Sorted gene genealogies and species-specific nonsynonymous substitutions point to putative postmating prezygotic isolation genes

Suegene Noh, Christopher Garcia, Daniel J Howard, Jeremy L Marshall

Not all genes contribute equally to reproductive isolation. In the *Allonemobius socius* complex of crickets, reproductive isolation is primarily accomplished via postmating prezygotic barriers. We show that two ~~ejaculate~~ protein-coding genes exhibit patterns of evolution consistent with a putative role as speciation genes. Both genes express male ejaculate proteins transferred to females during copulation ~~and~~ were previously identified through comparative proteomics. We found gene genealogies indicat~~ing~~ advanced degrees of lineage sorting, and fixed nonsynonymous substitutions and elevated ω values on the mutational steps separating species, between both pairs of species, on the haplotype networks of these genes compared to other candidate and control genes. At a contact zone between two members of the species complex, these genes maintained species-specificity of alleles despite ongoing gene flow. The putative speciation genes *arginine kinase* (AK) and *apolipoprotein A-1 binding protein* (APBP) are two of the first examples of sperm maturation, capacitation, and motility ~~related~~ proteins that show evidence of fixed nonsynonymous substitutions between species-specific alleles that may lead to reproductive isolation. Our results show that when speciation is ongoing and insufficient time has passed for nucleotide variation to accumulate, hypothesis testing based on haplotype networks and gene trees are more powerful than sequence-based population genetic metrics at detecting signatures of positive selection that may have led to speciation.

Sorted gene genealogies and species-specific nonsynonymous substitutions point to putative postmating prezygotic isolation genes

Suegene Noh*,[1,2], Christopher Garcia[3], Daniel J. Howard[3,4], Jeremy L. Marshall[2]

[1] Department of Biology, Washington University in St. Louis, St. Louis, Missouri 63130, USA

[2] Department of Entomology, Kansas State University, Manhattan, Kansas 66506, USA

[3] Department of Integrative Biology, University of Colorado Denver, Denver, Colorado 80217, USA

[4] Office of the Executive Vice President and Provost, New Mexico State University, Las Cruces, New Mexico 88003, USA

* Corresponding author: Suegene Noh, Department of Biology, Washington University in St. Louis, One Brookings Drive, Campus Box 1137, St. Louis, Missouri 63130, USA

E-mail: suegene.noh@gmail.com

Running title: Postmating prezygotic isolation genes

Key words: postmating prezygotic isolation, positive selection, lineage sorting, haplotype networks, ejaculate proteins

Word count: 5016

Data Archival Location: Sequences formatted as haplotypes are available from NCBI GenBank PopSets 372477483 (AK), 372477513 (APBP), 372477527 (EJAC-SP), 372477535 (GOT), 372477555 (SPAG6), 372477561 (SPI), 372477571 (ACG69).

Abstract

Not all genes contribute equally to reproductive isolation. In the *Allonemobius socius* complex of crickets, reproductive isolation is primarily accomplished via postmating prezygotic barriers. We show that two ejaculate protein-coding genes exhibit patterns of evolution consistent with a putative role as speciation genes. Both genes express male ejaculate proteins transferred to females during copulation and were previously identified through comparative proteomics. We found gene genealogies indicating advanced degrees of lineage sorting, and fixed nonsynonymous substitutions and elevated ω values on the mutational steps separating species, between both pairs of species, on the haplotype networks of these genes compared to other candidate and control genes. At a contact zone between two members of the species complex, these genes maintained species-specificity of alleles despite ongoing gene flow. The putative speciation genes *arginine kinase* (AK) and *apolipoprotein A-1 binding protein* (APBP) are two of the first examples of sperm maturation, capacitation, and motility related proteins that show evidence of fixed nonsynonymous substitutions between species-specific alleles that may lead to reproductive isolation. Our results show that when speciation is ongoing and insufficient time has passed for nucleotide variation to accumulate, hypothesis testing based on haplotype networks and gene trees are more powerful than sequence-based population genetic metrics at detecting signatures of positive selection that may have led to speciation.

Introduction

~~Not all genes contribute equally to reproductive isolation during speciation.~~ ‘Speciation’ (Wu, 2001; Wu & Ting, 2004; Nosil & Schluter, 2011), ‘isolation’ (Rieseberg, Church & Morjan, 2004), or ‘barrier’ (Noor & Feder, 2006) genes are expected to show very different patterns of evolution compared to genes that are not directly involved in reproductive isolation when species are still undergoing lineage sorting (Wu 2001). Therefore we expect to find putative speciation genes among those genes that become fixed for alternative alleles within each incipient species early in the process of divergence, with said alleles rarely crossing the species boundary in sympatry (Ting, Tsaur & Wu, 2000; Dopman et al., 2005).

Rapidly evolving reproductive proteins that can affect fertilization success have an important role in the evolution of postmating prezygotic reproductive isolation. Many reproductive genes are known to evolve rapidly in a variety of organisms (Civetta & Singh, 1998; Swanson & Vacquier, 2002; Clark, Aagaard & Swanson, 2006; Panhuis & Swanson, 2006; Snook et al., 2009). In *Drosophila* where some of the most extensive work has been done, genes that show male-biased expression evolve faster compared to female-biased and somatically expressed genes (Zhang, Hambuch & Parsch, 2004; Zhang & Parsch, 2005; Metta et al., 2006; Pröschel, Zhang & Parsch, 2006; Haerty et al., 2007), and seminal fluid proteins in particular tend to show an excess of nonsynonymous substitutions (Begun et al., 2000; Swanson et al., 2001; Wagstaff & Begun, 2005; Almeida & DeSalle, 2008). Similar patterns have also been observed in mice and primates (Clark & Swanson, 2005; Karn et al., 2008; Ramm et al., 2008; Turner, Chuong & Hoekstra, 2008; Dean et al., 2009). Using a proteomics approach on insect spermatophores to isolate male reproductive protein coding-genes that can directly interact with female counterparts has proved to be an efficient way of narrowing prospects in the search for putative speciation genes (Andrés, Maroja & Harrison, 2008; Marshall et al., 2011).

The male ejaculate proteome comprises sperm-expressed proteins and seminal fluid proteins. Sperm not only contribute half of the diploid genome, but are also involved in sperm-egg interactions including egg activation and deliver paternal factors during fertilization (Dorus et al., 2006). Seminal fluid proteins, the majority of which are produced by male accessory glands, contain conserved functional classes of peptides and pro-hormones that are involved in sperm binding, proteolysis, lipid metabolism, and immune function (Mueller et al., 2004; Chapman & Davies, 2004; Poiani, 2006; Avila et al., 2011). Once transferred into the female reproductive tract, these proteins can initiate a wide-range of physiological functions including increased egg production and oviposition, decreased receptivity, decreased lifespan, and increased feeding in females (reviewed in (Avila et al., 2011). The interacting female counterparts to these ejaculate proteins (EPs) are not well known (Ram, Ji & Wolfner, 2005; Ram & Wolfner, 2007; Snook et al., 2009) though genomic data is proving to be invaluable for identifying candidates (Findlay et al., 2014). The evolution of EPs has been hypothesized to be driven by one or more processes including female sperm preference, sperm competition, and sexual conflict (Mueller et al., 2004; Snook et al., 2009). Here, we show through multiple lines of evidence that two EP-coding genes in the *Allonemobius socius* complex of crickets show patterns of molecular evolution and gene genealogies consistent with a putative role as speciation genes.

The *A. socius* complex of ground crickets, *A. socius*, *A. fasciatus*, and *A. sp. nov.* Tex, represents a powerful system to explore the hypothesized link between EP divergence and reproductive isolation. Members of this complex are primarily isolated from one another by two postmating, prezygotic phenotypes – conspecific sperm precedence (Gregory & Howard, 1994; Howard et al., 1998a,b; Marshall, 2004) and the superior ability of conspecific males to induce females to lay eggs (Gregory & Howard, 1993; Howard et al., 1998b). Two other compelling features of this organismal system are species boundaries that remain intact in sympatry despite some gene flow (Howard, 1986; Howard & Waring, 1991; Traylor et al., 2008) and the very recent nature of divergence between these species (i.e., within the last 30,000 years; (Marshall, 2004, 2007). Indeed, divergence is so recent that few species-specific alleles have been identified; for example, only 2 of 17 allozyme markers (Howard, 1983), 2 of 5,400 AFLP markers (Howard et al., 2002), ~21 of 1,660 thorax proteins and ~33 of 922 ejaculate proteins (Marshall et al., 2011) and 1 of 16 randomly chosen reproductive genes spanning >7,500 bp of coding sequence (Marshall et al., *unpublished data*), yield evidence of species specificity. Taken together, the above data suggest that while there is sufficient genetic divergence to produce reproductive isolation and maintain species boundaries in sympatry, the vast majority of genes show no evidence of divergence and thus, no lineage sorting. In all, the *A. socius* complex represents a system whereby speciation is ongoing with relatively few genes contributing to the postmating, prezygotic reproductive isolation between species. Therefore, if we can identify those ejaculate and female reproductive tract genes that exhibit signatures of positive selection, and maintain species-specificity in sympatry, we will gain insight into the genes that contribute to reproductive isolation and ultimately are involved in driving speciation.

In this study, we expanded analyses from a previous study comparing EPs between the species *A. socius* and *A. fasciatus* (Marshall et al., 2011) by including more genes and an additional species, *A. sp. nov.* Tex (Traylor et al., 2008). Specifically, longer fragments of the five original proteins (ACG69, AK, APBP, EJAC-SP, SPI) plus two additional EPs (GOT, SPAG6) were compared for patterns of nucleotide variation, evidence of lineage-specific positive selection and different degrees of lineage sorting, and species-specificity of alleles in the contact zone between *A. socius* and *A. fasciatus*. These combined analyses point toward an important role for some but not all examined EPs during the evolution of reproductive isolation within this complex of crickets.

Methods

Background

Striped ground crickets of the *A. socius* complex inhabit moist grasslands across North America and do not show significant habitat isolation (Howard 1986). The three species *A. socius*, *A. fasciatus*, and *A. sp. nov.* Tex form two contact zones, one between *A. fasciatus* (north) and *A. socius* (south) from Illinois to New Jersey (Howard & Waring, 1991), and one between *A. sp. nov.* Tex (west) and *A. socius* (east) near the Louisiana – Texas state line (Traylor et al., 2008). *A. fasciatus* and *A. socius* seem to have diverged from a common ancestor approximately 30,000 years ago, and *A. sp. nov.* Tex seems to have subsequently diverged from *A. socius* approximately 24,000 years ago (Marshall, 2004, 2007). They have previously been shown to be isolated primarily via postmating prezygotic reproductive isolation (Howard et al., 2002; Marshall, 2004; Marshall & DiRienzo, 2012).

Population and gene sampling

Crickets were collected from each population and genotyped in the lab via allozymes (Isocytrate dehydrogenase and Hexokinase) to determine species identity (Howard, 1983, 1986). Sampling localities spanned the range of each species. *A. socius* populations were sampled near

Texarkana, AR (AR), Bottom, NC (Bot), Mt. Vernon, IL (IL), Pleasantville, NJ (Mi), Ruston, LA (LA), Gastonia, NC (NC), and Ardmore, OK (OK). *A. fasciatus* populations were sampled near Akron, OH (Akn), Frankfort, IL (FF), and New Paltz, NY (NP). *A. sp. nov.* Tex populations were sampled near Terrell, TX (Tx20), Royse City, TX (Tx30), and Gainesville, TX (Tx35). Contact zone populations of *A. fasciatus* and *A. socius* were sampled from two habitats at a single location in Kenna, WV. *A. fasciatus* was collected from a hillside habitat, which we call Kenna Hill (KH), and *A. socius* was collected along the base of hill near a creek which we call Kenna Creek (KC). We did not have samples from the contact zone between *A. socius* and *A. sp. nov.* Tex. General maintenance protocols followed Marshall et al (2009).

We dissected male accessory glands and testes from three individuals per allopatric population and 9 individuals per contact zone population. cDNA was synthesized from each tissue using RNA isolated via an Ambion RNAqueous-4PCR (#AM1914) kit and standard protocols for 1st strand cDNA synthesis. General PCR and sequencing procedures followed Marshall et al (2011). Standard PCR chemistry was followed with annealing temperatures between 50-60 °C depending on individual primer melting temperatures (primers used are shown in Supplementary Table 1). We compared nucleotide sequences of five candidate EP genes with two control EP genes. ~~Among~~ the five candidate genes, two were chosen based on species-specific proteome profiles (Marshall et al., 2011): 1) *arginine kinase* (AK), a phosphotransferase enzyme expressed in the sperm that may be involved in sperm motility, capacitation or the acrosome reaction (Strong & Ellington, 1993; Niksirat et al., 2015); 2) *apolipoprotein A-1 binding protein* (APBP), a phosphoprotein expressed in sperm and hypothesized to be involved in sperm capacitation (Jha et al., 2008). Two were chosen based on previous sequencing data showing species-specific molecular variation: 3) *ejaculate serine protease* (EJAC-SP), an abundant accessory gland-expressed serine protease previously shown to be involved in the induction of egg laying in successfully mated females (Marshall et al., 2009); 4) *aspartate aminotransferase* (GOT), a pyridoxal-phospate-dependent aminotransferase expressed in the testis and an allozyme historically used to diagnose species identity among *A. socius* complex crickets (Howard, 1983, 1986). The last candidate gene was chosen based on a review of sperm biology literature: 5) *sperm-associated antigen 6* (SPAG6), important for sperm flagellar motility and the structural integrity of the central apparatus (Neilson et al., 1999; Sapiro et al., 2002).

The control genes had non species-specific proteome profiles (Marshall et al., 2011) and were: 6) *serpine inhibitor* (SPI), a testis-expressed serine-type endopeptidase inhibitor; 7) *acg69* (ACG69), a protein of unknown function expressed in the accessory glands. Sequences formatted as haplotypes are available from NCBI GenBank PopSets 372477483 (AK), 372477513 (APBP), 372477527 (EJAC-SP), 372477535 (GOT), 372477555 (SPAG6), 372477561 (SPI), 372477571 (ACG69).

Sequence evolution-based analyses

Male biased genes have been shown to exhibit patterns of molecular evolution associated with relaxed selective constraints or strong positive selection, such as higher rates of nonsynonymous substitutions (Zhang, Hambuch & Parsch, 2004). We investigated multiple metrics of molecular sequence evolution to test for evidence of selection and a departure from neutral sequence evolution. We applied Tajima's D and Fu and Li's D tests to each gene to look for evidence of departure from neutral allelic distributions within species (Tajima, 1989; Fu & Li, 1993). We compared polymorphism within species to divergence between species using HKA tests (Hudson, Kreitman & Aguadé, 1987), and tested for differences in these ratios at each branching node of the species tree.

Next, we compared polymorphism and divergence between synonymous and nonsynonymous sites within each gene at each branching node of the species tree. We compared ω, the rate ratios of synonymous substitutions per synonymous site κ_a (d_N) and nonsynonymous substitutions per nonsynonymous site κ_s (d_S). We used McDonald-Kreitman tests to compare the ratio of nonsynonymous to synonymous intraspecific polymorphisms to the ratio of nonsynymous to synonymous fixed differences between species (McDonald, Kreitman & others, 1991). All tests were based on sequences aligned in BioEdit v.7.0.5.3 (Hall, 1999) and metrics calculated using DnaSP v.5.10.01 (Librado & Rozas, 2009). For HKA tests, we used the program hka provided by Jody Hey (Wang & Hey, 1996).

Gene genealogy-based analyses

Evolutionary relationships between species are tested with phylogenetic trees while hypotheses of intraspecific relationships benefit from haplotype network-based approaches (Posada & Crandall, 2001). Because our species are recently diverged, we used both tree-based and haplotype network-based analyses to detect interesting patterns of gene evolution.

We used statistical parsimony haplotype networks (Templeton, Crandall & Sing, 1992) of alleles from all three species to test for species-specificity of alleles. We used TCS (Clement, Posada & Crandall, 2000) to generate the haplotype networks using only allopatric individuals. Species-specific alleles were defined as those found only within each respective species. Common or shared alleles were those observed in more than one species. Once alleles were designated common or specific to a species, we turned to the *fasciatus* - *socius* contact zone. We looked at nine individuals each of contact zone *A. fasciatus* and *A. socius* and determined what types of allele these contact zone individuals possessed. As noted above, these individuals had previously been designated as fully (homozygous) *A. fasciatus* or *A. socius* based on allozymes. We used Fisher's exact tests with Freeman-Halton extensions for 2x3 contingency tables to determine the probability of observing the distribution of *fas* vs. *soc* vs. shared alleles for each gene.

We tested for lineage-specific positive selection on individual gene tree topologies using the Genetic Algorithm (GA) Branch method (Pond & Frost, 2005) via the Datamonkey webserver of the HyPhy package (Delport et al., 2010). GA Branch uses a genetic algorithm that allows estimates of the nonsynonymous to synonymous substitution rate ratio ($d_N/d_S = \kappa_a/\kappa_s = \omega$) to vary freely across branches within a phylogeny and compares models with different ω classes. Only allopatric individuals were included in the analysis and neighbor-joining trees used by GA Branch were generated natively within Datamonkey.

The genealogical sorting index (gsi) reflects the degree of lineage sorting of individual gene genealogies that occurs during speciation, with values ranging from zero (complete polyphyly) to 1 (complete monophyly) (Cummings, Neel & Shaw, 2008). We calculated gsi for each gene using the online server (www.genealogicalsorting.org) with gene trees including both allopatric and contact zone individuals. Sequences were phased in DnaSP prior to tree building for all genes except APBP, which had no heterozygous individuals. Sorting is more difficult to observe in phased data. We generated maximum likelihood gene trees with PhyML 3.0 (Guindon et al., 2010) via the Mobyle server (Neron et al., 2009). We used nearest neighbor interchange (NNI) tree search and HKY85 as our nucleotide substitution model. MEGA6 (Tamura et al., 2013) was used to visualize these trees.

Results

Sequence evolution-based analyses

We found a general lack of both synonymous and nonsynonymous nucleotide variation among all EP genes we investigated (Table 1). The Watterson estimator $\theta=4N_e\mu$ ranged from 0.001 to 0.011. Levels of θ in the EP candidate genes were approximately an order of magnitude lower than the control genes, although this difference was not statistically significant (*fas* - Mann-Whitney $U = 0$, $P = 0.051$; *soc* - Mann-Whitney $U = 4$, $P = 0.688$; Tex - Mann-Whitney $U = 4.5$, $P = 0.845$). In no cases were Tajima's D or Fu and Li's D tests significantly different from neutral expectations (Table 1) (all $P > 0.1$).

To compare polymorphism within species to divergence between species, we used a standard multilocus HKA test and HKA outlier tests for each branching event. We included all loci and performed 9999 rounds of coalescent simulations. The multilocus HKA test did not find a significant departure from neutral expectations for the first branching event between *A. fasciatus* and the two other species (X^2 $P = 0.916$). The outlier cell, which was *A. fasciatus* for polymorphism in ACG69, was not significantly different in its pattern of polymorphism to divergence ($P = 0.68$). The multilocus HKA test did find a significant departure from neutral expectations for the second branching event between *A. socius* and *A. sp. nov.* Tex. (X^2 $P = 0.012$). However, the outlier cell, which was polymorphism at GOT in *A. sp. nov.* Tex., was not significantly different in its pattern of polymorphism to divergence ($P = 0.06$).

We compared the rate ratios of nonsynonymous to synonymous substitutions $\omega = \kappa_a/\kappa_s$ at each branching event of the species tree. When ω is larger than 1 and the nonsynonymous substitution rate exceeds the synonymous substitution rate, positive or diversifying selection is inferred. When ω is smaller than 1, negative or purifying selection is inferred. However, ω = 1 is recognized as a conservative threshold because the average κ_a is expected to be much smaller than κ_s given the expectation of widespread purifying selection acting on functional genes (Nielsen, 2001). Therefore the value ω = 0.5 has been suggested as an alternate cutoff for the detection of positive selection as subsequent analyses generally indicate that such genes are indeed under positive selection (Swanson et al., 2004). In none of our genes did ω exceed 1, but in the older split between *A. fasciatus* and the other two species, ω exceeded 0.5 for the genes AK and APBP (Table 2).

We used McDonald-Kreitman tests to compare the ratio of nonsynonymous to synonymous intraspecific polymorphisms (P_N/P_S) to the ratio of nonsynonymous to synonymous fixed differences between species (D_N/D_S). Not all genes had fixed nonsynonymous substitutions between species and in these cases we were unable to apply the McDonald-Kreitman test. For those genes that were testable, we did not find significant differences in D_N/D_S compared to P_N/P_S at either branching event (Fisher's exact test $P = 0.07 \sim 1$) (Table 2).

Gene genealogy-based analyses

The statistical parsimony haplotype networks generated using allopatric individuals of all three species showed the presence of only species-specific alleles in AK, APBP, and GOT, while the other genes had alleles shared between two species each (Figure 1). Within the contact zone between *A. fasciatus* and *A. socius*, AK, APBP, EJAC-SP and SPAG6 had upwards of 16 species-specific alleles out of 18 possible alleles (approximately 88%) (Figure 2). In comparison, many GOT and SPI alleles were shared between the contact zone populations. Fisher's exact tests indicated the allelic distributions were nonrandom for all genes except ACG69.

The GA Branch method detected elevated ω classes in all genes except EJAC-SP and SPAG6 (Table 3). We mapped the substitution rate changes detected by GA Branch onto the haplotype networks (Figure 1). In AK and APBP, elevated ω were detected on mutational steps between both species pairs, between *A. fasciatus* and *A. socius* and between *A. socius* and *A. sp. nov.* Tex, and were associated with one or more fixed nonsynonymous substitutions. In GOT and SPI, elevated ω were detected on mutational steps between *A. fasciatus* and *A. sp. nov.* Tex, and also within *A. sp. nov.* Tex. Elevated ω were detected on several branches in ACG69 (Figure 1).

Comparisons of genealogical sorting index values based on maximum likelihood gene trees including all sampled individuals, both allopatric and contact zone, indicated that only AK and APBP showed advanced lineage sorting for all three species (Table 4, Supplementary Figures 1 & 2). Excluding *A. sp. nov.* Tex, which had high gsi-values overall most likely due to its limited range and lack of data from its contact zone with *A. socius*, the two control genes were unsorted across the species ranges. The remaining three candidate genes showed asymmetrical lineage sorting.

Discussion

Many reproductive genes, and in particular those that are male biased, are known to evolve rapidly, often exhibiting higher rates of nonsynonymous substitutions (Zhang, Hambuch & Parsch, 2004), reduced codon usage bias (Hambuch & Parsch, 2005), and evidence suggesting they are more likely to evolve by duplication (Ellegren & Parsch, 2007). However recent divergence can hinder the application of many metrics of molecular evolution that rely on sequence variation since not enough evolutionary time has passed to allow for differences to accumulate between incipient species. Thus data from relatively recently (~30,000 years) diverged species such as ours show a general lack of both synonymous and nonsynonymous nucleotide variation among all investigated genes (Table 1 & 2). In addition, our estimates of sequence variation were also at least an order of magnitude smaller compared to other known estimates from accessory gland protein coding genes in various other species groups (Mueller et al., 2005; Wagstaff & Begun, 2005; Almeida & DeSalle, 2008), including some *Gryllus* crickets whose species are of roughly similar age (Andrés et al., 2006). Therefore, while relatively young species offer an opportunity to observe the ongoing process of the genetics of species divergence, attempting to identify putative speciation genes based on DNA sequences requires an approach that takes into account gene trees and haplotype networks, along with species trees.

The ratio of nonsynonymous to synonymous substitution rates ω is commonly used to detect signatures of selection acting upon protein coding genes (Yang & Bielawski, 2000; Nielsen, 2001, 2005; Jensen, Wong & Aquadro, 2007). The original intended application of ω was for the analysis of sequence evolution among distantly related species, though in practice, it is not uncommonly applied to sequences between closely related populations of a species (Kryazhimskiy & Plotkin, 2008). This turns out to be problematic because when sequence evolution under selection was simulated over short evolutionary timescales, representative of genetic variation segregating within a species, vs. long evolutionary timescales, intraspecific ω behaved very differently from interspecific ω (Kryazhimskiy & Plotkin, 2008). In fact, Kryazhimskiy and Plotkin show that under positive selection, ω = 1 when selection was moderate but showed asymptotic behavior and eventually decreased below 1 as the selection coefficient increased. Over short timescales, its variance also increased as θ (= 2Nμ in the paper) became smaller, making it more difficult to accurately detect positive selection.

Another complication with using ω for short evolutionary timescales is that during initial sequence divergence ω could be higher than expected because slightly deleterious nonsynonymous mutations can persist in a population for generations due to genetic hitchhiking, and a time lag before they are removed by purifying selection (Rocha et al., 2006). Based on simulations, Rocha and colleagues showed this is why unexpectedly high ω values are frequently observed among closely related (1 – 2% sequence divergence) bacteria species. As evolutionary time progresses further, they show that synonymous mutations continued to accumulate and exceeded the initial overrepresentation of nonsynonymous mutations. Therefore if ω is estimated too soon after sequence divergence, one would expect to find high rates of false positive detection.

Finally, even at longer evolutionary timescales the assumption that κ_s is effectively neutral may need reconsideration, as a survey of 16 vertebrate genomes indicates that genes with high ω are more strongly influenced by small κ_s rather than large κ_a (Wolf et al., 2009). Similar patterns are observed in *Drosophila* species, where fast evolving genes show a negative correlation between κ_a and synonymous site polymorphism π_s (Andolfatto, 2007; Macpherson et al., 2007; Jensen & Bachtrog, 2010).Thus, the interaction between linkage and selection makes it challenging to distinguish between recurrent positive selection, background selection, and Hill-Robertson effects (Hill & Robertson, 1966; Charlesworth, 1994; Andolfatto, 2007; Charlesworth et al., 2009). Therefore in order to detect adaptive evolution due to positive selection, applying combinations of metrics including ω, Tajima's or Fu and Li's *D* and site frequency spectra, as well as estimates of population size and recombination rates seems necessary (Nielsen, 2005).

We failed to detect positive selection based on ω, and estimates of *D* for all genes compared here were not significantly different from neutral expectations (Tables 1 and 2). While population bottlenecks are thus likely to have contributed to sequence variation patterns since speciation in the *A. socius* complex is thought to coincide with glaciation history (Marshall, 2004, 2007), for our data sequence evolution-based tests are generally inconclusive as to demographic reasons for why our genes might lack nucleotide variation.

Instead, tests based on individual gene genealogies and haplotype networks ~~indicated~~ AK and APBP as putative speciation genes. Within the contact zone of *A. fasciatus* and *A. socius*, AK, APBP, and EJAC-SP show significantly nonrandom patterns of allelic distributions and had no shared alleles (Figure 2). When all allopatric and contact zone individuals were examined, only the genealogies of AK and APBP indicated that these genes were more advanced in their degree of lineage sorting compared to other candidate and control genes (Table 4, Supplementary Figures 1 & 2). These patterns fit models of ongoing speciation in the face of gene flow, where speciation genes are more likely to be fixed early on during lineage divergence (Wu, 2001). Incomplete lineage sorting and introgression have been suggested to be confounding factors in understanding speciation with ongoing gene flow (Machado & Hey, 2003; Broughton & Harrison, 2003; Payseur, 2010). However, speciation genes are more likely to become fixed for species-specific alleles early in the process of speciation and therefore are expected to be relatively exempt from incomplete sorting and subject to reduced introgression. Similar patterns have been observed in *Drosophila*, field crickets, and moths (Ting, Tsaur & Wu, 2000; Dopman et al., 2005; Maroja, Andrés & Harrison, 2009; Andrés et al., 2013; Larson et al., 2013). It is possible that these genes are not the direct targets but rather linked to targets of divergent selection. Because both genes were identified through comparative proteomics (Marshall et al 2011) this seems relatively unlikely, but the genomic regions around these genes should be investigated for evidence of selective sweeps to rule out this possibility.

Many studies of reproductive proteins report evidence of positive selection acting on a subset of the genes examined, in both males (Begun et al., 2000; Swanson et al., 2001; Clark & Swanson, 2005; Wagstaff & Begun, 2005; Andrés et al., 2006; Karn et al., 2008; Ramm et al., 2008; Almeida & DeSalle, 2008; Walters & Harrison, 2010) and females (Swanson et al., 2004; Panhuis & Swanson, 2006; Lawniczak & Begun, 2007; Prokupek et al., 2008; Kelleher & Markow, 2009; Kelleher, Clark & Markow, 2011). However, there are few examples of adaptive reproductive protein evolution leading to reproductive isolation outside of gamete recognition proteins (e.g. (Geyer & Palumbi, 2003; McCartney & Lessios, 2004; Springer & Crespi, 2007). Our putative speciation genes AK and APBP two of the first examples of sperm maturation and capacitation related proteins that show evidence of fixed nonsynonymous substitutions ~~between~~ species-specific alleles ~~leading to~~ reproductive isolation. In contrast to the other genes examined, fixed nonsynonymous substitutions and elevated ω values only on the mutational steps separating species on the haplotype network of APBP, and to a less exclusive extent in AK and GOT, indicate that these EPs may have evolved under positive selection and contribute to the reproductive isolation between these species (Table 3, Figure 1). We had previously observed this pattern between *A. fasciatus* and *A. socius* for both AK and APBP (Marshall et al 2011), but finding the same pattern in the mutational steps between *A. socius* and *A. sp. nov.* Tex with different species-specific nonsynonymous substitutions emphasizes the potential importance of these candidates.

Whether there are functional consequences to the species-specific nonsynonymous substitutions in AK and APBP needs to be investigated further. Since both candidates were identified by proteomic screens, we hypothesize that an interaction between each male EP and the female reproductive tract during capacitation is responsible for the postmating prezygotic isolation observed between the *A. socius* complex species. AK is a phosphagen kinase that catalyzes ATP-regeneration and energy transport in invertebrates and some protozoa (Ellington, 2001; Noguchi, Sawada & Akazawa, 2001; Uda et al., 2006). Insects and other ecdysozoans possess AK as their sole phosphagen system for cellular energy metabolism, and accordingly, arginine phosphate and its phosphagen kinase AK are found primarily in muscles, but also in sperm and compound eyes (Strong & Ellington, 1993; Kucharski & Maleszka, 1998; Ellington, 2001). The possible roles of AK as an EP can be related to sperm motility (Strong and Ellington 1993), capacitation, or the acrosome reaction (Niksirat et al., 2015). Two structural loops and several active sites near them are the proposed interaction interface of AK with the guanidinium groups of its substrates (Zhou et al., 1998; Pruett et al., 2003; Azzi et al., 2004; Clark, Davulcu & Chapman, 2012). As expected for an integral enzyme, the nonsynonymous substitutions we observed do not occur at these specific sites, though they may still influence its activity. APBP becomes phosphorylated during murine sperm capacitation and co-localizes with cholesterol during this process, but its specific function is unknown (Jha et al 2008). It does possess a Rossmann-like fold, indicating an enzymatic role. The nonsynonymous substitutions we observed in APBP occur in the Rossmann-like fold and are hypothesized to influence the activity of its binding site (Marshall et al., 2011).

Conclusions

A. socius complex crickets provide an excellent opportunity to identify patterns of evolution in speciation genes for two major reasons: speciation is incomplete as evidenced by ongoing gene flow in the field, and isolation is through a single type of reproductive isolation barrier (i.e., postmating prezygotic phenotypes). Therefore we looked for genes that contribute to postmating prezygotic isolation and exhibit fixed, or nearly fixed, nonsynonymous substitutions between species as putative speciation genes. We find that when speciation is ongoing, standard population genetics analyses based on θ and ω values are unable to detect signatures of positive selection contributing to fixed differences between species because insufficient time has passed for nucleotide variation to accumulate. Instead, hypothesis testing based on haplotype networks and gene trees proved to be more powerful at identifying putative postmating prezygotic isolation genes with fixed nonsynonymous substitutions between both pairs of species that may have led to speciation.

Acknowledgments

We thank Diana Huestis and Shanda Wheeler for their help in isolating RNA and screening individuals for species status with allozymes. This is contribution no. 12-015-J from the Kansas Agricultural Experiment Station. The authors declare no conflicts of interest.

References

Almeida FC, DeSalle R. 2008. Evidence of adaptive evolution of accessory gland proteins in closely related species of the *Drosophila repleta* group. *Molecular Biology and Evolution* 25:2043–2053.

Andolfatto P. 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. *Genome Research* 17:1755–1762.

Andrés JA, Maroja LS, Bogdanowicz SM, Swanson WJ, Harrison RG. 2006. Molecular evolution of seminal proteins in field crickets. *Molecular Biology and Evolution* 23:1574–1584.

Andrés JA, Larson EL, Bogdanowicz SM, Harrison RG. 2013. Patterns of transcriptome divergence in the male accessory gland of two closely related species of field crickets. *Genetics* 193:501–513.

Andrés JA, Maroja LS, Harrison RG. 2008. Searching for candidate speciation genes using a proteomic approach: seminal proteins in field crickets. *Proceedings of the Royal Society of London B: Biological Sciences* 275:1975–1983.

Avila FW, Sirot LK, LaFlamme BA, Rubinstein CD, Wolfner MF. 2011. Insect seminal fluid proteins: identification and function. *Annual review of entomology* 56:21–40.

Azzi A, Clark SA, Ellington WR, Chapman MS. 2004. The role of phosphagen specificity loops in arginine kinase. *Protein Science* 13:575–585.

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.

Broughton RE, Harrison RG. 2003. Nuclear gene genealogies reveal historical, demographic and selective factors associated with speciation in field crickets. *Genetics* 163:1389–1401.

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

Charlesworth B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genetical Research* 63:213–227.

Charlesworth B, Betancourt AJ, Kaiser VB, Gordo I. 2009. Genetic recombination and molecular evolution. *Cold Spring Harbor Symposium on Quantitative Biology* 74:177–186.

Civetta A, Singh RS. 1998. Sex and speciation: genetic architecture and evolutionary potential of sexual versus nonsexual traits in the sibling species of the *Drosophila melanogaster* complex. *Evolution*:1080–1092.

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

Clark SA, Davulcu O, Chapman MS. 2012. Crystal structures of arginine kinase in complex with ADP, nitrate, and various phosphagen analogs. *Biochemical and Biophysical Research Communications* 427:212–217.

Clark NL, Swanson WJ. 2005. Pervasive adaptive evolution in primate seminal proteins. *PLoS Genetics* 1:e35.

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

Cummings MP, Neel MC, Shaw KL. 2008. A genealogical approach to quantifying lineage divergence. *Evolution* 62:2411–2422.

Dean MD, Clark NL, Findlay GD, Karn RC, Yi X, Swanson WJ, MacCoss MJ, Nachman MW. 2009. Proteomics and comparative genomic investigations reveal heterogeneity in evolutionary rate of male reproductive proteins in mice (*Mus domesticus*). *Molecular Biology and Evolution* 26:1733–1743.

Delport W, Poon AF, Frost SD, Pond SLK. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* 26:2455–2457.

Dopman EB, Pérez L, Bogdanowicz SM, Harrison RG. 2005. Consequences of reproductive barriers for genealogical discordance in the European corn borer. *Proceedings of the National Academy of Sciences of the United States of America* 102:14706–14711.

Dorus S, Busby SA, Gerike U, Shabanowitz J, Hunt DF, Karr TL. 2006. Genomic and functional evolution of the *Drosophila melanogaster* sperm proteome. *Nature genetics* 38:1440–1445.

Ellegren H, Parsch J. 2007. The evolution of sex-biased genes and sex-biased gene expression. *Nature Reviews Genetics* 8:689–698.

Ellington WR. 2001. Evolution and the physiological roles of phosphagen systems. *Annual Review of Physiology* 63:289–325.

Findlay GD, Sitnik JL, Wang W, Aquadro CF, Clark NL, Wolfner MF. 2014. Evolutionary rate covariation identifies new members of a protein network required for *Drosophila melanogaster* female post-mating responses. *PLoS Genetics* 10:e1004108.

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

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

Gregory PJ, Howard DJ. 1993. Laboratory hybridization studies of *Allonemobius fasciatus* and *A. socius* (Orthoptera: Gryllidae). *Annals of the Entomological Society of America* 86:694–701.

Gregory PJ, Howard DJ. 1994. A post-insemination barrier to fertilization isolates two closely related ground crickets. *Evolution* 48:705–710.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* 59:307–321.

Haerty W, Jagadeeshan S, Kulathinal RJ, Wong A, Ram KR, Sirot LK, Levesque L, Artieri CG, Wolfner MF, Civetta A, others. 2007. Evolution in the fast lane: rapidly evolving sex-related genes in *Drosophila*. *Genetics* 177:1321–1335.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In: *Nucleic Acids Symposium Series*. 95–98.

Hambuch TM, Parsch J. 2005. Patterns of synonymous codon usage in *Drosophila melanogaster* genes with sex-biased expression. *Genetics* 170:1691–1700.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genetical Research* 8:269–294.

Howard DJ. 1983. Electrophoretic survey of eastern North American *Allonemobius* (Orthoptera: Gryllidae): evolutionary relationships and the discovery of three new species. *Annals of the Entomological Society of America* 76:1014–1021.

Howard DJ. 1986. A zone of overlap and hybridization between two ground cricket species. *Evolution*:34–43.

Howard DJ, Gregory PJ, Chu J, Cain ML. 1998a. Conspecific sperm precedence is an effective barrier to hybridization between closely related species. *Evolution* 52:511–516.

Howard DJ, Reece PG, Gregory PJ, Chu J, Cain ML. 1998b. The evolution of barriers to fertilization between closely related organisms. In: Howard DJ, Berlocher SH eds.

Endless Forms: Species and Speciation. New York, NY: Oxford University Press, 279–288.

Howard DJ, Marshall JL, Hampton DD, Britch SC, Draney ML, Chu J, Cantrell RG. 2002. The genetics of reproductive isolation: a retrospective and prospective look with comments on ground crickets. *American Naturalist* 159:S8–S21.

Howard DJ, Waring GL. 1991. Topographic diversity, zone width, and the strength of reproductive isolation in a zone of overlap and hybridization. *Evolution*:1120–1135.

Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116:153–159.

Jensen JD, Bachtrog D. 2010. Characterizing recurrent positive selection at fast-evolving genes in *Drosophila miranda* and *Drosophila pseudoobscura*. *Genome Biology and Evolution* 2:371–378.

Jensen JD, Wong A, Aquadro CF. 2007. Approaches for identifying targets of positive selection. *Trends in Genetics* 23:568–577.

Jha KN, Shumilin IA, Digilio LC, Chertihin O, Zheng H, Schmitz G, Visconti PE, Flickinger CJ, Minor W, Herr JC. 2008. Biochemical and structural characterization of apolipoprotein AI binding protein, a novel phosphoprotein with a potential role in sperm capacitation. *Endocrinology* 149:2108–2120.

Karn RC, Clark NL, Nguyen ED, Swanson WJ. 2008. Adaptive evolution in rodent seminal vesicle secretion proteins. *Molecular Biology and Evolution* 25:2301–2310.

Kelleher ES, Clark NL, Markow TA. 2011. Diversity-enhancing selection acts on a female reproductive protease family in four subspecies of *Drosophila mojavensis*. *Genetics* 187:865–876.

Kelleher ES, Markow TA. 2009. Duplication, selection and gene conversion in a *Drosophila mojavensis* female reproductive protein family. *Genetics* 181:1451–1465.

Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. *PLoS Genetics* 4:e1000304.

Kucharski R, Maleszka R. 1998. Arginine kinase is highly expressed in the compound eye of the honey-bee, *Apis mellifera*. *Gene* 211:343–349.

Larson EL, Andrés JA, Bogdanowicz SM, Harrison RG. 2013. Differential introgression in a mosaic hybrid zone reveals candidate barrier genes. *Evolution; International Journal of Organic Evolution* 67:3653–3661.

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

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451–1452.

Machado CA, Hey J. 2003. The causes of phylogenetic conflict in a classic *Drosophila* species group. *Proceedings of the Royal Society of London B: Biological Sciences* 270:1193–1202.

Macpherson JM, Sella G, Davis JC, Petrov DA. 2007. Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in *Drosophila*. *Genetics* 177:2083–2099.

Maroja LS, Andrés JA, Harrison RG. 2009. Genealogical discordance and patterns of introgression and selection across a cricket hybrid zone. *Evolution* 63:2999–3015.

Marshall JL. 2004. The *Allonemobius-Wolbachia* host-endosymbiont system: evidence for rapid speciation and against reproductive isolation driven by cytoplasmic incompatibility. *Evolution* 58:2409–2425.

Marshall JL. 2007. Rapid evolution of spermathecal duct length in the *Allonemobius socius* complex of crickets: species, population and *Wolbachia* effects. *PLoS One* 2:e720.

Marshall JL, Huestis DL, Hiromasa Y, Wheeler S, Oppert C, Marshall SA, Tomich JM, Oppert B, others. 2009. Identification, RNAi knockdown, and functional analysis of an ejaculate protein that mediates a postmating, prezygotic phenotype in a cricket. *PloS one* 4:e7537–e7546.

Marshall JL, Huestis DL, Garcia C, Hiromasa Y, Wheeler S, Noh S, Tomich JM, Howard DJ. 2011. Comparative proteomics uncovers the signature of natural selection acting on the ejaculate proteomes of two cricket species isolated by postmating, prezygotic phenotypes. *Molecular biology and evolution* 28:423–435.

Marshall JL, DiRienzo N. 2012. Noncompetitive gametic isolation between sibling species of a cricket: a hypothesized link between within-population incompatibility and reproductive isolation between species. *International Journal of Evolutionary Biology* 2012:593438.

McCartney MA, Lessios HA. 2004. Adaptive evolution of sperm bindin tracks egg incompatibility in neotropical sea urchins of the genus *Echinometra*. *Molecular Biology and Evolution* 21:732–745.

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

Metta M, Gudavalli R, Gibert J-M, Schlötterer C. 2006. No accelerated rate of protein evolution in male-biased *Drosophila pseudoobscura* genes. *Genetics* 174:411–420.

Mueller JL, Ripoll DR, Aquadro CF, Wolfner MF. 2004. Comparative structural modeling and inference of conserved protein classes in *Drosophila* seminal fluid. *Proceedings of the National Academy of Sciences* 101:13542–13547.

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

Neilson LI, Schneider PA, Van Deerlin PG, Kiriakidou M, Driscoll DA, Pellegrini MC, Millinder S, Yamamoto KK, French CK, Strauss JF. 1999. cDNA cloning and characterization of a human sperm antigen (SPAG6) with homology to the product of the *Chlamydomonas PF16* locus. *Genomics* 60:272–280.

Neron B, Menager H, Maufrais C, Joly N, Maupetit J, Letort S, Carrere S, Tuffery P, Letondal C. 2009. Mobyle: a new full web bioinformatics framework. *Bioinformatics* 25:3005–3011.

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

Nielsen R. 2005. Molecular signatures of natural selection. *Annual Review of Genetics* 39:197–218.

Niksirat H, James P, Andersson L, Kouba A, Kozák P. 2015. Label-free protein quantification in freshly ejaculated versus post-mating spermatophores of the noble crayfish *Astacus astacus*. *Journal of Proteomics* 123:70–77.

Noguchi M, Sawada T, Akazawa T. 2001. ATP-regenerating system in the cilia of *Paramecium caudatum*. *Journal of Experimental Biology* 204:1063–1071.

Noor MAF, Feder JL. 2006. Speciation genetics: evolving approaches. *Nature Reviews Genetics* 7:851–861.

Nosil P, Schluter D. 2011. The genes underlying the process of speciation. *Trends in Ecology & Evolution* 26:160–167.

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

Payseur BA. 2010. Using differential introgression in hybrid zones to identify genomic regions involved in speciation. *Molecular Ecology Resources* 10:806–820.

Poiani A. 2006. Complexity of seminal fluid: a review. *Behavioral Ecology and Sociobiology* 60:289–310.

Pond SLK, Frost SD. 2005. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. *Molecular Biology and Evolution* 22:478–485.

Posada D, Crandall KA. 2001. Intraspecific gene genealogies: trees grafting into networks. *Trends in Ecology & Evolution* 16:37–45.

Prokupek A, Hoffmann F, Eyun S, Moriyama E, Zhou M, Harshman L. 2008. An evolutionary expressed sequence tag analysis of *Drosophila* spermatheca genes. *Evolution* 62:2936–2947.

Pröschel M, Zhang Z, Parsch J. 2006. Widespread adaptive evolution of *Drosophila* genes with sex-biased expression. *Genetics* 174:893–900.

Pruett PS, Azzi A, Clark SA, Yousef MS, Gattis JL, Somasundaram T, Ellington WR, Chapman MS. 2003. The putative catalytic bases have, at most, an accessory role in the mechanism of arginine kinase. *Journal of Biological Chemistry* 278:26952–26957.

Ram KR, Ji S, Wolfner MF. 2005. Fates and targets of male accessory gland proteins in mated female *Drosophila melanogaster*. *Insect Biochemistry and Molecular Biology* 35:1059–1071.

Ramm SA, Oliver PL, Ponting CP, Stockley P, Emes RD. 2008. Sexual selection and the adaptive evolution of mammalian ejaculate proteins. *Molecular Biology and Evolution* 25:207–219.

Ram KR, Wolfner MF. 2007. Seminal influences: *Drosophila* Acps and the molecular interplay between males and females during reproduction. *Integrative and Comparative Biology* 47:427–445.

Rieseberg LH, Church SA, Morjan CL. 2004. Integration of populations and differentiation of species. *New Phytologist* 161:59–69.

Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH, Feil EJ. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of Theoretical Biology* 239:226–235.

Sapiro R, Kostetskii I, Olds-Clarke P, Gerton GL, Radice GL, III JFS. 2002. Male infertility, impaired sperm motility, and hydrocephalus in mice deficient in Sperm-Associated Antigen 6. *Molecular and Cellular Biology* 22:6298–6305.

Snook RR, Chapman T, Moore PJ, Wedell N, Crudgington HS. 2009. Interactions between the sexes: new perspectives on sexual selection and reproductive isolation. *Evolutionary Ecology* 23:71–91.

Springer SA, Crespi BJ. 2007. Adaptive gamete-recognition divergence in a hybridizing *Mytilus* population. *Evolution* 61:772–783.

Strong SJ, Ellington WR. 1993. Horseshoe crab sperm contain a unique isoform of arginine kinase that is present in the midpiece and flagellum. *Journal of Experimental Zoology* 267:563–571.

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *Proceedings of the National Academy of Sciences* 98:7375–7379.

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.

Swanson WJ, Vacquier VD. 2002. The rapid evolution of reproductive proteins. *Nature Reviews Genetics* 3:137–144.

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

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution* 30:2725–2729.

Templeton AR, Crandall KA, Sing CF. 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* 132:619–633.

Ting C-T, Tsaur S-C, Wu C-I. 2000. The phylogeny of closely related species as revealed by the genealogy of a speciation gene, *Odysseus*. *Proceedings of the National Academy of Sciences* 97:5313–5316.

Traylor T, Birand AC, Marshall JL, Howard DJ. 2008. A zone of overlap and hybridization between *Allonemobius socius* and a new *Allonemobius* sp. *Annals of the Entomological Society of America* 101:30–39.

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

Uda K, Fujimoto N, Akiyama Y, Mizuta K, Tanaka K, Ellington WR, Suzuki T. 2006. Evolution of the arginine kinase gene family. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics* 1:209–218.

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

Walters JR, Harrison RG. 2010. Combined EST and proteomic analysis identifies rapidly evolving seminal fluid proteins in *Heliconius* butterflies. *Molecular biology and evolution* 27:2000–2013.

Wang R-L, Hey J. 1996. The speciation history of *Drosophila pseudoobscura* and close relatives: inferences from DNA sequence variation at the *period* locus. *Genetics* 144:1113–1126.

Wolf JB, Künstner A, Nam K, Jakobsson M, Ellegren H. 2009. Nonlinear dynamics of nonsynonymous (*dN*) and synonymous (*dS*) substitution rates affects inference of selection. *Genome Biology and Evolution* 1:308–319.

Wu C-I. 2001. The genic view of the process of speciation. *Journal of Evolutionary Biology* 14:851–865.

Wu C-I, Ting C-T. 2004. Genes and speciation. *Nature Reviews Genetics* 5:114–122.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution* 15:496–503.

Zhang Z, Hambuch TM, Parsch J. 2004. Molecular evolution of sex-biased genes in *Drosophila*. *Molecular Biology and Evolution* 21:2130–2139.

Zhang Z, Parsch J. 2005. Positive correlation between evolutionary rate and recombination rate in *Drosophila* genes with male-biased expression. *Molecular Biology and Evolution* 22:1945–1947.

Zhou G, Somasundaram T, Blanc E, Parthasarathy G, Ellington WR, Chapman MS. 1998. Transition state structure of arginine kinase: implications for catalysis of bimolecular reactions. *Proceedings of the National Academy of Sciences* 95:8449–8454.

Table 1(on next page)

Nucleotide variation within each *A. socius* complex species

Table 1. Nucleotide variation within each *A. socius* complex species with Tajima’s *D*-values. Fu and Li’s *D*-values were similar (not shown)

Table 1. Nucleotide variation within each *A. socius* complex species with Tajima's *D*-values. Fu and Li's *D*-values were similar (not shown)

		within *A. fasciatus*					within *A. socius*					within *A. sp. nov.* Tex				
Gene	Length	n	π_s	π_a	θ_{fas}	*D*	n	π_s	π_a	θ_{soc}	*D*	n	π_s	π_a	θ_{Tex}	*D*
AK	1173	9	0.002	<0.001	0.001	0.975	15	0.004	0.001	0.002	-1.316	6	0.003	0.001	0.002	0.355
APBP	705	9	0.001	0	0.001	-1.088	15	0.005	0	0.001	-0.334	8	0	0	0	n/a
EJAC-SP	726	9	0	0	0	n/a	16	0.001	<0.001	0.001	-1.311	9	0.003	0	0.001	1.401
GOT	1122	9	0.002	0	<0.001	0.986	17	0	0	0	n/a	9	0.005	0.001	0.002	0.578
SPAG6	426	9	0	0	0	n/a	17	0	0	0	n/a	8	0.005	0	0.001	1.167
SPI	315	9	0.007	0	0.002	-1.362	16	0	0	0	n/a	9	0.008	0.001	0.002	0.196
ACG69	414	9	0.005	0.004	0.007	-1.286	14	0.021	0.009	0.011	0.264	7	0	0	0	n/a

Table 2(on next page)

Nucelotide variation at each branching node of the *A. socius* complex species tree.

Table 2. Nucelotide variation at each branching node of the *A. socius* complex species tree. (P_N: nonsynonymous polymorphisms; P_S: synonymous polymorphisms; D_N: nonsynonymous fixations; D_S: synonymous fixations; κ_s: rate of nonsynonymous substitutions per nonsynonymous site; κ_a: rate of synonymous substitutions per synonymous site; $\omega = \kappa_a / \kappa_s$)

Table 2. Nucelotide variation at each branching node of the *A. socius* complex species tree. (P_N: nonsynonymous polymorphisms; P_S: synonymous polymorphisms; D_N: nonsynonymous fixations; D_S: synonymous fixations; κ_s: rate of nonsynonymous substitutions per nonsynonymous site; κ_a: rate of synonymous substitutions per synonymous site; $\omega = \kappa_\mathrm{a} / \kappa_\mathrm{s}$)

		between fas & (soc+Tex)							between soc & Tex						
Gene	Length	P_N	P_S	D_N	D_S	κ_s	κ_a	ω	P_N	P_S	D_N	D_S	κ_s	κ_a	ω
AK	1173	3	12	2	0	0.006	0.003	0.557	2	9	1	2	0.011	0.002	0.206
APBP	705	1	4	1	0	0.005	0.003	0.523	0	3	1	0	0.007	0.002	0.278
EJAC-SP	726	1	3	0	0	0.008	0.001	0.131	1	2	0	1	0.014	0.002	0.123
GOT	1122	3	7	1	1	0.011	0.002	0.142	3	4	0	2	0.011	0.001	0.116
SPAG6	426	0	1	0	2	0.029	0	0	0	1	0	0	0.004	0	0
SPI	315	3	4	0	0	0.02	0.003	0.15	1	1	2	3	0.054	0.009	0.158
ACG69	414	7	6	0	0	0.024	0.009	0.374	7	6	0	0	0.021	0.007	0.317

Table 3(on next page)

Tests of lineage-specific positive selection

Table 3. Tests of lineage-specific positive selection that detects different ω classes along branches of a gene tree. The model with best c-AIC score is shown.

Table 3. Tests of lineage-specific positive selection that detects different ω classes along branches of a gene tree. The model with best c-AIC score is shown.

	Best model found by GA Branch		
Gene	c-AIC	Classes	ω classes
AK	3539.54	3	1: 0, 2: 0.148, 3: >>1
APBP	2018.52	2	1: 0, 2: >>1
EJAC-SP	2068.19	1	1: 0.079
GOT	3218.71	2	1: 0, 2: 0.545
SPAG6	1208.88	1	1: <0.001
SPI	955.82	2	1: 0, 2: 0.822
ACG69	1377	2	1: 0.081, 2: >>1

Table 4(on next page)

Genealogical sorting index values based on individual gene trees

Table 4. Genealogical sorting index values based on individual gene trees (see Supplementary Figure 1). Values range from zero (complete polyphyly) to one (complete monophyly).

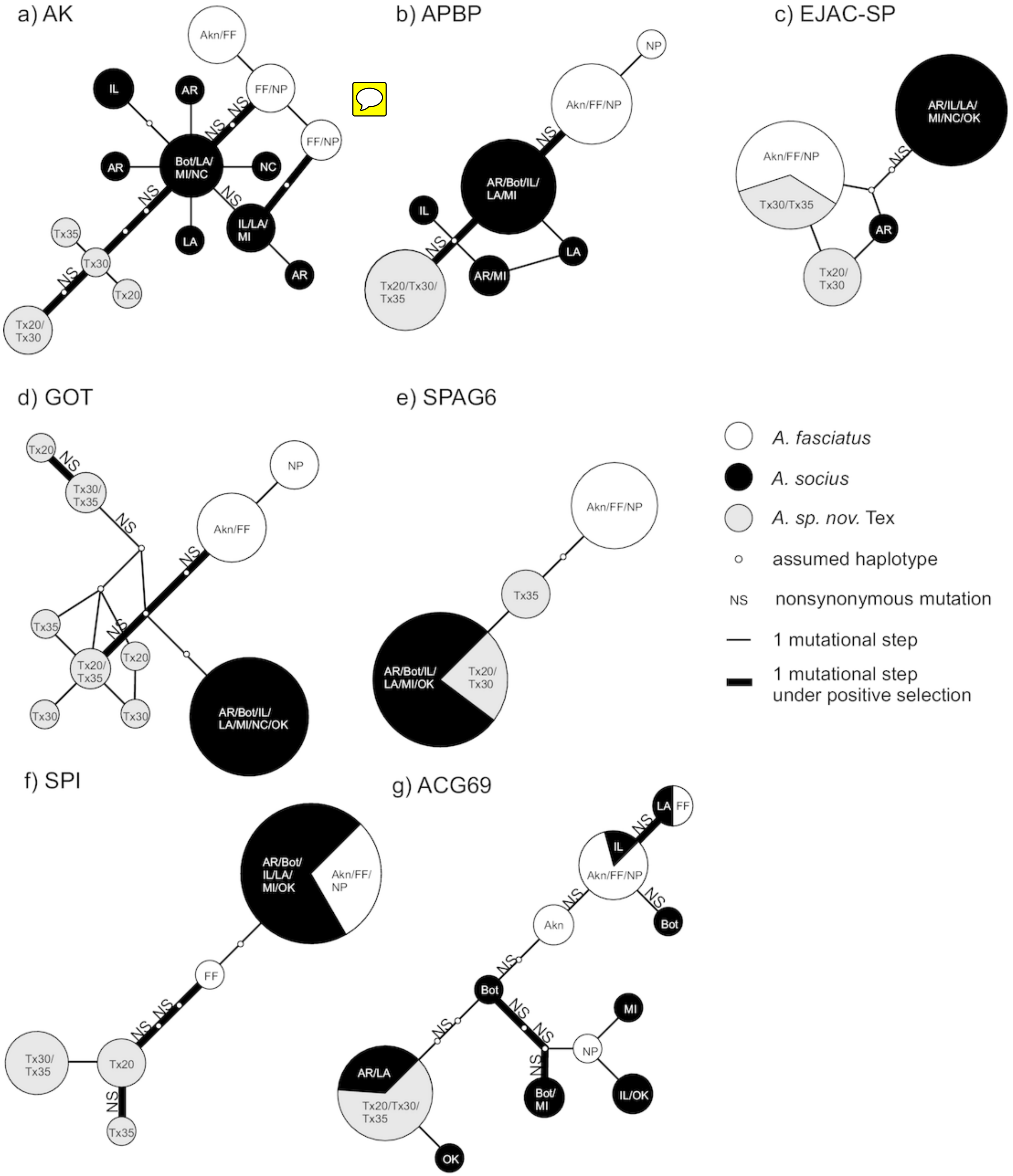
Table 4. Genealogical sorting index values based on individual gene trees (see Supplementary Figure 1). Values range from zero (complete polyphyly) to one (complete monophyly).

Gene	gsi-fas	P_{perm}	gsi-soc	P_{perm}	gsi-Tex	P_{perm}
AK	0.956	<0.001	0.919	<0.001	0.906	<0.001
APBP	1	<0.001	0.849	<0.001	1	<0.001
EJACSP	0.663	<0.001	0.732	<0.001	0.728	<0.001
GOT	0.919	<0.001	0.630	<0.001	0.934	<0.001
SPAG6	0.956	<0.001	0.670	<0.001	0.753	<0.001
SPI	0.140	0.001	0.339	<0.001	0.934	<0.001
ACG69	0.596	<0.001	0.05	0.644	0.917	<0.001

1

Statistical parsimony haplotype networks for all 7 genes from allopatric individual only

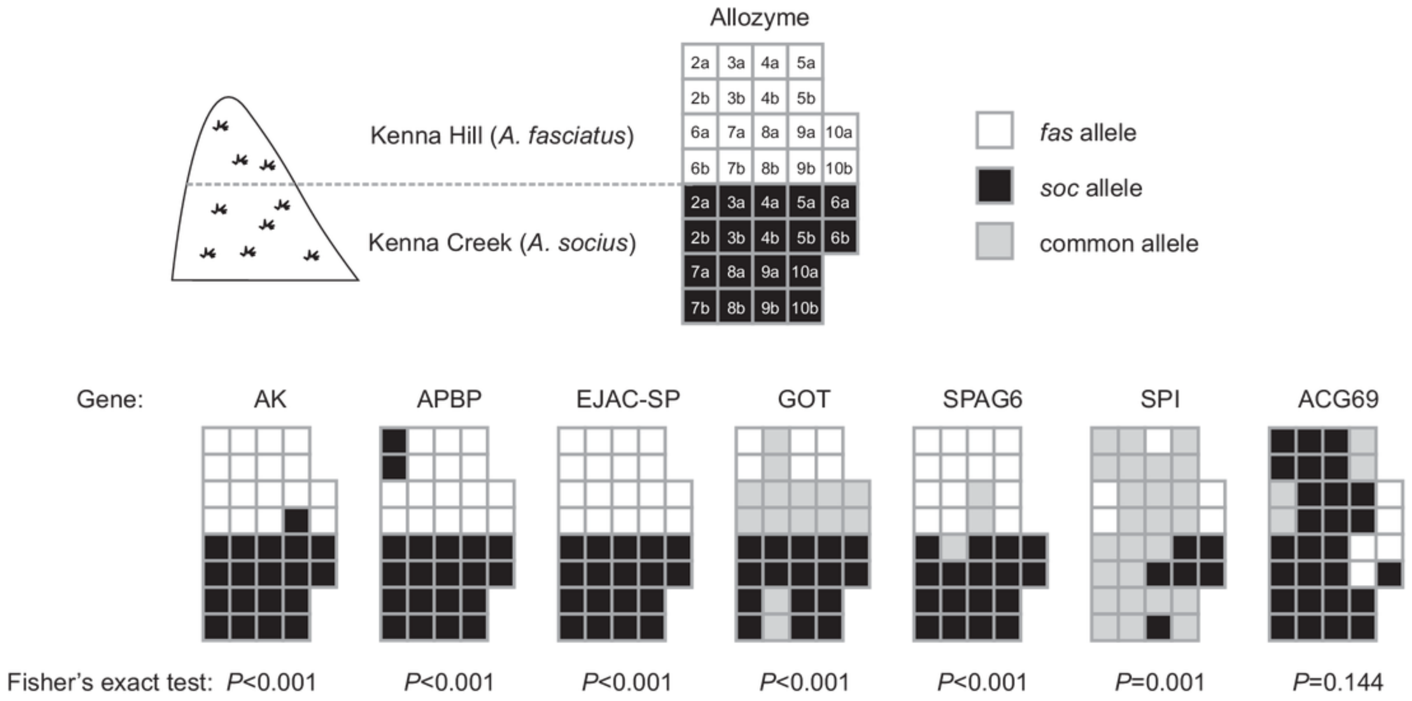
Figure 1. Statistical parsimony haplotype networks for all 7 genes (a-e: test; f-g: control) from allopatric individuals only, with nonsynonymous substitutions and mutational steps with elevated ω indicated. When more than two rate classes were detected (AK, see Table 1), only the largest rate class is indicated as a mutational step under positive selection. Population abbreviations are as in the main text.


a) AK
Akn/FF
FF/NP
FF/NP
IL
AR
AR
Bot/LA/MI/NC
NC
LA
IL/LA/MI
AR
Tx35
Tx30
Tx20
Tx20/Tx30
NS
b) APBP
NP
Akn/FF/NP
AR/Bot/IL/LA/MI
IL
LA
AR/MI
Tx20/Tx30/Tx35
NS
c) EJAC-SP
AR/IL/LA/MI/NC/OK
Akn/FF/NP
Tx30/Tx35
AR
Tx20/Tx30
NS
d) GOT
Tx20
Tx30/Tx35
NP
Akn/FF
Tx35
Tx20/Tx35
Tx20
Tx30
Tx30
AR/Bot/IL/LA/MI/NC/OK
NS
e) SPAG6
Akn/FF/NP
Tx35
AR/Bot/IL/LA/MI/OK
Tx20/Tx30
A. fasciatus
A. socius
A. sp. nov. Tex
assumed haplotype
NS nonsynonymous mutation
1 mutational step
1 mutational step under positive selection
f) SPI
AR/Bot/IL/LA/MI/OK
Akn/FF/NP
FF
Tx20
Tx30/Tx35
Tx35
NS
g) ACG69
LA
FF
IL
Akn/FF/NP
Akn
Bot
Bot
MI
NP
Bot/MI
IL/OK
AR/LA
Tx20/Tx30/Tx35
OK
NS

2

Distribution of species-specific vs. common (shared) alleles within the *A. fasciatus* and *A. socius* contact zone in Kenna, WV

Figure 2. Distribution of species-specific vs. common (shared) alleles within the *A. fasciatus* and *A. socius* contact zone in Kenna, WV. Nine individuals each with allozyme identities of pure (homozygous) *A. fasciatus* and *A. socius* had varying patterns of allelic identities for the seven genes. Numbers (2-9) indicate the sampled individual and letters (a & b) indicate the alleles within each individual.


Allozyme
2a 3a 4a 5a
2b 3b 4b 5b
6a 7a 8a 9a 10a
6b 7b 8b 9b 10b
2a 3a 4a 5a 6a
2b 3b 4b 5b 6b
7a 8a 9a 10a
7b 8b 9b 10b
Kenna Hill (*A. fasciatus*)
Kenna Creek (*A. socius*)
fas allele
soc allele
common allele
Gene:
AK
APBP
EJAC-SP
GOT
SPAG6
SPI
ACG69
Fisher's exact test: *P*<0.001
P<0.001
P<0.001
P<0.001
P<0.001
P=0.001
P=0.144