Extraordinary Sequence Divergence at Tsga8, an X-linked Gene Involved in Mouse Spermiogenesis

Jeffrey M. Good,*1,2 Dan Vanderpool,2 Kimberly L. Smith,1 and Michael W. Nachman1

1Department of Ecology and Evolutionary Biology, University of Arizona
2Division of Biological Sciences, The University of Montana

*Corresponding author: E-mail: jeffrey.good@mso.umt.edu.
Associate editor: John H. McDonald

Abstract

The X chromosome plays an important role in both adaptive evolution and speciation. We used a molecular evolutionary screen of X-linked genes potentially involved in reproductive isolation in mice to identify putative targets of recurrent positive selection. We then sequenced five very rapidly evolving genes within and between several closely related species of mice in the genus Mus. All five genes were involved in male reproduction and four of the genes showed evidence of recurrent positive selection. The most remarkable evolutionary patterns were found at Testis-specific gene a8 (Tsga8), a spermatogenesis-specific gene expressed during postmeiotic chromatin condensation and nuclear transformation. Tsga8 was characterized by extremely high levels of insertion–deletion variation of an alanine-rich repetitive motif in natural populations of Mus domesticus and M. musculus, differing in length from the reference mouse genome by up to 89 amino acids (27% of the total protein length). This population-level variation was coupled with striking divergence in protein sequence and length between closely related mouse species. Although no clear orthologs had previously been described for Tsga8 in other mammalian species, we have identified a highly divergent hypothetical gene on the rat X chromosome that shares clear orthology with the 5’ and 3’ ends of Tsga8. Further inspection of this ortholog verified that it is expressed in rat testis and shares remarkable similarity with mouse Tsga8 across several general features of the protein sequence despite no conservation of nucleotide sequence across over 60% of the rat-coding domain. Overall, Tsga8 appears to be one of the most rapidly evolving genes to have been described in rodents. We discuss the potential evolutionary causes and functional implications of this extraordinary divergence and the possible contribution of Tsga8 and the other four genes we examined to reproductive isolation in mice.

Key words: X chromosome, chromatin condensation, DNA binding, positive selection, male reproduction, Mus, reproductive isolation.

Introduction

Evolutionary theory predicts that natural selection will be more efficient on the X chromosome relative to the autosomes if new mutations are on average recessive (Charlesworth et al. 1987; Vicoso and Charlesworth 2009; Mank et al. 2010) because new mutations will be immediately exposed to selection in the hemizygous sex (e.g., males in mammals and Drosophila). Likewise, the X chromosome may differentially accumulate genes with sexually antagonistic fitness effects (Rice 1984) or so-called selfish genes that disrupt the normal meiotic segregation of chromosomes (Frank 1991; Hurst and Pomiankowski 1991; Presgraves 2008; Meiklejohn and Tao 2010). All of these models suggest that the X chromosome may be disproportionately involved in adaptive evolution and may differentially accumulate substitutions involved in reproductive isolation. Although support for faster X-linked evolution is mixed (Presgraves 2008; Mank et al. 2010), considerable data support the notion that patterns of molecular evolution on the X chromosome are unusual, including higher levels of X-linked protein divergence (Baines and Harr 2007; Begun et al. 2007), a higher incidence of positive selection (Bustamante et al. 2005; Torgerson and Singh 2006; Begun et al. 2007), and a significant enrichment of X-linked genes involved in reproductive functions (Wang et al. 2001; Lercher et al. 2003; Parisi et al. 2003; Khil et al. 2004; Dean et al. 2008).

At least some of the unusual evolutionary patterns found on the mammalian X chromosome can be ascribed to its unique behavior during spermatogenesis. The X chromosome is enriched for genes involved in the early stages of spermatogenesis (Wang et al. 2001; Khil et al. 2004), presumably reflecting positive selection for male-specific functions on the hemizygous chromosome (Khil et al. 2004). However, both sex chromosomes condense into a heterochromatic body midway through meiosis (Handel 2004), resulting in transcriptional inactivation and strong selection against X-linked genes involved in meiosis (Khil et al. 2004). Most X-linked genes remain inactive for the remainder of spermatogenesis (postmeiotic sex chromosome repression), save a relatively small subset of genes (Namekawa et al. 2006; Mueller et al. 2008) and microRNAs (Song et al. 2009) that are expressed in postmeiotic haploid cells (spermatids). The postmeiotic expression of these X-linked loci is noteworthy as it is during this time that the haploid germ cells undergo a complex developmental transition from round spermatids to elongated spermatozoa.
Good et al. (1990). Most male-specific aspects of sperm production occur during the postmeiotic stage of spermatogenesis (Eddy 2002) and this stage is where the strongest signatures of positive selection are found across mouse spermatogenesis (Good and Nachman 2005). A key component of this transition is the complete remodeling of chromatin through the replacement of conserved histones with transition proteins and protamines. Genes involved in spermatogenic DNA condensation are often rapidly evolving between mammalian species (Queralt et al. 1995; Wyckoff et al. 2000; Torgerson et al. 2002; Good and Nachman 2005; Turner et al. 2008; Martin-Coello et al. 2009). These observations suggest that X-linked spermatogenic genes may also be rapidly evolving due to positive selection over very short time scales, a hypothesis that could be tested through population-level comparisons among closely related species.

Here, we investigate the evolution of several X-linked genes involved in male reproduction in mice. Our motivation for this study was 2-fold. First, several evolutionary comparisons involving mouse and rat have revealed that many genes on the X chromosome are rapidly evolving (Torgerson and Singh 2006; Baines and Harr 2007; Good et al. 2010), and that the mouse X chromosome is enriched for certain classes of male reproductive genes (Wang et al. 2001; Khil et al. 2004; Dean et al. 2008). However, these patterns are largely restricted to genomic contrasts between mouse and rat and it is unclear to what extent this represents the action of positive natural selection (as opposed to relaxation of constraint). Moreover, mouse and rat diverged ~15 Ma so this comparison does not provide information about selection over short time scales. Second, the evolution of X-linked reproductive genes is relevant to mouse speciation. Several studies have mapped genes causing hybrid male sterility between different mouse species to the X chromosome (Guenet et al. 1995; Wyckoff et al. 2000; Torgerson et al. 2002; Good and Nachman 2005; Turner et al. 2008; Martin-Coello et al. 2009). These considerations led us to study the molecular population genetics of five genes that were 1) rapidly evolving in mouse–rat pairwise genome comparisons, 2) involved in male reproduction, and 3) located within the central region of the X chromosome. Our approach was to resequence these genes in several closely related species of mice and in population samples of M. domesticus and M. musculus. We found that four of the five loci show some evidence of positive selection. However, we focus the majority of our discussion on one extremely rapidly evolving gene, Testis-specific gene a8 (Tsga8), associated with chromatin condensation during spermiogenesis. Examination of the rat genome revealed a previously unidentified ortholog of Tsga8 that was so rapidly evolving that over 60% of the coding domain was unalignable. Despite this extreme nucleotide divergence, general biochemical features of the amino acid sequence composition remain intact, suggesting some conservation of protein function.

Materials and Methods

Sampling Strategy

We attempted to sequence the entire coding region of each locus for eight species of house mice: M. musculus, M. domesticus, M. castaneus, M. spicilegus, M. spreptus, M. caroli, M. cookii, and M. cervicolor. DNA samples were obtained through field collection, loan, or were purchased from the Jackson laboratory (Bar Harbor, ME). We also studied population samples of M. musculus and M. domesticus (supplementary table 1, Supplementary Material online). House mice show little evidence of population genetic structure across Europe (Baines and Harr 2007; Salcedo et al. 2007). Therefore, we followed previous sampling designs (Baines and Harr 2007; Salcedo et al. 2007; Geraldes et al. 2008) and used a single pooled collection of individuals from multiple localities across Europe to represent population-level variation in each species. For M. domesticus, all individuals were of standard karyotype and collected by Michael Nachman from multiple localities across Western Europe. Barbara Gibson collected M. musculus from multiple localities in the Slovak Republic, Poland, and Hungary. For a few loci, we examined additional M. musculus samples from the Czech Republic (provided by Jaroslav Pialek), Austria, Denmark, and Serbia.

Choice of Genes

Five genes were chosen for resequencing (table 1) using three criteria. First, we focused on genes that were rapidly evolving in pairwise comparisons with orthologous rat genes. Four of the loci (4933436101Rik, Testis-specific X-linked [Txa], melanoma antigen a9 [Magea9], and Probasin [Pbsn]) were selected from the top 7% of the most rapidly evolving X-linked genes based on the ratio of nonsynonymous to synonymous changes per site (dN/dS) versus rat (Dean et al. 2008). Tsga8 did not have an annotated ortholog in rat (but see below). Second, we focused on genes involved in male reproductive functions based on available expression data (Cunningham et al. 1998; Uchida et al. 2000; Schultz et al. 2003; Su et al. 2004; Namekawa et al. 2006). Tsga8, 4933436101Rik, and Txa are all expressed during specific stages of spermatogenesis (table 1) and Tsga8 has been hypothesized to play a role in chromatin condensation during spermiogenesis (also called Halap-X; Uchida et al. 2000). Pbsn is an odorant-binding protein primarily expressed in prostate, whereas Magea9 is a member of the Mage-A subfamily of genes that are primarily expressed in tumor cells and the male germ line (Chomez et al. 2001). Third, all five genes are located within the middle third of the X chromosome (60–100 Mb). This general region has been repeatedly...
implicated in reproductive isolation between \textit{M. musculus} and \textit{M. domesticus} (Oka et al. 2004; Payseur et al. 2004; Storchová et al. 2004; Good, Dean, et al. 2008). In particular, \textit{Tsga8} and \textit{Pbsn} have been identified as possible candidates for reproductive isolation based on their location in regions of very low introgression in the hybrid zone (Payseur et al. 2004) and/or putative overlap with mapped hybrid sterility quantitative trait loci (QTL) (Oka et al. 2004; Good, Dean, et al. 2008).

PCR, DNA Sequencing, and Data Preparation

For each gene, we amplified and sequenced the entire protein-coding region based on Ensembl annotation of the National Center for Biotechnology Information (NCBI) mouse genome build 36 (release 46, August 2007; www.ensembl.org). For two genes (\textit{Magea9} and \textit{493343610Rik}), a single amplicon containing the entire protein-coding domain was polymerase chain reaction (PCR) amplified and sequenced. For \textit{Tsga8}, we used several combinations of three forward and three reverse amplification primers to verify repeat variation (see below) and to help overcome sequence divergence within priming sites. For each individual, the consensus sequence of \textit{Tsga8} was verified with at least two independent sequencing reads spanning low-complexity regions. We generated two and four amplicons for \textit{Tsx} and \textit{Pbsn}, respectively. All primers were designed from the mouse genome sequence using Primer3.0 (Rozen and Skaletsky 2000) and are provided in \textit{Supplementary Table 2}, \textit{Supplementary Material} online. PCR products were prepared for sequencing with Qiagen QIAquick spin columns (Qiagen, Valencia, CA) or using the University of Arizona Genetics Core (UAGC) cleaning service. Cleaned PCR products were sequenced by UAGC using an ABI 3731 automated sequencer and big-dye terminator chemistry.

Chromatograms were edited and initially aligned using Sequencher (v4.1.4, Gene Codes Corp., Ann Arbor, MI). All sequence data are available through GenBank (accessions HQ619960–HQ620295). Haplotype phase was unambiguous for all sequences because we only used samples from inbred or male mice. For \textit{Tsga8}, a multiple sequence alignment of protein sequences was generated with \textsc{muscle} (Edgar 2004) and then further optimized manually using \textsc{mesquite} (version 2.74; Maddison and Maddison 2010). Manual adjustments were made in part based on the repeat structures inferred using dotplot analysis (Sonnhammer and Durbin 1995) as implemented with JDotter (Brodie et al. 2004). We used the programs \textsc{pepstats} and \textsc{charge} from the \textsc{EMBOSS} software suite (Rice et al. 2000) to calculate general properties of the amino acid sequence encoded by \textit{Tsga8}. To estimate isoelectric point, we used the Compute predicted isoelectric (pl)/Mw routine as implemented with the ExPaSy server (Gasteiger et al. 2003).

Evolutionary Genetic Analyses

We used a maximum likelihood (ML) framework as implemented in \textsc{codeml} (PAML 4.0; Yang 2007) to evaluate patterns of protein-coding evolution. First, we fit data from each locus to three alternative models of molecular evolution that allow heterogeneity in \(d_{\omega}d_{\delta}\) ratios across codons (\textit{M7}, \textit{M8}, and \textit{M8a}; Swanson et al. 2003). \textit{M7} and \textit{M8a} are related models that allow classes of sites to evolve neutrally or according to purifying selection (i.e., \(d_{\omega}d_{\delta}\) between 0 and 1), whereas \textit{M8} allows for an estimated proportion of sites to evolve due to positive selection (\(d_{\omega}d_{\delta} > 1\)). Positive selection was inferred if model \textit{M8} provided a significantly better fit to the data using a standard likelihood ratio test. We also tested for heterogeneity in \(d_{\omega}d_{\delta}\) between rat and mouse by comparing a model that assumed one \(d_{\omega}d_{\delta}\) ratio across the phylogeny with a two-ratio model that allowed for different \(d_{\omega}d_{\delta}\) values for species of \textit{Mus} and the lineage leading to rat (Yang 1998). ML phylogenies were estimated using Garli (version 0.96; Zwickl 2006) under the best-fit model of nucleotide substitution selected using the Akaike information criterion as implemented with \textsc{jmodeltest} (version 0.1; Posada 2008). Node support was estimated using 200 ML bootstrap replicates with five ML searches per bootstrapped data set.

We calculated multiple statistics of nucleotide variability within \textit{M. domesticus} and \textit{M. musculus} using \textsc{DnaSP} 5.10 (Rozas et al. 2003), including the average pairwise number of nucleotide differences, \(\pi\) (Nei and Li 1979), the proportion of segregating sites, \(\theta\) (Watterson 1975), and Tajima’s \(D\) or the normalized difference between \(\pi\) and \(\theta\) (Tajima 1989). We used the program \textsc{hudson-keirtnan-aguade} (HKA; written by Jody Hey) to conduct multi locus HKA tests to compare patterns of polymorphism to divergence across loci (HKA test; Hudson et al. 1987). To increase the power of the HKA test, we also included previously published noncoding data from four additional X-linked loci (\textit{Maoa}, \textit{Dmd}, \textit{Msn}, and \textit{Dach2}; \(\sim 3,800–5,200\) noncoding bp per locus; Salcedo et al. 2007). \textit{Maoa}, \textit{Dmd}, \textit{Msn}, and \textit{Dach2} were sampled using the same general panel of \textit{M. domesticus} samples \((n = 60–64\) males per locus) used in the current study and a more geographically restricted set of \textit{M. musculus} individuals primarily from the Czech Republic \((n = 18–22\) males per locus).

To contrast the level of \textit{Tsga8} protein sequence divergence between mouse and rat with genome-wide patterns, we used the Ensembl Biomart tool to retrieve the reciprocal protein sequence identities (i.e., percent identity of the
mouse and rat copies, respectively) for all one-to-one orthologous genes in mouse and rat (Genes 60, November 2010; www.ensembl.org). These reciprocal values reflect the percentage of all amino acids that match the orthologous copy in the other species and should be similar when the orthologous pairs are similar in length. Alternatively, large discrepancies in the total length of two copies of an orthologous pair can result in highly asymmetric levels of identity that would be misleading for the current analysis (i.e., the short variant showing high identity and the long variant showing low identity). To account for this, we calculated the ratio of the two values (% mouse ID/% rat ID) and removed all pairs falling in the lower 5% of the distribution. This filtering removes pairs where a short protein in rat induces a low overall percent identity in mouse. Note that there is not extreme overall length asymmetry between the mouse and rat genome copies of Tsga8, and that the rat copy is actually longer than the mouse variant used for this analysis (238 aa vs. 262 aa, respectively).

Expression of Tsga8 in Rat
We used a PCR assay to examine the expression of a putative rat ortholog of Tsga8 (hypothetical gene chrX.436.a; see Results and Discussion). Using primer3.0, we designed ratspecific Tsga8 primers that amplify a 303 bp from genomic DNA (spanning the first intron). Based on predicted exon–intron splice sites of chrX.436.a, these primers should yield an ~124 bp product amplified from cDNA derived from mature mRNA. For controls, we used published primers for the testis-specific transition protein 1 (Tnp1; Sluka et al. 2002) and the broadly expressed beta actin gene (Actb; Yamashita et al. 2005). Primers and PCR conditions are given in supplementary table 3, Supplementary Material online. Fresh testis, liver, heart, and kidney tissues were dissected from an adult rat (outbred strain Sprague Dawley) and immediately frozen on dry ice. Active spermatogenesis was verified by observation of motile sperm extracted from the caudal epididymis. Frozen tissues were homogenized with mortar and pestle, and total RNA was extracted using a Promega SV Total RNA Isolation System kit (Promega Corp., Madison, WI). Reverse transcription to cDNA libraries was performed using a Quanta qScript cDNA Supermix kit (Quanta Biosciences Inc., Gaithersburg, MD). Genomic DNA was extracted separately using the Macherey-Nagel Nucleospin Tissue Extraction kit (Macherey-Nagel, Duren, Germany).

Results and Discussion
An Evolutionary Screen for Targets of Positive Selection
We sequenced the entire coding region for each of the five X-linked genes in several species of house mice and downloaded available sequence from the rat genome. We then used an ML framework to test for evidence of positive selection influencing patterns of protein evolution at each gene (table 2). We conducted two kinds of comparisons: 1) a species-only comparison where we randomly chose a single haplotype to represent each species and 2) a unique-haplotype comparison where we included all unique haplotypes within M. domesticus and M. musculus. Three of the five genes (493343601Rik, Pbsn, and Tsx) provided a significantly better fit to a model of molecular evolution that allowed for a subset of codons to evolve rapidly (dN/dS > 1) for at least one of these two tests, suggesting these loci have experienced recurrent positive selection.

The signature of recurrent positive directional selection was strongest for 493343601Rik and Tsx and both showed high rates of nonsynonymous evolution between species of Mus (table 2). Pbsn had a high rate of nonsynonymous change along the branch leading to rat and appeared more constrained among sequenced species of Mus. Detecting positive selection by comparison of dN/dS ratios among sites requires fairly high levels of divergence (Anisimova et al. 2001) and some of our power comes from the relatively long branch leading to rat. None of these genes provide significant evidence for site-specific selection when considering only Mus species (all genes P > 0.05). Nevertheless, models allowing for different dN/dS between the Mus clade and the rat did not fit the data significantly better than models assuming a single rate of protein evolution across all lineages (all genes P < 0.05). Therefore, with the possible exception of Pbsn, the elevated dN/dS ratios do not appear to be driven exclusively by the longer branch leading to rat.

Two genes, Magea9 and Tsga8, did not show strong evidence of positive directional selection in our dN/dS analyses. The evolution of Magea9 appeared to be driven primarily by relaxed functional constraint. Of the seven species of Mus that we sequenced at Magea9, all except for M. domesticus had premature stop codons (supplementary fig. 1, Supplementary Material online). For M. musculus, M. castaneus, M. spretus, and M. spicilegus, the earlier stop codon was due to a single base pair deletion relative to M. domesticus that shifted the open reading frame, shortening the protein sequence by nine amino acids. This codon is deleted entirely in rat, making it difficult to judge the ancestral state. For M. cervicolor and M. cookii, we found multiple additional single base pair deletions that disrupted the open reading frames and severely truncated the putative protein sequences. Pseudogenes are relatively common in the Mage gene family (Chomez et al. 2001), and therefore loss of function at Magea9 is not surprising. The inferred amino acid sequence of Tsga8 was characterized by extreme length variation between species and within M. domesticus and M. musculus. This pattern of divergence is poorly suited for analysis with dN/dS based methods because positions with insertion–deletion (indel) variation are excluded. Nevertheless, the patterns of evolutionary divergence at Tsga8 were highly unusual and motivated a more in-depth analysis of this locus as described below.

High dN/dS pairwise ratios are often interpreted as signatures of adaptive evolution in genomic analyses. However, this is a very conservative test that rarely yields unequivocal evidence for positive selection (dN/dS > 1). There are many factors that influence rates of protein evolution (for a review see Pal et al. 2006) and it remains unclear how often high pairwise dN/dS values reflect a history of adaptive evolution. In our
Table 2. Rates of Protein Evolution for Five X-linked Genes.

<table>
<thead>
<tr>
<th>MGI Symbol</th>
<th>Species* (haps)</th>
<th>Length (bp)</th>
<th>dN/dS phylogeny</th>
<th>dN/dS Mus*</th>
<th>dN/dS Rat*</th>
<th>--2ΔL^ (sites)</th>
<th>dN/dS &gt; 1 (percentage of codons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>493343601Rik</td>
<td>8 (14)</td>
<td>1,221</td>
<td>0.96</td>
<td>1.09</td>
<td>0.85</td>
<td>11.17**</td>
<td>4.90 (8.2)</td>
</tr>
<tr>
<td>M. domesticus</td>
<td>8 (8)</td>
<td>1,221</td>
<td>0.96</td>
<td>1.09</td>
<td>0.85</td>
<td>10.17**</td>
<td>5.12 (7.2)</td>
</tr>
<tr>
<td>Magea9</td>
<td>6 (9)</td>
<td>723</td>
<td>0.82</td>
<td>0.72</td>
<td>0.87</td>
<td>1.95</td>
<td>10.07 (1.1)</td>
</tr>
<tr>
<td>Pbsn</td>
<td>5 (7)</td>
<td>516</td>
<td>0.77</td>
<td>0.53</td>
<td>0.80</td>
<td>4.44</td>
<td>17.59 (3.2)</td>
</tr>
<tr>
<td>Tsga8</td>
<td>5 (5)</td>
<td>516</td>
<td>0.84</td>
<td>∞</td>
<td>0.80</td>
<td>6.77**</td>
<td>26.08 (3.8)</td>
</tr>
<tr>
<td>Tsx</td>
<td>5 (2)</td>
<td>621</td>
<td>1.17</td>
<td>1.17</td>
<td>n.a.</td>
<td>3.74</td>
<td>4.10 (2.9)</td>
</tr>
<tr>
<td>T. us</td>
<td>5 (2)</td>
<td>429</td>
<td>1.68</td>
<td>1.96</td>
<td>1.55</td>
<td>11.51**</td>
<td>7.09 (14.5)</td>
</tr>
<tr>
<td>M. domesticus</td>
<td>9 (10)</td>
<td>429</td>
<td>1.67</td>
<td>1.89</td>
<td>1.57</td>
<td>11.64**</td>
<td>7.03 (16.2)</td>
</tr>
</tbody>
</table>

*a Number of species (including rat) and total number of coding region haplotypes. Variation in the total number of species across loci reflects PCR amplification failure or the exclusion of species with truncated coding sequence (Magea9). For each gene, we conducted two tests, shown on separate rows. The upper row for each gene includes all unique haplotypes from populations of Mus domesticus and M. musculus in addition to the single haplotype from each of the other species. The lower row for each gene includes one haplotype for each species. See supplementary table 2, Supplementary Material online, for a list of species included for each gene.
b Length of protein-coding alignment, excluding positions with indel variation.
c dN/dS estimated across the entire phylogeny.
d dN/dS estimated separately for all Mus species, and the branch leading to rat.
e Likelihood ratio test statistic comparing the difference in the likelihood (L) of M8 and M8a.
f M8 estimate of dN/dS for the positively selected site class and proportion of codons in this class.
**Significant to Bonferroni-corrected P < 0.01 based on \( \chi^2 \) test with df = 1.

study, three of the four genes chosen based on elevated pairwise dN/dS values between mouse and rat showed evidence of adaptive evolution over shorter timescales (table 2). Other studies in primates (Clark and Swanson 2005) and Drosophila (Swanson et al. 2004; Wagstaff and Begun 2005; Kelleher et al. 2007) have also found relatively high incidences of positive selection over short timescales at genes with high rates of pairwise protein divergence (i.e., dN/dS > 0.5) over deeper timescales.

Extraordinary Molecular Evolution at Tsga8 in Mice

The original cloning of Tsga8 in C57BL/6J laboratory mice (derived largely from M. domesticus) documented a 238 amino acid (aa) coding domain spanning two exons, including two alanine-rich repeat motifs in the middle of the second exon of Tsga8 (Uchida et al. 2000). This same sequence was later verified in the reference mouse genome, which also derives from C57BL/6J. We observed extensive indel variation within and between species in this low complexity, alanine-rich region. Strikingly, all individuals had amino acid sequences that were considerably longer (258–327 aa) than the mouse reference variant of Tsga8. The sources of this length variation were apparent when comparing pairwise dotplots between the reference mouse genome (C57BL/6J) and representative sequences from M. domesticus, M. musculus, and M. spretus (fig. 1). The first repetitive domain had four imperfect 15 aa [AAAAA-P/EAAAS(P/L)/ESS] repeats in the mouse genome. This region is expanded in M. domesticus, M. musculus, and M. spicilegus but shortened and partially truncated in M. spretus. A second repetitive domain of 18 aa [AAPE(A/V)AAA-PVAA(A/T)/PATP] occurs shortly after this first repeat and was found in between one to four complete copies. We used these amino acid motifs to help guide the generation of a multiple species alignment (supplementary fig. 2, Supplementary Material online).

To examine the evolution of the repeat variants in more detail, we generated a ML phylogeny from the entire sequenced region of Tsga8 (including the intron). Figure 2 shows the distribution of repeats across unique sequence haplotypes. We found five length variants within M. musculus (7 haplotypes) and six variants within M. domesticus (13 haplotypes), differing by as much as 45 aa in total length within species and 53 aa between species. We did not observe a single indel that disrupted the open reading frame of Tsga8 (all indels occurred in multiples of three nucleotides). Remarkably, the shortest protein variant in either species was still 36 aa longer than the allele found in the classic inbred strain C57BL/6J. Nevertheless, the reference mouse genome sequence clearly falls within the M. domesticus clade, suggesting that this rare short variant may be due to one or a few large deletion events. The overall mutational history of indel variation at Tsga8 is difficult to discern but it does not appear to be due to simple mutational events during replication (i.e., Tsga8 is not a simple microsatellite). Most length variants involved large indel events (>30 bp) that are quite rare in rodent genomes (Makova et al. 2004). It is possible that biased gene conversion during female meiosis has also contributed to the evolution of this region, given the existence of identical but nonadjacent repeats (fig. 2). Gene conversion can lead to homogenization of small tracts of DNA between alleles on homologous chromosomes (Chen et al. 2007). In some instances, interallelic gene conversion has been shown to cause very high levels of genetic diversity through the rapid generation of novel alleles (Yip 2002; von Salome et al. 2007).

Others have argued for positive selection acting directly on indel variation in reproductive genes (Podlaha and Zhang 2003; Podlaha et al. 2005; Schully and Hellberg 2006; Oliver et al. 2009). For example, the sperm calcium ion channel gene, Catsper1, shows considerable indel variation in primates (Podlaha and Zhang 2003) and mice (Podlaha et al. 2005). Variation has also been described in the number of repetitive zinc finger domains in the Prdm9 gene. Prdm9 encodes for a histone methyltransferase that is essential for male and female meiosis (Hayashi et al. 2005) and has been shown to cause hybrid male sterility between M. domesticus and M.
patterns of silent nucleotide variation for idly evolving genes.

...size of indels that we observed in mice at evolutionary timescale of our study (7–13 repeats) in new and old world mice (Salcedo et al. 2007). In stark contrast to divergence for each of three pairwise contrasts (M. domesticus–M. musculus; M. domesticus–M. spretus; M. musculus and M. spretus). The nine-locus HKA test was significant for the M. domesticus–M. musculus (χ² = 24.60, P = 0.031, 10,000 coalescent simulations) and marginally significant for contrasts between M. domesticus–M. spretus (χ² = 14.37, P = 0.053) and the M. musculus–M. spretus (χ² = 13.81, P = 0.068). For these analyses, we excluded all alignment positions with indel variation. We also excluded six individuals with rare Tsga8 length variants in order to maximize the number of aligned sites without an indel (we excluded M. musculus Haps 2, 5, 7; M. domesticus Haps 11; fig. 2). Nevertheless, we are cautious when interpreting quantitative comparisons of patterns of polymorphism and divergence at Tsga8 to other genes. The extreme length differences between species greatly reduced the number of alignable positions available for divergence comparisons and estimates of polymorphism could also be inflated by undetected errors in the protein alignment. As with our analysis of δN/dS at this locus, these difficulties underscore the limitations of most standard population-genetic frameworks for evaluating modes of molecular evolution given extensive length variation within and divergence between species.

Identification of a Highly Divergent Tsga8 Ortholog in Rat

The specific function of Tsga8 remains unknown but multiple lines of evidence suggest that Tsga8 plays a role in morphological reorganization of spermatids during postmeiotic spermatogenesis. Expression data across mouse spermatogenesis show that Tsga8 is highly expressed in round spermatids (Uchida et al. 2000). During this developmental transition, chromatin is restructured by replacement of histones with protamines through intermediate transition proteins (Ward and Coffey 1991). It has been hypothesized that Tsga8 may be involved in sperm chromatin condensation through interaction with DNA on the N-terminus and basic nuclear proteins on the acidic (and highly variable) C-terminus (Uchida et al. 2000). Interestingly, no annotated ortholog exists for Tsga8 in the rat genome and we were unable to sequence Tsga8 at evolutionary depths deeper than M. spicilegus and M. spretus (~2 Ma; Guenet and Bonhomme 2003), suggesting that it may be a recently evolved gene within Mus.

To explore this in more detail, we performed a BLAT query (Kent 2002) of the mouse Tsga8 region against the rat genome (version 3.4) using the University of California–San Cruz genome browser (Kuhn et al. 2009). The best match overlapped with a hypothetical protein (chrX.436.a; N-SCAN prediction, Gross and Brent 2006) in the middle of the rat X chromosome (~69.9 Mb; supplementary fig. 3, Supplementary Material online). Identifiable homology between the...
mouse and rat sequences spanned exon 1 (5′ UTR and CDS), the intron, and the beginning of exon 2. We calculated an uncorrected sequence divergence of 16.6% between mouse and rat CDS in this region, similar to the mouse–rat coding divergence found between the other four genes we sequenced (13.4–22.6%). There was no identifiable homology for 61.5% of the rat 262 aa coding domain (supplementary fig. 3, Supplementary Material online) spanning a large repetitive region of exon 2 (rat aa positions 70–230), but there was another alignable stretch of bases on the 3′ end of the gene. Somewhat paradoxically, this extreme divergence actually prevents us from effectively using standard divergence-based frameworks, such as dN/dS, to formally test if Tsga8 is evolving due to recurrent positive directional selection. Nevertheless, the mouse genome copy of Tsga8 shows 28.6% amino acid sequence identity (68 of 238 aa) to the rat Tsga8 (26% identity for rat to mouse, 68 of 262 aa), falling in the lower 0.1% of the genomic distribution (P = 0.00069) of amino acid sequence identity between 15,947 annotated 1:1 mouse–rat orthologs (fig. 3). Although this contrast is not a formal test of nonneutral molecular evolution, it does clearly demonstrate that the observed protein sequence divergence at Tsga8 is an extreme outlier between the mouse and rat genomes and thus less likely to reflect purely neutral divergence.

Despite the extreme divergence between the inferred Tsga8 protein sequences in mouse and rat, the predicted rat gene has the same general structure as the mouse copy of Tsga8: Coded on the negative strand, two exons separated by a single short intron, and a low-complexity region of highly acidic amino acids in the middle of the second exon.

**Fig. 2.** ML phylogeny for unique nucleotide haplotypes of Tsga8 based on the best-fit general time reversible + I model of sequence evolution. Node support is based on ML bootstraps and haplotype counts are given in parentheses. Per haplotype counts of the two imperfect repeat motifs are shown in dark (15 aa) and light gray (18 aa). Repeats with identical sequence are connected with lines and partial repeats are indicated with an (*). Partial repeats are all left justified and are not drawn to scale. Total protein lengths are given on the right margin.

**Rapid Evolution of Tsga8**

doi:10.1093/molbev/msq348

MBE
Dotplot analysis of the low-complexity region in rat revealed 19 imperfect valine-rich repeats (6–8 aa) with a general consensus sequence of AAVVVKEE (50% majority; supplementary fig. 4, Supplementary Material online). The overall amino acid composition differed considerably between the rat and the mouse copies of \textit{Tsga8} (supplemental fig. 4, Supplementary Material online). In particular, there were many fewer alanines in the rat version (40 vs. 82 residues), offset by a large increase in biochemically similar valines (55 vs. 11 residues). Remarkably, the overall charge and predicted isoelectric point appears strongly conserved between the two proteins (mouse \textit{Tsga8}, pI 5.4.65; rat \textit{Tsga8}, pI 5.4.62). The occurrence of a basic N-terminus and a highly acidic C-terminus is also strongly parallel between the two proteins (fig. 4), despite retaining identifiable sequence homology for only \textordmasculine{40} of the coding domain. Likewise, the rat transcript also retains a 5’ nuclear localization signal.

Thus, it appears that the rapidly evolving mouse and rat proteins have experienced some functional constraint on the underlying physical and chemical properties essential to their respective molecular functions.

Finally, we used a PCR assay to verify that the putative \textit{Tsga8} rat ortholog is indeed expressed in rat testis. Rat-specific \textit{Tsga8} primers were designed to amplify a 303 bp fragment from genomic DNA and an approximately 124 bp product from cDNA derived from mature mRNA (i.e., missing the first intron). In a survey of four adult rat tissues (heart, liver, kidney, and testis), \textit{Tsga8} expression was detected only in the testis (supplemental fig. 5, Supplementary Material online).

Further functional investigation of \textit{Tsga8} is warranted, given the exceptional levels of protein-coding variation segregating at \textit{Tsga8} within mice and the extreme divergence between apparent mouse and rat orthologs. If \textit{Tsga8} is indeed involved directly in chromatin condensation, then allelic variants at \textit{Tsga8} have the potential to influence the morphological form and function of sperm (Uchida et al. 2000). It is likely that \textit{Tsga8} transcripts are transported to adjacent Y-bearing spermatids through connective intracellular bridges as has been documented for many

![Fig. 3. Protein sequence identity between mouse and rat orthologous gene pairs. The distribution of percent sequence identity is based on 15,947 one-to-one orthologous pairs. Only ten genes show lower pairwise protein identity than \textit{Tsga8} (28.6%), placing \textit{Tsga8} in the lower 0.07% of this distribution. Our estimate of percent sequence identity for mouse \textit{Tsga8} is based on the number of identical amino acids found in the alignable portions of the gene (positions 1–70, 217–238) divided by the total length of the protein (238 aa).]
transcripts expressed during the haploid phase of spermiogenesis (Morales et al. 1998). However, only a single Tsga8 allele will be expressed during spermatogenesis within a given male because it is X-linked. Thus, allelic differences segregating within a population would directly translate into functional reproductive differences between individuals and thus could be the target of sexual selection related to sperm competition (Parker 1970; Swanson and Vacquier 2002) or antagonistic coevolutionary dynamics between males and females (Joseph and Kirkpatrick 2004). Sperm competition is common in mice (Dean et al. 2006) and evolutionary divergence driven by sperm competition is often considered under a model of recurrent positive directional selection (Swanson and Vacquier 2002). However, population-level interactions among competing males could also lead to the selective maintenance of multiple alleles within a population (Clark 2002). Such a mechanism could explain the high within-population diversity found at Tsga8. Interestingly, other proteins known to be involved in DNA binding (Ting et al. 1998; Barbash et al. 2003; Brideau et al. 2006; Oliver et al. 2009) and chromatin condensation (Queralt et al. 1995; Wyckoff et al. 2000; Good and Nachman 2005; Turner et al. 2008; Martin-Coello et al. 2009) have been shown to be under positive selection. In particular, several testis-specific genes (e.g., Hils1, Prm1, Prm2, and Trmp2) involved in sperm chromatin remodeling have been shown to be rapidly evolving in mammals (Queralt et al. 1995; Wyckoff et al. 2000; Torgerson et al. 2002; Turner et al. 2008; Martin-Coello et al. 2009). One possible mechanism driving this divergence is the runaway propagation of coevolution between interacting genic regions directly involved in DNA or protein binding. Regardless of the underlying evolutionary mechanisms, Tsga8 appears to be one of the most rapidly diverging protein-coding genes to have been described in mammals.

**Implications for Reproductive Isolation in Mice**

All five of the genes that we examined were chosen because they occur in regions of the X chromosome involved in reproductive isolation between *M. musculus* and *M. domesticus*. Of these, Tsga8 and 4933436I01Rik are the best candidates for contributing to reproductive isolation. Tsga8 is divergent between *M. domesticus* and *M. musculus* in both coding length and amino acid composition and occurs in the portion of the X chromosome that shows the lowest introgression across the European hybrid zone (Payseur et al. 2004). The phenotypic manifestation of hybrid male sterility in crosses between various species of mice is highly variable and appears to have a complex genetic and developmental basis (Forejt 1996; Storchová et al. 2004; Oka et al. 2007, 2010; Good, Handel, et al. 2008). Depending on the cross, hybrid male sterility may involve the disruption of genes acting at the mitotic, meiotic, or postmeiotic stages of spermatogenesis (Storchová et al. 2004; Good, Dean, et al. 2008; Good, Handel, et al. 2008; Mihola et al. 2009; Oka et al. 2010). Chromatin condensation is an important determinant of mature sperm head morphology (Toshimori and Ito 2003) and abnormal head morphology is one of the central postmeiotic phenotypes involved in hybrid male sterility between *M. domesticus* and *M. musculus* (Oka et al. 2004, 2007; Storchová et al. 2004; Good, Dean, et al. 2008). Meiotic inactivation and/or postmeiotic repression of the X chromosome appears to be disrupted in sterile hybrid males from crosses between a female *M. musculus* and a male *M. domesticus*, resulting in chromosome-wide over expression of genes expressed during the later postmeiotic stages of spermatogenesis.

**Fig. 4.** Sliding window analysis showing overall similarity of amino acid charge for Tsga8 in mouse and rat (3 aa window size). The pl points are shown for the N- and C-termini. The nuclear localization signal identified by Uchida et al. (2000) is shown for mouse (aa 64–70) and projected by alignment on the rat sequence (aa 63–69). Rat amino acid positions 70–230 are unalignable to the mouse.
was supported by an National Science Foundation (NSF) Integrative Graduate Education Research Traineeship grant Genomics Initiative (DGE0114420, J.M.G.), an NSF Doctoral Dissertation Improvement grant (DEB0608452, J.M.G.), an NSF grant (DEB0213013, M.W.N.), the National Institute of Health grant (1 R01 GM074245-01A1, M.W.N.), and start-up research funds from the University of Montana.

References


Dean MD, Ardlie KG, Nachman MW. 2006. The frequency of multiple paternity suggests that sperm competition is common in house mice (Mus domesticus). Mol Ecol. 15:4141–4151.


weight and fertility in an interspecies hybrid congenic strain for chromosome X. *Mamm Genome*. 12:45–51.


