doi:10.1111/evo.12011

RECURRENT AND RECENT SELECTIVE SWEEPS IN THE piRNA PATHWAY

Alfred Simkin,[1,2] Alex Wong,[3] Yu-Ping Poh,[1] William E. Theurkauf,[4] and Jeffrey D. Jensen[1,5,6]

[1]*Program in Bioinformatics & Integrative Biology, University of Massachusetts Medical School, Worcester, Massachusetts*

[2]*E-mail: alfred.simkin@umassmed.edu*

[3]*Department of Biology, Carleton University, Ottawa, Ontario, Canada*

[4]*Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, Massachusetts*

[5]*École Polytechnique Fédérale de Lausanne (EPFL), School of Life Sciences, Lausanne, Switzerland*

[6]*Swiss Institute of Bioinformatics (SIB), Switzerland*

Received March 7, 2012
Accepted October 25, 2012
Data Archived: Dryad doi:10.5061/dryad.1v8j7

Uncontrolled transposable element (TE) insertions and excisions can cause chromosome breaks and mutations with dramatic deleterious effects. The PIWI interacting RNA (piRNA) pathway functions as an adaptive TE silencing system during germline development. Several essential piRNA pathway proteins appear to be rapidly evolving, suggesting that TEs and the silencing machinery may be engaged in a classical "evolutionary arms race." Using a variety of molecular evolutionary and population genetic approaches, we find that the piRNA pathway genes *rhino*, *krimper*, and *aubergine* show patterns suggestive of extensive recurrent positive selection across *Drosophila* species. We speculate that selection on these proteins reflects crucial roles in silencing unfamiliar elements during vertical and horizontal transmission of TEs into naïve populations and species, respectively.

KEY WORDS: **Competition, maternal effect, molecular evolution, parasitism, polymorphism.**

piRNAs have been identified as the primary germline silencing agents for transposable elements (TEs; Brennecke et al. 2007). In the current model for TE silencing, 23 to 30nt piRNAs, in complex with PIWI clade Argonaute proteins, recognize and cleave complementary TE mRNAs. In *Drosophila*, piRNA silencing begins with a pool of preexisting piRNAs, termed primary piRNAs. piRNAs can replenish themselves, but appear to require preexisting maternally inherited piRNAs to prime the system. piRNAs are encoded by specialized 50 to 240 kb heterochromatic loci composed of nested TE fragments, termed piRNA clusters (Brennecke et al. 2007). The primary piRNAs that are complementary to TE sequences, termed "anti-sense stranded piRNAs," are bound by the PIWI protein Aubergine and cleave sense stranded TE transcripts, silencing expression and generating the precursors of so-called "sense strand" piRNAs that associate with the PIWI protein Argonaute 3 (Ago3). The Ago3-sense strand piRNA complexes cleave cluster transcripts, producing precursors for antisense stranded piRNAs (Gunawardane et al. 2007). The piRNA pathway is therefore composed of genetically defined proteins and clusters that require epigenetically inherited small RNAs to amplify and transmit silencing activity. The *Drosophila* RNAi and miRNA pathways, by contrast, do not appear to generate epigenetically heritable silencing activity.

Several recent studies have examined the evolution of small silencing RNA pathway proteins, including some with roles in piRNA silencing. Utilizing McDonald–Kreitman based approaches (Obbard et al. 2006), polymorphism-based composite likelihood tests of selection (e.g., CLSW; Kim and Stephan 2002) and Sweepfinder (Nielsen et al. 2005), and divergence-based analyses (e.g., phylogenetic analysis by maximum likelihood [PAML]; Yang 2007), these studies have shown that adaptive evolution is frequent within the RNAi pathway, which provides antiviral activity and is involved in transposon silencing. These observations suggest that viral infection and TE activity drive

EVOLUTION
INTERNATIONAL JOURNAL OF ORGANIC EVOLUTION

Table 1. piRNA proteins studied and putative functions. # annotated sp. describes the number of species correctly annotated on flybase (v. 5.29) as protein-coding orthologs.

Name	Annotation symbol	function	# Annotated sp.	Citation
Ago3	CG40300	piRNA binding/target cleavage	1	Li et al. (2009)
Armitage	CG11513	Helicase	10	Vagin et al. (2006)
Aubergine	CG6137	piRNA binding/target cleavage	10	Brennecke et al. (2007)
Krimper	CG15707	Tudor domain/nuage	12	Lim and Kai (2007)
Piwi	CG6122	piRNA binding/target cleavage	10	Brennecke et al. (2007)
Rhino	CG10683	Chromatin assembly	1	Klattenhoff et al. (2009)
SpnE	CG3158	Helicase	12	Vagin et al. (2006)
Squash	CG4711	Nuclease	12	Pane, Wehr, and Schüpbach (2007)
Vasa	CG3506	Helicase	4	Malone et al. (2009)
Zucchini	CG12314	Nuclease	12	Pane, Wehr, and Schüpbach (2007)

evolution of these small RNA-based silencing pathways. Numerous studies have demonstrated that certain classes of TEs have exhibited bursts of activity in the recent past (e.g., Yang et al. 2006; Diaz Gonzalez et al. 2010). However, Castillo et al. (2011) have used divergence estimates derived from PAML to show that positive selection in the piRNA pathway does not correlate with either the number of TE families or amount of TE sequence in the genome. This suggests that positive selection of piRNA pathway proteins is limited by purifying selection maintaining the core functionality of the piRNA pathway. Using both Sweepfinder and PAML, Kolaczkowski et al. (2011) find pervasive evidence of selection in both piRNA pathway and RNAi pathway genes, and note that RNAi pathway proteins with the strongest evidence of selection tend to be those that interact most directly with double-stranded RNA, consistent with the idea that selection is strongest at the interface between target RNA molecules and silencing machinery. Others have noted that there is strong evidence of recent positive selection centered on Argonaute2 in *D. melanogaster*, *D. simulans*, and *D. yakuba*, which the authors attribute to a longstanding arms race with viral and/or TE antagonists (Obbard et al. 2011).

We extend existing evolutionary approaches at higher resolution using multiple timescales, through an examination of the evolutionary pressures operating on 10 piRNA pathway proteins (Table 1) using a combination of divergence- and polymorphism-based methods, and find that selection is surprisingly nonuniform. However, we find strong evidence of positive selection within a core set of piRNA proteins in both divergence and polymorphism datasets, leading us to propose a model in which TEs recurrently exploit the same proteins as they first invade and then spread through natural populations.

Materials and Methods

We investigated the evolutionary history of 10 piRNA pathway genes, chosen by dramatically increased transposition rates in double or single knockout individuals (see Table 1) using two main approaches: divergence-based statistics were used to detect recurrent selection across multiple lineages, and polymorphism-based statistics were used to detect recently completed species-specific selective sweeps.

DIVERGENCE-BASED EVOLUTIONARY ANALYSIS

Our divergence-based analyses used three basic aspects of the PAML package (Yang 2007), referred to here as the sites test of selection, the branch test of selection, and the branch-sites test of selection. All of these divergence-based approaches give estimates of the ratio of nonsynonymous changes (dN) to synonymous changes (dS), with each test implemented to detect selective pressures under a different, narrow set of assumptions. Ratios of dN/dS greater than 1 can be attributed to positive selection driving nonsynonymous fixation, ratios near 1 are consistent with neutrality, and ratios smaller than 1 may be attributed to the action of purifying selection on nonsynonymous sites. In all cases, a likelihood ratio test (LRT) is used to compare a neutral null model with an alternative model allowing positive selection.

Divergence tests for positive selection were carried out among six closely related *Drosophila* species (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, and *D. ananassae*) using the PAML analysis package (Table 1, Yang 2007). Only those sequences no more divergent from *D. melanogaster* than *D. ananassae* were considered (roughly corresponding to a per site substitution rate of 1.0) to avoid the issue of saturation of synonymous sites that has been shown to cause dS to appear artificially small in highly divergent species, thus overestimating dN/dS in divergent clades (Stark et al. 2007; Clark et al. 2007). The protein-coding cDNA of all proteins among these six species was obtained from the flybase (release 5.29) website (Tweedie et al. 2009) and aligned using PRANK alignment software with the codon alignment option (Löytynoja and Goldman 2005).

Where a single ortholog was not annotated among all six species, best reciprocal blast annotated transcripts were chosen for analysis. Finally, where no best reciprocal blast hits were returned, syntenic alignments were collected from the UCSC genome browser as recommended by Kolaczkowski and Kern (pers. comm.). The Rhino protein sequences were obtained directly from GenBank annotations provided in the supplementary methods of Vermaak et al. (2005).

To examine positive selection on individual amino acids within a background of purifying selection, we utilized the sites model of PAML. If it is assumed that selective pressures do not vary across the phylogeny, PAML can use this model to estimate the distribution of selective constraints across the length of a protein. The sites model uses three pairs of null and alternative models, termed M1a versus M2a, M7 versus M8, and M8a versus M8 (see Supplementary Materials). In all three comparisons, the null model assumes all codons evolve only under purifying selection or neutrality, whereas the alternative allows the possibility of an additional class of codons which evolve under positive selection.

The branch-based method relies on a user-input phylogeny, and compares a model in which dN/dS values of each lineage are fixed at the same value to a model in which dN/dS scores vary across lineages. In addition to log likelihood scores, the null model produces an estimate of the fixed maximum likelihood dN/dS across all lineages, whereas the alternative estimates individual maximum likelihood estimates for each branch.

The branch-sites model separates all possible lineages into groups, such that one group is designated as "background" and constrained to evolve neutrally or under selective constraint, whereas the other is designated "foreground" and allowed to contain, in addition to neutral or constrained sites, sites with dN/dS values greater than one. This alternative model is measured against a more constrained model in which the foreground branch is also constrained to be neutral or negatively selected (dN/d$S \leq 1$).

To test the significance of the above methods within the context of the *Drosophila* genome, we obtained a full list of protein-coding annotated orthologs from Flybase. Only *D. melanogaster* genes with exactly one annotated ortholog in the species *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, and *D. ananassae* were included. cDNA of these orthologs was aligned using PRANK alignment software as above, and only sequences with greater than 20 alignable amino acids across all six species were retained. Sequences that did not complete in any PAML tests were removed from all analyses, and the resulting 9,005 genes were used as a rough proxy to estimate the genomic distribution of selective pressures. In permutation tests drawing random sets of genes from this genomic set to assess significance, the following criteria were used to assign individual genes to a putatively positively selected category. For the branch model, only genes that met the three criteria of rejecting equal dN/dS values across all lineages, containing one or more lineages with elevated dN/d$S >$ 0.5, and possessing dS values > 0.02 for these elevated dN/dS values were considered to be putatively positively selected (i.e., to avoid false inference owing to low dS values (Fig. S1). For the sites model, genes drawn from the permutation which rejected neutrality under at least one of the model pairs M1a versus M2a, M7 versus M8, and M8a versus M8 were examined for presence of individual amino acids with high probabilities of selection. For the branch-sites model, lineages that rejected neutrality were considered to be under putative positive selection, and if these lineages also possessed amino acids with a high probability (> 0.95) of selection, these amino acids were also considered to be putatively positively selected.

POLYMORPHISM-BASED EVOLUTIONARY ANALYSIS

As a complement to the purely divergence-based analyses, we investigated variation at the population level using (1) site frequency spectrum-based tests of selection (e.g., CLSW; Kim and Stephan 2002), (2) site frequency spectrum-based tests of neutrality (e.g., the H statistic; Fay and Wu 2000, as well as (3) the McDonald–Kreitman (1991) test.

The alignments studied here consist of syntenic assemblies from an average of 34 *D. melanogaster* individuals from North Carolina, an average of 6 *D. melanogaster* individuals from Malawi, and 6 *D. simulans* individuals from several inbred stocks, analyzed separately for each of the 10 piRNA pathway proteins. We used polymorphism data from the Drosophila Population Genomics Project (DPGP). Ambiguous nucleotides were conservatively analyzed by changing them to the most common nucleotide at the site among the population. These alignments were converted into ms format using *D. yakuba* as an outgroup ancestral state using an online recombination rate estimator derived from mapping studies (Fiston-Lavier et al. 2010) and an assumed population size of 1,000,000 individuals.

Using the frequency of polymorphisms across the sequence, CLSW assigns a likelihood to models of selection and neutrality, and performs a ratio test. The likelihood ratio scores from CLSW were compared to 1000 neutral simulations of identical θ ($= 4N_e\mu$), ρ ($= 4N_e r$), and length (in basepairs) parameters generated using the ssw simulation program (Kim and Stephan 2002)—simulations that were also used to assess significance of Fay and Wu's H-test (2002). Significant regions ($P < 0.05$) were further compared to 1000 selection simulations generated using ssw assuming a beneficially fixed mutation immediately prior to sampling ($\tau = 0.000001$ 2N generations). The goodness-of -fit statistic (GOF; Jensen et al. 2005) was used to determine whether selection alone was sufficient to explain the data.

The McDonald–Kreitman test is a comparison of nonsynonymous polymorphism to divergence compared with synonymous

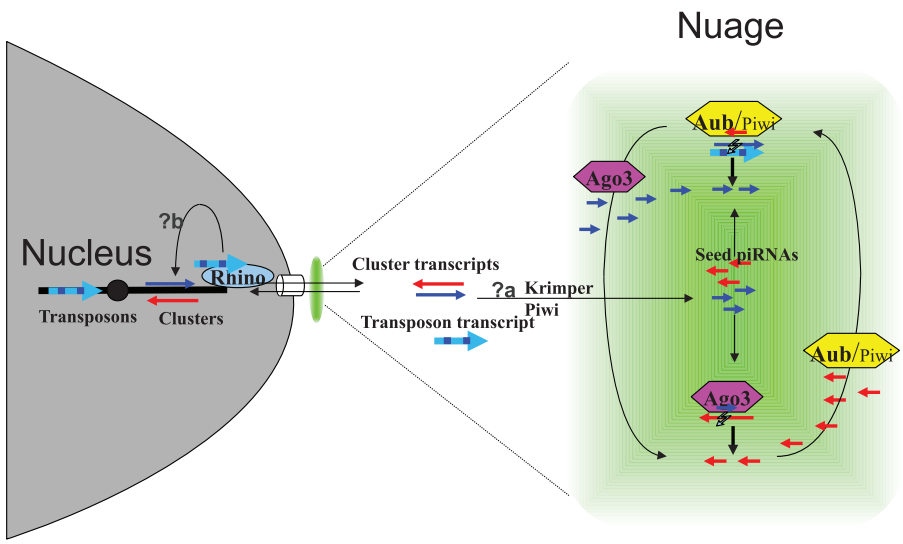

Figure 1. Transposon control by the piRNA pathway. Black arrows represent steps in pathways, whereas blue and red colored arrows represent sense and antisense RNA transcripts. Bidirectional cluster transcripts are produced from both strands of loci with homology to transposons. During hybrid dysgenic scenarios, these transcripts are cleaved by unknown mechanisms, which we speculate may be associated with modifications to PIWI and Krimper ?a that produce primary "seed" piRNAs. A large number of piRNA pathway proteins appear to localize to the nuage complex (green), an assembly of proteins situated between the nucleus and cytoplasm. piRNAs corresponding to sense strands (dark blue) are preferentially loaded onto Argonaute3, whereas piRNAs corresponding to antisense strands (red) are loaded onto Piwi and Aubergine. Once loaded, these Piwi family proteins are thought to mediate cleavage of complementary sequence from both transposons and clusters, silencing transposons and further amplifying both sense-stranded and antisense-stranded piRNA pools in a self-sustaining process that may be maintained across generations if these piRNAs are heritable. During horizontal transfer scenarios or activation of older transposons lacking cluster silencing, complementarity of clusters to transposons must be established, which we speculate might be associated with modifications to Rhino that favor the integration of these transposon transcripts into clusters ?b, thus expanding the repertoire of effective piRNAs.

polymorphism to divergence. The expectation under neutrality (for which synonymous sites here serve as a proxy) is that the rate of fixation is simply given by the neutral mutation rate. We utilized two population samples of *D. melanogaster* (deriving from Malawi and North Carolina), and determined divergence as compared to *D. simulans*.

Results

SPECIES-LEVEL ANALYSES

In the sites model of PAML, two of three tests of Ago3 rejected neutrality in favor of positive selection, but failed to localize this selection to any individual amino acid, whereas PIWI was identified as containing amino acids with a high posterior probability of positive selection but rejected neutrality in only one out of three model comparisons (M7 vs. M8). Only Rhino was significant across all three model comparisons and identified individual amino acids with a high posterior probability of positive selection (Table 2 columns 3 and 5–7), with approximately 0.011% of all genes in our genomic set showing a similar or more extreme pattern (Table 2).

In the branch model allowing dN/dS to vary among lineages, estimates of dN/dS are produced for each branch. Nearly all piRNA pathway proteins fit the varying dN/dS model significantly better than the one constraining selection to be identical along all lineages, suggesting that selective pressures vary across the phylogeny. Few proteins were inferred to have dN/d$S > 1.0$. However, even proteins experiencing positive selection on key amino acids can be expected to evolve under strong purifying selection along most of their length, producing average dN/dS values less than 1. The Rhino protein is one such example, showing clear evidence of positive selection within individual amino acids and rejecting neutrality in all three sites model comparisons, yet evolving generally under constraint, with dN/d$S < 1.0$ in the branch model in all lineages. We noted that the piRNA pathway proteins Rhino, Aubergine, and Krimper each had several lineages with dN/d$S > 0.5$, higher than we would expect for proteins with sterile loss of function phenotypes, which we would expect to evolve under strong selective constraint with dN/dS near 0. The other six piRNA proteins had no lineages with such elevated dN/dS (Fig. 2, Table 2).

To generate an empirical distribution, we compared these values to dN/dS estimates performed across 9,005 proteins with

Table 2. Divergence-based analyses. Rhino, Krimper, and Aubergine show the largest number of lineages under positive selection (1st and 2nd columns, out of 15 total pairwise comparisons and 9 total branches, respectively). Some piRNA pathway proteins were predicted to have individual amino acids under significant positive selection across the entire phylogeny.[1] Others showed significant evidence of lineage-specific positive selection (columns 5 and 6). Bonferroni corrected *P*-values are shown in parentheses, and comparisons to genomic distributions are shown in square brackets.

Name	Pairwise d*N*/d*S* >0.5	Branch model branches>0.5	Branch vs. Equal	Sites M7 vs. M8[2]	Branch-sites branches[3]	Branch-sites lineage *P*-values
Ago3	1/15	0/9	$P<0.001$	$P<0.004$[0.02]	none	Not significant
Armitage	0/15	1/9[0.07]	$P<0.001$	Not significant	none	Not significant
Aubergine	1/15	2/9[0.02]	$P<0.001$	$P<0.035$[0.06]	mel sim sec all, sim sec, sim sec all, mel[0.009]	0.04(0.47), 0.0008(0.009), 0.0009(0.01), 0.007(0.08)
Krimper	3/15	1/9[0.07]	$P<0.001$	Not significant	sim sec, sim sec all[0.1]	0.00005(0.0006), 0.0001(0.002)
Piwi	0/15	0/9	$P<0.002$	$P<0.037$[0.06]	none	Not significant
Rhino	6/15	3/9[0.009]	$P<0.001$	$P<0.017$[0.04]	yak ere, ana[0.1]	0.04(0.49), 0.006(0.07)
SpindleE	0/15	0/9	$P<0.001$	Not significant	ana	0.05(0.60)
Squash	0/15	0/9	$P<0.019$	Not significant	none	Not significant
Vasa	0/15	0/9	$P<0.001$	Not significant	none	Not significant
Zucchini	0/15	0/9	$P<0.001$	Not significant	mel sim sec, mel[0.1]	0.003(0.03), 0.03(0.36)

[1]Only Piwi and Rhino identified individual amino acids under positive selection. Piwi identified 56L as under positive selection under M8. Rhino identified 46S as under positive selection in M2a and M8.

[2]Two other sites tests were also performed (see section Materials and Methods). For M1a versus M2a, Rhino was the only significant protein ($P < 0.048$ with 1% of the genomic dataset more significant), whereas for M8a versus M8, Rhino and Argonaute3 were both significant ($P < 0.006$ and $P < 0.022$, with 1% and 2% of the genomic dataset more significant, respectively).

[3]Only Krimper and Aubergine localized positive selection to individual amino acids. Krimper analysis found six amino acids: 122T, 130S, 144E, 326E, 411S, and 416T, whereas Aubergine analysis found over 20 amino acids throughout the protein.

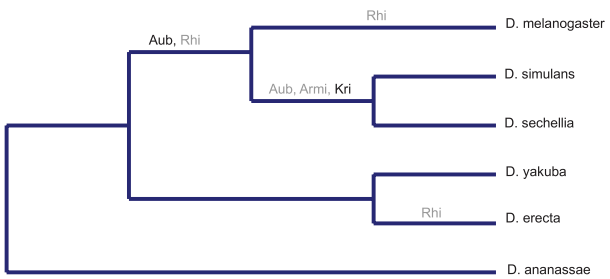

Figure 2. Branch model of PAML lineages under selection. Because this model cannot assign probabilities to individual lineages under selection, d*N*/d*S* cut-offs of 0.5 and 1.0 were used as proxies for positive selection or relaxed selective constraint, with an additional filter for d*S* > 0.02. When analyzed in this way, 3/9 lineages in Rhino, 1/9 lineages in Krimper, 2/9 lineages in Aubergine, and 1/9 lineages in Armitage show evidence of recurrent positive selection or relaxed selective constraint (d*N*/d*S* > 0.5, gray font, d*N*/d*S* > 1.0, black font). Kri = Krimper; Aub = Aubergine; Armi = Armitage; Rhi = Rhino; Zuc = Zucchini; SpnE = Spindle-E.

1:1 annotated orthologs among all six species, and found that only 7% of proteins possessed 1 or more lineages with d*N*/d*S* > 0.5. Performing a permutation test (10 proteins randomly chosen from our genomic distribution with 10,000 replicates), 0.4% of permutations contain 4 or more proteins with 1 or more lineages having d*N*/d*S* values > 0.5. Because only a subset of genomic proteins with elevated d*N*/d*S* are likely to be essential, our permutation results present an overestimate of the number of essential genomic proteins with similarly elevated d*N*/d*S*. Therefore, although we cannot rule out relaxed selective constraint as a cause, it is plausible that the statistically enriched elevations in d*N*/d*S* among our protein set are attributable to positive selection in a subset of amino acids (see also Swanson et al. 2004).

It is notable that the sites model assumes no variation in selective pressures across branches, an assumption which the branch model suggests is not valid, whereas the branch model cannot assign a probability of positive selection to a given branch or amino acid. Therefore, to distinguish between lineage-specific positive selection among individual amino acids and relaxed selective constraint, the branch-sites test of selection (Zhang et al. 2005) was implemented to allow for positive selection along an individual lineage while constraining the rest of the phylogeny to be evolving neutrally or under negative selection. The probability of such a scenario is assessed relative to a nearly identical model that assumes no positive selection on the lineage under examination or the rest of the phylogeny. By assuming selection along a single lineage, the probability of positive selection along a particular

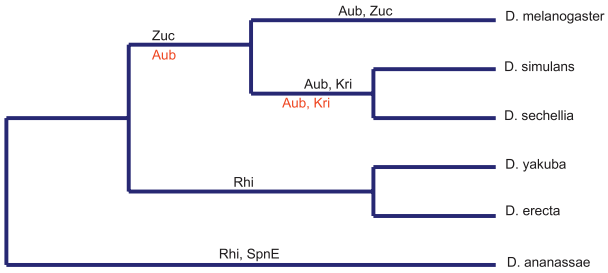

Figure 3. Branch-sites model of PAML lineages under positive selection. When imposing the criteria of a subset of branches allowing positive selection compared against a background of the remaining branches constrained to be under negative or neutral selection, the result is a group of piRNA proteins that, by definition, are not recurrently positively selected across the entire phylogeny. piRNA protein lineages rejecting neutrality with *P*-values < 0.05 are shown, and are most prominently represented by Aubergine. All lineages rejecting neutrality under these criteria also have d*N*/d*S* values > 1.0. Black represents single selective events, whereas red represents recurrent selection along an internal branch and all of its descendant lineages.

lineage can be directly measured in a way that is not possible in the branch model. The branch-sites model thus gains power relative to the branch model to detect the probability of lineage-specific selection at the expense of a loss of power to detect positive selection operating across more than one lineage, making these two tests, to some extent, complements of one another.

In contrast to the earlier analysis of sites positively selected across all lineages, which failed to find individual amino acids under selection in most piRNA proteins, the branch-specific tests were able to reject neutrality in favor of positive selection in several proteins, and identified individual amino acids with a high posterior probability of positive selection in Krimper and Aubergine (Fig. 3). This suggests that the elevations in d*N*/d*S* seen in the branch model in these proteins could be attributable to site and lineage-specific positive selection. Branch-sites tests consistently revealed the lineages *D. melanogaster*, *D. simulans*, and *D. sechellia* to be undergoing some combination of recurrent positive selection within these proteins. Other piRNA proteins also were identified as putatively positively selected, but without localization to any individual amino acids. SpindleE and Rhino showed evidence of positive selection within *D. ananassae*, Zucchini showed evidence of some recurrent positive selection in the *D. melanogaster*, *D. simulans*, and *D. sechellia* lineages, and Rhino showed evidence of positive selection in the ancestor of *D. yakuba* and *D. erecta*. Notably, Rhino—for which individual amino acids were identified as under recurrent positive selection in all three sites models—did not show evidence for positive selection on these amino acids in any individual lineage under the branch-sites model. This lack of overlap illustrates the respective power of the branch and branch-sites models to detect recurrent and lineage-specific positive selection acting on individual amino acids.

Within Aubergine, 24 sites were estimated in the branch-sites model to have a probability greater than 95% of being under positive selection. These amino acids do not appear to be centralized within one domain, but rather are dispersed across the length of the protein, similar to the findings of Kolaczkowski et al. (2011). Krimper had five sites estimated to be similarly positively selected within an annotated tudor domain and a nonsignificant Pfam match to a second unannotated tudor domain, as well as two additional amino acids outside of either domain.

When the branch-sites and branch results are summarized, it is notable that Rhino, Aubergine, and Krimper all have lineages with d*N*/d*S* > 0.5 and appear to have strong evidence of positively selected amino acids in the branch-sites or sites tests. Based on a genome-wide permutation test, the probability of such an observation in a set of 10 random proteins = 0.007. These observations suggest that a large portion of the piRNA pathway, as defined by the 10 proteins examined here, is shaped by recurrent positive selection.

POPULATION-LEVEL ANALYSES

In CLSW tests with the GOF correction, the North Carolina *D. melanogaster* group rejected neutrality in favor of selection in Aubergine, SpindleE, and Rhino (Table 3). Within *D. simulans*, Armitage, Vasa, and Zucchini all rejected neutrality in favor of selection (Table 3). In McDonald–Kreitman tests, Armitage, Aubergine, Krimper, and SpindleE all rejected neutrality in both *D. melanogaster* populations, whereas Vasa rejected neutrality in the Malawi population but not the North Carolina population. Performing Fay and Wu's *H*-test separately for each piRNA gene within each population, 12 such tests rejected equilibrium neutrality (Table S1).

SUMMARY

We find evidence of pervasive positive selection operating in the piRNA pathway—most notably within Rhino, Aubergine, and Krimper, which have strong divergence- and polymorphism-based evidence for both recurrent and recent strong positive selection. In earlier studies, these three proteins do not stand out relative to the rest of the piRNA pathway.

Here, we utilize a new alignment algorithm, PRANK, which has been shown recently (Fletcher and Yang 2010) to have a dramatically lower false positive rate in detecting positive selection in PAML branch-sites simulations while incurring only modest sacrifices in the detection of true positives. In addition, our use of the branch, sites, and branch-sites models allows for the detection of positive selection operating across the entire phylogeny. Also, using recent genomic resources from the DPGP we are able to evaluate two distinct population samples of *D. melanogaster*

Table 3. Polymorphism-based analyses. Only genes which reject neutrality in favor of positive selection are shown.

Gene	LR[4] value	LR *P*-value	GOF[5] score	GOF *P*-value	α[6]	X[7]
SpnE_NC	18.502	0	14.589	0.139	3913.87	6621
Aubergine_NC	10.145	0	−77.615	0.573	670.20	1865
Vasa_sim	5.478	0.001	99.155	0.234	1817.54	268
Armi_sim	4.563	0.001	93.992	0.731	964.41	2821
Zucchini_sim	1.950	0.007	26.697	0.223	162.90	394
Rhino_NC	4.616	0.022	−93.566	0.719	169.57	302

[4]LR denotes the natural log likelihood ratio of selection versus neutrality, as calculated in Kim and Stephan (2002). Because demographic parameters can affect these scores, neutral simulation is performed to assign empirical *P*-values, and selection is accepted as an alternative to neutrality when *P*-values are <0.05.

[5]GOF is a measure of the goodness of fit of the data to selection, as calculated in Jensen et al. (2005). Simulation under selection is performed to estimate empirical *P*-values, and selection is accepted as a viable alternative to demographic processes when *P*-values are <0.1.

[6]α denotes the selection strength, and is given by 2 Ns.

[7]X is the maximum likelihood location of the beneficial mutation in nucleotides.

Table 4. McDonald–Kreitman tests (only significant genes are shown).

Gene	Fixed nonsyn	Fixed syn	Poly nonsyn	Poly syn	Fisher table *P*	Fisher marg *P*	Chi square *P*	*G*-test *P*	G-test Williams Corr. *P*
Armi_MW	87	70	33	57	0.00545409	0.00188761	0.00455326	0.00436815	0.00450088
Armi_NC	83	70	38	64	0.0103282	0.00296076	0.00776147	0.0075214	0.00770929
Aubergine_MW	101	59	8	20	0.000814674	0.000537145	0.000632438	0.000638369	0.00071739
Aubergine_NC	99	58	11	28	0.0001194	6.78798e−05	8.65505e−05	8.03173e−05	8.97841e−05
Krimper_MW	119	46	57	48	0.00382677	0.00122766	0.00270824	0.00283003	0.00292413
Krimper_NC	123	47	58	55	0.000390862	0.000160545	0.000308591	0.000321512	0.000335574
SpnE_MW	87	95	24	54	0.0136117	0.00421318	0.0109429	0.0100884	0.0103694
SpnE_NC	88	96	30	56	0.0492191	0.0143502	0.0457688	0.044437	0.0451641
Vasa_MW	114	100	15	38	0.00121451	0.000571227	0.0011274	0.000944243	0.00100057

(Malawi and North Carolina). Finally, to characterize the relative mode and tempo of selection as compared to other coding regions, we examine a dataset of 9005 proteins in divergence analyses. This allows for a genomic distribution of selective pressures relative to which we can compare the results from the piRNA pathway.

Discussion

Although we see strong evidence of positive selection in piRNA proteins consistent with an evolutionary "arms race," it is difficult to account for the mechanism by which natural selection operates through a high substitution rate in TEs alone. Because the units conferring resistance to TEs are genetically inherited cluster insertions and epigenetically inherited mature piRNA pools that target TE transcripts, we would expect these variants to sweep through populations to confer resistance to a novel threat. Modifications to piRNA proteins that allow for these variants to occur, by contrast, would not be under selective pressure. Furthermore, because clusters have extensive sequence complementarity to the TEs they regulate, it is unlikely that a silenced TE could evade the piRNA pathway through mutation without destroying functionality. To explain the recurrent fixations we observe consistently in Rhino, Aubergine, and Krimper, we therefore speculate that the adaptive silencing mediated by the piRNA pathway is responding to a hypothetical class of TE-encoded inhibitors.

All 10 piRNA pathway proteins we examined have key roles in TE silencing and germline development. The clear prominence of only Krimper, Aubergine, and Rhino within our analyses is therefore quite unexpected, and is consistent with specific roles for these proteins in the adaptive response to novel elements. TEs are thought to spread through populations predominantly through direct inheritance. Therefore, most TEs will be transmitted with their silencing clusters. However, piRNAs are epigenetically inherited maternally, and amplification of the silencing RNA pool requires preexisting piRNAs. Paternally inherited TEs thus escape silencing in hybrids, leading to genetic instability and sterility. As hybrids age, however, piRNAs are produced de novo from paternal clusters, TEs are silenced and fertility is recovered (Khurana et al. 2011). If the initial failure to generate primary piRNAs from

inherited clusters is mediated in part by inhibitors encoded by the invading element, vertical TE spread may impose strong, recurrent selection within the host for genetic variants that evade these inhibitors and thus enhance de novo production of primary piRNAs from existing cluster transcripts, allowing for the observed re-establishment of fertility.

The evidence we see of positive selection acting on Krimper and Aubergine is consistent with a previous analysis of d*N*/d*S* conducted in *D. melanogaster* and *D. simulans* which found that these two proteins have the highest rates of amino acid substitution in the piRNA pathway (Obbard et al. 2009). These proteins may therefore represent promising candidates for further study of the adaptive response to new TEs. Aubergine is a PIWI protein that binds to mature piRNAs and has a direct role in the cleavage of piRNA precursors needed to amplify the primary piRNA pool (Brennecke et al. 2007). Krimper has a Tudor domain, and many Tudor domain proteins appear to directly bind dimethylated PIWI proteins (Siomi et al. 2010). This observation, the observed positive selection in both divergence based ((Figs. 2, 3), Table 2) and polymorphism based (Tables 3, 4) analyses, and the colocalization of Krimper and Aubergine in the nuage complex (Fig. 1), open the possibility that they may directly interact to process novel TEs into mature piRNAs in dysgenic scenarios even in the absence of preexisting guide RNAs, allowing them to activate their inherited clusters.

During horizontal transfer of TEs between species, by contrast, silencing appears to require insertion of invading elements into clusters, which generates piRNA precursors capable of initiating the amplification and silencing cycle. These occurrences may be surprisingly frequent, as evidenced by the introduction and fixation of the P-element within *D. melanogaster* over the course of the last 40 years, likely from *D. willistoni* (Anxolabéhére et al. 1988; Daniels et al. 1990), as well as the great diversity of TEs within and between Drosophila species (Yang et al. 2006; Clark et al. 2007; Diaz Gonzalez et al. 2010).

Rhino localizes to heterochromatin and is necessary for the production of piRNAs from dual strand clusters (Klattenhoff et al. 2009). Rhino could therefore play some role in directing transposition into clusters during horizontal transfer, perhaps through interaction with the transposition machinery. Alternatively, TEs may encode proteins that inhibit transposition into clusters to avoid silencing. Both models predict that Rhino will be under selection during horizontal transfer, but not vertical transfer, and are consistent with recent studies indicating that clusters, piRNA populations, and siRNA populations change dramatically and often globally on very short evolutionary timescales in response to changes in TE composition (Khurana et al. 2011; Rozhkov et al. 2010, 2011).

These evolutionary insights should help guide studies on the epigenetic and genetic functions of rapidly evolving piRNA pathway proteins in TE silencing within naïve populations and species. The sterile phenotype of piRNA mutants, the rapid accumulation of TEs in the Drosophila phylogeny, and the increasing number of studies demonstrating hybrid dysgenesis suggest that these events may be a strong and perhaps surprisingly frequent contributor to the complex interplay between piRNA pathway function and TE propagation.

Funding

This work was funded by a grant from the National Science Foundation [DEB-1002785], a faculty award from the Worcester Foundation to J. D. Jensen, a grant from the National Institutes of Health to Fen-Biao Gao (RO1 NS066586), and a grant from the National Institutes of Health to William Theurkauf (RO1 HD049116).

ACKNOWLEDGMENTS

We would like to thank Andrew Kern, Bryan Kolaczkowski, Nadia Singh, and members of the Jensen and Theurkauf Labs for numerous intellectual contributions.

LITERATURE CITED

Anxolabéhère, D., M. G. Kidwell, and G. Periquet. 1988. Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of Drosophila melanogaster by mobile P elements. Mol. Biol. Evol. 5:252–269.

Brennecke, J., A. A. Aravin, A. Stark, M. Dus, M. Kellis, R. Sachidanandam, and G. J. Hannon. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell 128:1089–1103.

Brower-Toland, B., S. D. Findley, L. Jiang, L. Liu, H. Yin, M. Dus, P. Zhou, S. C. R. Elgin, and H. Lin. 2007. Drosophila PIWI associates with chromatin and interacts directly with HP1a. Genes Dev. 21:2300–2311.

Castillo, D. M, J. C. Mell, K. S. Box, and J. P. Blumenstiel. 2011. Molecular evolution under increasing transposable element burden in Drosophila: a speed limit on the evolutionary arms race. BMC Evol. Biol. 11:258.

Clark, A. G., M. B. Eisen, D. R. Smith, C. M. Bergman, B. Oliver, T. A. Markow, T. C. Kaufman et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450:203–218.

Daniels, S. B., K. R. Peterson, L. D. Strausbaugh, M. G. Kidwell, and A. Chovnick. 1990. Evidence for horizontal transmission of the P transposable element between Drosophila species. Genetics 124:339–355.

Díaz-González, J., A. Domínguez, and J. Albornoz. 2010. Genomic distribution of retrotransposons 297, 1731, Copia, Mdg1 and Roo in the *Drosophila melanogaster* species subgroup. Genetica 138:579–586.

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

Fiston-Lavier, A. S., N. D. Singh, M. Lipatov, and D. A. Petrov. 2010. *Drosophila melanogaster* recombination rate calculator. Gene 463:18–20.

Fletcher, W., and Z. Yang. 2010. The effect of insertions, deletions and alignment errors on the branch-site test of positive selection. Mol. Biol. Evol. 27:2257–2267.

Gunawardane, L. S., K. Saito, K. M. Nishida, K. Miyoshi, Y. Kawamura, T. Nagami, H. Siomi, and M. C. Siomi. 2007. A slicer-mediated mechanism for repeat-associated siRNA 5, end formation in Drosophila. Science 315:1587–1590.

Huisinga, K. L, and S. C. R. Elgin. 2009. Small RNA-directed heterochromatin formation in the context of development: what flies might learn from fission yeast. Biochim. Biophys. Acta 1789:3–16.

Jensen, J. D., Y. Kim, V. B. DuMont, C. F. Aquadro, and C. D. Bustamante. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170:1401–1410.

Kazazian, H. H. 2004 Mobile elements: drivers of genome evolution. Science 303: 1626–1632.

Khurana, J. S., J. Wang, J. Xu, B. S. Koppetsch, T. C. Thomson, A. Nowosielska, C. Li, P. D. Zamore, Z. Weng, and W. E. Theurkauf. 2011. Adaptation to P element transposon invasion in *Drosophila melanogaster*. Cell 147:1551–1563.

Kim, Y., and W. Stephan. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160:765–777.

Klattenhoff, C., H. Xi, C. Li, S. Lee, J. Xu, J. S. Khurana, F. Zhang et al. 2009. The Drosophila HP1 homolog Rhino is required for transposon silencing and piRNA production by dual-strand clusters. Cell 138:1137–1149.

Kolaczkowski, B., D. N. Hupalo, and A. D. Kern. 2011. Recurrent adaptation in RNA-interference genes across the Drosophila phylogeny. Mol. Biol. Evol. 28:1033–1042.

Le Rouzic, A., and P. Capy. 2005. The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. Genetics 169:1033–1043.

Li, C., V. V. Vagin, S. Lee, J. Xu, S. Ma, H. Xi, H. Seitz, M. Horwich, M. Syrzycka, B. Honda, et al. 2009. Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. Cell 137: 509–521.

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

Lim, A. K., and T. Kai. 2007. Unique germ-line organelle, nuage, functions to repress selfish genetic elements in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA 104:6714–6719.

Löytynoja, A., and N. Goldman. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proc. Natl. Acad. Sci. USA 102: 10557–10562.

Malone, C. D., J. Brennecke, M. Dus, A. Stark, W. R. McCombie, R. Sachidanandam, and G. J. Hannon. 2009. Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. Cell 137:522–535.

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

Moshkovich, N., and E. P. Lei. 2010. HP1 recruitment in the absence of argonaute proteins in Drosophila. PLoS Genet. 6:e1000880.

Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3:418–426.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. Bustamante. 2005. Genomic scans for selective sweeps using SNP data. Genome Res. 15:1566–1575.

Obbard, D. J., F. M. Jiggins, D. L. Halligan, and T. J. Little. 2006. Natural selection drives extremely rapid evolution in antiviral RNAi genes. Curr. Biol. 16:580–585.

Obbard, D. J., K. H. J. Gordon, A. H. Buck, and F. M. Jiggins. 2009. The evolution of RNAi as a defence against viruses and transposable elements. Philos. Trans. R. Soc. Lond. 364:99–115.

Obbard, D. J., F. M. Jiggins, N. J. Bradshaw, and T. J. Little. 2011. Recent and recurrent selective sweeps of the antiviral RNAi gene argonaute-2 in three species of Drosophila. Mol. Biol. Evol. 28:1043–1056.

Pal-Bhadra, M., U. Bhadra, and J. A. Birchler. 2002. RNAi related mechanisms affect both transcriptional and posttranscriptional transgene silencing in Drosophila. Mol. Cell 9:315–327.

Pal-Bhadra, M., B. A. Leibovitch, S. G. Gandhi, M. Rao, U. Bhadra, J. A. Birchler, and S. C. R. Elgin. 2004. Heterochromatic silencing and HP1 localization in Drosophila are dependent on the RNAi machinery. Science (New York) 303:669–672.

Pane A., K. Wehr, and T. Schüpbach. 2007. Zucchini and squash encode two putative nucleases required for rasiRNA production in the Drosophila germline. Develop. Cell 12:851–862.

Przeworski, M. 2002. The signature of positive selection at randomly chosen loci. Genetics 160:1179–1189.

Rozhkov, N. V., A. A. Aravin, E. S. Zelentsova, N. G. Schostak, R. Sachidanandam, W. R. McCombie, G. J. Hannon, and M. B. Evgen'ev. 2010. Small RNA-based silencing strategies for transposons in the process of invading Drosophila species. RNA 16:1634–1145.

Rozhkov N. V., E. S. Zelentsova, N. G. Shostak, and M. B. Evgen'ev. 2011. Expression of Drosophila virilis retroelements and role of small RNAs in their intrastrain transposition. PloS One 6:e21883.

Siomi, M. C., T. Mannen, and H. Siomi. 2010. How does the royal family of Tudor rule the PIWI-interacting RNA pathway? Genes Develop. 24:636–646.

Stark, A., M. F. Lin, P. Kheradpour, J. S. Pedersen, L. Parts, J. W. Carlson, M. A. Crosby, M. D. Rasmussen, S. Roy, A. N. Deoras, et al. 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450:219–232.

Swanson, W. J., A. Wong, M. F. Wolfner, and C. F. Aquadro. 2004. Evolutionary expressed sequence tag analysis of Drosophila female reproductive tracts identifies genes subjected to positive selection. Genetics 168:1457–1465.

Tweedie, S., M. Ashburner, K. Falls, P. Leyland, P. McQuilton, S. Marygold, G. Millburn, D. Osumi-Sutherland, A. Schroeder, R. Sea, et al. 2009. FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Res. 37:D555–D559.

Vagin, V. V., A. Sigova, C. Li, H. Seitz, V. Gvozdev, and P. D. Zamore. 2006. A distinct small RNA pathway silences selfish genetic elements in the germline. Science 313:320–324.

Vermaak, D., S. Henikoff, and H. S. Malik. 2005. Positive selection drives the evolution of rhino, a member of the heterochromatin protein 1 family in Drosophila. PLoS Genet. 1:96–108.

Wong, W. S. W., Z. Yang, N. Goldman, and R. Nielsen. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168:1041–1051.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Yang, Z., and R. Nielsen. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. 46:409–418.

Yang, H. P., T. L. Hung, T. L. You, and T. H. Yang. 2006. Genomewide comparative analysis of the highly abundant transposable element DINE-1 suggests a recent transpositional burst in Drosophila Yakuba. Genetics 173:189–196.

Zhang J., R. Nielsen, and Z. Yang. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22:2472–2479.

Associate Editor: A. Cutter

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Figure S1. Genomic distribution of d*S* versus d*N*/d*S*.
Table S1. Fay and Wu's *H* tests.