Available online at www.sciencedirect.com
SCIENCE DIRECT®
Genomics 88 (2006) 34–43
GENOMICS
www.elsevier.com/locate/ygeno

# Remarkable expansions of an X-linked reproductive homeobox gene cluster in rodent evolution

Xiaoxia Wang, Jianzhi Zhang *

*Department of Ecology and Evolutionary Biology, University of Michigan, 1075 Natural Science Building, 830 North University Avenue, Ann Arbor, MI 48109, USA*

Received 8 November 2005; accepted 14 February 2006
Available online 30 March 2006

**Abstract**

*Rhox* is a recently identified cluster of 12 X-linked homeobox genes in mice. The expression pattern of *Rhox* genes during postnatal testis development corresponds to their chromosomal position, much like the colinear gene regulation of the *Hox* gene clusters during animal embryonic development. We here report the identification of 18 additional *Rhox* genes and 3 pseudogenes in mice. Comparative analyses of the mouse, rat, human, dog, cow, opossum, and chicken genomes suggest that the *Rhox* cluster originated in the common ancestor of primates and rodents. It subsequently underwent two remarkable expansions, first in the common ancestor of mice and rats and then in mice. Positive selection promoting amino acid substitutions was detected in some young *Rhox* genes, suggesting adaptive functional diversification. The recent expansions of the *Rhox* cluster provide an opportunity to study the mechanism and origin of colinear gene regulation, but they may also undermine the utility of mouse models for understanding the development and physiology of the human reproductive system.


*Keywords:* Homeobox; Rhox; Reproduction; Positive selection; Rodents; Gene duplication; Mouse; Rat

Homeobox-containing genes form a large gene superfamily that is found in animals, fungi, and plants [1,2]. Homeobox genes are characterized by the presence of a 180-nucleotide sequence encoding a conserved 60-amino-acid DNA-binding homeodomain [3]. As transcription factors, homeodomain-containing proteins play important roles in various developmental processes such as body-plan specification, pattern formation, and cell fate determination [3]. Among all homeobox genes, special attention has been given to animal *Hox* genes, which were first characterized in fruitflies [4–6] and subsequently found in all metazoans. These genes are organized into clusters on autosomes. The expression domain and time of activation are correlated with their relative positions in the cluster, exhibiting expression colinearity [7]. The evolution of the *Hox* cluster has been extensively studied and the duplications of the *Hox* genes and clusters are believed to have played important roles in the evolution of the animal body plan [8–12].

Recently, MacLean et al. [13] reported the discovery of a new homeobox gene cluster on the mouse X chromosome and named it *Rhox* (reproductive homeobox X-linked). The mouse *Rhox* gene cluster contains 12 genes, which are assembled into three subclusters, α, β, and γ, by their chromosomal locations. Phylogenetic analysis showed that the 12 *Rhox* genes form a new homeobox gene family, distinct from other known homeobox genes. The phenomenon of expression colinearity was also found among some *Rhox* genes within subclusters during postnatal testis development. The tissue-specific expression patterns of all *Rhox* genes suggested that they play important roles in the development of male and female reproductive tissues. This was further supported by reduced male fertility of *Rhox5*-knockout mice [13].

The expressional and functional examinations of individual *Rhox* genes started before the identification of the *Rhox* cluster, including studies of *Pem* (*Rhox5*) [14], *Psx* (*Rhox6*) [15], *Psx-2* (*Rhox9*) [16], and *Tox* (*Rhox8*) [17]. However, the discoveries of the clustering of *Rhox* genes and the expression colinearity provide a starting point for uncovering the functional relationships among the *Rhox* genes and for understanding the functional

* Corresponding author. Fax: +1 734 763 0544.
*E-mail address:* jianzhi@umich.edu (J. Zhang).

0888-7543/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2006.02.007

and evolutionary significance of this homeobox gene cluster. They also offer an opportunity to study the molecular mechanism of colinear regulation and its evolutionary origin. To address these questions, it is necessary first to identify all *Rhox* genes and uncover the evolutionary history of these genes. The availability of genome sequences of several mammalian species makes this goal achievable. Here we first report the identification of 21 additional *Rhox* genes and pseudogenes from the mouse genome. We then describe the search for *Rhox* genes in the rat, human, dog, cow, opossum, and chicken genomes. The genomic data led to the finding of remarkable expansions of the *Rhox* cluster in rodents. We also provide evidence for the action of positive selection in the divergence of young *Rhox* genes in mice.

## Results

### *New members of the mouse Rhox cluster*

We used the 12 known mouse Rhox protein sequences as queries to BLAST the mouse genome sequence. We were able to identify 11 of the 12 known *Rhox* genes on the X chromosome. *Rhox5*, a gene that has been well characterized in expression and function [13], cannot be found in the mouse genome sequence (Ensembl release 34, NCBI build 35), likely because the genome sequence is incomplete or misassembled. Among-strain genetic variation cannot be the reason, as the knockout experiment and genomic sequencing used the same mouse strain. Unexpectedly, we discovered 18 additional *Rhox* genes and 3 pseudogenes on the X chromosome that were not detected by MacLean et al. [13]. One of these genes (*Rhox4b*) is identical to a previously reported gene (*Ehox*) [18], but the remaining 20 genes/pseudogenes are newly discovered. No *Rhox* genes or pseudogenes were found in mouse autosomes. All 21 newly identified *Rhox* genes/pseudogenes share a similar gene structure with the 12 known *Rhox* genes. In particular, the two introns that break the homeobox in all known *Rhox* genes are also present in the 21 new genes/pseudogenes. The conceptually translated protein sequences of the 18 new *Rhox* genes have the signature of W and F residues at positions 48 and 49 of the homeodomain, as in all known homeodomains (Fig. 1). In addition, the 18 new Rhox proteins are invariant at positions 16, 34, and 53 of the homeodomain, which was found in the original 12 Rhox proteins [13]. These findings indicate that the 18 new *Rhox* genes are likely to be functional. *Rhox2, Rhox3*, and *Rhox4* each have seven

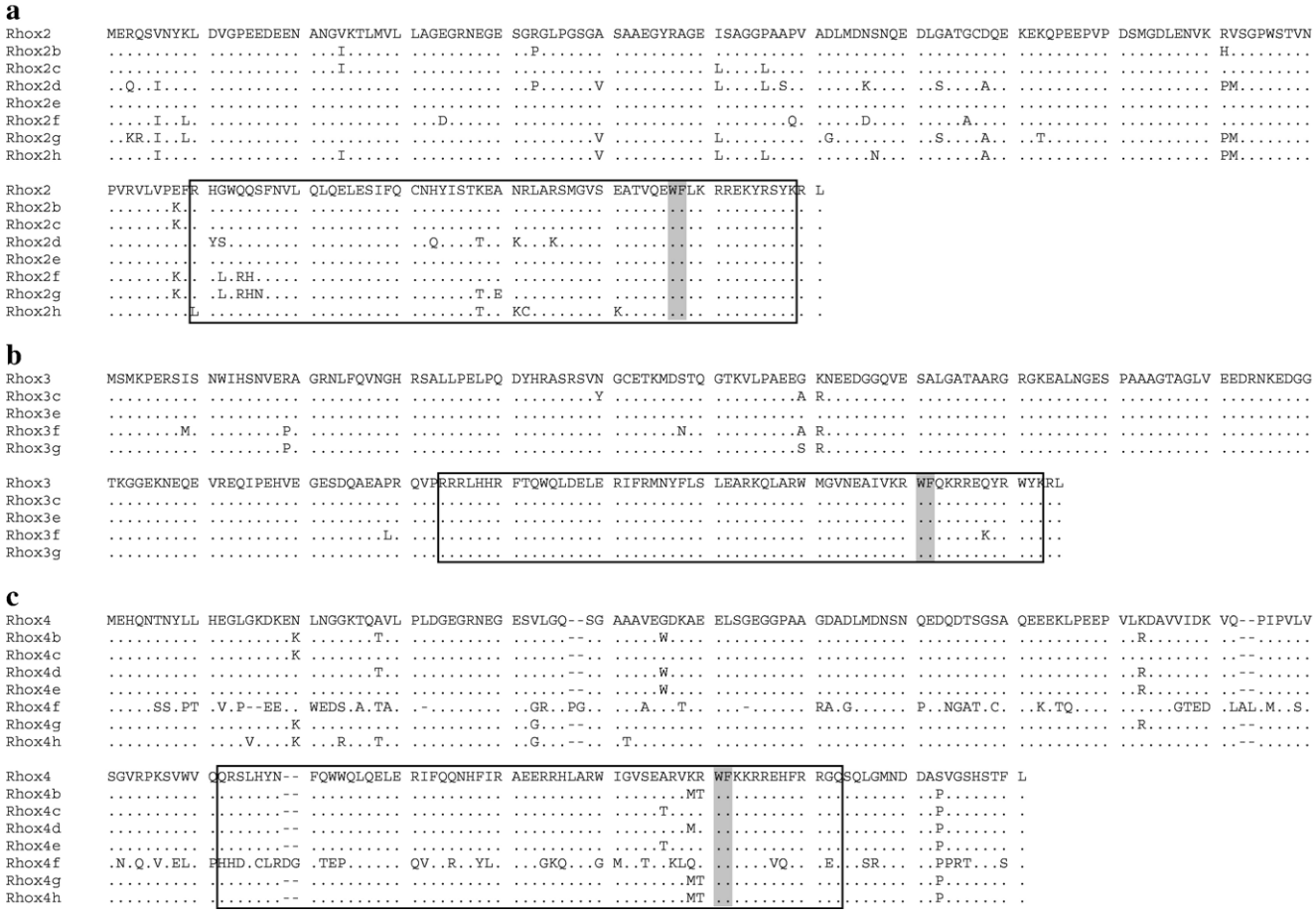


Fig. 1. The alignment of conceptually translated mouse (a) *Rhox2* genes, (b) *Rhox3* genes, and (c) *Rhox4* genes. "-" indicates an alignment gap and "." indicates identity to the first sequence. The homeodomain is boxed. Pseudogenes (*Rhox3bψ, Rhox3dψ*, and *Rhox3hψ*) are not presented. Conserved amino acid residues W and F within the homeodomain are shaded in gray.

paralogs among the newly identified genes (including the 3 pseudogenes that are most similar to *Rhox3*; Fig. 1). Therefore, we name these new genes and pseudogenes after their related genes by adding the letter *b* to *h* at the end of the gene names, with a ψ denoting a pseudogene. For example, *Rhox2b* to *Rhox2h* are given to the 7 genes most similar to *Rhox2*. It is interesting that all three *Rhox* pseudogenes we identified are related to *Rhox3* (*Rhox3b*ψ, *Rhox3d*ψ, and *Rhox3h*ψ). The pseudogenes are defined by the lack of open reading frames (ORFs), due to interruptions by premature stop codons. In the present case, the three pseudogenes also lack complete homeodomains and thus are nonfunctional. The premature stop codons in *Rhox3b*ψ and *Rhox3h*ψ are at the same position, indicating that the common ancestor of the two pseudogenes was already a pseudogene. Carrying a distinct premature stop codon, the pseudogenization of *Rhox3d*ψ likely resulted from an independent nonsense mutation. The genomic region harboring the 24 *Rhox2/3/4*-related genes spans ~313 kb, and the relative locations of these 21 new *Rhox* genes/pseudogenes on the X chromosome are shown in Fig. 2. Based on these locations, the most likely evolutionary scenario for the origin of the 24 *Rhox2/3/4*-related genes is a series of block duplications of a 3-gene unit such as {*Rhox2, Rhox3, Rhox4*}. It is interesting to note that the 17 genes proximal to the centromere and the 7 genes distal to the centromere have opposite directions of transcription, which can be explained by an inversion of the 7-gene block.

To rule out the possibility that the newly identified *Rhox2/3/4*-related genes are artifacts of sequencing and/or assembly errors in the mouse genome sequence, we used polymerase chain reaction (PCR) to amplify from the mouse genomic DNA *Rhox2-, Rhox3*-, and *Rhox4*-related genes and then sequenced the amplified genes. Because the genomic DNA was from an inbred mouse, no polymorphism was expected. Rather, sequence variations would indicate the presence of multiple paralogous genes that were amplified by the same pairs of PCR primers. Indeed, we predicted 35 variable sites from the mouse genome sequence and observed 33 of them from our sequencing results (Supplementary Figs. 1–3). For the remaining 2 sites, the fluorescent signals in the chromatograms do not unambiguously indicate variations. If we assume that these 2 variations in the mouse genome sequence are real, the principle reason for our failed detection may be that the variant nucleotides constitute a small fraction among all paralogs and thus are difficult to detect in PCR sequencing. Overall, our experiments support the computational predictions of the new *Rhox2/3/4*-related genes.

### *Evolution of mouse Rhox2/3/4-related genes*

Among all *Rhox2/3/4*-related genes, *Rhox4f* is abnormal. It has a relatively low sequence identity (~76% in nucleotide sequence) with other *Rhox4* genes, in contrast to the high sequence identity (94–100%) among the other *Rhox4* genes, *Rhox2* genes, and *Rhox3* genes. More interestingly, we found that the N-terminal half of *Rhox4f* is similar to that of other *Rhox4* genes, whereas the C-terminal half (including the homeodomain) is similar to *Rhox7* (Fig. 3). This chimera may have arisen from a gene conversion or recombination event.

As mentioned, the chromosomal locations of the 24 *Rhox2/3/4*-related genes strongly suggest block duplications of a 3-gene unit. If this hypothesis is correct, one expects identical topologies among the tree of *Rhox2*-related genes, that of *Rhox3*-related genes, and that of *Rhox4*-related genes. This pattern, however, was not observed (Fig. 4, Supplementary Fig. 4). Two possible scenarios may explain this observation. First, the homologous *Rhox* genes are subject to gene conversions as seen in *Rhox4f*, which would have distorted the true phylogenetic relationships among genes. In fact, we were able to detect statistically significant signals of gene conversion among *Rhox2* genes, *Rhox3* genes, and *Rhox4* genes, with Sawyer's method [19] (see Materials and methods). A less likely alternative scenario is that the duplications did not occur in units of 3 genes, which seems highly improbable given the pattern of gene numbers and locations in the X chromosome (Fig. 2).

To understand the selective pressures acting on the highly similar *Rhox2/3/4*-related genes, we compared the nonsynonymous ($d_N$) and synonymous ($d_S$) distances among the paralogous genes. Interestingly, different selective pressures appear to have acted on *Rhox2* genes, *Rhox3* genes, and *Rhox4* genes. Higher $d_N$ than $d_S$ was observed in 11 of 28 pair-wise comparisons of the eight *Rhox2* genes, with the ratio between mean $d_N$ and mean $d_S$ being 0.78 (Fig. 5a). Thus, overall, the divergences among *Rhox2*-related genes have been under weak purifying selection. For the five *Rhox3* genes, 2 of 10 pair-wise comparisons show $d_N > d_S$, with the ratio between mean $d_N$ and mean $d_S$ being 0.47 (Fig. 5b). Thus, *Rhox3*-related genes have been under stronger purifying selection. For the seven *Rhox4* genes (excluding *Rhox4f*), however, 20 of 21 pair-wise comparisons show $d_N > d_S$, with the ratio between mean $d_N$ and mean $d_S$ being 3.16 (Fig. 5c), suggesting the action of positive selection. To verify this result further, we adopted a phylogeny-based approach to avoid the dependence among the pair-wise distances [20]. We inferred the ancestral gene sequences at the interior nodes of the *Rhox4* gene tree (Fig. 4c) and counted the numbers of nonsynonymous ($n$) and synonymous ($s$) substitutions on each branch of the tree (Fig. 6). The total number of nonsynonymous substitutions throughout the tree is $\sum n = 24$, and the corresponding number of synonymous substitutions is $\sum s = 2$. The potential numbers of nonsynonymous and synonymous sites are $N = 464$ and $S = 151$, respectively. Hence, the rate of nonsynonymous substitution is $n/N = 24/464 = 0.052$, significantly greater than the rate of synonymous substitution $s/S = 2/151 = 0.013$ ($p = 0.027$, Fisher's exact test; [21]). This result is consistent with the result from the pair-wise analysis. It should be pointed out that the mean $d_S$ is lower in *Rhox4* genes than in *Rhox2* genes and *Rhox3* genes (Fig. 5), suggesting that gene conversion might be more frequent in *Rhox4* genes. Gene conversion itself is blind to synonymous and nonsynonymous changes and cannot cause $d_N > d_S$. If certain nonsynonymous differences between paralogs are advantageous to the organism,

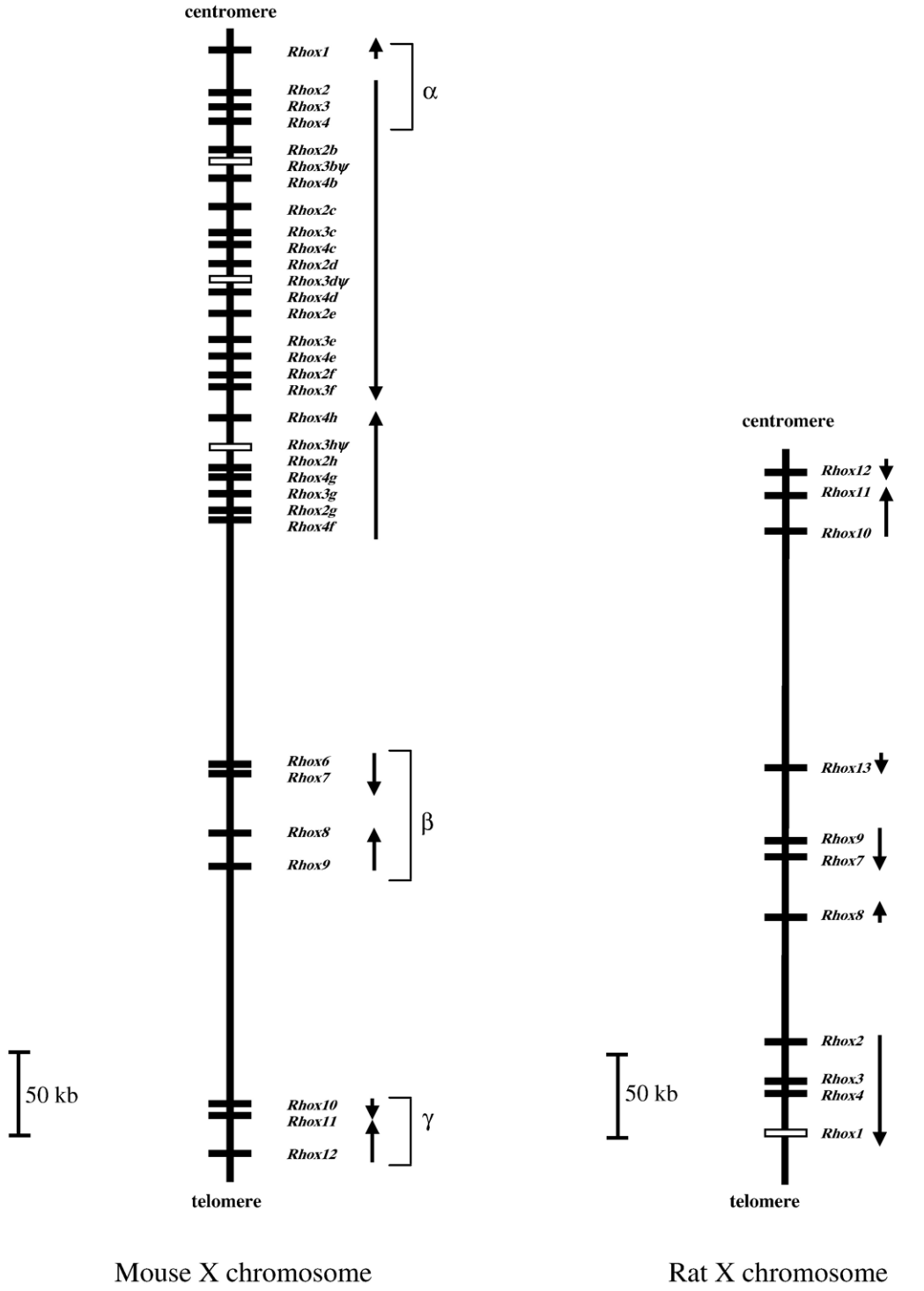


Fig. 2. The locations of *Rhox* genes on the mouse and rat X chromosomes. Putatively functional genes are indicated by solid boxes, whereas pseudogenes are depicted by open boxes. Transcriptional directions are shown by arrows. Three previously defined mouse *Rhox* subclusters [13], α, β, and γ, are shown. Previously reported mouse *Rhox5* and rat *Rhox5* cannot be found in the genome sequences and therefore are not depicted here. There are two rat *Rhox* genes (*Rhox14* and *Rhox15*) and three pseudogenes that are located on autosomes, and they are not depicted here. The gene maps are drawn to scale.

gene conversions homogenizing these nonsynonymous sites may be disfavored, while those homogenizing synonymous differences are neutral, resulting in $d_N > d_S$. Even in this improbable scenario, however, the nonsynonymous differences must be advantageous. Thus, the observation of $d_N > d_S$ suggests adaptive functional differences between paralogs.

### *Dating the block duplications of Rhox2/3/4-related genes*

To date the block duplications, we computed the pair-wise nucleotide distances (Kimura's two-parameter model) among all eight *Rhox2*-related genes using the intron sequences. The largest distance among the 28 pair-wise distances is

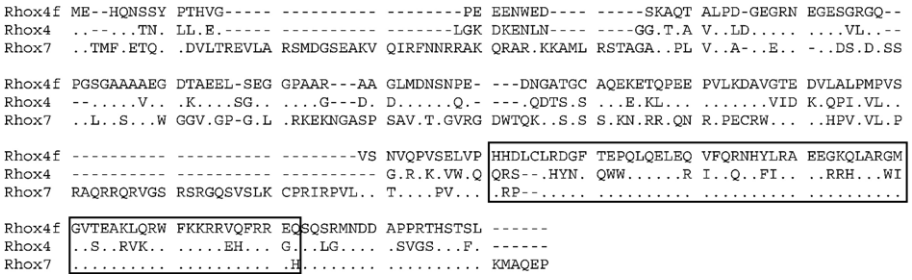


Fig. 3. The chimeric structure of mouse Rhox4f. Its N-terminal half is similar to that of mouse Rhox4, whereas the C-terminal half is similar to that of mouse Rhox7. The homeodomain is boxed.

0.068 ± 0.005. If we use the neutral substitution rate of $4.59 \times 10^{-9}$ per site per year for X-chromosomal sequences (see Materials and methods), the above distance suggests that the first duplication event occurred 7.4 million years (MY) ago. Similarly, the largest pair-wise nucleotide distance among all eight *Rhox3*-related genes is 0.032 ± 0.003 in the introns, corresponding to a duplication time of 3.5 MY ago. The largest pair-wise nucleotide distance among the seven *Rhox4*-related genes (excluding *Rhox4f*) is 0.049 ± 0.003 in the introns, corresponding to a duplication time of 5.3 MY ago. Because of gene conversions, these dates may be underestimated. Thus, the first block duplication likely occurred at least 7.4 MY ago.

### *Rhox genes in rats*

We used the mouse *Rhox* genes as queries to BLAST the rat genome and identified 16 *Rhox*-related genes (Supplementary Dataset 1), including 3 previously unannotated genes. These 3 new genes are named *Rhox13, Rhox14*, and *Rhox15*. The previously reported rat *Rhox5* (*Pem*) gene [22], however, was not found in the rat genome sequence, probably due to the incompleteness of the draft genome sequence or misassembly. Among the identified rat *Rhox* genes, 12 have intact ORFs with complete homeodomains, suggesting that these genes are functional, while 4 are pseudogenes with disrupted ORFs. If we assume that *Rhox5* exists in the rat X chromosome, the rat *Rhox* family contains 2 functional and 3 nonfunctional members on autosomes, in addition to 11 functional and 1 nonfunctional member on the X chromosome (Fig. 2). The autosomal genes are *Rhox14* on chromosome 18 and *Rhox*15 on chromosome 4. The 3 autosomal pseudogenes are located on chromosomes 3, 8, and 19, respectively. It is possible that the autosomal *Rhox* genes have been relocated from X during rat evolution.

Based on protein sequence, we reconstructed the phylogeny of the mouse and rat functional *Rhox* genes (Fig. 7). Each of the nine circles in the tree represents an ancestral *Rhox* gene that has at least one mouse and one rat descendant. This pattern indicates that the common ancestor of mice and rats had at least nine *Rhox* genes, although it is possible that the ancestor had more than nine genes, with some subsequently lost in one or both lineages. The tree also shows that many gene duplication events occurred after the mouse–rat separation. This is particularly obvious for *Rhox2/3/4*-related genes in mice. The orthologous relationships of the mouse and rat *Rhox* genes are complicated due to gene duplications and potential losses after the separation of the two species. Clear-cut one-to-one orthology can be inferred only for mouse and rat *Rhox8, Rhox10, Rhox11*, and *Rhox12* genes.

### *Rhox genes in nonrodent species*

To understand the origin of the rodent *Rhox* cluster, we searched for potential *Rhox* genes in the human, dog, cow, opossum, and chicken genome sequences. Two X-linked homeobox genes, *PEPP1* and *PEPP2*, were previously suggested to be the human counterparts of the mouse *Rhox* genes [13]. In addition to these two genes, we identified a new human gene that is closely related to *PEPP2* and named it *PEPP3*. PEPP3 and PEPP2 have only two amino acid differences and thus may have resulted from recent gene duplication (Supplementary Dataset 2). Wayne et al. [23] described a gene tentatively named *PEPP2B*, as they were unsure whether the gene actually exists. *PEPP2B* has never been annotated in the human genome sequence and is likely the same as *PEPP3*.

The dog genome sequence has a high (7.6) coverage and thus should be relatively complete. However, no unambiguous *Rhox* genes were identified using BLAST searches. One putative ortholog of human *PEPP2* was predicted by Genscan (http://genes.mit.edu/GENSCAN.html) from the sequences in the syntenic interval of the dog X chromosome. But, further phylogenetic analysis revealed that this dog gene is more closely related to other homeobox genes than to *Rhox* genes (data not shown). No *Rhox* genes were found in the cow, opossum, or chicken genome sequences. These results suggest that the *Rhox* gene (or cluster) originated in the common ancestor of primates and rodents and subsequently underwent remarkable expansions in rodents.

## Discussion

In this work, we identified 21 new genes and pseudogenes that are part of the mouse X-linked reproductive homeobox gene cluster *Rhox.* Including the previously identified 12 genes, the mouse *Rhox* cluster contains a total of 33 genes, including 30 putatively functional genes and 3 pseudogenes. This is substantively larger than any *Hox* gene cluster in mammals, which has at most 13 genes. Chromosomal locations of the mouse *Rhox* genes (Fig. 2) and comparative genomic analysis

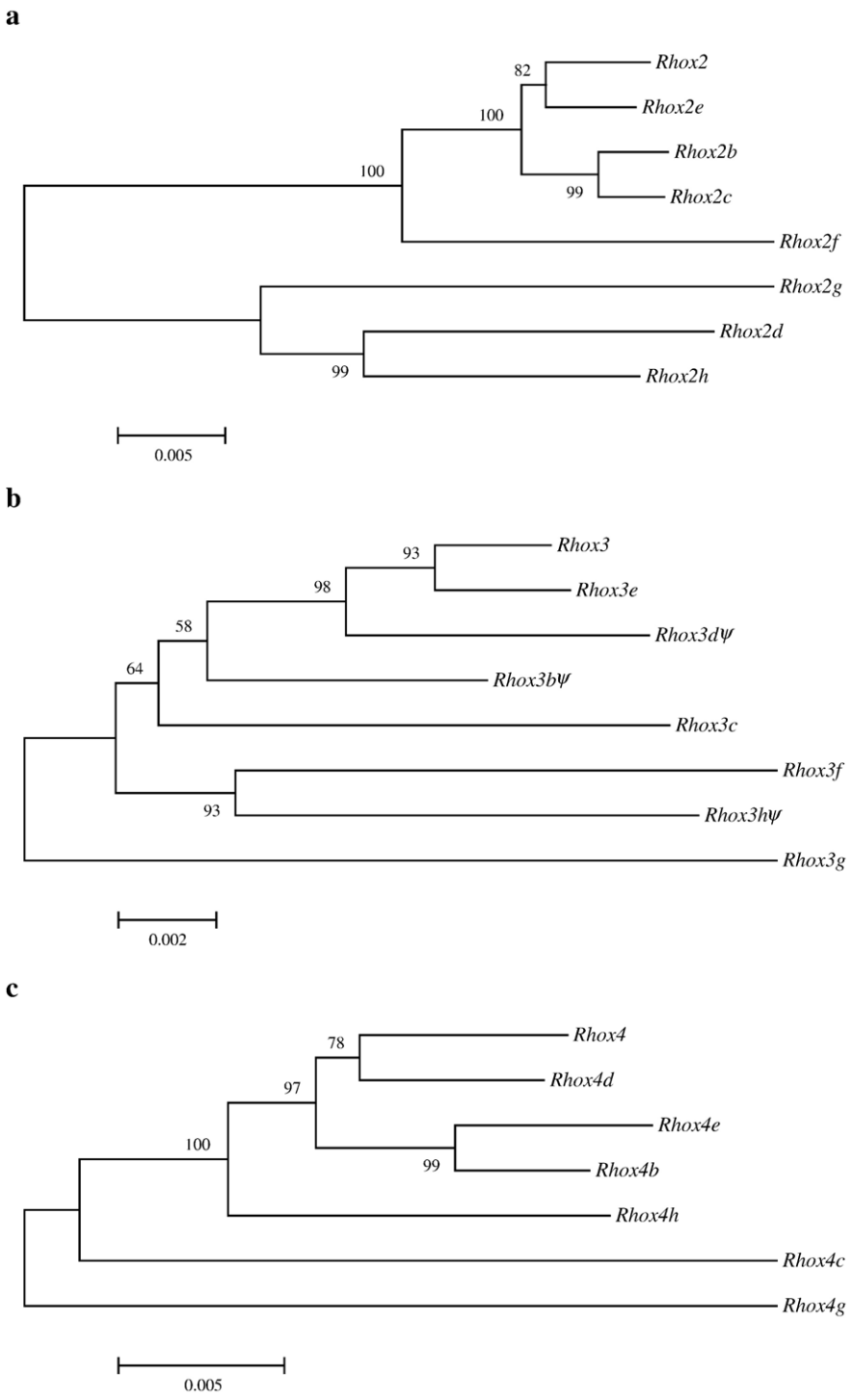


Fig. 4. Phylogenetic relationships of mouse (a) *Rhox2* genes, (b) *Rhox3* genes, and (c) *Rhox4* genes. The trees were reconstructed with the neighbor-joining method and Kimura's two-parameter distance, based on the entire gene sequences (exons and introns). A total of 3012, 3907, and 4436 nucleotide sites were used for the three trees, respectively, after the removal of alignment gaps. Bootstrap percentages (≥50) are shown on interior branches of the trees. The maximum-likelihood trees of these genes are presented in Supplementary Fig. 4.

(Fig. 7) suggest that the newly identified 21 *Rhox2/3/4*-related genes and pseudogenes arose by a series of block duplications of a three-gene unit in the mouse lineage since its separation from the rat lineage. Although only 3 of the 24 *Rox2/3/4*-related genes became pseudogenes, one wonders whether all the genes with ORFs have physiological functions in mice. Several lines of evidence suggest that the answer is yes. First, we detected the action of positive selection in *Rhox4*-related genes, which suggests adaptive functional diversification among the paralogs. Second, we estimated that for *Rhox2* genes, *Rhox3* genes, and *Rhox4* genes, the mean half-life is 2.65, 2.35, and 2.44 MY, respectively, based on an analysis using the PSEUDOGENE

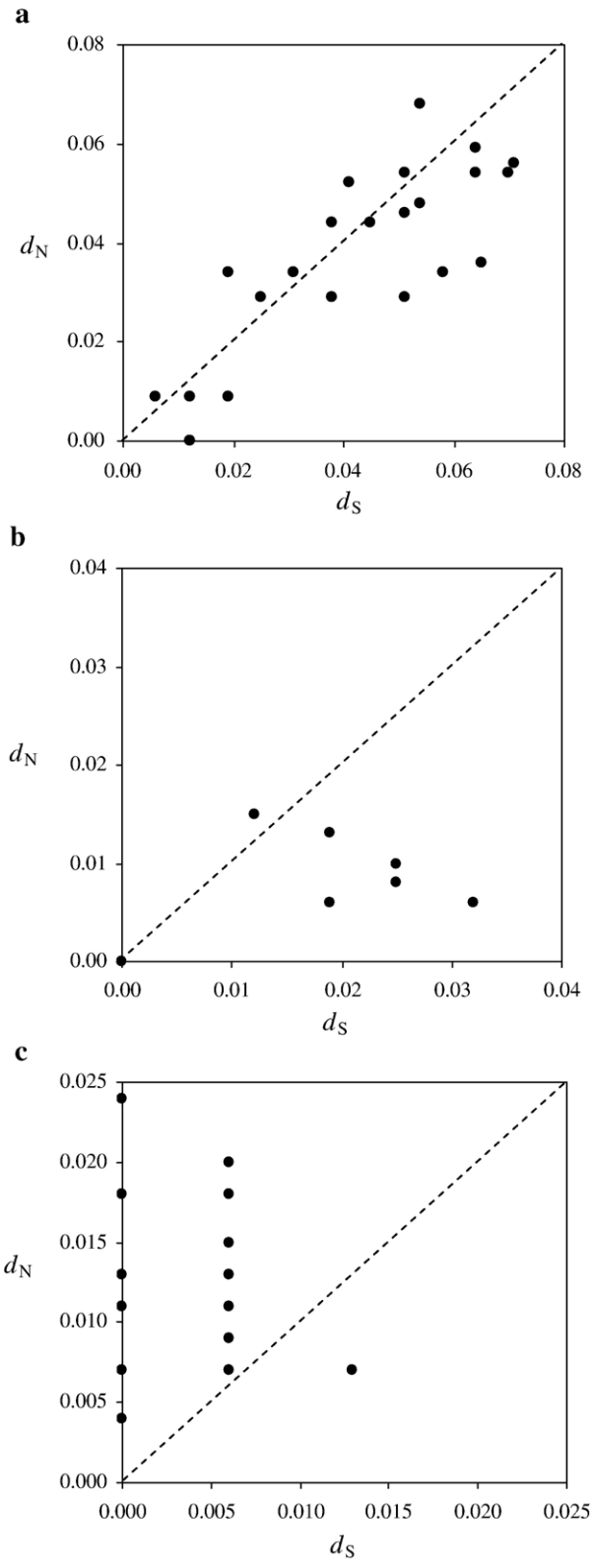


Fig. 5. Pair-wise synonymous ($d_S$) and nonsynonymous ($d_N$) nucleotide distances among mouse (a) *Rhox2* genes, (b) *Rhox3* genes, and (c) *Rhox4* genes (excluding *Rhox4f*). The dotted line indicates $d_N = d_S$.

program [24]. This means that under no functional constraints, 50% of *Rhox2/3/4*-related genes will become pseudogenes in ~2.5 MY. Because the block duplications started at least 7.4 MY ago, more than half of the duplicate genes should have lost ORFs if they had not been under functional constraint. In other words, the retention of the ORFs in 21/24 = 87.5% of *Rhox2/3/4*-related genes suggests that the majority of them are under functional constraint. Third, at least in *Rhox3* genes, and probably also in *Rhox2* genes, purifying selection against nonsynonymous substitutions can be detected, further suggesting that these duplicate genes are not without functional constraint. Finally, we used the 24 *Rox2/3/4*-related genes/pseudogenes as queries to search the mouse EST (expressed sequence tag) database. We were able to identify distinctive ESTs from *Rhox2, 2b, 2c, 2h, 3c, 4, 4e*, and *4g*, but not from the other genes. As expected, the 3 pseudogenes have no corresponding ESTs in the database. Our results strongly suggest that at least 6 of the newly identified *Rhox* genes are expressed.

MacLean et al. [13] showed in mice that within the *Rhox* subcluster α (*Rhox1* to *4*), there is colinearity between the physical location of a gene in the subcluster and the time of gene expression during postnatal testis development. Because the multiple *Rhox2, Rhox3*, and *Rhox4* genes in the mouse genome are highly similar in nucleotide sequence, it is likely that what MacLean et al. measured were total expression levels of multiple closely related genes (e.g., 8 *Rhox2* genes). If this hypothesis is correct, an interesting question is whether the 8 *Rhox2* genes have similar expression patterns. If they do, the colinear regulation is more complex than what MacLean et al. described or that of any *Hox* cluster, because the 8 *Rhox2* genes are not located next to each other in the chromosome. If the expression patterns of these closely related duplicate genes differ, the results of MacLean et al. would require reanalysis and reinterpretation.

Our genomic surveys identified 13 functional *Rhox* genes in the rat (if *Rhox5* is counted) and 3 in the human, but did not find any *Rhox* genes in the dog, cow, opossum, and chicken. Because primates and rodents are more closely related to each other than they are to dog, cow, opossum, and chicken, we infer that the *Rhox* genes originated in the common ancestor of

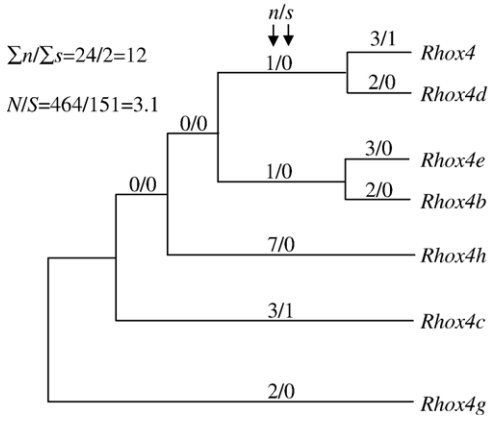


Fig. 6. Numbers of nonsynonymous (*n*) and synonymous (*s*) substitutions in the evolution of mouse *Rhox4* genes (excluding *Rhox4f*). The *n* and *s* values are shown on each branch. *N* and *S* are the potential numbers of nonsynonymous and synonymous sites, respectively. The tree topology is from Fig. 4c.

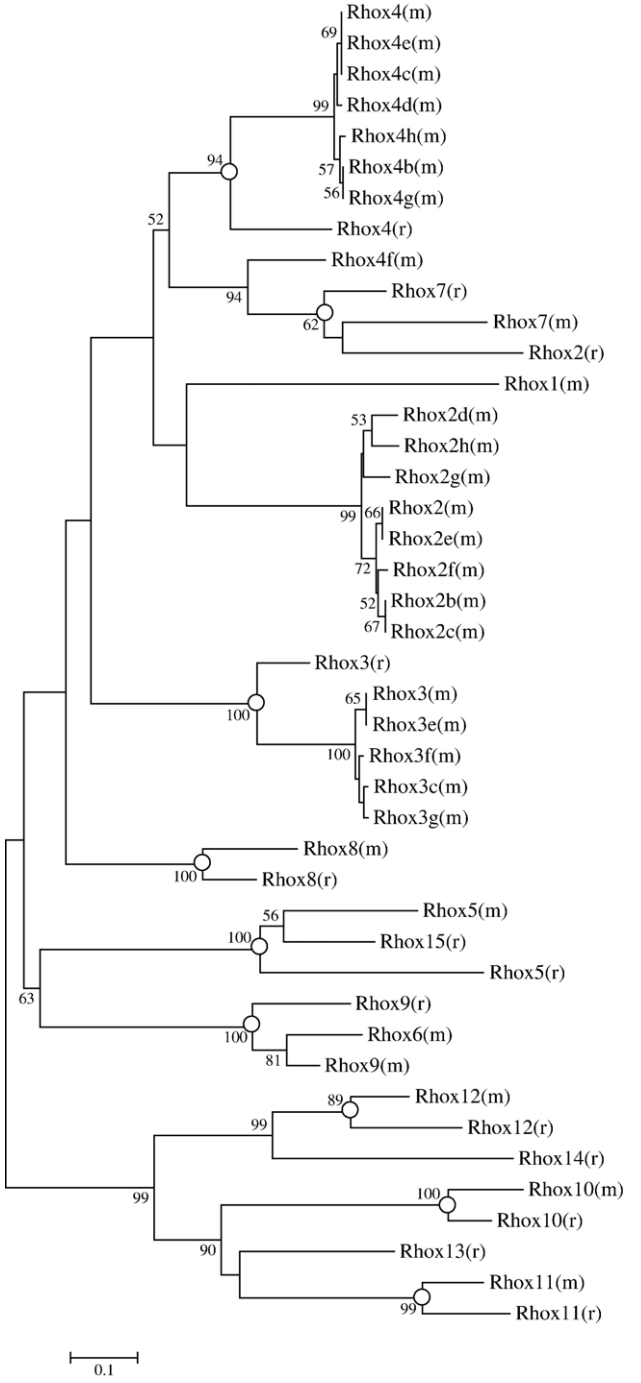


Fig. 7. Phylogenetic relationships of mouse and rat *Rhox* genes. The tree was reconstructed using the neighbor-joining method with Poisson-corrected protein distances. Bootstrap percentages (≥50) are shown on interior branches of the tree. The letters "m" and "r" in parentheses indicate mouse and rat genes, respectively. Circles indicate the nine genes that exist in the common ancestor of the mouse and rat. Note that nine is the minimal estimate. Rat *Rhox14* and *Rhox15* are autosomal. The likelihood tree of the same protein sequences is identical in topology to the tree shown here at all nodes with >70% bootstrap support.

primates and rodents. Our phylogenetic analysis suggests that the most recent common ancestor of primates and rodents had 1 or 2 *Rhox* genes, but the resolution of the gene tree is not high enough for us to determine whether it was 1 or 2 (data not shown). The *Rhox* cluster then underwent two dramatic expansions in rodents. The first expansion occurred before the separation of mice and rats and the second expansion took place in mice after their divergence from rats. The first expansion increased the *Rhox* gene number from 1 (or 2) to at least 9, whereas the second expansion further increased the number to 30 in mice. Grus et al. [25] recently identified several gene families that have great family-size variations among mammalian species. The top gene family on their list is the *V1r* vomeronasal receptor family, which has 8 functional members in the dog, but 187 in the mouse, with a variation of 23-fold. The *Rhox* family is also highly variable in size, as the mouse has 30 functional members, human has 3, and many mammals have none. The general pattern of great size variation of gene families involved in sensory, immune, and reproductive functions [25] appears to hold well.

The functional and evolutionary implications of the recent and massive *Rhox* family expansions are intriguing. All of the previously identified 12 mouse *Rhox* genes are expressed in male and/or female reproductive tissues and some have been shown to play important roles in mouse reproduction [13]. Because some newly duplicated *Rhox* genes have undergone adaptive functional changes, it is likely that the development and physiology of the mouse reproductive system have experienced rapid evolutionary changes in the past few million years. The rapid expansion of the *Rhox* cluster further suggests the possibility that even related rodent species such as the mouse and rat may have substantive differences in reproductive function. On a wider scale, the development of the reproductive system in rodents may differ from that in other mammals (including humans) because of the differences in *Rhox.* Thus, the utility of mouse models for studying mammalian reproduction may be limited.

The colinear gene regulation of members of the *Hox* cluster has been subjected to extensive studies. But, because the *Hox* cluster originated in the early stages of animal evolution, the exact steps leading to the formation of the cluster and the molecular mechanisms required for the establishment of colinear regulation are difficult to discern. The recent formation and expansion of the *Rhox* cluster thus provide an excellent opportunity for understanding the molecular mechanism and evolutionary origin of colinear gene regulation and for understanding the significance of such gene regulation. It would be particularly useful to acquire the structure of the *Rhox* cluster in other species of the *Mus* genus and study how a newly duplicated *Rhox* gene is co-opted into the colinear regulation. It is interesting to note that block duplication might be a convenient and necessary mechanism for generating new genes in the cluster without disrupting colinear regulation, as the colinearity can be retained within blocks.

X-linked testis-expressed homeobox genes were previously shown to exhibit a pattern of rapid evolution at the protein sequence level [26]. *Rhox* genes are no exception, as the average $d_N/d_S$ ratio between orthologous *Rhox* genes of mice and rats is 0.56, much higher than the mean ratio of about 0.1 for all mouse–rat orthologous genes [27]. In *Drosophila*, a rapidly

evolving X-linked testis-expressed homeobox gene (*OdsH*) is in part responsible for the reproductive isolation between *D. simulans* and *D. mauritiana* [28]. Because differential gene duplication or differential duplicate gene evolution in two populations may lead to hybrid incompatibility [29] and because *Rhox* genes function in reproduction, it is possible that the rapid duplication of *Rhox* genes has contributed to rapid formation of reproductive isolation and speciation in rodents, especially in the speciose *Mus* genus.

## Materials and methods

The mouse cDNA and protein sequences of the *Rhox1* to *Rhox12* genes were acquired from MGI (http://www.informatics.jax.org) and MacLean et al. [13]. Additional *Rhox* genes and pseudogenes were obtained by searching the mouse (*Mus musculus*), rat (*Rattus norvegicus*), human (*Homo sapiens*), dog (*Canis familiaris*), cow (*Bos taurus*), opossum (*Monodelphis domestica*), and chicken (*Gallus gallus*) genome sequences available at UCSC Genome Browser (http://genome.ucsc.edu), Ensembl (http://www.ensembl.org), and NCBI (http://www.ncbi.nih.gov) during June–October, 2005. The mouse Rhox2/3/4-related sequences were also used as queries to BLAST-search the mouse EST database in NCBI. BLASTN or TBLASTN [30] with default parameters (*E* value cutoff 0.01) were used for the searches. The exon/intron structures of newly identified genes were deduced by comparing their genomic sequences with the protein sequences of the known *Rhox* genes, using Wise2 (http://www.ebi.ac.uk/Wise2/index.html).

Fragment-specific primers (Supplementary Table 1) for part of the exon 2 in mouse *Rhox2-, Rhox3-*, and *Rhox4*-related genes were designed, respectively, according to the mouse genome sequences. These primers were designed to amplify all *Rhox2*-related genes, *Rhox3*-related genes, and *Rhox4*-related genes, respectively, with the exception of *Rhox4f*. PCRs were performed with MasterTaq (Eppendorf, Hamburg, Germany) under conditions recommended by the manufacturer. PCR products were separated on 1.5% agarose gels and purified using the Gel Extraction Kit (Qiagen, Valencia, CA, USA). Amplified DNA fragments were sequenced from both directions in an automated DNA sequencer using the dideoxy chain termination method. Sequencher (GeneCodes, Ann Arbor, MI, USA) was used to assemble the sequences and to identify nucleotide variations.

Rhox protein sequences were aligned by Clustal W [31] in MEGA 3.0 [32]. The nucleotide sequences were then aligned following the protein alignment. The neighbor-joining [33] and maximum likelihood methods (PAUP*; http://paup.csit.fsu.edu/about.html) were used in tree reconstruction. The bootstrap method [34] with 5000 replications was used to evaluate the reliability of the reconstructed trees. Phylogenetic analyses (except for the likelihood analysis) were conducted in MEGA 3.0. The number of synonymous substitutions per synonymous site ($d_S$) and the number of nonsynonymous substitutions per nonsynonymous site ($d_N$) between homologous genes were estimated using the modified Nei–Gojobori method [20]. The program PSEUDOGENE [24] was used to estimate the half-life of mouse *Rhox* genes under no selective constraint. We used the mutation rate of $2.86 \times 10^{-10}$ per site per year for ORF-disrupting indels, which was estimated from a mouse–rat genomic comparison [35]. It has been estimated that the mean $d_S$ between mouse–rat orthologous genes is 0.19 [27]. Under the assumption that *Mus* and *Rattus* diverged 18 million years ago [27], the genomic neutral substitution rate is $5.28 \times 10^{-9}$ per site per year. It has been reported that the neutral substitution rate on the X chromosome is about 87% of the genomic average in mice [36]. Thus, the neutral substitution rate in X-lined sequences is $4.59 \times 10^{-9}$ per site per year. This rate was used in the PSEUDOGENE analysis, as well as in the molecular dating of *Rhox* duplication events. We tested gene conversions among the eight *Rhox2* genes, eight *Rhox3* genes, and seven *Rhox4* genes (excluding *Rhox4f*) of mice using the program GENECONV [19]. The entire gene sequences, including all exons and introns, were used in the gene conversion test. Ancestral DNA sequences were reconstructed using the Bayesian method [37], which has been shown to be accurate for closely related sequences [38].

## Acknowledgments

We thank Soochin Cho, Wendy Grus, and three anonymous referees for valuable comments. This work was supported by research grants from the National Institutes of Health and the University of Michigan to J.Z.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ygeno.2006.02.007.

## References

[1] S. Banerjee-Basu, A.D. Baxevanis, Molecular evolution of the homeodomain family of transcription factors, Nucleic Acids Res. 29 (2001) 3258–3269.
[2] J. Nam, M. Nei, Evolutionary change of the numbers of homeobox genes in bilateral animals, Mol. Biol. Evol. 22 (2005) 2386–2394.
[3] W.J. Gehring, M. Affolter, T. Burglin, Homeodomain proteins, Annu. Rev. Biochem. 63 (1994) 487–526.
[4] E.B. Lewis, A gene complex controlling segmentation in Drosophila, Nature 276 (1978) 565–570.
[5] W. McGinnis, M.S. Levine, E. Hafen, A. Kuroiwa, W.J. Gehring, A conserved DNA sequence in homoeotic genes of the Drosophila Antennapedia and bithorax complexes, Nature 308 (1984) 428–433.
[6] M.P. Scott, A.J. Weiner, Structural relationships among genes that control development: sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci of Drosophila, Proc. Natl. Acad. Sci. USA 81 (1984) 4115–4119.
[7] D. Duboule, G. Morata, Colinearity and functional hierarchy among genes of the homeotic complexes, Trends Genet. 10 (1994) 358–364.
[8] N.M. Brooke, J. Garcia-Fernandez, P.W. Holland, The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster, Nature 392 (1998) 920–922.
[9] S.B. Carroll, Homeotic genes and the evolution of arthropods and chordates, Nature 376 (1995) 479–485.
[10] J. Garcia-Fernandez, P.W. Holland, Archetypal organization of the amphioxus Hox gene cluster, Nature 370 (1994) 563–566.
[11] C. Kappen, K. Schughart, F.H. Ruddle, Two steps in the evolution of Antennapedia-class vertebrate homeobox genes, Proc. Natl. Acad. Sci. USA 86 (1989) 5459–5463.
[12] J. Zhang, M. Nei, Evolution of Antennapedia-class homeobox genes, Genetics 142 (1996) 295–303.
[13] J.A. Maclean II, et al., Rhox: a new homeobox gene cluster, Cell 120 (2005) 369–382.
[14] S. Maiti, et al., The Pem homeobox gene: rapid evolution of the homeodomain, X chromosomal localization, and expression in reproductive tissue, Genomics 34 (1996) 304–316.
[15] Y.J. Han, A.R. Park, D.Y. Sung, J.Y. Chun, Psx, a novel murine homeobox gene expressed in placenta, Gene 207 (1998) 159–166.
[16] Y.J. Han, Y.H. Lee, J.Y. Chun, Identification and characterization of Psx-2, a novel member of the Psx (placenta-specific homeobox) family, Gene 241 (2000) 149–155.
[17] Y.L. Kang, H. Li, W.H. Chen, Y.S. Tzeng, Y.L. Lai, H.M. Hsieh-Li, A novel PEPP homeobox gene, TOX, is highly glutamic acid rich and specifically expressed in murine testis and ovary, Biol. Reprod. 70 (2004) 828–836.
[18] M. Jackson, J.W. Baird, N. Cambray, J.D. Ansell, L.M. Forrester, G.J. Graham, Cloning and characterization of Ehox, a novel homeobox gene essential for embryonic stem cell differentiation, J. Biol. Chem. 277 (2002) 38683–38692.
[19] S. Sawyer, Statistical tests for detecting gene conversion, Mol. Biol. Evol. 6 (1989) 526–538.

[20] J. Zhang, H.F. Rosenberg, M. Nei, Positive Darwinian selection after gene duplication in primate ribonuclease genes, Proc. Natl. Acad. Sci. USA 95 (1998) 3708–3713.
[21] J. Zhang, S. Kumar, M. Nei, Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes, Mol. Biol. Evol. 14 (1997) 1335–1338.
[22] K.A. Sutton, M.F. Wilkinson, The rapidly evolving Pem homeobox gene and Agtr2, Ant2, and Lamp2 are closely linked in the proximal region of the mouse X chromosome, Genomics 45 (1997) 447–450.
[23] C.M. Wayne, J.A. MacLean, G. Cornwall, M.F. Wilkinson, Two novel human X-linked homeobox genes, hPEPP1 and hPEPP2, selectively expressed in the testis, Gene 301 (2002) 1–11.
[24] J. Zhang, D.M. Webb, Evolutionary deterioration of the vomeronasal pheromone transduction pathway in catarrhine primates, Proc. Natl. Acad. Sci. USA 100 (2003) 8337–8341.
[25] W.E. Grus, P. Shi, Y.P. Zhang, J. Zhang, Dramatic variation of the vomeronasal pheromone receptor gene repertoire among five orders of placental and marsupial mammals, Proc. Natl. Acad. Sci. USA 102 (2005) 5767–5772.
[26] X. Wang, J. Zhang, Rapid evolution of mammalian X-linked testis-expressed homeobox genes, Genetics 167 (2004) 879–888.
[27] R.A. Gibbs, et al., Genome sequence of the Brown Norway rat yields insights into mammalian evolution, Nature 428 (2004) 493–521.
[28] C.T. Ting, S.C. Tsaur, M.L. Wu, C.I. Wu, A rapidly evolving homeobox at the site of a hybrid sterility gene, Science 282 (1998) 1501–1504.
[29] M. Lynch, A.G. Force, The origin of interspecific genomic incompatibility via gene duplication, Am. Nat. 156 (2000) 590–605.
[30] S.F. Altschul, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[31] J.D. Thompson, D.G. Higgins, T.J. Gibson, 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 (1994) 4673–4680.
[32] S. Kumar, K. Tamura, M. Nei, MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment, Brief Bioinform. 5 (2004) 150–163.
[33] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Mol. Biol. Evol. 4 (1987) 406–425.
[34] J. Felsenstein, Confidence-limits on phylogenies—an approach using the bootstrap, Evolution 39 (1985) 783–791.
[35] O. Podlaha, D.M. Webb, P.K. Tucker, J. Zhang, Positive selection for indel substitutions in the rodent sperm protein catsper1, Mol. Biol. Evol. 22 (2005) 1845–1852.
[36] R.H. Waterston, et al., Initial sequencing and comparative analysis of the mouse genome, Nature 420 (2002) 520–562.
[37] Z. Yang, S. Kumar, M. Nei, A new method of inference of ancestral nucleotide and amino acid sequences, Genetics 141 (1995) 1641–1650.
[38] J. Zhang, M. Nei, Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods, J. Mol. Evol. 44 (Suppl. 1) (1997) S139–S146.