UNIVERSITY OF CALIFORNIA, SAN DIEGO

EVOLUTION OF REPRODUCTIVE PROTEINS IN DEER MICE (*PEROMYSCUS*)

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Biology

by

Leslie McCue Turner

Committee in charge:

Professor Joshua R. Kohn, Chair
Professor Hopi E. Hoekstra, Co-Chair
Professor Peter Andolfatto
Professor Ronald S. Burton
Professor Victor D. Vacquier

2007

The dissertation of Leslie McCue Turner is approved, and it is acceptable in quality and form for publication on microfilm:

____

____

____

____
Co-Chair

____
Chair

University of California, San Diego

2007

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my advisor, Hopi Hoekstra, for being an excellent mentor and role model. I particularly appreciate her focus on professional development; she has invested not just in my dissertation project but also helped me in every way to launch a successful career as a scientist. My thesis committee, current members as well as former member John Huelsenbeck, provided valuable advice and insight, both at committee meetings and individually. The Hoekstra laboratory was always a fun and supportive as well as a scientifically stimulating research environment. All members of the lab were a joy to work with. I thank Lynne Mullen for always being available to give a second opinion on everything from PCR programs to how to word a sentence in an e-mail (or to go on a chocolate mission for the pig-mouse!). Edward Chuong, our "wondergrad", greatly improved my work by writing the bioinformatics scripts. Rachel Hirschmann, Jesse Weber, Cynthia Steiner, Anna Bree, Adrian Young, and Evan Kingsley provided helpful feedback on ideas and presentations, help in the field, and great company.

Before I came to UCSD, I was trained at Barnard College and Duke University. My advisors Janet Larkin and Marina Cords at Barnard and Theresa Pope at Duke provided valuable guidance and encouragement.

Finally, I would like to thank my amazing family and friends for supporting me through all these years of training!

Chapter 1, in full, is a reprint of the material as it appears in *Molecular Biology and Evolution* 2006: Turner LM, and Hoekstra HE. 2006. Adaptive evolution of

fertilization proteins within a genus: Variation in ZP2 and ZP3 in deer mice (*Peromyscus*). Mol. Biol. Evol. **23**:1656-1669.

Chapter 2 is in preparation for publication: Turner LM, and Hoekstra HE. In Preparation. Reproductive protein evolution within and between species: ZP3 sequence variation in *Peromyscus truei* and *P. gratus*.

Chapter 3 is in preparation for publication: Turner LM, Chuong EB, and Hoekstra HE. In Preparation. Comparative analysis of testis protein evolution in rodents.

CURRICULUM VITA

Education:

2007 Ph.D. in Biology
University of California, San Diego
Advisor: Dr. Hopi Hoekstra

2000 M.A. in Biological Anthropology and Anatomy
Duke University
Advisor: Dr. Theresa Pope

1997 A.B. in Biology
Barnard College, Columbia University
Advisor: Dr. Janet Larkin

Publications:

Woodruff, D.S. and **L.M. Turner**. In Review. The Indochinese-Sundaic zoogeographic transition: a description and analysis of terrestrial mammal species distributions. *Journal of Biogeography.*

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*). *Molecular Biology and Evolution*. 23(9): 1656-1669.

Howard, L.L., **L.M. Turner**, I.H. Stalis, and P.J. Morris. 2005. Serum gamma-glutamyltransferase as a prognostic indicator of neonatal viability in nondomestic ruminants. *Journal of Zoo and Wildlife Medicine* 36(2): 239 – 244.

Presentations:

2007 "Comparative analysis of testis protein evolution in rodents" Invited talk: Aquavit Meeting, Max Planck Institute, Plön, Germany (March)

2007 "Adaptive evolution of reproductive proteins in deer mice (*Peromyscus*)" Invited seminar: Institute for Genetics, University of Cologne, Germany (March)

2006 "Evolutionary patterns of reproductive proteins in deer mice (*Peromyscus*)" Talk: Meeting of Predoctoral Fellows in Biological Sciences, Howard Hughes Medical Institute, Bethesda, Maryland (September)

2006 "Evolution of testis-expressed proteins in *Peromyscus*" Poster: Genetics of Speciation Symposium, *American Genetics Association*, Vancouver, Canada (July)

2006 "Evolution of testis-expressed proteins in monogamous and promiscuous species of *Peromyscus*" Talk: *Society of Molecular Biology and Evolution* Meeting, Tempe, Arizona (May)

2006 "Comparative analysis of *V1ar* regulation and its contribution to mating system variation in deer mice (*Peromyscus*)" Poster: Society of Molecular Biology and Evolution Meeting, Tempe, Arizona (May)

2005 “Adaptive evolution of a fertilization protein in natural populations of deermice (*Peromyscus*)” Talk: Division of Biological Sciences Retreat, University of California, San Diego (September)

2005 “Adaptive evolution of a fertilization protein in natural populations of deermice (*Peromyscus*)” Poster: Gordon Research Conference: Fertilization and the Activation of Development, Plymouth, New Hampshire (July)

2005 “Reproductive protein evolution within a species: ZP3 sequence variation in *Peromyscus truei*” Talk: Evolution meeting, Fairbanks, Alaska (June)

2005 “Evolution of fertilization proteins in *Peromyscus*” Talk: Division of Biological Sciences Recruitment, University of California, San Diego (February and March)

2004 “Evolution of reproductive proteins in monogamous and promiscuous species of *Peromyscus*” Talk (Ernst Mayr Prize session): Evolution meeting, Fort Collins, Colorado (June)

ABSTRACT OF THE DISSERTATION

EVOLUTION OF REPRODUCTIVE PROTEINS IN DEER MICE (*PEROMYSCUS*)

by

Leslie McCue Turner

Doctor of Philosophy in Biology

University of California, San Diego, 2007

Professor Joshua R. Kohn, Chair

Professor Hopi E. Hoekstra, Co-Chair

This dissertation describes evolutionary patterns of female and male reproductive proteins and their potential contribution to speciation in deer mice (*Peromyscus*). Proteins involved in reproduction are among the most rapidly evolving

genes in many taxa. This striking pattern is of particular interest because reproductive proteins mediate species-specific fertilization, and thus changes in these proteins have the potential to contribute to reproductive isolation. In internally fertilized taxa, knowledge of the evolutionary dynamics of reproductive proteins in closely related species is limited primarily to seminal proteins expressed in accessory glands of *Drosophila*. Investigation of additional taxa and functional classes of proteins is necessary to determine if there is a general pattern of adaptive evolution of reproductive proteins between recently diverged species. In mammals, positive selection has been documented in male and female reproductive proteins among divergent species. The research presented here extends this work by investigating reproductive protein evolution within a mammalian genus.

Chapter 1 reports evidence that two egg-coat proteins, ZP2 and ZP3, have evolved under positive selection during diversification of the genus *Peromyscus* and identifies specific amino acid sites within these proteins that have been targets of selection.

Chapter 2 describes patterns of sequence variation of ZP3 within two sister species of *Peromyscus*. High levels of amino acid polymorphism in both species suggest that balancing selection might promote sequence divergence in ZP3 in these species.

Chapter 3 is a comparative analysis of testis protein evolution in three lineages of Muroid rodents: *Peromyscus* and the genetic model organisms *Mus* and *Rattus*. In each lineage, testis-expressed proteins evolve more rapidly, on average, than genes

with highest expression in another tissue. Genes with the highest rates of evolution serve a variety of functions. Five of eleven of these genes sequenced in six *Peromyscus* species show evidence for positive selection.

Together, these findings demonstrate that reproductive proteins evolve adaptively between closely related mammalian species, where reproductive isolation has evolved recently. Further, I identify positively selected egg and testis genes and specific amino acid sites that are targets of selection and are promising targets for future functional assays of allelic differences in fertilization potential.

Chapter 1: Adaptive evolution of fertilization proteins within a genus: Variation in ZP2 and ZP3 in deer mice (*Peromyscus*)

Abstract

Rapid evolution of reproductive proteins has been documented in a wide variety of taxa. In internally fertilized species, knowledge about the evolutionary dynamics of these proteins between closely related taxa is primarily limited to accessory gland proteins in the semen of *Drosophila.* Investigation of additional taxa and functional classes of proteins is necessary in order to determine if there is a general pattern of adaptive evolution of reproductive proteins between recently diverged species. We performed an evolutionary analysis of two egg coat proteins, ZP2 and ZP3, in fifteen species of deer mice (genus *Peromyscus*). Both of these proteins are involved in egg-sperm binding, a critical step in maintaining species-specific fertilization. Here, we show that *Zp2* and *Zp3* gene trees are not consistent with trees based on non-reproductive genes, *Mc1r* and *Lcat*, where species formed monophyletic clades. In fact, for both of the reproductive genes, intraspecific amino acid variation was extensive and alleles were sometimes shared across species. We document positive selection acting on ZP2 and ZP3 and identify specific amino acid sites that are likely targets of selection using both maximum likelihood approaches and patterns of parallel amino acid change. In ZP3, positively selected sites are clustered in and around the region implicated in sperm binding in *Mus*, suggesting changes may impact egg-sperm binding and fertilization potential. Finally, we identify lineages with significantly elevated rates of amino acid substitution using a

Bayesian mapping approach. These findings demonstrate that the pattern of adaptive reproductive protein evolution found at higher taxonomic levels can be documented between closely related mammalian species, where reproductive isolation has evolved recently.

Introduction

Over the past decade, a pattern of rapid evolution of proteins involved in reproduction has emerged from research in taxa ranging from diatoms to primates (Singh and Kulathinal 2000; Swanson and Vacquier 2002a; Swanson and Vacquier 2002b). Investigations of reproductive protein evolution have examined sperm-egg recognition proteins in marine invertebrates (Swanson and Vacquier 2002a; Swanson and Vacquier 2002b), accessory gland proteins in semen of *Drosophila* (Civetta and Singh 1995; Cirera and Aguade 1997; Tsaur and Wu 1997; Aguade 1999; Begun et al. 2000; Swanson et al. 2001a), proteins expressed in the female reproductive tract of *Drosophila* (Swanson et al. 2004), and male and female reproductive proteins in mammals (Wyckoff, Wang, and Wu 2000; Swanson et al. 2001b; Torgerson, Kulathinal, and Singh 2002; Jansa, Lundrigan, and Tucker 2003; Kingan, Tatar, and Rand 2003; Swanson, Nielsen, and Yang 2003; Dorus et al. 2004; Clark and Swanson 2005). Rapid divergence and positive selection have been documented in many of these reproductive proteins.

Adaptive evolution of egg and sperm interaction proteins, specifically, has been documented at several taxonomic levels in marine invertebrates (Swanson and

Vacquier 2002a; Galindo, Vacquier, and Swanson 2003; Geyer and Palumbi 2003; Mah, Swanson, and Vacquier 2005). In these broadcast spawners, maintenance of species-specific binding of gametes has been proposed as a possible explanation for their rapid divergence (Swanson and Vacquier 2002b). However, the selective pressures driving reproductive protein evolution may vary in different taxa (Swanson and Vacquier 2002b). For example, in internally fertilized species organisms have more control over which individuals exchange gametes, and the environment in which gametes interact is different (Eisenbach and Giojalas 2006). Thus, the selective pressures on reproductive proteins in internally fertilized species may differ substantially from those acting on species with external fertilization.

The seminal proteins produced by accessory glands of *Drosophila* (Acps) are the most well studied class of reproductive proteins in internally fertilized species. Rapid evolution and positive selection have been documented for many Acps, both between closely and distantly related species (Begun et al. 2000; Swanson et al. 2001a; Begun and Lindfors 2005; Mueller et al. 2005). Recent work has demonstrated that proteins expressed in the testes, ovaries and female reproductive tracts of *Drosophila* also evolve rapidly, although not as dramatically as Acps (Swanson et al. 2004; Jagadeeshan and Singh 2005). However, detailed examination of the evolution of reproductive proteins in different taxa and functional classes is necessary to determine if evolution of *Drosophila* Acps reflects a general pattern in internally fertilized species.

Research on the evolution of reproductive proteins in mammals has thus far focused primarily on identifying genes that have experienced positive selection by analyzing sequences from distantly related species. To our knowledge, there has only been one study addressing patterns of evolution of a reproductive protein within a mammalian genus (Jansa, Lundrigan, and Tucker 2003); the authors provided evidence that the egg protein ZP3 had experienced positive selection in the *Mus* genus, however, when their analysis was repeated without including outgroup sequences, there was no longer evidence for positive selection (see online supplementary information for details). Lack of significance may be due to limited sampling, therefore we decided to test extensively for positive selection on egg proteins in the evolution of a single genus.

Here, we extend previous work in mammals by documenting patterns of evolution of egg coat proteins in closely related species of deer mice (genus *Peromyscus*). As in *Drosophila*, pairs of *Peromyscus* taxa with varying degrees of reproductive isolation may be sampled, including populations, subspecies, sister species, species and species groups (Hooper 1968). In addition, sperm competition and sexual conflict have been proposed as important factors driving reproductive protein evolution (Wyckoff, Wang, and Wu 2000; Price et al. 2001; Swanson et al. 2001b; Torgerson, Kulathinal, and Singh 2002). *Peromyscus* has well documented variation in mating system (Kleiman 1977; Wolff 1989), thus we are also able to compare evolution of fertilization proteins between closely related species where the selective environment may differ.

The fertilization process, and specifically egg-sperm interactions, is better understood in mammals than in other internally fertilized species, providing a large number of candidate genes. We focused on two proteins that are directly involved in egg-sperm binding because this step of fertilization is critical to species-specific fertilization (Wassarman, Jovine, and Litscher 2001). The egg proteins ZP2 (zona pellucida glycoprotein 2, *Zp2*) and ZP3 (zona pellucida glycoprotein 3, *Zp3*) are two of the proteins that make up the zona pellucida, or egg coat, and they are both necessary for binding of the egg and sperm (Wassarman and Litscher 2001). We chose to focus initially on the egg component of this interaction because the identity and function of the sperm proteins involved are less well defined (Jansen, Ekhlasi-Hundrieser, and Toepfer-Petersen 2001).

The goal of this study was to determine patterns of evolution of ZP2 and ZP3 in *Peromyscus*. We identify differences in tree topologies and patterns of intraspecific variation between these egg coat proteins and non-reproductive proteins. We document positive selection acting on ZP2 and ZP3 and determine the spatial pattern and identity of amino acid sites under selection. Finally, we identify lineages with significantly elevated rates of amino acid substitution in ZP2 and ZP3. Together, these results suggest that positive selection is driving divergence of egg-coat proteins in closely related species, and allow us to nominate candidate amino acid sites that may contribute to reduced fertilization potential between sister taxa.

Materials and Methods

Extraction, amplification, and sequencing

To maximize genetic variation, one to three geographic locales for each of fifteen *Peromyscus* species were sampled (Table 1, see online supplementary information for details). For each locale, one to two individuals were included, for a total of 44 individuals (*Zp2*) and 48 individuals (*Zp3*). An additional two individuals of an outgroup species, *Onychomys torridus,* were sequenced for each gene. Genomic DNA was extracted from frozen or ethanol-preserved tissue samples (tail, liver, or kidney) using DNeasy tissue kits (Qiagen).

The entire genomic sequence of each reproductive gene and some 5' and 3' flanking sequence was determined in *Peromyscus polionotus*, totaling 12,755 bp for *Zp2* and 11,518 bp for *Zp3* (see online supplementary information for complete genomic sequences). Initially, 2 – 4 kb regions of each gene were amplified using primers designed in conserved regions, based on aligned exon sequences from mammalian species available in GenBank. Resulting PCR products were cloned (pGEM-T system, Promega) and sequenced using T7 and SP6 primers and internal sequencing primers. Sequences were edited and contigs assembled using SEQUENCHER (Gene Codes, Ann Harbor, MI). Once sequences were verified as the correct targets based on identity with *Mus* sequences, a genome walking approach (Universal GenomeWalker Kit, Clontech, BD Biosciences) was used to amplify and sequence 5' and 3' of cloned regions in the same *P. polionotus* individual until the entire genomic sequence was determined. The predicted amino acid sequence was aligned to *Mus* and *Rattus* sequences using CLUSTALW (Chenna et al. 2003).

After the entire *P. polionotus* nucleotide sequence for both fertilization genes was complete, *Peromyscus* specific primers were designed to amplify exons 8 – 10 of *Zp2* (2,102 bp) and exons 6 – 7 of *Zp3* (790 bp) (Fig. 1). These regions were chosen because they contain several sites identified as targets of selection in an analysis of divergent mammalian species (Swanson et al. 2001b). In addition, the region chosen for *Zp3* contains the sperm-combining region, which is necessary for ZP3's role in fertilization (Wassarman and Litscher 2001). PCR was performed under standard conditions (online supplementary material).

In order to determine whether phylogenies for the fertilization proteins are representative of species relationships, we sequenced two non-reproductive nuclear genes, the melanocortin-1 receptor (*Mc1r*) and lecithin cholesterol acyl transferase (*Lcat*). Mc1r is a G-protein coupled receptor involved in pigmentation (Barsh 1996). Lcat is an enzyme in the glycerophospholipid metabolism pathway (Kuivenhoven et al. 1997). An 869 bp fragment containing most of the single exon coding region of *Mc1r* and a 487 bp fragment containing most of exon 6 of *Lcat* were amplified under standard conditions (online supplementary material) using published primers (Robinson et al. 1997; Nachman, Hoekstra, and D'Agostino 2003).

PCR products were purified using a MinElute PCR purification kit (Qiagen) or a PerfectPrep PCR cleanup 96 kit (Eppendorf) if a single band was present. If multiple bands were present, PCR products were purified using the MinElute Gel

Extraction kit (Qiagen). Purified PCR products were directly sequenced on an ABI 3100 automated sequencer (Applied Biosystems) using both PCR amplification primers and internal sequencing primers. For *Zp2* and *Zp3*, if an individual was heterozygous at more than one site, PCR products were cloned (TOPO-TA, Invitrogen) and sequenced using T7 and T3 primers to determine phase. Base calls were confirmed by eye, and sequences were aligned in SEQUENCHER. Coding region sequences analyzed for each gene included: *Zp2* (381 bp, 127 aa), *Zp3* (228 bp, 76 aa), *Mc1r* (756 bp, 252 aa) and *Lcat* (445 bp, 148 aa). Sequences were deposited in GenBank (Accession numbers DQ482843 - 482899; DQ668051 - 668343).

Phylogenetic reconstruction

Bayesian gene trees were constructed using MRBAYES (Huelsenbeck and Ronquist 2001, GTR+(partitioned by position in codon, 10 million generation MCMC) both with and without outgroup sequences included. The first 500,000 generations were excluded as burn-in. *Mc1r* and *Lcat* were concatenated, and the data set was partitioned by both gene and position in codon; separate trees were generated using sequences from the same individuals included in the *Zp2* and *Zp3* data sets. We determined the appropriate model for each gene using hierarchical likelihood ratio tests comparing nested models (Huelsenbeck and Crandall 1997). Likelihoods of the resulting highest posterior probability tree were determined under alternative models available in MRBAYES (nst = 1, 2, 6) using PAUP*(v.4b10, Swofford 2002). Trees were rooted using outgroup (*Onychomys torridus*) sequences, if included. Neighbor-joining (NJ) and maximum likelihood (ML) trees were generated in PAUP* (GTR + Γ). In order to determine support values, ML analysis was repeated for 100 bootstrap data sets generated using the program SEQBOOT from the PHYLIP package (Felsenstein 2004). Because the gene data were partitioned by position in codon, we generated bootstrap data sets by resampling at the codon level rather than the nucleotide level.

Detection of positive selection

We tested for evidence of positive selection by comparing the nonsynonymous substitution rate (d_N) to the synonymous substitution rate (d_S). If a gene is evolving neutrally, $\omega = d_N/d_S$ is expected to equal one, whereas ω greater than one is considered strong evidence that a gene experiences positive selection. We used several maximum likelihood (ML) approaches to test for evidence of positive selection on these fertilization proteins. The first approach, developed by Nielsen and colleagues (hereafter referred to as NY models), involves comparisons of a neutral codon substitution model with ω constrained to be < 1 to a selection model where a class of sites has ω > 1 (Nielsen and Yang 1998; Yang et al. 2000). As neutral models are nested within the corresponding selection models, a likelihood ratio test (LRT) can be used to compare them. The test statistic -2 ΔlnL (ΔlnL = the difference in log likelihoods of the two models) follows a χ^2 distribution with degrees of freedom equal to the difference in number of parameters between models. In the specific models implemented, ω varies between codons as a discrete (neutral- M0, M1 selection- M3, M2) or beta distribution (neutral- M7, M8A selection- M8). We implemented models M0, M1, M2, M3, M7, and M8 (Wong et al. 2004) with the codeml program in PAML (v.3.14, Yang 2000). In order to account for uncertainty in the phylogeny, we performed the analysis using the ten most probable trees from MRBAYES as well as the NJ tree. Results of three model comparisons (M3 vs. M0, M2 vs. M1, M8 vs. M7) were consistent; here, we present data for the M8 vs. M7 comparison, as this comparison is considered a more stringent test of positive selection (Yang and Nielsen 2002). We performed an additional test comparing results from M8 to a modified version of the model where the selection class has ω set to 1 (model M8A, Swanson, Nielsen, and Yang 2003). This test rules out the possibility that the neutral model is rejected because of a poor fit of the beta distribution for neutral and negatively selected sites. The test statistic follows a 50:50 mix of a χ^2 distribution with one degree of freedom and a point mass at zero. Amino acid sites experiencing positive selection were identified using the Bayes empirical Bayes (BEB) procedure (Yang, Wong, and Nielsen 2005). The BEB procedure is a modified version of an empirical Bayes' approach (Nielsen and Yang 1998) that identifies the most likely ω class for each codon site. Those sites that are most likely to be in the positive selection class (ω > 1) are identified as likely targets of selection. The BEB procedure is an improvement over the previous approach as it takes into account sampling error in the ML estimates of parameters.

As our data include multiple alleles from each species, there is a possibility that recombination has occurred between alleles within species. In addition, if only a short time elapsed between speciation events, recombinant alleles from a polymorphic ancestor may have fixed in closely related species. Recombination can reduce the accuracy of the NY models (Anisimova, Nielsen, and Yang 2003) because different sites can have different phylogenetic histories. Specifically, differences in topology can result in patterns that look like recurrent substitution, and differences in branch lengths can result in variation in synonymous divergence among sites. We accounted for differences in topology in part by applying the NY models to multiple trees for each egg protein, however it is possible that phylogenetic histories for all sites were not sampled. In order to address the issue of differences in branch lengths between sites, we applied additional methods to test for positive selection.

While the NY models allow for variation in the nonsynonymous substitution rate, the synonymous rate is fixed across the sequence. Recently, several methods for detecting positive selection that allow for variation in synonymous rate have been proposed. These methods are new implementations of the three general classes of previous models, counting methods, fixed effects methods, and random effects methods. Counting methods map changes onto the phylogeny to estimate ω on a site-by-site basis. Kosakovsky Pond and Frost (Kosakovsky Pond and Frost 2005b) propose a version called the single-likelihood ancestor counting (SLAC) method, which calculates the number of nonsynonymous and synonymous substitutions that have occurred at each site using ML reconstructions of ancestral sequences. Kosakovky Pond and Frost additionally introduce a version of a fixed effect approach, which estimates ω on a site-by-site basis. Their fixed effect likelihood (FEL) method uses ML estimation and treats shared parameters (branch lengths, tree topology and nucleotide substitution rates) as fixed. The random effects likelihood method (REL) is similar to the NY model M3, however both nonsynonymous and synonymous rates vary as gamma distributions with three rate classes (Kosakovsky Pond and Frost 2005b; Kosakovsky Pond and Muse 2005). The SLAC and FEL methods were implemented using the web interface DATAMONKEY (Kosakovsky Pond and Frost 2005a) and the REL method was implemented in HYPHY (Kosakovsky Pond, Frost, and Muse 2005).

Mapping of amino acid substitutions

Nucleotide substitutions in both the reproductive genes *Zp2* and *Zp3* and non-reproductive genes *Mc1r* and *Lcat* were mapped onto the *Mc1r/Lcat* ML trees using maximum parsimony. Combined *Mc1r/Lcat* trees were used because of potential inaccuracies in the topology of gene trees for *Zp2* and *Zp3* due to parallel amino acid substitutions (see Results). In addition, substitutions in *Zp2* and *Zp3* were mapped using a Bayesian method (Nielsen 2002) with the program SIMMAP (Bollback 2006). Because, by definition, the parsimony method assumes that evolution has occurred in the fewest possible number of mutational steps, this approach provides a biased estimate. The degree to which parsimony underestimates the number of mutations depends on branch length and mutational parameters. The Bayesian method provides an advantage over parsimony because it accounts for uncertainty in the topology and model parameters by simulating mappings based on their probability of occurrence (Nielsen 2002). The Bayesian mappings were performed for *Zp2* and *Zp3* data sets that were modified such that, for sites that were variable within a species, only the derived state was included; this modification ensured that substitutions that were not fixed were not counted more than once and resulted in conservative estimates of the number of substitutions at these sites. Substitutions were mapped onto 1000 samples from posterior distributions of trees generated in MRBAYES based on both the data set for the gene and on the concatenated *Mc1r/Lcat* data sets (11 million generation MCMC, 1 million generation burn-in, GTR+Γ). We used the GTR+Γ model for mapping; mutational parameters were sampled from the posterior distribution for the *Zp2* and *Zp3* data sets. Ten realizations (mappings generated that are consistent with the data) were generated for each amino acid site for each of the 1000 trees for each data set.

We performed an additional mapping analysis to determine if there is significant variation in rate of substitution across lineages. For this analysis, we focused on the branches on the tree where substitutions occur. As with parsimony mapping, we mapped changes onto the ML trees based on the *Mc1r/Lcat* data from the same individuals. Using SIMMAP, we determined the mean total number of nonsynonymous and synonymous substitutions that occurred on each branch over 1000 realizations per codon. In order to determine if patterns of change were different from expectations (i.e. if there were no increase in rate of substitution for any particular branch), results for the observed data were compared to a null distribution based on 100 simulated datasets each generated from 1000 realizations for each codon with the same mutational parameters. The rate class and starting state for each codon realization was determined by passing from the tips to the root of the tree and determining conditional likelihoods of rate/state at each node. States at the tips of the tree were then simulated using that rate category (Bollback 2006). Observed values were considered significantly different from expected if they fell outside 95% of the probability density of the simulated distribution.

Results

Structure and sequence variation of egg proteins

Intron/exon structure for both *Zp2* and *Zp3* is conserved between *Peromyscus* and *Mus* (Fig. 1). Sequence identity between *Peromyscus polionotus* and *Mus* is 85% for *Zp2* and 84% for *Zp3*. Protein length is largely conserved with *Mus*; ZP2 is identical in length, and the few amino acid insertions/deletions in ZP3 (3 indels, 1 – 2 aa each) are very small. Conservation of length and ability to align the entire amino acid sequence suggest that these proteins probably retain the domain structures predicted in *Mus*. There are, however, some potentially important differences between *Peromyscus* and *Mus* ZP3. Namely, numerous gains and losses of glycosylation sites have occurred; these changes may have functional consequences, as some evidence indicates glycosylation is critical to ZP3 function (Chen, Litscher, and Wassarman 1998; but see Dean 2004). Three of six N-glycosylation sites found in *Mus* and *Rattus* were lost in *Peromyscus* and two N-glycosylation sites in different positions were gained. In addition, two O-glycosylation sites (*Mus* Ser-332 and Ser-334), which have been identified as essential for sperm-binding by ZP3 (Chen, Litscher, and Wassarman 1998; but see Dean 2004), are conserved in mouse, rat, hamster and human. One of these sites (331, homologous to *Mus* 332) has been lost in *Peromyscus.*

Phylogenetic reconstruction

Mc1r/Lcat trees produced by ML, Bayesian, and NJ methods were consistent with each other with the exception of lineages within *P. maniculatus*, thus, only the ML tree is presented (Fig. 2). Topologies were also consistent with published species trees based on morphological and molecular data (Avise, Smith, and Selander 1974; Rogers and Engstrom 1992; Tiemann-Boege et al. 2000). Species formed monophyletic groups with two exceptions: (1) a single *P. aztecus* individual fell outside the clade containing the other two *P. aztecus* individuals and *P. boylii* individuals (this individual was placed outside the clade for *Zp2* and *Zp3* as well, thus the taxonomic identity of that sample is uncertain), and (2) *P. leucopus* was paraphyletic. Similar to the *Mc1r/Lcat* phylogeny, the egg protein gene trees generated by different methods did not differ in topology, although some clades in the NJ tree were unresolved in the ML and Bayesian trees; thus, only the ML trees are presented (Fig. 2). Strikingly, gene trees of *Zp2* and *Zp3* were not consistent with each other, with *Mc1r/Lcat* trees, or with published phylogenies. The topology of these gene trees may reflect cases where the same amino acid substitution occurred independently in more than one lineage (see Results). For example, the ML tree for *Zp2* groups two alleles from *P. truei* with the *maniculatus* and *leucopus* species groups, a relationship not consistent with any other phylogeny. Exclusion of sites that changed in parallel in multiple lineages resulted in topologies for *Zp2* and *Zp*3 that were more similar to *Mc1r/Lcat* trees and published phylogenies (data not shown).

Both *Zp2* and *Zp3* experienced numerous amino acid substitutions during the evolution of the *Peromyscus* genus. Twenty-five of 127 amino acid sites (19.7%) in exons 8 – 10 of *Zp2* (Fig. 3a) and 22 of 76 sites (28.9%) in exons 6 – 7 of *Zp3* (Fig. 3b) were variable, with several sites having multiple substitutions. For comparison, only 13.9% of sites in *Mc1r* and 5.4% of sites in exon 6 of *Lcat* were variable. However, overall estimates of ω for *Zp2* (0.38) and *Zp3* (0.31) were less than one, indicating that if these genes experienced positive selection, selection acted on a subset of amino acid sites.

Intraspecific variation

Despite the limited number of alleles sampled for each species, we found extensive intraspecific amino acid variation in both *Zp2* and *Zp3*. For example, we identified four alleles of *Zp2* in three *P. aztecus* individuals with five variable amino acid sites. For *Zp3*, we found three alleles in three *P. truei* individuals, again with five variable amino acid sites. Alleles from a single species did not always form monophyletic groups, indicating alleles from different species were sometimes more similar than alleles within species.

Amino acid sites under selection

Results from all four ML approaches for detecting selection indicated that a proportion of amino acid sites of both egg proteins have evolved adaptively. For *Zp2*, the LRTs comparing NY selection model M8 to neutral models (M7 and M8A) were significant ($p < 0.05$), with 1 – 2% of sites in the positively selected class with a mean ω = 7.93 (range 7.77 – 9.27). The BEB procedure identified sites 239 and 321 as likely targets of positive selection. Results of the NY models for *Zp2* were consistent among analyses using the ten most probable Bayesian trees and the NJ tree; significance of the LRTs and sites identified as targets of positive selection did not differ. Estimates of ω and posterior probabilities were similar for the ten Bayesian trees but differed somewhat for the NJ tree, average values for the Bayesian trees are presented in Table 2. For *Zp3*, LRTs comparing M8 to M8A were significant for all trees, and LRTs comparing M8 to M7 were significant for four of the eleven trees and all tests had $P < 0.10$. Results averaged across the 10 most probable Bayesian trees are presented in Table 2. The non-significance of the LRT comparing M8 to M7 in some cases may be due to the fact that the comparison of the test statistic to a χ^2 with two degrees of freedom is an approximation causing the test to be conservative, particularly for short, closely related sequences (Anisimova, Bielawski, and Yang 2001). Parameter estimates indicate that 2 – 3% of sites are in the positively selected class with a mean ω = 5.26 (range 4.80 – 6.82). The BEB approach identified sites 343 and 345 as targets of positive selection for all analyses and site 316 in four of the eleven analyses.

The SLAC method did not identify any sites in ZP2 or ZP3 with evidence of positive selection significant at the $P < 0.05$ level, however both sites identified with the BEB method in ZP2 (239 and 321) and one of the sites in ZP3 (345) had $P < 0.20$ of positive selection (Table 2). Lack of significance at the 0.05 level is not surprising, as counting methods have low power with sequences of low divergence, and analyses of simulated datasets of similar size indicate that *P* values for the SLAC and FEL methods < 0.20 have a true Type I error rate of < 5% (Kosakovsky Pond and Frost 2005b). Using the FEL method, sites 239 and 321 of ZP2 and 345 of ZP3 were significant at the $P < 0.05$ level. An additional three sites in ZP3 were significant at the $P < 0.20$ level. The REL method also identified sites 239 and 321 of ZP2 as positively selected, although the posterior probability of selection for site 321 (0.76) was lower than with the BEB method (0.90). REL estimates of ω were higher than NY estimates for both sites. For ZP3, REL identified 15 sites with posterior probabilities > 0.5 of positive selection (290, 295, 296, 308, 316, 320, 324, 326, 329, 330, 335, 337, 340, 343, 345). However, in most cases high ω values were due to low d_S rather than high d_N; of the 15, only sites 343 and 345 (sites identified by all BEB analyses) were assigned to the class with the highest d_N. Those two sites and the third site identified in some of the BEB analyses (316) also had the highest estimates of ω. Thus, the results were relatively consistent with BEB, although as for ZP2, estimates of ω were higher for each site.

In summary, all four ML approaches identified sites 239 and 321 of ZP2 and site 345 of ZP3 as likely targets of selection. For ZP3, an additional two sites (316 and 343) were identified by some, but not all methods, as targets of selection.

Mapping of amino acid substitutions

The pattern of amino acid change based on parsimony mapping provides further evidence that *Zp2* and *Zp3* evolved under positive selection (Fig. 4). Eight amino acid sites in *Zp2* and seven sites in *Zp3* changed independently to the same amino acid in two or more *Peromyscus* lineages (Table 3). For example, site 239 in *Zp2* changed from arginine (R) to histidine (H) in three different *Peromyscus* lineages, and the reverse change occurred in four lineages. In *Zp3*, site 345 changed from arginine (R) to glutamine (Q) in three lineages. This change, although classified as conservative based on Grantham's distance, which takes into account amino acid size, hydrophobicity, charge and polarity, was a change from a positively charged to a non-charged residue. Parallel evolution at the amino acid sequence level can be interpreted as evidence of adaptive evolution (Zhang 2003); consequently, sites that have changed in parallel are likely targets of selection in addition to those identified with the ML approach.

There are two additional sites in *Zp3* (337 and 340) with patterns of substitution consistent with positive selection. Both of these sites had two substitutions in a single lineage over a relatively short period of time. For example, site 340 changed from glutamic acid (E) to aspartic acid (D) in the clade containing *californicus*/*eremicus/eva/fraterculus/crinitus*, and subsequently from aspartic acid (D) to alanine (A) in *P. crinitus*.

As previously mentioned, sequence comparisons between closely related species introduces the possibility of recombination, which could result in patterns that look like recurrent substitution. To address this potential problem we compared patterns of parallel substitution between the egg protein genes and the non-reproductive genes (*Mc1r* and *Lcat*). Recombination within extant species or a polymorphic ancestor is expected to generate similar patterns for both sets of genes, assuming they experience similar recombination rates, and would affect patterns of both nonsynonymous and synonymous substitutions in a similar manner. First, our results show that parallel amino acid substitutions were rare or absent in the non-reproductive genes. Second, the ratio of sites with parallel nonsynonymous substitutions to sites with parallel synonymous substitutions was significantly higher in reproductive genes than in non-reproductive genes (Fisher's exact test, $P = 0.002$). In addition, proportions of parallel sites did not differ significantly between genes within each class. These results support the supposition that positive selection, rather than recombination, is the cause of parallel patterns of amino acid substitutions in ZP2 and ZP3.

In order to determine potential functional consequences of amino acid substitutions, we examined the spatial pattern of nucleotide substitution in *Zp2* and *Zp3*. In addition to considering the location of adaptively evolving sites, overall patterns of synonymous and nonsynonymous change across the sequenced regions were determined through Bayesian mapping (Fig. 5). Amino acid substitutions in *Zp2* were not localized in any one region and there was no clustering of the sites identified as positively selected by the ML methods (239 and 321) or sites that experienced parallel changes (Fig. 3a). These results are consistent with the dispersed pattern of sites identified as positively selected in an analysis of *Zp2* in a diverse set of mammals (Swanson et al. 2001b). In addition, the specific functional roles of different domains of *Zp2* are not well characterized; therefore it is difficult to predict whether changes at these sites might impact egg-sperm binding. In contrast, amino acid substitutions in

Zp3 were concentrated in and around the region homologous to the *Mus* sperm-combining site (Fig. 5). In addition, sites identified as positively selected by more than one ML method (316, 343 and 345) and sites that have changed in parallel clustered in this region (Fig. 3b). Interestingly, these sites neighbor but are not the same as those sites identified as positively selected in *Zp3* in more divergent mammalian taxa (Swanson et al. 2001b).

For both *Zp2* and *Zp3*, the amount of amino acid change varied across lineages, both in absolute terms and in relation to the amount of synonymous change (Fig. 4). Differences are apparent when comparing the number of amino acid substitutions on each branch determined by parsimony to branch length determined by overall substitution. For example, in *Zp3*, parsimony mapping suggests there have been three amino acid substitutions in the *P. crinitus* lineage and four sites were variable within the species, yet only one substitution has occurred in the rest of the clade. In *Zp2*, a different pattern is observed for this clade: *P. crinitus* has not fixed any substitutions, but there have been five substitutions in the rest of the clade. In addition, we did not observe a consistent pattern of decrease in substitution rate in *Zp2* and *Zp3* in the two monogamous taxa, *P. polionotus* and *P. californicus*.

Bayesian mapping allowed us to test whether elevated rates of amino acid substitution in some lineages were significantly different from neutral expectations. In general, results from the Bayesian approach were in agreement with patterns inferred from parsimony results, but there were cases where two branches of similar length had the same number of nonsynonymous changes but one was significantly elevated and the other was not (Fig. 4). This discrepancy is the result of substitution parameters that are not taken into account by parsimony mapping, such as the rate category for the codon and the type of nucleotide substitution that occurred. For example, rates of change from C to T in *Zp3* were approximately five times higher than rates of change from A to C; branches with several substitutions that tend to be rare were more likely to be identified as having elevated rates by the Bayesian method. For *Zp2*, there were several branches that had high nonsynonymous rates but no amino acid substitutions as determined by parsimony. These were cases where mean rates for all realizations were very low, but non-zero due to a small proportion of realizations in SIMMAP that were inconsistent with parsimony. If the means for all simulated datasets were zero, then the very small values for the observed data were significantly elevated. Several branches had elevated nonsynonymous rates in both *Zp2* and *Zp3*, including the branches leading to the *boylii/aztecus* and the *truei/difficilis* lineages. This pattern suggests that, although these two egg proteins are involved in different stages of the fertilization process (Wassarman and Litscher 2001), selection may have acted on both proteins in the same lineages.

Discussion

In the past decade, rapid evolution of reproductive proteins has been documented in a wide variety of taxa (Swanson and Vacquier 2002b). In internally fertilized species, research on patterns of evolution of reproductive proteins in closely related taxa has been primarily limited to *Drosophila* species (but see Jansa, Lundrigan, and Tucker 2003). Here, we provide strong evidence that the egg proteins *Zp2* and *Zp3* have experienced positive selection in the *Peromyscus* genus. In addition, we identify specific amino acid sites in *Zp2* and *Zp3* that are likely targets of selection. In *Zp3*, these sites are clustered in and around the functionally important sperm-combining region. We show that some amino acids changed in parallel in multiple lineages, providing further support that these changes are adaptive and suggesting that the number of available pathways of adaptive evolution may be constrained. Finally, using a Bayesian method to map amino acid substitutions, we identify lineages with elevated rates of nonsynonymous change in both *Zp2* and *Zp3.* These data confirm that patterns of evolution of reproductive proteins across mammals are reflective of processes at lower taxonomic levels and suggest future avenues of investigation to characterize the potential functional consequences of amino acid change on fertilization potential.

For both ZP2 and ZP3, we have identified several species that have variation in amino acid sequence. In some cases, alleles were not monophyletic with respect to species. This pattern could be a result of incomplete lineage sorting, however, the same individuals were monophyletic or unresolved with respect to species for the autosomal, non-reproductive genes *Mc1r* and *Lcat*. This pattern suggests that selection may be maintaining divergent *Zp2* and *Zp3* alleles within species. However, more detailed intraspecific analysis is needed to confirm and explain this result. Similar patterns of extensive polymorphism and divergence have been found in other adaptively evolving reproductive proteins, including the sea urchin sperm protein bindin and *Drosophila* Acps*,* whereas other reproductive proteins, including abalone lysin, appear to have experienced selective sweeps resulting in very little intraspecific variation (Swanson and Vacquier 2002a).

In addition to the ML approaches, the parallel pattern of change at several sites in *Zp2* and *Zp3* provided evidence that these proteins have evolved adaptively. Parallel or convergent evolution at the amino acid sequence level can be interpreted as evidence of adaptive evolution; examples include lysozymes of cows and langurs (Stewart, Schilling, and Wilson 1987), butterfly and vertebrate opsins (Briscoe 2001), and HIV envelope protein genes between different lineages within a patient (Holmes et al. 1992). However, some sites that had parallel changes were not identified as targets of positive selection through the ML approaches, which assume that ω is consistent through time at a particular codon. The expectation that selective pressure remains constant is unrealistic; however, statistical methods that account for variation in ω both among codons in a sequence and through time require a large amount of variation and have thus far been applied successfully only to evolution of viral sequences, where rates of evolution are exceptionally high (Guindon et al. 2004). The parallel pattern of amino acid change allowed us to identify sites that are likely targets of positive selection, but where response to selection was limited to specific lineages and/or to specific times during the *Peromyscus* radiation.

This repeated pattern of amino acid change suggests that there may be a finite number of ways to change adaptively. If all substitutions that occurred multiple times were conservative in terms of amino acid properties, we might infer that this pattern is a result of the negative consequences of radical change, even in the context of positive selection. However, for both *Zp2* and *Zp3*, several repeated changes were not conservative, as defined by changes in charge or by Grantham's distance (Grantham 1974). Such non-conservative changes have been found to occur much less frequently than expected under neutrality (Li, Wu, and Luo 1984). Thus, the non-conservative changes we observed seem more likely to have consequences for protein structure and/or function.

In addition to parallel changes at single sites across taxa, we also observed correlated amino acid change in two sites that occurred in independent lineages. Two substitutions at sites 320 and 326 of *Zp3,* both from aspartic acid (D) to asparagine (N), occurred in the *P. aztecus* and *P. melanophrys* lineages (Fig. 3b). This pattern is intriguing as, in addition to the fact that these substitutions are charge changing, the change at site 326 created a potential N-glycosylation site. In fact, this site is also an N-glycosylation site in *Mus* and *Rattus*, and is known to be occupied in *Mus* (Boja et al. 2003). Evidence indicates that glycosylation of ZP3 in *Mus* is required for sperm-binding (Chen, Litscher, and Wassarman 1998; but see Dean 2004); consequently, changes at glycosylation sites may have a direct impact on egg-sperm binding.

Variation in the amount of nonsynonymous change that has occurred in different *Peromyscus* lineages suggests that the selective forces acting on these genes have not remained the same throughout the evolution of the genus. Patterns of variation within species and between members of a sister-species pair varied across taxa (Fig. 3). For example, there were differences in *Zp3* between the sister species *P.*

maniculatus and *P. polionotus* as well as variation within each species. In contrast, the amino acid sequence was identical for all samples of the sister species pair *P. leucopus* and *P. gossypinus*, which share a similar divergence time with *P. maniculatus*/ *P. polionotus* (Blair 1950).

Application of a Bayesian method for mapping nucleotide substitutions allowed us to identify variation in substitution rate both along the length of each gene and between lineages. Although these patterns can be inferred by examining sequence alignments and by parsimony mapping of substitutions, the Bayesian method provides a quantitative estimate of the amount of change that has occurred and allows statistical tests of elevated lineage-specific substitution rates. Using this approach, we identified several branches with significantly elevated rates of amino acid substitution in both ZP2 and ZP3.

In internally fertilized species, sperm competition and sexual conflict have been proposed as important factors driving reproductive protein evolution (Wyckoff, Wang, and Wu 2000; Price et al. 2001; Swanson et al. 2001b; Torgerson, Kulathinal, and Singh 2002). Thus, variation in rates of evolution between species with different mating systems is predicted. Specifically, monogamous species may have lower rates of reproductive protein evolution because of the lack of sperm competition and reduced sexual conflict. For example, rates of evolution of two genes encoding semen proteins that are components of the mating plug are correlated with female promiscuity in primates (Kingan, Tatar, and Rand 2003; Dorus et al. 2004) and rates of *Acp* evolution are higher in *Drosophila* species with higher remating rates

(Wagstaff and Begun 2005). Variation in mating system is found in *Peromyscus*; while most *Peromyscus* species are promiscuous, monogamy has evolved independently in two of the species sampled in this study, *P. californicus* and *P. polionotus* (Kleiman 1977; Foltz 1981; Ribble 1991; Ribble 2003). We did not find a consistent reduction of rate of evolution of *Zp2* or *Zp3* in the monogamous taxa. While *P. californicus* has had very little change in either of the proteins, the *P. polionotus* lineage has had multiple amino acid substitutions in both *Zp2* and *Zp3* (Fig. 4). Proteins involved in sperm morphology and performance may be more appropriate candidates to detect evidence of mating system effects on evolutionary rates.

Conclusions

Recent empirical and theoretical studies suggest that rapid evolution of reproductive proteins may play an important role in the evolution of reproductive isolation (Price et al. 2001; Servedio 2001; Coyne and Orr 2004). Allopatric populations that have limited overall phenotypic divergence may have significant divergence of reproductive proteins, leading to post-mating, pre-zygotic reproductive incompatibilities upon secondary contact (gametic isolation) and potentially resulting in reinforcement. In marine invertebrates, patterns of evolution of egg-sperm binding proteins suggest that changes have potentially contributed to reinforcement and reproductive isolation (Galindo, Vacquier, and Swanson 2003; Geyer and Palumbi 2003). However, in internally fertilized species, evidence for adaptive evolution of reproductive proteins within a genus is limited to *Drosophila* (Begun et al. 2000;

Swanson et al. 2001a; Swanson et al. 2004; Wagstaff and Begun 2005) and *Mus* (Jansa, Lundrigan, and Tucker 2003, but see online supplementary material).

Tests for positive selection that do not require population samples, and those which can identify specific sites subject to positive selection, generally have been applied to higher level taxa, where amino acid variation is more likely to be sufficient to significantly reject neutral models (Anisimova, Bielawski, and Yang 2002; Yang and Nielsen 2002). Here, we used a combination of several maximum likelihood approaches and parallel patterns of substitution to detect selection and identify the specific amino acid sites that are evolving adaptively. Our results documenting positive selection acting on *Zp2* and *Zp3* within a genus confirm that we can successfully extend work documenting adaptive evolution of reproductive proteins across mammals (Wyckoff, Wang, and Wu 2000; Swanson et al. 2001b; Torgerson, Kulathinal, and Singh 2002; Jansa, Lundrigan, and Tucker 2003; Swanson, Nielsen, and Yang 2003) to look at how these proteins have changed between closely related species, where isolating barriers act and have evolved recently. It is certainly possible that adaptive change of *Zp2* and *Zp3* did not contribute to reproductive isolation between *Peromyscus* species, either because changes were not sufficient to prevent fertilization or because other isolating barriers had evolved before the egg proteins had diverged sufficiently to cause incompatibilities. More detailed intraspecific analysis is necessary to determine if differences in *Zp2* or *Zp3* genotype correlate with incipient reproductive isolation between populations.

In order to confirm a role of reproductive protein evolution in gametic isolation, functional consequences of amino acid change on fertilization potential must be determined. Here, we identified specific amino acid sites likely to be targets of selection in *Zp2* and *Zp3.* Particularly appropriate for functional studies are the adaptively evolving sites in *Zp3*, which are clustered in and around the region known to be critical to successful egg-sperm binding in *Mus* (Wassarman and Litscher 2001). Evidence that sites in or around this region have evolved adaptively has been found previously in analysis of a taxonomically diverse set of mammals (Swanson et al. 2001b). Interestingly, the specific amino acid sites identified here are adjacent but not identical to the sites evolving adaptively across mammals, underscoring the value of examining evolutionary processes at multiple taxonomic levels. These data identify this region of ZP3, and the positively selected sites specifically, as promising targets for future functional assays of allelic differences in sperm-binding ability.

Acknowledgements

The authors wish to thank J.R. Kohn and V.D. Vacquier and three anonymous reviewers for helpful discussion and comments on this manuscript. J.P. Bollback generously provided a modified version of SIMMAP. J.P. Huelsenbeck and P. Andolfatto provided guidance with data analysis. E. Chuong and F. Kondrashov wrote perl scripts to summarize mapping and codon bootstrap results. R.J. Fisher, L.M. Mullen, and B.J. Haeck assisted with fieldwork. G. Arnaud, M. Kalcounis-Rueppell, M. Nachman, K. Nutt, J. Storz, A. Suazo, M. Wooten, the M.L. Bean Life

Science Museum (Brigham Young University), the Museum of Southwestern Biology (University of New Mexico), the Museum of Texas Tech University, the *Peromyscus* Genetic Stock Center, and the Museum of Vertebrate Zoology (University of California, Berkeley) kindly provided tissue samples. This research was supported by a Howard Hughes Medical Institute Predoctoral Fellowship, the *American Society of Mammalogists* Grants-in-Aid of Research, the University of California Reserve System Mildred Mathias Award (to LMT), NSF doctoral dissertation improvement grant DEB 0608030 (to LMT and HEH), and NSF DEB 0344710 (to HEH).

Chapter 1, in full, is a reprint of the material as it appears in *Molecular Biology and Evolution* 2006: Turner LM, and Hoekstra HE. 2006. Adaptive evolution of fertilization proteins within a genus: Variation in ZP2 and ZP3 in deer mice (*Peromyscus*). Mol. Biol. Evol. **23**:1656-1669. The dissertation author was the primary author and investigator of this paper.

Table 1.1 Samples of 15 *Peromyscus* species and the outgroup *Onychomys torridus*

Species	Sampling locations
P. aztecus	Michoacan, Mex; Guerrero, Mex; Chiapas, Mex
P. boylii	Culberson Co., TX; Nuevo Leon, Mex; San Luis Obispo Co., CA; Monterey Co., CA*(2)
P. californicus	Los Angeles Co., CA; San Diego Co., CA[a,b] (2); Monterey Co., CA[a,b] (2)
P. crinitus	Yuma Co., AZ* (2); Tooele Co., UT
P. difficilis	Lincoln Co., NM; Cibola Co., NM; Zacatecas, Mex
P. eremicus	Yuma Co., AZ; Dona Ana Co., NM; Soccorro Co., NM
P. eva	Baja California Sur, Mex[a,b] (2)
P. fraterculus	San Diego Co., CA
P. gossypinus	Decatur Co., TN; Bradley Co., AR; Bowie Co., TX
P. leucopus	Avery Co., NC; Antelope Co., NE; Lake Co., OH
P. maniculatus	Washtenaw Co., MI; Boulder Co., CO; Coconino Co., AZ; Mono Co., CA* (2); San Diego Co, CA
P. melanophrys	Zacatecas, Mex; Durango, Mex; Jalisco, Mex
P. mexicanus	Guanacaste, Costa Rica; Francisco Morazan, Honduras* (2); Selva Negra, Nicaragua
P. polionotus	Santa Rosa Co., FL (2); Marion Co., FL; Lee Co., AL[b]; Lake Co., FL
P. truei	Durango, Mex; Yavapai Co., AZ; Armstrong Co., TX
O. torridus	Kern Co., CA; Nye Co., NV

Note- One individual was sampled at each site, except for those sites indicated with (2), where two individuals were sampled. Both *Zp2* and *Zp3* were sequenced except for those samples indicated with [a]*Zp2* only, [b]*Zp3* only, and *one individual from site sequenced for *Zp3* only. See online supplementary information for sample sources and accession numbers.

Table 1.2 Positive selection on egg coat proteins in *Peromyscus*

Gene	N	L_C	S	d_N/d_S	Selection Parameters	P M8 vs. M7	P M8 vs. M8A	Positively Selected Sites (M8)	Positively Selected Sites (SLAC)	Positively Selected Sites (FEL)	Positively Selected Sites (REL)
Zp2	88	127	0.64	0.38	$p_s = 0.02$	**0.016**	**0.002**	239 (0.99)	239 (0.07)	239 (**0.03**)	239 (0.99)
					$\omega_s = 7.77$			ω = 5.52			ω = 13.55
								321 (0.90)	321 (0.16)	321 (**0.05**)	321 (0.76)
								ω = 5.13			ω = 10.57
Zp3	96	76	1.15	0.31	$p_s = 0.02$	0.059	**0.011**	316 (0.52)		316 (0.07)	316 (0.81)
					$\omega_s = 5.35$			ω = 2.09			ω = 3.41
										329 (0.20)	343 (0.70)
										337 (0.15)	ω = 3.69
								343 (0.64)			
								ω = 2.86			
								345 (0.99)	345 (0.06)	345 (**0.03**)	345 (0.98)
								ω = 4.50			ω = 13.65

N - number of alleles sequenced; L_C - length of sequence in codons; S - tree length in substitutions per codon; d_N/d_S - ratio averaged across all sites and lineages [estimated with PAML, M0 (Yang 2000)]; Selection Parameters – average values from M8 estimated in PAML using 10 most probable Bayesian trees, p_s - proportion of sites in the 'ω > 1' class, ω_s - ω estimate for that class; P – P value of likelihood ratio test comparing models listed, significant values (0.05 level) are **bold**; Positively Selected Sites (M8) – sites with posterior probability > 0.5 of being in positively selected class in any analysis given as amino acid position, posterior probability that the site experiences positive selection, and the average ML estimate of ω at the site. Positively Selected Sites (SLAC)/(FEL) – sites (and P values) identified using the single-likelihood ancestor counting (SLAC) and fixed effects likelihood (FEL) methods (Kosakovsky Pond and Frost 2005b) with P-values consistent with a Type I error rate less than 5% ($P < 0.20$). Positively Selected Sites (REL) – results of the random effects likelihood method for sites identified by BEB procedure are given as amino acid position, posterior probability that the site experiences positive selection, and the estimate of ω at the site.

Table 1.3 Parallel amino acid substitutions in ZP2 and ZP3

Change	Number of independent changes	Starting aa class	Ending aa class	Charge changing?	Grantham's distance	Type of change	Possible alternative aa substitutions
ZP2:							
S237N	2	P	P	N	46	C	5
R239H	**3**	+	+	**N**	**29**	**C**	**4**
H239R	**4**	+	+	**N**	**29**	**C**	**6**
A265V	2	NP	NP	N	64	MC	5
N297T	2	P	P	N	65	MC	6
R301S	2	+	P	Y	110	MR	4
D309N	2	-	P	Y	23	C	6
R321Q	**3**	**P**	+	**Y**	**43**	**C**	**3**
T353M	2	NP	P	N	81	MC	3
ZP3:							
A316T	2	NP	P	N	58	MC	5
D320N	2	-	P	Y	23	C	6
K324Q	2	+	P	Y	53	MC	5
D326N	2	-	P	Y	23	C	6
H335Q	2	+	P	Y	24	C	6
W343R	**2**	**NP**	+	**Y**	**101**	**MR**	**4**
R345Q	**3**	+	**P**	**Y**	**43**	**C**	**5**

Number of independent changes determined by parsimony mapping of ZP2 and ZP3 amino acid substitutions onto *Mc1r/Lcat* ML trees. Amino acid types: + - positively charged, - - negatively charged, P - polar, NP - non-polar. Grantham's distance between starting and ending amino acid (Grantham 1974). Types of change: C - conservative (Grantham's distance <50), MC- moderately conservative (51-100), MR- moderately radical (101-150), R - radical (>150) (Li, Wu, and Luo 1984). Possible alternative aa substitutions: number of possible amino acid substitutions given the starting codon and a single nucleotide mutation. Rows in *bold* are sites identified as likely targets of selection by the ML codon models/BEB procedure.

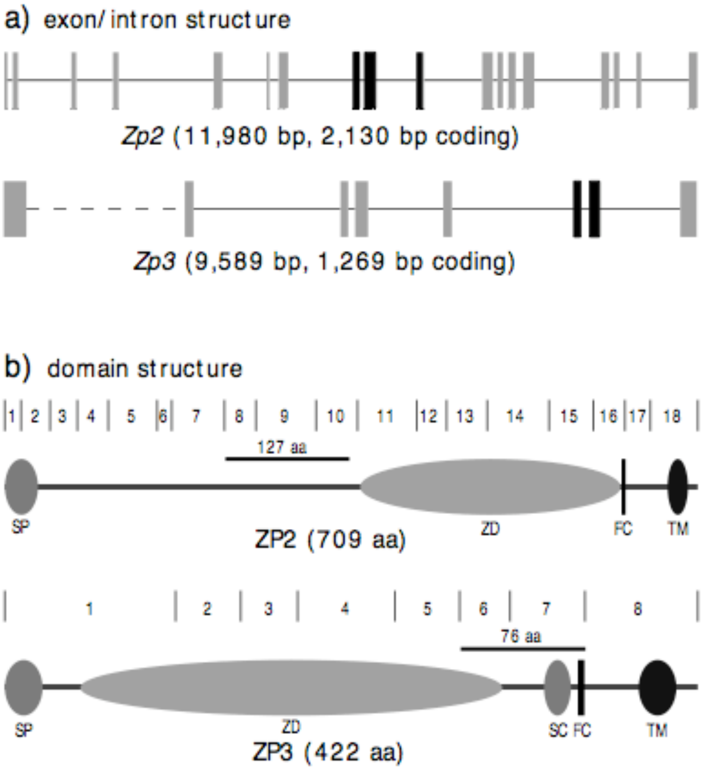


Figure 1.1 Gene and protein structure
(a) Exon/intron structure of fertilization protein genes in *Peromyscus polionotus* drawn to scale within each gene. Boxes indicate exons; areas amplified for interspecific analysis are in black. Intron 1 of *Zp3* is dashed because this region was not fully sequenced; length was estimated from *Mus*. (b) Protein structure with predicted functional domains based on alignment with *Mus* and *Rattus* (Akatsuka et al. 1998; Jovine et al. 2004). Domains are drawn to scale within each protein. Region amplified for interspecific analysis is indicated with a black bar. Exon boundaries are indicated above with tick marks and exon numbers between them. Abbreviations for domains are: SP – signal peptide, ZD – zona domain, FC – furin cleavage site, TM – transmembrane domain, and SC – sperm combining site

Figure 1.2 Maximum likelihood gene trees of (a) *Mc1r/Lcat*, (b) *Zp2* and (c) *Zp3*. In (a), individuals are identified by species name and location, and dashed lines indicate branches with a length of zero. In (b) and (c), only unique alleles are included and are indicated by species name; letters indicate different coding alleles and numbers indicate alleles that differ by synonymous substitutions only. Species are coded by color. Scale for each gene is indicated below each tree. Posterior probabilities (%) > 50 are given above and bootstrap percent values > 50 are given below each branch.

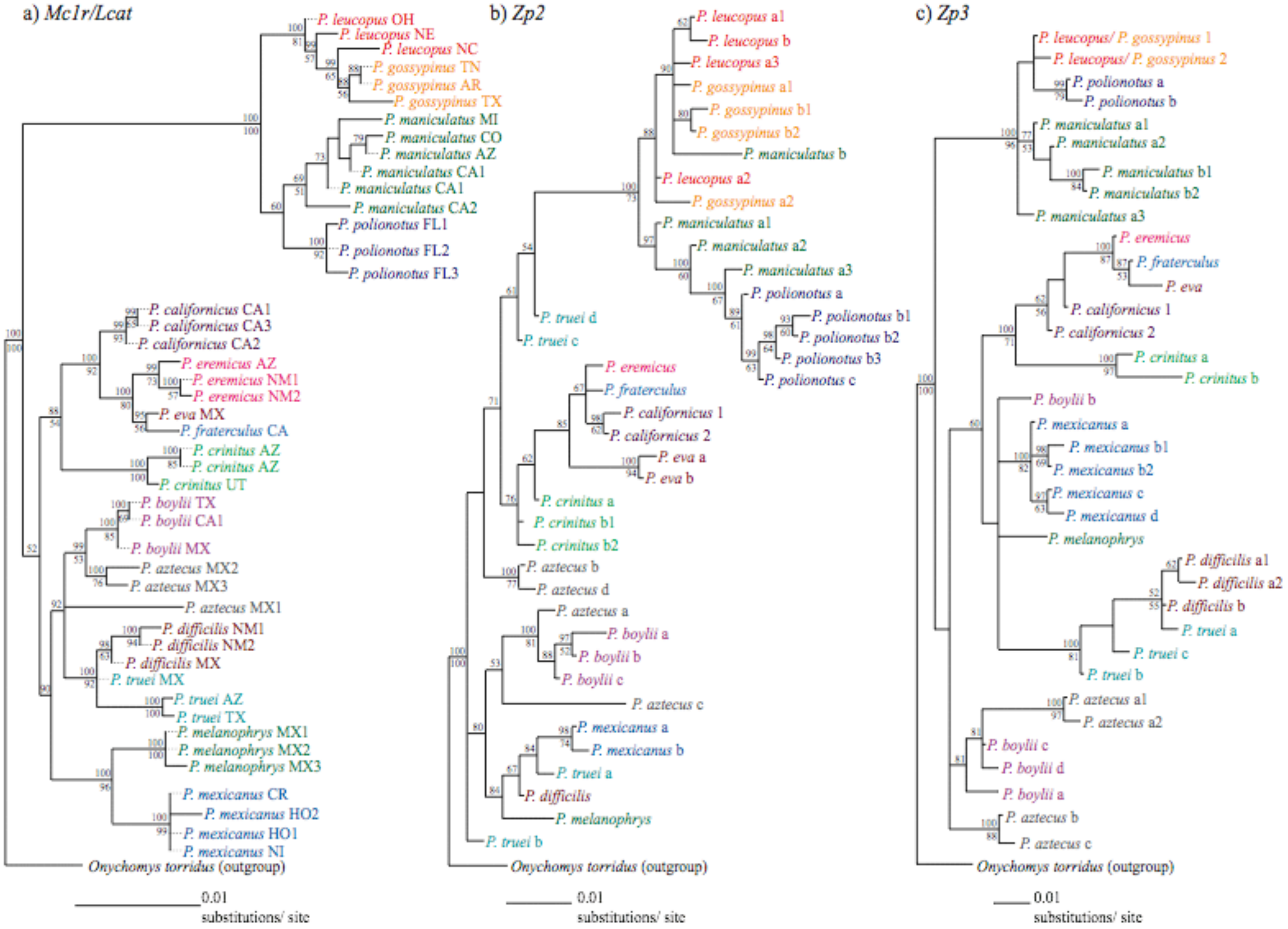
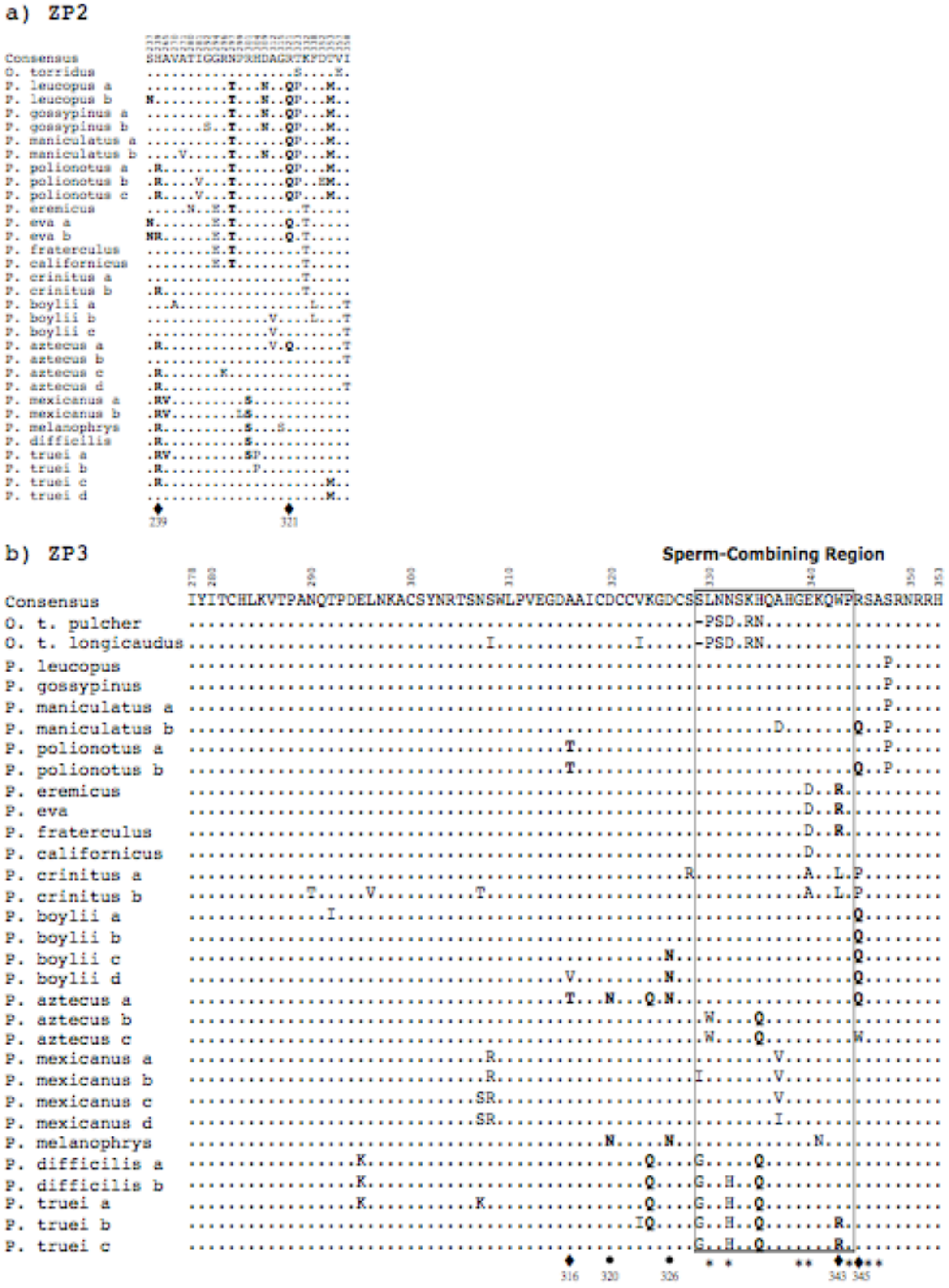

a) Mc1r/Lcat
P. leucopus OH
P. leucopus NE
P. leucopus NC
P. gossypinus TN
P. gossypinus AR
P. gossypinus TX
P. maniculatus MI
P. maniculatus CO
P. maniculatus AZ
P. maniculatus CA1
P. maniculatus CA1
P. maniculatus CA2
P. polionotus FL1
P. polionotus FL2
P. polionotus FL3
P. californicus CA1
P. californicus CA3
P. californicus CA2
P. eremicus AZ
P. eremicus NM1
P. eremicus NM2
P. eva MX
P. fraterculus CA
P. crinitus AZ
P. crinitus AZ
P. crinitus UT
P. boylii TX
P. boylii CA1
P. boylii MX
P. aztecus MX2
P. aztecus MX3
P. aztecus MX1
P. difficilis NM1
P. difficilis NM2
P. difficilis MX
P. truei MX
P. truei AZ
P. truei TX
P. melanophrys MX1
P. melanophrys MX2
P. melanophrys MX3
P. mexicanus CR
P. mexicanus HO2
P. mexicanus HO1
P. mexicanus NI
Onychomys torridus (outgroup)
0.01
substitutions/ site
b) Zp2
P. leucopus a1
P. leucopus b
P. leucopus a3
P. gossypinus a1
P. gossypinus b1
P. gossypinus b2
P. maniculatus b
P. leucopus a2
P. gossypinus a2
P. maniculatus a1
P. maniculatus a2
P. maniculatus a3
P. polionotus a
P. polionotus b1
P. polionotus b2
P. polionotus b3
P. polionotus c
P. truei d
P. truei c
P. eremicus
P. fraterculus
P. californicus 1
P. californicus 2
P. eva a
P. eva b
P. crinitus a
P. crinitus b1
P. crinitus b2
P. aztecus b
P. aztecus d
P. aztecus a
P. boylii a
P. boylii b
P. boylii c
P. aztecus c
P. mexicanus a
P. mexicanus b
P. truei a
P. difficilis
P. melanophrys
P. truei b
Onychomys torridus (outgroup)
0.01
substitutions/ site
c) Zp3
P. leucopus/ P. gossypinus 1
P. leucopus/ P. gossypinus 2
P. polionotus a
P. polionotus b
P. maniculatus a1
P. maniculatus a2
P. maniculatus b1
P. maniculatus b2
P. maniculatus a3
P. eremicus
P. fraterculus
P. eva
P. californicus 1
P. californicus 2
P. crinitus a
P. crinitus b
P. boylii b
P. mexicanus a
P. mexicanus b1
P. mexicanus b2
P. mexicanus c
P. mexicanus d
P. melanophrys
P. difficilis a1
P. difficilis a2
P. difficilis b
P. truei a
P. truei c
P. truei b
P. aztecus a1
P. aztecus a2
P. boylii c
P. boylii d
P. boylii a
P. aztecus b
P. aztecus c
Onychomys torridus (outgroup)
0.01
substitutions/ site

Figure 1.3 Amino acid sequence alignments of (a) variable amino acid sites in exons 8 – 10 (aa 236 – 362) of ZP2 and (b) all amino acid sites in exons 6 – 7 (aa 278 – 353) of ZP3; all sites are given for ZP3 in order to show spatial pattern of variable sites. In (a), numbering is consistent with *Mus* ZP2 and region in (b) aligns to aa 279 – 354 of Mus ZP3 (Boja et al. 2003). Alleles with identical amino acid sequences have been collapsed into a single haplotype, unless they are from different species. *Onychomys torridus* (*O. torridus*) sequences are included as the outgroup. Dots indicate identity with the consensus sequence. Diamonds indicate amino acid sites that have a posterior probability >50% of being under positive selection in Peromyscus in one or more BEB analysis; amino acid site numbers are below the diamonds. Amino acids that have been substituted independently in more than one *Peromyscus* lineage are in bold. In (b), the box delimits the sperm-combining region (Kinloch, Sakai, and Wassarman 1995), asterisks indicate sites that had a posterior probability >50% of being under positive selection in a phylogenetically diverse set of mammals (Swanson et al. 2001b) and closed circles indicate two sites that had correlated change in both *P. aztecus* and *P. melanophrys.*


a) ZP2
Consensus
SHAVATIGGRNPRHDAGRTKFDTVI
239
321
b) ZP3
Sperm-Combining Region
Consensus
IYITCHLKVTPANQTPDELNKACSYNRTSNSWLPVEGDAAICDCCVKGDCSSLNNSKHQAHGEKQWPRSASRNRRH
O. t. pulcher
O. t. longicaudus
P. leucopus
P. gossypinus
P. maniculatus a
P. maniculatus b
P. polionotus a
P. polionotus b
P. eremicus
P. eva
P. fraterculus
P. californicus
P. crinitus a
P. crinitus b
P. boylii a
P. boylii b
P. boylii c
P. boylii d
P. aztecus a
P. aztecus b
P. aztecus c
P. mexicanus a
P. mexicanus b
P. mexicanus c
P. mexicanus d
P. melanophrys
P. difficilis a
P. difficilis b
P. truei a
P. truei b
P. truei c
316
320
326
343
345

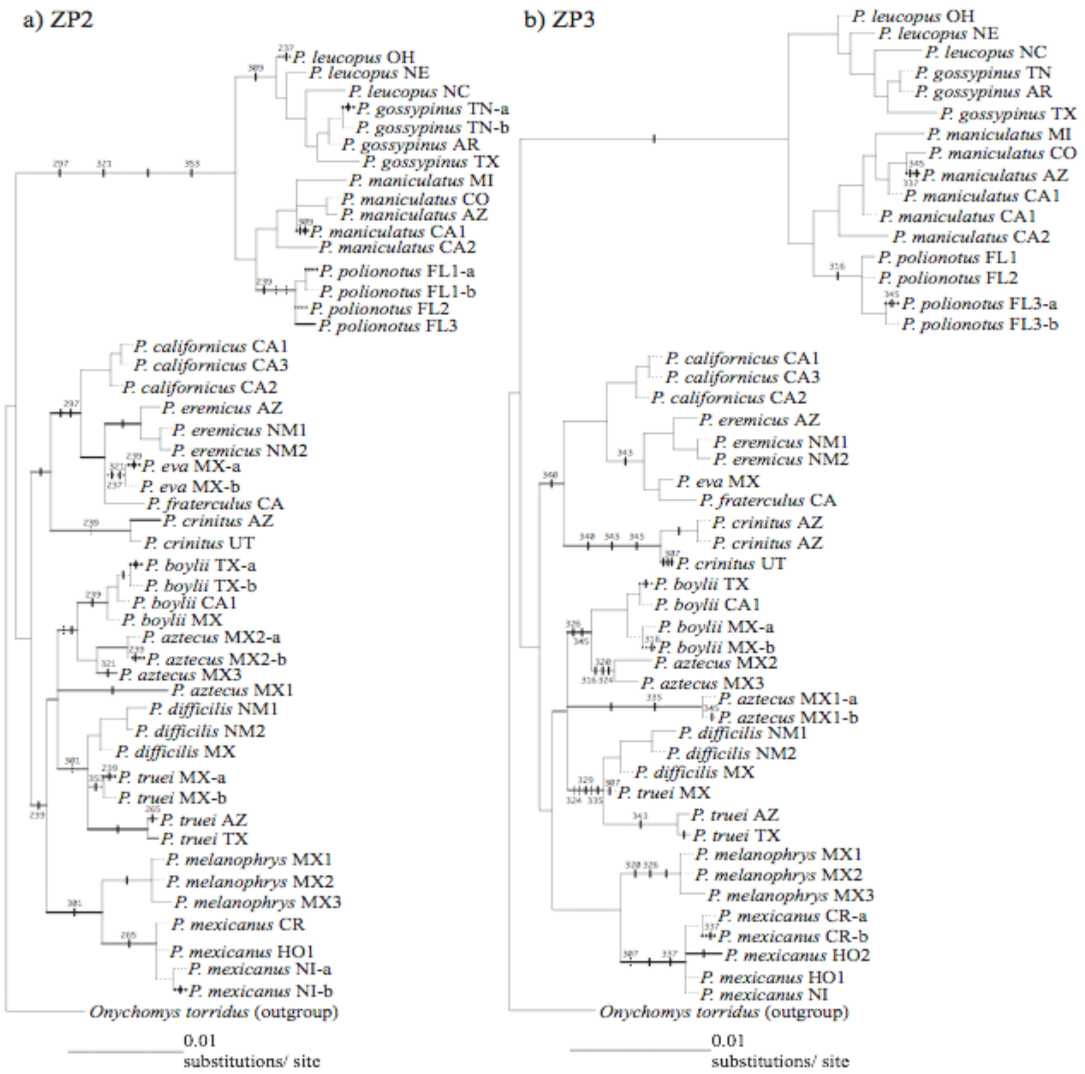
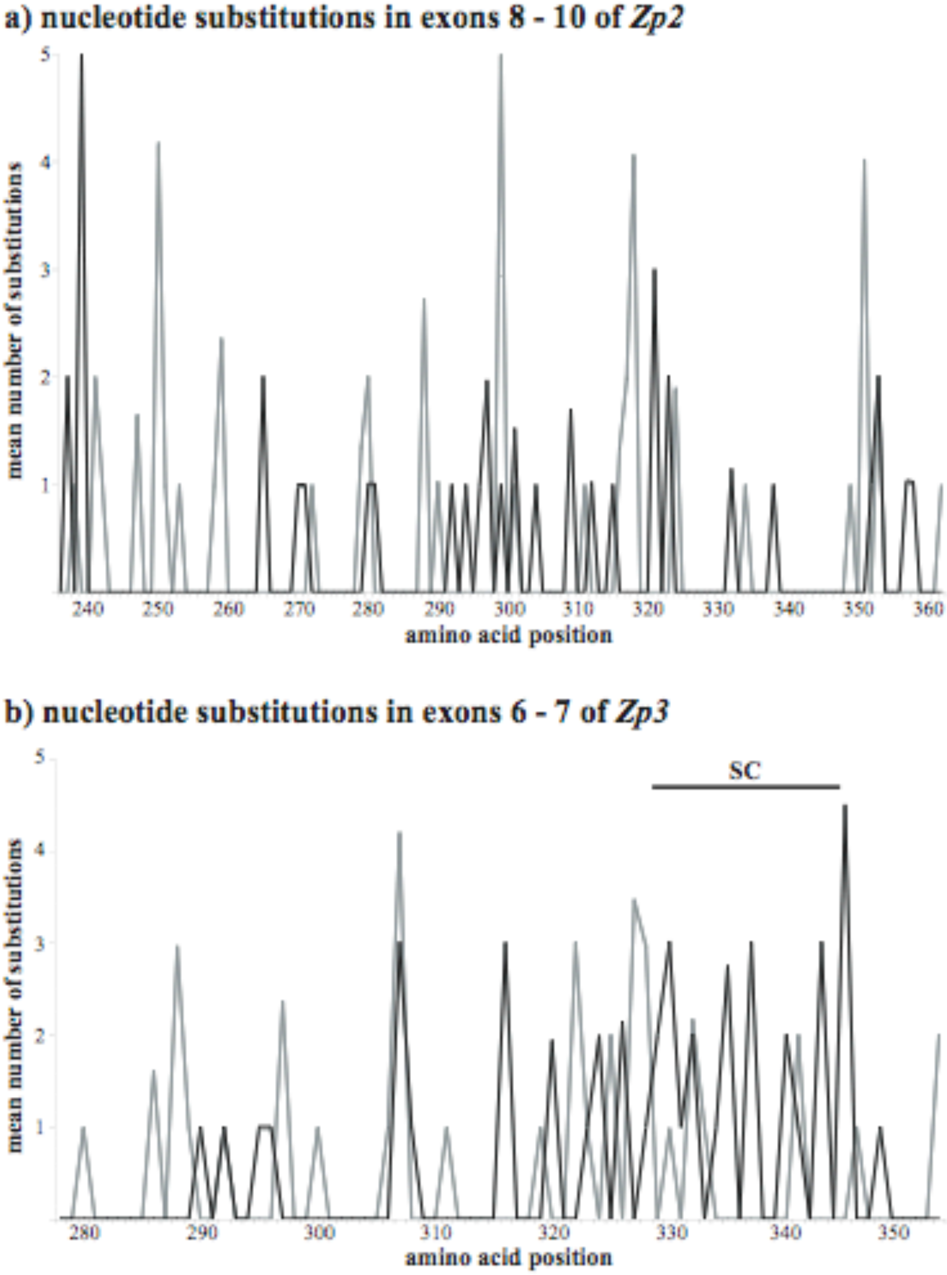


Figure 1.4 Amino acid substitutions in egg proteins mapped onto *Mc1r/Lcat* gene trees.

Substitutions were mapped using parsimony and a Bayesian method for the same individuals included in the (a) ZP2 and (b) ZP3 data sets. Scale for each gene is indicated below each tree. Dashed branches have a length of zero. Each tick mark represents a single amino acid substitution determined by parsimony. Substitutions at sites that have changed independently in more than one lineage have the amino acid position indicated. As determined by Bayesian analysis, branches with significantly higher nonsynonymous substitution rates are in bold black, and those where synonymous rates are also elevated are in bold gray.



a) nucleotide substitutions in exons 8 - 10 of *Zp2*

b) nucleotide substitutions in exons 6 - 7 of *Zp3*

Figure 1.5 Patterns of nucleotide substitution
(a) *Zp2* and (b) *Zp3*. Substitutions were mapped onto *Mc1r/Lcat* trees using the same individuals included in each gene data set. Bayesian mapping was performed using simmap. Mean numbers of substitutions are shown from 10 realizations per codon per tree for each of 1000 samples from a posterior distribution of trees generated in MRBAYES (10 million generation MCMC, 5 million generation burn-in, GTR +Γ). Gray lines indicate synonymous substitutions and black lines indicate nonsynonymous substitutions. The black bar labeled "SC" indicates the sperm-combining region of *Zp3* (Kinloch, Sakai, and Wassarman 1995).

Literature Cited

Aguade M. 1999. Positive selection drives the evolution of the Acp29AB accessory gland protein in *Drosophila*. Genetics **152**:543-551.

Anisimova M, Bielawski JP, and Yang Z. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol. Biol. Evol. **19**:950-958.

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

Anisimova M, Nielsen R, and Yang ZH. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics **164**:1229-1236.

Avise JC, Smith MH, and Selander RK. 1974. Biochemical polymorphism and systematics in the genus *Peromyscus* part 6: the *Peromyscus boylii* group. J. Mammal. **55**:751-763.

Barsh GS. 1996. The genetics of pigmentation: from fancy genes to complex traits. Trends Genet. **12**:299.

Begun DJ, and Lindfors HA. 2005. Rapid evolution of genomic *Acp* complement in the *melanogaster* subgroup of *Drosophila*. Mol. Biol. Evol. **22**:2010-2021.

Begun DJ, Whitley P, Todd BL, Waldrip-Dail HM, and Clark AG. 2000. Molecular population genetics of male accessory gland proteins in *Drosophila*. Genetics **156**:1879-1888.

Blair WF. 1950. Ecological factors in speciation of *Peromyscus*. Evolution **4**:253-275.

Boja ES, Hoodbhoy T, Fales HM, and Dean J. 2003. Structural characterization of native mouse zona pellucida proteins using mass spectrometry. J. Biol. Chem. **278**:34189-34202.

Bollback J. 2006. SIMMAP: Stochastic character mapping of discrete traits on phylogenies. BMC Bioinformatics **7**:88.

Briscoe AD. 2001. Functional diversification of lepidopteran opsins following gene duplication. Mol. Biol. Evol. **18**:2270-2279.

Chen J, Litscher ES, and Wassarman PM. 1998. Inactivation of the mouse sperm receptor, mZP3, by site-directed mutagenesis of individual serine residues located at the combining site for sperm. Proc. Natl. Acad. Sci. USA **95**:6193-6197.

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, and Thompson JD. 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. **31**:3497-3500.

Cirera S, and Aguade M. 1997. Evolutionary history of the sex-peptide (*Acp70A*) gene region in *Drosophila melanogaster*. Genetics **147**:189-197.

Civetta A, and Singh RS. 1995. High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. J. Mol. Evol. **41**:1085-1095.

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

Coyne JA, and Orr HA. 2004. Speciation. Sinauer Associates, Sunderland, Mass.

Dean J. 2004. Reassessing the molecular biology of sperm-egg recognition with mouse genetics. Bioessays **26**:29-38.

Dorus S, Evans PD, Wyckoff GJ, Choi SS, and Lahn BT. 2004. Rate of molecular evolution of the seminal protein gene *SEMG2* correlates with levels of female promiscuity. Nat. Genet. **36**:1326-1329.

Eisenbach M, and Giojalas LC. 2006. Sperm guidance in mammals - an unpaved road to the egg. Nat Rev Mol Cell Biol **7**:276.

Felsenstein J. 2004. PHYLIP (Phylogeny Inference Package) version 3.6. Department of Genome Sciences, University of Washington, Seattle, WA.

Foltz DW. 1981. Genetic evidence for long-term monogamy in a small rodent, *Peromyscus polionotus*. Am. Nat. **117**:665-675.

Galindo BE, Vacquier VD, and Swanson WJ. 2003. Positive selection in the egg receptor for abalone sperm lysin. Proc. Natl. Acad. Sci. USA **100**:4639-4643.

Geyer LB, and Palumbi SR. 2003. Reproductive character displacement and the genetics of gamete recognition in tropical sea urchins. Evolution **57**:1049-1060.

Grantham R. 1974. Amino-acid difference formula to help explain protein evolution. Science **185**:862-864.

Guindon S, Rodrigo AG, Dyer KA, and Huelsenbeck JP. 2004. Modeling the site-specific variation of selection patterns along lineages. Proc. Natl. Acad. Sci. USA **101**:12957-12962.

Holmes EC, Zhang LQ, Simmonds P, Ludlam CA, and Brown AJL. 1992. Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. Proc. Natl. Acad. Sci. USA **89**:4835-4839.

Hooper ET. 1968. Classification. Pp. 27 - 74 *in* J. A. King, ed. Biology of *Peromyscus* (Rodentia). American Society of Mammalogists.

Huelsenbeck JP, and Crandall KA. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst. **28**:437-466.

Huelsenbeck JP, and Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17**:754-755.

Jagadeeshan S, and Singh RS. 2005. Rapidly evolving genes of *Drosophila*: Differing levels of selective pressure in testis, ovary, and head tissues between sibling species. Mol. Biol. Evol. **22**:1793-1801.

Jansa SA, Lundrigan BL, and Tucker PK. 2003. Tests for positive selection on immune and reproductive genes in closely related species of the murine genus *Mus*. J. Mol. Evol. **56**:294-307.

Jansen S, Ekhlasi-Hundrieser M, and Toepfer-Petersen E. 2001. Sperm adhesion molecules: Structure and function. Cells Tissues Organs **168**:82-92.

Kingan SB, Tatar M, and Rand DM. 2003. Reduced polymorphism in the chimpanzee semen coagulating protein, semenogelin I. J. Mol. Evol. **57**:159-169.

Kleiman DG. 1977. Monogamy in mammals. Q. Rev. Biol. **52**:39-69.

Kosakovsky Pond SL, and Frost SDW. 2005a. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics **21**:2531-2533.

Kosakovsky Pond SL, and Frost SDW. 2005b. Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. **22**:1208-1222.

Kosakovsky Pond SL, Frost SDW, and Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics **21**:676-679.

Kosakovsky Pond SL, and Muse SV. 2005. Site-to-site variation of synonymous substitution rates. Mol. Biol. Evol. **22**:2375-2385.

Kuivenhoven JA, Pritchard H, Hill J, Frohlich J, Assmann G, and Kastelein J. 1997. The molecular pathology of lecithin cholesterol acyltransferase (LCAT) deficiency syndromes. J. Lipid Res. **38**:191-205.

Li WH, Wu CI, and Luo CC. 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J. Mol. Evol. **21**:58-71.

Mah SA, Swanson WJ, and Vacquier VD. 2005. Positive selection in the carbohydrate recognition domains of sea urchin sperm receptor for egg jelly (suREJ) proteins. Mol. Biol. Evol. **22**:533-541.

Mueller JL, Ram KR, McGraw LA, Bloch Qazi MC, Siggia ED, Clark AG, Aquadro CF, and Wolfner MF. 2005. Cross-species comparison of *Drosophila* male accessory gland protein genes. Genetics **171**:131-143.

Nachman MW, Hoekstra HE, and D'Agostino SL. 2003. The genetic basis of adaptive melanism in pocket mice. Proc. Natl. Acad. Sci. USA **100**:5268-5273.

Nielsen R. 2002. Mapping mutations on phylogenies. Syst. Biol. **51**:729-739.

Nielsen R, and Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**:929-936.

Price CSC, Kim CH, Gronlund CJ, and Coyne JA. 2001. Cryptic reproductive isolation in the *Drosophila simulans* species complex. Evolution **55**:81-92.

Ribble DO. 1991. The monogamous mating system of *Peromyscus californicus* as revealed by DNA fingerprinting. Beh. Ecol. Sociobiol. **29**:161-166.

Ribble DO. 2003. The evolution of social and reproductive monogamy in *Peromyscus*: evidence from *Peromyscus californicus* (the california mouse). Pp. 161 - 166 *in* U. H. Reichard, and C. Boesch, eds. Monogamy: mating strategies and partnerships in birds, humans, and other mammals. Cambridge University Press, Cambridge.

Robinson M, Catzeflis F, Briolay J, and Mouchiroud D. 1997. Molecular phylogeny of rodents, with special emphasis on murids: Evidence from nuclear gene *Lcat.* Mol. Phylogenet. Evol. **8**:423-434.

Rogers DS, and Engstrom MD. 1992. Evolutionary implications of allozymic variation in tropical *Peromyscus* of the *mexicanus* species group. J. Mammal. **73**:55-69.

Servedio MR. 2001. Beyond reinforcement: the evolution of premating isolation by direct selection on preferences and postmating, prezygotic incompatibilities. Evolution **55**:1909-1920.

Singh RS, and Kulathinal RJ. 2000. Sex gene pool evolution and speciation: A new paradigm. Genes Genet. Syst. **75**:119-130.

Stewart CB, Schilling JW, and Wilson AC. 1987. Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature **330**:401-404.

Swanson W, J., and Vacquier VD. 2002a. Reproductive protein evolution. Annu. Rev. Ecol. Syst. **33**:161-179.

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, and Aquadro CF. 2001a. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proc. Natl. Acad. Sci. USA **98**:7375-7379.

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

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

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

Swanson WJ, Yang Z, Wolfner MF, and Aquadro CF. 2001b. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc. Natl. Acad. Sci. USA **98**:2509-2514.

Swofford DL. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Mass.

Tiemann-Boege I, Kilpatrick CW, Schmidly DJ, and Bradley RD. 2000. Molecular phylogenetics of the *Peromyscus boylii* species group (Rodentia: Muridae) based on mitochondrial *cytochrome b* sequences. Mol. Phylogenet. Evol. **16**:366-378.

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

Tsaur S-C, and Wu CI. 1997. Positive selection and the molecular evolution of a gene of male reproduction, *Acp26Aa* of *Drosophila*. Mol. Biol. Evol. **14**:544-549.

Wagstaff BJ, and 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.

Wassarman PM, Jovine L, and Litscher ES. 2001. A profile of fertilization in mammals. Nat. Cell Bio. **3**:E59-E64.

Wassarman PM, and Litscher ES. 2001. Towards the molecular basis of sperm and egg interaction during mammalian fertilization. Cells Tissues Organs **168**:36-45.

Wolff JO. 1989. Social behavior. Pp. 271-291 *in* G. L. Kirkland, and J. N. Layne, eds. Advances in the study of *Peromyscus* (Rodentia). Texas Tech University Press, Lubbock, Texas.

Wong WSW, Yang Z, Goldman N, and Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics **168**:1041-1051.

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

Yang Z. 2000. Phylogenetic Analysis by Maximum Likelihood (PAML). University College, London.

Yang Z, and Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. **19**:908-917.

Yang Z, Nielsen R, Goldman N, and Pedersen AK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431-449.

Yang Z, Wong WSW, and Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. **22**:1107 - 1118.

Zhang J. 2003. Parallel functional changes in the digestive RNases of ruminants and colobines by divergent amino acid substitutions. Mol. Biol. Evol. **20**:1310-1317.

Chapter 2: Reproductive protein evolution within and between species: ZP3 sequence variation in *Peromyscus truei* and *P. gratus*

Abstract

Proteins involved in reproduction evolve rapidly in a variety of animal taxa. Most studies of reproductive protein evolution, however, focus on divergence between species. ZP3 is an egg coat protein involved in primary binding of egg and sperm and is essential for fertilization. Rapid evolution of ZP3 and evidence for positive selection have been shown previously in divergent mammals and in deer mice (*Peromyscus*). Here we investigate the molecular population genetics of ZP3 *within* two *Peromyscus* species. We do not find evidence of directional selection in the form of recent selective sweeps in either species. Amino acid polymorphism in ZP3 is high relative to silent polymorphism in both species; this pattern is consistent with balancing selection due to sexual conflict or to pathogen resistance, although we cannot rule out the formal possibility that these patterns result from relaxed levels of purifying selection on amino acid substitutions. Allelic variation in *P. truei* is similar between populations and across great geographic distances, thus we find no evidence for a role of reinforcement in promoting diversification of ZP3. While additional data are needed to identify the precise evolutionary forces responsible for patterns of sequence variation in ZP3 in *Peromyscus*, these results suggest that selection to maintain divergent alleles within species contributes to the previously identified

pattern of rapid amino acid substitution among divergent *Peromyscus* species in the sperm-combining region.

Introduction

One of the most striking genetic patterns seen in animals is that proteins involved in reproduction often evolve much more rapidly than do non-reproductive proteins (Singh and Kulathinal 2000; Swanson and Vacquier 2002b; Clark, Aagaard, and Swanson 2006). In mammals, evidence for rapid reproductive protein evolution is based largely on comparisons of widely divergent species (Queralt et al. 1995; Swanson et al. 2001; Torgerson, Kulathinal, and Singh 2002; Swanson, Nielsen, and Yang 2003; Glassey and Civetta 2004) or on comparisons of diverse species within orders (Retief et al. 1993; Wyckoff, Wang, and Wu 2000; Kingan, Tatar, and Rand 2003; Dorus et al. 2004; Clark and Swanson 2005; Podlaha et al. 2005). Only a few studies have focused on comparisons of closely related species (Jansa, Lundrigan, and Tucker 2003; Turner and Hoekstra 2006).

Intraspecific studies of reproductive protein variation have been limited to invertebrates. In these studies, patterns of genetic variation of fertilization proteins differ among taxa; in some cases (e.g. sea urchin bindin, Metz and Palumbi 1996; Drosophila accessory gland proteins, Begun et al. 2000; Wagstaff and Begun 2005) amino acid variation is extensive, while in others (e.g. abalone lysin, Lee, Ota, and Vacquier 1995) variation is limited or absent. In many cases, population-genetic data support a role for natural selection in driving the evolution of fertilization proteins, with directional selection operating in some instances and balancing selection in others. For example, there is evidence for recent positive selection on seminal proteins (accessory gland proteins, Acps) in several species of *Drosophila* (Begun et al. 2000; Wagstaff and Begun 2005) and on female reproductive tract proteins in *D. melanogaster* (Panhuis and Swanson 2006).

We know much less, however, about the patterns of variation in reproductive proteins within mammalian species, or about the evolutionary forces affecting these patterns. At present, surveys of intraspecific variation are limited to two sperm proteins in humans and in both cases, the spatial distribution of variation in these proteins suggests they may be subject to balancing selection as well as directional selection (Gasper and Swanson 2006; Hamm et al. 2007).

Proteins directly involved in gamete interaction are likely targets of natural and sexual selection. In mammals, the initial binding of sperm to the zona pellucida, or egg coat, is thought to be the critical recognition interaction between egg and sperm (Wassarman, Jovine, and Litscher 2001). The sperm protein(s) involved in this interaction have not been conclusively identified, but the egg protein ZP3 (Zona pellucida 3) is essential at this step (Wassarman, Jovine, and Litscher 2001). *In vitro* binding assays indicate that interaction with sperm occurs in a small region of ZP3 known as the “sperm-combining region” (Kinloch, Sakai, and Wassarman 1995). In the laboratory mouse, amino acid substitutions at critical sites in the sperm-combining region (Chen, Litscher, and Wassarman 1998) or replacement of the region with hamster sequence (Williams *et al.* 2006) reduces the affinity of sperm binding. ZP3 was long thought to be the single primary sperm-receptor protein, but a recent alternative model of egg-sperm interaction was proposed based on evidence from transgenic studies (Rankin *et al.* 2003; Dean 2004). In this model, specificity of binding is based on the three-dimensional structure of the zona pellucida, a matrix composed of ZP3 plus two additional proteins, ZP1 and ZP2.

Comparative sequence analysis of mammalian ZP3 shows evidence of positive selection. Moreover, a cluster of positively selected amino acid sites was identified in the sperm-combining region (Swanson *et al.* 2001). In previous work, we analyzed patterns of genetic variation of ZP3 in 15 species of deer mice (genus Peromyscus, Turner and Hoekstra 2006). We found evidence that ZP3 experienced positive selection during diversification of the genus. Further, despite small sample sizes, we observed substantial amino acid variation *within* some species.

Here, we expand our previous results by characterizing patterns of ZP3 sequence variation within a species that showed substantial amino acid variation in our earlier study. We sequenced regions of ZP3 and two non-reproductive genes in a large and geographically diverse sample of *Peromyscus truei* (the pinyon mouse) and one population of its sister species, *P. gratus* (Osgood's mouse). We find evidence for natural selection maintaining divergent ZP3 alleles in *P. truei*.

Materials and Methods

Extraction, amplification and sequencing

Tissues samples (liver, kidney, or tail) from 46 *Peromyscus truei* individuals from 14 localities and 10 *P. gratus* individuals from a single locale were obtained from museums or from our field collections (Table 1, Fig. 1a). Samples of *P. gratus* were identified as *P. truei* in our previous work (Turner and Hoekstra 2006), based on sample information from the museum and in agreement with the classification of *P. truei* by Hoffmeister (1951). However, Durish and colleagues (2004), show that northern populations and southern populations previously identified as *P. t. gratus* form reciprocally monophyletic clades for *Cytb*, in agreement with previously proposed elevation of *P. t. gratus* to species status based on karyotype data (Modi and Lee 1984). We find that *P. t. gratus* is reciprocally monophyletic with the rest of the *P. truei* samples for two genes (see below), and thus we change the designation of those samples to *P. gratus gentilis* here.

We extracted DNA from tissue using a DNeasy kit (Qiagen, Valencia, CA). A fragment containing exons 6 – 7 of *Zp3* (358 bp), which contains the region essential for sperm binding (Wassarman, Jovine, and Litscher 2001), was amplified and sequenced for all individuals using previously published primers and conditions (Turner and Hoekstra 2006). In addition, to control for demographic effects, we sequenced regions of the mitochondrial gene *Cytb* and the non-reproductive nuclear gene *Lcat* to look for evidence of population structure. Together, these two genes will provide *Cytb* should be more sensitive to population structure than an autosomal locus (because of its lower N_e). Like *Zp3*, *Lcat* is an autosomal locus, and thus the effects of demographic events on genetic variation in *Lcat* are expected to be more comparable to *Zp3*. *Lcat* is an enzyme involved in glycerophospholipid metabolism (Kuivenhoven *et al.* 1997) located on *Mus* chromosome 8 and thus not likely to be linked to *Zp3* (*Mus* chromosome 5). Partial *Cytb* sequences for many individuals were generously provided by J.L. Patton. For remaining samples, we amplified a 1,055 bp fragment using primers tCytbF1 (5'- CGA CCT CCC AAC TCC ATC CAA C-3') and tCytbR1 (5'- TGC CTG CCA TAG GTA TTA GGA C-3'). We sequenced two regions of *Lcat*; fragment one, containing exons 2 – 5 (664 bp), was amplified using conserved primers LCAT2FA (5'-ACA GAG GAC TTC TTC ACC ATC-3') and LCAT5RA (5'-AAT AGA GCA CAT GTA GGC AGC-3'). Fragment two, containing most of exon 6 (487 bp), was amplified as described previously (Turner and Hoekstra 2006) using published primers (Robinson *et al.* 1997).

We purified PCR products using a MinElute Gel Extraction Kit (Qiagen), a PerfectPrep PCR cleanup 96 kit (Eppendorf, Westbury, NY) or treatment with exonuclease I and shrimp alkaline phosphatase (Fermentas, Hanover, MD). We performed cycle sequencing using BigDye 3.1 chemistry (Applied Biosystems, Foster City, CA) and ran those products on an ABI 3100 genetic analyzer (Applied Biosystems). We aligned the sequences using SEQUENCHER (v. 4.2, Gene Codes, Ann Harbor, MI). If more than one heterozygous site was detected for nuclear genes, we determined phase by sequencing cloned products (TOPO-TA, Invitrogen) or inferred phase computationally using PHASE (v 2.1.1, Stephens, Smith, and Donnelly 2001; Stephens and Donnelly 2003).

Sequence analysis

To compare levels of nucleotide diversity within and among populations and species, we estimated summary statistics of polymorphism data using ANALYSERHKA (v 7, Haddrill *et al.* 2005). We calculated nucleotide diversity for silent sites, replacement sites and all sites. Diversity measures include π (Tajima 1983), which is based on the average pairwise difference between sequences and θ_W (Watterson 1975), which is based on the number of segregating sites. We excluded sites with alignment gaps or more than two variants within a species from further analysis.

To test for deviations from neutral evolution, we applied two statistical tests based on population parameters estimated from polymorphism data; both analyses test for a skew in the frequency distribution of polymorphisms, which indicates deviation from neutrality. Tajima's *D* test measures the difference between π and θ_W; negative values indicate an excess of rare mutations, a pattern consistent with a selective sweep, and positive values indicate an excess of high frequency mutations, consistent with balancing selection (Tajima 1989). Fay and Wu's *H* test measures the difference between π and θ_H, a measure of nucleotide diversity that gives greater weight to high frequency polymorphisms (Fay and Wu 2000). Negative values of *H* indicate an excess of high frequency derived mutations, consistent with hitchhiking. Tajima's *D* and Fay and Wu's *H* were implemented using ANALYSERHKA. We determined significance by comparing observed values to the results of 10,000 coalescent simulations performed in MS (Hudson 2002) under neutral conditions given the observed θ_W. For nuclear genes, we estimated the population recombination parameter (ρ; (Hudson 2001) and ran simulations with this empirical estimate of ρ and four additional ρ values such that ρ/θ_W equaled 0, 1, 10, and 50.

The additional values were included to determine the sensitivity of the results of *D* and *H* to variation in recombination rate; they represent a large range of recombination, from none up to values that are much greater than values in *Drosophila*, which has much higher levels of recombination than mammals.

We employed the McDonald-Kreitman (MK) test to look for evidence of balancing or directional selection on *Zp3* in *P. truei* and *P. gratus*. The MK test is based on the expectation under neutrality that the ratio of replacement to silent polymorphism should equal the ratio of replacement to silent divergence. A statistically significant excess of replacement divergent sites is interpreted as evidence for directional selecthion while excess replacement polymorphism is expected if a gene is subject to balancing selection (McDonald and Kreitman 1991).

To assess population structure within and between populations for each of the three genes, we used the program STRUCTURE, which clusters individuals based on their genotypes (v 2.2, Pritchard, Stephens, and Donnelly 2000). Analyses were run for 100,000 cycles and the first 30,000 cycles were discarded as burn-in. Analyses were run with increasing number of populations (*K*) until the likelihood value plateaued (likelihoods were similar for ≥ 3 values of *K*). For *P. truei,* although likelihood values plateaued at a maximum of $K = 3$, analyses for each gene were run with *K* values up to 6 because that is the number of geographic regions sampled. For *P. gratus*, analyses were run for $K = 1 – 4$ because likelihood values plateaued at a maximum of $K = 2$. Haplotype networks were constructed for each gene using TCS (v 1.21, Clement, Posada, and Crandall 2000) to visually represent the extent and type (i.e. nonsynonymous vs. synonymous) of variation between alleles.

Results

Population structure

Each of the three genes shows a unique pattern of intraspecific variation (Table 2, Fig. 2). Both *Lcat* and *Cytb* have fixed differences between *P. truei* and *P. gratus*, while *Zp3* does not. There is strong geographic structure in *Cytb* in *P. truei* (Fig. 1b,c); the species comprises two highly diverged clades ($D_{xy} = 0.049$), representing eastern and western populations. These mitochondrial clades meet in northern Arizona, where individuals representing both *Cytb* groups are found in close proximity. The location of the transition between clades does not coincide with any reported boundary between subspecies or other previously described geographic patterns of variation in *P. truei* (Hoffmeister 1951).

In contrast to the subdivision in *Cytb*, nucleotide variation in both nuclear genes has little or no geographic structure: both clusters identified by STRUCTURE analysis are represented in all *P. truei* populations sampled (Fig. 1b). If allele frequencies at *Zp3* differed substantially between the two divergent *Cytb* groups, we might infer that the alleles of the two *Zp3* clusters evolved in isolated populations and that the variation within clades resulted from recent neutral introgression after secondary contact. The relatively uniform distribution of genetic clusters between clades (Fig. 1b,d) and across great geographic distances indicates that either amino acid variation predates the mitochondrial split or that alleles spread rapidly in both directions due to selection following secondary contact.

Tests of neutrality

To test for deviations from neutral evolution, we applied two statistical tests based on population parameters estimated from polymorphism data (Table 3); both analyses test for a skew in the frequency distribution of polymorphism. Both Tajima's *D* and Fay and Wu's *H* are influenced by demography in different ways, thus significant results are difficult to interpret in the absence of knowledge of the demographic history of sampled populations (Haddrill *et al.* 2005). Unfortunately, as is the case with most species, we do not know the precise demographic history of *P. truei* or *P. gratus*. We control in part for demography by comparing patterns between loci; demographic events affect the entire genome in a similar way, while the signature of selection is limited to the selected site and sites in close proximity. However, any significant results must still be treated with caution, as demographic factors can also increase the variance in *D* and *H* across the genome (Nielsen 2001).

Neither the *D* nor *H* statistic was significantly extreme in the positive or negative direction for any class of sites of *Cytb* or *Zp3* in *P. truei* or *P. gratus* (Table 3), based on comparison of observed values to neutral coalescent simulations. For *Cytb*, both statistics were predominantly negative in *P. truei* and predominantly positive in *P. gratus*. Similarly, *D* and *H* values for all classes of sites of *Zp3* were negative for *P. truei* and all were positive for *P. gratus*. The consistently negative values of both statistics in *P. truei* result from a few low frequency polymorphisms (Fig. 2). In *P. gratus*, *D* and *H* are positive because most polymorphisms are at intermediate frequencies. For *Lcat*, no *D* or *H* values were significantly extreme in *P. gratus*, but in *P. truei,* Tajima's *D* was negative and significantly different from neutral expectations for replacement sites (D = -1.31, P = 0.035) and Fay and Wu's *H* was significantly negative for silent sites (H = -8.03, P = 0.009) and for all sites (H = -7.97, P = 0.012). The mixed results of the two tests and marginally significant *P* values for Tajima's *D* suggest that evidence for directional selection acting on *Lcat* is weak at best. There are neither high frequency replacement polymorphisms within *P. truei* nor fixed replacement differences between *P. truei* and *P. gratus* in the sequenced region of *Lcat*. Thus, any selection on *Lcat* has been on either a linked coding site that was not sequenced here or on a non-coding regulatory change.

Polymorphism in Zp3

Levels of replacement polymorphism are much higher in *Zp3* than in *Cytb* or *Lcat* (except when *Cytb* samples from the highly divergent eastern and western clades of *P. truei* are combined). In *P. gratus*, measures of nucleotide diversity for all classes of sites in *Zp3* are higher than for *P. truei* and for the nuclear gene *Lcat*; however, ratios of replacement to silent diversity are more than 10 times larger than the ratios in non-reproductive genes (θ_R/θ_S = 0.302, π_R/π_S = 0.311; Table 2). In *P. truei*, diversity at silent sites is slightly lower in *Zp3* than in the non-reproductive genes, but ratios of replacement to silent diversity observed are 6 – 40 times greater ($\theta_R/\theta_S = 0.750$, $\pi_R/\pi_S = 1.360$).

High levels of amino acid polymorphism can result from balancing selection that maintains divergent alleles or from relaxed levels of purifying selection. The McDonald-Kreitman (MK) test is commonly used to distinguish between these alternatives (McDonald and Kreitman 1991). However, the MK test lacks power to detect selection maintaining divergent alleles within species when both divergence and polymorphism are elevated at replacement sites relative to silent sites. We know that amino acid divergence is high between *P. truei* and *P. boylii* based on previous results; some branches along the lineage separating these species have elevated replacement substitution rates in *Zp3* relative to expected based on a simulated null distribution (Turner and Hoekstra 2006). MK tests do not indicate significant excess of replacement polymorphism or of replacement divergence for either species vs. an outgroup species (*P. boylii*). It is difficult to interpret this result because lack of significance in either direction could be explained in two ways. First, patterns of polymorphism within species may be the result of neutral processes, but small sample size prevented detection of the excess replacement divergence evident from our previous work. Second, both fixed and polymorphic amino acid variation may be favored, thus no difference is detected because both ratios reflect non-neutral processes. Thus, we cannot distinguish between balancing or relaxed selection on *Zp3* in *P. truei* or *P. gratus* from these sequence data alone.

Geographic distribution of allelic diversity in Zp3

To test for geographic variation in *Zp3* allelic diversity, we compared allele frequencies between populations of *P. truei* distributed across the species range (Fig. 2a). (We cannot describe geographic variation in *Zp3* in *P. gratus* because we sampled only a single population). Levels of allelic variation are similar across the species range (θ_R = 0.002 – 0.004). With the exception of a locale where only a single individual was sampled (TX), at least two haplotypes, differing by at least one amino acid site, were represented at each locale. In addition, there is haplotype sharing between populations across large geographic distances (Fig. 2d).

The sampled populations include localities both where *P. truei* is sympatric with other members of the *truei* species group (AZ-W, AZ-E, TX, Fig. 2a) and also localities where *P. truei* is the only species group member present (allopatric populations: CA-N, CA-S, UT) (Durish *et al.* 2004). Since allelic composition and levels of amino acid variation are similar between sympatric and allopatric populations, there is no evidence of increased divergence of *Zp3* between *P. truei* and sibling species in sympatric populations, as would be expected if reinforcement promoted divergence of *Zp3* in *P. truei*.

Discussion

Our survey of intraspecific variation of the female fertilization protein ZP3 reveals several intriguing patterns. While we do not find evidence of directional selection in the form of recent selective sweeps in either *P. truei* or *P. gratus*, we do find high ratios of replacement to silent polymorphism in *Zp3* relative to the ratios seen for two non-reproductive genes. This pattern is consistent with the maintenance of divergent alleles due to sexual conflict (Gavrilets and Waxman 2002) or to pathogen resistance (Roy and Kirchner 2000), although we cannot rule out the formal possibility of relaxed purifying selection. However, populations of *P. truei* sympatric with other members of the *truei* species group do not have reduced levels of polymorphism, derived substitutions, or unique haplotypes compared to allopatric populations. This result allows us to rule out the possibility that reinforcement is the driving force promoting ZP3 diversification.

Zp3 shows no evidence of selective sweeps

Simple population-genetic tests of neutrality in *Zp3* lack significance, indicating that recent selective sweeps acting on the sperm-combining region are unlikely. However, we cannot rule out the possibility of a sweep in *Zp3* at a more distant point in these species' histories. The signal detected by Tajima's *D* and Fay and Wu's *H* degrades rapidly with physical distance from the selected site and as time elapses after a sweep (Simonsen, Churchill, and Aquadro 1995; Przeworski 2002). Further, both tests have low power to detect "soft sweeps," that is, when selection is weak or acts on standing genetic variation rather than on a single new mutation (Hermisson and Pennings 2005). Thus, we can conclude only that if directional selection has acted on Zp3, it was neither strong nor recent.

Zp3 shows high levels of amino acid polymorphism

The importance of ZP3's role in fertilization leads us to believe that balancing selection rather than relaxed purifying selection might explain the patterns of nonsynonymous variation we observe in *P. truei* and *P. gratus*. ZP3 has sperm-binding activity in several divergent mammalian species, including hamster, human, pig and laboratory mouse (reviewed in McLeskey *et al.* 1997). In the laboratory mouse, amino acid mutations in the sperm-combining site (encompassed in the region sequenced here) inactivate ZP3 as a sperm receptor (Chen, Litscher, and Wassarman 1998). It seems unlikely that selection would be relaxed on a region of such functional importance. Further, evidence of positive selection has been documented in this region both within *Peromyscus* (Turner and Hoekstra 2006) and in a phylogenetically diverse sample of mammals (Swanson *et al.* 2001).

We need additional data to determine if patterns of variation observed in ZP3 are caused by relaxed purifying selection or selection to maintain divergent alleles. Demonstrating fitness consequences of replacement polymorphism would strongly implicate selection. Ideally, this would entail comparing the reproductive success of females having different *Zp3* genotypes in natural populations. However, evidence that variation in polymorphic sites of ZP3 affects female fertility in captive crosses or *in vitro* fertilization assays would strongly imply polymorphism contributes to fitness differences in nature. Alternatively, evidence of coevolution of ZP3 with its interacting sperm protein(s) would also confirm ZP3 has been the target of selection. As mentioned above, the sperm protein(s) that interact with ZP3 have not yet been conclusively identified, but there are some good candidates (e.g., zonadhesin, PKDREJ), thus preliminary assessment of the evidence for coevolution is possible by looking for correlated changes between amino acid sites in ZP3 and sperm genes.

Levels of allelic diversity in* Zp3 *are similar between populations of P. truei

The geographic distribution of allelic variation in *Zp3* in *P. truei* is inconsistent with reinforcement. For reinforcement to occur, diverging populations must hybridize in nature and experience incomplete postzygotic isolation (Coyne and Orr 2004). Hybridization of *P. truei* with sibling species has not been reported in nature (Dice 1968), however, laboratory crosses provide evidence of some postzygotic isolation between *P. truei* and sympatric populations of *P. difficilis* (sometimes referred to as *P. nasutus*). Crosses follow Haldane's rule: that is, hybrid female offspring are fertile and hybrid male offspring are partially or completely sterile (Haldane 1922; Dice 1968). Thus, although there are insufficient data to determine whether there has been ample opportunity for reinforcement in natural populations of *P. truei*, our results are inconsistent with reinforcement. It should be noted that the lack of a signature of reinforcement in *Zp3* is not evidence against reinforcement contributing to the evolution of reproductive isolation between *P. truei* and its sibling species. Reinforcing selection may have acted on pre-mating isolation (such as assortative mating) or on another male-female reproductive protein interaction, so that there may be no opportunity for selection to cause increased protein divergence at the gametic level (Lorch and Servedio 2005).

Further, we cannot describe geographic variation in *P. gratus* since we sampled a single population, but that site has very high diversity at *Zp3*, including five haplotypes differing at the amino acid sequence level among 20 sampled alleles. Additional geographic sampling of *P. gratus* will reveal whether this intriguing pattern is consistent among populations or if there is variation, perhaps due to reinforcement.

Conclusions

We conclude that while patterns of genetic variation in *Zp3* in *P. truei* and *P. gratus* are not consistent with strong purifying selection, with recent selective sweeps, or with reinforcement, we observe high levels of replacement polymorphism, a pattern that can result from selection acting to maintain divergent alleles. This type of selection is consistent with predictions of theories involving sexual conflict (Frank 2000; Gavrilets and Waxman 2002; Haygood 2004) as well as defense against pathogens (Roy and Kirchner 2000), two hypotheses that have been proposed to explain rapid rates of evolution of reproductive proteins (Swanson and Vacquier 2002a). However, additional data measuring differences between alleles in function or fitness are needed to (1) rule out the formal possibility of relaxed purifying selection, (2) identify the specific amino acid sites that are targets of selection, and (3) to determine the precise selective agent promoting ZP3 diversification.

Understanding the molecular interactions between ZP3 and sperm protein(s) will help us determine whether the patterns of polymorphism we observe reflect fitness differences among genotypes in nature. There has been recent progress in functionally characterizing two promising candidate proteins that interact with ZP3: zonadhesin (Bi *et al.* 2003) and PKDREJ (Sutton *et al.* 2006). Further, evolutionary analysis of both proteins within and between species of primates has identified promising regions of the proteins that may interact and coevolve with ZP3 (Gasper and Swanson 2006; Hamm et al. 2007).

Results of our intraspecific analysis provide insight into the pattern of repeated amino acid substitution in ZP3 that we first identified in our interspecific analysis and that provided evidence for positive selection (Turner and Hoekstra 2006). The lack of evidence for directional selection in both *P. truei* and *P. gratus* suggests that the interspecific pattern may not be the result of repeated selective sweeps within species, causing rapid turnover in alleles. Instead, at least in some species, selection may favor two or more alleles simultaneously; consequently, newly arising alleles may increase in frequency but not replace older alleles. An extreme example of this type of selection is self-incompatibility alleles in several families of plants; positive selection promotes amino acid divergence yet high levels of allelic variation are maintained through frequency dependent selection for such long periods of time that there is shared polymorphism between species of different genera (Vekemans and Slatkin 1994; Ishimizu et al. 1998; Takebayashi et al. 2003). In addition to the evidence required to determine the type of selection acting on *Zp3* in *P. truei* and *P. gratus*, data about the extent and distribution of amino acid variation within additional species of *Peromyscus* are necessary to evaluate whether there is widespread selection maintaining diversity in ZP3.

Lack of geographic structure of allelic variation in the sperm-combining region of ZP3 in *P. truei* suggests that this protein likely does not contribute to any current barriers to gene flow between populations. Measuring sequence variation in other sibling species and assessing functional differences between alleles will be an exciting next step in evaluating whether ZP3 may have contributed to the evolution or maintenance of isolating barriers between species.

Acknowledgements

We thank members of the Hoekstra Laboratory, P. Andolfatto, B. Payseur, J. Coyne, J. Jensen, J. Kohn and V. Vacquier for helpful discussion and comments on the manuscript. J. Patton generously provided help in the field, advice, and sequence data. J. Weber assisted with fieldwork. The Museum of Texas Tech University and the Museum of Vertebrate Zoology (University of California, Berkeley) kindly provided tissue samples. This research was supported by a Howard Hughes Medical Institute Predoctoral Fellowship and a University of California Reserve System Mildred Mathias Award (to LMT), National Science Foundation Doctoral Dissertation Improvement Grant DEB 0608030 (to LMT and HEH), and National Science Foundation grant DEB 0344710 (to HEH).

Chapter 2 is in preparation for publication: Turner LM, and Hoekstra HE. In Preparation. Reproductive protein evolution within and between species: ZP3 sequence variation in *Peromyscus truei* and *P. gratus*

Table 2.1 Samples of *Peromyscus* included in this study

Subspecies	Sampling location	*n*	Specimen numbers
P. t. montipinoris	Walker Pass, Kern Co., CA	3	LMT010 – 012[1]
P. t. truei	Williams Butte, Mono Co., CA	13	LMT007 – 009[1]; JNW007, 009, 011-012[1]; MVZ 208477 – 208481[2]
P. t. truei	Yavapai Co., AZ	1	TTU TK113804[3]
P. t. truei	Ryan, Coconino Co., AZ	2	MVZ 197284 – 5[2]
P. t. truei	Kaibab Plateau, Coconino Co., AZ	2	MVZ 197293 – 4[2]
P. t. truei	Tanner Tank, Coconino Co., AZ	2	MVZ 197295 – 6[2]
P. t. truei	Woodhouse Mesa, Coconino Co., AZ	12	MVZ 199469 – 80[2]
P. t. truei	Toroweap Valley, Mohave Co., AZ	1	MVZ 199467[2]
P. t. truei	Hack Canyon, Mohave Co., AZ	1	MVZ 199468[2]
P. t. truei	Little Colorado River, Coconino Co., AZ	1	MVZ 199481[2]
P. t. truei	Onion Creek, Grand Co., UT	1	MVZ 199482[2]
P. t. truei	Rock Canyon Corral, Grand Co., UT	3	MVZ 199483 – 5[2]
P. t. truei	0.5 mi E Rock Canyon Corral, Grand Co., UT	3	MVZ 199486 -8[2]
P. t. comanche	Armstrong Co., TX	1	TTU TK40209[3]
P. gratus gentilis[4]	Durango, Mexico	10	TTU TK48798, TK48826, TK48834, TK48839, TK48854, TK48856, TK48892 – 3, TK48899, TK48911[3]

[1] Hoekstra laboratory

[2] Museum of Vertebrate Zoology, University of California, Berkeley

[3] Museum of Texas Tech University

[4] Samples from this population were identified as *P. truei* in our previous work (Turner and Hoekstra 2006).

Table 2.2 DNA polymorphism and divergence.

				Silent						Replacement					
Locus	Species	Haplotypes	*n*	No. Sites	*S*	θ_W	π	D_{xy}	Fixed	No. Sites	*S*	θ_W	π	D_{xy}	Fixed
Cytb	*P. truei* (all)	31	46	236	72	0.0705	0.1019	0.380	53	724	10	0.0031	0.0034	0.007	3
	P. truei (W)	21	25	235	32	0.0349	0.0237	0.181	31	725	8	0.0029	0.0021	0.005	2
	P. truei (E)	10	21	237	9	0.0106	0.0097			723	1	0.0004	0.0002		
	P. gratus	8	10	239	20	0.0311	0.0327			721	2	0.0010	0.0008		
Lcat	*P. truei*	23	88	418	14	0.0066	0.0040	0.028	3	495	2	0.0008	0.0001	$7x10^{-5}$	0
	P. gratus	13	14	445	10	0.0064	0.0066			578	0	0	0		
Zp3	*P. truei*	10	92	177	5	0.0044	0.0020	0.034	0	176	3	0.0034	0.0026	0.017	0
	P. gratus	10	20	174	13	0.0211	0.0257			177	4	0.0064	0.0080		

		Total							
Locus	Species	No. Sites	*S*	θ_W	π	D_{xy}	Fixed	θ_{WR}/θ_{WS}	π_R/π_S
Cytb	*P. truei* (all)	964	82	0.0196	0.0275	0.099	56	0.042	0.033
	P. truei (W)	964	40	0.0107	0.0074	0.049	33	0.084	0.090
	P. truei (E)	964	10	0.0029	0.0026			0.036	0.026
	P. gratus	964	22	0.0081	0.0087			0.032	0.023
Lcat	*P. truei*	917	16	0.0035	0.0019	0.013	3	0.120	0.034
	P. gratus	1027	10	0.0028	0.0029			0	0
Zp3	*P. truei*	354	8	0.0039	0.0023	0.025	0	0.750	1.360
	P. gratus	352	17	0.0136	0.0167			0.302	0.311

Table 2.2 DNA polymorphism and divergence, continued.
n – number of alleles sampled; *S* – segregating sites; θ_W – Watterson's theta (per site); π – pairwise differences (per site); D_{xy} – average number of nucleotide substitutions (per site) between *P. truei* and *P. gratus* or, for *P. truei* (W) with *P. truei* (E); Fixed differences – number of fixed differences, comparisons as for D_{xy}; θ_{WR} – $\theta_{W,}$ replacement sites; θ_{WS} – $\theta_{W,}$ silent sites; π_R– π, replacement sites; π_S – π, silent sites

Table 2.3 Population-genetic tests of neutrality

Locus	Species	Silent		Replacement		Total	
		Tajima's *D* (*P*)	Fay & Wu's *H* (*P*)	Tajima's *D* (*P*)	Fay & Wu's *H* (*P*)	Tajima's *D* (*P*)	Fay & Wu's *H* (*P*)
Cytb	*P. truei* (all)	1.68 (0.971)	2.86 (0.237)	0.012 (0.572)	-2.66 (0.080)	1.51 (0.952)	0.16 (0.342)
	P. truei (W)	-1.20 (0.104)	2.41 (0.650)	-0.86 (0.218)	-2.32 (0.079)	-1.17 (0.112)	0.087 (0.328)
	P. truei (E)	-0.27 (0.443)	-1.09 (0.175)	-0.62 (0.189)	0.16 (0.669)	-0.36 (0.390)	-0.93 (0.189)
	P. gratus	0.38 (0.687)	3.29 (0.891)	-0.69 (0.244)	0.44 (0.771)	0.37 (0.680)	3.56 (0.893)
Lcat	*P. truei*	-1.10 (0.117	-8.03 **(0.009)**	-1.31 **(0.035)**	0.07 (0.468)	-1.28 (0.076)	-7.97 **(0.012)**
	P. gratus	0.17 (0.600)	1.45 (0.856)	0	0	0.17 (0.600)	1.45 (0.856)
Zp3	*P. truei*	-1.12 (0.117)	-1.11 (0.092)	-0.37 (0.399)	-0.78 (0.107)	-0.96 (0.173)	-1.88 (0.066)
	P. gratus	0.79 (0.826)	2.61 (0.964)	0.74 (0.786)	0.83 (0.897)	0.85 (0.843)	3.44 (0.976)

P = Significance determined by comparison of observed values to the results of 10,000 coalescent simulations performed in MS (Hudson 2002) under neutral conditions given the observed θ_W values for simulations with estimated ρ are shown; all tests significant with estimated ρ were also significant with $\rho = 0$. Significant *P* values (0.05 level) are in bold

Figure 2.1 Geographic distribution of sequence variation

(a) Ranges of *P. truei* (light gray) and *P. gratus* (dark gray); overlapping region is cross-hatched (Durish *et al.* 2004). Sample locations are indicated with filled squares, color indicates site grouping: red – northern California (CA-N), yellow – southern California (CA-S), purple – Arizona, western *Cytb* clade (AZ-W), pink – Arizona, eastern *Cytb* clade (AZ-E), green – Utah (UT), orange – Texas (TX), turquoise – Durango, Mexico (*P. gratus*). (b) Genotypic clustering within species for *Cytb*, *Lcat*, and *Zp3*; *Kt* = number of populations for *P. truei* and *Kg* = number of populations for *P. gratus*; each vertical bar represents a single individual; black lines separate geographic regions of sites. Haplotype networks for (c) *Cytb* and (d) *Zp3*; each circle represents a unique haplotype; the size of the circle is proportional to the number of alleles with the haplotype, with the exception of the most common allele of *Zp3* which is ¼ the scale of other alleles; color indicates geographic origins of sampled alleles (as in 1a); each tick mark represents a nucleotide difference, black marks indicate silent differences and red marks indicate nonsynonymous differences.

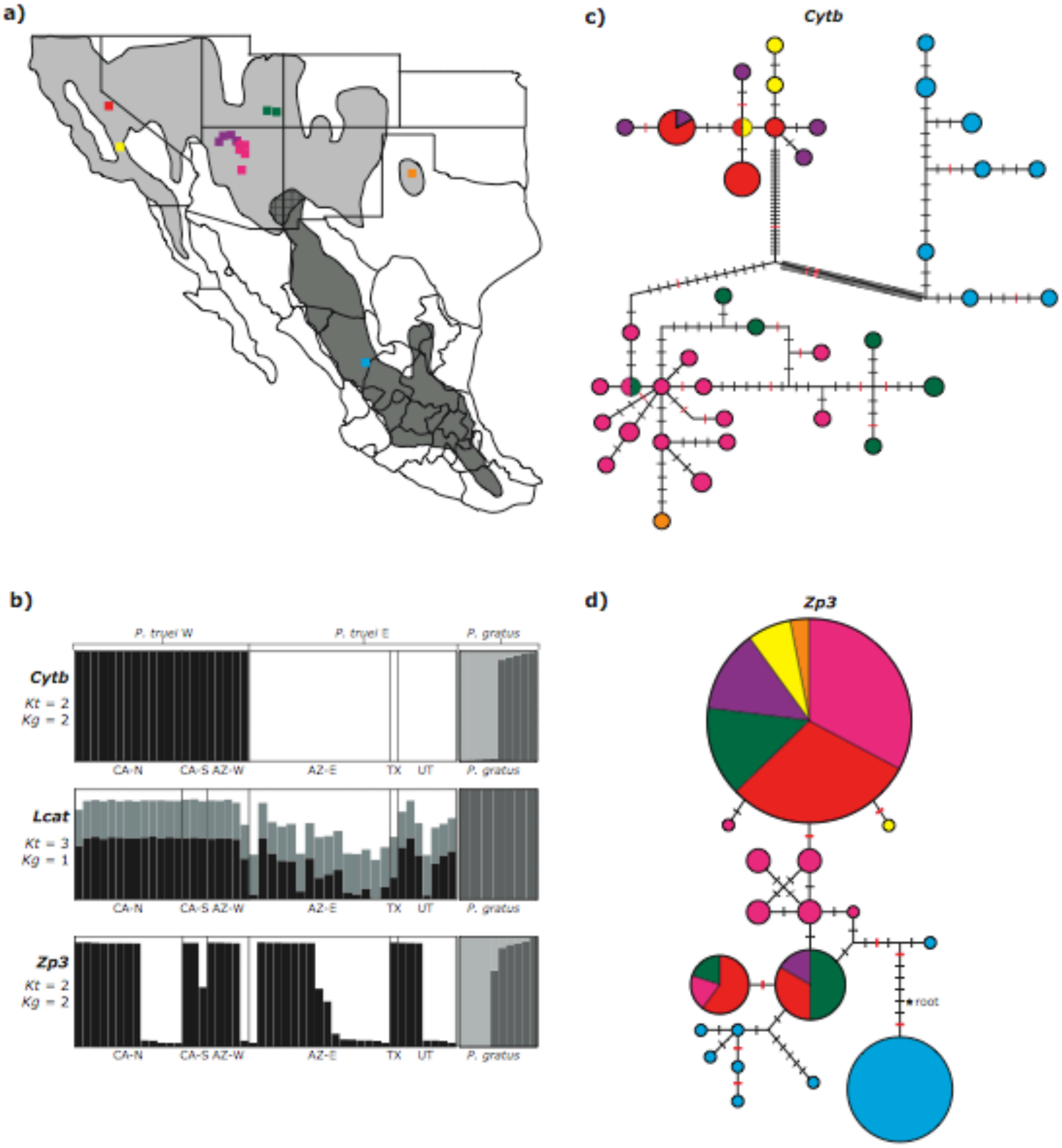

a)
c)
Cytb
b)
P. truei W
P. truei E
P. gratus
Cytb
Kt = 2
Kg = 2
CA-N
CA-S AZ-W
AZ-E
TX
UT
P. gratus
Lcat
Kt = 3
Kg = 1
CA-N
CA-S AZ-W
AZ-E
TX
UT
P. gratus
Zp3
Kt = 2
Kg = 2
CA-N
CA-S AZ-W
AZ-E
TX
UT
P. gratus
d)
Zp3
root

Figure 2.2 Tables of polymorphism
Polymorphic sites in (a) *Cytb*, (b) *Lcat*, and (c) *Zp3* in *P. truei* and *P. gratus*. Exon/intron structure of each gene, drawn to scale, is shown above each alignment. Boxes indicate exons; black fill for exons and bold lines for introns indicate region(s) we sequenced here. For each gene, all unique alleles found within each sampling location are shown (location names as in Fig. 1). Due to the large number of variable sites in *Cytb*, we show separate tables of polymorphism for *P. truei* and *P. gratus*. Consensus sequences are given as tCon and gCon, respectively. For *Lcat* and *Zp3* all sites that are polymorphic in either species are shown for each. Ancestral (Anc) sequences were inferred using parsimony through comparison to an outgroup sequence (*P. boylii*). Below each alignment we indicate the type of nucleotide substitution (I = intron, S = synonymous, N = nonsynonymous), the starting and ending amino acid, the amino acid position, the starting and ending amino acid type (+ = positively charged, - = negatively charged, P = polar, N = non-polar), and the type of change [C = conservative (Grantham's distance <50), MC = moderately conservative (51-100), MR = moderately radical (101-150), R = radical (>150), (Grantham, 1974; Li, Wu, and Luo 1984)].

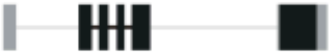
a) *Cytb*

1,144 bp

```
                                                                                    11111111
          1111122222233333333334444444455555555556666666667777788888889999999999999900000000
          2225501446800122244562445667901112344991123778990569901257880112344456679900012336
P. truei  1460379354039213729503149584810695739275840257393955873554059287025810100401570050
    tCon  TGTCACTTTCCCYTCCTTGTCATCTTYTRCARCCTCCTTCTATCCTACCTTTAGAGCATCATCTAAYATCATCTATTATGCG
CAN tru1  ....GTAC.T..T...C.AC...TC.TCG.GAT....CC....TTCGTTCCC....AGCT.CTC..TT.T.C...CC.....(1)
    tru2  ....GTAC.T..T...C.AC...TC.TCG.GAT....CC....TTCGTTCCC....AGCT.CTC..TT.T.CA..CC.....(5)
    tru3  ....GTAC.T..T...C.AC...TC.TCG.GAT....CC.....TCGTTCCC....AGCT.CTC..TT.T.C...CC.....(2)
    tru4  ....GTAC.T..T...CCAC...TC.TCG.GAT....CC....TTCGTTCCC..G.AGCT.CTC..TT.T.C...CC.....(5)

CAS tru1  ....GTAC.T..T...C.AC...TC.TCG.GAT....CC....TTCGTTCCC....AGCT.CTC..TT.T.C...CC.....(1)
    tru5  ....GTAC.T..T...C.AC...TC.TCG.GAT....CC.....TCGTTCCC....AGCT.CTCG.TT.T.C...CC.....(1)
    tru6  ....GTAC.T..T...C.AC...TC.TCG.GAT....CC.....TCGTTCCC...AAGCT.CTCG.TT.T.C...CC.....(1)

AZW tru4  ....GTAC.T..T...CCAC...TC.TCG.GAT....CC....TTCGTTCCC..G.AGCT.CTC..TT.T.C...CC.....(1)
    tru7  ....GTAC.T..T...CCAC...TC.TCG.GAT....CC....TTCGTTCCC..G.AGCT.CTC..TT.T.C...CC..C..(1)
    tru8  ....GTAC.T..TC..C.AC...TC.TCG.GAT....CC....TTCGTTCCC....AGCT.CTC..TT.T.C...CC..C..(1)
    tru9  ....GTAC.T..T...C.AC...TC.TCG.GAT....CC.....TCGTTCCCG...AGCT.CTC..TT.T.C...CC.....(1)
    tru10 ....GTAC.T..T...C.AC...TC.TCG.GAT..T.CC.....TCGTTCCC....AGCT.CTC..TT.T.C...CC.....(1)

AZE tru11 .A..........C..T..........C.A..G....T.....C.................T.....C.............T.(1)
    tru12 ...........TC.....A.......C.A..G........C.........................C.C.....G..G.CT.(1)
    tru13 ........C...C.............C.AT.G........C............A............C.C.....G..G.CT.(1)
    tru14 ............C.............C.A..G....T.....C.......................C............CT.(1)
    tru15 ............C.............C.A..G....T.....C.......................C...........C.T.(1)
    tru16 ............C............CC.AA.G....T.....C.......................C......C......T.(1)
    tru17 ..C.........C.............C.A..G....T....TC.......................C.............T.(2)
    tru18 C...........C.............C.A..G..C.T.....C.......................C.............TA(1)
    tru19 ............C.............T.A..A....T.....C.......................C...........C.T.(1)
    tru20 ............C.............C.A..G....T.....C.......................C.............T.(2)
    tru21 ..C.........C.............C.A..A....T....TC.......................C.............T.(1)
    tru22 ............C........G....C.A..G....T.....C.......................C.............T.(1)
    tru23 ............C.............C.A..G....T.....C.......................T......C......T.(2)
    tru24 ............C.............C.A..G....T..T..C.......................C...........C.T.(1)
    tru25 ............C.............C.A..G....T.....C.......................C.............T.(1)

TX  tru26 ...T......T.C.T...........T.A..G....T.....C.......................C.............T.(1)

UT  tru15 ............C.............C.A..G....T.....C.......................C...........C.T.(1)
    tru27 ........C...C.......A.C...C.AT.G........C........................GC.C.....G.....T.(1)
    tru28 .A......C...T.............C.AT.G........C........................GC.C.....G.......(2)
    tru30 ...........TC.............C.G..G........C.........................C.C.....G..G..T.(1)
    tru31 ...........TC.............C.G..G........C.........................C.C.....G.......(1)
    tru32 ........C...C.............C.AT.G.T....C.C........................GC.C.T...G.....T.(1)

    Type  SNSSSSSSNSSSSSSNSSNSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSNSSSSSSSSSSSSSSSNNSSSNSN

    aa     V      M      T  V                                       M               II   E A
           I      T      M  I                                       V               VT   Q T

    Pos    4      8      1  1                                       2               33   3 3
           2      2      0  1                                       9               33   4 5
                         8  7                                       2               44   4 4

    Type   N      N      P  N                                       N               NN   - N
           N      P      N  N                                       N               NP   P P

    Change C      M      M  C                                       C               CM   C M
                  C      C                                                           C     C
```

```
                           111111
           1223333455566799000000
           3460689204469602012255
P. gratus  2344646613663591570607
     gCon  ACTACTCGCCTTRTTTTRAYCT
     gra1  ............G....GGT..(2)
     gra2  ...G........A...CA.C.C(1)
     gra3  ....T...A...G....G.T..(2)
     gra4  .TC..CAT.ACCACCC.A.C..(1)
     gra5  .TC..CAA..CCA..C.A.C..(1)
     gra6  G...........G....GGT..(1)
     gra7  .T...C.....CA..C.A.C..(1)
     gra8  G..G........A...CA.CTC(1)

     Type  SSSNSSSSSNSSSSSSSSSSSS

     aa       M     F
              V     L

     Pos      6     1
              8     8
                    1

     Type     N     N
              N     N

     Change   C     C
```

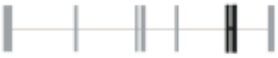
b) *Lcat*



3,831 bp

```
              111111111111111333333333333333
              11122233334444433444444445555
              11505812235667717023577891224
P. truei      37274177832020503633725791067
       Anc    CTGCGGGTGCGCTCGGAGGCAGCTGTCCT
       tCon   .........G......GT..T.TA....C
CAN    tru1   .............................(25
       tru2   .....A.......................(1)

CAS    tru1   .............................(4)
       tru3   ...............A.............(2)

AZW    tru1   .............................(7)
       tru3   ...............A.............(1)
       tru4   ..................A..........(1)

AZE    tru1   .............................(4)
       tru4   ..................A..........(1)
       tru5   .................G..A....C...(1)
       tru6   T................G..A........(1)
       tru7   ................AG...........(3)
       tru8   .................G...........(4)
       tru9   ...........................T.(1)
       tru10  ..A..............G...........(1)
       tru11  T............................(1)
       tru12  .................G..AA.......(3)
       tru13  .......................T....T(1)
       tru14  ...........T.................(1)
       tru16  .....A...........G...A.......(1)
       tru17  .....................A.......(1)
       tru18  .....A...........G..A........(1)
       tru19  .................G..A....CA..(2)
       tru20  ................AG..A.CT....T(2)

TX     tru7   ................AG...........(2)

UT     tru1   .............................(6)
       tru8   .................G...........(1)
       tru9   ...........................T.(1)
       tru19  .................G..A....CA..(1)
       tru21  ....................A........(1)
       tru22  .................G.........T.(2)
       tru23  .................G..A........(2)
P. gratus
       Anc    CTGCGGGTGCGCTCGGAGGCAGCTGTCCT
       gCon   ....A.AC.....A...............
       gra1   .C......A....T...............(1)
       gra2   .............TA..............(1)
       gra3   ...T....A....C...............(1)
       gra4   ...T........C................(1)
       gra5   ........A....C...............(1)
       gra6   .......T.....TA....T.........(1)
       gra7   ........A....C.........A.....(1)
       gra8   ......G......C...............(2)
       gra9   ........A.A..................(1)
       gra10  ............C................(1)
       gra11  .-.....T..A..................(1)
       gra12  .-.....T....C......T.........(1)
       gra13  .......T....C..........A.....(1)

       Type   IINSIIIIIIIIIIISSSNSSSSSSSSSS

       aa       R               R
                H               H

       Pos      1               3
                1               1
                3               0

       Type     +               +
                +               +

       Change   C               C
```

c) *Zp3*



9,589 bp

```
              999999999999999999999999
              111122222222222333333334
              235900234678889036667782
P. truei      843108290854583481256912
       Anc    CGGTGCGGTAACGCGTGCGATCGT
       tCon   .......C.GG....G....C..C
CAN    tru1   ........................(21)
       tru2   .......G..........ACT...(3)
       tru3   .......G...........CT...(2)

CAS    tru1   ........................(5)
       tru4   ................A.......(1)

AZW    tru1   ........................(9)
       tru3   .......G...........CT...(1)

AZE    tru1   ........................(23)
       tru2   .......G..........ACT...(1)
       tru5   .................T......(1)
       tru6   ...................C....(2)
       tru7   .......G...........C....(2)
       tru8   .......A...T.......C....(2)
       tru9   .......A....A......CT...(1)
       tru10  .......A...........CT...(2)

TX     tru1   ........................(2)

UT     tru1   ........................(10)
       tru2   .......G..........ACT...(1)
P. gratus
       Anc    CGGTGCGGTAACGCGTGCGATCGT
       gCon   .AAGATA..GG........C....
       gra1   ..........A.............(9)
       gra2   .GGTGCG..A....AG......A.(1)
       gra3   .GGTGCG..A....AG........(1)
       gra4   ...TGCG.G.A.A..G.......C(1)
       gra5   .G...................T..(1)
       gra6   T.GTGCG..A..A..G.......C(1)
       gra7   .GGTGCG..A....AG.......C(1)
       gra8   T.......................(3)
       gra9   .GGT.CG..A....AG.......C(1)
       gra10  .GGTGCG..A...TAG.......C(1)

       Type   SSNNIIIIIIIIIIIINSNNSSNN

       aa       EN            D VK  GW
                KK            N IQ  DR

       Pos      23            3 33  33
                90            1 22  24
                57            5 34  93

       Type     -P            - N+  PN
                ++            P NP  -+

       Change   MM            C CM  MM
                CC               C  CR
```

Figure 2.2 Tables of polymorphism, continued.

Literature Cited

Begun DJ, Whitley P, Todd BL, Waldrip-Dail HM, and Clark AG. 2000. Molecular population genetics of male accessory gland proteins in *Drosophila*. Genetics **156**:1879-1888.

Bi M, Hickox JR, Winfrey VP, Olson GE, and Hardy DM. 2003. Processing, localization and binding activity of zonadhesin suggest a function in sperm adhesion to the zona pellucida during exocytosis of the acrosome. Biochem. J. **375**:477-488.

Chen J, Litscher ES, and Wassarman PM. 1998. Inactivation of the mouse sperm receptor, mZP3, by site-directed mutagenesis of individual serine residues located at the combining site for sperm. Proc. Natl. Acad. Sci. USA **95**:6193-6197.

Clark NL, Aagaard JE, and Swanson WJ. 2006. Evolution of reproductive proteins from animals and plants. Reproduction **131**:11-22.

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

Clement M, Posada D, and Crandall KA. 2000. TCS: a computer program to estimate gene genealogies. Molecular Ecology **9**:1657-1659.

Coyne JA, and Orr HA. 2004. Speciation. Sinauer Associates, Sunderland, Mass.

Dean J. 2004. Reassessing the molecular biology of sperm-egg recognition with mouse genetics. Bioessays **26**:29-38.

Dice LR. 1968. Speciation. Pp. 75 - 97 *in* J. A. King, ed. Biology of *Peromyscus* (Rodentia). American Society of Mammalogists.

Dorus S, Evans PD, Wyckoff GJ, Choi SS, and Lahn BT. 2004. Rate of molecular evolution of the seminal protein gene *SEMG2* correlates with levels of female promiscuity. Nat. Genet. **36**:1326-1329.

Durish ND, Halcomb KE, Kilpatrick CW, and Bradley RD. 2004. Molecular systematics of the *Peromyscus truei* species group. J. Mammal. **85**:1160-1169.

Fay JC, and Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics **155**:1405-1413.

Frank SA. 2000. Sperm competition and female avoidance of polyspermy mediated by sperm-egg biochemistry. Evol. Ecol. Res. **2**:613-625.

Gasper J, and Swanson WJ. 2006. Molecular population genetics of the gene encoding the human fertilization protein zonadhesin reveals rapid adaptive evolution. Am. J. Hum. Genet. **79**:820-830.

Gavrilets S, and Waxman D. 2002. Sympatric speciation by sexual conflict. Proc. Natl. Acad. Sci. USA **99**:10533-10538.

Glassey B, and Civetta A. 2004. Positive selection at reproductive ADAM genes with potential intercellular binding activity. Mol. Biol. Evol. **21**:851-859.

Grantham R. 1974. Amino-acid difference formula to help explain protein evolution. Science **185**:862-864.

Haddrill PR, Thornton KR, Charlesworth B, and Andolfatto P. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. Genome Res. **15**:790-799.

Haldane JBS. 1922. Sex ratio and unisexual sterility in hybrid animals. Journal Of Genetics **12**:101-109.

Hamm D, Mautz BS, Wolfner MF, Aquadro CF, and Swanson WJ. 2007. Evidence of amino acid diversity–enhancing selection within humans and among primates at the candidate sperm-receptor gene *PKDREJ*. Am. J. Hum. Genet. **81**:44 - 52.

Haygood R. 2004. Sexual conflict and protein polymorphism. Evolution **58**:1414-1423.

Hermisson J, and Pennings PS. 2005. Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. Genetics **169**:2335-2352.

Hoffmeister DF. 1951. A taxonomic and evolutionary study of the pinon mouse. Illinois Biological Monographs **21**:1-104.

Hudson RR. 2001. Two-locus sampling distributions and their application. Genetics **159**:1805-1817.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18**:337-338.

Ishimizu T, Endo T, Yamaguchi-Kabata Y, Nakamura KT, Sakiyama F, and Norioka S. 1998. Identification of regions in which positive selection may operate in S-

RNase of Rosaceae: Implication for S-allele-specific recognition sites in S-RNase. Federation of European Biochemical Societies Letters **440**:337-342.

Jansa SA, Lundrigan BL, and Tucker PK. 2003. Tests for positive selection on immune and reproductive genes in closely related species of the murine genus *Mus*. J. Mol. Evol. **56**:294-307.

Kingan SB, Tatar M, and Rand DM. 2003. Reduced polymorphism in the chimpanzee semen coagulating protein, semenogelin I. J. Mol. Evol. **57**:159-169.

Kinloch RA, Sakai Y, and Wassarman PM. 1995. Mapping the mouse ZP3 combining site for sperm by exon swapping and site-directed mutagenesis. Proc. Natl. Acad. Sci. USA **92**:263-267.

Kuivenhoven JA, Pritchard H, Hill J, Frohlich J, Assmann G, and Kastelein J. 1997. The molecular pathology of lecithin cholesterol acyltransferase (LCAT) deficiency syndromes. J. Lipid Res. **38**:191-205.

Lee YH, Ota T, and Vacquier VD. 1995. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. Mol. Biol. Evol. **12**:231-238.

Li WH, Wu CI, and Luo CC. 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J. Mol. Evol. **21**:58-71.

Lorch PD, and Servedio MR. 2005. Postmating-prezygotic isolation is not an important source of selection for reinforcement within and between species in *Drosophila pseudoobscura* and *D. persimilis*. Evolution **59**:1039-1045.

McDonald JH, and Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature **351**:652-654.

McLeskey SB, Dowds C, Carballada R, White RR, and Sailing P. 1997. Molecules involved in mammalian sperm-egg interaction. International Review of Cytology **177**:57-113.

Metz EC, and Palumbi SR. 1996. Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. Mol. Biol. Evol. **13**:397-406.

Modi WS, and Lee MR. 1984. Systematic implications of chromosomal banding analyses of populations of *Peromyscus truei* (Rodentia, Muridae). P. Biol. Soc. Wash. **97**:716-723.

Nielsen R. 2001. Statistical tests of selective neutrality in the age of genomics. Heredity **86**:641-647.

Panhuis TM, and Swanson WJ. 2006. Molecular evolution and population genetic analysis of candidate female reproductive genes in *Drosophila*. Genetics **173**:2039-2047.

Podlaha O, Webb DM, Tucker PK, and Zhang J. 2005. Positive selection for indel substitutions in the rodent sperm protein CATSPER1. Mol. Biol. Evol. **22**:1845-1852.

Pritchard JK, Stephens M, and Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics **155**:945-959.

Przeworski M. 2002. The signature of positive selection at randomly chosen loci. Genetics **160**:1179-1189.

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

Rankin TL, Coleman JS, Epifano O, Hoodbhoy T, Turner SG, Castle PE, Lee E, Gore-Langton R, and Dean J. 2003. Fertility and taxon-specific sperm binding persist after replacement of mouse sperm receptors with human homologs. Developmental Cell **5**:33-43.

Retief JD, Winkfein RJ, Dixon GH, Adroer R, Queralt R, Ballabriga J, and Oliva R. 1993. Evolution of *protamine P1* genes in primates. J. Mol. Evol. **37**:426-434.

Robinson M, Catzeflis F, Briolay J, and Mouchiroud D. 1997. Molecular phylogeny of rodents, with special emphasis on murids: Evidence from nuclear gene *Lcat*. Mol. Phylogenet. Evol. **8**:423-434.

Roy BA, and Kirchner JW. 2000. Evolutionary dynamics of pathogen resistance and tolerance. Evolution **54**:51-63.

Simonsen KL, Churchill GA, and Aquadro CF. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141**:413-429.

Singh RS, and Kulathinal RJ. 2000. Sex gene pool evolution and speciation: A new paradigm. Genes Genet. Syst. **75**:119-130.

Stephens M, and Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am. J. Hum. Genet. **73**:1162-1169.

Stephens M, Smith NJ, and Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. **68**:978-989.

Sutton KA, Jungnickel MK, Ward CJ, Harris PC, and Florman HM. 2006. Functional characterization of PKDREJ, a male germ cell-restricted polycystin. J. Cell. Physiol. **209**:493-500.

Swanson W, J., and Vacquier VD. 2002a. Reproductive protein evolution. Annu. Rev. Ecol. Syst. **33**:161-179.

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

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

Swanson WJ, Yang Z, Wolfner MF, and Aquadro CF. 2001. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc. Natl. Acad. Sci. USA **98**:2509-2514.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**:585-596.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics **105**:437-460.

Takebayashi N, Brewer PB, Newbigin E, and Uyenoyama MK. 2003. Patterns of variation within self-incompatibility loci. Mol. Biol. Evol. **20**:1778-1794.

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

Turner LM, and Hoekstra HE. 2006. Adaptive evolution of fertilization proteins within a genus: Variation in ZP2 and ZP3 in deer mice (*Peromyscus*). Mol. Biol. Evol. **23**:1656-1669.

Vekemans X, and Slatkin M. 1994. Gene and allelic genealogies at a gametophytic self-incompatibility locus. Genetics **137**:1157-1165.

Wagstaff BJ, and 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.

Wassarman PM, Jovine L, and Litscher ES. 2001. A profile of fertilization in mammals. Nat. Cell Bio. **3**:E59-E64.

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

Williams Z, Litscher ES, Jovine L, and Wassarman PM. 2006. Polypeptide encoded by mouse ZP3 exon-7 is necessary and sufficient for binding of mouse sperm *in vitro*. J. Cell. Physiol. **207**:30-39.

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

Chapter 3: Comparative analysis of testis protein evolution in rodents

Abstract

Genes expressed in testes play important roles in male reproductive success, affecting spermatogenesis, sperm competition, and sperm-egg interaction. Testis-expressed proteins are thus likely to experience strong natural or sexual selection. Comparing testis proteins at different taxonomic levels can reveal which genes and functional classes are targets of selection, and whether these targets are similar across taxa. Here, we examine the evolution of testis-expressed proteins at two levels of divergence in Muroid rodents. First, we perform evolutionary analyses of expressed sequence tags (ESTs) from testis of the deer mouse (*Peromyscus maniculatus*), mouse (*Mus musculus*) and rat (*Rattus norvegicus*) to estimate rates of evolution and identify rapidly evolving genes. In all lineages, proteins with testis-specific expression evolve more rapidly than proteins with maximal expression in other tissues. Some genes show rate increases only in specific lineages, while others evolve rapidly in all lineages. Most of these rapidly evolving genes have not been identified previously as targets of natural or sexual selection. Genes with the highest rates of evolution have a variety of functional roles including signal transduction, DNA binding and egg-sperm interaction. Second, sequencing of a subset of these rapidly evolving genes in other *Peromyscus* species shows evidence for positive selection. Together, these results demonstrate rapid evolution of functionally diverse testis-expressed proteins in rodents that results from a combination of lineage-specific selection and selection common across mammals. Evidence for positive selection between 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 tend to evolve faster than other proteins, a pattern seen in diverse taxa (Singh and Kulathinal 2000; Swanson and Vacquier 2002a; Clark, Aagaard, and Swanson 2006). These rapidly evolving proteins serve diverse functions in both males and females and act at various stages of the fertilization process (Clark, Aagaard, and Swanson 2006). However, many questions remain unresolved, including: (1) Do proteins performing particular biological functions or participating in particular steps of fertilization evolve more rapidly than others? (2) Are the same proteins and amino acid sites targets of selection in different taxa? (3) Can the divergence of reproductive proteins contribute to speciation, causing reproductive isolation between closely related taxa?

In mammals, research on reproductive proteins has focused primarily on analyzing sequences 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, Aagaard, and Swanson 2006). However, a large number of proteins are involved in fertilization, many of whose functions are not yet well characterized (Jansen, Ekhlasi-

Hundrieser, and Toepfer-Petersen 2001; Tanphaichitr et al. 2007); choosing candidate genes based on function is difficult and likely to miss important targets of selection. A complementary approach is genome-wide analysis of reproductive proteins, which can characterize general patterns of evolution as well as identify rapidly evolving genes. This genomic approach has been used successfully to identify rapidly evolving male accessory gland proteins (Acps) in *Drosophila* (Swanson et al. 2001a) and crickets (Andres et al. 2006; Braswell et al. 2006), female reproductive tract proteins in *Drosophila* (Swanson et al. 2001b), and seminal proteins in primates (Clark and Swanson 2005).

Here we report a genomic analysis of testis-expressed proteins in three lineages of Muroid rodents. First, we identify rapidly evolving proteins by comparing expressed sequence tag (EST) sequences from testes of mouse (*Mus musculus*), rat (*Rattus norvegicus*), and deer mouse (*Peromyscus maniculatus*) to sequences of orthologues from the *Mus* or *Rattus* genomes. Comparisons of multiple species allowed us to test for differences in lineage-specific rates of evolution among proteins. Second, we determine whether these proteins are experiencing positive selection at a rate that could potentially cause reproductive isolation by sequencing the entire sequence in six species of *Peromyscus*. Together, these analyses identify a large number of rapidly evolving proteins, many of which have not previously been implicated as targets of selection.

Materials and Methods

Peromyscus maniculatus testis cDNA library construction and EST sequencing

A cDNA library was prepared from *P. maniculatus* testis tissue from a single adult male by Amplicon Express (Pullman, WA) using the ZAP-cDNA synthesis kit (Stratagene, La Jolla, CA). The library was amplified and phagemids excised and plated according to the manufacturer's protocol. Resulting colonies were grown overnight in Luria-Bertani/ ampicillin broth in deep well plates. Inserts were PCR amplified from bacterial cultures using T3 and T7 primers and presence of insert was verified on an agarose gel. PCR products from 4800 positive clones were sequenced from the 5' end using BigDye terminator (v. 3.1, Applied Biosystems, Valencia, CA) and samples were run on an ABI automated sequencer (3100, 3730, 3730XL, Applied Biosystems).

The program ALIGNER (CodonCode, Dedham, MA) was used to call bases (with embedded PHRED), trim vector sequences, and trim ends to maximize the region with an error rate < 0.05 (PHRED quality value > 13). Sequences < 90 bp in length were discarded. Remaining bases with quality values < 13 were changed to unknowns. ESTs were assembled into contigs using the CAP3 sequence assembly program (HUANG AND MADAN 1999).

Evolutionary EST analysis

We obtained sequences from testis cDNA libraries 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) in the NCBI dbEST database. Orthologues were identified 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). The following comparisons were made (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). Orthologues had 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. ω, 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. Non-overlapping ESTs matching the same RefSeq were concatenated.

For each orthologous pair, we estimated the ratio of nonsynonymous substitution rate to synonymous substitution rate ($d_N/d_S = \omega$) using maximum likelihood (ML) as implemented in the CODEML program from the PAML package (runmode -2, v 3.14,Yang 2000). Orthologous pairs with $d_S > 1.5$ were excluded from further analysis as these are unreliable due to estimation errors (Castillo-Davis, Hartl, and Achaz 2004). For pairs with ω values > 1, an additional model was run in PAML with ω fixed at 1. Likelihood ratio tests (LRT) comparing the estimated ω model to the fixed ω (neutral) model tested whether the estimated value of ω was significantly greater than 1. The test statistic for the LRT is the negative of twice the difference in log likelihoods between models (-2ΔlnL), and is χ^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 (Table 2), we performed a three-species comparison to identify which specific lineage(s) showed an elevated rate of amino acid change. Lineage-specific ω values were determined using the free ratios model in CODEML. A LRT comparing the free ratios model to the single ratio model was performed to determine whether there was significant evidence of variation in rates across lineages (test statistic = -2ΔlnL, χ^2, d.f. = 2).

Expression information for *Mus musculus* RefSeqs was obtained from the Genomics Institute of the Novartis Research Foundation (GNF) (Su et al. 2004; Walker et al. 2004) gcRMA-condensed (Wu and Irizarry 2005) dataset. Tissue specificity was classified 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. Remaining ESTs with maxT_S for testis were classified as testis specific (TS) and those with $\max T_S$ for another tissue were classified as non-testis-specific (NTS).

We compared rates of evolution between H, TS, and NTS ESTs. In general, tissue-specific genes evolve more rapidly than genes with broader expression, likely due to reduced pleiotropy (Duret and Mouchiroud 2000; Winter, Goodstadt, and Ponting 2004). Therefore we compared ω distribution of ESTs in each expression class using an analysis of covariance (ANCOVA) with level of tissue specificity ($\max T_S$) included as a covariate. ω and $\max T_S$ values were transformed towards normality and equal variances between groups; ω values were natural log-transformed and $\max T_S$ values were arcsine-transformed. In order to equalize variances among groups, ESTs with $\omega = 0$ were excluded. A greater proportion of NTS/H ESTs had $\omega = 0$, thus exclusion of these values resulted in a conservative test.

We classified functions of *Mus* homologues of ESTs using the PANTHER classification system (Thomas et al. 2003). Overrepresentation of functional processes in a particular group of genes (e.g. homologues of *Peromyscus* ESTs with $\omega > 0.5$) was determined relative to expected based on representation among total homologues of ESTs for each species. Significant over- or under-representation was determined using the binomial test (Cho and Campbell 2000) with Bonferroni correction for multiple tests.

Identification of protein domains in Peromyscus EST sequences

We used InterProScan (Zdobnov and Apweiler 2001) to search the InterPro combined protein database (Mulder et al. 2007) with all unique *Peromyscus* testis sequences (unigenes). InterProScan 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 species

Testis tissue from a single male from each of six species (*P. aztecus, P. californicus, P. eremicus, P. leucopus, P. maniculatus*, *P. polionotus*) was obtained from the *Peromyscus* Genetic Stock Center (see Supplementary material online for sample details). Tissue was excised from freshly sacrificed adult males and stored in RNAlater solution (Sigma, St. Louis, MO). RNA was extracted using the RNeasy kit (Qiagen, Valencia, CA) and cDNA was synthesized using a Superscript III RT kit (Invitrogen, Carlsbad, CA). All genes were amplified under standard PCR conditions, using primers designed by aligning *P. maniculatus* EST sequences to GenBank sequences from *Mus* and *Rattus*. Primer sequences are available from the authors upon request.

To determine species relationships, a 1,213 bp region of the mitochondrial genome (including *COIII* and *ND3*) was sequenced from one individual from each of the six *Peromyscus* species (see Supplementary material online for sample details). PCR amplification and sequencing were performed using published primers (Hoekstra, Drumm, and Nachman 2004).

PCR products were directly sequenced or cloned (TOPO-TA, Invitrogen). Cycle sequencing was performed with BigDye terminator (v. 3.1, Applied Biosystems, Valencia, CA) using PCR amplification primers and internal sequencing primers. Samples were run on an ABI 3100 automated sequencer (Applied Biosystems). Base calls were checked by eye and contigs were assembled in SEQUENCHER (Gene Codes, Ann Harbor, Michigan). Sequences were aligned using SEQUENCHER or MUSCLE (Edgar 2004). We excluded a large repetitive region from one of the genes (*Gm1276*, see Results) from further analysis because reliable alignment and inference of amino acid substitutions was not possible.

Analysis of Peromyscus testis-expressed gene sequences

Bayesian and maximum likelihood (ML) phylogenies of the six species were constructed, based on the mitochondrial sequences and 1,201 bp of the nuclear genes *Mc1r* and *Lcat* [from Genbank (Turner and Hoekstra 2006)]. Substitution model (GTR+G) was determined using MODELTEST 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. However, ML analyses were run both without partitions and with partitions (by gene); both methods resulted in identical topologies. Bayesian analysis was performed in MRBAYES with data partitioned by gene and codon position; substitution model parameters were estimated for each partition. Two runs were performed for 10 million generations; the first million generations were discarded as burn-in. The 99% credible set for the Bayesian analysis contained a single tree, identical in topology to the ML trees: (((*P. polionotus*, *P. maniculatus*), *P. leucopus*), *P. aztecus*, (*P. eremicus*, *P. californicus*)).

We used the codon-based ML method developed by Nielsen, Yang and colleagues (Nielsen and Yang 1998; Yang et al. 2000) to detect evidence for positive selection acting on gene pairs. This method employs a LRT to compare a neutral model, where ω for all sites is constrained to be < 1, to a selection model where a subset of sites have ω greater than one. As with the LRT described above, the test statistic is -2ΔlnL and is χ^2 distributed with degrees of freedom equal to the difference in number of estimated parameters between the neutral and selection models. We compared likelihoods of M8 (selection) to M7 (neutral), with two degrees of freedom. In M7, ω varies as a beta distribution between 0 and 1 and M8 adds a selection class with ω ≥ 1 (Nielsen and Yang 1998). This test has very low power and convergence problems with limited divergence or short sequences; thus, for this analysis, we only included sequences ≥ 50 codons and tree length > 0.11 (following Church et al. 2007). Models were implemented in CODEML.

We used the same codon-based ML method applied to gene pairs from the EST analysis to detect evidence for positive selection acting on testis genes in *Peromyscus*. We performed the following model comparisons (neutral vs. selection): M1a vs. M2a, M7 vs. M8 (described above), and M8A vs. M8. M1a has two site classes, the first with 0 < ω < 1 and the second with ω = 1. M2a adds an additional 'selection' class with ω ≥ 1. M8A is a modified version of M8 where ω for the selection class is constrained to equal one. The M8A vs. M8 comparison tests whether ω 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 -2ΔlnL test statistic is distributed as a 50:50 mixture of a point mass at zero and a χ^2 distribution with 1 degree of freedom (Swanson, Nielsen, and Yang 2003). Codon models were implemented in CODEML using the ML/Bayesian tree topology. Specific amino acid sites subject to positive selection were identified using the Bayes empirical Bayes (BEB) procedure (Yang, Wong, and Nielsen 2005).

In addition, we applied codon models to determine whether genes positively selected within *Peromyscus* have also evolved under positive selection among divergent species of mammals. For the five genes with evidence for positive selection (see Results), we identified homologues from other mammalian species in GenBank using BLAST (see supplemental material for species and accession numbers). To avoid significant results due to positive selection within *Peromyscus*, we included sequences from a single species, *Peromyscus maniculatus*, in these analyses. Amino acid sequences were aligned using MUSCLE and corresponding nucleotide alignments manually checked and adjusted in MEGA (Kumar et al. 2001). Sites with alignment gaps were excluded from further analysis. Neighbor-joining trees were constructed in PAUP* using model parameters determined by MODELTEST. Codon models were run 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). Sequence characteristics are described in detail in Glenn *et al.* (In Prep).

Evolutionary EST analysis

To identify the most rapidly evolving testis proteins, we identified orthologous genes in *Peromyscus*, *Mus* and *Rattus*. Orthologues in *Mus* and/or *Rattus* were identified for ~43% of unique *P. maniculatus* EST sequences (Table 1), resulting in 1,014 (PM) and 993 (PR) orthologous pairs. 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 were collapsed into 7,448 unigenes and we identified *Mus* homologues for 56% of these, for a total of 4,207 orthologous pairs.

For each EST-RefSeq comparison, ω for the vast majority of orthologue pairs was << 1, indicating that most genes experience purifying selection (Table 1). However, a small percentage of pairs (1.3 – 2.4%) have ω > 1, a signature of positive selection. Only a few of these comparisons have ω values significantly greater than one. This criterion for identifying positively selected genes is extremely stringent as pairwise ω values are averaged across all amino acid sites, and selection often targets only a few key sites in a protein. A literature survey of studies that used the ML approach of Yang and colleagues (Yang et al. 2000) to detect selection when ω varies across sites found significant evidence for positive selection acting on a subset of sites for most genes with overall $\omega > 0.5$ (Swanson et al. 2004). Of the Muroid orthologue pairs identified here, 7.5 – 12.2% have $0.5 < \omega < 1$. We include these genes in the rapidly evolving gene lists for further analyses.

A representative plot of d_N vs. d_S values for all pairs from the PM comparison is presented in Figure 1A. For three of the EST-RefSeq comparisons (PM, PR, RM), proportions of ESTs in three ω classes ($\omega < 0.5$, $0.5 < \omega < 1$, $\omega > 1$) are not significantly different ($P = 0.96$; Pearson's χ^2, d.f. = 4). The MR comparison has the largest proportion of ESTs in both rapidly evolving classes, resulting in a significant effect of species comparison on ω class when MR is included with the other three ($P < 0.001$; Pearson's χ^2, d.f. = 6). This is surprising since MR and RM compare sequences from the same species pair and approximately 1/3 of the total number of orthologous pairs between *Mus* and *Rattus* were identified in both comparisons. However, relatively high sequencing error rates in either *Mus* ESTs or *Rattus* RefSeqs could explain this difference. The latter case seems more likely as the *Rattus* genome sequence was completed more recently than the *Mus* genome, has lower sequencing coverage, and is less well annotated (Waterston et al. 2002; Gibbs et al. 2004).

In some cases, genes were identified as rapidly evolving in multiple species comparisons, whereas other genes were not rapidly evolving in all comparisons (Figure 2B). Overall, 64% (72/112) of rapidly evolving genes with orthologous pairs identified in multiple pairwise species comparisons were rapidly evolving in more than one comparison and 26% (29/112) were rapidly evolving in all comparisons. The genes that have ω > 0.5 in all species pairs (PM, PR, RM and/or MR) are listed in Table 2. Cases where genes were classified as rapidly evolving in some but not all species comparisons could reflect differences in evolutionary rate between lineages. However, differences could also result from ESTs aligning to different regions of the orthologous gene, which vary in evolutionary rate.

Codon-based ML methods are capable of detecting positive selection, even when it acts on a small proportion of amino acid sites (Nielsen and Yang 1998; Yang et al. 2000). These methods are not generally applied to pairwise sequence comparisons because when sample size is small, they have low power to detect selection and estimates of ω of the selected class can be unreliable. However, LRTs (comparing selection to neutral models) are conservative, particularly with small sample sizes. Thus, although power is low with few sequences, accuracy is high (Anisimova, Bielawski, and Yang 2001). Codon model comparisons (M2a vs. M1a) were recently employed to detect selection on genes from sequence pairs identified in EST analyses within sunflower and within lettuce (Church et al. 2007). We compared a different selection/neutral pair of codon models (M8 vs. M7), because this test is conservative (Anisimova, Bielawski, and Yang 2001). A substantial proportion of ESTs have significant evidence for positive selection acting on a subset of amino acid sites (Table 1). Proportions are much higher for the MR (20.6%) and RM (42.0%) comparisons than for PM (9.3%) and PR (8.7%). The latter two comparisons are between more divergent taxa (Figure 2A). One possible explanation for the difference in the proportion of ESTs with significant M8 vs. M7 could be that, for the less diverged sequences there have very few substitution events and thus more likelihood that the distribution of the few nonsynonymous and synonymous substitutions across sites is stochastic variation and not reflective of actual variation in selective pressure across sites.

Evolutionary rates of testis-specific genes

Genes expressed in the testis may also be expressed in other tissues; we used expression data from Mus to determine which genes were testis-specific and which had 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 ω 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 result may reflect stronger selection driving the rapid divergence of these genes. However, tissue-specific genes evolve more rapidly in general than genes with broader expression, likely due to reduced pleiotropy (Winter, Goodstadt, and Ponting 2004). To control for tissue-specificity, we used an ANCOVA analysis that incorporates expression class and level of tissue specificity (maxT_S) as covariates. This analysis confirmed a highly significant effect of testis-specific expression on ω in all species comparisons ($P < 0.0001$).

In addition, we compared the proportion of genes from each expression class between groups of ESTs with different ω values (ω < 0.5, ω > 0.5). The results of this analysis provide an intuitive demonstration of elevated evolutionary rates of testis-specific genes. For all four EST-RefSeq comparisons, there is a highly significant relationship between expression class and ω class ($P < 0.0001$, Pearson's χ^2, d.f. = 2). Specifically, proportions of testis-specific genes were significantly higher among rapidly evolving genes (ω > 0.5) 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 the PM comparison is depicted in Figure 1B.

We use expression data from mouse to categorize tissue specificity of genes, although it is possible that genes we define as testis-specific may not have had maximal expression in testis across Muroid lineages. Expression data are limited for *P. maniculatus* and testis was not represented among tissues in the GNF *Rattus* gene expression database, thus we are unable to evaluate the validity of the assumption that expression patterns have remained consistent among *Mus, Rattus,* and *Peromyscus*. Significant differences have been reported in level of expression of many genes in testis between two *Mus* species (Voolstra et al. 2007), however it is unclear how often changes in level of expression represent changes in tissue of maximal expression. Rates of expression profile divergence between mouse and human were inversely correlated with level of tissue specificity (Liao and Zhang 2006), thus genes with highly specific expression in *Mus* testis may be more likely to have been consistently testis-specific in all three lineages.

Functions of rapidly evolving genes

To determine if particular functional classes of genes tend to be rapidly evolving ($\omega > 0.5$), we used the PANTHER classification system to assign genes to particular functional categories. In all three pairwise species comparisons, genes unclassified for both biological process and molecular function are overrepresented in the rapidly evolving class ($P < 0.002$). 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). In contrast, several functional classes are underrepresented among rapidly evolving genes; nucleic acid binding proteins are underrepresented in both PM and MR/RM ($P < 0.01$) and several types are underrepresented in MR/RM [intracellular protein traffic ($P < 0.0001$); protein metabolism and modification ($P < 0.0001$); cell cycle ($P < 0.001$); nucleoside, nucleotide and nucleic acid metabolism ($P < 0.001$); protein biosynthesis ($P < 0.001$); general vesicle transport ($P < 0.005$); cytoskeletal proteins ($P < 0.01$); and select regulatory molecules ($P < 0.01$)].

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 ω values in the PM comparison (excluding hypothetical proteins) in preliminary runs of the EST screen. Three genes from the latter category (*Gsg1, H1fnt, Smcp*) had high ω values in preliminary screens, but much lower ω values in the final screen, subsequent to corrections of alignments or changes in *Mus* RefSeqs.

The amount of functional information available for the rapidly evolving genes varies. Some genes have known roles in sperm-egg interaction [*Acr* (Howes et al. 2001)*, Spa17* (Richardson, Yamasaki, and O'Rand 1994)*, Spag8 (Cheng et al. 2007)*] or spermatogenesis [*Hils1* (Yan et al. 2003b)]. 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. Finally, the majority of genes have no available functional information.

The sperm-egg binding gene *Acr* (Tanphaichitr et al. 2007) and the sperm motility gene *Smcp* (Nayernia et al. 2002) are of particular interest based on their known functions and previously reported evidence for rapid evolution (Swanson, Nielsen, and Yang 2003; Tanphaichitr et al. 2007). Knockouts of each of these genes have partially infertile phenotypes in *Mus* -/- males *Acr* knockouts never sire offspring in competitive mating trials with wild-type males (Jansen, Ekhlasi-Hundrieser, and Toepfer-Petersen 2001) and *Smcp* knockouts are infertile on an inbred background (Nayernia et al. 2002). Further, triple knockouts of *Acr* and/or *Smcp* in different combinations with other sperm genes have greatly reduced fertility (Nayernia et al. 2005). *Acr* has been identified as positively selected in a phylogenetically diverse sample of mammals (Swanson, Nielsen, and Yang 2003). Although ω estimates for *Smcp* were relatively low, this gene remains interesting as variation in length of a repetitive region may be selected (Hawthorne et al. 2006).

We looked for protein domain homology in all *Peromyscus* ESTs using InterProScan to get information about possible functions of genes when we could not identify orthologues in *Mus* or *Rattus* (non-matches). Unfortunately, this analysis was not informative; very few non-matches had any indication of homology.

Evolution of rapidly evolving genes in Peromyscus

For the most rapidly evolving genes identified in the genomic analysis, we sequenced most or all of the coding regions in several closely related *Peromyscus* species to determine whether there is evidence of rapid amino acid change that may contribute to reproductive isolation between sister taxa. Estimates of pairwise ω for coding regions sequenced in *P. maniculatus* vs. *Mus* homologues (Table 3) were consistent with ω values for the shorter ESTs in some cases (e.g. *Hils1*, *Gm1276*) and inconsistent in others (e.g. *Lrrc50*, *Acr*). Five of the eight genes classified as rapidly evolving (ω > 0.5) based on the EST screen also had ω > 0.5 for the full sequence. 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, ω within *Peromyscus* was not significantly correlated with pairwise ω estimates of *P. maniculatus* vs. *Mus* (*P* = 0.16), although the trend was positive (Rsq = 0.20). Some genes with high ω along the *Peromyscus-Mus* lineage had relatively high rates within *Peromyscus* (e.g. *Gm1276*), while others had lower average rates in *Peromyscus* (e.g. *Spa17*). We performed a three-species analysis of *P. maniculatus*, *Mus* and *Rattus* sequences for each gene in PAML to estimate lineage-specific values of ω. We thought that estimates of ω for the *Peromyscus* lineage might be better predictors of ω among

Peromyscus species, as they should not be influence by rates of evolution along the *Mus* lineage. However, there was significant evidence of variation in ω among lineages for three of the ten genes (*H1fnt*, *Smcp*, *Gsg1*). Further, the lineage-specific estimates were poorer predictors (*P* = 0.60, Rsq = 0.03) of ω within *Peromyscus* than the pairwise *P. maniculatus*-*Mus* estimates. This could be due to unreliability of lineage-specific ω estimates in PAML when differences are small.

Positive selection within Peromyscus

We used a ML approach to determine whether rapidly evolving genes in Muroid lineages have evidence of positive selection within *Peromyscus*. Results from the ML codon models indicate that 5 of the 11 genes have a subset of amino acid sites that are targets of positive selection in the *Peromyscus* genus (Table 3). For each of these genes, comparisons of M8 to M8A were significant, indicating that a proportion of sites was subject to selection and ω for the selected class was significantly greater than one. However, comparisons of M2 vs. M1 and M8 vs. M7 were only significant for one of these genes (*Gm1276*), a result not surprising given the low power of these tests when sequence divergence is limited and sample size is small (Anisimova, Bielawski, and Yang 2001). While the M8 vs. M8A comparison is more conservative than the other comparisons for distinguishing between positive selection and nearly neutral processes, the LRT is less conservative because there are fewer degrees of freedom and it is based on the true distribution of the test statistic rather than an approximation (Swanson, Nielsen, and Yang 2003).

For each gene sequenced in *Peromyscus*, we determined several estimates of evolutionary rate: pairwise ω of EST sequences vs. *Mus* and vs. *Rattus* RefSeqs, lineage-specific ω in *Peromyscus* determined through 3-species analysis, and overall ω determined through comparison of the full *P. maniculatus* sequence to *Mus*. However, none of these measures was a good predictor of which genes have evidence for positive selection within *Peromyscus.* The five genes with evidence for positive selection include the gene that had the highest estimate of ω vs. *Mus* in the EST screen (*Lrrc50*) as well as one gene with an EST that was not classified as rapidly evolving in the final screen (*Gsg1*). Similarly, positively selected genes include the gene with the highest lineage-specific estimate of ω (*Lrrc50*) and the second lowest value of the eleven genes (*Ddc8*). Overall measures of ω within *Peromyscus* for positively selected genes ranged from 0.22 (*Acr*) to 0.69 (*Gm1276*).

Amino acid alignments of the five positively genes are given in Figure 3 and 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 were less than 0.95. Unlike the previous method used to identify positively selected sites (naive empirical Bayes), the BEB procedure has low type I error rates (false positives) but also has low power to detect true positive sites with probability > 0.95 when sample size is small (Yang, Wong, and Nielsen 2005). We view these results as a preliminary indication of the spatial distribution of target sites along the length of the protein. For example, in *Acr* and *Lrrc50*, there are clusters of target sites whereas in *Gm1276* and *Ddc8*, target sites are scattered throughout the sequence. Evidence of functions of regions within these genes is lacking, consequently we are limited in our ability to infer the biological significance of these patterns.

In proacrosin, there is a cluster of target sites in the C-terminus. Evidence of 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, Harris, and Vazquez-Levin 2005), but in the mouse there is no evidence of the C-terminus binding to ZP2 (Howes et al. 2001). The C-terminus is cleaved during processing of proacrosin to the mature proteolytic form following ZP binding in boar (Mori et al. 1995), thus this region is unlikely to have a role in dissolution of the ZP.

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/β 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, but the specific function of this protein has not been characterized. The N-terminus of MS4A is located in the cytoplasm. One transmembrane (TM) domain of MS4A13 has been deleted; thus, unlike other family members, the C-terminus is extracellular. All positively selected sites in *Peromyscus* are located in this extracellular region.

Length variation in testis genes within Peromyscus

In addition to changes in amino acid sequence, changes in protein length have been implicated as a target of selection in reproductive proteins (Podlaha and Zhang 2003; Podlaha et al. 2005). We examined length variation in several testis genes to determine if length differences in *Peromyscus* may also be the target of selection. One gene, *Phf8,* has evolved premature stop codons between *Peromyscus* and the Murids (*Mus, Rattus*) and also within *Peromyscus*. In *P. aztecus*, 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. The *P. aztecus* sequence was excluded from further analysis, since the vast majority of it is untranslated. The *P. polionotus* stop is 180 codons upstream of the *Mus* stop (202 upstream of *Rattus*) and the rest of the species (*P. californicus*, *P. eremicus*, *P. leucopus*, *P. maniculatus*) have stops 62 codons upstream of *Mus*. PHF8 has an exceptionally high proportion of variable amino acid sites (26%) relative to the other proteins (Table 3), and ω is relatively high within *Peromyscus* (0.47) but there is no evidence for positive selection. Given the apparent loss of function in one species, high variability among remaining species, and occurrence of premature stop codons, it seems that in this case, high ω may result from relaxed functional constraint rather than positive selection acting on a subset of amino acid sites.

The gene with the strongest evidence for positive selection, *Gm1276*, contains a large repeat region that varies in length among *Peromyscus* species from 252 – 360 aa (Fig. 3). The region comprises 45 - 56 copies of a five aa repeat motif (consensus

PSQET) and 3 – 10 interspersed copies of an eight aa motif (PSEAYQDI). Repeat regions of the shorter motif 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, 330 aa respectively). The region is much shorter (≤ 125 aa) in other mammalian species sampled. The 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. Consequently, we can't make any inferences about how expansion/contraction of this region affects protein function.

Smcp also 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). Reported *Mus* and *Rattus* sequences have two additional copies of the repeat motif relative to *P. maniculatus*. The sequence determined here for *P. maniculatus* is identical to the published sequence. Length variation between *Peromyscus* species is limited, there are a few amino acid indels within repeats but all species have the same number of repeats. *Smcp* was included in our preliminary list of rapidly evolving genes but has low values of pairwise ω based on final alignments and there is no evidence for positive selection in *Peromyscus* (Tables 2 - 3). Therefore, this gene is not likely to be evolving rapidly within *Peromyscus*, either in terms of amino acid sequence or length.

Evolution of testis genes in divergent mammals

Most evidence for the rapid evolution of reproductive proteins in mammals comes from comparisons of divergent taxa. We wanted to determine if genes with evidence for positive selection in *Peromyscus* also show evidence for positive selection in comparisons among divergent mammalian taxa. For each of the five genes with evidence for positive selection in *Peromyscus*, we identified homologues in 5 – 10 additional mammalian species. For *Ddc8*, *Gsg1*, and *Lrrc50*, there is no evidence for positive selection from the LRTs comparing codon models. The M8 vs. M7 and M8A vs. M8 comparisons both identify a subset of sites of *Acr* that are positively selected ($P \leq 0.01$; $p_s = 0.07$, $\omega_s = 1.82$), 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, Nielsen, and Yang 2003). No amino acid sites were identified as positively selected both within this diverse sample and within *Peromyscus*, although two sites in the positive selection class (45, 48) in mammals are close to one positively selected site in *Peromyscus* (46). Unfortunately, no functional data exist for these specific sites, however they are only ~20 aa upstream of one of the two regions implicated in zona pellucida binding in *Mus* (Jansen, Ekhlasi-Hundrieser, and Toepfer-Petersen 2001). The other three positively selected sites in *Peromyscus* are clustered in the C-terminal region. This region cannot be aligned reliably between divergent species, thus we can't compare targets of selection in that region between these two taxonomic levels.

For *Gm1276*, all model comparisons provide strong evidence ($P < 0.0001$) for positive selection (M8: $p_s = 0.25$, ω_s. = 2.12). Similar to results within *Peromyscus*, the 13 sites with high probability (> 0.9) of being in the positively selected class are distributed along the length of the protein. One site (101) is in the intracellular loop between TM domains 2 and 3, another site (113) is in TM domain 3, and the remaining sites are in the C-terminal extracellular region. One of these sites (142) was also identified as positively selected in *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. However, in addition to the lack of evidence of functions of regions within this protein, it is not known whether this protein is present in sperm much less whether it affects fertilization success.

Discussion

In this study, we show that evolutionary rates of testis-specific proteins are consistently elevated in three Muroid lineages. We identify a list of rapidly evolving genes, many of which have not previously been implicated as targets of selection. These rapidly evolving genes are functionally diverse: there is little evidence that selection is focused on any particular biological process or stage of fertilization. The evolutionary pattern across taxa is variable: some genes have evidence for positive selection along a single lineage, whereas others are rapidly evolving in all three Muroid lineages as well as across divergent mammalian species. In addition, we find evidence for positive selection acting on five genes within closely related species of *Peromyscus*, raising the possibility that these genes may contribute to reduced fertilization potential between diverging species. This study contributes to a large body of evidence documenting a remarkable pattern of rapid evolution of reproductive proteins in animal taxa.

Functional roles of rapidly evolving genes

Identifying the functions of rapidly evolving genes may reveal whether a particular biological process or fertilization step is subject to strong selective pressure and may help to resolve which evolutionary forces (i.e. sperm competition, sexual conflict, pathogen defense) are the source of this pressure. A few of the genes with the highest rates of evolution in our analysis have well described roles in fertilization (Table 2). Two genes, *Hils1* and *H1fnt* are involved in DNA condensation during spermatogenesis, specifically the process of repackaging DNA onto testis-specific histones to produce the densely packed chromatin of sperm. Two other genes*, Acr* and *Spa17,* are zona pellucida (egg coat) binding proteins. Both of these classes of proteins have previously been identified as potentially important targets of selection. Three DNA packaging sperm proteins (*Prm1*, *Prm2, Tnp2*) have evidence for positive selection in mammals (Retief et al. 1993; Queralt et al. 1995; but see Clark and Civetta 2000; but see Rooney, Zhang, and Nei 2000; Wyckoff, Wang, and Wu 2000; Torgerson, Kulathinal, and Singh 2002). A lot of research effort has been focused on proteins involved in binding of egg and sperm, as this interaction is critical to species-specificity of fertilization (Wassarman, Jovine, and Litscher 2001). Numerous proteins on both the egg and sperm side of this interaction evolve rapidly and show signatures positive selection at a variety of levels of taxonomic divergence in mammals (Swanson et al. 2001b; Jansa, Lundrigan, and Tucker 2003; Swanson, Nielsen, and Yang 2003; Glassey and Civetta 2004; Good and Nachman 2005; Gasper and Swanson 2006; Podlaha, Webb, and Zhang 2006; Turner and Hoekstra 2006; Hamm et al. 2007).

In addition to well-characterized genes, many of the most rapidly evolving genes have gene ontology annotations. For these, we have an indication of the general function of the gene but it is not possible to identify involvement in a specific fertilization step. This group includes genes encoding proteins involved in protein binding (*Lrrc50*, *Phf8*) signal transduction (*Gm1276*, *4930596D02Rik*), and a variety of other molecular processes.

PANTHER classification of rapidly evolving genes highlights immunity proteins and KRAB box transcription factors as groups with high evolutionary rates between *Mus* and *Rattus*. High rates of evolution of proteins with immune function are not surprising, as these proteins tend to evolve rapidly in general (Roy and Kirchner 2000) and defense against pathogens has been proposed as a possible explanation for rapid reproductive protein evolution (Swanson and Vacquier 2002b). In contrast, transcription factors are often highly conserved (e.g. De Craene, van Roy, and Berx 2005; Wijchers, Burbach, and Smidt 2006), and are not generally expected to be common targets of positive selection. Zinc finger proteins with KRAB motifs are the largest family of transcription factors in mammals (Birtle and Ponting 2006);

prediction of which types of genes might be regulated by proteins in the rapidly evolving group is not straightforward. We can speculate that perhaps these transcription factors regulate genes specifically expressed during spermatogenesis.

Genes with unknown function are highly overrepresented in all comparisons; it is possible that the limited signal of particularly rapidly evolving protein classes is not truly reflective of functional diversity of rapidly evolving genes but because for some reason processes targeted by selection are not well-studied. Our results are consistent with a study that compared rates of evolution of genes expressed at different stages of spermatogenesis in the mouse (Good and Nachman 2005), which showed that rates of evolution are higher for genes expressed in late stages of spermatogenesis. These late-expressed genes serve a wide variety of functions but tend to be more testis-specific than genes expressed at earlier stages.

The identification of numerous rapidly evolving genes with unknown function, some of which have signatures of positive selection at multiple levels of taxonomic divergence, underscores the importance of combining analysis of candidate proteins that have well-described function with genomic approaches that facilitate identification of novel targets. Moreover, evolutionary analyses provide valuable data to researchers investigating the molecular processes of reproduction; for example, some of these previously undescribed genes subject to positive selection may play important roles in fertilization. Finally, for genes whose functions are known, knowledge of specific amino acid sites that are targets of selection can facilitate identification of key regions of functional importance (Yang 2005).

Targets of selection at various levels of divergence

A large number of proteins are involved in mammalian fertilization, and there is considerable functional redundancy between proteins, particularly on the male side (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 possible that the individual proteins targeted may differ between species. On the other hand, in some cases a common molecular basis has been found for natural phenotypes that require the interaction of numerous genes in complex pathways. For example, nearly 100 genes have been found to contribute to coat color phenotypes in laboratory mice (Barsh 1996), however, changes in a single gene, *Mc1r*, have been implicated in natural pigmentation variation in mammals, birds, and reptiles (Hoekstra 2006).

Since the majority of studies that have demonstrated positive selection on reproductive proteins in mammals have sampled divergent species, it is clear that there are some common targets. Further, some of the genes identified in these divergent 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 mammalian species (Swanson et al. 2001b) are evolving adaptively within Murids (Jansa, Lundrigan, and Tucker 2003) and within *Peromyscus* (Turner and Hoekstra 2006). 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).

However, in this study, 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 previously been identified as targets of selection in more divergent mammals (e.g. Spa17 Swanson, Nielsen, and Yang 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 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 pleiotropic effects of changes in genes with shared function. Simultaneous analysis of evolutionary patterns of the same genes at various taxonomic levels allowed us to identify cases where there are discrepancies in pattern between taxa and to identify novel targets of lineage-specific selection.

Success of EST screen

The evolutionary EST analysis we employed determines rates of evolution between a single species of *Peromyscus* and other Muroid rodents. Estimating ω for a large number of orthologous pairs is an appealing approach for identifying rapidly evolving genes because it is possible to quickly evaluate many genes, but we were unsure at the outset how successful this approach would be at identifying genes subject to positive selection within *Peromyscus*. High values of ω between *Peromyscus* and *Mus* or *Rattus* may result from rapid evolution solely within Murids or that occurred subsequent to the divergence of *Peromyscus* from Murids but before the diversification of the *Peromyscus* genus. This screen was successful, approximately half of the sequenced genes chosen based on high pairwise ω vs. *Mus* have evidence that a subset of amino acid have been subject to positive selection within *Peromyscus*. Moreover, four out of five positively selected genes have not been previously identified 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. However, as selection acts on a small proportion of amino acid sites in many genes, choosing genes based on ω values averaged across large regions certainly will miss important targets. EST analysis and other genomic approaches are complementary to choosing genes based on knowledge of their biological functions.

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.

Length variation

A comparison of evolutionary rates for mouse-human orthologous pairs demonstrated that sperm-specific proteins have exceptionally high rates of evolution, both in terms of amino acid substitution rate and variation in protein length (Torgerson, Kulathinal, and Singh 2002), suggesting both may be common responses to selection. Here, we found that the putative signal transduction protein Gm1276 has evolved rapidly in substitution rate as well as sequence length, both in a phylogenetically diverse sample of mammals and within *Peromyscus*. Length variation results from expansion and contraction of a large repeat region. This region expanded greatly sometime 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 region varies in length by more than 100 amino acids in six closely related *Peromyscus* species. These results suggest that this gene may consistently respond to selection through two different mechanisms of sequence evolution. However, the functional impact of length variation in Gm1276 is not clear because the role of the repeat region is unknown. Thus, although there is evidence that positive selection promotes amino acid substitution in this protein, functional data are required to evaluate whether length variation is adaptive or if the region has reduced constraint.

In contrast, we found no evidence for selection for amino acid substitutions or variation in repeat number between *Peromyscus* species in the sperm motility protein

Smcp, which was previously reported to evolve rapidly via changes in repeat number in diverse mammals including a difference between *P. maniculatus* and *Mus/Rattus*.

Implications for fertilization and reproductive isolation

We are ultimately interested in finding genes that cause and maintain reproductive isolation between species. 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". The initial motivation for this study was to identify testis proteins that are diverging rapidly in *Peromyscus* and potentially play a role in reducing fertilization success between diverging species. We have successfully identified five genes that are positively selected between these closely related species. Among these, *Acr* and *Gm1276* are particularly good candidates for further study, since some information about the structure and function of these proteins is also available.

Variation in expression pattern is another mechanism of divergence that could potentially be an important response to selection on testis proteins. Our analysis was limited to protein coding regions, so we are unable to compare rates of coding vs. regulatory change in testis proteins here. However, substantial differences in expression of testis proteins have been reported between species of house mice but not between subspecies that are partially reproductively isolated. This suggests that expression differences in testis do not contribute to reproductive isolation in *Mus* (Voolstra et al. 2007).

Conclusions

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; Riginos, Wang, and Abrams 2006) and determining the consequences of protein divergence on fertilization potential between species (Lyon and Vacquier 1999; Paumbi 1999; Levitan and Ferrell 2006). In mammals, however, a detailed understanding of the causes and consequences of the rapid divergence of reproductive proteins remains elusive. 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 (Coyne and Orr 2004).

Here, we identify a functionally diverse set of genes that are evolving rapidly in rodents. Most of these genes have not been previously considered potential targets of selection and the majority have unknown function. 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, within a superfamily, and within a genus whereas rapid evolution of other genes was limited to a single lineage. In addition, we identified five genes that are positively selected in closely related species of *Peromyscus*. These genes are particularly 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 between species.

Acknowledgements

We thank members of the Hoekstra Laboratory, J. Kohn, and V. Vacquier for helpful discussion. Julie Glenn and Adrienne Lewandowski sequenced the *P. maniculatus* ESTs. The *Peromyscus* Genetic Stock Center kindly provided tissue samples. This research was supported by a Howard Hughes Medical Institute Predoctoral Fellowship (to LMT), National Science Foundation doctoral dissertation improvement grant DEB 0608030 (to LMT and HEH), and National Science Foundation DEB 0344710 (to HEH).

Chapter 3 is in preparation for publication: Turner LM, Chuong EB, and Hoekstra HE. In Preparation. Comparative analysis of testis protein evolution in rodents.

Table 3.1 Evolutionary rates of testis-expressed genes.

Comparison	*N*	Homologues	Mean Alignment length (codons)	$\overline{d_N}$	$\overline{d_S}$	$\overline{\omega}$	Rapidly evolving 0.5 < ω < 1.0	Rapidly evolving ω > 1.0	significant M8 vs. M7/ total measured (%)	$\overline{\omega}$ H	$\overline{\omega}$ NTS	$\overline{\omega}$ TS
Peromyscus vs. *Mus*	2,364	1,014	157	0.06	0.35	0.19	76 (7.5%)	13 (0) (1.3%)	89/952 (9.3%)	0.13 (343)	0.23 (174)	0.30*** (226)
Peromyscus vs. *Rattus*	2,364	993	157	0.07	0.36	0.19	81 (8.2%)	14 (0) (1.4%)	81/926 (8.7%)	0.13 (325)	0.24 (165)	0.30*** (209)
Mus vs. *Rattus*	11,203	4,171	147	0.07	0.25	0.28	509 (12.2%)	99 (3) (2.4%)	720/3,502 (20.6%)	0.23 (1225)	0.23 (710)	0.33*** (667)
Rattus vs. *Mus*	7,448	4,207	200	0.06	0.24	0.23	342 (8.1%)	60 (0) (1.4%)	1650/3929 (42.0%)	0.19 (1530)	0.19 (936)	0.29*** (586)

N = number of unique ESTs compared to Refseqs. $\overline{d_N}$ = mean nonsynonymous substitution rate; $\overline{d_S}$ = mean nonsynonymous substitution rate; $\overline{\omega}$ = mean ω for all ESTs, H = housekeeping genes, NTS = non-testis-specific genes, T = testis-specific, number of ESTs in each category is indicated in parentheses. Number and percentage of rapidly evolving ESTs are given for two ω classes, 0.5 < ω < 1.0, and ω > 1.0. For the ω > 1.0 class, the number of ESTs with ω significantly greater than 1.0 are given in parentheses (as determined by a likelihood ratio test comparing estimated ω models vs. models with ω fixed at 1.0, implemented in PAML). Asterisks indicate significance (*** = P < 0.001) in an ANCOVA determining effects of expression class on ω, controlling for degree of tissue specificity of expression by including it as a covariate.

Table 3.2 Rapidly evolving testis-expressed genes.

Gene	Symbol	Chromosome (*Mus*)	Gene Ontology	Function	ω_{PM}	ω_{PR}	ω_{MR}	ω_{RM}
hypothetical protein LOC70900	*4921517D22Rik*[2]	13	unknown	unknown	2.28	1.20	0.73	0.63
hypothetical protein LOC238663	*4932411G14Rik*[2]	13	unknown	unkown	1.24	1.27	nd	0.72
hypothetical protein LOC242838	*4932412H11Rik*[2]	5	protein binding	unknown	1.21	1.37	1.44	nd
leucine rich region containing 50	*Lrrc50*[1,2]	8	protein binding	unknown	1.17	1.32	1.16	0.11
histone H1-like protein in spermatids 1	*Hils1*[1,2]	11	DNA binding; histone binding	DNA condensation during spermiogenesis (Yan et al. 2003a)	1.13	0.89	1.24	nd
hypothetical protein LOC210940	*4931408C20Rik*[2]	1	unknown	unknown	0.91	1.01	nd	0.87
gene model 1276	*Gm1276*[1,2]	19	receptor activity; signal transduction	unknown	0.88	0.82	nd	0.82
PHD finger protein 8	*Phf8*[1]	X	DNA binding; metal ion binding; protein binding; zinc ion binding	unknown	0.81	0.72	0.44	nd
chemokine-like factor isoform 1	*Cklf*[2]	8	cytokine activity; chemotaxis	unknown	0.81	0.54	0.55	nd
preproacrosin	*Acr*[1]	15	acrosin activity; amidase activity; fucose binding; hydrolase activity; mannose binding; peptidase activity; protein binding; serine-type endopeptidase activity	secondary binding to zona pellucida (ZP2), dispersal of acrosomal contents	0.80	0.40	0.23	0.17
hypothetical protein LOC71831 isoform 3	*1700007B14Rik*[2]	8	unknown	unknown	0.80	0.70	nd	0.63

Table 3.2 Rapidly evolving testis-expressed genes, continued.

coiled-coil-helix-coiled-coil-helix domain containing 6	*Chchd6*[2]	6	unknown	unknown	0.77	0.60	0.69	nd
hypothetical protein LOC75275	*4930563P21Rik*[2]	2	unknown	unknown	0.77	0.67	0.51	nd
hypothetical protein LOC78174	*4930503B16Rik*[2]	5	cytochrome-c oxidase activity; electron transport; mitochondrial respiratory chain	unknown	0.75	0.53	0.53	nd
sperm associated antigen 8	*Spag8*[1,2]	4	unknown	unknown	0.72	0.71	nd	0.63
hypothetical protein LOC381816	*4922502D21Rik*[2]	6	sugar binding	unknown	0.72	0.95	1.18	1.07
acrosome formation associated factor	*Afaf*[2]	4	unknown	acrosome formation during spermiogenesis (Li et al. 2006)	0.71	0.90	0.52	nd
CKLF-like MARVEL transmembrane domain containing 2A	*Cmtm2a*[2]	8	cytokine activity; protein binding; transcription corepressor activity; chemotaxis; negative regulation of transcription (DNA-dependent);	androgen receptor co-repressor involved in regulation of transcription (Jeong et al. 2004)	0.68	0.61	0.67	0.67
sperm autoantigenic protein 17	*Spa17*[1,2]	9	cAMP-dependent protein kinase regulator activity	zona pellucida binding	0.67	0.66	0.52	nd
similar to Protein C14orf32 homolog	*C130032J12Rik*[2]	14	unknown	unknown	0.67	0.52	nd	0.95

Table 3.2 Rapidly evolving testis-expressed genes, continued.

hypothetical protein LOC239036	*4930596D02Rik*[2]	14	calcium ion binding; guanyl-nucleotide exchange factor activity; regulation of small GTPase mediated signal transduction; small GTPase mediated signal transduction	unknown	0.64	0.63	0.51	0.75
hypothetical protein LOC271036 (CatSper$^{\beta}$)	*4932415G16Rik*[2]	12	unknown	part of CatSper1 ion channel protein complex, required for sperm hyperactivation (Liu et al. 2007)	0.63	0.81	0.57	nd
gene model 884	*Gm884*[2]	11	unkown	unknown	0.62	0.56	0.64	nd
hypothetical protein LOC73309	*1700047L15Rik*[2]	12	unknown	unknown	0.59	0.63	0.59	nd
hypothetical protein LOC67687 isoform 2	*1700011L22Rik*[2]	8	unknown	unknown	0.59	0.59	0.75	0.82
lysosomal-associated membrane protein 1	*Lamp1*[2]	8	unknown	release of spermatazoa from epithelium during spermatogenesis (Guttman, Takai, and Vogl 2004)	0.58	0.53	nd	0.54
testis specific protein Ddc8	*Ddc8*[1,2]	11	unknown	unknown	0.56	0.68	0.54	0.34
hypothetical protein LOC70980	*4931431F19Rik*[2]	7	unknown	unknown	0.55	0.54	0.61	0.56
spermatogenesis associated 3	*Spata3*[2]	1	apoptosis; spermatogenesis	unknown	0.54	0.72	1.16	nd
similar to kinesin-like motor protein C20orf23	*C20orf23*[2]	2	unknown	unknown	0.54	0.70	1.82	nd
germ cell-specific gene 1	*Gsg1*[1,2]	6	unknown	unknown	0.44	0.30	nd	0.36

Table 3.2 Rapidly evolving testis-expressed genes, continued.

histone H1 variant	*H1fnt*[1,3]	15	DNA binding	DNA condensation during spermiogenesis, essential for proper nuclear morphology (Martianov et al. 2005) (Tanaka et al. 2005)	0.18	0.14	0.74	0.52
sperm mitochondria-associated cysteine-rich protein	*Smcp*[1,3]	3	selenium binding	sperm motility(Nayernia et al. 2002)	0.14	nd	nd	nd

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(V. 3.14, 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_{PM} is between *P. maniculatus* EST and *M. musculus* RefSeq.

[1]candidate genes sequenced in additional *Peromyscus* species. [2]*ω* > 0.5 in all comparisons. [3]included because initial screen showed high *ω* values.

Table 3.3 Adaptive evolution of testis-expressed genes in *Peromyscus.*

			ω					
Gene	L_C	ω_{PM}	vs. *Mus*	*ω Peromyscus*	Variable aa sites (%)	M8 vs M8A	ω_s	p_s
Lrrc50	622	1.17	0.30	0.41	65 (10.4)	0.039*	3.90	0.05
Hils1	162	1.13	1.16	0.42	15 (9.3)	0.416	n/a	n/a
Gm1276	830	0.88	0.94	0.69	74 (8.9)	0.005**	10.12	0.02
Phf8	447	0.81	0.27	0.47	118 (26.4)	0.127	n/a	n/a
Acr	428	0.80	0.42	0.22	18 (4.2)	0.012*	7.55	0.01
Spag8	263	0.72	0.64	0.57	54 (20.5)	0.500	n/a	n/a
Spa17	147	0.67	0.59	0.16	4 (2.7)	0.283	n/a	n/a
Ddc8	539	0.56	0.47	0.55	50 (9.3)	0.042*	2.19	0.28
Gsg1	364	0.44	0.48	0.34	29 (8.0)	0.022*	2.18	0.17
H1fnt	304	0.18	0.56	0.28	21 (6.9)	0.225	n/a	n/a
Smcp	136	0.14	0.13	0.11	7 (5.1)	0.500	n/a	n/a

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

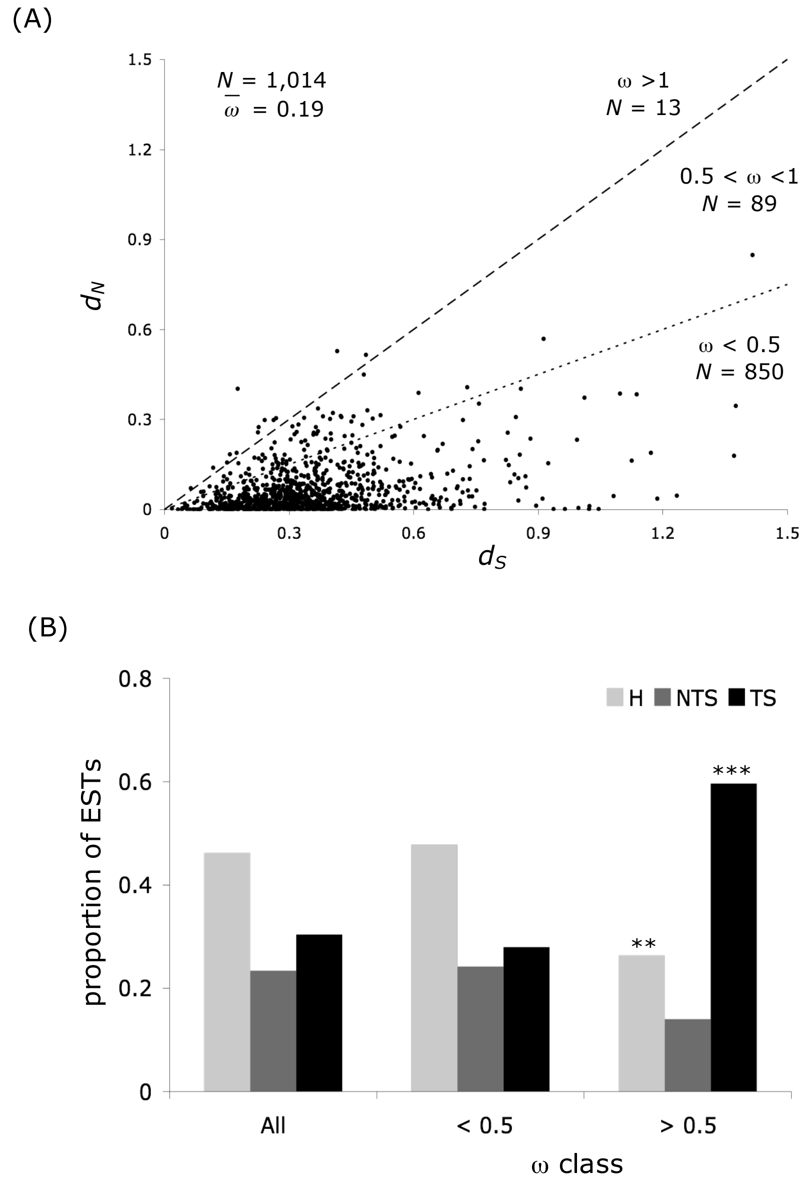


Figure 3.1 Evolutionary rates of testis-expressed ESTs.

(A) d_N vs. d_S estimated in paml(Yang 2000) where 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 ω 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 ω class.

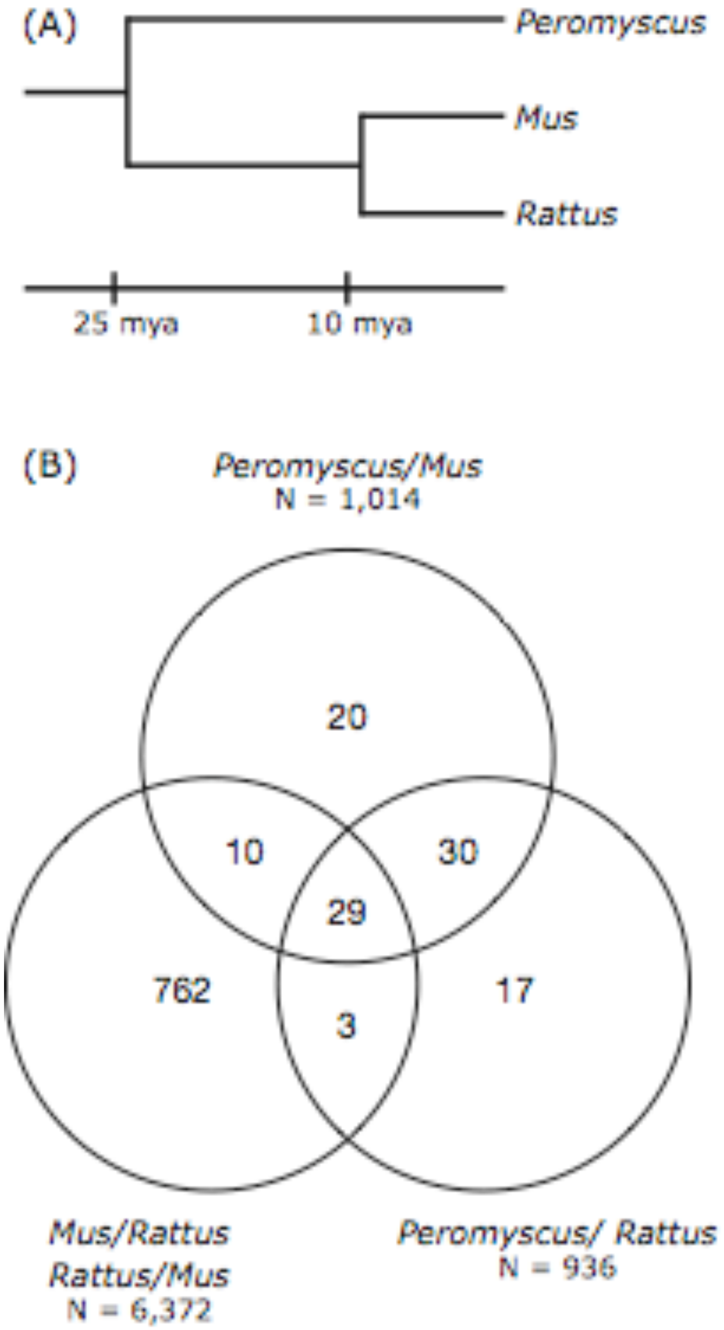


Figure 3.2 Rapidly evolving testis genes in three rodent lineages.
(A) Relationships and divergence times between *Peromyscus*, *Mus*, and *Rattus* (Steppan, Adkins, and Anderson 2004). (B) Numbers of testis genes identified as rapidly evolving (ω > 0.5) in one or more comparisons between rodent species.

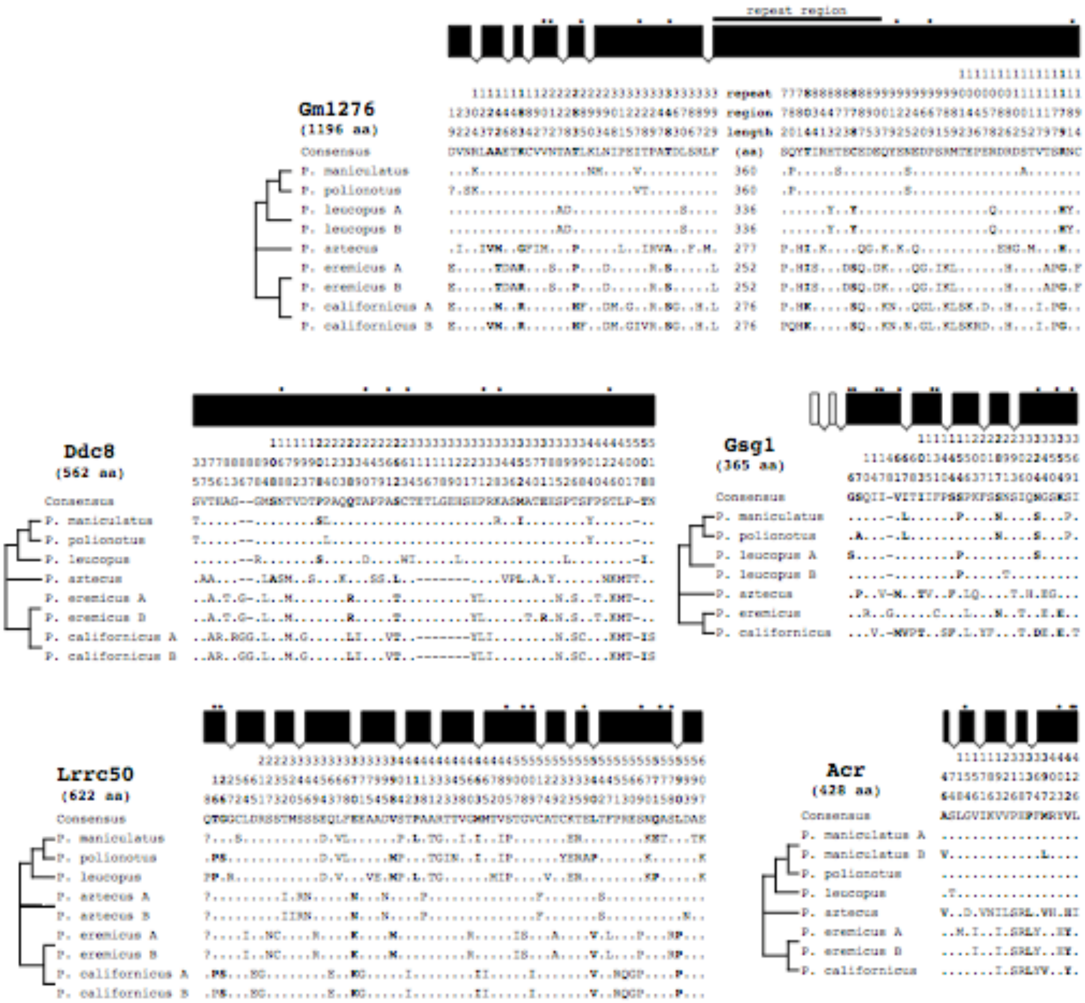


Figure 3.3 Positively selected testis genes in *Peromyscus*.
Alignments of variable amino acid sites are shown for each gene. Dots indicate identity with the consensus sequence. Amino acid sites identified as positively selected with BEB analysis are in bold. Relationships between species are shown to 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.

Literature Cited

Andres JA, Maroja LS, Bogdanowicz SM, Swanson WJ, and Harrison RG. 2006. Molecular evolution of seminal proteins in field crickets. Mol. Biol. Evol. **23**:1574-1584.

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

Barsh GS. 1996. The genetics of pigmentation: From fancy genes to complex traits. Trends Genet. **12**:299.

Birtle Z, and Ponting CP. 2006. Meisetz and the birth of the KRAB motif. Bioinformatics **22**:2841-2845.

Braswell WE, Andres JA, Maroja LS, Harrison RG, Howard DJ, and Swanson WJ. 2006. Identification and comparative analysis of accessory gland proteins in Orthoptera. Genome **49**:1069-1080.

Castillo-Davis CI, Hartl DL, and Achaz G. 2004. Cis-regulatory and protein evolution in orthologous and duplicate genes. Genome Res. **14**:1530-1536.

Cheng GY, Shi JL, Wang M, Hu YQ, Liu CM, Wang YF, and Xu C. 2007. Inhibition of mouse acrosome reaction and sperm-zona pellucida binding by anti-human sperm membrane protein 1 antibody. Asian J. Androl. **9**:23-29.

Cho RJ, and Campbell MJ. 2000. Transcription, genomes, function. Trends Genet. **16**:409-415.

Church S, Livingstone K, Lai Z, Kozik A, Knapp S, Michelmore R, and Rieseberg L. 2007. Using variable rate models to identify genes under selection in sequence pairs: Their validity and limitations for EST sequences. J. Mol. Evol. **64**:171.

Clark AG, and Civetta A. 2000. Evolutionary biology: Protamine wars. Nature **403**:261-263.

Clark NL, Aagaard JE, and Swanson WJ. 2006. Evolution of reproductive proteins from animals and plants. Reproduction **131**:11-22.

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

Coyne JA, and Orr HA. 2004. Speciation. Sinauer Associates, Sunderland, Mass.

De Craene B, van Roy F, and Berx G. 2005. Unraveling signalling cascades for the SNAIL family of transcription factors. Cell. Signal. **17**:535-547.

Duret L, and Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. **17**:68-74.

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

Furlong LI, Harris JD, and Vazquez-Levin MH. 2005. Binding of recombinant human proacrosin/acrosin to zona pellucida (ZP) glycoproteins. I. Studies with recombinant human ZPA, ZPB, and ZPC. Fertil. Steril. **83**:1780-1790.

Gasper J, and Swanson WJ. 2006. Molecular population genetics of the gene encoding the human fertilization protein zonadhesin reveals rapid adaptive evolution. Am. J. Hum. Genet. **79**:820-830.

Geyer LB, and Palumbi SR. 2003. Reproductive character displacement and the genetics of gamete recognition in tropical sea urchins. Evolution **57**:1049-1060.

Gibbs RAWeinstock GMMetzker ML et al. 2004. Genome sequence of the brown norway rat yields insights into mammalian evolution. Nature **428**:493-521.

Glassey B, and Civetta A. 2004. Positive selection at reproductive *Adam* genes with potential intercellular binding activity. Mol. Biol. Evol. **21**:851-859.

Good JM, and Nachman MW. 2005. Rates of protein evolution are positively correlated with developmental timing of expression during mouse spermatogenesis. Mol. Biol. Evol. **22**:1044-1052.

Guttman JA, Takai Y, and Vogl AW. 2004. Evidence that tubulobulbar complexes in the seminiferous epithelium are involved with internalization of adhesion junctions. Biol. Reprod. **71**:548-559.

Hamm D, Mautz BS, Wolfner MF, Aquadro CF, and Swanson WJ. 2007. Evidence of amino acid diversity–enhancing selection within humans and among primates at the candidate sperm-receptor gene *Pkdrej*. Am. J. Hum. Genet. **81**:44 - 52.

Hawthorne SK, Goodarzi G, Bagarova J, Gallant KE, Busanelli RR, Olend WJ, and Kleene KC. 2006. Comparative genomics of the sperm mitochondria-associated cysteine-rich protein gene. Genomics **87**:382-391.

Hoekstra HE. 2006. Genetics, development and evolution of adaptive pigmentation in vertebrates. Heredity **97**:222-234.

Hoekstra HE, Drumm KE, and Nachman MW. 2004. Ecological genetics of adaptive color polymorphism in pocket mice: Geographic variation in selected and neutral genes. Evolution **58**:1329-1341.

Howes E, Pascall JC, Engel W, and Jones R. 2001. Interactions between mouse ZP2 glycoprotein and proacrosin; a mechanism for secondary binding of sperm to the zona pellucida during fertilization. J. Cell Sci. **114**:4127-4136.

Huang XQ, and Madan A. 1999. CAP3: A DNA sequence assembly program. Genome Res. **9**:868-877.

Ishibashi K, Suzuki M, Sasaki S, and Imai M. 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.

Jansa SA, Lundrigan BL, and Tucker PK. 2003. Tests for positive selection on immune and reproductive genes in closely related species of the murine genus *Mus*. J. Mol. Evol. **56**:294-307.

Jansen S, Ekhlasi-Hundrieser M, and Toepfer-Petersen E. 2001. Sperm adhesion molecules: Structure and function. Cells Tissues Organs **168**:82-92.

Jeong BC, Hong CY, Chattopadhyay S, Park JH, Gong EY, Kim HJ, Chun SY, and Lee K. 2004. Androgen receptor corepressor-19 kda (ARR19), a leucine-rich protein that represses the transcriptional activity of androgen receptor through recruitment of histone deacetylase. Mol. Endocrinol. **18**:13-25.

Kumar S, Tamura K, Jakobsen IB, and Nei M. 2001. Mega2: Molecular evolutionary genetics analysis software. Bioinformatics **17**:1244-1245.

Levitan DR, and Ferrell DL. 2006. Selection on gamete recognition proteins depends on sex, density, and genotype frequency. Science **312**:267-269.

Li YC, Hu XQ, Zhang KY, Guo H, Hu ZY, Tao SX, Xiao LJ, Wang QZ, Han CS, and Liu YX. 2006. AFAF, a novel vesicle membrane protein, is related to acrosome formation in murine testis. FEBS Lett. **580**:4266-4273.

Liao BY, and Zhang JZ. 2006. Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. Mol. Biol. Evol. **23**:1119-1128.

Liu J, Xia J, Cho K-H, Clapham DE, and Ren D. 2007. CATSPERBETA, a novel transmembrane protein in the catsper channel complex. J. Biol. Chem. **282**:18945-18952.

Lyon JD, and Vacquier VD. 1999. Interspecies chimeric sperm lysins identify regions mediating species-specific recognition of the abalone egg vitelline envelope. Dev. Biol. **214**:151-159.

Martianov I, Brancorsini S, Catena R, Gansmuller A, Kotaja N, Parvinen M, Sassone-Corsi P, and Davidson I. 2005. Polar nuclear localization of H1T2, a histone H1 variant, required for spermatid elongation and DNA condensation during spermiogenesis. Proc. Natl. Acad. Sci. USA **102**:2808-2813.

Mori E, Kashiwabara S, Baba T, Inagaki Y, and Mori T. 1995. Amino acid sequences of porcine SP38 and proacrosin required for binding to the zona pellucida. Dev. Biol. **168**:575-583.

Mulder NJ, Apweiler R, Attwood TK et al. 2007. New developments in the Interpro database. Nucl. Acids Res. **35**:D224-228.

Nayernia K, Adham IM, Burkhardt-Gottges E, Neesen J, Rieche M, Wolf S, Sancken U, Kleene K, and Engel W. 2002. Asthenozoospermia in mice with targeted deletion of the sperm mitochondrion-associated cysteine-rich protein (*Smcp*) gene. Mol. Cell. Biol. **22**:3046-3052.

Nayernia K, Drabent B, Meinhardt A, Adham IM, Schwandt I, Mueller C, Sancken U, Kleene KC, and Engel W. 2005. Triple knockouts reveal gene interactions affecting fertility of male mice. Mol. Repro. Dev. **70**:406-416.

Nielsen R, and Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**:929-936.

Palumbi SR. 1999. All males are not created equal: Fertility differences depend on gamete recognition polymorphisms in sea urchins. Proc. Natl. Acad. Sci. USA **96**:12632-12637.

Pearson WR. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. **183**:63-98.

Podlaha O, Webb DM, Tucker PK, and Zhang J. 2005. Positive selection for indel substitutions in the rodent sperm protein CATSPER1. Mol. Biol. Evol. **22**:1845-1852.

Podlaha O, Webb DM, and Zhang J. 2006. Accelerated evolution and loss of a domain of the sperm-egg-binding protein SED1 in ancestral primates. Mol. Biol. Evol. **23**:1828-1831.

Podlaha O, and Zhang J. 2003. Positive selection on protein-length in the evolution of a primate sperm ion channel. Proc. Natl. Acad. Sci. USA **100**:12241-12246.

Posada D, and Crandall KA. 1998. Modeltest: Testing the model of DNA substitution. Bioinformatics **14**:817-818.

Queralt R, Adroer R, Oliva R, Winkfein RJ, Retief JD, and Dixon GH. 1995. Evolution of *Protamine p1* genes in mammals. J. Mol. Evol. **40**:601-607.

Retief JD, Winkfein RJ, Dixon GH, Adroer R, Queralt R, Ballabriga J, and Oliva R. 1993. Evolution of *Protamine p1* genes in primates. J. Mol. Evol. **37**:426-434.

Richardson RT, Yamasaki N, and O'Rand MG. 1994. Sequence of a rabbit sperm zona pellucida binding protein and localization during the acrosome reaction. Dev. Biol. **165**:688-701.

Riginos C, Wang D, and Abrams AJ. 2006. Geographic variation and positive selection on m7 lysin, an acrosomal sperm protein in mussels (*Mytilus* spp.). Mol. Biol. Evol. **23**:1952-1965.

Rooney AP, Zhang J, and Nei M. 2000. An unusual form of purifying selection in a sperm protein. Mol. Biol. Evol. **17**:278-283.

Roy BA, and Kirchner JW. 2000. Evolutionary dynamics of pathogen resistance and tolerance. Evolution **54**:51-63.

Singh RS, and Kulathinal RJ. 2000. Sex gene pool evolution and speciation: A new paradigm. Genes Genet. Syst. **75**:119-130.

Stajich JE, Block D, Boulez K et al. 2002. The bioperl toolkit: Perl modules for the life sciences. Genome Res. **12**:1611-1618.

Steppan S, Adkins R, and Anderson J. 2004. Phylogeny and divergence-date estimates of rapid radiations in muroid rodents based on multiple nuclear genes. Syst Biol **53**:533-553.

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

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, and Aquadro CF. 2001a. Evolutionary est analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proc. Natl. Acad. Sci. USA **98**:7375-7379.

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

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

Swanson WJ, and Vacquier VD. 2002b. Reproductive protein evolution. Annu. Rev. Ecol. Syst. **33**:161-179.

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

Swanson WJ, Yang Z, Wolfner MF, and Aquadro CF. 2001b. Positive darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc. Natl. Acad. Sci. USA **98**:2509-2514.

Swofford DL. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Mass.

Tanaka H, Iguchi N, Isotani A et al. 2005. HANP1/H1T2, a novel histone h1-like protein involved in nuclear formation and sperm fertility. Mol. Cell. Biol. **25**:7107-7119.

Tanphaichitr N, Carmona E, Khalil MB, Xu HB, Berger T, and Gerton GL. 2007. New insights into sperm-zona pellucida interaction: Involvement of sperm lipid rafts. Front. Biosci. **12**:1748-1766.

Thomas PD, Kejariwal A, Campbell MJ et al. 2003. Panther: A browsable database of gene products organized by biological function, using curated protein family and subfamily classification. Nucleic Acids Res. **31**:334-341.

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

Turner LM, and Hoekstra HE. 2006. Adaptive evolution of fertilization proteins within a genus: Variation in ZP2 and ZP3 in deer mice (*Peromyscus*). Mol. Biol. Evol. **23**:1656-1669.

Voolstra C, Tautz D, Farbrother P, Eichinger L, and Harr B. 2007. Contrasting evolution of expression differences in the testis between species and subspecies of the house mouse. Genome Res. **17**:42-49.

Walker JR, Su AI, Self DW, Hogenesch JB, Lapp H, Maier R, Hoyer D, and Bilbe G. 2004. Applications of a rat multiple tissue gene expression data set. Genome Res. **14**:742-749.

Wassarman PM, Jovine L, and Litscher ES. 2001. A profile of fertilization in mammals. Nat. Cell Bio. **3**:E59-E64.

Waterston RHLindblad-Toh KBirney E et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520-562.

Wijchers P, Burbach JPH, and Smidt MP. 2006. In control of biology: Of mice, men and foxes. Biochem. J. **397**:233-246.

Winter EE, Goodstadt L, and Ponting CP. 2004. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. Genome Res. **14**:54-61.

Wu ZJ, and Irizarry RA. 2005. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. J. Comput. Biol. **12**:882-893.

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

Yan W, Ma L, Burns KH, and Matzuk MM. 2003a. Hils1 is a spermatid-specific linker histone h1-like protein implicated in chromatin remodeling during mammalian spermiogenesis. Proc. Natl. Acad. Sci. USA **100**:10546-10551.

Yang Z. 2005. The power of phylogenetic comparison in revealing protein function. Proc. Natl. Acad. Sci. USA **102**:3179-3180.

Yang Z. 2000. Phylogenetic analysis by maximum likelihood (PAML). University College, London.

Yang Z, Nielsen R, Goldman N, and Pedersen AK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431-449.

Yang Z, Wong WSW, and Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. **22**:1107 - 1118.

Zdobnov EM, and Apweiler R. 2001. Interproscan - an integration platform for the signature-recognition methods in interpro. Bioinformatics **17**:847-848.