EVOLUTION OF REPRODUCTIVE TRACT INTERACTIONS IN CACTOPHILIC DROSOPHILA

by

Erin Sarah Kelleher

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF ECOLOGY AND EVOLUTIONARY BIOLOGY

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2009

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Erin Sarah Kelleher

entitled Evolution of Reproductive Tract Interactions in Cactophilic *Drosophila*

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

_____ Date: 12/8/08
Dr. Therese A. Markow

_____ Date: 12/8/08
Dr. Giovanni Bosco

_____ Date: 12/8/08
Dr. Carlos A. Machado

_____ Date: 12/8/08
Dr. Michael W. Nachman

_____ Date: 12/8/08
Dr. Willie J. Swanson

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: 12/8/08
Dissertation Director: Dr. Therese A. Markow

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Erin Sarah Kelleher

ACKNOWLEDGEMENTS

I would like to acknowledge my doctoral committee, Dr. Therese A. Markow, Dr. Giovanni Bosco, Dr. Carlos Machado, Dr. Michael Nachman, and Dr. Willie Swanson. Their guidance and generosity were integral to the development and completion of my dissertation research. I would particularly like to acknowledge my graduate advisor Dr. Therese A. Markow, for her contribution and support of my dissertation research, as well as her mentorship and commitment to my professional development.

I furthermore would like to acknowledge the many collaborators and peers whom have contributed to this research. Dr. Therese A. Markow provided lab space, reagents, *Drosophila* strains, computational tools, and insight on every study in my dissertation. Dr. Willie Swanson provided considerable insight in the design and analysis of my transcriptional study of *Drosophila arizonae* female reproductive tracts. Mr. James Pennington helped me design and implement biochemical assays of proteolytic activity on *D. arizonae* female reproductive tracts. Dr. Nathan Clark performed 3D structural analysis of functional sites and selected sites. Dr. Thomas Hartl provided assistance with mapping the novel paralog in the Mojave Desert population of *D. mojavensis*. My colleagues in the Markow Lab, Dr. Jeremy Bono, Dr. Vanessa Corby, Kimberly Franklin, Tamara Haselkorn, Brooke LaFlamme, Katie Massie, Dr. Luciano Matzkin, Dr. Laura Reed, Thomas Watts, as well as the Nachman Lab, Miguel Carneiro, Dr. Matthew Dean, Lisa Kent, Dr. Michael Nachman, Tovah Salcedo, Dr. Maria Sans-Fuentes, and Gabriela Wlasiuk, provided critical feedback on my research and a rich and exciting environment for pursuing scientific research. My peers in the department of Ecology and Evolutionary Biology, as well as Plant Sciences, Biochemistry, Molecular and Cellular Biology, and Insect Science, were both professionally and personally supportive throughout my dissertation.

Through my graduate career I was supported by fellowships from the NSF-IGERT program in Evolutionary, Functional and Computational Genomics at the University of Arizona, as well as a dissertation fellowship from the American Association of University of Women. I received research and travel support from the Graduate and Professional Student Council, the Center for Insect Science, the Department of Ecology and Evolutionary Biology, Women in Science and Engineering, Graduate Women in Science, the Society for the Study of Evolution, and the National Science Foundation.

DEDICATION

This dissertation research is dedicated with love to my parents William and Denise Kelleher, and my sister Caitlin Kelleher for their tireless support of all my pursuits.

TABLE OF CONTENTS

ABSTRACT

Reproductive traits evolve rapidly at the morphological, physiological and molecular levels, a taxonomically robust pattern that is thought to arise from sexual selection. In internally fertilizing organisms, female promiscuity results in competition between multiple male ejaculates for fertilizations in the same female reproductive tract, extending sexual selection past courtship and copulation. In this post-copulatory arena, biochemical interaction between male ejaculates and female reproductive tracts form a dynamic molecular interface that modulates female post-mating responses essential to reproductive fitness. Consistent with the hypothesis that these interactions are subject to sexual selection, sperm and seminal proteins are known to evolve rapidly in a broad range of taxa. The female role in this process, however, in terms of both molecular mechanisms and evolutionary dynamics, remains unclear.

The presented dissertation research examines the biochemical nature and evolutionary consequences of post-copulatory sexual selection in two sister-species of cactophilic *Drosophila*, *D. mojavensis* and *D. arizonae*. I first present data that female post-mating response in crosses between these to species is perturbed, severely reducing the reproductive output of heterospecific crosses. A breakdown of reproductive tract interactions in matings between divergent lineages suggests that male and female contributions to reproductive outcomes are coadapted. Next, I use a combination of bioinformatic analyses, comparative sequence analyses, and biochemical assays to elucidate candidate female reproductive tract proteins that may be involved in ejaculate-female dynamics. 241 candidate female reproductive proteins are identified, the most

intriguing of which are recently-duplicated secreted proteases. Finally, I explore the evolutionary history of two families of secreted proteases within geographically isolated populations of *D. mojavensis*. I show that both families evolve rapidly through a complex process involving gene duplication, gene conversion, pseudogenation and positive selection, a unique pattern never before documented in reproductive proteins.

Collectively, my dissertation research suggests that females are active participants in the evolution of reproductive tract interactions. Further exploration of how sexual reproduction coevolves between males and females, both in terms of interacting biomolecules, and dynamic evolutionary histories, remains an important challenge for future research.

CHAPTER 1: INTRODUCTION

Context of the Problem:

Darwin's original insight that traits increasing an individual's mating success experience strong directional selection was inspired by the overwhelming diversity he observed in male courtship displays and signaling traits in mammals, birds, fish and insects (1871). Subsequent research has shown that sexual selection affects a broad range of traits in both males and females, from molecules to morphology (Andersson 1994). Reproductive fitness, furthermore, is not solely determined by mating success: in organisms that fertilize internally, female promiscuity causes the ejaculates of multiple males to compete for fertilizations in a single reproductive tract (Parker 1970). Although selection in this post-copulatory arena partially reflects male-male competition (Parker 1970), a male's fertilization success also is determined by how effectively he interacts with, or exploits, the environment presented by the female reproductive tract (Eberhard 1996).

Biochemical interactions between male ejaculates and female reproductive tracts and oocytes present an intriguing forum for the mediation of intersexual dynamics. Sperm rely on molecular cues and responses from females to navigate through the reproductive tract, remain viable in this environment, and ultimately fertilize female gametes (Reviewed in Neubaum and Wolfner 1998). Seminal components also affect reproductive success by mediating post-copulatory physiological and behavioral changes in mated females (Reviewed in Woflner 2007; Robertson 2005; 2007). Consistent with the hypothesis that these interactions are subject to sexual selection, male seminal and sperm proteins have been observed to evolve rapidly in broad range of organisms, a

molecular mirror of the of the phenotypic diversity exhibited by many secondary sexual characters (Swanson and Vacquier 2002; Clark *et al* 2006; Panhuis *et al* 2006).

Adaptive evolution of male reproductive proteins could result from either intrasexual competition or intersexual selection, and likely reflects both. A truly comprehensive understanding of post-copulatory sexual selection, therefore, requires identification of the female interactors of male seminal proteins, as well as complimentary studies of the evolutionary history of these molecules. Although recent studies have begun to address these questions (Swanson *et al* 2004; Mack *et al* 2006; Panhuis and Swanson 2006; Turner and Hoekstra 2006; 2008; Calkins *et al* 2007; Yapici *et al* 2008), our current understanding of the female role in post-copulatory sexual selection remains sparse.

The presented dissertation research examines the biochemical nature and evolutionary consequences of ejaculate-female coevolution in two sister-species of cactophilic *Drosophila*, *D. mojavensis* and *D. arizonae*. Appendix A employs crosses between these two species to identify divergent post-copulatory processes that likely are evolving rapidly. Appendices B and C use a combination of experimental and computational approaches to identify female reproductive tract proteins that may be involved in this process. Appendices D and E explore evolutionary history of a subset of these proteins within *D. mojavensis* and throughout the *Drosophila repleta* species group. The focus on female proteins contributes to the handful of studies on these molecules, as well as providing a complement to ongoing research on male seminal proteins in this system (Wagstaff and Begun 2005; 2007; Almeida and DeSalle 2008).

Review of the Literature:

An array of studies indicate that females differentiate between competing ejaculates in the post-copulatory arena. Artificial selection experiments on the length of the *D. melanogaster* female sperm storage organ demonstrated correlated evolution of male sperm length, as well as biased fertilization in crosses between divergently selected lines (Miller and Pitnick 2002). These data suggest that females select for complimentary male traits, an assertion that is the supported by the frequent observation of correlated morphology between male genitalia or sperm and female reproductive tracts amongst closely related organisms (Reviewed in Eberhard 1996). Similarly, in cases where a female is multiply mated to both a conspecific and a heterospecific male, conspecific sperm often obtain the majority of fertilizations (Reviewed in Howard 1999; Markow *et al* 2007). A competitive advantage in the native environment implies that ejaculates and female reproductive tracts are coadapted, presumably through a shared history of intersexual selection. Reduced fertilization success or perturbed reproductive outcomes in heterospecific crosses, furthermore, sometimes are observed even in the absence of competing males (Reviewed in Howard 1999; Markow *et al* 2007). This result can only be explained by a break down of reproductive traits that are coadapted between the sexes.

Intersexual selection could result in reciprocal evolutionary change between males and females by two distinct mechanisms. First, cryptic female choice could empower females to bias fertilization success towards certain males based post-copulatory biochemical cues (Eberhard 1996). This may lead to cyclical evolution of male trait and female preference, consistent with traditional models of runaway sexual selection (Fisher, 1915; 1930; Lande 1981; Kirkpatrick 1982). Alternatively, sexual conflict, or a difference in the reproductive interests of the two sexes (Parker 1979), is

predicted to result in a coevolutionary arms race between males and females (Parker 1979; Rice 1996; Gavrilets 2000). At the molecular level, both cryptic female choice and sexual conflict could lead to ongoing coeovlution and directional selection on both male and female loci (Lande 1981; Kirkpatrick 1982; Gavrilets 2000). Under regimes of sexual conflict, however, female loci also may split into two divergent alleles, halting pursuit from males until they themselves diversify (Gavrilets 2000; Gavrilets and Waxan 2002; Hayashi, Vose and Gavrilets 2007).

The biochemistry and evolution of interacting reproductive proteins has been examined most extensively in the free-spawning marine invertebrates abalone, sea urchins, and oysters. Although fertilization in these organisms occurs outside of a female reproductive tract, it is largely dependent on biochemical interactions between sperm and female chemoattractants (Kaup *et al* 2006), and sperm-egg interactions (Reviewed in Mengerink and Vacquier 2001; Swanson and Vacquier 2002). Cryptic female choice and sexually antagonistic coevolution, therefore, are predicted to guide the divergence of these molecules in a manner analogous to internal fertilizers (Reviewed in Swanson and Vacquier 2002; Clark *et al* 2006; Panhuis *et al* 2006). Male gamete recognition proteins of all these organisms, as well as a female gamete recognition proteins in abalone, exhibit signatures of adaptive evolution in interspecific comparisons suggesting ongoing molecular coevolution (Yang *et al* 2000; Galindo *et al* 2003; Mah *et al* 2005; Aagaard *et al* 2006; Levitan and Ferell 2006; Clark *et al* 2007; Moy *et al* 2008). Male gamete recognition proteins, furthermore, often exhibit signatures of either directional or diversifying selection within discrete populations (Lee *et al* 1995; Metz and Palumbi 1996; Levitan and Ferell 2006; Clark *et al* 2007; Moy *et al* 2008; Springer *et al* 2008). Although, the selective force that partitions reproductive proteins into these alternate regimes is not well understood, diversifying selection in the sea urchin sperm protein bindin is associated with populations that experience more intense sexual conflict, as

predicted by theoretical models (Levitan and Ferrell 2006; Gavrilets and Waxman 2002; Haygood 2004 Hayashi, Vose and Gavrilets 2007). Similar to internal fertilizers, however, our current understanding of the coevolution of gamete recognition proteins in marine invertebrates is limited by a paucity of studies on the female proteins involved.

In internally fertilizing organisms, fruit-flies of the genus *Drosophila* have long served as an important model system for exploring the genetics and evolution of reproductive tract interactions (Reviewed in Markow 1996; 2002; Kubli 2003; Chapman and Davies 2004; Wolfner 2007). In the genetic model *D. melanogaster,* no fewer than 138 unique proteins in an array of biochemical classes are passed from males to females during copulation (Swanson *et al* 2001; Mueller *et al* 2005; Findlay *et al* 2008). These seminal fluid proteins play integral roles in the female post-mating response by modulating oogenesis, ovulation, immune response, sperm storage, female refractoriness, and feeding behavior (Reviewed in Wolfner 2007). Although pairs or groups of interacting proteins largely await identification (but see Yapici *et al* 2008), several male proteins either undergo proteolytic cleavage in mated females (Monsma *et al* 1990; Park and Wolfner 1995, Peng *et al* 2005), or localize to specific portions of the female reproductive tract (Bertram, Neubaum and Wolfner 1996; Heifetz *et al* 2000; Ravi Ram *et al* 2005), indicating that ejaculate–female interactions are mediated biochemically by females. Signatures of directional selection, as predicted under models of sexually antagonistic coevolution and cryptic female choice, have been observed amongst both male seminal proteins (Aguadé 1998; 1999; Begun *et al* 2000; Wong *et al* 2008), and female reproductive tract proteins (Swanson *et al* 2004; Panhuis and Swanson 2006; Lawniczak and Begun 2007) in these animals.

In complement to ongoing research on the biochemistry and evolution of reproductive tract interactions in *D. melanogaster*, recent studies have sought to explore the identity and evolutionary history of male seminal proteins of the *repleta* group species, *D. mojavensis* (Wagstaff and Begun 2005; 2007; Almeida and DeSalle 2008). Differences in reproductive biology between *D. mojavensis* and *D. melanogaster* have intriguing implications for post-copulatory intersexual selection. First, *D. mojavensis* females are three to five times more promiscuous than *D. melanogaster* (Reviewed in Markow 1996). Female promiscuity could influence the evolution of reproductive proteins by intensifying selection on post-copulatory traits or elevating sexual conflict (Parker 1979; Markow 2002). *Drosophila mojavensis* females, furthermore, are known to incorporate male-derived molecules into somatic tissues and oocytes (Markow and Ankney 1984). This nutritional benefit to copulation presents a dramatic contrast to the cost of mating incurred by *D. melanogaster* females (Chapman *et al* 1995; Pitnick and García-González 2002; Kuijper, Stewart and Rice 2006; Barnes *et al* 2008). Finally, *D. mojavensis* females exhibit an insemination reaction, an opaque mass of unknown composition that forms in the uterus after every copulation (Patterson 1946). This phenomenon is thought to protect the male's nutritional investment from cuckoldry by competing males (Markow and Ankney 1984; 1988; Pitnick, Spicer and Markow 1997), and may therefore present an example of sexual conflict (Knowles and Markow 2001).

At a physiological level, ejaculate-female coadaptation has been documented extensively in natural populations of *D. mojavensis* and its sister species *D. arizonae* (MRCA ~1.5 MYA, Matzkin 2004). Specifically, crosses between geographically isolated populations within both these species produce smaller eggs than intrapopulation

crosses (Pitnick *et al* 2003), a process known to be stimulated by several components of the male ejaculate in *D. melanogaster* (reviewed in Kubli 2003; Chapman and Davies 2004; Wolfner 2007). Additionally, the insemination reaction exhibits a larger size and duration in interpopulation crosses relative to intrapopulation crosses (Knowles and Markow 2001). Finally, desiccation resistance is higher in mated than unmated females (Knowles *et al* 2004) and the magnitude of this effect differs between inter- and intrapopulation crosses (Knowles *et al* 2005). These intriguing examples of ejaculate-female dynamics indicate that this will be an exciting system to explore the molecular basis of post-copulatory intersexual selection.

Explanation of the Dissertation Format:

Appendix A explores the contribution of ejaculate-female interactions to reproductive isolation between *D. mojavensis* and *D. arizonae*. Reproductive incompatibilities are discovered in four distinct post-copulatory processes: sperm storage, sperm viability, fertilization and oviposition. In conjunction with evidence for perturbed reproductive outcomes in interpopulation crosses within species (Knowles and Markow 2001; Pitnick *et al* 2003; Knowles *et al* 2005), these data suggest molecular coadaptation of male ejaculates and female reproductive tracts in this system.

To pinpoint female molecules involved in post-copulatory intersexual selection, Appendix B employs a comparative EST approach to identify rapidly-evolving female reproductive tract proteins that may interact with the male ejaculate. The study identified over 241 candidate female reproductive molecules, the most exciting of which were

recently-duplicated secreted proteases. A total of five lineage-specific protease gene families were discovered, three of which exhibited signatures of adaptive evolution. Appendix C uses a combination of bioinformatics and biochemical approaches to specifically explore the apparent adaptive expansion of secreted proteases. *D. arizonae* female reproductive tracts are shown to exhibit a wealth of secreted serine endoproteases with diverse predicted specificities. It furthermore is demonstrated that these tissues present a highly proteolytic environment, and that enzymatic activity is regulated by mating.

Appendices D and E explore the evolutionary history of two secreted serine endoproteases gene families within four geographically isolated populations of *D. mojavensis* and throughout the *repleta* species group. Deviations from neutrality consistent with both diversifying and directional selection were observed at these loci, consistent with models of cryptic female choice and sexual conflict (Lande 1981; Kirkpatrick 1982; Gavrilets 2000; Gavrilets and Waxman 2002; Haygood 2004; Hayashi, Gavrilets and Vose 2007). Each isolated population, furthermore, exhibits distinct signatures of selection, a possible indicator of unique coevolutionary trajectories within each geographic locale.

CHAPTER 2: PRESENT STUDY

The methods, results, and conclusions of this study are presented in the papers appended to this dissertation/thesis. The following is a summary of the most important findings in this document.

Appendix A employs dark-field and floursecent microscopy to compare conspecifically and heterospecifically mated *D. mojavensis* females from three geographic populations for a range of postcopulatory traits. I show that *D. mojavensis* females mated to *D. arizonae* males oviposit fewer eggs than conspecifically mated females, and furthermore, that the vast majority of oviposited eggs remain unfertilized. These reductions in fecundity and fertility in heterospecific crosses are associated with post-mating abnormalities in female reproductive tracts, including failure in sperm storage, reduced viability of stored sperm, and failure to degrade the insemination reaction mass. The data suggest that male and female contributions to reproduction are coadapted, consistent with models of intersexual selection. They furthermore highlight post-copulatory processes as an under-explored arena for the rise of isolating mechanisms that prevent gene flow between species.

To pinpoint genes involved in both intersexual coevolution and reproductive incompatibility, Appendix B employs a comparative EST approach to identify 649 unique proteins expressed in the *D. arizonae* lower female reproductive tract. Bioinformatics analyses are then used to identify 241 secreted or transmembrane proteins in an array of biochemical classes that are candidates for interaction with the male ejaculate. Interspecific comparisons between *D. arizonae* ESTs and their *D. mojavensis*

ortholog reveal that thirty-one of these proteins exhibit elevated amino acid substitution rates, making them candidates for molecular coevolution with the male ejaculate. The most exciting candidates revealed by the EST screen, however, were three gene families of secreted proteases. I use phylogenetic inference in conjunction with maximum-likelihood analyses of positive selection to show that these gene families are specific to the *repleta* species group, and that certain residues within these proteases have undergone adaptive evolution. Observation of adaptive evolution and gene duplication amongst female reproductive molecules mirrors studies of male seminal proteins in this lineage (Wagstaff and Begun 2007; Almeida and DeSalle 2008), indicating that females are active players in the evolution of reproductive tract interactions. Furthermore, preferential duplication of secreted proteases may suggest a lineage-specific expansion of female proteolytic capacity.

Appendix C compares the evolutionary dynamics, biochemical nature, and physiological significance of secreted female reproductive serine endoproteases between *D. arizonae* and its congener *D. melanogaster*. I show that *D. arizonae* secreted female reproductive serine endoproteases not only are enriched for recent duplicates, but they also encode a greater number of enzymes with a broader range of predicted specificities than *D. melanogaster*. Isolated lumen from *D. arizonae* lower female reproductive tracts, furthermore, exhibits significant tryspin-like and elastase-like serine endoprotease acitivity in biochemical assays, while no such activity is seen in *D. melanogaster*. Finally, trypsin and elastase-like activity in *D. arizonae* female reproductive tracts is negatively regulated by mating. I suggest that the intense proteolytic environment of the

D. arizonae female reproductive tract relates to the extraordinary reproductive physiology of this species.

Appendix D examines genetic variation for a gene family of five serine endoproteases identified in Appendices B and C, within four geographically isolated populations of *D. mojavensis* and throughout the *repleta* species group. An array of polymorphism and divergence based tests, as well as permutation-based analysis of gene conversion, are used to examine the evolutionary history of these loci. My data reveal dynamic patterns of pseudogenation, copy number variation, gene conversion, and selection within each geographic locale. I furthermore use phylogenetic inference, maximum-likelihood analyses of positive selection, and permutation-based analyses of gene conversion to show these patterns extend to three other *repleta* group species. This intriguing evolutionary history has never before been documented in a reproductive protein, and suggests this gene family evolves rapidly as a functionally redundant complex.

Appendix E uses the same approaches as Appendix D to examine a second five-paralog gene family of secreted female reproductive serine proteases within populations of *D. mojavensis*. Four of five paralogs in this gene family show evidence for the emergence of unusually structured haplotypes that suggest the retention of old polymorphism. These gene genealogies furthermore are accompanied by deviations from neutrality consistent with balancing selection. This study presents the first evidence that balancing selection, a predicted outcome of mathematical models of sexual conflict, is operating on female reproductive tract proteins.

REFERENCES

Aagard JE, Yi X, MacCoss MJ, Swanson WJ. 2006 .Rapidly evolving zona pellucida domain proteins are a major component of the vitelline envelope of abalone eggs. Proc. Natl. Acad. Sci. USA 103: 17302–17307.

Aguadé M. 1998. Different forces drive the evolution of the Acp26Aa and Acp26Ab accessory gland genes in the *Drosophila melanogaster* species complex. Genetics. 150:1079–1089.

Aguadé M. 1999. Positive selection drives the evolution of the Acp29AB accessory gland protein in *Drosophila*. *Genetics* 152:543–51.

Almeida FC, Desalle R.,2008. Evidence of adaptive evolution of accessory gland proteins in closely related species of the *Drosophila repleta* group. Mol. Biol. Evol. 25:2043–2053.

Andersson M. 1994. Sexual Selection. Princeton University Press; Princeton, NJ.

Barnes AI, Wigby S, Boone JM, Partridge L, Chapman T. 2008. Feeding, fecundity and lifespan in female *Drosophila melanogaster*. Proc. Biol. Sci. 275:1675–83.

Bertram MJ, Neubaum DM, Wolfner MF (1996) Localization of the Drosophila male accessory gland protein Acp36DE in the mated female suggests a role in sperm storage. Insect Biochem Mol Biol 26: 971–980.

Calkins JD, El-Hinn D, Swanson WJ. 2007. Adaptive evolution in an avian reproductive protein: ZP3. J. Mol. Evol. 2007 65:555–563.

Chapman T, Liddle LF, Kalb JM, Wolfner MF, Partridge L. 1995. Cost of mating in *Drosophila melanogaster* females is mediated by male accessory gland products. Nature 373:241–244.

Chapman T, Davies SJ. 2004. Functions and analysis of the seminal fluid proteins of male *Drosophila melanogaster* fruit flies. Peptides 25:1477-1490.

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

Clark NL, Findlay GD, Yi X, MacCoss MJ, Swanson WJ. 2007. Duplication and selection on abalone sperm lysin in an allopatric population. Mol. Biol. Evol. 24:2081–90.

Darwin C. 1871. The Descent of Man and Selection in Relation to Sex. John Murray: London.

Eberhard WG. 1996. Female Control: Sexual Selection by Cryptic Female Choice. Princeton, New Jersey: Princeton University Press.

Fisher RA. 1915. The evolution of sexual preference. Eugenics Review 7:115–123.

Fisher RA. 1930. The genetical theory of natural selection. Oxford: Clarendon Press.

Findlay GD, Yi X, Maccoss MJ, and Swanson WJ. 2008. Proteomics reveals novel *Drosophila* seminal fluid proteins transferred at mating. P.L.o.S. Biol. 6:e178.

Galindo BE, Vacquier VD, Swanson WJ. 2003. Positive selection in the egg receptor for abalone sperm lysin. Proc. Natl. Acad. Sci. U. S. A. 100 4639–4643.

Gavrilets S. 2000. Rapid evolution of reproductive barriers driven by sexual conflict. Nature 403:886–889.

Gavrilets S, Waxman D. 2002. Sympatric speciation by sexual conflict. Proc. Natl. Acad. Sci. U. S. A. 99:10533–10538.

Hayashi TI, Vose M, Gavrilets S. 2007. Genetic differentiation by sexual conflict. Evolution 61:516–29.

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

Heifetz Y, Lung O, Frongillo EA Jr, Wolfner MF. 2000. The *Drosophila* seminal fluid protein Acp26Aa stimulates release of oocytes by the ovary. Curr. Biol. 10: 99–102.

Howard DJ. 1999. Conspecific sperm and pollen precedence and speciation. Annu. Rev. Ecol. Syst. 30:109–132.

Kaupp, UB, Hildebrand E, Weyand I. 2006. Sperm chemotaxis in marine invertebrates-molecules and mechanisms. *J Cell Physiol.* 208:487–494.

Kirkpatrick M. 1982. Sexual selection and the evolution of female choice. Evolution 36:1–12.

Knowles LL, Markow TA. 2001. Sexually antagonistic coevolution of a postmating prezygotic reproductive character in desert *Drosophila*. Proc. Nat. Acad. Sci. U. S. A. 98:8692–8696.

Knowles LL, Hernandez BB, Markow TA. 2004. Exploring the consequences of postmating-prezygotic interactions between the sexes. Proc. Biol. Sci. 271 Suppl 5:S357–S359.

Knowles LL, Hernandez BB, Markow TA. 2005. Non-antagonistic interactions between the sexes revealed by the ecological consequences of reproductive traits. J. Evol. Biol. 18:156–161.

Kubli, E. (2003). Sex-peptides: seminal peptides of the Drosophila male. Cell. Mol. Life Sci. 60:1689-704.

Kuijper B, Stewart AD, Rice WR. 2006. The cost of mating rises nonlinearly with copulation frequency in a laboratory population of *Drosophila melanogaster*. J Evol Biol. 19:1795–1802.

Lande R. 1981. Models of speciation by sexual selection on polygenic traits. Proc. Natl. Acad. Sci. U. S. A. 1981 78:3721–3725.

Lawniczak MK, Begun DJ. 2007. Molecular population genetics of female-expressed mating-induced serine proteases in *Drosophila melanogaster*. Mol. Biol. Evol. 24:1944–1951.

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

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

Mack PD, Kapelnikov A, Heifetz Y, Bender M. 2006. Mating-responsive genes in reproductive tissues of female *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U. S. A. 103:10358–10363.

Mah SA, Swanson WJ, 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.

Markow TA, Ankney PF. 1984. *Drosophila* Males Contribute to Oogenesis in a Multiple Mating Species. Science 224:302–303.

Markow TA and Ankney PF. 1988. Insemination Reaction in *Drosophila* found in species whose males contribute material to oocytes before fertilization. Evolution 42:1097–1101.

Markow TA. 1996. Evolution of *Drosophila* mating systems. Evol. Biol. 29:73–106.

Markow TA. 2002. Perspective: female remating, operational sex ratio, and the arena of sexual selection in *Drosophila* species. Evolution 56:1725–1734.

Markow TA, Reed LK, Kelleher ES. 2007. Sperm fate and function in reproductive isolation in *Drosophila.* Soc. Reprod. Fertil. Suppl. 65:155-173.

Matzkin LM. 2004. Population genetics and geographic variation of alcohol dehydrogenase (Adh) paralogs and glucose-6-phosphate dehydrogenase (G6pd) in *Drosophila mojavensis*. Mol. Biol. Evol. 21:276–285.

Mengerink KJ, Vacquier VD. 2001. Glycobiology of sperm-egg interactions in deuterostomes. Glycobiology 11:37R-43R.

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

Monsma SA, Harada HA, Wolfner MF. 1990. Synthesis of two *Drosophila* male accessory gland proteins and their fate after transfer to the female during mating. Dev Biol 142:465–475.

Mueller JL, Ravi Ram K, McGraw LA, Bloch Qazi MC, Siggia ED, Clark AG, Aquadro CF, Wolfner MF. 2005. Cross-species comparison of Drosophila male accessory gland protein genes. Genetics 171:131–43.

Miller GT, Pitnick S. 2002. Sperm-female coevolution in *Drosophila*. Science 298:1230-1233.

Moy GW, Springer SA, Adams SL, Swanson WJ, Vacquier VD. 2008. Extraordinary intraspecific diversity in oyster sperm bindin. Proc. Natl. Acad. Sci. U. S. A. 105:1993–8.

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

Panhuis TM, Clark NL, Swanson WJ. 2006. Rapid evolution of reproductive proteins in abalone and *Drosophila*. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 361:261–268.

Park M, Wolfner MF. 1995. Male and female cooperate in the prohormonelike processing of a *Drosophila melanogaster* seminal fluid protein. Dev. Biol. 171: 694–702.

Parker GA. 1970. Sperm competition and its evolutionary consequences in the insects, Biological Reviews 45:525–567.

Parker, GA. 1979. Sexual selection and sexual conflict. In Blum MS, Blum NA, editors. Sexual selection and reproductive competition in insects. London: Academic Press. pp. 123–166.

Patterson JT. 1946. A new type of isolating mechanism *in Drosophila*. Proc. Nat. Acad. Sci. U. S. A. 32:202–208.

Peng J, Chen S, Busser S, Liu H, Honegger T, Kubli E. 2005. Gradual release of sperm bound sex-peptide controls female postmating behavior in Drosophila. Curr Biol 15:207–213.

Pitnick S, Spicer GS and Markow TA. 1997. Phylogenetic examination of female incorporation of ejaculate in *Drosophila*. Evolution 51:833–845.

Pitnick S, García-González F. 2002. Harm to females increases with male body size in *Drosophila melanogaster.* Proc. Biol. Sci. 269:1821–1828.

Pitnick S, Miller GT, Schneider K, Markow TA. 2003. Ejaculate-female coevolution in *Drosophila mojavensis*. Proc. Nat. Acad. Sci. U. S. A. 270:507–512.

Ravi Ram K, Ji S, Wolfner MF. 2005. Fates and targets of male accessory gland proteins in mated female *Drosophila melanogaster*. Insect Biochem Mol Biol 35: 1059–1071.

Rice WR. 1996. Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. Nature 381:232–234.

Robertson SA. 2005. Seminal plasma and male factor signalling in the female reproductive tract. Cell. Tiss. Res. 322:43–52.

Robertson SA. 2007. Seminal fluid signaling in the female reproductive tract: lessons from rodents and pigs. J. Anim, Sci. 85:E36–E44.

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

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

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

Turner LM, Hoekstra HE. 2006. Adaptive evolution of fertilization proteins within a genus: variation in ZP2 and ZP3 in deer mice (*Peromyscus*). Mol. Biol. Evol. 23:1656-69.

Turner LM, Hoekstra HE. 2008. Reproductive protein evolution within and between species: maintenance of divergent ZP3 alleles in *Peromyscus*. Mol. Ecol. 2008 17:2616–2628.

Wolfner MF. 2007. "S.P.E.R.M." (seminal proteins (are) essential reproductive modulators): the view from *Drosophila*. Soc. Reprod. Fertil. Suppl. 183-99.

Wagstaff BJ, Begun DJ. 2005. Molecular population genetics of accessory gland protein genes and testis-expressed genes in *Drosophila mojavensis* and *D. arizonae.* Genetics 171:1083–1010.

Wagstaff BJ. Begun DJ. 2007. Adaptive evolution of recently duplicated accessory gland protein genes in desert *Drosophila.* Genetics 177:1023–1030.

Yang Z, Swanson WJ, Vacquier VD. 2000. Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. Mol. Biol. Evol. 17:1446-1455.

Yapici N, Kim YJ, Ribeiro C, Dickson BJ. 2008. A receptor that mediates the post-mating switch in *Drosophila* reproductive behaviour. Nature 451:33-37.

APPENDIX A: REPRODUCTIVE TRACT INTERACTIONS CONTRIBUTE TO ISOLATION IN *DROSOPHILA*

Kelleher ES and Markow TA. 2007. Reproductive Tract Interactions Contribute to Isolation in *Drosophila*. Fly. 1:33-37.

ABSTRACT

The process of speciation requires the development of isolating mechanisms that act as barriers to gene flow between incipient species. Such mechanisms can occur at three different levels: precopulatory or behavioral isolation, postcopulatory-prezygotic isolation occurring in the female reproductive tract, or postzygotic isolation resulting in hybrid sterility or inviability. Only by extensively studying all three types of barriers in young species pairs can we begin to understand the evolution of early reproductive incompatibilities which may be important to the speciation process. Although precopulatory and postzygotic isolation have been well described it is only recently that the female reproductive tract has been intensely examined for possible mechanisms of reproductive isolation (For a review see refs 1, 2). The types of isolating mechanisms that develop at this level and their role in speciation, therefore, remain poorly understood.

INTRODUCTION

Polyandry, internal fertilization, and sperm storage have made *Drosophila* a popular system for the study of reproductive tract interactions, and there is a range of points along the postcopulatory–prezygotic (PCPZ) trajectory at which incompatibilities could arise. Males must transfer sperm successfully and the sperm must enter sperm storage organs, remain viable, and be able to fertilize eggs. Additionally, in many species of *Drosophila* females must be stimulated by mating to oviposit.[3] These postcopulatory processes rely on functional interactions between male and female morphology[4] and molecular biochemistry.[5-7] Such interactions are determinants of reproductive success, and therefore sexual selection and intersexual coevolution have caused them to become extremely divergent between species.[8-10] The morphology of sperm and sperm storage organs and the patterns of sperm transfer and storage show extreme variation across the genus.[8]Additionally, male seminal or accessory gland proteins, and female reproductive molecules are highly divergent between species and many show signatures of adaptive evolution at the molecular level.[11-17] Such coadapated divergence predicts failures of morphological and molecular interactions in heterospecific crosses.

In this study, we examined the role of reproductive tract interactions as isolating mechanisms between the cactophilic *Drosophila*, *D. mojavensis*, and its sister species *D. arizonae* (distributions shown in Figure 1). Because this species pair is young (~ 0.8 MY, 17), partially sympatric, and will hybridize in the laboratory, it provides an excellent opportunity for identifying early-acting barriers. Additionally, both precopulatory[19-22] and

postzygotic isolation[23,24] have been examined extensively. Several clues suggest that PCPZ isolation may also play an important role in restricting gene flow between these two species. First, there is a marked reduction in the proportion of heterospecifically mated *D. mojavensis* females that produce offspring.[25] Additionally, fertile heterospecific crosses produce very few hybrids, although the level of oviposition is normal.[23] Finally, the insemination reaction, a large white mass that forms in the uterus after mating in many *Drosophila*,[26] is reportedly more severe in heterospecific crosses.[25] Although the function of the reaction mass remains unknown, it may serve to delay female remating[27,28,29] and therefore be coevolving antagonistically between the sexes due to sexual conflict.[29]

We first examined both the fecundity and fertility of homospecifically and heterospecifically mated *D. mojavensis* females from three geographically isolated populations: Anza Borrego Desert, California (AB), Santa Catalina Island, California (CI), and Ensenada de los Muertos, Mexico (EN). Upon finding evidence that productivity of heterospecific crosses was severely reduced, we examined the reproductive tracts of mated females to identify specific incompatibilities. Evidence for incompatibilities in four distinct PCPZ processes was found: sperm storage, sperm viability, fertilization, and oviposition.

MATERIALS AND METHODS

Collection and rearing. *D. mojavensis* was collected from Ensenada de los Muertos, Mexico, in January 2001, Catalina Island, California, in April 2001, and Anza Borrego Desert, California, in March 1995 and April 2002. *D. arizonae* was collected from Peralta Canyon, Arizona, in April 1997 (Figure 1). For the strains collected in Anza Borrego, the March 1995 strain was used in the offspring viability and fertilization studies, while the 2002 strain was used in the microscopy study. Both species were reared on standard opuntia-banana medium (for recipe see http://stockcenter.arl.arizona.edu/), and have similar generation times of ~19 days.[30]

Offspring viability measures. Sexually mature flies no older than nine days post-eclosion were paired in individual vials and observed until copulation. Females were then isolated and transferred daily to fresh vials of opuntia banana medium. Daily oviposition and emerging adults were quantified. Two replicates were performed.

Percentage eggs fertilized. Flies were mass-mated and the resulting eggs were collected on agar plates. Although it was not possible to verify all eggs were oviposited by mated females for this portion of the study, *D. mojavensis* females require mating for oviposition.[3] Eggs were dechorionated in 2% hypochlorite, and their nucleic acid stained with 4',6-diamidino-2-phenylindole (DAPI). Prepared eggs were examined under a fluorescent microscope (200x) to determine if they were fertilized. Fertilized eggs are easily identified by the wiry appearance of the male pronucleus, adjacent to the micropyle.[31]

Microscopy of mated uteri. Sexually mature females no older than 12 days post-eclosion were observed to mate and then isolated on opuntia-banana medium for five days. Oviposition was quantified, as was total number of emerging adults from deposited oocytes. At 5 days post mating, whole lower female reproductive tracts, including the uterus, seminal receptacle, spermathecae, parovaria, and common oviduct were removed in PBS and mounted on a glass slide. Slides were observed with a Nikon E800 upright microscope under dark-field (200x). Digital images were taken with an attached camera and SPOT image software (http://www.diaginc.com/supdownloads.asp).

Scoring of phenotypes. Females dissected 5 days post-mating were scored for three different phenotypes: sperm storage, sperm viability, and severity of the insemination reaction mass. Sperm storage and viability refer only to the seminal receptacle, as *D. mojavensis* females do not store sperm in the spermathecae.[8] We chose to dissect flies 5 days post-mating because qualitative preliminary data indicated there were clear differences in the reproductive tracts of homospecifically and heterospecifically mated females at this time point. Females with one or more sperm in the seminal receptacle were scored as storing sperm. Females with one or more motile sperm were scored as having motile sperm. Females with any evidence of a reaction mass were scored as exhibiting a mass, while females with no evidence of a reaction mass were scored as no mass. We further scored the severity of the insemination reaction was from 1 to 6: 1 – clear uterus, 2 – fluid or debris present, 3 – small mass, 4 – large mass, 5 – condensed clog-like mass, 6 – clog-like mass with decomposing oocyte.

Statistical analysis.

For offspring oviposition and adult hatchability:

A model that included female population, crosstype, population x crosstype, and replicate found no evidence for a replicate effect ($F_{6,231} = 0.0239$, $p = 0.88$). Therefore the two replicates were pooled. Descriptive statistics of pooled data are represented in Figure 2.

For dissected reproductive tracts:

Chi-squared and Fisher's exact test were applied to 2 x 2 contingency tables to determine if the proportion of females who exhibited a given postcopulatory trait was independent of whether the female was mated to a *D. mojavensis* male or a *D. arizonae* male. Specifically, for each *D. mojavensis* population, proportions of females for a bivariate phenotype (for example, sperm and no sperm) were compared between homospecific and heterospecific crosses.

RESULTS

We assessed fecundity and fertility of heterospecific and homospecific crosses by quantifying oviposition and offspring production over a 7-day period. Approximately 50% of heterospecifically mated females failed to oviposit and were excluded from further analysis as possible instances of pseudocopulation. Heterospecifically mated *D. mojavensis* females from CI and EN that did oviposit laid significantly fewer eggs than

homospecifically mated females, while AB females laid significantly more (Figure 2). The more striking pattern, however, is that fertility, as measured by the ratio of viable adults to oviposited eggs, is reduced from 60–70% in homospecific to 4–16% in heterospecific matings (Figure 2). When fertilization success was examined by staining eggs for the presence of sperm heads, the low fertility of heterospecific crosses having normal levels of oviposition was found to result from fertilization failure rather than hybrid inviability (supplementary materials). These data clearly indicate the existence of isolating mechanisms that occur in the reproductive tracts of heterospecifically mated *D. mojavensis* females.

To identify the physical basis of the observed reductions in oviposition and fertilization, we examined the reproductive tracts of mated *D. mojavensis* females five days after copulation. Specifically, the presence and motility of sperm in the seminal receptacle and the presence and appearance of the insemination reaction were scored. Oviposition and offspring production were also quantified for each dissected female. Strong evidence for mismatches between several reproductive traits of the two species was found (Table 1).

Although all homospecifically mated females contained stored sperm, no sperm were seen in a significant portion of heterospecifically mated females. Since every female who failed to store sperm produced no offspring, this incompatibility resulted in a completely infertile cross. Additionally, only a small proportion of eggs oviposited by those heterospecifically mated females with sperm ever produced offspring. Clearly, problems in sperm storage alone cannot explain the low fertility of heterospecific crosses:

an additional incompatibility must occur later. The nature of this incompatibility remains unclear, but failures in sperm release from the receptacle, or in the timing or chemistry of the fertilization process, seem probable.

For every mating type, complete sperm mortality, as evidenced by a lack of motile sperm, occurred in some proportion of females examined (Table 1). Significant population variation in this proportion suggests different populations may experience different selective pressures for sperm longevity. Additionally, females from AB show a significant increase in mortality of stored heterospecific sperm. The increase in sperm death could result from two separate processes. First, the seminal receptacle could fail to provide a hospitable environment to *D. arizonae* sperm due to an intrinsic incompatibility in the environment provided and the metabolic requirements of the sperm. Alternatively, cryptic female choice could cause females to either undernourish undesired sperm or actively release spermicidal compounds.

All populations showed a significant increase in the presence of the insemination reaction in heterospecifically mated females (Table 1). Indeed, the proportion of heterospecifically mated females that still exhibited a reaction mass 5 days post-mating is strikingly high. The difference in appearance and location of the reaction mass between homospecific and heterospecific crosses, furthermore, is a compelling demonstration of PCPZ incompatibility. Five days postmating in homospecific crosses the mass was either absent, implying it had already been degraded by the female, or it appeared as an opaque fluid in the pocketed area of the uterus adjacent to the common oviduct (Figure 3a). In contrast, the reaction mass in many heterospecifically mated females appeared as a dense

gelatinous clog, implying that *D. mojavensis* females are inefficient at degrading the reaction mass induced by the seminal fluid of *D. arizonae* males. When this clog was observed to settle near the exit of the uterus, oviposition was blocked, as evidenced by the high incidence of decaying eggs in the uteri of these females (Figure 3b).

To quantify the relationship between the reaction mass and oviposition, we used a linear regression between the two variables. The severity of the reaction mass was scored from 1 to 6, in which a ranking of 1 denoted a clear uterus and a ranking of 6 denoted a clogged uterus with a decomposing oocyte. A strong negative correlation was found ($R^2 = 0.22$, $p < 0.001$), which indicates the reduction in oviposition in heterospecific crosses can be partially explained by the formation of more severe reaction masses in these females (Figure 4).

DISCUSSION

We present clear evidence that mismatches in reproductive tract interactions contribute to isolation in *Drosophila*. The identification of isolating mechanisms in the female reproductive tract that affect sperm storage, sperm viability, oviposition, and fertilization, in two closely related sister species with partially overlapping ranges indicate that PCPZ incompatibilities potentially play an important roles in speciation. The multitude of processes that are perturbed in the reproductive tracts of heterospecifically mated females indicates that incompatibilities at this level are extremely complex and likely involve the breakdown of several intersexual epistatic interactions. Although the

nature of these interactions remains unidentified, accessory gland proteins and female reproductive molecules are likely to play an integral role due to their function in mediating postcopulatory processes.

We hypothesize that PCPZ incompatibilities result from intersexual coevolution between the male ejaculate and female reproductive tract. Interpopulation differences in sperm mortality and reaction mass size seen here (table 1) are consistent with ejaculate-female coevolution. Indeed, there is evidence for coevolution of sperm and seminal receptacle size,[32] and reaction mass induction,[29] within populations of *D. mojavensis*. The insemination reaction mass is of particular interest, as sexually antagonistic coevolution of this trait is thought to result from sexual conflict over female remating.[29] The interference of the insemination reaction with oviposition (figure 4) therefore points to a role for sexual conflict in the evolution of reproductive isolation between species.

Differences in severity and presence of isolating mechanisms between populations shown here indicate that interpopulation variability within *D. mojavensis* is relevant to reproductive isolation from *D. arizonae*. An incompatibility that affected sperm longevity was found only in females from AB, which implies that some coevolutionary trajectories may result in incompatibilities, while others may not. Additionally, although all the populations showed a reduction in stored sperm and an increase in the incidence of a persistent insemination reaction in heterospecific crosses, significant variation between populations was found in the severity of these traits.

The incompatibilities we describe do not simply result in low productivity of heterospecific matings; they are extremely costly to females. Oviposition of unfertilized

eggs is a poor use of female resources invested in gamete production. Additionally, clogged uteri are likely to permanently sterilize females, having a severe effect on their lifetime reproductive output. Although we did not explicitly address this question, it follows that these costs would select for *D. mojavensis* females who discriminate against *D. arizonae* males in terms of mate choice. Intriguingly, there is strong evidence for reinforcement in sympatry when *D. mojavensis* females are mated with *D. arizonae* males[19-21] but not for the reciprocal cross.[22] As postzygotic isolation in this direction is relatively weak,[23,24] these results imply that reproductive tract interactions should be considered a possible driving force in the evolution of sympatric behavioural isolation, in addition to hybrid sterility and inviability. Further research into the relationship between PCPZ isolation and behavioural isolation will clarify relationships between types of isolating mechanisms and the speciation process as a whole.

Acknowledgements: The authors would like to acknowledge Tom Watts for assistance, Luciano Matzkin, Laura Reed, Michael Nachman, John Ormiston, and two anonymous reviewers for generous advice on this manuscript. This research was supported by grants from the National Science Foundation. E.S.K. was supported by the NSF-IGERT in Evolutionary, Functional, and Computational Genomics at the University of Arizona.

REFERENCES

1. Howard DJ. Conspecific sperm and pollen precedence and speciation. Annu Rev Ecol Syst 1999; 30:109-132.

2. Markow TA, Reed LK, Kelleher ES. Sperm fate and function in reproductive isolation in *Drosophila*. In: Roldan E, Gomiendo R. ed. Spermatology. Notthingham, UK: Nottingham University Press, 2006: *in press.*
3. Markow TA. Evolution of *Drosophila* Mating Systems. Evolutionary Biology; 29:73-106.

4. Miller GT, Pitnick S. Functional significance of seminal receptacle length in *Drosophila melanogaster*. J Evol Biol 2003; 16: 114–126.
5. Wolfner MF. The gifts that keep on giving: physiological functions and evolutionary dynamics of male seminal proteins in *Drosophila*. Heredity 2002; 88:85–93.

6. Chapman T, Davies SJ. Functions and analysis of the seminal fluid proteins of male *Drosophila melanogaster* fruit flies. Peptides 2004; 25:1477–1490.

7. Kubli E. Sex-peptides: seminal peptides of the *Drosophila* male. Cell Mol Life Sci 2003; 60:1689-704.

8. Pitnick S, Markow TA, Spicer GS. Evolution of multiple kinds of female sperm-storage organs in *Drosophila*. Evolution 1999; 53**:**1804**–**1822.

9. Miller GT, Pitnick S. Sperm–female co-evolution in *Drosophila*. Science 2002; 269: 1230–1233.

10. Clark AG, Aguade M, Prout T, Harshman LG, Langley CH. Variation in sperm displacement and its association with accessory gland protein loci in *Drosophila melanogaster*. Genetics 1995; 139:189–201.

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

12. Kern AD, Jones CD, Begun DJ. Molecular population genetics of male accessory gland proteins in the *Drosophila simulans* complex. Genetics 2004; 167:725–725.

13. Panhuis T, Swanson WJ. Molecular evolution and population genetics of candidate female reproductive genes in *Drosophila.* Genetics 2006; 173:2039-2047.

14. Schully SD, Hellberg ME. Positive Selection on Nucleotide Substitutions and Indels in Accessory Gland Proteins of the Drosophila pseudoobscura Subgroup. J Mol Evol 2006; 62:793-802.

15. Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proc Nat Acad Sci 2001; 13:7375–7379.

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

17. Wagstaff BJ, Begun DJ. Molecular population genetics of accessory gland protein genes and testes-expressed genes in *Drosophila mojavensis* and *D. arizonae*. Genetics 2005; 17:1083–1101.

18. Reed LK, Nyboer M, Markow TA Evolutionary relationships of *Drosophila mojavensis* geographic host races and their sister species *Drosophila arizonae.* Molecular Ecology, *in press.*

19. Wasserman M, Koepfer HR. Character displacement for sexual isolation between *Drosophila mojavensis* and *Drosophila arizonensis.* Evolution 1977; 31:812–823.

20. Markow, TA. Courtship behavior and control of reproductive isolation between *Drosophila mojavensis* and *Drosophila arizonensis.* Evolution 1981; 35:1022–1026.

21. Massie KR, Markow TA. Sympatry, allopatry and sexual isolation between *Drosophila mojavensis* and *D. arizonae.* Hereditas 2005; 142:51–54.

22. Massie KR. Sexual isolation between *Drosophila mojavensis* and *Drosophila arizonae*. Masters thesis, University of Arizona 2006.

23. Ruiz A, Heed WB, Wasserman M. Evolution of the *mojavensis* cluster of cactophilic *Drosophila* with description of two new species. J Hered 1990; 81:30–42.

24. Reed LK, Markow TA. Early events in speciation: Polymorphism for hybrid male sterility in *Drosophila*. Proc Nat Acad Sci 2004; 101:9009-9012.

25. Baker MR. A study of the isolating mechanisms found in *Drosophila arizonensis* and *Drosophila mojavensis*. University of Texas Publication 1947; 4752:78–115.

26. Patterson JT. A new type of isolating mechanism in *Drosophila*. Proc Nat Acad Sci 1946; 32:202–208.

27. Markow TA, Ankney P. Insemination reaction in *Drosophila*: A copulatory plug in species showing male contribution to offspring. Evolution 1988; 42:1097-1100.

28. Pitnick S, Spicer G, Markow TA. A phylogenetic examination of female incorporation of ejaculate in *Drosophila.* Evolution 1997; 51:833-845.

29. Knowles LL, Markow TA. Sexually antagonistic coevolution of a postmating-prezygotic reproductive character in desert *Drosophil*a. Proc Nat Acad Sci 2001; 98:8692–8626.

30. Markow TA, O'Grady PM. *Drosophila*: A guide to species identification and use. London, UK: Elsevier, 2006.

31. Karr TL. Intracellular sperm/egg interactions in *Drosophila*: a three-dimensional structural analysis of a paternal product in the developing egg. Mech Dev. 1991; 34:101-111.

29. Pitnick S, Miller GT, Schneider K, Markow TA. Ejaculate–female coevolution in *Drosophila mojavensis*. Proc Biol Sci 2003; 270:1507–1512.

TABLES

	Female population		
	Anza Borrego	**Santa Catalina Island**	**Ensenada de los Muertos**
N (homo)	23	21	20
N (hetero)	26	44	20
Reaction mass (homo)	9 (39%)	1 (5%)	5 (25%)
Reaction mass (hetero)	23 (88%)	23 (52.3%)	18 (90%)
P-value	**<0.001 (0.0003)*** **	**<0.001 (0.0001)*** **	**<0.001 (0.00003)*** **
Sperm storage (homo)	23 (100%)	21 (100%)	20 (100%)
Sperm storage (hetero)	14 (54%)	8 (18%)	11 (55%)
P-value	**NA (0.0001)*** **	**NA (7.1e-11)*** **	**NA (0.0006)*** **
Sperm motility (homo)	11 (48%)	14 (67%)	16 (80%)
Sperm motility (hetero)	1 (7%)	3 (38%)	8 (73%)
P-value	**<0.025 (0.01)***	<0.2 (0.15)	<1 (0.5)

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Table 1. Incidence of sperm storage, sperm mortality, and reaction mass. Incidence of the insemination reaction mass, stored sperm in the seminal receptacle, and motile sperm in the seminal receptacle for homospecifically and heterospecifically mated *D. mojavensis* females from Anza Borrego Desert, Santa Catalina Island, and Ensenada de los Muertos. *P*-values for X^2 and Fisher's exact test (parentheses) for differences between homospecific and heterospecific crosses. NA indicates X^2 was inappropriate to the data.

FIGURES

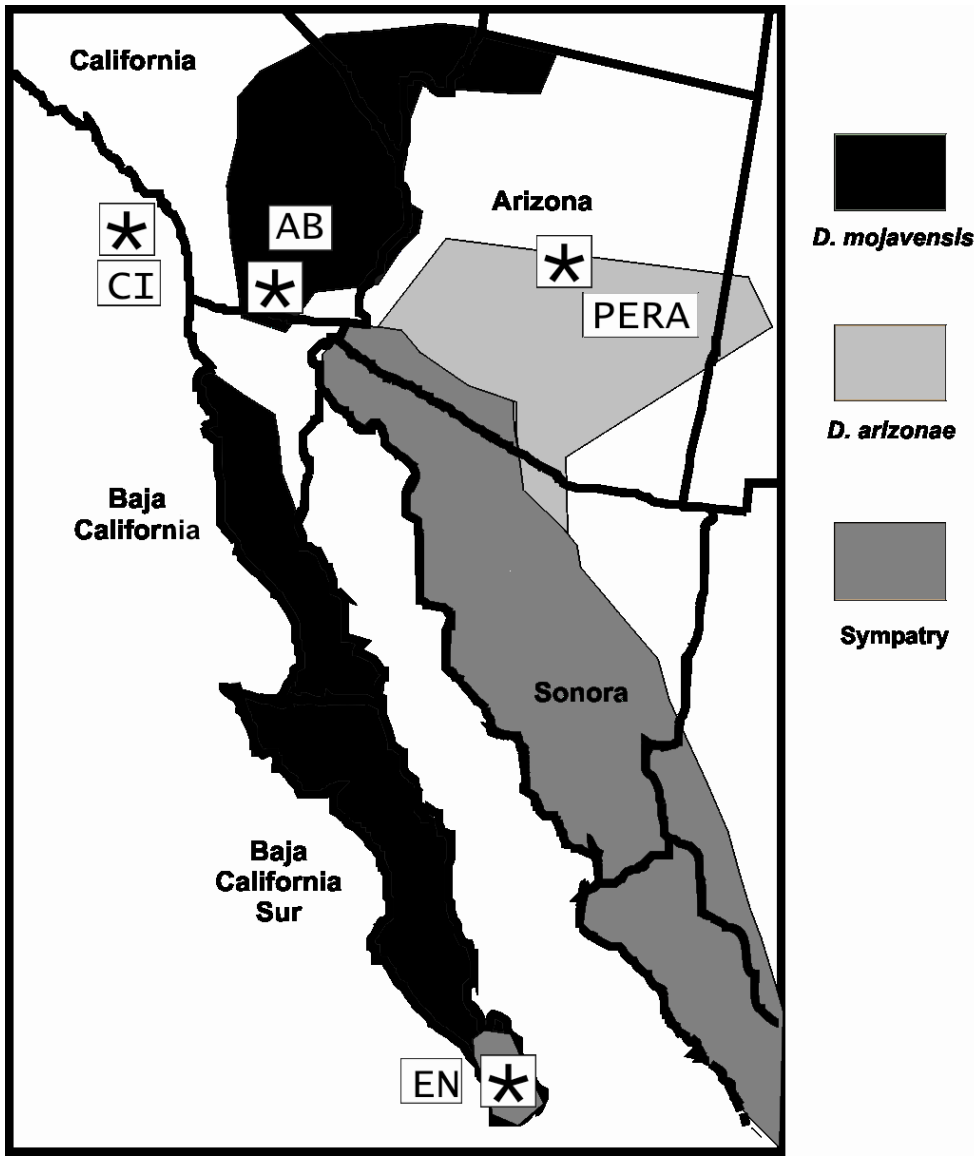


Figure 1. Species distributions of *D. mojavensis* and *D.arizonae*. Three allopatric and one sympatric population of *D. mojavensis* are indicated. One continuous population of *D. arizonae* is indicated.

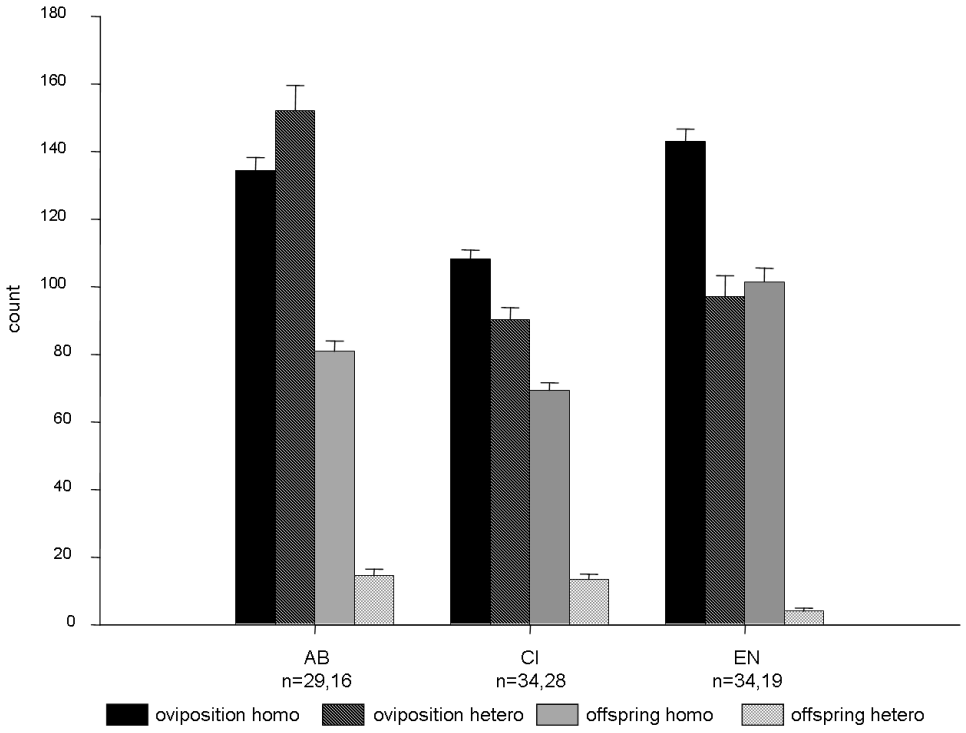


Figure 2. Reproductive output of homospecific and heterospecific crosses. Oviposition (average number of fertilized eggs) and offspring production (average number of viable adults) for homospecifically and heterospecifically mated *D. mojavensis* females from Anza Borrego Desert (AB), Santa Catalina Island (CI), and Ensenada de los Muertos (EN). D. arizonae males denoted by (A). Samples sizes for the homospecific and heterospecific cross are indicated, error bars indicate standard error (SE).

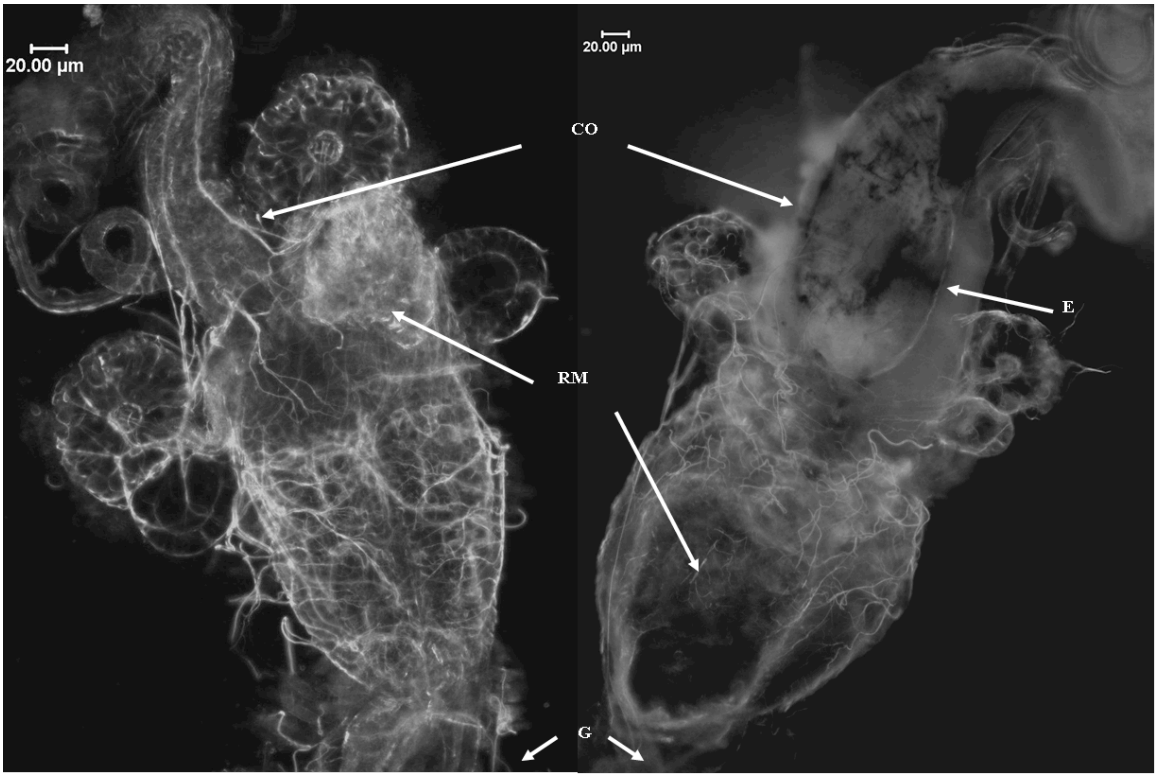


Figure 3. The reaction mass of a homospecifically and heterospecifically mated female. Reproductive tracts of homospecifically (left-panel) and heterospecifically (right-panel) mated *D. mojavensis* females from Santa Catalina Island five days post-copulation. Common oviducts (CO), reaction masses (RM), external genitalia (G), and eggs (E) are indicated.

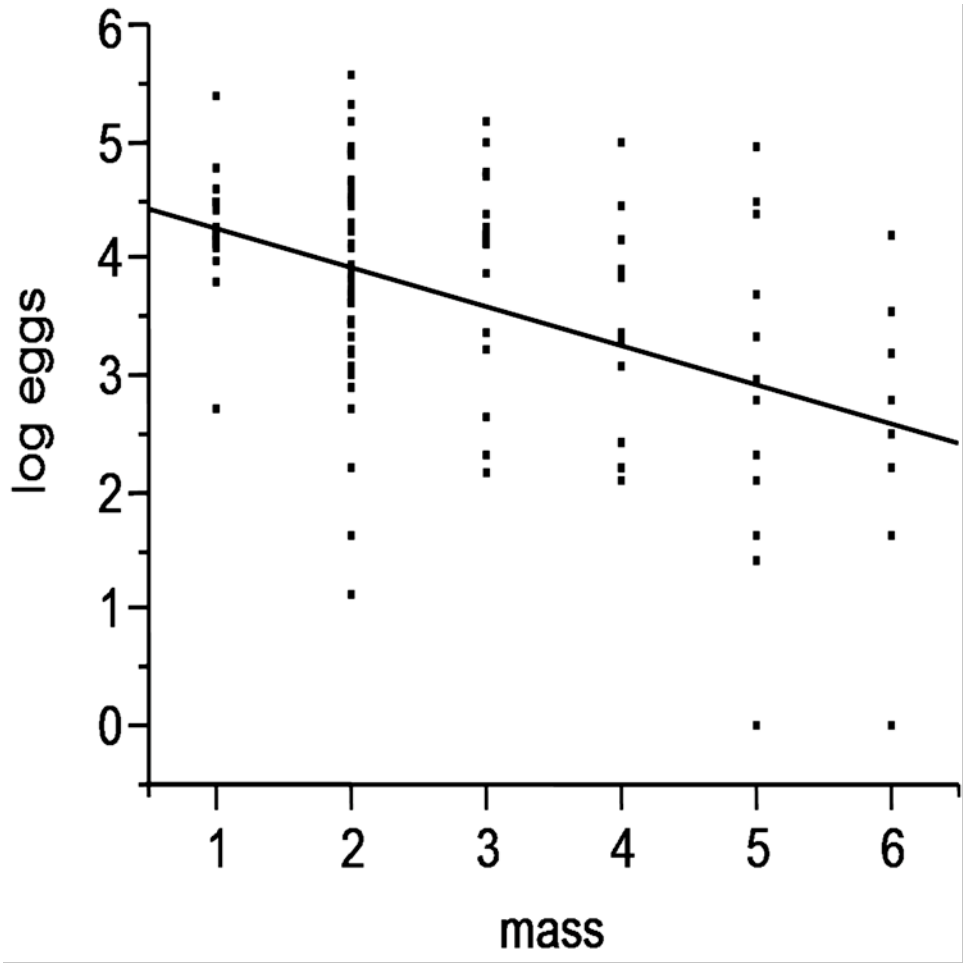
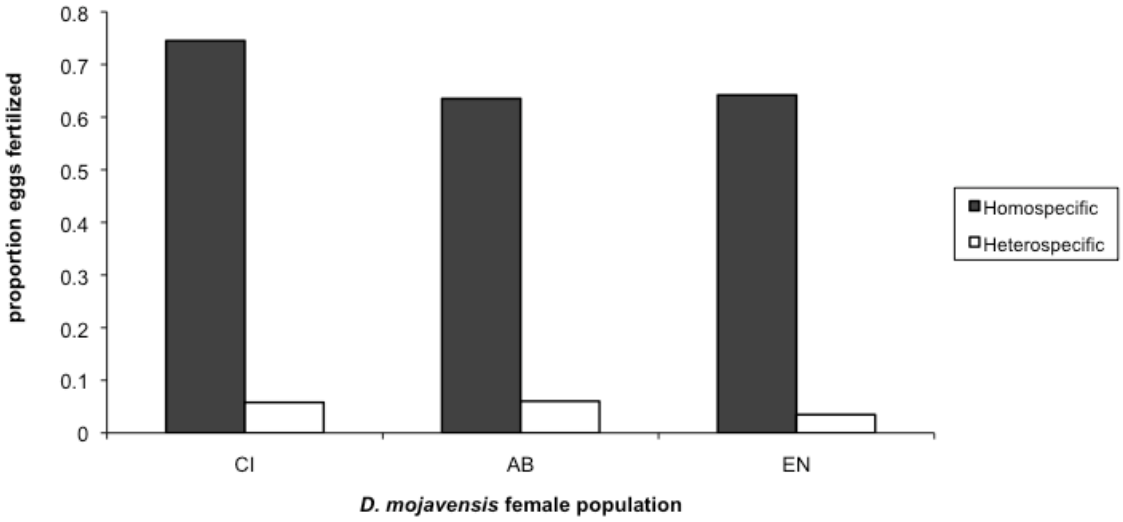


Figure 4. The negative correlation between the reaction mass and oviposition. Mass severity was ranked from 1 to 6. Log transformation of oviposition quantity. $F_{1,109}$ = 29.87, $p < 0.0001$, $R^2 = 0.22$.

SUPPLEMENTARY DATA



Supplementary Figure 1. Fertilization success. The proportion of oviposited eggs successfully fertilized in homospecifically and heterospecifically mated *D. mojavensis* females from Santa Catalina Island (CI), Anza Borrego Desert (AB), and Ensenada de los Muertos (EN). Fertilization success was determined by DAPI staining for the presence of a sperm head. Females from all populations showed a significant difference in fertilization success between the two cross types as determined by Pearson's *X2* ($p < 0.001$).

APPENDIX B: GENE DUPLICATION AND ADAPTIVE EVOLUTION OF DIGESTIVE PROTEASES IN *DROSOPHILA ARIZONAE* FEMALE REPRODUCTIVE TRACTS

**This appendix has been published as an open access article:

Kelleher ES, Swanson WJ, and Markow TA. 2007. Gene Duplication and Adaptive Evolution of Digestive Porteases in *Drosophila arizonae* Female Reproductive Tracts. P.L.o.S Genetics. 3:e148.

ABSTRACT

Background. It frequently has been postulated that intersexual coevolution between the male ejaculate and the female reproductive tract is a driving force in the rapid evolution of reproductive proteins. The dearth of research on female tracts, however, presents a major obstacle to empirical tests of this hypothesis.

Methodology/Principle Findings. Here we employ a comparative EST approach to identify 241 candidate female reproductive proteins in *Drosophila arizonae*, a *repleta* group species in which physiological ejaculate-female coevolution has been documented. Thirty one of these proteins exhibit elevated amino acid substitution rates, making them candidates for molecular coevolution with the male ejaculate. Strikingly, we also discovered 12 unique digestive proteases whose expression is specific to the *Drosophila arizonae* lower female reproductive tract. These enzymes belong to classes most commonly found in the gastrointestinal tracts of a diverse array of organisms. We show that these proteases are associated with recent, lineage-specific gene duplications in the *Drosophila repleta* species group, and exhibit strong signatures of positive selection.

Conclusions/Significance. Observation of adaptive evolution in several female reproductive tract proteins indicates they are active players in the evolution of reproductive tract interactions. Additionally, pervasive gene duplication, adaptive evolution, and rapid acquisition of a novel digestive function by the female reproductive

tract points to a novel coevolutionary mechanism of ejaculate-female interaction.

INTRODUCTION

Extensive research across a broad range of taxa has revealed that the proteins involved in sexual reproduction often evolve rapidly due to positive selection [reviewed in 1-3]. Although the selective forces that underlie this pattern remain unclear, it frequently has been postulated that adaptive evolution of reproductive proteins may result from intersexual coevolution [1-3]. Indeed, this has been demonstrated in the fertilization proteins of the free-spawning marine gastropod abalone, in which the male protein, lysin, and its female receptor, vitelline envelope receptor for lysin (VERL), both exhibit signatures of adaptive evolution [4-7]. In internally fertilizing organisms however, such as mammals or insects, the biochemical interactions between male and female reproductive proteins may be vastly more complex. Reproductive outcomes depend not only on interactions between male and female gamete proteins, but additionally on interactions between male seminal proteins and proteins in the lumen of a female's reproductive tract [8-11].

Fruit-flies of the genus *Drosophila* provide an important model system for exploring the function and evolution of reproductive tract interactions [reviewed in 9-12]. In *D. melanogaster*, the male ejaculate is comprised of just under 100 proteins, several of which are known to stimulate important processes in mated females such as ovulation, oogenesis, and sperm storage [reviewed in 9-11]. Several male proteins either undergo proteolytic cleavage in mated females [13-15], or localize to specific portions of the female reproductive tract [16-18], indicating that ejaculate-female interactions are

mediated biochemically by females. Between species, rapid changes in ejaculate composition frequently have resulted in lineage-specific seminal proteins [19- 21], many of which may be novel coding sequences [22]. Additionally, molecular evolutionary studies indicate that a significant portion of this ejaculate is subject to positive selection in the *melanogaster* [23-25], *obscura* [26], and *repleta* species groups [27].

By comparison, the female side of reproductive tract interactions has received little attention. Female reproductive tract proteins have been identified transcriptionally only in *D. simulans* [28], and their functions remain entirely unknown. Furthermore, although several female reproductive tract proteins [28, 29] and egg membrane proteins [30] show evidence of positive selection, these analyses largely have been confined to the *melanogaster* species group. It is unclear therefore, how diversity in female reproductive physiology and mating system across the genus [reviewed in 12, 31] is reflected in their reproductive proteins. This overall paucity of research on females presents a major obstacle to understanding the evolution of ejaculate-female interactions and the role of intersexual dynamics in the divergence of reproductive proteins.

Here we use a comparative Expressed Sequence Tag (EST) approach to characterize candidate female reproductive tract proteins in *D. arizonae. Drosophila arizonae* is a *repleta* group species that exhibits important differences from the *melanogaster* group in mating system and female physiology. *Drosophila arizonae* females remate daily, while *D. simulans* females wait several days before remating [12]. Female promiscuity may affect the evolution of reproductive proteins by increasing the number of competing male ejaculates [32]. Females of *D. arizonae* additionally exhibit

two remarkable post-mating physiological processes not seen in the *melanogaster* group. First, they incorporate peptide components of the male ejaculate into somatic tissues and oocytes [33], an adaptation which may help defray the cost of egg production during periods of resource limitation [34]. Second, they exhibit an insemination reaction, an opaque white mass of unknown biochemical composition that forms in the female uterus after copulation [35].

By comparing post-mating outcomes in inter and intra-population crosses, several studies have presented evidence for ejaculate-female coevolution in natural populations of *D. arizonae* and its sister species *D. mojavensis* (MRCA ~1.5 MYA [36]) [37-40]. Intrapopulation crosses of both species produce larger eggs than interpopulation crosses [37], a process known to be stimulated by several components of the male ejaculate in *D. melanogaster* [reviewed in 9-11]. Additionally, the insemination reaction exhibits a larger size and duration in interpopulation crosses relative to intrapopulation crosses, suggesting this trait is subject to sexually antagonistic coevolution [38]. Finally, desiccation resistance is higher in mated than unmated females [39], and the magnitude of this effect differs between inter- and intrapopulation crosses [40]. Such extensive evidence for physiological coevolution indicates this will be an exciting system to explore the molecular basis of reproductive tract interactions.

Our study identifies 241 candidate female reproductive proteins in *D. arizonae*, of which 31 show elevated rates of amino acid substitution suggestive of adaptive evolution. Unexpectedly, we also discovered three lineage-specific gene families of digestive proteases whose expression is specific to the lower female reproductive tract. These

proteins exhibit strong signatures of adaptive evolution, and selected sites cluster near functionally important amino acids. The implications of these findings for ejaculate-female interactions and intersexual coevolution are discussed.

RESULTS AND DISCUSSION

Functional Classes of Female Reproductive Proteins. We sequenced a total of 2,304 ESTs derived from the *D. arizonae* lower female reproductive tract (parovaria, oviduct, spermathecae, seminal receptacle, uterus) representing 649 unique proteins (for a complete list see Supplementary Materials online). Of particular interest are proteins found on cell surfaces or in the lumen of this tissue, which interact directly with the male ejaculate and likely play an integral role in reproductive tract interactions [28]. We therefore designate candidate female reproductive proteins as those that exhibit secreted signal sequences, or transmembrane domains. The gross functional composition of the 241 candidate female reproductive proteins identified in this study (Fig. 1) are similar to those of *D. simulans* [28], and include transport, signal transduction, and proteolysis.

Rapid Evolution of Female Reproductive Proteins. To explore the evolutionary histories our candidate female reproductive proteins, we calculated the ratio of replacement to silent substitutions (d_N/d_S) between our *D. arizonae* ESTs and their orthologs in the *D. mojavensis* genome. Candidate female reproductive proteins exhibit significantly larger d_N/d_S values than intracellular proteins in our data set (median test,

p>.0001), suggesting that these proteins evolve more rapidly than their intracellular counterparts. This elevated rate of amino acid substitution is predicted if adaptive evolution of secreted and transmembrane proteins is a frequent consequence of molecular coevolution with components of the male ejaculate.

Under strict neutrality, only $d_N/d_S >> 1$ can be considering robust evidence of adaptive evolution. While several of our candidate genes show $d_N/d_S > 1$, none of these tests is statistically significant (Table 1). A literature survey has shown, however, that 95% genes that exhibit a pairwise $d_N/d_S > 0.5$ contain a class of sites with $d_N/d_S >> 1$ [28]. Of 227 pairwise comparisons, 31 (14%) were identified with $d_N/d_S > 0.5$, indicating they are likely experiencing positive selection (Table 1). This result is largely independent of gene duplication, as the estimated frequency of adaptive evolution it is still 13% when recent duplicates are excluded from the data set.

On a functional level, several protein classes that commonly occur in seminal and fertilization proteins, including lipases, lectins, glycoproteins and proteases, are found in our candidates for adaptive evolution (Table 1). Roughly half of these 31 candidates, however, have no known function, and several others belong to functional classes that are not commonly represented among reproductive proteins. Proteins with unusual or unknown functions make excellent candidates for discovering genes which have acquired novel functions in a biochemical network which likely evolves rapidly. Future studies of these 31 candidates will yield significant insight into the function and evolution of reproductive tract interactions in the *repleta* species group.

Gene Duplication in Female Reproductive Proteins:

Gene duplication plays an integral role in the evolution of *D. arizonae* female reproductive tract proteins. Specifically, 47% (16) of all secreted proteases in *D. arizonae* female reproductive tracts have at least one closely related paralog that also is expressed in these same tissues. Duplication events have been extremely recent; as multiple, tandemly-duplicated paralogs in the *D. mojavensis* genome correspond to only a single gene in *D. virilis*, the most closely related fully sequenced outgroup (MRCA ~23 MYA reviewed in [41]). We therefore estimate that the duplication rate of secreted proteases expressed in *D. arizonae* tracts is 0.0298 (duplications per gene per MYR, see materials and methods), which is 21-fold higher than the genome wide estimate for *D. melanogaster* (.0014 [42]). Although the selective forces involved are yet obscure, such recent and pervasive gene duplication has not been seen in any class of reproductive protein yet studied, including *D. simulans* female reproductive proteins [28].

Four (of 16) duplicated proteases have resulted from two single gene duplication events. The remaining 12 duplicated proteases, however, are associated with small lineage-specific gene families. Each family contains 4-6 tandemly duplicated paralogs in the genome of *D. mojavensis* that are syntenic to a single ortholog in the genome of *D. virilis* (Fig. 2). For brevity, we hereafter refer to these three families of tandem duplicates as protease gene family 1, 2 and 3. Phylogenetic analysis of *D. arizonae* ESTs, and coding sequences from the genomes of *D. mojavensis*, *D. virilis*, and *D. grimshawi* (http://rana.lbl.gov/drosophila), reveals the majority of these tandem duplicates in the *D. mojavensis* genome have a *D. arizonae* ortholog that is expressed in the lower female

reproductive tract (Fig. 3). This strongly suggests that the gene duplication events relate in some way to the reproductive function of these proteases. Indeed, RT-PCR of all three gene families reveals that in adult *D. arizonae* these genes are exclusively expressed in the lower female reproductive tract (Fig. 4). Gene copies present in the *D. mojavensis* genome that do not correspond to *D. arizonae* ESTs are likely not highly expressed.

While the function of these duplicated proteins in *D. arizonae* female reproductive tracts is unknown, they are often similar or identical in their key amino acid residues to several families of digestive proteases found almost exclusively in gastrointestinal tracts (Table 2). Specifically, protease gene families 1 and 2 share appreciable homology with trypsin, chymotrypsin, and elastase, serine-endopeptidases commonly found in digestive tracts of both insects and mammals [reviewed in 43]. While, serine endopeptidases can also function in immune signaling cascades across a broad array of organisms, such proteases generally have secondary protein-protein interaction domains that allow for localized regulation of physiological responses [44]. No such domains are seen in either protease gene family 1 or 2, suggesting these proteases exhibit a primarily digestive function. Similar to the two families of serine endopeptidases, protease gene family 3 contains zinc-metalloendoproteases very similar to astacin, a prominent digestive enzyme in the crayfish midgut [reviewed in 45]. The reproductive tract-specific expression of these proteases, coupled with recent, lineage-specific gene duplications, suggest that *D. arizonae* female reproductive tracts recently have acquired a novel digestive function. Digestive enzymes in female reproductive tracts likely have important implications for male reproductive success, and therefore, the

evolution of the male ejaculate.

Adaptive Evolution of Digestive Proteases:

There is compelling evidence that directional selection has played an important role in the evolution of reproductive tract-specific secreted digestive proteases in *D. arizonae* females. All three families of digestive proteases exhibit a class of sites whose ratio of non-synonymous to synonymous substitutions (d_N/d_S) is significantly greater than the neutral expectation of 1 (Table 2). d_N/d_S values for these selected sites range from 2 to 11.96, indicating certain amino acids in these proteins have experienced strong positive selection. Notably, the two single gene duplication events show no evidence of adaptive evolution (Table 2), indicating that directional selection has been exclusive to the lineage-specific families of digestive proteases.

In order to interpret selection in terms of both duplication and speciation events, we used the PAML free ratios model [46] to estimate d_N/d_S along every branch in each of the three phylogenies (Fig. 3). Positive selection associated with three different speciation events suggests that ongoing changes in the biochemical environment of the female reproductive tract, including possible male contributions to this environment, have resulted in adaptive evolution in some of these proteins. A total of five gene duplication events are also immediately followed by a period of positive selection in one of the paralogous branches ($d_N/d_S > 1$), indicating neofunctionalization of a duplicate gene copy. The other seven duplication events however, are followed by elevated amino acid substitution rates ($d_N/d_S = .2\text{-}1$) but no evidence of adaptive evolution. This suggests that

relaxed constraint created by functional redundancy between paralogs has also played an important role in the evolution of these gene families.

Evidence for adaptive amino acid evolution in duplicated genes implies that selection has acted to diversify the paralogs functionally. Indeed, in all three of the protease gene families, polar, nonpolar, and charged amino acids are seen to inhabit the same selected site in different paralogs. This indicates that directional selection has resulted in recurrent and radical amino acid substitutions, likely affecting the structure and function of the encoded proteins. By mapping selected sites onto predicted molecular structures, it is possible to make more specific inferences about how the biochemical function of these enzymes has been impacted by adaptive evolution. In the two families of serine endopeptidases (protease gene families 1 and 2), positive selection clusters near the catalytic triad: the three amino acids essential for proteolytic function [reviewed in 46] (Fig. 5). Furthermore, in protease gene family 1, positive selection is found adjacent to, and in one case synonymous with, three amino acid sites known to effect substrate specificity [reviewed in 47]. Collectively, these data indicate that directional selection has acted to diversify the catalytic activity of both families of serine endoproteases, and that protease gene family 1 has concomitantly undergone adaptive evolution for increased breadth in substrate specificity. Future functional studies of these enzymes, particularly in terms of how they interact with the male ejaculate, will yield significant insight into the selective pressures that underlie diversification of these extraordinary gene families.

Implications:

Our most striking result was the observation of three lineage-specific radiations of secreted digestive proteases in *D. arizonae* female reproductive tracts. Although the biological significance of these gene duplications is yet unclear, they may relate to two unusual physiologies exhibited by both *D. arizonae* and *D. mojavensis* females. First, the insemination reaction must be degraded by females prior to oviposition or remating [35], a process which could require specialized digestive machinery. Second, female incorporation of ejaculate-derived protein, as observed in *D. arizonae* and *D. mojavensis*, could be facilitated by degrading seminal proteins and/or sperm into smaller fragments that are more easily absorbed.

Regardless of their physiological function, lower female reproductive-tract specific expression of digestive enzymes points to a novel form of ejaculate-female interaction, in which females may actively degrade, rather than process or activate [13-15], protein components of the male ejaculate. Digestion of seminal proteins or sperm would undoubtedly have important implications for male reproductive success, predicting an evolutionary response from males. Indeed, the association of these proteases with recent gene duplications and strong signatures of adaptive evolution suggests they are involved in an intersexual arms race. Exploring the male side of this interaction therefore, is an important avenue of future research.

The 31 candidates for adaptive evolution also have important implications for reproductive tract interactions and intersexual coevolution. Roughly half of these proteins have no known function or conserved domain, suggesting they are enriched for novel biochemical functions. Additionally, the candidates include several classes of proteins

that have not been implicated previously in reproductive tract interactions. Particularly intriguing are three transmembrane proteins with the conserved transporter domain MFS_1, for inorganic solutes (Table 1). Although the biochemical composition of the *Drosophila* ejaculate is largely unknown outside of its protein constituents, females of several species incorporate ejaculate-derived phosphorus into somatic tissues and oocytes [48]. It is unclear if these transporters underlie such a process in *D. arizonae*. Their presence and evolutionary history point, however, to non-peptide biochemical interactions in female reproductive tracts which also may evolve rapidly.

If divergence of reproductive proteins is driven by intersexual dynamics, particularly sexually antagonistic coevolution [49-51], species with more promiscuous mating systems are predicted to exhibit comparatively more adaptive evolution in their reproductive proteins. *Drosophila arizonae* is significantly more promiscuous than its previously examined congener *D. simulans* [28], and, consistent with the prediction, we find evidence that this difference in mating system may be reflected in the evolution of their female reproductive proteins. Specifically, we observed that candidate female reproductive proteins in our data set exhibit higher d_N/d_S values than intracellular proteins, while this effect was not seen in similar comparisons between *D. simulans* and *D. melanogaster* [28]. Additionally, the estimated frequency of adaptive evolution in *D. arizonae* female reproductive tract proteins (14%) is significantly higher (Fisher's Exact Test p = .003) than that of *D. simulans* (5%) [28]. Although the experimental approach for these two studies was quite similar, differences in divergence times between *D. arizonae* and *D. mojavensis* (~1.5 MYA, [36]), and *D. simulans* and *D. melanogaster* (~

3 MYA, [52]), could result in more stochastic influence on our measures of d_N/d_S. Firm conclusions about the effect of mating system on the evolution of female reproductive proteins therefore requires further empirical testing across a broader array of taxa.

Although the function and evolution of male seminal proteins have been researched extensively in both insects and mammals, our understanding of the female reproductive tract proteins with which they interact remains sparse. Our data, as well as previous research in the *melanogaster* group [28, 29], indicate that rapid evolution is common amongst female reproductive tract proteins. We furthermore present compelling evidence that differences in female physiology and possibly mating system between *Drosophila* species are reflected in their reproductive tract proteins. Our research indicates that female reproductive proteins are active players in reproductive tract interactions, and that rapid evolution of seminal proteins must be considered in terms of their relationship with female counterparts.

MATERIALS AND METHODS

Tissue harvesting. *D. arizonae* used in this study were collected in December, 2005 in Tucson, AZ by E.S.K. A total of 873 lower reproductive tracts (parovaria, oviduct, spermathecae, seminal receptacle, uterus) were dissected from mature adult females 9 days or older. In order to maximize transcriptional diversity obtained, dissected females were sampled from a diverse array of mating states. 662 of the females were from population bottles, while approximately 40 females were dissected from each of the

following treatments: virgin, homospecifically mated 4-8 hours post-copulation, homospecifically mated 24 hours post-copulation, heterospecifically (to *D. mojavensis*) mated 4-8 hours post-copulation, and heterospecifically mated 24 hours post-copulation.

Library Construction. The harvested tracts were pooled into four separate aliquots of TRIZOL® reagent (Invitrogen) and total RNA was extracted according to manufacturer instructions. Quality of these samples was verified with an Agilent 2100 bioanalyzer, at which point they were pooled. mRNA enrichment was achieved by binding poly-A tails on Oligotex® (Qiagen) spin columns. Quality of enriched mRNA was verified on with an Agilent 2100 bioanalyzer, and the total yield (1.5 μg) was used for library construction with the Cloneminer® cDNA library construction kit (Invitrogen). Approximately 300,000 CFUs were obtained with an estimated insert size of 1kb. Of these clones, 10,000 were picked with a QBOT (Genetix) operated by the Arizona Genomics Institute. 1,920 of these clones were sequenced bidirectionally, and an additional 384 were sequenced exclusively from their 5' ends. All sequencing was done on at the Arizona Genomics Institute on an ABI 3700 DNA analyzer with big-dye terminator chemistry. All sequences for this study are available under GenBank Accession Nos. EV41299147751410 to EV41383447752253

Sequence Data Analysis. Base calling and assembly were implemented in Phred and Phrap [53]. All bases with a Phred quality score below 20 (99% accurate) were excluded from further analysis. The estimated frequency of sequencing errors in included bases

was .04%. BLASTN [54] (e-value = .01) against the GLEANR coding sequence annotations (from CAF1 assembly http://rana.lbl.gov/drosophila/) of the *D. mojavensis* genome was used to identify orthologs of *D. arizonae* ESTs. For ESTs with no good BLASTN hit to annotated coding sequence, BLASTN (e-value = .01) was implemented against the complete CAF1 assembly of the *D. mojavensis* genome. ESTs with BLAST hits in the *D. mojavensis* genome that contained long open reading frames (ORF) were used to annotate additional genes in *D. mojavensis* by eye. No examples of ESTs with long open reading frames but no good BLASTN hit the *D. mojavensis* genome were identified.

Translations of these coding sequences were used to identify secreted proteins and cell surface receptors using SignalP [55], and transmembrane proteins using TMHMM [56]. Conserved protein family (Pfam) domains were identified with hmmpfam [57]. Gene Ontology (GO) terms [58] were obtained from FlyBase for *D. melanogaster* homologs, or based on conserved Pfam domains if no *D. melanogaster* homolog was found. For explicit definitions of GO terms see http://www.geneontology.org/.

In total, the *D. arizonae* ESTs corresponded to 649 unique proteins in the D. mojavensis genome. The orthologous genes were aligned using CLUSTALW [59] and alignment accuracy was verified by eye. Maximum-likelihood estimates of non-synonymous substitutions rate (d_N), synonymous substitution rate (d_S), and the ratio of non-synonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site (d_N/d_S), were obtained from PAML [46]. For duplicated genes, only reciprocally monophyletic homologs were compared in pairwise analyses.

Sequence Analysis of Multigene Families. Sequence data for *D. arizonae* was obtained from the EST library, while sequences from *D. mojavensis*, *D. virilis* and *D. grimshawi* were obtained from their unpublished, publicly available genomes (http://rana.lbl.gov/drosophila/). GENECONV was used to test for gene conversion between paralogs, using the method of Sawyer [60]. No examples of gene conversion were found. Phylogenetic reconstruction of multigene families was implemented in Mr. Bayes v3.0b4. Netsted maximum-likelihood models of codon evolution were implemented in the codeml program of PAML [46] and compared using likelihood ratio tests. Two tests of positive selection were performed. In the first test the neutral model (M1) is compared with the selection model, in which a class of sites is permitted to exhibit d_N/d_S (ω) > 1 (M2). In the second test, a beta distribution of site classes in which the most rapidly evolving is fixed to ω =1 (M8a) is compared to a similar model in which the most rapidly evolving site class is permitted to exhibit $\omega > 1$ (M8) [61]. Multiple initial values of ω were used to ensure convergence on the likelihood optima. For the second test, critical values of the test statistic are determined from Wong *et al* [62]. Lineage-specific selection patterns of d_N/d_S were determined by implementing branch-specific models [63].

Determination of Duplication Rate. A total of 34 secreted proteases were identified in *D. arizonae* female reproductive tracts. Using BLASTNhomology and maximum-likelihood phylogenetic reconstruction implemented in PAUP* we determined these 34

proteins correspond to 37 orthologs in the genome of *D. mojavensis*, and 23 orthologs in the genome of *D. virilis* (http://rana.lbl.gov/drosophila/). Assuming no gene conversion or gene loss, the total copy number of these genes was 23 at the divergence of the *D. mojavensis* and *D. virilis* lineages. Duplication rate can therefore be estimated by the following exponential growth equation:

$C_M = C_A 2^{rt}$

Where C_M is copy number of *D. mojavensis* (37), C_A is the ancestral copy number (23), t is the divergence time between *D. mojavensis* and *D. virilis* (t=23 MYA [41]), and r is the estimated rate of duplication per gene per million years.

RT-PCR. *Drosophila arizonae* RNA was extracted from 20 whole males, 70 reproductively mature females from population bottles lacking their lower reproductive tracts, and 70 lower reproductive tracts preserved in TRIZOL® (Invitrogen) according to manufacturer instructions. Purified RNA was treated with DNAseI (Gibco), and reverse transcribed with the iScript cDNA synthesis kit (Bio-Rad). Resultant cDNA was diluted to 10 ng/µL, and used as template for standard PCR using universal primers, with *D. arizonae* genomic DNA as a positive control. Primer sequences are as follows:
Dmoj\GLEANR_8528-F 5'-AAGAAGCGCACCAAGCACTTCATC-3',
Dmoj\GLEANR_8528-TCTGTTGTCGATACCCTTGGGCTT-3',
protease gene family 1 -F1 5'-ATGTGGAATCTAAGCCCAGCCAA-3',

protease gene family 1 -F2 5'-RTAGATGGCAGTTGCTYCTYGTG-3',

protease gene family 1 -R1 5'-GATGYGATACCAATCACRGTGCT-3',

protease gene family 1 -R2 5'-ACGATRCCAATCACRGTGCYAGA-3',

protease gene family 2 -F1 5'-CTCAAACCGCARTAGYTRTCCT-3',

protease gene family 2 -F2 CTTCAAGCCGCMGTWGCTGTCCT-3',

protease gene family 2 -R1 5'-CACCRCTGTGYTYCCTRATCCATTC-3',

protease gene family 2 -R2 5'-CACCGCWGTGCTCYYTGATCCATT-3',

protease gene family 3 -F1 5'-TGAAACCGATCCCAGACTTATAGC-3',

protease gene family 3 -F2 5'-ATGAAACCGATCCCGAGTTGATAG-3',

protease gene family 3 -R1 5'-ATCAGCCATGCTCAATTCTTGTCG-3',

protease gene family 3 -R2 5'-ATCAGCCCAGCTTAATTCTAGTCG-3'.

Structural Modeling. 3D structure was predicted by SWISS-MODEL [64], and visualized by Deep View. Selected sites were determined from Bayes Emperical Bayes calculation [65] implemented under M8 in PAML [46].

Acknolwedgements. The authors would like to acknowledge Luciano Matzkin and James Pennington for helpful discussion, and Jeff Good, Matt Dean, Gabriela Wlasiuk, and four anonymous reviewers for generous comments on this manuscript. This research was funded by the University of Arizona, and the NSF-IGERT program in Evolutionary, Functional and Computational Genomics at the University of Arizona. E.S.K. was supported by an NSF-IGERT fellowship in Evolutionary, Functional and Computational

Genomics at the University of Arizona. W.J.S. was supported by NIH grant HD42563.

REFERENCES

1. Swanson WJ, Vacquier VD (2002) The rapid evolution of reproductive proteins. Nat Rev Genet 3: 137–44.
2. Panhuis TM, Clark NL, Swanson WJ (2006) Rapid evolution of reproductive proteins in abalone and *Drosophila*. Philos Trans R Soc Lond B Biol Sci 361: 261–8.
3. Clark NL, Aagaard JE ,Swanson WJ (2006) Evolution of reproductive proteins from animals and plants. Reproduction 131: 11–22.
4. Lee YH, Ota T, Vacquier VD (1995) Positive selection is a general phenomenon in the evolution of abalone sperm lysin. Mol Biol Evol 12: 231–238.
5. Yang Z, Swanson WJ, Vacquier VD (2000) Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. Mol Biol Evol 17: 1446–55.
6. Galindo BE, Moy GW, Swanson WJ,Vacquier VD (2002) Full-length sequence of VERL, the egg vitelline envelope receptor for abalone sperm lysin. Gene 288: 111–117.
7. Galindo BE, Vacquier VD, Swanson WJ (2003) Positive selection in the egg receptor for abalone sperm lysin. Proc Natl Acad Sci U S A. 100: 4639–4643.
8. Roberston SA (2007) Seminal fluid signaling in the female reproductive tract: lessons from rodents and pigs. J Anim Sci 85: E36-44.
9. Wolfner MF (2002) The gifts that keep on giving: physiological functions and evolutionary dynamics of male seminal proteins in *Drosophila.* Heredity 88: 85–93.
10. Kubli E (2003) Sex-peptides: seminal peptides of the *Drosophila* male. Cell Mol Life Sci 60: 1689–1704.
11. Chapman T, Davies SJ (2004) Functions and analysis of the seminal fluid proteins of male *Drosophila melanogaster* fruit flies. Peptides 25: 1477–1490.
12. Markow TA (1996) Evolution of *Drosophila* Mating Systems. Evol Biol 29: 73–106.
13. Monsma SA, Harada HA, Wolfner MF (1990) Synthesis of two *Drosophila* male accessory gland proteins and their fate after transfer to the female during mating. Dev Biol 142: 465–475.
14. Park M, Wolfner MF (1995) Male and female cooperate in the prohormone-like processing of a *Drosophila melanogaster* seminal fluid protein. Dev Biol 171: 694–702.
15. Peng J, Chen S, Busser S, Liu H, Honegger T, Kubli E (2005) Gradual release of sperm bound sex-peptide controls female postmating behavior in Drosophila. Curr Biol 15: 207–213.
16. Bertram MJ, Neubaum DM, Wolfner MF (1996) Localization of the *Drosophila* male accessory gland protein Acp36DE in the mated female suggests a role in sperm storage. Insect Biochem Mol Biol 26: 971–980.
17. Heifetz Y, Lung O, Frongillo EA Jr, Wolfner MF (2000) The *Drosophila* seminal fluid protein Acp26Aa stimulates release of oocytes by the ovary. Curr Biol 10: 99–102.
18. Ravi Ram K, Ji S, Wolfner MF (2005) Fates and targets of male accessory gland

proteins in mated female *Drosophila melanogaster*. Insect Biochem Mol Biol 35: 1059–1071.
19. Civetta A, 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.
20. Begun DJ, Lindfors HA (2005) Rapid evolution of genomic Acp complement in the *melanogaster* subgroup of *Drosophila*. Mol Biol Evol 22**:** 2010–2021.
21. Mueller JL, Ravi Ram K, McGraw LA, Bloch Qazi MC, Siggia ED, Clark AG, Aquadro CF, Wolfner MF (2005) Cross-species comparison of *Drosophila* male accessory gland protein genes. Genetics 171: 131–143.
22. Begun DJ, Lindfors HA, Thompson ME, Holloway AK (2006) Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. Genetics 172: 1675–1681.
23. Begun DJ, Whitley P, Todd BL, Waldrip-Dail HM, Clark AG (2000) Molecular population genetics of male accessory gland proteins in *Drosophila*. Genetics 156: 1879–1888.
24. Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF (2001) Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proc Nat Acad Sci U S A 13: 7375–7379.
25. Kern AD, Jones CD, Begun DJ (2004) Molecular population genetics of male accessory gland proteins in the *Drosophila simulans* complex. Genetics 167: 725–725.
26. Schully SD, Hellberg ME (2006) Positive Selection on Nucleotide Substitutions and Indels in Accessory Gland Proteins of the *Drosophila pseudoobscura* Subgroup. J Mol Evol 62: 793–802.
27. Wagstaff BJ, Begun DJ (2005) Molecular population genetics of accessory gland protein genes and testis-expressed genes in *Drosophila mojavensis* and *D. arizonae.* Genetics 171: 1083–101.
28. Swanson WJ, Wong A, Wolfner MF, Aquadro CF (2004) Evolutionary expressed sequence tag analysis of *Drosophila* female reproductive tracts identifies genes subjected to positive selection. Genetics 168: 1457–1465.
29. Panhuis T, Swanson WJ (2006) Molecular evolution and population genetics of candidate female reproductive genes in *Drosophila*. Genetics 173: 2039–2047.
30. Jagadeeshan S, Singh RS (2007) Rapid evolution of outer egg membrane proteins in the *Drosophila melanogaster* subgroup: a case of ecologically driven evolution of female reproductive traits. *Mol Biol Evol* 24: 929–938.
31. Markow TA (2002) Female remating, operational sex ratio, and the arena of sexual selection in *Drosophila*. *Evolution* 56: 1725–1734.
32. Dorus S, Evans PD, Wyckoff GJ, Choi SS, Lahn BT (2004) Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. Nat Genet 36: 1326–1329.
33. Markow TA, Ankney P (1988) Insemination reaction in *Drosophila*: A copulatory plug in species showing male contribution to offspring. Evolution 42**:** 1097–1100.

34. Markow TA, Gallagher PD, Krebs RA (1990) Ejaculate-derived nutritional contribution and female reproductive success in *Drosophila mojavensis* (Patterson and Crow. Func Ecol 4: 67–73.
35. Patterson JT (1946) A new type of isolating mechanism in *Drosophila*. Proc Nat Acad Sci U S A 32: 202–208.
36. Matzkin LM (2004) Population genetics and geographic variation of alcohol dehydrogenase (Adh) paralogs and glucose-6-phosphate dehydrogenase (G6pd) in *Drosophila mojavensis*. Mol Biol Evol 21: 276–285.
37. Pitnick S, Miller GT, Schneider K, Markow TA (2003) Ejaculate-female coevolution in *Drosophila mojavensis.* Proc Nat Acad Sci U S A 270: 507–1512.
38. Knowles LL, Markow TA (2001) Sexually antagonistic coevolution of a postmating prezygotic reproductive character in desert *Drosophila*. Proc Nat Acad Sci U S A 98: 8692–8696.
39. Knowles LL, Hernandez BB, Markow TA (2005) Exploring the consequences of postmating-prezygotic interactions between the sexes. Proc Biol Sci 271 Suppl 5: S357–S359.
40. Knowles LL, Hernandez BB, Markow TA (2005) Non-antagonistic interactions between the sexes revealed by the ecological consequences of reproductive traits. J Evol Biol 18:156–161.
41. Powell JR (1997) Progress and Prospects in Evolutionary Biology: The *Drosophila* Model. New York: Oxford University Press. 576 p.
42. Gu Z, Cavalcanti A, Chen FC, Bouman P, Li WH (2002) Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. Mol Biol Evol 19: 256–62.
43. Neurath H (1984) Evolution of proteolytic enzymes. *Science* 224: 350–357.
44. Ross J, Jiang H, Kanost MR, Wang Y (2003). Serine proteases and their homologs in the *Drosophila melanogaster* genome: an initial analysis of sequence conservation and phylogenetic relationships. Gene 304:117–31.
45. Stocker W, Zwilling R (1995) Astacin. Methods Enzymol 248: 305–25.
46. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555–556.
47. Sprang SR, Fletterick RJ, Graf L, Rutter WJ, Craik CS (1988) Studies of specificity and catalysis in trypsin by structural analysis of site-directed mutants. Crit Rev Biotechnol 8: 225–36.
48. Markow TA, Coppola A, Watts TD (2001) How *Drosophila* males make eggs: it is elemental. Proc Biol Sci 268: 1527–1532.
49. Rice WR (1996) Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. Nature 381: 232–4.
50. Gavrilets S (2000) Rapid evolution of reproductive barriers driven by sexual conflict. Nature 403: 886–889.
51. Hayashi TI, Vose M, Gavrilets S (2007) Genetic differentiation by sexual conflict. Evolution. 61: 516–29.
52. Hey J, Kliman RM (1993) Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* species complex. Mol Biol Evol 10: 804–822.

53. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8: 186–194.
54. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215 :403–410.
55. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0 J Mol Biol 340: 783–795.
56. Sonnhammer EL, von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 6: 175–182.
57. Eddy, S. R. (1998) Profile hidden Markov models. Bioinformatics 14:755–763.
58. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.
59. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research 22: 4673-4680.
60. Sawyer SA (1989) Statistical tests for detecting gene conversion. Mol Biol Evol 6: 526–538.
61. Swanson WJ, Nielsen R, Yang Q (2003) Pervasive adaptive evolution in mammalian fertilization proteins. Mol Biol Evol 20: 18–20.
62. Wong WS, Yang Z, Goldman N, 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.
63. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15: 568–573.
64. Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res 31: 3381–3385.
65. Yang Z, Wong WS, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol 22: 1107–1118.

TABLES

D. mojavensis CDS	SS	TM	d_N	d_S	d_N/d_S	*D. melanogaster* CDS	*D. melanogaster* function	conserved domain
GLEANR_9982	S	0	0.03	0.02	1.83	CG7443-PA	unknown function	no
anon-EST:Kelleher15	S	1	0.03	0.02	1.38	no hits	NA	no
GLEANR_5396	S	1	0.02	0.02	1.30	Gp150-PD	cell adhesion/signal transduction	LRR_1
GLEANR_4627	Q	10	0.03	0.03	1.28	CG30344-PA	unknown function	MFS_1
anon-EST:Kelleher5	S	0	0.05	0.04	1.20	CG10472	proteolysis	trypsin
GLEANR_17128	S	2	0.06	0.06	1.08	no hits	NA	no
GLEANR_5358	Q	1	0.01	0.01	1.04	CG30415-PA	unknown function	no
GLEANR_1967	S	1	0.05	0.05	0.96	CG7778-PA	unknown function	no
GLEANR_896	S	0	0.11	0.12	0.89	CG31954-PA	proteolysis	trypsin
GLEANR_17617	S	4	0.02	0.03	0.82	CG4729-PA	metabolism	Acyltransferase
GLEANR_3367	A	0	0.02	0.03	0.82	lectin-46Cb-PA	sugar binding	Lectin_C
GLEANR_5037	A	4	0.01	0.02	0.80	CG15098-PA	unknown function	no
GLEANR_9029	A	2	0.01	0.01	0.80	Nep2-PA	proteolysis	peptidase_M13_N, Peptidase_M13
GLEANR_13559	Q	12	0.02	0.03	0.79	CG10960-PB	metabolism/ transport	Sugar_tr, MFS_1
GLEANR_2703	S	0	0.09	0.12	0.77	CG15254-PA	proteolysis	astacin
GLEANR_9617	S	0	0.07	0.09	0.74	CG3739-PA	proteolysis	peptidase_S28
GLEANR_12094	S	0	0.01	0.02	0.73	CG6409-PA	GPI anchor biosynthesis	no
GLEANR_10002	S	0	0.04	0.05	0.70	CG9418-PA	DNA binding	no
GLEANR_1209	A	1	0.04	0.06	0.69	no hits	NA	no
GLEANR_6649	S	0	0.01	0.02	0.67	l(1)G0193-PB	unknown function	no
GLEANR_7394	S	0	0.02	0.03	0.66	CG17271-PA	Ca2+ Binding	no
GLEANR_10	S	0	0.03	0.04	0.65	CG5630-PA	no known	no

683							funciton	
GLEANR_60							unknown	
54	Q	2	0.01	0.02	0.62	CG4627-PA	function	no
GLEANR_27							metabolism/tran	
75	Q	10	0.03	0.05	0.59	CG4726-PA	sport	MFS_1
anon-								
EST:Kelleher								
9	S	0	0.04	0.06	0.59	CG10472	proteolysis	trypsin
GLEANR_12								
45	Q	1	0.03	0.05	0.57	no hits	NA	Lamp
GLEANR_16								
396	S	0	0.05	0.09	0.55	no hits	NA	no
GLEANR_28						CG14536-	unknown	
81	Q	1	0.03	0.05	0.54	PB	function	ubiquitin
GLEANR_82								
58	S	0	0.03	0.06	0.53	CG3734-PA	proteolysis	peptidase_S28
GLEANR_70							lipid	
51	S	0	0.05	0.09	0.52	CG6283-PA	metabolism	Lipase
GLEANR_36						CG18067-	unknown	
13	S	0	0.02	0.04	0.51	PA	function	no

Table 1. Candidate Female Reproductive Proteins. SS: S=secreted, A=anchor, Q=quiescent as predicted by SignalP 3.0 [55], TM: number of identified transmembrane domains [56] d_N: estimated non-synonymous substitutions per non-synonymous site d_S: estimated synonymous substitutions per synonymous site, d_N/d_S: estimated ratio non-synonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site, calculated in PAML [46], *D. melanogaster* CG: best tblastx hit in the *D. melanogaster* genome, *D. melanogaster* function: from FlyBase annotations, Conserved domain: Pfam conserved domain predicted from hmmpfam [57].

gene family name	enzyme class	sequences analyzed	LRT(M1 vsM2)	ω	p	LRT(M 8 avsM8)	ω	p
protease gene family 1	elastase/ chymotrypsin	*mojavensis*(5), *arizonae*(6)	**103.96*** **	5.39	0.08	**91.6`****	4.63	0.09
protease gene family 2	trypsin	*mojavensis*(5), *arizonae*(4), *virilis*(1), *grimshawi*(1)	**12.41*****	2.86	0.04	**11.52****	2.00	0.11
protease gene family 3	astacin	*mojavensis*(4), *arizonae*(2), *virilis*(1), *grimshawi*(1)	0.00	1.00	0.11	**15.39****	11.96	0.01
Dmoj\ GLEANR_ 12324\12325	serine carboxypeptidase	*mojavensis*(2), *arizonae*(2), *virilis*(1)	0.00	1.00	0.05	0.00	1.00	0.00
Dmoj\ GLEANR _8258\8259	serine-type peptidase/lipase	*mojavensis*(2), *arizonae*(2), *virilis*(1)	0.00	1.00	0.05	0.01	1.00	0.02

Table 2. PAML analysis of positive selection in duplicated proteases. Gene families are identified by their assigned name. Enzyme class is determined from hmmpfam [57] and SWISS-MODEL [64]. Species analyzed are indicated, followed by number of paralogs per species in parentheses *D. mojavensis*, *D. virilis* and *D. grimshawi* sequences were obtained from their published genomes (http://rana.lbl.gov/drosophila), while all *D. arizonae* sequences were found in the library. LRT denotes the value of the likelihood ratio test between the two models, followed by an indication of the statistical significance of the test. ω corresponds to the estimated highest estimated dN/dS of all site classes, and p corresponds to the proportion of sites in this class.

FIGURES

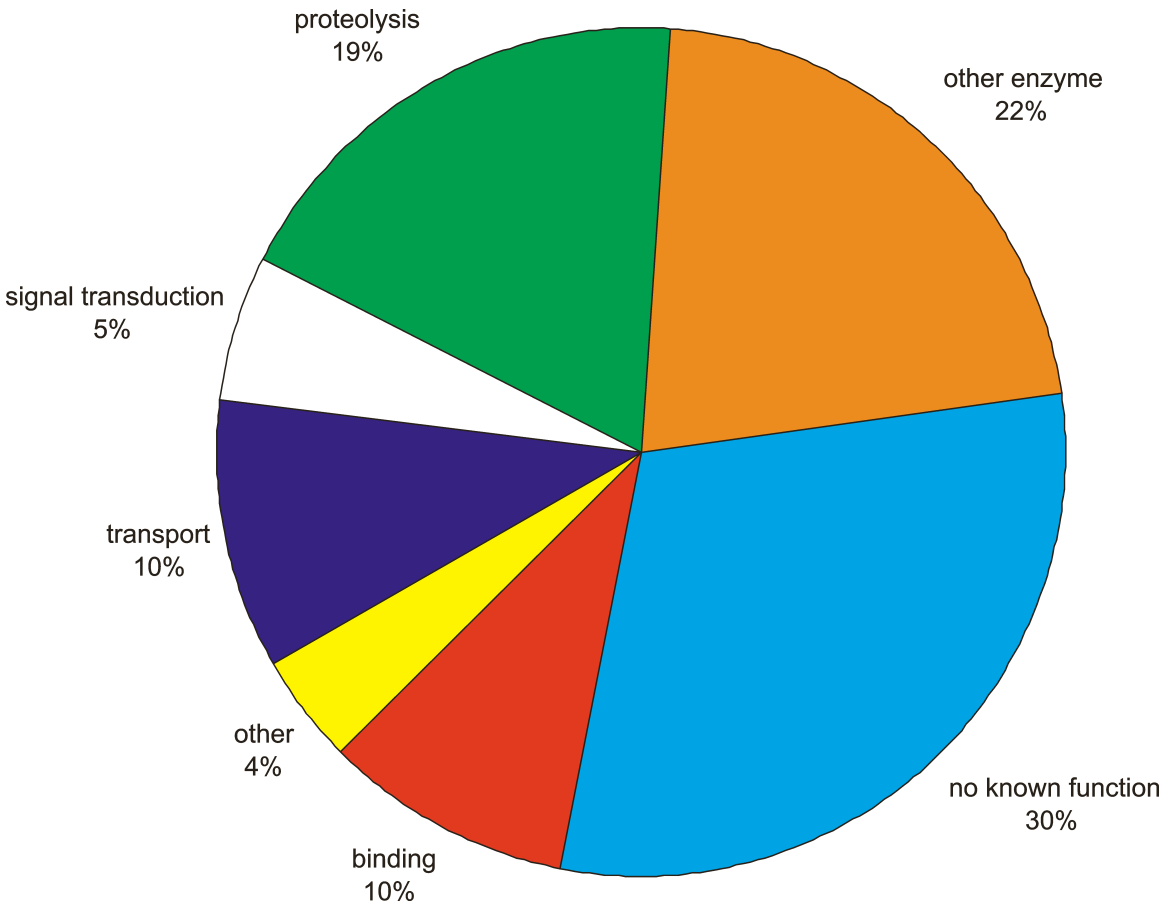


Figure 1. Functional Composition of Candidate Female Reproductive Proteins. Functional composition of 241 secreted and transmembrane proteins in *D. arizonae* female reproductive tracts based on Gene Ontology (GO) terms [58].

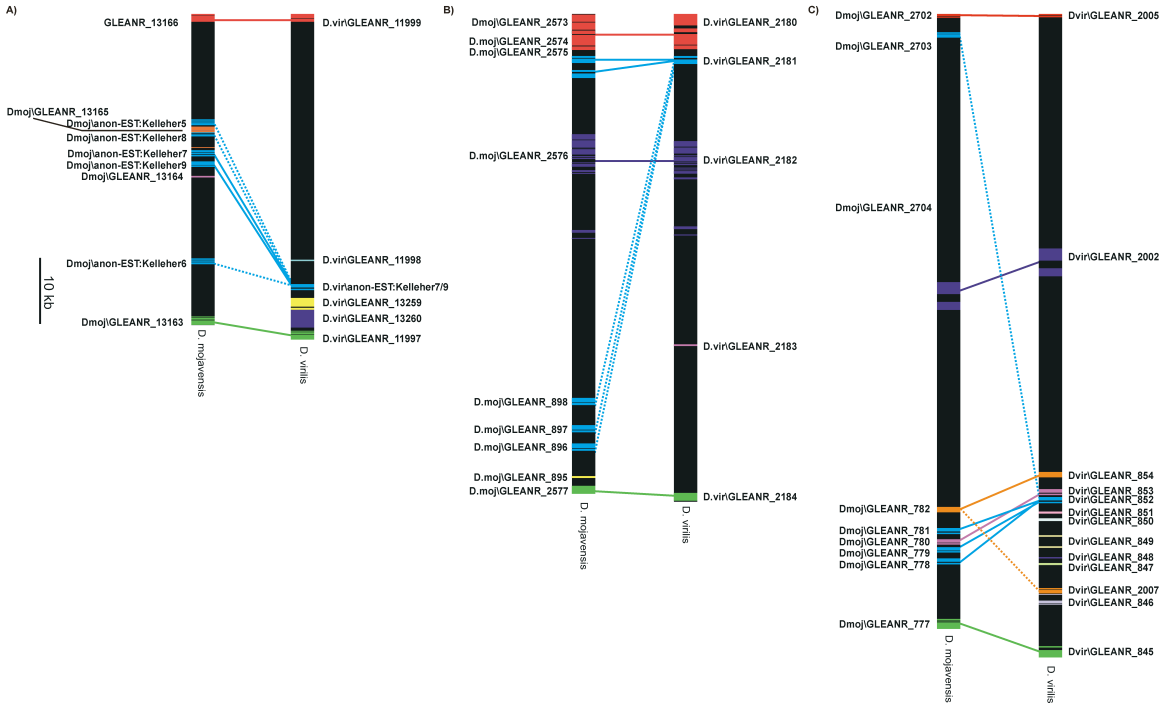


Figure 2. Distribution of Three Protease Gene Families in *D. mojavensis* and *D. virilis* Genomes. Syntenic regions of A) protease gene family 1 - *D. mojavensis* chromosome 4 (scaffold_6680 bp 10216565-10169309) and *D. virilis* chromosome 3 (scaffold_13049 bp 10558802-10608251) B) protease gene family 2 - *D. mojavensis* chromosome 3 (scaffold_6500 bp 18241557- 18296199) and *D. virilis* chromosome 4 (scaffold_12963 bp 15263878-15319561) C) protease gene family 3 - *D. mojavensis* chromosome 3 (scaffold_6500 bp 20970182- 21063420) and *D. virilis* chromosome 4 (scaffold_12963 bp 12250368-12347919). Colored blocks indicate individual exons, where each gene is indicated by a different color. Orthologous genes are the same color in both species, and connected by colored lines. Solid lines indicate orthologs with the same orientation, while dotted lines indicate inverted orthologs. Multiple, tandemly duplicated copies in the genome of *D. mojavensis* correspond to a single gene in the genome of *D. virilis.* Annotation and assembly obtained from unpublished *Drosophila* genomes (http://rana.lbl.gov/drosophila/).

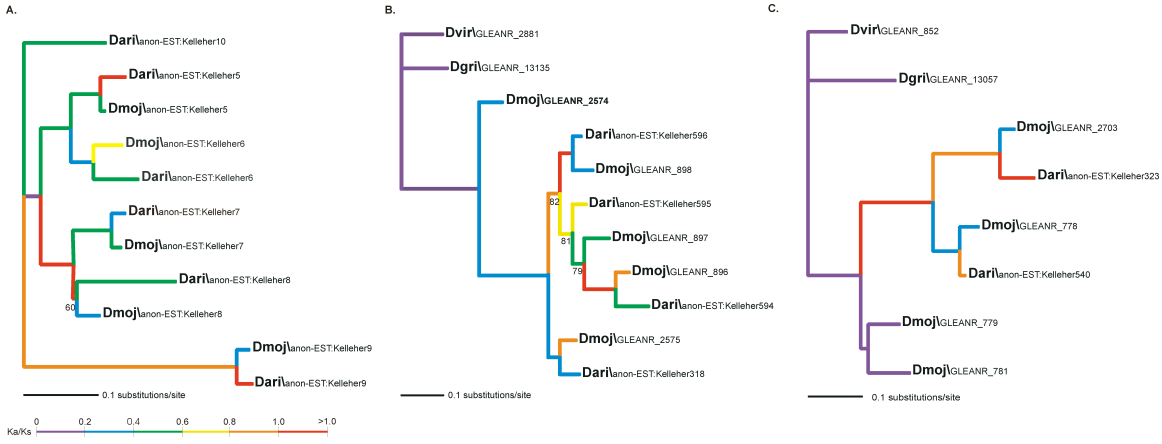


Figure 3. Bayesian phylogenies of A. protease gene family 1 B. protease gene family 2 C. protease gene family 3. A is midpoint rooted, as *D. virilis* sequence was too divergent to make an appropriate outgroup. Grey taxon name denotes a pseudogene. Branch colors indicate Ka/Ks values calculated in the codeml package of PAML [46].

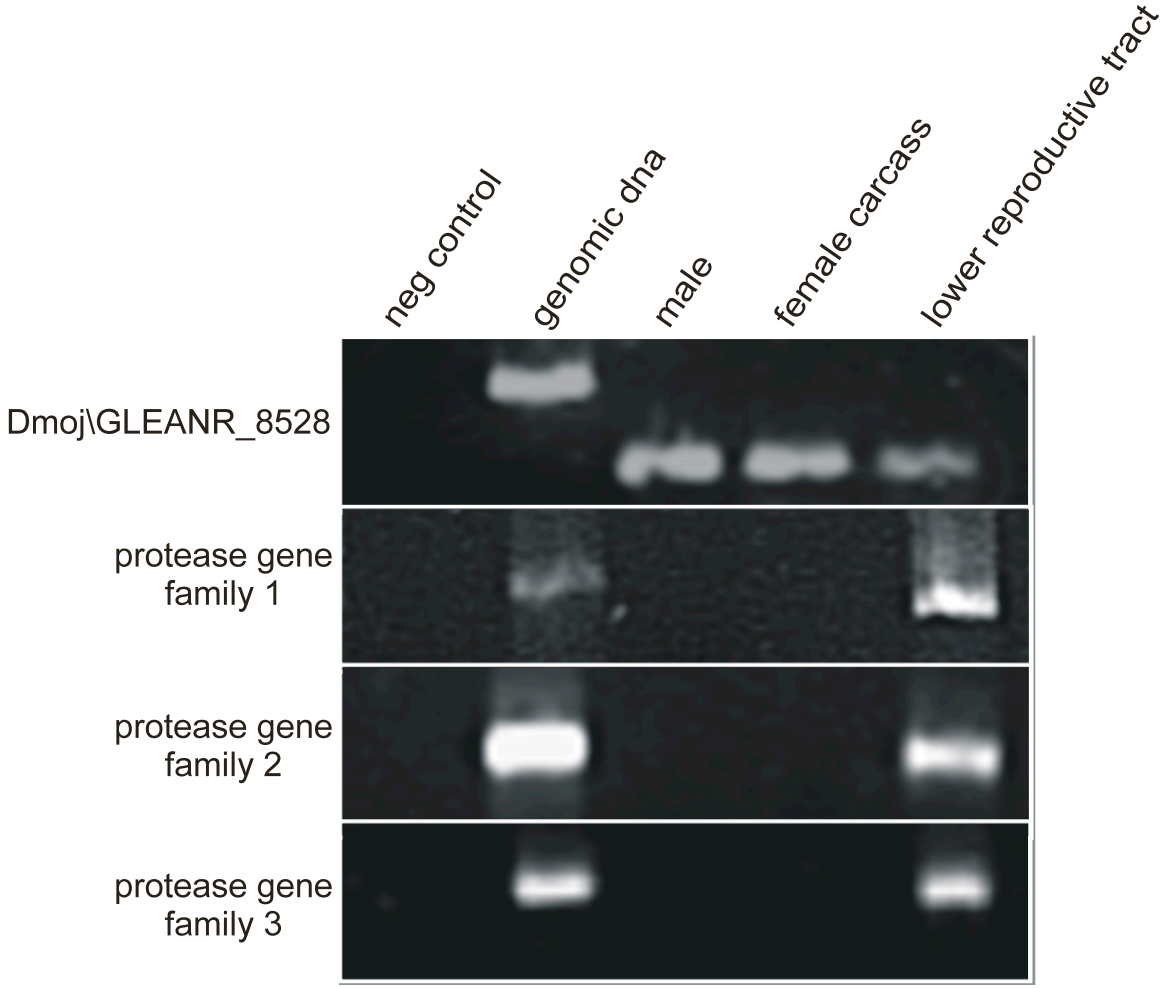


Figure 4. RT-PCR of Three Gene Families.
Universal primers for each gene family were used to amplify genomic DNA, and cDNA from males, female carcasses (no lower reproductive tract), and lower reproductive tracts (for complete gels see Supplementary Materials).

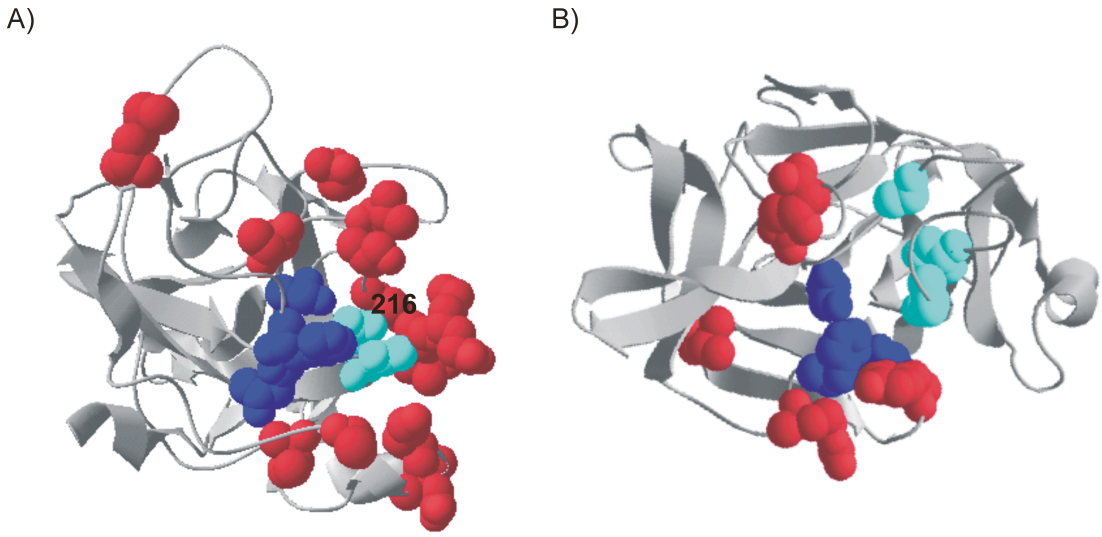


Figure 5. Structural models generated in SWISS-MODEL. A) protease gene family 1 B) protease gene family 2. The blue amino acids comprise the catalytic triad of the active site. The aquamarine amino acids are determinants of substrate specificity [47]. The red amino acids indicate positively selected sites. The labeled amino acid in panel A, 216, is a positively selected amino acid that is also a determinant of substrate specificity.

SUPPLEMENTARY DATA

D. mojavensis CDS	*D. arizonae* EST	*D. melanogaster* homolog	Signal P	TMHMM	ka	ks	ka/ks	prot % ID	CDS % ID	conserved domain	
scaffold_6680.1507	Kelleher1	no hits	Q	0	0.00	0.04	0.00	100.00	98.92	SAM_1	SAM_2
scaffold_6500.1871	Kelleher2	CG31680-PA	Q	0	0.03	0.05	0.56	93.75	96.76	peptidase_A17	
scaffold_6498.43	Kelleher3	bt	Q	0	0.00	0.05	0.00	100.00	98.55	I-set	ig
NM_176740	Kelleher4	Ranbp16-PB	Q	1	0.00	0.08	0.00	97.06	99	IBN_N	
Elastase-1	Kelleher5	CG10472	S	0	0.05	0.04	1.20	89.67	95.2	trypsin	
Elastase-2	Kelleher6	CG10472	S	0	0.08	0.17	0.44	86.52	91.26	trypsin	
Elastase-3	Kelleher7	CG10472	S	0	0.03	0.07	0.36	94.46	96.68	trypsin	
Elastase-4	Kelleher8	CG10472	S	0	0.14	0.31	0.47	77.74	84.78	trypsin	
Elastase-5	Kelleher9	CG10472	S	0	0.04	0.06	0.59	92.80	96.13	trypsin	
Elastase-6	Kelleher10	CG-10472	S	0	no clear ortholog					trypsin	
ESK-CONTIG103-MOJ	Kelleher11	mt:ND3-PA									
ESK-CONTIG221-MOJ	Kelleher12	mt:ND5-PA									
ESK-CONTIG256-MOJ	Kelleher13	mt:ND6-PA									
ESK-CONTIG256-MOJ	Kelleher14	mt:ND6-PA									
ESK-CONTIG2cg09-MOJ	Kelleher15	no hits	S	1	0.03	0.02	1.38	92.54	97.01	no	
ESK-CONTIG306-MOJ	Kelleher16	Argk-PC	Q	0	0.00	0.01	0.00	100.00	99.7	ATP-gua_PtransN	ATP-gua_Ptrans
ESK-CONTIG355-MOJ	Kelleher17	mt:ND4-PA									
ESK-CONTIG359-MOJ	Kelleher18	mt:ND1-PA									
ESK-CONTIG369-MOJ	Kelleher19	mt:Cyt-b-PA									
ESK-CONTIG376-MOJ	Kelleher20	mt:CoII-PA									
ESK-CONTIG381-MOJ	Kelleher21	mt:CoI-PA									
GLEANR_10002	Kelleher22	CG9418-PA	S	0	0.04	0.05	0.70	93.41	95.97	no	
GLEANR_10052	Kelleher23	CG5903-PA	S	0	0.00	0.02	0.00	100.00	99.48	no	
GLEANR_10094	Kelleher24	SNF4Agamma-PF	Q	0	0.00	0.10	0.02	99.61	97.29	no	
GLEANR_10104	Kelleher25	CG18519-PB	Q	0	0.01	0.03	0.37	97.67	98.45	Ald_Xan_dh_C2	Fer2_2
GLEANR_10183	Kelleher26	RpL13A-PB	Q	0	0.00	0.01	0.00	100.00	99.69	Ribosomal_L13	
GLEANR_10263	Kelleher27	Surf6-PA	Q	0	0.02	0.06	0.34	96.55	96.93	SURF6	
GLEANR_10285	Kelleher28	RpL27-PA	Q	0	0.00	0.01	0.00	100.00	99.75	Ribosomal_L	KOW

										27e	
GLEANR_10316	Kelleher29	CG10340-PA	Q	0	0.01	0.08	0.14	97.63	97.63	ATP11	
GLEANR_10340	Kelleher30	RpS8-PC	Q	0	0.00	0.04	0.00	100.00	99.04	Ribosomal_S 8e	
GLEANR_10359	Kelleher31	CG1746-PC	Q	2	0.00	0.01	0.00	100.00	99.52	ATP-synt_C	
GLEANR_10408	Kelleher32	desat1-PC	Q	4	0.00	0.03	0.00	100.00	99	FA_desaturas e	
GLEANR_1051	Kelleher33	CG31715-PA	Q	0	0.00	0.04	0.00	100.00	98.89	Ank	
GLEANR_10514	Kelleher34	RpS30-PB	Q	0	0.00	0.03	0.00	100.00	99.24	Ribosomal_S 30	
GLEANR_1053	Kelleher35	RpS27A-PA	Q	0	0.00	0.01	0.00	100.00	99.79	ubiquitin	Ribosomal_S 27
GLEANR_10535	Kelleher36	Sap-r-PB	S	1	0.02	0.11	0.13	96.67	96.3	SapB	SapB_2
GLEANR_1054	Kelleher37	Klp31E-PA	Q	0	0.01	0.07	0.20	96.99	97.24	Kinesin	
GLEANR_10568	Kelleher38	no hits	S	0	0.02	0.06	0.38	95.45	96.97	Cystatin	
GLEANR_10598	Kelleher39	Hrb87F-PB	Q	0	0.01	0.05	0.20	98.20	97.9	RRM_1	
GLEANR_10600	Kelleher40	sqd-PA	Q	0	0.00	0.05	0.00	100.00	99.03	RRM_1	
GLEANR_10617	Kelleher41	CG31357-PA	Q	0	0.02	0.04	0.48	97.06	97.79	no	
GLEANR_10632	Kelleher42	Surf4-PB	Q	6	0.00	0.04	0.00	100.00	98.69	SURF4	
GLEANR_10646	Kelleher43	Mlc1-PB	Q	0	0.00	0.01	0.00	100.00	99.78	no	
GLEANR_10655	Kelleher44	Atpalpha-PA	Q	8	0.00	0.04	0.00	100.00	98.99	E1-E2_ATPase	Cation_ATPa se_C
GLEANR_10658	Kelleher45	ATPsyn-d-PA	Q	0	0.00	0.03	0.00	100.00	99.23	Mt_ATP-synt_D	
GLEANR_10660	Kelleher46	CG17119-PB	S	7	0.02	0.05	0.30	96.65	97.45	PQ-loop	
GLEANR_10683	Kelleher47	CG5630-PA	S	0	0.03	0.04	0.65	93.42	96.93	no	
GLEANR_10717	Kelleher48	mod(mdg4)-PP	Q	0	0.00	0.05	0.04	99.58	98.45	BTB	FLYWCH
GLEANR_10729	Kelleher49	CG3308-PA	Q	0	0.00	0.03	0.00	100.00	98.92	TatD_DNase	
GLEANR_10731	Kelleher50	Rab1-PB	Q	0	0.00	0.03	0.00	100.00	98.97	Ras	Miro
GLEANR_10845	Kelleher51	RpL30-PA	Q	0	0.00	0.03	0.00	100.00	99.03	Ribosomal_L 7Ae	
GLEANR_10867	Kelleher52	no hits	Q	0						Septin	
GLEANR_10946	Kelleher53	RpS19a-PC	Q	0	0.00	0.01	0.30	99.36	99.57	Ribosomal_S 19e	
GLEANR_10949	Kelleher54	no hits	Q	0	0.00	0.00	0.21	100.00	100	no	
GLEANR_10995	Kelleher55	RpL36-PC	Q	0	0.00	0.02	0.00	100.00	99.71	Ribosomal_L 36e	
GLEANR_10997	Kelleher56	MED18-PA	Q	0	0.00	0.07	0.00	100.00	98.48	no	
GLEANR_11012	Kelleher57	sesB-PA	Q	3	0.00	0.00	0.19	100.00	100	Mito_carr	
GLEANR_11026	Kelleher58	RpS10b-PC	Q	0	0.00	0.04	0.00	100.00	98.96	S10_plectin	
GLEANR_11027	Kelleher59	CG14207-PA	Q	0	0.00	0.02	0.00	100.00	99.35	HSP20	
GLEANR_11031	Kelleher60	RpS6-PB	Q	0	0.00	0.04	0.00	100.00	99.19	Ribosomal_S 6e	
GLEANR_11114	Kelleher61	CG3415-PA	Q	0	0.01	0.04	0.25	98.03	98.25	no	
GLEANR_11115	Kelleher62	Pros28.1-PA	Q	0	0.00	0.05	0.00	100.00	98.76	MaoC_dehyd ratas	adh_short
GLEANR_11116	Kelleher63	eas-PE	Q	0	0.02	0.03	0.57	96.58	98.01	Choline_kina se	APH
GLEANR_11116	Kelleher64	eas-PE	Q	0	0.02	0.03	0.57	96.58	98.01	Proteasome	
GLEANR_11120	Kelleher65	CG18624-PB	S	1	0.00	0.00	0.02	100.00	100	no	
GLEANR_11316	Kelleher66	no hits	Q	0	0.01	0.01	0.40	98.92	99.28	rve	
GLEANR_11376	Kelleher67	Pka-C3-PB	Q	0	0.01	0.05	0.16	98.12	98.33	Pkinase	Pkinase_Tyr
GLEANR_11397	Kelleher68	CG18081-PA	Q	0	0.00	0.01	0.00	100.00	99.57	no	
GLEANR_11412	Kelleher69	Cat-PA	Q	0	0.00	0.08	0.00	100.00	98.31	Catalase	
GLEANR_11421	Kelleher70	CG14184-PA	Q	0	0.02	0.03	0.61	96.15	98.29	no	
GLEANR_11519	Kelleher71	CG11593-PA	Q	0	0.00	0.08	0.00	100.00	98.3	no	
GLEANR_11538	Kelleher72	CG7888-PC	Q	1 0	0.00	0.01	0.25	99.34	99.34	Aa_trans	
GLEANR_11541	Kelleher73	SuUR-PA	Q	0	0.02	0.06	0.40	94.81	97.01	no	

GLEANR_11581	Kelleher74	Mpcp-PA	Q	0	0.00	0.03	0.08	99.47	98.94	Mito_carr	
GLEANR_11639	Kelleher75	CG32444-PA	Q	0	0.00	0.11	0.00	100.00	98.53	Aldose_epim	
GLEANR_11758	Kelleher76	RpS12-PF	Q	0	0.00	0.02	0.00	100.00	99.76	Ribosomal_L7Ae	
GLEANR_11784	Kelleher77	no hits	Q	0	0.02	0.03	0.92	94.74	97.81	no	
GLEANR_11823	Kelleher78	RpS4-PB	Q	0	0.00	0.02	0.00	100.00	99.35	Ribosomal_S4e	RS4NT
GLEANR_11906	Kelleher79	VhaM9.7-2-PA	A	2	0.00	0.08	0.06	98.88	98.13	ATP_synt_H	
GLEANR_11970	Kelleher80	mRpL15-PA	Q	0	0.01	0.08	0.07	98.86	97.35	no	
GLEANR_11986	Kelleher81	CG9674-PD	Q	0	0.00	0.07	0.00	100.00	98.47	Glu_synthase	GATase_2
GLEANR_12010	Kelleher82	CG32277-PA	S	0	0.00	0.02	0.13	99.55	99.26	trypsin	
GLEANR_12014	Kelleher83	CG32276-PB	A	1	0.00	0.08	0.00	100.00	97.92	RAMP4	
GLEANR_12020	Kelleher84	Adk1-PA	Q	0	0.00	0.02	0.00	100.00	99.61	ADK	
GLEANR_12031	Kelleher85	CG32473-PC	A	11	0.01	0.05	0.15	98.41	98.06	Laminin_A	
GLEANR_12051	Kelleher86	CG5687-PA	S	3	0.00	0.09	0.00	100.00	98.14	SSF	
GLEANR_1208	Kelleher87	CG31886-PA	A	1	0.01	0.00	NA	97.44	99.15	no	
GLEANR_1209	Kelleher88	no hits	A	1	0.04	0.06	0.69	93.10	95.4	no	
GLEANR_12094	Kelleher89	CG6409-PA	S	0	0.01	0.02	0.73	97.55	98.77	no	
GLEANR_12118	Kelleher90	CG6767-PA	Q	0	0.00	0.03	0.00	100.00	99.13	Pribosyltran	
GLEANR_12123	Kelleher91	RpS9-PA	Q	0	0.00	0.04	0.00	100.00	99.15	Ribosomal_S4	S4
GLEANR_12149	Kelleher92	PGRP-LF-PA	A	1	0.01	0.03	0.32	98.20	98.5	Amidase_2	
GLEANR_1216	Kelleher93	Ef2b-PA	Q	0	0.00	0.05	0.00	100.00	98.86	GTP_EFTU	EFG_IV
GLEANR_12170	Kelleher94	CG5150-PA	S	1	0.00	0.11	0.01	99.78	97.86	Alk_phosphatase	
GLEANR_12175	Kelleher95	Ubp64E-PC	Q	0	0.00	0.01	0.00	100.00	99.67	UCH	
GLEANR_12218	Kelleher96	CG8042-PA	Q	0	0.01	0.04	0.17	99.06	98.59	UBX	
GLEANR_12238	Kelleher97	CG33054-PB	Q	0	0.00	0.04	0.00	100.00	98.82	A1pp	
GLEANR_1224	Kelleher98	Crc-PA	Q	0	0.03	0.01	2.04	94.37	97.65	bZIP	bZIP_2
GLEANR_12314	Kelleher99	CG13901-PA	Q	0	0.00	0.03	0.07	99.50	98.84	no	
GLEANR_12324	Kelleher100	CG3344-PA	S	0	0.01	0.05	0.13	98.69	98.17	peptidase_S10	
GLEANR_12325	Kelleher101	CG3344-PA	S	0	0.02	0.09	0.24	95.36	95.95	peptidase_S10	
GLEANR_12326	Kelleher102	CG3371-PA	Q	0	0.01	0.02	0.29	98.85	99.23	no	
GLEANR_1234	Kelleher103	TepIV-PA	S	0	0.01	0.06	0.12	98.39	98.21	A2M_comp	A2M
GLEANR_12346	Kelleher104	no hits	Q	0	0.00	0.00	0.19	100.00	100	PAX	Homeobox
GLEANR_12378	Kelleher105	Syx7-PA	Q	1	0.01	0.03	0.19	98.68	98.68	SNARE	Syntaxin
GLEANR_12436	Kelleher106	CAH2-PA	S	0	0.00	0.04	0.05	99.54	99.09	Carb_anhydrase	
GLEANR_12445	Kelleher107	CG7630-PA	Q	1	0.01	0.05	0.10	98.89	98.52	no	
GLEANR_1245	Kelleher108	no hits	Q	1	0.03	0.05	0.57	95.45	96.46	Lamp	
GLEANR_12468	Kelleher109	CG8560-PA	S	0	0.01	0.04	0.26	97.33	98.22	peptidase_M14	
GLEANR_12471	Kelleher110	Sh3beta-PB	Q	0	0.00	0.08	0.00	100.00	98.5	SH3BGR	
GLEANR_12479	Kelleher111	RpL8-PB	QS	0	0.00	0.01	0.00	100.00	99.59	Ribosomal_L2	Ribosomal_L2_C
GLEANR_12488	Kelleher112	no hits		1						Ribosomal_L2_C	Ribosomal_L2
GLEANR_12492	Kelleher113	CG7015-PA	Q	0	0.00	0.03	0.00	100.00	99.12	CSD	
GLEANR_12502	Kelleher114	CG10359-PA	S	0	0.01	0.05	0.12	98.54	98.37	Fibrinogen_C	
GLEANR_1252	Kelleher115	fok-PA	Q	0	0.01	0.02	0.23	98.82	98.82	no	
GLEANR_12553	Kelleher116	CG17737-PA	Q	0	0.00	0.00	0.22	100.00	100	SUI1	
GLEANR_12608	Kelleher117	loj-PB	S	2	0.00	0.00	0.00	100.00	100	EMP24_GP25L	
GLEANR_1263	Kelleher118	Ance-PB	S	0	0.00	0.03	0.13	99.27	98.94	peptidase_M2	

GLEANR_12687	Kelleher119	CG18778-PA	S	0	0.01	0.03	0.27	98.15	98.77	Chitin_bind_4	
GLEANR_12742	Kelleher120	RpL28-PD	Q	0	0.01	0.02	0.34	98.61	98.84	Ribosomal_L28e	
GLEANR_12749	Kelleher121	mge-PA	Q	0	0.00	0.00	0.29	100.00	100	Tom22	
GLEANR_12750	Kelleher122	CG32744-PA	Q	0						ubiquitin	
GLEANR_12762	Kelleher123	no hits	Q	0						no	
GLEANR_12765	Kelleher124	CG1146-PC	A	1	0.01	0.04	0.33	96.97	97.98	no	
GLEANR_12819	Kelleher125	no hits	Q	0	0.00	0.08	0.00	100.00	98.04	HMG_box	
GLEANR_12820	Kelleher126	CG8583-PA	A	3	0.01	0.04	0.29	97.26	98.17	DnaJ	Sec63
GLEANR_12829	Kelleher127	CG18417-PA	Q	0	0.04	0.10	0.36	94.00	95.25	peptidase_M14	
GLEANR_12879	Kelleher128	RpL14-PA	Q	0	0.00	0.00	0.00	100.00	100	Ribosomal_L14e	
GLEANR_12884	Kelleher129	CG6416-PA	Q	0	0.00	0.00	0.32	100.00	100	PDZ	
GLEANR_12910	Kelleher130	Adgf-A-PB	S	0	0.00	0.02	0.00	100.00	99.51	A_deaminase	A_deaminase_N
GLEANR_12931	Kelleher131	CG14820-PA	S	0	0.01	0.06	0.16	97.83	97.99	peptidase_M14	Propep_M14
GLEANR_12984	Kelleher132	CG13887-PB	A	3	0.00	0.03	0.13	99.17	98.62	Bap31	
GLEANR_12991	Kelleher133	CG8177-PJ	Q	1 1	0.01	0.06	0.24	96.67	97.78	HCO3_cotransp	Band_3_cyto
GLEANR_13009	Kelleher134	CG9153-PA	Q	1	0.00	0.09	0.03	99.35	97.82	HECT	RCC1
GLEANR_13018	Kelleher135	eIF-2beta-PA	Q	0	0.00	0.03	0.00	100.00	98.67	eIF-5_eIF-2B	
GLEANR_13023	Kelleher136	CG10638-PA	Q	0	0.02	0.08	0.26	94.81	96.75	Aldo_ket_red	
GLEANR_1304	Kelleher137	CG10470-PA	S	2	0.00	0.06	0.00	100.00	98.32	DUF841	
GLEANR_13041	Kelleher138	Hsp26-PA	Q	0	0.01	0.10	0.13	96.77	96.77	HSP20	
GLEANR_13067	Kelleher139	Eip71CD-PB	Q	0	0.00	0.09	0.04	99.22	97.67	PMSR	
GLEANR_13121	Kelleher140	DnaJ-1-PB	Q	0	0.00	0.01	0.00	100.00	99.57	DnaJ	DnaJ_C
GLEANR_13126	Kelleher141	CG10592-PA	S	0	0.01	0.09	0.08	98.31	97.75	Alk_phosphatase	
GLEANR_13129	Kelleher142	sinu-PA	A	4	0.00	0.05	0.00	100.00	98.71	Clc-like	
GLEANR_13130	Kelleher143	no hits	S	0	0.04	0.19	0.23	92.59	92.06	Collagen	
GLEANR_13131	Kelleher144	no hits	S	0	0.00	0.03	0.08	99.39	98.99	no	
GLEANR_13148	Kelleher145	UGP-PC	Q	0	0.00	0.07	0.00	100.00	98.37	UDPGP	
GLEANR_1316	Kelleher146	NC2beta-PA	Q	0	0.00	0.04	0.00	100.00	98.69	CBFD_NFYB_HMF	
GLEANR_1319	Kelleher147	yellow-c-PA	S	0	0.00	0.03	0.09	99.35	99.13	MRJP	
GLEANR_13204	Kelleher148	RpS17-PB	Q	0	0.00	0.05	0.00	100.00	98.77	Ribosomal_S17e	
GLEANR_13209	Kelleher149	Nc73EF-PE	S	0	0.00	0.04	0.00	100.00	99.04	E1_dh	Transket_pyr
GLEANR_13220	Kelleher150	no hits	S	8						DUF300	
GLEANR_13227	Kelleher151	no hits	Q	0						zf-MIZ	
GLEANR_13246	Kelleher152	CG7324-PA	Q	0	0.00	0.07	0.00	100.00	98.49	TBC	GRAM
GLEANR_13248	Kelleher153	CG11309-PB	Q	0	0.01	0.03	0.42	97.22	98.15	Abhydrolase_1	
GLEANR_13271	Kelleher154	CG4769-PA	A	0	0.00	0.02	0.00	100.00	99.66	Cytochrom_C1	
GLEANR_13283	Kelleher155	Hsp83-PA	Q	0	0.00	0.03	0.00	100.00	99.25	HSP90	HATPase_c
GLEANR_1329	Kelleher156	CG11034-PA	S	0	0.01	0.08	0.06	98.80	97.87	DPPIV_N	Peptidase_S9
GLEANR_13330	Kelleher157	CG6020-PA	S	0	0.00	0.05	0.06	99.41	98.42	Epimerase	3Beta_HSD
GLEANR_13351	Kelleher158	CG7369-PA	Q	0	0.00	0.02	0.00	100.00	99.27	RasGEF	RasGEF_N
GLEANR_13378	Kelleher159	RpL10Ab-PA	Q	0	0.00	0.01	0.00	100.00	99.74	Ribosomal_L1	
GLEANR_13389	Kelleher160	Pros54-PA	Q	0	0.00	0.09	0.05	99.10	97.6	UIM	
GLEANR_13391	Kelleher161	CG7597-PA	Q	0	0.01	0.03	0.29	98.05	98.44	Pkinase	Pkinase_Tyr
GLEANR_13409	Kelleher162	RpL26-PA	Q	0	0.00	0.02	0.00	100.00	99.78	KOW	
GLEANR_1342	Kelleher163	CG9140-PA	Q	0	0.00	0.04	0.00	100.00	98.63	Complex1_51K	

GLEANR_13497	Kelleher164	CG32195-PA	Q	01	0.02	0.02	0.69	96.20	98.31	DUF227	
GLEANR_13559	Kelleher165	CG10960-PB	Q	2	0.02	0.03	0.79	95.70	97.85	Sugar_tr	MFS_1
GLEANR_13569	Kelleher166	no hits	Q	0	0.00	0.00	0.22	100.00	100	GRASP55_65	
GLEANR_13571	Kelleher167	Su(Tpl)-PA	Q	0	0.01	0.05	0.18	98.10	98.1	Occludin_ELL	
GLEANR_13573	Kelleher168	Rab8-PA	Q	0	0.00	0.03	0.12	99.27	99.03	Ras	Miro
GLEANR_13585	Kelleher169	CG5684-PA	Q	0	0.00	0.01	0.00	100.00	99.81	CAF1	
GLEANR_13590	Kelleher170	ApepP-PA	Q	0	0.00	0.07	0.05	99.16	98.03	peptidase_M24	
GLEANR_13593	Kelleher171	no hits	Q	0	0.01	0.01	0.55	98.96	99.31	IMD	SH3_1
GLEANR_13675	Kelleher172	eIF4AIII-PA	Q	0	0.00	0.00	0.00	100.00	99.84	DEAD	Helicase_C
GLEANR_13676	Kelleher173	RpLP0-PA	Q	0	0.00	0.00	0.08	100.00	100	Ribosomal_L10	Ribosomal_60s
GLEANR_13679	Kelleher174	no hits	S	2	0.00	0.00	0.00	100.00	100	DOMON	Cu2_monoox_C
GLEANR_1368	Kelleher175	Eno-PA	Q	0	0.00	0.05	0.00	100.00	98.86	Enolase_C	Enolase_N
GLEANR_1374	Kelleher176	CaBP1-PA	S	0	0.00	0.09	0.04	99.23	97.18	Thioredoxin	
GLEANR_13794	Kelleher177	CG4446-PA	Q	0	0.00	0.05	0.00	100.00	98.82	no	
GLEANR_1388	Kelleher178	Akap200-PD	Q	0	0.00	0.03	0.06	99.57	98.72	no	
GLEANR_13880	Kelleher179	Prestin-PA	Q	9	0.00	0.08	0.05	99.10	97.6	Sulfate_transp	STAS
GLEANR_14090	Kelleher180	eIF-5A-PB	Q	0	0.00	0.03	0.00	100.00	99.39	eIF-5a	KOW
GLEANR_14091	Kelleher181	ATPsyn-beta-PA	S	0	0.00	0.02	0.00	100.00	99.42	ATP-synt_ab	ATP-synt_ab_C
GLEANR_14114	Kelleher182	Crk-PA	Q	0	0.01	0.05	0.15	98.29	98.48	SH3_2	SH2
GLEANR_14162	Kelleher183	bt-PD	Q	0	0.00	0.08	0.06	99.08	97.25	I-set	ig
GLEANR_14192	Kelleher184	RpS3A-PA	Q	0	0.00	0.02	0.00	100.00	99.38	Ribosomal_S3Ae	
GLEANR_14207	Kelleher185	no hits	Q	0	0.04	0.16	0.23	90.99	94.59	Ank	Gar1
GLEANR_1467	Kelleher186	CG6860-PB	S	1	0.02	0.10	0.18	95.95	96.4	LRR_1	
GLEANR_14694	Kelleher187	Ubi-p63E-PB	Q	0	0.00	0.07	0.00	100.00	97.93	ubiquitin	
GLEANR_14790	Kelleher188	Pros45-PA	Q	0	0.00	0.03	0.05	99.66	98.97	AAA	AAA_2
GLEANR_14827	Kelleher189	l(1)G0230-PA	Q	0	0.00	0.08	0.03	99.36	97.88	ATP-synt_DE_N	ATP-synt_DE
GLEANR_14844	Kelleher190	CG10958-PA	Q	0	0.00	0.10	0.03	99.29	97.62	no	
GLEANR_14860	Kelleher191	RpS28b-PA	Q	0	0.00	0.00	0.47	96.00	100	Ribosomal_S28e	
GLEANR_14919	Kelleher192	CG10260-PB	Q	01	0.00	0.03	0.00	100.00	99.22	PI3_PI4_kinase	PI3Ka
GLEANR_1493	Kelleher193	hoe1-PC	Q	1	0.00	0.11	0.00	100.00	97.73	CitMHS	
GLEANR_15004	Kelleher194	RpLP2-PB	Q	0	0.01	0.02	0.40	97.98	98.65	Ribosomal_60s	
GLEANR_1522	Kelleher195	CG4887-PA	Q	0	0.01	0.02	0.60	98.09	98.94	G-patch	RRM
GLEANR_15404	Kelleher196	CG12576-PA	S	2						no	
GLEANR_15414	Kelleher197	Idgf4-PA	S	0	0.00	0.03	0.07	99.46	99.09	Glyco_hydro_18	
GLEANR_1543	Kelleher198	no hits	A	2						Reticulon	
GLEANR_15461	Kelleher199	Ag5r2-PA	Q	0	0.02	0.09	0.29	94.55	96.36	SCP	
GLEANR_15469	Kelleher200	CG6842-PA	Q	0	0.00	0.02	0.00	100.00	99.36	AAA	MIT
GLEANR_1549	Kelleher201	cni-PA	A	3	0.00	0.04	0.00	100.00	98.94	Cornichon	
GLEANR_15506	Kelleher202	no hits	S	0	0.00	0.00	NA	100.00	100	Troponin	
GLEANR_15520	Kelleher203	CG7033-PA	S	0	0.00	0.05	0.05	99.43	98.67	Cpn60_TCP1	
GLEANR_15547	Kelleher204	Hex-A-PA	Q	0	0.00	0.01	0.00	100.00	99.7	Hexokinase_2	Hexokinase_1
GLEANR_15585	Kelleher205	CG4949-PA	Q	0	0.00	0.03	0.00	100.00	99.07	no	
GLEANR_15663	Kelleher206	His3.3B-PB	Q	0	0.00	0.02	0.00	100.00	99.51	Histone	
GLEANR_15674	Kelleher207	CG5703-PA	Q	0	0.00	0.11	0.00	100.00	97.62	Complex1_24kDa	

GLEANR_15754	Kelleher208	CG11417-PA	Q	0	0.00	0.11	0.00	100.00	98.42	NUC153	
GLEANR_1576	Kelleher209	Pdsw-PA	Q	0	0.00	0.03	0.10	99.38	99.18	no	
GLEANR_15807	Kelleher210	no hits	Q	0	0.01	0.02	0.64	97.10	98.55	no	
GLEANR_15851	Kelleher211	skpA-PE	Q	0	0.01	0.00	NA	98.18	99.39	Skp1	Skp1_POZ
GLEANR_15933	Kelleher212	Cbp80-PB	Q	0	0.00	0.04	0.00	100.00	99.29	MIF4G	
GLEANR_15954	Kelleher213	CG17841-PA	Q	6	0.02	0.10	0.16	96.77	96.06	no	
GLEANR_15960	Kelleher214	RpL17-PB	Q	0	0.00	0.01	0.00	100.00	99.64	Ribosomal_L22	
GLEANR_15970	Kelleher215	Gclc-PA	Q	0	0.00	0.03	0.04	99.71	99.24	GCS	
GLEANR_15985	Kelleher216	CG12065-PA	Q	0	0.00	0.06	0.03	99.62	98.74	no	
GLEANR_15988	Kelleher217	CG9099-PA	Q	0	0.00	0.01	0.00	100.00	99.52	SUI1	
GLEANR_15995	Kelleher218	CG17754-PC	Q	0	0.00	0.02	0.00	100.00	99.42	Kelch	BACK
GLEANR_15999	Kelleher219	no hits	Q	0	0.03	0.01	2.59	93.85	97.44	RA	PH
GLEANR_16002	Kelleher220	CG1640-PF	Q	0	0.00	0.03	0.15	99.07	99.07	Aminotran_1_2	
GLEANR_16030	Kelleher221	ran-PA	Q	0	0.00	0.03	0.00	100.00	99.65	Ras	Miro
GLEANR_16139	Kelleher222	l(1)G0289-PA	S	1	0.01	0.03	0.41	97.12	98.08	PSI	
GLEANR_16147	Kelleher223	CG3446-PA	A	1	0.01	0.03	0.20	98.95	98.95	GRIM-19	
GLEANR_16173	Kelleher224	RpL22-PA	Q	0	0.01	0.03	0.19	99.10	98.49	Ribosomal_L22e	
GLEANR_16236	Kelleher225	CG10992-PA	S	0	0.00	0.03	0.15	99.17	98.9	peptidase_C1	Propeptide_C1
GLEANR_16280	Kelleher226	no hits	S	1	0.01	0.03	0.23	99.25	98.26	no	
GLEANR_16311	Kelleher227	CG6461-PA	S	0	0.01	0.00	NA	98.15	99.38	G_glu_transpept	
GLEANR_16330	Kelleher228	CG11642-PC	Q	8	0.00	0.06	0.00	100.00	98.82	TRAM1	LAG1
GLEANR_16345	Kelleher229	CG17633-PA	S	0	0.02	0.05	0.43	96.01	97.26	peptidase_M14	
GLEANR_1635	Kelleher230	RpL37A-PB	Q	0	0.00	0.00	0.00	100.00	100	Ribosomal_L37ae	
GLEANR_16396	Kelleher231	no hits	S	0	0.05	0.09	0.55	91.49	94.33	no	
GLEANR_16445	Kelleher232	CG33254-PA	S	1	0.00	0.00	0.00	100.00	100	no	
GLEANR_16491	Kelleher233	up-PG	Q	0	0.00	0.04	0.00	100.00	98.33	no	
GLEANR_16517	Kelleher234	RpS14a-PB	Q	0	0.00	0.04	0.00	100.00	98.55	Ribosomal_S11	
GLEANR_16533	Kelleher235	RpS5a-PA	Q	0	0.00	0.01	0.00	100.00	99.56	Ribosomal_S7	
GLEANR_16593	Kelleher236	CkIIbeta-PA	Q	0	0.00	0.02	0.00	100.00	99.31	CK_II_beta	
GLEANR_16625	Kelleher237	CG32560-PA	Q	0	0.01	0.13	0.08	97.55	96.93	RasGAP	C2
GLEANR_16630	Kelleher238	CG7536-PA	Q	4	0.00	0.03	0.05	99.66	99.09	EXS	SPX
GLEANR_16645	Kelleher239	no hits	Q	0	0.04	0.03	1.10	91.30	96.38	no	
GLEANR_16654	Kelleher240	Yp3-PA	S	0	0.01	0.11	0.09	97.78	97.22	Lipase	
GLEANR_16690	Kelleher241	CG9691-PA	S	0	0.00	0.03	0.00	100.00	99.49	no	
GLEANR_16781	Kelleher242	sta-PD	Q	0	0.00	0.02	0.00	100.00	99.51	Ribosomal_S2	
GLEANR_16783	Kelleher243	no hits	S	2	0.03	0.06	0.49	97.14	96.19	Mpv17_PMP22	
GLEANR_1678a	Kelleher244	CLIP-190-PA	Q	0	0.00	0.01	0.00	100.00	99.71	CAP_GLY	
GLEANR_1678b	Kelleher245	CLIP-190-PD	Q	0	0.02	0.04	0.44	97.20	97.67	CAP_GLY	
GLEANR_16799	Kelleher246	CG11160-PA	Q	0	0.01	0.05	0.14	98.42	98.31	no	
GLEANR_16816	Kelleher247	exd-PC	Q	0	0.00	0.02	0.00	100.00	99.31	PBC	Homeobox
GLEANR_16857	Kelleher248	CG1637-PA	Q	0	0.00	0.05	0.09	99.01	98.68	Metallophos	
GLEANR_16919	Kelleher249	RpL7A-PC	Q	0	0.00	0.03	0.00	100.00	99.29	Ribosomal_L7Ae	
GLEANR_1696	Kelleher250	no hits	Q	0						zf-AD	
GLEANR_16960	Kelleher251	CG17896-PB	Q	0	0.00	0.04	0.00	100.00	98.9	Aldedh	
GLEANR_16971	Kelleher252	RpL35-PB	Q	0	0.00	0.01	0.00	100.00	99.63	Ribosomal_L29	
GLEANR_16975	Kelleher253	no hits	S	0						Tsg	

GLEANR_16977	Kelleher254	CG2650-PA	Q	0	0.01	0.04	0.21	98.25	98.39	DUF233	
GLEANR_17001	Kelleher255	mRpL49-PA	Q	0	0.01	0.05	0.26	97.01	98	Img2	
GLEANR_17046	Kelleher256	no hits	Q	0	0.02	0.08	0.23	95.35	97.67	Pkinase	Pkinase_Tyr
GLEANR_17067	Kelleher257	CG7846-PA	Q	0	0.00	0.10	0.00	100.00	98.55	Actin	
GLEANR_17128	Kelleher258	no hits	S	2	0.06	0.06	1.08	90.54	94.14	no	
GLEANR_17146	Kelleher259	RpS15Aa-PA	Q	0	0.00	0.01	0.00	100.00	99.74	Ribosomal_S8	
GLEANR_17148	Kelleher260	CG14235-PA	Q	0	0.00	0.09	0.00	100.00	99.13	COX6B	
GLEANR_17198	Kelleher261	VhaAC39-PA	Q	0	0.00	0.03	0.00	100.00	99.17	vATP-synt_AC39	
GLEANR_1720	Kelleher262	CG32744-PA	Q	0	0.00	0.07	0.00	100.00	98.18	ubiquitin	Ribosomal_L40e
GLEANR_1725	Kelleher263	TepII-PC	S	1	0.02	0.05	0.37	96.10	97.19	A2M_comp	A2M_N
GLEANR_17319	Kelleher264	CG4169-PA	Q	0	0.00	0.03	0.00	100.00	99.13	peptidase_M16	Peptidase_M16_C
GLEANR_17322	Kelleher265	Argk-PA	Q	0	0.00	0.06	0.00	100.00	98.81	ATP-gua_Ptrans	ATP-gua_PtransN
GLEANR_17329	Kelleher266	Pka-R1-PG	Q	0	0.00	0.03	0.00	100.00	99.29	cNMP_binding	
GLEANR_17341	Kelleher267	CG1299-PA	S	0	0.01	0.03	0.24	98.80	98.8	trypsin	
GLEANR_17342	Kelleher268	Ero1L-PB	S	1	0.00	0.05	0.04	99.53	98.58	ERO1	
GLEANR_17381	Kelleher269	no hits	A	1	0.00	0.03	0.00	100.00	98.96	no	
GLEANR_17461	Kelleher270	CG12272-PA	Q	0	0.00	0.02	0.26	99.05	99.05	no	
GLEANR_17471	Kelleher271	Baldspot-PA	Q	6	0.00	0.04	0.07	99.34	98.56	ELO	
GLEANR_17547	Kelleher272	RpL23A-PA	Q	0	0.00	0.02	0.00	100.00	99.26	Ribosomal_L23	Ribosomal_L23eN
GLEANR_17554	Kelleher273	CG11526-PB	Q	0	0.00	0.03	0.00	100.00	99.63	N1221	
GLEANR_17603	Kelleher274	Syx8-PA	Q	1	0.02	0.10	0.20	95.65	96.23	SNARE	
GLEANR_17614	Kelleher275	fax-PA	Q	0	0.00	0.04	0.00	100.00	99.32	no	
GLEANR_17617	Kelleher276	CG4729-PA	S	4	0.02	0.03	0.82	95.24	97.78	Acyltransferase	
GLEANR_17618	Kelleher277	CG4729-PA	A	4	0.01	0.05	0.22	98.02	98.02	Acyltransferase	
GLEANR_1762	Kelleher278	CG10373-PA	Q	3	0.01	0.03	0.18	98.72	98.72	PRA1	
GLEANR_1784	Kelleher279	CG31919-PC	Q	0	0.00	0.02	0.00	100.00	99.49	no	
GLEANR_1803	Kelleher280	CG8498-PA	Q	0	0.00	0.02	0.00	100.00	99.25	ACBP	
GLEANR_1820	Kelleher281	CG10602-PB	Q	0	0.00	0.12	0.00	100.00	97.05	peptidase_M1	
GLEANR_1828	Kelleher282	CG11455-PA	Q	0	0.00	0.06	0.07	99.01	98.35	no	
GLEANR_1835	Kelleher283	CG11555-PA	Q	0	0.04	0.06	0.65	93.20	95.79	no	
GLEANR_1850	Kelleher284	CG17331-PA	Q	0	0.00	0.09	0.00	100.00	98.96	Proteasome	
GLEANR_1862	Kelleher285	Ugt36Bc-PB	S	1	0.01	0.09	0.14	97.35	97.13	UDPGT	
GLEANR_1893	Kelleher286	NLaz-PA	S	6	0.00	0.05	0.00	100.00	98.58	Lipocalin_2	
GLEANR_1934	Kelleher287	CG33129-PA	Q	0	0.00	0.02	0.27	99.07	99.38	no	
GLEANR_1967	Kelleher288	CG7778-PA	S	1	0.05	0.05	0.96	90.20	95.1	no	
GLEANR_1988	Kelleher289	CG6746-PA	Q	6	0.01	0.09	0.09	98.24	97.36	PTPLA	
GLEANR_1989	Kelleher290	CG31900-PA	S	1	0.01	0.21	0.05	97.50	94.58	no	
GLEANR_2008	Kelleher291	RpS26-PB	Q	0	0.00	0.17	0.00	100.00	96.75	Ribosomal_S26e	
GLEANR_2020	Kelleher292	beta'Cop-PA	Q	0	0.00	0.06	0.00	100.00	98.26	Coatomer_WDAD	WD40
GLEANR_2023	Kelleher293	CG10237-PA	Q	0	0.00	0.07	0.03	99.52	98.56	CRAL_TRIO	
GLEANR_2026	Kelleher294	CG17549-PA	Q	0	0.01	0.06	0.18	97.61	97.77	no	
GLEANR_2065	Kelleher295	Idgf2-PA	S	0	0.01	0.03	0.17	98.95	98.6	Glyco_hydro_18	
GLEANR_2097	Kelleher296	l(2)35Di-PA	Q	1	0.00	0.07	0.00	100.00	98.55	no	
GLEANR_2102	Kelleher297	no hits	S	0	0.01	0.06	0.17	97.93	98.16	TGF_beta	

GLEANR_2103	Kelleher298	nrv2-PA	Q	1	0.00	0.07	0.00	100.00	98.4	Na_K-ATPase	
GLEANR_2121	Kelleher299	no hits	Q	0	0.00	0.12	0.00	100.00	98.61	no	
GLEANR_2135	Kelleher300	TfIIS-PA	Q	0	0.00	0.03	0.00	100.00	99.06	TFIIS_M	TFIIS_C
GLEANR_2153	Kelleher301	CG15173-PA	Q	0	0.01	0.07	0.10	98.53	97.55	TPR_2	TPR_1
GLEANR_2160	Kelleher302	CG17294-PA	Q	0	0.00	0.03	0.00	100.00	98.99	Hydrolase	
GLEANR_2277	Kelleher303	no hits	Q	2	0.01	0.03	0.35	98.26	98.55	CD36	
GLEANR_2283	Kelleher304	no hits	Q	0						Ribosomal_L21e	
GLEANR_2286	Kelleher305	no hits	Q	0	0.00	0.05	0.03	99.58	99.17	Ribosomal_L18p	
GLEANR_2289	Kelleher306	no hits	Q	1	0.01	0.06	0.18	97.87	97.87	no	
GLEANR_2297	Kelleher307	retm-PA	Q	0	0.00	0.02	0.00	100.00	99.57	MSF1	CRAL_TRIO
GLEANR_2334	Kelleher308	Gpdh-PB	Q	0	0.00	0.02	0.00	100.00	99.4	NAD_Gly3P_dh_C	NAD_Gly3P_dh_N
GLEANR_2339	Kelleher309	eIF-4a-PD	Q	0	0.00	0.03	0.04	99.74	99.23	DEAD	Helicase_C
GLEANR_2371	Kelleher310	Pka-C1-PB	Q	0	0.00	0.03	0.00	100.00	99.53	Pkinase	Pkinase_Tyr
GLEANR_2380	Kelleher311	RpL13-PA	Q	0	0.00	0.05	0.00	100.00	98.78	Ribosomal_L13e	
GLEANR_2398	Kelleher312	RpL7-PA	Q	0	0.00	0.01	0.00	100.00	99.6	Ribosomal_L30	Ribosomal_L30_N
GLEANR_2406	Kelleher313	Pect-PD	Q	0	0.00	0.02	0.00	100.00	99.39	CTP_transf_2	
GLEANR_2450	Kelleher314	PRL-1-PA	Q	0	0.00	0.03	0.00	100.00	99.35	no	
GLEANR_2458	Kelleher315	ref(2)P-PA	Q	0	0.03	0.04	0.73	94.87	97.44	ZZ	
GLEANR_2521	Kelleher316	CG4598-PA	S	0	0.00	0.04	0.00	100.00	98.86	ECH	
GLEANR_2563	Kelleher317	for-PJ	Q	0	0.00	0.06	0.00	100.00	99	Pkinase	cNMP_binding
GLEANR_2575	Kelleher318	CG31954-PA	S	0	0.07	0.14	0.48	89.29	92.02	trypsin	
GLEANR_2607	Kelleher319	porin-PC	Q	0	0.00	0.05	0.09	99.04	98.4	Porin_3	
GLEANR_2624	Kelleher320	no hits	S	0							
GLEANR_2678	Kelleher321	CG9894-PA	Q	0	0.00	0.02	0.00	100.00	98.96	no	
GLEANR_2689	Kelleher322	La-PB	Q	0	0.00	0.17	0.00	100.00	97.66	La	
GLEANR_2703	Kelleher323	CG15254-PA	S	0	0.09	0.12	0.77	85.60	91.2	astacin	
GLEANR_2708	Kelleher324	CG6115-PA	Q	0	0.00	0.06	0.00	100.00	98.8	Complex1_LYR	
GLEANR_2757	Kelleher325	RpL27A-PA	Q	01	0.00	0.02	0.00	100.00	99.33	L15	
GLEANR_2775	Kelleher326	CG4726-PA	Q	0	0.03	0.05	0.59	93.33	96.51	MFS_1	
GLEANR_2786	Kelleher327	cl-PA	Q	0	0.00	0.05	0.00	100.00	99.14	DUF953	
GLEANR_2844	Kelleher328	CG10882-PA	Q	0	0.00	0.15	0.00	100.00	96.07	Sec23_trunk	Sec23_helical
GLEANR_2852	Kelleher329	CG8891-PA	Q	0	0.00	0.08	0.03	99.43	97.54	Ham1p_like	
GLEANR_2877	Kelleher330	Cyt-c-p-PA	Q	0	0.00	0.04	0.00	100.00	99.07	Cytochrom_C	
GLEANR_2879	Kelleher331	Rack1-PA	Q	0	0.00	0.05	0.00	100.00	98.67	WD40	
GLEANR_2881	Kelleher332	CG14536-PB	Q	1	0.03	0.05	0.54	95.71	96.67	ubiquitin	
GLEANR_2914	Kelleher333	CG10570-PA	Q	0	0.05	0.06	0.81	89.83	95.48	no	
GLEANR_3019	Kelleher334	RpL24-PA	Q	0	0.00	0.04	0.00	100.00	98.92	Ribosomal_L24e	
GLEANR_3021	Kelleher335	CG10026-PB	Q	0	0.00	0.05	0.06	99.25	98.63	CRAL_TRIO	CRAL_TRIO_N
GLEANR_3062	Kelleher336	CG4968-PA	Q	0	0.01	0.05	0.26	97.20	97.82	no	
GLEANR_3068	Kelleher337	Pten-PD	Q	0	0.02	0.18	0.09	96.15	95.73	no	
GLEANR_3081	Kelleher338	CG4972-PA	S	2	0.01	0.10	0.11	97.52	97.11	no	
GLEANR_3129	Kelleher339	Pros35-PA	Q	0	0.01	0.07	0.12	97.86	98.01	Proteasome	
GLEANR_3152	Kelleher340	no hits	Q	0							
GLEANR_3271	Kelleher341	garz-PB	Q	0	0.00	0.09	0.00	100.00	98.03	Sec7	
GLEANR_3274	Kelleher342	slik-PE	Q	0	0.00	0.01	0.00	100.00	99.52	Pkinase	Pkinase_Tyr
GLEANR_3281	Kelleher343	CG9890-PA	Q	0	0.00	0.02	0.11	99.57	99.28	zf-C2H2	

GLEANR_3353	Kelleher344	RpL31-PA	Q	0	0.00	0.05	0.00	100.00	98.92	Ribosomal_L31e	
GLEANR_3367	Kelleher345	lectin-46Cb-PA	A	0	0.02	0.03	0.82	95.45	97.73	Lectin_C	
GLEANR_3369	Kelleher346	CG7712-PA	Q	0	0.00	0.04	0.00	100.00	99.19	Complex1_LYR	
GLEANR_3375	Kelleher347	no hits	Q	0	0.05	0.06	0.79	91.00	95.33	no	
GLEANR_3383	Kelleher348	shot-PA	Q	0	0.00	0.05	0.00	100.00	98.67	Spectrin	CH
GLEANR_3389	Kelleher349	CG8207-PA	Q	0	0.01	0.04	0.12	98.88	98.5	NTP_transferase	Hexapep
GLEANR_3395	Kelleher350	Scs-fp-PA	S	01	0.00	0.08	0.02	99.59	98.08	FAD_binding_2	Succ_DH_flav_C
GLEANR_3427	Kelleher351	Picot-PB	Q	0	0.00	0.03	0.11	99.15	99.15	MFS_1	
GLEANR_3432	Kelleher352	CG8306-PA	Q	3	0.00	0.11	0.00	100.00	97.66	NAD_binding_4	Sterile
GLEANR_3458	Kelleher353	CG10320-PA	Q	1	0.00	0.04	0.00	100.00	98.77	NDUF_B12	
GLEANR_3480	Kelleher354	hrg-PB	Q	0	0.00	0.03	0.06	99.60	98.92	PAP_central	PAP_RNA-bind
GLEANR_3500	Kelleher355	CG7777-PA	Q	6	0.00	0.08	0.00	100.00	98.28	MIP	
GLEANR_3506	Kelleher356	no hits	A	1	0.00	0.02	0.00	100.00	99.22	no	
GLEANR_3511	Kelleher357	CG4802-PA	Q	0	0.00	0.08	0.00	100.00	98.02	Mtap_PNP	
GLEANR_3554	Kelleher358	no hits	Q	0	0.03	0.04	0.72	93.02	96.9	no	
GLEANR_3580	Kelleher359	CG9172-PA	Q	0	0.00	0.03	0.00	100.00	99.23	Oxidored_q6	
GLEANR_3597	Kelleher360	pAbp-PA	Q	0	0.00	0.00	0.31	100.00	100	RRM_1	PABP
GLEANR_3606	Kelleher361	CG13430-PA	S	0	0.01	0.04	0.32	97.35	98.23	trypsin	
GLEANR_3613	Kelleher362	CG18067-PA	S	0	0.02	0.04	0.51	95.48	97.44	no	
GLEANR_3658	Kelleher363	CG11807-PA	Q	0	0.01	0.07	0.16	97.96	97.28	PX	LRR_1
GLEANR_3717	Kelleher364	CG13868-PA	Q	0	0.00	0.04	0.00	100.00	98.61	no	
GLEANR_3718	Kelleher365	CG11200-PB	S	0	0.01	0.04	0.19	98.33	98.33	adh_short	
GLEANR_3724	Kelleher366	CG13335-PA	Q	0	0.01	0.07	0.09	98.51	97.76	no	
GLEANR_3765	Kelleher367	CG1665-PA	Q	0	0.01	0.02	0.61	97.26	98.63	MOSC	MOSC_N
GLEANR_3771	Kelleher368	CG30010-PA	Q	0	0.01	0.07	0.19	97.04	97.28	no	
GLEANR_3775	Kelleher369	RpL11-PA	Q	0	0.00	0.01	0.00	100.00	99.64	Ribosomal_L5_C	Ribosomal_L5
GLEANR_3781	Kelleher370	CG5597-PA	Q	0	0.00	0.07	0.00	100.00	98.11	no	
GLEANR_3787	Kelleher371	Tal-PA	Q	0	0.00	0.07	0.05	99.41	97.45	Transaldolase	
GLEANR_3795	Kelleher372	RpL12-PC	Q	0	0.00	0.04	0.00	100.00	98.97	Ribosomal_L11	Ribosomal_L11_N
GLEANR_3797	Kelleher373	PebIII-PA	S	0	0.01	0.10	0.14	96.83	96.83	OS-D	
GLEANR_3801	Kelleher374	CG3209-PB	A	3	0.00	0.05	0.00	100.00	98.99	Acyltransferase	
GLEANR_3825	Kelleher375	CG10527-PA	Q	0	0.00	0.07	0.00	100.00	98.69	no	
GLEANR_3858	Kelleher376	FK506-bp2-PA	Q	0	0.00	0.02	0.00	100.00	99.26	FKBP_C	
GLEANR_3867	Kelleher377	Ggamma1-PC	Q	0	0.00	0.03	0.00	100.00	99.05	G-gamma	
GLEANR_3886	Kelleher378	RpS11-PC	Q	0	0.00	0.04	0.00	100.00	99.34	Ribosomal_S17	
GLEANR_3894	Kelleher379	CG8309-PA	Q	0	0.00	0.02	0.00	100.00	99.06	PCI	
GLEANR_3933	Kelleher380	Hsc70-5-PA	S	0	0.00	0.00	0.39	100.00	100	HSP70	
GLEANR_3934	Kelleher381	Cyt-b5-PA	Q	1	0.01	0.03	0.33	98.18	98.79	Cyt-b5	
GLEANR_3938	Kelleher382	RpL29-PA	Q	0	0.00	0.01	0.00	100.00	99.56	Ribosomal_L29e	
GLEANR_3940	Kelleher383	CG9485-PB	Q	0	0.00	0.09	0.00	100.00	98.13	GDE_C	
GLEANR_3941	Kelleher384	RpS15-PA	Q	0	0.00	0.02	0.00	100.00	99.55	Ribosomal_S19	
GLEANR_3958	Kelleher385	CG9812-PB	S	0	0.00	0.02	0.19	99.07	99.07	no	
GLEANR_3967	Kelleher386	CG13551-PA	Q	0	0.00	0.02	0.22	99.07	99.38	IATP	
GLEANR_4030	Kelleher387	CG1600-PC	Q	0	0.00	0.03	0.08	99.49	98.98	ADH_N	
GLEANR_4039	Kelleher388	l(2)k05713-PB	S	0	0.00	0.00	0.10	100.00	100	DAO	efhand

GLEANR_4061	Kelleher389	emp-PB	A	2	0.00	0.03	0.13	98.99	98.88	CD36	
GLEANR_4064	Kelleher390	RpL19-PA	Q	0	0.00	0.01	0.00	100.00	99.67	Ribosomal_L19e	
GLEANR_4112	Kelleher391	CG13585-PB	Q	1	0.00	0.05	0.09	99.04	98.72	no	
GLEANR_4113	Kelleher392	CG32625-PA	Q	0	0.01	0.03	0.42	97.08	98.25	UPF0224	
GLEANR_4173	Kelleher393	CG30219-PA	Q	0	0.01	0.13	0.08	97.52	96.69	no	
GLEANR_4230	Kelleher394	Rab2-PA	Q	0	0.00	0.04	0.00	100.00	98.93	Ras	Miro
GLEANR_4236	Kelleher395	l(2)k03203-PA	Q	0	0.00	0.01	0.00	100.00	99.54	no	
GLEANR_4245	Kelleher396	bic-PA	Q	0	0.00	0.03	0.00	100.00	99.24	NAC	
GLEANR_4270	Kelleher397	no hits	Q	0	0.00	0.03	0.00	100.00	98.67	HTH_psq	
GLEANR_4271	Kelleher398	psq-PL	Q	0	0.00	0.03	0.00	100.00	98.94	BTB	
GLEANR_4286	Kelleher399	CG12505-PA	Q	0	0.01	0.05	0.25	97.42	98.11	Retrotrans_gag	
GLEANR_4292	Kelleher400	CG15083-PA	A	4	0.01	0.02	0.31	98.64	98.87	no	
GLEANR_4353	Kelleher401	CG12918-PA	S	0	0.01	0.08	0.09	98.40	97.7	no	
GLEANR_4356	Kelleher402	Spn6-PA	Q	0	0.00	0.07	0.00	100.00	98.64	Serpin	
GLEANR_438	Kelleher403	Rsf1-PA	Q	0	0.00	0.06	0.00	100.00	98.81	RRM_1	
GLEANR_4384	Kelleher404	CG4670-PA	S	1	0.00	0.02	0.00	100.00	99.56	Evr1_Alr	Thioredoxin
GLEANR_4397	Kelleher405	Tsp42Ea-PB	A	4	0.00	0.10	0.00	100.00	97.78	Tetraspannin	
GLEANR_440	Kelleher406	CG5390-PA	S	0	0.01	0.08	0.11	98.08	97.82	trypsin	
GLEANR_4400	Kelleher407	Tsp42Ed-PA	A	4	0.01	0.00	NA	98.39	99.46	Tetraspannin	
GLEANR_4406	Kelleher408	Tsp42Ej-PA	A	4	0.00	0.09	0.04	99.24	97.98	Tetraspannin	
GLEANR_4444	Kelleher409	CG16936-PA	S	0	0.01	0.04	0.28	97.31	98.21	GST_N	GST_C
GLEANR_4476	Kelleher410	cathD-PA	S	0	0.00	0.00	0.07	100.00	100	Asp	A1_Propeptide
GLEANR_4482	Kelleher411	CG18812-PC	Q	0	0.00	0.03	0.00	100.00	99.22	A1pp	
GLEANR_4507	Kelleher412	CG6406-PB	Q	0	0.01	0.03	0.34	98.08	98.72	no	
GLEANR_4515	Kelleher413	ERp60-PA	S	0	0.00	0.06	0.08	99.00	98	Thioredoxin	
GLEANR_4542	Kelleher414	CG3957-PA	Q	0	0.00	0.05	0.00	100.00	98.84	WD40	
GLEANR_4546	Kelleher415	CG7997-PA	S	0	0.02	0.08	0.19	96.59	97	Melibiase	
GLEANR_4581	Kelleher416	CG1884-PB	Q	0	0.01	0.06	0.08	98.92	98.02	Not1	
GLEANR_4584	Kelleher417	ced-6-PA	Q	0	0.01	0.04	0.21	97.67	98.45	PID	
GLEANR_4622	Kelleher418	Act57B-PA	Q	01	0.00	0.00	0.29	100.00	100	Actin	
GLEANR_4627	Kelleher419	CG30344-PA	Q	0	0.03	0.03	1.28	94.85	96.91	MFS_1	
GLEANR_4633	Kelleher420	Mys45A-PA	Q	0	0.00	0.08	0.03	99.42	98.55	SDA1	NUC130_3NT
GLEANR_4634	Kelleher421	Pmm45A-PA	Q	0	0.02	0.04	0.42	96.83	97.74	PGM_PMM_I	PGM_PMM_II
GLEANR_4635	Kelleher422	CG11127-PA	A	1	0.01	0.09	0.09	98.26	97.25	no	
GLEANR_4644	Kelleher423	no hits	Q	0	0.02	0.03	0.57	96.58	98.01	DAGK_acc	DAGK_cat
GLEANR_4684	Kelleher424	CG12479-PA	Q	2	0.01	0.05	0.15	98.63	98.17	no	
GLEANR_4717	Kelleher425	Spn4-PC	Q	0	0.01	0.02	0.61	97.26	98.63	Serpin	
GLEANR_4733	Kelleher426	IM4-PA	S	0	0.00	0.00	0.21	100.00	100	no	
GLEANR_4758	Kelleher427	Hil-PB	Q	0	0.00	0.09	0.00	100.00	98.58	LIM	
GLEANR_4819	Kelleher428	Nup62-PA	S	0	0.01	0.08	0.13	97.71	97.4	Nsp1_C	
GLEANR_4840	Kelleher429	Amy-p-PA	S	0	0.00	0.08	0.03	99.35	98.91	Alpha-amylase	Alpha-amylase_C
GLEANR_4849	Kelleher430	Ef1alpha48D-PB	Q	0	0.00	0.01	0.00	100.00	99.78	GTP_EFTU	GTP_EFTU_D3
GLEANR_488	Kelleher431	oho23B-PA	Q	0	0.00	0.00	0.12	100.00	100	Ribosomal_S21e	
GLEANR_4880	Kelleher432	CanB2-PA	Q	0	0.00	0.03	0.00	100.00	99.07	efhand	
GLEANR_4894	Kelleher433	RpS24-PA	Q	0	0.00	0.03	0.00	100.00	99.49	Ribosomal_S24e	
GLEANR_4958	Kelleher434	Mp20-PB	Q	0	0.00	0.04	0.13	98.91	98.91	CH	Calponin
GLEANR_5002	Kelleher435	GstE6-PA	Q	0	0.01	0.02	0.63	97.16	98.42	GST_N	GST_C

GLEANR_5037	Kelleher436	CG15098-PA	A	4	0.01	0.02	0.80	97.24	98.71	no	
GLEANR_5105	Kelleher437	CG9436-PA	Q	0	0.01	0.02	0.27	98.61	99.07	Aldo_ket_red	
GLEANR_5127	Kelleher438	Vha36-PA	Q	0	0.00	0.07	0.00	100.00	98.64	ATP-synt_D	
GLEANR_5191	Kelleher439	blw-PA	Q	0	0.00	0.02	0.00	100.00	99.32	ATP-synt_ab	ATP-synt_ab_C
GLEANR_5224	Kelleher440	Glycogenin-PA	Q	0	0.00	0.04	0.00	100.00	99.15	no	
GLEANR_5246	Kelleher441	CG4692-PB	Q	1	0.00	0.00	0.46	100.00	100	no	
GLEANR_5301	Kelleher442	nito-PB	Q	0	0.00	0.03	0.00	100.00	99.13	SPOC	RRM_1
GLEANR_5307	Kelleher443	RpS23-PA	Q	0	0.00	0.05	0.00	100.00	99.05	Ribosomal_S12	
GLEANR_5328	Kelleher444	Amph-PA	Q	0	0.00	0.00	0.00	100.00	100	BAR	SH3
GLEANR_5351	Kelleher445	yellow-d-PA	S	1	0.01	0.06	0.18	97.50	98.12	MRJP	
GLEANR_5358	Kelleher446	CG30415-PA	Q	1	0.01	0.01	1.04	97.56	98.78	no	
GLEANR_5359	Kelleher447	eIF2B-delta-PC	Q	0	0.00	0.03	0.00	100.00	99.05	IF-2B	
GLEANR_5384	Kelleher448	14-3-3zeta-PD	S	0	0.00	0.01	0.00	100.00	99.63	14-3-3	
GLEANR_5392	Kelleher449	RpL23-PA	Q	0	0.00	0.02	0.00	100.00	99.26	Ribosomal_L14	
GLEANR_5396	Kelleher450	Gp150-PD	S	1	0.02	0.02	1.30	94.77	97.87	LRR_1	
GLEANR_5420	Kelleher451	no hits	Q	0	0.00	0.00	0.14	100.00	100	no	
GLEANR_5421	Kelleher452	CG8446-PA	Q	0	0.00	0.04	0.00	100.00	99.14	BPL_LipA_LipB	
GLEANR_5437	Kelleher453	Cam-PB	Q	0	0.00	0.03	0.00	100.00	99.66	efhand	
GLEANR_5447	Kelleher454	CG10306-PA	Q	0	0.00	0.02	0.11	99.50	99.33	SAC3_GANP	
GLEANR_5477	Kelleher455	CG2556-PA	Q	0							
GLEANR_5490	Kelleher456	Treh-PF	Q	0	0.02	0.06	0.24	96.55	97.41	Trehalase	
GLEANR_5491	Kelleher457	CG7686-PA	Q	0	0.00	0.09	0.03	99.40	97.99	LTV	
GLEANR_5512	Kelleher458	CG3136-PB	Q	0	0.01	0.06	0.17	98.21	97.97	bZIP	bZIP_2
GLEANR_5517	Kelleher459	no hits	Q	0	0.00	0.03	0.00	100.00	99.05	Ribosomal_L38e	
GLEANR_5519	Kelleher460	Mlp60A-PA	Q	0	0.00	0.05	0.09	98.91	98.55	LIM	
GLEANR_5548	Kelleher461	Ca-P60A-PA	Q	71	0.00	0.05	0.06	99.31	98.62	E1-E2_ATPase	Cation_ATPase_C
GLEANR_5550	Kelleher462	CG5594-PB	Q	0	0.00	0.03	0.00	100.00	99.17	AA_permease	
GLEANR_5560	Kelleher463	Spt5-PB	Q	0	0.00	0.05	0.00	100.00	98.46	Supt5	KOW
GLEANR_5561	Kelleher464	betaTub56D-PB	Q	0	0.00	0.03	0.00	100.00	99.15	Tubulin	Tubulin_C
GLEANR_5567	Kelleher465	par-1-PE	Q	0	0.00	0.03	0.13	99.21	98.77	Pkinase	Pkinase_Tyr
GLEANR_5586	Kelleher466	Gapdh1-PB	Q	0	0.00	0.01	0.00	100.00	99.71	Gp_dh_C	Gp_dh_N
GLEANR_5605	Kelleher467	Sec61beta-PA	Q	1	0.00	0.10	0.00	100.00	96.94	Sec61_beta	
GLEANR_5629	Kelleher468	CG9394-PA	S	0	0.01	0.07	0.10	98.12	98.75	GDPD	CBM_20
GLEANR_5674	Kelleher469	Cyp6a13-PA	S	1	0.00	0.05	0.00	100.00	98.67	p450	
GLEANR_5690	Kelleher470	CG17922-PA	Q	5	0.00	0.05	0.06	98.90	98.65	cNMP_binding	Ion_trans
GLEANR_5692	Kelleher471	CG30197-PA	S	1	0.00	0.06	0.00	100.00	98.85	WAP	
GLEANR_5738	Kelleher472	CG30371-PA	S	0	0.01	0.06	0.12	98.26	98.09	trypsin	CUB
GLEANR_5826	Kelleher473	RpS16-PA	Q	0	0.00	0.03	0.00	100.00	99.32	Ribosomal_S9	
GLEANR_5841	Kelleher474	CG6550-PA	Q	1	0.01	0.08	0.10	98.17	97.71	UPF0004	Radical_SAM
GLEANR_5917	Kelleher475	CG3502-PA	S	0	0.01	0.10	0.10	97.54	97.09	peptidase_M1	
GLEANR_5950	Kelleher476	Chit-PA	S	0	0.00	0.01	0.29	99.08	99.39	Glyco_hydro_18	
GLEANR_5956	Kelleher477	Ef1beta-PA	Q	0	0.00	0.04	0.06	99.52	98.41	EF1_GNE	
GLEANR_5957	Kelleher478	CG6421-PA	S	1	0.01	0.07	0.07	98.73	98.09	Destabilase	
GLEANR_5983	Kelleher479	CG8241-PA	Q	0	0.00	0.05	0.00	100.00	98.81	HA2	DUF1605
GLEANR_6000	Kelleher480	Cp1-PA	S	0	0.01	0.08	0.14	97.35	97.25	peptidase_C1	Inhibitor_I29
GLEANR_6017	Kelleher481	CG1623-PB	Q	0	0.00	0.06	0.00	100.00	98.64	no	

GLEANR_6033	Kelleher482	CG12384-PA	Q	0	0.00	0.12	0.00	100.00	96.75	no	
GLEANR_6052	Kelleher483	Aats-val-PA	Q	0	0.01	0.02	0.36	98.72	99.15	tRNA-synt	Anticodon_1
GLEANR_6054	Kelleher484	CG4627-PA	Q	2	0.01	0.02	0.62	96.99	98.5	no	
GLEANR_6067	Kelleher485	Vha16-PB	Q	4	0.00	0.00	0.45	100.00	100	ATP-synt_C	
GLEANR_6084	Kelleher486	Ngp-PA	Q	0	0.00	0.02	0.00	100.00	99.34	NGP1NT	MMR_HSR1
GLEANR_6085	Kelleher487	RpL18A-PA	Q	0	0.00	0.03	0.00	100.00	99.09	Ribosomal_L18ae	
GLEANR_618	Kelleher488	VhaSFD-PA	Q	0	0.01	0.00	NA	98.70	99.57	V-ATPase_H	
GLEANR_624	Kelleher489	Msp-300-PD	Q	0	0.03	0.04	0.69	94.92	97.01	no	
GLEANR_631	Kelleher490	CG31705-PB	S	0	0.03	0.14	0.18	95.33	94.08	no	
GLEANR_6443	Kelleher491	CG3192-PB	Q	1	0.00	0.04	0.00	100.00	99.22	NDUF_B8	
GLEANR_6456	Kelleher492	CG2924-PC	Q	0	0.00	0.00	0.01	100.00	100	no	
GLEANR_6493	Kelleher493	Rbp2-PC	Q	0	0.00	0.02	0.00	100.00	99.33	RRM_1	
GLEANR_6518	Kelleher494	sog-PA	Q	1	0.00	0.03	0.13	98.96	98.96	CHRD	VWC
GLEANR_6529	Kelleher495	CG32744-PA	Q	0	0.00	0.13	0.00	100.00	97.48	ubiquitin	
GLEANR_6554	Kelleher496	RpL37A-PA	Q	0	0.00	0.00	0.04	100.00	100	Ribosomal_L37e	
GLEANR_6649	Kelleher497	l(1)G0193-PB	S	0	0.01	0.02	0.67	97.56	98.37	no	
GLEANR_665	Kelleher498	RpLP1-PA	S	0	0.00	0.03	0.00	100.00	99.11	Ribosomal_60s	
GLEANR_6678	Kelleher499	CG33178-PA	A	3	0.01	0.02	0.29	98.76	98.96	MAPEG	
GLEANR_6702	Kelleher500	no hits	Q	0							
GLEANR_6725	Kelleher501	CG13403-PA	S	0	0.01	0.07	0.19	97.30	97.07	no	
GLEANR_6739	Kelleher502	Cyp1-PA	Q	0	0.00	0.02	0.00	100.00	99.6	Pro_isomerase	
GLEANR_6743	Kelleher503	arm-PA	Q	0	0.00	0.00	0.70	100.00	100	Arm	HEAT
GLEANR_6749	Kelleher504	regucalcin-PC	Q	0	0.00	0.05	0.05	99.38	98.76	SGL	
GLEANR_6792	Kelleher505	Ctr1A-PA	Q	3	0.00	0.01	0.00	100.00	99.56	Ctr	
GLEANR_6984	Kelleher506	CG10469-PA	S	0	0.01	0.03	0.46	97.44	98.53	trypsin	
GLEANR_705	Kelleher507	Acon-PB	Q	0	0.00	0.06	0.00	100.00	98.77	Aconitase	Aconitase_C
GLEANR_7051	Kelleher508	CG6283-PA	S	0	0.05	0.09	0.52	90.29	94.51	Lipase	
GLEANR_709	Kelleher509	Hel25E-PB	Q	0	0.00	0.09	0.00	100.00	97.78	DEAD	Helicase_C
GLEANR_722	Kelleher510	smi21F-PB	Q	0	0.00	0.00	0.38	100.00	100	no	
GLEANR_7235	Kelleher511	CG7048-PA	Q	0	0.00	0.02	0.16	99.40	99.4	Prefoldin	
GLEANR_7296	Kelleher512	RpS3-PA	Q	0	0.00	0.02	0.00	100.00	99.56	Ribosomal_S3_C	KH_2
GLEANR_7308	Kelleher513	CG13822-PA	S	0	0.01	0.11	0.07	98.29	96.72	GILT	
GLEANR_7312	Kelleher514	B52-PD	Q	0	0.00	0.02	0.00	100.00	99.16	RRM_1	
GLEANR_7314	Kelleher515	CG2781-PA	Q	6	0.01	0.04	0.13	98.78	98.78	ELO	
GLEANR_7317	Kelleher516	CG31522-PA	Q	5	0.00	0.00	0.00	100.00	100	ELO	
GLEANR_7353	Kelleher517	CG33722-PC	Q	0	0.01	0.12	0.06	98.65	96.4	UBX	RBD
GLEANR_7392	Kelleher518	CG17273-PA	Q	0	0.00	0.06	0.04	99.56	97.94	Adenylsucc_synt	
GLEANR_7393	Kelleher519	RpS20-PA	Q	0	0.00	0.02	0.00	100.00	99.22	Ribosomal_S10	
GLEANR_7394	Kelleher520	CG17271-PA	S	0	0.02	0.03	0.66	95.29	98.04	no	
GLEANR_7441	Kelleher521	Gcn2-PA	Q	0	0.00	0.11	0.02	99.47	97.7	Pkinase	RWD
GLEANR_746	Kelleher522	vir-1-PC	S	0	0.00	0.08	0.03	99.56	97.96	no	
GLEANR_747	Kelleher523	CG6579-PA	S	0	0.01	0.08	0.08	98.35	98.07	no	
GLEANR_7487	Kelleher524	TflIA-L-PA	Q	0	0.00	0.03	0.00	100.00	98.92	TFIIA	
GLEANR_7501	Kelleher525	CG11876-PD	S	0	0.00	0.07	0.05	99.35	97.84	Transket_pyr	Transketolase_C
GLEANR_7521	Kelleher526	no hits	S	1	0.01	0.02	0.36	98.89	98.89	no	
GLEANR_7541	Kelleher527	Aph-4-PB	S	0	0.00	0.09	0.02	99.55	98.33	Alk_phosphatase	
GLEANR_7556	Kelleher528	ATPsyn-gamma-PC	S	0	0.00	0.03	0.00	100.00	99.08	ATP-synt	

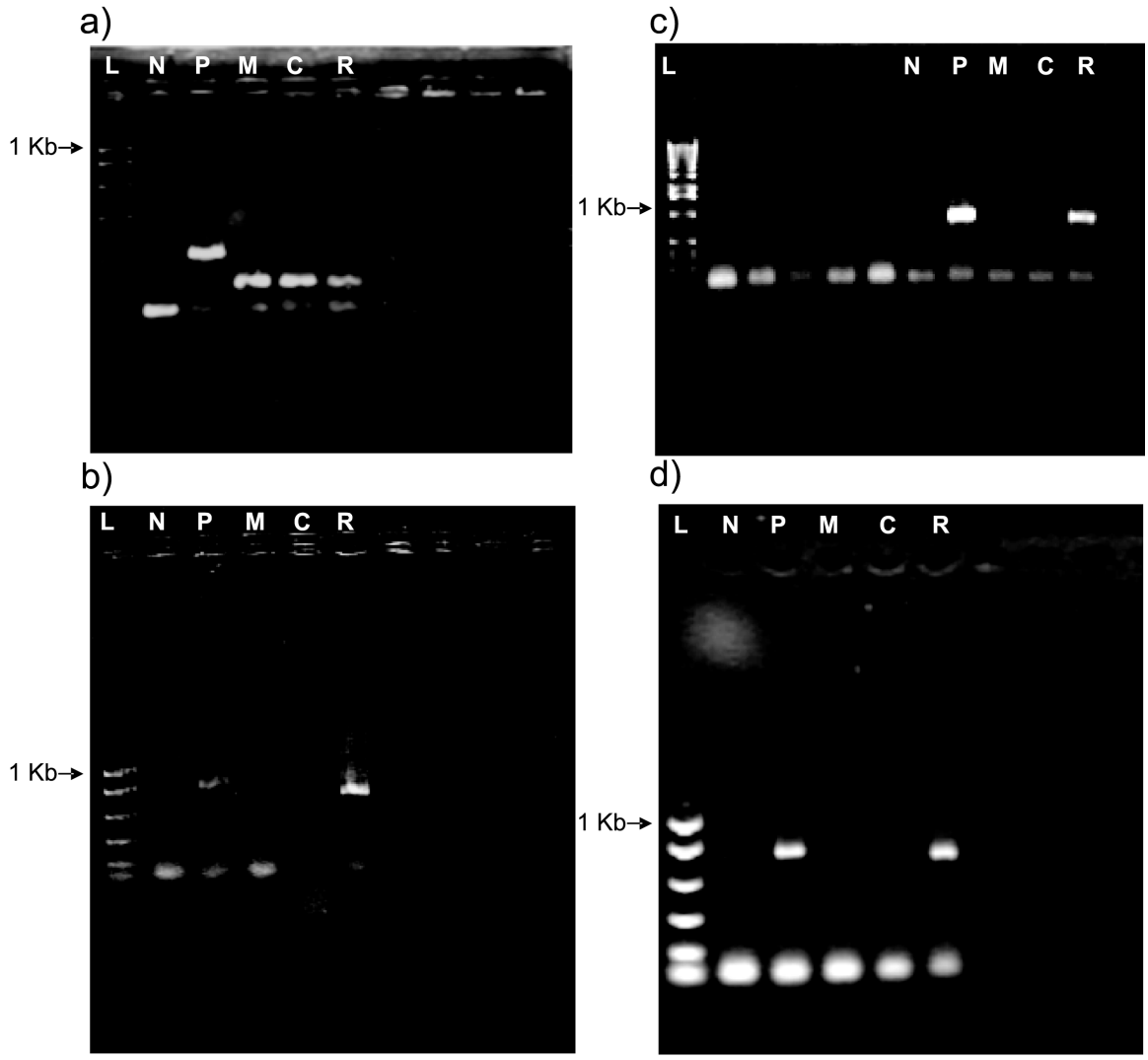
GLEANR_7572	Kelleher529	SP1029-PC	S	0	0.00	0.05	0.10	98.92	98.71	peptidase_M1	
GLEANR_7573	Kelleher530	Ef1gamma-PA	Q	0	0.00	0.04	0.00	100.00	99.07	EF1G	GST_N
GLEANR_7597	Kelleher531	PyK-PA	Q	0	0.00	0.04	0.00	100.00	99.11	PK	PK_C
GLEANR_7606	Kelleher532	epsin-like-PB	Q	0	0.00	0.01	0.00	100.00	99.44	ENTH	
GLEANR_7617	Kelleher533	CG10214-PA	Q	0	0.03	0.07	0.47	93.39	96.14	Exonuc_X-T	
GLEANR_7620	Kelleher534	CG10221-PA	S	1	0.00	0.03	0.00	100.00	99.06	Sel1	
GLEANR_7653	Kelleher535	RhoGAP92B-PA	Q	0	0.00	0.02	0.00	100.00	99.42	RhoGAP	BAR
GLEANR_770	Kelleher536	RpL36A-PA	Q	0	0.00	0.06	0.00	100.00	99.03	Ribosomal_L44	
GLEANR_7742	Kelleher537	CG5103-PA	Q	0	0.00	0.04	0.06	99.39	98.77	Transketolase_N	Transket_pyr
GLEANR_7755	Kelleher538	CG1161-PA	S	2	0.01	0.06	0.12	98.40	98.05	TMEM9	
GLEANR_7775	Kelleher539	CG10550-PB	Q	0	0.01	0.07	0.14	97.73	97.73	DUF227	
GLEANR_778	Kelleher540	CG15254-PA	S	0	0.03	0.09	0.31	93.65	96.03	astacin	
GLEANR_7826	Kelleher541	CG6666-PA	Q	3	0.01	0.12	0.06	98.18	97.27	Sdh_cyt	
GLEANR_7870	Kelleher542	no hits	Q	0						Fe-S_biosyn	
GLEANR_7892	Kelleher543	mRpL40-PA	S	0	0.01	0.08	0.09	98.38	97.84	no	
GLEANR_7906	Kelleher544	CG9836-PA	Q	0	0.01	0.07	0.09	98.70	98.05	NifU_N	
GLEANR_7910	Kelleher545	CG5023-PA	Q	0	0.00	0.17	0.00	100.00	97.92	CH	Calponin
GLEANR_7933	Kelleher546	RpS29-PA	Q	0	0.03	0.02	1.12	96.43	97.62	Ribosomal_S14	
GLEANR_8002	Kelleher547	CG5823-PA	Q	1	0.00	0.00	0.10	100.00	100	UQ_con	
GLEANR_8020	Kelleher548	CG2943-PA	S	1	0.01	0.03	0.19	98.62	98.62	DUF1620	
GLEANR_8029	Kelleher549	CG6359-PA	Q	0	0.00	0.02	0.00	100.00	99.28	PX	
GLEANR_8079	Kelleher550	RpL34a-PA	Q	0	0.00	0.04	0.00	100.00	99.16	Ribosomal_L34e	
GLEANR_8095	Kelleher551	Hsc70-4-PA	Q	0	0.00	0.07	0.00	100.00	98.65	HSP70	MreB_Mbl
GLEANR_8098	Kelleher552	Set-PA	Q	0	0.00	0.02	0.00	100.00	99.17	NAP	
GLEANR_8102	Kelleher553	Tm1-PB	Q	0	0.00	0.00	NA	99.15	99.72	Tropomyosin	
GLEANR_8104	Kelleher554	Tm2-PB	Q	0	0.00	0.02	0.00	100.00	99.41	Tropomyosin	
GLEANR_8135	Kelleher555	kuk-PB	Q	0	0.01	0.09	0.13	97.69	96.91	no	
GLEANR_8159	Kelleher556	RpS27-PA	A	0	0.00	0.00	0.20	100.00	100	Ribosomal_S27e	
GLEANR_8165	Kelleher557	VhaPPA1-1-PA	A	5	0.00	0.07	0.06	99.09	97.88	ATP-synt_C	
GLEANR_8170	Kelleher558	CG7523-PA	Q	4	0.00	0.01	0.00	100.00	99.54	no	
GLEANR_8204	Kelleher559	CG3153-PA	S	0	0.00	0.08	0.00	100.00	98.53	E1_DerP2_DerF2	
GLEANR_8221	Kelleher560	alphaTub84D-PA	Q	0	0.00	0.00	0.33	100.00	100	Tubulin	Tubulin_C
GLEANR_8254	Kelleher561	CG11858-PA	Q	0	0.01	0.08	0.09	96.43	98.2	no	
GLEANR_8258	Kelleher562	CG3734-PA	S	0	0.0298	0.0559	0.53	94.59	96.47	peptidase_S28	
GLEANR_8259	Kelleher563	CG3734-PA	S	0	0.02	0.09	0.20	96.58	96.58	peptidase_S28	
GLEANR_8270	Kelleher564	Octbeta2R-PA	S	7	0.01	0.10	0.09	97.95	97.95	7tm_1	
GLEANR_8284	Kelleher565	ttk-PD	Q	0	0.00	0.01	0.00	98.79	99.59	BTB	zf-C2H2
GLEANR_8303	Kelleher566	cher-PC	Q	0	0.00	0.07	0.00	100.00	98.49	Filamin	
GLEANR_8337	Kelleher567	Rm62-PE	Q	0	0.00	0.00	0.07	100.00	100	DEAD	Helicase_C
GLEANR_835	Kelleher568	CG9336-PA	S	0	0.02	0.07	0.23	96.69	97.25	no	
GLEANR_8352	Kelleher569	Mdh-PA	S	0	0.01	0.06	0.10	98.83	98.05	Malic_M	malic
GLEANR_8393	Kelleher570	CG9796-PA	S	2	0.00	0.05	0.04	99.55	98.79	GILT	
GLEANR_8424	Kelleher571	CG6124-PA	S	0	0.01	0.09	0.13	97.65	96.86	C_tripleX	EGF
GLEANR_8516	Kelleher572	Mlc2-PA	Q	0	0.00	0.00	0.00	100.00	100	no	
GLEANR_8528	Kelleher573	RpL32-PC	Q	0	0.00	0.02	0.00	100.00	99.5	Ribosomal_L32e	
GLEANR_8553	Kelleher574	CG5677-PA	A	1	0.00	0.02	0.14	99.42	99.03	SPC22	
GLEANR_8593	Kelleher575	CG12268-PA	Q	2	0.01	0.04	0.36	97.22	98.15	NAD_bindin	Sterile

										g_4	
GLEANR_8594	Kelleher576	Spn5-PB	S	0	0.00	0.06	0.03	99.52	98.65	Serpin	
GLEANR_8630	Kelleher577	CG1234-PA	Q	0	0.00	0.07	0.00	100.00	97.82	CBF	NOC3p
GLEANR_8647	Kelleher578	Kap-alpha3-PA	Q	0	0.00	0.02	0.00	100.00	99.37	Arm	IBB
GLEANR_8689	Kelleher579	hth-PF	Q	0	0.00	0.00	0.23	100.00	100	Homeobox	
GLEANR_8733	Kelleher580	CG16749-PA	S	0	0.03	0.07	0.39	94.70	96.34	trypsin	
GLEANR_8752	Kelleher581	CkIIalpha-PC	Q	0	0.00	0.03	0.00	100.00	99.02	Pkinase	Pkinase_Tyr
GLEANR_8787	Kelleher582	Zeelin1-PC	Q	0	0.00	0.02	0.00	100.00	99.8	no	
GLEANR_8796	Kelleher583	RpS13-PA	Q	0	0.00	0.04	0.00	100.00	98.9	Ribosomal_S13_N	Ribosomal_S15
GLEANR_8798	Kelleher584	Vha26-PA	Q	0	0.00	0.01	0.00	100.00	99.72	vATP-synt_E	
GLEANR_8820	Kelleher585	RpL3-PA	Q	0	0.00	0.05	0.00	100.00	98.84	Ribosomal_L3	
GLEANR_8841	Kelleher586	PP2A-B'-PB	Q	0	0.01	0.03	0.25	98.46	98.97	B56	
GLEANR_8848	Kelleher587	slmb-PA	Q	0	0.00	0.00	0.43	100.00	100	WD40	F-box
GLEANR_8856	Kelleher588	Rbp1-like-PA	Q	0	0.00	0.04	0.00	100.00	99.03	RRM_1	
GLEANR_8869	Kelleher589	Ahcy89E-PC	Q	0	0.01	0.05	0.11	98.77	98.36	AdoHcyase	AdoHcyase_NAD
GLEANR_8893	Kelleher590	no hits	Q	0	0.00	0.00	0.40	100.00	100	TBC	
GLEANR_8925	Kelleher591	Qm-PA	Q	0	0.00	0.05	0.00	100.00	98.62	Ribosomal_L10e	
GLEANR_8928	Kelleher592	CG32230-PB	S	1	0.00	0.04	0.00	100.00	99.2	no	
GLEANR_8934	Kelleher593	Atg2-PA	Q	0	0.01	0.04	0.29	97.62	98.41	no	
GLEANR_896	Kelleher594	CG31954-PA	S	0	0.11	0.12	0.89	83.71	90.03	trypsin	
GLEANR_897	Kelleher595	CG31954-PA	S	0	not orthologs					trypsin	
GLEANR_898	Kelleher596	CG31954-PA	S	0	0.05	0.12	0.44	90.00	93.57	trypsin	
GLEANR_9029	Kelleher597	Nep2-PA	A	2	0.01	0.01	0.80	97.56	98.86	peptidase_M13_N	Peptidase_M13
GLEANR_9057	Kelleher598	fau-PC	Q	0	0.00	0.01	0.00	100.00	99.72	no	
GLEANR_9073	Kelleher599	CG7218-PA	Q	5	0.01	0.07	0.17	97.30	97.3	DUF747	
GLEANR_9114	Kelleher600	CG11852-PA	S	0	0.02	0.05	0.43	95.37	97.22	DUF233	
GLEANR_9115	Kelleher601	no hits	S	0	0.03	0.00	NA	93.75	97.92	DUF233	
GLEANR_9116	Kelleher602	CG11854-PA	Q	0	0.02	0.06	0.34	96.22	97.06	DUF233	
GLEANR_9125	Kelleher603	RpS25-PB	Q	0	0.00	0.00	0.99	100.00	100	Ribosomal_S25	
GLEANR_915	Kelleher604	no hits	Q	0	0.00	0.05	0.06	99.28	98.55	no	
GLEANR_9150	Kelleher605	Crc-PA	S	0	0.00	0.05	0.00	100.00	98.91	Calreticulin	
GLEANR_9151	Kelleher606	no hits	Q	0	0.00	0.04	0.00	100.00	99	Ribosomal_L15e	
GLEANR_9161	Kelleher607	bai-PA	S	2	0.00	0.05	0.00	100.00	98.71	no	
GLEANR_9167	Kelleher608	CG9602-PA	Q	0	0.00	0.02	0.00	100.00	99.29	UQ_con	
GLEANR_9206	Kelleher609	CG13631-PA	Q	0	0.00	0.00	NA	100.00	100	no	
GLEANR_9214	Kelleher610	Timp-PA	S	0	0.00	0.03	0.00	100.00	99.58	TIMP	
GLEANR_9216	Kelleher611	CG12814-PA	Q	1	0.00	0.00	0.31	100.00	100	Zona_pellucida	
GLEANR_9240	Kelleher612	mfas-PN	S	0	0.00	0.05	0.00	100.00	98.88	Fasciclin	
GLEANR_9295	Kelleher613	alpha-Est1-PA	Q	0	0.01	0.06	0.19	98.77	97.94	COesterase	
GLEANR_9302	Kelleher614	CG31472-PA	Q	0	0.01	0.08	0.12	97.95	97.03	Pyridox_oxidase	
GLEANR_9312	Kelleher615	Rab7-PA	Q	0	0.00	0.06	0.00	100.00	98.48	Ras	Miro
GLEANR_9348	Kelleher616	no hits	Q	0						Ank	KH_1
GLEANR_9365	Kelleher617	CycG-PD	Q	0	0.01	0.06	0.11	98.72	97.86	Cyclin_N	
GLEANR_9366	Kelleher618	RpL6-PB	Q	0	0.00	0.03	0.10	99.18	99.18	Ribosomal_L6e	Ribosomal_L6e_N
GLEANR_9384	Kelleher619	no hits	S	0						SCP	
GLEANR_9388	Kelleher620	CG2185-PA	Q	0	0.00	0.02	0.00	100.00	99.47	efhand	
GLEANR_9393	Kelleher621	CG8031-PA	Q	0	0.00	0.07	0.02	99.64	98.22	UPF0103	

GLEANR_9418	Kelleher622	Vha100-2-PA	Q	6	0.00	0.06	0.00	100.00	98.61	V_ATPase_I	
GLEANR_9468	Kelleher623	RpL4-PA	Q	0	0.00	0.06	0.00	100.00	98.48	Ribosomal_L4	
GLEANR_9471	Kelleher624	Gp93-PA	S	0	0.00	0.04	0.07	99.43	99.05	HSP90	HATPase_c
GLEANR_9479	Kelleher625	RpS7-PA	Q	0	0.00	0.03	0.00	100.00	99.31	Ribosomal_S7e	
GLEANR_9488	Kelleher626	CG32473-PC	S	1	0.02	0.05	0.32	96.73	97.55	peptidase_M1	
GLEANR_9490	Kelleher627	CG8790-PA	Q	0	0.01	0.03	0.31	97.78	98.52	Mito_carr	
GLEANR_9493	Kelleher628	CG8863-PA	Q	0	0.00	0.14	0.00	100.00	97.38	DnaJ	DnaJ_CXXCXGXG
GLEANR_9495	Kelleher629	Act87E-PB	Q	0	0.00	0.11	0.00	100.00	97.78	Actin	
GLEANR_951	Kelleher630	CG13124-PA	Q	0	0.00	0.06	0.00	100.00	98.23	MIF4G	
GLEANR_952	Kelleher631	sop-PA	Q	0	0.00	0.04	0.00	100.00	98.15	Ribosomal_S5	Ribosomal_S5_C
GLEANR_9531	Kelleher632	GstD1-PA	Q	0	0.00	0.06	0.04	99.48	98.63	GST_N	GST_C
GLEANR_9617	Kelleher633	CG3739-PA	S	0	0.07	0.09	0.74	87.93	93.1	peptidase_S28	
GLEANR_9629	Kelleher634	Pcmt-PA	Q	0	0.00	0.12	0.00	100.00	97.33	PCMT	
GLEANR_9656	Kelleher635	gfzf-PC	Q	0	0.00	0.03	0.07	99.55	98.95	FLYWCH	GST_N
GLEANR_9657	Kelleher636	CG10068-PA	Q	0	0.00	0.06	0.00	100.00	98.44	ELMO_CED12	
GLEANR_9666	Kelleher637	beta4GalNAcTB-PA	A	1	0.01	0.13	0.09	97.42	96.56	Galactosyl_T_2	
GLEANR_9714	Kelleher638	no hits	Q	0							
GLEANR_975	Kelleher639	Grp1-PA	Q	0	0.00	0.02	0.00	100.00	99.61	Sec7	PH
GLEANR_9752	Kelleher640	Vha13-PA	Q	0	0.01	0.05	0.15	98.29	98.58	V-ATPase_G	
GLEANR_9765	Kelleher641	Msr-110-PC	Q	0	0.00	0.01	0.31	99.54	99.54	SWIRM	Myb_DNA-binding
GLEANR_9777	Kelleher642	CG17931-PA	Q	0	0.00	0.00	0.39	100.00	100	4F5	
GLEANR_9778	Kelleher643	CG10311-PA	A	4	0.00	0.04	0.00	100.00	98.89	no	
GLEANR_978	Kelleher644	CG7231-PB	Q	0	0.01	0.07	0.08	98.72	98.29	no	
GLEANR_980	Kelleher645	CG31758-PA	S	0	0.03	0.15	0.18	94.44	94.44	Kazal_2	Kazal_1
GLEANR_9845	Kelleher646	CG17836-PB	Q	0	0.02	0.04	0.54	95.68	97.12	no	
GLEANR_9848	Kelleher647	Cyp9f2-PA	S	1	0.01	0.11	0.05	98.76	97.31	p450	
GLEANR_9898	Kelleher648	m1-PA	S	0	0.01	0.11	0.06	98.59	97.18	no	
GLEANR_9937	Kelleher649	mod-PA	Q	0	0.01	0.06	0.08	98.79	98.38	RRM_1	
GLEANR_9965	Kelleher650	betaTub97EF-PA	Q	0	0.00	0.03	0.00	100.00	99.56	Tubulin	Tubulin_C
GLEANR_9982	Kelleher651	CG7443-PA	S	0	0.03	0.02	1.83	95.00	97.5	no	

Supplementary Table 1: Female Reproductive ESTs Identified in this Study. *D. mojavensis* CDS: coding sequence from GLEANR annotations (http://rana.lbl.gov/~venky/caf1), *D. arizonae* EST: Dari\anon-EST assignment for GenBank (http://www.ncbi.nlm.nih.gov/sites/entrez). *D. melanogaster* homolog identified by BLAST. SignalP: S=secreted, A=anchor, Q=quiescent as predicted by SignalP 3.0 [55], TMHMM: number of identified transmembrane domains [56], Ka: estimated non-synonymous substitutions per non-synonymous site Ks: estimated

synonymous substitutions per synonymous site, Ka/Ks: estimated ratio non-synonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site, PROT %ID: Protein % identity, CDS %ID: coding sequence % identity calculated in PAML[46], Conserved domain: pfam conserved domain predicted from hmmpfam [57].



Supplementary Figure 1. RT-PCR of a) Dmoj\GLEANR_8528 49 b) protease gene family 1 c) protease gene family 2 d) protease gene family 3. 1 Kb Markers are indicated. L: DNA ladder, N: negative control M: whole male cDNA C: female carcass (no lower reproductive tract) cDNA R: lower female reproductive tract cDNA

APPENDIX C: PROTEASE GENE DUPLICATION AND PROTEOLYTIC ACTIVITY IN *DROSOPHILA* FEMALE REPRODUCTIVE TRACTS

**This appendix has been submitted for publication:

Kelleher ES and Pennington JE. 2009. Protease Gene Duplication and Proteolytic Activity in *Drosophila* Female Reproductive Tracts. Molecular Biology and Evolution. *submitted*.

ABSTRACT

Secreted proteases play integral roles in sexual reproduction in a broad range of taxa. In the genetic model *Drosophila melanogaster*, these molecules are thought to process peptides and activate enzymes inside female reproductive tracts, mediating critical post-mating responses. A recent study of female reproductive tract proteins in the cactophilic fruit-fly *D. arizonae*, identified pervasive, lineage-specific gene duplication amongst secreted proteases. Here we compare the evolutionary dynamics, biochemical nature, and physiological significance of secreted female reproductive serine endoproteases (SFRSEs) between *D. arizonae* and its congener *D. melanogaster*. We show that *D. arizonae* lower female reproductive tract (LFRT) proteins are significantly enriched for recently-duplicated secreted proteases, particularly serine endoproteases, relative to *D. melanogaster*. Isolated lumen from *D. arizonae* LFRTs, furthermore, exhibits significant tryspin-like and elastase-like serine endoprotease acitivity, while no such activity is seen in *D. melanogaster*. Finally, trypsin and elastase-like activity in *D. arizonae* female reproductive tracts is negatively regulated by mating. We propose that the intense proteolytic environment of the *D. arizonae* female reproductive tract relates to the extraordinary reproductive physiology of this species, and that ongoing gene duplication amongst these proteases is an evolutionary consequence of sexual conflict.

INTRODUCTION

In internally fertilizing organisms, sexual reproduction is mediated by an elaborate series of interactions between the male ejaculate and the female reproductive tract. This interface extends far beyond gamete fusion, playing essential roles in sperm fate (Reviewed in Neubaum and Wolfner 1998), as well as female behavior and physiology (Reviewed in Wolfner 2007; Roberston 2007). Although reproductive tract interactions are fundamental to fertilization and organismal fitness, male ejaculates and female reproductive tracts are observed to evolve rapidly at both the morphological (Pitnick *et al* 1999; Brennan *et al* 2007; Marshall 2007) and biochemical levels (reviewed in Swanson and Vacquier 2002; Clark, Aagaard and Swanson 2006; Panhuis, Clark and Swanson 2006). This exceptional divergence often is hypothesized to be a consequence of a coevolutionary chase between males and females driven by sexual conflict, or a difference in the reproductive interests of the two sexes (Parker 1979; Rice 1996; Gavrilets 2000).

The molecular underpinnings of ejaculate-female dynamics remain poorly understood, however, proteases have emerged as prominent reproductive players in both insects (Swanson *et al* 2001; 2004; Braswell *et al* 2006; Sirot *et al* 2008), and mammals (reviewed in Dacheux, Gatti and Dacheux 2003). In *Drosophila melanogaster*, proteolysis thought to modulate female post-mating response by processing or activating male-derived peptides and enzymes (Monsma, Harada and Wolfner 1990; Park and Wolfner 1995; Peng *et al* 2005; Ravi Ram, Sirot and Wolfner 2006; Pilpel *et al* 2008). Population-genetic and divergence-based analyses, furthermore, reveal a high frequency of adaptive evolution amongst both male and female reproductive tract proteases and protease homologs, suggesting an exciting role for this class of enzymes in intersexual

coevolution (Swanson *et al* 2004; Panhuis and Swanson 2006; Haerty *et al* 2007; Lawniczack and Begun 2007; Findlay *et al* 2008; Wong *et al* 2008; Prokupek *et al* 2008).

A recent EST screen of the *D. arizonae* lower female reproductive tract (LFRT: uterus, spermathecae, seminal receptacle, parovaria, common oviduct) identified five lineage-specific protease gene families in which two or more paralogs are expressed in the LFRT (Kelleher, Swanson and Markow 2007). Recurrent duplication of independent loci with similar biochemical functions, in conjunction with evidence of positive selection in three of these gene families, points to an adaptive expansion of proteolytic capacity in the *D. arizonae* lineage (Kelleher, Swanson and Markow 2007). It also may suggest intense sexual conflict, as mathematical models have shown that rapid diversification is an important female "strategy" in sexually antagonistic coevolution (Gavrilets and Waxman 2002; Hayashi, Vose and Gavrilets 2007).

D. arizonae females exhibit two specialized physiological processes that could necessitate enhanced proteolytic capacity in the LFRT. First, *D. arizonae* incorporate significant quantities of male-derived protein into somatic tissues and oocytes (Markow and Ankney 1988; Pitnick, Spicer and Markow 1997). Proteases could play a critical role in this process by degrading sperm and/or seminal proteins into smaller peptides that are more easily absorbed. Second, *D. arizonae* females form an insemination reaction, an opaque white mass of unknown biochemical composition, after every copulation (Patterson 1946). Females must degrade this mass in order to oviposit or remate (Knowles and Markow 2001), a process which could involve proteolysis.

In this study, we compare the evolutionary history, biochemical nature, and physiological significance of secreted female reproductive serine endoproteases (SFRSEs) between *D. arizonae* and its congener *D. melanogaster*. *D. melanogaster* exhibits neither ejaculate incorporation nor an insemination reaction (Markow and

Ankney 1984, 1988; Pitnick, Spicer and Markow 1997), making it ideal for interspecific comparison with *D. arizonae*. First, we explicitly test the hypothesis that secreted proteases expressed in *D. arizonae* LFRTs have experienced a high frequency of recent gene duplication when compared to *D. melanogaster*. We show that *D. arizonae* LFRTs are significantly enriched for recently-duplicated secreted proteases, particularly serine endoproteases. Serine endoproteases comprise an enzymatic class that is particularly well studied in terms of catalytic function (Reviewed in Polgar 1995), key residues that determine substrate specificity (Perona and Craik 1995), and availability of synthetic substrates and inhibitors for biochemical assays. We therefore explore differences in serine endoprotease complement between *D. arizonae* and *D. melanogaster* LFRTs using both bioinformatic approaches and *in vitro* assays. *D. arizonae* female reproductive tracts are shown to encode a greater number of enzymes in a broader range of specificities relative to *D. melanogaster*, as well as enhanced proteolytic activity that is regulated by mating. We discuss our results in terms of differences in reproductive biology between *D. arizonae* and *D. melanogaster*.

MATERIALS AND METHODS

Gene duplication analyses. Protein sequences from candidate LFRT proteins for *D. melanogaster* (150 annotated candidates, Swanson *et al* 2004), and *D. mojavensis* (234 annotated candidates, Kelleher, Swanson and Markow 2007) were obtained from flybase (http://www.flybase.org). It was necessary to use *D. mojavensis*, the closely related sister species of *D. arizonae* (MRCA = ~1.5 MYA, Matzkin *et al* 2004), for this analysis, as no fully sequenced genome is available for *D. arizonae*. Swanson *et al* (2004), and Kelleher, Swanson and Markow (2007) used almost identical experimental approaches for

identifying candidate LFRT proteins, and therefore present comparable datasets between *D. arizonae* and *D. melanogaster*.

Drosophila melanogaster serine endoproteases and serine endoprotease homologs (204 proteases and protease homologs, Ross *et al* 2003) were obtained from flybase (http://www.flybase.org). It was necessary to identify candidate serine endoproteases in the *D. mojavensis* genome *de novo*, using the same approach as Ross *et al* (2003). Briefly, *Manduca sexta* PAP (Jiang *et al* 1998) was used to query the GLEANR protein annotations of *D. mojavensis* (http://rana.lbl.gov/drosophila/) using PSI-BLAST (e-value=1, Altschul *et al* 1997). Every 20th sequence was retained for a second iteration of PSI-BLAST. Conserved serine endoprotease domains were confirmed with hmmpfam (Eddy 1998). The complete list of 167 candidate *D. mojavensis* identified in this study is presented in supplementary table 1.

To examine the frequency of recent duplicates amongst both candidate LFRT proteins, and candidate serine endorptoeases, additional paralogs were identified in the genomes of *D. mojavensis* and *D. melanogaster* using blastP (e = .001, Altschul 1995). For each protein and blast hit pair, coding sequences were aligned in ClustalW (Thompson 1994), and % protein identity and corrected synonymous divergence (d_s) were calculated in PAML (Yang, 1997). Recent duplicates were defined as proteins with greater than 50% identity, where $d_s < 0.5$, and are presented in Supplementary Table 2 (LFRT proteins), and Supplementary Table 3 (candidate proteases).

Functional Enrichment. Significantly over-represented gene ontology terms (GO terms, Ashburner *et al* 2000) in recently duplicated *D. arizonae/mojavensis* LFRT proteins were identified in Fatigo (Al-Shahrour *et al* 2004). GO annotations for the *D. melanogaster*

homolog of each LFRT protein was used, as there is no existing GO annotation dataset for *D. mojavensis*. Over-represented GO terms were identified with Fisher's Exact Test, after correcting for multiple measure based on the false discovery rate (Benjamini and Hochberg 1995).

SFRSE annotation. We searched data sets from previous expression studies of *D. melanogaster* (Swanson *et al* 2004; Mack *et al* 2006; Lawniczak and Begun 2007), and *D. arizonae* (Kelleher, Swanson and Markow 2007) LFRTs to identify SFRSEs in both these species (Table 2). Conservation of the catalytic triad, necessary for proteolytic function (Polgar, 2005), was verified in *D. arizonae* ESTs where possible, or in the ortholog of its sister species, *D. mojavensis* (http://rana.lbl.gov/drosophila/) when the relevant sequence was not present in the EST. Secondary domains in these proteases were identified previously (Kelleher, Swanson and Markow 2007), and CLIP domains were identified by eye as in Jiang and Kanost (2000). *D. arizonae* female reproductive tract protease ESTs were translated and aligned to porcine elastase to identify key substrate specificity residues, as in Perona and Craik (1995). Catalytic function, secondary domains, and substrate specificity for *D. melanogaster* female reproductive tract proteases were adapted from Ross *et al* (2003).

Stocks and Fly Husbandry. The *D. melanogaster* Oregon-R strain was obtained from T.A. Hartl at the University of Arizona, and reared on standard cornmeal

media. The *D. arizonae* strain was collected in Tucson, AZ in 12/2005 by E.S.K., and reared on opuntia banana media (http://stockcenter.arl.arizona.edu/).

Tissue Harvesting. For assays of proteolytic activity in *D. arizonae* and *D. melanogaster* LFRTs, and *D. arizonae* male seminal vesicles and accessory glands (SVAG), tissue was harvested from adults reared in population bottles in order to achieve the maximum diversity of mating states. LFRTs were removed from *D. melanogaster*, $\geq$ 1 day post-eclosion, while LFRTs and SVAGs were removed from *D. arizonae* $\geq$ 9 days post eclosion to ensure reproductive maturity (Reviewed in Markow 1996).

For comparisons of proteolytic activity between virgin and mated *D. arizonae* LFRTs, virgin males and females were isolated within 24 hours of eclosion, and aged separately for 9-12 days. For each cohort of females, 50% were mated at densities of approximately 10 females and 20 males per vial, while the remaining 50% were retained as virgins. After 2 hours of unrestricted mating, the females were separated and their LFRTs removed within 2 hours. We did not verify that all females had mated, however, most dissected females exhibited an insemination reaction indicative of recent copulation (Patterson, 1946). Virgin females were dissected concurrently to minimize differences between the two treatments.

All dissections were performed in 1X Phosphate Buffer Solution (PBS) on a glass slide. Tissue was harvested directly into trypsin assay buffer on ice (50 mM Tris, 10 mM $CaCl_2$, pH 7), and stored at -20ºC. Dissections were performed with care

to prevent contamination from closely-associated gut tissue (see supplementary figure 1).

Colorimetric Assays of Proteolytic Activity in *D. arizonae* and *D. melanogaster* Female Reproductive Tissues. Chromogenic *p*-Nitroanilide substrate for trypsin, Bz-DL-Arg-*p*NA · HCl (DL-BApNA, Sigma), was prepared as a 100 mM stock solution in Dimethyl Sulfoxide (DMSO). Colorimetric *p*-Nitroanilide substrate for elastase, Boc-Ala-Ala-Pro-Ala-*p*NA (BAAPApNA, Calbiochem), was prepared as a 2 mM stock solution in trypsin assay buffer. Diisoflourphosphate (DFP, Calbiochem) serine protease inhibitor was prepared as a 1 M stock solution in isopropyl alcohol. 4-(2-Aminoethyl) benzenesulfonyl fluoride hydrochloride (AEBSF, Sigma-Aldrich) serine protease inhibitor was prepared as a 1 M stock solution in deionized water.

For both species, 9 replicates of 100 individually dissected LFRTs were centrifuged at 1000 x g for 3 minutes, to release only the soluble fraction. The supernatant of all 9 replicates was pooled, and then split into 9 replicate aliquots. These aliquots formed three technical replicates of three treatments: 1) chromogenic substrate at final concentration 3.3 mM (trypsin) or 1 mM (elastase) 2) 60 second preincubation with AEBSF at final concentration 6.66 mM, followed by addition of the chromogenic substrate at final concentration 3.3 mM (trypsin) or 1 mM (elastase) 3) 60 second preincubation with DFP at final concentration 6.66 mM, followed by addition of the chromogenic substrate at final concentration 3.3 mM (trypsin) or 1 mM (elastase).

Trypsin assays were allowed to incubate for 20 minutes at room temperature, while elastase assays were allowed to incubate for 10 minutes at room temperature. For all experiments, activity was measured as an increase in absorbance at 405 nm, as detected by a Cary 50 Bio UV spectrophotometer (Varian, Palo Alto, CA), compared to a standard control of 3.3 mM trypsin substrate or 1 mM elastase substrate in assay buffer.

Colorimetric Assays of Proteolytic Activity in *D. arizonae* male reproductive tissues. Reagents, protein isolation, and reaction conditions were as in assays of LFRTs (above). Supernatant from 10 replicates of 100 individually dissected SVAGs was pooled and split into 10 replicate aliquots. These 10 aliquots formed three technical replicates of three different treatments (as above), plus a control containing only reproductive tract protein in assay buffer. This control was necessary, as *D. arizonae* testes are pigmented. Activity of all nine assays was measured as an increase in absorbance at 405 nm above this control.

Colorimetric Assays of Proteolytic Activity Virgin vs Mated *D. arizonae* Lower Female Reproductive Tracts. Stock solutions, reaction conditions, and activity measurements were as in other assays (above), however, both the DL-BApNA (ICN Biomedicals) and the BAAPApNA (Bachem) were ordered from a different supplier. Supernatant from four biological replicates of 100 virgin LFRTs and 100 mated LFRTs were compared for trypsin and elastase-like activity.

Evolutionary Analyses. Maximum-likelihood estimates of pairwise d_N/d_S between *D. melanogaster* and *D. simulans* coding sequences, and between *D. arizonae* ESTs and *D. mojavensis* coding sequences were generated in PAML (Yang 1997). Although the divergence times between *D. melanogaster* and *D. simulans* (~3 MYA, Hey and Kliman 1993) and *D. arizonae* and *D. mojavensis* (~1.5 MYA, Matzkin 2004) are slightly different, this should not affect our estimate of d_N/d_S, as the difference in divergence time will effect both site classes equally.

RESULTS

***D. arizonae* SFRSEs are enriched for recently duplicated serine endoproteases.** To explicitly test the hypothesis that the *D. arizonae**D. mojavensis* lineage has experienced exceptional duplication of SFRSEs, we first compared the frequency of recent duplicates between *D. arizonae*/*mojavensis* and *D. melanogaster* LFRT proteins. While only 3 (of 150, Swanson *et al* 2004) *D. melanogaster* LFRT proteins have a highly similar paralog (d_s < 0.5) in the *D. melanogaster* genome, a total of 19 *D. arizonae/mojavensis* LFRT proteins (of 234, Kelleher, Swanson and Markow 2007) have a highly similar paralog in the *D. mojavensis* genome (Table 1, Supplementary Table 2). *D. arizonae/mojavensis* LFRT proteins as a whole, therefore, are considerably enriched for recent duplicates relative to *D. melanogaster* (two-tailed Fisher's Exact Test, p = 0.01). We note this is likely a conservative estimate, as six recent duplicates identified in Kelleher, Swanson and

Markow (2007) remain unannotated, and thus were excluded from the comparison. There is no evidence that *D. mojavensis* experiences elevated turnover in gene families with respect to other *Drosophila* species, including *D. melanogaster* (Hahn, Han and Han 2007). It is unlikely, therefore, that the increased frequency of recent duplicates is a genome-wide phenomenon in *D. mojavensis* or *D. arizonae*.

To identify classes of proteins that are prevalent amongst recent duplicates, we tested for over-representation of molecular function gene ontology terms (GO terms, Ashburner *et al* 2000) relative to our complete list of annotated and unannotated *D. arizonae/mojavensis* LFRT proteins (241 total genes, Kelleher, Swanson and Markow 2007). Five interrelated terms were significantly over-represented in recent duplicates after correction for multiple testing: hydrolase activity, peptidase activity, serine-type peptidase activity, endopeptidase activity, and serine-type endopeptidase activity. Recently duplicated *D. arizonae* LFRT proteins, therefore, are significantly enriched for secreted proteases, particularly serine endoproteases. *D. arizonae* LFRT proteins as a whole, moreover, are not enriched in recent duplicates relative to *D. melanogaster* when all proteases are excluded from the data (two-tailed Fisher's Exact Test $p = 0.75$). Thus, the high frequency of recent duplicates observed in *D. arizonae* LFRT protein largely is due to preferential duplication of secreted proteases in this lineage.

The observed preferential duplication could be exclusive to those serine endoproteases that are expressed in lower female reproductive tracts, or could be general to all serine endoproteases in the *D. mojavensis* genome. We therefore

examined whether there was a higher frequency of recent duplicates (d_s < 0.5) amongst *D. mojavensis* serine endoproteases (167 total, supplementary table 1) relative to *D. melanogaster* (204 total, Ross *et al* 2003). *D. mojavensis* serine endoproteases are significantly enriched for recent duplicates (31 of 167) relative to *D. melanogaster* (16 of 204, two-tailed Fisher's Exact Test p = 0.003). This enrichment is not significant however, when LFRT proteins and their close paralogs are excluded from the data set (two-tailed Fisher's Exact Test p = 0.08), suggesting that the enrichment of recent duplicates largely is driven by the preferential duplication of LFRT proteins. Indeed, recently duplicated *D. mojavensis* serine endoproteases are significantly enriched for LFRT proteins and their close paralogs (two-tailed Fisher's Exact Test p= 1.29 x 10^-5).

An elevated frequency of recent duplicates amongst serine endoproteases points to an adaptive expansion of proteolytic capacity in *D. arizonae* LFRTs. As an enzymatic class, serine endoproteases are exceedingly well described in terms of defining how key amino acid residues affect catalytic function (Reviewed in Polgar 2005) and substrate specificity (Perona and Craik 1995). Synthetic substrates and inhibitors for these proteases, furthermore, are readily available. The remainder of this study, therefore, focuses on a comparison of the secreted female reproductive serine endoprotease (SFRSE) complement between *D. arizonae* and *D. melanogaster*.

D. arizonae **LFRTs are enriched for digestive serine endoproteases.**

Comparisons of the nature, number, and specificity of SFRSEs suggest dramatic

enhancement of *D. arizonae* proteolytic capacity relative to *D. melanogaster* (Table 2). Almost twice as many SFRSEs are found in *D. arizonae* LFRTs (15), as in *D. melanogaster* LFRTs (8), despite, multiple examinations of female reproductive tract proteins in the latter species including two high-throughput transcriptional studies (Swanson *et al* 2004; Mack *et al* 2006; Panhuis and Swanson 2006; Lawniczack and Begun 2007). All but two of these *D. arizonae* SFRSEs , furthermore, lack secondary protein-protein interaction domains (Table 2). The presence of such domains is significant, as they are common to insect serine endoproteases involved in physiological responses and developmental cascades, and generally are absent in proteases whose primary function is nutritional digestion (Ross *et al* 2003).

Serine endoproteases make effective digestive enzymes because they exhibit no absolute specificity in terms of recognizing the three dimensional structure of their substrate. Rather, these enzymes show preferences for cleaving the scissile bond of a specific amino acid or set of amino acids, as determined by three key residues in the substrate-binding pocket (Perona and Craik, 1995). Examination of these residues in *D. arizonae* SFRSEs suggests a broad range of specificities including all three major classes of digestive enzymes, trypsin, chymotrypsin, and elastase, as well as several proteases with unpredictable specificity. *D. melanogaster* SFRSEs, by comparison, present no evidence for chymotrypsin or elastase-like activity, suggesting a narrower range of putative substrates.

***D. arizonae* LFRTs exhibit significant trypsin-like and elastase-like serine endoprotease activity.** Our evolutionary and bioinformatic analyses suggest that recent gene duplication has enriched *D. arizonae* LFRTs for digestive serine endoproteases with a broad range of specificities including trypsin, chymotrypsin and elastase (Table 1). To test this hypothesis directly, we used chromogenic *p*-Nitroanilide substrates to detect proteolytic activity in isolated LFRT lumens. While chymotrypsin activity was not detected in *D. arizonae* LFRTs (data not shown), significant levels of trypsin and elastase-like activity were exhibited by lumen isolated from these tissues (Figure 1). This activity decreased when isolated lumen was pre-incubated with the serine endoprotease inhibitors AEBSF (trypsin; $F_{1,6}$ = 102.57, p = 5.29 x 10^{-5}, elastase: $F_{1,6}$ = 41.04, p = 6.82 x 10^{-4}) and DFP (trypsin; $F_{1,6}$ = 184.64, p = 9.86 x 10^{-6}, elastase: $F_{1,6}$ = 4140.83, p = 9.47 x 10^{-10}), as expected if trypsin and elastase-like activities are due to serine endoproteases (Figure 1).

To determine if trypsin and elastase-like serine endoproteases could be derived from males during mating, we assayed *D. arizonae* seminal vesicles and accessory glands (SVAGs) for serine endoprotease activity. While the spectrophotometer detects absorbance at 405 nm, this value was not significantly different in assays pre-incubated with serine endoprotease inhibitors. Because these assays were not controlled for the inherent yellow pigment of *p*-Nitroanilide stock solution (see materials and methods) we end this represents background absorbance from the chromogenic substrate rather than enzyme activity. These absorbance values, furthermore, are similar to values seen in blank solution

containing only assay buffer and chromogenic substrate (not shown). Although male-derived proteases could become activated only inside females (Ravi Ram, Sirot and Wolfner 2006), our data provide no evidence that trypsin or elastase-like activity in *D. arizonae* female reproductive tracts originates in the male ejaculate.

D. melanogaster LFRTs exhibit fewer serine endoproteases than *D. arizonae*, and no predicted elastase-like serine endoproteases (Table 1). Consistent with this observation, our enzyme assays detect minimal trypsin or elastase-like activity in isolated LFRT lumen (Figure 1). Enzyme activity, furthermore, was not significantly reduced upon pre-incubation with serine endoprotease inhibitors (Figure 1), providing no evidence for serine endoprotease activity. While it remains possible that the relative magnitude of detected activity would differ under other assay conditions, these data suggest that proteolytic capacity may present a significant physiological difference between *D. arizonae* and *D. melanogaster*.

Serine endoprotease activity in *D. arizonae* female reproductive tracts is negatively regulated by mating. To further elucidate the interaction between female proteases and the male ejaculate, we measured differences in trypsin and elastase-like activity in matched cohorts of virgin and recently mated (<4 hours post-copulation) *D. arizonae* females. Virgin females exhibit significant trypsin and elastase-like acitivity, suggesting that the proteolytic activity detected here does not primarily originate in the male ejaculate. Both trypsin and elastase-like activity, furthermore, were significantly reduced in mated female LFRT lumens when

compared to virgins (trypsin; $F_{1,6}$ = 100.18, p = 5.76 x 10^{-5}, elastase: $F_{1,6}$ = 8.44, p = 0.027, Figure 2), the opposite relationship of what would be expected if proteolytic acitivty was derived from males.

Reduced proteolytic activity in mated females when compared to virgins suggests that, SFRSEs are negatively regulated by the male ejaculate. While it is possible that reduced activity could reflect competition between male-derived substrates and synthetic substrates for access to proteases, the magnitude of the observed decrease, particularly for trypsin-like enzymes makes this explanation unlikely. Synthetic substrates are expected to be in considerable molar excess to proteases and endogenous substrates, minimizing the effect of dilution by endogenous molecules.

Some *D. melanogaster* and *D. arizonae* SFRSEs evolve rapidly. Evolutionary rates of SFRSEs could serve as a metric to detect important differences in SFRSE dynamics between *D. arizonae* and *D. melanogaster*. We therefore estimated the ratio of replacement to silent substitutions (d_N/d_S) in both *D. arizonae* and *D. melanogaster* SFRSEs by comparing to their ortholog in the *D. simulans* and *D. mojavensis* genomes, respectively (table 3). Modest discrepancies between our results and previously reported values (Swanson *et al* 2004), likely arise from the use of a *D. simulans* EST, rather than the full length coding sequence, in the previous study. We find no evidence for a difference in d_N/d_S between *D. melanogaster* and *D. arizonae* SFRSEs ($F_{1,22}$ = .13, p = .72), suggesting similar selective regimes in both

lineages. We furthermore note that both data sets exhibit a high average d_N/d_S (*D. melanogaster* = .43, *D. arizonae* = .48), and several proteases with $d_N/d_S > 0.5$, suggestive of adaptive evolution (Swanson *et al* 2004). Indeed, several of these proteins have been shown to experience positive selection in previous studies (Panhuis and Swanson 2006; Lawniczak and Begun 2007; Kelleher, Swanson and Markow 2007; Kelleher and Markow 2009).

DISCUSSION

Our previous observation of lineage-specific gene families of secreted proteases in *D. arizonae* LFRT proteins suggested a recent, adaptive expansion of female reproductive proteolytic capacity (Kelleher, Swanson and Markow 2007; Kelleher and Markow 2009). The data presented here indicate that *D. arizonae* LFRT proteins are enriched for recent duplicates relative to its congener *D. melanogaster*, and that this enrichment reflects preferential duplication of secreted proteases, particularly serine endoproteases. We furthermore show that *D. arizonae* female reproductive tracts exhibit a larger more diverse complement of serine endoproteases in their LFRTs, as well as considerable trypsin and elastase like serine endoprotease activity that is regulated by mating. Collectively, our data suggest that SFRSEs exhibit divergent evolutionary dynamics and physiological functions between these two lineages.

D. arizonae LFRT proteins are enriched for recently-duplicated serine endoproteases when compared to those of *D. melanogaster*. This pattern reflects preferential duplication of serine enroproteases expressed in the LFRT, rather than an elevated duplication rate in this enzymatic class as a whole. Intriguingly, male seminal proteins in the *repleta* species group also exhibit a high frequency of recent duplicates, although these paralogs are not clearly biased towards a particular functional class (Wagstaff and Begun 2007; Almeida and DeSalle 2008 A; 2008B). Accelerated gene duplication rates, therefore, may be an important aspect of reproductive protein evolution within the *repleta* species group.

Although the selective force that underlies the exceptional frequency of gene duplications amongst *repleta* species group reproductive proteins remains unclear, it is interesting to speculate that this pattern may arise from sexual conflict. Mathematical models of sexually antagonistic coevolution between interacting male and female molecules have predicted it is adaptive for females to diversify in the face of pursuit by a male locus, and that male proteins may in turn diversify in response to females (Gavrilets and Waxman 2002; Hayashi, Vose and Gavrilets 2007). Although these models predict the rise of two divergent alleles at a single locus (Gavrilets and Waxman 2002; Hayashi, Vose and Gavrilets 2007), duplication and diversification of such loci would produce the same ultimate result. Intriguingly, *D. arizonae* females are three to five times more promiscuous than *D. melanogaster* (Reviewed Markow 1996), indicating this lineage will experience comparatively more intense sexual conflict (Parker, 1979).

The adaptive significance of preferential duplication of SFRSEs in the *D. arizonae*/*D. mojavensis* lineage is yet unclear. The bioinformatics analysis presented in this study, however, indicates that *D. arizonae* presents a larger number of SFRSEs in a broader range of predicted specificities than *D. melanogaster*. The majority of these proteins lack secondary protein-protein interaction domains, furthermore, suggesting their primary function is digestive (Ross *et al* 2003). Consistent with this hypothesis, isolated lumen from *D. arizonae* LFRTs exhibits considerable trypsin and elastase-like serine endoprotease activity reminiscent of gastrointestinal tracts (Billingsley and Hecker 1991; Oppert, Hartzer and Zuercher 2002; Zhu, Zeng and Oppert 2003), while no such activity is detected in *D. melanogaster*. The intense proteolytic environment presented by the *D. arizonae* female reproductive tract, therefore, may represent an important physiological difference from *D. melanogaster*.

Mated *D. arizonae* LFRTs exhibited significantly lower enzyme activity than virgin LFRTs, particularly for trypsin-like enzymes. This result appears counterintuitive; if female proteases cleave or degrade substrates in the male ejaculate, mating is predicted to be a positive regulator of proteolytic activity. If it is adaptive for males to avoid degradation of ejaculatory components due to sexual conflict, however, they may seek to negatively regulate female proteases. Mechanistically, this could be accomplished at either the transcriptional level, or through protease inhibitors in the male ejaculate (Wagstaff and Begun 2005; Kelleher *et al* 2009).

We previously have hypothesized that duplicated digestive proteases in *D. arizonae* LFRTs may be required to facilitate incorporation of ejaculate-derived protein,

degradation of the insemination reaction, or both, in mated *D. arizonae* females (Kelleher, Swanson and Markow 2007). Adaptive male avoidance of female proteases is easy to envision in the context of this specialized reproductive physiology. If females are digesting important seminal proteins or sperm for their own nutritional purposes, this could be extremely costly to males. Alternatively, males may want to encumber female degradation of the ejaculate-induced insemination reaction. Indeed, the reaction mass is thought to be a male "strategy" to delay female remating and ensure paternity (Markow and Ankney 1984, 1988; Pitnick, Spicer, and Markow 1997), and male-female conflict over the size and duration of the insemination reaction previously has been proposed (Knowles and Markow 2001).

Acknolwedgements. The authors would like to acknowledge Roger Miesfeld for generous use of equipment and reagents, and Therese Markow, Willie Swanson, Jeremy Bono, Vanessa Corby-Harris and three anonymous reviewers for helpful comments that significantly improved the manuscript. This research was funded by a Doctoral Dissertation Improvement Grant to E.S.K., and N.I.H. grant AI31951 to R.L.M. E.S.K. was supported by an NSF-IGERT research traineeship in Evolutionary, Functional and Computational Genomics at the University of Arizona, and a Dissertation Fellowship from the American Association of University Women.

REFERENCES

Al-Shahrour F, Díaz-Uriarte R, Dopazo J. 2004. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics. 20:578–80.

Al-Shahrour F, Minguez P, Tárraga J, Medina I, Alloza E, Montaner D, Dopazo J. 2007. FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. Nucleic Acids Res. 35:W91–6.

Almeida FC, DeSalle R. 2008a. Evidence of adaptive evolution of accessory gland proteins in closely related species of the *Drosophila repleta* group. Mol. Biol. Evol. 25:2043–2053.

Almeida FC, DeSalle R. 2008b. Orthology, Function, and Evolution of Accessory Gland Proteins in the *Drosophila repleta* Group. Genetics *epub ahead of print.*

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Ashburner M, Ball CA, Blake JA, Botstein D, et al. (20 co-authors). 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25:25–29.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B. 57:289–300.

Billingsley PF, Hecker H. 1991. Blood digestion in the mosquito, *Anopheles stephensi* Liston (Diptera: Culicidae): activity and distribution of trypsin, aminopeptidase, and alpha-glucosidase in the midgut. J. Med. Entomol. 28:865–871.

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

Brennan PL, Prum RO, McCracken KG, Sorenson MD, Wilson RE, Birkhead TR. 2007. Coevolution of male and female genital morphology in waterfowl. PLoS ONE. 2:e418.

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

Dacheux JL, Gatti JL, Dacheux F. 2003. Contribution of epididymal secretory proteins for spermatozoa maturation. Microsc. Res. Tech. 61:7–17.

Eddy SR. 1998. Profile hidden Markov models. Bioinformatics 14:755– 763.

Gavrilets S. 2000. Rapid evolution of reproductive barriers driven by sexual conflict. Nature 403:886–889.

Gavrilets S, Waxman D. 2002. Sympatric speciation by sexual conflict. Proc. Natl. Acad. Sci. U. S. A. 99:10533–10538.

Graf L, Hegyi G, Liko I, Hepp J, Medzihradszky K, Craik CS, and Rutter WJ (1988) Structural and functional integrity of specificity and catalytic sites of trypsin *Intl.* J. Peptide and Protein Research 32:512–518.

Hayashi TI, Vose M, Gavrilets S. 2007. Genetic differentiation by sexual conflict. Evolution 61:516–29.

Haerty W, Jagadeeshan S, Kulathinal RJ, et al. (11 co-authors). 2007. Evolution in the fast lane: rapidly evolving sex-related genes in *Drosophila*. 177:1321–1335.

Hahn MW, Han MV, Han SG. 2007. Gene family evolution across 12 *Drosophila* genomes. P.L.o.S. Genet. 3(11):e197.

Hey J, Kliman RM (1993) Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* species complex. Mol Biol Evol 10: 04–822.

Jiang H, Wang Y, Kanost MR. 1998. Pro-phenol oxidase activating proteinase from an insect, *Manduca sexta*: a bacteria-inducible protein similar to *Drosophila* easter. Proc. Natl. Acad. Sci. USA 95:12220–12225.

Jiang H, Kanost MR. 2000. The clip-domain family of serine proteinases in arthropods. Insect Biochem. Mol. Biol. 30:95–105.

Kelleher ES, Swanson WJ, Markow TA. 2007. Gene duplication and adaptive evolution of digestive proteases in *Drosophila arizonae* female reproductive tracts. P.L.o.S. Genet. 3:1541-1549.

Kelleher ES, Markow TA. 2009. Duplication, Selection, and Gene Conversion in a *Drosophila mojavensis* Female Reproductive Protein Family. *Genetics. online early.*

Kelleher ES, Watts TD, LaFlamme BA, Haynes PD, Markow TA. 2009. Proteomic analysis of *Drosophila mojavensis* male accessory glands suggests novel classes of seminal fluid proteins. *Journal of Insect Biochemistry and Molecular Biology, in press.*

Knowles LL, Markow TA. 2001. Sexually antagonistic coevolution of a postmating prezygotic reproductive character in desert *Drosophila*. Proc. Nat. Acad. Sci. U. S. A. 98:8692–8696.

Lawniczak MK, Begun DJ. 2007. Molecular population genetics of female-expressed mating-induced serine proteases in *Drosophila melanogaster*. Mol. Biol. Evol. 24:1944–1951.

Mack PD, Kapelnikov A, Heifetz Y, Bender M. 2006. Mating-responsive genes in reproductive tissues of female *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U. S. A. 103:10358–63.

Markow TA, Ankney PF. 1984. *Drosophila* Males Contribute to Oogenesis in a Multiple Mating Species. Science 224:302–303.

Markow TA and Ankney PF. 1988. Insemination Reaction in *Drosophila* found in species whose males contribute material to oocytes before fertilization. Evolution 42:1097–1101.

Markow TA. 1996. Evolution of *Drosophila* mating systems. Evol. Biol. 29:73–106.

Marshall JL. 2007. Rapid evolution of spermathecal duct length in the *Allonemobius socius* complex of crickets: species, population and *Wolbachia* effects. P.L.o.S. ONE. 2(1):e720.

Matzkin LM. 2004. Population genetics and geographic variation of alcohol dehydrogenase (Adh) paralogs and glucose-6-phosphate dehydrogenase (G6pd) in *Drosophila mojavensis*. Mol. Biol. Evol. 21:276–285.

Monsma SA, Harada HA, Wolfner MF. 1990. Synthesis of two *Drosophila* male accessory gland proteins and their fate after transfer to the female during mating. Dev. Biol. 142:465–475.

Neubaum DM, Wolfner MF. 1999. Wise, winsome, or weird? Mechanisms of sperm storage in female animals. Curr. Top. Dev. Biol. 41:67–97.

Neurath H, 1984. Evolution of proteolytic enzymes. Scienc*e* 224:350–357.

Oppert B, Hartzer K, Zuercher M. 2002. Digestive proteinases in *Lasioderma serricorne* (Coleoptera: Anobiidae). Bull. Entomol. Res. 92:331–336.

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

Panhuis TM, Clark NL, Swanson WJ. 2006. Rapid evolution of reproductive proteins in abalone and *Drosophila*. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 361:261–8.

Park M, Wolfner MF. 1995. Male and female cooperate in the prohormone-like processing of a *Drosophila melanogaster* seminal fluid protein. Dev. Biol. 171:694–702.

Parker, GA. 1979 Sexual selection and sexual conflict. In Blum MS, Blum NA, editors. Sexual selection and reproductive competition in insects. London: Academic Press. pp. 123–166.

Patterson JT. 1946. A new type of isolating mechanism *in Drosophila*. Proc. Nat. Acad. Sci. U. S. A. 32:202–208.

Peng J, Chen S, Busser S, Liu H, Honegger T, Kubli E. 2005. Gradual release of sperm bound sex-peptide controls female postmating behavior in *Drosophila*. Curr Biol 15:207–213.

Perona, JJ, Craik CS. 1995. Structural basis of substrate specificity in the serine proteases. Protein Sci. 4:337–360.

Pilpel N, Nezer I, Applebaum SW, Heifetz Y. 2008. Mating-increases trypsin in female *Drosophila* hemolymph. Insect. Biochem. Mol. Biol. 38:320–30.

Pitnick S, Spicer GS and Markow TA. 1997. Phylogenetic examination of female incorporation of ejaculate in *Drosophila*. Evolution 51:833–845.

Pitnick S, Markow TA, Spicer GS. 1999. Evolution of multiple kinds of female sperm-storage organs in *Drosophila*. Evolution 53:18041-1822.

Polgar, L. 2005. The catalytic triad of serine peptidases. Cell. Mol. Life. Sci. 62:2161–2172.

Powell JR. 1997. Progress and prospects in evolutionary biology: The *Drosophila* model. New York: Oxford University Press.

Prokupek A, Hoffmann F, Eyun SI, Moriyama E, Zhou M, Harshman L. 2008. An evolutionary expressed sequence tag analysis of *Drosophila* spermatheca genes. Evolution 62:2936–2647.

Ravi Ram K, Sirot LK, Wolfner MF. 2006. Predicted seminal astacin-like protease is required for processing of reproductive proteins in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U. S. A. 103:18674–18679.

Rice, WR. 1996. Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. Nature 381:232–234.

Robertson SA. 2007. Seminal fluid signaling in the female reproductive tract: lessons from rodents and pigs. J. Anim, Sci. 2007 85:E36–44.

Ross J, Jiang H, Kanost MR, Wang Y. 2003. Serine proteases and their homologs in the *Drosophila melanogaster* genome: An initial analysis of sequence conservation and phylogenetic relationships. Gene 304:117–131.

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

Sirot LK, Poulson RL, McKenna MC, Girnary H, Wolfner MF, Harrington LC. 2008. Identity and transfer of male reproductive gland proteins of the dengue vector mosquito, *Aedes aegypti*: potential tools for control of female feeding and reproduction. Insect Biochem. Mol. Biol. 38:176–89.

Spofford, JB, 1969. Heterosis and evolution of duplications. Am. Nat. 103:407–432.

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

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

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

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673–4680.

Wagstaff BJ, Begun DJ. 2005. Molecular population genetics of accessory gland protein genes and testis-expressed genes in *Drosophila mojavensis* and *D. arizonae.* Genetics 171:1083–1010.

Wagstaff BJ. Begun DJ. 2007. Adaptive evolution of recently duplicated accessory gland protein genes in desert *Drosophila.* Genetics 177:1023–1030.

Wolfner MF. 2007. "S.P.E.R.M." (seminal proteins (are) essential reproductive modulators): the view from *Drosophila*. Soc. Reprod. Fertil. Suppl. 183–99.

Wong A, Turchin MC, Wolfner MF, Aquadro CF. 2008. Evidence for positive selection on *Drosophila melanogaster* seminal fluid protease homologs. Mol. Biol. Evol. 25:497–506.

Yang Z .(997. PAML: A program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555–556.

Zhu YC, Zeng F, Oppert B. 2003. Molecular cloning of trypsin-like cDNAs and comparison of proteinase activities in the salivary glands and gut of the tarnished plant bug *Lygus lineolaris* (Heteroptera: Miridae*).* Insect Biochem. Mol. Biol. 33:889–899.

TABLES

Candidate Female Reproductive Tract Protein	**Functional Class**
D. melanogaster	
IM10-PA	defense response
CG30035-PB	carbohydrate transport
scpr-C-PA	CRISP
D. mojavensis	
Dmoj\GLEANR_12010	serine endoprotease
Dmoj\GLEANR_12324	serine protease
Dmoj\GLEANR_12325	serine protease
Dmoj\GLEANR_1234	protease inhibitor
Dmoj\GLEANR_12931	metalloprotease
Dmoj\GLEANR_13880	sulfate transport
Dmoj\GLEANR_2575	serine endoprotease
Dmoj\GLEANR_2703	metalloprotease
Dmoj\GLEANR_3081	unknown function
Dmoj\GLEANR_4546	glycosyl hydrolase
Dmoj\GLEANR_5037	unknown function
Dmoj\GLEANR_6725	unknown function
Dmoj\GLEANR_6984	serine endoprotease
Dmoj\GLEANR_7051	lipase
Dmoj\GLEANR_778	metalloprotease
Dmoj\GLEANR_896	serine endoprotease
Dmoj\GLEANR_897	serine endoprotease
Dmoj\GLEANR_898	serine endoprotease
Dmoj\GLEANR_9617	serine protease

Table 1. Recent Duplicates in *D. melanogaster* and *D. mojavensis* LFRT proteins. Annotated candidate LFRT proteins from *D. melanogaster* (Swanson *et al* 2004) and *D. arizonae* (Kelleher, Swanson and Markow 2007) with recent duplicates in the *D. melanogaster* and *D. mojavensis* genomes are identified. Functional class is based on GO terms from flybase (http://flybase.org/), and conserved domains.

CDS	189	216	226	predicted specificity	secondary domain
D. arizonae					
Dari/anon-EST:Kelleher5	Lys	Lys	Thr	elastase?	
Dari/anon-EST:Kelleher6	Thr	Gly	Ala	chymotrypsin	
Dari/anon-EST:Kelleher7	Ser	Gly	Arg	unknown	
Dari/anon-EST:Kelleher8	Ser	Val	Asn	elastase	
Dari/anon-EST:Kelleher10	Thr	Gly	Ala	chymotrypsin	
Dari/anon-EST:Kelleher82	Thr	?	?	unknown	
Dari/anon-EST:Kelleher267	?	?	?	unknown	2 CLIP
Dari/anon-EST:Kelleher318	Asp	Gly	Thr	unknown	
Dari/anon-EST:Kelleher361	Asp	?	?	unknown	
Dari/anon-EST:Kelleher472	Gly	Gly	Gly	unknown	CUB
Dari/anon-EST:Kelleher506	Met	Gly	Asp	elastase?	
Dari/anon-EST:Kelleher580	Lys	?	?	unknown	
Dari/anon-EST:Kelleher594	Asp	Gly	Gly	trypsin	
Dari/anon-EST:Kelleher595	Asp	Gly	Gly	trypsin	
Dari/anon-EST:Kelleher596	Gly	Ala	Ala	unknown	
D. melanogaster					
Dmel/CG3066	Asp	Gly	Gly	trypsin	CLIP
Dmel/Tequila	Asp	Gly	Gly	trypsin	CBM_14\SCSR\Ldl_recept_a
Dmel/CG16705	Asp	Gly	Gly	trypsin	CLIP
Dmel/CG17012	Gly	Thr	Thr	unknown	
Dmel/CG17240	Asp	Gly	Gly	trypsin	
Dmel/CG17239	Asp	Gly	Gly	trypsin	
Dmel/CG17234	Ser	Val	Arg	unknown	
Dmel/CG14642	Ser	Gly	Ser	trypsin	

Table 2. Secreted Female Reproductive Serine Endoproteases in *D. melanogaster* and *D. arizonae*. For each protease, key residues for substate specificity 189, 216, 226, as well as predicted specificity as in Perona and Craik (1995). Secondary protein-protein interaction domains were identified by eye (CLIP domains) or from previous reports

(Ross *et al* 2003; Kelleher, Swanson and Markow 2007). More details on protein domains can be found at (http://pfam.sanger.ac.uk/). ? indicates the relevant site was not included in the EST sequence.

D. arizonae EST	*D. mojavensis* CDS	d_N	d_S	d_N/d_S
Dari\anon-EST:Kelleher5	Dmoj\anon-EST:Kelleher5	0.05	0.04	1.20
Dari\anon-EST:Kelleher5	Dmoj\anon-EST:Kelleher6	0.08	0.17	0.44
Dari\anon-EST:Kelleher8	Dmoj\anon-EST:Kelleher8	0.14	0.31	0.47
Dari\anon-EST:Kelleher7	Dmoj\anon-EST:Kelleher7	0.03	0.07	0.36
Dari\anon-EST:Kelleher10	no ortholog			
Dari\anon-EST:Kelleher82	Dmoj\GLEANR_12010	0.00	0.02	0.13
Dari\anon-EST:Kelleher267	Dmoj\GLEANR_17341	0.01	0.03	0.24
Dari\anon-EST:Kelleher318	Dmoj\GLEANR_2575	0.07	0.14	0.48
Dari\anon-EST:Kelleher361	Dmoj\GLEANR_3606	0.01	0.04	0.32
Dari\anon-EST:Kelleher472	Dmoj\GLEANR_5738	0.01	0.06	0.12
Dari\anon-EST:Kelleher506	Dmoj\GLEANR_6984	0.01	0.03	0.46
Dari\anon-EST:Kelleher580	DmojGLEANR_8733	0.03	0.07	0.39
Dari\anon-EST:Kelleher594	Dmoj\GLEANR_896	0.11	0.12	0.89
Dari\anon-EST:Kelleher596	Dmoj\GLEANR_898	0.05	0.12	0.44
Dari\anon-EST:Kelleher595	Dmoj\GLEANR_897	0.10	0.13	0.83
		mean d_N/d_S= 0.48 $\pm$ .075		

D. melanogaster CDS	*D. simulans* CDS	d_N	d_S	d_N/d_S
Dmel/CG3066	Dsim/GLEANR_3734	0.02	0.12	0.16
Dmel/Tequila	Dsim/GLEANR_14168,14169	0.02	0.14	0.12
Dmel/CG16705	Dsim/GLEANR_4787	0.02	0.18	0.09
Dmel/CG17012	Dsim/GLEANR_6593	0.13	0.13	0.92
Dmel/CG17240	Dsim/GLEANR_6596	0.07	0.12	0.60
Dmel/CG17239	Dsim/GLEANR_6595	0.08	0.11	0.69
Dmel/CG17234	Dsim/GLEANR_6882	0.07	0.10	0.73
Dmel/CG14642	Dsim/GLEANR_3486	0.03	0.14	0.18
		mean d_N/d_S= 0.44 $\pm$ .10		

Table 3. Protein Evolution of Secreted Female Reproductive Serine Endoproteases

Evolutionary rates were calculated between *D. melanogaster* and *D. arizonae* and their orthologs in the *D. simulans* and *D. mojavensis* genomes in PAML (Yang 1997). d_N = non-synonymous substitutions per non-synonymous site, d_S = synonymous substitutions per non-synonymous site, d_N/d_S = ratio non-synonymous substitutions per non-synonymous site to synonymous substitutions per non-synonymous site.

FIGURES

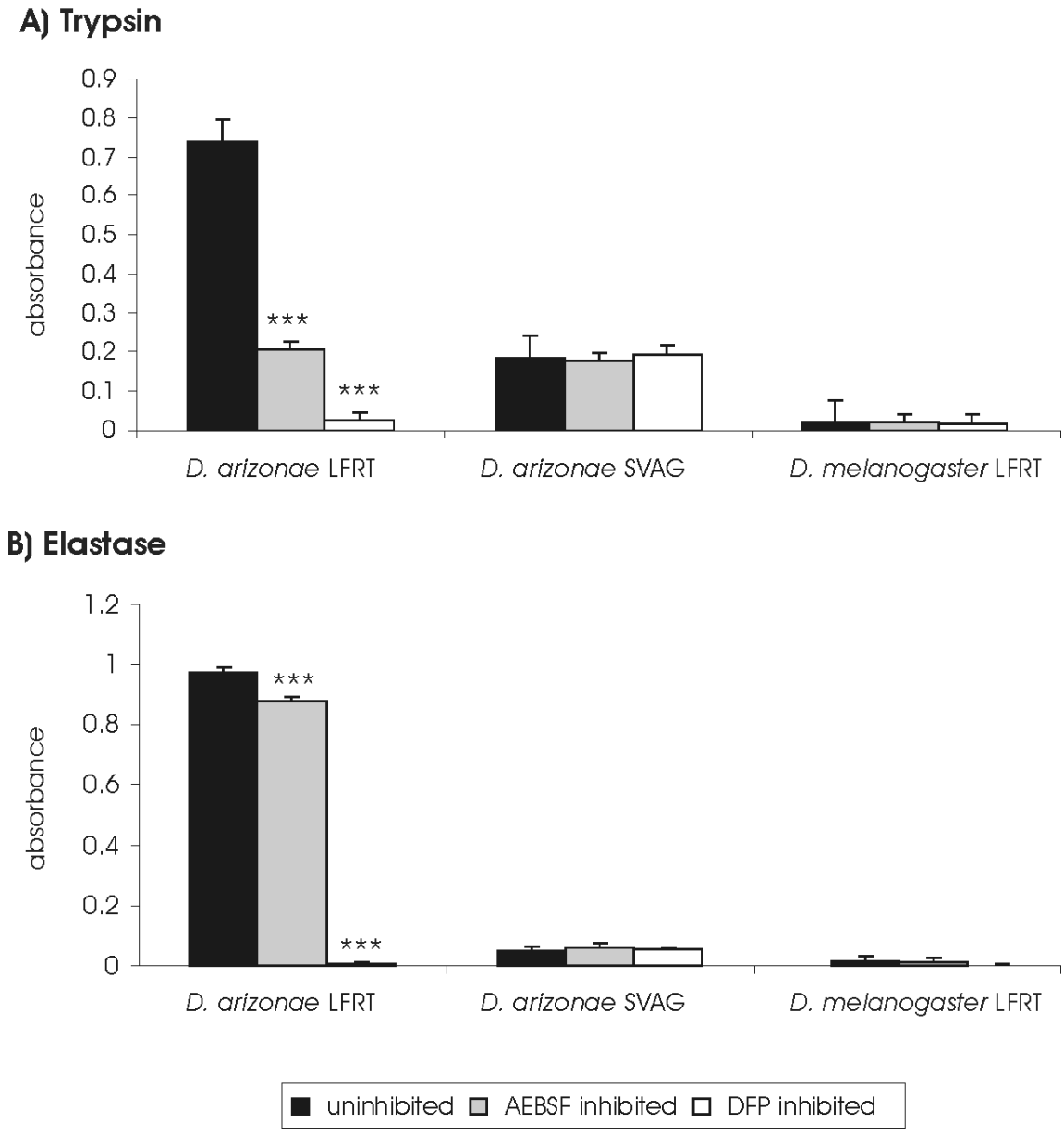


Figure 1. Serine endoprotease activity in the reproductive tissues of *D. arizonae* females and males, and *D. melanogaster* females. Activity is measured as absorbance of the chromogenic A) trypsin and B) elastase substrate at 405nm. Enzyme activity activity is decreased by preincubation with serine endoprotease inhibitors indicating the active protease utilize serine in their active sites.

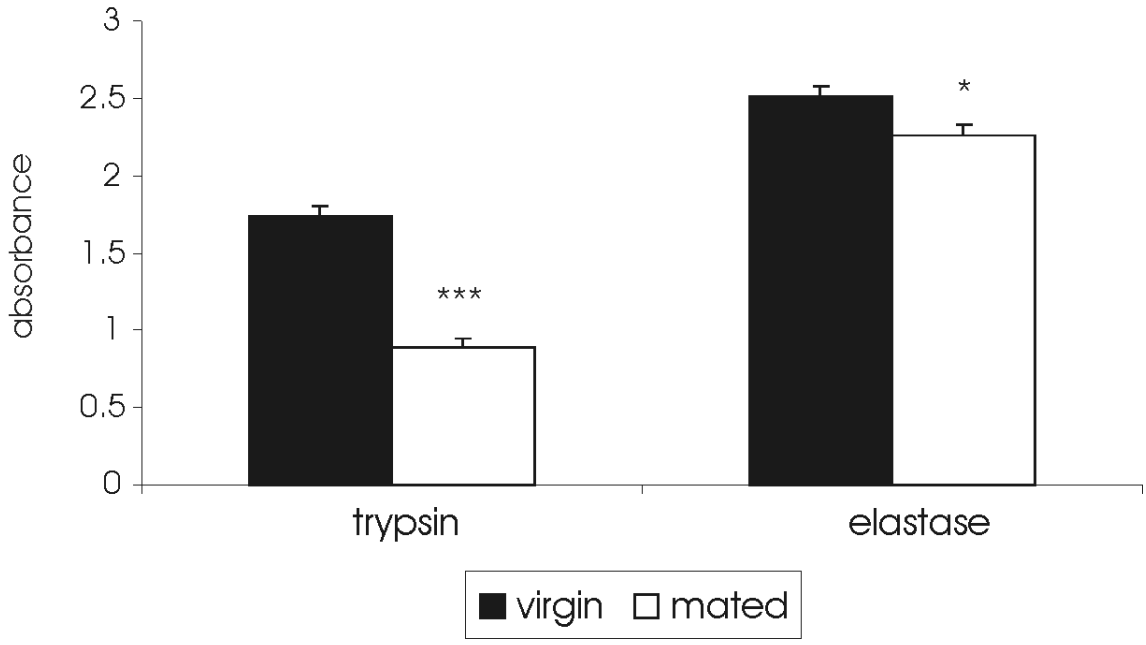


Figure 2. Serine endoprotease activity *D. arizonae* lower reproductive tracts is dependent on female mating status. Activity is absorbance of the chromogenic substrate at 405nm.

SUPPLEMENTARY DATA

***D. mojavensis* GLEANR annotation**

GLEANR_17599_1
GLEANR_13152_1
GLEANR_9833_1
GLEANR_15527_1
GLEANR_10225_1
GLEANR_10447_1
GLEANR_10448_1
GLEANR_10449_1
GLEANR_10578_1
GLEANR_10590_1
GLEANR_1070_1
GLEANR_10747_1
GLEANR_11413_1
GLEANR_11532_1
GLEANR_11868_1
GLEANR_12007_1
GLEANR_12008_1
GLEANR_12009_1
GLEANR_12010_1
GLEANR_12011_1
GLEANR_12012_1
GLEANR_12013_1
GLEANR_12092_1
GLEANR_12345_1
GLEANR_12399_1
GLEANR_12400_1
GLEANR_12691_1
GLEANR_12922_1
GLEANR_13195_1
GLEANR_13196_1
GLEANR_13197_1
GLEANR_13198_1
GLEANR_13279_1
GLEANR_13513_1
GLEANR_13613_1
GLEANR_13708_1
GLEANR_15519_1
GLEANR_15557_1
GLEANR_15984_1
GLEANR_16066_1
GLEANR_16067_1

GLEANR_16068_1
GLEANR_16100_1
GLEANR_16233_1
GLEANR_16317_1
GLEANR_16582_1
GLEANR_16583_1
GLEANR_16584_1
GLEANR_16585_1
GLEANR_16586_1
GLEANR_16735_1
GLEANR_16773_1
GLEANR_16786_1
GLEANR_17203_1
GLEANR_17303_1
GLEANR_17340_1
GLEANR_17341_1
GLEANR_17432_1
GLEANR_17433_1
GLEANR_17434_1
GLEANR_17436_1
GLEANR_17437_1
GLEANR_17438_1
GLEANR_17440_1
GLEANR_17464_1
GLEANR_17466_1
GLEANR_17467_1
GLEANR_17468_1
GLEANR_17578_1
GLEANR_17587_1
GLEANR_17703_1
GLEANR_1954_1
GLEANR_2015_1
GLEANR_2240_1
GLEANR_2241_1
GLEANR_2258_1
GLEANR_2415_1
GLEANR_2574_1
GLEANR_2575_1
GLEANR_2655_1
GLEANR_2656_1
GLEANR_3285_1
GLEANR_3286_1
GLEANR_3287_1
GLEANR_3348_1

GLEANR_3349_1
GLEANR_3350_1
GLEANR_3351_1
GLEANR_3606_1
GLEANR_3624_1
GLEANR_3673_1
GLEANR_3744_1
GLEANR_4035_1
GLEANR_4335_1
GLEANR_4364_1
GLEANR_4365_1
GLEANR_4366_1
GLEANR_4368_1
GLEANR_4369_1
GLEANR_440_1
GLEANR_5188_1
GLEANR_5189_1
GLEANR_5236_1
GLEANR_5260_1
GLEANR_5326_1
GLEANR_5449_1
GLEANR_5602_1
GLEANR_5682_1
GLEANR_5683_1
GLEANR_5738_1
GLEANR_5739_1
GLEANR_5754_1
GLEANR_5949_1
GLEANR_6035_1
GLEANR_6036_1
GLEANR_6037_1
GLEANR_6038_1
GLEANR_6039_1
GLEANR_6040_1
GLEANR_6041_1
GLEANR_6524_1
GLEANR_6984_1
GLEANR_7193_1
GLEANR_7398_1
GLEANR_7399_1
GLEANR_7499_1
GLEANR_7584_1
GLEANR_7639_1
GLEANR_7676_1

GLEANR_7677_1
GLEANR_7679_1
GLEANR_7854_1
GLEANR_7969_1
GLEANR_800_1
GLEANR_8231_1
GLEANR_8253_1
GLEANR_8299_1
GLEANR_8300_1
GLEANR_8301_1
GLEANR_8353_1
GLEANR_8733_1
GLEANR_896_1
GLEANR_897_1
GLEANR_898_1
GLEANR_9077_1
GLEANR_9148_1
GLEANR_9271_1
GLEANR_9299_1
GLEANR_9354_1
GLEANR_9355_1
GLEANR_9356_1
GLEANR_9391_1
GLEANR_9475_1
GLEANR_9476_1
GLEANR_9523_1
GLEANR_9585_1
GLEANR_9586_1
GLEANR_9587_1
GLEANR_9588_1
GLEANR_9589_1
GLEANR_9677_1
GLEANR_9678_1
GLEANR_9679_1
GLEANR_9680_1
GLEANR_9964_1
GLEANR_9988_1
GLEANR_9989_1

Supplementary Table 1. Serine endoproteases identified in the GLEANR annotations of the *D. mojavensis* genome (http://rana.lbl.gov/drosophila/).

FRTP	paralog	KA	KS	KA/KS	protein %ID	CDS %ID
dmoj_GLEANR_3081	dmoj_GLEANR_3082	0.0021	0	99	99.58	99.86
dmoj_GLEANR_3081	dmoj_GLEANR_3083	0.0038	0	99	99.23	99.74
dmoj_GLEANR_1234	dmoj_GLEANR_11311	0.0075	0.0032	2.3427	98.67	99.38
dmoj_GLEANR_6984	dmoj_GLEANR_12691	0	0.0218	0.001	100	99.52
dmoj_GLEANR_6725	dmoj_GLEANR_6724	0.0242	0.0574	0.4214	95.8	96.74
dmoj_GLEANR_13880	dmoj_GLEANR_11380	0.0779	0.1	0.7789	85.05	92.21
dmoj_GLEANR_1234	dmoj_GLEANR_1233	0.0345	0.1001	0.3449	94.52	95.27
dmoj_GLEANR_12931	dmoj_GLEANR_12932	0.0304	0.1123	0.2703	94.23	95.35
dmoj_GLEANR_4546	dmoj_GLEANR_4547	0.035	0.1161	0.3019	92.74	95
dmoj_GLEANR_5037	dmoj_GLEANR_5036	0.1854	0.1485	1.2483	73.02	85.01
dmoj_GLEANR_897	dmoj_GLEANR_896	0.1271	0.1782	0.7134	83.33	87.5
dmoj_GLEANR_896	dmoj_GLEANR_897	0.1271	0.1782	0.7134	83.33	87.5
dmoj_GLEANR_12325	dmoj_GLEANR_12324	0.0579	0.2085	0.2776	89.82	91.48
dmoj_GLEANR_12324	dmoj_GLEANR_12325	0.0579	0.2085	0.2776	89.82	91.48
dmoj_GLEANR_898	dmoj_GLEANR_897	0.1341	0.2114	0.634	79.93	86.37
dmoj_GLEANR_897	dmoj_GLEANR_898	0.1341	0.2114	0.634	79.93	86.37
dmoj_GLEANR_898	dmoj_GLEANR_896	0.1894	0.2289	0.8272	74.24	82.95
dmoj_GLEANR_896	dmoj_GLEANR_898	0.1894	0.2289	0.8272	74.24	82.95
dmoj_GLEANR_898	dmoj_GLEANR_2575	0.1253	0.2292	0.5464	79.29	86.55
dmoj_GLEANR_2575	dmoj_GLEANR_898	0.1253	0.2292	0.5464	79.29	86.55
dmoj_GLEANR_13880	dmoj_GLEANR_81	0.1469	0.2408	0.6101	73.95	85.15
dmoj_GLEANR_897	dmoj_GLEANR_2575	0.1371	0.2963	0.4626	78.81	84.63
dmoj_GLEANR_2575	dmoj_GLEANR_897	0.1371	0.2963	0.4626	78.81	84.63
dmoj_GLEANR_9617	dmoj_GLEANR_8260	0.1104	0.2981	0.3702	82.57	86.38
dmoj_GLEANR_2575	dmoj_GLEANR_896	0.202	0.3017	0.6695	73.86	80.93
dmoj_GLEANR_12010	dmoj_GLEANR_12011	0.1991	0.3052	0.6523	71.21	80.93
dmoj_GLEANR_896	dmoj_GLEANR_2575	0.2012	0.3112	0.6468	73.86	80.81
dmoj_GLEANR_13880	dmoj_GLEANR_6176	0.1382	0.3118	0.4433	79.45	84.25
dmoj_GLEANR_13880	dmoj_GLEANR_6373	0.0913	0.339	0.2693	84.12	86.47
dmoj_GLEANR_7051	dmoj_GLEANR_8436	0.1205	0.3549	0.3396	79.88	85.45
dmoj_GLEANR_13880	dmoj_GLEANR_11323	0.2366	0.3658	0.6466	69.08	78.07
dmoj_GLEANR_13880	dmoj_GLEANR_15318	0.2464	0.3749	0.6571	67.83	77.39
dmoj_GLEANR_778	dmoj_GLEANR_2703	0.185	0.4029	0.4591	71.26	80.97
dmoj_GLEANR_2703	dmoj_GLEANR_778	0.185	0.4029	0.4591	71.26	80.97
dmoj_GLEANR_7051	dmoj_GLEANR_8434	0.1713	0.4237	0.4042	75	82.41
dmoj_GLEANR_13880	dmoj_GLEANR_14459	0.2285	0.4402	0.519	70.22	77.21
dmoj_GLEANR_896	dmoj_GLEANR_2574	0.2259	0.4521	0.4996	70.08	77.53
dmoj_GLEANR_2575	dmoj_GLEANR_2574	0.1738	0.4552	0.3819	74.91	80.24
dmoj_GLEANR_897	dmoj_GLEANR_2574	0.1732	0.467	0.3708	75.76	80.18
dmoj_GLEANR_13880	dmoj_GLEANR_282	0.1407	0.4876	0.2885	76.54	81.48
dmoj_GLEANR_13880	dmoj_GLEANR_6975	0.1302	0.4956	0.2627	78.76	81.86

Supplementary Table 2A. Candidate Recently Duplicated Lower Female Reproductive Tract Proteins in the *D. mojavensis* genome. Ka: estimated non-synonymous substitutions per non-synonymous site Ks: estimated synonymous substitutions per synonymous site, Ka/Ks: estimated ratio non-synonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site, PROT %ID: Protein % identity, CDS %ID: coding sequence % identity calculated in PAML (Yang 1997).

FRTP	Paralog	Ka	Ks	Ka/Ks	protein %ID	CDS %ID
CG30035-PB	CG8234	0.0587	0.1977	0.2971	88.93	91.46
IM10-PA	CG33470	0.0039	0.0039	1.0024	99.61	99.61
scpr-C-PA	scpr-B	0.0031	0.0656	0.0469	99.24	98.73
scpr-C-PA	scpr-A	0.0375	0.2813	0.1332	93.89	93

Supplementary Table 2B. Candidate Recently Duplicated Lower Female Reproductive Tract Proteins in the *D. melanogaster* genome. Ka: estimated non-synonymous substitutions per non-synonymous site Ks: estimated synonymous substitutions per synonymous site, Ka/Ks: estimated ratio non-synonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site, PROT %ID: Protein % identity, CDS %ID: coding sequence % identity calculated in PAML (Yang 1997).

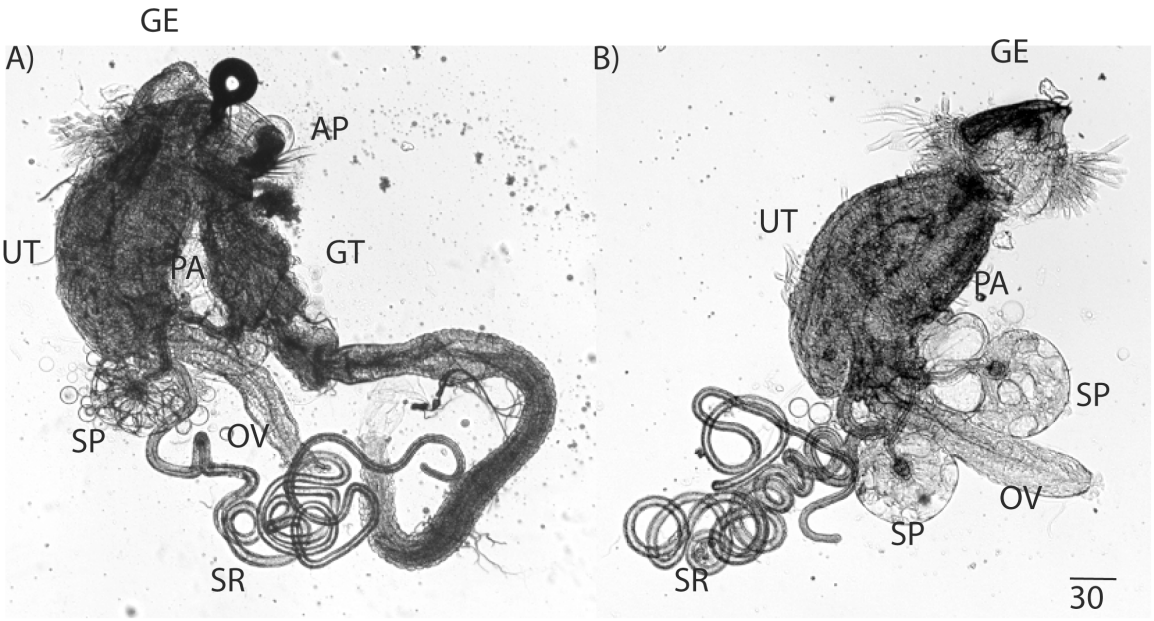
protease	paralog	Ka	Ks	Ka/Ks	protein %ID	CDS %ID
dmoj_GLEANR_12010	dmoj_GLEANR_12011	0.1991	0.3052	0.6523	71.21	80.93
dmoj_GLEANR_12011	dmoj_GLEANR_12010	0.1991	0.3052	0.6523	71.21	80.93
dmoj_GLEANR_12691	dmoj_GLEANR_6984	0	0.0218	0.001	100	99.52
dmoj_GLEANR_13196	dmoj_GLEANR_13197	0.1077	0.4615	0.2335	81.65	84.14
dmoj_GLEANR_13197	dmoj_GLEANR_13196	0.1077	0.4615	0.2335	81.65	84.14
dmoj_GLEANR_15984	dmoj_GLEANR_16100	0.051	0.0958	0.5318	91.9	94.13
dmoj_GLEANR_16100	dmoj_GLEANR_15984	0.051	0.0958	0.5318	91.9	94.13
dmoj_GLEANR_16735	dmoj_GLEANR_17599	0.0026	0.0138	0.1918	99.38	99.52
dmoj_GLEANR_17434	dmoj_GLEANR_17435	0.0451	0.1498	0.3009	90.2	94.12
dmoj_GLEANR_17436	dmoj_GLEANR_17437	0.0118	0.0001	99	98.64	99.09
dmoj_GLEANR_17437	dmoj_GLEANR_17436	0.0118	0.0001	99	98.64	99.09
dmoj_GLEANR_17437	dmoj_GLEANR_17468	0.071	0.3668	0.1937	89.53	88.76
dmoj_GLEANR_17466	dmoj_GLEANR_17467	0.0331	0.1721	0.1922	93.56	93.94
dmoj_GLEANR_17467	dmoj_GLEANR_17466	0.0331	0.1721	0.1922	93.56	93.94
dmoj_GLEANR_17468	dmoj_GLEANR_17437	0.071	0.3668	0.1937	89.53	88.76
dmoj_GLEANR_17599	dmoj_GLEANR_16735	0.0026	0.0138	0.1918	99.38	99.52
dmoj_GLEANR_2240	dmoj_GLEANR_2241	0	0	0.0344	100	100
dmoj_GLEANR_2241	dmoj_GLEANR_2240	0	0	0.001	100	100
dmoj_GLEANR_2574	dmoj_GLEANR_896	0.2259	0.4521	0.4996	70.08	77.53
dmoj_GLEANR_2574	dmoj_GLEANR_2575	0.1738	0.4552	0.3819	74.91	80.24
dmoj_GLEANR_2574	dmoj_GLEANR_897	0.1732	0.467	0.3708	75.76	80.18
dmoj_GLEANR_2575	dmoj_GLEANR_898	0.1253	0.2292	0.5464	79.29	86.55
dmoj_GLEANR_2575	dmoj_GLEANR_897	0.1371	0.2963	0.4626	78.81	84.63
dmoj_GLEANR_2575	dmoj_GLEANR_896	0.202	0.3017	0.6695	73.86	80.93
dmoj_GLEANR_2575	dmoj_GLEANR_2574	0.1738	0.4552	0.3819	74.91	80.24
dmoj_GLEANR_2655	dmoj_GLEANR_2656	0.053	0.134	0.3954	88.68	93.71
dmoj_GLEANR_2656	dmoj_GLEANR_2655	0.053	0.134	0.3954	88.68	93.71
dmoj_GLEANR_3285	dmoj_GLEANR_3286	0.2317	0.2989	0.7752	66.54	79.66
dmoj_GLEANR_3285	dmoj_GLEANR_3287	0.2279	0.322	0.7078	66.93	79.53
dmoj_GLEANR_3286	dmoj_GLEANR_3287	0.0625	0.0773	0.8089	88.58	93.83
dmoj_GLEANR_3286	dmoj_GLEANR_3285	0.2317	0.2989	0.7752	66.54	79.66
dmoj_GLEANR_3287	dmoj_GLEANR_3286	0.0625	0.0773	0.8089	88.58	93.83
dmoj_GLEANR_3287	dmoj_GLEANR_3285	0.2279	0.322	0.7078	66.93	79.53
dmoj_GLEANR_3349	dmoj_GLEANR_3350	0.0049	0.007	0.6978	98.83	99.48
dmoj_GLEANR_3349	dmoj_GLEANR_6036	0.0096	0.122	0.0786	98.05	97.27
dmoj_GLEANR_3350	dmoj_GLEANR_3349	0.0049	0.007	0.6978	98.83	99.48
dmoj_GLEANR_3350	dmoj_GLEANR_6036	0.0144	0.1222	0.1181	96.88	96.88
dmoj_GLEANR_6036	dmoj_GLEANR_3349	0.0096	0.122	0.0786	98.05	97.27
dmoj_GLEANR_6036	dmoj_GLEANR_3350	0.0144	0.1222	0.1181	96.88	96.88
dmoj_GLEANR_6984	dmoj_GLEANR_12691	0	0.0218	0.001	100	99.52
dmoj_GLEANR_896	dmoj_GLEANR_897	0.1271	0.1782	0.7134	83.33	87.5
dmoj_GLEANR_896	dmoj_GLEANR_898	0.1894	0.2289	0.8272	74.24	82.95
dmoj_GLEANR_896	dmoj_GLEANR_2575	0.2012	0.3112	0.6468	73.86	80.81

dmoj_GLEANR_896	dmoj_GLEANR_2574	0.2259	0.4521	0.4996	70.08	77.53
dmoj_GLEANR_897	dmoj_GLEANR_896	0.1271	0.1782	0.7134	83.33	87.5
dmoj_GLEANR_897	dmoj_GLEANR_898	0.1341	0.2114	0.634	79.93	86.37
dmoj_GLEANR_897	dmoj_GLEANR_2575	0.1371	0.2963	0.4626	78.81	84.63
dmoj_GLEANR_897	dmoj_GLEANR_2574	0.1732	0.467	0.3708	75.76	80.18
dmoj_GLEANR_898	dmoj_GLEANR_897	0.1341	0.2114	0.634	79.93	86.37
dmoj_GLEANR_898	dmoj_GLEANR_896	0.1894	0.2289	0.8272	74.24	82.95
dmoj_GLEANR_898	dmoj_GLEANR_2575	0.1253	0.2292	0.5464	79.29	86.55

Supplementary Table 3A. Candidate Recently Duplicated Serine Endoproteases in the *D. mojavensis* genome. Ka: estimated non-synonymous substitutions per non-synonymous site Ks: estimated synonymous substitutions per synonymous site, Ka/Ks: estimated ratio non-synonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site, PROT %ID: Protein % identity, CDS %ID: coding sequence % identity calculated in PAML (Yang 1997).

protease	paralog	Ka	Ks	Ka/Ks	protein %ID	CDS %ID
alphaTry-PA	CG30025-PA GB_protein	0.0605	0.395	0.1531	88.54	89.59
alphaTry-PA	deltaTry-PA GB_protein	0.0586	0.3994	0.1467	88.93	89.72
alphaTry-PA	CG30031-PA GB_protein	0.0586	0.3994	0.1467	88.93	89.72
alphaTry-PA	gammaTry-PA GB_protein	0.0586	0.3994	0.1467	88.93	89.72
betaTry-PA	CG30025-PA GB_protein	0.0724	0.4793	0.151	88.14	88.27
CG1304-PA	Ser6-PA GB_protein	0.1109	0.4132	0.2683	81.85	84.56
CG18477-PA	CG31780-PB GB_protein	0	0	0.1493	100	100
CG18478-PA	CG31827-PA GB_protein	0	0	0.1104	100	100
CG18557-PA	CG18557-PA GB_protein	0	0	0.4483	100	100
CG30415-PA	CG30415	0	0	0.2863	100	100
gammaTry-PA	deltaTry-PA GB_protein	0	0	0.4605	100	100
gammaTry-PA	CG30031-PA GB_protein	0	0	99	100	100
gammaTry-PA	CG30025-PA GB_protein	0.0016	0.0213	0.0768	99.6	99.47
gammaTry-PA	alphaTry-PA GB_protein	0.0586	0.3994	0.1467	88.93	89.72
grass-PA	grass-PA GB_protein	0	0	0.1882	100	100
Jon99Cii-PA	Jon99Ciii-PA GB_protein	0	0.0235	0.001	100	99.62
Jon99Ciii-PA	Jon99Cii-PA GB_protein	0	0.0235	0.001	100	99.62
Jon99Fi-PA	Jon99Fii-PA GB_protein	0.003	0.1197	0.0249	99.25	98.13
Jon99Fii-PA	Jon99Fi-PA GB_protein	0.003	0.1197	0.0249	99.25	98.13
olf186-F-PE	olf186-F-PB GB_protein	0.2255	0.4958	0.4547	74.64	77.3
Ser6-PA	CG1304-PA GB_protein	0.1109	0.4132	0.2683	81.85	84.56
sphinx1-PB	sphinx2-PB	0.1118	0.2989	0.3741	82.13	86.52

Supplementary Table 3B. Candidate Recently Duplicated Serine Endoproteases in the *D. melanogaster* genome. Ka: estimated non-synonymous substitutions per non-synonymous site Ks: estimated synonymous substitutions per synonymous site, Ka/Ks: estimated ratio non-synonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site, PROT %ID: Protein % identity, CDS %ID: coding sequence % identity calculated in PAML (Yang 1997).



Supplementary Figure 1. Dissected female reproductive tracts of *D. arizonae*. A) The Drosophila lower female reproductive tract is intimately associated with the gastrointestinal tract. The two are furthermore attached by a thin layer of chitin connecting the external genitalia to the anal plate. B) By breaking this attachment and removing female reproductive tracts directly by the ovipositor, we are able to obtain whole female reproductive tracts with no contamination from gut tissue. We furthermore never remove or disrupt the gut from the body cavity of the fly, thus minimizing the possibility of contamination from gastrointestinal proteases in our assays. GE = external genitalia, UT = uterus, SP = spermathecae SR - seminal receptacle, OV = oviduct, PA = parovaria, GT = gastrointesinal tract, AP = anal plate. Scale bar = 30 micrometers.

APPENDIX D: DUPLICATION, SELECTION, AND GENE CONVERSION IN A *DROSOPHILA MOJAVENSIS* FEMALE REPRODUCTIVE PROTEIN FAMILY

Kelleher ES and Markow TA. 2009. Duplication, Selection and Gene Conversion in a Drosophila mojavensis Female Reproductive Protein Family. Genetics. 181:1451-1465.

ABSTRACT

Protein components of the *Drosophila* male ejaculate, several of which evolve rapidly, are critical modulators of reproductive success. Recent studies of female reproductive tract proteins indicate they also are extremely divergent between species, suggesting that reproductive molecules may coevolve between the sexes. Our current understanding of intersexual coevolution, however, is severely limited the paucity of genetic and evolutionary studies on the female molecules involved. Physiological evidence of ejaculate-female coadaptation, paired with a promiscuous mating system, makes *D. mojavensis* an exciting model system in which to study the evolution of reproductive proteins. Here we explore the evolutionary dynamics of a five paralog gene family of female reproductive proteases within populations of *D. mojavensis* and throughout the *repleta* species group. We show that the proteins have experienced ongoing gene duplication and adaptive evolution, and further exhibit dynamic patterns of pseudogenation, copy number variation, gene conversion, and selection within geographically isolated populations of *D. mojavensis*. The integration of these patterns in a single gene family has never before been documented in a reproductive protein.

INTRODUCTION

In internally fertilizing organisms, female reproductive tracts are the arena for a dynamic molecular interface between the sexes. Ejaculate-female interactions are essential to sperm fate and fertilization, guiding sperm through the female reproductive tract, preserving them in this environment, and ultimately mediating gamete fusion (Reviewed in NEUBAUM and WOLFNER 1999). Reproductive tract interactions also modulate critical post-mating changes in female behavior and physiology, such as upregulating immune response, reformatting the female reproductive tract, and delaying female remating (Reviewed in WOLFNER 2007; ROBERTSON 2007).

Despite the significance of ejaculate-female interactions for overall fitness, the male molecules involved in these processes exhibit dynamic evolutionary histories. Seminal proteins and sperm proteins have been observed to evolve rapidly in a broad range of taxa (Reviewed in SWANSON and VACQUIER 2002; CLARK *et al* 2006; PANHUIS *et al* 2006). Similarly, lineage-specific gene duplications have been documented in *Drosophila* seminal fluid proteins (CIRERA and AGUADÉ 1998; WAGSTAFF and BEGUN 2007; FINDLAY *et al* 2008; ALMEIDA and DESALLE 2008a; 2008b), as well as fertilization proteins in both *Drosophila* and abalone (LOPPIN *et al* 2005; CLARK *et al* 2007). Finally, *Drosophila* male ejaculates are known to undergo a high frequency of lineage-specific changes in seminal fluid content, by functionally co-opting existing genes and acquiring novel genes from non-coding sequence (BEGUN and LINDFORS 2005; MUELLER *et al* 2005; BEGUN *et al* 2006; FINDLAY *et al* 2008).

The rapid evolution of male ejaculates frequently is postulated to arise from molecular coevolution with interacting proteins in the female reproductive tract (PARKER 1979; EBERHARD 1996; SWANSON and VACQUIER 2002). If this is the case, female reproductive molecules also are expected to evolve rapidly. Recent evidence of adaptive evolution in *Drosophila* female reproductive tract proteins is consistent with this prediction (SWANSON *et al* 2004; PANHUIS and SWANSON 2006; KELLEHER *et al* 2007; LAWNICZAK and BEGUN 2007; PROKUPEK *et al* 2008). Compared to the preponderance of studies of male ejaculates, however, the dynamics of female proteins remain largely unexplored.

Two, non-mutually exclusive mechanisms are hypothesized to result in reciprocal evolutionary change between male and female reproductive molecules. First, cryptic female choice could empower females to bias fertilization success towards certain males based on post-copulatory biochemical cues (EBERHARD 1996). Cryptic female choice may lead to cyclical evolution of male trait and female preference, consistent with traditional models of runaway sexual selection (FISHER, 1915; 1930). Alternatively, sexual conflict, or a difference in the reproductive interests of the two sexes (PARKER 1979), is predicted to result in an evolutionary arms race between males and females (RICE 1996; GAVRILETS 2000).

In this study, we explore the dynamics of a female reproductive tract protein gene family in the cactophilic fruit-fly *D. mojavensis*. A promiscuous mating system (reviewed in MARKOW 1996), as well as extensive evidence of ejaculate-female biochemical coadaptation (KNOWLES and MARKOW 2001; PITNICK *et al* 2003; KNOWLES *et al* 2005; KELLEHER and MARKOW 2007) makes *D. mojavensis* an extraordinary system for the study of reproductive molecules. Specifically, interpopulation crosses exhibit significant

differences from intrapopulation crosses in egg size (PITNICK *et al* 2003), a mating-dependent increase in female desiccation resistance (KNOWLES *et al* 2005), and the size and duration of the insemination reaction, an opaque mass that forms in the uterus after copulation (KNOWLES and MARKOW 2001). Similarly, interspecific crosses between *D. mojavensis* and its sister-species *D. arizonae* (MRCA ~ 0.7 MYA, REED *et al* 2007, MATZKIN 2008), exhibit considerable sperm mortality, failure in sperm storage, reduced oviposition, and aberrant insemination reactions, consistent with a breakdown in coadapted gene complexes (KELLEHER and MARKOW 2007).

The gene family examined here is one of five lineage-specific protease gene families identified from *D. arizonae* female reproductive tracts, and encodes five serine-endoprotease paralogs: Dmoj\GLEANR_2575 (GI17776), Dmoj\GLEANR_2574 (GI17775), DmojGLEANR_896 (GI23802), Dmoj\GLEANR_897 (GI23804), and Dmoj\GLEANR_898 (GI23805) (Figure 1, KELLEHER *et al* 2007). Although the specific function of these enzymes remains unknown, they are predicted secreted proteins expressed only in the lower female reproductive tract, implying specialized interaction with the male ejaculate (KELLEHER *et al* 2007). Serine endoprotease activity in *D. arizonae* female reproductive tracts, furthermore, is regulated by mating, pointing to a direct relationship between reproduction and proteolytic function (KELLEHER and PENNINGTON, *submitted*).

If female reproductive tract proteases are coevolving with the male ejaculate, two predictions follow about their evolutionary dynamics. First, the coevolutionary trajectory within each population should exert unique selective pressures on the proteins involved.

To explore this hypothesis we compare patterns of variation and deviations from neutrality at these loci between the four geographically isolated populations of *D. mojavensis*: Baja Peninsula, Catalina Island, Mainland Sonora and Mojave Desert (REED *et al* 2007; MACHADO *et al* 2007, Figure 2). Second, ongoing coevolution with interacting proteins predicts a history of adaptive evolution across the *repleta* species group. We therefore examine patterns of divergence at these loci from five *repleta* group species and two outgroups. We discuss our results in terms of our predictions, as well as the emerging role of gene duplication in reproductive protein evolution.

MATERIALS AND METHODS

Flies. *Drosophila mojavensis* were collected from Catalina Island (2001), Mojave Desert (2002), Baja Peninsula (2002), and Mainland Sonora (2007) by J. Bono, L. Reed, and L. Matzkin. *Drosophila arizonae* were collected in Tucson, Arizona (2000) by L. Matzkin. *Drosophila navajoa*, *D. mettleri*, and *D. mayaguana* were obtained from the Tucson *Drosophila* Stock Center, now located at the University of California at San Diego. All flies used in population analyses were maintained as isofemale lines. Between 7 and 14 isofemale lines were sampled for each population and locus (supplementary table 1).

Sequencing. Genomic DNA was isolated from whole flies using the DNeasy Kit (Qiagen) according to manufacturer instructions. For *D. mojavensis* and *D. arizonae*, standard PCR was performed using internal, paralog-specific primers (Figure 1). In cases

where gene conversion obscured paralog identity (GLEANR_896 and GLEANR_897), additional flanking primers were used to ensure gene-specific amplification. For *D. navajoa*, *D. mettleri*, and *D. mayaguana* universal primers for the entire gene family were used to amplify and clone PCR products. Cloned PCR products were sequenced using M13F and M13R primers. All sequencing was performed on an ABI 3700 DNA sequencer with Big Dye Terminator chemistry. *Drosophila grimshawi* and *D. virilis* sequences were obtained from their sequenced genomes (http://rana.lbl.gov/drosophila/). Primers and PCR conditions are available from the authors upon request. Base-calling and assembly were performed in Sequencher 4.8.

Inverse Polymerase Chain Reaction (PCR). Genomic DNA from a single Mojave Desert isofemale line was digested with each of four restriction enzymes according to manufactuter instructions (New England Biolabs): Aci I, Mbo I, Mse I, and Taq I. Digested fragments were then incubated with ~20 units DNA ligase (Fermentas) at 17 C overnight to generate circularized DNA. Circularized DNA was then used for standard PCR with inverted primers specific to the novel paralog. Primers and PCR conditions are available from the authors upon request.

Reverse Transcriptase Polymerase Chain Reaction. Total RNA was extracted from 20 adult males, 20 adult female reproductive tracts (oviduct, spermathecae, seminal receptacle, parovaria, uterus), and 20 adult female carasses (no female reproductive tract) from a Mojave Desert isofemale line using TRIZOL reagent (Invitrogen), according to

manufacturer instructions. RNA was treated with Dnase I (NEB) and reverse-transcribed with the iScript cDNA synthesis kit (Roche). Resultant cDNA was diluted to 5 ng/μL for all three samples, and used as template for standard PCR with ribosomal protein 32 (control) and paralog-specific (experimental) primers. Quantity of resultant product was compared on a 1% agarose gel stained with ethidium bromide. Primers and PCR conditions are available from the authors upon request.

Polymorphism Analyses. Haplotypes were phased in Arlequin (http://lgb.unige.ch/arlequin/software/), and subsequent polymorphism analyses, estimation of population parameters, and tests of selection were performed in DNAsp (ROZAS and ROZAS, 1995) and SITES (http://lifesci.rutgers.edu/~heylab/ProgramsandData/Programs/SITES/SITES). Sample sizes, sequence lengths, estimates of polymorphism, site frequency spectra tests, and McDonald Kreitman tests (MCDONALD and KREITMAN, 1991) for all loci are presented in supplementary table 1. Significance of site frequency spectra statistics was assessed by coalescent simulations under the conservative assumption of no recombination. For tests requiring an outgroup, one or more *D. arizonae* orthologs were used.

Gene conversion was detected by GENECONV (http://www.math.wustl.edu/~sawyer/geneconv/) within an alignment of all unique haplotypes for all paralogs using the method of SAWYER (1989). Briefly, gene conversion tracts between pairs of sequences are identified by stretches of complete identity interspersed between two regions of considerable mismatch, or one region of mismatch

and the end of the alignment. Statistical significance of these fragments is determined by permutation tests. Neighbor-joining gene trees (SAITOU and NEI 1987) were constructed in Paup*4.0b10 (SWOFFORD 2000).

HKA tests. Polymorphism data from all 10 random loci in MACHADO *et al* (2007) were partitioned into the four geographic populations of *D. mojavensis* and a single *D. arizonae* outgroup sequence. Polymorphism and divergence for these loci was measured in DNAsp (ROZAS and ROZAS, 1995), and neutrality was assessed by the method of HUDSON, KREITMAN and AGUADÉ (1987), implemented in HKA (http://lifesci.rutgers.edu/~heylab/heylabsoftware.htm#HKA). For the complete set of 10 loci, significant deviations from neutrality were detected in all four populations of *D. mojavensis*. To identify a neutral sample, loci with large deviations from expected values were sequentially removed until the P-value of the HKA test was > 0.1. The neutral sample was then compared against experimental loci using HKA.

Phylogenetic Analyses. Consensus sequences were used to eliminate mutations introduced by cloning or Taq DNA polymerase. Sequences were additionally screened by eye to identify PCR recombinants. No such chimeric sequences were found. Phylogenetic relationships were inferred with Mr. Bayes (http://mrbayes.csit.fsu.edu/authors.php).

Codon-Based Analyses of Adaptive Evolution. Netsted maximum-likelihood models of codon evolution were implemented in the codeml program of PAML (YANG 1997), and

compared using likelihood ratio tests. Two tests of positive selection were performed. In the first test the neutral model (M1) is compared with the selection model, in which a class of sites is permitted to exhibit d_N/d_S (ω) > 1 (M2). In the second test, a beta distribution of site classes in which the most rapidly evolving is constrained to $\omega \leq 1$ (M7) was compared to a similar model in which the most rapidly evolving site class is permitted to exhibit $\omega > 1$ (M8). Multiple initial values of ω were used to ensure convergence on the likelihood optima.

Two additional tests were implemented to determine if specific branches on the phylogeny had experienced adaptive evolution. First, a free-ratios model, in which each branch is allowed to have a different d_N/d_S, was compared to a model where the d_N/d_S of the branch of interest was fixed to 1 (YANG 1998). Second, a branch site model, in which the branch of interest is allowed a rapidly evolving class of sites, $\omega > 1$, was compared to a similar model in which ω is fixed to 1 (YANG *et al* 2005).

Three Dimensional (3D) Modeling. Bayes Empirical Bayes positively selected sites predicted under M8 (YANG 1997; YANG *et al* 2005), catalytic sites (Reviewed in POLGAR 2005), and protease inhibitor sites (Reviewed in SRINIVASAN *et al* 2006) were mapped to a predicted 3D model for GLEANR_898 obtained from Swiss-Model (SCHWEDE *et al* 2003).

We tested for an association between positively selected sites and protease inhibitor sites using a permutation test previously implemented in CLARK *et al* (2007). The test statistic was the mean distance from each selected site to the nearest inhibitor

site. Each permutation identified a random set of selected sites, equal in number to those observed, and calculated the statistic for that set. Buried, core sites were not considered for random sets, because they evolve at a relatively slower rate than surface sites and are rarely inferred as positively selected. This exclusion makes the test more conservative. Buried sites were those with 10% or less surface exposure per residue as calculated by GETAREA (FRACZKIEWICZ and BRAUN 1998). A p-value was determined as the fraction of random permutations with a mean distance equal to or lower than the observed mean distance between selected and inhibitor sites. The test for clustering of positively selected sites was similar except that the test statistic was the mean pairwise distance between all selected sites as described in CLARK and SWANSON (2005).

RESULTS

A Novel Gene-Duplicate in the Mojave Desert Population. Consistent, reproducible heterozygosity in sequence data for GLEANR_896 in multiple individuals from seven isofemale lines derived from the Mojave Desert population suggested the acquisition of a novel paralog. Flanking sequence upstream of the novel paralog generated by inverse-PCR identified a breakpoint with the repetitive element dmoj_2 (http://insects.eugenes.org/species/cgi-bin/gbrowse/dmoj/). Although this repetitive element made subsequent inverse-PCR uninformative, test PCRs pairing a primer on the breakpoint with multiple primers in the coding sequences of GLEANR_896 and GLEANR_897 amplified an approximately 2kb fragment between the breakpoint and the

3’ end of GLEANR_897. The sequence of this fragment included an additional breakpoint between dmoj_2 and the 3’ flanking sequence of GLEANR_897. We thus hypothesize that the new paralog maps to the intergenic sequence between GLEANR_897 and GLEANR_898 (Figure 1).

Using the breakpoint between the new paralog and dmoj_2 we were able to design paralog-specific primers and obtain sequence for 13 of 14 sampled isofemale lines from the Mojave Desert. We were unable to amplify the new paralog from any isofemale lines from Mainland Sonora, Catalina Island, or the Baja Peninsula. Southern blots further confirmed that this paralog is absent from all sampled isofemale lines from these three localities (not shown).

To determine if and where the new paralog is expressed we performed semi-quantitative RT-PCR on sexually mature adult males, sexually mature lower female reproductive tracts, and sexually mature female carcasses lacking their female reproductive tracts (supplementary figure 1). Similar to the other five paralogs, the novel paralog was expressed exclusively in females, with enriched expression in lower female reproductive tracts (supplementary figure 1). Resultant cDNA was sequenced to verify paralog identity. Collectively, these data indicate that the Mojave Desert population recently has acquired a novel paralog, whose expression pattern suggests female-specific reproductive function.

Ectopic Recombination. Ectopic recombination, through both non-allelic homologous recombination and gene conversion, facilitates exchange of genetic information between

paralogous members of a multigene family. It is critical to describe ectopic recombination in population data, as this process can significantly alter patterns of polymorphism in duplicated genes (INNAN 2003, THORNTON 2007). We employed GENECONV (SAWYER, 1989) to identify pairs of divergent paralogous haplotypes that share regions of complete identity, indicative of gene conversion (Figure 3). No gene conversion tracts were detected between the most basal duplicate, GLEANR_2574, and any other paralog, suggesting this paralog evolves independently (Figure 3). Significant fragments, however, were detected for at least one haplotype of all other paralogs in the gene family (Figure 3).

The highest frequency of significant converted fragments, as well as the longest average fragment length, were observed between the adjacent, closely related duplicates GLEANR_896, GLEANR_897 and the new paralog (Figure 3, also see table of polymorphism, supplementary table 2). Gene genealogies of GLEANR_896 and GLEANR_897 haplotypes, furthermore, revealed that these loci are not reciprocally monophyletic, suggesting extensive ectopic recombination between paralogous lineages (Figure 2, supplementary table 2). In contrast, no recombination is detected between genetically and physically distant paralogs GLEANR_896 and GLEANR_2575 (Figure 3). Ectopic recombination, therefore, is negatively associated with both phylogenetic and physical distance.

In many cases, it was impossible to infer the directionality of gene conversion, in terms of a donor and recipient paralog. For GLEANR_896 and GLEANR_897, however, putatively ancestral haplotypes group with the *D. arizonae* ortholog, while converted

haplotypes group with the alternate paralog (Figure 2). Ancestral haplotypes, furthermore, are found in all four populations, while converted haplotyes are population-specific. Thus, converted haplotypes of GLEANR_896 have been recipients of genetic variation from ancestral GLEANR_897 donors, and reciprocally, converted haplotypes of GLEANR_897 have been recipients of genetic variation from ancestral GLEANR_896 donors (Figure 2). The approximate gene conversion tract length was 518 bp for GLEANR_896 conversion haplotypes, and 443 bp for GLEANR_897 conversion haplotypes (of ~700 aligned bases), based on visual examination of polymorphic sites (see supplementary table 2).

Ectopic recombination involving the genetically more distant paralogs, GLEANR_898 and GLEANR_2575, was not extensive enough to degrade allelic monophyly. Gene genealogies of converted and unconverted regions were therefore compared separately to determine if the evolutionary history of these two portions of the gene could be confidently inferred (Figure 4). In two cases, gene conversion tracts from a set of recipient haplotypes grouped with all haplotypes from a donor paralog with high bootstrap support (Figure 4), indicating the direction of gene conversion.

To explore the contribution of genetic exchange between paralogs to genetic variation within populations, we estimated nucleotide diversity (π) for both the complete set of sampled alleles from a given population, as well as for the sample with all recipient alleles excluded. In all cases, our estimate of π was lower when recipient alleles were excluded (Table 1). In four cases, furthermore, the observed decrease was greater than

two standard deviations, indicating that ectopic recombination contributes significantly to standing variation within populations (Table 1).

Segregating Pseudogenes. Functional redundancy between recent duplicates is predicted to result in relaxed evolutionary constraint at individual paralogs, allowing for the acquisition of deleterious mutations or complete loss of function (OHNO 1970; HUGHES 1994; FORCE *et al* 1999). Consistent with this prediction, we found evidence of three distinct pseudogene haplotypes in two different paralogs, GLEANR_2575 and GLEANR_898. In the Baja Peninsula population, one premature stop codon and one frame-shift deletion are found in GLEANR_2575. These mutations occur prior to the first of three amino acid residues that comprise the catalytic triad (reviewed in POLGAR 2005), as well as residues that determine substrate binding affinity (SPRANG *et al* 1988), thus rendering the protease completely non-functional. Both alleles were resequenced to verify the mutations did not reflect amplification or sequencing errors. One converted allele of GLEANR_897 sampled from Mainland Sonora also contained a frame shift deletion, although insufficient DNA remained for resequencing of this individual. This frame shift occurs between the second and third amino acids in the catalytic triad, but prior to all residues that determine substrate binding affinity, and likely also renders the protease non-functional.

Pseudogene haplotypes often reflect relaxed purifying selection, but can also be maintained as balanced polymorphisms (HEXTER 1968; WIESENFELD 1968), or sweep rapidly through populations in cases of adaptive gene loss (STEDMAN *et al* 2004; WANG

et al 2006). GLEANR_2575 alleles sampled from the Baja Peninsula and GLEANR_897 alleles from Mainland Sonora do not exhibit deviations from neutrality in McDonald-Kreitman tests (MCDONALD and KREITMAN 1991), nor do they show a significant skew in the site frequency spectra (not shown). There is no evidence, therefore, that pseudogene haplotypes observed here confer a selective advantage.

Deviations from Neutrality at GLEANR_898. Standard McDonald-Kreitman (MCDONALD and KREITMAN, 1991) tests for GLEANR_898 indicate an excess of non-synonymous polymorphism, relative to divergence, in the Baja Peninsula, Catalina Island, and Mainland Sonora populations (Table 2). Intriguingly, both Mainland Sonora and Catalina Island exhibit segregating conversion alleles at this locus (Table 1). Although Catalina Island no longer exhibits a deviation from neutrality when segregating conversion alleles are excluded from the analysis, the G-test for Mainland Sonora remains significant (Table 2).

Balancing or diversifying selection is one possible explanation for an excess of non-synonymous polymorphism in a McDonald-Kreitman framework. In general, these selective regimes are accompanied by other patterns, such as an excess of intermediate frequency polymorphism, the appearance of two well-differentiated haplogroups that exhibit significant linkage disequilibrium, or an excess of polymorphism relative to divergence when compared to other loci (HUDSON *et al* 1987). No excess of intermediate frequency polymorphism is observed either in Mainland Sonora or the Baja Peninsula for GLEANR_898, as Tajima's *D* (TAJIMA 1989) is slightly negative in both cases (Table 2).

Two well-differentiated haplotypes groups, furthermore, are not apparent in gene genealogies of this locus (not shown). Z_{ns}, a measure of the correlation in allele frequencies across polymorphic sites (KELLY 1997), does not indicate significant linkage-disequilibrium at this locus (not shown). Finally, an HKA test (HUDSON *et al* 1987) detects no excess of polymorphism, relative to neutral loci (not shown). The data, therefore, provide little evidence that balancing selection is operating on the GLEANR_898 locus in Mainland Sonora or the Baja Peninsula.

An alternate explanation for the observed excess of replacement variation is that these sites represent weakly deleterious variants that contribute only to polymorphism, but not to divergence. If so, the majority of these variants should be segregating at low frequency (KIMURA 1983, NACHMAN *et al* 1998). Site frequency spectra for silent and replacement sites in GLEANR_898 were therefore compared separately for the Mainland Sonora and Baja California populations (Table 3). In both populations, Tajima's *D* (TAJIMA 1989) is slightly more positive for replacement sites than for silent sites, the opposite of what is expected for mildly deleterious variants (Table 3). The observed excess of non-synonymous polymorphism, therefore, does not appear to arise from weakly deleterious mutations.

A third explanation for the observed deviation from neutrality is that a recent relaxation in functional constraint may allow for the acquisition of replacement mutations that were not tolerated under the previous selective regime (TAKAHATA 1993). Although this scenario is difficult to verify empirically, it is plausible for a multigene family that may be undergoing antagonistic molecular coevolution. The degree of ectopic

recombination, as well as the frequency of segregating pseudogenes, suggests that the paralogs sampled here are at least partially functionally redundant. If coevolving interactors change their evolutionary "strategy", paralogs with formerly critical function could experience relaxed selective constraint.

Evolutionary History of the Novel Paralog. Neighbor joining analysis indicates that the novel paralog found in the Mojave Desert is most similar to converted alleles of GLEANR_897 from Mainland Sonora and the Baja Peninsula (Figure 2). Because the new duplicate is a chimera of GLEANR_896 and GLEANR_897, but is not nested between these two paralogs (Figure 1), the conversion haplotype and the gene duplication could not have resulted from a single event of non-allelic homologous recombination. The new paralog, therefore, likely has arisen via tandem duplication of a segregating conversion allele of GLEANR_897. Although it is impossible to determine the history of this gene with confidence, Figure 5 outlines a mechanism for the creation of the Mojave Desert chromosome with the fewest mutational steps. First, a gene conversion event from GLEANR_896 to GLEANR_897 creates a converted GLEANR_897 allele. Second, unequal crossing over, mediated by homologous or repetitive flanking sequence, results in a tandem gene duplication event. Third, this duplicated chromosome rises to high frequency in the Mojave Desert population.

It is intriguing that the duplication event in the Mojave Desert population unites the converted and unconverted haplotypes of GLEANR_897 on a single chromosome. This result is reminiscent of models in which two alleles are maintained as a balanced

polymorphism, and a subsequent gene duplication experiences immediate directional selection due to heterosis (SPOFFORD, 1969, OHNO 1970, OTTO and YONG 2002; WALSH 2003; PROULX and PHILLIPS 2006). If GLEANR_897 converted and unconverted haplotypes represent a balanced polymorphism, the GLEANR_897 converted haplotype should have arisen by a single ancestral gene conversion event, prior to the divergence of the Mainland Sonora, Baja Peninsula and Mojave Desert populations (0.45-0.68 MYA, REED *et al* 2007; MATZKIN 2008).

Although, all GLEANR_897 haplotypes group together with high-bootstrap support (Figure 2), this is not necessarily indicative of a single mutational origin for the converted haplotype. If ectopic recombination between paralogs is more frequent, or more frequently tolerated, in certain genetic regions, similar chimeric haplotypes could be generated continuously by gene conversion. If so, a considerable number of shared polymorphisms are expected between GLEANR_897 converted alleles and GLEANR_896 ancestral alleles within the converted region. The number of private and shared polymorhisms in converted GLEANR_897 alleles and GLEANR_896 ancestral alleles within the converted region are presented in Table 4. In the Mainland Sonora population only one polymorphism is shared between converted and ancestral alleles (Table 4), suggesting that converted alleles are not continuously sampling genetic variation from ancestral haplotypes. In Baja Peninsula, where eight shared polymorphisms are seen, the polymorphisms are associated with only two ectopic recombination events. Collectively, therefore, the data do not suggest a high frequency of gene conversion from GLEANR_896 ancestral alleles to GLEANR_897 converted

alleles. This result is in stark contrast to GLEANR_896 converted and GLEANR_897 ancestral alleles, which exhibit a high frequency of shared polymorphisms indicative of ongoing gene conversion (Table 4).

If the GLEANR_897 converted haplotype is an old balanced polymorphism, it is predicted to have acquired and maintained its own set of genetic variation. Consistent with this hypothesis, this haplogroup exhibits one silent and two replacement polymorphisms, fixed or at high frequency (>60%) amongst these alleles, which are not present in any other haplotype of GLEANR_896 or GLEANR_897 in *D. mojavensis*. A third amino acid variant, fixed in the GLEANR_897 haplogroup, is present in only one sampled haplotype of GLEANR_896 and was entirely absent from unconverted haplotypes of GLEANR_897. Intriguingly, the three amino acid variants, also found in the new paralog from the Mojave Desert, are shared with *D. arizonae* GLEANR_896. Sites that are shared with an outgroup are inferred to represent the ancestral state. Thus, the conversion tract in GLEANR_897 converted haplotypes appears to be derived from an ancestral allele of GLEANR_896 that is no longer segregating in any *D. mojavensis* population. Ancestral variation is expected if the converted haplotype resulted from an ancient gene conversion event that occurred prior to the radiation of the Mainland Sonora, Baja Peninsula, and Mojave Desert populations.

The confounding nature of gene conversion makes it problematic to present a compelling argument that the maintenance of GLEANR_897 converted and ancestral haplotypes is the result of balancing selection. Extensive gene conversion generates slightly positive values of Tajima's *D* (TAJIMA 1989), and furthermore makes this

statistic extremely conservative because the variance of the test statistic is over estimated (INNAN 2003). Similarly, HKA tests are inappropriate assessments of balancing selection for duplicates undergoing gene conversion because recombining paralogs are on average more polymorphic than single-copy loci (INNAN 2003; THORNTON 2007). Nonetheless, our data do suggest that the two haplotypes are old, have been retained in two of four geographically isolated populations of *D. mojavensis* for at least .45 MY, and have duplicated in a third population. The degree of linkage disequilibrium in both Mainland Sonora (Z_{ns} = 0.69 p = 0.01) and the Baja Peninsula (Z_{ns} = 0.82, p= 0.00), furthermore, indicate little recombination has occurred between haplogroups during this time. Determining the role of natural selection in maintaining the GLEANR_897 converted and ancestral polymorphism will present an important challenge for future studies.

Although our data suggest that gene duplication was preceded by allelic divergence between the GLEANR_897 ancestral and GLEANR_897 converted haplotypes, GLEANR_897 converted haplotypes are separated from the new paralog by an average of 20 nucleotide differences (~3%). The majority of these differences are inside the gene conversion tract. To explore if gene duplication may have been followed by a period of adaptive evolution, we estimated the corrected ratio of non-synonymous to synonymous divergence (d_N/d_S) for this branch in the portion of the alignment contiguous with the conversion tract (YANG 1998). Although the branch leading to the novel paralog does exhibit d_N/d_S of 1.25, consistent with adaptive evolution, this value does not provide a significantly better fit to the data than a model where the value is fixed to 1 (p = 1.00). A branch-site model, in which only a subset of sites on this branch were hypothesized to experience positive selection, similarly did not provide a significantly better fit to the data than a model that does not incorporate adaptive evolution (p = 1.00, YANG *et al* 2005).

Although these analyses provide little evidence for adaptive protein evolution following gene duplication, it is important to remember that their statistical power is extremely limited for branches where few changes have occurred.

Directional Selection. Although segregation of deleterious mutations clearly suggests relaxed purifying selection at some loci in this multigene family, we also find evidence for positive directional selection, a frequent observation amongst reproductive proteins (Reviewed in SWANSON and VACQUIER 2002; CLARK *et al* 2006; PANHUIS *et al* 2006). Catalina Island flies show an excess of low frequency polymorphism at GLEANR_898 and GLEANR_897, a possible indicator of recent directional selection (Table 5). Similarly, Mojave Desert flies exhibit an excess of low frequency polymorphism at GLEANR_898, and no segregating sites at GLEANR_897, GLEANR_896, or the new paralog (Table 4, Supplementary Table 1). A reanalysis of 7 autosomal and 3 sex-linked random loci sampled in MACHADO *et al* (2007) does not detect any significant skew towards positive or negative values in site frequency spectra tests for either of these populations (Supplementary Table 3). The observed excess of rare polymorphism, therefore, does not appear to result from demographic processes such as a recent population expansion. Gene conversion, furthermore, is known to skew Tajima's *D* marginally positive (INNAN 2003; THORNTON 2007), making the observation of significantly negative values highly unexpected.

To further test the hypothesis of directional selection, polymorphism and divergence between our experimental loci, and a group of loci that behave neutrally (MACHADO *et al* 2007, see materials and methods) were compared by the HKA test

(HUDSON *et al* 1987, Table 5). When including GLEANR_896 and GLEANR_897 in the data set, no deviations from neutrality were detected for the Catalina Island population (Table 5). It is important to note, however, that the HKA test is extremely conservative for duplicate genes experiencing ectopic recombination, as the expected level of polymorphism is higher than for single copy loci (INNAN 2003, THORNTON 2007). For the Mojave Desert population, GLEANR_897, as well as a test that included GLEANR_898, GLEANR_897 and GLEANR_896, both showed an excess of divergence consistent with directional selection. Although we cannot infer the causative mutation responsible for these patterns, it is intriguing that the selective sweep is associated with a chromosome harboring a novel duplicate. The novel duplicate could be adaptive because of its specific sequence, or alternatively, simply because it represents an additional gene copy.

Duplication and Adaptive Evolution in the *repleta* Species Group. To further elucidate the evolutionary history of this gene family, we sequenced paralogs across five *repleta* group species, *D. mojavensis, D. arizonae, D. navajoa*, *D. mayaguana*, and *D. mettleri*. Sequence data from the *D. grimshawi* and *D. virilis* genomes provided appropriate outgroups. Bayesian phylogenetic inference of 22 orthologs and paralogs indicates that the genes exist as a single copy in *D. grimshawi* and *D. virilis*, whereas three or more copies exist in all *repleta* group species (Figure 6). The radiation of the gene family, therefore, appears lineage-specific to the *repleta* species group. *D. mojavensis*, *D. navajoa*, *D. mayaguana*, and *D. mettleri*, furthermore, all exhibit two paralogs that are more closely related to each other than to any other sequence in the

alignment. This pattern, common to multigene families, suggests either ongoing gain and loss of individual paralogs, or concerted evolution by extensive ectopic recombination (Reviewed in NEI and ROONEY 2005). GENECONV detected a significant fragment in at least one paralog from *D. mojavensis*, *D. arizonae*, *D. mettleri*, and *D. mayaguana*, indicating ectopic recombination contributes to divergence of this multigene family. No significant fragments are found between lineage-specific paralogs from *D. navajoa* or *D. mayaguana* however, suggesting these are authentic lineage-specific duplicates. Observation of a novel paralog and segregating pseudogenes in the polymorphism data further supports the assertion that lineage-specific gain and loss is an ongoing process in the evolution of this gene family.

To determine if the gene family has experienced positive selection within the *repleta* species group, we implemented maximum-likelihood codon based models in PAML (YANG 1997). For this analysis, all nodes with a posterior probability <90 (Figure 6) were collapsed to polytomies to prevent spurious results due to inaccuracy in the tree topology. For two different tests of positive selection, a model that allowed for a class of sites that evolves adaptively ($d_N/d_S > 1$) provided a significantly better fit to the data than a model that did not (Table 6). The detected signature of adaptive evolution is consistent with our previous analysis (KELLEHER *et al* 2007).

Two aspects of our data could lead to an incorrect inference of adaptive evolution in this type of analysis. First, sequences from *D. navajoa*, *D. mayaguana*, and *D. mettleri* were obtained from cloned PCR products, meaning there could be mutations in the aligment that have been introduced by Taq DNA polymerase. All cloned sequences in the

alignment, however, are a consensus of three or more colonies except *D. mettleri*-1 and *D. mettleri*-2*,* and should therefore be free of PCR introduced mutations. A reanalysis of the data with *D. mettleri-1* and *D. mettleri*-2 excluded still yields highly significant test, indicating the inference of adaptive evolution is not the result of PCR error (Table 6).

The observed gene conversion in our alignment could also lead to spurious results in codon-based analysis of adaptive evolution, as recombination is known to cause false positives for this class of tests (ANISIMOVA *et al* 2003). To avoid this problem, two subsets of the alignment that included only one of a pair or group of sequences with evidence for gene conversion were created (Table 6). Analyses of these pruned alignments were still highly significant, indicating that the observed adaptive evolution is independent of gene conversion.

Depending on the data set, likelihood analysis suggests that between 3 and 13% of sites are experiencing positive selection, with an estimated d_N/d_S between 1.7 and 3.02 (Table 6). Bayes Empricial Bayes selected sites (YANG *et al* 2005), furthermore, are remarkably congruent between different data sets and different models (Table 6, Figure 7). Selected sites, shown in black, often are observed to be closely associated with sites important to protease inhibitor susceptibility and resistance (Figure 7, Reviewed in SRINIVASAN *et al* 2006). Indeed, three selected sites and protease inhibitor interaction sites occur at the same residue: a statistically significant excess (Fisher's Exact Test, $p = 0.0085$).

To further explore if selected sites and protease inhibitor interaction sites are associated in three dimensional space, we compared the average pairwise distance

between each selected site and the closest protease inhibitor interaction site to 10^6 sets of randomly sampled sites. Selected sites are significantly closer to protease inhibitor interaction sites than expected by chance ($p = 0.02220$), indicating that these two groups of sites are physically associated within the structure of the protein. This result does not reflect a spurious association of a cluster of selected sites with a single protease inhibitor site, as selected sites are not significantly clustered with each other ($p = 0.31839$).

DISCUSSION

Several aspects of our data suggest that the protease gene family examined here evolves non-independently as a functionally redundant complex. First, we observed ectopic recombination between five of six paralogs within this gene family. Although our data do not indicate if ectopic recombination is a source of adaptive genetic variation, in many cases conversion tracts were segregating at intermediate or high frequency, indicating that these mutations are not significantly deleterious. Considerable interchange of divergent sequence implies functional overlap between the encoded proteins.

Paralogs with partially or completely overlapping functions are expected to experience relaxed evolutionary constraint (OHNO 1970, HUGHES 1994; FORCE *et al* 1999). Consistent with this prediction, we find two indicators of relaxed constraint at three different loci in this multigene family. First, GLEANR_898 exhibits an excess of replacement polymorphism but no evidence for balancing selection or the segregation of weakly deleterious mutations. This deviation from neutrality, therefore, may indicate that

a recent relaxation in functional constraint has allowed for the accumulation of mutations that were not tolerated in the previous selective regime (TAKAHATA 1993; NACHMAN 1998). Second, we discovered three distinct pseudogene haplotpes in two different paralogs. In all three cases, the relevant mutations likely rendered the protein completely non-functional. The prevalence of such haplotypes in our sample would suggest that purifying selection is relatively weak.

Although relaxed constraint may imply these proteases have little or no important function, evidence for adaptive evolution within this gene family would suggest otherwise. Our analysis of divergence across the *repleta* species group asserts that these genes are evolving rapidly and adaptively, consistent with a critical role in organismal fitness. The Mojave Desert population, furthermore, exhibited an elevated ratio of divergence to polymorphism in GLEANR_897, as well as an excess of rare variants at the adjacent GLEANR_898, indicative of recent directional selection in this genomic region. Although we found no compelling evidence of adaptive evolution in the remaining three populations, this may reflect the limited framework for detecting deviations from neutrality in the complex scenario of multiple paralogs undergoing gene conversion (INNAN 2003; THORNTON 2007).

We propose that the observed pattern of relaxed constraint paired with positive directional selection reflects an intriguing evolutionary mechanism employed by *repleta* group females. By tolerating a larger array of genetic variation, generated by single base-pair mutations, ectopic recombination, and gene duplication, females can more rapidly explore adaptive space to generate novel advantageous variants. This strategy long has

been hypothesized to explain the complex evolutionary histories of vertebrate MHC alleles, and their role in immune response, although the empirical data remain controversial (Reviewed in MARTINSOHN *et al* 1999; NEI and ROONEY 2005). Interestingly, several single copy reproductive proteins exhibit a similar pattern of elevated polymorphism within populations, but rapid, adaptive evolution between species (SWANSON *et* al 2001; GALINDO *et al* 2003; TURNER and HOEKSTRA 2006; 2008; GASPER and SWANSON 2006; HAMM *et al* 2007; MOY *et al* 2008).

Mathematical models of sexual conflict predict that females can halt the evolutionary chase of a male interactor by splitting into two divergent haplogroups (GAVRILETS and WAXMAN 2002, HAYASHI *et al* 2007). Although our data provide no compelling evidence of balancing selection, it is easy to envision how a complex of paralogs that duplicate and recombine could be adaptive in the context of sexually antagonistic coevolution. Determining the relative roles of sexual conflict and cryptic female choice in shaping the evolutionary history of the proteases examined here, however, will require a significantly more detailed understanding of their biochemical and physiological functions.

If the gene family examined here is engaged in an evolutionary dynamic with components of the male ejaculate, its history within populations is expected to be a unique reflection of this coevolutionary trajectory. Consistent with this prediction, the patterns of pseudogenation, duplication, gene conversion, and adaptive evolution exhibited by the female reproductive proteases examined in this study are largely population-specific. Ectopic recombination between GLEANR_896 and GLEANR_897

is biased in opposite directions between the Mainland Sonora and Catalina Island populations. Pseudogene haplotypes, and acquisition of a novel paralog also were confined to a single population. Finally, all deviations from neutrality were population-specific, as predicted if the selective pressure experienced by this gene family is determined by a distinct intersexual dynamic.

The identities of male interactors for the female proteases examined here remain obscure, however, it is intriguing that positively selected sites in this gene family are significantly associated with residues known to determine protease inhibitor susceptibility (Reviewed in SRINIVASAN *et al* 2006). Protease inhibitors are found in the male ejaculates of both *D. mojavensis* (WAGSTAFF and BEGUN 2005), and *D. melanogaster* (SWANSON *et al* 2001; FINDLAY *et al* 2008). Consistent with the hypothesis that male protease inhibitors regulate female proteases, trypsin and elastase-like serine endoprotease activity in *D. arizonae* female reproductive tracts is observed to decrease after mating (KELLEHER and PENNINGTON, *submitted*). Adaptive evolution of female proteases, therefore, may reflect molecular coevolution with protease inhibitors in the male ejaculate, as previously suggested for *D. melanogaster* reproductive proteases and inhibitors (WONG *et al* 2008).

We previously have hypothesized that the proteases examined here may play a role in the degradation of the insemination reaction in mated females (KELLEHER *et al* 2007). This opaque mass that fills the uterus after mating (PATTERSON 1946) differs in severity between the four populations of *D. mojavensis* (KNOWLES and MARKOW 2001). Male and female contributions to this process, furthermore, are thought coevolve

antagonistically between the sexes (KNOWLES and MARKOW 2001). It is exciting, therefore, that the evolutionary history of the novel paralog is correlated with insemination reaction mass size differences between populations. Specifically, the Mojave Desert population exhibits the largest reaction mass in intrapopulation crosses (KNOWLES and MARKOW 2001), as well as a gene duplication event that engendered permanent heterozygosity for the converted and unconverted alleles of GLEANR_897. This chromosomal region, furthermore, is associated with a recent selective sweep. Similarly, the Mainland Sonora and Baja Peninsula populations exhibit intermediate reaction mass sizes (KNOWLES and MARKOW 2001), and evidence of an old polymorphism between converted and unconverted GLEANR_897 haplogroups. Finally, the Catalina Island population exhibits the smallest reaction mass (KNOWLES and MARKOW 2001), and evidence for neither an old polymorphism nor a novel paralog. Future genetic studies of these proteins will shed light on their potential role in intersexual dynamics and determination of reaction mass size.

CONCLUSIONS:

Extensive research in a broad range of taxa has demonstrated that proteins involved in sexual reproduction evolve rapidly (SWANSON and VACQUIER 2002; CLARK *et al* 2006; PANHUIS *et al* 2006). The complex history exhibited by the protease gene family examined here, however, includes pseudogenation, duplication, gene conversion, and positive selection. Although many of these processes previously have been observed

in reproductive proteins (AGUADÉ 1998; CIRERA and AGUADÉ 1998; SWANSON and VACQUIER 1999), their integration in a single gene family represents a novel and intriguing observation in the study of reproductive protein evolution. The divergence of these genes between four well-structured populations of *D. mojavensis* with evidence of ejaculate-female coadaptation (KNOWLES and MARKOW 2001; PITNICK *et al* 2003; KNOWLES *et al* 2005; KELLEHER and MARKOW 2007), furthermore, suggests an exciting role for gene family evolution in the mediation of intersexual dynamics. Documenting this unique evolutionary history in a female reproductive protein highlights the under-explored "female-side" of reproductive tract interactions.

Gene duplication recently has emerged as an integral aspect of reproductive protein evolution in *Drosophila*. Lineage-specific duplicates are common amongst *Drosophila* reproductive proteins, (CIRERA and AGUADÉ 1998; LOPPIN *et al* 2005; DORUS *et al* 2008; FINDLAY *et al* 2008), particularly within the *repleta* species group (KELLEHER *et al* 2007; WAGSTAFF and BEGUN 2007; ALMEIDA and DESALLE 2008A; 2008B). Genome wide patterns of gene gain and loss across twelve *Drosophila* genomes, furthermore, indicates proteins involved in sexual reproduction turn over more rapidly than other functional classes (HAHN *et al* 2007). Finally, *D. melanogaster* genes with copy number polymorphism are enriched for proteins expressed in the male accessory gland (DOPMAN and HARTL 2007), the primary site for production of seminal fluid protein in *Drosophila* (Reviewed in WOLFNER 2002). Elucidating the role of gene family evolution in determining reproductive success, mediating intersexual dynamics, or both, presents an exciting avenue for future research.

Acknowledgements. The authors would like to acknowledge Michael Nachman and Giovanni Bosco for use of equipment and reagents, Nathan Clark for generously analyzing structural data, Tom Hartl for technical assistance, and Kevin Thornton and Matthew Dean for helpful discussion. Michael Nachman, Matthew Hahn, Luciano Matzkin, Willie Swanson, and the members of the Nachman Lab provided helpful comments that significantly improved the manuscript. This research was funded by a National Science Foundation Doctoral Dissertation Improvement Grant to E.S.K., and the Center for Insect Science at the University of Arizona. E.S.K. was supported by an NSF-IGERT research traineeship in Evolutionary, Functional and Computational Genomics at the University of Arizona, and a Dissertation Fellowship from the American Association of University Women.

REFERENCES

AGUADÉ, M., 1998 Different forces drive the evolution of ACP26Aa and ACP26Ab accessory gland genes in the *Drosophila melanogaster* species complex. Genetics. **150:** 1079–1089.

ALMEIDA, F. C., and R. DESALLE, 2008A Evidence of adaptive evolution of accessory gland proteins in closely related species of the *Drosophila repleta* group. Mol. Biol. Evol. **25:** 2043–2053.

ALMEIDA, F. C., and R. DESALLE, 2008B Orthology, Function, and Evolution of Accessory Gland Proteins in *the Drosophila repleta* Group. Genetics *epub ahead of print*.

ANISIMOVA, M., R. NIELSEN, and Z. YANG, 2003 Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics **164:** 1229–1236.

BEGUN, D. J., and H. A. LINDFORS, 2005 Rapid evolution of genomic Acp complement in the melanogaster subgroup of *Drosophila*. Mol. Bio.l Evol. **22:** 2010–2021.

CIRERA, S., and M. AGUADÉ, 1998 Molecular evolution of a duplication: the sex-peptide (Acp70A) gene region of *Drosophila subobscura* and *Drosophila madeirensis*. Mol. Biol. Evol. **15:** 988–996.

CLARK N. L., and W. J. SWANSON, 2005. Pervasive adaptive evolution in primate seminal proteins. P. L. o. S. Genet.b e35.

CLARK, N. L., J. E. AAGAARD, and W. J. SWANSON, 2006 Evolution of reproductive proteins from animals and plants. Reproduction **131:** 11–22.

CLARK, N. L., G. D. FINDLAY, X.YI, M. J. MACCOSS, and W. J. SWANSON, 2007 Duplication and selection on abalone sperm lysin in an allopatric population. Mol. Biol. Evol. **24:** 2081–2090.

DOPMAN, E. B., and D. L. HARTL, 2007 A portrait of copy-number polymorphism in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U. S. A. **104:** 19920–19925.

EBERHARD, W. G., 1996 Female Control: Sexual Selection by Cryptic Female Choice.. Princeton University Press, Princeton, New Jersey.

FISHER, R. A., 1915 The evolution of sexual preference. Eugenics Review **7:**115–123.

FISHER, R. A. 1930 The genetical theory of natural selection. Clarendon Press, Oxford.

FINDLAY, G. D ., X. YI, M. J. MACCOSS, and W. J. SWANSON, 2008 Proteomics reveals novel *Drosophila* seminal fluid proteins transferred at mating. P. L. o. S. Biol. **6:** e178.

FORCE, A., M. LYNCH, F. B. PICKETT, A. AMORES, Y. L. YAN, and J. POSTLETHWAIT, 1999 Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151:** 1531–1545.

FRACZKIEWICZ, R., and W. BRAUN, 1998 Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. J. Comp. Chem. **19:** 319.

GALINDO, B. E, V. D. VACQUIER, and W. J. SWANSON, 2003 Positive selection in the egg receptor for abalone sperm lysin. Proc. Natl. Acad. Sci. U. S. A. **100**: 4639–4643.

GASPER J., and W. J. SWANSON, 2006 Molecular population genetics of the gene encoding the human fertilization protein zonadhesin reveals rapid adaptive evolution. Am. J. Hum. Genet. **79:** 82–30.

GAVRILETS, S., 2000 Rapid evolution of reproductive barriers driven by sexual conflict. Nature **403:** 886–889.

GAVRILETS, S., and D. WAXMAN, 2002 Sympatric speciation by sexual conflict. Proc. Natl. Acad. Sci. U. S. A. **99:** 10533–10538.

HAHN, M. W., M. V. HAN, and S. G. HAN, 2007 Gene family evolution across 12 *Drosophila* genomes. P. L. o. S. Genet. **3:** e197.

HAMM, D., B. S. MAUTZ, M. F. WOLFNER, C. F. AQUADRO, and W. J. SWANSON, 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.

HAYASHI, T. I., M. VOSE, and S. GAVRILETS, 2007 Genetic differentiation by sexual conflict. Evolution **61:** 516–529.

HEXTER, A., 1968 Selective advantage of the sickle-cell trait. Science **160:** 436–437.

HUDSON, R. R., M. KREITMAN, and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

HUGHES, A. L. 1994. The evolution of functionally novel proteins after gene duplication. Proc. Biol. Sci. **256:** 119–24.

INNAN, H., 2003 The coalescent and infinite-site model of a small multigene family. Genetics **163:** 803–810.

KELLEHER, E. S., and T. A. MARKOW, 2007 Reproductive Tract Interactions Contribute

to Isolation in *Drosophila*. Fly **1:** 33–37.

KELLEHER, E. S., W. J. SWANSON, and T. A. MARKOW, 2007 Gene Duplication and Adaptive Evolution of Digestive Proteases in *Drosophila* Female Reproductive Tracts. P.L.o.S. Genet. **3:** 1541–1549.

KELLY J. K., 1997 A test of neutrality based on interlocus associations. Genetics 1997 **146:** 1197–1206.

KIMURA, M., 1983 The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

KNOWLES. L. L., and T. A. MARKOW, 2001 Sexually antagonistic coevolution of a postmating prezygotic reproductive character in desert *Drosophila*. Proc. Nat. Acad. Sci. U. S. A. **98:** 8692–8696.

KNOWLES, L. L., B. B. HERNANDEZ, and T. A. MARKOW, 2005 Non-antagonistic interactions between the sexes revealed by the ecological consequences of reproductive traits. J. Evol. Biol. **18:** 156–161.

LOPPIN, B., D. LEPETIT, S. DORUS, P. COUBLE, and T. L. KARR, 2005 Origin and neofunctionalization of a *Drosophila* paternal effect gene essential for zygote viability. Curr. Biol. 2005 **15:** 87–93.

MACHADO, C. A., L. M. MATZKIN, L. K. REED, and T. A. MARKOW, 2007 Multilocus nuclear sequences reveal intra- and interspecific relationships among chromosomally polymorphic species of cactophilic *Drosophila*. Mol. Ecol. 2007 **16:** 3009–3024.

MATZKIN, L. M., 2008 The molecular basis of host adaptation in cactophilic *Drosophila*: molecular evolution of a glutathione S-transferase gene (GstD1) in *Drosophila mojavensis*. Genetics **178:** 1073–1083.

MARKOW, T. A., 1996 Evolution of *Drosophila* mating systems. Evol. Biol. **29:** 73–106.

MARTINSOHN, J. T., A. B. SOUSA, L. A. GUETHLEIN, and J. C. HOWARD, 1999 The gene conversion hypothesis of MHC evolution: a review. Immunogenetics **50:** 168–200.

MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive protein evolution at the Adh locus in *Drosophila*. Nature **351:** 652-654.

MOY, G. W., S. A. SPRINGER, S. L. ADAMS, W. J. SWANSON, and V.D. VACQUIER, 2008 Extraordinary intraspecific diversity in oyster sperm bindin. Proc. Natl. Acad. Sci. U. S. A. **105:** 1993–8.

NACHMAN, M. W., 1998 Deleterious mutations in animal mitochondrial DNA. Genetica **102/103:** 61–69.

NEI, M., and A. P. ROONEY, 2005 Concerted and birth-and-death evolution of multigene families. Annu. Rev. Genet. **39:** 121–52.

NEUBAUM, D. M., and M. F. WOLFNER, 1999 Wise, winsome, or weird? Mechanisms of sperm storage in female animals. Curr. Top. Dev. Biol. **41:** 67–97.

OHNO, S, 1970Evolution by gene duplication. Springer-Verlag, New York.

OTTO, S. P., and P. YONG, 2002 The evolution of gene duplicates. Adv. Genet. **46:** 451–83.

PARKER, G. A., 1979 Sexual selection and sexual conflict, pp. 123–166 in *Sexual selection and reproductive competition in insects*, edited by M. S. Blum and N. A. Blum. Academic Press, London, U. K.

PANHUIS, T. M., and W. J. SWANSON, 2006 Molecular evolution and population genetic analysis of candidate female reproductive genes in *Drosophila*. Genetics **173:** 2039–2047.

PANHUIS, T. M., N. L. CLARK, and W. J. SWANSON, 2006 Rapid evolution of reproductive proteins in abalone and *Drosophila*. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **361:** 261–268.

PERONA, J. J., and C. S. CRAIK, 1995 Structural basis of substrate specificity in the serine proteases. Protein Sci. **4:** 337–360.

PITNICK, S., G. T. MILLER, K. SCHNEIDER, and T. A. MARKOW, 2003 Ejaculate-female coevolution in *Drosophila mojavensis*. Proc. Nat. Acad. Sci. U. S. A. **270:** 507–512.

POLGAR, L., 2005 The catalytic triad of serine peptidases. Cell. Mol. Life. Sci. **62:** 2161–2172.

PROULX, S. R., and R. C. PHILLIPS., 2006 Allelic divergence precedes and promotes gene duplication. Evolution **60:** 881–892.

PROKUPEK, A, F. HOFFMANN, S. I. EYUN, E. MORIYAMA, M. ZHOU, AND L. HARSHMAN 2008 An evolutionary expressed sequence tag analysis of *Drosophila* spermatheca genes. Evolution. **62:** 2936–2947.

REED, L. K., M. NYBOER, and T. A. MARKOW, 2007 Evolutionary relationships of *Drosophila mojavensis* geographic host races and their sister species *Drosophila arizonae*. Mol. Ecol. **16:** 1007–22.

RICE, W. R., 1996 Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. Nature **381:** 232–234.

ROBERTSON, S. A., 2007 Seminal fluid signaling in the female reproductive tract: lessons from rodents and pigs. J. Anim, Sci. **85:** E36–44.

ROZAS, J., and R. ROZAS, 1995 DnaSP, DNA sequence polymorphism: an interactive program for estimating Population Genetics parameters from DNA sequence data. Comput. Applic. Biosci. **11:** 621–625.

SAITOU N, and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

SAWYER, S. A., 1989 Statistical tests for detecting gene conversion. Mol. Biol. Evol. **6:** 526–538.

SCHWEDE, T., J. KOPP, N. GUEX, and M. C. PEITSCH, 2003 SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res **31:** 3381**–**3385.

SPOFFORD, J. B., 1969 Heterosis and evolution of duplications. Am. Nat. **103:** 407–432.

SPRANG, S. R., R. J. FLETTERICK, L. GRAF, W. J. RUTTER, C. S. CRAIK, 1988 Studies of specificity and catalysis in trypsin by structural analysis of site-directed mutants. Crit. Rev. Biotechnol. **8:** 225–236.

SRINIVASAN, A., A. P. GIRI, and V. S. GUPTA, 2006 Structural and functional diversities in *lepidopteran* serine proteases. Cell. Mol. Biol. Lett. **11:** 132–54.

STEDMAN, H. H., B. W. KOZYAK, A. NELSON, D. M. THESIER, L. T. SU *et al* 2004 Myosin gene mutation correlates with anatomical changes in the human lineage. Nature **428:** 415–418.

SWANSON, W. J., and V. D. VACQUIER, 1998 Concerted evolution in an egg receptor for a rapidly evolving abalone sperm protein. **281:** 710–712.

SWANSON, W. J., and V. D. VACQUIER, 2002 The rapid evolution of reproductive proteins. Nat. Rev. Genet. **3:** 137–144.

SWANSON W. J., C. F. AQUADRO, and V. D. VACQUIER, 2001 Polymorphism in abalone fertilization proteins is consistent with the neutral evolution of the egg's receptor for lysin (VERL) and positive darwinian selection of sperm lysin. Mol Biol Evol. **18:** 376–83.

SWANSON, W. J., A. WONG, M. F. WOLFNER, and C. F. AQUADRO, 2004 Evolutionary expressed sequence tag analysis of *Drosophila* female reproductive tracts identifies genes subjected to positive selection. Genetics **168:**1457–1465.

SWOFFORD, D. L., 2000 PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–95.

TAKAHATA, N., 1993 Relaxed natural selection in human populations during the Pleistocene. Jpn. J. Genet. **68:** 539–47.

THORNTON, K. R., 2007 The neutral coalescent process for recent gene duplications and copy-number variants. Genetics **177:** 987–1000.

THORNTON, K. and M. LONG, 2005 Excess of amino acid substitutions relative to polymorphism between X-linked duplications in *Drosophila melanogaster*. Mol. Biol. Evol **22:** 273-284.

TURNER L. M., and H. E. HOEKSTRA, 2006 Adaptive evolution of fertilization proteins within a genus: variation in ZP2 and ZP3 in deer mice (Peromyscus). Mol. Biol. Evol**. 23:** 1656–69.

TURNER L. M., and H. E. HOEKSTRA, 2008 Reproductive protein evolution within and between species: maintenance of divergent ZP3 alleles in Peromyscus. Mol. Ecol. **17:** 2616–28.

WAGSTAFF B. J., and D. J. BEGUN, 2005 Molecular population genetics of accessory gland protein genes and testis-expressed genes in *Drosophila mojavensis* and *D. arizonae.* Genetics **171:** 1083–1010.

WAGSTAFF, B. J., and D. J. BEGUN, 2007 Adaptive evolution of recently duplicated accessory gland protein genes in desert *Drosophila.* Genetics **177:** 1023–1030.

WANG, X., W. E. GRUS, and J. ZHANG, 2006 Gene losses during human origins. P. L. o. S. Biol. 2006 **4:** e52.

WOLFNER, M. F. 2002 The gifts that keep on giving: physiological functions and evolutionary dynamics of male seminal proteins in *Drosophila*. Heredity **88:** 85–93.

WOLFNER, M. F., 2007 "S.P.E.R.M." (seminal proteins (are) essential reproductive modulators): the view from *Drosophila*. Soc. Reprod. Fertil. **Suppl:** 183–199.

WONG, A., M. C. TURCHIN, M. F. WOLFNER, C. F. AQUADRO, 2008 Evidence for positive selection on *Drosophila melanogaster* seminal fluid protease homologs. Mol. Biol. Evol. **25:** 497–506.

WIESENFELD, S. L., 1968 Selective advantage of the sickle-cell trait. Science. **160:** 437.

YANG, Z., 1997 PAML: A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13:** 555–556.

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

YANG, Z., W. S. WONG, and R. NIELSEN, 2005 Bayes empirical bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. **22:** 1107–1118.

TABLES

population	*donor*	*recipient*	*Nc*	*Nnc*	*π*	*SD(π)*	*π(nc)*
Baja Peninsula	GLEANR_898	GLEANR_897	7	5	0.0689	0.0127	0.0553
Baja Peninsula	GLEANR_896	GLEANR_897	4	8	0.0689	0.0127	0.0110***
Baja Peninsula	GLEANR_897	GLEANR_896	7	6	0.0787	0.0082	0.01378***
Mainland Sonora	GLEANR_2575	GLEANR_898	1	6	0.0136	0.0026	0.0108
Mainland Sonora	GLEANR_897	GLEANR_898	1	11	0.0564	0.0180	0.0453
Mainland Sonora	GLEANR_897	GLEANR_896	9	12	0.0564	0.0180	0.0240
Catalina Island	GLEANR_2575	GLEANR_898	1	6	0.0085	0.0039	0.0000*
Catalina Island	GLEANR_897	GLEANR_896	5	2	0.0725	0.0238	0.00673*

Table 1. Ectopic Recombination Contributes to Genetic Variation. For individual paralogs, nucleotide diversity was estimated for the complete data set (π), as well as for the data set with conversion alleles excluded (π(nc)). Nc = the number of sampled recipient conversion alleles. Nnc = the number of sampled alleles that were not recipients of gene conversions. * denotes greater than two standard deviations below π. ** denotes greater than 3 standard deviations below π. *** denotes greater than 4 standard deviations below π.

		Standard MK Test			*Standard MK Test* (no conversion)			
		Polymorphic	***Fixed***	***Test***	***Polymorphic***	***Fixed***	***Test***	***Tajima's D***
Baja Peninsula	Syn.+nc	9	35	*G-test*				-0.69
	Non-Syn.	13	12	**				NS
Catalina Island	Syn.+nc	14	31	*G-test*	0	15	NA	-1.69
	Non-Syn.	13	10	*	0	7		**
Mojave Desert	Syn.+nc	6	35	*G-test*				-1.43
	Non-Syn.	4	12	NS				*
Mainland Sonora	Syn.+nc	12	31	*G-test*	3	17	*G-test*	-0.10
	Non-Syn.	18	12	**	13	7	**	NS

Table 2. Deviations from Neutrality in GLEANR_898. McDonald-Kreitman tests utilized *D. arizonae* as an outgroup. Lineage-specific McDonald-Kreitman tests were polarized with Dmoj\GLEANR_2575. * denotes $p < .05$. ** denotes $p<.01$.

	Tajima's D	
	Baja Peninsula	***Mainland Sonora***
All	-0.69	-0.10
Silent	-0.88	0.70
Replacement	-0.48	-0.35

Table 3. Estimates of *Tajima's* D for Silent and Replacement Sites in GLEANR_898.

haplogroup	population	length converted region	ancestral *S*	shared *S*	converted *S*	% ancestral shared	% converted shared
GLEANR_897 converted	Baja Peninsula	443 bp	20	8	3	28.57%	72.73%
GLEANR_897 converted	Mainland Sonora	443 bp	22	1	4	4.30%	20.00%
GLEANR_896 converted	Baja Peninsula	518 bp	3	6	2	66.67%	80.00%
GLEANR_896 converted	Catalina Island	518 bp	1	2	1	66.67%	66.67%

Table 4. Private and Shared Polymorphisms in Ancestral and Converted Haplogroups of GLEANR_896 and GLEANR_897. Ancestral *S* = private polymorphisms in the ancestral (or donor) haplogroup in the converted region. For GLEANR_897 converted, the ancestral haplogroup is GLEANR_896 ancestral, and for GLEANR_896 the ancestral haplogroup is GLEANR_897 ancestral. Shared *S* = number of shared polymorphisms between ancestral and converted haplogroups. Converted *S* = private polymorphisms in the converted (or recipient) haplogroup within the converted region. % ancestral shared = the percentage of polymorphisms in the ancestral haplogroup that are shared with the converted haplogroup. % converted shared = the percentage of polymorphisms in the converted haplogroup that are shared with the ancestral haplogroup.

population	locus	inheritance	intraspecific length	*S*	interspecific length	D	*Tajima's D*
Catalina Island	996†	autosomal	856	2	827	38.50	-0.71
	5239†	autosomal	870	1	870	13.25	-0.61
	5246†	autosomal	872	0	849	22.00	NA
	A4125†	autosomal	880	4	871	49.00	-0.78
	X100†	sex-linked	875	0	849	46.00	NA
	GLEANR_898	autosomal	710	21	710	54.29	-1.69**
	GLEANR_897	autosomal	682	34	682	88.28	-1.45*
					neutral $X^2 = 2.77$		p = 0.54
					neutral + 898 $X^2 = 8.8493$		p = 0.12
					neutral + 897 $X^2 = 8.3066$		p = 0.14
					neutral + 898 + 897 $X^2 = 9.23$		p = 0.16
Mojave Desert	996†	autosomal	856	1	827	40.25	-0.61
	1343†	autosomal	886	1	869	9.25	-0.61
	5239†	autosomal	870	3	870	14.75	-0.75
	5246†	autosomal	870	1	850	16.25	-0.61
	A4115†	autosomal	824	2	824	16.50	-0.71
	A4125†	autosomal	917	4	908	48.00	0.65
	X100†	sex-linked	911	3	890	47.50	0.17
	GLEANR_898	autosomal	710	4	710	49.22	-1.43*
	GLEANR_897	autosomal	691	0	691	100.91	NA
	GLEANR_896	autosomal	697	0	697	75.00	NA
					neutral $X^2 = 2.59$		p = 0.84
					neutral + 898 $X^2 = 3.10$		p = 0.87
					neutral + 897 $X^2 = 14.06$		p = 0.05
					neutral + 896 $X^2 = 11.21$		p = 0.13
					neutral + 898 + 897 $X^2 = 22.40$		p = 0.008

Table 5. HKA and Site-Frequency Spectra Analysis of GLEANR_898, GLEANR_897, and GLEANR_896. *S* = number of segregating sites. *D* = Divergence from *D. arizonae* ortholog. X^2 and p-values for multiple HKA tests performed in HKA (http://lifesci.rutgers.edu/~heylab/heylabsoftware.htm#HKA) are reported. * denotes p < .05. ** denotes p<.01. † denotes sequences from MACHADO *et al* (2007).

Data Set	*M1*	*M2*	*LRT*	*p*	*P(s)*	*ω*	*BEB selected sites*
full alignment	-7424.47	-7388.35	72.24	1.91E-17	0.08	2.80	68, 132, 133, 135, 253
exclude *D. mettleri-1*, *D. mettleri-2*	-6642.72	-6619.94	45.55	1.49E-11	0.07	2.63	68, 132, 133, 135, 253
exclude *D. mettleri-1*, *D. mettleri-2*, Dmoj\GLEANR_897, *D. arizonae-4, D. arizonae-5,* Dvir\GLEANR_2181	-5726.11	-5703.86	44.49	2.55E-11	0.08	2.79	68, 132, 133, 135, 253
exclude *D. mettleri-1*, *D. mettleri-2*, Dmoj\GLEANR_896, Dmoj\GLEANR_898, Dmoj\NEW_PARALOG, *D.arizonae-1, D. arizonae-2, D.arizonae-5, D. mayaguana-1*	-5265.49	-5258.06	14.88	1.15E-04	0.03	3.02	253

Data Set	*M7*	*M8*	*LRT*	*p*	*P(s)*	*ω*	*BEB selected sites*
full alignment	-7426.01	-7376.89	98.23	3.72E-23	0.13	2.11	68, 73, 112, 113, 132, 133, 135, 179, 184, 187, 204, 208, 211, 209, 253
exclude *D. mettleri-1*, *D. mettleri-2*	-6643.22	-6608.53	69.37	8.16E-17	0.13	1.92	68, 73, 112, 132, 133, 135, 179, 187, 204, 253
exclude *D. mettleri-1*, *D. mettleri-2*, Dmoj\GLEANR_897, *D. arizonae-4, D. arizonae-5,* Dvir\GLEANR_2181	-5734.58	-5698.60	71.96	2.19E-17	0.14	2.08	68, 73, 112, 113, 132, 133, 135, 179, 187, 204, 208, 209, 253, 257
exclude *D. mettleri-1*, *D. mettleri-2*, Dmoj\GLEANR_896, Dmoj\GLEANR_898, Dmoj\NEW_PARALOG, *D.arizonae-1, D. arizonae-2, D.arizonae-5, D. mayaguana-1*	-5269.09	-5253.85	30.48	3.38E-08	0.11	1.70	68, 112, 113, 133, 135, 179, 187, 253

Table 6. Maximum-likelihood Codon-Based Analysis of Positive Selection in the *Repleta* Species Group. M1, M2, M7 and M8 denote codon models implemented in PAML (YANG 1997). LRT = the value of the likelihood ratio test between nested models. *p* = the probability of the LRT under a chi-square distribution. *P(s)* = proportion of sites in the positively selected site class. *ω* = estimated *dN/dS* of the positively-selected site class. BEB selected sites = Bayes Empirical Bayes predicted selected sites for the given selection model (YANG *et al* 2005).

population	*donor*	*recipient*	*Nc*	*Nnc*	*π*	*SD(π)*	*π(nc)*
Baja Peninsula	GLEANR_898	GLEANR_897	7	5	0.0689	0.0127	0.0553
Baja Peninsula	GLEANR_896	GLEANR_897	4	8	0.0689	0.0127	0.0110***
Baja Peninsula	GLEANR_897	GLEANR_896	7	6	0.0787	0.0082	0.01378***
Mainland Sonora	GLEANR_2575	GLEANR_898	1	6	0.0136	0.0026	0.0108
Mainland Sonora	GLEANR_897	GLEANR_898	1	11	0.0564	0.0180	0.0453
Mainland Sonora	GLEANR_897	GLEANR_896	9	12	0.0564	0.0180	0.0240
Catalina Island	GLEANR_2575	GLEANR_898	1	6	0.0085	0.0039	0.0000*
Catalina Island	GLEANR_897	GLEANR_896	5	2	0.0725	0.0238	0.00673*

Table 1. Ectopic Recombination Contributes to Genetic Variation. For individual paralogs, nucleotide diversity was estimated for the complete data set (π), as well as for the data set with conversion alleles excluded (π(nc)). Nc = the number of sampled recipient conversion alleles. Nnc = the number of sampled alleles that were not recipients of gene conversions. * denotes greater than two standard deviations below π. ** denotes greater than 3 standard deviations below π. *** denotes greater than 4 standard deviations below π.

		Standard MK Test			*Standard MK Test* (no conversion)			
		Polymorphic	***Fixed***	***Test***	***Polymorphic***	***Fixed***	***Test***	***Tajima's D***
Baja Peninsula	Syn.+nc	9	35	*G-test*				-0.69
	Non-Syn.	13	12	**				NS
Catalina Island	Syn.+nc	14	31	*G-test*	0	15	NA	-1.69
	Non-Syn.	13	10	*	0	7		**
Mojave Desert	Syn.+nc	6	35	*G-test*				-1.43
	Non-Syn.	4	12	NS				*
Mainland Sonora	Syn.+nc	12	31	*G-test*	3	17	*G-test*	-0.10
	Non-Syn.	18	12	**	13	7	**	NS

Table 2. Deviations from Neutrality in GLEANR_898. McDonald-Kreitman tests utilized *D. arizonae* as an outgroup. Lineage-specific McDonald-Kreitman tests were polarized with Dmoj\GLEANR_2575. * denotes $p < .05$. ** denotes $p<.01$.

	Tajima's D	
	Baja Peninsula	***Mainland Sonora***
All	-0.69	-0.10
Silent	-0.88	0.70
Replacement	-0.48	-0.35

Table 3. Estimates of *Tajima's* D for Silent and Replacement Sites in GLEANR_898.

haplogroup	population	length converted region	ancestral *S*	shared *S*	converted *S*	% ancestral shared	% converted shared
GLEANR_897 converted	Baja Peninsula	443 bp	20	8	3	28.57%	72.73%
GLEANR_897 converted	Mainland Sonora	443 bp	22	1	4	4.30%	20.00%
GLEANR_896 converted	Baja Peninsula	518 bp	3	6	2	66.67%	80.00%
GLEANR_896 converted	Catalina Island	518 bp	1	2	1	66.67%	66.67%

Table 4. Private and Shared Polymorphisms in Ancestral and Converted Haplogroups of GLEANR_896 and GLEANR_897. Ancestral *S* = private polymorphisms in the ancestral (or donor) haplogroup in the converted region. For GLEANR_897 converted, the ancestral haplogroup is GLEANR_896 ancestral, and for GLEANR_896 the ancestral haplogroup is GLEANR_897 ancestral. Shared *S* = number of shared polymorphisms between ancestral and converted haplogroups. Converted *S* = private polymorphisms in the converted (or recipient) haplogroup within the converted region. % ancestral shared = the percentage of polymorphisms in the ancestral haplogroup that are shared with the converted haplogroup. % converted shared = the percentage of polymorphisms in the converted haplogroup that are shared with the ancestral haplogroup.

population	locus	inheritance	intraspecific length	*S*	interspecific length	D	*Tajima's D*
Catalina Island	996†	autosomal	856	2	827	38.50	-0.71
	5239†	autosomal	870	1	870	13.25	-0.61
	5246†	autosomal	872	0	849	22.00	NA
	A4125†	autosomal	880	4	871	49.00	-0.78
	X100†	sex-linked	875	0	849	46.00	NA
	GLEANR_898	autosomal	710	21	710	54.29	-1.69**
	GLEANR_897	autosomal	682	34	682	88.28	-1.45*
					neutral $X^2 = 2.77$		p = 0.54
					neutral + 898 $X^2 = 8.8493$		p = 0.12
					neutral + 897 $X^2 = 8.3066$		p = 0.14
					neutral + 898 + 897 $X^2 = 9.23$		p = 0.16
Mojave Desert	996†	autosomal	856	1	827	40.25	-0.61
	1343†	autosomal	886	1	869	9.25	-0.61
	5239†	autosomal	870	3	870	14.75	-0.75
	5246†	autosomal	870	1	850	16.25	-0.61
	A4115†	autosomal	824	2	824	16.50	-0.71
	A4125†	autosomal	917	4	908	48.00	0.65
	X100†	sex-linked	911	3	890	47.50	0.17
	GLEANR_898	autosomal	710	4	710	49.22	-1.43*
	GLEANR_897	autosomal	691	0	691	100.91	NA
	GLEANR_896	autosomal	697	0	697	75.00	NA
					neutral $X^2 = 2.59$		p = 0.84
					neutral + 898 $X^2 = 3.10$		p = 0.87
					neutral + 897 $X^2 = 14.06$		p = 0.05
					neutral + 896 $X^2 = 11.21$		p = 0.13
					neutral + 898 + 897 $X^2 = 22.40$		p = 0.008

Table 5. HKA and Site-Frequency Spectra Analysis of GLEANR_898, GLEANR_897, and GLEANR_896. *S* = number of segregating sites. *D* = Divergence from *D. arizonae* ortholog. X^2 and p-values for multiple HKA tests performed in HKA (http://lifesci.rutgers.edu/~heylab/heylabsoftware.htm#HKA) are reported. * denotes p < .05. ** denotes p<.01. † denotes sequences from MACHADO *et al* (2007).

Data Set	*M1*	*M2*	*LRT*	*p*	*P(s)*	*ω*	*BEB selected sites*
full alignment	-7424.47	-7388.35	72.24	1.91E-17	0.08	2.80	68, 132, 133, 135, 253
exclude *D. mettleri-1*, *D. mettleri-2*	-6642.72	-6619.94	45.55	1.49E-11	0.07	2.63	68, 132, 133, 135, 253
exclude *D. mettleri-1*, *D. mettleri-2*, Dmoj\GLEANR_897, *D. arizonae-4, D. arizonae-5,* Dvir\GLEANR_2181	-5726.11	-5703.86	44.49	2.55E-11	0.08	2.79	68, 132, 133, 135, 253
exclude *D. mettleri-1*, *D. mettleri-2*, Dmoj\GLEANR_896, Dmoj\GLEANR_898, Dmoj\NEW_PARALOG, *D.arizonae-1, D. arizonae-2, D.arizonae-5, D. mayaguana-1*	-5265.49	-5258.06	14.88	1.15E-04	0.03	3.02	253

Data Set	*M7*	*M8*	*LRT*	*p*	*P(s)*	*ω*	*BEB selected sites*
full alignment	-7426.01	-7376.89	98.23	3.72E-23	0.13	2.11	68, 73, 112, 113, 132, 133, 135, 179, 184, 187, 204, 208, 211, 209, 253
exclude *D. mettleri-1*, *D. mettleri-2*	-6643.22	-6608.53	69.37	8.16E-17	0.13	1.92	68, 73, 112, 132, 133, 135, 179, 187, 204, 253
exclude *D. mettleri-1*, *D. mettleri-2*, Dmoj\GLEANR_897, *D. arizonae-4, D. arizonae-5,* Dvir\GLEANR_2181	-5734.58	-5698.60	71.96	2.19E-17	0.14	2.08	68, 73, 112, 113, 132, 133, 135, 179, 187, 204, 208, 209, 253, 257
exclude *D. mettleri-1*, *D. mettleri-2*, Dmoj\GLEANR_896, Dmoj\GLEANR_898, Dmoj\NEW_PARALOG, *D.arizonae-1, D. arizonae-2, D.arizonae-5, D. mayaguana-1*	-5269.09	-5253.85	30.48	3.38E-08	0.11	1.70	68, 112, 113, 133, 135, 179, 187, 253

Table 6. Maximum-likelihood Codon-Based Analysis of Positive Selection in the *Repleta* Species Group. M1, M2, M7 and M8 denote codon models implemented in PAML (YANG 1997). LRT = the value of the likelihood ratio test between nested models. *p* = the probability of the LRT under a chi-square distribution. *P(s)* = proportion of sites in the positively selected site class. *ω* = estimated *dN/dS* of the positively-selected site class. BEB selected sites = Bayes Empirical Bayes predicted selected sites for the given selection model (YANG *et al* 2005).

FIGURES

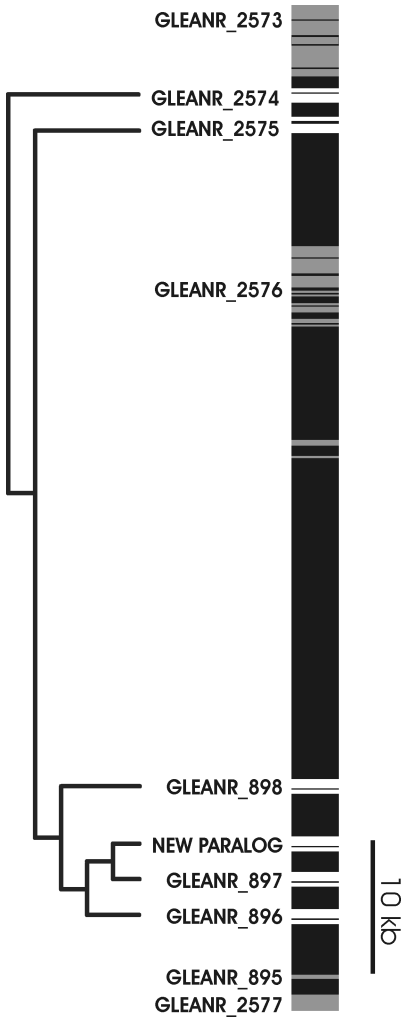


Figure 1. Genomic arrangement of the female reproductive tract protease gene family examined in this study. The exon structure of 6 paralogs (white) and neighboring coding sequences (grey) are indicated along an ~50kb region of *D. mojavensis* chromosome 3. The position of novel paralog, not present in the sequenced strain of *D. mojavensis*, was determined by PCR. Scale is indicated by 10kb size marker.

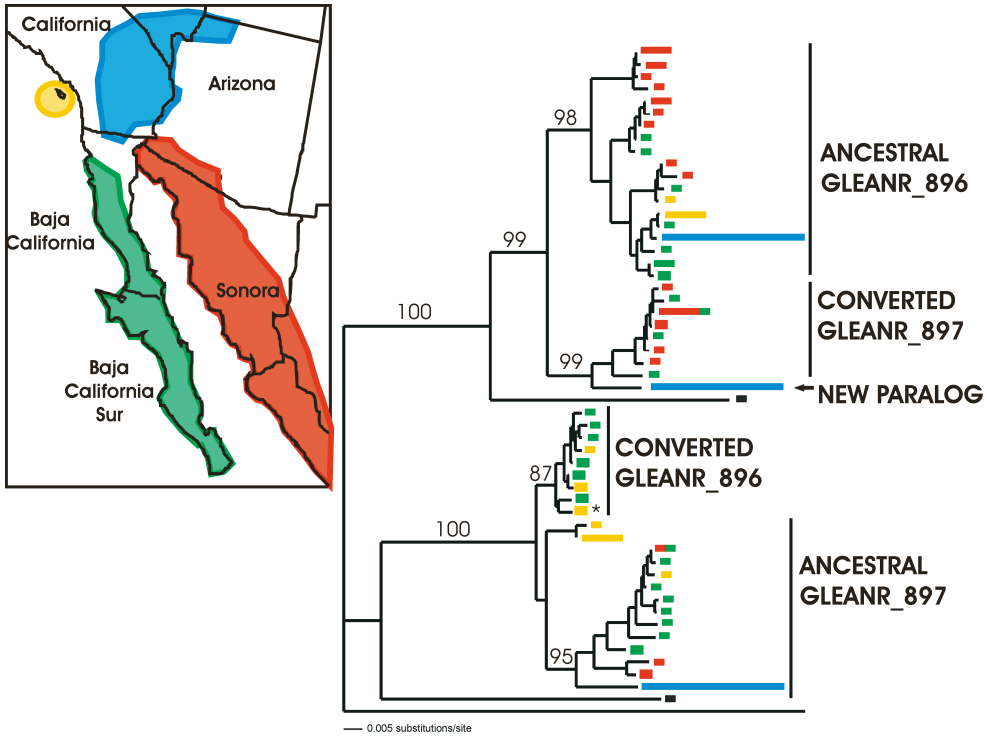


Figure 2. Neighbor-joining analysis sampled GLEANR_896, GLEANR_897, and new paralog haplotypes. Bar length indicates number of sampled individuals corresponding to each haplotype, and bar color is indicative of geographic locality. * denotes a GLEANR_897 ancestral allele that does not group with the remainder of its haplogroup. Neighbor-joining bootstrap values are indicated above the relevant branch.

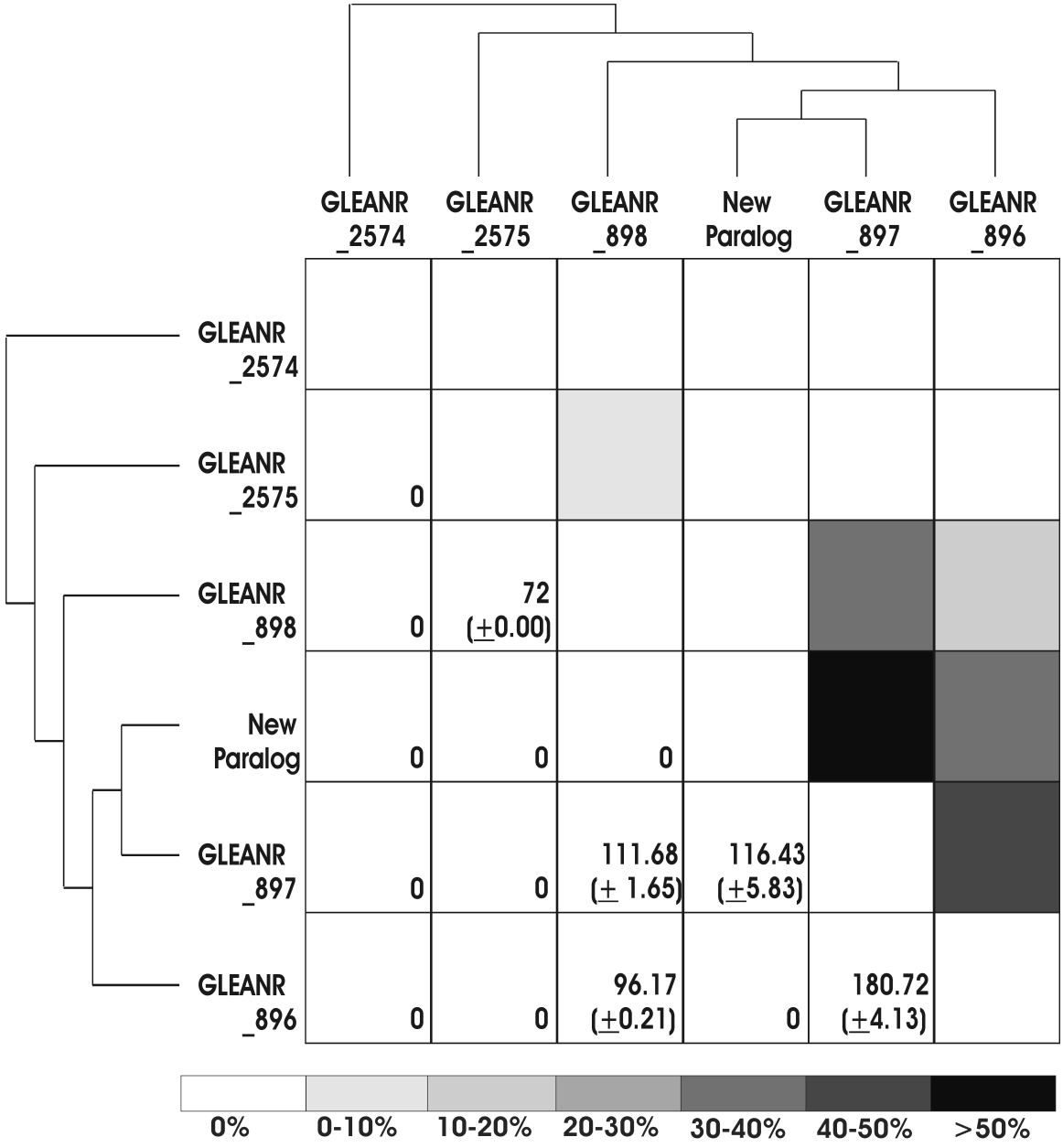


Figure 3. Ectopic Recombination. An alignment of all unique haplotypes was used to detect significant fragments of complete identity in GENECONV, based on the method of SAWYER (1989). Branching relationships are from KELLEHER *et al* (2007) and this publication. Note that there is some ambiguity concerning the placement of GLEANR_897. The percentage of pairwise comparisons between paralogs that show evidence of gene conversion is indicated by gray shading in the boxes above the diagonal. The average length of identified conversion tracts between paralogs, and the standard deviation of this estimate, are indicated in the boxes below the diagonal.

Figure 4. Directional Gene Conversion in GLEANR_2575, GLEANR_898 and GLEANR_897. Black boxes denote individual haplotypes of donor paralogs. Grey boxes

indicate unconverted haplotypes of the recipient paralog. * denotes converted haplotypes of the recipient paralog. A) Neighbor joining analysis of GLEANR_897 (grey) and GLEANR_898 (black) haplotypes, excluding a 52 bp gene conversion tract. B) Neighbor joining analysis of GLEANR_897 (grey) and GLEANR_898 (black) haplotypes, including only a 52 bp gene conversion tract. C) Neighbor joining analysis of GLEANR_898 (grey) and GLEANR_2575 (black) haplotypes, excluding a 72 bp gene conversion tract. D) Neighbor joining analysis of GLEANR_898 (grey) and GLEANR_2575 (black) haplotypes, including only a 72 bp gene conversion tract.

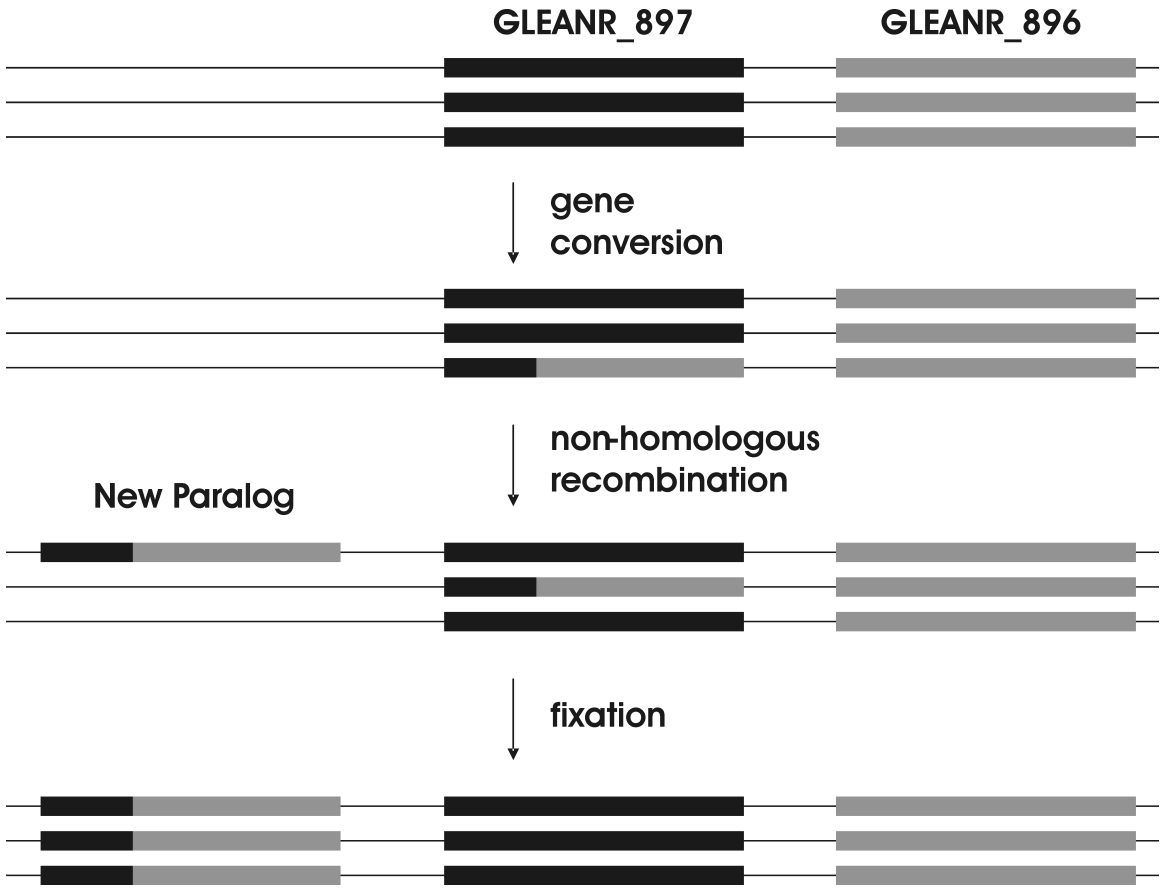


Figure 5. Hypothesized mechanism for the origin of a new paralog in the Mojave Desert population. Two tandem duplicates, GLEANR_897 (black) and GLEANR_896 (grey) are indicated on a chromosome. A gene conversion events results in a novel allele GLEANR_897. Unequal crossing over between and ancestral and converted chromosome then results in a novel tandem duplicate. The duplicated chromosome is then fixed in the Mojave Desert.

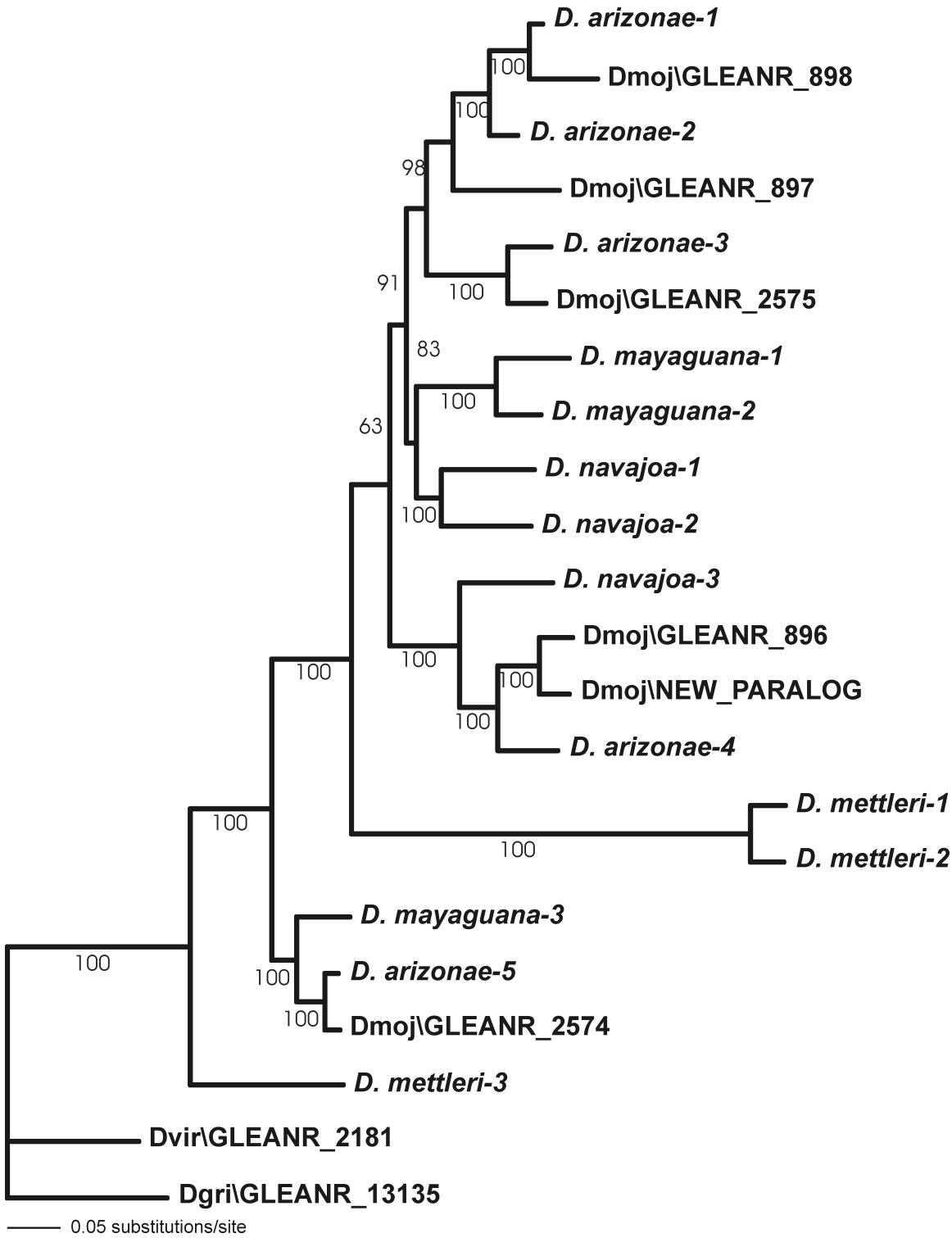


Figure 6. Bayesian phylogeny of 22 orthologs and paralogs from 7 *Drosophila* species. Posterior probabilities are indicated.

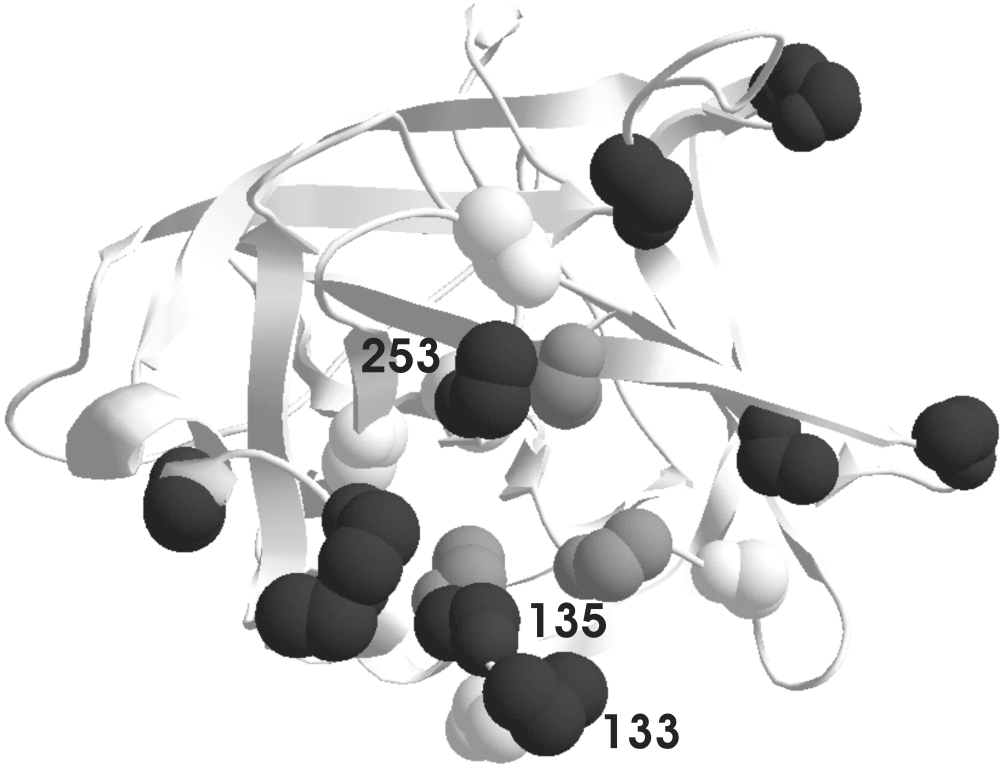


Figure 7. Predicted 3D Structure of GLEANR_898. Bayes Empirical Bayes selected sites identified under M8 (YANG 1997; YANG *et al* 2005) identified with at least two data sets are shown in black. Sites that are determinants of protease inhibitor susceptibility are shown in white (Reviewed in SRINIVASAN *et al* 2006). Sites in grey comprise the catalytic triad (Reviewed in POLGAR 2005). Selected sites 133, 135, and 253 also are determinants of inhibitor susceptibility (SRINIVASAN *et al* 2006).

SUPPLEMENTARY DATA

				synonymous			*non-synonymous*			*synonymous and non-coding*		
locus	***population***	***N***	***L***	***S***	***π***	***θ***	***S***	***π***	***θ***	***S***	***π***	***θ***
GLEANR_2574												
	all populations	31	678	10	0.0267	0.0175	2	0.0010	0.0011	12	0.0238	0.0147
	Baja Peninsula	8	678	10	0.0317	0.0270	2	0.0014	0.0016	12	0.0247	0.0227
	Catalina Island	7	678	5	0.0113	0.0143	0	0.0000	0.0000	7	0.0108	0.0140
	Mainland Sonora	8	678	8	0.0238	0.0217	2	0.0020	0.0025	9	0.0184	0.0171
	Mojave Desert	8	678	2	0.0075	0.0054	0	0.0000	0.0000	4	0.0105	0.0078
GLEANR_2575												
	all populations	30	716	12	0.0080	0.0198	19	0.0075	0.0096	15	0.0086	0.0179
	Baja Peninsula	7	716	10	0.0230	0.0267	12	0.0088	0.0098	12	0.0202	0.0231
	Catalina Island	7	748	1	0.0018	0.0025	2	0.0016	0.0016	1	0.0013	0.0019
	Mainland Sonora	8	745	1	0.0016	0.0024	5	0.0027	0.0037	1	0.0011	0.0018
	Mojave Desert	8	748	1	0.0033	0.0024	3	0.0025	0.0022	1	0.0024	0.0017
GLEANR_896												
	all populations	47	698	40	0.0656	0.0536	50	0.0462	0.0254	50	0.0745	0.0590
	Baja Peninsula	13	721	35	0.0979	0.0741	51	0.0646	0.0328	52	0.1128	0.0786
	Catalina Island	7	713	27	0.0833	0.0817	46	0.0596	0.0412	44	0.0986	0.0917
	Mainland Sonora	13	725	8	0.0167	0.0169	17	0.0150	0.0108	10	0.0145	0.0151
	Mojave Desert	14	700	0	0.0000	0.0000	0	0.0000	0.0000	0	0.0000	0.0000
GLEANR_897												
	all populations	45	630	36	0.0946	0.0709	67	0.0738	0.0383	54	0.1027	0.0698
	Baja Peninsula	12	749	33	0.0848	0.0743	59	0.0583	0.0388	50	0.0998	0.0796
	Catalina Island	7	683	10	0.0253	0.0300	20	0.0130	0.0186	10	0.0174	0.0206
	Mainland Sonora	12	711	32	0.0648	0.0747	54	0.0472	0.0364	50	0.0798	0.0816
	Mojave Desert	14	692	0	0.0000	0.0000	0	0.0000	0.0000	0	0.0000	0.0000
GLEANR_898												
	all populations	28	710	12	0.0160	0.0205	26	0.0109	0.0134	13	0.0117	0.0158
	Baja Peninsula	7	710	4	0.0088	0.0108	11	0.0083	0.0090	5	0.0076	0.0096
	Catalina Island	7	710	10	0.0181	0.0271	8	0.0066	0.0446	10	0.0133	0.1076
	Mainland	7	710	8	0.0187	0.0217	16	0.0137	0.0131	8	0.0133	0.0116

	Sonora Mojave Desert	7	710	2	0.0038	0.0054	2	0.0012	0.0016	2	0.0027	0.0039
New Paralog	Mojave Desert	13	708	0	0.0000	0.0000	0	0.0000	0.0000	0	0.0000	0.0000

Supplementary Table 1A. Sample Sizes, Gene Length (L) and Estimates of Polymorphism for the 6 loci examined in this study.

locus	*population*	*Tajima's D*	*Fu and Li's D**	*Fu and Li's F**	*Fu and Li's D*	*Fu and Li's F*	*Fay and Wu's H*
GLEANR_896	Baja California	2.14481***	1.1678*	1.6395**	1.5623**	2.15***	9.5897
	Catalina Island	0.77983	1.38019*	1.38024*	1.81718*	1.89352*	-58.57143*
	Mainland Sonora	0.3538	0.60062	0.61141	0.96036	0.91672	-5.1749
	Mojave Desert	NA	NA	NA	NA	NA	NA
GLEANR_897	Baja California	1.22577	0.83208	1.06786	1.4657*	1.71211*	-48.5455*
	Catalina Island	-1.4457*	-1.47053	-1.61834*	-1.54863	-1.87781	-4
	Mainland Sonora	0.383	0.96358	0.92463	1.20599	1.17418	13.606
	Mojave Desert	NA	NA	NA	NA	NA	NA
GLEANR_898	Baja California	-0.69364	-0.50141	0.59937	-0.5696	-0.7055	-1.80952
	Catalina Island	-1.6882**	-1.79119**	-1.94968**	-0.6698	-1.12661	-10.238*
	Mainland Sonora	-0.10218	-0.08191	-0.09541	0.86332	0.79221	-5.238
	Mojave Desert	-1.434*	-1.5099*	-1.61727*	-1.1639	-1.44536*	-0.71429
GLEANR_2575	Baja California	-0.69634	-0.70234	-0.7735	-0.44227	-0.6036	-2
	Catalina Island	-0.30178	-0.519	-0.50749	0.2358	0.13564	-1.38095
	Mainland Sonora	-1.128	-1.0969	-1.20302	-0.98409	-1.21123	-0.19048
	Mojave Desert	0.88922	0.56807	0.70611	1.33922	1.45748	-0.8578
GLEANR_2574	Baja California	0.2969	0.29395	0.32717	1.1288	1.10279	-3.42857
	Catalina Island	-1.20798	-1.1933	-1.30643	0.48375	0.09782	-4.95238
	Mainland Sonora	0.05487	0.08364	0.08535	0.223715	0.22936	-1.21429
	Mojave Desert	-0.71512	1.31251	1.54060*	1.33922	1.68171*	-0.28571

Supplementary Table 1B. Site-Frequency Spectra for 5 loci examined in this study.

		Standard MK Test		
GLEANR_2574		***Polymorphic***	***Fixed***	***Test***
Baja Peninsula	Syn.+nc	23	5	*G-test*
	Non-Syn.	7	2	NS
Catalina Island	Syn.+nc	18	6	*G-test*
	Non-Syn.	5	2	NS
Mojave Desert	Syn.+nc	16	7	*G-test*
	Non-Syn.	5	2	NS
Mainland Sonora	Syn.+nc	20	7	*G-test*
	Non-Syn.	8	2	NS

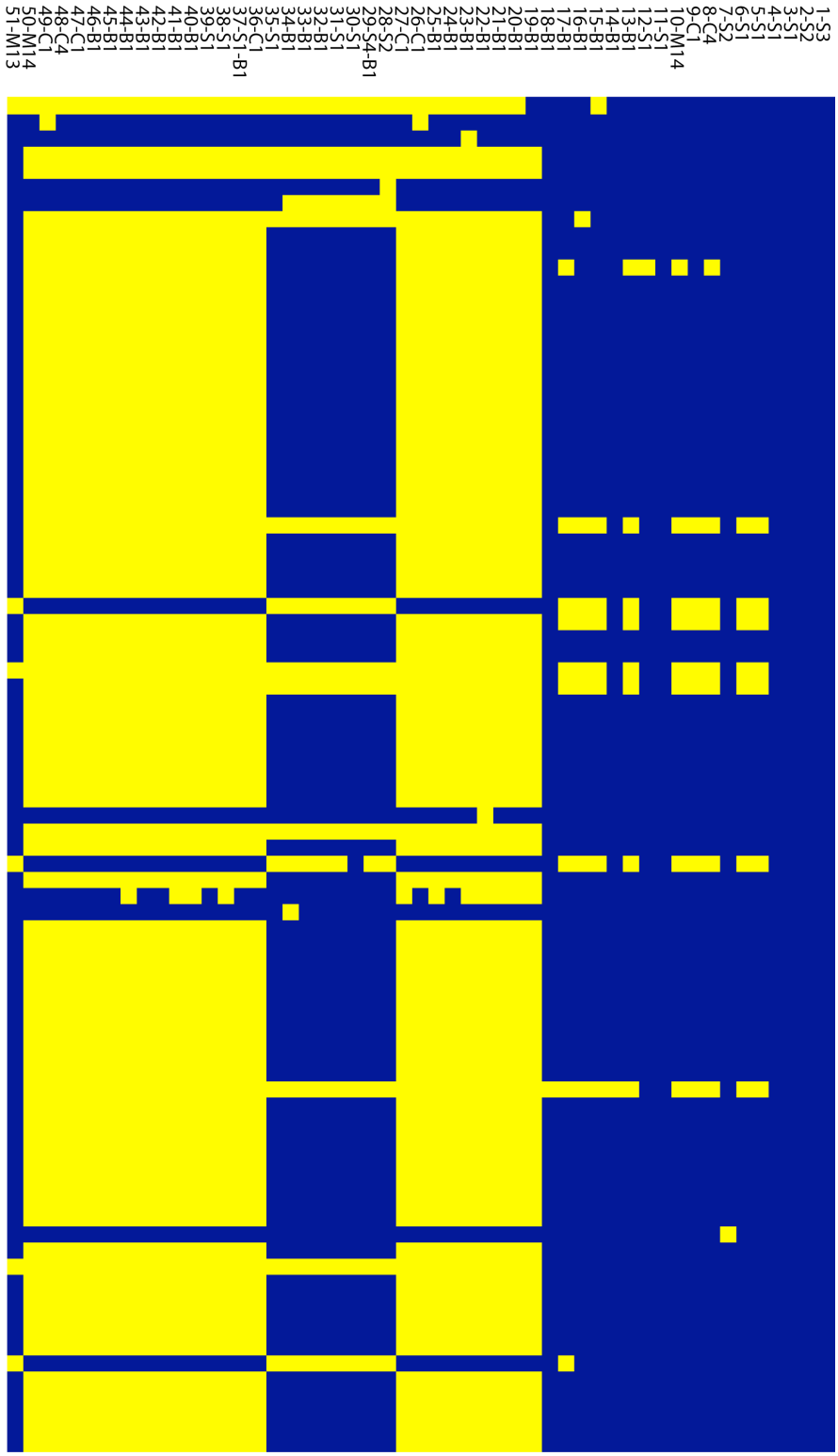
		Standard MK Test		
GLEANR_2575		***Polymorphic***	***Fixed***	***Test***
Baja Peninsula	Syn.+nc	16	27	*G-test*
	Non-Syn.	17	21	NS
Catalina Island	Syn.+nc	17	18	*G-test*
	Non-Syn.	13	21	NS
Mojave Desert	Syn.+nc	18	17	*G-test*
	Non-Syn.	22	13	NS
Mainland Sonora	Syn.+nc	17	19	*G-test*
	Non-Syn.	16	18	NS

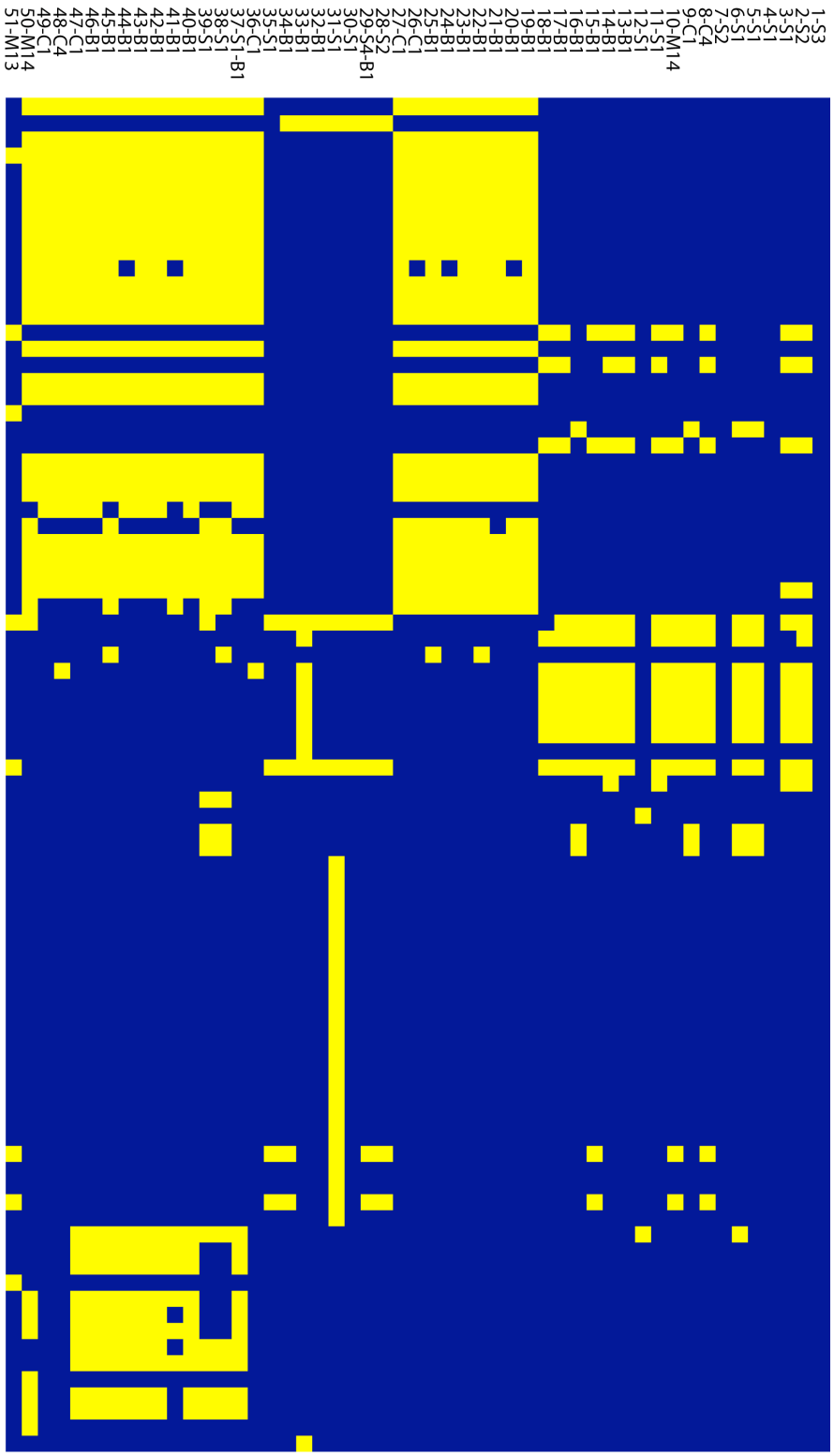
		Standard MK Test		
GLEANR_898		***Polymorphic***	***Fixed***	***Test***
Baja Peninsula	Syn.+nc	9	35	G-test
	Non-Syn.	13	12	**
Catalina Island	Syn.+nc	14	31	G-test
	Non-Syn.	13	10	*
Mojave Desert	Syn.+nc	6	35	G-test
	Non-Syn.	4	12	NS
Mainland Sonora	Syn.+nc	12	31	G-test
	Non-Syn.	18	12	**

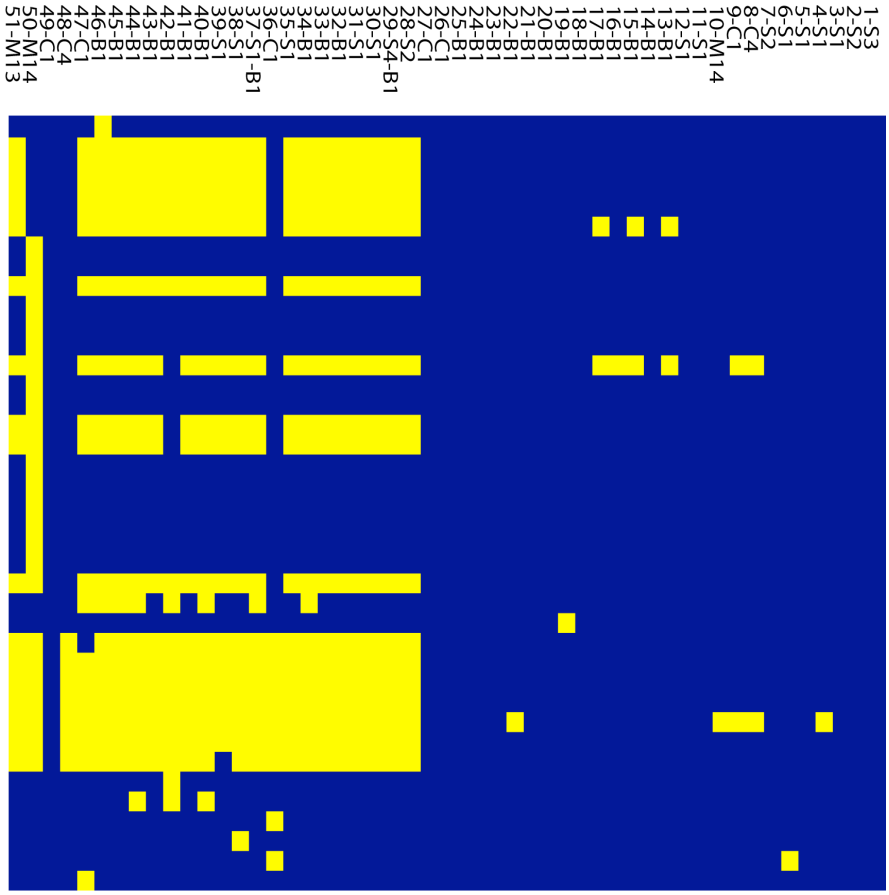
Supplementary Table 1C. Standard McDonald-Krietman Tests for 3 loci examined in this study.

GLEANR_896 /897		*Modified MK Test* *Polymorphic*	*Fixed*	*Test*	*Lineage-Specific MK Test (896)* *Polymorphic*	*Fixed*	*Test*	*Lineage-Specific MK Test (897)* *Polymorphic*	*Fixed*	*Test*
Baja Peninsula	Syn.+nc	61	4	*G-test*	52	2	*G-test*	50	2	*G-test*
	Non-Syn.	67	11	NS	51	5	NS	59	5	NS
Catalina Island	Syn.+nc	53	1		46	1	*G-test*	10	0	
	Non-Syn.	69	1	NA	50	1	NS	20	0	NA
Mojave Desert	Syn.+nc	0	49		0	22		0	22	
	Non-Syn.	0	61	NA	0	15	NA	0	15	NA
Mainland Sonora	Syn.+nc	10	58	*G-test*	10	4	*G-test*	50	5	*G-test*
	Non-Syn.	16	68	NS	17	12	NS	54	4	NS

Supplementary Table 1D. Modified MK Test (Thornton and Long 2005) of GLEANR_896 and GLEANR_897.


1-S3
2-S2
3-S1
4-S1
5-S1
6-S1
7-S2
8-C4
9-C1
10-M14
11-S1
12-S1
13-B1
14-B1
15-B1
16-B1
17-B1
18-B1
19-B1
20-B1
21-B1
22-B1
23-B1
24-B1
25-B1
26-C1
27-C1
28-S2
29-S4-B1
30-S1
31-S1
32-B1
33-B1
34-B1
35-S1
36-C1
37-S1-B1
38-S1
39-S1
40-B1
41-B1
42-B1
43-B1
44-B1
45-B1
46-B1
47-C1
48-C4
49-C1
50-M14
51-M13


1-S3
2-S2
3-S1
4-S1
5-S1
6-S1
7-S2
8-C4
9-C1
10-M14
11-S1
12-S1
13-B1
14-B1
15-B1
16-B1
17-B1
18-B1
19-B1
20-B1
21-B1
22-B1
23-B1
24-B1
25-B1
26-C1
27-C1
28-S2
29-S4-B1
30-S1
31-S1
32-B1
33-B1
34-B1
35-S1
36-C1
37-S1-B1
38-S1
39-S1
40-B1
41-B1
42-B1
43-B1
44-B1
45-B1
46-B1
47-C1
48-C4
49-C1
50-M14
51-M13



Supplementary Table 2. 208 polymorphic sites in 682 aligned bases of GLEANR_896, GLEANR_897, and the New Paralog. All unique halpotypes of D. mojavensis are included. Ancestral GLEANR_896 = haplotypes 1-18. Converted GLEANR_896 = haplotypes 19-26. Converted GLEANR_897 = haplotypes 28-35. Ancestral GLEANR_897 = haplotypes 36-50. New Paralog = haplotype 51. Letter following the haplotype number indicates the population from which it was derived. Baja Peninsula = B. Catalina Island = C. Mainland Sonora = S. Mojave Desert = M. Subsequent number indicates number of sampled alleles that correspond to the haplotype. Polymorphic sites include both SNPs and indels.

Locus	*N*	*S*	π	θ	Tajima's D	Fu and Li's D*	Fu and Li's F*	Fu and Li's D	Fu and Li's F	Fay and Wu's H
996	3	3	0.0023	0.0023	NA	NA	NA	0.0000	0.0000	0.0000
1343	3	11	0.0083	0.0083	NA	NA	NA	-0.1745	-0.1756	0.3333
3196	3	12	0.0092	0.0092	NA	NA	NA	-0.9165	-0.9165	2.0000
5239	3	11	0.0084	0.0084	NA	NA	NA	-1.3027	-1.3027	2.6667
5246	2	9	0.0108	0.0171	NA	NA	NA	NA	NA	NA
5307	3	1	0.0007	0.0007	NA	NA	NA	-0.7071	-0.7071	0.3333
A4115	3	8	0.0064	0.0064	NA	NA	NA	-0.4083	-0.4083	0.6667
A4125	3	6	0.0044	0.0044	NA	NA	NA	0.2789	0.2789	-0.3333
M491	3	21	0.0180	0.0180	NA	NA	NA	-0.8239	-0.8239	2.6667
X100	3	13	0.0096	0.0096	NA	NA	NA	-1.0074	-1.0074	2.3330
average			0.0078	0.0084				-0.5624	-0.5625	1.1851
standard error			0.0046	0.0053				0.4898	0.4897	1.1451

Supplementary Table 3A. Site Frequency Spectra Statistics for the Baja Peninsula

Locus	*N*	*S*	π	θ	Tajima's D	Fu and Li's D*	Fu and Li's F*	Fu and Li's D	Fu and Li's F	Fay and Wu's H
996	8	5	0.0025	0.0025	0.0005	0.1265	0.0107	-0.0531	-0.0478	1.0000
1343	8	19	0.0081	0.0081	-0.4904	-0.4961	-0.5292	-0.5513	-0.5977	0.4286
3196	8	16	0.0054	0.0071	-1.2100	-1.0977	-1.2497	-0.8388	-1.1350	-4.8570
5239	8	19	0.0069	0.0085	-0.9914	-0.8740	-1.0025	-1.1941	-1.3843	2.7143
5246	8	10	0.0033	0.0047	-1.5123	-1.5656	-1.7210	-2.2525*	-2.48228**	1.0714
5307	8	11	0.0056	0.0046	1.1020	0.5784	0.7766	0.5038	0.7914	0.3571
A4115	7	13	0.0054	0.0064	-0.8566	-0.9738	-1.0412	-1.2655	-1.4037	0.9524
A4125	8	3	0.0015	0.0013	0.7117	0.3007	0.4335	0.1655	0.3411	-0.1429
M491	8	20	0.0096	0.0106	-0.4782	-0.4718	-0.5266	-0.4450	-0.5516	0.5714
x100	8	21	0.0074	0.0094	-1.1000	-1.1596	-1.2806	-1.3813	-1.5950	-2.1419
average			0.0056	0.0063	-0.4825	-0.5633	-0.6130	-0.5622	-0.6203	-0.0046
standard error			0.0026	0.0030	0.8524	0.7031	0.8058	0.6690	0.8413	2.0885

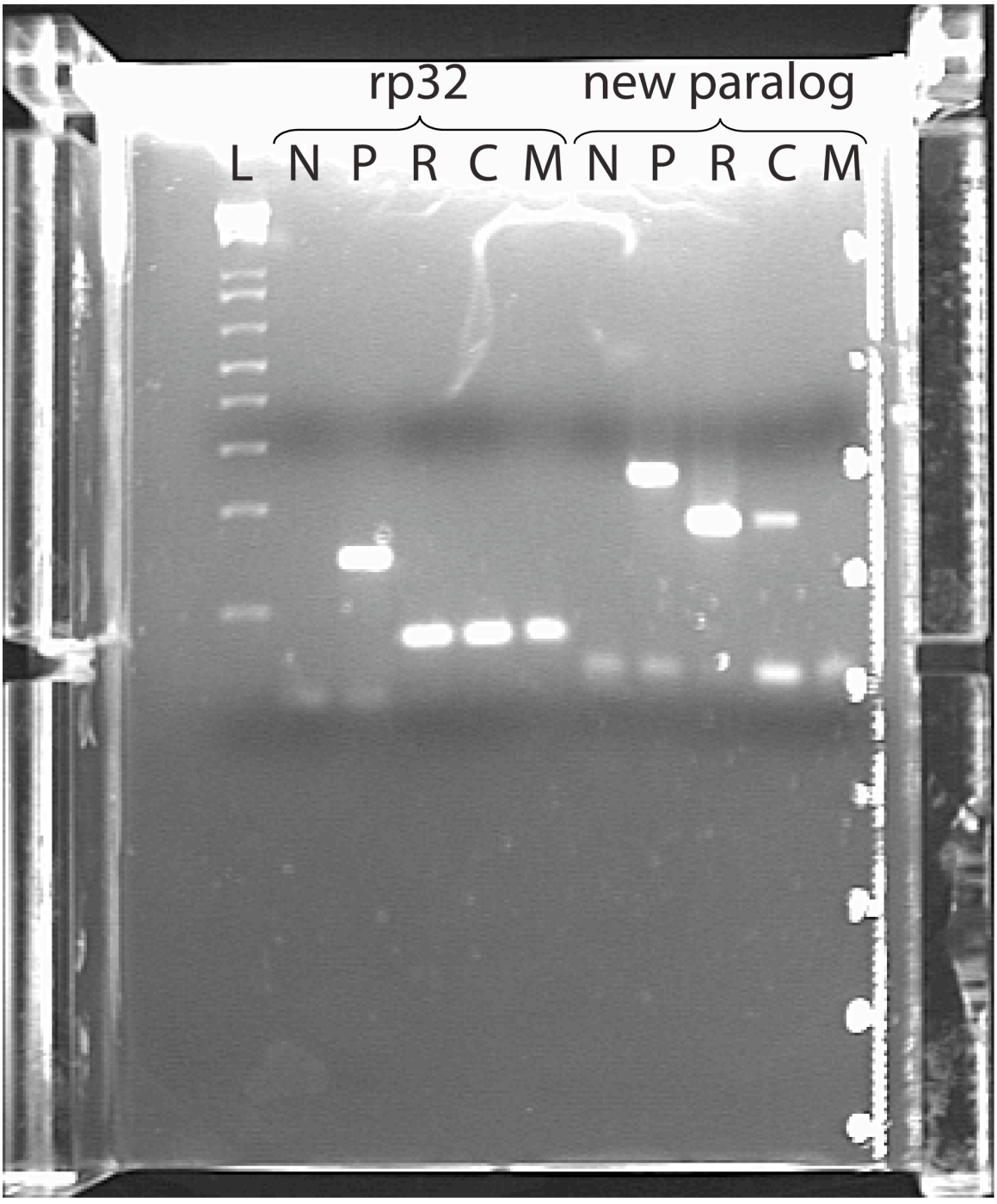
Supplementary Table 3B. Site Frequency Spectra Statistics for Mainland Sonora.

Locus	*N*	*S*	π	θ	Tajima's D	Fu and Li's D*	Fu and Li's F*	Fu and Li's D	Fu and Li's F	Fay and Wu's H
996	4	2	0.0012	0.0013	-0.7099	-0.7099	-0.6043	1.4408	1.2788	-2.0000
1343	3	19	0.0130	0.0121	0.7933	0.7933	0.8203	0.5580	0.7127	1.6667
3196	4	4	0.0023	0.0025	-0.7801	-0.7801	-0.7205	-1.5068	-1.5971	1.3330
5239	4	1	0.0006	0.0006	-0.6124	-0.6124	-0.4787	-0.9129	-0.2176	0.3333
5246	4	0	0.0000	0.0000	NA	NA	NA	NA	NA	NA
5307	4	4	0.0025	0.0024	0.6501	0.6501	0.6004	0.1507	0.2662	0.6667
A4115	4	9	0.0063	0.0060	0.5222	0.5222	0.5185	-0.0406	0.0713	1.6667
A4125	4	4	0.0023	0.0025	-0.7801	-0.7801	-0.7205	0.1507	0.0000	-1.3333
M491	4	6	0.0049	0.0042	1.6621	1.6621	1.5977	1.3965	1.6725	0.3333
x100	4	0	0.0000	0.0000	NA	NA	NA	NA	NA	NA
average			0.0033	0.0032	0.0932	0.0932	0.1266	0.1546	0.2733	0.3333
standard error			0.0040	0.0037	0.9346	0.9346	0.8750	1.0210	1.0012	1.3569

Supplementary Table 3C. Site Frequency Spectra Statistics for Catalina Island.

Locus	*N*	*S*	π	θ	Tajima's D	Fu and Li's D*	Fu and Li's F*	Fu and Li's D	Fu and Li's F	Fay and Wu's H
996	4	1	0.0006	0.0006	-0.6124	-0.6124	-0.3787	-0.9129	-0.9757	0.3333
1343	4	1	0.0006	0.0006	-0.6124	-0.6124	-0.4787	-0.9129	-0.9757	0.3333
3196	8	5	0.0029	0.0031	-0.7968	-0.7968	-0.7530	0.5125	0.3390	-2.3330
5239	4	3	0.0017	0.0019	-0.7545	-0.7545	-0.6747	0.6441	0.4887	-1.6667
5246	4	1	0.0006	0.0006	-0.6124	-0.6124	-0.4787	-0.9129	-0.9757	0.3333
5307	4	11	0.0081	0.0066	2.23308**	2.3308**	2.2464*	2.23388**	2.61315*	0.0000
A4115	4	2	0.0012	0.0013	-0.7099	-0.7099	-0.6043	-1.2007	-1.2788	0.6667
A4125	4	4	0.0025	0.0024	0.6501	0.6501	0.6004	0.9794	1.0747	-0.6667
M491	4	9	0.0069	0.0064	0.8602	0.8602	0.8539	1.2987	1.4253	-1.3333
X100	4	3	0.0018	0.0018	0.1677	0.1677	0.1499	-1.2007	1.2788	0.6667
average			0.0027	0.0025	-0.2689	-0.2689	-0.1960	-0.1895	0.0445	-0.3666
standard error			0.0025	0.0021	0.6122	0.6122	0.5533	0.9651	1.0354	1.0159

Supplementary Table 3D. Site Frequency Spectra Statistics for Mojave Desert.



Supplementary Figure 1. RT-PCR. L= Ladder. N = negative control. P = positive control. R = lower female reproductive tracts. C = female carcasses. M = males. Expression of rp32 (control) and the new paralog (experimental) were assessed. The new

paralog exhibits female-specific expression, with enriched expression in lower female reproductive tracts.

APPENDIX E: FEMALE REPRODUCTIVE PROTEASE EVOLUTION SUGGESTS SEXUAL CONFLICT IN GEOGRAPHICALLY ISOLATED POPULATIONS OF *DROSOPHILA MOJAVENSIS*

**This appendix is in preparation:

Kelleher ES, Clark NL and Markow TA. 2009. Female Reproductive Protease Evolution Suggests Sexual Conflict in Geographically Isolated Populations of *Drosophila mojavensis.* Molecular Biology and Evolution. *In Preparation.*

ABSTRACT

Protein components of the *Drosophila* male ejaculate are critical modulators of reproductive success, several of which are known to evolve rapidly. Recent evidence of adaptive evolution in female reproductive tract proteins suggests this pattern may reflect sexual selection at the molecular level. Mathematical models of sexual selection predict two distinct outcomes of sexual selection for the female molecules involved. First, runaway selection or ongoing coevolutionary chase can result in strong directional selection. Second, in cases of sexual conflict, females can differentiate into two groups, paralyzing males and effectively halting coevolution. Here we explore the evolutionary dynamics of a five paralog gene family of female reproductive proteases within populations of *D. mojavensis*. Remarkably, four of five paralogs show evidence for the emergence of unusually structured haplotypes that suggest the retention of old polymorphism. These gene genealogies furthermore are accompanied by deviations from neutrality consistent with balancing selection. Our study provides the first evidence of this predicted outcome of sexual conflict in *Drosophila*.

INTRODUCTION

Extensive research in a broad range of taxa has demonstrated that the molecular interface that underlies sexual reproduction is extraordinarily dynamic. Reproductive proteins evolve rapidly in a host biologically distinct organisms, including plants, diatoms and humans (Reviewed in Swanson and Vacquier 2002; Clark, Aagaard and Swanson 2006; Panhuis, Clark and Swanson 2006). Molecules with reproductive functions also diverge between species by lineage-specific gene duplications, evolution of novel proteins, and regulatory changes (Begun and Lindfors 2005; Mueller *et al* 2005; Begun *et al* 2006; Findlay *et al* 2008). Within species, reproductive proteins show evidence of two contrasting selective regimes. Reduced polymorphism or elevated divergence at many reproductive protein loci indicates they have experienced positive directional selection (Lee, Ota and Vacquier 1995; Aguadé 1998; 1999; Clark *et al* 2007; Calkins, El-Hinn and Swanson 2007). In contrast, other reproductive proteins exhibit remarkable intraspecific diversity generated by balancing selection (Metz and Palumbi 1996; Gasper and Swanson 2006; Levitan and Ferrell 2006; Hamm *et al* 2007), alternative splicing (Moy *et al* 2008; Springer *et al* 2008), copy number variation (Dopman and Hartl 2007), and gene conversion (Kelleher and Markow 2009).

It frequently has been postulated that the unique patterns of interspecific divergence and intraspecific variation observed amongst reproductive proteins are the result of intersexual coevolution between interacting male and female

reproductive molecules. This reciprocal evolutionary change can be explained by two non-mutually exclusive mechanisms. First, cryptic female choice could empower females to bias fertilization success towards certain males (Eberhard 1996), leading to cyclical evolution of male trait and female preference (Fisher, 1915; 1930). Alternatively, sexual conflict, or a difference in the reproductive interests of the two sexes (Parker 1979), is predicted to result in an evolutionary arms race between males and females (Rice 1996; Gavrilets 2000). Notably, although positive directional selection observed amongst reproductive proteins is an outcome of many sexual selection models (Fisher 1915; 1930; Lande 1981; Kirkpatrick 1982; Gavrilets 2000), diversifying selection is only predicted under regimes that incorporate sexual conflict (Gavrilets 2002; Haygood 2004; Hayashi, Gavrilets and Vose 2007).

Fruit flies of the genus *Drosophila* have long been an important model system for exploring the genetics and evolution of sexual reproduction (Reviewed in Markow 1996; 2002). In these animals, fertilization and reproductive fitness are dependent not only on sperm egg fusion, but also on a complex network of biochemical interactions between male seminal proteins and female reproductive tracts (Reviewed in Kubli 2003; Chapman and Davies 2004; Reviewed in Wolfner 2007). Male seminal proteins modulate an array of critical reproductive outcomes in mated females such as sperm storage, ovulation and oviposition, and female refractoriness (Reviewed in Kubli 2003; Chapman and Davies 2004; Wolfner 2007). Some seminal proteins, furthermore, are known to have negative effects on female fitness (Lung *et al* 2002; Wigby and Chapman 2005; Mueller, Page and Wolfner 2007), creating the potential for sexual conflict and sexually antagonistic coevolution. Consistent with predictions

of intersexual coevolution, male seminal proteins (Aguadé 1998; 1999; Begun *et al* 2000; Swanson *et al* 2001: Wong *et al* 2008), and female reproductive tract proteins (Swanson *et al* 2004; Panhuis and Swanson 2006; Lawniczak and Begun 2007; Prokupek *et al* 2008) are known to evolve rapidly in these organisms.

Although the genus *Drosophila* exhibits remarkable interspecific variation in reproductive biology (Reviewed in Markow 1996; 2002), little is known about the evolution of reproductive proteins outside the *melanogaster* group, particularly for females. Several recent studies, however, have sought to identify reproductive proteins and describe their evolutionary dynamics within the *repleta* group species, *D. mojavensis* (Wagstaff and Begun 2005; 2007; Kelleher, Swanson and Markow 2007; Almeida and Desalle 2008A, 2008B; Kelleher and Markow 2009; Kelleher *et al* 2009). Differences in reproductive biology between *D. mojavensis* and *D. melanogaster* have intriguing implications for reproductive protein evolution. First, *D. mojavensis* females are three to five times more promiscuous than *D. melanogaster* (Reviewed in Markow 1996). Female promiscuity could influence the evolution of reproductive proteins by intensifying selection on post-copulatory traits or elevating sexual conflict (Parker 1979; Markow 2002). *D. mojavensis* females, furthermore, are known to incorporate male-derived molecules into somatic tissues and oocytes (Markow and Ankney 1984). This nutritional benefit to copulation presents a dramatic contrast to the cost of mating incurred by *D. melanogaster* females (Chapman *et al* 1995; Pitnick and García-González 2002; Kuijper, Stewart and Rice 2006; Barnes *et al* 2008). Finally, *D. mojavensis* females

exhibit an insemination reaction, an opaque mass of unknown composition that forms in the uterus after every copulation (Patterson 1946). This phenomenon is thought to protect the male's nutritional investment from cuckoldry by competing males (Markow and Ankney 1984; 1988; Pitnick, Spicer and Markow 1997), and furthermore, may coevolve antagonistically between the sexes (Knowles and Markow 2001).

The current study explores the evolutionary history of a female reproductive tract gene family in four geographically isolated populations of *D. mojavensis*; Baja Peninsula, Catalina Island, Mainland Sonora, and the Mojave Desert. The family encodes five serine-endoprotease paralogs, exclusively expressed in the lower female reproductive tract (Figure 1, Kelleher, Swanson and Markow 2007*).* Although their specific function remains unknown, activity of this class of enzymes in female reproductive tracts is negatively regulated by mating, suggesting potential susceptibility to protease inhibitors in the male ejaculate (Kelleher and Pennington *submitted*). Molecules that may interact biochemically with components of the male ejaculate are exciting candidates for elucidating the molecular bases of ejaculate-female coevolution.

Previous studies suggest that the four geographic populations of *D. mojavnesis* are highly structured (Machado *et al* 2007), and that male and female contributions to reproductive outcomes are coadapted within them (Knowles and Markow 2001; Pitnick *et al* 2003; Knowles, Hernandez and Markow 2005; Kelleher and Markow 2007). If coadaptation is a reflection of molecular coevolution, the male

and female reproductive proteins involved should exhibit unique signatures of selection within each population. The promiscuous mating system of this species (Markow 1996), in conjunction with evidence of sexually antagonistic coevolution (Knowles and Markow 2001), furthermore, would predict that sexual conflict plays an important role in *D. mojavensis* reproductive protein evolution. We discuss our data in terms of these predictions.

MATERIALS AND METHODS

Fly Strains.

D. mojavensis were collected from Catalina Island (2001), Mojave Desert (2002), Baja Peninsula (2002), and Mainland Sonora (2007) by J. Bono, L. Reed, and L. Matzkin. *D. arizoznae,* the sister species of *D. mojavensis,* were collected in Tucson, Arizona (2000) by L. Matzkin. A third, closely related species, *D. navajoa*, was obtained from the Tucson *Drosophila* Stock Center. All flies used in population analyses were maintained as isofemale lines. Between 7 and 8 isofemale lines were sampled from each population for each locus.

Loci and Primer Design.

The genomic arrangement and phylogenetic relationships of the protease gene family examined in this study are presented in Figure 1. Although the genes remain unannotated, our previous study showed that they were expressed in *D.*

arizonae, and that orthologous sequences were present in the *D. mojavensis* genome (Kelleher, Swanson and Markow 2007). Intron-exon splice sites were inferred from *D. arizonae* ESTs in Kelleher, Swanson and Markow (2007). For simplicity, we refer to these genes as female reproductive protease A-E, or FRP-A-E. *D. mojavensis* (http://rana.lbl.gov/drosophila/) orthologs further were aligned to available *D. arizonae* ESTs to generate paralog-specific primers that amplified the majority of the coding sequence for each locus. All paralogs were reciprocally monophyletic (not shown). Further, heterozygosity in sampled isofemale lines was quite low, except for flies that recently had been introduced to the lab (<5 generations). We are confident, therefore, that each set of primers amplified a unique genomic location.

Sequencing.

Genomic DNA was isolated from whole flies using the DNeasy Kit (Qiagen) according to manufacturer instructions. Standard PCR was performed using internal, paralog-specific primers (Figure 1). All sequencing was performed on an ABI 3700 DNA sequencer with Big Dye Terminator chemistry. Primers and PCR conditions are available from the authors upon request. Base-calling and assembly were performed in Sequencher 4.8.

Polymorphism Analyses.

Haplotypes were phased in Arlequin (http://lgb.unige.ch/arlequin/software/), and a single haplotype for each individual

was retained for subsequent analyses. Polymorphism analyses, estimation of population parameters, and tests of selection were performed in DNAsp (Rozas *et al* 2003). Sample sizes, sequence lengths and estimates of polymorphism are presented in supplementary table 1. Significance of site frequency spectra statistics was assessed by coalescent simulations under the conservative assumption of no recombination. For tests requiring an outgroup, one or more *D. arizonae* orthologs were used for FRP-A, FRP-B, and FRP-C. For FRP-D we used FRP-E as an outgroup, and vice versa, due to uncertainty surrounding the identity of the *D. arizonae* ortholog. Using paralogs as an outgroup is known to be a conservative approach for McDonald-Kreitman tests (McDonald and Kreitman 1991; Thornton and Long 2005; Thornton 2007). We note, however, that using the putative *D. arizonae* ortholog had no effect on the outcome of the test. Tests were polarized with the appropriate sequence from *D. navajoa.*

Gene conversion was detected by GENECONV within an alignment of all unique haplotypes for all paralogs using the method of Sawyer (1989). Briefly, gene conversion tracts between pairs of sequences are identified by considerable stretches of complete identity interspersed between two regions of considerable mismatch, or one region of mismatch and the end of the alignment. Statistical significance of these fragments is determined by permutation tests. Neighbor-joining gene trees (Saitou and Nei 1987) were constructed in Paup*4.0b10 (Swofford 2000).

Three Dimensional (3D) Modeling.

Serine endoprotease catalytic sites (Reviewed in Polgar 2005), and protease inhibitor sites (Reviewed in Srinivasan, Giri, and Gupta 2006), as well as previously identified Bayes Empirical Bayes positively selected sites (Kelleher, Swanson and Markow 2007; Yang, Wong and Nielsen 2005), were mapped to a predicted 3D model for FRP-C obtained from Swiss-Model (Schwede *et al* 2003).

To test for an association between positively selected sites and protease inhibitor sites, we implemented a permutation analysis previously described in Clark *et al* (2007). Briefly, the distance from each selected site to the nearest functional site was calculated, and their mean value compared to a distribution of distances between random pairs of sites. Buried, core sites with 10% or less surface exposure per residue, as calculated by GETAREA (Fraczkiewicz and Braun 1998), were not considered for random sets. This exclusion makes the test more conservative, because these sites evolve slowly relative to surface sites and rarely are inferred as positively selected. Statistical significance was determined as the fraction of random permutations with a mean distance equal to or lower than the observed mean distance between selected and inhibitor sites.

RESULTS

Multiple Paralogs Evolve Non-Independently Through Gene Conversion.

Gene conversion and non-allelic homologous recombination results in non-independent evolutionary histories of paralogous loci. Describing this process is critical, as it leads to complex genealogies and unusual patterns of polymorphism not seen for single copy genes (Innan 2003A; Thornton 2007). Gene conversion tracts between pairs of paralogous haplotypes were identified as fragments of complete identity flanked by regions of significant mismatch, using the method of Sawyer (1989). No significant fragments were detected between FRP-A and any other paralog, indicating this locus evolves independently (Figure 2). In contrast, there is evidence of gene conversion in at least one pairwise comparison between all other paralogs in the examined gene family (Figure 2). The lack of detectable gene conversion between FRP-A and the other paralogs may suggest that gene conversion does not occur, but could also suggest that it is deleterious. Indeed, all sampled haplotypes of this paralog exhibit the same replacement changes in the three residues of the catalytic triad (reviewed in Polgar 2005), suggesting it has acquired a divergent, non-proteolytic function. In contrast, all other paralogs have retained a catalytic triad, indicating that their biochemical functions may be more similar.

In terms of both tract length and frequency, the most extensive gene conversion was observed between paralogs FRP-C and FRP-D. These paralogs are neither physically adjacent, nor are they genetically more similar to each other than to the remainder of the gene family. Conversely, minimal gene conversion was observed between adjacent paralogs, or between genetically similar pairs FRP-D

and FRP-E, and FRP-B and FRP C. There is no evidence, therefore for an association between phylogenetic or physical distance and gene conversion.

Examination of the frequency that a given site is found within a significant fragment reveals that gene conversion is non-randomly distributed along the chromosome (Figure 3). Gene conversion is frequent in the 5' end of the gene, peaks near the center, and is entirely absent from the 3'end. Intriguingly, previously described selected-sites (Kelleher, Swanson and Markow 2007) are highly concentrated at the 3' end of the gene (Figure 3), suggesting a possible negative-association between gene conversion and adaptive evolution. Consistent with this hypothesis, we observed that positively selected sites exhibited a significantly lower frequency of gene conversion than non-selected sites (Wilcoxon Rank Sum Test, $p = 0.0016$).

Female Reproductive Proteases Exhibit Unusually Structured Haplotypes.

A previous examination of genome-wide variation in *D. mojavensis* indicates that haplotypes are structured between four geographically isolated populations, Baja Peninsula, Catalina Island, Mainland Sonora, and the Mojave Desert (Machado *et al* 2007, Figure 4A). Specifically, a concatenated genealogy of 10 nuclear loci revealed a well-supported reciprocally monophyletic clade for each locality (Machado *et al* 2007). To explore the relationships between female reproductive protease haplotypes sampled here, gene genealogies were constructed for each of the five paralogs (Figure 4B-F). Well-supported clades (bootstrap values >95%)

containing individuals from multiple localities were observed for FRP-A, FRP-C, FRP-D, and FRP-E, providing little evidence for a geographical structuring of sampled haplotypes. The genealogies of FRP-A and FRP-D are particularly unusual, as haplotypes sort into to two divergent clades, suggesting ancestral polymorphism. To determine if these results represent a genuine difference from the loci examined in Machado *et al* (2007), we constructed gene genealogies for each of the 10 loci sampled in that study individually. No well-supported clades grouping haplotypes from multiple localities were observed (not shown), indicating that the genealogies of the duplicated proteases examined here are distinct from those of putatively neutral loci.

To provide a quantitative measure of the unusual haplotype structure, we estimated linkage-disequilibrium within each locus and population, using *Zns* (Kelly 1997) and *Za* (Rozas *et al* 2001) (Table 1). *Zns* measures the correlation of pairwise-linkage disequilibrium, r^2, for all polymorphic sites within a locus (Kelly 1997), while *Za* estimates the correlation in r^2 values between adjacent polymorphic sites only (Rozas *et al* 2001). For both statistics, values that approach 1 indicate high linkage disequilibrium and low asymmetry in the frequency of polymorphic sites, a pattern that could result either from cryptic population structure, or from natural selection acting on a polymorphism that is closely linked to neutral sites in this region. All five paralogs exhibit a degree of linkage disequilibrium that is inconsistent with a standard neutral model in at least one population (Table 1). These significant values may result from natural selection, as no evidence for cryptic

population structure has been seen for other loci examined from these same geographic populations (Machado *et al* 2007; Reed *et al* 2007; Matzkin *et al* 2008).

If gene conversion introduces multiple linked polymorphisms within the same conversion tract, it could lead to strong linkage disequilibrium between polymorphic sites and significant values of *Zns* and *Za*. Because no gene conversion was detected between FRP-A and any other paralog, this phenomenon cannot explain the degree of linkage disequilibrium observed at this locus. To elucidate the contribution of gene conversion to linkage disequilibrium at the other four loci, we identified sites with evidence of gene conversion within each population and excluded them from the analysis. In several cases, significant linkage disequilibrium was no longer detected when sites of gene conversion were excluded (Table 1). The majority of values, however, remained significant (Table 1), suggesting that forces other than gene conversion are responsible for the observed deviations from neutrality. We cannot, however, rule out the possibility that GENECONV failed to detect older gene conversion events that nonetheless contribute to linkage disequilibrium.

Site-Frequency Spectra Suggest Both Directional and Balancing Selection.

One explanation for the unusual patterns of haplotype structure and linkage disequilibrium observed in the female reproductive proteases examined here is that these loci are subject to balancing selection. The pattern is particularly compelling for FRP-A and FRP-D, where the presence of two well-structured haplogroups

shared between two or more populations suggests an old, and possibly balanced, polymorphism. To elucidate additional deviations from neutrality that could result from selection, four statistics were used to detect skews in the site-frequency spectra. Tajima's *D* (Tajima 1989) detects an excess of intermediate or low frequency polymorphisms suggestive of balancing or directional selection, respectively. Similarly, Fu and Li's *F**, and Fu and Li's *F* (Fu and Li 1993), look for an excess of recent polymorphisms indicative of a selective sweep, or old polymorphisms indicative of balancing selection, by assigning variation to branches on a gene genealogy. In contrast, Fay and Wu's *H* (2000) detects an excess of high frequency derived polymorphisms characteristic of genetic hitchhiking associated with a recent selective sweep. Although the site-frequency spectra can be affected by demographic processes, these statistics tend to be slightly negative and close to zero at putatively neutral loci for all four populations of *D. mojavensis* (Machado *et al* 2007; Kelleher and Markow 2009). Significantly positive or negative values, therefore, cannot be attributed to demographic history.

There was a general trend towards positive values of *D, F,* and *F**, amongst the five paralogs (Table 2). This result is not unexpected, as gene conversion is predicted to create a marginally positive skew in the site frequency spectra of duplicate loci (Innan 2003A; Thornton 2007). Significantly positive values do not result from neutral gene conversion, however, as an over-estimation of the variance makes these statistics quite conservative for duplicate loci undergoing concerted evolution (Innan 2003A; Thornton 2007). In four combinations of populations and

loci, FRP-A (Mojave Desert), FRP-C (Mainland Sonora), FRP-D (Baja Peninsula), and FRP-E (Catalina Island), these statistics indicated a significant excess of intermediate-frequency or old polymorphisms (Table 2), suggesting that selection acts to retain genetic variation relative to a neutral model. All values remained significant when sites undergoing gene conversion were excluded from the alignment (not shown), further confirming that the observed deviations from neutrality are not the result of gene conversion.

At FRP-E we observed a significantly negative value of Fay and Wu's H (Fay and Wu 2000) in Mainland Sonora. This signature of genetic hitchhiking suggests a partial or complete selective sweep at, or adjacent to, this locus. The significance of this result can be difficult to interpret, however, as partial selective sweeps may be associated with more complex scenarios of balancing selection (Meiklejohn *et al* 2004).

McDonald Kreitman Tests Suggest Both Directional and Balancing Selection.

The McDonald-Kreitman (MK) test compares the ratio of silent to replacement polymorphism within populations to silent and replacement divergence between species (McDonald and Kreitman 1991). Positive directional is predicted to result in significant excess of replacement divergence (McDonald and Kreitman 1991). In contrast, an excess of replacement polymorphism may indicate balancing selection, segregation of mildly deleterious variation, or recently relaxed constraint (Nachman 1998). MK tests detected deviations from neutrality for FRP-A

(Table 3), FRP-D (Table 4), and FRP-E (Table 4), but not for FRP-B or FRP-C (not shown).

FRP-A exhibited an excess of replacement polymorphism in the Mojave Desert population in a standard MK test (Table 3). The lineage-specific test was not significant (Table 3), however, this likely reflects a lack of power resulting from the paucity of fixed differences on the *D. mojavensis* branch. Although balancing selection is only one possible interpretation of this result, other deviations from neutrality at this locus point to this same selective regime. The Mojave Desert population also showed and excess of old polymorphism (Table 2), and significant linkage disequilibrium (Table 1) at FRP-A. The presence of two differentiated haplogroups in the Baja Peninsula, Mainland Sonora, and Mojave Desert (Figure 4), as well as considerable linkage disequilibrium in the Baja Peninsula and Mojave Desert (Table 2), further support the assertion that FRP-A may harbor a balanced ancestral polymorphism.

In contrast to the other populations, a lineage-specific MK test for Catalina Island suggests an excess of replacement divergence at FRP-A consistent with directional selection (Table 3). This population, furthermore, exhibits only one haplogroup of FRP-A (Figure 3B). It is apparent, therefore, that Catalina Island flies have undergone a shift from the ancestral selective regime of FRP-A.

For FRP-D, a lineage-specific excess of replacement divergence was observed for the Baja California and Mainland Sonora populations (Table 4). This signature of directional selection was inconsistent with the presence of two differentiated haplogroups in both these populations (Figure 3, Table 1), as well as the excess of

both intermediate-frequency and old polymorphism in the Baja Peninsula (Table 2). The converse result is observed for FRP-E, where an excess of high-frequency derived polymorphisms suggests a recent selective sweep (Table 2), whereas a lineage-specific MK test suggests an excess of replacement polymorphism characteristic of balancing selection (Table 4). Although these deviations from neutrality may appear contradictory, it is important to remember that they are sensitive to different evolutionary signatures and time-scales. While site-frequency spectra may suggest recent selective events at a given locus, MK tests will be more sensitive to the history of the locus since divergence from the outgroup. Deviations from neutrality in opposite directions, therefore, may indicate that the evolutionary history of these loci has been more complex than simple models of directional or diversifying selection.

Selected Sites are Structurally-Associated with Determinants of Protease Inhibitor Susceptibility.

The five paralogs examined in this study are serine endoproteases. Serine endoprotease activity in *Drosphila arizonae* female reproductive tracts is negatively regulated by mating, suggesting susceptibility of female proteases to inhibitors in the male ejaculate (Kelleher and Pennington *submitted*). If female proteases interact and coevolve with male protease inhibitors, it is predicted that selected sites will cluster near residues that determine susceptibility to protease inhibitors. Consistent with this hypothesis, selected sites (Kelleher, Swanson and Markow 2007) often are

observed to be closely-associated with sites important to protease inhibitor susceptibility and resistance (Figure 5, Reviewed in Srinivasan, Giri and Gupta 2006). We furthermore observed that 3 of 14 selected sites are also determinants of inhibitor susceptibility, a marginally significant excess (Fisher's Exact Test p = 0.058).

To determine if selected sites and protease inhibitor interaction sites are associated in three dimensional space, we compared the average pairwise distance between each selected site and the closest protease inhibitor interaction site to 10^6 sets of randomly sampled sites. Selected sites are significantly closer to protease inhibitor interaction sites than expected by chance (p = 0.02220), indicating that these two groups of sites are physically associated within the structure of the protein.

DISCUSSION

Although there is considerable empirical evidence that male reproductive proteins experience both directional and diversifying selective regimes in natural populations, the degree to which these patterns extend to their female interactors, as predicted under models of intersexual coevolution, remains largely unexplored. The female reproductive tract gene family examined here exhibits a history of directional selection, balancing selection, and gene conversion, a result that is entirely consistent with molecular coevolution. Evidence for balancing selection at

these loci is particularly intriguing, as this mode of evolution is an explicit prediction of sexual conflict and sexually antagonistic coevolution (Gavrilets and Waxman 2002; Hayashi, Vose and Gavrilets 2007). Geographically isolated populations of *D. mojavensis*, furthermore, differed in terms of signatures of selection at individual loci, as predicted if adaptation results from a unique coevolutionary trajectory.

Gene Conversion.

Paralogous members of a multi-gene family can evolve non-independently through interlocus gene conversion and recombination (Reviewed in Liao 1999). All paralogs examined in this study, except FRP-A, exhibited considerable evidence for interlocus gene conversion (Figure 2). Phylogenetic and physical distance, furthermore, were not associated with the frequency of gene conversion between paralogs. This result contrasts previous studies in yeast, which suggest that gene conversion is negatively associated with both phylogenetic and physical distance (Drouin 2002).

We observed a strong negative association between the frequency of gene conversion and sites inferred to have experienced adaptive evolution (Figure 3). This suggests that gene conversion may interfere with the process of adaptive evolution, or may be costly in genetic regions that have experienced positive selection since gene duplication. This observation is consistent with evolutionary models suggesting that divergence of members of multigene families is determined

an antagonistic process between diversifying force of selection, and the homogenizing force of gene conversion (Walsh 1987; Innan 2003B).

Female Reproductive Proteases Experience Both Directional and Balancing Selection.

Three different loci examined in this study showed evidence of directional selection in at least one population of *D. mojavensis*. An excess of derived polymorphisms characteristic of genetic hitchhiking and a recent selective sweep was observed in Mainland Sonora for FRP-E. Similarly, an excess of replacement divergence, indicative of directional selection, was observed in Catalina Island at FRP-A, and in the Baja Peninsula and Mainland Sonora at FRP-D. These observations mirror previous population surveys of *D. melanogaster* female reproductive tract proteins, which report a high frequency of positive directional selection (Swanson *et al* 2004; Panhuis and Swanson 2006; Lawniczak and Begun 2007).

Although directional selection may have played a role in shaping the observed patterns of variation, a preponderance of evidence for balancing selection suggests this selective regime may have dominated the evolutionary history of this gene family. Four of the five proteases exhibited a haplotype structure uncharacteristic of neutral loci, two of which indicated the maintenance of two distinct haplogroups in multiple populations. Haplotype structure of non-geographic origin may suggest the maintenance of old variation under a balancing selective regime. All five loci we examined exhibited significant linkage disequilibrium, furthermore, as expected if these loci are linked to a

balanced polymorphism (Kreitman 1983; Kelly 1997). Finally, deviations from neutrality at four of the five loci were consistent with balancing selection. Site frequency spectra tests for four different loci exhibit an excess of intermediate-frequency or old polymorphisms, suggesting selective retention of genetic variation relative to a neutral model. McDonald Kreitman tests at two different loci, furthermore, exhibit an excess of replacement polymorphisms consistent with balancing selection.

Isolated Populations Experience Different Selective Regimes.

Previous research in *D. mojavensis* has provided extensive evidence that male and female contributions to reproductive outcomes are coadapted within the populations examined here (Knowles and Markow 2001; Pitnick *et al* 2003; Knowles, Hernandez and Markow 2005; Kelleher and Markow 2007). If coadaptation is a reflection of molecular coevolution, the male and female reproductive proteins involved are predicted to exhibit unique signatures of selection within each of the four populations. Exactly this pattern is seen in the female reproductive tract proteases examined in this study. Deviations from neutrality, in terms of linkage disequilibrium, site-frequency spectra, and MK tests, are invariably confined to a single population or group of populations. Additionally, populations often showed opposing selective regimes at the same locus. For example, while FRP-A showed evidence of balancing selection in the Mojave Desert, Baja Peninsula, and Mainland Sonoran populations, this same locus appears to have experienced directional selection in Catalina Island. Similar examples of population-

specific selection have been seen in marine invertebrate fertilization proteins, and are thought to reflect differences in the degree of sexual conflict (Levitan and Ferrell 2006; Clark *et al* 2007).

Sexual Selection and Sexual Conflict.

The unique evolutionary patterns exhibited by *D. mojavensis* female reproductive tract proteins bear a striking resemblance to mathematical models of sexual selection (Fisher 1915; 1930; Lande 1981; Kirkpatrick 1982; Gavrilets 2000: Gavrilets and Waxman 2002; Hayashi, Gavrilets and Vose 2007) Traditional sexual selection, analogous to cryptic female choice in the post-copulatory arena, may drive a runaway process that exerts directional selection on the loci involved (Fisher 1915; 1930; Lande 1981; Kirkpatrick 1982). In contrast, under sexual conflict, female loci can either "run-away", or they can diverge into distinct alleles, effectively halting coevolutionary chase from a male locus (Gavrilets 2002; Hayashi, Gavrilets and Vose 2007). Thus, directional selection amongst the proteases examined here could result from either cryptic female choice or sexually antagonistic coevolution. The emergence of divergent haplotypes maintained by balancing selection, however, is only predicted by selective regimes that incorporate sexual conflict.

Although the physiological and biochemical underpinnings of intersexual coevolution remain unexplored, it is compelling that the female proteases examined in this study are negatively regulated by mating (Kelleher and Pennington *submitted*). This observation is consistent with sexual conflict: if female proteases

are costly to males, males may seek to regulate them through protease inhibitors in the male ejaculate (Wagstaff and Begun 2005; Kelleher *et al* 2009). Our structural analysis revealed that previously inferred selected sites are clustered with sites that determine susceptibility to protease inhibitors (Figure 5), as expected if female proteases coevolve antagonistically with male protease inhibitors.

Balancing Selection and Gene Duplication.

Balancing selection in *D. mojavensis* female reproductive tract proteins is reminiscent of another emergent pattern of reproductive protein evolution in this lineage: gene duplication (Kelleher, Swanson and Markow 2007; Wagstaff and Begun 2007; Almeida and DeSalle 2008; Kelleher and Markow 2009). Although there are a few reports of lineage-specific duplications in *D. melanogaster* seminal fluid proteins (Cirera and Aguade, Findlay *et al* 2008), recent duplicates occur with high frequency amongst both male seminal proteins (Wagstaff and Begun 2007; Almeida and DeSalle 2008), and female reproductive tract proteins (Kelleher, Swanson and Markow 2007; Kelleher and Markow 2009) in the *repleta* species group. It is exciting to speculate that the selective forces that underlie gene duplication and balancing selection may not be independent. Several models have suggested that if a balanced polymorphism is maintained by overdominant selection, a gene duplication event that unites two functionally diverged alleles on the same chromosome will immediately experience a selective advantage due to heterosis (Spofford, 1969, Ohno 1970, Otto and Yong 2002; Walsh 2003; Proulx and

Phillips 2006). If female diversification is a strategy in antagonistic coevolution, therefore, balancing selection and gene duplication may be iterative steps in this process.

Conclusion.

Molecular coevolution of male and female reproductive molecules is an explicit prediction of sexual selection (Fisher 1915; 1930; Lande 1981; Kirkpatrick 1982; Gavrilets 2000; Gavrilets and Waxman 2002; Hayashi, Gavrilets and Vose 2007), which is supported by the evolutionary dynamics of reproductive proteins in a broad range of taxa (Swanson and Vacquier 2002; Clark, Aagard and Swanson 2006; Panhuis, Clark and Swanson 2006). Empirical evidence of interacting male and female reproductive proteins with similar selective regimes, however, remains confined to the free spawning marine gastropod abalone (Lee, Ota and Vacuier 1995; Yang, Swanson and Vacquier 2000; Galindo *et al* 2002; Galindo, Swanson and Vacquier 2003). Identifying biochemical interactions between male and female reproductive proteins, and exploring how these interactions translate to evolutionary patterns, remains an important and persistent biological challenge.

Acknowledgements.

The authors would like to acknowledge Michael Nachman for generous use of equipment, Willie Swanson for helpful discussion, and Jeremy Bono, and Stephen Schaeffer for generous comments on the manuscript. This research was funded by a

National Science Foundation Doctoral Dissertation Improvement Grant to E.S.K., and the Center for Insect Science and the University of Arizona. E.S.K. was supported by an NSF-IGERT research traineeship in Evolutionary, Functional and Computational Genomics at the University of Arizona, and a Dissertation Fellowship from the American Association of University Women.

REFERENCES

Aguadé M. 1998. Different forces drive the evolution of the Acp26Aa and Acp26Ab accessory gland genes in the *Drosophila melanogaster* species complex. Genetics 150:1079–1089.

Aguadé M. 1999. Positive selection drives the evolution of the Acp29AB accessory gland protein in *Drosophila*. Genetics 152:543–51.

Almeida FC, Desalle R. 2008A. Evidence of adaptive evolution of accessory gland proteins in closely related species of the *Drosophila repleta* group. Mol. Biol. Evol. 25:2043–2053.

Almeida FC, DeSalle R. 2008B. Orthology, Function, and Evolution of Accessory Gland Proteins in the *Drosophila repleta* Group. Genetics 181:235–245.

Barnes AI, Wigby S, Boone JM, Partridge L, Chapman T. 2008. Feeding, fecundity and lifespan in female *Drosophila melanogaster*. Proc. Biol. Sci. 275:1675–83.

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

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

Begun DJ, Lindfors HA, Thompson ME, Holloway AK. 2006. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. Genetics 172:1675–1681.

Calkins JD, El-Hinn D, Swanson WJ. 2007. Adaptive evolution in an avian reproductive protein: ZP3. J. Mol. Evol. 65:555–563.

Cirera S, Aguadé M. 1998. Molecular evolution of a duplication: the sex-peptide (Acp70A) gene region of *Drosophila subobscura* and *Drosophila madeirensis*. Mol. Biol. Evol. 15:988–996.

Chapman T, Liddle LF, Kalb JM, Wolfner MF, Partridge L. 1995. Cost of mating in *Drosophila melanogaster* females is mediated by male accessory gland products. Nature 373:241–244.

Chapman T, Davies SJ. 2004. Functions and analysis of the seminal fluid proteins of male *Drosophila melanogaster* fruit flies. Peptides 25:1477–1490.

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

Clark NL, Findlay GD, Yi X, MacCoss MJ, Swanson WJ. 2007. Duplication and selection on abalone sperm lysin in an allopatric population. Mol. Biol. Evol. 24:2081–90.

Dopman EB, Hartl DL. 2007. A portrait of copy-number polymorphism in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U. S. A. 104:19920–19925.

Drouin G. 2002. Characterization of the gene conversions between the multigene family members of the yeast genome. J. Mol. Evol. 55:14–23.

Eberhard WG. 1996. Female Control: Sexual Selection by Cryptic Female Choice. Princeton, New Jersey: Princeton University Press.

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

Fisher RA. 1915. The evolution of sexual preference. Eugenics Review 7:115–123.

Fisher RA. 1930. The genetical theory of natural selection. Oxford: Clarendon Press.

Findlay GD, Yi X, Maccoss MJ, and Swanson WJ. 2008. Proteomics reveals novel *Drosophila* seminal fluid proteins transferred at mating. P.L.o.S. Biol. 6:e178.

Fraczkiewicz R, Braun W. 1998. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. J. Comp. Chem. 19:319.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics 1993 133:693–709.

Galindo BE, Moy GW, Swanson WJ,Vacquier VD. 2002. Full-length sequence of VERL, the egg vitelline envelope receptor for abalone sperm lysin. Gene 288:111–117.

Galindo BE, Vacquier VD, Swanson WJ. 2003. Positive selection in the egg receptor for abalone sperm lysin. Proc. Natl. Acad. Sci. U. S. A. 100 4639–4643.

Gasper J, Swanson WJ. 2006. Molecular population genetics of the gene encoding the human fertilization protein zonadhesin reveals rapid adaptive evolution. Am. J. Hum. Genet. 79:82–30.

Gavrilets S. 2000. Rapid evolution of reproductive barriers driven by sexual conflict. Nature 403:886–889.

Gavrilets S, Waxman D. 2002. Sympatric speciation by sexual conflict. Proc. Natl. Acad. Sci. U. S. A. 99:10533–10538.

Hamm D, Mautz BS, Wolfner MF, Aquadro CF, 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.

Hayashi TI, Vose M, Gavrilets S. 2007. Genetic differentiation by sexual conflict. Evolution 61:516–29.

Haygood R. 2004. Sexual conflict and protein polymorphism. Evolution 58:1414–23.

Innan H. 2003A. The coalescent and infinite-site model of a small multigene family. Genetics 163:803–810.

Innan H. 2003B. A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. Proc. Natl. Acad. Sci. U. S. A. 100:8793-8798.

Kelleher ES, Markow TA. 2007. Reproductive Tract Interactions Contribute to Isolation in *Drosophila*. Fly 1:33–37.

Kelleher ES, Swanson WJ. Markow TA. 2007. Gene duplication and adaptive evolution of digestive proteases in *Drosophila arizonae* female reproductive tracts. P.L.o.S. Genet. 3:1541–1549.

Kelleher ES, Markow TA. 2009. Duplication, Selection and Gene Conversion in a *Drosophila mojavensis* Female Reproductive Protein Family. Genetics. 181:1451-1465.

Kelleher ES, Watts TD, Laflamme BA, Haynes PA, Markow TA. 2009. Proteomic analysis of *Drosophila mojavensis* male accessory glands suggests novel classes of seminal fluid proteins. Insect Biochem. Mol. Biol. *Epub ahead of print.*

Kelly JK. 1997. A test of neutrality based on interlocus associations. Genetics 146:1197–1206.

Kirkpatrick M. 1982. Sexual selection and the evolution of female choice. Evolution 36:1–12.

Kuijper B, Stewart AD, Rice WR. 2006. The cost of mating rises nonlinearly with copulation frequency in a laboratory population of *Drosophila melanogaster*. J. Evol. Biol. 19:1795–1802.

Knowles LL, Markow TA. 2001. Sexually antagonistic coevolution of a postmating prezygotic reproductive character in desert *Drosophila*. Proc. Nat. Acad. Sci. U. S. A. 98:8692–8696.

Knowles LL, Hernandez BB, Markow TA. 2005. Non-antagonistic interactions between the sexes revealed by the ecological consequences of reproductive traits. J. Evol. Biol. 18:156–161.

Kubli E. 2003. Sex-peptides: seminal peptides of the *Drosophila* male. Cell. Mol. Life. Sci. 60:1689–1704.

Kreitman M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature 304:412-417.

Lande R. 1981. Models of speciation by sexual selection on polygenic traits. Proc. Natl. Acad. Sci. U. S. A. 1981 78:3721–3725.

Lawniczak MK, Begun DJ. 2007. Molecular population genetics of female-expressed mating-induced serine proteases in *Drosophila melanogaster*. Mol. Biol. Evol. 24:1944–1951.

Lee YH, Ota T, Vacquier VD. 1995. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. Mol. Biol. Evol. 12:231–8.

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

Liao D. 1999. Concerted evolution: molecular mechanism and biological implications. Am. J. Hum. Genet. 64:24-30.

Machado CA, Matzkin LM, Reed LK, Markow TA. 2007. Multilocus nuclear sequences reveal intra- and interspecific relationships among chromosomally polymorphic species of cactophilic *Drosophila*. Mol. Ecol. 2007 16:3009–3024.

Markow TA, Ankney PF. 1984. *Drosophila* Males Contribute to Oogenesis in a Multiple Mating Species. Science 224:302–303.

Markow TA and Ankney PF. 1988. Insemination Reaction in *Drosophila* found in species whose males contribute material to oocytes before fertilization. Evolution 42:1097–1101.

Markow TA. 1996. Evolution of *Drosophila* mating systems. Evol. Biol. 29:73–106.

Markow TA. 2002. Perspective: female remating, operational sex ratio, and the arena of sexual selection in *Drosophila* species. Evolution 56:1725–1734.

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

Meiklejohn CD, Kim Y, Hartl DL, Parsch J. 2004. Identification of a locus under complex positive selection in *Drosophila simulans* by haplotype mapping and composite-likelihood estimation. Genetics 168:265–279.

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

Mueller JL, Page JL, Wolfner MF. 2007. An ectopic expression screen reveals the protective and toxic effects of *Drosophila* seminal fluid proteins. Genetics. 175:777–83.

Moy GW, Springer SA, Adams SL, Swanson WJ, Vacquier VD. 2008. Extraordinary intraspecific diversity in oyster sperm bindin. Proc. Natl. Acad. Sci. U. S. A. 105:1993–8.

Mueller JL, Ravi Ram K, McGraw LA, Bloch Qazi MC, Siggia ED, Clark AG, Aquadro CF, Wolfner MF. 2005. Cross-species comparison of Drosophila male accessory gland protein genes. Genetics 171:131–43.

Nachman MW. 1998. Deleterious mutations in animal mitochondrial DNA. Genetica 102/103: 61–69.

Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.

Otto SP, Yong P. 2002. The evolution of gene duplicates. Adv. Genet. 46:451–83.

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

Panhuis TM, Clark NL, Swanson WJ. 2006. Rapid evolution of reproductive proteins in abalone and *Drosophila*. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 361:261–8.

Parker, GA. 1979. Sexual selection and sexual conflict. In Blum MS, Blum NA, editors. Sexual selection and reproductive competition in insects. London: Academic Press. pp. 123–166.

Patterson JT. 1946. A new type of isolating mechanism *in Drosophila*. Proc. Nat. Acad. Sci. U. S. A. 32:202–208.

Perona JJ, Craik CS. 1995. Structural basis of substrate specificity in the serine proteases. Protein Sci. 4:337–360.

Pitnick S, Spicer GS and Markow TA. 1997. Phylogenetic examination of female incorporation of ejaculate in *Drosophila*. Evolution 51:833–845.

Pitnick S, García-González F. 2002. Harm to females increases with male body size in *Drosophila melanogaster*. Proc. Biol. Sci. 269:1821–8.

Pitnick S, Miller GT, Schneider K, Markow TA. 2003. Ejaculate-female coevolution in *Drosophila mojavensis*. Proc. Nat. Acad. Sci. U. S. A. 270:507–512.

Polgar L. 2005. The catalytic triad of serine peptidases. Cell. Mol. Life. Sci. 62:2161–2172.

Proulx SR, Phillips RC. 2006. Allelic divergence precedes and promotes gene duplication. Evolution 60:881–892.

Prokupek A, Hoffmann F, Eyun SI, Moriyama E, Zhou M, Harshman L. 2008. An evolutionary expressed sequence tag analysis of *Drosophila* spermatheca genes. Evolution 62:2936–2647.

Reed LK, Nyboer M, Markow TA. 2007. Evolutionary relationships of *Drosophila mojavensis* geographic host races and their sister species *Drosophila arizonae*. Mol. Ecol. 16:1007–1022.

Rice WR. 1996. Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. Nature 381:232–234.

Rozas J, Gullaud M, Blandin G, Aguadé M. DNA variation at the rp49 gene region of *Drosophila simulans*: evolutionary inferences from an unusual haplotype structure. Genetics 158:1147–55.

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

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

Sawyer SA. 1989 Statistical tests for detecting gene conversion. Mol Biol Evol 6:526–538.

Schwede T, Kopp J, Guex N, Peitsch MC. 2003. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res. 31:3381–3385.

Spofford, JB. 1969. Heterosis and evolution of duplications. Am. Nat. 103:407–432.

Springer SA, Moy GW, Friend DS, Swanson WJ, Vacquier VD. 2008. Oyster sperm bindin is a combinatorial fucose lectin with remarkable intra-species diversity. Int. J. Dev. Biol. 52:759-768.

Srinivasan A, Giri AP, Gupta VS. 2006. Structural and functional diversities in *lepidopteran* serine proteases. Cell. Mol. Biol. Lett. 11:132–154.

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proc. Nat. Acad. Sci. U. S. A. 13:7375–7379.

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

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

Swofford DL. 2000. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sunderland, Massachusetts: Sinauer Associates.

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

Thornton K, Long M. 2005. Excess of amino acid substitutions relative to polymorphism between X-linked duplications in *Drosophila melanogaster*. Mol. Biol. Evol. 22:273-284.

Thornton KR. 2007. The neutral coalescent process for recent gene duplications and copy-number variants. Genetics 177:987–1000.

Wagstaff BJ, Begun DJ. 2005. Molecular population genetics of accessory gland protein genes and testis-expressed genes in *Drosophila mojavensis* and *D. arizonae.* Genetics 171:1083–1010.

Wagstaff BJ. Begun DJ. 2007. Adaptive evolution of recently duplicated accessory gland protein genes in desert *Drosophila.* Genetics 177:1023–1030.

Walsh B. 2003. Population-genetic models of the fates of duplicate genes. Genetica 118:279–94.

Walsh JB. 1987. Sequence-dependent gene conversion: can duplicated genes diverge fast enough to escape conversion? Genetics. 117:543-557.

Wigby S, Chapman T. 2005. Sex peptide causes mating costs in female *Drosophila melanogaster.* Curr. Biol. 2005 15:316–321.

Wolfner MF. 2007. "S.P.E.R.M." (seminal proteins (are) essential reproductive modulators): the view from *Drosophila*. Soc. Reprod. Fertil. Suppl. 183–99.

Wong A, Turchin MC, Wolfner MF, Aquadro CF. 2008. Evidence for positive selection on *Drosophila melanogaster* seminal fluid protease homologs. Mol. Biol. Evol. 25:497–506.

Yang Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555–556.

Yang Z, Swanson WJ, Vacquier VD. 2000. Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. Mol. Biol. Evol. 17:1446–1455.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22:1107–1118.

TABLES

	all		*no conversion*	
	Zns	***Za***	***Zns***	***Za***
FRP-A				
Baja Peninsula	.71*	0.81**	NA	NA
Catalina Island	0.42	0.52	NA	NA
Mainland Sonora	0.47	0.53	NA	NA
Mojave Desert	0.81**	0.90***	NA	NA
FRP-B				
Baja Peninsula	0.62	0.74*	0.49	0.60
Catalina Island	0.77**	0.96***	0.78*	0.92*
Mainland Sonora	0.39	0.6	0.31	0.47
Mojave Desert	.77**	0.91**	0.74*	0.94**
FRP-C				
Baja Peninsula	0.38	0.54	0.36	0.54
Catalina Island	NA		NA	NA
Mainland Sonora	0.73	0.75*	1.00***	1.00***
Mojave Desert	NA		NA	NA
FRP-D				
Baja Peninsula	0.91**	0.96***	0.90***	0.96***
Catalina Island	NA	NA	NA	NA
Mainland Sonora	0.67	0.80*	0.31	0.24
Mojave Desert	NA	NA	NA	NA
FRP-E				
Baja Peninsula	0.85**	0.85**	0.83**	0.84**
Catalina Island	0.71	0.95**	NA	NA
Mainland Sonora	0.38	0.47	NA	NA
Mojave Desert	0.91**	0.93**	NA	NA

Table 1. Linkage Disequilibrium. *Zns* (Kelly 1997) and *Za* (Rozas 2001) were calculated for all combinations of populations and loci in DnaSP (Rozas and Rozas 1995). Values for all sites, as well as values where sites with evidence for gene conversion are reported. NA indicates the statistic was incalculable, or there was no evidence for gene conversion. * denotes $p < 0.05$. ** denotes $p < 0.01$. *** denotes $p < 0.001$.

	Tajima's D	*Fu and Li's F**	*Fu and Li's F*	*Fay and Wu's H*
FRP-A				
Baja Peninsula	-0.28	0.13	0.21	-6.93
Catalina Island	0.69	1.05	1.49	0.19
Mainland Sonora	0.42	0.42	0.52	0.14
Mojave Desert	1.30	1.02	1.625*	-3.71
FRP-B				
Baja Peninsula	-0.76	0.08	-0.63	-5.71
Catalina Island	0.47	0.64	0.87	-2.61
Mainland Sonora	0.5	0.82	0.88	1.52
Mojave Desert	0.46	0.69	1.05	-5.14
FRP-C				
Baja Peninsula	-1.3	-1.49	-1.7	-2.48
Catalina Island	0.21	0	-0.18	0.38
Mainland Sonora	1.57*	1.395	1.6554	-0.9048
Mojave Desert	-1.359	-1.506	0.01839	0.01
FRP-D				
Baja Peninsula	1.60*	1.37*	1.67	-1.10
Catalina Island	-1.01	-1.10	-1.34	0.24
Mainland Sonora	0.42	0.37	0.61	-.24
Mojave Desert	0.42	0.37	0.47	-1.5
FRP-E				
Baja Peninsula	1.19	1.04	1.55	1.62
Catalina Island	1.52*	1.62*	1.84*	2.52
Mainland Sonora	-0.99	-1.30	-0.51	-5.71*
Mojave Desert	0.82	0.66	0.75	3.24

Table 2. Site Frequency Spectra. Tajima's *D* (Tajima 1983) Fu and Li's *F** (Fu and Li 1993), Fu and Li's *F* (1993), and Fay and Wu's *H* (Fay and Wu 2000) were calculated for all combinations of populations and loci in DnaSP (Rozas *et al* 2003). * denotes $p < 0.05$.

		Standard MK Test			*Lineage-Specific MK Test*		
		Polymorphic	***Fixed***	***Test***	***Polymorphic***	***Fixed***	***Test***
Baja	Syn.+nc	19	7	*G-test*	12	1	*G-test*
Peninsula	Non-Syn.	26	6	NS	19	1	NS
Catalina	Syn.+nc	10	9	*G-test*	6	2	*G-test*
Island	Non-Syn.	8	5	NS	1	4	*
Mainland	Syn.+nc	17	7	*G-test*	10	1	*G-test*
Sonora	Non-Syn.	32	5	NS	25	1	NS
Mojave	Syn.+nc	11	8	*G-test*	7	2	*G-test*
Desert	Non-Syn.	26	5	*	19	1	NS

Table 3. MK-Tests of FRP-A. Silent and replacement variation within populations of *D. mojavensis* were compared to divergence from the *D. arizonae* ortholog. Lineage-specific test was polarized with *D. navajoa* outgroup. * denotes $p < .05$.

		FRP-D Lineage-Specific MK Test			*FRP-E Lineage-Specific MK Test*		
		Polymorphic	***Fixed***	***Test***	***Polymorphic***	***Fixed***	***Test***
Baja	Syn.+nc	20	3	*G-test*	16	12	*G-test*
Peninsula	Non-Syn.	8	11	**	25	9	NS
Catalina	Syn.+nc	0	13	*Fisher's*	7	11	*G-test*
Island	Non-Syn.	1	16	NS	9	12	NS
Mainland	Syn.+nc	27	3	*G-test*	3	12	*G-test*
Sonora	Non-Syn.	23	13	**	8	6	*
Mojave	Syn.+nc	0	11	*Fisher's*	18	10	*G-test*
Desert	Non-Syn.	1	14	NS	21	9	NS

Table 4. Modified MK-Tests of FRP-D and FRP-E. Silent and replacement variation within populations of *D. mojavensis* were compared to divergence from opposing paralog, as in Thornton and Long (2005). Lineage-specific test was polarized with *D. navajoa* outgroup, and the type of contingency test is indicated. * denotes $p < .05$.

FIGURES

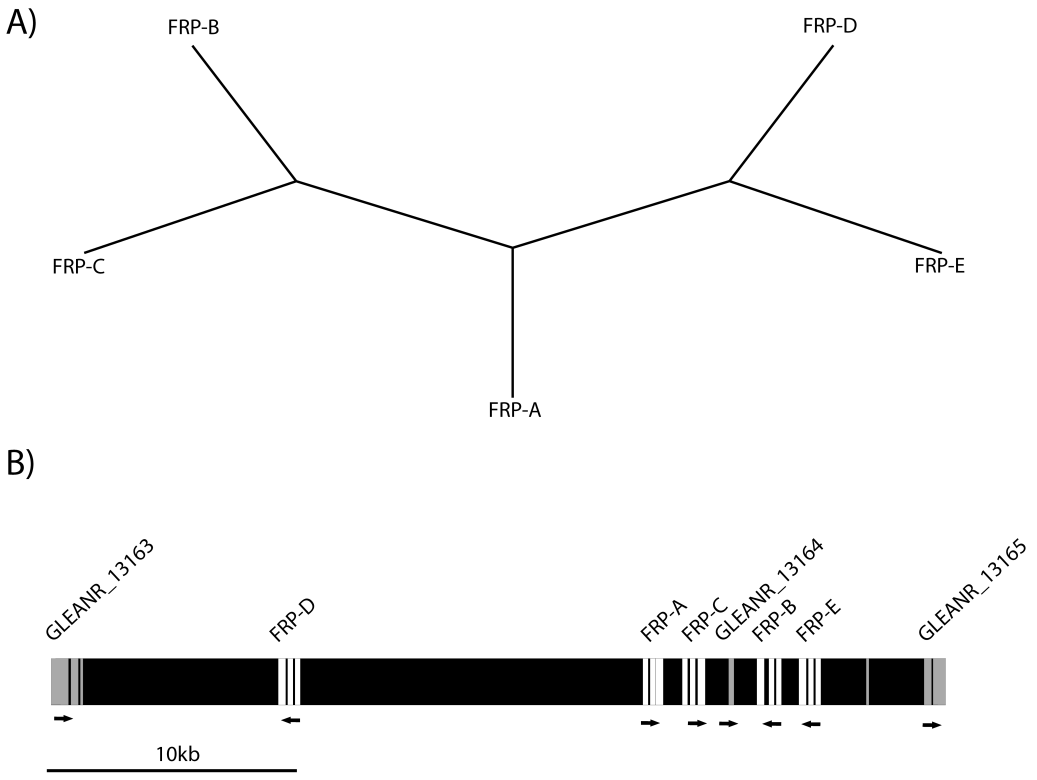


Figure 1. Phylogenetic Relationships and Genomic Arrangement of the Paralogs Examined in this Study. A) Evolutionary relationships between paralogs adapted from Kelleher, Swanson and Markow (2007) B) Genomic arrangement of the protease gene family on *D. mojavensis* chromosome 4 (scaffold_6680 bp 10216565-10169309, http://rana.lbl.gov/drosophila/). White blocks indicate individual exons of duplicated paralogs, while grey blocks indicate individual exons of flanking and interspersed coding sequences. Arrows indicate the direction of transcription.

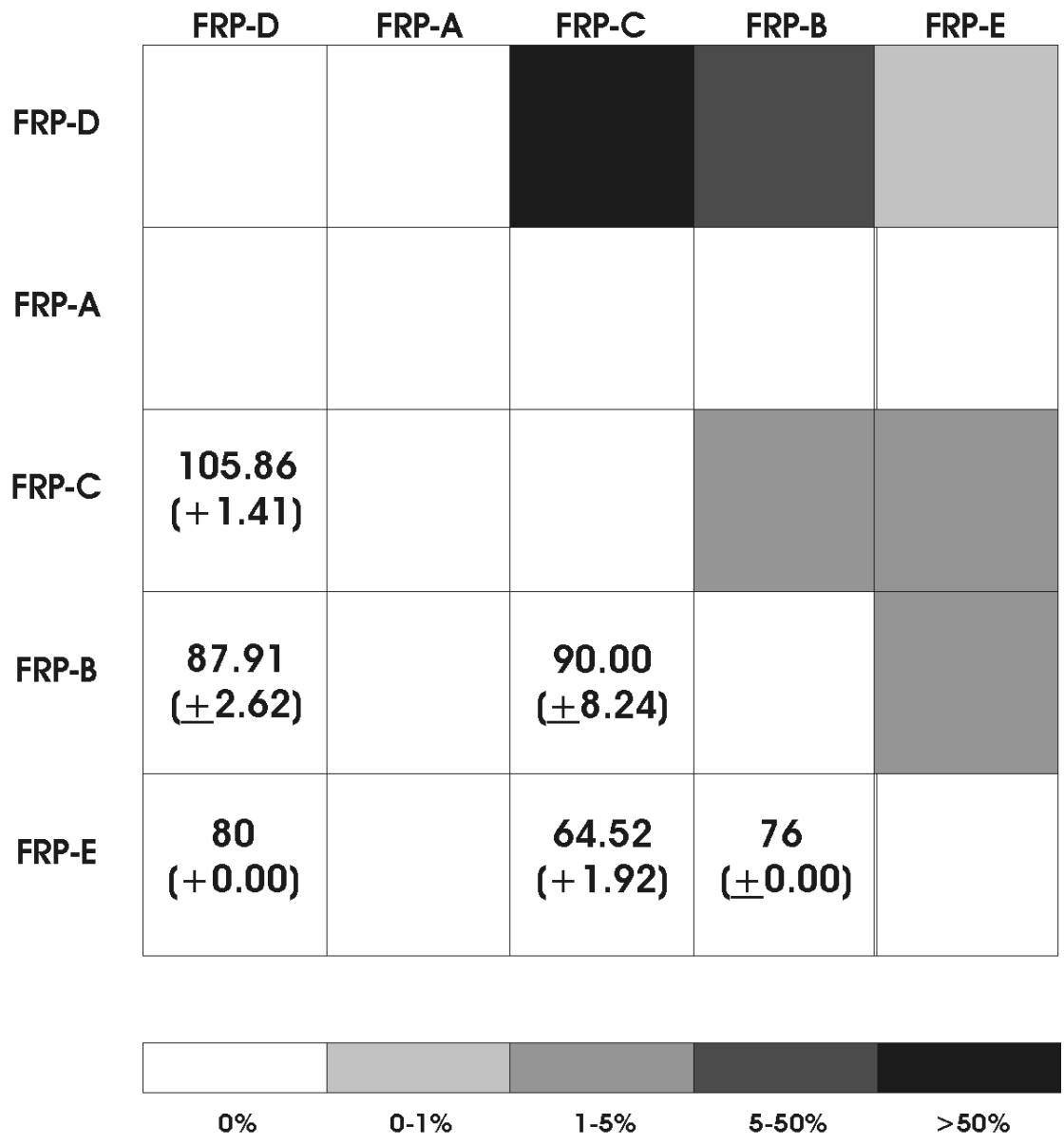


Figure 2. Ectopic Recombination. An alignment of all unique haplotypes was used to detect significant fragments of complete identity in GENECONV, based on the method of Sawyer (1989). The percentage of pairwise comparisons between paralogs that show evidence of gene conversion is indicated by gray shading in the

upper right. The average length of identified conversion tracts between paralogs, and the standard deviation of this estimate, are indicated in the lower left.

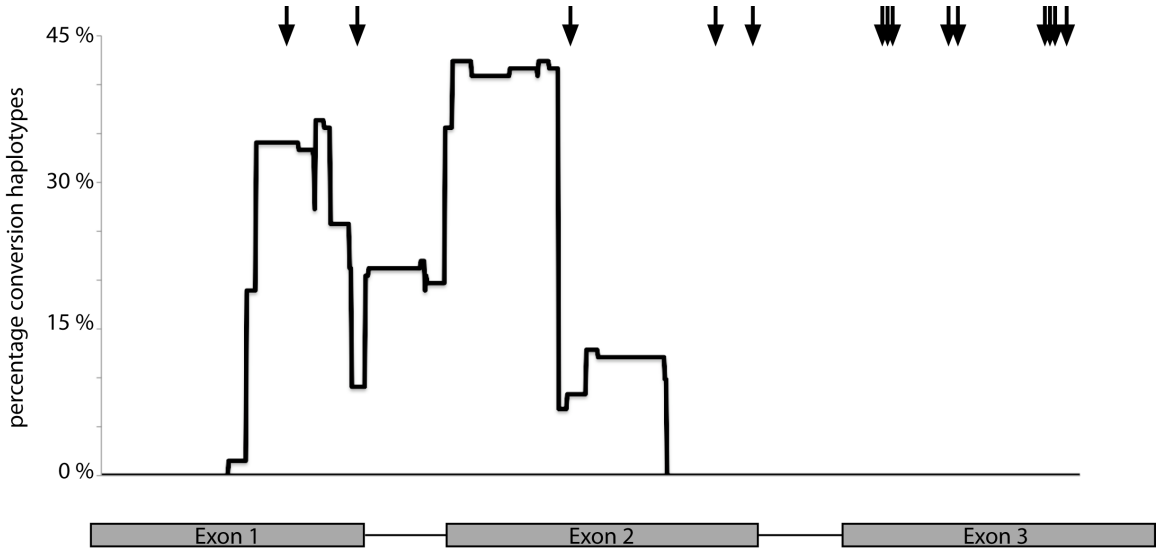


Figure 3. Sliding Window Analysis of Gene Conversion. Y-axis denotes the percentage of haplotypes from the full alignment that show evidence of gene conversion at a particular site. The intron-exon structure is shown to scale along the X-axis, with a total length of ~1 kb. Black arrows denote selected sites from Kelleher, Swanson and Markow (2009).

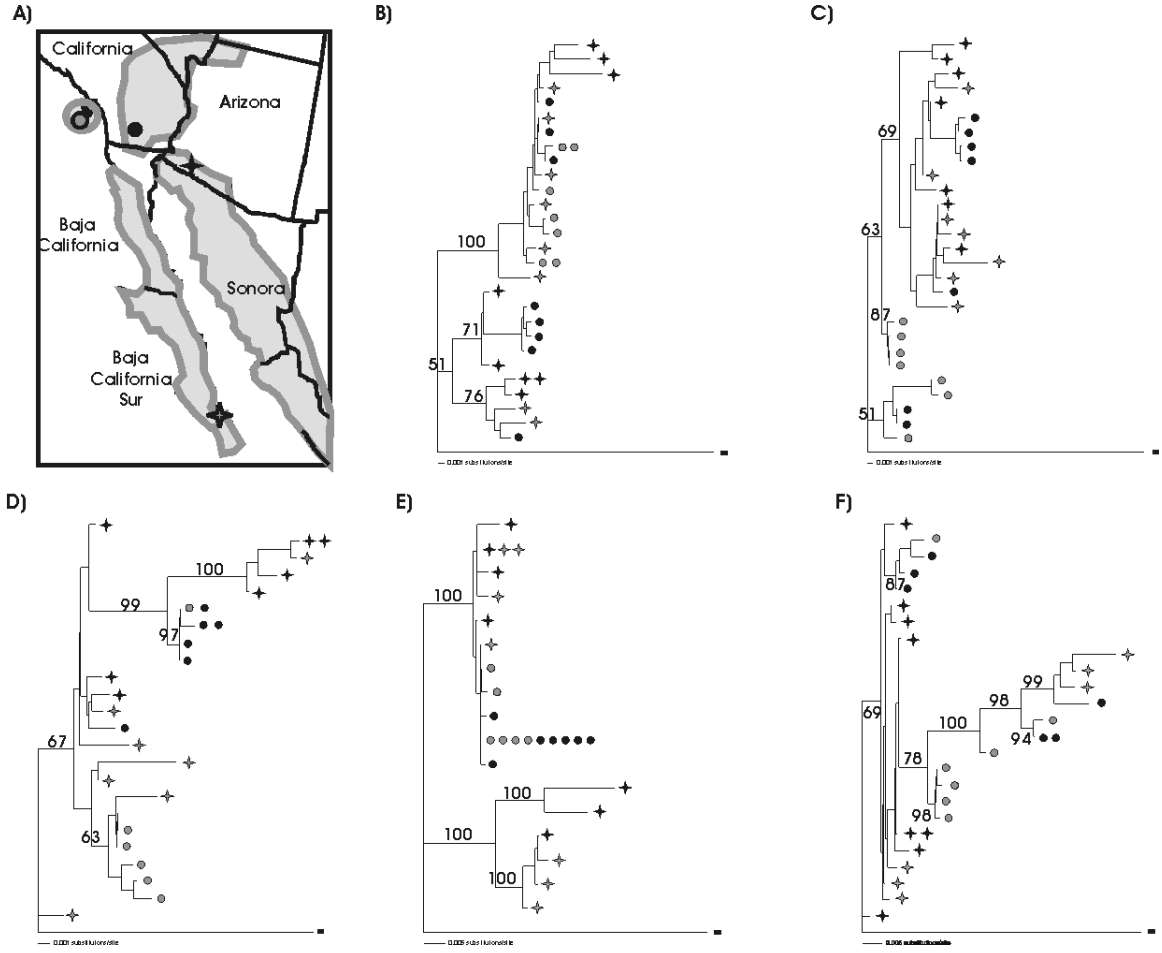


Figure 4. Female Reproductive Proteases Exhibit Unusual Haplotype Structures. A) Geographic distribution and collection sites of four isolated populations of *D. mojavensis*: Baja Peninsula (grey star), Catalina Island (grey circle), Mainland Sonora (black star), and Mojave Desert (black circle). Gene genealogies of FRP-A (B), FRP-B (C), FRP-C (D), FRP-D (E), and FRP-E (F). All genealogies were inferred by neighbor-joining in Paup*4.0b10(Swofford 2000). Symbols indicate population of origin, and number of symbols indicates individual sampled alleles that correspond to that haplotype. Bootstrap values are indicated.

Figure 5. Predicted 3D Structure of FRP-C. Bayes Empirical Bayes selected sites identified under M8 (Yang 1997; Yang, Wong, and Nielsen 2005) were identified in Kelleher, Swanson and Markow (2007). Sites that are determinants of protease inhibitor susceptibility are shown in white (Reviewed in Srinivasan, Giri and Gupta 2006). Sites in grey comprise the catalytic triad (Reviewed in Polgar 2005). Selected sites 124, 244, and 246 also are determinants of inhibitor susceptibility.