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 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.

1 **Sorted gene genealogies and species-specific nonsynonymous substitutions point to putative**

2 **postmating prezygotic isolation genes**

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

5 [1] Department of Biology, Washington University in St. Louis, St. Louis, Missouri 63130, USA

6 [2] Department of Entomology, Kansas State University, Manhattan, Kansas 66506, USA

7 [3] Department of Integrative Biology, University of Colorado Denver, Denver, Colorado 80217,

8 USA

9 [4] Office of the Executive Vice President and Provost, New Mexico State University, Las Cruces,

10 New Mexico 88003, USA

11 * Corresponding author: Suegene Noh, Department of Biology, Washington University in St.

12 Louis, One Brookings Drive, Campus Box 1137, St. Louis, Missouri 63130, USA

13 E-mail: suegene.noh@gmail.com

15 **Running title:** Postmating prezygotic isolation genes

16 **Key words:** postmating prezygotic isolation, positive selection, lineage sorting, haplotype

17 networks, ejaculate proteins

18 **Word count:** 5016

19 **Data Archival Location:** Sequences formatted as haplotypes are available from NCBI GenBank

20 PopSets 372477483 (AK), 372477513 (APBP), 372477527 (EJAC-SP), 372477535 (GOT),

21 372477555 (SPAG6), 372477561 (SPI), 372477571 (ACG69).

23   **Abstract**

24   Not all genes contribute equally to reproductive isolation. In the *Allonemobius socius* complex of

25   crickets, reproductive isolation is primarily accomplished via postmating prezygotic barriers. We

26   show that two ejaculate protein-coding genes exhibit patterns of evolution consistent with a

27   putative role as speciation genes. Both genes express male ejaculate proteins transferred to

28   females during copulation and were previously identified through comparative proteomics. We

29   found gene genealogies indicating advanced degrees of lineage sorting, and fixed

30   nonsynonymous substitutions and elevated ω values on the mutational steps separating species,

31   between both pairs of species, on the haplotype networks of these genes compared to other

32   candidate and control genes. At a contact zone between two members of the species complex,

33   these genes maintained species-specificity of alleles despite ongoing gene flow. The putative

34   speciation genes *arginine kinase* (AK) and *apolipoprotein A-1 binding protein* (APBP) are two

35   of the first examples of sperm maturation, capacitation, and motility related proteins that show

36   evidence of fixed nonsynonymous substitutions between species-specific alleles that may lead to

37   reproductive isolation. Our results show that when speciation is ongoing and insufficient time

38   has passed for nucleotide variation to accumulate, hypothesis testing based on haplotype

39   networks and gene trees are more powerful than sequence-based population genetic metrics at

40   detecting signatures of positive selection that may have led to speciation.

Introduction

44 Not all genes contribute equally to reproductive isolation during speciation. 'Speciation' (Wu,

45 2001; Wu & Ting, 2004; Nosil & Schluter, 2011), 'isolation' (Rieseberg, Church & Morjan,

46 2004), or 'barrier' (Noor & Feder, 2006) genes are expected to show very different patterns of

47 evolution compared to genes that are not directly involved in reproductive isolation when species

48 are still undergoing lineage sorting (Wu 2001). Therefore we expect to find putative speciation

49 genes among those genes that become fixed for alternative alleles within each incipient species

50 early in the process of divergence, with said alleles rarely crossing the species boundary in

51 sympatry (Ting, Tsaur & Wu, 2000; Dopman et al., 2005).

52   Rapidly evolving reproductive proteins that can affect fertilization success have an

53 important role in the evolution of postmating prezygotic reproductive isolation. Many

54 reproductive genes are known to evolve rapidly in a variety of organisms (Civetta & Singh, 1998;

55 Swanson & Vacquier, 2002; Clark, Aagaard & Swanson, 2006; Panhuis & Swanson, 2006;

56 Snook et al., 2009). In *Drosophila* where some of the most extensive work has been done, genes

57 that show male-biased expression evolve faster compared to female-biased and somatically

58 expressed genes (Zhang, Hambuch & Parsch, 2004; Zhang & Parsch, 2005; Metta et al., 2006;

59 Pröschel, Zhang & Parsch, 2006; Haerty et al., 2007), and seminal fluid proteins in particular

60 tend to show an excess of nonsynonymous substitutions (Begun et al., 2000; Swanson et al.,

61 2001; Wagstaff & Begun, 2005; Almeida & DeSalle, 2008). Similar patterns have also been

62 observed in mice and primates (Clark & Swanson, 2005; Karn et al., 2008; Ramm et al., 2008;

63 Turner, Chuong & Hoekstra, 2008; Dean et al., 2009). Using a proteomics approach on insect

64 spermatophores to isolate male reproductive protein coding-genes that can directly interact with

65 female counterparts has proved to be an efficient way of narrowing prospects in the search for

66 putative speciation genes (Andrés, Maroja & Harrison, 2008; Marshall et al., 2011).

67   The male ejaculate proteome comprises sperm-expressed proteins and seminal fluid

68 proteins. Sperm not only contribute half of the diploid genome, but are also involved in sperm-

69 egg interactions including egg activation and deliver paternal factors during fertilization (Dorus

70 et al., 2006). Seminal fluid proteins, the majority of which are produced by male accessory

71 glands, contain conserved functional classes of peptides and pro-hormones that are involved in

72 sperm binding, proteolysis, lipid metabolism, and immune function (Mueller et al., 2004;

73 Chapman & Davies, 2004; Poiani, 2006; Avila et al., 2011). Once transferred into the female

74 reproductive tract, these proteins can initiate a wide-range of physiological functions including

75 increased egg production and oviposition, decreased receptivity, decreased lifespan, and

76 increased feeding in females (reviewed in (Avila et al., 2011). The interacting female

77 counterparts to these ejaculate proteins (EPs) are not well known (Ram, Ji & Wolfner, 2005;

78 Ram & Wolfner, 2007; Snook et al., 2009) though genomic data is proving to be invaluable for

79 identifying candidates (Findlay et al., 2014). The evolution of EPs has been hypothesized to be

80 driven by one or more processes including female sperm preference, sperm competition, and

81 sexual conflict (Mueller et al., 2004; Snook et al., 2009). Here, we show through multiple lines

82 of evidence that two EP-coding genes in the *Allonemobius socius* complex of crickets show

83 patterns of molecular evolution and gene genealogies consistent with a putative role as speciation

84 genes.

85   The *A. socius* complex of ground crickets, *A. socius*, *A. fasciatus*, and *A. sp. nov.* Tex,

86 represents a powerful system to explore the hypothesized link between EP divergence and

87 reproductive isolation. Members of this complex are primarily isolated from one another by two

88   postmating, prezygotic phenotypes – conspecific sperm precedence (Gregory & Howard, 1994;

89   Howard et al., 1998a,b; Marshall, 2004) and the superior ability of conspecific males to induce

90   females to lay eggs (Gregory & Howard, 1993; Howard et al., 1998b). Two other compelling

91   features of this organismal system are species boundaries that remain intact in sympatry despite

92   some gene flow (Howard, 1986; Howard & Waring, 1991; Traylor et al., 2008) and the very

93   recent nature of divergence between these species (i.e., within the last 30,000 years; (Marshall,

94   2004, 2007). Indeed, divergence is so recent that few species-specific alleles have been identified;

95   for example, only 2 of 17 allozyme markers (Howard, 1983), 2 of 5,400 AFLP markers (Howard

96   et al., 2002), ~21 of 1,660 thorax proteins and ~33 of 922 ejaculate proteins (Marshall et al.,

97   2011) and 1 of 16 randomly chosen reproductive genes spanning >7,500 bp of coding sequence

98   (Marshall et al., *unpublished data*), yield evidence of species specificity. Taken together, the

99   above data suggest that while there is sufficient genetic divergence to produce reproductive

100   isolation and maintain species boundaries in sympatry, the vast majority of genes show no

101   evidence of divergence and thus, no lineage sorting. In all, the *A. socius* complex represents a

102   system whereby speciation is ongoing with relatively few genes contributing to the postmating,

103   prezygotic reproductive isolation between species. Therefore, if we can identify those ejaculate

104   and female reproductive tract genes that exhibit signatures of positive selection, and maintain

105   species-specificity in sympatry, we will gain insight into the genes that contribute to reproductive

106   isolation and ultimately are involved in driving speciation.

107   In this study, we expanded analyses from a previous study comparing EPs between the

108   species *A. socius* and *A. fasciatus* (Marshall et al., 2011) by including more genes and an

109   additional species, *A. sp. nov.* Tex (Traylor et al., 2008). Specifically, longer fragments of the

110   five original proteins (ACG69, AK, APBP, EJAC-SP, SPI) plus two additional EPs (GOT,

111 SPAG6) were compared for patterns of nucleotide variation, evidence of lineage-specific

112 positive selection and different degrees of lineage sorting, and species-specificity of alleles in the

113 contact zone between *A. socius* and *A. fasciatus*. These combined analyses point toward an

114 important role for some but not all examined EPs during the evolution of reproductive isolation

115 within this complex of crickets.

116

117 **Methods**

118 *Background*

119   Striped ground crickets of the *A. socius* complex inhabit moist grasslands across North

120 America and do not show significant habitat isolation (Howard 1986). The three species *A.*

121 *socius*, *A. fasciatus*, and *A. sp. nov.* Tex form two contact zones, one between *A. fasciatus* (north)

122 and *A. socius* (south) from Illinois to New Jersey (Howard & Waring, 1991), and one between *A.*

123 *sp. nov.* Tex (west) and *A. socius* (east) near the Louisiana – Texas state line (Traylor et al.,

124 2008). *A. fasciatus* and *A. socius* seem to have diverged from a common ancestor approximately

125 30,000 years ago, and *A. sp. nov.* Tex seems to have subsequently diverged from *A. socius*

126 approximately 24,000 years ago (Marshall, 2004, 2007). They have previously been shown to be

127 isolated primarily via postmating prezygotic reproductive isolation (Howard et al., 2002;

128 Marshall, 2004; Marshall & DiRienzo, 2012).

129

130 *Population and gene sampling*

131   Crickets were collected from each population and genotyped in the lab via allozymes

132 (Isocytrate dehydrogenase and Hexokinase) to determine species identity (Howard, 1983, 1986).

133 Sampling localities spanned the range of each species. *A. socius* populations were sampled near

134 Texarkana, AR (AR), Bottom, NC (Bot), Mt. Vernon, IL (IL), Pleasantville, NJ (Mi), Ruston,

135 LA (LA), Gastonia, NC (NC), and Ardmore, OK (OK). *A. fasciatus* populations were sampled

136 near Akron, OH (Akn), Frankfort, IL (FF), and New Paltz, NY (NP). *A. sp. nov.* Tex populations

137 were sampled near Terrell, TX (Tx20), Royse City, TX (Tx30), and Gainesville, TX (Tx35).

138 Contact zone populations of *A. fasciatus* and *A. socius* were sampled from two habitats at a

139 single location in Kenna, WV. *A. fasciatus* was collected from a hillside habitat, which we call

140 Kenna Hill (KH), and *A. socius* was collected along the base of hill near a creek which we call

141 Kenna Creek (KC). We did not have samples from the contact zone between *A. socius* and *A. sp.*

142 *nov.* Tex. General maintenance protocols followed Marshall et al (2009).

143  We dissected male accessory glands and testes from three individuals per allopatric

144 population and 9 individuals per contact zone population. cDNA was synthesized from each

145 tissue using RNA isolated via an Ambion RNAqueous-4PCR (#AM1914) kit and standard

146 protocols for 1st strand cDNA synthesis. General PCR and sequencing procedures followed

147 Marshall et al (2011). Standard PCR chemistry was followed with annealing temperatures

148 between 50-60 °C depending on individual primer melting temperatures (primers used are shown

149 in Supplementary Table 1). We compared nucleotide sequences of five candidate EP genes with

150 two control EP genes. Among the five candidate genes, two were chosen based on species-

151 specific proteome profiles (Marshall et al., 2011): 1) *arginine kinase* (AK), a phosphotransferase

152 enzyme expressed in the sperm that may be involved in sperm motility, capacitation or the

153 acrosome reaction (Strong & Ellington, 1993; Niksirat et al., 2015); 2) *apolipoprotein A-1*

154 *binding protein* (APBP), a phosphoprotein expressed in sperm and hypothesized to be involved

155 in sperm capacitation (Jha et al., 2008). Two were chosen based on previous sequencing data

156 showing species-specific molecular variation: 3) *ejaculate serine protease* (EJAC-SP), an

157   abundant accessory gland-expressed serine protease previously shown to be involved in the

158   induction of egg laying in successfully mated females (Marshall et al., 2009); 4) *aspartate*

159   *aminotransferase* (GOT), a pyridoxal-phospate-dependent aminotransferase expressed in the

160   testis and an allozyme historically used to diagnose species identity among *A. socius* complex

161   crickets (Howard, 1983, 1986). The last candidate gene was chosen based on a review of sperm

162   biology literature: 5) *sperm-associated antigen 6* (SPAG6), important for sperm flagellar motility

163   and the structural integrity of the central apparatus (Neilson et al., 1999; Sapiro et al., 2002).

164   The control genes had non species-specific proteome profiles (Marshall et al., 2011) and

165   were: 6) *serpine inhibitor* (SPI), a testis-expressed serine-type endopeptidase inhibitor; 7) *acg69*

166   (ACG69), a protein of unknown function expressed in the accessory glands. Sequences formatted

167   as haplotypes are available from NCBI GenBank PopSets 372477483 (AK), 372477513 (APBP),

168   372477527 (EJAC-SP), 372477535 (GOT), 372477555 (SPAG6), 372477561 (SPI), 372477571

169   (ACG69).

170

171   *Sequence evolution-based analyses*

172   Male biased genes have been shown to exhibit patterns of molecular evolution associated

173   with relaxed selective constraints or strong positive selection, such as higher rates of

174   nonsynonymous substitutions (Zhang, Hambuch & Parsch, 2004). We investigated multiple

175   metrics of molecular sequence evolution to test for evidence of selection and a departure from

176   neutral sequence evolution. We applied Tajima's D and Fu and Li's D tests to each gene to look

177   for evidence of departure from neutral allelic distributions within species (Tajima, 1989; Fu & Li,

178   1993). We compared polymorphism within species to divergence between species using HKA

179 tests (Hudson, Kreitman & Aguadé, 1987), and tested for differences in these ratios at each

180 branching node of the species tree.

181   Next, we compared polymorphism and divergence between synonymous and

182 nonsynonymous sites within each gene at each branching node of the species tree. We compared

183 ω, the rate ratios of synonymous substitutions per synonymous site κ_a (d_N) and nonsynonymous

184 substitutions per nonsynonymous site κ_s (d_S). We used McDonald-Kreitman tests to compare the

185 ratio of nonsynonymous to synonymous intraspecific polymorphisms to the ratio of

186 nonsynymous to synonymous fixed differences between species (McDonald, Kreitman & others,

187 1991). All tests were based on sequences aligned in BioEdit v.7.0.5.3 (Hall, 1999) and metrics

188 calculated using DnaSP v.5.10.01 (Librado & Rozas, 2009). For HKA tests, we used the program

189 hka provided by Jody Hey (Wang & Hey, 1996).

190

191 *Gene genealogy-based analyses*

192   Evolutionary relationships between species are tested with phylogenetic trees while

193 hypotheses of intraspecific relationships benefit from haplotype network-based approaches

194 (Posada & Crandall, 2001). Because our species are recently diverged, we used both tree-based

195 and haplotype network-based analyses to detect interesting patterns of gene evolution.

196   We used statistical parsimony haplotype networks (Templeton, Crandall & Sing, 1992) of

197 alleles from all three species to test for species-specificity of alleles. We used TCS (Clement,

198 Posada & Crandall, 2000) to generate the haplotype networks using only allopatric individuals.

199 Species-specific alleles were defined as those found only within each respective species.

200 Common or shared alleles were those observed in more than one species. Once alleles were

201 designated common or specific to a species, we turned to the *fasciatus - socius* contact zone. We

202   looked at nine individuals each of contact zone *A. fasciatus* and *A. socius* and determined what

203   types of allele these contact zone individuals possessed. As noted above, these individuals had

204   previously been designated as fully (homozygous) *A. fasciatus* or *A. socius* based on allozymes.

205   We used Fisher's exact tests with Freeman-Halton extensions for 2x3 contingency tables to

206   determine the probability of observing the distribution of *fas* vs. *soc* vs. shared alleles for each

207   gene.

208   We tested for lineage-specific positive selection on individual gene tree topologies using

209   the Genetic Algorithm (GA) Branch method (Pond & Frost, 2005) via the Datamonkey

210   webserver of the HyPhy package (Delport et al., 2010). GA Branch uses a genetic algorithm that

211   allows estimates of the nonsynonymous to synonymous substitution rate ratio ($d_N/d_S = \kappa_a/\kappa_s = \omega$)

212   to vary freely across branches within a phylogeny and compares models with different ω classes.

213   Only allopatric individuals were included in the analysis and neighbor-joining trees used by GA

214   Branch were generated natively within Datamonkey.

215   The genealogical sorting index (gsi) reflects the degree of lineage sorting of individual

216   gene genealogies that occurs during speciation, with values ranging from zero (complete

217   polyphyly) to 1 (complete monophyly) (Cummings, Neel & Shaw, 2008). We calculated gsi for

218   each gene using the online server (www.genealogicalsorting.org) with gene trees including both

219   allopatric and contact zone individuals. Sequences were phased in DnaSP prior to tree building

220   for all genes except APBP, which had no heterozygous individuals. Sorting is more difficult to

221   observe in phased data. We generated maximum likelihood gene trees with PhyML 3.0 (Guindon

222   et al., 2010) via the Mobyle server (Neron et al., 2009). We used nearest neighbor interchange

223   (NNI) tree search and HKY85 as our nucleotide substitution model. MEGA6 (Tamura et al.,

224   2013) was used to visualize these trees.

225

226 **Results**

227 *Sequence evolution-based analyses*

228   We found a general lack of both synonymous and nonsynonymous nucleotide variation

229 among all EP genes we investigated (Table 1). The Watterson estimator $\theta=4N_e\mu$ ranged from

230 0.001 to 0.011. Levels of θ in the EP candidate genes were approximately an order of magnitude

231 lower than the control genes, although this difference was not statistically significant (*fas* -

232 Mann-Whitney $U = 0$, $P = 0.051$; *soc* - Mann-Whitney $U = 4$, $P = 0.688$; Tex - Mann-Whitney U

233 $= 4.5$, $P = 0.845$). In no cases were Tajima's D or Fu and Li's D tests significantly different from

234 neutral expectations (Table 1) (all $P > 0.1$).

235   To compare polymorphism within species to divergence between species, we used a

236 standard multilocus HKA test and HKA outlier tests for each branching event. We included all

237 loci and performed 9999 rounds of coalescent simulations. The multilocus HKA test did not find

238 a significant departure from neutral expectations for the first branching event between *A.*

239 *fasciatus* and the two other species ($X^2 P = 0.916$). The outlier cell, which was *A. fasciatus* for

240 polymorphism in ACG69, was not significantly different in its pattern of polymorphism to

241 divergence ($P = 0.68$). The multilocus HKA test did find a significant departure from neutral

242 expectations for the second branching event between *A. socius* and *A. sp. nov.* Tex. ($X^2 P =$

243 0.012). However, the outlier cell, which was polymorphism at GOT in *A. sp. nov.* Tex., was not

244 significantly different in its pattern of polymorphism to divergence ($P = 0.06$).

245   We compared the rate ratios of nonsynonymous to synonymous substitutions $\omega = \kappa_a/\kappa_s$ at

246 each branching event of the species tree. When ω is larger than 1 and the nonsynonymous

247 substitution rate exceeds the synonymous substitution rate, positive or diversifying selection is

248   inferred. When ω is smaller than 1, negative or purifying selection is inferred. However, $\omega = 1$ is

249   recognized as a conservative threshold because the average κ_a is expected to be much smaller

250   than κ_s given the expectation of widespread purifying selection acting on functional genes

251   (Nielsen, 2001). Therefore the value $\omega = 0.5$ has been suggested as an alternate cutoff for the

252   detection of positive selection as subsequent analyses generally indicate that such genes are

253   indeed under positive selection (Swanson et al., 2004). In none of our genes did ω exceed 1, but

254   in the older split between *A. fasciatus* and the other two species, ω exceeded 0.5 for the genes

255   AK and APBP (Table 2).

256   We used McDonald-Kreitman tests to compare the ratio of nonsynonymous to

257   synonymous intraspecific polymorphisms (P_N/P_S) to the ratio of nonsynymous to synonymous

258   fixed differences between species (D_N/D_S). Not all genes had fixed nonsynonymous substitutions

259   between species and in these cases we were unable to apply the McDonald-Kreitman test. For

260   those genes that were testable, we did not find significant differences in D_N/D_S compared to

261   P_N/P_S at either branching event (Fisher's exact test $P = 0.07 \sim 1$) (Table 2).

262

263   *Gene genealogy-based analyses*

264   The statistical parsimony haplotype networks generated using allopatric individuals of all

265   three species showed the presence of only species-specific alleles in AK, APBP, and GOT, while

266   the other genes had alleles shared between two species each (Figure 1). Within the contact zone

267   between *A. fasciatus* and *A. socius*, AK, APBP, EJAC-SP and SPAG6 had upwards of 16

268   species-specific alleles out of 18 possible alleles (approximately 88%) (Figure 2). In comparison,

269   many GOT and SPI alleles were shared between the contact zone populations. Fisher's exact

270   tests indicated the allelic distributions were nonrandom for all genes except ACG69.

271   The GA Branch method detected elevated ω classes in all genes except EJAC-SP and

272   SPAG6 (Table 3). We mapped the substitution rate changes detected by GA Branch onto the

273   haplotype networks (Figure 1). In AK and APBP, elevated ω were detected on mutational steps

274   between both species pairs, between *A. fasciatus* and *A. socius* and between *A. socius* and *A. sp.*

275   *nov.* Tex, and were associated with one or more fixed nonsynonymous substitutions. In GOT and

276   SPI, elevated ω were detected on mutational steps between *A. fasciatus* and *A. sp. nov.* Tex, and

277   also within *A. sp. nov.* Tex. Elevated ω were detected on several branches in ACG69 (Figure 1).

278   Comparisons of genealogical sorting index values based on maximum likelihood gene

279   trees including all sampled individuals, both allopatric and contact zone, indicated that only AK

280   and APBP showed advanced lineage sorting for all three species (Table 4, Supplementary

281   Figures 1 & 2). Excluding *A. sp. nov.* Tex, which had high gsi-values overall most likely due to

282   its limited range and lack of data from its contact zone with *A. socius*, the two control genes were

283   unsorted across the species ranges. The remaining three candidate genes showed asymmetrical

284   lineage sorting.

285

286   **Discussion**

287   Many reproductive genes, and in particular those that are male biased, are known to

288   evolve rapidly, often exhibiting higher rates of nonsynonymous substitutions (Zhang, Hambuch

289   & Parsch, 2004), reduced codon usage bias (Hambuch & Parsch, 2005), and evidence suggesting

290   they are more likely to evolve by duplication (Ellegren & Parsch, 2007). However recent

291   divergence can hinder the application of many metrics of molecular evolution that rely on

292   sequence variation since not enough evolutionary time has passed to allow for differences to

293   accumulate between incipient species. Thus data from relatively recently (~30,000 years)

294   diverged species such as ours show a general lack of both synonymous and nonsynonymous

295   nucleotide variation among all investigated genes (Table 1 & 2). In addition, our estimates of

296   sequence variation were also at least an order of magnitude smaller compared to other known

297   estimates from accessory gland protein coding genes in various other species groups (Mueller et

298   al., 2005; Wagstaff & Begun, 2005; Almeida & DeSalle, 2008), including some *Gryllus* crickets

299   whose species are of roughly similar age (Andrés et al., 2006). Therefore, while relatively young

300   species offer an opportunity to observe the ongoing process of the genetics of species divergence,

301   attempting to identify putative speciation genes based on DNA sequences requires an approach

302   that takes into account gene trees and haplotype networks, along with species trees.

303   The ratio of nonsynonymous to synonymous substitution rates ω is commonly used to

304   detect signatures of selection acting upon protein coding genes (Yang & Bielawski, 2000;

305   Nielsen, 2001, 2005; Jensen, Wong & Aquadro, 2007). The original intended application of ω

306   was for the analysis of sequence evolution among distantly related species, though in practice, it

307   is not uncommonly applied to sequences between closely related populations of a species

308   (Kryazhimskiy & Plotkin, 2008). This turns out to be problematic because when sequence

309   evolution under selection was simulated over short evolutionary timescales, representative of

310   genetic variation segregating within a species, vs. long evolutionary timescales, intraspecific ω

311   behaved very differently from interspecific ω (Kryazhimskiy & Plotkin, 2008). In fact,

312   Kryazhimskiy and Plotkin show that under positive selection, $\omega = 1$ when selection was

313   moderate but showed asymptotic behavior and eventually decreased below 1 as the selection

314   coefficient increased. Over short timescales, its variance also increased as θ ($= 2N\mu$ in the paper)

315   became smaller, making it more difficult to accurately detect positive selection.

316   Another complication with using ω for short evolutionary timescales is that during initial

317   sequence divergence ω could be higher than expected because slightly deleterious

318   nonsynonymous mutations can persist in a population for generations due to genetic hitchhiking,

319   and a time lag before they are removed by purifying selection (Rocha et al., 2006). Based on

320   simulations, Rocha and colleagues showed this is why unexpectedly high ω values are frequently

321   observed among closely related (1 – 2% sequence divergence) bacteria species. As evolutionary

322   time progresses further, they show that synonymous mutations continued to accumulate and

323   exceeded the initial overrepresentation of nonsynonymous mutations. Therefore if ω is estimated

324   too soon after sequence divergence, one would expect to find high rates of false positive

325   detection.

326   Finally, even at longer evolutionary timescales the assumption that κ_s is effectively

327   neutral may need reconsideration, as a survey of 16 vertebrate genomes indicates that genes with

328   high ω are more strongly influenced by small κ_s rather than large κ_a (Wolf et al., 2009). Similar

329   patterns are observed in *Drosophila* species, where fast evolving genes show a negative

330   correlation between κ_a and synonymous site polymorphism π_s (Andolfatto, 2007; Macpherson et

331   al., 2007; Jensen & Bachtrog, 2010).Thus, the interaction between linkage and selection makes it

332   challenging to distinguish between recurrent positive selection, background selection, and Hill-

333   Robertson effects (Hill & Robertson, 1966; Charlesworth, 1994; Andolfatto, 2007; Charlesworth

334   et al., 2009). Therefore in order to detect adaptive evolution due to positive selection, applying

335   combinations of metrics including ω, Tajima's or Fu and Li's *D* and site frequency spectra, as

336   well as estimates of population size and recombination rates seems necessary (Nielsen, 2005).

337   We failed to detect positive selection based on ω, and estimates of *D* for all genes

338   compared here were not significantly different from neutral expectations (Tables 1 and 2). While

339   population bottlenecks are thus likely to have contributed to sequence variation patterns since

340   speciation in the *A. socius* complex is thought to coincide with glaciation history (Marshall, 2004,

341   2007), for our data sequence evolution-based tests are generally inconclusive as to demographic

342   reasons for why our genes might lack nucleotide variation.

343   Instead, tests based on individual gene genealogies and haplotype networks indicated AK

344   and APBP as putative speciation genes. Within the contact zone of *A. fasciatus* and *A. socius*,

345   AK, APBP, and EJAC-SP show significantly nonrandom patterns of allelic distributions and had

346   no shared alleles (Figure 2). When all allopatric and contact zone individuals were examined,

347   only the genealogies of AK and APBP indicated that these genes were more advanced in their

348   degree of lineage sorting compared to other candidate and control genes (Table 4, Supplementary

349   Figures 1 & 2). These patterns fit models of ongoing speciation in the face of gene flow, where

350   speciation genes are more likely to be fixed early on during lineage divergence (Wu, 2001).

351   Incomplete lineage sorting and introgression have been suggested to be confounding factors in

352   understanding speciation with ongoing gene flow (Machado & Hey, 2003; Broughton &

353   Harrison, 2003; Payseur, 2010). However, speciation genes are more likely to become fixed for

354   species-specific alleles early in the process of speciation and therefore are expected to be

355   relatively exempt from incomplete sorting and subject to reduced introgression. Similar patterns

356   have been observed in *Drosophila*, field crickets, and moths (Ting, Tsaur & Wu, 2000; Dopman

357   et al., 2005; Maroja, Andrés & Harrison, 2009; Andrés et al., 2013; Larson et al., 2013). It is

358   possible that these genes are not the direct targets but rather linked to targets of divergent

359   selection. Because both genes were identified through comparative proteomics (Marshall et al

360   2011) this seems relatively unlikely, but the genomic regions around these genes should be

361   investigated for evidence of selective sweeps to rule out this possibility.

362   Many studies of reproductive proteins report evidence of positive selection acting on a

363   subset of the genes examined, in both males (Begun et al., 2000; Swanson et al., 2001; Clark &

364   Swanson, 2005; Wagstaff & Begun, 2005; Andrés et al., 2006; Karn et al., 2008; Ramm et al.,

365   2008; Almeida & DeSalle, 2008; Walters & Harrison, 2010) and females (Swanson et al., 2004;

366   Panhuis & Swanson, 2006; Lawniczak & Begun, 2007; Prokupek et al., 2008; Kelleher &

367   Markow, 2009; Kelleher, Clark & Markow, 2011). However, there are few examples of adaptive

368   reproductive protein evolution leading to reproductive isolation outside of gamete recognition

369   proteins (e.g. (Geyer & Palumbi, 2003; McCartney & Lessios, 2004; Springer & Crespi, 2007).

370   Our putative speciation genes AK and APBP two of the first examples of sperm maturation and

371   capacitation related proteins that show evidence of fixed nonsynonymous substitutions between

372   species-specific alleles leading to reproductive isolation. In contrast to the other genes examined,

373   fixed nonsynonymous substitutions and elevated ω values only on the mutational steps

374   separating species on the haplotype network of APBP, and to a less exclusive extent in AK and

375   GOT, indicate that these EPs may have evolved under positive selection and contribute to the

376   reproductive isolation between these species (Table 3, Figure 1). We had previously observed

377   this pattern between *A. fasciatus* and *A. socius* for both AK and APBP (Marshall et al 2011), but

378   finding the same pattern in the mutational steps between *A. socius* and *A. sp. nov.* Tex with

379   different species-specific nonsynonymous substitutions emphasizes the potential importance of

380   these candidates.

381   Whether there are functional consequences to the species-specific nonsynonymous

382   substitutions in AK and APBP needs to be investigated further. Since both candidates were

383   identified by proteomic screens, we hypothesize that an interaction between each male EP and

384   the female reproductive tract during capacitation is responsible for the postmating prezygotic

385   isolation observed between the *A. socius* complex species. AK is a phosphagen kinase that

386   catalyzes ATP-regeneration and energy transport in invertebrates and some protozoa (Ellington,

387   2001; Noguchi, Sawada & Akazawa, 2001; Uda et al., 2006). Insects and other ecdysozoans

388   possess AK as their sole phosphagen system for cellular energy metabolism, and accordingly,

389   arginine phosphate and its phosphagen kinase AK are found primarily in muscles, but also in

390   sperm and compound eyes (Strong & Ellington, 1993; Kucharski & Maleszka, 1998; Ellington,

391   2001). The possible roles of AK as an EP can be related to sperm motility (Strong and Ellington

392   1993), capacitation, or the acrosome reaction (Niksirat et al., 2015). Two structural loops and

393   several active sites near them are the proposed interaction interface of AK with the guanidinium

394   groups of its substrates (Zhou et al., 1998; Pruett et al., 2003; Azzi et al., 2004; Clark, Davulcu &

395   Chapman, 2012). As expected for an integral enzyme, the nonsynonymous substitutions we

396   observed do not occur at these specific sites, though they may still influence its activity. APBP

397   becomes phosphorylated during murine sperm capacitation and co-localizes with cholesterol

398   during this process, but its specific function is unknown (Jha et al 2008). It does possess a

399   Rossmann-like fold, indicating an enzymatic role. The nonsynonymous substitutions we

400   observed in APBP occur in the Rossmann-like fold and are hypothesized to influence the activity

401   of its binding site (Marshall et al., 2011).

402

403   **Conclusions**

404   *A. socius* complex crickets provide an excellent opportunity to identify patterns of

405   evolution in speciation genes for two major reasons: speciation is incomplete as evidenced by

406   ongoing gene flow in the field, and isolation is through a single type of reproductive isolation

407   barrier (i.e., postmating prezygotic phenotypes). Therefore we looked for genes that contribute to

408   postmating prezygotic isolation and exhibit fixed, or nearly fixed, nonsynonymous substitutions

409   between species as putative speciation genes. We find that when speciation is ongoing, standard

410   population genetics analyses based on θ and ω values are unable to detect signatures of positive

411   selection contributing to fixed differences between species because insufficient time has passed

412   for nucleotide variation to accumulate. Instead, hypothesis testing based on haplotype networks

413   and gene trees proved to be more powerful at identifying putative postmating prezygotic

414   isolation genes with fixed nonsynonymous substitutions between both pairs of species that may

415   have led to speciation.

416

417   **Acknowledgments**

418   We thank Diana Huestis and Shanda Wheeler for their help in isolating RNA and

419   screening individuals for species status with allozymes. This is contribution no. 12-015-J from

420   the Kansas Agricultural Experiment Station. The authors declare no conflicts of interest.

422   **References**

423   Almeida FC, DeSalle R. 2008. Evidence of adaptive evolution of accessory gland proteins in

424   closely related species of the *Drosophila repleta* group. *Molecular Biology and Evolution*

425   25:2043–2053.

426   Andolfatto P. 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the

427   *Drosophila melanogaster* genome. *Genome Research* 17:1755–1762.

428   Andrés JA, Maroja LS, Bogdanowicz SM, Swanson WJ, Harrison RG. 2006. Molecular

429   evolution of seminal proteins in field crickets. *Molecular Biology and Evolution*

430   23:1574–1584.

431   Andrés JA, Larson EL, Bogdanowicz SM, Harrison RG. 2013. Patterns of transcriptome

432   divergence in the male accessory gland of two closely related species of field crickets.

433   *Genetics* 193:501–513.

434   Andrés JA, Maroja LS, Harrison RG. 2008. Searching for candidate speciation genes using a

435   proteomic approach: seminal proteins in field crickets. *Proceedings of the Royal Society*

436   *of London B: Biological Sciences* 275:1975–1983.

437   Avila FW, Sirot LK, LaFlamme BA, Rubinstein CD, Wolfner MF. 2011. Insect seminal fluid

438   proteins: identification and function. *Annual review of entomology* 56:21–40.

439   Azzi A, Clark SA, Ellington WR, Chapman MS. 2004. The role of phosphagen specificity loops

440   in arginine kinase. *Protein Science* 13:575–585.

441   Begun DJ, Whitley P, Todd BL, Waldrip-Dail HM, Clark AG. 2000. Molecular population

442   genetics of male accessory gland proteins in *Drosophila*. *Genetics* 156:1879–1888.

443   Broughton RE, Harrison RG. 2003. Nuclear gene genealogies reveal historical, demographic and

444   selective factors associated with speciation in field crickets. *Genetics* 163:1389–1401.

445   Chapman T, Davies SJ. 2004. Functions and analysis of the seminal fluid proteins of male

446   *Drosophila melanogaster* fruit flies. *Peptides* 25:1477–1490.

447   Charlesworth B. 1994. The effect of background selection against deleterious mutations on

448   weakly selected, linked variants. *Genetical Research* 63:213–227.

449   Charlesworth B, Betancourt AJ, Kaiser VB, Gordo I. 2009. Genetic recombination and

450   molecular evolution. *Cold Spring Harbor Symposium on Quantitative Biology* 74:177–

451   186.

452   Civetta A, Singh RS. 1998. Sex and speciation: genetic architecture and evolutionary potential of

453   sexual versus nonsexual traits in the sibling species of the *Drosophila melanogaster*

454   complex. *Evolution*:1080–1092.

455   Clark NL, Aagaard JE, Swanson WJ. 2006. Evolution of reproductive proteins from animals and

456   plants. *Reproduction* 131:11–22.

457   Clark SA, Davulcu O, Chapman MS. 2012. Crystal structures of arginine kinase in complex with

458   ADP, nitrate, and various phosphagen analogs. *Biochemical and Biophysical Research*

459   *Communications* 427:212–217.

460   Clark NL, Swanson WJ. 2005. Pervasive adaptive evolution in primate seminal proteins. *PLoS*

461   *Genetics* 1:e35.

462   Clement M, Posada D, Crandall KA. 2000. TCS: a computer program to estimate gene

463   genealogies. *Molecular ecology* 9:1657–1659.

464   Cummings MP, Neel MC, Shaw KL. 2008. A genealogical approach to quantifying lineage

465   divergence. *Evolution* 62:2411–2422.

466   Dean MD, Clark NL, Findlay GD, Karn RC, Yi X, Swanson WJ, MacCoss MJ, Nachman MW.

467   2009. Proteomics and comparative genomic investigations reveal heterogeneity in

468   evolutionary rate of male reproductive proteins in mice (*Mus domesticus*). *Molecular*

469   *Biology and Evolution* 26:1733–1743.

470   Delport W, Poon AF, Frost SD, Pond SLK. 2010. Datamonkey 2010: a suite of phylogenetic

471   analysis tools for evolutionary biology. *Bioinformatics* 26:2455–2457.

472   Dopman EB, Pérez L, Bogdanowicz SM, Harrison RG. 2005. Consequences of reproductive

473   barriers for genealogical discordance in the European corn borer. *Proceedings of the*

474   *National Academy of Sciences of the United States of America* 102:14706–14711.

475   Dorus S, Busby SA, Gerike U, Shabanowitz J, Hunt DF, Karr TL. 2006. Genomic and functional

476   evolution of the *Drosophila melanogaster* sperm proteome. *Nature genetics* 38:1440–

477   1445.

478   Ellegren H, Parsch J. 2007. The evolution of sex-biased genes and sex-biased gene expression.

479   *Nature Reviews Genetics* 8:689–698.

480   Ellington WR. 2001. Evolution and the physiological roles of phosphagen systems. *Annual*

481   *Review of Physiology* 63:289–325.

482   Findlay GD, Sitnik JL, Wang W, Aquadro CF, Clark NL, Wolfner MF. 2014. Evolutionary rate

483   covariation identifies new members of a protein network required for *Drosophila*

484   *melanogaster* female post-mating responses. *PLoS Genetics* 10:e1004108.

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

486   Geyer LB, Palumbi SR. 2003. Reproductive character displacement and the genetics of gamete

487   recognition in tropical sea urchins. *Evolution* 57:1049–1060.

488   Gregory PJ, Howard DJ. 1993. Laboratory hybridization studies of *Allonemobius fasciatus* and *A.*

489   *socius* (Orthoptera: Gryllidae). *Annals of the Entomological Society of America* 86:694–

490   701.

491 Gregory PJ, Howard DJ. 1994. A post-insemination barrier to fertilization isolates two closely

492   related ground crickets. *Evolution* 48:705–710.

493 Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New

494   algorithms and methods to estimate maximum-likelihood phylogenies: assessing the

495   performance of PhyML 3.0. *Systematic Biology* 59:307–321.

496 Haerty W, Jagadeeshan S, Kulathinal RJ, Wong A, Ram KR, Sirot LK, Levesque L, Artieri CG,

497   Wolfner MF, Civetta A, others. 2007. Evolution in the fast lane: rapidly evolving sex-

498   related genes in *Drosophila*. *Genetics* 177:1321–1335.

499 Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis

500   program for Windows 95/98/NT. In: *Nucleic Acids Symposium Series*. 95–98.

501 Hambuch TM, Parsch J. 2005. Patterns of synonymous codon usage in *Drosophila melanogaster*

502   genes with sex-biased expression. *Genetics* 170:1691–1700.

503 Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genetical*

504   *Research* 8:269–294.

505 Howard DJ. 1983. Electrophoretic survey of eastern North American *Allonemobius* (Orthoptera:

506   Gryllidae): evolutionary relationships and the discovery of three new species. *Annals of*

507   *the Entomological Society of America* 76:1014–1021.

508 Howard DJ. 1986. A zone of overlap and hybridization between two ground cricket species.

509   *Evolution*:34–43.

510 Howard DJ, Gregory PJ, Chu J, Cain ML. 1998a. Conspecific sperm precedence is an effective

511   barrier to hybridization between closely related species. *Evolution* 52:511–516.

512 Howard DJ, Reece PG, Gregory PJ, Chu J, Cain ML. 1998b. The evolution of barriers to

513   fertilization between closely related organisms. In: Howard DJ, Berlocher SH eds.

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

515   288.

516   Howard DJ, Marshall JL, Hampton DD, Britch SC, Draney ML, Chu J, Cantrell RG. 2002. The

517   genetics of reproductive isolation: a retrospective and prospective look with comments on

518   ground crickets. *American Naturalist* 159:S8–S21.

519   Howard DJ, Waring GL. 1991. Topographic diversity, zone width, and the strength of

520   reproductive isolation in a zone of overlap and hybridization. *Evolution*:1120–1135.

521   Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on

522   nucleotide data. *Genetics* 116:153–159.

523   Jensen JD, Bachtrog D. 2010. Characterizing recurrent positive selection at fast-evolving genes

524   in *Drosophila miranda* and *Drosophila pseudoobscura*. *Genome Biology and Evolution*

525   2:371–378.

526   Jensen JD, Wong A, Aquadro CF. 2007. Approaches for identifying targets of positive selection.

527   *Trends in Genetics* 23:568–577.

528   Jha KN, Shumilin IA, Digilio LC, Chertihin O, Zheng H, Schmitz G, Visconti PE, Flickinger CJ,

529   Minor W, Herr JC. 2008. Biochemical and structural characterization of apolipoprotein

530   AI binding protein, a novel phosphoprotein with a potential role in sperm capacitation.

531   *Endocrinology* 149:2108–2120.

532   Karn RC, Clark NL, Nguyen ED, Swanson WJ. 2008. Adaptive evolution in rodent seminal

533   vesicle secretion proteins. *Molecular Biology and Evolution* 25:2301–2310.

534   Kelleher ES, Clark NL, Markow TA. 2011. Diversity-enhancing selection acts on a female

535   reproductive protease family in four subspecies of *Drosophila mojavensis*. *Genetics*

536   187:865–876.

537   Kelleher ES, Markow TA. 2009. Duplication, selection and gene conversion in a *Drosophila*

538   *mojavensis* female reproductive protein family. *Genetics* 181:1451–1465.

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

540   4:e1000304.

541   Kucharski R, Maleszka R. 1998. Arginine kinase is highly expressed in the compound eye of the

542   honey-bee, *Apis mellifera*. *Gene* 211:343–349.

543   Larson EL, Andrés JA, Bogdanowicz SM, Harrison RG. 2013. Differential introgression in a

544   mosaic hybrid zone reveals candidate barrier genes. *Evolution; International Journal of*

545   *Organic Evolution* 67:3653–3661.

546   Lawniczak MK, Begun DJ. 2007. Molecular population genetics of female-expressed mating-

547   induced serine proteases in *Drosophila melanogaster*. *Molecular Biology and Evolution*

548   24:1944–1951.

549   Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA

550   polymorphism data. *Bioinformatics* 25:1451–1452.

551   Machado CA, Hey J. 2003. The causes of phylogenetic conflict in a classic *Drosophila* species

552   group. *Proceedings of the Royal Society of London B: Biological Sciences* 270:1193–

553   1202.

554   Macpherson JM, Sella G, Davis JC, Petrov DA. 2007. Genomewide spatial correspondence

555   between nonsynonymous divergence and neutral polymorphism reveals extensive

556   adaptation in *Drosophila*. *Genetics* 177:2083–2099.

557   Maroja LS, Andrés JA, Harrison RG. 2009. Genealogical discordance and patterns of

558   introgression and selection across a cricket hybrid zone. *Evolution* 63:2999–3015.

559   Marshall JL. 2004. The *Allonemobius-Wolbachia* host-endosymbiont system: evidence for rapid

560   speciation and against reproductive isolation driven by cytoplasmic incompatibility.

561   *Evolution* 58:2409–2425.

562   Marshall JL. 2007. Rapid evolution of spermathecal duct length in the *Allonemobius socius*

563   complex of crickets: species, population and *Wolbachia* effects. *PLoS One* 2:e720.

564   Marshall JL, Huestis DL, Hiromasa Y, Wheeler S, Oppert C, Marshall SA, Tomich JM, Oppert

565   B, others. 2009. Identification, RNAi knockdown, and functional analysis of an ejaculate

566   protein that mediates a postmating, prezygotic phenotype in a cricket. *PloS one* 4:e7537–

567   e7546.

568   Marshall JL, Huestis DL, Garcia C, Hiromasa Y, Wheeler S, Noh S, Tomich JM, Howard DJ.

569   2011. Comparative proteomics uncovers the signature of natural selection acting on the

570   ejaculate proteomes of two cricket species isolated by postmating, prezygotic phenotypes.

571   *Molecular biology and evolution* 28:423–435.

572   Marshall JL, DiRienzo N. 2012. Noncompetitive gametic isolation between sibling species of a

573   cricket: a hypothesized link between within-population incompatibility and reproductive

574   isolation between species. *International Journal of Evolutionary Biology* 2012:593438.

575   McCartney MA, Lessios HA. 2004. Adaptive evolution of sperm bindin tracks egg

576   incompatibility in neotropical sea urchins of the genus *Echinometra*. *Molecular Biology*

577   *and Evolution* 21:732–745.

578   McDonald JH, Kreitman M, others. 1991. Adaptive protein evolution at the *Adh* locus in

579   *Drosophila*. *Nature* 351:652–654.

580   Metta M, Gudavalli R, Gibert J-M, Schlötterer C. 2006. No accelerated rate of protein evolution

581   in male-biased *Drosophila pseudoobscura* genes. *Genetics* 174:411–420.

582 Mueller JL, Ripoll DR, Aquadro CF, Wolfner MF. 2004. Comparative structural modeling and

583   inference of conserved protein classes in *Drosophila* seminal fluid. *Proceedings of the*

584   *National Academy of Sciences* 101:13542–13547.

585 Mueller JL, Ram KR, McGraw LA, Qazi MB, Siggia ED, Clark AG, Aquadro CF, Wolfner MF.

586   2005. Cross-species comparison of *Drosophila* male accessory gland protein genes.

587   *Genetics* 171:131–143.

588 Neilson LI, Schneider PA, Van Deerlin PG, Kiriakidou M, Driscoll DA, Pellegrini MC,

589   Millinder S, Yamamoto KK, French CK, Strauss JF. 1999. cDNA cloning and

590   characterization of a human sperm antigen (SPAG6) with homology to the product of the

591   *Chlamydomonas PF16* locus. *Genomics* 60:272–280.

592 Neron B, Menager H, Maufrais C, Joly N, Maupetit J, Letort S, Carrere S, Tuffery P, Letondal C.

593   2009. Mobyle: a new full web bioinformatics framework. *Bioinformatics* 25:3005–3011.

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

595   647.

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

597   218.

598 Niksirat H, James P, Andersson L, Kouba A, Kozák P. 2015. Label-free protein quantification in

599   freshly ejaculated versus post-mating spermatophores of the noble crayfish *Astacus*

600   *astacus*. *Journal of Proteomics* 123:70–77.

601 Noguchi M, Sawada T, Akazawa T. 2001. ATP-regenerating system in the cilia of *Paramecium*

602   *caudatum*. *Journal of Experimental Biology* 204:1063–1071.

603 Noor MAF, Feder JL. 2006. Speciation genetics: evolving approaches. *Nature Reviews Genetics*

604   7:851–861.

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

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

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

611 Poiani A. 2006. Complexity of seminal fluid: a review. *Behavioral Ecology and Sociobiology*
612   60:289–310.

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

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

617 Prokupek A, Hoffmann F, Eyun S, Moriyama E, Zhou M, Harshman L. 2008. An evolutionary
618   expressed sequence tag analysis of *Drosophila* spermatheca genes. *Evolution* 62:2936–
619   2947.

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

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

625 Ram KR, Ji S, Wolfner MF. 2005. Fates and targets of male accessory gland proteins in mated
626   female *Drosophila melanogaster*. *Insect Biochemistry and Molecular Biology* 35:1059–
627   1071.

628   Ramm SA, Oliver PL, Ponting CP, Stockley P, Emes RD. 2008. Sexual selection and the

629   adaptive evolution of mammalian ejaculate proteins. *Molecular Biology and Evolution*

630   25:207–219.

631   Ram KR, Wolfner MF. 2007. Seminal influences: *Drosophila* Acps and the molecular interplay

632   between males and females during reproduction. *Integrative and Comparative Biology*

633   47:427–445.

634   Rieseberg LH, Church SA, Morjan CL. 2004. Integration of populations and differentiation of

635   species. *New Phytologist* 161:59–69.

636   Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH, Feil EJ. 2006. Comparisons

637   of dN/dS are time dependent for closely related bacterial genomes. *Journal of Theoretical*

638   *Biology* 239:226–235.

639   Sapiro R, Kostetskii I, Olds-Clarke P, Gerton GL, Radice GL, III JFS. 2002. Male infertility,

640   impaired sperm motility, and hydrocephalus in mice deficient in Sperm-Associated

641   Antigen 6. *Molecular and Cellular Biology* 22:6298–6305.

642   Snook RR, Chapman T, Moore PJ, Wedell N, Crudgington HS. 2009. Interactions between the

643   sexes: new perspectives on sexual selection and reproductive isolation. *Evolutionary*

644   *Ecology* 23:71–91.

645   Springer SA, Crespi BJ. 2007. Adaptive gamete-recognition divergence in a hybridizing *Mytilus*

646   population. *Evolution* 61:772–783.

647   Strong SJ, Ellington WR. 1993. Horseshoe crab sperm contain a unique isoform of arginine

648   kinase that is present in the midpiece and flagellum. *Journal of Experimental Zoology*

649   267:563–571.

650 Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF. 2001. Evolutionary EST

651   analysis identifies rapidly evolving male reproductive proteins in *Drosophila*.

652   *Proceedings of the National Academy of Sciences* 98:7375–7379.

653 Swanson WJ, Wong A, Wolfner MF, Aquadro CF. 2004. Evolutionary expressed sequence tag

654   analysis of *Drosophila* female reproductive tracts identifies genes subjected to positive

655   selection. *Genetics* 168:1457–1465.

656 Swanson WJ, Vacquier VD. 2002. The rapid evolution of reproductive proteins. *Nature Reviews*

657   *Genetics* 3:137–144.

658 Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA

659   polymorphism. *Genetics* 123:585–595.

660 Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary

661   Genetics Analysis version 6.0. *Molecular Biology and Evolution* 30:2725–2729.

662 Templeton AR, Crandall KA, Sing CF. 1992. A cladistic analysis of phenotypic associations

663   with haplotypes inferred from restriction endonuclease mapping and DNA sequence data.

664   III. Cladogram estimation. *Genetics* 132:619–633.

665 Ting C-T, Tsaur S-C, Wu C-I. 2000. The phylogeny of closely related species as revealed by the

666   genealogy of a speciation gene, *Odysseus*. *Proceedings of the National Academy of*

667   *Sciences* 97:5313–5316.

668 Traylor T, Birand AC, Marshall JL, Howard DJ. 2008. A zone of overlap and hybridization

669   between *Allonemobius socius* and a new *Allonemobius* sp. *Annals of the Entomological*

670   *Society of America* 101:30–39.

671 Turner LM, Chuong EB, Hoekstra HE. 2008. Comparative analysis of testis protein evolution in

672   rodents. *Genetics* 179:2075–2089.

673   Uda K, Fujimoto N, Akiyama Y, Mizuta K, Tanaka K, Ellington WR, Suzuki T. 2006. Evolution

674   of the arginine kinase gene family. *Comparative Biochemistry and Physiology Part D:*

675   *Genomics and Proteomics* 1:209–218.

676   Wagstaff BJ, Begun DJ. 2005. Molecular population genetics of accessory gland protein genes

677   and testis-expressed genes in *Drosophila mojavensis* and *D. arizonae*. *Genetics*

678   171:1083–1101.

679   Walters JR, Harrison RG. 2010. Combined EST and proteomic analysis identifies rapidly

680   evolving seminal fluid proteins in *Heliconius* butterflies. *Molecular biology and*

681   *evolution* 27:2000–2013.

682   Wang R-L, Hey J. 1996. The speciation history of *Drosophila pseudoobscura* and close relatives:

683   inferences from DNA sequence variation at the *period* locus. *Genetics* 144:1113–1126.

684   Wolf JB, Künstner A, Nam K, Jakobsson M, Ellegren H. 2009. Nonlinear dynamics of

685   nonsynonymous (dN) and synonymous (dS) substitution rates affects inference of

686   selection. *Genome Biology and Evolution* 1:308–319.

687   Wu C-I. 2001. The genic view of the process of speciation. *Journal of Evolutionary Biology*

688   14:851–865.

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

690   Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends in*

691   *Ecology & Evolution* 15:496–503.

692   Zhang Z, Hambuch TM, Parsch J. 2004. Molecular evolution of sex-biased genes in *Drosophila*.

693   *Molecular Biology and Evolution* 21:2130–2139.

694   Zhang Z, Parsch J. 2005. Positive correlation between evolutionary rate and recombination rate

695   in *Drosophila* genes with male-biased expression. *Molecular Biology and Evolution*

696   22:1945–1947.

697   Zhou G, Somasundaram T, Blanc E, Parthasarathy G, Ellington WR, Chapman MS. 1998.

698   Transition state structure of arginine kinase: implications for catalysis of bimolecular

699   reactions. *Proceedings of the National Academy of Sciences* 95:8449–8454.

700

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)

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

2 shown)

Gene	Length	within *A. fasciatus*					within *A. socius*					within *A. sp. nov.* Tex				
		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

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; $\omega = \kappa_a / \kappa_s$)

1 Table 2. Nucelotide variation at each branching node of the *A. socius* complex species tree. (P_N: nonsynonymous polymorphisms; P_S:

2 synonymous polymorphisms; D_N: nonsynonymous fixations; D_S: synonymous fixations; κ_s: rate of nonsynonymous substitutions per

3 nonsynonymous site; κ_a: rate of synonymous substitutions per synonymous site; $\omega = \kappa_a / \kappa_s$)

Gene	Length	between fas & (soc+Tex)							between soc & Tex						
		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

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.

1   Table 3. Tests of lineage-specific positive selection that detects different ω classes along

2   branches of a gene tree. The model with best c-AIC score is shown.

Gene	Best model found by GA Branch		
	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

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).

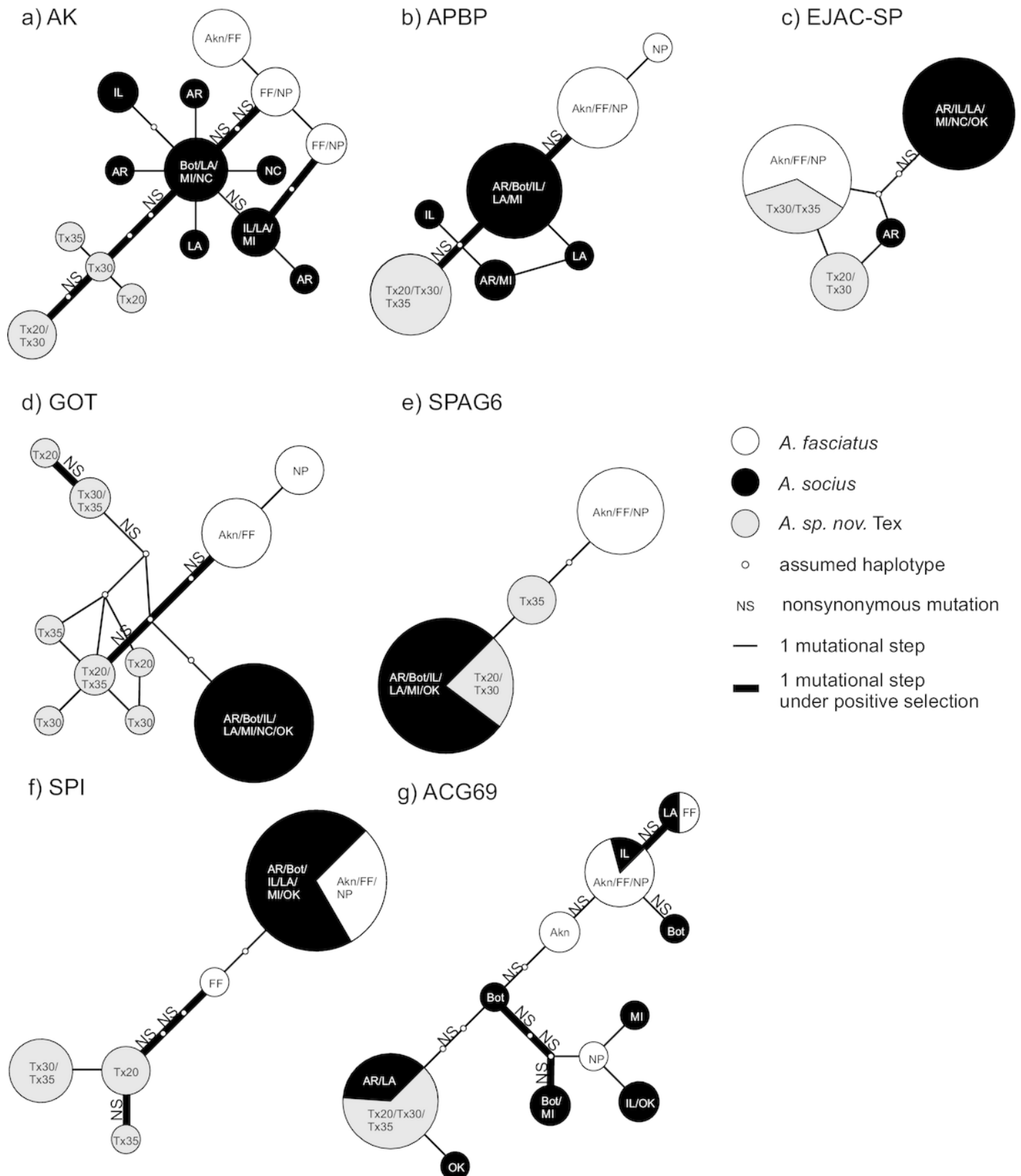
1   Table 4. Genealogical sorting index values based on individual gene trees (see Supplementary

2   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.

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.