Recent and Recurrent Selective Sweeps of the Antiviral RNAi Gene Argonaute-2 in Three Species of Drosophila

Darren J. Obbard,*1,2 Francis M. Jiggins,3 Nicholas J. Bradshaw,4 and Tom J. Little1,2

1Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK
2Centre for Immunity, Infection and Evolution, University of Edinburgh, Edinburgh, UK
3Department of Genetics, University of Cambridge, Downing Street, Cambridge, UK
4Molecular Medicine Centre, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, UK

*Corresponding author: E-mail: darren.obbard@ed.ac.uk.

Associate editor: John H. McDonald

All sequences have been submitted to GenBank as aligned sets with accession numbers HQ262575–HQ262947, DQ228766–DQ228773, and DQ228784–DQ228791.

Abstract

Antagonistic host–parasite interactions can drive rapid adaptive evolution in genes of the immune system, and such arms races may be an important force shaping polymorphism in the genome. The RNA interference pathway gene Argonaute-2 (AGO2) is a key component of antiviral defense in Drosophila, and we have previously shown that genes in this pathway experience unusually high rates of adaptive substitution. Here we study patterns of genetic variation in a 100-kbp region around AGO2 in three different species of Drosophila. Our data suggest that recent independent selective sweeps in AGO2 have reduced genetic variation across a region of more than 50 kbp in Drosophila melanogaster, D. simulans, and D. yakuba, and we estimate that selection has fixed adaptive substitutions in this gene every 30–100 thousand years. The strongest signal of recent selection is evident in D. simulans, where we estimate that the most recent selective sweep involved an allele with a selective advantage of the order of 0.5–1% and occurred roughly 13–60 Kya. To evaluate the potential consequences of the recent substitutions on the structure and function of AGO2, we used fold-recognition and homology-based modeling to derive a structural model for the Argonaute-2 protein, and this suggests that recent substitutions in D. simulans are overrepresented at the protein surface. In summary, our results show that selection by parasites can consistently target the same genes in multiple species, resulting in areas of the genome that have markedly reduced genetic diversity.

Key words: RNAi, virus, arms race, selective sweep, AGO2, Drosophila.

Background

When natural selection replaces one allele with another, the hitchhiking of nearby variants through the population leaves a characteristic footprint in genetic diversity (Maynard Smith and Haigh 1974; Braverman et al. 1995). Such ‘selective sweeps’ will significantly shape genomic variation, with the impact depending on the selective advantage of new mutations, and the frequency with which they arise (e.g., Maynard Smith and Haigh 1974; Durrett and Schweinsberg 2004; Pennings and Hermisson 2006; Hermisson and Pfaffelhuber 2008). Many such sweeps have been identified in nature and have contributed to our understanding of both the process and targets of recent selection (e.g., Nielsen et al. 2007). In particular, studies in Drosophila have provided a plethora of examples, including genes thought to be involved in adaptation to new or anthropogenic environments (Schlenke and Begun 2004; Pool et al. 2006) and genes likely to be engaged in intragenomic conflict, such as male–female conflict or meiotic drive (Nurminsky et al. 1998; Derome et al. 2004; Holloway and Begun 2004; Presgraves et al. 2009).

These findings reflect the broader observation that conflict within and between genomes may be an important driver of adaptive molecular evolution (Begun et al. 2007; Haerty et al. 2007; Nielsen et al. 2007; Obbard, Welch, et al. 2009; Singh et al. 2009; Slote et al. 2010). The conflict that occurs between host and parasite—requiring continual innovation on both sides as hosts evolve to resist their parasites and parasites evolve to evade host resistance—is thought to be of particular importance (e.g., Hurst and Smith 1999; Woolhouse et al. 2002; Schlenke and Begun 2003) and recent studies using Drosophila have confirmed this at the genome-wide scale, both by analyzing rates of nonsynonymous substitution across the Drosophila phylogeny and by inferring patterns of adaptive substitution for different components of the Drosophila immune system (Sackton et al. 2007; Obbard, Welch et al. 2009).

A key component of innate immunity in plants, fungi, and invertebrates is antiviral RNA interference (RNAi)
Obbard et al. shallow within-species genealogy associated with recent selective sweeps in (MEGA v. 3.1, and the total length of amplified sequence per individual was approximately 13.5 kbp. Gene trees were constructed using neighbor joining that the amplified fragment from clade is

\[1044\]

mentary table S1, Supplementary Material

We analyzed polymorphism data at Choice of Loci

Materials and Methods

Choice of Loci

We analyzed polymorphism data at AGO2 and eight flanking loci, spread over a approximately 60-kbp region either side of AGO2 in three species of Drosophila (see supplementary table S1, Supplementary Material online). For clarity, we refer to loci by the name of the D. melanogaster ortholog throughout the article. These loci were chosen based on position, with more closely spaced markers being chosen near the putative site of selection where the gradient in diversity is expected to be steepest. In D. melanogaster, this region has an intermediate recombination rate (ca. 1.9cM Mbp⁻¹). Known or hypothetical coding sequences were chosen so that synonymous and short-intron sites could be used for analysis as these are least likely to be under appreciable selection in Drosophila (e.g., Halligan and Keightley 2006). Following initial analysis of this region, 35 short haplotypes (1,325 bp) were additionally obtained from the center of AGO2 in D. simulans.

In D. melanogaster, the wider region spans the range 3L:15,492,244–15,615,246 (genome release r5.26), whereas in D. simulans, it spans 3L:14,833,076–14,954,658 (release r1.3). Relative spacing was taken from D. simulans genome release r1.3; however, incomplete assembly of the D. simulans genome (e.g., several hundred bases of AGO2 appear in unplaced fragment chrU_M_6024) means that absolute locus positions can only be treated as approximate. In D. yakuba, this region falls within synteny block 42 (as defined by Ranz et al. 2007) and loci span the range 3L:18,007,466–18,163,218. However, sequences nearly identical to some of these loci additionally appear in unplaced fragments, and absolute positions should again be treated as provisional.

Origin of Accessions

For the analysis of AGO2 and eight surrounding loci, flies were sourced as described in Obbard et al. (2006) and Jiggins and Kim (2007). Briefly, 21 East African D. simulans haplotypes (Nairobi, Kenya as described in Dean and Ballard 2004) were obtained either from lines that been inbred by sib mating for six to nine generations (a subset of these AGO2 sequences, but not other loci, are reported in

Fig. 1. Genomic positions and gene trees for loci surrounding AGO2. In each tree, the upper clade is Drosophila melanogaster and the lower clade is D. simulans. The size and position of amplified regions are shown by white boxes, and gray boxes show the corresponding genes (note that the amplified fragment from Yellow-k partially overlaps locus CG7945). The total length of the surveyed region was approximately 123 kbp, and the total length of amplified sequence per individual was approximately 13.5 kbp. Gene trees were constructed using neighbor joining (MEGA v. 3.1, Kumar et al. 2004), based on coding sites only and were rooted using D. yakuba. All trees are drawn to the same scale. The shallow within-species genealogy associated with recent selective sweeps in AGO2 is clear in both species.

(Ding and Voinnet 2007; Obbard, Gordon, et al. 2009). This defence mechanism is particularly well studied in Drosophila (Galiana-Arnoux et al. 2006; van Rij et al. 2006; Sabin et al. 2009; Saleh et al. 2009), where it evolves under unusually strong selective pressure: we have previously found that the antiviral RNAi genes Argonaute-2 (AGO2), Dicer-2, and R2D2 each show elevated rates of adaptive evolution over the long term in both Drosophila melanogaster and D. simulans (Obbard et al. 2006; Obbard, Welch, et al. 2009). Because many positive-sense RNA viruses express viral suppressors of RNAi (VSRs) that block antiviral RNAi, it has been hypothesized that this rapid adaptive evolution in RNAi genes is likely to be due to a molecular arms race with VSRs (Obbard et al. 2006; Marques and Carthew 2007; van Rij and Andino 2008; Obbard, Goldon, et al. 2009).

Here we examine the wider impact on the Drosophila genome of selective sweeps in one of these antiviral RNAi genes (AGO2) and discuss the evidence that this gene has experienced recurrent and recent selection in multiple species. To do this, we surveyed DNA sequence variation in natural populations of D. melanogaster, D. simulans, and D. yakuba across a 120-kbp region around AGO2 (fig. 1). Multiple lines of evidence suggest a recent selective fixation in, or near, AGO2 in all three species, and using these data, we investigate the timing of selective sweeps and the strength of selection associated with the most recently fixed alleles in D. simulans.
Obbard et al. 2006) or by employing simulans-specific primers on artificial melanogaster–simulans interspecies hybrids. Twelve West African haplotypes were obtained from D. melanogaster (collected by B. Ballard and S. Charlat in Franceville, Gabon in 2002) using third chromosomes previously made isogenic by standard crosses to the balancer stock TM6/Sb. Seven to eleven D. yakuba (Gabon) sequences were obtained for each locus from isofemale lines that had been inbred by sib mating for six to nine generations. In addition, 35 shorter AGO2 haplotypes were also obtained from partially inbred North American and Madagascan D. simulans lines provided by P. Andolfatto and P. Haddrill. The Madagascan accessions are described in Dean and Ballard (2004) and North American accessions were collected in California (by A. Clark in 1999).

Drosophila erecta sequences (used to provide an outgroup for D. yakuba) were derived from published D. erecta genome sequence (r1.3). The D. sechellia AGO2 sequence (used to infer putative sites of recent selection in D. simulans AGO2) was derived from genome shotgun sequence chromatograms deposited in the NCBI Trace archive (Clark et al. 2007), complemented by new sequencing from a D. sechellia line provided by the Drosophila species stock center (University of California San Diego).

PCR and DNA Sequencing
Polymerase chain reaction (PCR) primers for AGO2 were designed from the published genome sequences of D. melanogaster, D. simulans, and D. yakuba. The 5' end of AGO2 was not sequenced as glutamine-rich repeat regions make alignment ambiguous and sequencing problematic. “Universal” PCR primers for the flanking loci were designed using consensus sequences from all three species. PCR failed for locus CG12031 in D. yakuba, and this region was not sequenced in this species. Where it was necessary to obtain D. simulans sequences from artificial simulans–melanogaster hybrids, a single simulans-specific PCR primer was paired with a universal primer for each locus. All primer sequences and reaction conditions are available from the authors on request. After PCR, unincorporated primers and dNTPs were removed using exonuclease I and shrimp alkaline phosphatase, and the products were then sequenced in both directions using BigDye v3.1 (Applied Biosystems) and using a ABI capillary sequencer (Gene Pool facility, University of Edinburgh). The sequence chromatograms were inspected by eye to confirm the validity of all variants within and between species and assembled using SeqManII (DNAsstar Inc., Madison).

Tests for Selection
Unless otherwise specified, we used DNAsp (Rozas et al. 2003) to calculate summary statistics. Pairwise Hudson–Kreitman–Aguadé (HKA) tests (Hudson et al. 1987) between AGO2 and each neighboring locus were performed on synonymous sites only, using the program “HKA” (Hey 2004). Significance was assessed by coalescent simulation (as implemented in HKA), making the conservative assumptions that loci were unlinked and that no recombination occurred within loci. We ran 10,000 replicates for each test. In addition, we tested for a departure from neutrality in AGO2 as compared with all neighboring loci using the likelihood ratio test by Wright and Charlesworth (2004), based on the HKA approach. Starting parameters were taken from the standard HKA tests, and the Markov chain was run for 500,000 iterations. Each run was repeated three times to confirm convergence. Haplo-type-based tests (K: number of haplotypes, M: frequency of the commonest haplotype, H_{S}, haplotype diversity, and the haplotype configuration) were performed using coalescent simulations implemented in “haploconfig” (Innan et al. 2005), conditional on estimated \( \theta \) (Watterson 1975). To allow for some uncertainty in estimations of the local recombination rate, we followed Innan et al. (2005) in using a uniform distribution of recombination rates spanning the range \( \pm 50\% \) either side of the map-based estimate for this region in place of a single recombination rate estimate. Where we wished to separate the effect of selection on the D. melanogaster and D. simulans lineages, we reconstructed hypothetical ancestral sequences using a maximum-likelihood codon-based approach (PAML; Yang 2007) and D. yakuba and D. erecta as outgroups. Tajima’s D statistic (Tajima 1989) was calculated for synonymous and silent sites in the short central region of D. simulans AGO2 using DNAsp (Rozas et al. 2003).

Selective Sweep Models in D. melanogaster and D. simulans
We used three different approaches to model the selective sweep process. First, we applied the method of Kim and Stephan (2002) to AGO2 and the eight flanking loci, using the program CLSW (http://yuseobkim.net/Programs/KimCLA0906/). This approach calculates the composite likelihood ratio (CLR) for a single sweep model versus a standard neutral model, based on the full site frequency spectrum and assuming that sites are independent. We inferred the ancestral or derived status of variants by maximum likelihood (PAML; Yang 2007) and applied test LR1, which uses the unfolded frequency spectrum. Statistical significance was inferred by comparing the CLR to an empirical null distribution derived from 1,000 standard neutral coalescent simulations, implemented in “ms” (Hudson 2002). To reduce the time needed to run the neutral simulations, the number of potentially recombining sites was taken to be one-tenth of the sequence length, rather than the number of bases in the sequence.

To mitigate the potential impact of segregating deleterious variants that will skew allele frequencies, we limited our analysis to short introns and four-fold degenerate positions, both which should more closely approximate neutrality (e.g., Halligan and Keightley 2006). This was done by simulating the whole 120-kbp region as if all sites were neutral and later retaining only those variants that fall in positions which correspond to our four-fold and short-intron sites. For D. melanogaster, we ran simulations assuming that \( \theta = 0.008 \) (the average for the loci analyzed here), the recombination rate \( r = 1.92 \) cM Mbp\(^{-1}\) (updated
for locations in genome release 5 from Singh, Arndt, and Petrov 2005, and the mutation rate $\mu = 3.5 \times 10^{-9}$ bp$^{-1}$ generation$^{-1}$ (Keightley et al. 2009). For D. simulans, we used $\theta = 0.0237$ (the average for the loci analyzed here), $r = 4.2$ cM Mbp$^{-1}$ (third chromosome rate, taken from Wall et al. 2002), and assumed that the mutation rate was identical to that of D. melanogaster. Because some demographic processes can lead to a high rate of false positives (Jensen et al. 2005), we also compared CLR statistics to null distributions generated under various demographic scenarios simulated using "ms" (Hudson 2002). Rather than simultaneously infer the demographic parameters from our data, or using a demographic model inferred from a different data set (e.g., Li and Stephan 2006), we chose to simulate six simple population growth models which cover the range of likely scenarios for African populations (Haddrill et al. 2005; Li and Stephan 2006): 10-fold and 2.5-fold step-change increases in population size, each occurring 10, 50, and 100 kya (assuming 10 generations per year). We did not include a bottleneck scenario, for which there is little evidence in African populations (Haddrill et al. 2005; Li and Stephan 2006).

Second, we applied the maximum likelihood method of Li and Stephan (test L1 in Li and Stephan 2005) to both synonymous and nonsynonymous sites in AGO2 and eight flanking loci using the program MOSY (http://www.zi.biolologie.uni-muenchen.de/~li/mosy/). This method differs from the approach of Kim and Stephan (2002) in that it conditions on the expected branch lengths of the genealogy and uses data from the compact site frequency spectrum that considers only the frequencies 1, 2, and $\geq 3$, in place of the full-site frequency spectrum. Significance was inferred by comparison of the likelihood ratio with an empirical null distribution derived from constant-size neutral coalescent simulations performed with MOSY. For the neutral simulations, $\theta$ was taken to be the average of the loci analyzed here ($\theta = 0.0038$ and 0.0135 for D. melanogaster and D. simulans, respectively) and $N_e$ was estimated from previously published synonymous site diversity in these populations (Obbard, Welch, et al. 2009) and the neutral mutation rate in D. melanogaster (Keightley et al. 2009) (estimated $N_e = 1.9 \times 10^6$ for D. simulans and $1.3 \times 10^6$ for D. melanogaster). The time since the sweep occurred ($t$), the strength of selection ($s$), and the location of the site under selection were set as free parameters to be estimated. To reduce the time taken to generate null distributions, only 100 (D. simulans) and 200 (D. melanogaster) simulation replicates were performed.

Third, to improve our estimate of the timing and geographic scope of the sweep in D. simulans, we applied the Bayesian rejection sampling algorithm of Przeworski (2003) to estimate $s$, the selective coefficient, and $T$, the time since fixation of the selected allele (in units of $4N_e$ generations). This approach aims to sample the posterior distribution of $T$ and $s$ conditional on three summary statistics (the number of segregating sites, the number of haplotypes, and Tajima’s $D$) by simulating a sweep model under random draws from the parameter priors and retaining those draws in which the simulated summary statistics deviate from the observed statistics by less than a specified threshold. Based on the analyses above, the selected site was taken to be immediately adjacent to the sequenced region (parameter $K = 1$), and priors were gamma distributed with mean mutation rate $\mu = 3.5 \times 10^{-9}$, recombination rate $r = 4.2 \times 10^{-8}$, and $N_e = 1.9 \times 10^6$ (as used in MOSY, above). $T$ was sampled from a uniform prior over the interval $[0,1]$ and $s$ from a uniform prior over the interval $[50/N_e,0.05]$. The acceptance threshold was set to 0.1, and 500 samples from the posterior were used to estimate parameters $T$ and $s$; all other options were set to the defaults. This analysis was performed on 56 short (1,325 bp, including 288 bp of intronic sequence) D. simulans haplotypes in addition to the sequences used in the analyses above. The additional sequences comprised 11 accessions from California, which is thought to have been colonized relatively recently by D. simulans (Capy and Gibert 2004), and 24 accessions from Madagascar, which is argued to be its ancestral range (Dean and Ballard 2004).

Timer the Most Recent Selective Sweep in D. simulans

In addition to estimating the time since fixation under a sweep model, we also used BEAST (Drummond and Rambaut 2007) to estimate the age of common ancestry for the 56 D. simulans AGO2 haplotypes under a simple gene-tree model. This was done for all sites, and also for a subset of the data comprising only short-intron and four-fold degenerate positions, as these should more closely conform to the neutral substitution rate. We assumed a strict molecular clock with the D. melanogaster mutation rate of $3.5 \times 10^{-8}$ bp$^{-1}$ yr$^{-1}$ (i.e., 10 generations per year) and that no recombination had occurred within this region since the common ancestor of the sequences. This corresponds to a model in which all the mutations seen have arisen since the sweep started. If some of the segregating variants actually predate the sweep and are present because of recombination into the selected background during the sweep process, or if recombination has occurred between variants since the sweep completed, then this approach will tend to overestimate the time that has elapsed. We used an HKY substitution model with no rate heterogeneity between sites, and default priors were used for all parameters except tree shape, which followed a Yule process. We ran the Markov chain for $10^8$ steps, recording a total of 1,000 states, of which we discarded the first 10% as burn-in. Traces suggested the chain mixed well and had reached stationarity. The effective sample size for each parameter estimate was $\geq 500$.

McDonald–Kreitman Tests

The proportion (or number) of nonsynonymous substitutions attributable to positive selection rather than genetic drift can be estimated from counts of polymorphisms and substitutions at synonymous and nonsynonymous sites (reviewed in Eyre-Walker 2006). This forms the basis of the McDonald–Kreitman (MK) test (McDonald and
Structural Modeling of AGO2

To better understand the nature of recent selection on AGO2, we used fold-recognition and homology modeling to build a three-dimensional structural model of the AGO2 protein and pinpoint recent amino acid substitutions within the structure of this protein. We excluded the extreme 5’ end of the gene as the length-variable glutamine-rich repeats cannot be aligned between species. Residue positions are given relative to the start of our alignment (residue 431 of FlyBase CG7439-PB). A fold-recognition search (Bennett-Lovsey et al. 2008) identified the closest structural homolog to full-length *D. simulans* AGO2 as the Argonauta protein from *Pyrococcus furiosus* (PHYRE E value = 3.15 × 10^{-15}, estimated precision = 100%). Its highest resolution structure (PDB ID: 1U04, Song et al. 2004) was used as a template, along with part of a related structure (PDB ID: 1Z26) and the experimentally determined PAZ domain from *D. melanogaster* AGO2 (PDB ID: 1R62, Song et al. 2003). Target-template alignment was based on a multiple sequence alignment of individual domains of the *Drosophila* proteins (N-terminal/stalk, PAZ, anchor, mid, and PIWI), using the program PROMALS3D (Pei et al. 2008) to improve indel positioning. After generating initial models of individual domains using MODELLER 9.7 (Sali and Blundell 1993), alignments were subjected to further manual editing based on predicted (PSIPRED 2.4 McGuffin et al. 2000) and known (STRIDE Frishman and Argos 1995) secondary structure. Some strongly predicted secondary structure elements absent in the template were restrained during model building. Four 14–28 residue segments lacking template-guided restraints were modeled on other protein segments (from PDB IDs: 2GJU, 2WZI, 313L, and 3KZ1) with primary and secondary structures similar to those predicted for *Drosophila* AGO2. The extreme C-terminus was not modeled due to lack of an appropriate template. Twenty models of *D. simulans* AGO2 were generated. From the five with the lowest objective function score (Sali and Blundell 1993), the best was selected based on feasible domain–domain interaction and orientation within the intact overall structure, valid stereochemistry using a Ramachandran plot (selected representative model dihedral angle statistics: favored 86.6%; allowed 8.8%; outliers 4.6%; Morris et al. 1992), coarse packing quality (WHATIF average quality control score: −1.508, Vriend and Sander 1993), and the model validation program, ProQ (LGscore: 2.659, MaxSub: 0.209, Wallner and Elofsson 2003). Although actual position and orientation of side chains might be tentative given the low target–template sequence identity, the derived model should be suitable for inferring residues that are surface exposed or buried, as determined using GETAREA (Fraczkiewicz and Braun 1998). We also confirmed that the mid-domain of our model was similar to one derived from the recently published human AGO2 Mid-domain (Frank et al. 2010) (supplementary fig. S1, Supplementary Material online).

Results

Diversity Is Reduced around AGO2

In all three species (*D. melanogaster*, *D. simulans*, and *D. yakuba*), we found a substantial reduction in genetic diversity surrounding the antiviral gene AGO2 (*figs* 1 and 2 and supplementary table S2, Supplementary Material online). This effect is reflected in the shallow within-species gene trees seen for AGO2 and the closest neighboring loci in both *D. melanogaster* and *D. simulans* (*fig. 1*). In each case, there is a ∼100-kbp valley of reduced genetic diversity centered on AGO2, and in all three species, AGO2 has a lower genetic diversity than any of the other loci analyzed (*fig. 2*). Across the nine loci analyzed in *D. melanogaster*, Watterson’s (1975) θ at synonymous sites ranged from 0.004 at AGO2 to 0.025 at the edge of the sampled region; in *D. simulans* from 0.003 at AGO2 up to 0.069; and in *D. yakuba* from 0.006 up to 0.054 (*fig. 2* and supplementary table S2 in supporting information, Supplementary Material online). The genetic variation observed at AGO2 is substantially less than is typical for other genes in the genome. In previous studies, the mean θ_{syn} was 0.018 and 0.038 in the same populations of *D. melanogaster* and *D. simulans*, respectively (chromosome 3L, data from Obbard, Welch, et al. 2009) and 0.030 in *D. yakuba* (autosomal loci with intermediate recombination, Llopart et al. 2005).

This reduction in genetic variation is suggestive of recent selective sweeps of new advantageous AGO2 alleles but might also be due to differences in mutation rate. To test whether the reduction in diversity is significant, we used the HKA test, which identifies differences in diversity between loci given the divergence between species for those loci, thereby correcting for mutation rate (Hudson et al. 1987). In *D. melanogaster*, HKA tests found a significant reduction in synonymous site diversity at AGO2 relative to each of the three loci in the 5’ flank of AGO2 and one
on the 3’ flank (fig. 2 and supplementary table S3, Supplementary Material online). In *D. simulans*, all the flanking loci had significantly higher diversity than AGO2. In *D. yakuba*, the diversity of the two loci immediately neighboring AGO2 was not significantly higher than that of AGO2, but the remaining six loci were significantly higher. An HKA-based maximum likelihood approach to identify differences in diversity (Wright and Charlesworth 2004) also found that AGO2 had significantly reduced diversity in all three species (table 1). These tests are highly conservative as they use neighboring loci as the neutral standard, despite the fact that they too appear to have been affected by the sweep (fig. 2).

**AGO2 and Neighboring Loci Display Unusual Haplotype Structure**

Following a selective sweep, linked alleles from neighboring sites spread together, increasing the extent to which polymorphisms from different loci co-occur within individuals (i.e., linkage disequilibrium) and reducing the number of haplotypes (e.g., Kim and Nielsen 2004; Innan et al. 2005). We tested *D. simulans* and *D. melanogaster* for deviations in haplotype structure from the standard neutral model but did not include *D. yakuba* in this analysis because of our small sample size (*n* = 7) and a lack of information on the local recombination rate. In *D. simulans*, four different tests detected unusual haplotype structure at AGO2 and a neighboring locus (table 2). Both loci had significantly fewer haplotypes than expected under a neutral model (Innan et al. 2005) (K, table 2) and correspondingly lower haplotype diversity *Hd*. In addition, the frequency of the most common haplotype (*M*) was significantly higher than expected and the haplotype configuration contained too many high-frequency haplotypes to be compatible with a standard neutral model (Innan et al. 2005). Within the *D. melanogaster* data set, AGO2 deviated from a neutral model only under the haplotype configuration test (Innan et al. 2005), displaying too many intermediate-frequency haplotypes (table 2). Three out of four tests additionally identified a departure for a third locus in *D. simulans* (table 2).

**Model Fitting Identifies a Recent Selective Sweep in *D. simulans***

To obtain estimates of the strength of selection, the timing of the sweeps, and the location of the site under selection in *D. simulans* and *D. melanogaster*, we applied the CLR method of Kim and Stephan (2002), which uses information from genetic diversity and the site-frequency spectrum. For *D. melanogaster*, this method found no evidence for a recent selective sweep (fig. 3). This was true under both the constant-size population model (CLR = 5.0

![Fig. 2. Genetic diversity around AGO2. Genetic diversity at all sites (upper row) and synonymous sites (middle row) is considerably reduced around AGO2 (positioned at zero on the x axis) in all three species. This is also reflected in the diversity/divergence ratio at synonymous sites (lower row: synonymous site diversity within species, *θ* syn, divided by divergence between species, *K* s). Loci that are significantly different from AGO2 in individual pairwise HKA tests (Hudson et al. 1987) are marked on the diversity/divergence graphs with asterisks: *P* < 0.05, **P** < 0.01, ***P*** < 0.001.](image-url)

### Table 1. HKA Likelihood Ratio Tests.

<table>
<thead>
<tr>
<th></th>
<th>2ΔlnL</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Drosophila melanogaster</strong> (vs. ancestral sequence)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No selection</td>
<td>1</td>
<td>−49.2</td>
</tr>
<tr>
<td>Selection on AGO2</td>
<td>0.22</td>
<td>−46.5</td>
</tr>
<tr>
<td>D. simulans (vs. ancestral sequence)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No selection</td>
<td>1</td>
<td>−62.6</td>
</tr>
<tr>
<td>Selection on AGO2</td>
<td>0.08</td>
<td>−56.7</td>
</tr>
<tr>
<td>D. yakuba (vs. D. erecta)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No selection</td>
<td>1.00</td>
<td>−53.3</td>
</tr>
<tr>
<td>Selection on AGO2</td>
<td>0.22</td>
<td>−50.6</td>
</tr>
</tbody>
</table>

**Note**—*k* is the estimated reduction in diversity due to selection on AGO2 (Wright and Charlesworth 2004). *ln* is the log-likelihood of the model, and 2ΔlnL is the log-likelihood test statistic.
Recent Selection on Argonaute-2 in Drosophila · doi:10.1093/molbev/msq280

Table 2. Haplotype Configuration Tests.

<table>
<thead>
<tr>
<th></th>
<th>K (95%)</th>
<th>M (95%)</th>
<th>H_d (Haplotype configuration)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Drosophila melanogaster</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG7275</td>
<td>12 (7, 12)</td>
<td>1 (1, 5)</td>
<td>0.917 (12, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)</td>
<td>ns</td>
</tr>
<tr>
<td>yellow-k</td>
<td>11 (7, 12)</td>
<td>2 (1, 5)</td>
<td>0.903 (10, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)</td>
<td>ns</td>
</tr>
<tr>
<td>CrebA</td>
<td>11 (6, 12)</td>
<td>2 (1, 5)</td>
<td>0.903 (10, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)</td>
<td>ns</td>
</tr>
<tr>
<td>AGO2</td>
<td>7 (7, 12)</td>
<td>3 (1, 4)</td>
<td>0.820 (4, 1, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)</td>
<td>0.025</td>
</tr>
<tr>
<td>CG7739</td>
<td>9 (5, 10)</td>
<td>2 (2, 6)</td>
<td>0.875 (6, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)</td>
<td>ns</td>
</tr>
<tr>
<td>CG6498</td>
<td>7 (6, 11)</td>
<td>5 (2, 5)</td>
<td>0.764 (5, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)</td>
<td>ns</td>
</tr>
<tr>
<td>CG12301</td>
<td>9 (6, 11)</td>
<td>2 (2, 5)</td>
<td>0.875 (6, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)</td>
<td>ns</td>
</tr>
<tr>
<td>RhoGAP71E</td>
<td>8 (4, 9)</td>
<td>3 (2, 7)</td>
<td>0.847 (5, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)</td>
<td>ns</td>
</tr>
<tr>
<td>CG7372</td>
<td>9 (6, 12)</td>
<td>3 (1, 5)</td>
<td>0.861 (7, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)</td>
<td>ns</td>
</tr>
<tr>
<td><strong>D. simulans</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG7275</td>
<td>14 (14, 21)</td>
<td>4 (1, 4)</td>
<td>0.903 (10, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)</td>
<td>ns</td>
</tr>
<tr>
<td>yellow-k</td>
<td>16 (14, 20)</td>
<td>3 (2, 4)</td>
<td>0.921 (13, 1, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)</td>
<td>ns</td>
</tr>
<tr>
<td>CrebA</td>
<td>20 (13, 20)</td>
<td>2 (2, 5)</td>
<td>0.948 (19, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)</td>
<td>ns</td>
</tr>
<tr>
<td>AGO2</td>
<td>8* (10, 18)</td>
<td>13** (2, 8)</td>
<td>0.594** (6, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>CG7739</td>
<td>6** (9, 17)</td>
<td>14** (1, 8)</td>
<td>0.530** (3, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>CG6498</td>
<td>18 (12, 20)</td>
<td>3 (2, 5)</td>
<td>0.934 (16, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)</td>
<td>ns</td>
</tr>
<tr>
<td>CG12301</td>
<td>14 (14, 21)</td>
<td>6** (1, 4)</td>
<td>0.875* (11, 2, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)</td>
<td>0.011</td>
</tr>
<tr>
<td>RhoGAP71E</td>
<td>17 (12, 19)</td>
<td>4 (2, 6)</td>
<td>0.921 (15, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)</td>
<td>ns</td>
</tr>
<tr>
<td>CG7372</td>
<td>18 (16, 21)</td>
<td>2 (1, 4)</td>
<td>0.939 (15, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)</td>
<td>ns</td>
</tr>
</tbody>
</table>

Note.—K (95%) is the number of haplotypes (95% bounds under neutrality from simulation); M (95%) the frequency of the most common haplotype (95% bounds under neutrality from simulation), and H_d the haplotype diversity (as in Depaulis and Veuille 1998). Asterisks denote significance. The haplotype configuration is a vector that records the frequency of haplotypes occurring n times in the sample where n = (1,2,3,...,x) and x is the sample size (Innan et al. 2005). Note that the fragment from Yellow-k partly overlaps locus CG7945.

vs., nominal 5% significance threshold of CLR = 21) and all the population-growth models (observed CLR = 5.0 vs. significance thresholds of CLR = 21–25 for the six population-growth models). For D. simulans, there was evidence for a recent selective sweep (fig. 3) under both the constant-size model (CLR = 36.7 vs. 5% significance threshold of CLR = 10) and the population-growth models (CLR = 36.7 vs. significance thresholds of CLR = 10–12 across the six growth models) (see Materials and Methods for details). This approach identified position 59041 of the analyzed region, which is within the coding sequence of AGO2 (approximately genomic position 3L:14,892,118) as the most likely focal site of the sweep and the strength of selection (2N_e s) as 21482, implying a selective advantage of roughly 0.8%. The detection of a sweep in D. simulans but not in D. melanogaster could in principle result from the lower power in D. melanogaster (n = 12) compared with D. simulans (n = 21). However, an identical analysis of this region in the Drosophila Population Genomics Project data set (n = 37, derived from North American

![Figure 3](https://example.com/figure3.png)

**Fig. 3.** Composite likelihood profile. The CLR between a standard neutral model and selective sweep model, considering each site in turn (Li and Stephan 2005). The region surrounding AGO2 is shown for Drosophila melanogaster (upper panel) and D. simulans (lower panel). Gray regions are those for which sequence data are available, and the thin horizontal line shows the most stringent 5% significance threshold for this statistic derived a range of plausible population-growth scenarios (see Materials and Methods). The maximum likelihood estimate of the focal site for the sweep is given by a vertical dashed line; note that under this model there is no significant evidence of a recent sweep in D. melanogaster but that D. simulans shows strong evidence of a sweep in AGO2.
samples, http://www.dpgp.org/) gives an extremely similar CLR profile and similarly fails to detect a significant sweep in *D. melanogaster*, suggesting that our result is not merely due to reduced power in *D. melanogaster* (data not shown).

As a second approach to fit a selective sweep model, we applied the method of Li and Stephan (test L1 in Li and Stephan 2005). This method also found no evidence for a recent selective event in *D. melanogaster* as the difference in log-likelihood between the neutral and selective models was $\Delta \log L = 16.9$, as compared to a nominal 5% significance threshold from constant-population neutral simulation of $\Delta \log L = 36.5$. However, there was again evidence for a recent selective sweep in *D. simulans*, where $\Delta \log L = 151.3$ compared with a nominal 5% significance threshold from constant-population neutral simulation of $\Delta \log L = 93.6$. The inferred site of selection was position 60787 bp of the analyzed region, which falls very close to the 3’ end of AGO2, and the estimated strength of selection was $s = 0.3\%$.

Both these approaches assume a single recent sweep has occurred at an unknown location within the sequenced region, and both use information from the site frequency spectrum to draw inferences. However, recurrent sweeps and soft sweeps (from standing variation) alter the expected site frequency spectrum (e.g., Kim 2006; Pennings and Hermisson 2006), and it is unclear how well single-sweep analyses perform if recurrent selection means that patterns of genetic diversity had not reached equilibrium prior to the most recent sweep. The unknown outcome of erroneously applying single-sweep models means that these tests should be treated with some caution. However, if this effect does result in reduced power, this may explain why they failed to detect selection on AGO2 in *D. melanogaster*, despite other analyses being consistent with recent selection.

**Date and Scale of the Selective Sweep in *D. simulans***

The maximum likelihood method of Li and Stephan (2005) dated the sweep in *D. simulans* to 0.05 $\times 4N_e$ generations ago. This corresponds to approximately 38 kya, assuming 10 generations per year and $N_e \sim 1.9 \times 10^6$, and suggests that the sweep occurred much more recently than *D. simulans*’ common ancestry with *D. sechellia* or *D. mauritiana* (~250 kya, see McDermott and Kliman 2008) and possibly before *D. simulans*’ expansion out of Africa into Europe, which is itself thought to be more recent than the spread of *D. melanogaster* (Capy and Gibert 2004) that occurred ~10–16 kya (Stephan and Li 2007).

To improve estimates of the timing and geographic scope of the sweep, we sequenced a short central region (1.3 kbp) adjacent to the putative site of selection from 11 Californian and 24 Madagascan accessions, in addition to those sampled from Kenya. Across the fifty-six 1325 bp haplotypes, we found only 16 silent- and synonymous-site polymorphisms ($\pi_s = 0.0035$) and 6 nonsynonymous polymorphisms ($\pi_a = 0.0006$), and Tajima’s $D$ statistic (silent and synonymous sites only) was $D = -1.96$ ($P < 0.008$, assessed by coalescent simulation assuming no recombination). All three populations had individually low diversity ($\pi_{silent} = 0.0018, 0.0019, 0.0054$, respectively, for Kenya, California, and Madagascar), and although there was some genetic divergence between populations, it was extremely low and there were no fixed differences: Hudson’s (2000) $S_{nn} = 0.49$ and $K_{st} = 0.07$, $P < 0.01$ for both. These observations suggest that the sweep is likely to have affected the whole extant *D. simulans* population.

Using these 56 *D. simulans* partial AGO2 sequences, we also inferred the approximate date of the sweep using the rejection-sampling algorithm of Przeworski (2003) and a tree-based approach implemented in BEAST (Drummond and Rambaut 2007). The selective sweep model (Przeworski 2003) estimated the time since fixation to be 13.5 kya (95% Highest Posterior Density (HPD) interval 2.1–45) and $s = 0.01$ (95% HPD interval: 0.0006–0.046), and the BEAST analysis estimated it to be 45 kya (95% HPD interval: 25–70) when using all sequenced sites or 57 kya (95% HPD interval: 16 to 105) when limited to putatively neutral sites. Note that the BEAST values are expected to be overestimates (see Materials and Methods).

**Sites of Recent Substitution in *D. simulans* AGO2**

Both model-fitting approaches (Kim and Stephan 2002; Li and Stephan 2005) suggest the most likely target of selection during the most recent sweep falls close to, or within, the coding sequence of AGO2; however, given these data, neither approach can infer the position with sufficient resolution to determine which site within AGO2 was substituted. Nonetheless, all analyses indicate the sweep occurred much more recently than the split between *D. simulans* and *D. sechellia* (estimated to be 250 kya), suggesting that the selected site will appear as a substitution on the *D. simulans* lineage alone. We used a codon-based phylogenetic model (PAML) to infer the ancestral *D. simulans*–*D. sechellia* sequence by maximum likelihood, and thereby identify 8 candidate nonsynonymous substitutions that have been fixed in *D. simulans* since its common ancestor with *D. sechellia*. However, there appear to be no clear patterns in the structural locations of these recent substitutions (fig. 4).

Argonaute proteins are defined by the presence of PAZ and PIWI domains, which are thought to interact with the 3’ end of single-stranded RNA and to catalyze Argonaute endonuclease activity, respectively. Two of the eight candidates for recent substitution in the *D. simulans* lineage appear in the region 5’ of the PAZ domain, two fall within the PAZ domain, three between the PAZ and PIWI domains, and one near the center of the PIWI domain. The majority of the substitutions (M106L, A185G, D194E, I238V, P404S, S428P) are conservative and do not alter residue charge or size (though P404S is located in an alpha-helix, which may be altered by the substitution). The most 5’ site (K15T) and the site near the center of the PIWI domain (K636T) are more radical substitutions in terms of residue size and charge, and the loss of charge...
Recent Selection on Argonaute-2 in Drosophila

The surface structure of Drosophila AGO2 derived from published archean and Drosophila Argonaute structures by fold-recognition and homology modeling (see Materials and Methods). Moving down the figure, the four panels are successive 90° rotations about the vertical axis. The PAZ domain is indicated in green, the PIWI domain is indicated in blue, and the amino acid substitutions that occurred in D. simulans since the split from D. sechellia (ca. 250 kya) are shown in red. The two remaining substitutions at L106 and S404 are buried within the structure (see also supplementary fig. S2B, Supplementary Material online).

Fig. 4. Recent amino acid substitutions in D. simulans AGO2. The surface structure of Drosophila AGO2 derived from published archean and Drosophila Argonaute structures by fold-recognition and homology modeling (see Materials and Methods). Moving down the figure, the four panels are successive 90° rotations about the vertical axis. The PAZ domain is indicated in green, the PIWI domain is indicated in blue, and the amino acid substitutions that occurred in D. simulans since the split from D. sechellia (ca. 250 kya) are shown in red. The two remaining substitutions at L106 and S404 are buried within the structure (see also supplementary fig. S2B, Supplementary Material online).

Long-Term Adaptive Evolution in AGO2

To explore the possibility that strong directional selection on protein-coding loci other than AGO2 may contribute to the low diversity seen in this region, we studied the action of long-term selection using an approach based on the MK test (McDonald and Kreitman 1991). Following Welch (2006, see also Obbard, Welch, et al. 2009), we estimated the number of adaptive substitutions per nonsynonymous site that have occurred in each of the sequenced loci, for each of the three lineages. For each species, the best-supported model was one in which AGO2 had a higher rate of adaptive substitution than the other genes (Akaike weights of 1.00 for D. melanogaster, 0.84 for D. simulans and 0.98 for D. yakuba; see table 3 for MK data and see supplementary table S3, Supplementary Material online, for details of model selection). The only other model to receive appreciable support was for D. simulans, in which all genes had different rates (Akaike weight 0.16). These results corroborate previous analyses, suggesting that AGO2 experiences an unusually high rate of adaptive evolution (Obbard et al. 2006; Obbard, Welch, et al. 2009) and suggest
that, in addition to having the lowest diversity of all the genes in this region (above), AGO2 has also experienced the strongest selection for change over the long term.

Maximum likelihood estimates for the total number of adaptive substitutions at nonsynonymous sites in the sequenced region of AGO2 are 48, 75, and 63 for D. melanogaster, D. simulans, and D. yakuba, respectively (measuring from the common ancestor of D. melanogaster and D. simulans) and from D. erecta. If we assume the most recent common ancestor of D. melanogaster and D. simulans lived 2.3 Ma (Russo et al. 1995), and a total of 7.2 My of evolution separate D. yakuba and D. erecta, then these numbers correspond to rates of 0.011, 0.017, and 0.005 adaptive substitutions per nonsynonymous site per million years in AGO2 or sweeps occurring roughly every 50, 30, and 110 thousand years.

Discussion

Evidence for Recurrent Selection on AGO2

We found that genetic diversity was greatly reduced around AGO2 in three species of Drosophila. There is compelling evidence that the reduced variation is caused by selection on AGO2 itself, as in all the species, the valley of low diversity was centered on AGO2, and in D. simulans, a model-fitting approach identified AGO2 as the target of selection. It is hard to envisage any process other than recurrent natural selection that could account for our results. For example, local variation in mutation rate is accounted for by reference to outgroup divergence, and gene conversion and changes in population size are unlikely to affect the same candidate gene (chosen a priori) across different species. Furthermore, sequences were derived from populations that do not appear to have experienced a strong bottleneck, which may lead to a spurious inference of selection (Haddrill et al. 2005). Finally, there is also evidence for long-term selection on AGO2 since these species shared a common ancestor as, in lineages leading to all three species, there is a significant excess of amino acid substitutions in AGO2 relative to neighboring genes, suggesting that the protein-coding sequence of AGO2 has experienced an elevated rate of adaptive substitution (table 3).

Two further independent observations from other studies also support recent selection on AGO2. First, while screening for transposable element (TE) insertions in derived D. melanogaster populations, Gonzalez et al. (2008) identified a potential selective sweep associated with an S-element insertion in the first intron of AGO2. Because the insertion allele is associated with altered expression levels of AGO2, it was argued that this may represent an adaptive substitution in AGO2 (above, and Obbard et al. 2006; Obbard, Welch, et al. 2009), it is hard to exclude the possibility that

Table 3. Interspecífic Divergence and McDonald–Kreitman Analysis.

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Ln</th>
<th>Ls</th>
<th>$K_a$</th>
<th>$K_s$</th>
<th>$K_a/K_s$</th>
<th>Ds</th>
<th>Ps</th>
<th>Dn</th>
<th>Pn</th>
<th>a</th>
<th>a/bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. melanogaster</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG7275</td>
<td>12</td>
<td>941</td>
<td>280</td>
<td>0.002</td>
<td>0.054</td>
<td>0.042</td>
<td>9</td>
<td>21</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>yellow-k</td>
<td>12</td>
<td>882</td>
<td>252</td>
<td>0.005</td>
<td>0.059</td>
<td>0.082</td>
<td>9</td>
<td>16</td>
<td>3</td>
<td>6</td>
<td>−6</td>
<td>−0.007</td>
</tr>
<tr>
<td>CrebA</td>
<td>12</td>
<td>761</td>
<td>247</td>
<td>0.003</td>
<td>0.04</td>
<td>0.079</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>3</td>
<td>−3</td>
<td>−0.004</td>
</tr>
<tr>
<td>AGO2</td>
<td>12</td>
<td>1884</td>
<td>567</td>
<td>0.032</td>
<td>0.068</td>
<td>0.476</td>
<td>34</td>
<td>54</td>
<td>4</td>
<td>48</td>
<td>0.025</td>
<td></td>
</tr>
<tr>
<td>CG7739</td>
<td>12</td>
<td>1055</td>
<td>322</td>
<td>0.009</td>
<td>0.082</td>
<td>0.11</td>
<td>23</td>
<td>7</td>
<td>9</td>
<td>3</td>
<td>4</td>
<td>0.004</td>
</tr>
<tr>
<td>CG6498</td>
<td>12</td>
<td>1074</td>
<td>336</td>
<td>0.004</td>
<td>0.038</td>
<td>0.103</td>
<td>11</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>0.001</td>
</tr>
<tr>
<td>CG12301</td>
<td>12</td>
<td>1064</td>
<td>275</td>
<td>0.022</td>
<td>0.027</td>
<td>0.819</td>
<td>7</td>
<td>5</td>
<td>21</td>
<td>8</td>
<td>8</td>
<td>0.008</td>
</tr>
<tr>
<td>RhoCAP71E</td>
<td>12</td>
<td>537</td>
<td>171</td>
<td>0.001</td>
<td>0.062</td>
<td>0.017</td>
<td>8</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>−3</td>
<td>−0.006</td>
</tr>
<tr>
<td>CG7372</td>
<td>12</td>
<td>974</td>
<td>265</td>
<td>0.017</td>
<td>0.065</td>
<td>0.263</td>
<td>15</td>
<td>5</td>
<td>14</td>
<td>11</td>
<td>−3</td>
<td>−0.003</td>
</tr>
<tr>
<td>D. simulans</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG7275</td>
<td>21</td>
<td>941</td>
<td>280</td>
<td>0.005</td>
<td>0.081</td>
<td>0.061</td>
<td>8</td>
<td>56</td>
<td>4</td>
<td>7</td>
<td>2</td>
<td>0.002</td>
</tr>
<tr>
<td>yellow-k</td>
<td>21</td>
<td>882</td>
<td>251</td>
<td>0.008</td>
<td>0.071</td>
<td>0.11</td>
<td>7</td>
<td>43</td>
<td>3</td>
<td>17</td>
<td>−3</td>
<td>−0.003</td>
</tr>
<tr>
<td>CrebA</td>
<td>21</td>
<td>761</td>
<td>247</td>
<td>0.001</td>
<td>0.031</td>
<td>0.018</td>
<td>6</td>
<td>8</td>
<td>0</td>
<td>2</td>
<td>−1</td>
<td>−0.001</td>
</tr>
<tr>
<td>AGO2</td>
<td>21</td>
<td>1889</td>
<td>565</td>
<td>0.041</td>
<td>0.067</td>
<td>0.632</td>
<td>37</td>
<td>6</td>
<td>75</td>
<td>1</td>
<td>75</td>
<td>0.040</td>
</tr>
<tr>
<td>CG7739</td>
<td>21</td>
<td>1058</td>
<td>328</td>
<td>0.011</td>
<td>0.04</td>
<td>0.27</td>
<td>11</td>
<td>10</td>
<td>11</td>
<td>3</td>
<td>10</td>
<td>0.009</td>
</tr>
<tr>
<td>CG6498</td>
<td>21</td>
<td>1072</td>
<td>338</td>
<td>0.002</td>
<td>0.033</td>
<td>0.064</td>
<td>6</td>
<td>26</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>CG12301</td>
<td>21</td>
<td>1121</td>
<td>295</td>
<td>0.009</td>
<td>0.051</td>
<td>0.179</td>
<td>10</td>
<td>36</td>
<td>4</td>
<td>33</td>
<td>−7</td>
<td>−0.006</td>
</tr>
<tr>
<td>RhoCAP71E</td>
<td>21</td>
<td>531</td>
<td>171</td>
<td>0.007</td>
<td>0.046</td>
<td>0.162</td>
<td>2</td>
<td>28</td>
<td>1</td>
<td>15</td>
<td>−4</td>
<td>−0.008</td>
</tr>
<tr>
<td>CG7372</td>
<td>21</td>
<td>948</td>
<td>258</td>
<td>0.028</td>
<td>0.077</td>
<td>0.359</td>
<td>7</td>
<td>64</td>
<td>6</td>
<td>94</td>
<td>−26</td>
<td>−0.027</td>
</tr>
<tr>
<td>D. yakuba</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG7275</td>
<td>7</td>
<td>939</td>
<td>282</td>
<td>0.014</td>
<td>0.149</td>
<td>0.091</td>
<td>29</td>
<td>37</td>
<td>11</td>
<td>3</td>
<td>4</td>
<td>0.005</td>
</tr>
<tr>
<td>yellow-k</td>
<td>8</td>
<td>880</td>
<td>254</td>
<td>0.019</td>
<td>0.166</td>
<td>0.113</td>
<td>34</td>
<td>13</td>
<td>14</td>
<td>9</td>
<td>−5</td>
<td>−0.006</td>
</tr>
<tr>
<td>CrebA</td>
<td>7</td>
<td>759</td>
<td>249</td>
<td>0.017</td>
<td>0.091</td>
<td>0.183</td>
<td>21</td>
<td>4</td>
<td>12</td>
<td>2</td>
<td>7</td>
<td>0.010</td>
</tr>
<tr>
<td>AGO2</td>
<td>11</td>
<td>1882</td>
<td>566</td>
<td>0.04</td>
<td>0.215</td>
<td>0.186</td>
<td>11</td>
<td>8</td>
<td>70</td>
<td>4</td>
<td>63</td>
<td>0.033</td>
</tr>
<tr>
<td>CG7739</td>
<td>8</td>
<td>1056</td>
<td>327</td>
<td>0.015</td>
<td>0.19</td>
<td>0.079</td>
<td>52</td>
<td>12</td>
<td>15</td>
<td>1</td>
<td>13</td>
<td>0.012</td>
</tr>
<tr>
<td>CG6498</td>
<td>8</td>
<td>1009</td>
<td>323</td>
<td>0.01</td>
<td>0.15</td>
<td>0.068</td>
<td>37</td>
<td>19</td>
<td>10</td>
<td>1</td>
<td>8</td>
<td>0.008</td>
</tr>
<tr>
<td>RhoCAP71E</td>
<td>7</td>
<td>557</td>
<td>178</td>
<td>0.015</td>
<td>0.184</td>
<td>0.079</td>
<td>26</td>
<td>12</td>
<td>8</td>
<td>0</td>
<td>8</td>
<td>0.014</td>
</tr>
<tr>
<td>CG7372</td>
<td>7</td>
<td>1034</td>
<td>283</td>
<td>0.112</td>
<td>0.245</td>
<td>0.458</td>
<td>59</td>
<td>22</td>
<td>96</td>
<td>17</td>
<td>58</td>
<td>0.056</td>
</tr>
</tbody>
</table>

Note—n is the number of alleles sampled; Ln and Ls the number of nonsynonymous and synonymous sites, respectively; $K_a$ and $K_s$ are the nonsynonymous and synonymous divergence; Ds, Ps, Dn, and Pn are counts of fixed differences (D) and polymorphisms (P) at synonymous and nonsynonymous sites; and “a” is the maximum-likelihood estimate of the number of nonsynonymous adaptive substitutions per gene under the model described in Obbard et al. (2006) and Materials and Methods (above). Divergence is measured from their common ancestor in the case of D. melanogaster and D. simulans and from D. erecta in the case of D. yakuba.
the TE-insertion hitchhiked to its presently high (but not fixed) frequency on a selected amino acid variant. Second, in a genome-wide survey of changes in expression level, Graze et al. (2009) found AGO2 expression to significantly differ between D. simulans and D. melanogaster (see also McManus et al. 2010), and changes in expression tend to be correlated with rapid evolution (Nuzhdin et al. 2004).

These results add to a growing literature showing that parasite-mediated selection is an important cause of molecular evolution. However, the extent to which this process determines variation in extant phenotype remains unclear as selection can fix alleles with selective advantages much smaller than can be measured in the laboratory. Our analyses suggest that the most recently fixed D. simulans AGO2 allele had a selective advantage of somewhat less than 1% and may therefore have had experimentally measurable phenotypic effects. However, although we estimate 90–100% of nonsynonymous substitutions in AGO2 have been driven by selection (compared with a genome average of ca. 45%, Bierne and Eyre-Walker 2004; Welch 2006; Obbard, Welch, et al. 2009), this only corresponds to a selective sweep in this gene once every 30–100 thousand years, with the most recent sweep in D. simulans occurring 13–57 thousand years ago. Thus, although this gene is one of the most strongly selected in the Drosophila immune system (Obbard, Welch, et al. 2009), such arms races are unlikely to be driving allelic replacement with a frequency likely to be observed on an ecological timescale.

Likely Selective Agents
RNA interference mediated by the Dicer2-R2D2-AGO2 pathway is a major antiviral defence mechanism in Drosophila (e.g., Galiana-Arnoux et al. 2006; van Rij et al. 2006) and other insects (e.g., Campbell et al. 2008), and several insect viruses—including Drosophila C virus—carry genes that actively suppress RNAi (Chao et al. 2005; van Rij et al. 2006; Wang et al. 2006; Nayak et al. 2010). VSRs act in several ways: by sequestering short-interfering RNAs (siRNAs), by competing with siRNAs for the active sites of RNAi pathway genes, by blocking cell-to-cell movement of siRNAs, or by destabilizing or degrading key proteins in the pathway (Diaz-Pendon and Ding 2008). The last class includes VSRs that interact directly or indirectly with Argo
ta proteases (e.g., Csorba et al. 2010) including one that interacts with D. melanogaster AGO2 (Nayak et al. 2010). Our structural model of Drosophila AGO2 suggests that recent substitutions in D. simulans have primarily occurred at the protein surface. Although this is true of many proteins, should these substitutions indeed be the result of recent selective sweeps, then these data may be indicative of AGO2 evolving to alter its ability to interact with other molecules, potentially VSRs.

Nevertheless, the Dicer2 and AGO2 not only mediate antiviral defence but also target transcripts from TEs (see Obbard and Finnegan 2008 for references), particularly during colonization by a new TE (Rozhkov et al. 2010). Like viruses, TEs are costly to their hosts, and although no TE-encoded suppressors of RNAi have been characterized, they may exist (Blumenstiel and Hartl 2005), and evolutionary conflict with TEs has the potential to drive a host–parasite molecular arms race (Lee and Langley 2010; Lu and Clark 2010). It is therefore striking that the Piwi-interacting RNAi pathway, which modulates TE transcript levels in germline tissues and is thought to target heterochromatin formation to TE insertions (Klattenhoff and Theurkauf 2008), also contains several genes which show a high rate of adaptive substitution (reviewed in Obbard, Gordon, et al. 2009). These include the heterochromatin protein Rhino, the putative exonucleases Maelstrom and Krimper, the helicases Spindle-E and Armitage, and Piwi-family Argonaute proteins Aubergine and Piwi (Vermaak et al. 2005; Heger and Ponting 2007; Obbard, Gordon, et al. 2009; Obbard, Welch, et al. 2009). Therefore, it remains possible that the recent selection on AGO2 is associated with its role in TE suppression (potentially during the invasion of new TEs, Rozhkov et al. 2010) rather than its antiviral function. Nevertheless, because we expect host–TE conflict to be limited to reproductive tissues (Charlesworth and Langley 1986), the opportunity for TEs to be the driving force in recurrent selection at AGO2 may depend on the relative importance of the Dicer2–AGO2 pathway in suppressing germline versus somatic TE expression.

Finally, the distinction between viruses and retrotransposons such as gypsy (an “endogenous retroviruses”) is a subtle one (Huszar and Imler 2008; Llorens et al. 2008), and there may be considerable mechanistic overlap in their control. For example, Drosophila lacking functional piwi-interacting (pi) RNA pathway genes Armitage, Piwi, or Aubergine appear to be compromised in their ability to resist the double-stranded RNA birnavirus DXV (Zambon et al. 2006), and piRNAs derived from viruses including DAV, DXV, DCV, and Nora have been reported from Drosophila cell culture that expresses Piwi (Wu et al. 2010). If it is confirmed that components of the piRNA pathway can play a role in antiviral defence, then viruses may be the selective force in both the piRNA and the Dicer2-R2D2-AGO2 pathways.

Conclusions
The evolution of many of some the most rapidly evolving genes in animal genomes appears to be driven by evolutionary arms races, where there is a battle of adaptation and counter-adaptation with parasites (e.g., Hurst and Smith 1999; Schlenke and Begun 2003; Sacston et al. 2007; Obbard, Welch, et al. 2009), or during sexual reproduction (Begun et al. 2007; Haerty et al. 2007; Suttle et al. 2010). In line with this, we have previously shown that the antiviral RNAi pathway of Drosophila, which is known to be targeted by viral suppressor molecules, contains three genes under exceptionally strong selection (Dcr-2, AGO2, and R2D2) (Obbard et al. 2006). Here we have shown that selection on AGO2 has left its mark across a large genomic region in three different species of Drosophila (figs 1 and 2). As far as we are aware, comparable examples—such as artificial selection and drug or pesticide resistance (e.g.,
Schlenke and Begun 2004)—tend to stem from recent human action. Thus, the finding of recent selective sweeps in AGO2 represents a particularly interesting example of how recurrent parasite-mediated selection may have a significant impact on the genome.

Supplementary Material

Supplementary figures S1 and S2 and tables S1, S2, and S3 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Bill Ballard, Sylvain Charlat, Penny Haddrill, and Peter Andolfatto for the original fly collections and/or permission to use their fly lines; Kang-Wook Kim for performing DNA extractions; Jill Lovell and the Edinburgh Gene Pool for DNA sequencing; Dinesh Soares for help and advice with structural modeling; Haipeng Li, Pavlos Pavlidis, and Molly Przeworski for advice on the use of their software and/or help with computational approaches to selective sweep analysis. We thank John Welch, Andrea Bettancourt, Pavlos Pavlidis, John McDonald, and an anonymous reviewer for comments which improved the manuscript considerably. We also thank Nick Barton for comments on the first version of this article and Toby Johnson for performing a selective-sweep analysis presented in that version. Author contributions: T.J.L. and F.M.J. initiated this study of diptherian immunity genes. D.J.O. and F.M.J. designed this analysis of AGO2 and D.J.O. performed all PCR, DNA sequencing, and sequence analyses. F.M.J. provided fly stocks, screened D. melanogaster for inversions, and performed crosses. N.J.B. performed the structural modeling and its interpretation. D.J.O., F.M.J., and T.J.L. wrote the article, with contributions from N.J.B. This work was funded by Wellcome Trust (grant 073210/Z/03/Z to T.J.L.). D.J.O. is currently funded by Wellcome Trust RCD Fellowship (085064/Z/08/Z) and F.M.J. was supported by a Royal Society research fellowship and Wellcome Trust (grant WT081279MA). N.J.B. was funded by Wellcome Trust (grant 088179/A/09/Z).

References


Presgraves DC, Gerard PR, Cherukuri A, Lyttle TW. 2009. Large-scale selective sweep among segregation distorter chromosomes in...


