Molecular Biology and Evolution

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. Nachman[1]

[1]Department of Ecology and Evolutionary Biology, University of Arizona

[2]Division 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.

Research article

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

(Russell 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. 1990; Elliott et al. 2001, 2004; Oka et al. 2004, 2007; Storchová et al. 2004; ; Good, Dean, et al. 2008). The X chromosome also shows reduced gene flow relative to the autosomes across a natural hybrid zone between *Mus domesticus* and *M. musculus* (Tucker et al. 1992; Dod et al. 1993; Munclinger et al. 2002; Macholán et al. 2007; Teeter et al. 2008). In particular, the center of the X chromosome has been shown to harbor one or more male sterility factors (Oka et al. 2004; Storchová et al. 2004; Good, Dean, et al. 2008) and shows exceptionally low levels of gene flow across the hybrid zone (Payseur et al. 2004).

These considerations led us to study the molecular populations 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. spretus*, *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 (*4933436I01Rik*, *Testis-specific X-linked* [*Tsx*], *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 (d_N:d_S) 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*, *4933436I01Rik*, and *Tsx* 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

Table 1. Five Sequenced X-linked Genes.

MGI Symbol	Ensembl ID	Position (Mb)[a]	d_N:d_S[b]	Primary Tissue[c]
4933436I01Rik	ENSMUSG00000025288	65.17	0.85	Testis (RS)[d]
Magea9	ENSMUSG00000046301	70.11	0.67	—
Pbsn	ENSMUSG00000000003	75.08	0.79	Prostate
Tsga8	ENSMUSG00000035522	80.73	n.a.	Testis (RS)[e]
Tsx	ENSMUSG00000031329	100.6	1.34	Testis[f]

[a] Physical position based NCBI mouse build 37 (Ensembl 57, May 2010).
[b] Rate of protein evolution based on pairwise comparison with rat one-to-one orthologs (Dean et al. 2008), no ortholog of *Tsga8* has been annotated in the rat genome.
[c] Based on relative expression in adult testis versus 60 other tissues (Su et al. 2004).
[d] RS = postmeiotic expression in round spermatid germ cells (Namekawa et al. 2006).
[e] Postmeiotic expression in round spermatid germ cells (Uchida et al. 2000).
[f] Spermatogenic expression in spermatocytes (Schultz et al. 2003) and Sertoli cells (Cunningham et al. 1998).

implicated in reproductive isolation between *M. musculus* and *M. domesticus* (Oka et al. 2004; Payseur et al. 2004; Storchová et al. 2004; Good, Dean, et al. 2008). In particular, *4933436I01Rik*, *Tsga8*, and *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 (*Magea9* and *4933436I01Rik*), a single amplicon containing the entire protein-coding domain was polymerase chain reaction (PCR) amplified and sequenced. For *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 *Tsga8* was verified with at least two independent sequencing reads spanning low-complexity regions. We generated two and four amplicons for *Tsx* and *Pbsn*, respectively. All primers were designed from the mouse genome sequence using Primer3.0 (Rozen and Skaletsky 2000) and are provided in supplemental table 2, 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 *Tsga8*, a multiple sequence alignment of protein sequences was generated with MUSCLE (Edgar 2004) and then further optimized manually using 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 Pepstats and Charge from the EMBOSS software suite (Rice et al. 2000) to calculate general properties of the amino acid sequence encoded by *Tsga8*. To estimate isoelectric point, we used the Compute predicted isoelectric (pI)/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 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_N:d_S ratios across codons (M7, M8, and M8a; Swanson et al. 2003). M7 and M8a are related models that allow classes of sites to evolve neutrally or according to purifying selection (i.e., d_N:d_S between 0 and 1), whereas M8 allows for an estimated proportion of sites to evolve due to positive selection (d_N:$d_S > 1$). Positive selection was inferred if model M8 provided a significantly better fit to the data using a standard likelihood ratio test. We also tested for heterogeneity in d_N:d_S between rat and mouse by comparing a model that assumed one d_N:d_S ratio across the phylogeny with a two-ratio model that allowed for different d_N:d_S values for species of *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 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 *M. domesticus* and *M. musculus* using DnaSP 5.10 (Rozas et al. 2003), including the average pairwise number of nucleotide differences, π (Nei and Li 1979), the proportion of segregating sites, θ (Watterson 1975), and Tajima's *D* or the normalized difference between π and θ (Tajima 1989). We used the program Hudson–Kreitman–Aguade (HKA; written by Jody Hey) to conduct multilocus 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 (*Maoa*, *Dmd*, *Msn*, and *Dach2*; ~3,800–5,200 noncoding bp per locus; Salcedo et al. 2007). *Maoa*, *Dmd*, *Msn*, and *Dach2* were sampled using the same general panel of *M. domesticus* samples (n = 60–64 males per locus) used in the current study and a more geographically restricted set of *M. musculus* individuals primarily from the Czech Republic (n = 18–22 males per locus).

To contrast the level of *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 rat-specific *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 (*4933436I01Rik*, *Pbsn*, and *Tsx*) provided a significantly better fit to a model of molecular evolution that allowed for a subset of codons to evolve rapidly (d_N:$d_S > 1$) 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 *4933436I01Rik* 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 d_N:d_S 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 d_N:d_S 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 d_N:d_S 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 d_N:d_S 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 d_N:d_S 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 d_N:d_S 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 (d_N:$d_S > 1$). 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 d_N:d_S values reflect a history of adaptive evolution. In our

Table 2. Rates of Protein Evolution for Five X-linked Genes.

MGI Symbol	Species[a] (haps)	Length (bp)[b]	d_N:d_S phylogeny[c]	d_N:d_S *Mus*[d]	d_N:d_S Rat[d]	$-2\Delta L$[e] (sites)	d_N:d_S > 1 (percentage of codons)[f]
4933436I01Rik	8 (14)	1,221	0.96	1.09	0.85	11.11**	4.90 (8.2)
	8 (8)	1,221	0.96	1.09	0.85	10.17**	5.12 (7.2)
Magea9	6 (9)	723	0.82	0.72	0.87	1.95	10.07 (1.1)
	6 (6)	723	0.81	0.68	0.87	1.99	10.10 (1.1)
Pbsn	5 (7)	516	0.77	0.53	0.80	4.44	17.59 (3.2)
	5 (5)	516	0.84	∞	0.80	6.77**	26.08 (3.8)
Tsga8	5 (15)	621	1.17	1.17	n.a.	3.74	4.10 (2.9)
	5 (5)	621	0.91	0.91	n.a.	0	—
Tsx	9 (10)	429	1.68	1.96	1.55	11.51**	7.09 (14.5)
	9 (9)	429	1.67	1.89	1.57	11.64**	7.03 (16.2)

[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 population samples 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] d_N:d_S estimated across the entire phylogeny.

[d] d_N:d_S 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 d_N:d_S for the positively selected site class and proportion of codons in this class.

**Significant to Bonferroni-corrected $P < 0.01$ based on χ^2 test with df = 1.

study, three of the four genes chosen based on elevated pairwise d_N:d_S 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., d_N:$d_S > 0.5$) 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 aas) 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 [AAAAAPEAAAS(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)AAAPEVAA(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.*

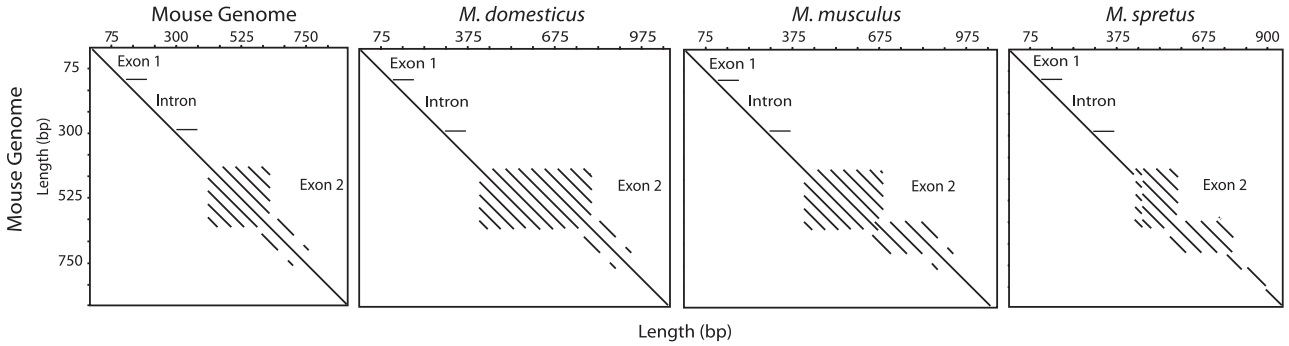

Fig. 1. Dotplots comparing the mouse genome sequence of *Tsga8* with itself and three representative mouse species. Pairwise comparisons were generated using JDotter with a 10 bp sliding window. The boundaries between exons 1 and 2 and intron 1 are indicated along the diagonal. Off diagonal lines indicate regions of repetitive similarity.

musculus (Mihola et al. 2009). The causative mutation for hybrid male sterility is likely a single 28 aa indel in the repetitive zinc finger domain, which has been hypothesized to influence DNA binding affinity (Mihola et al. 2009). *Prdm9* also shows considerable variation in the number of zinc finger domains (7–13 repeats) in new and old world mice (~25 My; Oliver et al. 2009). When considering the much shorter evolutionary timescale of our study (~2 My), the number and size of indels that we observed in mice at *Tsga8* is much more extreme than what has been described for either of these rapidly evolving genes.

Next, we compared nucleotide variation at *Tsga8* with the other four reproductive genes that we sequenced (table 3). Patterns of silent nucleotide variation for *4933436I01Rik*, *Pbsn*, *Magea9*, and *Tsx* were similar to previously documented X-linked patterns (Salcedo et al. 2007) for *M. domesticus* (average $\pi = 0.03$; $\theta = 0.06$) and *M. musculus* ($\pi = 0.02$; $\theta = 0.03$). Several loci showed significantly negative values of Tajima's *D* (table 3), consistent with a recent population expansion of house mice (Salcedo et al. 2007). In stark contrast to all other loci, patterns of silent nucleotide variation at *Tsga8* were much higher in *M. domesticus* and *M. musculus*. If higher levels of polymorphism were driven by a higher mutation rate at *Tsga8*, then we would also expect a proportionally higher level of divergence. However, average levels of silent divergence relative to *M. spretus* appear normal for this locus. To test if *Tsga8* harbored an excess of nucleotide polymorphism, we used a multilocus HKA test to compare patterns of silent site polymorphism across our five sequenced loci and published data from four additional X-linked loci (*Maoa*, *Dmd*, *Msn*, and *Dach2*; ~3,800–5,200 kb per locus; Salcedo et al. 2007) 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* ($\chi^2 = 24.60$, $P = 0.031$, 10,000 coalescent simulations) and marginally significant for contrasts between *M. domesticus*–*M. spretus* ($\chi^2 = 14.37$, $P = 0.053$) and the *M. musculus*–*M. spretus* ($\chi^2 = 13.81$, $P = 0.068$). 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* Hap 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 d_N:d_S 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; Schultz et al. 2003) and immunohistological assays have localized *Tsga8* to the nucleoplasm just prior to nuclear reorganization (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

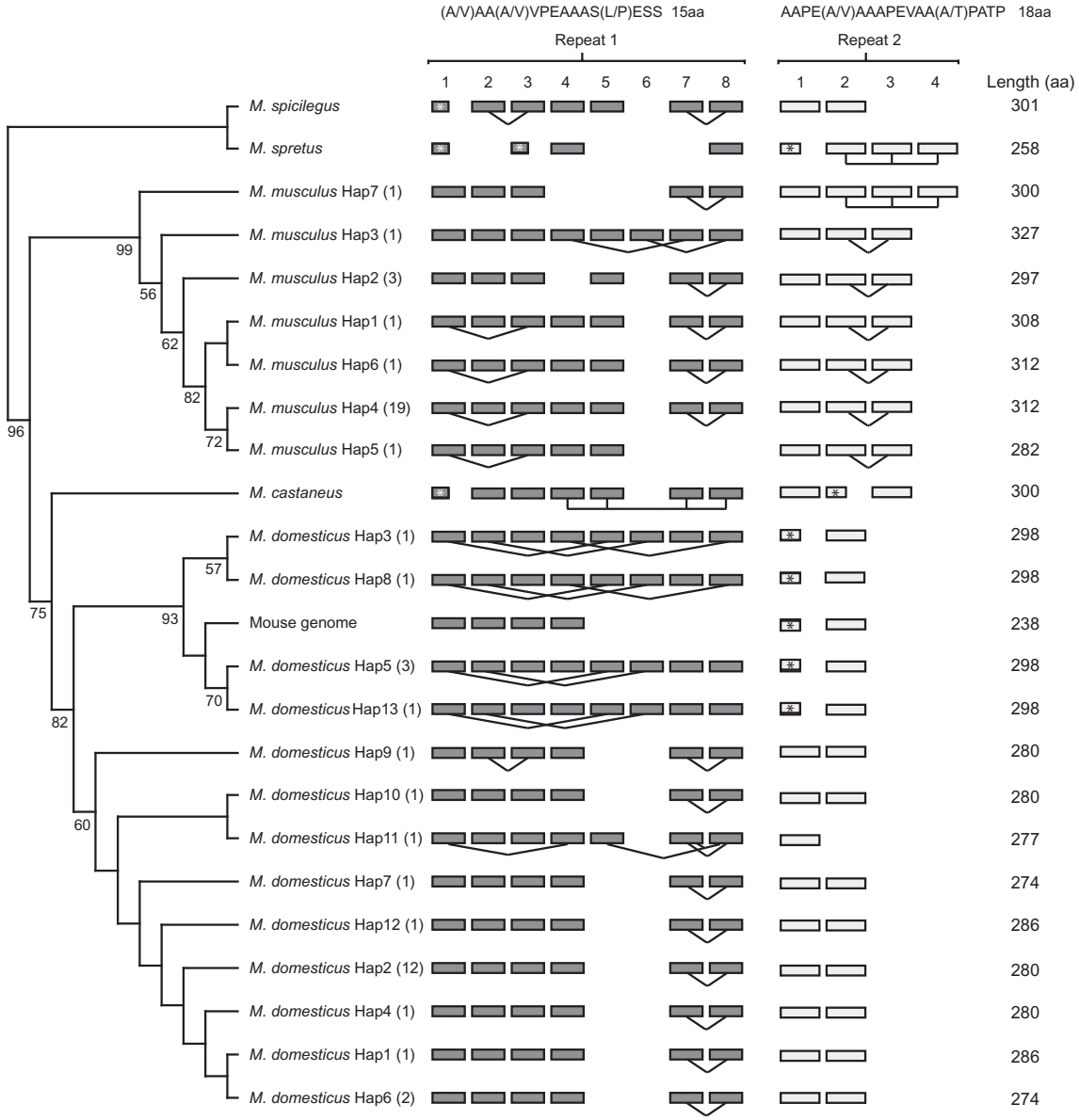

Fig. 2. ML phylogeny for unique nucleotide haplotypes of *Tsga8* based on the best-fit general time reversible + *I* model of sequence evolution. Node support is based on ML bootstraps and haplotype counts are given in parentheses. Per haplotype counts of the two imperfect repeat motifs are shown in 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.

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 d_N:d_S, 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.

Table 3. Nucleotide Variation within *M. domesticus* and *M. musculus*.

							π (%)		θ (%)			D_{XY}[e] (%)	
Locus/Sample	*n*	All[a]	Coding[a]	Silent[a] Sites	Hap.[b]	*S*[c]	All	Silent[d]	All	Silent[d]	Tajima's *D*[d]	All	Silent[d]
4933436I01Rik													
Mus domesticus	32	1,649	1,287	416	7	6	0.038	0.03	0.090	0.119	−1.504*	1.58	1.94
M. musculus	39	1,649	1,287	416	3	3	0.015	—	0.043	—	—	1.76	2.16
Magea9													
M. domesticus	32	794	729	110	3	2	0.109	—	0.063	—	—	2.75	3.64
M. musculus	44	799	708	112	2	1	0.056	—	0.029	—	—	2.47	2.68
Tsga8													
M. domesticus	26	981	804	335	11	12	0.210	0.202	0.321	0.391	−1.37	1.77	2.12
M. musculus	22	1,101	924	368	4	7	0.072	0.049	0.174	0.149	−1.52*	2.05	1.63
Pbsn													
M. domesticus	31	1,995	522	1,297	7	8	0.068	0.041	0.100	0.058	−0.689	1.18	1.39
M. musculus	27	2,649	522	1,951	5	5	0.017	0.019	0.049	0.053	−1.714*	1.07	1.14
Tsx													
M. domesticus	28	3,740	432	3,078	8	9	0.040	0.032	0.062	0.05	−1.075	1.34	1.40
M. musculus	26	3,751	432	3,089	8	6	0.049	0.060	0.042	0.051	0.510	1.32	1.41

[a] Based on aligned regions, excluding indel variation.
[b] Number of unambiguously resolved haplotypes considering all sites.
[c] Total number of polymorphic sites.
[d] Based on noncoding and 4-fold degenerate synonymous coding sites, excluding conserved regions flanking introns (20 bp) and indel variation.
[e] Average percent of nucleotide substitutions per site based on comparison with *M. spretus*.
$P < 0.05$ based on 10,000 coalescent simulations.

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 *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 *Tsga8*, pI = 4.65; rat *Tsga8*, pI = 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 ~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 *Tsga8* rat ortholog is indeed expressed in rat testis. Rat-specific *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), *Tsga8* expression was detected only in the testis (supplemental fig. 5, Supplementary Material online).

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

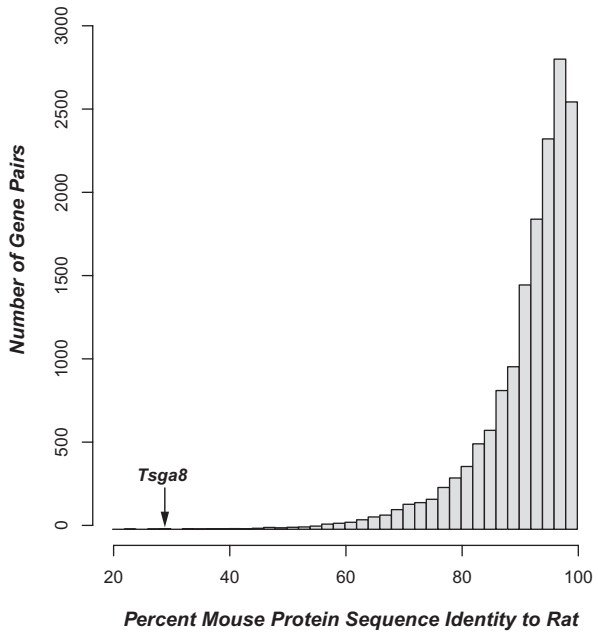

Fig. 3. Protein sequence identity between mouse and rat orthologous gene pairs. The distribution of percent sequence identity is based on 15,947 one-to-one orthologous pairs. Only ten genes show lower pairwise protein identity than *Tsga8* (28.6%), placing *Tsga8* in the lower 0.07% of this distribution. Our estimate of percent sequence identity for mouse *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).

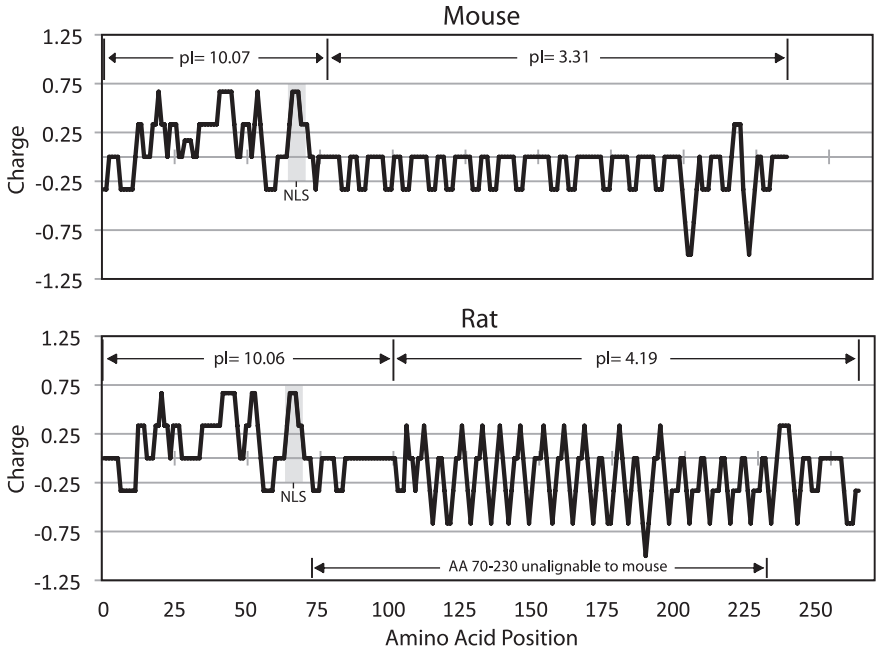

Fig. 4. Sliding window analysis showing overall similarity of amino acid charge for *Tsga8* in mouse and rat (3 aa window size). The pI points are shown for the N- and C-termini. The nuclear localization signal identified by Uchida et al. (2000) is shown for mouse (aa 64–70) and projected by alignment on the rat sequence (aa 63–69). Rat amino acid positions 70–230 are unalignable to the mouse.

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 *Trnp2*) 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

(Good et al. 2010). *Tsga8* is among 32 postmeiotic genes that appear to be misexpressed in sterile hybrid males (Good et al. 2010) and thus may contribute to F_1 sterility. However, overexpression of X-linked genes in sterile hybrid males appears to be a chromosome-wide effect, and it is difficult to evaluate the relative contribution of individual genes to the F_1 breakdown of spermatogenesis. *Tsga8* is not within the confidence intervals of any major QTL identified in a recent mapping study of abnormal head morphology on the *M. musculus* X chromosome (Good, Dean, et al. 2008), although extensive linkage to sterility spanned most of the X chromosome in this study. It is also possible that *Tsga8* contributes to a phenotype that influences reproductive isolation in nature but that has not been considered in laboratory mapping experiments.

Based on available experimental data, *4933436I01Rik* remains a strong candidate for contributing to X-linked hybrid male sterility between *M. domesticus* and *M. musculus*. This gene occurs in a relatively narrow interval on the *M. musculus* X chromosome (8.44 Mb) that contains one or more QTL of major effect on abnormal sperm head morphology and reduced testis size (Good, Dean, et al. 2008). *4933436I01Rik* is a testis-specific postmeiotic-expressed gene that is rapidly evolving across lineages of *Mus* (table 2), including three fixed nonsynonymous changes that map to the *M. musculus* lineage. Furthermore, *4933436I01Rik* is one of nine genes expressed during spermatogenesis in this QTL interval (Good, Dean, et al. 2008), and one of three postmeiotically expressed loci within this region that also are misexpressed in sterile F_1 hybrid males (Good et al. 2010). Evolutionary divergence due to positive selection is thought to play an important role in the evolution of hybrid incompatibilities (Coyne and Orr 2004; Presgraves 2010) and there is strong evidence that *4933436I01Rik* has been subject to recurrent positive directional selection within murid rodents (table 3). Nevertheless, the current data do not provide any statistical evidence for recent positive selection on *4933436I01Rik* in *M. musculus*. Further characterization of *4933436I01Rik* and other genes within this region combined with refined mapping information will be necessary to determine if any of these loci are directly involved in speciation between *M. domesticus* and *M. musculus*.

Supplementary Material

Supplementary figures S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Bret Payseur, Matt Dean, Miguel Carneiro, Gabriela Wlasiuk, Armando Geraldes, Matt Saunders, Jim Krenz, and Tovah Salcedo for critical input on data generation and analysis. Carlos Machado, Therese Markow, and two anonymous reviewers provided comments on this manuscript. We are grateful to Mike Hammer, Barbara Gibson, and Jaroslav Pialek for providing samples of mice. This research 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

Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol*. 18:1585–1592.

Baines JF, Harr B. 2007. Reduced X-linked diversity in derived populations of house mice. *Genetics* 175:1911–1921.

Barbash DA, Siino DF, Tarone AM, Roote J. 2003. A rapidly evolving MYB-related protein causes species isolation in *Drosophila*. *Proc Natl Acad Sci U S A*. 100:5302–5307.

Begun DJ, Holloway AJ, Stevens K, et al. (13 co-authors). 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol*. 5:e310.

Brideau NJ, Flores HA, Wang J, Maheshwari S, Wang X, Barbash DA. 2006. Two Dobzhansky-Muller genes interact to cause hybrid lethality in *Drosophila*. *Science* 314:1292–1295.

Brodie R, Roper RL, Upton C. 2004. JDotter: a Java interface to multiple dotplots generated by dotter. *Bioinformatics* 20:279–281.

Bustamante CD, Fledel-Alon A, Williamson S, et al. (14 co-authors). 2005. Natural selection on protein-coding genes in the human genome. *Nature* 437:1153–1157.

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

Chen JM, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP. 2007. Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet*. 8:762–775.

Chomez P, De Backer O, Bertrand M, De Plaen E, Boon T, Lucas S. 2001. An overview of the *MAGE* gene family with the identification of all human memebers of the family. *Cancer Res*. 61:5544–5551.

Clark AG. 2002. Sperm competition and the maintenance of polymorphism. *Heredity* 88:148–153.

Clark NL, Swanson WJ. 2005. Pervasive adaptive evolution in primate seminal proteins. *PLoS Genet*. 1:335–342.

Coyne JA, Orr HA. 2004. Speciation. Sunderland (MA): Sinauer Associates, Inc.

Cunningham DB, Segretain D, Arnaud D, Rogner UC, Avner P. 1998. The mouse *Tsx* gene is expressed in sertoli cells of the adult testis and transiently in premeiotic germ cells during puberty. *Dev Biol*. 204:345–360.

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.

Dean MD, Good JM, Nachman MW. 2008. Adaptive evolution of proteins secreted during sperm maturation: an analysis of the mouse epididymal transcriptome. *Mol Biol Evol*. 25:383–392.

Dod B, Jermiin LS, Boursot P, Chapman VH, Nielsen JT, Bonhomme F. 1993. Counterselection on sex chromosomes in the *Mus musculus* European hybrid zone. *J Evol Biol*. 6:529–546.

Eddy EM. 2002. Male germ cell gene expression. *Rec Prog Hormone Res*. 57:103–128.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792–1797.

Elliott RW, Miller DR, Pearsall RS, Hohman C, Zhang YK, Poslinski D, Tabaczynski DA, Chapman VM. 2001. Genetic analysis of testis

weight and fertility in an interspecies hybrid congenic strain for chromosome X. *Mamm Genome*. 12:45–51.
Elliott RW, Poslinski D, Tabaczynski D, Hohman C, Pazik J. 2004. Loci affecting male fertility in hybrids between *Mus macedonicus* and C57BL/6. *Mamm Genome*. 15:704–710.
Forejt J. 1996. Hybrid sterility in the mouse. *Trends Genet*. 12:412–417.
Frank SA. 1991. Divergence of meiotic drive suppression systems as an explanation for sex-biased hybrid sterility and inviability. *Evolution* 45:262–267.
Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. 2003. ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res*. 31:3784–3788.
Geraldes A, Basset P, Gibson B, Smith KL, Harr B, Yu HT, Bulatova N, Ziv Y, Nachman MW. 2008. Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. *Mol Ecol*. 17:5349–5363.
Good JM, Dean MD, Nachman MW. 2008. A complex genetic basis to X-linked hybrid male sterility between two species of house mice. *Genetics* 179:2213–2228.
Good JM, Giger T, Dean MD, Nachman MW. 2010. Widespread over-expression of the X chromosome in sterile F1 hybrid mice. *PLoS Genet*. 6:e1001148.
Good JM, Handel MA, Nachman MW. 2008. Asymmetry and polymorphism of hybrid male sterility during the early stages of speciation in house mice. *Evolution* 62:50–65.
Good JM, Nachman MW. 2005. Rates of protein evolution are positively correlated with developmental timing of expression during mouse spermatogenesis. *Mol Biol Evol*. 22:1044–1052.
Gross SS, Brent MR. 2006. Using multiple alignments to improve gene prediction. *J Comput Biol*. 13:379–393.
Guenet J-L, Nagamine C, Simon-Chazottes D, Montagutelli X, Bonhomme F. 1990. *Hst-3*: an X-linked hybrid sterility gene. *Genet Res*. 163–165.
Guenet JL, Bonhomme F. 2003. Wild mice: an ever-increasing contribution to a popular mammalian model. *Trends Genet*. 19:24–31.
Handel MA. 2004. The XY body: a specialized meiotic chromatin domain. *Exp Cell Res*. 296:57–63.
Hayashi K, Yoshida K, Matsui Y. 2005. A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. *Nature* 438:374–378.
Hudson RR, Kreitman M, Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116:153–159.
Hurst LD, Pomiankowski A. 1991. Causes of sex-ratio bias may account for unisexual sterility in hybrids: a new explanation of Haldane's rule and related phenomena. *Genetics* 128: 841–858.
Joseph SB, Kirkpatrick M. 2004. Haploid selection in animals. *Trends Ecol Evol*. 19:592–597.
Kelleher ES, Swanson WJ, Markow TA. 2007. Gene duplication and adaptive evolution of digestive proteases in *Drosophila arizonae* female reproductive tracts. *PLoS Genet*. 3:1541–1549.
Kent WJ. 2002. BLAT—The BLAST-like alignment tool. *Genome Res*. 12:656–664.
Khil PP, Smirnova NA, Romanienko PJ, Camerini-Otero RD. 2004. The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. *Nat Genet*. 36:642–646.
Kuhn RM, Karolchik D, Zweig AS, et al. (22 co-authors). 2009. The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res*. 37:D755–D761.
Lercher MJ, Urrutia AO, Hurst LD. 2003. Evidence that the human X chromosome is enriched for male-specific but not female-specific genes. *Mol Biol Evol*. 20:1113–1116.
Macholán M, Munclinger P, Sugerková M, Dufková P, Bímová B, Bozíková E, Zima J, Piálek J. 2007. Genetic analysis of autosomal and X-linked markers across a mouse hybrid zone. *Evolution* 61:746–771.
Maddison WP, Maddison DR. 2010. Mesquite: a modular system for evolutionary analysis. Version 2.74 http://mesquiteproject.org.
Makova K, Yang S, Chiaromonte F. 2004. Insertion and deletions are male biased too: a whole-genome analysis in rodents. *Genome Res*. 14:567–573.
Mank JE, Vicoso B, Berlin S, Charlesworth B. 2010. Effective population size and the faster-X effect: empirical results and their interpretation. *Evolution* 64:663–674.
Martin-Coello J, Dopazo H, Arbiza L, Ausio J, Roldan ERS, Gomendio M. 2009. Sexual selection drives weak positive selection in protamine genes and high promoter divergence, enhancing sperm competitiveness. *Proc R Soc B Biol Sci*. 276:2427–2436.
Meiklejohn CD, Tao Y. 2010. Genetic conflict and sex chromosome evolution. *Trends Ecol Evol*. 25:215–223.
Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J. 2009. A mouse speciation gene encodes a meiotic histone H3 methyltransferase. *Science* 323:373–375.
Morales CR, Wu XQ, Hecht NB. 1998. The DNA/RNA-binding protein, TB-RBP, moves from the nucleus to the cytoplasm and through intercellular bridges in male germ cells. *Dev Biol*. 201:113–123.
Mueller JL, Mahadevaiah SK, Park PJ, Warburton PE, Page DC, Turner JMA. 2008. The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression. *Nat Genet*. 40:794–799.
Munclinger P, Boziková E, Sugerková M, Piálek J, Macholán M. 2002. Genetic variation in house mice (*Mus*, Muridae, Rodentia) from the Czech and Slovak Republics. *Folia Zool*. 51:81–92.
Namekawa SH, Park PJ, Zhang L-F, Shima JE, McCarrey JR, Griswold MD, Lee JT. 2006. Postmeiotic sex chromatin in the male germline of mice. *Curr Biol*. 16:660–667.
Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A*. 76:5269–5273.
Oka A, Aoto T, Totsuka Y, et al. 2007. Disruption of genetic interaction between two autosomal regions and the X chromosome causes reproductive isolation between mouse strains derived from different subspecies. *Genetics* 175:185–197.
Oka A, Mita A, Sakurai-Yamatani N, Yamamoto A, Takagi N, Takano-Shimizu T, Toshimori K, Moriwaki K, Shiroishi T. 2004. Hybrid breakdown caused by substitution of the X chromosome between two mouse subspecies. *Genetics* 166:913–924.
Oka A, Mita A, Takada Y, Koseki H, Shiroishi T. 2010. Reproductive isolation in hybrid mice due to spermatogenesis defects at three meiotic stages. *Genetics* 186:339–351.
Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, Phadnis N, Beatson SA, Lunter G, Malik HS, Ponting CP. 2009. Accelerated evolution of the *Prdm9* speciation gene across diverse metazoan taxa. *PLoS Genet*. 5:e1000753.
Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet*. 7:337–348.
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.
Parker GA. 1970. Sperm competition and its evolutionary consequences in the insects. *Biol Rev*. 45:525–567.
Payseur BA, Krenz JG, Nachman MW. 2004. Differential patterns of introgression across the X chromosome in a hybrid zone between two species of house mice. *Evolution* 58:2064–2078.

Podlaha O, Webb DM, Tucker PK, Zhang JZ. 2005. Positive selection for indel substitutions in the rodent sperm protein *Catsper1*. *Mol Biol Evol*. 22:1845–1852.

Podlaha O, Zhang JZ. 2003. Positive selection on protein-length in the evolution of a primate sperm ion channel. *Proc Natl Acad Sci U S A*. 100:12241–12246.

Posada D. 2008. jModelTest: phylogenetic model averaging. *Mol Biol Evol*. 25:1253–1256.

Presgraves DC. 2008. Sex chromosomes and speciation in *Drosophila*. *Trends Genet*. 24:336–343.

Presgraves DC. 2010. The molecular evolutionary basis of species formation. *Nat Rev Genet*. 11:175–180.

Queralt R, Adroer R, Oliva R, Winkfein RJ, Retief JD, Dixon GH. 1995. Evolution of protamine P1 genes in mammals. *J Mol Evol*. 40:601–607.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European molecular biology open software suite. *Trends Genet*. 16:276–277.

Rice WR. 1984. Sex chromosomes and the evolution of sexual dimorphism. *Evolution* 38:735–742.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19:2496–2497.

Rozen S, Skaletsky HJ. 2000. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, editor. Bioinformatics methods and protocols: methods in molecular biology. Totowa (NJ): Humana Press. p. 365–386.

Russell LD, Ettlin RA, Sinha Hikin AP, Clegg ED. 1990. Histological and histopathological evaluation of the testis. Clearwater (FL): Cache River Press.

Salcedo T, Geraldes A, Nachman MW. 2007. Nucleotide variation in wild and inbred mice. *Genetics* 177:2277–2291.

Schully SD, Hellberg ME. 2006. Positive selection on nucleotide substitutions and indels in accessory gland proteins of the *Drosophila pseudoobscura* subgroup. *J Mol Evol*. 62:793–802.

Schultz N, Hamra FK, Garbers DL. 2003. A multitude of genes expressed solely in meiotic or postmeiotic spermatogenic cells offers a myriad of contraceptive targets. *Proc Natl Acad Sci U S A*. 100:12201–12206.

Sluka P, O'Donnell L, Stanton PG. 2002. Stage-specific expression of genes associated with rat spermatogenesis: characterization by laser-capture microdissection and real-time polymerase chain reaction. *Biol Reprod*. 67:820–828.

Song R, Ro S, Michaels JD, Park C, McCarrey JR, Yan W. 2009. Many X-linked microRNAs escape meiotic sex chromosome inactivation. *Nat Genet*. 41:488–493.

Sonnhammer ELL, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene Combis*. 167:1–10.

Storchová R, Gregorová S, Buckiová D, Kyselová V, Divina P, Forejt J. 2004. Genetic analysis of X-linked hybrid sterility in the house mouse. *Mamm Genome*. 15:515–524.

Su AI, Wiltshire T, Batalov S, et al. (13 co-authors). 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*. 101:6062–6067.

Swanson WJ, Nielsen R, Yang QF. 2003. Pervasive adaptive evolution in mammalian fertilization proteins. *Mol Biol Evol*. 20:18–20.

Swanson WJ, Vacquier VD. 2002. The rapid evolution of reproductive proteins. *Nat Rev Genet*. 3:137–144.

Swanson WJ, Wong A, Wolfner MF, Aquadro CF. 2004. Evolutionary expressed sequence tag analysis of *Drosophila* female reproductive tracts identifies genes subjected to positive selection. *Genetics* 168:1457–1465.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.

Teeter KC, Payseur BA, Harris LW, Bakewell MA, Thibodeau LM, O'Brien JE, Krenz JG, Sans-Fuentes MA, Nachman MW, Tucker PK. 2008. Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Res*. 18:67–76.

Ting CT, Tsaur SC, Wu ML, Wu C-I. 1998. A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science* 282:1501–1504.

Torgerson DG, Kulathinal RJ, Singh RS. 2002. Mammalian sperm proteins are rapidly evolving: evidence of positive selection in functionally diverse genes. *Mol Biol Evol*. 19:1973–1980.

Torgerson DG, Singh RS. 2003. Sex-linked mammalian sperm proteins evolve faster than autosomal ones. *Mol Biol Evol*. 20:1705–1709.

Torgerson DG, Singh RS. 2006. Enhanced adaptive evolution of sperm-expressed genes on the mammalian X chromosome. *Heredity* 96:39–44.

Toshimori K, Ito C. 2003. Formation and organization of the mammalian sperm head. *Arch Histol Cytol*. 66:383–396.

Tucker PK, Sage RD, Warner J, Wilson AC, Eicher EM. 1992. Abrupt cline for sex chromosomes in a hybrid zone between two species of mice. *Evolution* 46:1146–1163.

Turner LM, Chuong EB, Hoekstra HE. 2008. Comparative analysis of testis protein evolution in rodents. *Genetics* 179:2075–2089.

Uchida K, Tsuchida J, Tanaka H, et al. (11 co-authors). 2000. Cloning and characterization of a complementary deoxyribonucleic acid encoding haploid-specific alanine-rich acidic protein located on chromosome-X. *Biol Reprod*. 63:993–999.

Vicoso B, Charlesworth B. 2009. Effective population size and the faster-X effect: an extended model. *Evolution* 63:2413–2426.

von Salome J, Gyllensten U, Bergstrom TF. 2007. Full-length sequence analysis of the HLA-DRB1 locus suggests a recent origin of alleles. *Immunogenetics* 59:261–271.

Wagstaff BJ, Begun DJ. 2005. Molecular population genetics of accessory gland protein genes and testis-expressed genes in *Drosophila mojavensis* and *D. arizonae*. *Genetics* 171:1083–1101.

Wang PJ, McCarrey JR, Yang F, Page DC. 2001. An abundance of X-linked genes expressed in spermatogonia. *Nat Genet*. 27:422–426.

Ward WS, Coffey DS. 1991. DNA packaging and organization in the mammalian spermatozoa: comparison with somatic cells. *Biol Reprod*. 44:569–574.

Watterson GA. 1975. Number of segregating sites in genetic models without recombination. *Theor Popul Biol*. 7:256–276.

Wyckoff GJ, Wang W, Wu CI. 2000. Rapid evolution of male reproductive genes in the descent of man. *Nature* 403:304–309.

Yamashita S, Wakazono K, Nomoto T, Tsujino Y, Kuramoto T, Ushijima T. 2005. Expression quantitative trait loci analysis of 13 genes in the rat prostate. *Genetics* 171:1231–1238.

Yang ZH. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol*. 15:568–573.

Yang ZH. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24:1586–1591.

Yip SP. 2002. Sequence variation at the human ABO locus. *Ann Human Genet*. 66:1–27.

Zwickl DJ. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Austin (TX): The University of Texas at Austin.