF* | | | | | | | | |
| | | Chimpanzee | |PW..... |P..... |W.....P..... |V.....W..... | |-V..... |M..... |P..... | | | | | | | | | | |
| | | Gorilla | |P..... |RM..... | | | |-V..... |VA.G..... |P..... |T..... |R..... | | | | | | | | |
| | | Orangutan | |G..... |R.M.P..... |M..... |L..... |G..... |G.MA.L..... |V..... |M..... |PR..... |P..... |R..... |G..... | | | | | | |
| | | Gibbon | |H.M..... |G..... |VH.M.P..... |VHM..... | | |V..... |LV.M..... |P..... | |R..... |G..... | | | | | | |
| | | Rhesus monkey | |-R..... |VRM..... |R..... |P..... |VPM..... |R..... |V.....R..... | |M..... |PP..... |RI..... |R..... |G..... | | | | | |
| | | Baboon | |-R..... |VRM..... |R..... |P..... |T..... |VPM..... |R..... |V.....R..... | |M..... |PP..... |RI..... |R..... |G..... | | | | |
| | | Green monkey | |-R..... |VRM..... |R..... |P..... | |VPM..... |R..... |V.....R..... | |M..... |PP..... |RI..... |R..... |G..... | | | | |
| | | Talapoia | |-R..... |VRM..... |R..... |P..... | |VPM..... |R..... |V.....R..... | |M..... |PP..... |RI..... |R..... |G..... | | | | |
| | | Langur | |MVHM..... | |P..... |S.....M..... |R..... |VVPMQ..... |R..... |VHM..... |R..... |V.....R..... | |M..... |PP..... |RI..... |R..... |TG..... | | |
| | | Marmoset | | | | | | | | | |M..... |G.APM..... |F..... |G.A..... |FN..... |VG..... | | | | |
| | | Tamarin | | | | | | | | | |M..... |G.APM..... |F..... |G.A..... |FN..... |VG..... | | | | |
| | | Owl monkey | |-R..... |TQAR..... |RM..... |QAR..... |GM..... |A..... |VVPM..... |A..... |M..... | |VV..... | | | | | | | |
| | | Squirrel monkey | | | | | | | | | |M..... |APM..... |DI.Q..... |LG.A..... |FN..... |G.HVG..... |Y..... | | | |
| | | Woolly monkey | | | | | | | | | |M..... |APM..... |HI..... |G.A..... |FN..... |G.VR..... | | | | |

B

| | | Homeodomain | | | | | | | | | | | | | | | | | |
|----------------|--|-------------|------------|-------------|-----------|------------|-----------|----------|------------|-------------|---------------|-----------|-------------|--|--|--|--|--|--|
| Mus musculus | | VWFQNRRAKW | RRLRRAQAFR | NMVFVMSPP | VGVLDDHYG | PIPIVEVIWK | CYPMVPRMH | PQMPLPPR | PPGFRMPFPF | RPPPLPFVWF | PFVPPDAHPI | NAAREYNPF | FFFPFPPFF | | | | | | |
| Mus cookii | | | | | | | | | |L..... |H.V..... | |L..... | | | | | | |
| Mus cervicolor | | | | | | | | | |L..... | | |L..... | | | | | | |
| Mus spretus | | | | | | | | | | | | |N..... | | | | | | |
| Mus musculus | | NFPNPNPNP | NPNPNPNPNP | QNPFAGPK*- | | | | | | | | | | | | | | | |
| Mus cookii | | | |RYR Y* | | | | | | | | | | | | | | | |
| Mus cervicolor | | | |RYR Y* | | | | | | | | | | | | | | | |
| Mus spretus | | | |RYR Y* | | | | | | | | | | | | | | | |

Figure 3. Protein sequence alignment for exon 4 of *ESX1* (*Esx1*) in (A) 12 primates and (B) 4 *Mus* species. ‘.’ indicates identity to the first sequence in each alignment. ‘-’ indicates an alignment gap and ‘*’ indicates a stop codon. The partial homeodomain region is indicated. ‘OW’, Old World; ‘NW’, New World.

then computed for species pairs within each of the three groups. It happened that each of the three groups has 5 species. Of the 30 pairwise comparisons, 23 (76.7%) show $d_N > d_S$. This finding is similar to the above result when all 15 sequences are aligned and analyzed together.

To test positive selection in primate *ESX1* more rigorously, we conducted a phylogeny-based analysis (22). We inferred the ancestral sequences for the C-terminus at all interior nodes of the primate tree (Fig. 5) using PAML (30) and then counted the numbers of non-synonymous and synonymous substitutions on each tree branch. We found that the ratio of the total number of non-synonymous substitutions and that of synonymous substitutions over all branches of the tree equals $n/s = 139/38 = 3.66$, significantly greater than the expected value of $N/S = 177/93 = 1.90$ ($P = 0.002$, Fisher’s exact test). Here N and S are the numbers of non-synonymous and synonymous sites, respectively, in the C-terminus.

We also conducted a likelihood-based analysis to detect positive selection on individual codons within the C-terminus using PAML. We compared the null model M8a with the alternative model M8. M8a, introduced by Swanson *et al.* (31), assumes that the d_N/d_S ratio of individual codons follows a beta distribution between 0 and 1, with an extra class of codons with fixed d_N/d_S of 1. M8 is identical to M8a except for the presence of an additional class of codons with any d_N/d_S . M8 is found to fit the data significantly better than M8a ($\chi^2 = 49.63$, $df = 2$, $P < 10^{-10}$). M8 suggests

that ~66% of codons in the C-terminus have been subject to positive selection with $d_N/d_S = 4.04$. Analysis using another pair of models, M1a and M2a, also supports a large proportion of codons under positive selection ($\chi^2 = 49.77$, $df = 2$, $P < 10^{-10}$). Taken together, various analyses provide strong evidence that positive selection has acted in the C-terminus of primate *ESX1* to promote amino acid substitutions. We note that although sequence alignment is not easy for *ESX1*, alignment errors cannot render d_N significantly greater than d_S , because even when the alignment is completely random, d_N is expected to be equal to d_S .

Purifying selection on rodent *Esx1*

To test whether positive selection on *ESX1* extends to non-primate mammals, we turn to rodents. We first obtained the *Esx1* gene sequence of *Mus musculus* from GenBank and determined the sequence of exons 2 and 4 of *Esx1* from *M. spretus*. Coding for only two and 15 amino acids, respectively, exons 1 and 3 are not studied here. *Esx1* possesses a unique PF/PN motif at its C-terminus, consisting of proline-phenylalanine (PF) tandem repeats followed by proline-asparagine (PN) tandem repeats (Fig. 1B). An earlier study found that the PF/PN motif can inhibit both nuclear localization and DNA binding activity of the *Esx1* protein (14). A comparison of exons 2 and 4 sequences of *M. musculus* and *M. spretus* shows strong purifying selection on *Esx1*, as d_N (0.006) is significantly lower than d_S (0.032) ($P = 0.01$,

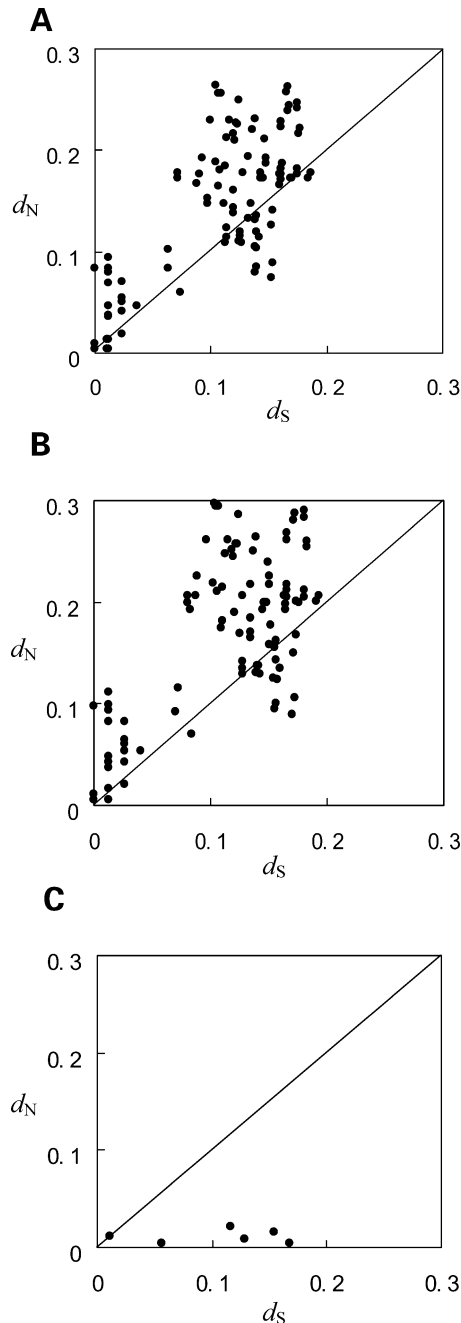


Figure 4. Pairwise synonymous (d_S) and non-synonymous (d_N) nucleotide distances for (A) the entire exon 4 of *ESX1* among 15 primates, (B) the C-terminal non-homeodomain region of *ESX1* among 15 primates and (C) the exon 4 of *Esx1* among 4 *Mus* species.

Fisher's exact test). Exon 4 was also sequenced in *M. cervicolor* and *M. cookii* (Fig. 3B). Significantly lower d_N than d_S is observed in all pairwise comparisons among the four *Mus* species with the exception of the comparison between *M. cervicolor* and *M. cookii*, probably owing to the small number of substitutions involved (Fig. 4C). Sequence length variation is observed in the PN/PF motif but not in the proline-rich region. Overall, our results suggest that *Esx1* has been subject to purifying selection in the *Mus* genus of rodents.

DISCUSSION

In this work, we provide evidence for positive selection acting in the C-terminus region of the homeodomain-containing protein *ESX1* during primate evolution as well as in human populations. In adult humans, *ESX1* is primarily expressed in testis. A previous study showed that *ESX1* is proteolytically processed into a 45-kDa N-terminal fragment (including the homeodomain) and a 20-kDa C-terminal fragment. The C-terminal fragment is found in cytoplasm and can inhibit the degradation of cyclin A and B1, causing cell-cycle arrest in human cells (32). Cyclins are a family of proteins controlling transitions through different phases of the cell cycle. Thus, it has been proposed that the C-terminal fragment of *ESX1* plays a role in spermatogenesis, functioning as a checkpoint in male germ cell division (32). In contrast, the N-terminal fragment, including the homeodomain, functions as a transcriptional repressor in nucleus (32,33). Our observations of conserved sequences in the N-terminal fragment but rapid sequence changes in the C-terminal fragment are explainable by the distinct functions of the two regions. The finding of positive selection in the C-terminus of primate *ESX1* suggests that even in the recent past of human and primate evolution, spermatogenesis has been subject to adaptive modifications (34). Because different species reach sexual maturity at different age, the optimal time of germ cell division may also vary among species. The observed positive selection on *ESX1* may reflect such adaptations in individual species. In general, our finding is consistent with many reports of rapid evolution of proteins involved in animal male reproduction (35). Furthermore, mammalian sperm proteins on the X chromosome have been found to evolve faster than those on autosomes (36). Thus, the rapid evolution of *ESX1* is likely related to its role in spermatogenesis as well as its location in the X chromosome.

Interestingly, male mice with null *Esx1* are fertile, indicating that *Esx1* is not essential for spermatogenesis in mice (19). The observation of purifying selection acting on the C-terminal region of *Esx1* in mice may be explained by the fact that the gene function has changed between primates and rodents. It is likely that *Esx1* is more important for placenta development rather than spermatogenesis in mice (19). Biochemical studies also showed that the nuclear localization of mouse *Esx1* is regulated by the presence of the PF/PN motif (14), which is lacking in primate *ESX1*, further suggesting functional differences between primate *ESX1* and rodent *Esx1*. To examine the C-terminal sequence of *ESX1* in other mammals, we TBLASTN-searched the GenBank with human *ESX1* and mouse *Esx1* as queries. We found putative orthologous *Esx1* genes in rat, dog and cow. In horse, only a partial sequence was identified by WISE2 (<http://www.ebi.ac.uk/Wise2/>). We did not find *Esx1* orthologs in opossum and chicken genome sequences. The estimated d_N/d_S ratio between mouse and rat in the C-terminal region of *Esx1* is significantly lower than 1 ($P < 0.01$), consistent with our findings in the *Mus* genus. Rat *Esx1* has a similar domain structure as mouse *Esx1*, with the exception that all of the PF repeats are replaced by PN repeats in the PF/PN motif. In contrast, the C-terminus of cow and horse *Esx1* proteins is similar to that of primates, with the proline-rich region but not the

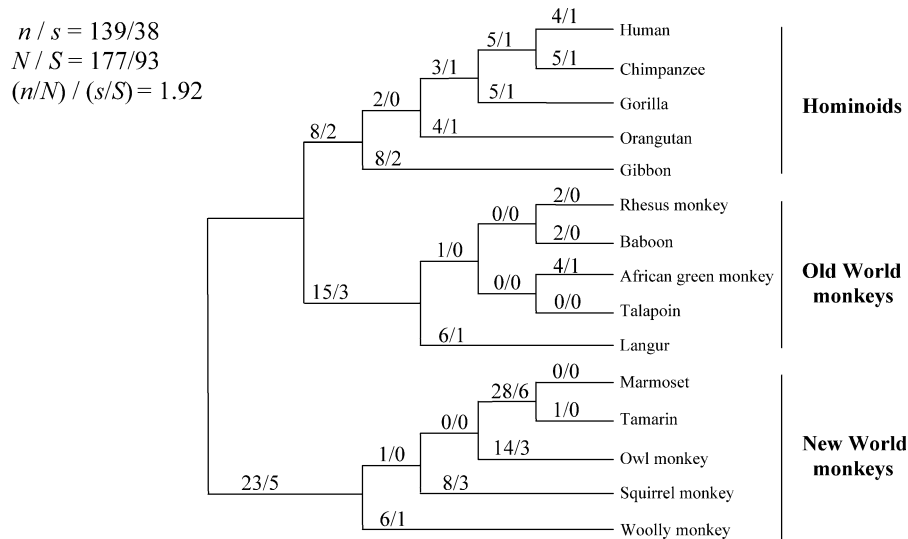


Figure 5. Numbers of synonymous (s) and non-synonymous (n) substitutions in the evolution of primate *ESX1*. Shown on each branch are the n and s values for the C-terminal non-homeodomain region. The numbers of non-synonymous (N) and synonymous (S) sites for the same region are also given.

PF/PN motif. The putative *Esx1* in dog has neither the proline-rich region nor the PF/PN motif at its C-terminus. It seems likely that the PF/PN motif was acquired by *Esx1* in rodent evolution.

The different evolutionary patterns of primate *ESX1* and rodent *Esx1* suggest that the utility of the mouse model for studying human reproduction may be limited. Previous studies also reported several other reproduction-related genes that show substantive human–mouse differences. For example, *SED1*, a protein involved in sperm-egg binding in mice, has lost an important protein–protein binding domain in ancestral primates, which was accompanied by rapid sequence changes in another domain by positive selection (37). In another example, three human X-linked homeobox genes, *PEPP1*, *PEPP2* and *PEPP3*, correspond to a cluster to 30 *Rhox* genes in mouse, due to dramatic expansions of the gene cluster in rodent evolution (9). The mouse *Rhox* genes are expressed in male and female reproductive tissues and at least one of them (*Rhox5*) is involved in male reproduction, evident from reduced fertilities of *Rhox5*-knockout mice (7).

Esx1 is paternally imprinted in mouse placenta and is functionally important to placenta morphogenesis and fetal growth (17,19). In contrast, *ESX1* is not imprinted in human placenta (20). Imprinting is an important regulatory pathway involved in the development and function of the placenta in eutherian mammals. The imprinting of *Esx1* is consistent with the general phenomenon in mice that the paternally derived X chromosome is preferentially inactivated in placental tissues of female embryos (38,39). Recently, Monk *et al.* (40) reported that several human orthologs of mouse placenta-imprinted genes are un-imprinted. In addition, an earlier investigation revealed a widespread reduction in the maintenance of imprinting in humans (41). If imprinted genes tend to be involved in intra-genomic conflict and hence evolve rapidly by arms race (42), our observation of

rapid evolution of the un-imprinted primate *ESX1* but slow evolution of imprinted rodent *Esx1* is unexpected. While a change in spermatogenesis function might explain the unexpected evolutionary pattern for *ESX1/Esx1*, in the future, it would be interesting to test the genomic conflict hypothesis by comparing the evolutionary rates of all mouse imprinted genes with those of their un-imprinted human orthologs.

MATERIALS AND METHODS

DNA samples

One individual from each of 12 primate species, 32 male humans, 1 *Mus spretus*, 1 *Mus cookie* and 1 *Mus cervicolor* were surveyed. The 12 primate species include three hominoids (gorilla *Gorilla gorilla*, orangutan *Pongo pygmaeus* and gibbon *Hylobates lar*), four Old World monkeys (green monkey *Cercopithecus aethiops*, langur *Pygathrix nemaeus*, talapoin *Miopithecus talapoin* and baboon *Papio hamadryas*) and five New World monkeys (marmoset *Callithrix jacchus*, tamarin *Saguinus oedipus*, owl monkey *Aotus trivirgatus*, squirrel monkey *Saimiri sciureus* and woolly monkey *Lagothrix lagotricha*). The animal DNA samples were from Wang and Zhang (12) and Podlaha *et al.* (43), whereas the human DNA samples were purchased from Coriell (<http://ccr.coriell.org/>).

Gene amplification and DNA sequencing

The amplified *ESX1* regions in different species and the primers used for amplification are described in Supplementary Material, Table S2. Primers were designed according to the published human (NT_011651) and mouse (NM_007957) sequences. Polymerase chain reactions (PCRs) were performed with MasterTaq or TripleMasterTaq under conditions recommended by the manufacturer (Eppendorf, Hamburg,

Germany). Dimethyl sulfoxide (DMSO) was used in PCR amplification and DNA sequencing of exon 4. Amplified exon 4 sequences from 12 primates were cloned into PCR4TOPO vector (Invitrogen) and then sequenced from both directions. Other PCR products were purified and directly sequenced from both directions. The dideoxy chain termination method was used in DNA sequencing by an automated sequencer. Sequencher (GeneCodes) was used to assemble the sequences and identify DNA polymorphisms in humans.

Human population genetic analysis

The *DH* test was conducted by program DH.jar (26). The population recombination rate used in the test was estimated to be $R = 3Nr = 3 \times 10\,000 \times (0.18 \times 10^{-6} \times 723) = 4$ per sequence per generation. Here $N = 10\,000$ is the effective population size of humans, 0.18×10^{-6} is the pedigree-based recombination rate per generation per nucleotide at the *ESX1* locus (44) and 723 is the number of nucleotides of the human *ESX1* exon 4 sequence. For samples of African, Caucasian and Asian origins, we used $N = 10\,000$, 4000 and 4000, respectively, as their effective population sizes (45). The chimpanzee *ESX1* sequence was used as the outgroup in computing *DH* except for one site where the gorilla sequence was used as the outgroup because the chimpanzee sequence is different from both human alleles. *P*-values in the *DH* test and *H* test were estimated using 100 000 replications of coalescent simulation.

Evolutionary analysis

The coding sequences of human *ESX1* and mouse *Esx1* were obtained from GenBank with accession numbers AY114148 and NM_007957, respectively. Clustal W (46) was used to conduct sequence alignment for the primates and the *Mus* species, respectively. MEGA3 (47) was used for the phylogenetic analysis. Pairwise synonymous (d_S) and non-synonymous (d_N) distances were calculated using the modified Nei-Gojobori method (48), with estimated transition/transversion ratios. Based on the phylogeny of 15 primates, we inferred ancestral *ESX1* sequences at all interior nodes of the tree by using the likelihood method under the M8 model in PAML3.15 (30). The number of synonymous (s) and non-synonymous (n) substitutions on each branch of the tree was then counted. The number of synonymous (S) and non-synonymous (N) sites was also estimated by PAML.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS

We thank Kai Zeng for assistance with the *DH* test and Meg Bakewell, Soochin Cho and Ondrej Podlaha for valuable comments. This work was supported by research grants from the University of Michigan and National Institutes of Health to J.Z.

Conflict of Interest statement. None declared.

REFERENCES

- Gehring, W.J., Affolter, M. and Burglin, T. (1994) Homeodomain proteins. *Annu. Rev. Biochem.*, **63**, 487–526.
- Nam, J. and Nei, M. (2005) Evolutionary change of the numbers of homeobox genes in bilateral animals. *Mol. Biol. Evol.*, **22**, 2386–2394.
- McGinnis, W., Hart, C.P., Gehring, W.J. and Ruddle, F.H. (1984) Molecular cloning and chromosome mapping of a mouse DNA sequence homologous to homeotic genes of *Drosophila*. *Cell*, **38**, 675–680.
- Zhang, J. and Nei, M. (1996) Evolution of Antennapedia-class homeobox genes. *Genetics*, **142**, 295–303.
- Ting, C.T., Tsaur, S.C., Wu, M.L. and Wu, C.I. (1998) A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science*, **282**, 1501–1504.
- Sutton, K.A. and Wilkinson, M.F. (1997) Rapid evolution of a homeodomain: evidence for positive selection. *J. Mol. Evol.*, **45**, 579–588.
- Maclean, J.A., II, Chen, M.A., Wayne, C.M., Bruce, S.R., Rao, M., Meistrich, M.L., Macleod, C. and Wilkinson, M.F. (2005) RhoX: a new homeobox gene cluster. *Cell*, **120**, 369–382.
- MacLean, J.A., II, Lorenzetti, D., Hu, Z., Salerno, W.J., Miller, J. and Wilkinson, M.F. (2006) RhoX homeobox gene cluster: recent duplication of three family members. *Genesis*, **44**, 122–129.
- Wang, X. and Zhang, J. (2006) Remarkable expansions of an X-linked reproductive homeobox gene cluster in rodent evolution. *Genomics*, **88**, 34–43.
- Morris, L., Gordon, J. and Blackburn, C.C. (2006) Identification of a tandem duplicated array in the RhoX alpha locus on mouse chromosome X. *Mamm. Genome*, **17**, 178–187.
- Jackson, M., Watt, A.J., Gautier, P., Gilchrist, D., Driehaus, J., Graham, G.J., Keebler, J., Prugnonne, F., Awadalla, P. and Forrester, L.M. (2006) A murine specific expansion of the RhoX cluster involved in embryonic stem cell biology is under natural selection. *BMC Genomics*, **7**, 212.
- Wang, X. and Zhang, J. (2004) Rapid evolution of mammalian X-linked testis-expressed homeobox genes. *Genetics*, **167**, 879–888.
- Fohn, L.E. and Behringer, R.R. (2001) ESX1L, a novel X chromosome-linked human homeobox gene expressed in the placenta and testis. *Genomics*, **74**, 105–108.
- Yan, Y.T., Stein, S.M., Ding, J., Shen, M.M. and Abate-Shen, C. (2000) A novel PF/PN motif inhibits nuclear localization and DNA binding activity of the ESX1 homeoprotein. *Mol. Cell. Biol.*, **20**, 661–671.
- Figueiredo, A.L., Salles, M.G., Albano, R.M. and Porto, L.C. (2004) Molecular and morphologic analyses of expression of ESX1L in different stages of human placental development. *J. Cell. Mol. Med.*, **8**, 545–550.
- Murthi, P., Doherty, V.L., Said, J.M., Donath, S., Brennecke, S.P. and Kalionis, B. (2006) Homeobox gene ESX1L expression is decreased in human pre-term idiopathic fetal growth restriction. *Mol. Hum. Reprod.*, **12**, 335–340.
- Li, Y., Lemaire, P. and Behringer, R.R. (1997) Esx1, a novel X chromosome-linked homeobox gene expressed in mouse extraembryonic tissues and male germ cells. *Dev. Biol.*, **188**, 85–95.
- Branford, W.W., Zhao, G.Q., Valerius, M.T., Weinstein, M., Birkenmeier, E.H., Rowe, L.B. and Potter, S.S. (1997) Spx1, a novel X-linked homeobox gene expressed during spermatogenesis. *Mech. Dev.*, **65**, 87–98.
- Li, Y. and Behringer, R.R. (1998) Esx1 is an X-chromosome-imprinted regulator of placental development and fetal growth. *Nat. Genet.*, **20**, 309–311.
- Grati, F.R., Sirchia, S.M., Gentilin, B., Rossella, F., Ramoscelli, L., Antonazzo, P., Cavallari, U., Bulfamante, G., Cetin, I., Simoni, G. et al. (2004) Biparental expression of ESX1L gene in placentas from normal and intrauterine growth-restricted pregnancies. *Eur. J. Hum. Genet.*, **12**, 272–278.
- Bakewell, M.A., Shi, P. and Zhang, J. (2007) More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc. Natl Acad. Sci. USA*, **104**, 7489–7494.
- Zhang, J., Kumar, S. and Nei, M. (1997) Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. *Mol. Biol. Evol.*, **14**, 1335–1338.
- Consortium, C.S.a.A. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.

24. Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
25. Fay, J.C. and Wu, C.I. (2000) Hitchhiking under positive Darwinian selection. *Genetics*, **155**, 1405–1413.
26. Zeng, K., Fu, Y.X., Shi, S. and Wu, C.I. (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*, **174**, 1431–1439.
27. Akey, J.M., Eberle, M.A., Rieder, M.J., Carlson, C.S., Shriver, M.D., Nickerson, D.A. and Kruglyak, L. (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.*, **2**, e286.
28. Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
29. Li, W. (1997) *Molecular Evolution*, Sinauer, Sunderland, Mass.
30. Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.
31. Swanson, W.J., Nielsen, R. and Yang, Q. (2003) Pervasive adaptive evolution in mammalian fertilization proteins. *Mol. Biol. Evol.*, **20**, 18–20.
32. Ozawa, H., Ashizawa, S., Naito, M., Yanagihara, M., Ohnishi, N., Maeda, T., Matsuda, Y., Jo, Y., Higashi, H., Kakita, A. *et al.* (2004) Paired-like homeodomain protein ESXR1 possesses a cleavable C-terminal region that inhibits cyclin degradation. *Oncogene*, **23**, 6590–6602.
33. Yanagihara, M., Ishikawa, S., Naito, M., Nakajima, J., Aburatani, H. and Hatakeyama, M. (2005) Paired-like homeoprotein ESXR1 acts as a sequence-specific transcriptional repressor of the human K-ras gene. *Oncogene*, **24**, 5878–5887.
34. Wyckoff, G.J., Wang, W. and Wu, C.I. (2000) Rapid evolution of male reproductive genes in the descent of man. *Nature*, **403**, 304–309.
35. Swanson, W.J. and Vacquier, V.D. (2002) The rapid evolution of reproductive proteins. *Nat. Rev. Genet.*, **3**, 137–144.
36. Torgerson, D.G. and Singh, R.S. (2003) Sex-linked mammalian sperm proteins evolve faster than autosomal ones. *Mol. Biol. Evol.*, **20**, 1705–1709.
37. Podlaha, O., Webb, D.M. and Zhang, J. (2006) Accelerated evolution and loss of a domain of the sperm-egg-binding protein SED1 in ancestral primates. *Mol. Biol. Evol.*, **23**, 1828–1831.
38. West, J.D., Frels, W.I., Chapman, V.M. and Papaioannou, V.E. (1977) Preferential expression of the maternally derived X chromosome in the mouse yolk sac. *Cell*, **12**, 873–882.
39. Wagschal, A. and Feil, R. (2006) Genomic imprinting in the placenta. *Cytogenet. Genome Res.*, **113**, 90–98.
40. Monk, D., Arnaud, P., Apostolidou, S., Hills, F.A., Kelsey, G., Stanier, P., Feil, R. and Moore, G.E. (2006) Limited evolutionary conservation of imprinting in the human placenta. *Proc. Natl Acad. Sci. USA*, **103**, 6623–6628.
41. Morison, I.M., Ramsay, J.P. and Spencer, H.G. (2005) A census of mammalian imprinting. *Trends Genet.*, **21**, 457–465.
42. Haig, D. (1993) Genetic conflicts in human pregnancy. *Q. Rev. Biol.*, **68**, 495–532.
43. Podlaha, O., Webb, D.M., Tucker, P.K. and Zhang, J. (2005) Positive Selection for indel substitutions in the rodent sperm protein Catsper1. *Mol. Biol. Evol.*, **22**, 1845–1852.
44. Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsson, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G. *et al.* (2002) A high-resolution recombination map of the human genome. *Nat. Genet.*, **31**, 241–247.
45. Tenesa, A., Navarro, P., Hayes, B.J., Duffy, D.L., Clarke, G.M., Goddard, M.E. and Visscher, P.M. (2007) Recent human effective population size estimated from linkage disequilibrium. *Genome Res.*, **17**, 520–526.
46. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (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.
47. Kumar, S., Tamura, K. and Nei, M. (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform.*, **5**, 150–163.
48. Zhang, J., Rosenberg, H.F. and Nei, M. (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl Acad. Sci. USA*, **95**, 3708–3713.