Origins of New Male Germ-line Functions from X-Derived Autosomal Retrogenes in the Mouse

Meng-Shin Shiao,† Pavel Khil,‡ R. Daniel Camerini-Otero,‡ Toshihiko Shiroishi,§ Kazuo Moriwaki,|| Hon-Tsen Yu,†¶ and Manyuan Long**

*Department of Ecology and Evolution, University of Chicago; †Institute of Zoology, National Taiwan University, Taipei, Taiwan; ‡Genetics and Biochemistry Branch, National Institute of Diabetes and Digestive and Kidney Disease, National Institutes of Health, Bethesda, MD; §National Institute of Genetics, Mishima, Shizuoka-ken, Japan; ||RIKEN BioResources Center, Tsukuba-shi, Ibaraki, Japan; and ¶Department of Life Science, National Taiwan University, Taipei, Taiwan

Recent literature demonstrates that retrogenes tend to leave the X chromosome and integrate onto the autosomes and evolve male-biased expression patterns. Several selection-based evolutionary mechanisms have been proposed to explain this observation. Testing these selection-based models requires examining the evolutionary history and functional properties of new retrogenes, particularly those that show evidence of directional movement between the X and the autosomes (X-related retrogenes). This includes autosomal retrogenes with parental paralogs on the X chromosome (X-derived autosomal retrogenes) and those retrogenes integrated onto the X chromosomes (X-linked retrogenes). In order to understand why retrogenes tend to move nonrandomly in genomes, we examined the expression patterns and evolutionary mechanisms concerning gene pairs having young retrogenes—originating less than 20 MYA (after mouse–rat split). We demonstrate that these X-derived autosomal retrogenes evolved a more restricted male-biased expression pattern: they are expressed exclusively or predominantly in the testis, in particular, during the late stages of spermatogenesis. In contrast, the parental counterparts have relatively broad expression patterns in various tissues and spermatogenetic stages. We further observed that positive selection is targeting these X-derived autosomal retrogenes with novel male-biased expression patterns. This suggests that such retrogenes evolved new male germ-line functions that may be complementary to the functions of the parental paralogs, which themselves contribute little during spermatogenesis. Such evolutionary changes may be beneficial to the populations. Furthermore, most identified X-related retrogenes have recruited novel adjacent sequences as their untranslated regions (UTRs), suggesting that these UTRs, acquired de novo, may play an important role in establishing new regulatory mechanisms to carry out the new male germ-line functions.

Introduction

Biological diversity relies on the eventual emergence of duplicate genes in genomes. Among the various mechanisms of generating new genes, retropositions have been proposed as one of the major mechanisms in a variety of species (Long et al. 2003; Marques et al. 2005; Wang et al. 2006; Bai et al. 2007). Retrogenes originate at new genomic locations by insertions of reverse-transcribed mRNAs of progenitor or parental functional genes. Generations of retrogenes thus represent various movement patterns due to the different genomic locations of retrogenes and their parental paralogs, that is, an autosomal retrogene with a parental paralog on the X chromosome demonstrates the movement pattern from the X chromosome to the autosomes. Recently, a nonrandom distribution of retrogenes has been reported in the genomes of several species: an excess of autosomal retrogenes generated from counterparts on the X chromosome were observed in *Drosophila* (Betran et al. 2002; Dai et al. 2006; Bai et al. 2007), human, and mouse (Emerson et al. 2004). Compared with retrogenes that moved between autosomes, a significantly higher frequency of retrogene moved out of the X chromosome was identified in those species.

The observed nonrandom distribution of retrogenes (Betran et al. 2002; Emerson et al. 2004) was shown to be associated with a particular expression profile. Retrogenes, especially those that moved out of the X chromosome, were often identified as being expressed specifically or predominantly in the male germ line (male-biased expression patterns; Betrán et al. 2002; Marques et al. 2005). Several mechanisms have been proposed to explain this phenomenon. First, an excess of retrogenes generated from parental genes on the X chromosome could result from a potentially higher overall expression level of genes on the X chromosome (X-linked genes) in the testis. Thus, genes on the X chromosome have a higher probability to become parental genes for retroposition onto autosomes. However, the premise of this hypothesis was not supported by the expression data of the human genome (Emerson et al. 2004). An alternative mechanism was proposed based on the investigation of the evolutionary forces that are acting on newly generated retrogenes. Among various evolutionary forces, natural selection is the main cause in the retention and evolution of retrogenes in genomes: the selection-driven hypothesis suggests that autosomal retrogenes can serve as the functional complement for their counterparts on the X chromosome. During this process, mutations in the regulatory regions that lead to male-specific functions will be selected and a pattern of male-specific expression will be conferred on new genes, as shown in *Drosophila* (Betrán et al. 2002).

Recent studies have proposed that the abnormal function of an X-derived autosomal retrogene would result in severe defects in the male reproductive system in mammals, suggesting a critical role of autosomal counterparts for those genes on the X chromosome (Rohozinski and Bishop 2004; Rohozinski et al. 2006). Additionally, these authors demonstrated a special case whereby the emergence of 2 autosomal retrogenes derived from a single parental paralog on the X chromosome, occurring independently in the lineages of human and mouse. These 2 autosomal retrogenes, *Utp14b* (mouse) and *UTP14C* (human), evolved

Key words: mouse, retroposition, spermatogenesis, male functions.

E-mail: mlong@uchicago.edu; ayu@ntu.edu.tw.

Mol. Biol. Evol. 24(10):2242–2253. 2007
doi:10.1093/molbev/msm153
Advance Access publication July 23, 2007

male-biased expression patterns and were revealed to have crucial male-related functions. *Utp14b* was identified as acquiring testis-specific function and becoming expressed throughout spermatogenesis, whereas the parental paralog on the X can only be detected in the early spermatogenetic stages. The absence of the normal function of *Utp14b* in mouse will result in juvenile spermatogonial depletion, which is believed to be related to male sterility (Rohozinski and Bishop 2004). In humans, *UTP14C* has also been identified as being necessary for normal male fertility (Rohozinski et al. 2006). These findings further support the importance of autosomal retrogenes in the male-related functions.

Several selection-based models have been proposed to be responsible for the biased movements of retrogenes in genomes. First, meiotic sex chromosome inactivation (MSCI) in the male germ line is likely to be an important force in driving genes off the X chromosome (Betrán et al. 2002, 2004; Emerson et al. 2004; Khil et al. 2004; Dai et al. 2006). The X chromosome in the male germ line undergoes inactivation in the late meiotic stages, resulting in the down-regulation of most genes on the X chromosome. Based on this model, we expect that a duplicate retrogene on an autosome with male-biased functions useful in late meiosis would be favored by natural selection due to its functional substitution for the X-linked parental. Second, a sexual antagonism driving X inactivation model (SAXI model; Wu and Xu 2003) was posited differently from the classic model (Rice 1984). Unlike Rice's classic model that predicts a disproportionate number of male-biased genes on the X chromosome, which is inconsistent with the recent observed genomic distributions (Reinke et al. 2000; Betran et al. 2002; Parisi et al. 2003; Ranz et al. 2003; Khil et al. 2004), the SAXI model proposes that the longer sojourn time of the X chromosome in females would lead to feminization of the X and result in movements of the male-biased genes off the X. Genomic expression analysis showed that the mouse X chromosome appears to contain a slight excess of female-specific genes, providing support for the prediction of feminization of the X by sexual antagonism (Khil et al. 2004). The third model shows that a sex-related mutation that is dominant or partially dominant has a higher fixation probability on the autosomes than on the X (Charlesworth et al. 1987). Therefore, male-biased retrogenes that are dominant or partially dominant will have a higher fixation probability on the autosomes than on the X chromosome in the populations. The 3 hypotheses described above all predict a male-biased expression pattern of autosomal retrogenes that were derived from the X chromosome.

Although these hypotheses have been tested or discussed in comparative analyses of sequence and expression data, a direct test of evolutionary mechanisms using genetic variation within a species is necessary. Further understanding of the observed gene movement requires detailed studies on the expression patterns and polymorphism distributions within gene regions in the populations. In attempt to characterize early events during the origination and evolution of these retrogenes, we focused in mouse on a group of young retrogenes showing evidence of movement between the X chromosome and the autosomes (X-related retrogenes that emerged after mouse–rat split). We examined expression differentiation between retrogenes and their parental paralogs. Additionally, we investigated the selection mechanisms involved in the preservation of retrogenes and the divergence in expression between retrogenes and parental paralogs. Finally, we identified the novel structures of retrogenes by obtaining full-length cDNAs.

Material and Methods

Selection of Retrogene Candidates

To identify duplicate retrogenes in mouse genome with the most recent origins, we first conducted the whole-genome database analysis (Materials and Methods in [Emerson et al. 2004]). First, only gene pairs that are reciprocal best hits to each other were selected. Then, further criteria were applied in fishing candidate retrogene pairs: 1) the overlap length is more than 70% to each gene; 2) more than 50% identity in amino acid sequences; 3) moving between chromosomes; 4) one gene possesses one exon, whereas the other one are with multiple exons; 5) both of them have known transcripts. A total of 64 pair of genes was retrieved based on the criteria. These gene pairs include retrogenes with all movement patterns: both the retrogene and the parental paralog are on the autosomes, or one of them is on the X chromosome and the other one is on the autosomes (supplementary table S1, Supplementary Material online).

In order to elucidate the nonrandom movements of retrogenes between the X chromosomes and the autosomes, we conducted different approaches to select retrogene pairs originated or moved onto the X chromosome (X-related retrogenes) with most recent origins. We first applied K_s thresholds filtering for candidate X-related retrogenes. The K_s values of all X-related gene pairs that range from 0.0000 to 1.2112 imply various ages of retrogenes in the mouse genome. In order to select retrogenes with the most recent origins, we thus only preserved retrogenes with K_s ranging from 0.02 to 0.6. The threshold $K_s = 0.02$ ensure the coding regions of retrogenes to be distinguishable from their parental genes. A total of 11 retrogenes with evidence of movements between X chromosome and autosome were identified (table 1).

Then, we carried out Blast searches of retrogene sequences against the genomes of human and rat and phylogenetic screening from various closely related rodent species by polymerase chain reaction (PCR). Blast search was performed by using 11 retrogene genomic sequences as probe and Blast against whole genome of rat and human. A retrogene's origination will be posited in the lineage of mammals' common ancestor if the homologs can be found in both the human and rat genomes. If the homologs can only be identified in the rat but not in the human genome, the retrogenes are most likely lineage specific to rodent species. For those retrogenes have no homologs in either rat or human genome, we further conducted PCR sequencing to screen genomes of 7 rodent species including house mouse, *Mus musculus*. Primer pairs for amplification and sequencing were designed specific to retrogenes in the mouse genome. We chose 4 species including rat within

Table 1
The Movement of Retrogenes between the X Chromosome and the Autosomes ($0.02 < K_s < 0.6$)

Retrogene			Parental Gene						
Ensembl Gene ID	Gene Name	Location	Ensembl Gene ID	Gene Name	Location	Length[a]	K_a/K_s	K_s	Distributions
ENSMUSG00000063724	*XP_484661.1*	X	ENSMUSG00000025290	*Rps24*	14	390	0.0921	0.0439	Mammals
ENSMUSG00000055936	*AU015836*	X	ENSMUSG00000034203	*Chchd4*	6	351	0.3474	0.5041	Rodent
ENSMUSG00000058670	*Dmtf1-R*	X	ENSMUSG00000042508	*Dmtf1*	5	693	0.4126	0.5483	Rodent
ENSMUSG00000049576	*Zfa*	10	ENSMUSG00000000103	*Zfx*	X	2226	0.5147	0.0373	*Mus*
ENSMUSG00000059695	*MusT-R*	12	ENSMUSG00000067647	*MusT1*	X	351	0.7751	0.0890	*Mus*
ENSMUSG00000047995	*Cypt9*	9	ENSMUSG00000033856	*Cypt2*	X	459	1.0021	0.0997	*Mus*
ENSMUSG00000026063	*Ny-sar-97-R*	1	ENSMUSG00000042433	*Ny-sar-97*	X	654	0.2374	0.1633	*Mus–Apodemus*
ENSMUSG00000050035	*Fhl4*	10	ENSMUSG00000023092	*Fhl1*	X	837	0.4363	0.4231	Rodent
ENSMUSG00000042668	*Tspan7-R*	7	ENSMUSG00000058254	*Tspan7*	X	738	0.4734	0.5093	*Mus–Apodemus*
ENSMUSG00000059395	*4921504I05Rik*	13	ENSMUSG00000016409	*2610020O08Rik*	X	1185	0.5395	0.5273	Mammals
ENSMUSG00000039224	*D1Pas1*	1	ENSMUSG00000000787	*Ddx3x*	X	1980	0.0522	0.5417	Mammals

NOTE.—Mammals: retrogenes sequences are shared by mouse, rat, and human, using Blast search; Rodents: retrogenes sequences are shared by mouse and rat, using Blast search; *Mus*: retrogenes sequences are only shared by closely related species to mouse, using PCR screening; *Mus–Apodemus*: retrogenes sequences are shared by genus *Mus* and *Apodemus semotus* but not other rodent species.

[a] Number of overlapping base pairs between retrogene and parental gene.

Murinae those are closely related to mouse and rat (*Mus spretus*, *Mus caroli*, *Apodemus semotus*, and *Rattus norvegicus*) and 2 species outside Murinae as outgroup (*Meriones unguiculatus* and *Phodopus campebelli*). Taken all methods together, 3 retrogenes are proposed as originating before mammals' divergence, 3 of them are shared by rodent species (include mouse and rat), and 5 of them originated recently after mouse–rat split. Among the 5 retrogenes with the most recent origins, 3 of them are *Mus* lineage specific, whereas 2 of them are shared by closely related species in genus *Apodemus* (table 1).

Expression Analyses and Gene Structure Identifications

The spatial expression patterns of candidate genes were identified by performing reverse transcriptase–PCR (RT–PCR) in testis, ovary, brain, and liver. Commercial premade cDNAs from 4 tissues were purchased from Ambion, Inc. (Austin, TX).

The temporal expression patterns of candidate genes were analyzed by using quantitative RT–PCR (qRT–PCR). Two mutant strains were selected in the study: homozygous knock out of *Spo11* (*Spo11*$^{-/-}$) (Romanienko and Camerini-Otero 2000) and *mei1* mutant (Libby et al. 2002, 2003), which have similar phenotype: spermatogenesis of those individuals with homozygous mutation is arrested in the early stage of meiosis (Romanienko and Camerini-Otero 2000; Libby et al. 2002, 2003). Two wild types derived from the same genetic background with mutant strains were also analyzed as comparisons. The expression levels are represented by threshold cycles for detection, $C_t \sim -\log_2 (C_0/A)$; where A is constant and C_0 is starting target concentration. C_t around 30 cycles indicates the gene has very low expression level. Primers of RT–PCR and qRT–PCR for each gene are listed in supplementary table S2 (Supplementary Material online). For details on RNA purification and qRT–PCR procedures see Smirnova et al. (2006).

Gene structures are identified by rapid amplification of cDNA ends (RACE). RACE-ready testicle cDNA was also purchased from Ambion. Nested PCR steps followed First-Choice RACE-Ready cDNA Instruction Manual. Primers designed for amplifying both ends are listed in supplementary table S2 (Supplementary Material online).

DNA Sequence Analyses

Gene surrounding and flanking regions of candidate genes in the mouse natural populations were amplified by PCR from genomic DNA. Population samples were collected from South Asia including China, Philippine, Indonesia, and Malaysia. A total of 48 samples were used in this study. The final sample numbers shown in table 3 vary due to the sequencing quality. Primers are designed specific to the house mouse genomic sequences released in Ensembl database (http://www.ensembl.org/index.html) and are specific to each gene surrounding regions (supplementary table S2, Supplementary Material online). Besides, although heterozygous sites were identified in the population, cloning would be carried out for heterozygous individuals. We randomly picked up 2 colonies and performed sequencing. Furthermore, we manually generate duplicate identical sequences for the rest of homozygous samples.

To measure skewness in the frequency spectrum of polymorphic sites, Tajima's *D* test (Tajima 1989) was obtained in these regions. In this standard test of the null hypothesis of the theory of neutral evolution (Kimura 1983), the deviation of the allele frequencies in each polymorphic site from the prediction of neutrality was estimated and tested. Besides, haplotype diversity test (*H* test) and haplotype number test (*K* test; Depaulis and Veuille 1998) for the hypothesis of neutrality was applied and calculated for *Tspan7-R* and *Tspan7*. All the tests were carried out in DnaSP 4.0 package (Rozas et al. 2003). Probability values for all statistical tests were obtained out of 10,000 coalescence simulation (Hudson 1990). Furthermore, *Tspan7-R* sequences in *M. caroli* were amplified by the primers designed specific to *M. musculus Tspan7-R*. *M. caroli* genomic DNA was purchased from the Jackson Laboratory.

Results

X-Related Young Retrogenes Evolved More Restricted Male Functions

The studies of young retrogenes address the most fundamental question of how novel genes emerge in genomes by uncovering the early history of retrogenes. The previous work has investigated retrogenes of all ages in the mouse genome (Emerson et al. 2004). In this study, we focus on the analyses of retrogenes with the most recent origins, that is, emerged after mouse–rat split, and with the movement patterns between the X chromosome and the autosomes in the mouse genome. This includes autosomal retrogenes with parental paralogs on the X chromosome (X-derived autosomal retrogenes) and those on the X chromosome with counterparts on the autosomes (X-linked retrogenes).

We applied K_s threshold filtering, Blast searching, and PCR screening to select X-related retrogenes with the most recent origins (see Material and Methods). Among 11 X-related retrogenes selected by the above criteria, we finally determined 5 retrogenes that are lineage specific to genus *Mus* or are shared between genus *Mus* and genus *Apodemus*. This indicates the young age of these X-related retrogenes, which emerged after mouse–rat split, about 20 Myr (O'hUigin and Li 1992; table 1). The 5 retrogenes are as follows: 1700019M22Rik (we designated as *Mus*-specific testicular retrogene, *MusT-R*), cysteine-rich perinuclear theca 9 (*Cypt9*), autosomal zinc finger protein (*Zfa*), *Ny-sar-97*-derived retrogene (*Ny-sar-97-R*), and tetraspanin 7–derived retrogene (*Tspan7-R*). Interestingly, none of the X-linked retrogene was identified as younger than the mouse–rat split, suggesting an early antedating origin of the mammalian sex chromosomes. All 5 newly evolved retrogenes, including a well-characterized one, *Zfa* (Ashworth et al. 1990; Luoh and Page 1994; Banks et al. 2003; Kumar et al. 2004), are derived from parental paralogs on the X chromosome. This further supports the previous conclusion (Emerson et al. 2004) that the directional gene movement, from the X chromosome to the autosomes, is a recent and ongoing process.

Although none of the X-linked retrogenes were determined as being generated recently, we chose one X-linked retrogene pair, cyclin D–binding myb-like transcription factor 1 (*Dmtf1*) and its derived retrogene (*Dmtf1-R*), as a reference to be examined together with the other 4 X-derived autosomal retrogenes pairs. Besides, *Zfa* and *Zfx* were not included in this study because the expression patterns of these 2 genes have been reported (Ashworth et al. 1990; Mardon et al. 1990; Erickson et al. 1993).

To detect the functionality of retrogenes, we applied K_a/K_s-based selection analyses in the orthologous comparisons between species. Three retrogenes (*Ny-sar-97-R*, *Tspan7-R*, and *Dmtf1-R*) show significantly strong selective constraints from K_a/K_s comparisons between mouse and other rodent species (data not shown). K_a/K_s ratios of *Cypt9* between the rodent species demonstrate a sign of selective constraint (K_a/K_s ranging from 0.01 to 0.4 but not significant, data not shown). The lack of significance between gene *Cypt9* and its orthologs may be due to the recent origination of the retrogene and the short length of coding regions. Another young retrogene, *MusT-R*, has high K_a/K_s ranging from 0.6 to 1.8 between mouse and the closely related rodent species (data now shown). Although we do not have strong evidence for the selective constraint on the sequence substitutions in this case (the retrogene copy originated less than 5 MYA), the coding sequence retains in frame without immature stop codon, suggesting that *MusT-R* in the mouse genome may be functional.

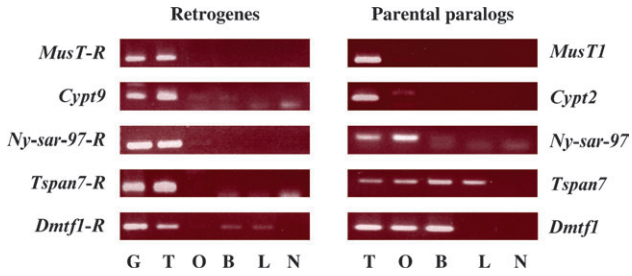

FIG. 1.—Spatial expression patterns of the retrogenes and the parental paralogs. The left panel shows the amplified signals of retrogenes from genomic DNA (G) and expression patterns of retrogenes from testis (T), ovary (O), brain (B), liver (L), and negative control (N). One the right panel shows expression patterns of correspondent parental paralogs.

The expression patterns for these 5 gene pairs were then analyzed from 2 aspects: 1) we examined the spatial expression patterns by carrying out RT–PCR from 4 different mouse tissues: testes, ovary, brain, and liver. Retrogenes are most likely expressed in these 4 tissues based on the summary of expression patterns of known retrogenes (Emerson et al. 2004; table S5, Supplementary Material online). 2) We defined temporal expression patterns of each gene in the mouse male germ line by estimating the expression discrepancy between mutant strains and wild type. In the 2 mutant strains we selected, *Spo11*$^{-/-}$ knockout (Romanienko and Camerini-Otero 2000) and *mei1* meiotic mutant mice (Libby et al. 2002, 2003), spermatogenesis is blocked early in meiosis. When a retrogene is expressed predominantly in the late stages of spermatogenesis, the mRNA concentration will be significantly higher in the wild type than in the mutant strains due to the lack of meiotic cells in mutants. In contrast, expression level will be the same or higher in the mutant strains for a gene expressed in the early stages or expressed both early and late stages in spermatogenesis. In addition, we used a housekeeping gene, β-*actin*, on the autosome as a control, which is universally expressed during spermatogenesis.

We observed that retrogenes are expressed with a more restricted pattern than parental paralogs. Retrogenes are all expressed predominantly in testis and also late in the spermatogenetic stages. In contrast, the parental paralogs have various spatial and temporal expression patterns (figs. 1 and 2). Among the 5 retrogenes, 4 X-derived autosomal retrogenes are all exclusively expressed in the testis. The only X-linked retrogene, *Dmtf1-R*, is expressed predominantly in testis although low level of mRNA could also be detected in the brain and liver. In contrast, 5 parental paralogs are expressed in a relatively broad pattern: 2 genes, 1700019M22Rik (we designated it as *Mus*-specific testicular gene 1, *MusT1*) and cysteine-rich perinuclear theca 2 (*Cypt2*), are testis-specific expressed. The other 3 parental

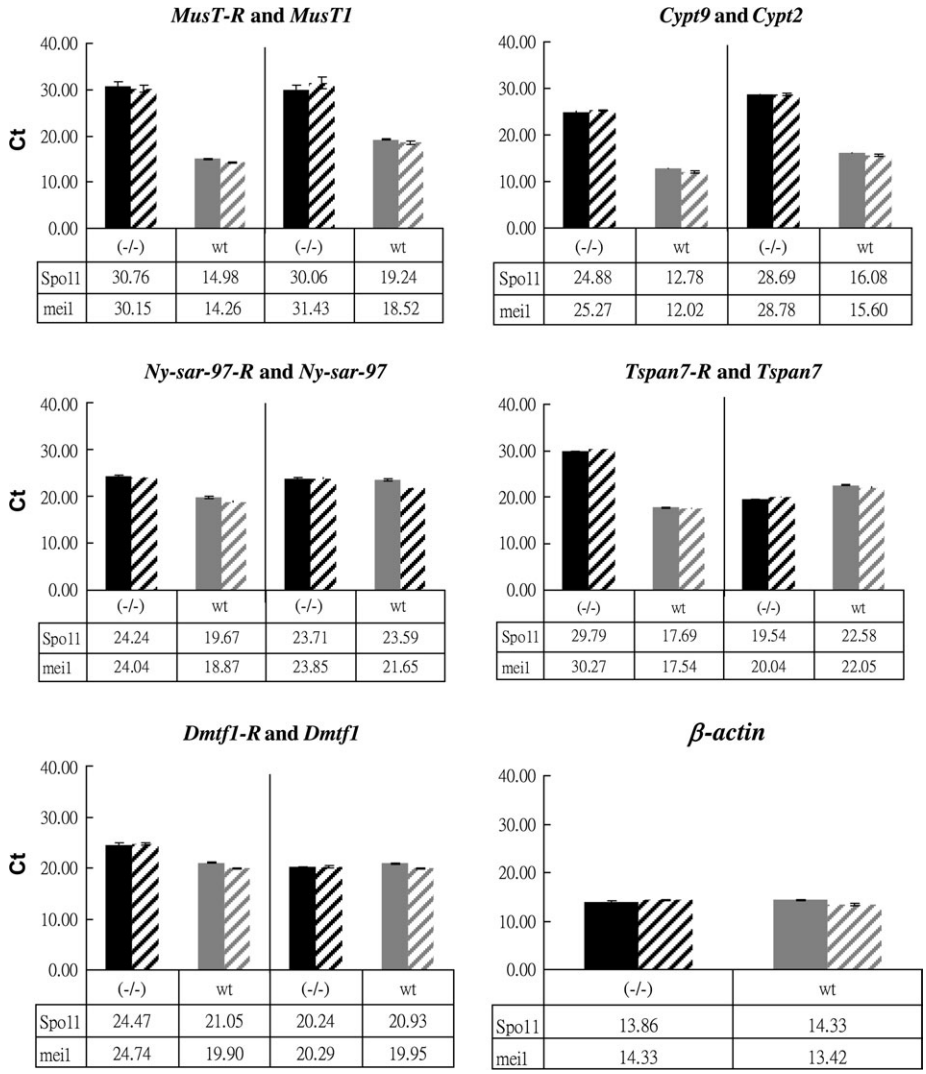

FIG. 2.—Temporal expression patterns of retrogene pairs in spermatogenesis between mutant (−/−) and wild-type (wt) strains. Histograms demonstrate average C_t values (also listed below the bars) and standard deviation for each gene. C_t values represent the threshold cycles for detecting mRNA concentration, $C_t \sim -\log_2 (C_0/A$; where A is constant and C_0 is starting target concentration). Black bars represent 2 mutant strains (solid bars—*Spo11* knockout, slashed bars—*mei1* mutant) and gray bars indicate wild-type strains with identical genetic background (solid bars—wild type for *Spo11* knockout, slashed bars—wild type for *meu1* mutant. Two wild-type strains used are of different genetic background). Retrogenes and parental paralogs are demonstrated separately in left and right panels of each chart for comparisons. Housekeeping gene, β-*actin*, was amplified simultaneously as control.

paralogs, sarcoma antigen *Ny-sar-97* (*Ny-sar-97*), tetraspanin 7 (*Tspan7*), and cyclin D–binding myb-like transcription factor 1 (*Dmtf1*), are expressed ubiquitously (being expressed in more than 2 tissues; fig. 1).

Furthermore, the entire 5 retrogenes evolved the late expression pattern during spermatogenesis regardless of the locations, whereas the parental genes, again, are expressed in a more various pattern. The expression is detected in the early stage (*Tspan7*), in late stage (*MusT1* and *Cypt2*), or throughout all spermatogenetic stages (*Ny-sar-97* and *Dmtf1*) for parental paralogs (fig. 2). Expression patterns are summarized in table 2.

A "novel" expression pattern can be defined by comparing the expression patterns between the retrogenes and the parental paralogs. For example, a retrogene expressed only in the male germ line (male biased) can be stated as evolving novel expression patterns, whereas its parental paralog is expressed ubiquitously among tissues that were analyzed. Based on this definition, we found that not all the retrogenes have evolved novel, particularly male-biased, expression patterns (table 2). Retrogene, *Ny-sar-97-R*, is expressed exclusively in testis and in the late spermatogenetic stage, whereas parental paralog, *Ny-sar-97*, is expressed in both testis and ovary and throughout entire spermatogenesis. In addition, *Tspan7-R* and *Dmtf1-R* are both expressed exclusively in the male germ line and in the late stages of spermatogenesis, whereas their parental paralogs are expressed ubiquitously in spatial and temporal. Taken together, we identified 3 retrogenes with diverged novel expression patterns from their parental paralogs either among different tissues or spermatogenetic stages.

Table 2
Spatial and Temporal Expression Patterns

		Expression Patterns				Expression Patterns	
Retrogenes	Location	Tissue	Spermatogenesis Stage	Parental Paralogs	Location	Tissue	Spermatogenesis Stage
MusT-R	Ch 12	Testis	Late	*MusT1*	Ch X	Testis	Late
Cypt9	Ch 9	Testis	Late	*Cypt2*	Ch X	Testis	Late
Ny-sar-97-R	Ch 1	Testis	Late	*Ny-sar-97*	Ch X	Gonad exclusive	No difference
Tspan7-R	Ch 7	Testis	Late	*Tspan7*	Ch X	Ubiquitous	Early
Dmtf1-R	Ch X	Testis/brain/liver	Late	*Dmtf1*	Ch 5	Ubiquitous	No difference
Control							
β-*actin*		No difference					

Selection Is Targeting the Autosomal Retrogenes with Novel Male-Biased Functions and Their Parental Paralogs

To understand the evolutionary forces acting on gene pairs, we conducted a population genetic approach for the natural populations of the subspecies *Mus musculus castaneus*. The subspecies was identified as genetic unique and distantly related to other subspecies of *M. musculus* (Yonekawa et al. 1981). The subspecies thus provides a good source for studying evolutionary forces. Genetic variations of DNA sequences in the natural populations can be estimated by 2 different parameters: the number of segregating sites (S) and the average number of nucleotide differences by pairwise comparison (π). Tajima's D tests were performed by estimating the differentiations between these 2 parameters of genetic variations (Tajima 1989). If a strong selection against changes is acting on the sequences, there will be an excess of polymorphic sites toward rare alleles (e.g., singleton; Kimura 1983).

The polymorphic distributions of the 5 retrogenes and the 5 parental paralogs were obtained from the house mouse natural populations in Asia for statistical analyses. Considering that due to the short length of retrogenes, that is, *MusT-R* and *Cypt9* (354 and 629 bp, respectively), may lack power in statistical tests, we thus designed one of the primers in the upstream or downstream gene surrounding regions to amplify longer sequences including the coding regions. This will result in a sequence length of about 650 bp for *MusT-R* and 830 bp for *Cypt9*. Besides, to rule out selection sweep from the flanking regions and demographic effects in the evolution history of the house mice populations, we further examined the intergenic flanking regions those are 3–10 kb away from the gene surrounding regions. The intergenic flanking regions in one chromosome represent the evolution history of the entire chromosome, as for the history of the populations. If a population has been subject to bottleneck effects followed by rapid expansions, a distribution of rare alleles would be expected from the examined intergenic flanking regions.

Remarkably, our results reveal that those autosomal retrogenes that moved out of the X chromosome and evolved novel male-biased expression patterns are subject to positive selection (tables 3 and 4). No selection force was determined for those retrogenes which they and their parental paralogs have similar expression patterns, that is, *MusT-R* and *MusT1*. Among the 5 retrogenes examined in this study, natural selection is only detectable for 2 retrogenes, *Ny-sar-97-R* and *Tspan7-R*. Surprisingly, for these 2 gene pairs, both retrogenes and parental paralogs are subject to natural selection. The selection effects on these 2 gene pairs were revealed either by the access of rare polymorphic site (table 3) or by the nonuniform distribution of polymorphic sites in the population (table 4). Overall, neither the gene pairs with identical expression patterns nor those with X-linked retrogenes showed significant evidence of being subject to selection.

Table 3
Tajima's *D* and Estimated Parameters of Retrogenes and Their Parental Copies

		5′ Flanking				Gene				3′ Flanking			
	Location	*D*	*S*	*L*	*N*	*D*	*S*	*L*	*N*	*D*	*S*	*L*	*N*
MusT-R	Ch 12	n/a	0	1044	16	−1.1492**	1	649	13	0.3701	6	580	17
MusT1	Ch X	—	—	—	—	−1.1492**	1	676	13	—	—	—	—
Cypt9	Ch 9	1.5476	1	868	18	−1.0600	6	833	26	−0.4662	2	1483	13
Cypt2	Ch X	n/a	0	1128	11	−0.4430	2	589	27	−0.0283	3	1177	12
Ny-sar-97-R	Ch 1	2.2922	5	1048	10	−1.5341*	11	897	32	−0.7100	19	719	13
Ny-sar-97	Ch X	n/a	0	1649	30	−1.6488*	6	1073	34	−1.5074**	2	984	30
Tspan7-R	Ch 7	−1.5622*	3	765	10	0.2208	9	729	24	−0.7065	7	633	26
Tspan7	Ch X	0.6117	3	875	14	1.3083	11	1288	12	0.9940	4	1216	14
Dmtf1-R	Ch X	0.0831	8	1116	20	−0.1816	3	912	31	n/a	0	1115	11
Dmtf1	Ch 5	0.9087	9	773	30	−0.1890	4	735	21	−1.5074	7	1756	13

NOTE.—n/a, not applicable.

[a] The lack of *MusT1* flanks is due to large proportions of repetitive sequences in both 5′ and 3′ flanking regions.

* $P<0.05$.

** $P<0.01$.

Table 4
Statistic Tests for Haplotype Structures of Gene Pair *Tspan7-R* versus *Tspan7*

Test	Parameters	Observed Values	
		Tspan7-R	*Tspan7*
H test	Haplotype diversity (H_d)	0.493*	0.719
K test	Haplotype number (*K*)	5	4*

* $P < 0.05$.

A total of 11 polymorphic sites were detected in the retrogene *Ny-sar-97-R* in the house mouse populations (table 5). A statistically significant Tajima's *D* ($D = -1.5809$, $P < 0.05$, table 3) in the gene surrounding regions of *Ny-sar-97-R* indicates a deviation from the prediction of neutrality. This suggests that the gene is likely subject to a directional positive selection. Alternatively, the polymorphism distribution could also result from a demographic process such as a recent bottleneck effect in the population. However, this effect could be rejected by the observation of an excess of polymorphisms with intermediate frequencies in the 5′ flanking region of gene *Ny-sar-97-R* (Tajima's $D = 2.2922$, $P < 0.01$) and no departure from neutrality in the 3′ flank (Tajima's $D = -0.7100$, $P > 0.05$; table 3). Taken together, the data indicate that the selection force is unique to the gene region that is not associated with the evolutionary history of a particular chromosome or hitchhiking effects from the flanking regions. As for the counterpart on the X, *Ny-sar-97*, the same evolutionary force was noticed by 6 polymorphic sites (Tajima's $D = -1.6488$, $P < 0.05$). Due to the general low polymorphism on the X chromosome from our results, we combined 2 regions flank the gene *Tspan7* to represent intergenic regions on X. There are total 2092 bp with 8 segregating sites defined within the joint region. The neutrality cannot be rejected by the polymorphism distributions (Tajima's $D = 1.0963$, $P > 0.05$). Therefore, we suggest that the X chromosome in the *M. musculus castaneous* natural population is not subject to demographic effects. We thus conclude that the retrogene, *Ny-sar-97-R*, together with its parent, *Ny-sar-97* on the X, are subject to natural selection that is unique to the gene surrounding regions.

The evidence of natural selection, most likely balancing selection, on *Tspan7-R* and *Tspan7* was detected from nonuniformly distributed polymorphic sites in the population (tables 4 and 5). Based on the assumptions of the *K* test

Table 5
Polymorphic Distributions of *Ny-sar-97-R/Ny-sar-97* and *Tspan7-R/Tspan7*

Ny-sar-97-R		*Ny-sar-97*		*Tspan7-R*		*Tspan7*	
	1 24566677880 85225778267 39076131898		**1 391 555930 012509**		**444457788 077772706 826739772**		**1111111 45590000112 95615677368 37693707769**
MG656.1	CACGAGTCTAC	MG728	CATTAT	MG794.1	ATTTAGCAA	MG5120	AGAACAACTAA
MG656.2	....	MG421	.....	MG902.1	....	MG5121	....
MG727.1	....	MG450	...A..	MG906.1	....	MG504	.A.........
MG727.2	....	MG5058	.....	MG794.2	.C.......	MG712	.A.........
MG728.1	....	MG5059	.....	MG452.1	....CA...	MG902	.A.........
MG728.2	....	MG5122	.....	MG452.2	....CA...	MG5058	.A.........
MG733.1	....	MG5123	.....	MG795.1	G.C.C.GGG	MG5059	.A.........
MG733.2	....	MG5124	....C.	MG795.2	G.C.C.GGG	MG793	.AGG..TTCG.
MG709.2	....	MG5202	ATC...	MG797.1	G.C.C.GGG	MG794	.AGG..TTCG.
MG711.1	....	MG5203	.....	MG797.2	G.C.C.GGG	MG903	.AGG..TTCG.
MG711.2	....	MG733	.....	MG902.2	G.C.C.GGG	MG904	.AGG..TTCG.
MG712.1	....	MG503	.....	MG904.1	G.C.C.GGG	MG906	.AGG..TTCG.
MG712.2	....	MG534	.....	MG904.2	G.C.C.GGG	MG907	.AGG..TTCG.
MG5120.1	....	MG734	.....	MG906.2	G.CCC.GGG	MG97	TA.GGGT.C.G
MG5120.2	....	MG709	.....A	MG907.1	G.C.C.GGG	MG421	TA.GGGT.C.G
MG5121.1	....	MG710	.....	MG907.2	G.C.C.GGG	MG450	TA.GGGT.C.G
MG5121.2	....	MG902	.....	MG2103.1	G.C.C.GGG	MG452	TA.GGGT.C.G
MG793.1	....	MG903	.....	MG2103.2	G.C.C.GGG	MG795	TA.GGGT.C.G
MG793.2	....	MG904	.....	MG2104.1	G.C.C.GGG	MG797	TA.GGGT.C.G
MG794.1	....	MG906	.....	MG2104.2	G.C.C.GGG	MG2104	TA.GGGT.C.G
MG794.2	....	MG907	.....	MG504.1	G.C.C.GGG	MG5122	TA.GGGT.C.G
MG795.1	....	MG908	.....A	MG504.2	G.C.C.GGG	MG5123	TA.GGGT.C.G
MG795.2	....	MG909	.....	MG532.1	G.C.C.GGG	MG5203	TA.GGGT.C.G
MG503.1	......C....	MG939	.....	MG532.2	G.C.C.GGG		
MG503.2	......C....	MG925	.....A				
MG734.1	....C......	MG951	.....A				
MG734.2	....C......	MG2103	.....				
MG709.1	.....T....T	MG2107	.....A				
MG710.1	.G...T..C..	MG2110	.....A				
MG450.1	T..T.T.T.G.	Yu695	.....				
MG450.2	T..T.T.T.G.	Yu738	.....				
MG710.2	T.TT.T.T...	Yu756	.....				
		Yu776	.....				
		Yu782	.....				

Table 6
Haplotypes of *Tspan7-R*

	444457788 077773706 937840883
M.caroli	ATTTCGCAA
Haplotype group A	
MG794.1	....A....
MG902.1	....A....
MG906.1	....A....
MG794.2	.C.......
MG452.1	.....A...
MG452.2	.....A...
Haplotype group B	
MG795.1	G.C...GGG
MG795.2	G.C...GGG
MG797.1	G.C...GGG
MG797.2	G.C...GGG
MG902.2	G.C...GGG
MG904.1	G.C...GGG
MG904.2	G.C...GGG
MG906.2	G.CC..GGG
MG907.1	G.C...GGG
MG907.2	G.C...GGG
MG2103.1	G.C...GGG
MG2103.2	G.C...GGG
MG2104.1	G.C...GGG
MG2104.2	G.C...GGG
MG504.1	G.C...GGG
MG504.2	G.C...GGG
MG532.1	G.C...GGG
MG532.2	G.C...GGG

and the *H* test, an unexpected low haplotype number or haplotype diversity is evidence for positive selection on certain haplotypes (Depaulis and Veuille 1998; Wang et al. 2004). That is, polymorphic sites would not be able to recombine randomly under selection forces and will result in an unexpected low haplotype number or haplotype diversity. A total of 5 haplotypes from 9 polymorphic sites were detected within *Tspan7-R* with a haplotype diversity of 0.493. Although haplotype number of *Tspan7-R* does not show evidence of skewness due to the 2 singletons in individuals MG749.2 and MG906.2, the haplotype diversity is indeed significantly lower than expected ($P < 0.05$, table 4). This is in congruence with a slight excess of polymorphic sites with intermediate frequencies, as indicated by positive Tajima's *D* value (0.2208, table 3). The flanking regions of retrogene *Tspan7-R* also suggest that the evolution force is acting solely on the gene surrounding regions. Similarly, only 4 haplotypes were defined from 11 polymorphic sites in parental gene, *Tspan7*, which is also unexpectedly low ($P < 0.05$, table 4). It is likely that these haplotypes of *Tspan7* are also under balanced positive selection as its retrogene counterpart, which is consistent with the excess of intermediate-frequent polymorphisms, as indicated by a positive Tajima's *D* value (1.3083, $P > 0.05$).

A further comparison between *M. musculus castaneus* and *M. caroli* indicates that the retrogene, *Tspan7-R*, is mainly composed of 2 haplotype groups, A and B (table 6). We sequenced both chromosomes from every wild-type individual due to the heterozygosity in the gene surrounding region of *Tspan7-R*. A total of 8 polymorphic sites were identified in the gene surrounding region that correspond to 2 haplotype groups with frequencies 0.25 (haplotype group A) and 0.75 (haplotype group B), respectively, in the populations. In order to identify the derivation of these alleles, we further obtained retrogene sequence in *M. caroli*, which diverged from *M. musculus* in 3–4 Myr. The 8 polymorphic sites were retrieved from *Tspan7-R* in *M. caroli* and compared with polymorphic sites discovered from the populations of *M. musculus castaneus*. Interestingly, the haplotype group B with more frequent alleles showed the evidence of deriving from the common ancestral allele shared by group A and *M. caroli* (table 6).

Retrogenes Acquired Untranslated Regions De Novo

In this study, we demonstrate that retrogenes evolve diverged or novel expression patterns from their parental paralogs in genomes. The results have been stated in the previous studies (Zhang et al. 2004; Babaya et al. 2006). Here, we further show that the novelties in the expression patterns of retrogenes might result from newly acquired structures of 5′ or 3′ untranslated regions (UTRs). We identified UTRs of retrogenes by performing both 5′ and 3′ RACE to obtain full-length cDNAs then followed by mapping onto genomic sequences of the mouse genome. Although none of the 5 retrogenes analyzed in this study form chimeric protein structures with existing coding sequences as observed in literature (Long et al. 1999), 4 retrogenes show evidence of recruiting adjacent chromosome-specific sequences as novel UTRs that are absent in their parental paralogs (fig. 3; supplementary fig. S1, Supplementary Material online).

The novel gene structures were identified as follows (fig. 3): 1) *MusT-R* recruited ~30 bp from the 5′ adjacent genomic region as the 5′ UTR, whereas the homologous 5′ UTR to the parental paralog is lost during the duplication process. 2) Locus *Ny-sar-97-R* recruited ~150 bp from the upstream genomic region as a novel 5′ UTR. Besides, the 3′ UTR is shortened in comparison with the parental paralog. 3) For the locus *Tspan7-R*, novel UTRs were recruited from both upstream and downstream adjacent neighboring regions with length ~400 bp each side. 4) *Dmtf1-R* transcripts recruited several hundred base pairs from the 3′ downstream genomic region as both new coding regions and UTRs. The 5′ sequence of *Dmtf1-R* is based on the prediction from the Ensembl database due to the difficulty of RACE. Besides, although *Cypt9* did not acquire novel UTRs, it diverged from the parental paralog by 5 in-frame indels within the coding regions, which resulted in a shortened peptide sequence.

Discussion

Selection-Based Mechanisms Play Critical Role in Driving Genes Out of the X and in Evolving Male-Biased Functions

Our data reveal a correlation between the selection and the expression patterns. Among the 5 retrogenes, 3 retrogenes were identified as evolving a novel male-biased expression pattern. However, only retrogenes generated from

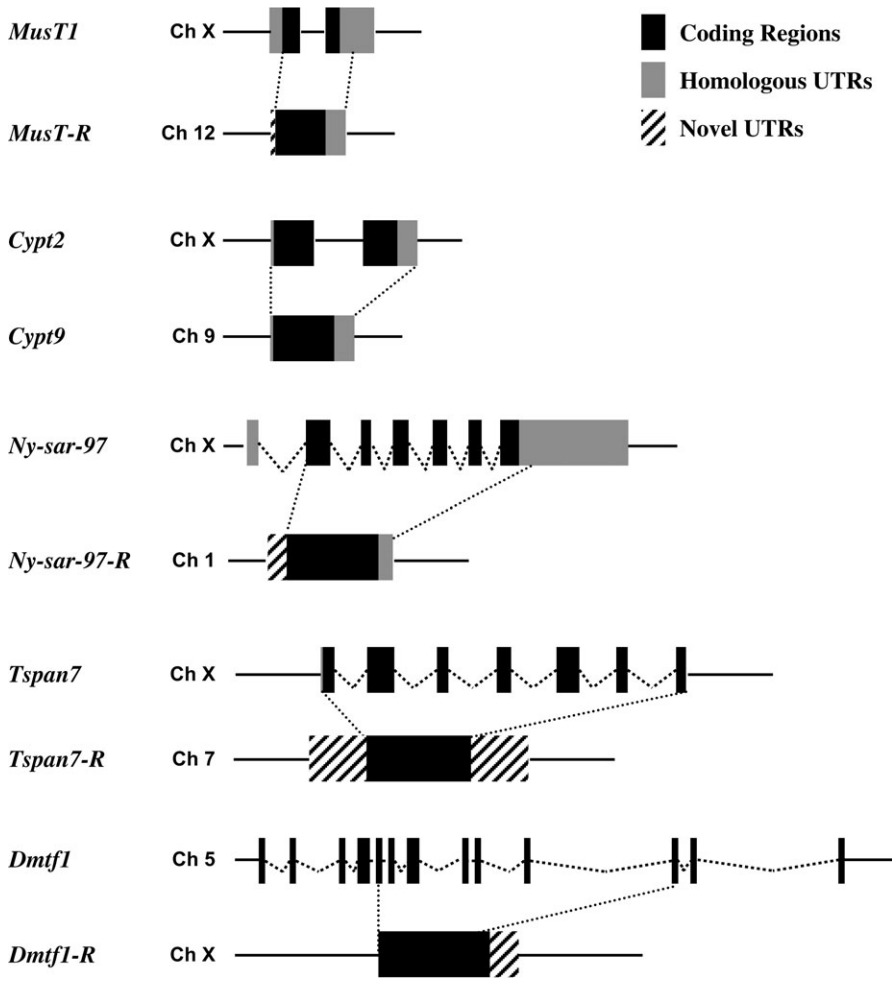

FIG. 3.—Retrogene structures identified by RACE. Gene structures of parental paralogs followed the prediction of Ensembl database. Coding regions are represented in black boxes. UTRs for retrogenes are distinguished by 2 categories: derived from the parental genes (in gray boxes) or recruited de novo from the adjacent genomic regions (in slashed boxes). The correspondent sequences between the retrogenes and the parental paralogs are indicated by orange dotted lines. To be noted that the 5′ UTR of *Dmtf1-R* is referred to Ensembl genome database due to the difficulty of amplification.

the counterparts on the X chromosome, *Ny-sar-97-R* and *Tspan7-R*, are subject to positive selection. This correlation implies a possible causal relationship that the novel male-biased functions in young retrogenes may be a target of positive selection. Furthermore, the novel male expression patterns of X-derived autosomal retrogenes may complement the parental genes that contribute less to male-related functions.

Two models, MSCI and SAXI, were proposed to be the major mechanisms driving genes out of the X chromosome, and a population genetic process also suggests a higher fixation probability on the autosomes for a male-biased dominant mutation. Three models all predict a male-biased expression pattern of autosomal retrogenes that were derived from the genes on the X chromosome. The next question will be which selection-based model is responsible for the selection process of new retrogenes? From our data, we revealed no direct evidence to favor one model over the others at this moment. The MSCI and SAXI models could both be responsible for the nonrandom movement of X-derived retrogenes. However, we have no information about the dominance degree for these 2 new genes; thus, we are not able to test the third population process hypothesis yet.

We observed that retrogenes, *Ny-sar-97-R* and *Tspan7-R*, are expressed exclusively in testis and late spermatogenetic stage exclusively in the early stage, whereas their retrogene counterparts on the autosomes, *Ny-sar-97-R* and *Tsapn7-R*, are expressed exclusively in testis and late spermatogenetic stage (fig. 2, table 2). This suggests that the origins of retrogenes, *Ny-sar-97-R* and *Tsapn7-R*, may result from a possibility to avoid MSCI, which is not necessarily incorporated by the other 2 selection models. Moreover, previous literature demonstrated that the expression of the X-derived autosomal retrogene, *Zfa*, also starts from the late spermatogenetic stage, whereas its X-linked parental paralog, *Zf*x, is continuous expressed (Erickson et al. 1993). These suggest that MSCI may play a critical role in driving genes out of the X. Besides, one of the parental gene on the X, *Ny-sar-97*, was identified as being expressed more predominantly in the ovary (fig. 1), which is consistent with the feminized property of genes on the X as predicted by SAXI model. Taken together, our study demonstrates a possibility that MSCI and SAXI

may both play critical role in driving retrogenes out of the X chromosome in mouse.

We further propose that the novel male-biased expression patterns are targets to natural selection. Different mechanisms have been proposed to explain the restricted expression patterns of retrogenes, particularly those X-derived autosomal ones (Marques et al. 2005). Among those hypotheses, our data suggest that natural selection plays an important role in reinforcing the novel male-biased expression patterns for retrogenes. Although the retrogenes in our studies show the male-biased expression patterns, they may not all be favored in the populations for 2 reasons: 1) they do not contribute novel functions to the genome (i.e., *MusT-R* and *Cypt9* showed identical expression patterns with parental paralogs); 2) those retrogenes moved onto the X chromosome may potentially suffer from MSCI. Consequently, only X-derived autosomal retrogenes developed a restricted male-biased function that can compensate the downregulation of their X-linked parental paralogs that are likely to be fixed in the populations. This indicates that in the evolutionary process, the male-biased expressions are subject to natural selection for the novel functions in genomes.

Derived Haplotype Are More Frequent and May Be Adapted in the Populations

It is commonly held that the adaptations of newly emerged genes are driven by directional selection, which is revealed by skewed polymorphisms in the gene surrounding regions in a population. However, the polymorphic distributions of the X-derived autosomal retrogene, *Tspan7-R*, reflect a history of adapted selection, most likely balancing selection, in the mouse natural populations.

Our studies demonstrate the first case that young retrogenes may evolve adaptively in the form of favored haplotypes in the population. In the previous literature, retrogenes were mostly identified as evolving directionally, which were represented by an excess of polymorphic sites with unexpected low frequencies (Long and Langley 1993; Betran et al. 2002). We suggest that the derived haplotypes, group B, might contribute beneficially to the populations. A recent study in *Drosophila* also observed an adaptive haplotype in the population resulted from a transposable element insertion in the genome (Aminetzach et al. 2005). The insertion of transposable element, *Doc1420*, resulted in 2 major haplotypes in regard to the possession of *Doc1420*. The *Doc1420*-containing chromosomes are much more frequent, around 80%, in the recent populations than in the ancestral African *Drosophila* populations. The population genetic analyses in both gene surrounding regions and flanking regions provide evidence of specific evolution force, natural selection, of *Doc1420*-containing alleles. Further genetic studies found that individuals that carry novel haplotypes with *Doc1420* insertion have higher resistance to an organophosphate pesticide. The increased resistance to pesticide may be responsible for the rapid spread of *Doc1420*-containing alleles in certain populations. Therefore, we suggest that the derived alleles (haplotype group B) of *Tspan7-R* may contribute beneficially to the populations and result in being subject to positive selection in the populations.

The De Novo Acquisition of UTRs May Play an Important Role in New Functions

How retrogenes obtain novel regulatory mechanisms is still controversial. Ohno (1970) proposed that novel or diverged functions play important roles in retaining duplicate genes in genomes. However, the mechanisms contributing to novel expression patterns of duplicate genes remain interesting questions. Previous studies on retrogenes demonstrated that retrosequences could regain their functions in several ways: 1) by forming chimeric structures with an existing functional gene (Long et al. 1999; Wang et al. 2002); 2) by being coexpressed by neighboring genes without disturbing their original function (Betrán et al. 2002; Bradley et al. 2004); 3) by recruiting adjacent novel regions as regulatory sequences, that is, transcription factors or 5′ and 3′ UTRs (Jacobs et al. 1998; Rohozinski and Bishop 2004); 4) by being regulated by transposable elements nearby (Marino-Ramirez et al. 2005); 5) by carrying homologous regulatory sequences with the parental paralog due to aberrant transcription (Soares et al. 1985; Wentworth et al. 1986; McCarrey 1990). Based on the 5 mechanisms addressed above, retrogenes would most likely evolve novel expression patterns by acquiring new regulatory machineries.

Our data suggest that the de novo recruitments of UTRs may lead to the novel expression patterns of retrogenes. We found that these young retrogenes have mostly recruited nearby genomic regions to form chimeric gene structures to some extent, and these novel sequences were recruited mostly as 5′ or 3′ UTRs. The recruitment of novel sequences as UTRs or introns has also been reported in the human genome (Vinckenbosch et al. 2006). It is believed that UTRs may form duplexes with their antisense sequences, and the ratio of duplexes is essential for posttranscriptional regulation of mRNA levels (Lipman 1997) and is tissue specific for some genes (Li et al. 1996). UTRs of certain genes have been proposed to be critical in regulating mRNA activities. In *Drosophila*, both 5′ and 3′ UTRs of *oskar*, a gene involved in cell differentiation, were identified to be crucial for mRNA activation and localization in the early developmental stages (Gunkel et al. 1998). The mRNA will be activated through interaction between 5′ and 3′ UTRs. In human, UTRs were shown to be related to posttranscriptional regulation of gene expression through microRNA (miRNA) regulations (Lewis et al. 2005; Xie et al. 2005). These miRNA regulatory motifs are most often located in the 3′ UTRs of genes. Furthermore, retrogenes with known functions were also identified as acquiring novel UTRs (Jacobs et al. 1998; Rohozinski and Bishop 2004) or evolving modified UTRs from the parental genes (Lagace et al. 2001; Betran et al. 2002) in mammals. In mouse, 2 novel exons were recruited in ADP ribosylation factor-like 4 (*Arl4*) gene as 5′ UTRs. These 2 exons are alternatively spliced in different tissues (Jacobs et al. 1998): the first exon appears predominantly in cDNA clones from mouse testis, whereas the second exon is primarily expressed in fat tissue. Besides, previous studies of

mechanisms that gave rise to new expression factors have focused on 5′ *cis*-upstream regulatory sequences (Betran et al. 2002; Betran and Long 2003). In this study, the widely acquired new UTRs in retrogenes as we observed further suggest that these novel UTRs may provide more possible evolutionary routes for new genes.

Conclusions

In this study, by examining retrogenes with the most recent origins, we show that 2 X-derived autosomal retrogenes with novel male-biased expression patterns are subject to positive Darwinian selection, indicating a possible advantage in evolving functional complementary counterparts on the autosomes for certain genes on the X chromosome. The novel expression patterns are revealed by a more restricted male expression pattern for the X-related retrogenes, whereas their parental paralogs are expressed more ubiquitously. Based on the results, we suggest that both MSCI and SAXI may play critical roles in driving genes out of the X chromosomes. Besides, we demonstrate the first case that the young retrogene is subject to positive selection in the form of 2 major haplotypes. However, which haplotype contributes more beneficially to the population is still unknown. Finally, we propose that these de novo acquired UTRs may play important roles in establishing novel functions of certain retrogenes.

Supplementary Material

Supplementary tables S1 and S2 and fig. S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

National Science Foundation, USA (MCB-9977990) and National Institutes of Health, USA (R01GM65429-01A1 and R01GM078070-01A1) support M.L., and National Science Council, Taiwan provided grant to H.T.Y. (912311B002055, 922311B002009, and 932311B002007). Goodwill Foundation, Taiwan granted a fellowship to M.S.S. Members in the laboratory of M.L. and H.T.Y. provided valuable discussions. We thank K. Bullaughey, Y. S. Lin, B. Y. Liao, C. Fan, M. Moreira, M. Vibranovski, and S. Sun for reading of the manuscript.

Literature Cited

Aminetzach YT, Macpherson JM, Petrov DA. 2005. Pesticide resistance via transposition-mediated adaptive gene truncation in *Drosophila*. Science. 309:764–767.

Ashworth A, Skene B, Swift S, Lovell-Badge R. 1990. *Zfa* is an expressed retroposon derived from an alternative transcript of the *Zfx* gene. EMBO J. 9:1529–1534.

Babaya N, Nakayama M, Moriyama H, Gianani R, Still T, Miao D, Yu L, Hutton JC, Eisenbarth GS. 2006. A new model of insulin-deficient diabetes: male NOD mice with a single copy of *Ins1* and no *Ins2*. Diabetologia. 49:1222–1228.

Bai Y, Casola C, Feschotte C, Betrán E. 2007. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila. Genome Biol. 8:R11.

Banks KG, Johnson KA, Lerner CP, Mahaffey CL, Bronson RT, Simpson EM. 2003. Retroposon compensatory mechanism hypothesis not supported: *zfa* knockout mice are fertile. Genomics. 82:254–260.

Betrán E, Emerson JJ, Kaessmann H, Long M. 2004. Sex chromosomes and male functions. Where do new genes go? Cell Cycle. 3:837–875.

Betrán E, Long M. 2003. *Dntf-2r*, a young Drosophila retroposed gene with specific male expression under positive Darwinian selection. Genetics. 164:977–988.

Betrán E, Thornton K, Long M. 2002. Retroposed new genes out of the X in *Drosophola*. Genome Res. 12:1854–1859.

Betrán E, Wang W, Jin L, Long M. 2002. Evolution of the phosphoglycerate mutase processed gene in human and chimpanzee revealing the origin of a new primate gene. Mol Biol Evol. 19:654–663.

Bradley J, Baltus A, Skaletsky H, Royce-Tolland M, Dewar K, Page DC. 2004. An X-to-autosome retrogene is required for spermatogenesis in mice. Nat Genet. 36:872–876.

Charlesworth B, Coyne JA, Barton NH. 1987. The relative rates of evolution of sex chromosomes and autosomes. Am Nat. 130:113–146.

Dai H, Yoshimatsu TF, Long M. 2006. Retrogene movement within- and between-chromosomes in the evolution of *Drosophila* genomes. (Special Volume for "6th Anton Dohrn Workshop: genome Evolution). Gene. 385:96–102.

Depaulis F, Veuille M. 1998. Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol Biol Evol. 15:1788–1790.

Emerson JJ, Kaessmann H, Betrán E, Long M. 2004. Extensive gene traffic on the mammalian X chromosome. Science. 303:537–540.

Erickson R, Zwingman T, Ao A. 1993. Gene expression, X-inactivation, and methylation during spermatogenesis: the case of *Zfa*, *Zfx*, and *Zfy* in mice. Mol Reprod Dev. 35:114–120.

Gunkel N, Yano T, Markussen FH, Olsen LC, Ephrussi A. 1998. Localization-dependent translation requires a functional interaction between the 5′ and 3′ ends of oskar mRNA. Genes Dev. 12:1652–1664.

Hudson RR. 1990. Gene genealogies and the coalescent process. New York: Oxford University Press.

Jacobs S, Schurmann A, Becker W, Bockers TM, Copeland NG, Jenkins NA, Joost HG. 1998. The mouse ADP-ribosylation factor-like 4 gene: two separate promoters direct specific transcription in tissues and testicular germ cell. Biochem J. 335:259–265.

Khil PP, Smirnova NA, Romanienko PJ, Camerini-Otero RD. 2004. The mouse X chromosome is enriched for sex-bias genes not subject to selection by meiotic sex chromosome inactivation. Nat Genet. 36:642–646.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Kumar RA, Chan KL, Wong AH, Little KQ, Rajcan-Separovic E, Abrahams BS, Simpson EM. 2004. Unexpected embryonic stem (ES) cell mutations represent a concern in gene targeting: lessons from "fierce" mice. Genesis. 38:51–57.

Lagace M, Xuan JY, Young SS, McRoberts C, Maier J, Rajcan-Separovic E, Korneluk RG. 2001. Genomic organization of the X-linked inhibitor of apoptosis and identification of a novel testis-specific transcript. Genomics. 77:181–188.

Lewis BP, Burge CB, Bartel DP. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 120:15–20.
Li AW, Seyoum G, Shiu RP, Murphy PR. 1996. Expression of the rat BFGF antisense RNA transcript is tissue-specific and developmentally regulated. Mol Cell Endocrinol. 118:113–123.
Libby BJ, De La Fuente R, O'Brien MJ, Wigglesworth K, Cobb J, Inselman A, Eaker S, Handel MA, Eppig JJ, Schimenti JC. 2002. The mouse meiotic mutation mei1 disrupts chromosome synapsis with sexually dimorphic consequences for meiotic progression. Dev Biol. 242:174–187.
Libby BJ, Reinholdt LG, Schimenti JC. 2003. Positional cloning and characterization of Mei1, a vertebrate-specific gene required for normal meiotic chromosome synapsis in mice. Proc Natl Acad Sci USA. 100:15706–15711.
Lipman D. 1997. Making (anti)sense of non-coding sequence conservation. Nucleic Acids Res. 25:3580–3583.
Long M, Betrán E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. Nat Rev Genet. 4:865–875.
Long M, Langley CH. 1993. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. Nature. 260:91–95.
Long M, Wang W, Zhang J. 1999. Origin of new genes and source for N-terminal domain of the chimerical gene, *jingwei*, in *Drosophila*. Gene. 238:135–141.
Luoh SW, Page DC. 1994. The structure of the *Zfx* gene on the mouse X chromosome. Genomics. 19:310–319.
Mardon G, Luoh SW, Simpson EM, Gill G, Brown LG, Page DC. 1990. Mouse Zfx protein is similar to *Zfy-2*: each contains an acidic activating domain and 13 zinc fingers. Mol Cell Biol. 10:681–688.
Marino-Ramirez L, Lewis KC, Landsman D, Jordan IK. 2005. Transposable elements donate lineage-specific regulatory sequences to host genomes. Cytogenet Genome Res. 110:333–341.
Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of young human genes after a burst of retroposition in primates. PLoS Biol. 3:1–10.
McCarrey JR. 1990. Molecular evolution of the human *Pgk-2* retroposon. Nucleic Acids Res. 18:949–955.
Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.
O'hUigin C, Li WH. 1992. The molecular clock ticks regularly in muroid rodents and hamsters. J Mol Evol. 35:377–384.
Parisi M, Nuttall R, Naiman D, Bouffard G, Malley J, Andrews J, Eastman S, Oliver B. 2003. Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. Science. 299:697–700.
Ranz JM, Castillo-Davis CI, Meiklejohn CD, Hartl DL. 2003. Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. Science. 300:1742–1745.
Reinke V, Smith HE, Nance J, et al. 2000. A global profile of germline gene expression in *C. elegans*. Mol Cell. 6:605–616.
Rice WR. 1984. Sex chromosomes and the evolution of sexual dimorphism. Evolution. 38:735–742.
Rohozinski J, Bishop CE. 2004. The mouse juvenile spermatogonial depletion (jsd) phenotype is due to a mutation in the X-derived retrogene, mUtp14b. Proc Natl Acad Sci USA. 32:11695–11700.
Rohozinski J, Lamb DJ, Bishop CE. 2006. *UTP14c* is a recently acquired retrogene associated with spermatogenesis and fertility in man. Biol Reprod. 74:644–651.
Romanienko PJ, Camerini-Otero RD. 2000. The mouse *Spo11* gene is required for meiotic chromosome synapsis. Mol Cell. 6:975–987.
Rozas J, Sánchez-DelBarrio JC, Messegyer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics. 19:2496–2497.
Smirnova NA, Romanienko PJ, Khil PP, Camerini-Otero RD. 2006. Gene expression profiles of Spo11-/- mouse testes with spermatocytes arrested in meiotic prophase I. Reproduction. 132:67–77.
Soares MB, Schon E, Henderson A, Karathanasis SK, Cate R, Zeitlin S, Chirgwin J, Efstratiadis A. 1985. RNA-mediated gene duplication: the rat preproinsulin I gene is a functional retroposon. Mol Cell Biol. 5:2090–2103.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585–595.
Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci USA. 103:3220–3225.
Wang W, Brunet FG, Nevo E, Long M. 2002. Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. Proc Natl Acad Sci USA. 99:4448–4453.
Wang W, Thornton K, Emerson JJ, Long M. 2004. Nucleotide variation and recombination along the fourth chromosome in *Drosophila* simulans. Genetics. 166:1783–1794.
Wang W, Zheng H, Fan C, et al. (14 co-authors). 2006. High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell. 18:1791–1802.
Wentworth BM, Schaefer IM, Villa-Komaroff L, Chirgwin JM. 1986. Characterization of the two nonallelic genes encoding mouse preproinsulin. J Mol Evol. 23:305–312.
Wu C-I, Xu EY. 2003. Sexual antagonism and X inactivation—the SAXI hypothesis. Trends Genet. 19:243–247.
Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature. 434:338–345.
Yonekawa H, Moriwaki K, Gotoh O, Hayashi JI, Watanabe J, Miyashita N, Petras ML, Tagashira Y. 1981. Evolutionary relationships among five subspecies of Mus musculus based on restriction enzyme cleavage patterns of mitochondrial DNA. Genetics. 98:801–816.
Zhang J, Dean AM, Brunet F, Long M. 2004. Evolving protein functional diversity in new genes of *Drosophila*. Proc Natl Acad Sci USA. 101:16246–16250.

Takashi Gojobori, Associate Editor

Accepted July 16, 2007