COMPARATIVE ANALYSIS OF TESTIS PROTEIN EVOLUTION IN RODENTS

LESLIE M. TURNER¹, EDWARD B. CHUONG¹,* AND HOPI E. HOEKSTRA²

¹Division of Biological Sciences, University of California at San Diego, La Jolla, CA 92093
²Department of Organismic and Evolutionary Biology and The Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138

Correspondence (and present address):
Leslie M. Turner
Department of Evolutionary Genetics
Max Planck Institute for Evolutionary Biology
August Thienemann Str. 2
24306 Ploen
Germany
Tel: +49 (0)4522 763 287
email: turner@evolbio.mpg.de

*Present address:
Edward Chuong
Department of Genetics
Stanford University
Stanford, CA 94305-5120

Sequence data from this article have been deposited with the GenBank Data Libraries under accession nos. XXXX-XXXX.

Running head: Testis protein evolution in rodents

Key words: EST, fertilization, Muroid rodents, Peromyscus, positive selection, reproductive isolation
ABSTRACT

Genes expressed in testes are critical to male reproductive success, affecting spermatogenesis, sperm competition, and sperm-egg interaction. Comparing the evolution of testis proteins at different taxonomic levels can reveal which genes and functional classes are targets of natural and sexual selection, and whether the same genes are targets among taxa. Here we examine the evolution of testis-expressed proteins at different levels of divergence among three rodents, mouse (Mus musculus), rat (Rattus norvegicus) and deer mouse (Peromyscus maniculatus), to identify rapidly evolving genes. Comparison of expressed sequence tags (ESTs) from testes suggests that proteins with testis-specific expression evolve more rapidly on average than proteins with maximal expression in other tissues. Genes with the highest rates of evolution have a variety of functional roles including signal transduction, DNA binding, and egg-sperm interaction. Most of these rapidly evolving genes have not been identified previously as targets of selection in comparisons among more divergent mammals. To determine if these genes are evolving rapidly among closely related species, we sequenced eleven of these genes in six Peromyscus species and found evidence for positive selection in five of them. Together, these results demonstrate rapid evolution of functionally diverse testis-expressed proteins in rodents, including the identification of amino acids under lineage-specific selection in Peromyscus. Evidence for positive selection among closely related species suggests that changes in these proteins may have consequences for reproductive isolation.
INTRODUCTION

One of the most striking patterns in molecular evolution is that reproductive proteins evolve faster than other protein classes, a pattern consistent across diverse taxa (Singh and Kulathinal 2000; Swanson and Vacquier 2002; Clark et al. 2006). These rapidly evolving proteins serve diverse functions in both males and females and act at various stages of the fertilization process ranging from navigation of sperm through the female reproductive tract through egg-sperm fusion (Clark et al. 2006). Many questions, however, remain unresolved: (1) Do proteins involved in particular biological functions or participating in specific steps of fertilization evolve more rapidly than others? (2) Are the same proteins and amino acid sites targets of selection in different taxa? (3) Does divergence in reproductive proteins contribute to reproductive isolation between closely related taxa?

In mammals, research on reproductive protein evolution has focused primarily on sequence analysis of candidate genes chosen because of their role in fertilization. This approach has identified positive selection (mainly based on relative rates of nonsynonymous versus synonymous change) acting on genes involved in sperm motility, semen coagulation, sperm-egg binding, and sperm-egg fusion (Clark et al. 2006). The functions of numerous proteins involved in fertilization, however, are unknown (Jansen et al. 2001; Tanphaichitr et al. 2007); therefore candidate gene approaches are likely to miss important targets of selection. In contrast, a genome-wide analysis of reproductive proteins can characterize general patterns of evolution as well as identify rapidly evolving genes. Such genomic approaches have been particularly useful in identifying rapidly evolving male accessory gland proteins (Acps) in Drosophila (Swanson et al. 2001b) and crickets (Andres et al. 2006; Braswell et al. 2006), female reproductive tract
proteins in *Drosophila* (Swanson et al. 2001a), and seminal proteins in primates (Clark and Swanson 2005).

We use a genomic approach to characterize reproductive protein evolution in three rodents, mouse (*Mus musculus*), rat (*Rattus norvegicus*), and deer mouse (*Peromyscus maniculatus*). Rodents are an excellent system for investigating mammalian reproductive protein evolution. Fertilization is better characterized in *Mus* than any other mammal, due to its importance as a model in human reproductive health research (Nixon et al. 2007). Both *Mus* and *Rattus* have complete genome sequences that are well annotated, enabling broad-scale comparisons and links to protein function. In contrast, *Peromyscus* has been studied extensively in the wild, and is a speciose genus with well-documented diversity in reproductive morphology, physiology, ecology, behavior, and mating system (Kleiman 1977; Wolf 1989; Dewey and Dawson 2001). Because *Peromyscus* exhibit a range of reproductive isolation among populations, subspecies, sister species, species and species groups (Dice 1933; Liu 1953; Maddock and Dawson 1974), we can gain a more in-depth understanding of patterns of reproductive protein evolution and their potential consequences for speciation (Coyne and Orr 2004). Moreover, the intensity of sperm competition and sexual conflict, two selective forces that may drive reproductive protein divergence, is expected to differ among species of *Peromyscus* with different mating systems. Comparisons of reproductive protein evolution between these species can thus inform our understanding of the nature of selection acting on fertilization.

Here we report a comparative genomic analysis of testis-expressed proteins in these rodents. First, we identify rapidly evolving proteins by comparing expressed sequence tag (EST) sequences from testes of *M. musculus*, *R. norvegicus*, and *P. maniculatus* to orthologous sequences from the *Mus* and *Rattus* genomes. Comparisons of multiple species allow us to test
for differences in lineage-specific rates of evolution. Second, we test for evidence of positive selection on these proteins at a timescale relevant to reproductive isolation by sequencing a subset of rapidly evolving genes in six *Peromyscus* species. Together, these analyses identify a large number of rapidly evolving proteins, many of which have not been implicated previously as targets of selection, and specific amino acid sites that may play a role in reproductive isolation among rodents.

**MATERIALS AND METHODS**

*Peromyscus maniculatus* testis cDNA library construction and EST sequencing: We had a cDNA library prepared by Amplicon Express (Pullman, WA) using *P. maniculatus* testis tissue from a single adult male with the ZAP-cDNA synthesis kit (Stratagene, La Jolla, CA). EST sequencing is described in Weston Glenn *et al.* (in press, GenBank accession numbers EV469380 - 472065). Briefly, the cDNA library was amplified, and phagemids excised according to the manufacturer’s protocol. Resulting colonies were grown overnight in Luria-Bertani/ampicillin broth in deep well plates. PCR-amplified inserts from bacterial cultures of 4800 clones were sequenced from the 5’ end using BigDye terminator (v. 3.1, Applied Biosystems, Valencia, CA) and run on an ABI automated sequencer (3100, 3730, 3730XL, Applied Biosystems). We made base calls (with embedded PHRED, EWING and GREEN 1998; EWING *et al.* 1998) and trimmed both vector sequences and sequence ends from ESTs to reduce error rate to < 0.05 (PHRED quality value > 13) using the program ALIGNER (CodonCode, Dedham, MA). We discarded sequences < 90 bp in length and changed remaining bases with quality values < 13 to unknowns. We assembled ESTs into contigs using the CAP3 sequence assembly program (HUANG AND MADAN 1999).
**Evolutionary EST analysis:** We obtained testis cDNA sequences from *Mus musculus* (Stratagene mouse testis library, 6,068 sequences; RIKEN full-length enriched mouse testis cDNA library, 14,000 sequences) and *Rattus norvegicus* (NIH_MGC_238 library, 13,046 sequences) from the NCBI dbEST database. We identified orthologues by pairwise comparison of ESTs to transcript libraries from the NCBI Reference Sequence (RefSeq) database (nuclear chromosomal cDNA only, downloaded December, 2006) using FASTX (v. 3.3, default settings, PEARSON 1990). We made the following four comparisons (EST vs. RefSeq): *P. maniculatus* vs. *M. musculus* (PM), *P. maniculatus* vs. *R. norvegicus* (PR), *M. musculus* vs. *R. norvegicus* (MR), and *R. norvegicus* vs. *M. musculus* (RM). We made the reciprocal MR and RM comparisons to increase the sample size of orthologue pairs and to assess the effects of differences between EST and genome sequence sources on evolutionary rate estimates. We defined orthologues as sequence pairs that have a minimum of 40% sequence identity for > 20% of EST length. If multiple RefSeqs met these criteria, the most likely orthologue was determined as either: (1) the sequence with the greatest amino acid identity (% sequence identity * alignment length), or (2) the sequence with the lowest divergence at synonymous sites ($d_S$). There were few discrepancies between these criteria, and most of these were matches to alternate isoforms. In these few cases, we used the first criterion, amino acid identity, because it is more conservative; estimates of rate of evolution (i.e. $\omega$, defined below) for orthologous pairs based on amino acid identity were the same or lower than estimates determined for best match based on $d_S$. We concatenated non-overlapping ESTs matching the same RefSeq.

For each orthologous pair, we estimated the rate of evolution as the ratio of nonsynonymous substitution rate to synonymous substitution rate ($d_N/d_S = \omega$). $\omega$ values for
neutrally evolving genes are expected to equal one whereas $\omega$ values < 1 indicate purifying selection and $\omega$ > 1 is considered strong evidence for positive selection. This test is stringent, as pairwise $\omega$ values are averaged across all amino acid sites. A literature survey of studies that used a maximum likelihood (ML) approach (Yang et al. 2000) to detect selection showed that most genes with overall $\omega$ > 0.5 show evidence for positive selection acting on a subset of amino acid sites (Swanson et al. 2004). We therefore classify all genes identified here with $\omega$ > 0.5 as “rapidly evolving”.

We estimated $\omega$ using ML as implemented in the CODEML program from the PAML package (runmode -2, v 3.14, Yang 2000). We excluded orthologue pairs with $d_s$ > 1.5 from further analysis as these are unreliable due to estimation errors (Castillo-Davis et al. 2004). For pairs with estimated $\omega$ values > 1, we ran an additional model in PAML with $\omega$ fixed at 1. To determine whether the estimated value of $\omega$ was significantly greater than one, we compared the estimated $\omega$ model to the fixed $\omega$ (neutral) model using a likelihood ratio test (LRT). The test statistic for the LRT is the negative of twice the difference in log likelihoods between models ($-2\Delta \ln L$), and is $\chi^2$ distributed with degrees of freedom equal to the difference in number of estimated parameters (in this case 1). Alignment of ESTs, identification of orthologues, and implementation of models in PAML were automated using perl, Bioperl (v. 1.5, Stajich et al. 2002), and PHP scripts.

For rapidly evolving genes with orthologues identified in all three rodents, we performed a three-species comparison to identify lineage-specific increases in the rate of amino acid change. We estimated lineage-specific $\omega$ values using the free ratios model in CODEML and performed a LRT comparing the free ratios model to the single ratio model to test whether there is significant evidence of rate variation across lineages (test statistic = $-2\Delta \ln L$, $\chi^2$, d.f. = 2).
We obtained expression information for *Mus musculus* RefSeqs from the Genomics Institute of the Novartis Research Foundation (GNF) gcRMA-condensed dataset (Wu and Irizarry 2005). We classified expression class solely based on the *Mus* data because testis expression data are not available for *Rattus* or *Peromyscus*; we thus assumed that maximal tissue of expression is the same for all three lineages. We classified tissue specificity following Winter et al. (2004); tissue specificity (T_S) is defined as the expression of a given gene in one tissue relative to total expression of that gene in all tissues. Genes with maximum T_S (maxT_S) < 0.08 are considered “housekeeping” (H) genes. We classified the remaining ESTs with maxT_S for testis as testis-specific (TS) and those with maxT_S for another tissue are classified as non-testis-specific (NTS). Expression patterns from three genes with different values of maxT_S are provided as examples in supplemental Figure S1.

We then compared rates of evolution between H, TS, and NTS ESTs. Because tissue-specific genes evolve more rapidly than genes with broader expression, likely due to reduced pleiotropy (Duret and Mouchiroud 2000; Winter et al. 2004), we compared the ω distribution of ESTs in each expression class using an analysis of covariance (ANCOVA) with level of tissue specificity (maxT_S) as a covariate. We transformed ω and maxT_S values towards normality and equal variances between groups; ω values were natural log-transformed and maxT_S values were arcsine-transformed. To equalize variances among groups, we excluded ESTs with ω = 0. A greater proportion of NTS/H ESTs have ω = 0, thus their exclusion results in a conservative test.

We identified putative secreted proteins (containing signal peptides) using Signal P (v. 3.0, Nielsen and Krogh 1998; Bendtsen et al. 2004) and cell-surface proteins (with predicted transmembrane domains) using TMHMM (Sonnhammer et al. 1998; Krogh et al. 2001). We
tested whether average $\omega$ differed significantly between extracellular (contain a signal peptide and/or a transmembrane domain) and intracellular protein genes using permutation tests (10,000 permutations).

We determined the function of Mus EST homologues using the PANTHER classification system (Thomas et al. 2003). We tested whether any particular biological process or molecular function was overrepresented among rapidly evolving genes by comparing the proportion in the rapidly evolving group relative to expected based on representation among total EST homologues for each species. We tested the significance of overrepresentation using the binomial test (Cho and Campbell 2000) with Bonferroni correction for multiple tests.

**Identification of protein domains in Peromyscus EST sequences:** We searched for all unique Peromyscus testis sequences (unigenes) in the InterPro-combined protein database (Mulder et al. 2007) using InterProScan (Zdobnov and Apweiler 2001), which uses a variety of search algorithms to identify homology between six-frame translations of input nucleotide sequences and known protein domains. This method allows for identification of domains in all ESTs, including those that do not have orthologues identified in Mus or Rattus.

**Additional sequencing in Peromyscus:** We obtained testis tissue from a single male from each of six Peromyscus species (P. aztecus, P. californicus, P. eremicus, P. leucopus, P. maniculatus, P. polionotus) from the Peromyscus Genetic Stock Center (supplemental Table S1). Tissue was excised from freshly sacrificed adult males and stored in RNAlater solution (Sigma, St. Louis, MO). We extracted RNA using the RNeasy kit (Qiagen, Valencia, CA) and synthesized cDNA using a Superscript III RT kit (Invitrogen, Carlsbad, CA). We amplified all genes under standard
PCR conditions, using primers designed by aligning *P. maniculatus* EST sequences to GenBank sequences from *Mus* and *Rattus*. To determine species relationships, we sequenced a 1,213 bp region of the mitochondrial genome (including *COIII* and *ND3*) from one individual from each of the six *Peromyscus* species (supplemental Table S1) using published primers (HOEKSTRA *et al.* 2004).

We directly sequenced or cloned (TOPO-TA, Invitrogen) PCR products. We performed cycle sequencing with BigDye terminator (v. 3.1, Applied Biosystems) and ran products on an ABI 3100 automated sequencer (Applied Biosystems). We checked base calls by eye, assembled contigs, and aligned sequences in SEQUENCHER (Gene Codes, Ann Harbor, Michigan). We used MUSCLE (default parameter settings, EDGAR 2004) when sequence alignments were ambiguous. A large repetitive region from one of the genes (*Gm1276*, see Results) was excluded because reliable alignment was not possible.

**Analysis of *Peromyscus* testis-expressed gene sequences:** To determine species relationships, we constructed Bayesian and maximum likelihood (ML) phylogenies of the six species, based on the mitochondrial sequences and 1,201 bp of the nuclear genes *Mc1r* and *Lcat* [from Genbank (TURNER and HOEKSTRA 2006)]. We identified the most appropriate substitution model (GTR+Γ) using MODELTREE v. 3.7 (POSADA and CRANDALL 1998). A partition homogeneity test implemented in PAUP* (SWOFFORD 2002) was not significant, indicating no conflicts between data partitions. We performed Bayesian analysis in MRBAYES (v. 3.1, HUELSNBECK AND RONQUIST 2001), with data partitioned by gene and codon position. We performed two runs for 10 million generations and discarded the first million generations as burn-in. The 99% credible set for the Bayesian analysis contains a single tree, identical in topology to the ML tree: (((*P.
polionotus, P. maniculatus), P. leucopus, P. aztecus, (P. eremicus, P. californicus)). This topology is consistent with published species trees of Peromyscus (Turner and Hoekstra 2006; Bradley et al. 2007).

Using this ML/Bayesian tree, we implemented the codon-based ML method (Nielsen and Yang 1998; Yang et al. 2000) to detect positive selection in Peromyscus testis genes. This method employs a LRT to compare a neutral model, where \( \omega \) for all sites is constrained to be < 1, to a selection model where a subset of sites have \( \omega \) greater than one (test statistic = \(-2\Delta \ln L, \chi^2\)). We performed the following model comparisons (neutral vs. selection): M1a vs. M2a, M7 vs. M8, and M8A vs. M8. M1a has two site classes, the first with \( 0 < \omega < 1 \) and the second with \( \omega = 1 \). M2a adds an additional ‘selection’ class with \( \omega \geq 1 \). In M7, \( \omega \) varies as a beta distribution between 0 and 1, and M8 adds a selection class with \( \omega \geq 1 \). M8A is a modified version of M8 where \( \omega \) for the selection class is constrained to equal one. The M8A vs. M8 comparison tests whether \( \omega \) is significantly greater than one, providing a control for false positives resulting from a poor fit of the data to the beta distribution. For this comparison, the test statistic is distributed as a 50:50 mixture of a point mass at zero and a \( \chi^2 \) distribution with one degree of freedom (Swanson et al. 2003). We implemented codon models in CODEML. Specific amino acid sites subject to positive selection were identified using the Bayes empirical Bayes (BEB) procedure (Yang et al. 2005).

In addition, we applied codon models to determine whether genes positively selected within Peromyscus have evidence for positive selection among divergent species of mammals. For the five genes with evidence for positive selection within Peromyscus (see Results), we identified homologues from other mammals in GenBank using BLAST (see supplemental Table S2 for species and accession numbers). To avoid significant results due to positive selection
within *Peromyscus*, we included sequence from only a single species, *Peromyscus maniculatus*, in these analyses. We aligned amino acid sequences using default parameter settings in MUSCLE, adjusted the corresponding nucleotide alignments in MEGA (Kumar et al. 2001), and excluded sites with alignment gaps. We constructed neighbor-joining trees in PAUP* using model parameters determined in MODELTEST. We ran codon models in CODEML, as above.

**RESULTS**

**EST sequencing:** Sequencing of 4,800 ESTs from the *P. maniculatus* testis cDNA library resulted in 3,840 quality sequences > 90 bp in length. After removal of redundant sequences and assembly of overlapping sequences into contigs there was a total of 2,364 unigenes (446 contigs, 1,918 singlets).

**Evolutionary EST analysis:** To identify the most rapidly evolving testis proteins, we compared orthologous genes in *Peromyscus, Mus* and *Rattus*. We found orthologues in *Mus* and/or *Rattus* for ~43% of unique *P. maniculatus* EST sequences (Table 1), resulting in 1,014 (PM) and 993 (PR) orthologous pairs. The 20,068 *Mus* EST sequences included 11,203 unigenes; we identified orthologues in *Rattus* for 37% of unigenes, for a total of 4,171 pairs. 13,046 *Rattus* ESTs collapsed into 7,448 unigenes and we found *Mus* orthologues for 56% of these, for a total of 4,207 orthologous pairs. The lower proportion of *Mus* ESTs with identified *Rattus* orthologues is not surprising because the *Rattus* genome sequence was completed more recently than the *Mus* genome, is therefore less well annotated, and has lower sequencing coverage (Waterston et al. 2002; Gibbs et al. 2004).
For each EST-RefSeq comparison, estimates of $\omega$ for the vast majority of orthologue pairs are $<< 1$, consistent with the action of purifying selection (Table 1). A small percentage of pairs (1.3 – 2.4%), however, have $\omega > 1$ (a signature of positive selection) and three of these pairs have $\omega$ values significantly greater than one. These three genes are all from the MR comparison and include a hypothetical protein of unknown function ($LOC691850$) and two microtubule associated proteins: a signaling protein involved in spermatogenesis (Mast2, WALDEN and COWAN 1993), and a protein with microtubule motor activity and a lipid-binding domain whose function in testis is unknown (Stard9, KANNO et al. 2007). An additional 7.5 – 12.2% of orthologue pairs have $0.5 < \omega < 1$. All rapidly evolving genes ($\omega > 0.5$) are listed in supplemental Table S3.

A representative plot of $d_N$ vs. $d_S$ values for all pairs from the PM comparison is presented in Figure 1A. Proportions of ESTs in three $\omega$ classes ($\omega < 0.5$, $0.5 < \omega < 1$, $\omega > 1$) are not significantly different among three of the EST-RefSeq comparisons (PM, PR, RM, $P = 0.96$; Pearson’s $\chi^2$, d.f. = 4). The MR comparison has the largest proportion of ESTs in both rapidly evolving classes (12.2%, $0.5 < \omega < 1$; 2.4%, $\omega > 1$), resulting in a significant effect of species comparison on $\omega$ class when MR is included ($P < 0.001$; Pearson’s $\chi^2$, d.f. = 6). Some genes are rapidly evolving in all species comparisons, whereas other genes appear to be rapidly evolving in just one or two lineages (Figure 2B). Overall, 44% (72/163) of rapidly evolving genes with orthologous pairs identified in multiple species comparisons are rapidly evolving in more than one comparison and 18% (29/163) are rapidly evolving in all comparisons. The genes that have $\omega > 0.5$ in all comparisons (PM, PR, RM and/or MR) are listed in Table 2.
Testis-specific genes evolve rapidly: Genes expressed in testes may also be expressed in other tissues; we used expression data from *Mus* to determine which genes are testis-specific and which have broader patterns of expression. Using these data, we tested if rates of protein evolution are correlated with expression pattern. Expression data were available for 62 - 73% of *Mus* orthologues. For all species comparisons, mean $\omega$ values for testis-specific genes were higher than overall means and means for other expression classes, indicating that testis-specific genes evolve more rapidly on average than non-testis-specific and housekeeping genes (Table 1). This analysis shows that there is a highly significant effect of testis-specific expression on $\omega$ in all species comparisons (ANCOVA, $P < 0.0001$).

In addition, we compared the proportion of genes from each expression class between groups of ESTs with different $\omega$ values ($\omega < 0.5$ versus $\omega > 0.5$). The results provide an intuitive demonstration of the elevated evolutionary rate of testis-specific genes. For all four EST-RefSeq comparisons, there is a highly significant relationship between expression class and $\omega$ class (Pearson’s $\chi^2$, d.f. = 2, $P < 0.0001$). Specifically, the proportion of testis-specific genes was significantly higher among rapidly evolving genes than expected based on the proportion of all genes that are testis-specific (two-tailed binomial test, $P < 0.0001$). Overrepresentation of testis-specific genes among rapidly evolving genes from one comparison (PM) is depicted in Figure 1B.

Rapidly evolving genes are functionally diverse: To determine if particular functional classes of genes tend to be rapidly evolving, we used the PANTHER classification system to assign genes to functional categories. In all three pairwise species comparisons, genes unclassified for both biological process and molecular function are overrepresented in the rapidly evolving class ($P < 0.0001$).
In addition, defense and immunity proteins ($P < 0.002$) and KRAB box transcription factors ($P < 0.005$) are overrepresented in the *Mus-Rattus* comparison (MR/RM).

The list of rapidly evolving genes (Table 2) includes genes with $\omega > 0.5$ in all three species comparisons (PM, PR, RM or MR), and 11 genes chosen for sequencing in additional *Peromyscus* species because they had the highest pairwise $\omega$ values in the PM comparison (excluding hypothetical proteins) in initial EST comparisons. Three genes from the latter category (*Gsg1, Hlfnt, Smcp*) had high $\omega$ estimates in preliminary screens, but much lower $\omega$ estimates in the final screen, subsequent to corrections of alignments or changes in *Mus* RefSeqs.

The amount of functional information available for these rapidly evolving genes varies. Some genes have known roles in sperm-egg interaction [Acr (Howes et al. 2001), Spa17 (Richardson et al. 1994), Spag8 (Cheng et al. 2007)] or spermatogenesis (Hils1, Yan et al. 2003). Another set of genes have inferred functions based on domain homology; a wide variety of functions is represented including receptor activity, DNA binding and protein binding. However, the majority of genes have no available functional information.

We also compared values of $\omega$ for secreted (contain a signal peptide) and cell-surface proteins (contain a transmembrane domain) to $\omega$ values of intracellular proteins. Extracellular proteins are more likely to interact with foreign proteins, such as those of pathogens, the female reproductive tract and gametes, or sperm/seminal proteins from another male and thus may experience stronger selection than intracellular proteins (Swanson et al. 2001b; Clark and Swanson 2005). We find that extracellular proteins have higher $\omega$ values on average than intracellular proteins for all species comparisons (Table 1), and this difference is statistically significant for the PM ($P = 0.037$), RM ($P = 0.003$), and MR ($P = 0.017$) comparisons.
We looked for protein domain homology in all *Peromyscus* ESTs using InterProScan to obtain information about possible functions of genes when we could not identify orthologues in *Mus* or *Rattus* (non-matches). However, this analysis was not informative; very few non-matches had any indication of homology. Seventeen non-matches have predicted signal protein sequences or transmembrane domains, indicating they are extracellular and may interact with foreign proteins.

**Evidence for adaptive evolution of testis genes in *Peromyscus***: For the 11 most rapidly evolving genes identified in the preliminary genomic analysis, we sequenced most or all of the coding regions in six *Peromyscus* species to determine whether there is evidence of adaptive change between closely related taxa. We specifically looked for evidence of rapid amino acid change as indicated by high values of \( \omega \), evidence for positive selection on a subset of amino acid sites, and changes in protein length.

Estimates of pairwise \( \omega \) for the entire coding regions sequenced in *P. maniculatus* vs. *Mus* homologues (Table 3) were consistent with \( \omega \) values for the shorter ESTs in some instances (e.g. *Hils1, Gm1276*) and inconsistent in others (e.g. *Lrrc50, Acr*). Thus, the EST screen identified both genes with high rates of evolution across their entire length as well as genes with rapidly evolving regions. However, \( \omega \) within *Peromyscus* is not significantly correlated with pairwise \( \omega \) estimates of *P. maniculatus* vs. *Mus* \((P = 0.16)\), although there is a non-significant positive trend \((R^2 = 0.20)\). We performed a three-species analysis of *P. maniculatus, Mus* and *Rattus* sequences for each gene in PAML to estimate lineage-specific values of \( \omega \). We tested the hypothesis that the estimate of \( \omega \) for the *Peromyscus* lineage is a better predictor of \( \omega \) among *Peromyscus* species, as this value is not affected by evolution along the *Mus* lineage. However,
there is significant evidence of variation in $\omega$ among lineages for only 3 of the 11 genes ($H1fnt$, $Smcp$, $Gsg1$), and the lineage-specific estimates were actually poorer predictors ($P = 0.60$, $R^2 = 0.03$) of $\omega$ within Peromyscus than the pairwise $P. maniculatus$-$Mus$ estimates. Three of the 11 genes ($Gm1276$, $Spag8$, and $Ddc8$) have $\omega > 0.5$ within Peromyscus, suggesting they are rapidly evolving and may be subject to positive selection.

Using a ML approach, we tested whether any of the 11 genes have statistical evidence for positive selection within Peromyscus. Results from the ML codon models indicate that 5 of the 11 genes show evidence for positive selection in the Peromyscus genus (Table 3). For each of these genes, comparisons of M8 to M8A are significant, indicating that a proportion of sites are subject to selection and $\omega$ for the selected class is significantly greater than one. In addition, comparisons of M2 vs. M1 as well as M8 vs. M7 are also significant for one of these genes ($Gm1276$).

For each gene sequenced in Peromyscus, we determined several estimates of evolutionary rate: pairwise $\omega$ of EST sequences vs. Mus and vs. Rattus RefSeqs, lineage-specific $\omega$ in Peromyscus determined through three-species analysis, and overall $\omega$ determined through comparison of the full $P. maniculatus$ sequence to Mus. Surprisingly, none of these measures was a good predictor of which genes have evidence for positive selection within Peromyscus. For example, of the five positively selected genes, one gene has the highest estimate of $\omega$ vs. Mus from the EST screen ($Lrre50$), whereas another has an $\omega$ estimate < 0.5 ($Gsg1$). Overall measures of $\omega$ within Peromyscus for positively selected genes ranged from 0.22 ($Acr$) to 0.69 ($Gm1276$). Further, none of these genes have remarkably high levels of amino acid variation (range 4.2 – 10.4% variable aa sites), highlighting how genes can be subject to positive selection even when overall variability is low.
Amino acid alignments of these five genes are given in Figure 3, and amino acid sites assigned to the positive selection class using the BEB procedure are indicated. With the exception of two sites in Acr (397, 412), the posterior probabilities of assignment of the sites to the positive selection class are less than 0.95. These results provide a preliminary indication of the spatial distribution of target sites along the protein. For example, in Acr and Lrrc50, target sites are clustered, whereas in Gm1276 and Ddc8, sites are scattered.

In addition to changes in amino acid sequence, changes in protein length in response to selection have been described in reproductive proteins (PODLAHA and ZHANG 2003; PODLAHA et al. 2005; HAWTHORNE et al. 2006). We examined length variation to determine if there is evidence for this type of change in Peromyscus.

One gene, Phf8, has premature stop codons, some of which evolved along the lineage between Peromyscus and the Murids (Mus, Rattus) and some evolved within Peromyscus. In P. aztecsus, there are numerous frameshifts and stop codons; the predicted protein product would be only 66 aa long (full length protein is 908 aa in Mus), and thus unlikely to be functional. We therefore excluded the P. aztecsus sequence from further analysis. The remaining species have premature stops 180 (P. polionotus) or 62 (P. californicus, P. eremius, P. leucopus, P. maniculatus) codons upstream of the Mus stop codon. PHF8 has an exceptionally high proportion of variable amino acid sites (26%) relative to the other proteins (Table 3), and \( \omega \) is relatively high within Peromyscus (0.47) but there is no evidence for positive selection. Thus, high \( \omega \) in Phf8 may result from relaxed purifying selection rather than positive selection.

The gene with the highest \( \omega \) value, Gm1276, contains a large repeat region that varies in length among Peromyscus species from 252 – 360 aa (Fig. 3). Homologous repeat regions are present in all available mammalian sequences. In Mus and Rattus, the regions are similar in
length to the longer sequences found in *Peromyscus* species (360 aa and 330 aa, respectively) but the region is much shorter (≤ 125 aa) in other mammals.

*Smcp* has a repeat region that varies both in length and motif sequence in diverse mammals, including *P. maniculatus*, *Mus*, and *Rattus* (Hawthorne *et al.* 2006). Length variation among *Peromyscus* species is limited; there are a few amino acid indels within repeats, but all six species have the same number of repeats. For *Smcp*, estimates of pairwise ω based on final alignments are low, and there is no evidence for positive selection in *Peromyscus* either in terms of amino acid sequence or length variation (Tables 2 and 3).

**Rapid evolution of testis genes can be lineage specific:** Most evidence for the rapid evolution of reproductive proteins in mammals rests on comparisons of highly divergent taxa. We tested if genes with evidence for positive selection in *Peromyscus* (*Ddc8*, *Gsg1*, *Lrrc50*, *Acr*, and *Gm1276*) also show evidence for positive selection in comparisons among divergent mammals.

For each of the five genes with evidence for positive selection in *Peromyscus*, we identified homologues in 5 – 10 additional mammals. For *Ddc8*, *Gsg1*, and *Lrrc50*, there is no evidence for positive selection in higher-level comparisons (see supplemental Table S4). The M8 vs. M7 and M8A vs. M8 comparisons both identify a subset of sites in *Acr* that are positively selected (*P* ≤ 0.01; *p*ₚ = 0.07, *ω*ₚ = 1.82), however the M2 vs. M1 comparison is not significant (*P* = 0.22).

Such mixed results are similar to previous analysis of *Acr* sequences from a smaller sample of mammals (Swanson *et al.* 2003). For *Gm1276*, all model comparisons provide strong evidence (*P* < 0.0001) for positive selection (M8: *p*ₚ = 0.25, *ω*ₚ = 2.12). The 13 sites with high probability (> 0.9) of being in the positively selected class are distributed along the length of the protein,
similar to the pattern we observe within *Peromyscus*. One site (142) is positively selected at both taxonomic levels.

**DISCUSSION**

Here, we identify a functionally diverse set of genes that are evolving rapidly in rodents, most of which have not been identified previously as targets of selection or functionally characterized. Evolutionary analysis of the same genes at different taxonomic depths often yields different patterns; some genes have evidence for positive selection across divergent mammalian taxa whereas rapid evolution of other genes is lineage-specific. These results show that there is not a common set of targets of positive selection among mammalian reproductive proteins, and instead studies of gametic isolation will require species-specific examination. In addition, we find evidence for positive selection acting on five genes within *Peromyscus*, raising the possibility that these genes may contribute to reduced fertilization potential between these closely related species. This study contributes to a growing body of evidence documenting a remarkable pattern of rapid evolution of reproductive proteins in animals.

**Functional roles of rapidly evolving genes:** Identifying the functions of rapidly evolving genes may reveal whether selection targets a particular biological process or fertilization step which in turn may suggest evolutionary forces (i.e. sperm competition, sexual conflict, pathogen defense) responsible for the rapid evolution of testis-expressed proteins. We find that extracellular proteins are evolving more rapidly, on average, than intracellular proteins (Table 1), suggesting that interaction with foreign proteins (such as female reproductive proteins or proteins from other males) may result in stronger selection. Some of the most rapidly evolving genes in our analysis
have well described roles in fertilization (Table 2). One gene, *Hils1*, is involved in DNA condensation during spermatogenesis, specifically repackaging DNA onto testis-specific histones to produce the densely packed chromatin of sperm. Evidence for positive selection has been reported previously for three other DNA packaging sperm proteins [*Prm1*, *Prm2*, *Tnp2*; (Retief et al. 1993; Queralt et al. 1995; Wyckoff et al. 2000; Torgerson et al. 2002; but see Clark and Civetta 2000; Rooney et al. 2000)], suggesting the process of DNA condensation may be a common target of selection.

Two other genes, *Acr* and *Spa17*, are zona pellucida (egg coat) binding proteins. Much research effort has focused on proteins involved in egg and sperm binding, as this interaction is critical to species-specific fertilization (Wassarman et al. 2001). Numerous gamete-binding proteins from both egg and sperm evolve rapidly and show evidence for positive selection at various levels of divergence in mammals (Swanson et al. 2001a; Jansa et al. 2003; Swanson et al. 2003; Glassey and Civetta 2004; Good and Nachman 2005; Gasper and Swanson 2006; Podlaha et al. 2006; Turner and Hoekstra 2006, in press; Hamm et al. 2007). *Acr* is of particular interest because knockouts have partially infertile phenotypes in *Mus*; homozygous males never sire offspring in competitive mating trials with wild-type males (Jansen et al. 2001).

In addition to these well-characterized genes, many of the most rapidly evolving genes identified here have gene ontology annotations. For these, we have an indication of the general function of the gene but not its involvement in a specific fertilization step. This group includes genes encoding proteins involved in signal transduction (*Gm1276, 4930596D02Rik*) and protein binding (*Lrrc50, Phf8*). The functional diversity among rapidly evolving genes identified here is consistent with a study that compared rates of evolution of genes expressed at different stages of
spermatogenesis in the mouse, which showed that rates of evolution are higher for genes expressed in late stages of spermatogenesis that serve a wide variety of functions (GOOD and NACHMAN 2005). These patterns in rodents are also consistent with previous studies demonstrating adaptive evolution of sperm proteins with diverse functions across more distantly related mammals (TORGERSON et al. 2002; SWANSON et al. 2003).

We identified several rapidly evolving genes with unknown function, some of which have signatures of positive selection at multiple levels of taxonomic divergence. This result underscores the importance of combining analysis of proteins that have well-described function with genomic approaches that facilitate identification of novel targets. Moreover, evolutionary analyses provide valuable data about the molecular processes of reproduction; for example, some of the positively selected genes identified here may play important roles in fertilization.

**Targets of selection at various levels of divergence:** A large number of proteins participate in the processes of mammalian fertilization, and there is considerable functional redundancy between proteins, particularly in males (TANPHAICHITR et al. 2007). Even in the unlikely case that there is one predominant form of selection acting on one particular step of fertilization, it is likely that the individual proteins targeted by selection differ between species.

Since the majority of studies that have demonstrated positive selection on mammalian reproductive proteins have sampled divergent species, it is clear that there are some common targets. Further, some of the genes identified in these analyses have subsequently been found to be subject to positive selection in more closely related taxa. For example, egg coat proteins under selection across divergent mammals (SWANSON et al. 2001a) are evolving adaptively within Murines (JANSA et al. 2003; SWANN et al. 2007) and within *Peromyscus* (TURNER and
HOEKSTRA 2006, in press). In addition, almost all rapidly evolving seminal proteins identified through a comparison of human and chimpanzee sequences have evidence for positive selection when sequenced in a more diverse sample of primate species (CLARK and SWANSON 2005). In this study, however, evidence for rapid evolution and positive selection for some testis proteins is limited when we investigate evolutionary patterns at three taxonomic levels: within a genus (*Peromyscus*), within a superfamily (Muroidea), and within an order (Mammalia). For example, some genes with high rates of evolution between *Peromyscus* and *Mus* have been identified previously as targets of selection in more divergent mammals (e.g. *Spa17*, SWANSON et al. 2003; *Hils1*, GOOD and NACHMAN 2005), but have no evidence for positive selection within *Peromyscus*. The inverse pattern, rapid evolution in closely related taxa but not divergent taxa, was also evident; three (*Ddc8, Gsg1, Lrrc50*) of five genes that are rapidly evolving in Muroidea and positively selected within *Peromyscus* have no evidence for positive selection among diverse mammals. Variation in evolutionary pattern across taxonomic levels might result from variation in the selective agent between taxa, differences in levels of redundancy of genes serving different functions during the fertilization processes, or different degrees of pleiotropy of genes with shared function. Clearly, however, studies that wish to link rapid reproductive protein evolution to gametic isolation will have to examine patterns of nucleotide variation between closely related species as well as test the functional effects of amino acid or expression differences on a case-by-case basis.

**Success of EST screen at identifying positively selected genes in *Peromyscus***: The evolutionary EST analysis we employed here determines evolutionary rates between a single species of *Peromyscus* and two other rodents. High values of $\omega$ between *Peromyscus* and *Mus* or
Rattus, however, may result from rapid evolution solely within Murids or preceding the diversification of the Peromyscus genus. Approximately half of the sequenced genes chosen based on high pairwise $\omega$ in comparisons of Peromyscus vs. Mus show evidence for positive selection within Peromyscus. Moreover, four out of five positively selected genes have not been identified previously as targets of selection in mammals. Therefore, this approach is a promising one for identifying new genes likely to be rapidly evolving in taxa without sequenced genomes.

Since close to half of the genes identified in the EST screen that we sequenced in multiple Peromyscus species have evidence for positive selection, it is likely that there are additional targets of selection among the remaining genes with high rates of evolution. Further analysis of rapidly evolving genes among closely related species of Mus and Rattus will likely yield similar success in identifying targets of selection within those genera. However, as selection acts on a small proportion of amino acid sites in many genes, choosing genes based on $\omega$ values averaged across large regions certainly will miss important targets (Hughes 2007). EST analysis and other genomic approaches are complementary to choosing genes based on knowledge of their biological functions.

**Adaptive change in amino acid sequence and length of testis proteins:** Detailed characterization of 11 testis proteins within Peromyscus allowed us to extend our analysis of the functional targets and lineage-specificity of positive selection to the amino acid level. For two proteins that are subject to positive selection in Peromyscus, proacrosin and Gm1276, functional information is available, even about specific domains. Further, both of these proteins have evidence for positive selection among divergent mammals, allowing us to compare the specific targets of selection among taxa.
In *Peromyscus*, we identified a cluster of selection target sites in the C-terminus of proacrosin (Fig. 3). Evidence for the function of this region differs between species. In boar and human, this region is implicated in binding to the ZP (Mori *et al.* 1995; Furlong *et al.* 2005), but in the mouse there is no evidence of the C-terminus binding to ZP2 (Howes *et al.* 2001). In boar, the C-terminus is cleaved during processing of proacrosin to the mature proteolytic form following ZP binding (Mori *et al.* 1995), thus this region is unlikely to have a role in dissolution of the ZP. Although we identified positively selected sites in both a diverse sample of 11 mammals and within *Peromyscus*, these sites differed. However, two sites in the positive selection class (45, 48) in mammals (supplemental Table S4) neighbor one of the positively selected sites in *Peromyscus* (46). These three sites are ~20 aa upstream of one of the two regions implicated in zona pellucida binding in *Mus* (Jansen *et al.* 2001). The other three positively selected sites in *Peromyscus* are clustered in the C-terminal region, which unfortunately cannot be aligned reliably between divergent species.

The human homologue of Gm1276 is MS4A13 (also known as NYD-SP21). This protein is a member of the membrane-spanning four-domains (MS4A) family, which is part of the CD20/ß subunit of high affinity IgE receptor superfamily (Ishibashi *et al.* 2001). These plasma membrane-bound proteins interact with other cell surface proteins in oligomeric complexes that have signal transduction functions in a variety of tissues. Gm1276/ MS4A13 has highly specific expression in testis in both mouse and human. All eight positively selected sites in *Peromyscus* are located in the extracellular C terminus (Fig. 3). In the analysis of Gm1276 sequences from eight divergent mammals, we identified 13 positively selected sites including one site (101) in the intracellular loop between TM domains 2 and 3, another site (113) in TM domain 3, and 11 sites in the C-terminal extracellular region. One of these C-terminal sites (142) is positively
selected in both this analysis and within *Peromyscus*. As MS4A13 is a putative signaling protein with receptor activity, we can speculate that substitutions (and/or length variation) in the long extracellular domain might affect ligand binding.

A comparison of mouse-human orthologous pairs showed that sperm-specific proteins have exceptionally high rates of evolution, both amino acid substitution rate and variation in protein length (Torgerson *et al.* 2002), suggesting both may be responses to selection. We found that Gm1276 has evolved rapidly in protein length as well as amino acid substitution rate, both in a phylogenetically diverse sample of mammals and within *Peromyscus*. Length variation results from the expansion and contraction of a large repeat region. This repeat region is in the C-terminal extracellular region, downstream of the region homologous to other MS4A family members; querying the InterPro database with the repeat sequence failed to identify any homologous protein domain. This region expanded some time between the divergence of Muroids from other mammalian lineages and the time of the divergence of Cricetids (including *Peromyscus*) from Murids (including *Mus* and *Rattus*). In addition, this repeat varies in length by more than 100 amino acids within *Peromyscus*. These results suggest that this gene may consistently respond to selection through two different mechanisms of sequence evolution; however, although there is evidence that positive selection promotes amino acid substitution in this protein, additional functional data are required to evaluate whether length variation is the result of selection or reduced constraint.

For both *Acr* and *Gm1276*, patterns of amino acid change within *Peromyscus* and across divergent mammals show that selection has repeatedly targeted the same protein regions, and sometimes the same amino acid sites. Taken together, the results from our genome-wide and gene-level analyses reveal that selection on testis proteins ranges widely in scope - from effects
seen in a single lineage to effects common across lineages, even those that have been diverging for more than 100 million years.

**Implications for fertilization and reproductive isolation:** Ultimately, we are interested in finding genes that cause or maintain reproductive isolation between species. In marine invertebrates, through a combination of detailed analysis of evolutionary patterns within and between recently diverged species and functional characterization of positively selected genes, great progress has been made in identifying the selective forces promoting divergence of sperm proteins (Geyer and Palumbi 2003; Levitan and Ferrell 2006; Rigos et al. 2006) and determining the consequences of protein divergence on fertilization potential between species (Lin and Vacquier 1999; Palumbi 1999; Coyne and Orr 2004; Levitan and Ferrell 2006). In mammals, however, a detailed understanding of the causes and consequences of the rapid divergence of reproductive proteins remains elusive (Turner and Hoekstra in press). Progress towards this goal requires the identification and comparison of evolutionary dynamics of these proteins across a range of taxonomic levels as well as experimental assessment of the influence of allelic variation on fertilization success in natural populations with incomplete or recently evolved isolating barriers (Yan et al. 2003). The timescale of change in reproductive proteins relative to other factors (e.g. ecological specialization, postzygotic isolation) promoting divergence determines whether reproductive genes may be “speciation genes.” Here, we have successfully identified five testis protein genes that are evolving rapidly in Peromyscus and potentially play a role in reducing fertilization success between diverging species. Among these, Acr and Gm1276 are strong candidates for intraspecific and functional analysis to identify specific selective forces driving rapid divergence of male reproductive proteins and to assess
their contributions to reproductive isolation. Moreover, *Peromyscus* subspecies and species pairs with evidence for reduced fertility can be crossed in the laboratory, providing the opportunity to investigate the effects of allelic variation in these proteins *in vivo*.

**ACKNOWLEDGEMENTS**

We thank members of the Hoekstra Laboratory, J. Kohn, and V. Vacquier for helpful discussion, H. Fisher and two anonymous reviewers for useful comments on the manuscript, and J. Weston Glenn and M. Dewey for early access to the *P. maniculatus* EST sequences. The *Peromyscus* Genetic Stock Center (PGSC) kindly provided tissue samples. This research was supported by a Howard Hughes Medical Institute Predoctoral Fellowship and National Science Foundation doctoral dissertation improvement grant DEB 0608030 (to LMT), National Science Foundation DEB 0344710 (to HEH), National Institutes of Health P40-RR14279/ RO1-M069601, and U.S. Department of Energy DE-FC09-07SR22506 (to PGSC).
LITERATURE CITED


BENDTSEN, J. D., H. NIELSEN, G. VON HEIJNE and S. BRUNAK, 2004 Improved prediction of

BRADLEY, R. D., N. D. DURISH, D. S. ROGERS, J. R. MILLER, M. D. ENGSTROM et al., 2007
Toward a molecular phylogeny for peromyscus: Evidence from mitochondrial

BRASWELL, W. E., J. A. ANDRES, L. S. MAROJA, R. G. HARRISON, D. J. HOWARD et al., 2006
Identification and comparative analysis of accessory gland proteins in Orthoptera.
Genome 49: 1069-1080.

CASTILLO-DAVIS, C. I., D. L. HARTL and G. ACHAZ, 2004 Cis-regulatory and protein evolution

CHENG, G. Y., J. L. SHI, M. WANG, Y. Q. HU, C. M. LIU et al., 2007 Inhibition of mouse
acrosome reaction and sperm-zona pellucida binding by anti-human sperm membrane

CHO, R. J., and M. J. CAMPBELL, 2000 Transcription, genomes, function. Trends Genet. 16: 409-
415.

263.

CLARK, N. L., and W. J. SWANSON, 2005 Pervasive adaptive evolution in primate seminal


DICE, L. R., 1933 Fertility relationships between some of the species and subspecies of mice in the genus Peromyscus. J. Mammal. 14: 298-305.


ISHIBASHI, K., M. SUZUKI, S. SASAKI and M. IMAI, 2001 Identification of a new multigene four-transmembrane family (MS4A) related to CD20, HTm4 and beta subunit of the high-affinity IgE receptor. Gene 264: 87-93.


Liu, T. T., 1953 The measurement of fertility and its use as an index of reproductive isolation among certain laboratory stocks of Peromyscus. Contributions from the Laboratory of Vertebrate Biology 59: 1 - 12.


SWANSON, W. J., A. G. CLARK, H. M. WALDRIP-DAIL, M. F. WOLFNER and C. F. AQUADRO,
2001b Evolutionary EST analysis identifies rapidly evolving male reproductive proteins

SWOFFORD, D. L., 2002 *PAUP* (* Phylogenetic Analysis Using Parsimony (*and Other

TANAKA, H., N. IGUCHI, A. ISOTANI, K. KITAMURA, Y. TOYAMA *et al.*, 2005 HANP1/H1T2, a
novel histone H1-like protein involved in nuclear formation and sperm fertility. Mol.

into sperm-zona pellucida interaction: involvement of sperm lipid rafts. Front. Biosci. **12**:
1748-1766.

THOMAS, P. D., A. KEJARIWAL, M. J. CAMPBELL, H. Y. MI, K. DIEMER *et al.*, 2003 PANTHER:
a browsable database of gene products organized by biological function, using curated

TORGERSON, D. G., R. J. KULATHINAL and R. S. SINGH, 2002 Mammalian sperm proteins are

TURNER, L. M., and H. E. HOEKSTRA, 2006 Adaptive evolution of fertilization proteins within a
genus: Variation in ZP2 and ZP3 in deer mice (*Peromyscus*). Mol. Biol. Evol. **23**: 1656-
1669.

TURNER, L. M., and H. E. HOEKSTRA, in press Reproductive protein evolution within and between


Table 1. Evolutionary rates of testis-expressed genes.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>N</th>
<th>Kc</th>
<th>dN</th>
<th>dS</th>
<th>( \overline{\omega} )</th>
<th>( \overline{\omega} )</th>
<th>IC</th>
<th>EC</th>
<th>H</th>
<th>NTS</th>
<th>TS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peromyscus vs. Mus</td>
<td>2,364</td>
<td>1,014</td>
<td>157</td>
<td>0.06</td>
<td>0.35</td>
<td>(0.17-0.20)</td>
<td>(7.5%)(1.3%)</td>
<td>0.19</td>
<td>0.17</td>
<td>(0.17-0.19)</td>
<td>(0.12-0.16)</td>
</tr>
<tr>
<td>Peromyscus vs. Rattus</td>
<td>2,364</td>
<td>993</td>
<td>157</td>
<td>0.07</td>
<td>0.36</td>
<td>(0.18-0.21)</td>
<td>(8.2%)(1.4%)</td>
<td>0.19</td>
<td>0.17</td>
<td>(0.17-0.18)</td>
<td>(0.11-0.15)</td>
</tr>
<tr>
<td>Comparison</td>
<td>N</td>
<td>es</td>
<td>( \bar{L}_C )</td>
<td>( \bar{d}_N )</td>
<td>( \bar{d}_S )</td>
<td>( \bar{\omega} )</td>
<td>( \bar{\omega} )</td>
<td>IC</td>
<td>EC</td>
<td>H</td>
<td>NTS</td>
</tr>
<tr>
<td>----------------</td>
<td>----</td>
<td>-----</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----</td>
<td>----</td>
<td>---</td>
<td>-----</td>
</tr>
<tr>
<td><strong>Mus vs. Rattus</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.27</td>
<td>0.33*</td>
<td>0.23</td>
<td>0.23</td>
<td>0.33***</td>
<td></td>
</tr>
<tr>
<td>Mus</td>
<td>11,203</td>
<td>4,171</td>
<td>147</td>
<td>0.07</td>
<td>0.25</td>
<td>0.28</td>
<td>509</td>
<td>99 (3)</td>
<td>0.25-0.31</td>
<td>0.29</td>
<td>(0.21-0.25)</td>
</tr>
<tr>
<td>Rattus</td>
<td>7,448</td>
<td>4,207</td>
<td>200</td>
<td>0.06</td>
<td>0.24</td>
<td>0.23</td>
<td>342</td>
<td>60 (0)</td>
<td>0.22-0.23</td>
<td>0.27</td>
<td>(0.18-0.20)</td>
</tr>
</tbody>
</table>

\( N \) = number of unique ESTs compared to Refseqs. \( \bar{d}_N \) = mean nonsynonymous substitution rate; \( \bar{d}_S \) = mean nonsynonymous substitution rate; \( \bar{\omega} \) = mean \( \omega \) for all ESTs. Number and percentage of rapidly evolving ESTs are given for two \( \omega \) classes, \( 0.5 < \omega < 1.0 \).
1.0, and $\omega > 1.0$. For the $\omega > 1.0$ class, the number of ESTs with $\omega$ significantly greater than 1.0 are given in parentheses (as determined by a likelihood ratio test comparing estimated $\omega$ models vs. models with $\omega$ fixed at 1.0, implemented in PAML). Mean $\omega$ values and 95% confidence intervals (based on 1000 bootstrap replicates, in parentheses) are given for: IC = intracellular proteins; EC = extracellular protein genes (with a predicted signal sequence and/or transmembrane domain); H = housekeeping genes; NTS = non-testis-specific genes; T = testis-specific genes. The number of ESTs in each category is given in parentheses. We compared $\bar{\omega}$ for EC vs. IC genes using permutation tests (10,000 permutations, $^* = P < 0.05$, $^{**} = P < 0.01$) and H, NTS, TS genes using an ANCOVA determining effects of expression class on $\omega$, controlling for degree of tissue specificity of expression ($^{***} = P < 0.0001$).
**Table 2.** Rapidly evolving testis-expressed genes.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Symbol</th>
<th>Chromosome (Mus)</th>
<th>Gene Ontology</th>
<th>Function</th>
<th>(\omega_{PM})</th>
<th>(\omega_{PR})</th>
<th>(\omega_{MR})</th>
<th>(\omega_{RM})</th>
</tr>
</thead>
<tbody>
<tr>
<td>hypothetical protein</td>
<td>4921517D22Rik(^2)</td>
<td>13</td>
<td>unknown</td>
<td>unknown</td>
<td>2.28</td>
<td>1.20</td>
<td>0.73</td>
<td>0.63</td>
</tr>
<tr>
<td>LOC70900</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hypothetical protein</td>
<td>4932411G14Rik(^2)</td>
<td>13</td>
<td>unknown</td>
<td>unknown</td>
<td>1.24</td>
<td>1.27</td>
<td>nd</td>
<td>0.72</td>
</tr>
<tr>
<td>LOC238663</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hypothetical protein</td>
<td>4932412H11Rik(^2)</td>
<td>5</td>
<td>protein binding</td>
<td>unknown</td>
<td>1.21</td>
<td>1.37</td>
<td>1.44</td>
<td>nd</td>
</tr>
<tr>
<td>LOC242838</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>leucine rich region</td>
<td>Lrrc5(^{1,2})</td>
<td>8</td>
<td>protein binding</td>
<td>unknown</td>
<td>1.17</td>
<td>1.32</td>
<td>1.16</td>
<td>0.11</td>
</tr>
<tr>
<td>containing 50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>histone H1-like protein in</td>
<td>Hils(^{1,2})</td>
<td>11</td>
<td>DNA binding; histone binding</td>
<td>DNA condensation during spermiogenesis (Li et al. 2006)</td>
<td>1.13</td>
<td>0.89</td>
<td>1.24</td>
<td>nd</td>
</tr>
<tr>
<td>spermatids 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hypothetical protein</td>
<td>4931408C20Rik(^2)</td>
<td>1</td>
<td>unknown</td>
<td>unknown</td>
<td>0.91</td>
<td>1.01</td>
<td>nd</td>
<td>0.87</td>
</tr>
<tr>
<td>LOC210940</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene model 1276</td>
<td>Gm1276(^{1,2})</td>
<td>19</td>
<td>receptor activity; signal transduction</td>
<td>unknown</td>
<td>0.88</td>
<td>0.82</td>
<td>nd</td>
<td>0.82</td>
</tr>
<tr>
<td>PHD finger protein 8</td>
<td>Phf8(^1)</td>
<td>X</td>
<td>DNA binding; metal ion binding; protein binding; zinc ion binding</td>
<td>unknown</td>
<td>0.81</td>
<td>0.72</td>
<td>0.44</td>
<td>nd</td>
</tr>
<tr>
<td>Gene</td>
<td>Symbol/Chromosome</td>
<td>Chromosome (Mus)</td>
<td>Gene Ontology</td>
<td>Function</td>
<td>( \omega_{PM} )</td>
<td>( \omega_{PR} )</td>
<td>( \omega_{MR} )</td>
<td>( \omega_{RM} )</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>-------------------------</td>
<td>------------------</td>
<td>------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>chemokine-like factor isoform 1</td>
<td>( Cklf^2 )</td>
<td>8</td>
<td>cytokine activity; chemotaxis</td>
<td>unknown</td>
<td>0.81</td>
<td>0.54</td>
<td>0.55</td>
<td>nd</td>
</tr>
<tr>
<td>preproacrosin</td>
<td>( Acr^1 )</td>
<td>15</td>
<td>acrosin activity; amidase activity; fucose binding; hydrolase activity; mannose binding; peptidase activity; protein binding; serine-type endopeptidase activity</td>
<td>secondary binding to zona pellicuda (ZP2), dispersal of acrosomal contents</td>
<td>0.80</td>
<td>0.40</td>
<td>0.23</td>
<td>0.17</td>
</tr>
<tr>
<td>hypothetical protein LOC71831 isoform 3</td>
<td>( 1700007B14Rik^2 )</td>
<td>8</td>
<td>unknown</td>
<td>unknown</td>
<td>0.80</td>
<td>0.70</td>
<td>nd</td>
<td>0.63</td>
</tr>
<tr>
<td>coiled-coil-helix-coiled-coil-helix domain containing 6</td>
<td>( Chchd6^2 )</td>
<td>6</td>
<td>unknown</td>
<td>unknown</td>
<td>0.77</td>
<td>0.60</td>
<td>0.69</td>
<td>nd</td>
</tr>
<tr>
<td>hypothetical protein LOC75275</td>
<td>( 4930563P21Rik^2 )</td>
<td>2</td>
<td>unknown</td>
<td>unknown</td>
<td>0.77</td>
<td>0.67</td>
<td>0.51</td>
<td>nd</td>
</tr>
<tr>
<td>Gene</td>
<td>Symbol</td>
<td>Chromosome (Mus)</td>
<td>Gene Ontology</td>
<td>Function</td>
<td>ωPM</td>
<td>ωPR</td>
<td>ωMR</td>
<td>ωRM</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-------------------------</td>
<td>------------------</td>
<td>-------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>hypothetical protein</td>
<td>LOC78174</td>
<td>5</td>
<td>cytochrome-c oxidase activity; electron transport; mitochondrial respiratory chain</td>
<td>unknown</td>
<td>0.75</td>
<td>0.53</td>
<td>0.53</td>
<td>nd</td>
</tr>
<tr>
<td>sperm associated antigen 8</td>
<td>Spag81,2</td>
<td>4</td>
<td>unknown</td>
<td>unknown</td>
<td>0.72</td>
<td>0.71</td>
<td>nd</td>
<td>0.63</td>
</tr>
<tr>
<td>hypothetical protein LOC381816</td>
<td><em>Afaf</em>2</td>
<td>4</td>
<td>sugar binding</td>
<td>acrosome formation during spermiogenesis (JEONG et al. 2004)</td>
<td>0.71</td>
<td>0.90</td>
<td>0.52</td>
<td>nd</td>
</tr>
<tr>
<td>CKLF-like MARVEL transmembrane domain containing 2A</td>
<td>Cmtm2a2</td>
<td>8</td>
<td>cytokine activity; protein binding; transcription</td>
<td>androgen receptor co-repressor involved in regulation of transcription (Liu et al. 2007)</td>
<td>0.68</td>
<td>0.61</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>sperm autoantigenic protein 17</td>
<td>Spa171,2</td>
<td>9</td>
<td>cAMP-dependent protein kinase regulator activity</td>
<td>zona pellucida binding</td>
<td>0.67</td>
<td>0.66</td>
<td>0.52</td>
<td>nd</td>
</tr>
<tr>
<td>Gene</td>
<td>Symbol</td>
<td>Chromosome (Mus)</td>
<td>Gene Ontology</td>
<td>Function</td>
<td>ωPM</td>
<td>ωPR</td>
<td>ωMR</td>
<td>ωRM</td>
</tr>
<tr>
<td>------</td>
<td>--------</td>
<td>----------------</td>
<td>---------------</td>
<td>----------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>similar to Protein C14orf32 homolog</td>
<td>C130032J12Rik&lt;sup&gt;2&lt;/sup&gt;</td>
<td>14</td>
<td>unknown</td>
<td>unknown</td>
<td>0.67</td>
<td>0.52</td>
<td>nd</td>
<td>0.95</td>
</tr>
<tr>
<td>hypothetical protein</td>
<td>4930596D02Rik&lt;sup&gt;2&lt;/sup&gt;</td>
<td>14</td>
<td>calcium ion binding; guanyl-nucleotide exchange factor activity; regulation of small GTPase mediated signal transduction; small GTPase mediated signal transduction</td>
<td>unknown</td>
<td>0.64</td>
<td>0.63</td>
<td>0.51</td>
<td>0.75</td>
</tr>
<tr>
<td>hypothetical protein</td>
<td>4932415G16Rik&lt;sup&gt;2&lt;/sup&gt;</td>
<td>12</td>
<td>unknown</td>
<td>part of CatSper&lt;sup&gt;β&lt;/sup&gt; ion channel protein complex, required for sperm hyperactivation (GUTTMAN et al. 2004)</td>
<td>0.63</td>
<td>0.81</td>
<td>0.57</td>
<td>nd</td>
</tr>
<tr>
<td>gene model 884</td>
<td>Gm884&lt;sup&gt;2&lt;/sup&gt;</td>
<td>11</td>
<td>unknown</td>
<td>unknown</td>
<td>0.62</td>
<td>0.56</td>
<td>0.64</td>
<td>nd</td>
</tr>
<tr>
<td>hypothetical protein</td>
<td>1700047L15Rik&lt;sup&gt;2&lt;/sup&gt;</td>
<td>12</td>
<td>unknown</td>
<td>unknown</td>
<td>0.59</td>
<td>0.63</td>
<td>0.59</td>
<td>nd</td>
</tr>
<tr>
<td>LOC73309</td>
<td>1700011L22Rik&lt;sup&gt;2&lt;/sup&gt;</td>
<td>8</td>
<td>unknown</td>
<td>unknown</td>
<td>0.59</td>
<td>0.59</td>
<td>0.75</td>
<td>0.82</td>
</tr>
</tbody>
</table>

L.M. Turner, E.B. Chuong, and H.E. Hoekstra
<table>
<thead>
<tr>
<th>Gene</th>
<th>Symbol</th>
<th>Chromosome (Mus)</th>
<th>Gene Ontology</th>
<th>Function</th>
<th>$\omega_{PM}$</th>
<th>$\omega_{PR}$</th>
<th>$\omega_{MR}$</th>
<th>$\omega_{RM}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>lysosomal-associated membrane protein 1</td>
<td>$Lamp1^2$</td>
<td>8</td>
<td>unknown</td>
<td>release of spermatozoa from epithelium during spermatogenesis (MARTIANOV et al. 2005)</td>
<td>0.58</td>
<td>0.53</td>
<td>nd</td>
<td>0.54</td>
</tr>
<tr>
<td>testis specific protein Ddc8</td>
<td>$Ddc8^{1,2}$</td>
<td>11</td>
<td>unknown</td>
<td>unknown</td>
<td>0.56</td>
<td>0.68</td>
<td>0.54</td>
<td>0.34</td>
</tr>
<tr>
<td>hypothetical protein LOC70980</td>
<td>$4931431F19Rik^2$</td>
<td>7</td>
<td>unknown</td>
<td>unknown</td>
<td>0.55</td>
<td>0.54</td>
<td>0.61</td>
<td>0.56</td>
</tr>
<tr>
<td>spermatogenesis associated 3</td>
<td>$Spata3^2$</td>
<td>1</td>
<td>apoptosis; spermatogenesis</td>
<td>unknown</td>
<td>0.54</td>
<td>0.72</td>
<td>1.16</td>
<td>nd</td>
</tr>
<tr>
<td>similar to kinesin-like motor protein C20orf23</td>
<td>$C20orf23^2$</td>
<td>2</td>
<td>unknown</td>
<td>unknown</td>
<td>0.54</td>
<td>0.70</td>
<td>1.82</td>
<td>nd</td>
</tr>
<tr>
<td>germ cell-specific gene 1</td>
<td>$Gsg1^{1,2}$</td>
<td>6</td>
<td>unknown</td>
<td>unknown</td>
<td>0.44</td>
<td>0.30</td>
<td>nd</td>
<td>0.36</td>
</tr>
<tr>
<td>Gene</td>
<td>Symbol</td>
<td>me (Mus)</td>
<td>Gene Ontology</td>
<td>Function</td>
<td>ω&lt;sub&gt;PM&lt;/sub&gt;</td>
<td>ω&lt;sub&gt;PR&lt;/sub&gt;</td>
<td>ω&lt;sub&gt;MR&lt;/sub&gt;</td>
<td>ω&lt;sub&gt;RM&lt;/sub&gt;</td>
</tr>
<tr>
<td>------------------------------</td>
<td>----------</td>
<td>----------</td>
<td>---------------</td>
<td>--------------------------------------------------------------------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>histone H1 variant</td>
<td>H1fnt&lt;sup&gt;1,3&lt;/sup&gt;</td>
<td>15</td>
<td>DNA binding</td>
<td>DNA condensation during spermiogenesis, essential for proper nuclear morphology (TANAKA et al. 2005)</td>
<td>0.18</td>
<td>0.14</td>
<td>0.74</td>
<td>0.52</td>
</tr>
<tr>
<td>sperm mitochondria-associated cysteine-rich protein</td>
<td>Smcp&lt;sup&gt;1,3&lt;/sup&gt;</td>
<td>3</td>
<td>selenium binding</td>
<td>sperm motility(v. 3.14, YANG 2000)</td>
<td>0.14</td>
<td>nd</td>
<td>nd</td>
<td>nd</td>
</tr>
</tbody>
</table>

Gene names, symbols, and gene ontology (GO) terms are indicated for *Mus* homologues for the most rapidly evolving proteins. All ω values were estimated in paml (YANG 2000). *P* = *P. maniculatus*, *M* = *M. musculus*, *R* = *R. norvegicus*. ω indicates pairwise ω between EST and RefSeq for the species pair, e.g. ω<sub>PM</sub> is between *P. maniculatus* EST and *M. musculus* RefSeq.

<sup>1</sup>candidate genes sequenced in additional *Peromyscus* species. <sup>2</sup>ω > 0.5 in all comparisons. <sup>3</sup>included because initial screen showed high ω values.
Table 3. Adaptive evolution of testis-expressed genes in *Peromyscus.*

<table>
<thead>
<tr>
<th>Gene</th>
<th>$L_C$</th>
<th>$\omega_{PM}$ vs. $\omega$</th>
<th>$\omega$</th>
<th>Variable aa sites (%)</th>
<th>M8 vs M8A</th>
<th>$\omega_s$</th>
<th>$p_s$</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Lrrc50</em></td>
<td>622</td>
<td>1.17</td>
<td>0.30</td>
<td>0.41</td>
<td>65 (10.4)</td>
<td>0.039*</td>
<td>3.90</td>
</tr>
<tr>
<td><em>Hils1</em></td>
<td>162</td>
<td>1.13</td>
<td>1.16</td>
<td>0.42</td>
<td>15 (9.3)</td>
<td>0.416</td>
<td>n/a</td>
</tr>
<tr>
<td><em>Gm1276</em></td>
<td>830</td>
<td>0.88</td>
<td>0.94</td>
<td>0.69</td>
<td>74 (8.9)</td>
<td>0.005**</td>
<td>10.12</td>
</tr>
<tr>
<td><em>Phf8</em></td>
<td>447</td>
<td>0.81</td>
<td>0.27</td>
<td>0.47</td>
<td>118 (26.4)</td>
<td>0.127</td>
<td>n/a</td>
</tr>
<tr>
<td><em>Acr</em></td>
<td>428</td>
<td>0.80</td>
<td>0.42</td>
<td>0.22</td>
<td>18 (4.2)</td>
<td>0.012*</td>
<td>7.55</td>
</tr>
<tr>
<td><em>Spag8</em></td>
<td>263</td>
<td>0.72</td>
<td>0.64</td>
<td>0.57</td>
<td>54 (20.5)</td>
<td>0.500</td>
<td>n/a</td>
</tr>
<tr>
<td><em>Spa17</em></td>
<td>147</td>
<td>0.67</td>
<td>0.59</td>
<td>0.16</td>
<td>4 (2.7)</td>
<td>0.283</td>
<td>n/a</td>
</tr>
<tr>
<td><em>Ddc8</em></td>
<td>539</td>
<td>0.56</td>
<td>0.47</td>
<td>0.55</td>
<td>50 (9.3)</td>
<td>0.042*</td>
<td>2.19</td>
</tr>
<tr>
<td><em>Gsg1</em></td>
<td>364</td>
<td>0.44</td>
<td>0.48</td>
<td>0.34</td>
<td>29 (8.0)</td>
<td>0.022*</td>
<td>2.18</td>
</tr>
<tr>
<td><em>Hlfnt</em></td>
<td>304</td>
<td>0.18</td>
<td>0.56</td>
<td>0.28</td>
<td>21 (6.9)</td>
<td>0.225</td>
<td>n/a</td>
</tr>
<tr>
<td><em>Smcp</em></td>
<td>136</td>
<td>0.14</td>
<td>0.13</td>
<td>0.11</td>
<td>7 (5.1)</td>
<td>0.500</td>
<td>n/a</td>
</tr>
</tbody>
</table>

$L_C =$ length of sequence analyzed in codons; $\omega_{PM} = \omega$ of the *Peromyscus* EST vs. Mus homologue; $\omega$ vs. Mus = pairwise $\omega$ for the entire *P. maniculatus* sequence vs. the Mus homologue; $\omega$ *Peromyscus* = $\omega$ in *Peromyscus* sample, averaged across all sites and lineages [estimated with PAML, M0 (Yang 2000)]; M8 vs M8A = $P$ value of likelihood ratio test; $\omega_s =$ $\omega$ estimate for ‘$\omega > 1$’ class; $p_s =$ proportion of sites in the ‘$\omega > 1$’ class for M8. *$P < 0.05$, **$P < 0.01$.}
Figure legends

Fig. 1. Evolutionary rates of testis-expressed ESTs. (A) $d_N$ vs. $d_S$ estimated in PAML (Steppan et al. 2004), each point represents the respective substitution rate for a given *Peromyscus maniculatus* testis EST vs. its *Mus musculus* homolog. (B) Proportion of ESTs in each expression class among all ESTs and among ESTs grouped by $\omega$ value; H = housekeeping, NTS = non-testis-specific, TS = testis-specific; ** = $P < 0.01$ and *** = $P < 0.001$ in a two-tailed binomial test for under- or overrepresentation of ESTs of an expression type in the given $\omega$ class.

Fig. 2. Rapidly evolving testis genes in three rodent lineages. (A) Relationships and divergence times between *Peromyscus*, *Mus*, and *Rattus*. (B) Rapidly evolving ($\omega > 0.5$) testis genes with homologues identified in more than one species comparison are shown, including genes identified as rapidly evolving in one or multiple comparisons. Total number of EST pairs (N) and total number of rapidly evolving genes (RE) are given for each species comparison.

Fig. 3. Positively selected testis genes in *Peromyscus*. Alignments of variable amino acid sites are shown for each gene. Dots indicate identity with the consensus sequence. Amino acid sites identified as positively selected with BEB analysis are indicated with arrows and are in bold. Relationships between species are shown to the left of each alignment (based on maximum likelihood analysis of 1,213 bp of the mitochondrial genome and 1,201 bp of the nuclear genes *Mc1r* and *Lcat*). Exon/intron structure of each gene is
shown above the alignment. Boxes indicate exons, and are drawn to scale within each gene; open boxes indicate noncoding exons and filled boxes coding exons. Asterisks above indicate the positions of the positively selected sites.
(A) $N = 1,014$  
$\omega = 0.19$  
$\omega > 1$  
$N = 13$  
$0.5 < \omega < 1$  
$N = 89$  
$\omega < 0.5$  
$N = 850$

(B) H NTS TS

proportion of ESTs

All  < 0.5 $\omega$ class  > 0.5

Figure 1
Figure 2.
Figure 3
Figure 1

(A) Scatter plot showing the relationship between $d_w$ and $d_S$.

1. For $\omega > 1$, $N = 13$.
2. For $0.5 < \omega < 1$, $N = 89$.
3. For $\omega < 0.5$, $N = 850$.

(B) Bar chart showing the proportion of ESTs in different $\omega$ classes.

- All: $N = 1,014$, $\omega = 0.19$.
Figure 2.
Figure 3