JOURNAL OF Evolutionary Biology

doi: 10.1111/jeb.13021

Genetic differentiation and adaptive evolution at reproductive loci in incipient *Drosophila* species

F. C. ALMEIDA*† & R. DESALLE*‡

*Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY, USA
†Department of Biology, New York University, New York, NY, USA
‡Division of Invertebrate Zoology, American Museum of Natural History, New York, NY, USA

Keywords:

accessory gland proteins;
Bayesian model selection;
candidate genes;
Drosophila repleta group;
positive selection;
reproductive isolation;
speciation.

Abstract

Accessory gland proteins (Acps) are part of the seminal fluid of male *Drosophila* flies. Some Acps have exceptionally high evolutionary rates and evolve under positive selection. Proper interactions between Acps and female reproductive molecules are essential for fertilization. These observations lead to suggestions that fast evolving Acps could be involved in speciation by promoting reproductive incompatibilities between emerging species. To test this hypothesis, we used population genetics data for three sibling species: *D. mayaguana*, *D. parisiena* and *D. straubae*. The latter two species are morphologically very similar and show only incipient reproductive isolation. This system allowed us to examine Acp evolution at different time frames with respect to speciation and reproductive isolation. Comparing data of 14 Acp loci with data obtained for other genomic regions, we found that some Acps show extraordinarily high levels of divergence between *D. mayaguana* and its two sister species *D. parisiena* and *D. straubae*. This divergence was likely driven by adaptive evolution at several loci. No fixed nucleotide differences were found between *D. parisiena* and *D. straubae*, however. Nevertheless, some Acp loci did show significant differentiation between these species associated with signs of positive selection; these loci may be involved in this early phase of the speciation process.

Introduction

A speciation gene has been defined as one that contributes to reproductive isolation, having ecological, sexual or post-mating effects (Wu & Ting, 2004). In a recent review, Nosil & Schluter (2011) suggested the term should exclusively refer to genes that contribute to the speciation process itself, not just contemporary reproductive isolation, which could have evolved after speciation. In this way, a speciation gene should fulfil three main criteria: i) the gene is involved in contemporary reproductive isolation, ii) the locus must have diverged during incipient speciation, and iii) divergence at the locus must have had a significant effect on reproductive isolation ('effect size') (Nosil & Schluter, 2011). Most candidate speciation genes identified so far were first identified based on criterion i, including genes involved in hybrid sterility or inviability, genes that contribute to ecological differences between sister species, and genes involved in premating sexual isolation and gametic incompatibility (a compilation of examples can be found in Nosil & Schluter, 2011). In most of these cases, however, testing these candidate genes for criteria ii and iii may not be possible because of the criterion's time dependence. Here, we focused on criterion ii, or the 'the magnifying glass' approach (Via, 2009), by using a system that includes incipient species to identify candidate speciation genes among those encoding accessory gland proteins (Acps): fast-evolving, reproduction-related loci that due to these characteristics could potentially participate in the evolution of reproductive isolation between incipient species.

Correspondence: Francisca C. Almeida, Departamento de Ecología, Instituto de Ecología, Genética y Evolución de Buenos Aires (CONICET), Av. Intendente Güiraldes s/n, Ciudad Universitaria, Pabellón II, Buenos Aires 1428, Argentina.
Tel.: +54 11 45763300; fax: +54 11 45763354; e-mail: falmeida@nyu.edu

Prezygotic reproductive isolation is an important aspect of the speciation process (Civetta & Singh, 1998; Orr *et al.*, 2007; Matute & Coyne, 2009). Although behaviour is the main mechanism in premating isolation, proteins involved in fertilization are also likely to be associated with some instances of prezygotic postmating reproductive isolation (Ritchie, 2007). Such proteins are sometimes called gamete recognition proteins because they can ensure specificity and fertilization efficiency, with obvious consequences for reproductive isolation (Vacquier, 1998; Swanson & Vacquier, 2002; Manier *et al.*, 2013). Loci that may affect prezygotic reproductive isolation, such as the fertilization proteins of marine invertebrates with external fertilization, represent interesting candidates for speciation genes (Palumbi, 2009; Hart *et al.*, 2014). Their importance is particularly evident in cases of speciation in the absence of a strong barrier to migration because in this scenario prezygotic isolation is predicted to occur before postzygotic isolation (Coyne & Orr, 1997).

In *Drosophila*, some proteins expressed in the male accessory glands (accessory gland proteins – Acps) fulfil the expected requisites for fertilization proteins that may be involved in early reproductive isolation and species divergence. Besides their fundamental role in fertilization, some Acps have extraordinarily high substitution rates accompanied by evidence of positive selection in the divergence of orthologs of closely related species (Aguadé *et al.*, 1992; Aguadé, 1999; Begun *et al.*, 2000; Swanson *et al.*, 2001; Swanson, 2003; Wagstaff & Begun, 2005; Almeida & DeSalle, 2008). Acps are produced in the male accessory glands and are passed to the female reproductive tract together with sperm during insemination. Inside the female reproductive tract, Acps participate in several necessary steps of fertilization such as sperm storage and the induction of female responses to insemination such as egg-laying stimulation (Wolfner, 2002; Chapman & Davies, 2004; Ram & Wolfner, 2007). The fundamental role of seminal proteins in determining fertilization has been demonstrated in studies of interspecific crosses between *Drosophila suzukii* and *D. pulchrella* (Fuyama, 1983), and *D. mojavensis* and *D. arizonae* (Kelleher & Markow, 2007). The fast evolutionary rates observed in Acps in molecular studies have been attributed to selective regimes imposed by sperm competition, cryptic female choice and sexual conflict, all of which are mechanisms of sexual selection made evident by empirical studies on *Drosophila* reproduction (reviewed in Ram & Wolfner, 2007; Avila *et al.*, 2011).

An appropriate system to identify candidate speciation genes meeting Nosil & Schluter's (2011) criterion ii would be one involving incipient species. Incipient speciation is characterized by a pair of populations that are still able to interbreed, but with a certain degree of differentiation, thus showing the potential to speciate. An interesting case of incipient speciation is found in a cactophilic *Drosophila* species subcluster endemic to the Caribbean. The *Drosophila mayaguana* subcluster is composed of three morphocryptic species, *D. mayaguana*, *D. parisiena* and *D. straubae*. The *D. mayaguana* triad belongs to the *D. repleta* species group, which presents high female remating rates favouring sperm competition and cryptic female choice. Accordingly, in this species group, Acps show even stronger evidence of evolution by positive selection as compared to other *Drosophila* groups (Wagstaff & Begun, 2005; Almeida & DeSalle, 2009).

Among the three species, *D. mayaguana* is reproductively isolated from the other two, whereas *D. parisiena* and *D. straubae* can produce fertile hybrids in the laboratory and likely in nature (Wasserman, 1992; Wasserman & Wasserman, 1992). Nevertheless, laboratory experiments showed that both copulation and production of viable offspring are significantly less efficient (i.e. occurs less frequently) in interspecific crosses when compared to control, intraspecific crosses (Wasserman, 1992). *Drosophila parisiena* differs from *D. straubae* in chromosomal rearrangements and male genitalia morphology (Heed & Grimaldi, 1991; Wasserman & Wasserman, 1992). Moreover, there seems to be an ecological differentiation between the three species: it was recorded in the field that *D. parisiena* shows high specificity in the type of cactus exploited, whereas *D. straubae* and *D. mayaguana* may use various cactus species. A molecular dating of the *repleta* group suggests the split of the *mayaguana* triad is very recent, likely within the last million year (Oliveira *et al.*, 2012). Using hypervariable sequences from five different regions of the nuclear and mitochondrial genomes, O'Grady *et al.* (2002) uncovered only one fixed site difference between *D. mayaguana* and the other two species. In that study, no fixed difference at the nucleotide sequence level, however, was found between *D. parisiena* and *D. straubae*, which are believed to be sister species (Fig. 1a). This result is compatible with the incipient nature of this split.

Here, we characterize differentiation and the dynamics of selection on 14 Acp loci in three time frames: before, during and after speciation. Our objective was to survey these loci in search of one or more Acps that may meet criterion ii for a candidate speciation gene. We hypothesize that some Acps may be involved in early species divergence due to its adaptive, accelerated evolution. In this way, we searched for genes that show increased divergence levels between closely related and incipient species that also show signatures of positive selection both between and within species.

Materials and methods

Samples

Samples of *D. mayaguana* (11), *D. parisiena* (18) and *D. straubae* (9) from several localities were obtained

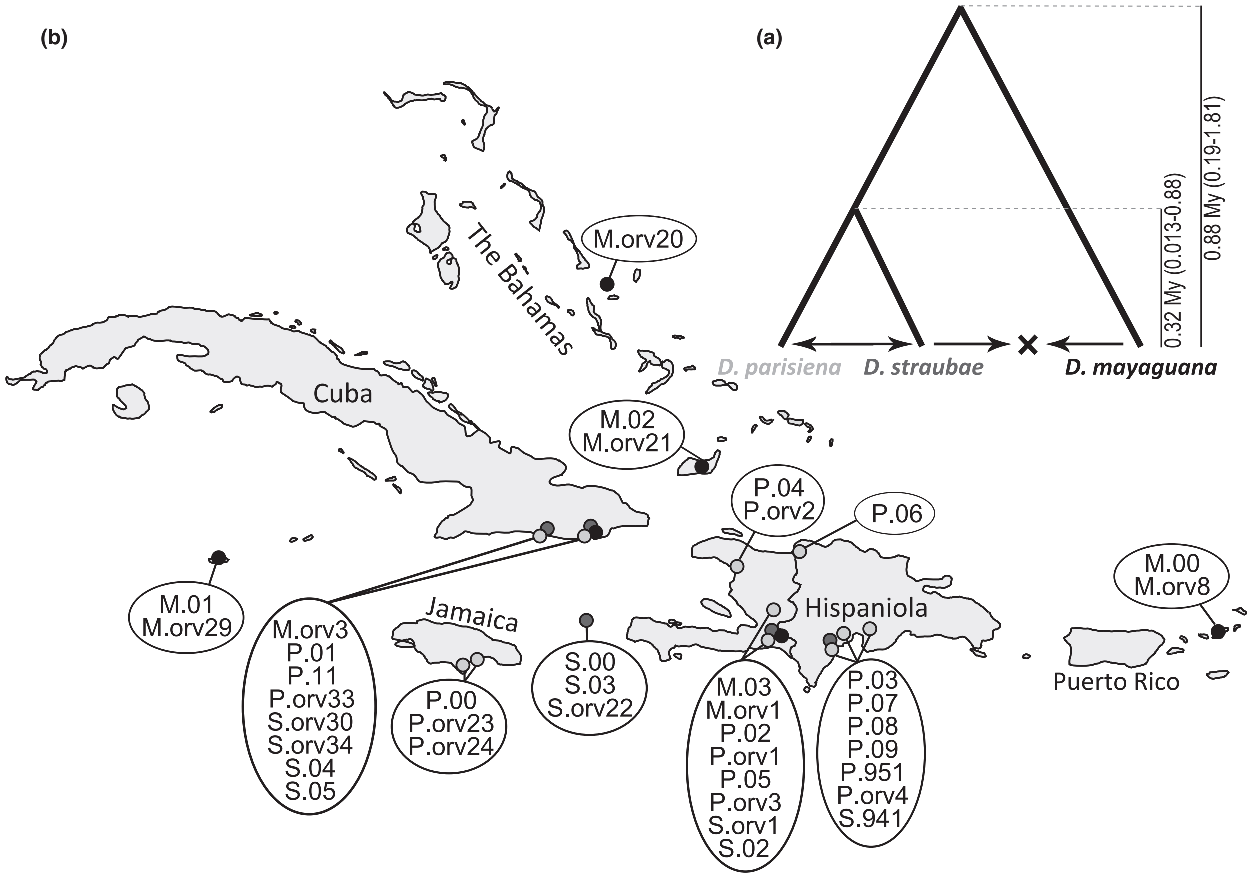

Fig. 1 (a) Relationships and divergence time of the species in the *mayaguana* triad. Arrows represent gene flow. (b) Collection sites of the samples used in this study. M., P. and S. prefix samples of *Drosophila mayaguana*, *Drosophila parisiena* and *Drosophila straubae*, respectively.

from the Tucson Stock Center and the Ambrose Monell Cryo Collection at the American Museum of Natural History (Table S1, Fig. 1b). Samples from both sources were originally obtained from isofemale lines established by W. Heed and M. Wasserman in collections made in the 1980s. DNA was extracted from one to three individual flies (pooled together) per line using the DNeasy (QIAGEN) kit.

Loci and molecular data

Fourteen genes obtained in an EST library of accessory glands of *D. mayaguana* (Almeida & DeSalle, 2009) were sequenced for this study (Table S2). According to a previous study (Almeida & DeSalle, 2008), eight of these genes had significant results in site model tests for positive selection based on ω, the ratio between the rate of nonsynonymous substitution (d_N) and the rate of synonymous substitution (d_S). One of the genes included here, *may82*, also obtained in the accessory gland cDNA library of *D. mayaguana*, was not classified as an Acp for lack of a signal peptide sequence, which is a marker of secretory genes. Nevertheless, the cDNA sequence did not seem to be complete at its 5′ end, so it is possible *may82* is in fact an Acp. Some loci included here, however, were not included in the Almeida & DeSalle (2008) study due to amplification failure in species other than those of the *mayaguana* subcluster. Of these 14 genes, 6 did not show similarities to any *D. mojavensis* Acp or any non-Acp gene expressed in the accessory glands of that species and for this reason they carry the prefix 'may'. Two additional non-Acp genes were included in some of the analyses for comparison: the nuclear gene *hunchback*, which had been previously sequenced in the *mayaguana* species by O'Grady *et al.* (2002), and the cytochrome oxidase I (*COI*) gene, widely used as a molecular barcoding gene for showing divergence even between closely related species.

Gene sequences were amplified from total genomic DNA using the primers listed in the Table S3. Forward and reverse sequences were obtained for each PCR product with an ABI 3730 automated sequencer. Sequences were edited with SEQUENCHER v4.15 (Gene Codes) and multiple alignments were obtained with MAFFT v5 (Katoh *et al.*, 2005). Codon alignment and intron trimming were done in MACCLADE v4 (Maddison

& Maddison, 2000), using the *D. mayaguana* cDNA sequence as a guide. A variable number of individuals were sequenced per locus and species (Table S2). All sequences were deposited in the GenBank with accession numbers KY087045-KY087490.

Species divergence

Divergence at Acp loci was assessed with two approaches. First, we performed a population aggregation analysis (PAA; Davis & Nixon, 1992; DeSalle *et al.*, 2005), which is an alternative to the use of trees in species delimitation and consists of identifying fixed differences between species using the nucleotide alignment. This analysis is a very simple way to assess fixed site differences between species. Then, we carried out phylogenetic analyses for each locus separately. Trees were obtained using the maximum likelihood (ML) and maximum parsimony (MP) methods. MP heuristic tree searches were performed using 200 random stepwise additions and tree bisection reconnection (TBR) swapping. These analyses were run with PAUP* (Swofford, 2003). ML trees were obtained with the RAXML program (Stamatakis, 2006), using the GTRGAMMA substitution model and 10 independent runs. Sequences of at least two other species were used as outgroups, which, depending on availability, could be *D. mojavensis*, *D. arizonae*, *D. mulleri*, *D. navojoa* and/or *D. huyalasi* (mostly from Almeida & DeSalle, 2008). For the COI locus, we also built a network using the statistical parsimony method (Templeton *et al.*, 1992) as implemented in the program TCS (Clement *et al.*, 2000).

Selection tests and intraspecific variation

To characterize the genetic divergence between species, we obtained maximum likelihood estimates of divergence rates at synonymous (d_S) and nonsynonymous (d_N) sites and ω (d_N/d_S) with the program PAML (Yang, 1997, 1998). The role of positive selection in the divergence between the species of the *mayaguana* triad was assessed with the McDonald–Kreitman test (McDonald & Kreitman, 1991; Gibbs & O'Bryan, 2007) and the significance of departures from neutrality was obtained with the G-test. Direction of selection was estimated with the neutrality index (NI; Rand & Kann, 1996) and the direction of selection statistic (DoS; Stoletzki & Eyre-Walker, 2011).

As an additional test for adaptive evolution, we applied likelihood-based tests for detecting sites and phylogenetic branches under positive selection implemented in the package HYPHY (Kosakovsky Pond *et al.*, 2005) on the Datamonkey webserver (Delport *et al.*, 2010). Namely, we performed the REL (Kosakovsky Pond & Frost, 2005; Kosakovsky Pond *et al.*, 2011) and MEME (Murrell *et al.*, 2012) analyses, which detect individual sites under diversifying or episodic positive selection, and the BRS (branch-site REL, Kosakovsky Pond *et al.*, 2011) analysis, which detects lineages (branches) under diversifying selection. Because we were dealing with population samples with little sequence variation, we limited these analyses to the most variable loci studied herein: *Acp7*, *mayAcp58* and *mayAcp74*.

The role of selection in intraspecific variation was assessed with four neutrality tests based on the allele frequency spectrum: Tajima's D (Tajima, 1989), Fu and Li's D, Fu and Li's F (Fu & Li, 1993), and Fay and Wu's H (Fay & Wu, 2000). Statistical significance of neutrality tests was assessed using 1000 coalescent simulations (Hudson, 1990). Intraspecific nucleotide variation was characterized in the three species by θ_w, an estimate of the population mutation rate (4Nμ; Watterson, 1975), and the nucleotide diversity (π; Nei & Li, 1979). For both estimates, we used the Jukes and Cantor correction (Lynch & Crease, 1990). Intralocus recombination was assessed using the method described by Pond *et al.* (2006). Population differentiation was estimated with N_{ST} (Lynch & Crease, 1990) and the statistical significance of population structure was obtained with a permutation test on K_{ST} (Hudson *et al.*, 1992). Analyses of intraspecific sequence variation, neutrality tests, population differentiation, recombination and the MK test for positive selection were performed using DNASP (Rozas *et al.*, 2003).

Model selection and Approximate Bayesian Computation

To test our results using an alternative approach, we employed Bayesian model selection (reviewed in Bertorelle *et al.*, 2010) to determine whether a demographic model with selection fit the data obtained for the locus *Acp7* better than a neutral model. For that we first simulated nucleotide variation data using a simple coalescent model under two scenarios: neutral (no selection) and selection in one population (in this case *D. parisiena*). In both cases, we set the effective population size (N_e) to 100 000 and assumed it to be the same for both species. We also assumed equal probability of migration from one species to the other and set the migration rate to Nm = 2 following the results obtained with *COI* and other Acps apparently not affected by selection in these species. Finally, the parameter θ_w was set to that empirically observed in the *Acp7* locus. The neutral model data were simulated 1 million times. The selection model had the same parameters as the neutral model, except for the appearance of a new, positively selected, dominant mutation in *D. parisiena* at 0.05 ($4N_e$) generations after its split from *D. straubae*. We ran 1 million selection model simulations, with selection coefficients randomly distributed between 0.001 and 0.1. Simulations were run on the program *msms* (Ewing & Hermisson, 2010).

We then used the program *MSABC* (Pavlidis *et al.*, 2010) to obtain summary statistics for the simulated data sets and the observed data for the *Acp7* locus. We used eight statistics for the model selection: *D. parisiena*'s S, θ_W, Tajima's *D* and ZnS (Kelly, 1997), F_{ST}, the fraction of shared and private polymorphism between *D. parisiena* and *D. straubae*, and Fay and Wu's H. Finally, Bayesian model selection was performed to obtain the posterior probability of each model (neutral and selection) given the observed data using a tolerance of 0.01. The test was run with the 'abc' package implemented in R (Csilléry *et al.*, 2010, 2011), using the *postpr* function and logistic regression. We then used approximate Bayesian computation (ABC) to estimate the selection coefficient with highest posterior probability to have generated the observed data, using a regression–rejection approach. All commands used in these analyses are available in the supplementary material.

Table 1 Number of fixed differences between *Drosophila mayaguana* and *Drosophila parisiena* and estimates of ω, d_N and d_S between species.

Locus	Diff *may* × *par**	ω†	d_N	d_S	Site model ω‡
Acp1	23 (11)	0.417	0.029	0.069	**0.686**
mayAcp2a	15 (4)	0.344	0.029	0.083	n.a.
mayAcp2b	21 (3)	1.767	0.066	0.037	n.a.
Acp7	43 (9)	1.128	0.104	0.092	**1.275**
Acp19	20 (8)	1.130	0.041	0.036	**0.896**
Acp25	20 (5)	1.213	0.044	0.037	**0.640**
Acp45	26	0.775	0.060	0.078	**0.800**
mayAcp57	14 (7)	0.272	0.020	0.074	0.179
mayAcp58	33 (5)	0.570	0.088	0.155	**0.635**
mayAcp63	18 (10)	0.064	0.007	0.113	0.163
mayAcp65	18	0.954	0.076	0.080	**0.789**
mayAcp74	159	1.334	0.267	0.201	**0.770**
mayAcp77	24 (2)	0.358	0.037	0.102	n.a.
may82	53 (4)	1.408	0.122	0.086	n.a.
hunchback	1	0.049	0.003	0.054	**0.236**
COI	0	0.001	0.0002	0.227	n.a.

*Between parentheses are the numbers of fixed substitutions in noncoding regions.

†Estimates obtained from comparisons between samples M.00 and P.00 unless these sequences were not available, in which case other sequences were randomly chosen.

‡From Almeida & DeSalle (2008). Bold values highlight the loci for which statistical tests for presence of positively selected sites were significant with $P < 0.01$; n.a. = not available.

Results

Species divergence at Acp loci

A PAA comparison between *D. mayaguana* and the two other species (*D. parisiena* and *D. straubae*) revealed fixed nucleotide substitutions in all Acp loci analysed (a total 370 sites across loci), whereas no fixed differences were observed at the *COI* locus (Table 1). O'Grady *et al.* (2002), studying 5 hypervariable regions in the same species, found only one fixed substitution (at the *hunchback* locus) across 1385 bp. A comparison between *D. straubae* and *D. parisiena*, however, revealed no fixed difference along the 9709 bp of Acp genes analysed, at the *COI* locus, nor at the hypervariable regions surveyed in O'Grady *et al.* (2002). The Acp gene trees agreed with the PAA: *D. mayaguana* was usually monophyletic, whereas *D. parisiena* and *D. straubae* were always paraphyletic in relation to one another (Figs 2, S1 and S2). Different reconstruction methods gave very similar results. The phylogenetic analysis of the *COI* gene, however, retrieved a paraphyletic *D. mayaguana*, with two clades: one including all the Bahamas sequences, closer to the *D. parisiena* / *D. straubae* clade, and another, more basal, including the remaining sequences. A network analysis of the *COI* sequences revealed similar results (Fig. S3). These latter results and the low number of fixed differences at a mitochondrial locus attest to the recent divergence of the species triad, whereas the high number of fixed substitutions between *D. mayaguana* and *D. parisiena* at Acps is in agreement with fast evolutionary rates and divergence at these loci.

Positive selection in the divergence of *Drosophila mayaguana*

To assess positive selection in the first split within the *mayaguana* triad (i.e. between *D. mayaguana* and the ancestor of *D. parisiena* + *D. straubae*), we estimated ω (d_N/d_S) and employed the MK test. Considering the results of both analyses, 10 of 14 Acp loci appear to have had adaptive evolution since that split. We estimated ω between *D. mayaguana* and *D. parisiena* samples (Tables 1 and Table S4) and found that 6 of the 14 genes analysed had $\omega > 1$ and two others had $\omega > 0.6$ (a number that suggests at least some sites are under positive selection, Almeida & DeSalle, 2008). The *hunchback* and *COI* loci had considerably lower ω values when compared to the ω values of most Acp genes, largely reflecting a low d_N, which is in agreement with purifying selection acting in these non-Acp loci. The neutrality index (NI < 1) and the direction of selection (DoS > 0) statistics showed an excess of fixed nonsynonymous substitutions in all but four loci. A significant deviation from the null hypothesis of neutral evolution was detected with the MK test, contrasting *D. mayaguana* with the other two species (pooled together), in four genes: *mayAcp2a*, *Acp25*, *mayAcp57* and *may82* (Table 2). To check whether selection-driven differentiation between *D. parisiena* and *D. straubae* could be biasing the test results, we reran the MK analysis using only *D. parisiena* and *D. mayaguana*. The results did not change for most genes, except for *mayAcp2b*, which this time showed significant deviation from neutrality ($P = 0.015$). Interestingly, the results of the MK test do not totally agree with the ω values.

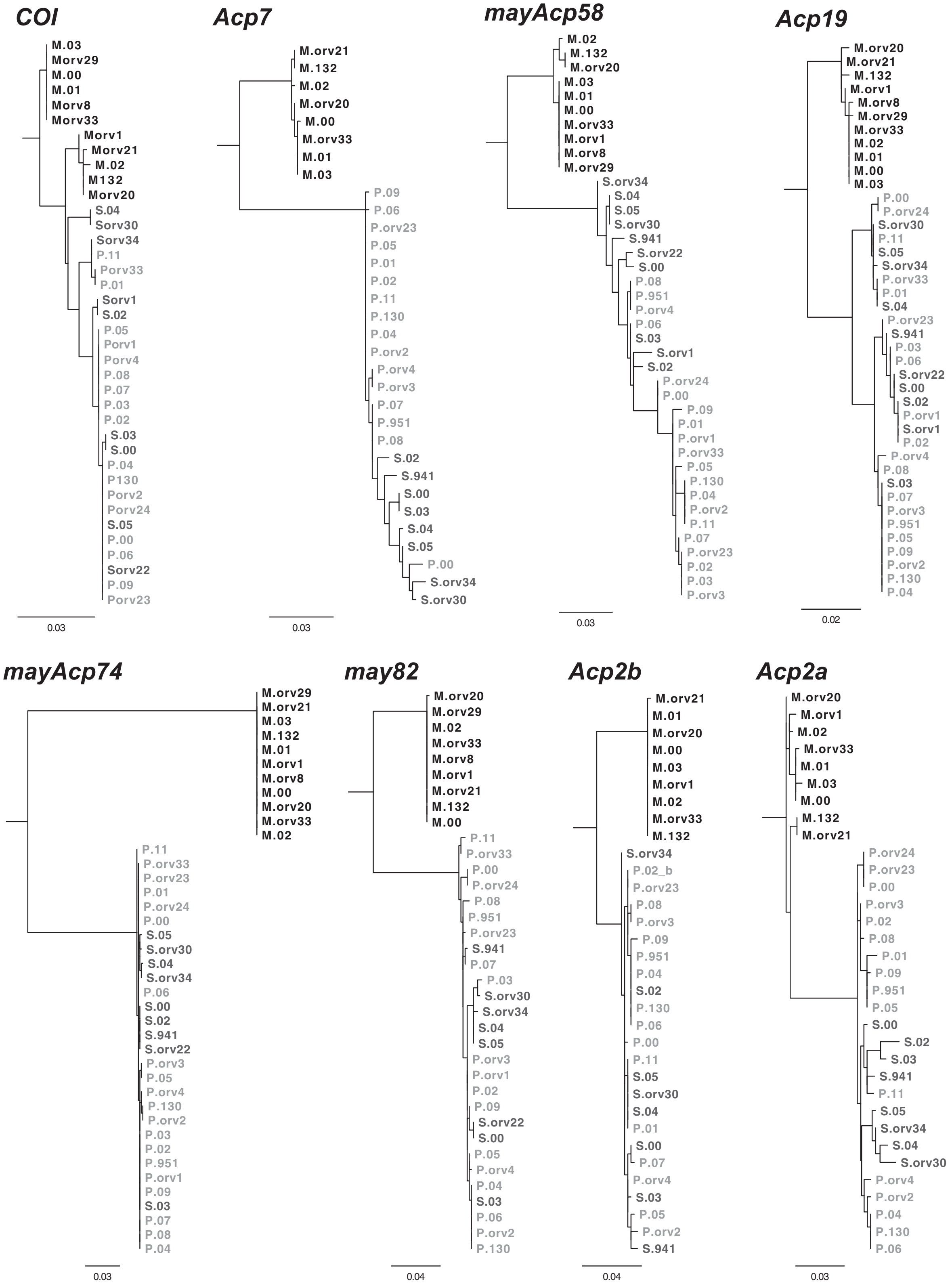

Fig. 2 ML gene trees of some of the loci analysed in this study. M., P. and S. prefix samples of *Drosophila mayaguana*, *Drosophila parisiena* and *Drosophila straubae*, respectively. Remaining gene trees are in the Figs S1 and S2.

}

Table 2 McDonald–Kreitman (MK) test between *Drosophila mayaguana* and *Drosophila parisiena* + *Drosophila straubae*. Shown are the number of silent (Sil.) and replacement (Rep.) substitutions among polymorphic and fixed sites, significance values as obtained with the G-test, neutrality index (NI) and the direction of selection index (DoS).

	Polymorphic		Fixed				
Locus	Sil.	Rep.	Sil.	Rep.	G-test P	NI	DoS
Acp1	10	13	5	7	0.918	0.93	0.02
Acp2a	25	5	7	6	**0.048**	0.23	0.30
Acp2b	8	7	5	16	0.069*	0.27	0.30
Acp7	17	16	17	24	0.388	0.67	0.10
Acp19	10	18	6	13	0.769	0.83	0.04
Acp25	12	1	7	12	**0.001**	0.05	0.56
Acp45	3	8	8	20	0.935	1.07	−0.01
mayAcp57	23	2	8	8	**0.002**	0.09	0.42
mayAcp58	21	21	10	23	0.083	0.44	0.20
mayAcp63	1	2	7	1	0.152	14	−0.54
mayAcp65	4	6	10	13	0.852	1.15	−0.04
mayAcp74	0	9	39	122	n.a.	n.a.	−0.24
mayAcp77	3	6	10	12	0.532	1.67	−0.12
may82	16	8	13	40	**0.000**	0.16	0.42
COI	33	1	0	0	n.a.	n.a.	n.a.

*$P = 0.015$ if only *D. parisiena* samples were included in the test, instead of both *D. parisiena* and *Drosophila straubae*.
In bold are significant *P* values at alpha=0.05.

Table 3 Population subdivision statistics in comparisons between *Drosophila parisiena* and *Drosophila straubae*, including the number of polymorphic sites found in the two species (total), the number of fixed sites in *D. straubae* (*str*), the number of fixed sites in *D. parisiena* (*par*) and the number of polymorphisms shared by the two species (Shared).

locus	N_{ST}	*Nm**	K_{ST}†	Total	*str*	*par*	Shared
Acp1	0.159	1.33	**0.082**	35	9	15	11
Acp2a	0.237	0.81	**0.135**	27	9	11	7
Acp2b	0.089	2.56	0.043	14	5	3	6
Acp7	0.388	0.39	**0.265**	28	6	15	7
Acp19	0.077	2.98	0.040	23	1	3	19
Acp25	0.041	5.86	0.019	9	0	5	4
mayAcp58	0.512	0.24	**0.321**	35	14	13	8
mayAcp74	0.307	0.56	**0.184**	9	3	5	1
may82	0.110	2.03	**0.056**	20	5	6	9
all Acps	0.314	0.547	**0.184**	200	52	76	72
COI	0.104	2.16	0.060	19	0	10	9

*Nm estimates were obtained from Nst.
†Bold values were significant according to permutation tests; $P < 0.001$ for all but the *may82* locus that had $P < 0.05$.

Divergence between *Drosophila parisiena* and *Drosophila straubae*

Ideally, a good candidate for a speciation gene should show complete monophyly between two incipient species that are not, respectively, monophyletic in other loci. Such a pattern was not observed in any of the genes analysed herein. The locus that showed the most obvious species clustering pattern in the phylogenetic analysis was *Acp7*, although the loci *Acp1*, *Acp2a*, *mayAcp58* and *mayAcp74* also showed a trend in this direction (Figs 2, S1 and S2). Assuming that there is gene flow between *D. parisiena* and *D. straubae*, as suggested by evidence of hybridization, the absence of fixed nucleotide substitutions and the phylogenetic results, we estimated genetic differentiation between these two species using population genetic statistics (Table 3). The degree of differentiation at the nucleotide level (N_{ST}) was variable among loci. Differentiation was significant at six loci as measured by permutation tests on K_{ST}. The highest N_{ST} values were observed at the *Acp7*, *mayAcp58* and may*Acp7*4 loci.

Intraspecific DNA sequence variation and evolution

To assess the role of selection within species, we first estimated intraspecific variation for each species of the *D. mayaguana* subcluster (Table S5). Levels of nucleotide diversity (π) in Acp genes were quite low, but comparable to the diversity found in *D. sechellia*, also an island endemic (Kern *et al.*, 2004). *Drosophila straubae* and *D. parisiena* had similar intraspecific diversity at most loci analysed. Exceptions were the *Acp7* and the *COI* loci. Considering all three species, five of the loci (*Acp2b*, *Acp7*, *Acp19*, *mayAcp58* and *mayAcp74*) presented higher nonsynonymous than synonymous diversities, suggesting current intraspecific adaptive evolution. This would explain why the MK test was not significant for some loci even though they had $\omega > 1$, reinforcing the argument that the MK test may not be accurate for genes with intraspecific selection (Hughes, 2007). As in the interspecific comparisons (ω), even at the Acp loci where the nonsynonymous/synonymous ratio was smaller than 1, this ratio was much higher than the one observed at the *COI* locus.

Next, we applied neutrality tests based on the allele frequency spectrum (Tajima's D, Fu and Li's D and F, and Fay and Wu's H) to the intraspecific data (Table 4). First, we used an algorithm to detect intralocus recombination as it could bias the results of these tests; recombination was detected in only two loci (*Acp2a* and *mayAcp58*); therefore, we concluded it had little influence in our overall results. The results of the Fu and Li's F-test were practically identical to those obtained with the Fu and Li's D tests; thus, only the results of the latter test are presented. Significant, negative Tajima's D and Fu and Li's D, as found at the *Acp1* and *Acp7* loci in *D. parisiena*, are generally interpreted as footprints of a selective sweep. Selective sweeps are in accordance with the particularly low levels of sequence variation in this species at these loci (Table S5). Fu and Li's D and Fay and Wu's H were significant for *D. parisiena* at the *COI* locus. The positive Fu and Li's D statistics suggests balancing selection or population subdivision. A negative Fay and Wu's H is generally interpreted as a sign of positive selection, but it has been shown to be highly affected by population subdivision because it relies on the frequency of high-frequency substitutions (Li, 2011). In fact, excluding *D. parisiena* samples from Cuba (they clustered with *D. straubae* samples from the same islands in the gene tree), the neutrality tests had very different, nonsignificant results (Fu and Li's D = 0.666, n.s.; Fay and Wu's H = −0.076, n.s.). In the absence of other evidence that *COI* is evolving under selection and the geographic clustering revealed in the *COI* gene tree (Fig. 2) and network (Fig. S3), we interpret these results as a consequence of the population structure observed at this mitochondrial locus.

In *D. straubae*, deviations from neutrality were observed at the *Acp2a*, *Acp19*, *mayAcp58* and *mayAcp74* loci. For the former two loci, the test statistics suggested populational structure (positive Tajima's D in the sliding window analysis), which is in agreement with the geographic clustering of individuals in their gene trees. Given the amount of nonsynonymous polymorphism observed at *Acp19*, it is possible that differential selection at this locus in isolated population might be contributing to the population structure in addition to genetic drift. In an opposite trend, *mayAcp58* and *mayAcp74* showed departure from neutrality with a negative Fay and Wu's H, which is a sign of ongoing positive selection. Additionally, the locus *may82* had borderline significant negative Tajima's D and Fay and Wu's H in both *D. parisiena* and *D. straubae*.

The observed differences among loci in the results of neutrality tests would be expected if other factors besides demography (such as selection) were also acting on the evolution of some Acps. However, because often the influence of different factors in these tests can not be sorted out, final conclusions on the role of selection were made based on additional evidence presented herein.

Selection and the divergence between *D. parisiena* and *D. straubae*

In the population differentiation analysis, three loci, *Acp7*, *mayAcp58* and *mayAcp74* showed particularly high levels of differentiation between *D. parisiena* and *D. straubae* (Table 3). Two of these loci also showed a tendency to cluster samples by species in the phylogenetic analyses and all three presented signs of evolution by positive selection (summarized in Table 5). To evaluate the hypothesis that positive selection at these loci plays a role in the divergence of *D. parisiena* and *D. straubae*, we performed a more detailed analysis of the nucleotide differences between the two sister species.

Table 4 Results of neutrality tests based on allele frequency spectrum applied to *Drosophila parisiena* and *Drosophila straubae* samples.

Taxa	Locus	N*	Tajima's D	Fu and Li's D	Fay and Wu's H	Sliding window Tajima's D†
D. parisiena	*Acp1*	16	−1.118	−0.888	−0.92	**354–528 (−)**
	Acp2a	15	−0.154	−0.665	0.200	n.s.
	Acp2b	16	−0.293	−0.81	−0.683	n.s.
	Acp7	16	−**1.584**	−**2.15**	−0.050	**558–719 (−)**
	Acp19	16	0.248	0.451	−4.067	n.s.
	mayAcp25	15	−0.601	−1.522	−0.133	n.s.
	mayAcp58	16	1.008	1.118	0.667	n.s.
	mayAcp74	16	0.353	−1.05	−0.752	n.s.
	may82	16	−1.282	−1.51	−3.233	n.s.
	COI	18	0.077	**1.499**	−**6.065**	n.s.
D. straubae	*Acp1*	8	−0.747	−1.055	0.786	n.s.
	Acp2a	8	−0.156	−0.406	2.000	251–400 (+)
	Acp2b	8	−0.413	0.045	0.143	n.s.
	Acp7	8	−0.028	0.419	−1.500	n.s.
	Acp19	8	1.037	0.419	1.929	**51–275 (+)**
	mayAcp25	8	1.091	1.219	0.714	n.s.
	mayAcp58	8	−0.501	−0.016	−5.286	n.s.
	mayAcp74	8	0.015	1.339	−**2.714**	n.s.
	may82	7	−1.01	−0.238	−2.238	440–628 (−)
	COI	8	0.552	1.106	−2.929	n.s.

Bold values had $P < 0.05$, and underlined values had $P < 0.10$.

*Number of samples included in the analysis.

†Numbers refer to the nucleotide position range where the test was significant and signs between parentheses indicate whether the test statistics was positive (+) or negative (−).

mayAcp58 – Differentiation and high levels of nonsynonymous variation

mayAcp58 presented high intraspecific variation (Table S5), the highest levels of differentiation between *D. parisiena* and *D. straubae* (Table 3), and signs of positive selection in *D. straubae* (Tables 4 and 5). In the sliding window analysis (Fig. S5), the region with highest Fay and Wu's H (nucleotide positions 434 to 583) had mostly nonsynonymous polymorphisms (7 of 8), including three sites that are fixed in *D. straubae*, two sites that are fixed in *D. parisiena* and three sites that are polymorphic in both species but with marked frequency difference (Fig. S6). Two nonsynonymous substitutions located within the conserved domain (positions 113 to 460) also showed marked frequency difference between *D. parisiena* and *D. straubae*. Evidence of positively selected codons both within (sites 115 and 320) and outside (sites 485, 488, and 500) the conserved domain was obtained with the REL and MEME analyses (Table 5).

mayAcp74 – Strong evidence of selection at two levels

In *mayAcp74*, 9 sites (in 7 codons), all showing nonsynonymous variation, were polymorphic in *D. parisiena* and *D. straubae*. Codon bias cannot explain the lack of variation in synonymous sites since very low bias was found in this gene in both *D. mayaguana* (ENC = 57.09) and *D. parisiena* (ENC = 58.47). All but one polymorphic site were found within the conserved trypsin domain. These substitutions involve 2 amino acids sites that are 4 and 7 amino acids apart from the enzyme active site and one site that is 2 amino acids apart from a binding site. Six of the 7 polymorphic codons were identified by either the REL or MEME analyses as evolving under positive selection (Table 5). Confirming previous findings (Almeida and DeSalle, 2008), the BSR analysis identified the branch separating *D. mayaguana* from the other two species with evidence of episodic diversifying selection (average $\omega = 1.54$, $\omega_3 = 6.0$, $p_3 = 0.32$, corrected $P < 0.0002$). The BSR analysis also identified several branches within the *D. straubae* and *D. parisiena* clade as having at least some codons with $\omega > 1$; these results, however, lacked statistical support, probably due to the low sequence variability in this part of the tree. The overall low levels of intraspecific variation are in accordance with this gene being under strong selection.

Acp7 – Selective sweep in *Drosophila parisiena*?

Acp7 showed the strongest pattern of taxonomic structuring among *D. parisiena* and *D. straubae* samples in the phylogenetic analysis (Fig. 2), significant differentiation between these species (Table 3), increased nonsynonymous polymorphism and signs of a recent selective sweep in *D. parisiena* (Tables 4 and 5). More interestingly, there is a nearly fixation of one nonsynonymous substitution (from serine to glycine) differentiating the two species. This substitution is derived in *D. parisiena* when polarized with *D. mayaguana*, which

Table 5 Summary of tests and statistics to detected signs of positive selection at Acp loci.

	Mean ω	MK* (*P*)	REL BF > 50† (BF > 20)	MEME‡ $P < 0.1$	Tajima's D§	Fu and Li's D	Fay and Wu's H	$\pi_N > \pi_S$
Acp1	0.615		–¶	–	par–			
mayAcp2a	0.352	<0.05	–	–	str+			
mayAcp2b	2.586	<0.1	–	–	may–			par, str
Acp7	1.152		2 (7)	4	par–	par–		may, par, str
Acp19	1.471		–	–	str+			par, str
Acp25	1.070	<0.01	–	–				
mayAcp45	0.755		–	–	–	–	–	–
mayAcp57	0.172	<0.01	–	–	–	–	–	–
mayAcp58	0.725	<0.01	5	2			str-	may
mayAcp63	0.069		–	–	–	–	–	–
mayAcp65	0.910		–	–	–	–	–	–
mayAcp74	1.318	–	(1)	5			str-	par, str
mayAcp77	0.349		–	–	–	–	–	–
may82	1.268	<0.001	–	–	par-, str-, may-	par-		
COI	0.001		–	–	may+	par+	par-	

*McDonald–Kreitman test, as shown in Table 2; absence of an entry means the test was nonsignificant (as for the other tests shown in this table).

†Number of sites with Bayes factor (BF) ? 50 and BF > 20 (between parentheses) to be under selection as estimated with the random effects likelihood method.

‡Number of sites with $P < 0.10$ in the mixed effects model of evolution test for detecting sites under episodic selection.

¶For all the frequency spectrum neutrality tests, the sign of the test statistics is given together with the species for which the test was significant: may = *D. mayaguana*, par = *D. parisiena*, and str = *D. straubae*.

§A dash means the analysis was not applied to that locus.

is in agreement with a selective sweep in *D. parisiena*. Further, independent evidence of positive selection at this locus was obtained with the REL and MEME analyses, which identified a total of 9 putatively selected codons (Table 5), although the BSR test failed to detect positive selection in the branch separating *D. parisiena* from *D. straubae* (see Discussion).

As an alternative approach to test the hypothesis of the involvement of positive selection in the differentiation between *D. parisiena* and *D. straubae*, we employed a Bayesian model selection analysis. Assuming equal population sizes, the test supported the selection model over the neutral model with a posterior probability (PP) of 0.996. Although neutrality was clearly rejected, it was not possible to estimate with confidence the selection coefficient most likely to have generated the observed data. Depending on the tolerance level (0.01, 0.005 or 0.001) applied, the mode of the estimated selection coefficients ranged between 0.009 and 0.041 and the cross-validation gave very high error rates at all three tolerance levels (Fig. S4). Although the much lower genetic variation observed in *D. parisiena* as compared to *D. straubae* at *Acp7* supports a selective sweep in the former species, it could have been instead a consequence of historically smaller population sizes in that species. A smaller N_e in *D. parisiena* would affect the statistics used in the ABC. Thus, it was not surprising that assuming that *D. parisiena* had an effective population 1/3 that of *D. straubae* in the model selection analysis, the selection model (PP = 0.575) was only slightly favoured over the neutral one. Although the *COI* locus showed a pattern of genetic variation similar to that shown by *Acp7*, all other loci analysed showed similar levels of variation in the two species, suggesting similar population sizes (Table S5).

Discussion

Divergence and differentiation in the *mayaguana* triad

The divergence of the *mayaguana* triad is very recent, with estimates for the age of the *D. mayaguana* split at around 0.88 million years (95% CI 0.19–1.81; Oliveira *et al.*, 2012). In such cases, incongruence between species delimitations and gene trees is very common (e.g. Wang *et al.*, 1997; Ting *et al.*, 2000; Belfiore *et al.*, 2008) as a result of incomplete lineage sorting (i.e. ancient shared polymorphism and deep coalescence) and/or introgression by hybridization (Pamilo & Nei, 1988; Funk & Omland, 2003). In fact, paraphyly and even polyphyly between closely related species of *Drosophila* have been observed in several species groups (e.g. Wang *et al.*, 1997; Kliman *et al.*, 2000). Nevertheless, in genes involved in early species divergence, lineage sorting should be faster and introgression should be less common (Wu & Ting, 2004; Nosil & Schluter, 2011). Thus, closely related species will attain monophyly faster at these genes. According to these expectations, most Acp loci showed *D. mayaguana* as monophyletic. In contrast, the *COI* gene tree shown here and 2 other non-Acp gene trees obtained by O'Grady *et al.* (2002) recovered a paraphyletic *D. mayaguana* in relation to [*D. parisiena* + *D. straubae*]. Although this result suggests that Acps may be involved in early species divergence, it does not provide direct evidence that Acps were involved in speciation itself.

On the other hand, no surveyed locus, allele or SNP has supported the monophyly of *D. parisiena* and *D. straubae*, which is in agreement with previous observations that they can produce fully viable hybrids in the laboratory and that hybridization likely occurs in natural populations (Wasserman & Wasserman, 1992). This pattern characterizes them as incipient species, a necessary scenario to test criterion ii (Nosil & Schluter, 2011) for candidate speciation genes. Three Acp loci, however, showed evidence of genetic differentiation between species, with Nm < 1, accompanied by evidence of selection. Had these two species been fully separated species, they should have particular loci where different alleles were completely fixed in both species, characterizing reproductive isolation (e.g. Turner & Hahn, 2010). This is obviously not the case, but would a $F_{ST} > 0.3$ between *D. parisiena* and *D. straubae,* as estimated for some Acp loci (*Acp7*, *mayAcp* 58 and *mayAcp74*) be enough to characterize these genes as outliers? Variation in F_{ST} is naturally expected under neutral regimes due to drift and variable mutation rates (Nosil *et al.*, 2009). This is especially true in insular species with isolated populations, as in the case of the present triad of species. It would be necessary to have genomic F_{ST} averages for comparisons, but this information is not available for these three species. However, demonstrating that relatively high F_{ST} is associated with divergent selection would provide an additional source of evidence, since this kind of selective regime is another hallmark of speciation genes, particularly in animals (Wu & Ting, 2004; Nosil *et al.*, 2009; Via, 2009; Nosil & Schluter, 2011).

The role of selection in early divergence at Acp loci

When incipient species are found in sympatry, in the absence of any temporal or microclimatic barrier that may preclude gene flow, differentiation can only be maintained by strong disruptive selection, either in ecological or reproductive traits (Dieckmann & Doebeli, 1999; Higashi *et al.*, 1999; Nosil *et al.*, 2009). Several Acp loci, including some of the ones analysed herein, have been previously shown to have elevated ω in multispecies comparisons between members of the *Drosophila mulleri* species complex (Almeida & DeSalle, 2008). In addition to the loci previously reported, we found two other Acp loci with elevated ω among the four newly analysed and showed that some Acp loci have experienced adaptive evolution even in the short amount of time since the *mayaguana* triad started to diverge. Signs of non-neutral evolution in Acps came, for some loci, from interspecific comparisons (between *D. mayaguana* and the other two species), whereas for others the footprint of selection was found within species or, yet, at both levels (summarized in Table 5).

In agreement with a strong role of selection in the evolution of Acps, divergence at most nuclear, Acp loci was more profound than divergence at the mitochondrial *COI*. According to neutral expectations, the time required for lineage sorting in neutral genes is positively correlated with the effective population size of the species, which is smaller for mitochondrial genes (Pamilo & Nei, 1988; Birky *et al.*, 1989). Therefore, under an entirely neutral process, mitochondrial loci should show monophyletic species clustering four times faster than nuclear genes. In fact, the MK tests point to a role of selection in the divergence of *D. mayaguana* from its sister species in several Acp loci. Moreover, several loci showed high rates of intraspecific nonsynonymous substitution, which suggest that selection at Acps is an ongoing process and therefore could contribute in the early processes of species divergence rather than happening in a later stage, after reproductive isolation had taken place.

In the case of the incipient species *D. parisiena* and *D. straubae*, three loci (*Acp7*, *Acp58* and *mayAcp74*) showed particularly interesting patterns of genetic variation that suggest selection might be involved in their early divergence (Table 5). In particular, *Acp7* showed significant differentiation between *D. parisiena* and *D. straubae*, accompanied by several signs of positive selection within *D. parisiena* (Table 5). Interestingly, the locus *Acp7* also showed high levels of differentiation between *D. mojavensis* and *D. mojavensis baja* – a phylogenetically related pair of incipient species – accompanied by high nonsynonymous substitution rates (Wagstaff & Begun, 2005). *Acp7* has no predicted protein domain (InterPro), but a *D. arizonae* locus with high similarity was identified as transcriptional regulatory protein AlgP.

A problem to address is that many of the selection tests employed here are affected by other deviations from neutrality. Exceptions are the model-based tests, also employed here on the three above-mentioned loci. Two of these tests (MEME and REL), which are used to determine whether some sites on an alignment are under selection, had significant results in all three loci. Although they corroborate the other selection tests employed herein, they do not indicate where on the phylogeny positive selection has been acting on these genes. The third test, the branch-site model test BSR, was promising because it allows detecting particular branches on the phylogeny that have been influenced by diversifying selection. It is reasonable to posit that speciation gene should show high ω values on branches in a phylogeny that separate species that have recently formed. The high ω values, in fact, are evident for *Acp74*, on the branch separating *D. mayaguana* from the other two species. Nevertheless, the failure to detect high ω on branches separating *D. parisiena* and *D. straubae* is neither surprising nor invalidates the hypothesis that positive selection on these loci might be contributing to the divergence of these species. First, model-based tests are highly dependent on the phylogeny, which is quite unstable for these species as evidenced by lack of congruence for the phylogenies of the multiple genes used in this study. Second, these tests are known to have reduced power on short branches (Anisimova *et al.*, 2001; Kosakovsky Pond *et al.*, 2011), which is the case within the [*D. parisiena* + *D. straubae*] clade. In fact, if only a single amino acid change is involved in the divergence, as could be the case for *Acp7*, it would definitely not be detected by this test.

Possible scenarios

How could Acp divergence have evolved in the incipient species of the *mayaguana* triad? One possibility is that these two species started diverging in a period when the incipient species had allopatric distributions, such as in different islands. Alternatively, ecological speciation driven by host shift could have occurred in sympatry/parapatry (Via, 2009; Nosil & Schluter, 2011). Both scenarios would provide assortative mating favourable to divergent Acp evolution. If the ecological differentiation observed between *D. parisiena* and *D. straubae* affects hybrid fitness, Acp evolution could potentially be involved in reinforcement (Dobzhansky, 1940; Servedio, 2001), similarly to the reproductive character displacement observed in reproductive proteins of marine invertebrates (Geyer & Palumbi, 2003). The reinforcement hypothesis prediction that sympatric populations should show more pronounced reproductive isolation has in fact been observed in *D. parisiena* and *D. straubae* (Wasserman & Wasserman, 1992). It is also important to notice that the scenario analysed here may be transient and that eventually *D. parisiena* and *D. straubae* may merge back into a single species due to continual gene flow.

Conclusions

Despite limitations of the methods and samples, we believe we found some candidate Acp loci that could be involved in the early speciation process. These loci show signs of adaptive evolution associated with genetic differentiation between incipient species. A definitive test would include testing for the other speciation gene criteria, namely the effect of nucleotide variation on cross-species fertility (criteria i and iii of

Nosil & Schluter, 2011). This could be accomplished by experimentally testing whether the different alleles of the candidate genes revealed in our analysis differently affect cross-species mating efficiency. Genetic transformation would be very helpful in this task, by allowing directed changes in one gene at a time. Additionally, it would be desirable to evaluate hybrid fitness to test the reinforcement hypothesis.

Acknowledgments

We thank Julio A. Santillán for his help in the laboratory and Julio Rozas for comments on the manuscript. Funds for this research were provided by the Sackler Institute for Comparative Genomics (American Museum of Natural History), the Cullman Program in Molecular Systematics, and National Science Foundation (DEB 0129105 to R.D.).

F.C.A. was supported by the Henry McCraken Fellowship (New York University). [Correction added on 03 February 2017, after first online publication: The Acknowledgements was previously omitted and this has been corrected in this current version.]

References

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

Aguadé, M., Miyashita, N. & Langley, C.H. 1992. Polymorphism and divergence in the *Mst26A* male accessory gland gene region in *Drosophila*. *Genetics* **132**: 755–770.

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

Almeida, F.C. & DeSalle, R. 2009. Orthology, function, and evolution of accessory gland proteins in the *Drosophila repleta* group. *Genetics* **181**: 235–245.

Anisimova, M., Bielawski, J.P. & Yang, Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.* **18**: 1585–1592.

Avila, F.W., Sirot, L.K., LaFlamme, B.A., Rubinstein, C.D. & Wolfner, M.F. 2011. Insect seminal fluid proteins: identification and function. *Ann. Rev. Entomol.* **56**: 21–40.

Begun, D.J., Whitley, P., Todd, B.L., Waldrip-Dail, H.M. & Clark, A.G. 2000. Molecular population genetics of male accessory gland proteins in *Drosophila*. *Genetics* **156**: 1879–1888.

Belfiore, N.M., Liu, L. & Moritz, C. 2008. Multilocus phylogenetics of a rapid radiation in the genus *Thomomys* (Rodentia: Geomyidae). *Syst. Biol.* **57**: 294–310.

Bertorelle, G., Benazzo, A. & Mona, S. 2010. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Mol. Ecol.* **19**: 2609–2625.

Birky, C.W.J., Fuerst, P. & Maruyama, T. 1989. Organelle gene diversity under migration, mutation and drift: equilibrium expectations, approach to equilibrium, effects of heteroplasmic cells, and comparison to nuclear genes. *Genetics* **121**: 613–627.

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

Civetta, A. & Singh, R.S. 1998. Sex-related genes, directional sexual selection, and speciation. *Mol. Biol. Evol.* **15**: 901–909.

Clement, M., Posada, D. & Crandall, K.A. 2000. TCS: a computer program to estimate gene genealogies. *Mol. Ecol.* **9**: 1657–1660.

Coyne, J.A. & Orr, H.A. 1997. "Patterns of speciation in *Drosophila*" revisited. *Evolution* **51**: 295–303.

Csilléry, K., Blum, M.G., Gaggiotti, O.E. & François, O.E. 2010. Approximate Bayesian Computation (ABC) in practice. *Trends Ecol. Evol.* **25**: 410–418.

Csilléry, K., François, O.E. & Blum, M.G. 2011. abc: an R package for Approximate Bayesian Computation (ABC). in arXiv:1106.2793v1, ed.

Davis, J.I. & Nixon, K.C. 1992. Populations, genetic variation, and the delimitation of phylogenetic species. *Syst. Biol.* **41**: 421–435.

Delport, W., Poon, A.F., Frost, S.D.W. & Kosakovsky Pond, S.L. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26**: 2455–2457.

DeSalle, R., Egan, M.G. & Siddall, M. 2005. The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Phil. Trans. R. Soc. B* **360**: 1905–1916.

Dieckmann, U. & Doebeli, M. 1999. On the origin of species by sympatric speciation. *Nature* **400**: 354–357.

Dobzhansky, T. 1940. Speciation as a stage in evolutionary divergence. *Am. Nat.* **74**: 312–321.

Ewing, G. & Hermisson, J. 2010. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**: 2064–2065.

Fay, J.C. & Wu, C.I. 2000. Hitchhiking under positive Darwinian selection. *Genetics* **155**: 1405–1413.

Fu, Y.X. & Li, W.H. 1993. Statistical tests of neutrality of mutations. *Genetics* **133**: 693–709.

Funk, D.J. & Omland, K.E. 2003. Species level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annu. Rev. Ecol. Evol. Syst.* **34**: 397–423.

Fuyama, Y. 1983. Species-specificity of paragonial substances as an isolating mechanism in *Drosophila*. *Experientia* **39**: 190–192.

Geyer, L.B. & Palumbi, S.R. 2003. Reproductive character displacement and the genetics of gamete recognition in tropical sea urchins. *Evolution* **57**: 1049–1060.

Gibbs, G. & O'Bryan, M. 2007. Cysteine rich secretory proteins in reproduction and venom. *Soc. Reprod. Fertil. Suppl.* **65**: 261–267.

Hart, M.W., Sunday, J.M., Popovic, I., Learning, K.J. & Konrad, C.M. 2014. Incipient speciation of sea star populations by adaptive gamete recognition coevolution. *Evolution* **68**: 1294–1305.

Heed, W.B. & Grimaldi, D.A. 1991. Revision of the morphocrytpic, Caribbean *mayaguana* species subcluster in the *Drosophila repleta* group (Diptera: Drosophilidae). *Am. Mus. Novit.* **2999**: 1–10.

Higashi, M., Takimoto, G. & Yamamura, N. 1999. Sympatric speciation by sexual selection. *Nature* **402**: 523–526.

Hudson, R.R. 1990. Gene genealogies and the coalescent process. In: *Oxford Surveys in Evolutionary Biology* (D. Futuyma & J. Antonovics, eds), pp. 1–44. University Press, Oxford.

Hudson, R.R., Boos, D.D. & Kaplan, N.L. 1992. A statistical test for detecting geographic subdivision. *Mol. Biol. Evol.* **9**: 138–151.

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.

Katoh, K., Misawa, K., Toh, H. & Miyata, T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**: 511–518.

Kelleher, E.S. & Markow, T.A. 2007. Reproductive tract interactions contribute to isolation in *Drosophila*. *Fly* **1**: 33–37.

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

Kern, A.D., Jones, C.D. & Begun, D.J. 2004. Molecular population genetics of male accessory gland proteins in the *Drosophila simulans* complex. *Genetics* **167**: 725–735.

Kliman, R.M., Andolfatto, P., Coyne, J.A., Depaulis, F., Kreitman, M., Berry, A.J. *et al.* 2000. The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics* **156**: 1913–1931.

Kosakovsky Pond, S.L. & Frost, S.D.W. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* **22**: 1208–1222.

Kosakovsky Pond, S.L., Frost, S.D. & Muse, S.V. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**: 676–679.

Kosakovsky Pond, S.L., Murrell, B., Fourment, M., Frost, S.D.W., Delport, W. & Scheffler, K. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Mol. Biol. Evol.* **28**: 3033–3043.

Li, H. 2011. A new test for detecting recent positive selection that is free from the confounding impacts of demography. *Mol. Biol. Evol.* **28**: 365–375.

Lynch, M. & Crease, T.J. 1990. The analysis of population survey data on DNA sequence variation. *Mol. Biol. Evol.* **7**: 377–394.

Maddison, D. & Maddison, W. 2000. *MacClade 4: Analysis of Phylogeny and Character Evolution*. Sinauer Associates, Sunderland, MA.

Manier, M.K., Lüpold, S., Belote, J.M., Starmer, W.T., Berben, K.S., Ala-Honkola, O. *et al.* 2013. Postcopulatory sexual selection generates speciation phenotypes in *Drosophila*. *Curr. Biol.* **23**: 1853–1862.

Matute, D.R. & Coyne, J.A. 2009. Intrinsic reproductive isolation between two sister species of *Drosophila*. *Evolution* **64**: 903–920.

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

Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K. & Kosakovsky Pond, S.L. 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* **8**: e1002764.

Nei, M. & Li, W.-H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl Acad. Sci. USA* **76**: 5269–5273.

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

Nosil, P., Harmon, L.J. & Seehausen, O. 2009. Ecological explanations for (incomplete) speciation. *Trends Ecol. Evol.* **24**: 145–156.

O'Grady, P.M., Durando, C.M., Heed, W.B., Wasserman, M., Etges, W. & DeSalle, R. 2002. Genetic divergence within the *Drosophila mayaguana* subcluster, a closely related triad of Caribbean species in the *repleta* species group. *Hereditas* **136**: 240–245.

Oliveira, D.C.S.G., Almeida, F.C., O'Grady, P., Armella, M.A., DeSalle, R. & Etges, W.J. 2012. Monophyly, divergence times, and evolution of host plant use inferred from a revised phylogeny of the *Drosophila repleta* species group. *Mol. Phylogen. Evol.* **64**: 533–544.

Orr, H.A., Masly, J.P. & Phadnis, N. 2007. Speciation in *Drosophila*: from phenotypes to molecules. *J. Hered.* **98**: 103–110.

Palumbi, S.R. 2009. Speciation and the evolution of gamete recognition genes: pattern and process. *Heredity* **102**: 66–76.

Pamilo, P. & Nei, M. 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* **5**: 568–583.

Pavlidis, P., Jensen, J.D. & Stephan, W. 2010. Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics* **185**: 907–922.

Pond, S.L.K., Posada, D., Gravenor, M.B., Woelk, C.H. & Frost, S.D. 2006. Automated phylogenetic detection of recombination using a genetic algorithm. *Mol. Biol. Evol.* **23**: 1891–1901.

Ram, K.R. & Wolfner, M.F. 2007. Seminal influences: *Drosophila* Acps and the molecular interplay between males and females during reproduction. *Integr. Comp. Biol.* **47**: 427–445.

Rand, D.M. & Kann, L.M. 1996. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. *Mol. Biol. Evol.* **13**: 735–748.

Ritchie, M.G. 2007. Sexual selection and speciation. *Annu. Rev. Ecol. Evol. Syst.* **38**: 79–102.

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

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

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690.

Stoletzki, N. & Eyre-Walker, A. 2011. Estimation of the neutrality index. *Mol. Biol. Evol.* **28**: 63–70.

Swanson, W.J. 2003. Sex peptide and the sperm effect in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **100**: 9643–9644.

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

Swanson, W.J., Clark, A.G., Waldrip-Dail, H.M., Wolfner, M.F. & Aquadro, C.F. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *Proc. Natl Acad. Sci. USA* **98**: 7375–7379.

Swofford, D.L. 2003. *PAUP*. Phylogenetic Anlysis Using Parsimony (*and Other Methods)*. Sinauer Associates, Sunderland, MA.

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

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

Ting, C.T., Tsaur, S.C. & Wu, C.I. 2000. The phylogeny of closely related species as revealed by the genealogy of a speciation gene, Odysseus. *Proc. Natl Acad. Sci. USA* **97**: 5313–5316.

Turner, T.L. & Hahn, M.W. 2010. Genomic islands of speciation or genomic islands and speciation?. *Mol. Ecol.* **19**: 848–850.

Vacquier, V.D. 1998. Evolution of gamete recognition proteins. *Science* **281**: 1995–1998.

Via, S. 2009. Natural selection in action during speciation. *Proc. Natl Acad. Sci. USA* **106**: 9939–9946.

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

Wang, R.L., Wakeley, J. & Hey, J. 1997. Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives. *Genetics* **147**: 1091–1106.

Wasserman, M. 1992. Cytological evolution of the *Drosophila repleta* species group. In: *Drosophila Inversion Polymorphism* (C.B. Krimbas & J.R. Powell, eds), pp. 455–525. CRC Press, Boca Raton, FL.

Wasserman, M. & Wasserman, F. 1992. Inversion polymorphism in island species of *Drosophila*. In: *Evolutionary Biology* (M.K. Hecht, ed.), pp. 351–381. Plenum Press, New York, NY.

Watterson, G.A. 1975. On the number of segregating sites in genetic models without recombination. *Theor. Popul. Biol.* **7**: 256–276.

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

Wu, C.I. & Ting, C.T. 2004. Genes and speciation. *Nat. Rev. Genetics* **5**: 114–122.

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

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

Supporting information

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix S1 Summary of the commands used in the Bayesian model selection and ABC analyses.

Table S1 Fly samples used. Tucson Stock, refers to the Tucson Drosophila Stock Center and AMCCAMNH refers to the Ambrose Monell Cryo Collection at the American Museum of Natural History.

Table S2 Characteristics of the loci sequenced for this study.

Table S3 Amplification primer sequences.

Table S4 Summary statistics of multiple pairwise Nei-Gojobori (NG) estimates of dN/dS between *D. parisiena* or *D. straubae* samples and *D. mayaguana* samples.

Table S5 Intraspecific variation in the whole sequence and at non-synonymous and synonymous sites of the coding region.

Figure S1 ML gene trees of Acp loci.

Figure S2 MP gene trees.

Figure S3 Network of the *COI* locus, including samples of *D. mayaguana* (M), *D. parisiena* (P), and *D. straubae* (S).

Figure S4 A) Density plot of the posterior probability of the selection coefficient (s) as obtained in the ABC analysis of the *Acp7* locus using different tolerance levels (see methods). B) Same as A, but including neutral simulations (s = 0) to illustrate rejection of neutrality (tolerance = 0.001).

Figure S5 Sliding window analysis of polymorphism (π), population divergence (Dxy), and Fay and Wu's H test in *D. straubae* and *D. parisiena* for the locus *mayAcp58*.

Figure S6 Alignment of *D. straubae* and *D. parisiena* sequences of the *mayAcp58* locus.

Received 23 June 2016; revised 14 November 2016; accepted 15 November 2016