PROCEEDINGS
OF
THE ROYAL
SOCIETY B

Proc. R. Soc. B (2008) **275**, 1975–1983
doi:10.1098/rspb.2008.0423
Published online 21 May 2008

Searching for candidate speciation genes using a proteomic approach: seminal proteins in field crickets

Jose A. Andrés*, **Luana S. Maroja and Richard G. Harrison**

Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY 14853, USA

In many animals, male seminal proteins influence gamete interactions and fertilization ability and are probably involved in barriers to gene flow between diverging lineages. Here we use a proteomic approach to identify seminal proteins that are transferred to females during copulation and that may be involved in fertilization barriers between two hybridizing field crickets (*Gryllus firmus* and *Gryllus pennsylvanicus*). Analyses of patterns of divergence suggest that much of the field cricket genome has remained undifferentiated following the evolution of reproductive isolation. By contrast, seminal protein genes are highly differentiated. Tests of selection reveal that positive selection is likely to be responsible for patterns of differentiation. Together, our observations suggest that some of the loci encoding seminal proteins may indeed play a role in fertilization barriers in field crickets.

Keywords: reproductive isolation; fertilization barriers; gene genealogies; selection; *Gryllus*

1. INTRODUCTION

Identifying genetic changes that contribute to reproductive isolation is one of the central problems in evolutionary biology. Although almost all of the genes of an organism potentially can contribute to intraspecific variation and population differentiation, only relatively few genes, so-called 'speciation' or 'barrier' genes (Wu 2001; Noor & Feder 2006), are presumably involved in barriers to gene flow between diverging lineages. Historically, studies of speciation have focused primarily on the reduced fitness of hybrids (postzygotic isolation) and on premating barriers to gene exchange, especially behavioural isolation (Coyne & Orr 2004). More recently, attention has been drawn to fertilization barriers as components of reproductive isolation and speciation. For example, in several groups of marine external fertilizers (sea urchins, abalones, turban snails and mussels) gamete recognition proteins have been identified and shown to be evolving rapidly under directional selection (Metz & Palumbi 1996; Hellberg & Vacquier 1999; Yang *et al.* 2000; Swanson & Vacquier 2002; Riginos *et al.* 2006).

In many animals fertilization is internal, and zygote formation is mediated/facilitated not only by sperm and egg proteins but also by proteins present in the seminal fluid (seminal proteins). In insects and primates, a subset of seminal proteins is among the most rapidly evolving proteins and, like gamete recognition proteins in marine invertebrates, exhibit a clear signature of positive selection (Clark *et al.* 2006). The rapid adaptive evolution of these proteins (Clark & Swanson 2005; Mueller *et al.* 2005; Andrés *et al.* 2006) may be an important component of reproductive isolation during the early stages of the speciation process (Andrés & Arnqvist 2001; Coyne & Orr 2004).

Loci encoding seminal proteins are likely to be involved in the barriers that generate new species because these genes determine traits that influence sperm activation, gamete interactions and ovulation (Peitz 1988; Tram & Wolfner 1999; Viscuso *et al.* 2001; Fry & Wilkinson 2004). In polyandrous species, seminal proteins play a key role in postmating sexual selection (Fiumera *et al.* 2005). Postcopulatory competition for the fertilization of female eggs (sperm competition and selective fertilization) may lead to rapid coevolution between seminal proteins (those transferred from males to females during copulation) and proteins of the female reproductive tract. Independent episodes of rapid coevolution (e.g. in allopatric populations) could lead to reproductive divergence and ultimately to speciation (Howard 1999; Andrés & Arnqvist 2001; Coyne & Orr 2004; Fricke *et al.* 2006).

Genealogical analyses and tests of selection can be used to assess whether genes encoding seminal proteins are potentially involved in reproductive isolation. Because genomes are genealogical or historical mosaics (Wu 2001), gene genealogies for closely related species vary with genomic region (Ting *et al.* 2000; Machado & Hey 2003; Dopman *et al.* 2005). Regions of the genome contributing to adaptive divergence and reproductive isolation will diverge more rapidly owing to the selective sorting of ancestral polymorphisms and/or the selective disadvantage of introgressed alleles (Putman *et al.* 2007). Hence, gene genealogies for loci that have been subject to recent selective sweeps, including those that contribute to reproductive isolation, are more likely to reveal closely related species as differentiated or exclusive (i.e. monophyletic) groups; at most other loci, shared ancestral polymorphism will persist and/or introgression will erase differences that are accumulated in allopatry (Ting *et al.* 2000).

The closely related field crickets *Gryllus firmus* and *Gryllus pennsylvanicus*, which come into contact in a well-characterized hybrid zone in eastern North America

* Author for correspondence (jaa53@cornell.edu).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2008.0423 or via http://journals.royalsociety.org.

Received 27 March 2008
Accepted 24 April 2008

(Harrison & Bogdanowicz 1997; Ross & Harrison 2002), provide an important model system for investigating the origin of barriers to gene exchange. The net mtDNA sequence divergence between this species pair is less than 0.5% (Willett *et al.* 1997), and surveys of variation in allozymes and nuclear gene introns have failed to reveal genome regions for which the two species exhibit fixed differences or are exclusive groups (Harrison 1979; Broughton & Harrison 2003). The estimated time since divergence for *G. firmus* and *G. pennsylvanicus* is only $0.1N_e$; therefore, we expect ancestral polymorphisms to segregate in most regions of the genome (Broughton & Harrison 2003). Loci for which the two species are highly differentiated or exclusive groups probably mark genome regions that have experienced recent selective sweeps and/or harbour barrier genes.

Field and laboratory studies of the two cricket species (Harrison 1983, 1985, 1986; Harrison & Rand 1989) have identified both pre- and postmating barriers to gene exchange. Premating barriers include ecological (habitat) isolation, temporal isolation and mate choice. However, premating barriers are not complete, and adults of the two species, together with individuals of mixed ancestry, are found together at hybrid zone localities. The only known postmating barrier is a one-way incompatibility between *G. firmus* females and *G. pennsylvanicus* males. That is, heterospecific males do not trigger normal oviposition behaviour in *G. firmus* females, and the eggs produced, which do not develop, are indistinguishable from unfertilized eggs (Harrison 1983; L. S. Maroja, J. A. Andres & R. G. Harrison 2008, unpublished data). Furthermore, an evolutionarily expressed sequence tag screen of the male cricket accessory gland has identified a set of genes encoding secreted proteins that show male-biased gene expression (Andrés *et al.* 2006; Braswell *et al.* 2006). These secreted proteins, some of which exhibit rapid evolution and evidence of positive selection, are putative seminal proteins.

Here we use a proteomic approach to provide unambiguous identification of genes encoding proteins that are transferred from male to female crickets during copulation. The approach relies on comparison of peptide sequences from tryptic digests of spermatophore homogenates with peptide sequences generated *in silico* from a translation of cricket accessory gland expressed sequence tag (EST) database. Matching of peptides within the spermatophore with those generated *in silico* allows definitive identification of genes that encode seminal proteins. Direct identification of seminal proteins is more reliable than a purely bioinformatics approach and avoids the possibility of missing rapidly evolving genes. The goal is to target a particular class of genes that may well be subject to positive selection and also may play a role in reproductive isolation. We then compare genealogies for genes encoding seminal proteins with genealogies for genes that are not related to reproduction. Our observations suggest that loci encoding seminal proteins may indeed play a role in barriers to gene exchange in field crickets.

2. MATERIAL AND METHODS

(a) *Protein sample preparation*

Two independent samples of male spermatophores ($n=6$ males per sample), the proteinaceous capsule containing the ejaculate, were harvested directly from *G. pennsylvanicus* males and homogenized in 100 µl of ice-cold phosphate buffer. The two independent samples were centrifuged (14 000*g* at 4°C) to separate seminal fluid from most of the sperm and spermatophore debris. Samples were stored at −80°C, and aliquots of each were sent to the Genome BC Proteomics Centre (University of Victoria, Canada), where proteomic analyses were carried out as described below.

Seminal fluid proteins were solubilized by adding 9.5 M urea, 50 mM NH_4HCO_3 and 0.2% SDS to the sample prior to sonication. Solubilized proteins were subject to disulphide reduction and sulphydryl alkylation (200 mM DTT and 200 mM iodoacetamide). Digestion was carried out overnight at 37°C using 20 µg of trypsin (Promega). The digested samples were cleaned using a cation exchange Cartridge Kit for cICAT (Applied Biosystems).

(b) *Strong cation exchange chromatography*

Samples were brought up to 2 ml with 10 mM KPO_4 (pH = 2.7), 25% ACN buffer and injected onto a Polysulphoethyl A strong cation exchange chromatography (SCX) column (Poly LC, Columbia, MD). The flow rate was set to 0.5 ml min^{-1}. After equilibration, a 0–35% gradient of 10 mM KH_2PO_4, 25% ACN, 0.5 M KCl was applied for 30 min. Each collected SCX fraction was reduced and transferred to autosampler vials (Dionex/LC Packings, Amsterdam).

(c) *One-dimensional reversed-phase chromatography with online mass spectrometry*

Liquid chromatography-mass spectrometry/mass spectrometry (LC-MS/MS) analyses of the SCX seminal fluid protein fractions were performed in a Hybrid Quadruple-TOF LC–MS/MS mass spectrometer (QStar Pulsar I, MDS Sciex), equipped with a nanoelectrospray ionization source (Proxeon, Odense, Denmark) fitted with a 10 µm fused-silica emitter tip (New Objective, Woburn, MA). Chromatographic separation was achieved on C18AQ Nano LC and a Zorbax C18 guard column (Agilent Technologies). The mobile phase consisted of 98 : 2 (v/v) water/acetonitrile with 0.1% formic acid. Data were acquired automatically using Analyst QS v. 1.1 software service pack (ABI MDS SCIEX, Concord, Canada). Curtain gas was set at 23, nitrogen was used as the collision gas, and the ionization tip voltage was 2700 V.

(d) *Mass spectrometry data analyses*

Analyst v. 1.1 software was used to view the information dependent acquisition file, and a built-in Mascot script (1.6b16 ABI—Matrix Science Limited) was used to create the peak lists. Spectra with less than 10 peaks were discarded. MS/MS data were centroided but not de-isotoped. Data were analysed using Mascot v. 2.0 (Matrix Science Limited). Carbamidomethyl cysteine was used as a fixed modification, oxidation of methionine was selected as a variable modification, and up to one missed cleavage was allowed. Spectrometry data were searched against deduced partial proteomes of *G. firmus* and *G. pennsylvanicus* accessory glands. The combined database contained 1435 putative protein entries with 32 048 residues. Details of the development of the EST database supporting partial proteomes have been described elsewhere (Andrés *et al.* 2006).

(e) *Genetic analyses: sampling*

(i) *Populations*

We sampled two allopatric populations of each of the cricket species (*G. firmus*: Guilford, CT and Essex, MD;

G. pennsylvanicus: Ithaca, NY and Scranton, PA). The Guilford and Ithaca populations have previously been used to characterize barriers to gene exchange between these species (Harrison 1983, 1986). Less is known about the Essex and Scranton populations, but based on geography and cricket phenotype, they appear to represent populations that are predominantly *G. firmus* and *G. pennsylvanicus*, respectively. Four randomly selected individuals per population ($n=32$ alleles) were used for genetic analyses. Additional individuals were analysed for *AG-0005F.*

(ii) *Genes*

In total, we analysed polymorphism and divergence data for the entire coding regions of six genes, all of which are expressed in cricket accessory glands: *peptidyl-prolyl isomerase-1* (*PPase*), *guanlyate kinase-1* (*GuKc*), *ubiquinone biosynthesis protein COQ7-1* (*COQ7*) and three seminal proteins (*AG-0005F*, *AG-0308F* and *AG-0334P*). *PPase* is an enzyme that accelerates protein folding by catalysing the *cis–trans* isomerization of proline peptide bonds; *GuKc* catalyses the ATP-dependent phosphorylation of GMP into GDP; and *COQ7* is a central metabolic regulatory protein. These proteins, which do not show male-biased expression, have not been identified in the spermatophore and probably do not have a function in cricket reproduction. The genes that encode these proteins are therefore considered as 'control genes'. Of the three seminal proteins *AG-0308F* is a serine protease, whereas the other two seminal proteins are biochemically uncharacterized. Total RNA from accessory glands of individual crickets was extracted using Trizol Reagent (Invitrogen), reverse transcribed and PCR amplified using gene-specific primers (Andrés *et al.* 2006; see the electronic supplementary material S1). The PCR products were sequenced using an ABI 3100 automatic sequencer (Applied Biosystems, Foster City, CA). Initial visual identification of single nucleotide polymorphisms were confirmed using Tracediff and Hetscan of the Staden package (Staden 1996). Individual alleles were reconstructed using the Phase algorithm (Stephens *et al.* 2001). Sequences have been deposited in GenBank (EU669672–EU669818).

(f) *Genealogical analyses*

For each dataset, an optimal substitution model was determined by hierarchical likelihood ratio tests using Modeltest v. 3.7 (Posada & Crandall 1998; see the electronic supplementary material S2), and the evolutionary relationship among alleles was estimated by maximum-likelihood (ML) methods using PAUP v. 4.0b10 (Swofford 2003). Nodal support was based on 1000 heuristic non-parametric bootstrap replicates. To assess the agreement among loci, we used a partition of variance test with 1000 heuristic replicates. To test if *G. firmus* and *G. pennsylvanicus* constitute exclusive (monophyletic) groups, we performed exclusivity tests by comparing the increment in the posterior log-likelihood ($\Delta\ln L$) distribution between unconstrained gene genealogies and genealogies constrained to represent the two species as monophyletic groups (*Gf* and *Gp*) at each of the sampled loci. The posterior distribution of $\Delta\ln L$ was obtained by calculating the difference of 10 000 paired ($\ln L_{\text{unconstrained}}-\ln L_{\text{constrained}}$) samples. If the 95% credible interval is entirely positive (does not include 0), we have evidence to reject genealogical exclusivity. Because there is evidence of intragenic recombination for some loci, exclusivity tests were performed using the longest non-recombining region of each locus. Non-recombining regions were inferred using the algorithms and methods implemented in the IMgc software package (Woerner *et al.* 2007). Log-likelihood distributions were estimated using BEAST v.1.3 (Drummond & Rambaut 2003*a*) with the appropriate substitution models and default prior parameters for all scale operators. Convergence and effective sample sizes were monitored in Tracer v. 1.3 (Drummond & Rambaut 2003*b*). Analyses were run for 10 million generations, with parameters sampled for every 1000 generations.

For each locus at which alleles code for different protein products, we also obtained ML genealogies using the JTT, PAM and PMB amino acid substitution models as implemented in the Phylip v. 3.6 PROML module (Felsentein 2005), with heterogeneity of evolutionary rates among sites estimated using Tree-Puzzle v. 5.2 (Schmidt *et al.* 2002). Nodal support for protein genealogies was based on 100 non-parametric bootstrap (BOOTSEQ) replicates.

Genealogical patterns were also analysed using neighbour-joining trees based on ML distances and statistical parsimony networks as implemented in Tcs v. 1.21 (Clement *et al.* 2000). The statistical parsimony approach allows reticulations to reflect uncertainty generated by recombination and homoplasy. For each dataset, the algorithm estimates (with 95% statistical confidence) the maximum number of differences among haplotypes.

(g) *Test of selection*

The relative rate of fixation of non-synonymous (d_N) and synonymous (d_S) substitutions provides an estimate of the selection pressures acting on a given protein. For any set of amino acid residues, when $d_N/d_S=\omega=1$, a neutral model of evolution cannot be rejected, whereas $\omega<1$ indicates purifying selection and $\omega>1$ indicates positive selection.

Although the selection parameter ω is commonly calculated using phylogenetic likelihood methods (Goldman & Yang 1994), these methods are unreliable in the presence of intragenic recombination because this process leads to not one, but multiple evolutionary trees along the gene sequence (Anisimova *et al.* 2003; Wilson & McVean 2006). In this paper we used the method recently developed by Wilson & McVean (2006) to calculate ω in the presence of recombination. This method relaxes the assumption of a single common history for all codons, and performs Bayesian inferences of ω using a population genetics approximation to the coalescent with recombination (Hudson 1983; Li & Stephens 2003). One disadvantage of this method is that it does not provide estimates of d_N and d_S.

Using OmegaMap (Wilson & McVean 2006) we estimated the selection parameter (ω), recombination rate (ρ), transition–transversion ratio (κ) and the rate of synonymous transversion μ for each gene. We used improper inverse prior distributions for all parameters with means $\omega=1$, $\rho=0.07$, $\kappa=3.6$ and $\mu=0.3$. Both ω and ρ were modelled as constants (i.e. all sites are assumed to share common values). The frequency of codons was assumed to be equal, and the number of alignment orderings was set to 10. The number of iterations of the Markov chain Monte Carlo (MCMC) algorithm was 250 000 with a burn-in of 25 000 and a thinning of 1000. For each gene, two independent convergent runs were merged to provide the posterior distributions of the estimated parameters. The effective sample size for the estimated parameters was always more than 100,

Table 1. Field cricket seminal proteins identified by shotgun proteomics (two-dimensional LC/LC–MS/MS). (Gene expression pattern is as described in Andrés *et al.* (2006; M, male-biased expression; M/F, expressed in both sexes). All proteins have an associated probability of being correctly identified above 99.9% (based on Mascot MudPIT scores). NP is the number of digested peptides that significantly matched each protein. SC represents the percentage of protein sequence identified in the sample. Only peptide hits of eight amino acids or longer and showing a 99% probability of being correct were accepted for protein identification. In the case of single-peptide hits, the probability threshold was increased to 99.9%.)

locus	expression	functional homology	NP	SC (%)
AG-0315F	M	unknown	37	25
AG-0312F	M	unknown	14	16
AG-0076F	M	unknown	22	15
AG-0241F	M	unknown	6	22
AG-0313F	M	unknown	35	20
AG-0001F	M	unknown	8	16
AG-0020F	M	unknown	11	22
AG-0334P	M	unknown	3	6
AG-308F	M	serine protease	4	7
AG-508F	M	serine protease	2	7
AG-159F	M	serine protease	1	4
AG-0197P	M	unknown	1	4
AG-0157F	M	nucleotidase	1	11
AG-0085F	M/F	unknown	2	3
AG-0005F	M	unknown	2	4
AG-0099F	—	unknown	2	6
AG-0115F	—	unknown	1	4
AG-0161F	M	unknown	1	11
AG-0144F	—	unknown	1	16
AG-0055F	M	unknown	1	7
AG-0090F	M/F	unknown	3	13
AG-0042F	M	unknown	9	13

suggesting that the MCMC chains were run long enough to obtain accurate estimates of the parameters.

Statistical differences in the selection parameter (ω) among loci were inferred using Bayesian pairwise comparisons. If the two given loci A and B are under the same evolutionary constraints, the difference between the posterior distribution of ω estimated from each locus should not be distinguishable from zero. Thus, the 95% credible interval of the distribution of $\omega_A - \omega_B$ should include zero. Otherwise, these two ω values are significantly different.

3. RESULTS

(a) *Proteomic analyses*

We combined data from two accessory gland EST libraries (Andrés *et al.* 2006) with shotgun (LC/LC–MS/MS) peptide sequencing of proteins extracted from field cricket spermatophores (table 1). This dataset represents a pool of two mass spectrometry analytical runs on two independently extracted protein samples ($n=6$ males/sample). Comparing the spectrometry data of both runs against our *Gryllus* accessory gland proteome database ($n=1435$ putative proteins), we identified 30 genes that are expressed in the male accessory gland. Of these genes, 22 appear to encode proteins that are components of the spermatophore coat or the seminal fluid (table 1); the other eight are probably associated with residual sperm in our sample (see the electronic supplementary material S3). Most of the identified genes are expressed exclusively or more strongly in males and encode proteins of unknown function, i.e. proteins that lack any functional homology or structural similarity to biochemically characterized proteins (table 1). Among these 'unknown' proteins are six (encoded by *AG-0001F*, *AG-0076F*, *AG-0312F*, *AG-0313F*, *AG-0315F* and *AG-0317F*) that have a highly repetitive primary structure of small amino acid residues (Ala, Leu) and a predicted helical structure. These are probably proteins of the spermatophore coat (Andrés *et al.* 2006). Three loci (*AG-0159F*, *AG-0308F* and *AG-0508F*) encode proteins with predicted serine protease activity; proteases are known to be an important component of the seminal fluid in other insects (e.g. Mueller *et al.* 2005). One locus codes for a presumed nucleotidase and the remaining 12 genes encode seminal fluid proteins for which structures/functions have not been identified. Moreover, by expanding our search to include general peptide databases (MSDB), we have also identified additional 'housekeeping proteins' (see the electronic supplementary material S3). Many of these proteins (e.g. β-actin, γ-actin, α-tubulin, ATP-synthetase) are probably components of sperm.

(b) *Genealogical analyses*

In order to contrast genealogical patterns in seminal proteins with those for control genes, we examined patterns of molecular variation for the entire coding regions of six genes in allopatric populations of *G. firmus* and *G. pennsylvanicus*. The three control genes (*COQ7*, *GuKc* and *PPase*) are expressed in accessory gland, but have not been identified as components of the seminal fluid and are not probably involved in reproductive functions. The three putative seminal protein genes were chosen from the list in table 1 and include a serine protease (*AG-0308F*) and two genes encoding proteins of unknown function (*AG-0005F* and *AG-0334P*). Sequence datasets ranged from 432 bp for *PPase* to 957 bp for *COQ7*. Excluding autapomorphies, *PPase*, *COQ7* and *GuKc* contained very few segregating sites (*PPase*, 0/432; *COQ7*, 1/568: *GuKc* 5/597), whereas two of the seminal protein genes have a higher number of phylogenetically informative sites (24/924 and 9/958 for *AG-0005F* and *AG-0334P*, respectively).

For all variable genes, inferred genealogies showed an incongruent phylogenetic signal (partition homogeneity test, $p<0.01$). For each locus, we used a Bayesian framework to test the hypothesis that *G. firmus* and *G. pennsylvanicus* form exclusive genealogical groups (see §2). The hypothesis of exclusivity was rejected for variable control genes; *COQ7* and *GuKc* did not show evidence of species-specific clustering (and *PPase* exhibited no variation). For *COQ7* and *GuKc*, the credible interval of the difference between the posterior distribution of the likelihood scores of unconstrained and constrained (i.e. reciprocally monophyletic) genealogies is entirely positive (does not overlap 0; table 2). Thus, there is over 5% chance that the true value of the difference between unconstrained and monophyletic genealogies is 0.

In addition to examining patterns of variation for three accessory gland expressed control genes, we also reanalysed data from Broughton & Harrison (2003) on patterns of variation in *G. firmus* and *G. pennsylvanicus* for four

Table 2. Posterior means of the differences between the posterior distributions of unconstrained and constrained (reciprocal monophyly) gene genealogies and their 95% credible intervals. (The observed gene genealogies are consistent with genealogical exclusivity between *G. firmus* and *G. pennsylvanicus* if the credible interval of the log-likehood difference overlaps zero (values shown in italics). Data for intron sequences are from Broughton & Harrison (2003).)

seminal proteins	*AG-0005F*	*AG-0308F*	*AG-0334P*	
	4.128	7.013	*0.879*	
	(*−0.995–18.610*)	(0.040–14.985)	(*−6.617–8.562*)	
'control genes' exons	*COQ7*	*GuKc*		
	7.121	29.987		
	(1.041–13.406)	(21.380–38.814)		
'control genes' introns	*calmodulin*	*cytochrome c*	*EF1α*	*Pgi*
	7.568	15.957	11.887	36.680
	(1.502–13.948)	(7.343–24.933)	(3.891–19.912)	(25.351–48.072)

nuclear gene introns (table 2). Data from these control genes (*EF-1α*, *calmodulin*, *phosphoglucose isomerase* and *cytochrome c*) reveal extensive haplotype sharing and also fail to support an hypothesis of exclusivity for the two cricket species.

In contrast to the results for control genes, two of the seminal protein genes (*AG-0005F* and *AG-0334P*) exhibit nearly fixed differences between the cricket species, and the gene genealogies of these two loci do not differ significantly from one showing an exclusive relationship between the two species (table 2). The ML tree for *AG-0334P* (figure 1) reveals two major, well-supported clades. One clade includes all haplotypes from *G. firmus* populations plus one haplotype from a single *G. pennsylvanicus* population. The second clade includes all but three haplotypes from *G. pennsylvanicus* (figure 1). The ML tree for *AG-0005F* also shows significant divergence between the two species, with only haplotypes from *G. firmus* in one well-supported clade, and all haplotypes from *G. pennsylvanicus* populations, together with three haplotypes from *G. firmus* populations, in the second clade (figure 1). Correspondingly, the estimated levels of molecular divergence between species for *AG-0334P* and *AG-0005F* are higher than for those genes that show non-exclusive gene genealogies (table 3). Nucleotide sequence data for the third seminal protein gene, *AG-0308F*, showed a pattern of molecular variation similar to that for control genes, with a very small proportion of informative sites (3/666) and no evidence of exclusivity.

For none of the loci were the results significantly affected by the type of phylogenetic analysis applied. For all loci, ML and maximum parsimony trees had consistent topologies (Wilcoxon signed-rank test p values ranged from 0.629 to 1). Distance-based analyses (which group alleles on the basis of overall similarity reflecting approximate patterns of relatedness among alleles) and parsimony networks (reflecting uncertainty generated by recombination) also yielded the same results as the ML gene genealogies.

Comparisons of amino acid sequences within *AG-0005F* or *AG-0334P* revealed substantial differences between alleles' characteristic of the two species. With the exception of a single allele found in the Scranton, PA population, *AG-0334P* is unambiguously sorted by species. The clades supporting 'exclusivity' of *G. firmus* and *G. pennsylvanicus* have high bootstrap support values (71 and 91%, respectively). *AG-0334P* has two principal allelic forms that are characterized by three radical and three conservative amino acid differences (Asn75Thr, His85Leu, Ile166Asn, Gly168Cys, Asp191Glu and Glu205Lys; see the electronic supplementary material S4). The ML genealogy for *AG-0005F* also shows two clades (94% bootstrap support) associated with two species characteristic allelic forms of the protein. These alleles are highly divergent; aside from the *G. pennsylvanicus* alleles found in presumed *G. firmus* populations, there are five fixed radical amino acid substitutions between the two species (Phe117Val, Ile167Thr, Lys182Gln, Gln228Leu(His) and Leu280Phe; see the electronic supplementary material S5). The topologies are robust to model choice, and different amino acid substitution models (JTT, PAM or PMB) yielded the same ML topologies.

(c) *Test of selection*

Seminal protein loci that show substantial differentiation between *G. firmus* and *G. pennsylvanicus* (*AG-0005F* and *AG-0334P*) exhibit an elevated rate of non-synonymous substitutions (d_N) compared with loci that fail to show evidence of species-specific clustering (*AG-0308F*, *COQ7* and *GuKc*). The 95% credible intervals of the posterior probability distribution of the selection parameter (ω) for the pair of genes showing strong haplotype frequency differences between species include values greater than 1, as expected if positive selection plays an important role in the diversification of these genes (table 4). Pairwise comparisons show that the ω values for these genes are generally higher than for genes showing no evidence of differentiation between species (table 5).

4. DISCUSSION

During the speciation process, genomes of diverging lineages will be mosaics with respect to molecular genealogy (Ting *et al.* 2000). Under a neutral model of divergence, stochastic lineage sorting that leads to exclusivity or reciprocal monophyly is expected to occur over long periods of time in organisms such as field crickets with large effective population sizes and only a single generation each year. Throughout most of the genome, diverging lineages may continue to share alleles or haplotypes owing to introgression and/or the retention of ancestral polymorphisms. By contrast, the genealogies of speciation/barrier genes and genes responsible for lineage-specific adaptations are more likely to reveal differentiation or exclusivity between species.

Many recent studies of closely related species (flies, mice, sunflowers, moths and butterflies) have demonstrated genealogical discordance and/or differential

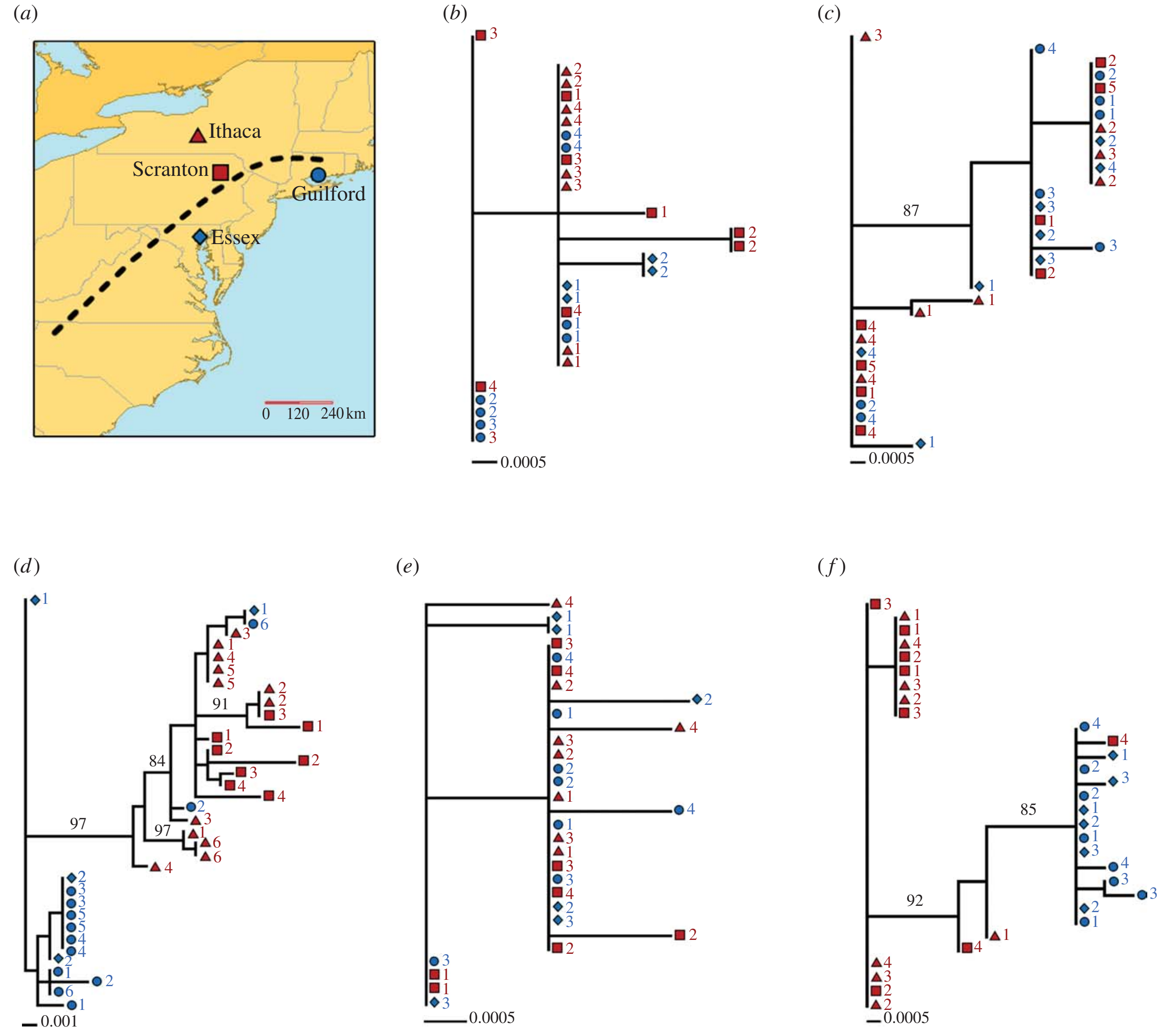

Figure 1. Contrasting DNA gene genealogies between control ((*b*) *COQ7* and (*c*) *GuKc*; *PPase* not shown owing to absence of variation) and seminal protein genes ((*d*) *AG-0005F*, (*e*) *AG-0308F* and (*f*) *AG-0334P*). (*a*) The map shows the approximate location of the hybrid zone (dashed line) between *G. firmus* (blue) and *G. pennsylvanicus* (red), and the location of the populations sampled in this study. Numbers on the branches represent bootstrap support values over 75%. Numbers after symbols represent the identity of the individuals sequenced. Scale bars represent divergence measured as substitutions per nucleotide. Note the scale differences between *AG-0005F* and all other loci.

Table 3. Divergence estimates between *G. firmus* and *G. pennsylvanicus*. (*k* is the average number of nucleotide differences between populations; *Da* is an estimate of the net number of nucleotide substitutions per site between species; and *Dxy* is the average number of substitutions per site between the two species.)

locus	*k*	*Da*	*Dxy*
AG-0005F	10.653	0.0084	0.0156
AG-0308F	0.696	<0.0001	0.0002
AG-0334P	6.884	0.0056	0.0072
COQ7	0.833	<0.0001	0.0014
GuKc	2.446	0.0003	0.0041

introgression (Rieseberg *et al.* 1999; Beltran *et al.* 2002; Machado & Hey 2003; Ting *et al.* 2000; Panithanarak *et al.* 2004; Dopman *et al.* 2005; Payseur & Nachman 2005). These studies, which represent genome scans or scans of the sex chromosome, have also defined genome regions that show little evidence of introgression or for which closely related species are exclusive groups (Machado & Hey 2003; Panithanarak *et al.* 2004; Dopman *et al.* 2005). One pattern that has emerged is that genome regions associated with fixed differences in chromosome arrangements introgress less (Rieseberg *et al.* 1999; Noor *et al.* 2001; Ortiz-Barrientos *et al.* 2002) and may show accelerated rates of protein evolution (Navarro & Barton 2003). In the absence of a linkage map for field crickets, we cannot examine patterns of differentiation at defined genomic regions. Instead, we have used a proteomic approach to target a particular class of genes (seminal proteins), in a search for regions of exclusivity or differentiation within the cricket genome.

Because the cricket species *G. pennsylvanicus* and *G. firmus* have diverged very recently, we expect that most regions of the genome will continue to reveal shared

Table 4. Detection of selection in the presence of intragenic recombination. (Values represent the posterior probability means of the transition–transversion ratio (κ), the rate of synonymous transversion (μ), recombination rate (ρ) and selection parameter (ω). ω values in parentheses represent the 95.5% credible intervals of the posterior probability distributions. Values in italics represent ω values with credible intervals that include estimates consistent with positive selection acting on that locus (i.e. $\omega > 1$).)

locus	κ	μ	ρ	ω
AG-0005F	2.465	0.166	0.075	*0.602 (0.343–1.173)*
AG-0308F	4.541	0.054	11.576	0.043 (0.010–0.123)
AG-0334P	0.858	0.055	0.017	*1.091 (0.201–2.702)*
COQ7	10.917	0.059	10.78	0.216 (0.017–0.613)
GuKc	2.462	0.072	0.018	0.325 (0.037–0.736)

Table 5. Posterior mean of the difference between the estimated ω values for each pair of loci (e.g. $\omega_{AG\text{-}0005F}-\omega_{AG\text{-}0308F}$) and its 95% credible interval. (Two ω values are significantly different if the credible interval of their difference does not include zero $^{*}p<0.05$, $^{**}p<0.01$.)

	AG-0005F	*AG-0308F*	*AG-0334P*	*COQ7*	*GuKc*
AG-0005F		0.460** (0.197–0.975)	−0.591 (−2.273 to 0.296)	0.369* (0.132–1.417)	0.203 (−0.512 to 0.738)
AG-0308F			−1.071** (−3.754 to 0.179)	−0.214 (−0.605 to 0.074)	−0.280 (−0.992 to 0.014)
AG-0334P				0.895** (0.321–1.973)	0.790* (0.043–3.500)
COQ7					0.914 (−0.477 to 5.147)

ancestral polymorphisms (Broughton & Harrison 2003). Indeed, previous efforts to identify diagnostic differences between the two species have generally been unsuccessful and most genomic regions that have been surveyed exhibited evidence of shared polymorphisms (Harrison & Bogdanowicz 1997; Broughton & Harrison 2003).

Because field cricket females are highly promiscuous, and because fertilization barriers are partly responsible for reproductive isolation between *G. firmus* and *G. pennsylvanicus* (Harrison 1983), genes encoding seminal proteins are likely to be under strong diversifying selection and may be involved in barriers to gene exchange. If so, genes encoding seminal proteins should be strongly differentiated, and their genealogies should be congruent with the reproductive differences between these two species.

The gene genealogies of the 'control loci' examined in this study (*COQ7*, *GuKc* and *PPase*) confirmed what we learned from previous genealogical analyses of four nuclear gene introns (Broughton & Harrison 2003). Although levels of variation differ substantially among these loci, *G. firmus* and *G. pennsylvanicus* do not form exclusive groups at any locus and much of the polymorphism appears to predate lineage splitting (i.e. shared alleles do not show geographical structure). Thus, as expected, a large portion of the field cricket genome has remained undifferentiated after the evolution of reproductive isolation. By contrast, two of the three loci that encode seminal proteins (*AG-0005F* and *AG-0334P*) are highly divergent between species, and sequence differences at these loci include radical amino acid substitutions that could have functional significance for the reproductive biology of the crickets. Given the recent divergence of the two cricket species (estimated to be $0.1N_e$ by Broughton & Harrison 2003) random sorting of ancestral polymorphisms is improbable. Thus, the observed pattern suggests that *AG-0005F* and *AG-0334P* might play an important role in species differentiation. However, the extent to which these two genes represent outliers in the cricket genome needs to be confirmed by a more extensive survey of both 'control' and seminal protein genes.

During the early stages of the speciation process, positive selection is likely to be responsible for the origin of genomic differentiation between diverging lineages. Lineage-specific selective sweeps will produce regions of exclusivity surrounding loci under selection. Our results suggest that the pattern of differentiation observed for *AG-0005F* and *AG-0334P* may be a consequence of rapid evolution due to selection acting on these loci. The selection coefficient parameter (ω) is significantly higher at these loci than those loci showing lack of genealogical exclusivity, and the 95% credible intervals of ω include values above one. One possible scenario is that directional postmating sexual selection acting on these seminal proteins has led to the accumulation of amino acid substitutions in allopatric populations, which have relatively recently come into secondary contact. Available evidence certainly supports the notion that the two cricket species diverged in allopatry and that the current hybrid zone represents the coming together of previously isolated populations (Willett *et al.* 1997). Alternatively, since seminal proteins are exclusively expressed in the male accessory gland, they might be free from antagonistic pleiotropic effects and evolve more rapidly than control genes that are likely to have broader spatial and temporal patterns of expression (Duret & Mouchiroud 2000; Winter *et al.* 2004). The ability of our tests to detect positive selection is relatively limited (see Hughes 2007), and the estimated distribution of ω values cannot unequivocally rule out a scenario of lack of constraint.

For both *AG-0005F* and *AG-0334P*, alleles characteristic of one species are found at low frequency in the other species. For *AG-0005F*, two crickets from Guilford and one cricket from Essex each carry one *G. pennsylvanicus* allele; for *AG-0334F*, a single Scranton individual carries a *G. firmus* allele. This pattern could be explained either by persistence of shared ancestral polymorphisms or by recent hybridization and introgression. The similarity of the 'foreign' alleles in one species to those currently found in the other species argues in favour of recent introgression. The consequences of the presence of foreign alleles at *AG-0005F* and *AG-0334P* for the reproductive phenotype of the individual crickets are not yet clear. That the two species share alleles, however, does not preclude a role for the protein products of *AG-0005F* and *AG-0334P* in barriers to gene exchange. Given our allele sampling, more extensive population genetic analyses are necessary to assess the relative extent of allele sharing at these loci. Furthermore, the effects of variation in amino acid sequence can be assessed by testing the reproductive phenotype of knock-down (RNAi) males as well as by pairing males and females with different allelic combinations.

A proteomics approach for identifying candidate speciation/barrier genes certainly shows great promise. Using a combination of genomic, proteomic, phylogenetic and population genetic techniques we have been able to identify candidate barrier genes in field crickets, but functional analyses are obviously required before we can understand the role of seminal proteins in reproductive isolation.

We thank Steve Bogdanowicz for his help in the laboratory and Jonathan Richmond for his comments on statistical methods and critical reading of the manuscript. This research was supported by NSF grants DEB-0111613 (Willie Swanson, PI) and DEB-0639904 to R.G.H.

REFERENCES

Andrés, J. A. & Arnqvist, G. 2001 Genetic divergence of the seminal signal-receptor system in houseflies: the footprints of sexually antagonistic coevolution? *Proc. R. Soc. B* **268**, 399–405. (doi:10.1098/rspb.2000.1392)

Andrés, J. A., Maroja, L. S., Bogdanowicz, S. M. & Harrison, R. G. 2006 Molecular evolution of seminal proteins in field crickets. *Mol. Biol. Evol.* **23**, 1574–1584. (doi:10.1093/molbev/msl020)

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

Beltran, M., Jiggins, C. D., Bull, V., Linares, M., Mallet, J., McMillan, W. O. & Berminghan, E. 2002 Phylogenetic discordance at the species boundary: comparative gene genealogies among rapidly radiating *Heliconius* butterflies. *Mol. Biol. Evol.* **19**, 2176–2190.

Braswell, W. E., Andrés, J. A., Maroja, L. S., Harrison, R. G., Howard, D. J. & Swanson, W. J. 2006 Identification and comparative analysis of accessory gland proteins in Orthoptera. *Genome* **49**, 1069–1080. (doi:10.1139/G06-061)

Broughton, R. E. & Harrison, R. G. 2003 Nuclear gene genealogies reveal historical, demographic and selective factors associated with speciation in field crickets. *Genetics* **163**, 1389–1401.

Clark, N. L. & Swanson, W. J. 2005 Pervasive adaptive evolution in primate seminal proteins. *PLoS Genet.* **1**, e35. (doi:10.1371/journal.pgen.0010035)

Clark, N. L., Aagaard, J. E. & Swanson, W. J. 2006 Evolution of reproductive proteins from animals and plants. *J. Reprod. Fertil.* **131**, 11–22.

Clement, M., Posada, D. & Crandall, K. A. 2000 TCS: a computer program to estimate gene genealogies. *Mol. Ecol.* **9**, 1657–1659. (doi:10.1046/j.1365-294x.2000.01020.x)

Coyne, J. A. & Orr, H. A. 2004 *Speciation*. Sunderland, MA: Sinauer Associates.

Dopman, E. B., Peréz, L., Bogdanowicz, S. M. & Harrison, R. G. 2005 Consequences of reproductive barriers for genealogical discordance in the European corn borer. *Proc. Natl Acad. Sci. USA* **102**, 14 706–14 711. (doi:10.1073/pnas.0502054102)

Drummond, A. J. & Rambaut, A. 2003*a BEAST: Bayesian evolutionary analysis sampling trees*, v. 1.3. Oxford, UK: University of Oxford.

Drummond, A. J. & Rambaut, A. 2003*b Tracer*, v. 1.3. Oxford, UK: University of Oxford.

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

Felsentein, J. 2005 *Phylip*, v. 3.6. Seattle, WA: University of Washington.

Fiumera, A. C., Dumont, B. L. & Clark, A. G. 2005 Sperm competition ability in *Drosophila melanogaster* associated with variation in male reproductive proteins. *Genetics* **169**, 243–257. (doi:10.1534/genetics.104.032870)

Fricke, C., Arnqvist, G. & Amaro, N. 2006 Female modulation of reproductive rate and its role in postmating prezygotic isolation in *Callosobruchus maculatus*. *Funct. Ecol.* **20**, 360–368. (doi:10.1111/j.1365-2435.2006.01102.x)

Fry, C. L. & Wilkinson, G. S. 2004 Sperm survival in female stalk-eyed flies depends on seminal fluid and meiotic drive. *Evolution* **58**, 1622–1626.

Goldman, N. & Yang, Z. 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736.

Harrison, R. G. 1979 Speciation in North American field crickets: evidence from electrophoretic comparisons. *Evolution* **33**, 1009–1023. (doi:10.2307/2407463)

Harrison, R. G. 1983 Barriers to gene exchange between closely related cricket species. I. Laboratory hybridization studies. *Evolution* **37**, 245–251. (doi:10.2307/2408333)

Harrison, R. G. 1985 Barriers to gene exchange between closely related cricket species. II. Life cycle variation and temporal isolation. *Evolution* **39**, 244–259. (doi:10.2307/2408360)

Harrison, R. G. 1986 Pattern and process in a narrow hybrid zone. *Heredity* **56**, 337–349. (doi:10.1038/hdy.1986.55)

Harrison, R. G. & Bogdanowicz, S. M. 1997 Patterns of variation and linkage disequilibrium in a field cricket hybrid zone. *Evolution* **51**, 493–505. (doi:10.2307/2411122)

Harrison, R. G. & Rand, D. M. 1989 Mosaic hybrid zones and the nature of species boundaries. In *Speciation and its consequences* (eds D. Otte & J. A. Endler), pp. 111–133. Sunderland, MA: Sinauer Associates.

Hellberg, M. E. & Vacquier, V. D. 1999 Rapid evolution of fertilization selectivity and lysin cDNA sequences in teguline gastropods. *Mol. Biol. Evol.* **16**, 839–848.

Howard, D. J. 1999 Conspecific sperm and pollen precedence and speciation. *Annu. Rev. Ecol. Syst.* **30**, 109–132. (doi:10.1146/annurev.ecolsys.30.1.109)

Hudson, R. R. 1983 Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**, 183–201. (doi:10.1016/0040-5809(83)90013-8)

Hughes, A. L. 2007 Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* **99**, 364–373. (doi:10.1038/sj.hdy.6801031)

Li, N. & Stephens, M. 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233.

Machado, C. A. & Hey, J. 2003 The causes of phylogenetic conflict in a classic *Drosophila* species group. *Proc. R. Soc. B* **270**, 1193–1202. (doi:10.1098/rspb.2003.2333)

Metz, E. C. & Palumbi, S. R. 1996 Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. *Mol. Biol. Evol.* **13**, 397–406.

Mueller, J. L., Ram, K. R., McGraw, L. A., Bloch Qazi, M. C., Siggia, E. D., Clark, A. G., Aquadro, C. F. & Wolfner, M. F. 2005 Cross-specific comparison of *Drosophila* male accessory gland protein genes. *Genetics* **171**, 131–143. (doi:10.1534/genetics.105.043844)

Navarro, A. & Barton, N. H. 2003 Chromosomal speciation and molecular divergence accelerated evolution in rearranged chromosomes. *Science* **300**, 321–324. (doi:10.1126/science.1080600)

Noor, M. A. F. & Feder, J. L. 2006 Speciation genetics: evolving approaches. *Nat. Rev. Genet.* **7**, 851–861. (doi:10.1038/nrg1968)

Noor, M. A. F., Grams, K. L., Bertucci, L. A. & Reiland, J. 2001 Chromosomal inversions and the reproductive isolation of species. *Proc. Natl Acad. Sci. USA* **98**, 12 084–12 088. (doi:10.1073/pnas.221274498)

Ortiz-Barrientos, D., Reiland, J., Hey, J. & Noor, M. A. F. 2002 Recombination and the divergence of hybridizing species. *Genetica* **116**, 167–178. (doi:10.1023/A:1021296829109)

Panithanarak, T., Hauffe, H. C., Dallas, J. F., Glover, A., Ward, R. G. & Searle, J. B. 2004 Linkage-dependent gene flow in a house mouse chromosomal hybrid zone. *Evolution* **58**, 184–192.

Payseur, B. A. & Nachman, M. W. 2005 The genomics of speciation: investigating the molecular correlates of X chromosome introgression across the hybrid zone between *Mus domesticus* and *Mus musculus*. *Biol. J. Linn. Soc.* **84**, 523–534. (doi:10.1111/j.1095-8312.2005.00453.x)

Peitz, B. 1988 Effects of seminal vesicle fluid components on sperm motility in the house mouse. *J. Reprod. Fertil.* **88**, 169–176.

Posada, D. & Crandall, K. A. 1998 Modeltest: testing the model of DNA substitution. *Bioinformatics* **14**, 817–818. (doi:10.1093/bioinformatics/14.9.817)

Putman, A. S., Scriber, J. M. & Andolfatto, P. 2007 Dicordant divergence times among Z-chromosome regions between two ecologically distinct swallowtail butterfly species. *Evolution* **61**, 912–927. (doi:10.1111/j.1558-5646.2007.00076.x)

Rieseberg, L. H., Whitton, J. & Gardner, K. 1999 Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics* **152**, 713–727.

Riginos, C., Wang, D. & Abrams, A. J. 2006 Geographic variation and positive selection on M7 lysin, an acrosomal sperm protein in mussels (*Mytilus* spp.). *Mol. Biol. Evol.* **23**, 1952–1965. (doi:10.1093/molbev/msl062)

Ross, C. L. & Harrison, R. G. 2002 A fine-scale analysis of the mosaic hybrid zone between *Gryllus firmus* and *Gryllus pennsylvanicus*. *Evolution* **56**, 2296–2312.

Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. 2002 Tree-Puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502–504. (doi:10.1093/bioinformatics/18.3.502)

Staden, R. 1996 The Staden sequence analysis package. *Mol. Biotechnol.* **5**, 233–241.

Stephens, M., Smith, N. J. & Donnelly, P. 2001 A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–989. (doi:10.1086/319501)

Swanson, W. J. & Vacquier, V. D. 2002 The rapid evolution of reproductive proteins. *Nat. Rev. Genet.* **3**, 137–144. (doi:10.1038/nrg733)

Swofford, D. L. 2003 *PAUP**, v. 4.0B10. Sunderland, MA: Sinauer.

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*. *Proc. Natl Acad. Sci. USA* **97**, 5313–5316. (doi:10.1073/pnas.090541597)

Tram, U. & Wolfner, M. F. 1999 Male seminal fluid proteins are essential for sperm storage in *Drosophila melanogaster*. *Genetics* **153**, 833–844.

Viscuso, R., Narcisi, L., Sottile, L. & Violetta Brundo, M. 2001 Role of male accessory glands in spermatodesm reorganization in Orthoptera Tettigonioidea. *Tissue Cell* **33**, 33–39. (doi:10.1054/tice.2000.0147)

Willett, C. S., Ford, M. J. & Harrison, R. G. 1997 Inferences about the origin of a field cricket hybrid zone from a mitochondrial DNA phylogeny. *Heredity* **79**, 484–494.

Wilson, D. J. & McVean, G. 2006 Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics* **172**, 1411–1425. (doi:10.1534/genetics.105.044917)

Winter, E. E., Goodstadt, L. & Ponting, C. P. 2004 Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res.* **14**, 54–61. (doi:10.1101/gr.1924004)

Woerner, A. E., Murray, P. C. & Hammer, M. F. 2007 Recombination-filtered genomic datasets by information maximization. *Bioinformatics* **23**, 1851–1853. (doi:10.1093/bioinformatics/btm253)

Wu, C. I. 2001 The genic view of the process of speciation. *J. Evol. Biol.* **14**, 851–865. (doi:10.1046/j.1420-9101.2001.00335.x)

Yang, Z., Swanson, W. J. & Vacquier, V. D. 2000 Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Mol. Biol. Evol.* **17**, 1446–1455.