RNA interference pathways display high rates of adaptive protein evolution across multiple invertebrates

William H. Palmer*[1,2]

Jarrod Hadfield[1]

Darren J. Obbard[1,2]

[1] **Institute of Evolutionary Biology University of Edinburgh, Kings Buildings, West Mains Road, Edinburgh, UK**

[2] **Centre for Infection, Evolution and Immunity, University of Edinburgh, Kings Buildings, West Mains Road, Edinburgh, UK**

***Author for correspondence**

Abstract

Conflict between organisms can lead to reciprocal adaptation that manifests itself as an increased evolutionary rate in the genes mediating the conflict. This adaptive signature has been observed in RNA interference (RNAi) pathway genes involved in the suppression of viruses and transposable elements in *Drosophila melanogaster*, suggesting that a subset of *Drosophila* RNAi genes may be locked into an arms race with these parasites. However, it is not known whether rapid evolution of RNAi genes is a general phenomenon across invertebrates, or which RNAi genes generally evolve adaptively. Here we use population genomic data from eight invertebrate species to infer rates of adaptive sequence evolution, and to test for past and ongoing selective sweeps in RNAi genes. We assess rates of adaptive protein evolution across species by using a formal meta-analytic framework to combine data across species, and by implementing a multispecies generalised linear mixed model of mutation counts. In all species, we find that RNAi genes display a greater rate of adaptive protein substitution than other genes, and that this is primarily mediated by positive selection acting on the subset of genes that are most likely to defend against viruses and transposable elements. In contrast, evidence for recent selective sweeps is broadly spread across functional classes of RNAi genes and differs substantially among species. Finally, we identify genes that exhibit elevated adaptive evolution across the analysed insect species combined, perhaps due to concurrent parasite-mediated arms races.

Introduction

RNA-interference mechanisms include a diverse group of pathways, united by their use of Argonaute-family proteins complexed with short (20-30 nt) RNA molecules to guide the targeting of longer RNA molecules through sequence complementarity (Carmell, et al., 2002; Meister, 2013). These pathways regulate multiple biological processes that can be divided into three distinct subpathways in arthropods and nematodes, each represented by a characteristic class of small RNAs: the micro-RNA (miRNA), the short-interfering RNA (siRNA), and the piwi-interacting RNA (piRNA) pathways. The miRNA pathway processes endogenously-encoded foldback hairpins which, once in their mature miRNA form, regulate gene expression and coordinate developmental processes (Alvarez-Garcia & Miska, 2005; Chen, et al., 2014; Ha & Kim, 2014). The siRNA pathway has two distinct roles, depending on the endogenous or exogenous origin of its substrate. First, the endo-siRNA pathway processes endogenously encoded dsRNA to regulate processes such as TE defense (Kawamura, et al., 2008; Czech, et al., 2008; Ghildiyal, et al., 2008) chromosomal segregation (Hall, et al., 2003; Huang, et al., 2015), and heterochromatin formation (Deshpande, et al., 2005). Second, the exo-siRNA (or viRNA) functions primarily as a form of antiviral immunity (Wang, et al., 2006; Bronkhorst & van Rij, 2014). The piRNA pathway forms a defence against transposable elements (TEs) in the germ line, and piRNAs are derived from endogenously-encoded piRNA clusters of inactivated TE sequences and from active TEs (Klattenhoff & Theurkauf, 2008; Thomson & Lin, 2009; Czech, et al., 2016).

Nevertheless, within this simple framework there is substantial variation among species, and RNAi-pathway components seem to be evolutionarily labile. For example, in nematodes the mechanism and function of the piRNA pathway is not well conserved: primary piRNA-like small RNAs are encoded by short distinct loci instead of the clusters observed in flies and mammals, and mediate the biogenesis of a separate endo-siRNA population transcribed by an RNA-dependent RNA Polymerase (RdRP) and processed by Dicer (Duchaine, et al., 2006; Das, et al., 2008). Further, only one of the five major clades of nematode have retained Piwi-subfamily proteins — the canonical effector of the piRNA pathway —and instead rely solely on the (RDRP-produced) endo-siRNAs (Sarkies, et al., 2015). The piRNA pathway can also take on entirely new roles, for example, multiple duplications of *piwi* in *Aedes* mosquitoes has allowed the piRNA pathway to adopt an antiviral role in the somatic tissues (Morazzani, et al., 2012), while other *piwi* duplicates maintain the ancestral function (Miesen, et al., 2015; Miesen, et al., 2016).

The role of RNAi pathways in mediating inter-genomic (host-virus) and intra-genomic (host-TE, segregation distortion) (Ferree & Barbash, 2007) conflict suggests that they may be a hotspot of adaptive protein evolution. This has been well studied in *Drosophila*, where RNAi pathway genes show elevated rates of adaptive protein evolution (Obbard, et al., 2006; Obbard, et al., 2009), signatures of selective sweeps (Obbard, et al., 2011; Kolaczkowski, et al., 2011; Lewis, et al., 2016), and sites with elevated protein evolution across the *Drosophila* phylogeny (Vermaak, et al., 2005; Heger & Ponting, 2007; Kolaczkowski, et al., 2011). For example, a comparison of the antiviral RNAi genes *AGO2*, *Dcr-2*, and *r2d2* to their miRNA functional counterparts with no known role in conflict (the paralogs *AGO1*, *Dcr-1*, and *loqs*) shows a striking difference in rates of protein evolution, as well as a greater rate of adaptive amino-acid substitution (Obbard, et al., 2006). In addition, evolutionary rates of piRNA pathway genes involved in transcriptional silencing are elevated and highly correlated with other piRNA pathway genes across the *Drosophila* phylogeny (Blumenstiel, et al., 2016).

Although some antiviral and anti-TE RNAi pathway genes clearly display elevated rates of adaptive protein evolution in *Drosophila*, the generality of this pattern remains to be elucidated. Here we apply both traditional McDonald-Kreitman (McDonald & Kreitman, 1991) and SnIPRE-style (Eilertson, et al., 2012) analyses, and selective sweep-based analyses (Nielsen, et al., 2005; Pavlidis, et al., 2013) to publicly-available genome-scale data from 6 insects and 2 nematodes. By combining estimates across species, we investigate the specific RNAi subpathways that may be the target of elevated positive selection. This allows us to estimate the rates of adaptation across species, thereby improving single gene estimates and allowing us to identify genes that are undergoing parallel adaptation across the taxa analysed. Finally, we summarise the evidence for recently completed and ongoing selective sweeps in RNAi genes across these eight taxa. We conclude that rapid evolution of RNAi genes is a general phenomenon in these invertebrates, although evidence for recent sweeps is highly contingent on the focal species.

Materials and Methods

Selection of genes for analysis

Genes implicated in the RNAi pathway of either *Drosophila melanogaster* or *Caenorhabditis elegans* were used to find homologues in six insects and two nematode species (Table S1, Table S2). For the six insect species, these were further classified as miRNA, piRNA, siRNA, or viRNA. Although the viRNA pathway is not widely regarded as separate from siRNA, we make this distinction based on the hypothesis that these genes may be evolving adaptively in response to viruses, as these genes have direct experimental evidence of an antiviral role in *D. melanogaster*. We also split the piRNA pathway genes among three functional categories: post-transcriptional silencing effectors, transcriptional silencing effectors, and biogenesis machinery. A gene was considered a biogenesis factor if piRNA levels decrease upon loss-of-function, an effector if piRNA pathway function is compromised without reducing piRNA levels, and a transcriptional silencing effector if the effector is involved in transcriptional silencing (Table S1). Finally, we selected 65 piRNA genes in *D. melanogaster* with known tissue-specificity to calculate rates of adaptation in the germline versus the somatic follicle cells (Table S3). This gene list contains the core of the piRNA pathway and genes independently validated in two of the three recent screens for piRNA pathway constituents (Handler, et al., 2013; Czech, et al., 2013; Muerdter, et al., 2013).

Homologs of the *D. melanogaster* and *C. elegans* genes were identified using a two-step process. First, a hidden Markov Model (HMMer) (Eddy, 2008)) was used to find best reciprocal best-hits for a gene of interest using predicted protein sets (if available) or UniProtKB. If no hit was found, then Exonerate was used to identify unannotated homologues in the genome using the model 'protein2genome' (Slater & Birney, 2005). If exonerate was unable to model a homologue, then this gene was classified as missing, either due to gene loss or an incomplete genome assembly. We defined genes as duplicates (paralogues) if multiple regions of a genome shared a best hit to a reference gene, and these regions showed substantial sequence divergence between them (i.e. they were not obviously a mis-assembly duplicate or allelic). Because of the large divergence times between insects and nematodes and the complexity of RNAi pathways in nematodes, and hence the associated difficulty in assigning homology in the two nematodes, we restricted our gene-level analyses to only the insect species.

Population genomic data

We utilised previously published population genomic data for *Drosophila melanogaster* (Lack et al, 2015)*, Drosophila pseudoobscura* (Pseudobase) (McGaugh, et al., 2012)*, Anopheles gambiae* (The *Anopheles gambiae* 1000 Genomes Consortium (2014): Ag1000G phase 1 AR2 data release. MalariaGEN.)*, Heliconius melpomene* (Kronforst, et al., 2013)*, Bombyx mandarina* (Xia, et al., 2009)*, Apis mellifera* (Harpur, et al., 2014)*, Pristionchus pacificus* (Rödelsperger, et al., 2014)*,* and *Caenorhabditis briggsae* (Thomas, et al., 2015) for our analyses (Table S4). For both *Drosophila* species, we used previously-published haplotype data (haploid sequencing of *D. melanogaster*, inbred lines of *D. pseudoobscura*). For the other taxa we obtained raw sequencing reads from EBI ENA (identifiers provided in Table S4) and mapped them to the most recent reference genome for each species using Bowtie2 (Langmead & Salzberg, 2012) with default settings. We used GATK's HaplotypeCaller on each individual separately (DePristo, et al., 2011) to call variants in a 200 kb region surrounding each gene of interest. For high coverage datasets (*A. mellifera, H. melpomene, C. briggsae, A. gambiae,* and *P. pacificus*) we excluded sites with a read depth lower than 5, but we reduced this threshold to 2 for the low-coverage *B. mandarina*. After mapping and filtering sites we created two randomly resolved pseudohaplotype sequences per individual (i.e. without any linkage information) from the sites that remained, and these were used for downstream analyses (none of which depend on linkage information). Only one haplotype was sampled from each *C. briggsae* and *P. pacificus* individual as the sequenced individuals were reported to be highly homozygous. In *H. melpomene*, we occasionally observed long stretches of high divergence shared by multiple individuals. We assumed these to be possible cases of either contamination, inversions that have recently risen to a high frequency, or introgression (Pardo-Diaz, et al., 2012), and removed these haplotypes.

To calculate divergence between genes, and to polarise mutations for sweep analyses, we used the outgroup species *Drosophila simulans, Drosophila miranda, Heliconius hecale, Bombyx huttoni, Anopheles christyi* and *Anopheles melas, Apis cerana, Caenorhabditis nigoni,* and *Pristionchus exspectatus*, respectively (Table S4). Outgroups were chosen based on their divergence from the ingroup species (*ca.* 1-10% divergence of all sites) and on the availability of genomic data. For *A. gambiae* we tested outgroups with low (*An. melas*) and high (*An. christyi*) divergence times, as most *Anopheles* species are too close or too divergent to provide a robust outgroup for MK tests (Obbard, et al., 2007), and our results remain qualitatively the same for both outgroups (*A. melas* used for the presented analyses). For *D. simulans* (FlyBase, r2.02)*, D. miranda* (Pseudobase, MSH22 strain)*, A. melas* (VectorBase, CM1001059 strain, AmelC1 assembly)*, A. christyi* (VectorBase, ACHKN1017 strain, AchrA1 assembly)*, B. huttoni* (Sackton, et al., 2014) (BioProject PRJNA198873), and *P. exspectatus* (WormBase, Bioproject PRJEB6009), the outgroup reference assemblies were publicly available and used as provided. However, the *Caenorhabditis nigoni* reference assembly sequence (caenorhabditis.bio.ed.ac.uk/home/download) is contaminated with the more divergent nematode *Caenorhabditis afra* (Thomas, et al., 2015), and *Caenorhabditis nigoni* is the only current suitable outgroup for C. briggsae. We therefore applied a sliding window across the alignments between *C. nigoni* and *C. afra*, and excluded regions that were greater than 6 standard deviations from the mean divergence. Published reference assemblies were not available for *Apis cerana* and *Heliconius hecale*. To generate outgroup sequences for these species we iteratively remapped reads (*H. hecale:* ERR260306; *A. cerana:* SRR957079) to the respective *Apis mellifera* and *Heliconius melpomene* references, each time updating the previous reference with homozygous nonreference calls. These reads were mapped with Bowtie2 and then remapped with the divergent alignment software, Stampy (Lunter & Goodson, 2011). Homozygous nonreference calls (enriched for sites divergent between the ingroup and outgroup) were made with GATK's HaplotypeCaller, with the heterozygosity parameter set to the expected divergence between species. Such sequences will not perfectly reflect the true outgroup sequence, and are expected to be biased toward the ingroup, downwardly biasing estimates of divergence in high-divergence regions. However, we confirmed that this approach works well by iteratively mapping *D. simulans* to *D. melanogaster*, and comparing the result with the known *D. simulans* assemblies (K_S= 0.10 for iterative mapping vs K_S=0.12 for the true assembly), and while bias probably remains, it is unlikely to spuriously elevate the inferred rates of one class of genes relative to the other. More generally, our approach to mapping, filtering, and variant calling may be prone to such biases, but they are unlikely to differentially affect gene classes of different function.

For MK analyses, target sequences were aligned as amino acids using MUSCLE (Edgar, 2004), and then each examined by eye to remove putative mis-alignments. Within-species data was aligned first, and then a consensus sequence of this alignment used to align against the outgroup sequence. Synonymous and nonsynonymous substitutions between species were inferred using codeml from the PAML package using the YN00 model (Yang & Nielsen, 2000), which estimates substitution rates using an approximation to maximum likelihood methods, while accounting for base composition differences between codon positions and differences in transition/transversion rates.

Rates of adaptive evolution by pathway

To estimate the rate of adaptive protein evolution in different functional classes of gene, and to test for differences in rate between classes, we used two different approaches derived from the McDonald-Kreitman test ('MK framework') (McDonald & Kreitman, 1991). The MK framework combines polymorphism and divergence data from putatively neutral (synonymous) and potentially selected (nonsynonymous) variants to infer an excess of nonsynonymous fixations —beyond that expected under model of neutrality and constraint—that can be attributed to positive selection. We first used an explicit population-genetic model to estimate the number of adaptive nonsynonymous substitutions per site (DFE-alpha) (Eyre-Walker & Keightley, 2009). This approach has the advantage that it provides direct estimates of the parameters of interest, and explicitly models changes in population size and the distribution of deleterious fitness effects, which might otherwise bias estimates (Keightley & Eyre-Walker, 2007; Eyre-Walker & Keightley, 2009). However, as currently implemented, this method does not allow data to be directly combined between species. Therefore, to obtain more precise homologue- and pathway-based estimates we combined per-gene point estimates from DFE-alpha using a linear mixed model (including their estimated uncertainty; i.e. a meta-analysis). Our second approach used an extension of the SnIPRE model (Eilertson, et al., 2012), which re-frames the MK test as a linear model in which polymorphism and substitution counts are predicted by synonymous or nonsynonymous state. Although this model does not explicitly consider the same underlying population-genetic processes, it does permit a straightforward extension to natively include gene, homologue, pathway, and host species as predictors in the model, and therefore provides a direct test of the questions of interest (although at a cost of potentially less accurate or arbitrarily-scaled parameter estimates). We have re-implemented the SnIPRE model using the Bayesian Generalised Linear Mixed Modelling R package MCMCglmm (Hadfield, 2010), and the code is provided in S1 text.

DFE-alpha analyses

DFE-alpha (Eyre-Walker & Keightley, 2009) infers ω_A (the number of adaptive nonsynonymous substitutions per nonsynonymous site, relative to the number of synonymous substitutions per synonymous site), while simultaneously modelling the distribution of deleterious fitness effects and population size changes (Keightley & Eyre-Walker, 2007; Eyre-Walker & Keightley, 2009). The ω_A statistic is closely related to the more widely reported α statistic (the proportion of nonsynonymous substitutions that are adaptive (Charlesworth, 1994; Fay, et al., 2001; Smith & Eyre-Walker, 2002; Bierne & Eyre-Walker, 2004; Welch, 2006), but differs in that ω_A is expected to be less dependent on effective population size and therefore better for cross-species comparisons (because the denominator, dS, should be less affected by the efficacy of selection, and thus effective population size (Gossmann, et al., 2010; Gossmann, et al., 2012; Kousathanas, et al., 2014). DFE-alpha utilises the observed site frequency spectrum (SFS) for putatively neutral synonymous sites and potentially selected nonsynonymous sites, and maximises the likelihood of observing these spectra given the distribution of deleterious fitness effects (DFE) for nonsynonymous sites and a step-change in effective population size (Eyre-Walker & Keightley, 2009). The 'excess' nonsynonymous divergence attributable to adaptive substitution is then inferred, given the maximum likelihood estimate of the DFE and the observed divergence (Eyre-Walker & Keightley, 2009). We inferred ω_A for: (i) each RNAi gene and each position-matched control gene (i.e. those with no known RNAi-pathway role falling within the same 200 Kbp interval); (ii) each RNAi subpathway and their set of control genes, and; (iii) all RNAi pathway genes together, by pooling polymorphism and divergence data across genes within classes. We then compared this grouped polymorphism and divergence data in pathways of interest against control genes. We estimated the parameters of the nominal change in population size (the relative population size change parameter N_2, and the time of the population size change, t_2) for all genes treated together within species, and then fixed these estimates for pathway and individual gene estimates. Conditional on this species-wide estimate of demographic history, the DFE was estimated separately for RNAi and control genes. We obtained confidence intervals for estimates of α and ω_A by bootstrapping genes within classes (1000 draws), and we tested for differences in rate between gene classes by randomly permuting genes 1000 times between classes. To test for differences in the DFE between RNAi and control genes we performed a likelihood ratio test between a model in which parameters of the DFE were estimated for all genes together, and one in which we allowed the DFE parameters to be estimated separately for RNAi and control genes.

Pooling polymorphism and divergence data across genes allows calculation of pathway-specific ω_A within a species, but cannot readily give cross-species estimates. Therefore, we also calculated ω_A for individual genes in each species, and analysed these estimates across species. In general, such estimates are extremely poor unless samples sizes are extremely large (e.g. hundreds of alleles are sampled, or genes are very large) (Keightley & Eyre-Walker, 2010). However, if the selective pressure acting on genes is consistent across species, for example as is assumed by many phylogenetic approaches to detecting selection (Yang, 2007), we can acquire more accurate estimates of the relative rate of adaptive evolution by combining information across species. We therefore used a formal meta-analytic approach to combine small-group and single-gene estimates across species using MCMCglmm (Hadfield, 2010), by constructing linear mixed models. These models were used to estimate average gene-level ω_A of various pathways and homologues, and variation among gene-level ω_A estimates.

The first three models took the same form, only distinguished by the pathways among which genes were divided. In Model 1A the genes were classified as either 'control' or 'RNAi', in Model 1B the RNAi class was expanded into four levels: 'miRNA', 'siRNA', 'viRNA', and 'piRNA' and in Model 1C the piRNA class was further split into three functional categories: 'effectors of transcriptional silencing', 'effectors of post-transcriptional silencing', and 'biogenesis factors'. The model for the estimate of ω_A (i.e. $\widehat{\omega}_A$) for homologue *k* in gene class *l* in species *m* had the form:

$$\widehat{\omega}_{A:klm} = \beta_0 + \beta_{Class:l} + u_{Organism:m} + m_{klm} + \varepsilon_{klm} \qquad [1]$$

where β_0 is the intercept, $\beta_{Class:l}$ is a fixed effect associated with gene class *l*, $u_{Organism:m}$ is a random effect associated with species *m*, m_{klm} is the sampling error associated with each estimate, and ε_{klm} is the between observation error after accounting for measurement error, which was allowed to vary by gene class (i.e. pathway). The variance of the sampling errors was obtained by bootstrapping genes by codon, and this sampling error variance was fixed at that value in the analysis. All species effects were assumed to come from a single normal distribution but the errors were assumed to come from independent normal distributions with different variances for each gene class.

Model 2 extended Model 1 by including homologue as a random effect ($u_{Hom:kl}$) in order to identify homologues with elevated adaptation across lineages:

$$\widehat{\omega}_{A:klm} = \beta_0 + \beta_{Class:l} + u_{Organism:m} + u_{Hom:kl} + \ m_{klm} + \varepsilon_{klm} \qquad [2]$$

where the homologue effects were assumed to come from independent normal distributions with different variances for each gene class. In this model the cross-species average ω_A for a homologue *k* in gene class *l* is given by $\overline{\omega}_{A:kl} = \ \beta_0 + u_{Class:l} + u_{Hom:kl}$. However, if genes are misclassified with respect to the gene class they belong, then $\overline{\omega}_{A:kl}$ is likely to biased in general, and particularly so for misclassified genes. An arguably more conservative approach is to only use information from homologous genes to estimate the cross-species (i.e. remove the class effects from the model; this approach is provided as Model 2B in S1 text) and have $\overline{\omega}_{A:kl} = \ \beta_0 + u_{Hom:kl}$. See S1 text for R code and a full description of the models used.

SnIPRE-like analysis

The meta-analytic approach to cross-species analysis above has the advantage of utilising DFE-alpha estimates that are inferred under an explicit population-genetic model. However, it has the disadvantage that it conditions on point estimates from a model, rather than using the available data directly. We have therefore taken advantage of the Poisson linear mixed model approach to MK analyses 'SnIPRE' proposed by Eilertson et al. (2012), which models the counts of mutations in four classes: synonymous within-species polymorphisms, nonsynonymous within-species polymorphisms, between-species synonymous differences (divergence) and between-species nonsynonymous differences. By fitting 'nonsynonymous' and 'divergent' as main effects, selection can be inferred from their interaction, which records the excess contribution of nonsynonymous mutations to between-species divergence. This excess can be assessed at the level of individual genes (by treating gene identity as a random effect) or can be expressed as a function of other fixed or random effects such as gene class and species. Although this approach does not directly provide parameter estimates that are interpretable in simple population-genetic terms, such as ω_A, it has the advantage of extending naturally to provide comparisons between species and gene classes while still using raw count data directly. Here we combine polymorphism and divergence data from several species to test whether RNAi genes have higher rates of adaptive substitution than our set of control genes, whether these rates vary between different subclasses of RNAi gene, and whether these rates vary between different homologues. We fit these models with the R package MCMCglmm (Hadfield, 2010) and the code is provided in the S1 text. In their single-species and single-class analysis Eilertson et al. (2012) used the generalised linear mixed model with the fixed effect part of the model as:

$$log(\mu_{ijk}) = \beta_0 + \beta^N i + \beta^D j + \beta^{ND} ij + \beta_{length} x_{ik} \quad [3]$$

where μ_{ik} is the expected number of mutations in gene k in one of the four classes indexed by *i* = 0,1 and *j* = 0,1 where *i* = 1 indicates nonsynonymous (*N*) and *k* = 1 divergent (*D*). This model estimates the intercept β_0 (the density of synonymous polymorphisms), β^N (the genome-wide difference between a mutation being nonsynonymous versus synonymous), β^D (the genome-wide difference between a mutation being a substitution versus a polymorphism), and β^{ND} (the interaction effect describing any genome-wide excess or dearth of nonsynonymous substitutions). x_{ik} is the logarithm of the number of sites in gene *k* where a synonymous (*i* = 0) or a nonsynonymous (*i* = 1) mutation could occur and the fixed effect β_{length} models how the number of observed mutations changes as a function of the number of sites. Eilertson et al. (2012) also fitted a random effect structure that models between-gene mutation patterns after accounting for the fixed effects:

$$log(\mu_{ijk}) = \beta_0 + \beta^N i + \beta^D j + \beta^{ND} ij + \beta_{length} x_{ik} + \varepsilon_k + \varepsilon_k^N i + \varepsilon_k^D j + \varepsilon_k^{ND} ij \; [4]$$

where the additional terms denoted ε are the gene-specific random deviations from each of the first four fixed effect terms described above. The four gene-specific random deviations were assumed to come from a multivariate normal distribution with estimated (co)variance matrix. Eilertson et al. (2012) define the selection effect of gene *k* as $\beta^{ND} + \varepsilon_k^{DG}$, where a positive effect is evidence for positive selection, and (in Bayesian terms) the posterior probability that the effect exceeds zero can be directly assessed.

Here we extend the *SnIPRE*-like model of Eilertson et al. (2012) to accommodate multiple species and to allow the evolutionary parameters to differ between different classes of gene. To this end we allowed the four fixed effects to vary by species and by gene class (control, piRNA, siRNA, miRNA and viRNA) to give the fixed effect model:

$$\beta_0 + \beta^N i + \beta^D j + \beta^{ND} ij + \beta_{length} x_{iklm} + \beta_{Class:l} + \beta^N_{Class:l} i + \beta^D_{Class:l} j + \beta^{ND}_{Class:l} ij + \beta_{Organism:m} + \beta^N_{Organism:m} i + \beta^D_{Organism:m} j + \beta^{ND}_{Organism:m} ij \quad [5]$$

From this, we calculated the estimated selection effect for a specific pathway as $\beta^{ND} + \beta^{ND}_{Class:l}$. The random effect portion of the model included homologue-specific effects and gene-specific effects and had the form

$$u_{Hom:k} + u^N_{Hom:k} i + u^D_{Hom:k} j + u^{ND}_{Hom:k} ij + \varepsilon_{klm} + \varepsilon^N_{klm} i + \varepsilon^D_{klm} j + \varepsilon^{ND}_{klm} ij \qquad [6]$$

In addition to the four gene effects, the four homologue effects were also assumed to come from a multivariate normal distribution with estimated (co)variance matrix. We used this model to calculate the selection effect for homologue *k* in gene class *l* as $\beta^{ND} + \beta^{ND}_{Class:l} + u^{ND}_{Hom:kl}$ and each gene as $\beta^{ND} + \beta^{ND}_{Class:l} + u^{ND}_{Hom:klm} + \varepsilon^{ND}_{klm}$. We estimated β_{length} rather than fixing it at one, as in Eilertson et al. (2012), although the posterior mean of β_{length} was close to one, supporting the assumption of Eilertson et al. (2012). In addition, we also fitted the SnIPRE model without assuming genes belong to known pathways, analogous to model 2. The code to fit these models is provided in the S1 text.

Selective sweep analysis

The recent spread of a positively selected allele leaves characteristic patterns of diversity and allele frequencies in the genomic region surrounding the selected site, and these can be used to detect recent adaptive substitutions (e.g. Maynard Smith & Haigh, 1974; Barton, 1998; Nielsen, et al., 2005). We used SweeD (Pavlidis, et al., 2013; derived from Sweepfinder, Nielsen, et al., 2005) to search for evidence of recent selective sweeps in the regions surrounding RNAi genes. The algorithm scans the genome and at a user-defined interval calculates the composite likelihood of the observed site frequency spectrum (SFS) under a model of a selective sweep centred on that site, versus a standard neutral model. The ratio of the two composite likelihoods (CLR) is then used as a test statistic, with significance assessed by coalescent simulation (see Figure S1 and Text S2). We used this method to scan 200 kb (or less if the reference genome contig was less than 200 kb) surrounding each gene of interest in each species. For each focal region, we polarised the SFS by parsimony between the outgroup reference genome and the ingroup consensus sequence, which we aligned with LastZ ungapped alignment (Harris, 2007). We did not assume an ancestral state for fixed differences that were invariant in our ingroup (i.e. these sites were folded). This will make the analysis more robust to possible errors during contig alignment, because misalignment would manifest itself as regions of increased divergence between species. We included invariant sites in the analysis, as a characteristic signature of a recent sweep is a lack of diversity, and so including invariant sites in Sweepfinder analyses can greatly improve statistical power (Nielsen, et al., 2005). This comes with a risk of increased false positives (Huber et al, 2016), but including these sites should not differentially affect RNAi and control genes, unless there is a consistent difference in mutation rates between these two classes of genes. The SweeD analysis provides CLR values for equidistant points across the genome, with CLR values forming a "peak" in areas with high support for a sweep. To assess whether RNAi genes have experienced more sweeps than control genes in 6 of our 8 species (*B. mandarina* and *P. pacificus* were not tested because the published genome assemblies are unannotated), we counted the number of RNAi and control genes that overlapped significant peaks in the CLR statistic (based on the significance threshold provided by coalescent simulation, Figure S1, S2 text). If consecutive peaks occurred within 1 kb of each other, we classified them as a single broad peak, such that the contig was split into "sweep-positive" and "sweep-negative" areas. We then classified all genes along the contig as to whether they overlapped a "sweep-positive" area or not, and whether or not they were an RNAi gene. We used a binomial test to assess whether RNAi or control classes had more sweep-positive genes than expected given the summed gene length for each class.

To test whether sweeps were enriched in any particular subpathway, we normalised the maximum CLR statistic in a gene by the expected significance threshold from coalescent simulations and modelled these values ($\widetilde{CLR}$) using the following linear mixed model:

$$\widetilde{CLR}_{klm} = \beta_0 + \beta_{Class:l} + u_{Organism:m} + \varepsilon_{klm} \qquad [7]$$

Here, $\beta_{Class:l}$ is a fixed effect for the pathway each gene is assigned (miRNA, siRNA, piRNA or viRNA), $u_{Organism:m}$ is a random effect for species *m* and ε_{klm} is the error term.

In the four organisms for which we have haplotype information (*D. melanogaster, D. pseudoobscura, P. pacificus, C. briggsae*), we additionally tested for ongoing or soft sweeps using the haplotype-based nSL statistic (Ferrer-Admetlla, et al., 2014). The nSL statistic is similar to the more widely used iHS statistic (Voight, et al., 2006), except that distance is measured in polymorphic sites rather than the genetic map distance (Ferrer-Admetlla, et al., 2014). This genome scan calculates the average number of consecutive polymorphisms associated with either the ancestral or derived allele at each polymorphic site along the contig across all pairwise comparisons. Areas with long range linkage disequilibrium will therefore be identified through SNPs with extreme nSL values.

Results

Evidence of genome-wide adaptive substitution in insects, but not nematodes

The position-matched 'control' genes (that lack RNAi-related function) included in our analyses allowed us to estimate the average genome-wide rate of adaptation, assuming that proximity to RNAi gene has no effect on their rate of adaptive evolution. Our analysis broadly agrees with previous ones, suggesting a substantial fraction of amino-acid substitution is adaptive across insect species (Figure 1). All insect species shared similar estimates (ω_A from 0.02 to 0.05) except for *D. pseudoobscura,* which exhibited an extremely high ω_A value of 0.16 [0.05,0.32] (95% bootstrap confidence interval) adaptive nonsynonymous substitutions per synonymous substitution per site. Although we only sampled two nematode lineages, it is notable that both ω_A estimates were negative (*C. briggsae*: -0.20 [-0.25, -0.15]; *P. pacificus:* -0.24 [-0.27, -0.21]. This is consistent with the previously noted high ratio of nonsynonymous to synonymous polymorphism (π_A/π_S) ratio in these species, and perhaps suggests population structure and local adaptation (Rödelsperger, et al., 2014; Thomas, et al., 2015). We also calculated α, or the proportion of adaptive substitutions for each species, which reflect the same patterns observed for ω_A (Figure S2).

The cross-species SnIPRE-like model provides a formal comparison of adaptive divergence in the insect species. The structure of the model forces comparison relative to one species, for which we chose *D. melanogaster*. *Anopheles gambiae* and *Bombyx mandarina* had levels of putatively adaptive nonsynonymous divergence that were indistinguishable from those of *D. melanogaster* (MCMCp = 0.489 and MCMCp=0.616, respectively). Consistent with the DFE-alpha estimates of $\omega_{A,}$ *A. mellifera* and *H. melpomene* had significantly less adaptive nonsynonymous divergence than *D. melanogaster* (MCMCp = 0.04 and MCMCp < $3\text{x}10^{-4}$, respectively), whereas *D. pseudoobscura* had an increased excess of nonsynonymous divergence (MCMCp = 0.0005). Other species-specific SnIPRE parameters can be found in S1 text.

RNAi genes consistently display more adaptive protein substitution than other genes

For each focal species we estimated the distribution of fitness effects of new mutations using DFE-alpha for RNAi pathway and non-RNAi ('control') genes, by pooling polymorphism and divergence data for each gene class. We fitted two models, one in which RNAi and control genes share a single DFE, and second in which each class of gene had a separate DFE. We then compared these models using a likelihood ratio test. In *D. melanogaster, D. pseudoobscura*, *H. melpomene, A. mellifera,* and *C. briggsae*, models in which control and RNAi genes have separate DFE parameters fitted the data significantly better than a model in which the two classes share a single DFE (Figure 1). Although there is no clear or universal trend, the DFE of control genes generally seemed slightly shifted towards more deleterious mutations than RNAi genes. For example, in most lineages (not *D. pseudoobscura* or *A. gambiae*), the estimated DFEs had a higher proportion of strongly deleterious mutations in control genes than RNAi genes, which suggests less constraint in RNAi genes. However, the overall shape of the DFE is quite different between species, either indicating that in these species gene function may play a smaller role than other factors in patterns of polymorphism, or that the DFE is estimated with low precision.

We then compared rates of adaptive amino acid substitution in RNAi genes to those in the non-RNAi control genes in each lineage, by pooling polymorphism and divergence data for the two classes as input to DFE-alpha (Figure 1). In every species tested, the point-estimate of class-wide ω_A was greater in RNAi genes than control genes. Although the effect was often small, the difference was individually significant in *D. melanogaster, D. pseudoobscura, H. melpomene,* and *P. pacificus.* To quantify the overall difference, we analysed individual gene estimates of ω_A in a linear mixed model framework (i.e. a meta-analysis) to estimate across-species rates of adaptive evolution in control and RNAi genes (model 1 in S1 text, Figure 1). We found the cross-species ω_A was significantly greater for RNAi genes than control genes, estimated as ω_A = 0.062 [0.049, 0.078] (95% Highest posterior density) versus ω_A = 0.01 [0.0009, 0.019] ($p <$ 0.001). In addition, the residual gene-level variance was also much greater (MCMCp <0.001) for RNAi genes (0.0037, [0.0022, 0.0051]) than control genes (0.0003, [0.0001, 0.0004]), implying that ω_A is more variable in this class than among genes in general and consistent with a subset of RNAi genes or pathways undergoing extreme rates of adaptive amino acid substitution (Figure 1).

Adaptive rates are high in piRNA and viRNA pathways

The higher rate of adaptive substitution seen in RNAi genes as a whole could result from slightly elevated positive selection across all components, or to a subset of the genes or pathway being substantially elevated. The higher gene-level variance seen in RNAi genes (above) suggests the latter, and to test this we pooled polymorphism and divergence data by sub-pathway for each insect species to calculate rates of adaptation in miRNA, siRNA, viRNA (i.e. confirmed antiviral siRNA in *D. melanogaster*), and piRNA pathways (Figure 2). In each species, the piRNA pathway exhibited a significantly greater rate of adaptive amino acid substitution than control genes, and miRNA pathway genes showed similar rates to control genes. Rates of adaptation for the siRNA (both endo-siRNA and viRNA) pathway were greater in only a subset of lineages. The magnitude of rates and proportion of adaptive lineages increased upon removing endo-siRNA genes and restricting the analysis to viRNA genes only. For all subsequent analyses, we analysed these pathways separately to test the hypothesis that the core antiviral RNAi genes have elevated rates of adaptive evolution.

To formalise the effect of pathway (miRNA, piRNA, non-antiviral endo-siRNA, viRNA) while accounting for variability in adaptation across species (model 2 in S1 text, Figure 2), we performed a meta-analysis of ω_A estimates in individual genes from DFE-alpha, fitting pathway as a fixed effect. The piRNA, viRNA, and endo-siRNA pathways were each significantly different from control genes (control ω_A =0.01 [0.002,0.018]; piRNA MCMCp < 0.001; viRNA MCMCp = 0.002; siRNA MCMCp = 0.004; for MCMCp value calculation, see the S1 text), with cross-species estimates of ω_A of 0.08 [0.06,0.10], 0.18 [0.06, 0.30] and 0.03 [0.01,0.05], respectively. The viRNA pathway ω_A estimate was not significantly greater than the piRNA pathway (MCMCp = 0.07), but was greater than the endo-siRNA pathway (MCMCp = 0.01), and the miRNA pathway (MCMCp < 0.001). The ω_A estimate for the piRNA pathway was significantly greater than the endo-siRNA (MCMCp = 0.002) and the miRNA pathways (MCMCp < 0.001). Consistent with our analysis of pooled polymorphism and divergence data, the rate of adaptive evolution in the miRNA pathway (ω_A = 0.01 [-0.001, 0.02]; MCMCp=0.09) was not significantly different from control genes. Our linear models included pathway-specific error variances, which were lower for control genes (3 [2,4] x10^{-4}) and miRNA pathway genes (7 [2,12] x10^{-4}) than for endo-siRNA (13 [4,22] x10^{-4}), piRNA (66 [37,97] x10^{-4}), and viRNA pathway genes (0.04 [0.007, 0.86]), consistent with a great variation in adaptive rates in these pathways.

We repeated the subpathway-level analysis using a SnIPRE-like model (Eilertson, et al., 2012) to estimate the average selection effect within subpathways across organisms, without making any explicit assumptions about the DFE. Although SnIPRE can be used to provide estimates of population genetic parameters, we limit our discussion to the “selection effect” statistic, where negative values are consistent with constraint and positive values with adaptive protein evolution, and magnitude reflects the strength of positive or negative selection. Consistent with our analysis of DFE-alpha estimates, the SnIPRE model identified a mean positive selective effect estimated across species (selective effect=0.25 [0.02, 0.46] 95% HPD interval, MCMCp = 0.03), with large variance among genes (Figure 3). Again, viRNA, endo-siRNA, and piRNA pathway-level selection effects were significantly elevated compared to control genes (viRNA: 1.10 [0.63, 1.57] MCMCp < 5x10^{-4}, non-antiviral siRNA: 0.96 [0.44, 1.52] MCMCp = 0.02, piRNA: 0.63 [0.44, 0.84] MCMCp < 3x10^{-4}), with the viRNA pathway exhibiting a significantly larger effect than the piRNA (MCMCp = 0.006), but not the endo-siRNA (MCMCp = 0.66). In agreement with the DFE alpha analysis, the miRNA pathway was not significantly different from control genes (MCMCp = 0.07), and had a selection effect of 0.53 [0.20, 0.86].

Adaptation is elevated in all major piRNA pathway functions, but is most enriched in transcriptional silencing

Rapid adaptation in *Drosophila* piRNA pathway genes has been hypothesized to be the result of fluctuating selection for increased TE defence and decreased off-target genic silencing (Blumenstiel et al 2016). A prediction of this hypothesis is that genes involved in transcriptional silencing would be under increased positive selection. We tested this prediction by further dividing the piRNA pathway into effectors (e.g. PIWIs), biogenesis factors (e.g. adapter proteins), and transcriptional silencing factors, and using single-gene polymorphism and divergence data to estimate ω_A and the selection effect for each piRNA functional category (Model 3). We found all piRNA functional groups are significantly greater than control genes (MCMCp < 0.001) (Figure 2C), and that transcriptional silencing genes (ω_A = 0.16 [0.08-0.25]) have greater adaptive rates than effectors (MCMCp = 0.04, ω_A = 0.08 [0.04-0.13]) and biogenesis factors (MCMCp = 0.03, ω_A = 0.08 [0.05-0.11]). This result holds when excluding *Drosophila* transcriptional silencing factors *rhino*, *deadlock*, and *cutoff*, which are products of recent gene duplication or *de novo* formation (Figure S3), and may not have evolutionary rates that are directly comparable to other genes.

We also estimated the average selection effect for each functional process of the piRNA pathway using the SnIPRE approach. Similar to the DFE-alpha meta-analysis, we find that all piRNA functional categories have elevated positive selection relative to control genes (biogenesis: MCMCp=0.018, effector: MCMCp=0.012, transcriptional silencing: MCMCp=0.0004), that transcriptional silencing factors had the largest average selection effect of 0.92 [0.58, 1.31], and that genes involved in transcriptional silencing were significantly greater than biogenesis factors (selection effect: 0.53, [0.29, 0.78], MCMCp = 0.027) (Figure 3B). In contrast to the DFE-alpha meta-analysis, however, genes involved in transcriptional silencing were not significantly greater than effector genes (0.78 [0.40, 1.19], MCMCp = 0.68), and pathway-level point estimates of these selection effects were much closer (Figure 2C, Figure 3B).

Individual genes in the piRNA and viRNA pathway show elevated adaptation

The higher overall rates of adaptive protein substitution seen in RNAi genes may result from the engagement of some genes in an evolutionary arms race (e.g. with viral suppressors of RNAi), a response to the selection imposed by the invasion of novel parasites (e.g. transposable elements), or a trade-off between the specificity and sensitivity of genome defense (Aravin, et al., 2007; Obbard, et al., 2006; Blumenstiel, et al., 2016). We used a linear mixed model to combine single-gene estimates of ω_A from DFE-alpha across multiple species to identify candidate arms race genes in the RNAi pathways, fitting subpathway as a fixed effect, with homologue and organism as random effects, and subpathway-specific error variances. We found little variation among genes in a subpathway after accounting for subpathway, and in most cases there was not enough information to differentiate individual genes from the pathway mean (Figure 4A). Although a model that accounts for pathway is statistically preferable if pathways are meaningful, any errors in assigning 'pathway' membership would introduce bias to the estimates for misclassified genes. We therefore also estimated homologue-specific effects in a model that excludes the subpathway effect (model 2B in the S1 text). This model finds significant evidence for positive selection in fewer genes (Figure S4A) including 13 of 22 piRNA genes, 2 of 3 viRNA genes, and no genes in the siRNA or miRNA pathway.

We also performed this homologue-level analyses using the SnIPRE approach. Similar to the DFE-alpha meta-analysis, we found very little information after accounting for subpathway (Figure 4B), resulting in low among-gene variation within RNAi subpathways. When we excluded subpathway effects, we found a similar result to the homologue-level DFE-alpha meta-analysis without subpathway, except fewer piRNA pathway genes are nominally significant (6/22 genes). Notably, *maelstrom, eggless, piwi* (including *aub*), *AGO2,* and *Dcr-2* were found to have significantly elevated positive selection across all four homologue-level analyses (i.e. with or without imposing a subpathway classification).

MK tests are commonly used to test for positive selection in individual genes. SnIPRE selection effects can be used to perform an analogous test for selection, except the approach can gain power by taking in the genome-wide distribution of polymorphism and divergence patterns by fitting gene as a random effect (Eilertson et al, 2010). We find that 36% of RNAi genes show nominally 'significant' positive selection. In contrast, only 5% of selection effects in control genes are significantly positive (Table S5). At the pathway level, 40% of piRNA genes, 44% of viRNA genes, 26% of non-antiviral siRNA pathway genes, and 25% of miRNA pathway genes have significantly positive selection effects (Table S5). No gene had positive selection effects in every lineage, although *armitage, capsuleen, cutoff, tudor, vasa, vretano,* and *Yb* homologs were identified in over half the lineages.

Selective sweeps are detectable across functional classes of RNAi genes

Recent positive selection is expected to leave a characteristic mark in the genome, including a SFS skewed towards low and high frequency alleles and a local reduction in polymorphism (Maynard Smith & Haigh, 1974; Barton, 1998; Nielsen, et al., 2005). As RNAi genes show elevated rates of adaptive evolution, we speculated that they may also exhibit more evidence of recent selective sweeps. Using SweeD, we found that many of the insect lineages do show evidence for sweeps in a subset of RNAi genes (Figure 5, Figures S5-S12). We tested whether RNAi genes have undergone more recent sweeps than surrounding genes by classifying nominally significant peaks as either occurring near (within 1 KB) an RNAi gene or not, and using a binomial test to determine whether more sweeps than expected occur in RNAi genes (given their length). In four of the six species tested (*D. melanogaster, D. pseudoobscura, A. mellifera,* and *A. gambiae*) there were significantly more detectable sweep signals in RNAi genes than in surrounding non-RNAi genes (*D. melanogaster p* = 0.0006; *A. mellifera p* = 0.015*; A. gambiae p* = 0.0001*; D. pseudoobscura p* = $7x10^{-5}$). However, we find no difference among subpathways in the frequency with which we detected recent sweeps. In addition, none of the genes exhibited a significant CLR peak across all organisms tested, although *spn-E* and *vig* display significant evidence of recent sweeps in five of the six insect lineages. It was notable that 34% of the variation in the per-gene maximum CLR test statistic was attributable to species, consistent with either sample size or demographic history playing a substantial role in our power to detect sweeps.

Sweep signatures were the most pronounced in *A. mellifera*, in both the CLR magnitude and breadth of the genomic region affected (Figure 5, Figure S10). These were associated with large regions devoid of any polymorphism, despite the high rate of recombination seen in honeybees (Beye, et al., 2006), which is expected to narrow the region affected by a nearby sweep. We also searched for evidence of haplotype structure, as would be expected during an ongoing or soft selective sweeps using the nSL statistic (data not shown). However, there were no strong signals in any of the RNAi genes for which we had haplotype information.

Discussion

Using both DFE-alpha and SnIPRE-like McDonald-Kreitman framework analyses we identify elevated rates of adaptive evolution in RNAi-pathway genes across six insects and two nematodes. In most species, the RNAi-pathway genes are also more likely to display evidence of a recent selective sweep. These results generalise the findings of previous analyses in *Drosophila*, and are consistent with these genes being engaged in an arms race across the invertebrates. Across species, we find that genes involved in the suppression of viruses and transposable elements show the highest rates of adaptive evolution, and those in the miRNA pathway the lowest (not significantly different from non-RNAi-genes). There is substantial variation in rates among RNAi but the antiviral genes *AGO2* and *Dcr-2* and the piwi-pathway genes *maelstrom, eggless, piwi*, *aub, armitage, capsuleen, cutoff, tudor, vasa, vretano, spn-E, vig* and *Yb* show consistently strong signatures of long-term and/or recent positive selection.

Identification of rapidly evolving pathways by DFE-alpha and SnIPRE

Estimating rates of adaptive protein evolution in an MK-framework (McDonald and Kreitman, 1991) can be biased by past population size changes and slightly deleterious mutations that segregate at low frequencies. We compare adaptive rates between different classes of RNAi genes, accounting for these biases by explicitly modelling the DFE and demographic history using DFE-alpha (Eyre-Walker and Keightley, 2009), or by modelling the genome-wide patterns of polymorphism and divergence with SnIPRE (Eilertson, et al., 2012). Most of the qualitative results of each of these analyses are in agreement, however, SnIPRE and DFE-alpha analyses disagree on the relative differences in the rate of adaptive evolution among subpathways. For example, the DFE-alpha meta-analysis provides low point estimates for the endo-siRNA and miRNA pathways relative to the piRNA and viRNA, but SnIPRE identifies the endo-siRNA selection effect as higher than the piRNA, and piRNA genes closer to the miRNA. This incongruence could reflect differences in the DFE between subpathways. For example, genes in the miRNA and endo-siRNA pathways are highly conserved and have low rates of protein evolution, while mechanisms of piRNA pathway function are surprisingly diverse across animals (e.g. Morazzani, et al., 2012; Sarkies, et al., 2015). These differences in constraint could lead to an underestimation of miRNA and endo-siRNA pathway adaptation and overestimation of piRNA adaptation in the DFE-alpha analyses, and indicate that estimating the DFE separately for each subpathway may improve estimates.

Adaptive protein evolution across species is enriched in specific functional pathways

We found large differences in rates of adaptative protein substitution between insects and nematodes, but less variation among insect species. In an analysis of variance, we find that species explained only 11% of the variation in gene-level estimates of ω_A. In contrast, gene and pathway explained 42% of the variation in gene-level ω_A estimates. The elevated rate and among-gene variation seen in piRNA and viRNA pathway genes across species could be caused by rapid adaptation in the same subset of genes in a pathway, or in a random selection of genes in a pathway. Homologue-level analysis of ω_A and selection effects (Figure 4, Figure S4) indicates it is probably both, as subsets of homologues within pathways show consistent evidence for elevated adaptive protein evolution, but homologous genes also exhibit high variances across species.

Potential Drivers of Adaptation in the viRNA pathway

It seems likely that the elevated rates of adaptive protein evolution we detect in the viRNA and piRNA pathways are a result of recurrent selection mediated by viruses and/or TEs. First, it is well established that defensive pathways show high rates of adaptive evolution, presumably as a consequence of antagonistic coevolution with parasites (Stenseth & Maynard Smith, 1984; Buckling & Rainey, 2002; Paterson, et al., 2010; Brockhurst, et al., 2014). For example, a recent analysis of virus-interacting proteins estimated that 30% of adaptive protein changes in mammals are driven by viruses (Enard, et al., 2016). Second, for the viRNA pathway genes at least, viral suppressors of RNAi are strong candidates to be the driving agent. Many RNA and DNA viruses of invertebrates are known to have proteins or structural RNAs which actively block RNAi function (Li, et al., 2002; Van Rij, et al., 2006; Nayak, et al., 2010; van Mierlo, et al., 2012; Bronkhorst, et al., 2014), and these can evolve rapidly and can be highly host-specific, consistent with an arms-race scenario (van Mierlo, et al., 2014). We find that *AGO2* and *Dcr-2* display consistently elevated rates of adaptive protein substitution across insect species, with additional limited evidence of elevated adaptation in *hen1*, all of which have previously been identified as targets of active suppression by viral proteins (viral suppressors of RNAi; VSRs) (Van Rij, et al., 2006; Vogler, et al., 2007; Nayak, et al., 2010; van Mierlo, et al., 2012; van Cleef, et al., 2014), lending credibility to the hypothesis that viruses may play a major role in driving the observed rapid evolution in RNAi genes.

Potential Drivers of Adaptation in the piRNA pathway

Whereas an arms-race between antiviral RNAi genes and viral suppressors of RNAi is intuitive, the observed rapid adaptive evolution of piRNA pathway genes is currently harder to explain. Similar to viruses, TEs are costly for their hosts and could in principle select for increased suppression (Charlesworth, et al., 1994). However, piRNA-generating clusters ostensibly provide an adaptive defence that can arise on much faster time scales than fixation of advantageous mutations, reminiscent of acquired immunity (Brennecke, et al., 2007; Khurana, et al., 2011; Mohn, et al., 2015; Han, et al., 2015). The adaptive response in piRNA genes could be mediated by at least three non-exclusive mechanisms: (i) direct piRNA pathway suppression by TEs or by off-target VSRs, (ii) recurrent “retuning” of piRNA machinery after a novel TE invasion (Lee and Langley et al, 2012; Yi et al, 2014), or (iii) fluctuating selection on the sensitivity to detect transposon sequences and specificity to exclude off-target genic silencing (i.e. the “genomic auto-immune hypothesis”) (Blumenstiel, et al., 2016). Besides the global de-repression of transposons upon invasion of the Penelope retroelement in *D. virilis* (Petrov, et al., 1995; Evgen'ev, et al., 1997; Rozhkov, et al., 2010; Blumenstiel, et al., 2016), there is limited evidence for (i), and the mechanism underlying this phenomenon still awaits elucidation. The latter two hypotheses are not mutually exclusive, and both posit that piRNA adaptation occurs in response to recurrent horizontal transfer of new TEs into the genome, a common occurrence in insects (Peccoud, et al., 2017). In (ii), the piRNA pathway evolves to optimise defence against the current suite of transposons, becoming “less adapted” for dealing with historic, obsolete ones. This would result in a Red Queen-like scenario, but instead of antagonistic coevolution with one parasite, the piRNA pathway must defend against a constant recycling of TE lineages. As the germline cells face a higher TE diversity than somatic tissues, this is broadly supported by our observation that piRNA pathway genes with primarily germline function (Handler, et al., 2013; Czech, et al., 2013; Muerdter, et al., 2013) have higher rates of adaptive protein evolution than those functioning in the somatic layer of cells surrounding the *Drosophila* ovary (Figure S13), . The genomic autoimmunity hypothesis (iii) goes further, and proposes piRNA pathway adaptation to TE invasions results in increased piRNA function and associated off-target genic effects, which are then selected against after the TE is supressed (Blumenstiel, et al., 2016). It could be argued that our analysis of adaptive rates in piRNA functions lends broad support for this, in that genes mediating transcriptional silencing show the greatest adaptive rates across species in the piRNA pathway, with additional evidence for rapid adaptation in biogenesis factors, whose rates are expected to be correlated with the transcriptional machinery (Blumenstiel, et al., 2016). However, our pathway-level and homologue-level analyses also find signals of elevated adaptation in effector genes, which have rates that covary to a lesser degree with other piRNA factors (Blumenstiel, et al., 2016). This does not refute the genomic autoimmunity hypothesis, but may suggest additional selective forces acting on the piRNA pathway independent of genes underlying a trade-off between sensitivity and specificity. Nevertheless, our results would also fit within the context of (ii), in a scenario where the transcriptional machinery has a greater evolutionary potential than the rest of the piRNA pathway.

Acknowledgements

We thank the authors of the previous studies who made their data publicly available. We thank Rob Ness and Peter Keightley for helpful advice on the analysis.

Funding

WHP is supported by the Darwin Trust of Edinburgh and this work in DJO’s laboratory was partly supported by a Wellcome Trust strategic award to the Centre for Immunity, Infection and Evolution (WT095831). JDH is supported by a Royal Society Research Fellowship (UF150696).

References

Alvarez-Garcia, Ines, and Eric A Miska. 2005. "MicroRNA functions in animal development and human disease." *Development (Cambridge, England)* 132 (21): 4653-62.

Aravin, A. A., G. J. Hannon, and J. Brennecke. 2007. "The Piwi-piRNA Pathway Provides an Adaptive Defense in the Transposon Arms Race." *Science* 318 (5851): 761-764.

Barton, N. H. 1998. "The effect of hitch-hiking on neutral genealogies." *Genetical Research* 72 (2): S0016672398003462.

Beye, Martin, Irene Gattermeier, Martin Hasselmann, Tanja Gempe, Morten Schioett, John F Baines, David Schlipalius, et al. 2006. "Exceptionally high levels of recombination across the honey bee genome." *Genome research* 16 (11): 1339-44.

Bierne, Nicolas, and Adam Eyre-Walker. 2004. "The genomic rate of adaptive amino acid substitution in Drosophila." *Molecular biology and evolution* 21 (7): 1350-60.

Blumenstiel, Justin P, Alexandra A Erwin, and Lucas W Hemmer. 2016. "What Drives Positive Selection in the Drosophila piRNA Machinery? The Genomic Autoimmunity Hypothesis." *The Yale journal of biology and medicine* (Yale Journal of Biology and Medicine) 89 (4): 499-512.

Brennecke, Julius, Alexei A Aravin, Alexander Stark, Monica Dus, Manolis Kellis, Ravi Sachidanandam, and Gregory J Hannon. 2007. "Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila." *Cell* 128 (6): 1089-103.

Brockhurst, Michael A., Tracey Chapman, Kayla C. King, Judith E. Mank, Steve Paterson, and Gregory D. D. Hurst. 2014. "Running with the Red Queen: the role of biotic conflicts in evolution." *Proceedings of the Royal Society of London B: Biological Sciences* 281 (1797).

Bronkhorst, Alfred W, and Ronald P van Rij. 2014. "The long and short of antiviral defense: small RNA-based immunity in insects." *Current Opinion in Virology* 7: 19-28.

Bronkhorst, Alfred W, Koen W R van Cleef, Hanka Venselaar, and Ronald P van Rij. 2014. "A dsRNA-binding protein of a complex invertebrate DNA virus suppresses the Drosophila RNAi response." *Nucleic acids research* (Oxford University Press) 42 (19): 12237-48.

Buckling, Angus, and Paul B Rainey. 2002. "Antagonistic coevolution between a bacterium and a bacteriophage." *Proceedings. Biological sciences* (The Royal Society) 269 (1494): 931-6.

Carmell, Michelle A, Zhenyu Xuan, Michael Q Zhang, and Gregory J Hannon. 2002. "The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis." *Genes & development* 16 (21): 2733-42.

Charlesworth, B. 1994. "The effect of background selection against deleterious mutations on weakly selected, linked variants." *Genetical research* 63 (3): 213-27.

Charlesworth, B, P Sniegowski, and W Stephan. 1994. "The evolutionary dynamics of repetitive DNA in eukaryotes." *Nature* 371 (6494): 215-20.

Chen, Ya-Wen, Shilin Song, Ruifen Weng, Pushpa Verma, Jan-Michael Kugler, Marita Buescher, Sigrid Rouam, and Stephen M Cohen. 2014. "Systematic study of Drosophila microRNA functions using a collection of targeted knockout mutations." *Developmental cell* 31 (6): 784-800.

Czech, Benjamin, Colin D Malone, Rui Zhou, Alexander Stark, Catherine Schlingeheyde, Monica Dus, Norbert Perrimon, et al. 2008. "An endogenous small interfering RNA pathway in Drosophila." *Nature* (Nature Publishing Group) 453 (7196): 798-802.

Czech, Benjamin, Gregory J. Hannon, S. Houwing, et al., V.V. Vagin, et al., A.A. Aravin, et al. 2016. "One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing." *Trends in Biochemical Sciences* (Elsevier) 41 (4): 324-337.

Czech, Benjamin, Jonathan B Preall, Jon McGinn, and Gregory J Hannon. 2013. "A transcriptome-wide RNAi screen in the Drosophila ovary reveals factors of the germline piRNA pathway." *Molecular cell* 50 (5): 749-61.

Das, Partha P, Marloes P Bagijn, Leonard D Goldstein, Julie R Woolford, Nicolas J Lehrbach, Alexandra Sapetschnig, Heeran R Buhecha, et al. 2008. "Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the Caenorhabditis elegans germline." *Molecular cell* 31 (1): 79-90.

DePristo, Mark A, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, et al. 2011. "A framework for variation discovery and genotyping using next-generation DNA sequencing data." *Nature genetics* 43 (5): 491-8.

Deshpande, Girish, Gretchen Calhoun, and Paul Schedl. 2005. "Drosophila argonaute-2 is required early in embryogenesis for the assembly of centric/centromeric heterochromatin, nuclear division, nuclear migration, and germ-cell formation." *Genes & development* 19 (14): 1680-5.

Duchaine, Thomas F, James A Wohlschlegel, Scott Kennedy, Yanxia Bei, Darryl Conte, Kaming Pang, Daniel R Brownell, et al. 2006. "Functional proteomics reveals the biochemical niche of C. elegans DCR-1 in multiple small-RNA-mediated pathways." *Cell* 124 (2): 343-54.

Eddy, Sean R. 2008. "A probabilistic model of local sequence alignment that simplifies statistical significance estimation." *PLoS computational biology* (Public Library of Science) 4 (5): e1000069.

Edgar, Robert C. 2004. "MUSCLE: multiple sequence alignment with high accuracy and high throughput." *Nucleic acids research* 32 (5): 1792-7.

Eilertson, Kirsten E, James G Booth, and Carlos D Bustamante. 2012. "SnIPRE: selection inference using a Poisson random effects model." *PLoS computational biology* (Public Library of Science) 8 (12): e1002806.

Enard, David, Le Cai, Carina Gwennap, Dmitri A Petrov, GR. Abecasis, A. Auton, LD. Brooks, et al. 2016. "Viruses are a dominant driver of protein adaptation in mammals." *eLife* (eLife Sciences Publications Limited) 5: 56-65.

Evgen'ev, M. B., H. Zelentsova, N. Shostak, M. Kozitsina, V. Barskyi, D.-H. Lankenau, and V. G. Corces. 1997. "Penelope, a new family of transposable elements and its possible role in hybrid dysgenesis in Drosophila virilis." *Proceedings of the National Academy of Sciences* 94 (1): 196-201.

Eyre-Walker, Adam, and Peter D Keightley. 2009. "Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change." *Molecular biology and evolution* 26 (9): 2097-108.

Fay, Justin C., Gerald J. Wyckoff, and Chung-I Wu. 2001. "Positive and Negative Selection on the Human Genome." *Genetics* 158 (3): 1227-1234.

Ferree, Patrick M, and Daniel A Barbash. 2007. "Distorted sex ratios: a window into RNAi-mediated silencing." *PLoS biology* (Public Library of Science) 5 (11): e303.

Ferrer-Admetlla, Anna, Mason Liang, Thorfinn Korneliussen, and Rasmus Nielsen. 2014. "On detecting incomplete soft or hard selective sweeps using haplotype structure." *Molecular Biology and Evolution* 31 (5): 1275-1291.

Ghildiyal, Megha, Hervé Seitz, Michael D. Horwich, Chengjian Li, Tingting Du, Soohyun Lee, Jia Xu, et al. 2008. "Endogenous siRNAs Derived from Transposons and mRNAs in Drosophila Somatic Cells." *Science* 320 (5879): 1077-1081.

Gossmann, Toni I, Bao-Hua Song, Aaron J Windsor, Thomas Mitchell-Olds, Christopher J Dixon, Maxim V Kapralov, Dmitry A Filatov, and Adam Eyre-Walker. 2010. "Genome wide analyses reveal little evidence for adaptive evolution in many plant species." *Molecular biology and evolution* 27 (8): 1822-32.

Gossmann, Toni I, Peter D Keightley, and Adam Eyre-Walker. 2012. "The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes." *Genome biology and evolution* 4 (5): 658-67.

Ha, Minju, and V Narry Kim. 2014. "Regulation of microRNA biogenesis." *Nature reviews. Molecular cell biology* (Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.) 15 (8): 509-524.

Hadfield, Jarrod D. 2010. "MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package." *Journal of Statistical Software* 33 (2): 1-22.

Hall, Ira M, Ken-Ichi Noma, and Shiv I S Grewal. 2003. "RNA interference machinery regulates chromosome dynamics during mitosis and meiosis in fission yeast." *Proceedings of the National Academy of Sciences of the United States of America* 100 (1): 193-8.

Han, Bo W, Wei Wang, Chengjian Li, Zhiping Weng, and Phillip D Zamore. 2015. "piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production." *Science (New York, N.Y.)* 348 (6236): 817-21.

Handler, Dominik, Katharina Meixner, Manfred Pizka, Kathrin Lauss, Christopher Schmied, Franz Sebastian Gruber, and Julius Brennecke. 2013. "The genetic makeup of the Drosophila piRNA pathway." *Molecular cell* 50 (5): 762-77.

Harpur, Brock A, Clement F Kent, Daria Molodtsova, Jonathan M D Lebon, Abdulaziz S Alqarni, Ayman A Owayss, and Amro Zayed. 2014. "Population genomics of the honey bee reveals strong signatures of positive selection on worker traits." *Proceedings of the National Academy of Sciences of the United States of America* 111 (7): 2614-9.

Harris, Robert S. 2007. "Improved pairwise alignment of genomic dna." (Pennsylvania State University).

Heger, Andreas, and Chris P Ponting. 2007. "Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes." *Genome research* 17 (12): 1837-49.

Huang, Chuan, Xiaolin Wang, Xu Liu, Shuhuan Cao, and Ge Shan. 2015. "RNAi pathway participates in chromosome segregation in mammalian cells." *Cell Discovery* (Nature Publishing Group) 1: 15029.

Kawamura, Yoshinori, Kuniaki Saito, Taishin Kin, Yukiteru Ono, Kiyoshi Asai, Takafumi Sunohara, Tomoko N Okada, Mikiko C Siomi, and Haruhiko Siomi. 2008. "Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells." *Nature* (Nature Publishing Group) 453 (7196): 793-7.

Keightley, Peter D, and Adam Eyre-Walker. 2010. "What can we learn about the distribution of fitness effects of new mutations from DNA sequence data?" *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 365 (1544): 1187-93.

Keightley, Peter D., and Adam Eyre-Walker. 2007. "Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies." *Genetics* 177 (4): 2251-2261.

Khurana, Jaspreet S, Jie Wang, Jia Xu, Birgit S Koppetsch, Travis C Thomson, Anetta Nowosielska, Chengjian Li, Phillip D Zamore, Zhiping Weng, and William E Theurkauf. 2011. "Adaptation to P element transposon invasion in Drosophila melanogaster." *Cell* 147 (7): 1551-63.

Klattenhoff, Carla, and William Theurkauf. 2008. "Biogenesis and germline functions of piRNAs." *Development (Cambridge, England)* 135 (1): 3-9.

Kolaczkowski, Bryan, Daniel N Hupalo, and Andrew D Kern. 2011. "Recurrent adaptation in RNA interference genes across the Drosophila phylogeny." *Molecular biology and evolution* 28 (2): 1033-42.

Kousathanas, Athanasios, Daniel L Halligan, and Peter D Keightley. 2014. "Faster-X adaptive protein evolution in house mice." *Genetics* 196 (4): 1131-43.

Kronforst, Marcus R, Matthew E B Hansen, Nicholas G Crawford, Jason R Gallant, Wei Zhang, Rob J Kulathinal, Durrell D Kapan, and Sean P Mullen. 2013. "Hybridization reveals the evolving genomic architecture of speciation." *Cell reports* 5 (3): 666-77.

Langmead, Ben, and Steven L Salzberg. 2012. "Fast gapped-read alignment with Bowtie 2." *Nature methods* 9 (4): 357-9.

Lewis, Samuel H., Claire L. Webster, Heli Salmela, and Darren J. Obbard. 2016. "Repeated Duplication of Argonaute2 Is Associated with Strong Selection and Testis Specialization in Drosophila." *Genetics* 204 (2).

Li, Hongwei, Wan-Xiang Li, and Shou-Wei Ding. 2002. "Induction and suppression of RNA silencing by an animal virus." *Science* 296 (5571): 1319-21.

Lunter, Gerton, and Martin Goodson. 2011. "Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads." *Genome research* 21 (6): 936-9.

Maynard Smith, J, and J Haigh. 1974. "The hitch-hiking effect of a favourable gene." *Genetical research* 23 (1): 23-35.

McDonald, J H, and M Kreitman. 1991. "Adaptive protein evolution at the Adh locus in Drosophila." *Nature* 351 (6328): 652-4.

McGaugh, Suzanne E., Caiti S. S. Heil, Brenda Manzano-Winkler, Laurence Loewe, Steve Goldstein, Tiffany L. Himmel, and Mohamed A. F. Noor. 2012. "Recombination Modulates How Selection Affects Linked Sites in Drosophila." Edited by Nick H. Barton. *PLoS Biology* (Public Library of Science) 10 (11): e1001422.

Meister, Gunter. 2013. "Argonaute proteins: functional insights and emerging roles." *Nature reviews. Genetics* (Nature Publishing Group) 14 (7): 447-59.

Miesen, Pascal, Alasdair Ivens, Amy H Buck, and Ronald P van Rij. 2016. "Small RNA Profiling in Dengue Virus 2-Infected Aedes Mosquito Cells Reveals Viral piRNAs and Novel Host miRNAs." *PLoS neglected tropical diseases* 10 (2): e0004452.

Miesen, Pascal, Erika Girardi, and Ronald P van Rij. 2015. "Distinct sets of PIWI proteins produce arbovirus and transposon-derived piRNAs in Aedes aegypti mosquito cells." *Nucleic acids research* 43 (13): 6545-56.

Mohn, Fabio, Dominik Handler, and Julius Brennecke. 2015. "piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis." *Science (New York, N.Y.)* 348 (6236): 812-7.

Morazzani, Elaine M., Michael R. Wiley, Marta G. Murreddu, Zach N. Adelman, and Kevin M. Myles. 2012. "Production of Virus-Derived Ping-Pong-Dependent piRNA-like Small RNAs in the Mosquito Soma." Edited by Shou-Wei Ding. *PLoS Pathogens* (Public Library of Science) 8 (1): e1002470.

Muerdter, Felix, Paloma M Guzzardo, Jesse Gillis, Yicheng Luo, Yang Yu, Caifu Chen, Richard Fekete, and Gregory J Hannon. 2013. "A genome-wide RNAi screen draws a genetic framework for transposon control and primary piRNA biogenesis in Drosophila." *Molecular cell* 50 (5): 736-48.

Nayak, Arabinda, Bassam Berry, Michel Tassetto, Mark Kunitomi, Ashley Acevedo, Changhui Deng, Andrew Krutchinsky, John Gross, Christophe Antoniewski, and Raul Andino. 2010. "Cricket paralysis virus antagonizes Argonaute 2 to modulate antiviral defense in Drosophila." *Nature structural & molecular biology* (Nature Publishing Group) 17 (5): 547-554.

Nielsen, R., Scott Williamson, Yuseob Kim, Melissa J Hubisz, Andrew G Clark, and Carlos Bustamante. 2005. "Genomic scans for selective sweeps using SNP data." *Genome Research* 15 (11): 1566-1575.

Obbard, D J, Y-M Linton, F M Jiggins, G Yan, and T J Little. 2007. "Population genetics of Plasmodium resistance genes in Anopheles gambiae: no evidence for strong selection." *Molecular ecology* 16 (16): 3497-510.

Obbard, Darren J., Francis M. Jiggins, Daniel L. Halligan, and Tom J. Little. 2006. "Natural selection drives extremely rapid evolution in antiviral RNAi genes." *Current Biology* 16 (6): 580-585.

Obbard, Darren J., Francis M. Jiggins, Nicholas J. Bradshaw, and Tom J. Little. 2011. "Recent and recurrent selective sweeps of the antiviral RNAi gene argonaute-2 in three species of drosophila." *Molecular Biology and Evolution* 28 (2): 1043-1056.

Obbard, Darren J., John J. Welch, Kang Wook Kim, and Francis M. Jiggins. 2009. "Quantifying adaptive evolution in the Drosophila immune system." *PLoS Genetics* 5 (10).

Pardo-Diaz, Carolina, Camilo Salazar, Simon W Baxter, Claire Merot, Wilsea Figueiredo-Ready, Mathieu Joron, W Owen McMillan, and Chris D Jiggins. 2012. "Adaptive introgression across species boundaries in Heliconius butterflies." *PLoS genetics* (Public Library of Science) 8 (6): e1002752.

Paterson, Steve, Tom Vogwill, Angus Buckling, Rebecca Benmayor, Andrew J. Spiers, Nicholas R. Thomson, Mike Quail, et al. 2010. "Antagonistic coevolution accelerates molecular evolution." *Nature* (Nature Publishing Group) 464 (7286): 275-278.

Pavlidis, P., D. Zivkovic, A. Stamatakis, and N. Alachiotis. 2013. "SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes." *Molecular Biology and Evolution* 30 (9): 2224-2234.

Peccoud, Jean, Vincent Loiseau, Richard Cordaux, and Clément Gilbert. 2017. "Massive horizontal transfer of transposable elements in insects." *Proceedings of the National Academy of Sciences of the United States of America* (National Academy of Sciences) 114 (18): 4721-4726.

Petrov, D. A., J. L. Schutzman, D. L. Hartl, and E. R. Lozovskaya. 1995. "Diverse transposable elements are mobilized in hybrid dysgenesis in Drosophila virilis." *Proceedings of the National Academy of Sciences* 92 (17): 8050-8054.

Rödelsperger, Christian, Richard A Neher, Andreas M Weller, Gabi Eberhardt, Hanh Witte, Werner E Mayer, Christoph Dieterich, and Ralf J Sommer. 2014. "Characterization of genetic diversity in the nematode Pristionchus pacificus from population-scale resequencing data." *Genetics* 196 (4): 1153-65.

Rozhkov, Nikolay V, Alexei A Aravin, Elena S Zelentsova, Natalia G Schostak, Ravi Sachidanandam, W Richard McCombie, Gregory J Hannon, and Michael B Evgen'ev. 2010. "Small RNA-based silencing strategies for transposons in the process of invading Drosophila species." *RNA (New York, N.Y.)* 16 (8): 1634-45.

Sackton, Timothy B, Russell B Corbett-Detig, Javaregowda Nagaraju, Lakshmi Vaishna, Kallare P Arunkumar, and Daniel L Hartl. 2014. "Positive selection drives faster-Z evolution in silkmoths." *Evolution; international journal of organic evolution* 68 (8): 2331-42.

Sarkies, Peter, Murray E. Selkirk, John T. Jones, Vivian Blok, Thomas Boothby, Bob Goldstein, Ben Hanelt, et al. 2015. "Ancient and Novel Small RNA Pathways Compensate for the Loss of piRNAs in Multiple Independent Nematode Lineages." Edited by Laurence D Hurst. *PLOS Biology* (Public Library of Science) 13 (2): e1002061.

Slater, Guy St C, and Ewan Birney. 2005. "Automated generation of heuristics for biological sequence comparison." *BMC bioinformatics* 6 (1): 31.

Smith, Nick G C, and Adam Eyre-Walker. 2002. "Adaptive protein evolution in Drosophila." *Nature* 415 (6875): 1022-4.

Stenseth, Nils Chr., and J. Maynard Smith. 1984. "Coevolution in Ecosystems: Red Queen Evolution or Stasis?" *Evolution* 38 (4): 870.

Thomas, Cristel G, Wei Wang, Richard Jovelin, Rajarshi Ghosh, Tatiana Lomasko, Quang Trinh, Leonid Kruglyak, Lincoln D Stein, and Asher D Cutter. 2015. "Full-genome evolutionary histories of selfing, splitting, and selection in Caenorhabditis." *Genome research* 25 (5): 667-78.

Thomson, Travis, and Haifan Lin. 2009. "The biogenesis and function of PIWI proteins and piRNAs: progress and prospect." *Annual review of cell and developmental biology* 25: 355-76.

van Cleef, Koen W R, Joël T van Mierlo, Pascal Miesen, Gijs J Overheul, Jelke J Fros, Susan Schuster, Marco Marklewitz, Gorben P Pijlman, Sandra Junglen, and Ronald P van Rij. 2014. "Mosquito and Drosophila entomobirnaviruses suppress dsRNA- and siRNA-induced RNAi." *Nucleic acids research* 42 (13): 8732-44.

van Mierlo, Joël T, Alfred W Bronkhorst, Gijs J Overheul, Sajna A Sadanandan, Jens-Ola Ekström, Marco Heestermans, Dan Hultmark, Christophe Antoniewski, and Ronald P van Rij. 2012. "Convergent evolution of argonaute-2 slicer antagonism in two distinct insect RNA viruses." *PLoS pathogens* (Public Library of Science) 8 (8): e1002872.

van Mierlo, Joël T., Alfred W. Bronkhorst, Gijs J. Overheul, Sajna A. Sadanandan, Jens Ola Ekström, Marco Heestermans, Dan Hultmark, Christophe Antoniewski, and Ronald P. van Rij. 2012. "Convergent Evolution of Argonaute-2 Slicer Antagonism in Two Distinct Insect RNA Viruses." *PLoS Pathogens* 8 (8).

van Mierlo, Joël T., Gijs J. Overheul, Benjamin Obadia, Koen W. R. van Cleef, Claire L. Webster, Maria-Carla Saleh, Darren J. Obbard, and Ronald P. van Rij. 2014. "Novel Drosophila Viruses Encode Host-Specific Suppressors of RNAi." Edited by David S. Schneider. *PLoS Pathogens* 10 (7): e1004256.

Van Rij, Ronald P., Maria Carla Saleh, Bassam Berry, Catherine Foo, Andrew Houk, Christophe Antoniewski, and Raul Andino. 2006. "The RNA silencing endonuclease Argonaute 2 mediates specific antiviral immunity in Drosophila melanogaster." *Genes and Development* 20 (21): 2985-2995.

Vermaak, Danielle, Steven Henikoff, and Harmit S Malik. 2005. "Positive selection drives the evolution of rhino, a member of the heterochromatin protein 1 family in Drosophila." *PLoS genetics* (Public Library of Science) 1 (1): 96-108.

Vogler, Hannes, Rashid Akbergenov, Padubidri V Shivaprasad, Vy Dang, Monika Fasler, Myoung-Ok Kwon, Saule Zhanybekova, Thomas Hohn, and Manfred Heinlein. 2007. "Modification of small RNAs associated with suppression of RNA silencing by tobamovirus replicase protein." *Journal of virology* 81 (19): 10379-88.

Voight, Benjamin F, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K Pritchard. 2006. "A map of recent positive selection in the human genome." *PLoS biology* (Public Library of Science) 4 (3): e72.

Wang, Xiao-Hong, Roghiyh Aliyari, Wan-Xiang Li, Hong-Wei Li, Kevin Kim, Richard Carthew, Peter Atkinson, and Shou-Wei Ding. 2006. "RNA Interference Directs Innate Immunity Against Viruses in Adult Drosophila." *Science* 312 (5772).

Welch, John J. 2006. "Estimating the genomewide rate of adaptive protein evolution in Drosophila." *Genetics* 173 (2): 821-37.

Xia, Qingyou, Yiran Guo, Ze Zhang, Dong Li, Zhaoling Xuan, Zhuo Li, Fangyin Dai, et al. 2009. "Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx)." *Science (New York, N.Y.)* (NIH Public Access) 326 (5951): 433-6.

Yang, Z, and R Nielsen. 2000. "Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models." *Molecular biology and evolution* 17 (1): 32-43.

Yang, Z. 2007. "PAML 4: Phylogenetic Analysis by Maximum Likelihood." *Molecular Biology and Evolution* (Oxford University Press) 24 (8): 1586-1591.

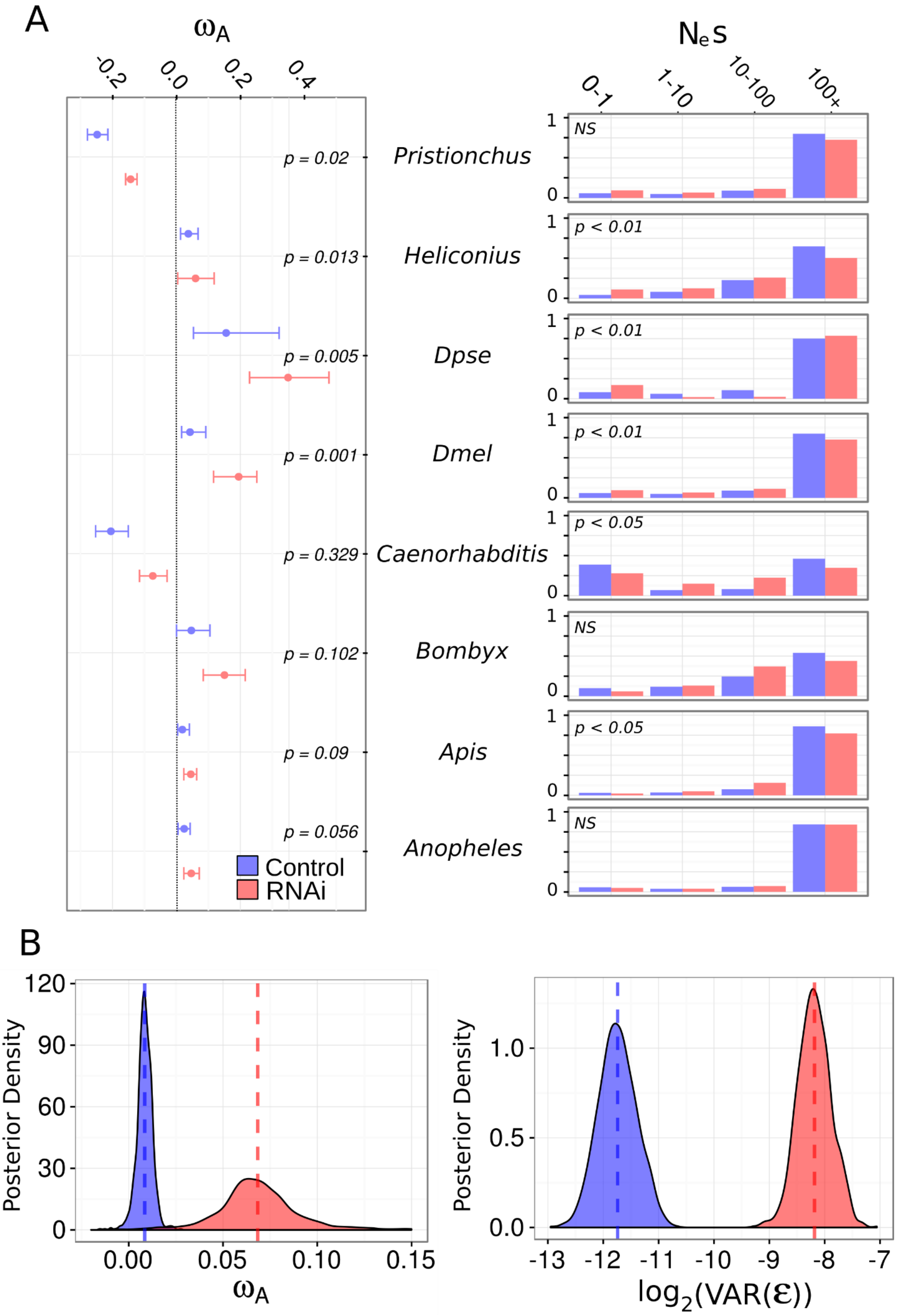

A
ω_A
-0.2
0.0
0.2
0.4
$N_e s$
0-1
1-10
10-100
100+
p = 0.02
Pristionchus
NS
p = 0.013
Heliconius
p < 0.01
p = 0.005
Dpse
p < 0.01
p = 0.001
Dmel
p < 0.01
p = 0.329
Caenorhabditis
p < 0.05
p = 0.102
Bombyx
NS
p = 0.09
Apis
p < 0.05
p = 0.056
Anopheles
NS
1
0
Control
RNAi
B
Posterior Density
120
90
60
30
0
0.00
0.05
0.10
0.15
ω_A
Posterior Density
1.0
0.5
0.0
-13
-12
-11
-10
-9
-8
-7
log2(VAR(ε))

Figure 1: ω_A the DFE differ between RNAi genes and other genes

(A) Left: For each species, ω_A estimates and bootstrap confidence intervals for control (i.e. non-RNAi; blue) and RNAi (red) genes are plotted, with 95% bootstrap confidence intervals. Significance was determined by permutation. Right: The estimated discretised DFE for each species, with the proportion of mutation with $N_e s$ values in each category given for non-RNAi (blue) and RNAi (red) genes. (B) The posterior distribution of estimated ω_A for RNAi (red) versus control (blue) genes, showing that RNAi genes have much great ω_A estimates (left) and greater residual gene-level variation (right), indicating RNAi genes display higher rates adaptive amino-acid substitution, but are more variable.

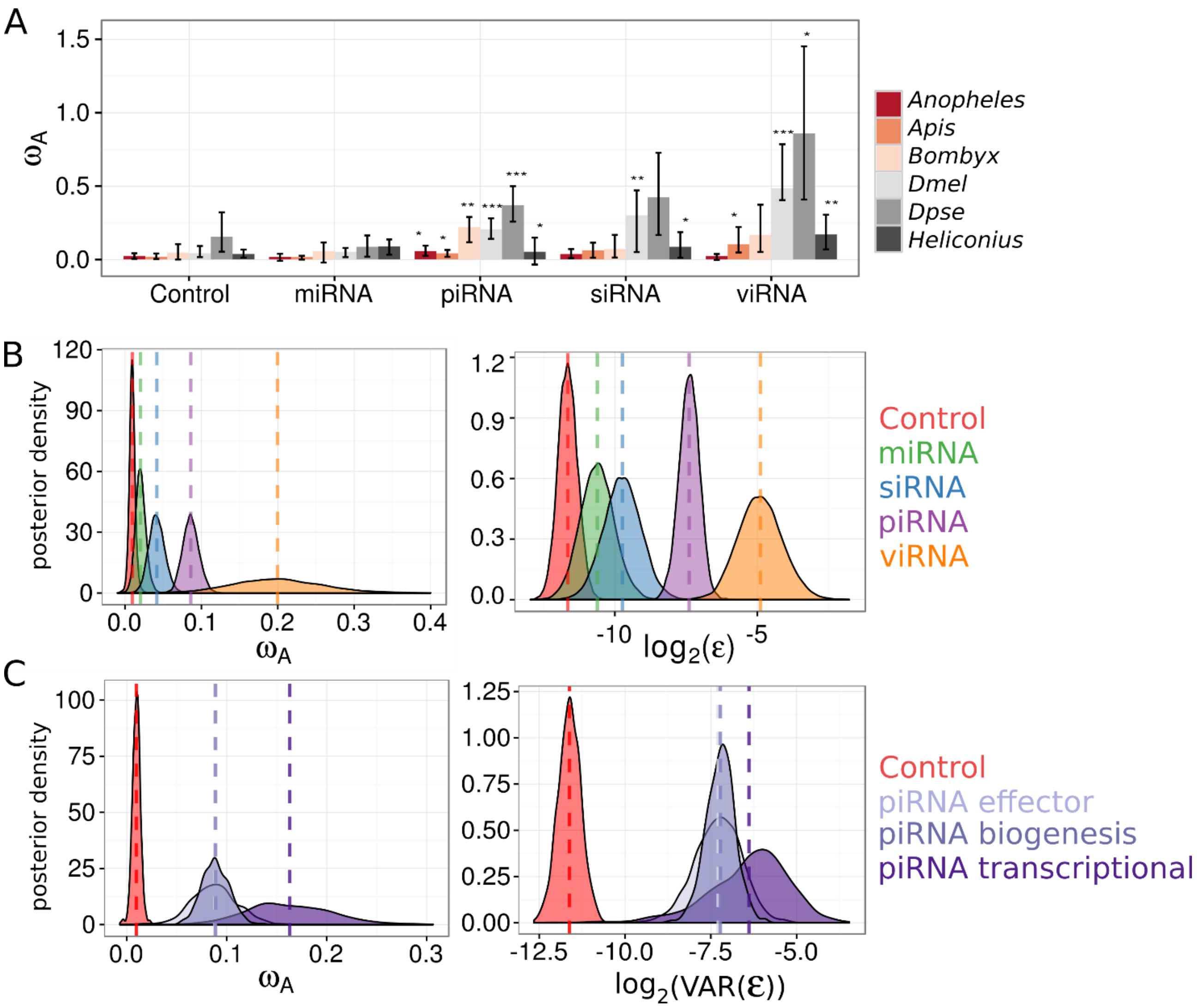


Figure 2: DFE-alpha estimates of ω_A for each subpathway

(A) ω_A estimates from pooled polymorphism and divergence data across insect RNAi subpathways using DFE-alpha. ω_A was estimated for each subpathway in each organism and confidence intervals obtained by bootstrapping across genes. Significance was assessed by permutation tests between sub-pathway and control genes for each organism ($p < 0.05$*, $p<0.01$**, $p<0.001$***). (B) Individual-gene DFE-alpha ω_A estimates were analysed using a linear mixed model in MCMCglmm, and show that (left) the viRNA pathway exhibits the fastest rate of adaptive protein substitution, followed by the piRNA pathway, and that among-gene variance shows the same pattern (right). (C) Individual gene DFE-alpha ω_A estimates were analysed in MCMCglmm, except that the piRNA pathway was further split into genes involved in transcriptional silencing, piRNA biogenesis, or piRNA-mediated effectors of silencing. The posterior distributions of these three effect sizes versus control genes are plotted. All three piRNA functions are targets of elevated positive selection and have large residual variances, although genes mediating transcriptional silencing have greater point estimates for both.

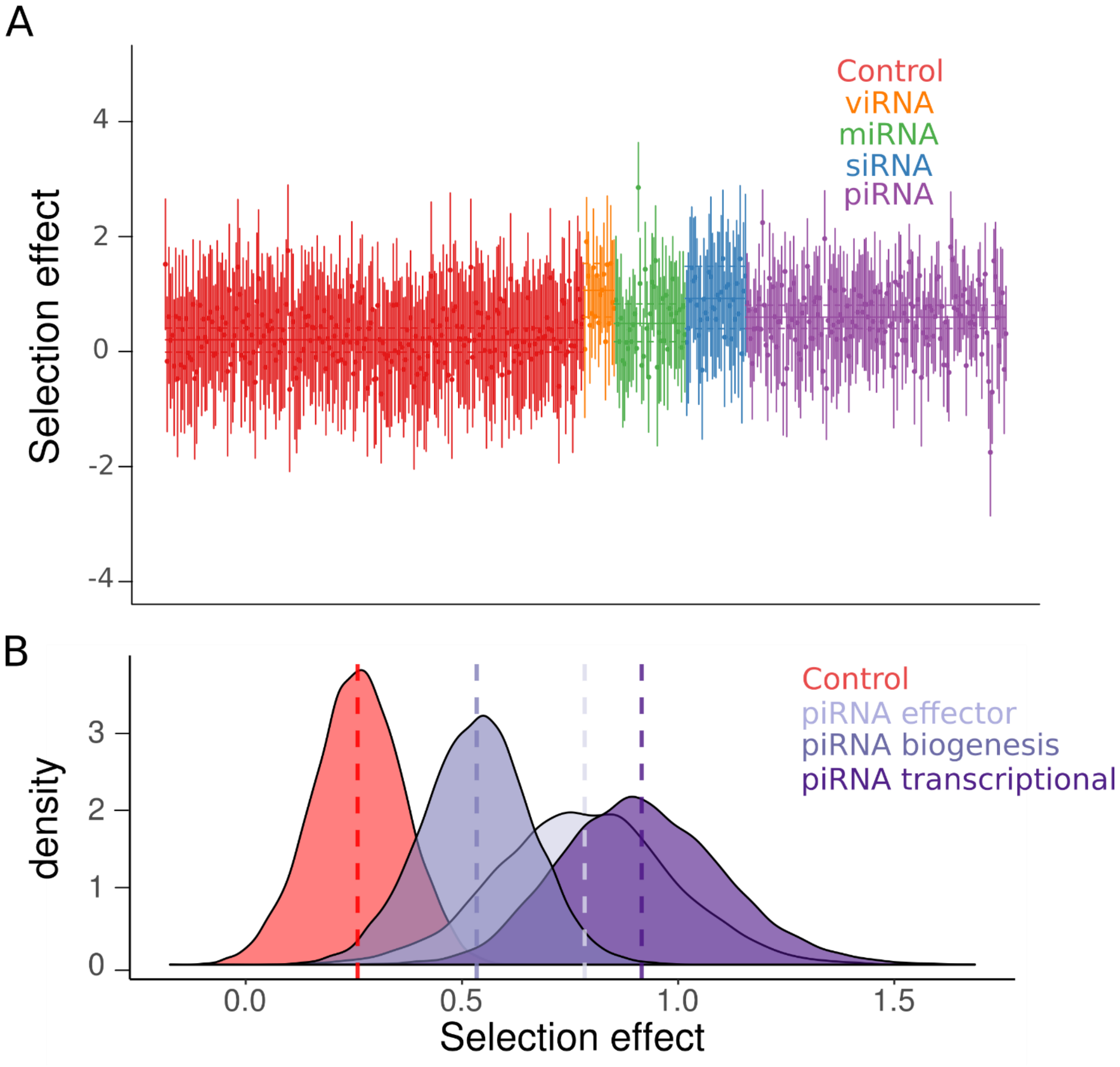

A
Selection effect
4
2
0
-2
-4
Control
viRNA
miRNA
siRNA
piRNA
B
density
3
2
1
0
0.0
0.5
1.0
1.5
Selection effect
Control
piRNA effector
piRNA biogenesis
piRNA transcriptional

Figure 3 SnIPRE-like selection effects

(A) SnIPRE 'selection effect' with 95% confidence intervals (species-level effects removed) are plotted for each gene in each species, coloured according to the gene's role in the RNAi pathway. Solid horizontal lines signify the mean selection effect for each RNAi subpathway (or control genes) with dotted lines signifying the 95% confidence intervals for the subpathway mean. SnIPRE and DFE-alpha analyses are consistent in suggesting that the viRNA, endo-siRNA, and piRNA pathway have more adaptive amino-acid substitutions than control genes. (B) We also performed a SnIPRE analysis after dividing the piRNA pathway into three functional classes, as in Figure 2. The posterior distribution of selection effects associated with each piRNA function are plotted. Similar to DFE-alpha, SnIPRE identifies all three pathways as significantly elevated relative to control genes, however in the SnIPRE analysis transcriptional silencing genes have a significantly greater adaptive rate than biogenesis factors.

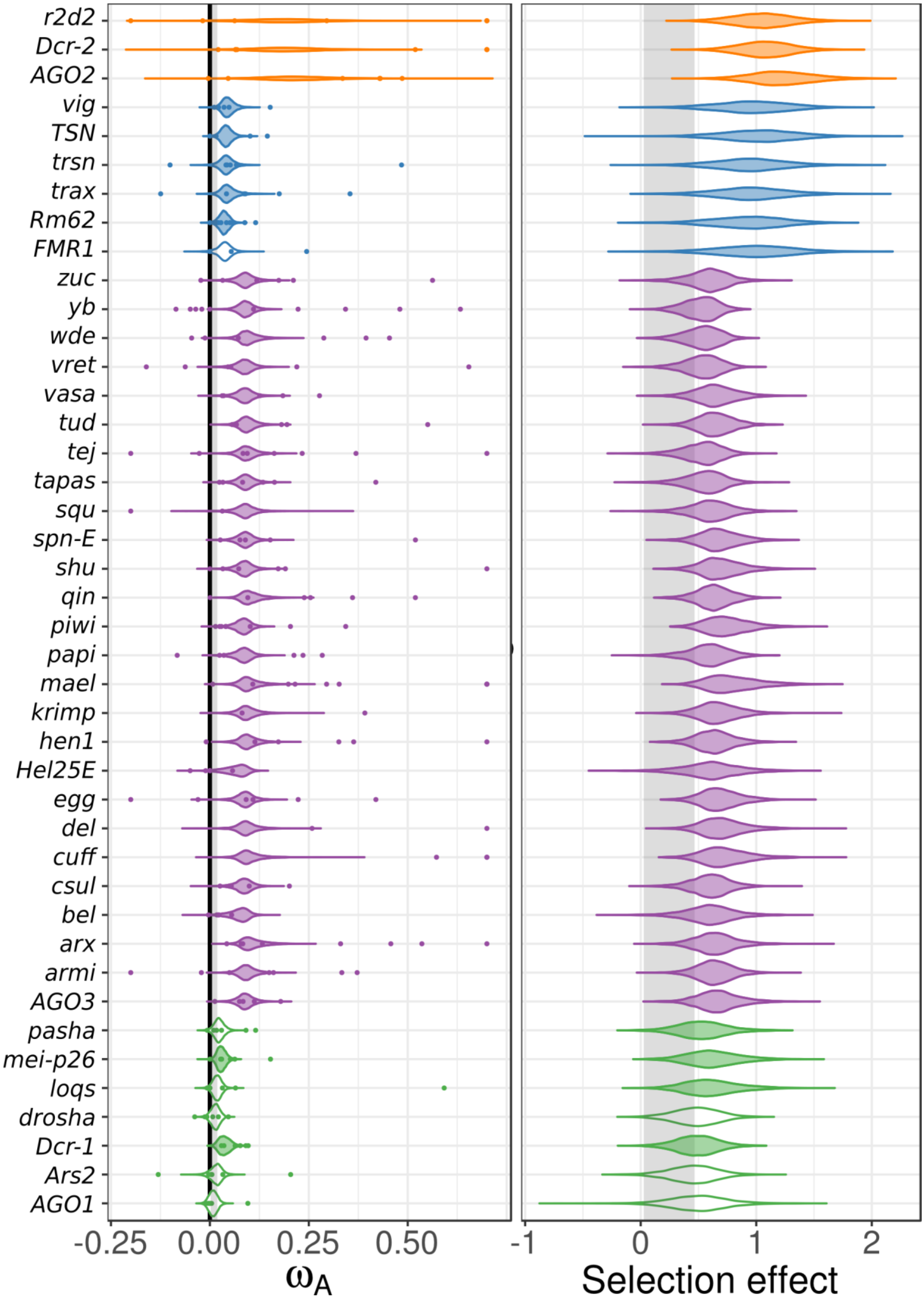


Figure 4 Cross-species homologue-level estimates of ω_A and selection effects

(Left) Individual homologue ω_A estimates (coloured points) were calculated using DFE-alpha and analysed using a linear mixed model with subpathway as fixed effect and species and homologue as a random effect (estimate uncertainty was included by incorporating bootstrap intervals as measurement error variance). The posterior distributions of the cross-species estimate for ω_A for each homologue are plotted, and shaded if significantly different from the control gene distribution (region shaded grey). Single-gene estimates of ω_A > 0.75 are plotted at 0.75 for clarity. (Right) The analogous analysis performed using SnIPRE, with the posterior distribution of homologue-level selection effects plotted. Both analyses find little variation among homologues after accounting for subpathway, and homologue-level analyses generally mirror pathway-specific analyses. See Figure S4 for the equivalent models that exclude the fixed effect of pathway.

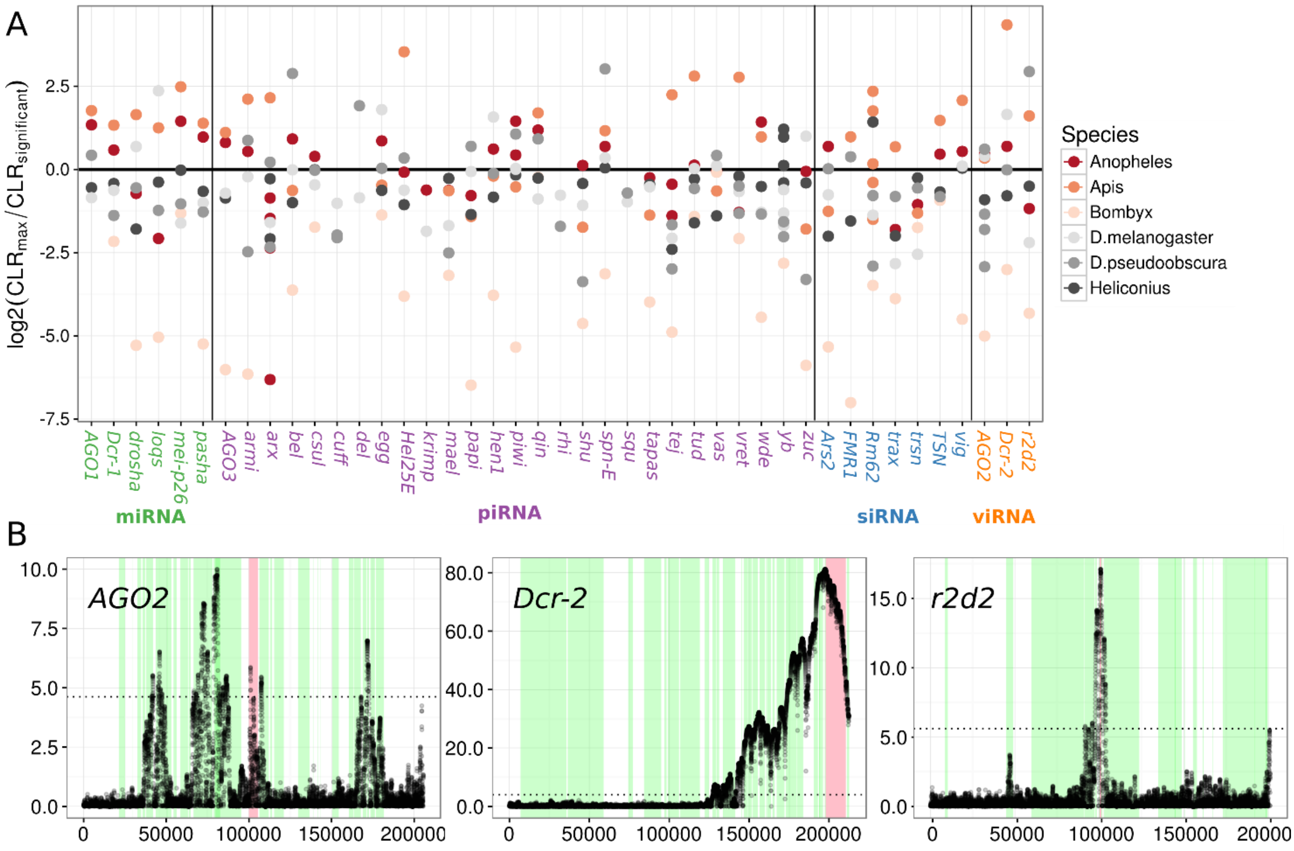


Figure 5 Selective sweeps in RNAi genes and example SweeD plots

(A) Points indicate the $\log_2$ ratio of the maximum observed CLR value (from SweeD) in the named gene to the CLR 95% significance threshold inferred from simulation. Values above 0 indicate there was a 'significant' CLR peak in a genic region and colours indicate species. (B) The viRNA pathway in *Apis mellifera* shows strong evidence for recent sweeps. For each of the three viRNA pathway genes the CLR statistic is plotted across a 200 kb region. The dotted line is the significance threshold estimated through neutral simulations under a published demographic history. Red regions denote the focal gene and green regions highlight surrounding genes. In Apis, both *Dcr2* and *R2D2* show strong evidence for sweeps with the surrounding region of *Dcr2* being devoid of polymorphism, indicating this sweep was recent and rapid. *AGO2* also shows a significant peak, but this is narrow and only marginally significant.

Gene	Flybase ID	Pathway	References
AGO1	FBgn0262739	miRNA	Okamura et al, 2004
AGO2	FBgn0087035	viRNA	Wang et al, 2006; van Rij et al, 2006
AGO3	FBgn0250816	piRNA - effector	Brennecke et al, 2007; Gunawardane et al, 2007
armi	FBgn0041164	piRNA - biogenesis	Saito et al, 2010
Ars2	FBgn0033062	miRNA	Gruber et al, 2009
arx	FBgn0036826	piRNA - effector	Ohtani et al, 2013; Donertas et al, 2013
bel	FBgn0263231	piRNA	Lo et al, 2016
csul	FBgn0015925	piRNA - effector	Kirino et al, 2009
cuff	FBgn0260932	piRNA - transcriptional	Pane et al, 2011; Mohn et al, 2014
Dcr-1	FBgn0039016	miRNA	Jiang et al, 2005; Saito et al, 2005
Dcr-2	FBgn0034246	viRNA	Wang et al, 2006; Galiana-Arnoux et al, 2006
del	FBgn0086251	piRNA - transcriptional	Mohn et al, 2014
drosha	FBgn0026722	miRNA	Lee et al, 2003
egg	FBgn0086908	piRNA - transcriptional	Rangan et al, 2011
FMR1	FBgn0086908	siRNA	Ishizuka et al, 2002
Hel25E	FBgn0014189	piRNA - biogenesis	Zhang et al, 2012
hen1	FBgn0033686	piRNA - biogenesis	Saito et al, 2007; Horwich et al, 2007
krimp	FBgn0034098	piRNA	Sato et al, 2015
loqs	FBgn0032515	miRNA	Jiang et al, 2005; Saito et al, 2005
mael	FBgn0016034	piRNA - transcriptional	Sienski et al, 2012
Mei-p26	FBgn0026206	miRNA	Neumuller et al, 2008
papi	FBgn0031401	piRNA - biogenesis	Saxe et al, 2013
pasha	FBgn0039861	miRNA	Yeom et al, 2006
piwi	FBgn0004872/FBgn0000146	piRNA - effector	Brennecke et al, 2007
qin	FBgn0263974	piRNA - biogenesis	Zhang et al, 2011
r2d2	FBgn0031951	viRNA	Liu et al, 2003
rhi	FBgn0004400	piRNA - transcriptional	Klatenhoff et al, 2009; Mohn et al, 2014
rm62	FBgn0003261	siRNA	Csink et al, 1994
shu	FBgn0003401	piRNA - biogenesis	Olivieri et al, 2012; Preall et al, 2012
spn-E	FBgn0003483	piRNA - biogenesis	Malone et al, 2009
squ	FBgn0267347	piRNA - effector	Haase et al, 2010
tapas	FBgn0027529	piRNA - biogenesis	Patil et al, 2014
tejas	FBgn0033921	piRNA - biogenesis	Patil et al, 2010
trax	FBgn0038327	siRNA	Liu et al, 2009
trsn	FBgn0033528	siRNA	Liu et al, 2009
TSN	FBgn0035121	siRNA	Caudy et al, 2003
tudor	FBgn0003891	piRNA - biogenesis	Nishida et al, 2009
vasa	FBgn0283442	piRNA - biogenesis	Xiol et al, 2014
vig	FBgn0024183	siRNA	Caudy et al, 2002
vret	FBgn0263143	piRNA - biogenesis	Zamparini et al, 2011
wde	FBgn0027499	piRNA - transcriptional	Koch et al, 2009
Yb	FBgn0000928,FBgn0051755,FBgn0037205	piRNA - biogenesis	Saito et al, 2010
zuc	FBgn0261266	piRNA - biogenesis	Mohn et al, 2015; Han et al, 2015

Table S1: Insect genes analysed, FlyBase identifiers, and subpathway involvement

Gene	Wormbase ID	Gene	Wormbase ID
ACR-11	WBGene00000050	WAGO-4	WBGene00010263
AIN-2	WBGene00015007	LAM-2	WBGene00016913
C04F12.1	WBGene00007297	C35E7.8	WBGene00016460
CDC-25.1	WBGene00000386	EGO-1	WBGene00001214
DRH-3	WBGene00008400	EKL-1	WBGene00009052
DRSH-1	WBGene00009163	PPW-1	WBGene00004093
ERI-6	WBGene00016561	PPW-2	WBGene00004094
ERI-7	WBGene00016566	PRG-1	WBGene00004178
F57C9.7	WBGene00019013	PRG-2	WBGene00004179
GEI-11	WBGene00001568	RDE-2	WBGene00004324
HAF-6	WBGene00001816	SAGO-1	WBGene00019666
HPO-24	WBGene00011945	SAGO-2	WBGene00018921
MEL-26	WBGene00003209	SMG-2	WBGene00004880
MUT-16	WBGene00003508	WAGO-1	WBGene00011061
MUT-2	WBGene00003499	WAGO-2	WBGene00018862
PTR-2	WBGene00004217	WAGO-5	WBGene00022877
RDE-10	WBGene00021634	Y23H5A.3	WBGene00021270
RRF-1	WBGene00004508	ERGO-1	WBGene00019971
RRF-2	WBGene00004509	NRDE-2	WBGene00011333
RSD-6	WBGene00004684	PHO-1	WBGene00004020
ULP-5	WBGene00006740	PID-1	WBGene00017549
W01A8.5	WBGene00012167	PIR-1	WBGene00011967
PASH-1	WBGene00011908	RRF-3	WBGene00004510
MEL-47	WBGene00017132	VIG-1	WBGene00006924
TSN-1	WBGene00006626	WAGO-11	WBGene00021711
XRN-2	WBGene00006964	ALG-3	WBGene00011910
CUL-2	WBGene00000837	ALG-4	WBGene00006449
DCR-1	WBGene00000939	CGH-1	WBGene00000479
ERI-9	WBGene00016143	MUT-7	WBGene00003504
HENN-1	WBGene00015349	TBP-1	WBGene00006542
HPL-2	WBGene00001996	Y37D8A.16	WBGene00012554
NHL-2	WBGene00003598	COGC-2	WBGene00000585
NRDE-1	WBGene00007577	ERI-3	WBGene00021103
RDE-4	WBGene00004326	RDE-11	WBGene00023421
SID-2	WBGene00004796	C14B1.7	WBGene00007578
ZK418.8	WBGene00022737	C27H6.3	WBGene00007785
CSR-1	WBGene00017641	DCS-1	WBGene00000940
DRH-1	WBGene00001090	HRDE-1	WBGene00007624
EPI-1	WBGene00001328	MUT-14	WBGene00003507
ERI-1	WBGene00001332	NRDE-3	WBGene00019862
ERI-5	WBGene00021419	PRDE-1	WBGene00008995
LAM-1	WBGene00002247	RDE-12	WBGene00010280
RSD-2	WBGene00004681	WAGO-10	WBGene00020707
F20A1.9	WBGene00017620	AIN-1	WBGene00015547
MUT-15	WBGene00011323	ALG-1	WBGene00000105
RDE-1	WBGene00004323	ALG-2	WBGene00000106
SID-1	WBGene00004795		

Table S2 Nematode Genes and WormBase identifiers

Gene	Location
BoYb	Germline
CG9925	Germline
csul	Germline
cuff	Germline
del	Germline
tud	Germline
vls	Germline
AGO3	Germline
aub	Germline
Hel25E	Germline
krimp	Germline
qin	Germline
rhi	Germline
spn-E	Germline
squ	Germline
tejas	Germline
acn	Germline/Follicle cells
arp6	Germline/Follicle cells
CG3689	Germline/Follicle cells
CG3909	Germline/Follicle cells
CG7504	Germline/Follicle cells
CG9754	Germline/Follicle cells
droj2	Germline/Follicle cells
gasz	Germline/Follicle cells
his2av	Germline/Follicle cells
his33a	Germline/Follicle cells
mago	Germline/Follicle cells
mino	Germline/Follicle cells
tsu	Germline/Follicle cells
wde	Germline/Follicle cells
armi	Germline/Follicle cells
egg	Germline/Follicle cells
mael	Germline/Follicle cells
hen1	Germline/Follicle cells
shu	Germline/Follicle cells
vret	Germline/Follicle cells
adgf-a	Follicle cells
arx	Follicle cells
asf1	Follicle cells
atu	Follicle cells
caf1105	Follicle cells
CG14749	Follicle cells
CG4294	Follicle cells
CG5222	Follicle cells
CG5491	Follicle cells
CG6479	Follicle cells
CG7950	Follicle cells
CG8569	Follicle cells
CG8949	Follicle cells
ebi	Follicle cells
hdac3	Follicle cells
hmgd	Follicle cells
l3neo38	Follicle cells
mep1	Follicle cells
nup214	Follicle cells
nup54	Follicle cells
nup58	Follicle cells
omd	Follicle cells
patr-1	Follicle cells
sbr	Follicle cells
SoYb	Follicle cells
tfiis	Follicle cells
top1	Follicle cells
veli	Follicle cells
Yb	Follicle cells

Table S3: Larger set of piRNA-implicated genes used to calculate rates of adaptive evolution in the germline and soma of *D. melanogaster*

Organism	Outgroup	Sampled alleles	Data Acquisition	Reference genome
Drosophila melanogaster	Drosophila simulans	197	http://www.dpgp.org/dpgp3/DPGP3.html	Flybase Release 5
Drosophila pseudoobscura	Drosophila miranda	11	http://pseudobase.biology.duke.edu/	Flybase Release 2.9
Anopheles gambiae	Anopheles christyii/melas	24	SAMEA1964392,SAMEA2055794, SAMEA2055810,SAMEA2055935, SAMEA2056025,SAMEA2056029, SAMEA2056081,SAMEA2056178, SAMEA2056188,SAMEA2056339, SAMEA2056343,SAMEA2058436	AgamP3, PEST strain
Heliconius melpomene	Heliconius hecale	20	SRR1057594,SRR1057595,SRR1057596, SRR1057597,SRR1057598,SRR1057599, SRR1057600,SRR1057601,SRR1057602, SRR1057603	Release 1.1
Bombyx mandarina	Bombyx huttoni	22	SRR020026,SRR020027,SRR020758,SRR020028, SRR020029,SRR020030,SRR020770,SRR020771, SRR020031,SRR020032,SRR020759,SRR020035, SRR020034,SRR020033,SRR020760,SRR020036, SRR020761,SRR020773,SRR020774,SRR020775, SRR020776,SRR020802,SRR020037,SRR020772, SRR020039,SRR020762,SRR020777,SRR020778, SRR020779,SRR020780,SRR020781,SRR020041, SRR020040,SRR020042,SRR020763,SRR020764, SRR020043,SRR020765,SRR020044,SRR020766	silkgenome (Bombyx mori)
Apis mellifera	Apis cerana	78	SRR957058,SRR957059,SRR957060,SRR957061, SRR957062,SRR957063,SRR957064,SRR957065, SRR957066,SRR957067,SRR957068,SRR957069, SRR957070,SRR957071,SRR957072,SRR957073, SRR957074,SRR957075,SRR957076,SRR957077, SRR957078,SRR957080,SRR957081,SRR957082, SRR957083,SRR957084,SRR957085,SRR957086, SRR957087,SRR957088,SRR957089,SRR957090, SRR957091,SRR957092,SRR957093,SRR957094, SRR957095,SRR957096,SRR957097	Release 4.5
Caenorhabditis briggsae	Caenorhabditis negoni	36	SRR1792917,SRR1792918,SRR1792919, SRR1792920,SRR1792921,SRR1792922, SRR1792923,SRR1792931,SRR1792933, SRR1792934,SRR1792937,SRR1792942, SRR1792945,SRR1792947,SRR1792950, SRR1792953,SRR1792955,SRR1792961, SRR1792963,SRR1792964,SRR1792967, SRR1792970,SRR1792972,SRR1792974, SRR1792976,SRR1792978,SRR1792980, SRR1792983,SRR1792992,SRR1792996, SRR1793002,SRR1793004,SRR1793005, SRR1793006,SRR1793010,SRR1793012	CB4
Pristionchus pacificus	Pristionchus exspectatus	46	SRR543703,SRR545197,SRR545198,SRR545199, SRR545200,SRR545201,SRR545202,SRR546092, SRR546098,SRR546100,SRR546101,SRR546241, SRR546242,SRR546246,SRR546251,SRR546267, SRR546268,SRR546269,SRR546271,SRR546272, SRR546274,SRR546276,SRR546395,SRR546396, SRR546477,SRR546479,SRR546484,SRR546486, SRR546488,SRR546489,SRR546490,SRR546491, SRR546493,SRR546494,SRR546498,SRR546499, SRR546500,SRR546501,SRR546502,SRR546503, SRR546504,SRR546507,SRR546508,SRR546509, SRR546510,SRR546511	Pristionchus Freeze 1

Table S4 Accession numbers for public data and genome assembly used for each species

Species	Gene	Duplicate	Selection Effect	95% HPD interval
D. melanogaster	arx	A	1.16	[0.157-2.09]
D. melanogaster	piwi	A	1.11	[0.294-1.94]
D. melanogaster	csul	A	1.53	[0.693-2.33]
D. melanogaster	Dcr-1	A	1.21	[0.441-2]
D. melanogaster	Dcr-2	A	1.03	[0.151-1.89]
D. melanogaster	krimp	A	1.12	[0.174-1.94]
D. melanogaster	qin	A	1.04	[0.421-1.61]
D. melanogaster	r2d2	A	1.59	[0.625-2.66]
D. melanogaster	tapas	A	1.35	[0.639-2.03]
D. melanogaster	tej	A	0.826	[0.21-1.4]
D. melanogaster	tud	A	0.816	[0.0691-1.51]
D. melanogaster	vas	A	1.63	[0.975-2.33]
D. melanogaster	vret	A	0.995	[0.211-1.72]
D. melanogaster	Yb	C	1.2	[0.472-1.93]
D. melanogaster	SoYb	B	0.78	[0.262-1.3]
D. pseudoobscura	AGO2	A	1.07	[0.307-1.86]
D. pseudoobscura	armi	B	2.66	[2.09-3.25]
D. pseudoobscura	armi	A	1.22	[0.622-1.82]
D. pseudoobscura	arx	B	1.01	[0.0294-2.11]
D. pseudoobscura	piwi	A	0.952	[0.043-1.81]
D. pseudoobscura	csul	A	0.921	[0.00777-1.83]
D. pseudoobscura	cuff	A	1.26	[0.41-2.12]
D. pseudoobscura	cuff	B	1.33	[0.458-2.16]
D. pseudoobscura	Dcr-2	A	1.48	[0.762-2.27]
D. pseudoobscura	del	A	1.33	[0.413-2.17]
D. pseudoobscura	drosha	A	1.62	[0.788-2.47]
D. pseudoobscura	egg	A	0.949	[0.249-1.63]
D. pseudoobscura	loqs	A	1.33	[0.301-2.42]
D. pseudoobscura	mael	A	1.22	[0.0958-2.36]
D. pseudoobscura	mael	B	1.22	[0.0897-2.36]
D. pseudoobscura	mei-p26	A	1.66	[0.852-2.38]
D. pseudoobscura	papi	A	0.79	[0.0114-1.56]
D. pseudoobscura	pasha	A	1.2	[0.247-2.14]
D. pseudoobscura	qin	A	1.56	[0.975-2.16]
D. pseudoobscura	r2d2	A	1.29	[0.22-2.37]
D. pseudoobscura	rm62	B	1.33	[0.247-2.29]
D. pseudoobscura	rm62	A	1.09	[0.0839-2.01]
D. pseudoobscura	shu	A	1.57	[0.667-2.48]
D. pseudoobscura	spn-E	A	1.97	[1.31-2.59]
D. pseudoobscura	squ	A	1.19	[0.26-2.09]
D. pseudoobscura	tej	A	1.13	[0.171-2]
D. pseudoobscura	trsn	A	1.54	[0.622-2.56]
D. pseudoobscura	tud	A	0.912	[0.343-1.53]
D. pseudoobscura	vas	A	1.16	[0.28-2.04]
D. pseudoobscura	vig	A	1.49	[0.337-2.65]
D. pseudoobscura	vret	B	1.48	[0.703-2.3]
D. pseudoobscura	wde	A	0.876	[0.251-1.5]
D. pseudoobscura	Yb	A	1.75	[0.871-2.74]
D. pseudoobscura	Yb	B	0.737	[0.144-1.31]
D. pseudoobscura	zuc	A	1.43	[0.649-2.15]
Heliconius	shu	A	0.756	[0.0191-1.51]
Heliconius	Yb	A	1.03	[0.483-1.57]
Apis	AGO2	A	1.55	[0.76-2.37]
Apis	piwi	A	1.04	[0.306-1.79]
Apis	Dcr-2	A	0.965	[0.195-1.72]
Apis	rm62	C	1.12	[0.02-2.13]
Apis	tud	A	0.941	[0.253-1.61]
Apis	vas	A	1.46	[0.467-2.48]
Anopheles	AGO3	A	0.934	[0.103-1.85]
Anopheles	armi	A	0.82	[0.0614-1.57]
Anopheles	arx	A	1.39	[0.376-2.33]
Anopheles	piwi	B	1.26	[0.569-1.92]
Anopheles	csul	A	1.04	[0.141-1.91]
Anopheles	Dcr-1	A	0.772	[0.101-1.43]
Anopheles	drosha	A	1.54	[0.721-2.37]
Anopheles	loqs	A	1.68	[0.794-2.54]
Anopheles	mei-p26	A	1.08	[0.234-1.89]
Anopheles	papi	A	0.98	[0.203-1.79]
Anopheles	tej	A	0.868	[0.169-1.53]
Anopheles	trax	A	1.48	[0.433-2.48]
Anopheles	tud	A	1.37	[0.716-2.03]
Anopheles	vig	A	1.29	[0.196-2.4]
Anopheles	wde	A	0.957	[0.338-1.58]
Anopheles	zuc	A	1.42	[0.535-2.31]
Bombyx	AGO2	A	1.41	[0.633-2.18]
Bombyx	Dcr-1	A	2.95	[2.14-3.75]
Bombyx	egg	A	2.06	[1.26-2.92]
Bombyx	Hel25E	A	1.18	[0.0422-2.35]
Bombyx	mael	A	1.04	[0.324-1.75]
Bombyx	papi	A	1.15	[0.11-2.18]
Bombyx	shu	A	1.31	[0.64-1.97]
Bombyx	tapas	A	0.989	[0.233-1.74]
Bombyx	trsn	A	1.73	[0.595-2.94]
Bombyx	vig	A	1.72	[0.432-2.92]
Bombyx	vret	A	0.714	[0.0928-1.33]
Bombyx	Yb	A	1.34	[0.778-1.87]

Table S5 Genes with significantly elevated selection effects

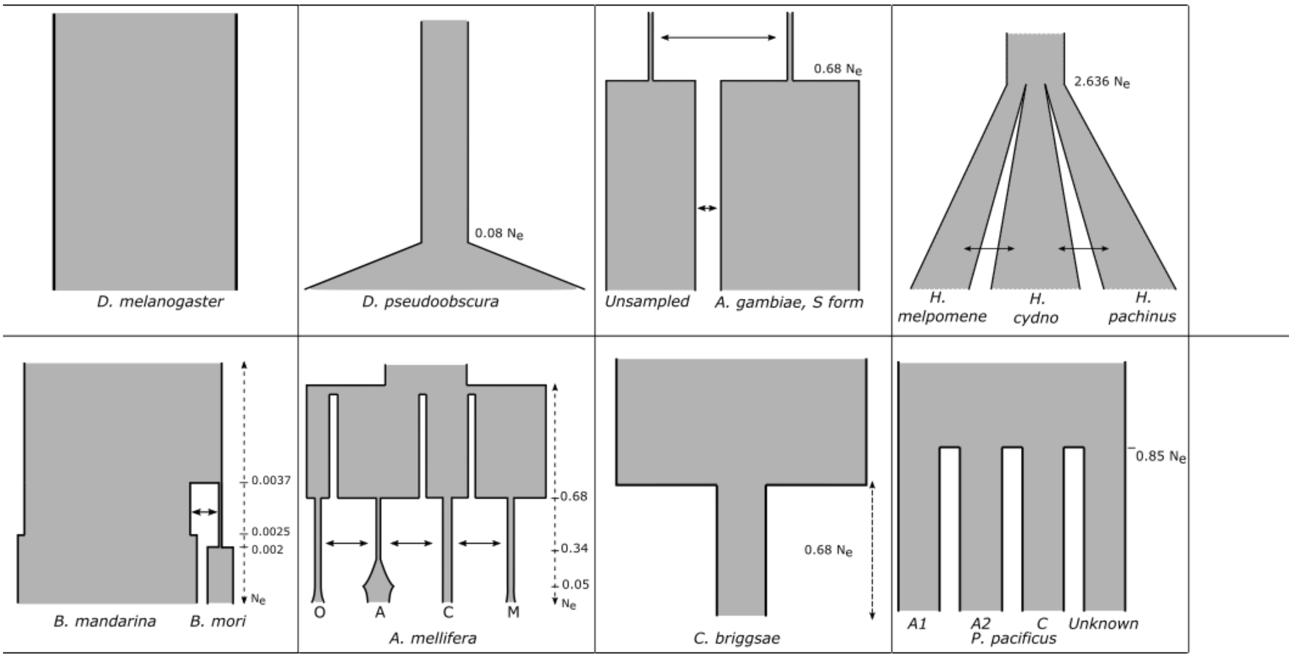


Figure S1 Demographic scenarios simulated for SweeD analysis

Coalescent simulations were performed using ms for demographic scenarios for each species which are supported by other studies. The African (Zambia) *D. melanogaster* were assumed to have a constant population size. *D. pseudoobscura* has recently undergone a population expansion 0.08 Ne generations ago. *A. gambiae* shares migrants with some other unknown, unsampled subpopulation which split 0.68 Ne generations ago. *Heliconius* species in Costa Rica split 2.636 Ne generations ago and have shared migrants since. *Bombyx mandarina* went through a small bottleneck when *B. mori* split, and shared migrants during that bottleneck (but not after). *Apis mellifera* have four subpopulations which have gone through multiple population expansions and bottlenecks, with all subpopulations sharing migrants until they join 0.68 Ne generations ago. *Caenorhabditis briggsae "*tropical samples" have undergone a population bottleneck 0.68 Ne generations ago. Finally, *Pristionchus pacificus* were sampled from four subpopulations, which split 0.85 Ne generations ago.

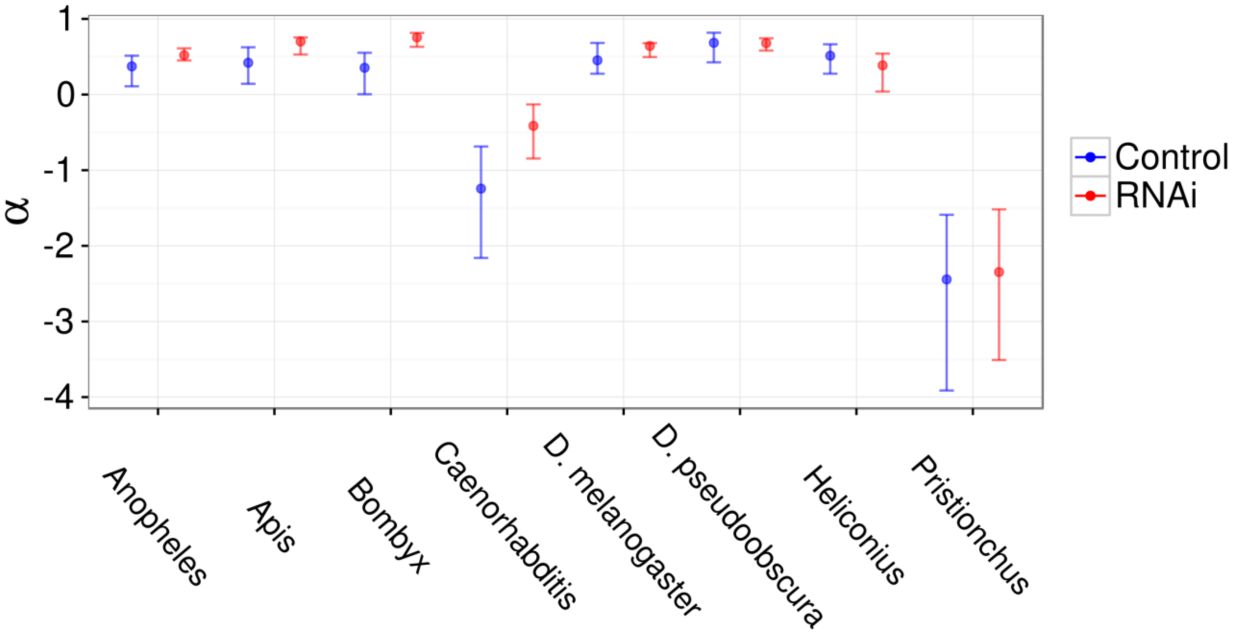


Figure S2: Alpha values for RNAi genes

For each species, α, or the proportion of adaptive substitutions was estimated from pooled polymorphism and divergence data using DFE-alpha for RNAi genes and position-matched control genes. α estimates for control genes are fairly constant across insect species, but are negative in the two nematode species. In all species except *H. melpomene*, the RNAi gene estimates are greater in RNAi genes than control genes.

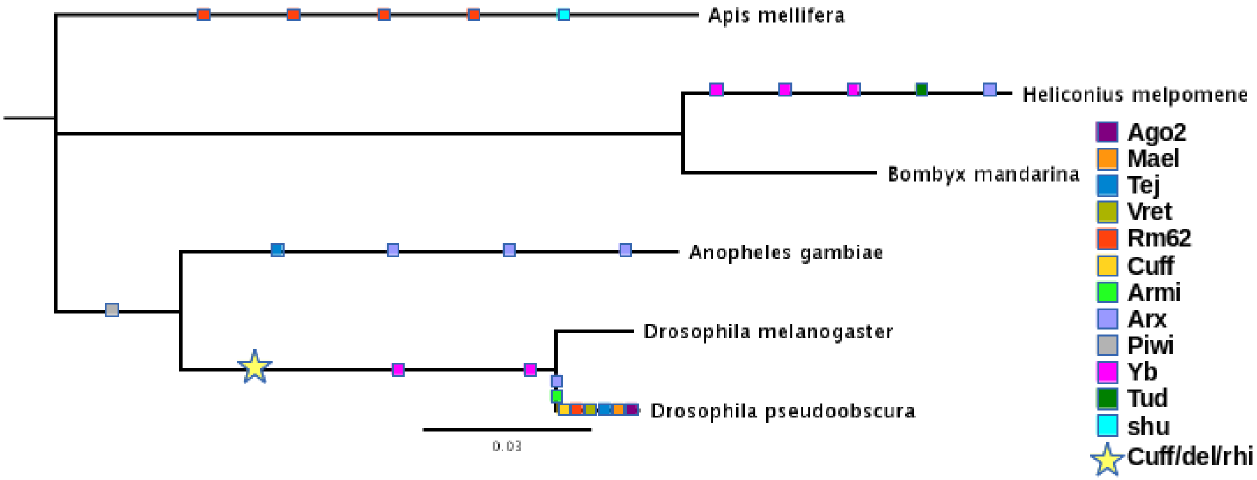


Figure S3 Possible duplications in RNAi pathway

Relationships of the insect species sampled, including coloured squares where possible gene duplications have occurred. Our search for RNAi genes in insect species other than *D. melanogaster* identified numerous duplications, and also some genes which were specific to *Drosophila*. Of note, *D. pseudoobscura* harboured duplications in *asterix, armitage, cutoff, rm62, vretano, tejas, maelstrom,* in addition to the multiple *AGO2* duplications reported previously (Lewis et al, 2016; Lewis et al, 2016), perhaps indicating an extensive addition to RNAi related pathways. *Asterix* was further duplicated three times in *Anopheles* and once in *Heliconius*, and *A. mellifera* also has five distinct copies of *rm62*. Furthermore, *yb* duplications have occurred independently in the lineage leading to *H. melpomene* and the one leading to the *Drosophila* species. The piRNA cluster transcriptional complex composed of cutoff, deadlock, and rhino were only observed in the two *Drosophila* species (represented by a star), and thus have likely either been lost in the other species or have evolved in since the split between *Anopheles* and *Drosophila*.

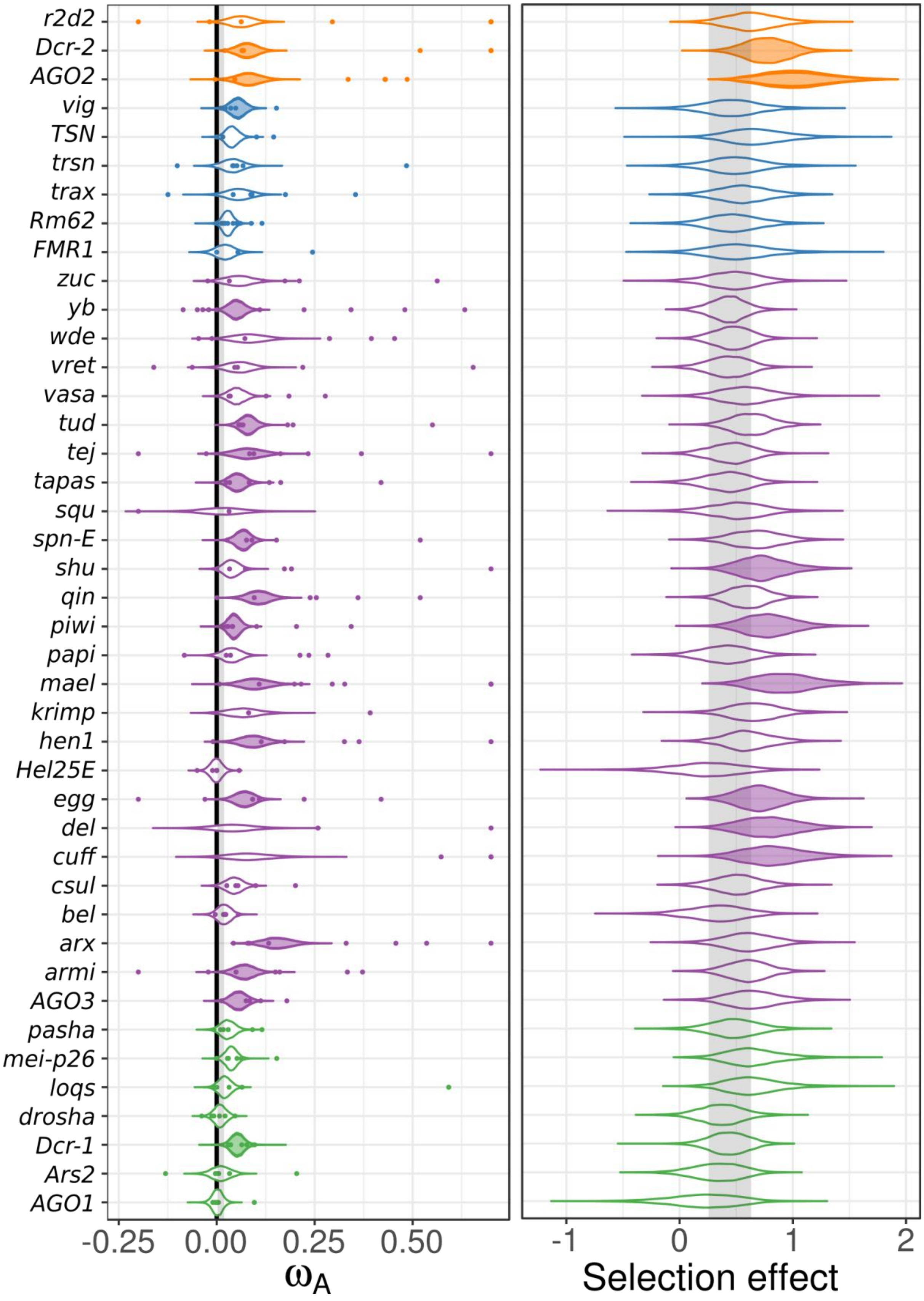

r2d2
Dcr-2
AGO2
vig
TSN
trsn
trax
Rm62
FMR1
zuc
yb
wde
vret
vasa
tud
tej
tapas
squ
spn-E
shu
qin
piwi
papi
mael
krimp
hen1
Hel25E
egg
del
cuff
csul
bel
arx
armi
AGO3
pasha
mei-p26
loqs
drosha
Dcr-1
Ars2
AGO1
-0.25
0.00
0.25
0.50
ω_A
-1
0
1
2
Selection effect

Figure S4 Cross-species homologue-level estimates of ω_A and selection effects without pathway assumptions

(Left) Individual gene ω_A estimates (coloured points) were calculated using DFE-alpha and analyses using a linear mixed model with species and gene as random effects (estimate uncertainty was included by incorporating bootstrap intervals as measurement error variance), but without subpathway as fixed effect (see Figure 4). The posterior distributions of the cross-species estimate for ω_A for each gene are plotted, and shaded if the MCMCp < 0.05 when tested against the control gene distribution (shaded grey region). Single-gene estimates of $\omega_A > 0.75$ are plotted at 0.75 for clarity. (Right) The analogous analysis, except performed using SnIPRE, with the posterior distribution of homologue-level selection effects plotted. Both analyses find *AGO2, Dicer-2, piwi, maelstrom,* and *eggless* as having elevated protein substitution.

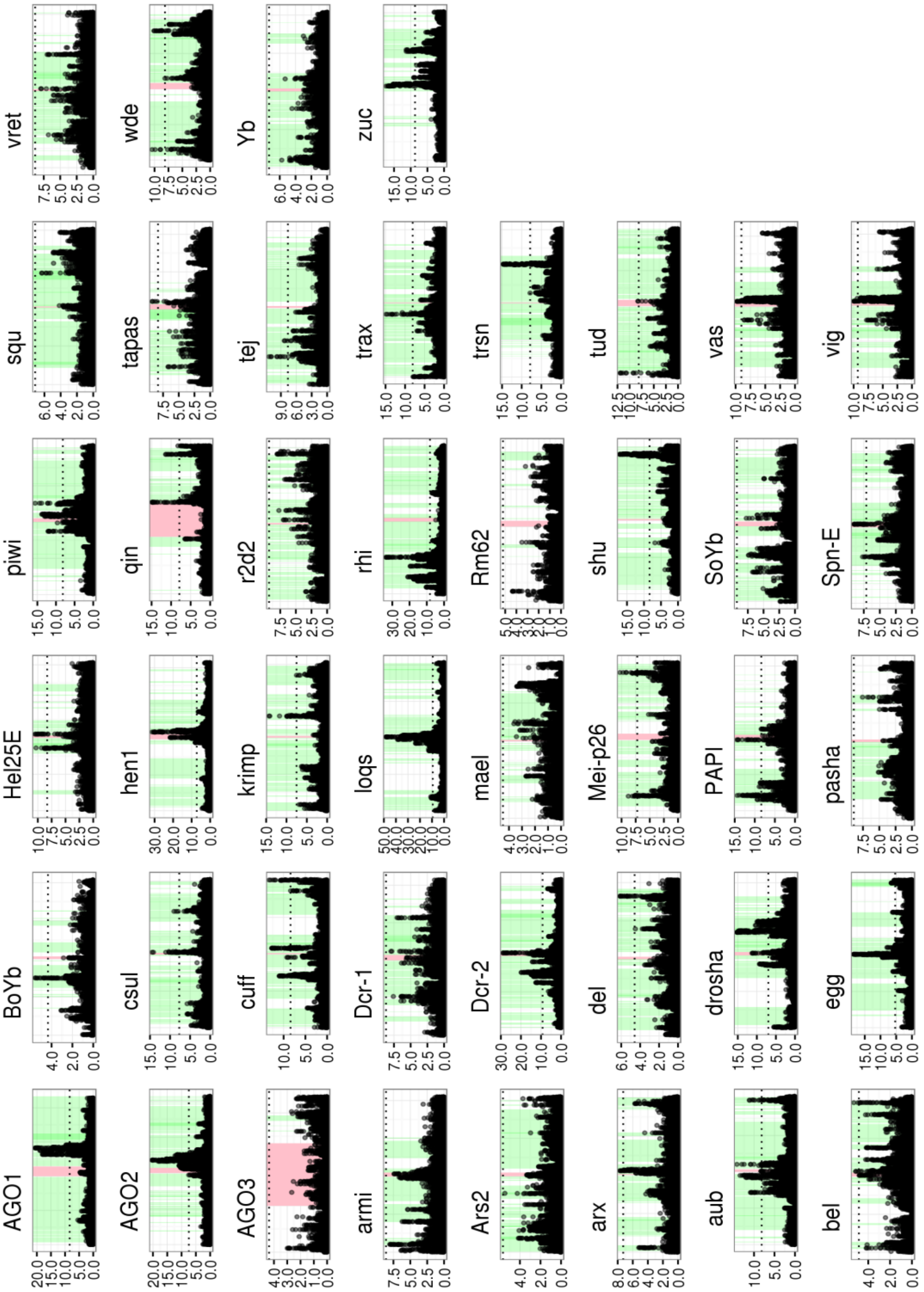


Figure S5: Drosophila melanogaster sweeps

For each *D. melanogaster* gene, the CLR statistic was plotted across a 200 kb region including the gene of interest. Each panel represents a region of the *D. melanogaster* genome, with red-shaded regions being the gene of interest and green-shaded regions being other genes along the chromosome. The horizontal dotted lines in each panel are significance thresholds calculated through neutral coalescent simulations.

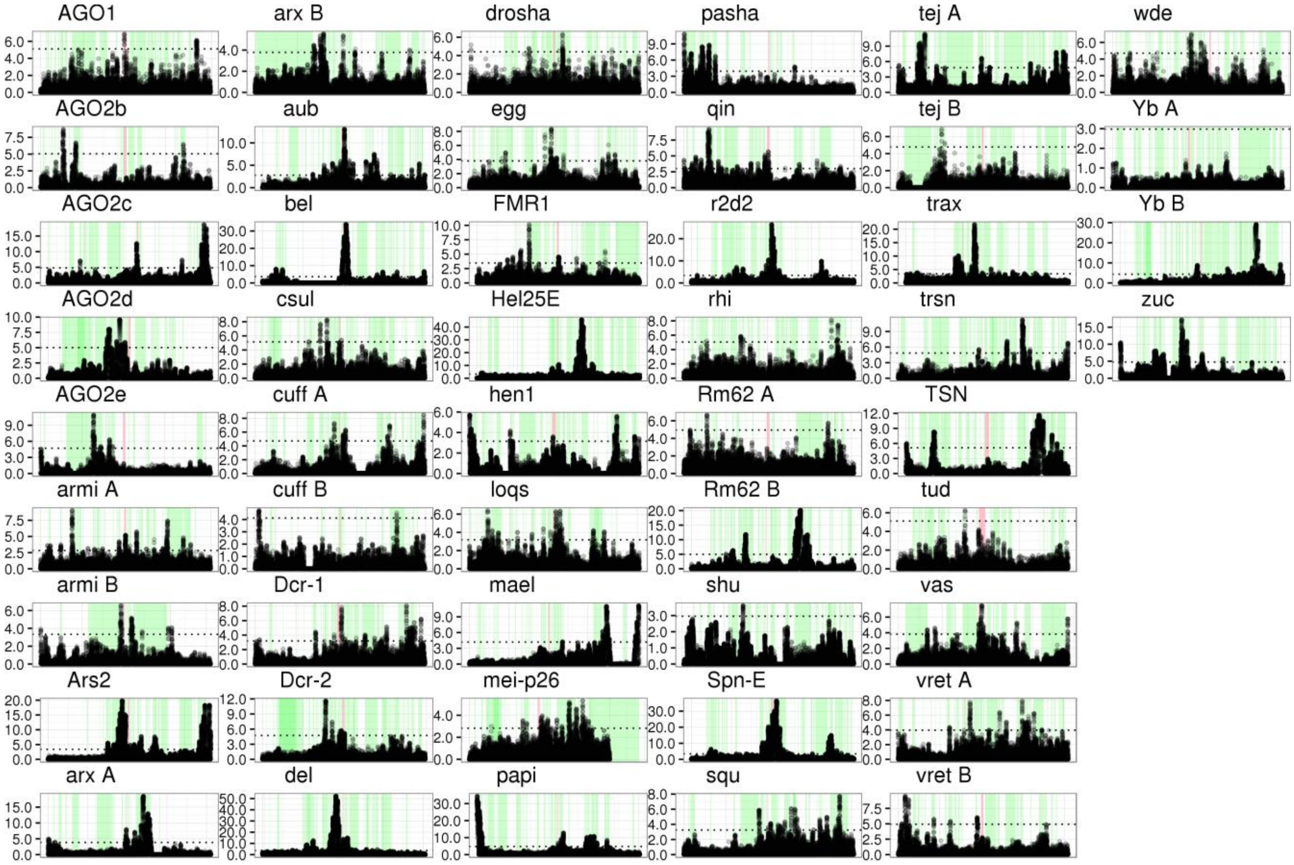


Figure S6: Drosophila pseudoobscura sweeps

For each *D. pseudoobscura* gene, the CLR statistic was plotted across a 200 kb region including the gene of interest. Each panel represents a region of the *D. pseudoobscura* genome, with red-shaded regions being the gene of interest and green-shaded regions being other genes along the chromosome. The horizontal dotted lines in each panel are significance thresholds calculated through neutral coalescent simulations.

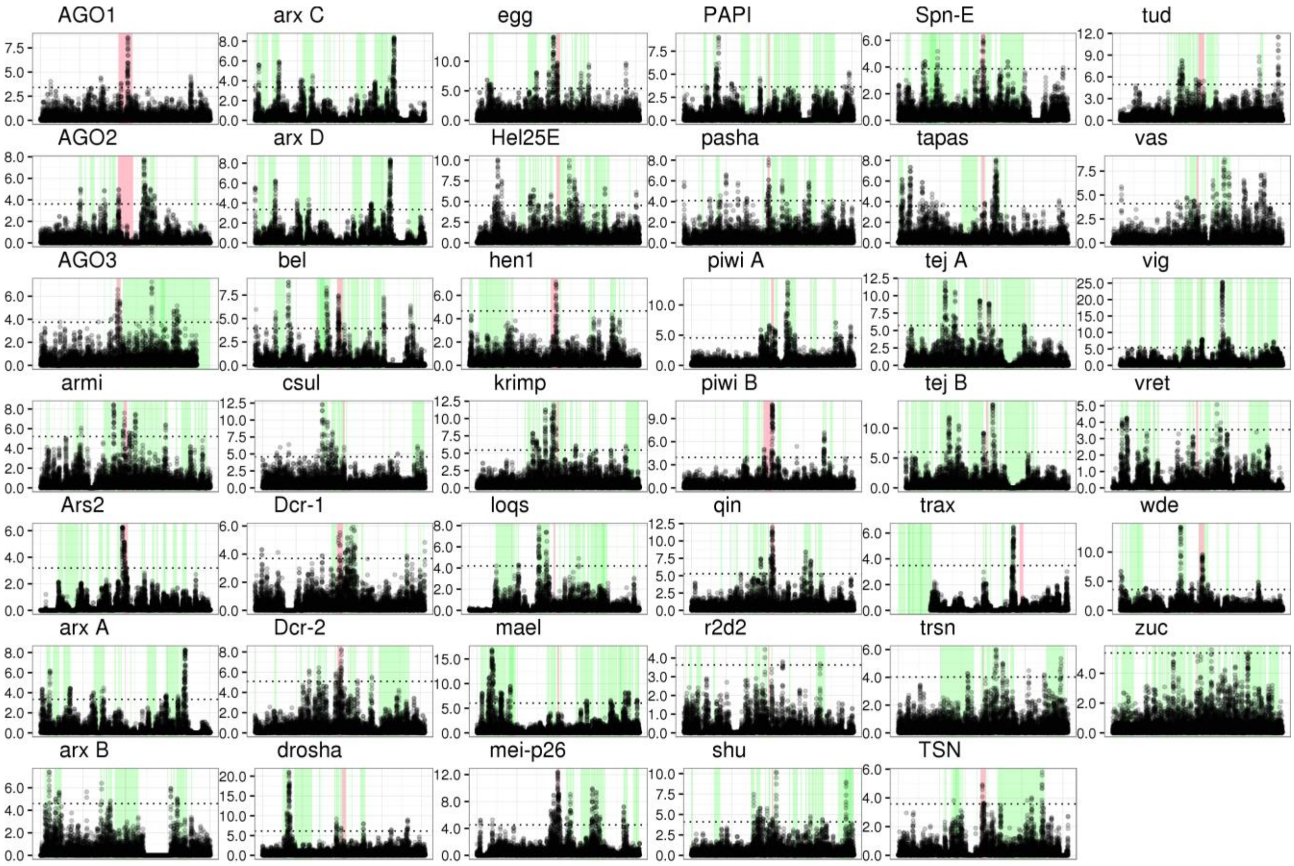


Figure S7: Anopheles gambiae sweeps

For each *A. gambiae* gene, the CLR statistic was plotted across a 200 kb region including the gene of interest. Each panel represents a region of the *A. gambiae* genome, with red-shaded regions being the gene of interest and green-shaded regions being other genes along the chromosome. The horizontal dotted lines in each panel are significance thresholds calculated through neutral coalescent simulations.

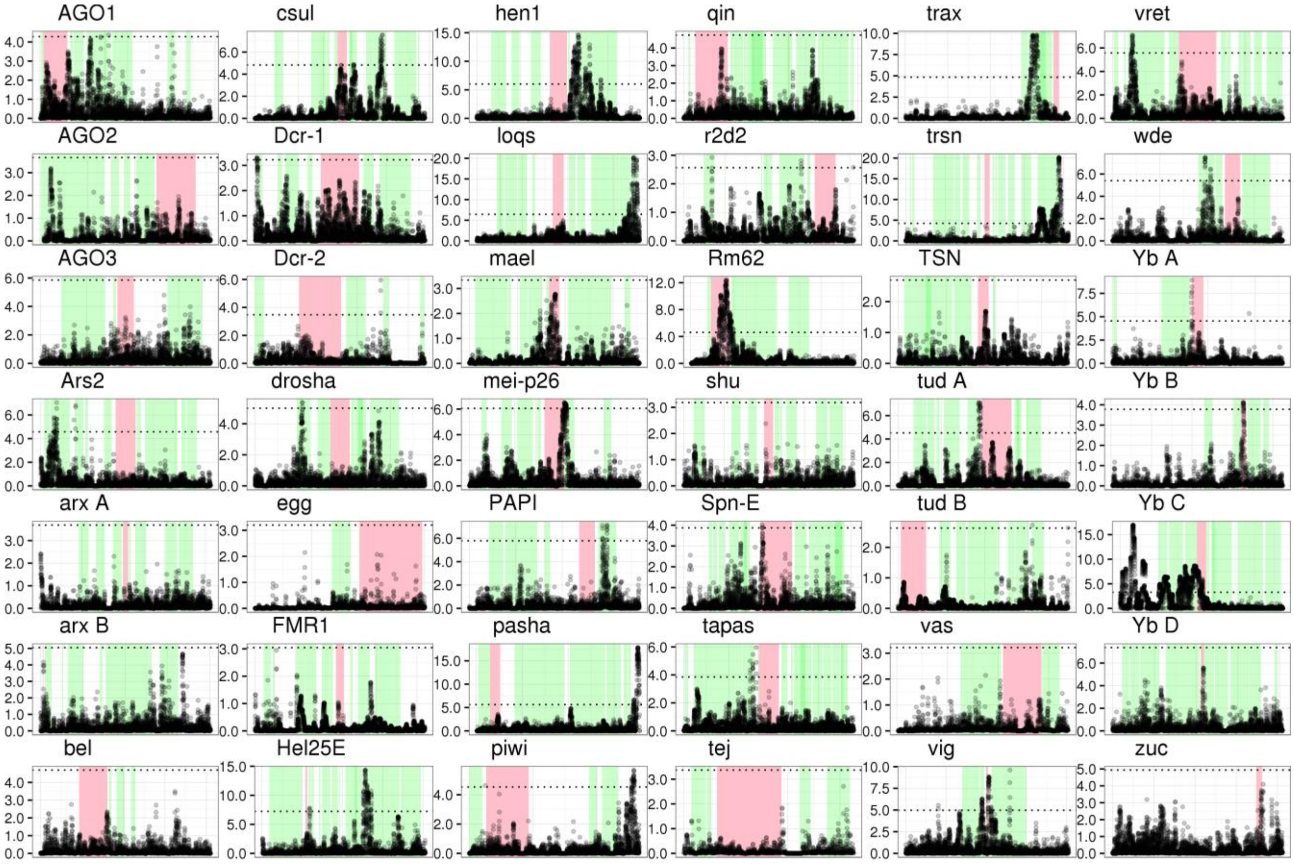


Figure S8: Heliconius melpomene sweeps

For each *H. melpomene* gene, the CLR statistic was plotted across a 200 kb region including the gene of interest. Each panel represents a region of the *H. melpomene* genome, with red-shaded regions being the gene of interest and green-shaded regions being other genes along the chromosome. The horizontal dotted lines in each panel are significance thresholds calculated through neutral coalescent simulations.

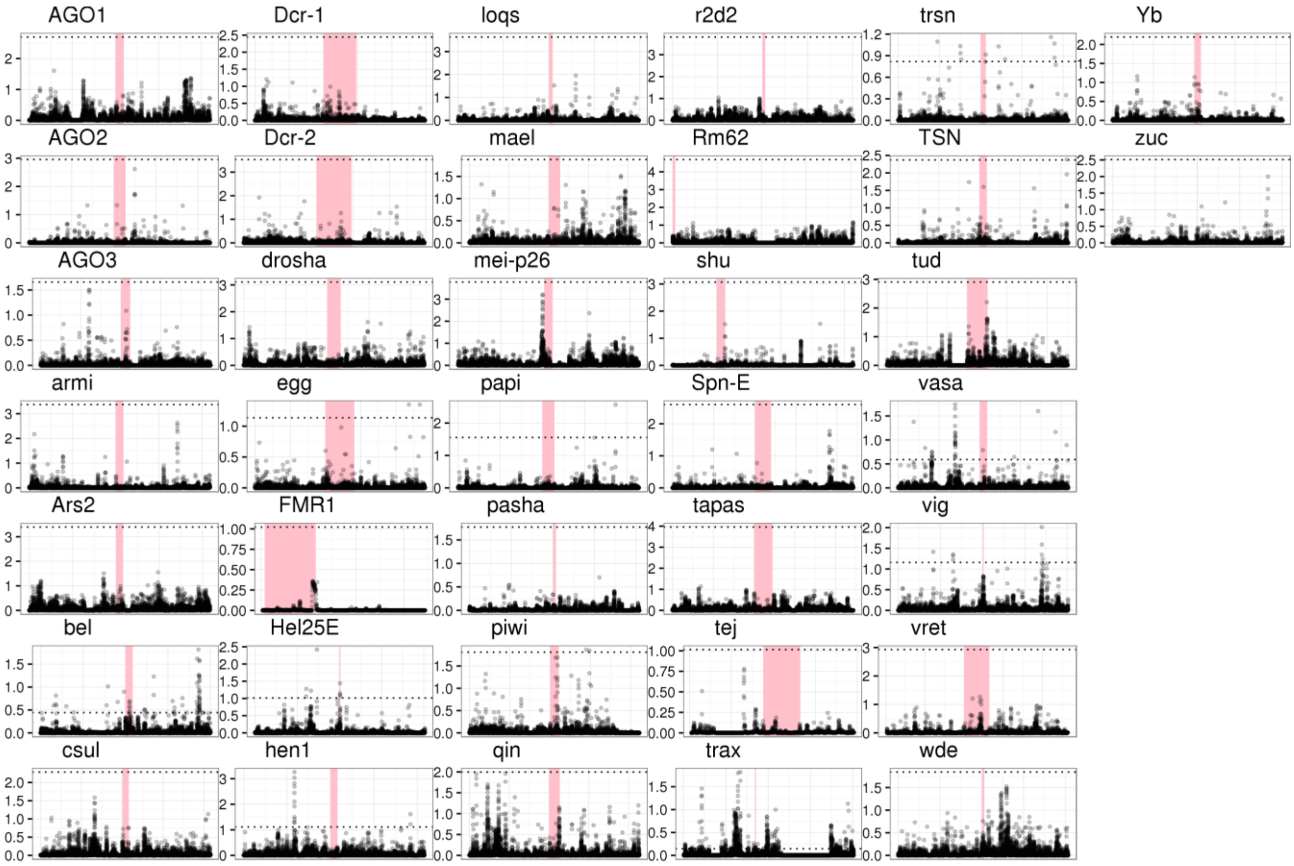


Figure S9: Bombyx mandarina sweeps

For each *B. mandarina* gene, the CLR statistic was plotted across a 200 kb region including the gene of interest. Each panel represents a region of the *B. mandarina* genome, with red-shaded regions being the gene of interest. The horizontal dotted lines in each panel are significance thresholds calculated through neutral coalescent simulations. The *Bombyx* genome used did not have an associated gff file, and so positions of nearby genes were not included.

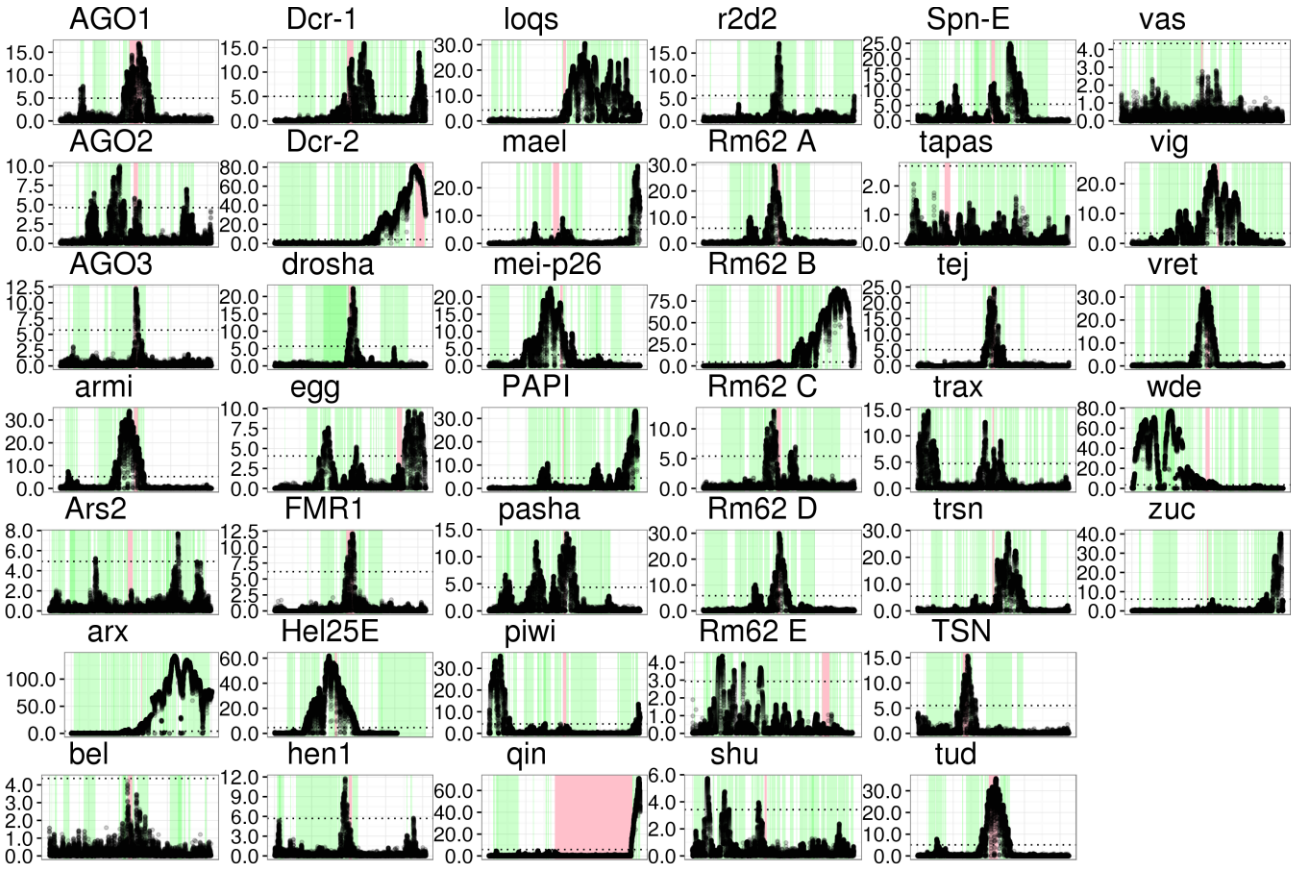


Figure S10: Apis mellifera sweeps

For each *A. mellifera* gene, the CLR statistic was plotted across a 200 kb region including the gene of interest. Each panel represents a region of the *A. mellifera* genome, with red-shaded regions being the gene of interest. The horizontal dotted lines in each panel are significance thresholds calculated through neutral coalescent simulations.

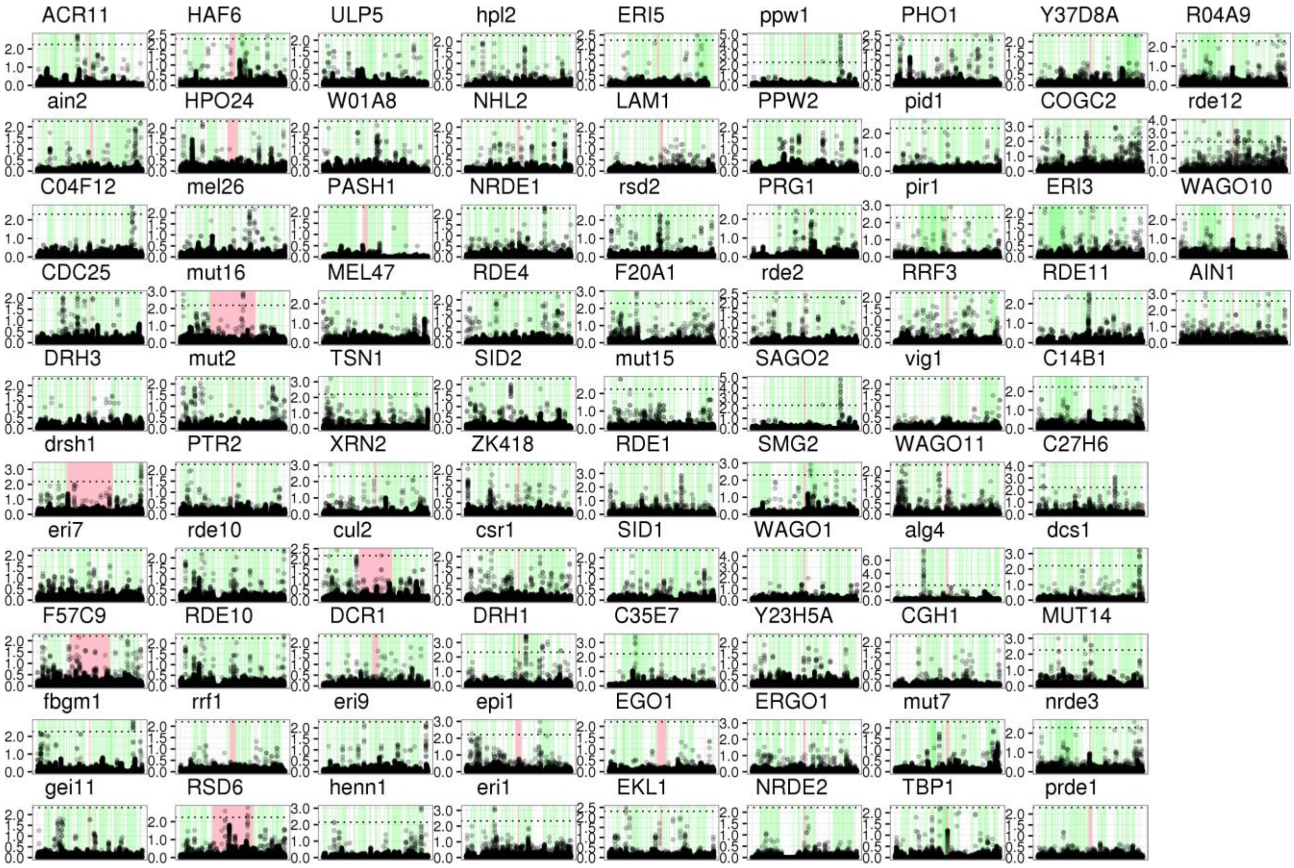


Figure S11: Caenorhabditis briggsae sweeps

For each *C. briggsae* gene, the CLR statistic was plotted across a 200 kb region including the gene of interest. Each panel represents a region of the *C. briggsae* genome, with red-shaded regions being the gene of interest. The horizontal dotted lines in each panel are significance thresholds calculated through neutral coalescent simulations.

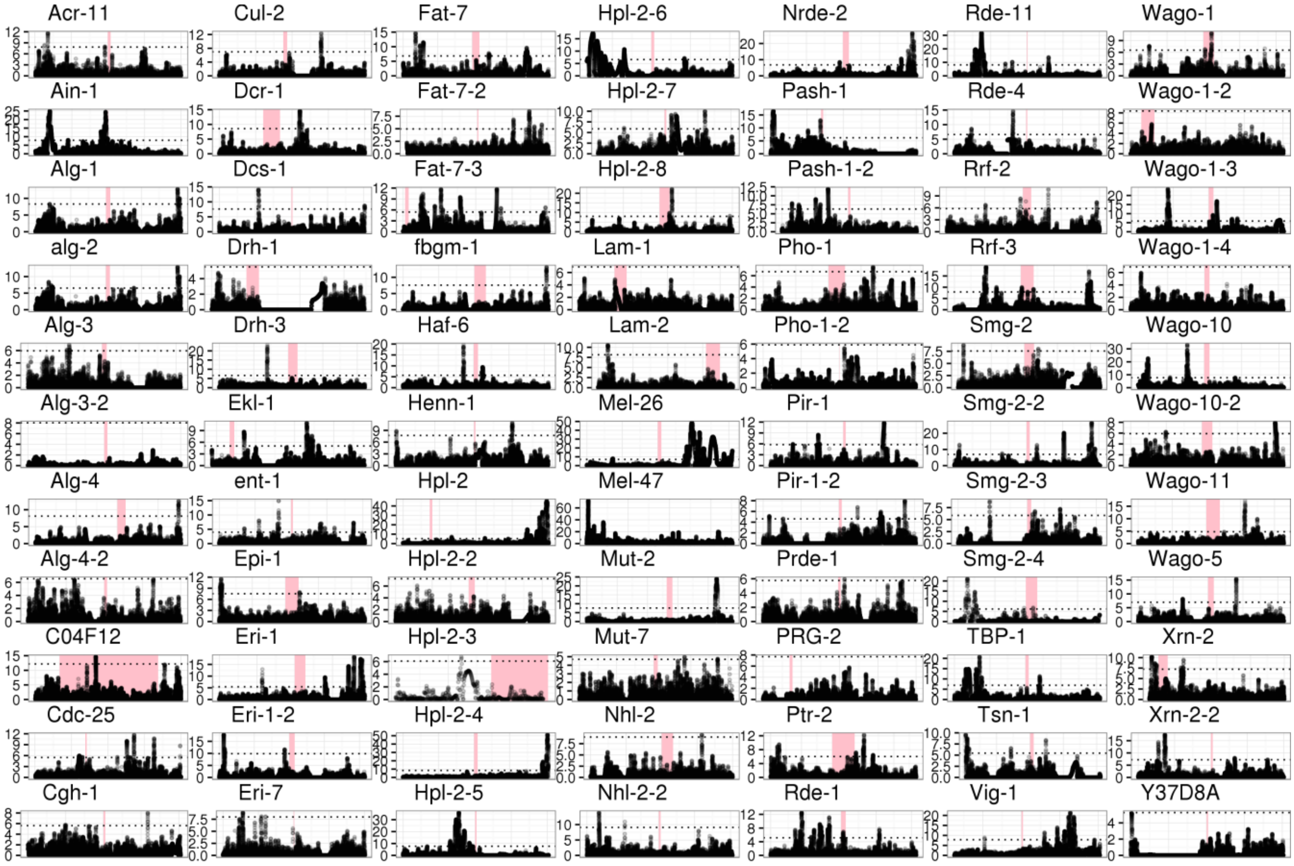
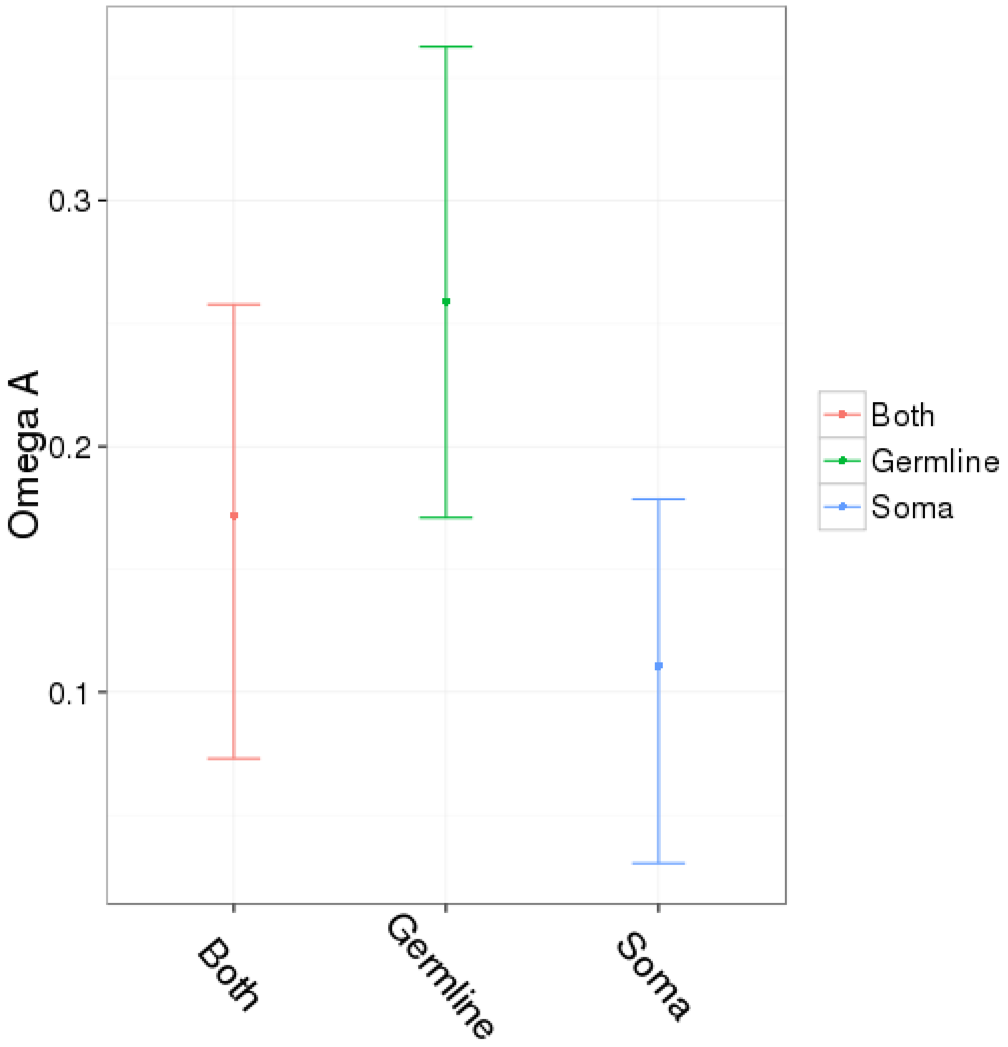


Figure S12: *Pristionchus pacificus* sweeps

For each *P. pacificus* gene, the CLR statistic was plotted across a 200 kb region including the gene of interest. Each panel represents a region of the *P. pacificus* genome, with red-shaded regions being the gene of interest. The horizontal dotted lines in each panel are significance thresholds calculated through neutral coalescent simulations. The *Pristionchus* genome used did not have an associated gff file, so positions of nearby genes were not included.



Supp Figure 13: Germline and somatic piRNA pathway genes

Polymorphism and divergence from a larger set of piRNA pathway genes (those identified in two of three recent piRNA pathway screens, plus the core piRNA pathway) (Handler et al, 2013; Czech et al, 2013; Muerdter et al, 2013) in *D. melanogaster* were pooled based on whether they are active in the germline, soma, or both and used to calculate ω_A. Confidence intervals were obtained by bootstrapping by gene 1000 times. Genes active in germline TE suppression show higher rates of adaptive protein evolution than those active in the somatic follicle cells, with genes active in both having an intermediate adaptive rate.

S1 Text: Supplementary R code for models

DFE-alpha meta-analysis

Data set up:

```
library(MCMCglmm)

## Loading required package: Matrix

## Loading required package: coda

## Loading required package: ape

#########################################################
# Data Upload for gene-level omega.A linear mixed models ----------------------
---------------------------------------
dat<-read.table("dfe-alpha-ind_withVIRNA.csv", sep=",", header=TRUE)
dat$omega_A[which(dat$omega_A==-Inf)]<-NA #Remove genes that can't be estimated
. The results are unaffected if these values are set to zero.
dat<-subset(dat, !is.na(omega_A) & !is.na(omega_se) )
```

Model 1A: Comparison of RNAi and control genes

```
prior.1A=list(R=list(V=diag(2),
                      nu=0.002),
               G=list(G1=list(V=diag(1),
                              nu=1,
                              alpha.mu=c(0),
                              alpha.V=diag(1))))

model.1A <-MCMCglmm(omega_A~Class,
                   random= ~Organism,
                   rcov = ~idh(Class):units,
                   mev=dat$omega_se^2,
                   prior=prior.1A,
                   data=dat, verbose = FALSE)
```

Model 1A models gene-level ωA estimates as a gaussian response with gene class as a fixed effect and species as a random effect. For the random effects (random= ~Organism), we assume all (co)variances among organisms are equal. We also estimate separate error variances for each gene class (rcov = ~idh(Class):units), allowing us to test whether the variance of the adaptive rate of RNAi and control genes differ. The idh() function specifies that the residual variance associated with each class of genes is independent, and sets the off-diagonals of the covariance matrix to zero. We specify the sampling error associated with each estimate of ωA (mev=dat$omega_se^2) obtained by bootstrapping by codon and rerunning DFE-alpha on the new codon set.

```
summary(model.1A)

##
##  Iterations = 3001:12991
##  Thinning interval  = 10
##  Sample size  = 1000
##
##  DIC: -1537.862
##
##  G-structure:  ~Organism
##
##          post.mean  l-95% CI  u-95% CI eff.samp
## Organism 8.451e-05 1.994e-10 0.0002989    518.6
##
##  R-structure:  ~idh(Class):units
##
##                    post.mean  l-95% CI  u-95% CI eff.samp
## ClassControl.units 0.0003228 0.0001796 0.0004817    705.2
## ClassRNAi.units    0.0036677 0.0021694 0.0052551    596.8
##
##  Location effects: omega_A ~ Class
##
##             post.mean  l-95% CI  u-95% CI eff.samp  pMCMC
## (Intercept) 0.0095622 0.0002037 0.0188204     1000  0.028 *
## ClassRNAi   0.0530510 0.0382375 0.0662674     1000 <0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The command summary(model.1A) prints some aspects of the MCMC chain, the DIC score, the variance components for the G-structure (random effects), the error variance estimates, and the estimates for the fixed effects. The RNAi class of genes is estimated to be 0.05 greater than control genes, and this is signficant (pMCMC < 0.001) We also test whether the variance is significantly greater for RNAi genes using the posterior distributions for the error variances saved in the VCV object.

```
head(model.1A$VCV)[,c("ClassControl.units", "ClassRNAi.units")]

## Markov Chain Monte Carlo (MCMC) output:
## Start = 3001
## End = 3061
## Thinning interval = 10
##      ClassControl.units ClassRNAi.units
## [1,]       0.0002867517     0.002768112
## [2,]       0.0002296532     0.004182672
## [3,]       0.0002213186     0.002911138
## [4,]       0.0002727306     0.003119912
## [5,]       0.0002874699     0.002956077
## [6,]       0.0003067956     0.004262320
## [7,]       0.0002863971     0.002101608
```

To test for significantly different variances, we subtract one posterior distribution from the other and ask what proportion overlaps zero.

```
iterations.less.than.zero <- length(which(model.1A$VCV[,"ClassControl.units"] -
model.1A$VCV[,"ClassRNAi.units"] > 0))
total.chain.length <- nrow(model.1A$VCV)
pMCMC <- iterations.less.than.zero/total.chain.length
pMCMC*2

## [1] 0
```

Therefore, in every iteration of the chain (1000 sampled iterations, thinning interval of 10), the error variance associated with RNAi genes was greater than control genes (MCMCp < 0.001).

Model 1B: Comparison of RNAi subpathways (piRNA, siRNA, viRNA, miRNA)

```
prior.1B=list(R=list(V=diag(5),
                     nu=0.002),
              G=list(G1=list(V=diag(1),
                             nu=1,
                             alpha.mu=c(0),
                             alpha.V=diag(1))))
model.1B <-MCMCglmm(omega_A~Subclass,
                    random= ~Organism,
                    data=dat,
                    mev=dat$omega_se^2,
                    rcov = ~idh(Subclass):units,
                    prior=prior.1B, verbose = FALSE)
```

Model 1B is similar to model 1A, except the RNAi class has now been divided into four subpathways (miRNA, siRNA, piRNA, viRNA). The summary of the model output is the following:

```
summary(model.1B)

## 
##  Iterations = 3001:12991
##  Thinning interval  = 10
##  Sample size  = 1000 
## 
##  DIC: -1529.198 
## 
##  G-structure:  ~Organism
## 
##          post.mean  l-95% CI  u-95% CI eff.samp
## Organism 7.741e-05 4.421e-11 0.0003035    673.8
## 
##  R-structure:  ~idh(Subclass):units
## 
##                       post.mean  l-95% CI  u-95% CI eff.samp
## SubclassControl.units 0.0003252 0.0001822 0.0004842    553.1
## Subclassmi.units      0.0006925 0.0002427 0.0012781    990.4
## Subclasspi.units      0.0063899 0.0037257 0.0094438    661.8
## Subclasssi.units      0.0012648 0.0003426 0.0024599   1000.0
## Subclassvi.units      0.0391034 0.0081908 0.0842066   1000.0
##
##  Location effects: omega_A ~ Subclass
##
##              post.mean   l-95% CI   u-95% CI eff.samp  pMCMC
## (Intercept)  0.0096087  0.0002142  0.0178675    902.3  0.036 *
## Subclassmi   0.0108465 -0.0007559  0.0246557   1000.0  0.082 .
## Subclasspi   0.0794998  0.0586724  0.1002338   1000.0 <0.001 ***
## Subclasssi   0.0301952  0.0087547  0.0480110   1000.0  0.002 **
## Subclassvi   0.1863027  0.0690058  0.3184570   1000.0  0.002 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The subclass effects (parameterised as "mi", "pi", "si", and "vi"), along with their pMCMC values are listed under location effects. Differences in residual variances between subpathways can be tested in a similar manner to model 1A. We tested whether certain subpathways were greater than others as previously done with variance components in Model 1A, except using the posteriors for the fixed effects stored in the Sol object. For example:

```
iterations.less.than.zero <- length(which(model.1B$Sol[,"Subclassmi"] - model.1
B$Sol[,"Subclassvi"] > 0))
total.chain.length <- nrow(model.1B$Sol)
pMCMC <- iterations.less.than.zero/total.chain.length
pMCMC*2

## [1] 0.002
```

We conclude that the viRNA pathway has a significantly greater rate of adaptive protein evolution (MCMCp = 0.002).

Model 1C: Comparison of RNAi subpathways, with the piRNA split into effectors, biogenesis factors, and transcriptional silencing factors.

```
dat$Subclass_pi <- factor(dat$Subclass_pi, levels=c("Control", "mi", "si","effe
ctor", "transcriptional", "biogenesis","vi"))

prior.1C=list(R=list(V=diag(7), nu=0.002),
               G=list(G1=list(V=diag(1), nu=1, alpha.mu=c(0), alpha.V=diag(1)))
)
model.1C <-MCMCglmm(omega_A~Subclass_pi,
                   random= ~Organism,
                   rcov = ~idh(Subclass_pi):units,
                   mev=dat$omega_se^2,
                   data=dat,
                   prior=prior.1C, verbose = FALSE, nitt = 50000)
```

Again, Model 1C is structurally identical to Model 1A and Model 1B, with the only difference being the number of factor levels which genes are grouped into.

```
summary(model.1C)

##
##  Iterations = 3001:49991
##  Thinning interval  = 10
##  Sample size  = 4700
##
##  DIC: -1533.084
##
##  G-structure:  ~Organism
##
##          post.mean  l-95% CI  u-95% CI eff.samp
## Organism 6.068e-05 1.341e-11 0.0002115     2955
##
##  R-structure:  ~idh(Subclass_pi):units
##
##                                 post.mean  l-95% CI  u-95% CI eff.samp
## Subclass_piControl.units        0.0003235 0.0001818 0.0004742     3475
## Subclass_pimi.units             0.0005880 0.0002091 0.0010588     4254
## Subclass_pisi.units             0.0012564 0.0003701 0.0024096     4342
## Subclass_pieffector.units       0.0071395 0.0017279 0.0144296     4700
## Subclass_pitranscriptional.units 0.0158533 0.0003308 0.0397985     2898
## Subclass_pibiogenesis.units     0.0069787 0.0032832 0.0112018     3689
## Subclass_pivi.units             0.0391355 0.0080485 0.0820324     4700
##
##  Location effects: omega_A ~ Subclass_pi
##
##                           post.mean  l-95% CI  u-95% CI eff.samp   pMCMC
## (Intercept)                0.009687  0.001916  0.018081     4132 0.02383
## Subclass_pimi              0.010512 -0.001300  0.022172     4700 0.06638
## Subclass_pisi              0.030278  0.011611  0.051474     4315 0.00383
## Subclass_pieffector        0.079908  0.036203  0.125608     4700 < 2e-04
## Subclass_pitranscriptional  0.154907  0.074015  0.242532     4013 < 2e-04
## Subclass_pibiogenesis      0.078657  0.054069  0.106494     4700 < 2e-04
## Subclass_pivi              0.189578  0.067866  0.309402     5537 0.00340
##
## (Intercept)               *
## Subclass_pimi             .
## Subclass_pisi             **
## Subclass_pieffector       ***
## Subclass_pitranscriptional ***
## Subclass_pibiogenesis     ***
## Subclass_pivi             **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All three piRNA pathways are significantly greater than control genes. Significance between subpathways and variance components was assessed as above.

Model 2A: Comparison of RNAi homologues (with subpathway as a fixed effect)

```
prior.2A=list(R=list(V=diag(5), nu=0.002),
             G=list(G1=list(V=diag(4),
                            nu=1,
                            alpha.mu=rep(0,4),
                            alpha.V=diag(4)),
                    G2=list(V=diag(1),
                            nu=1,
                            alpha.mu=rep(0,1),
                            alpha.V=diag(1))))
model.2A <-MCMCglmm(omega_A~Subclass,
                   random=~idh(at.level(Subclass, c('mi', 'pi', 'si', 'vi'))):G
ene + Organism,
                   rcov =~idh(Subclass):units,
                   data=dat, mev=dat$omega_se^2, pr=TRUE, prior = prior.2A, ver
bose = FALSE)
```

The second model is similar to the first, but for two differences. First, the effect of each RNAi homologue is estimated across species, specifying idh(at.level(Subclass, c('mi', 'pi', 'si', '.vi'))) so that the homologue effect is not estimated for the (nonhomologous) control genes. Second, we specify pr=TRUE, so that the random effects are stored along with the fixed effects in the model output.

```
summary(model.2A)

## 
##  Iterations = 3001:12991
##  Thinning interval  = 10
##  Sample size  = 1000 
## 
##  DIC: -1542.535 
## 
##  G-structure:  ~idh(at.level(Subclass, c("mi", "pi", "si", "vi"))):Gene
## 
##                                                     post.mean  l-95% CI
## at.level(Subclass, c("mi", "pi", "si", "vi"))1.Gene 0.0004272 1.394e-11
## at.level(Subclass, c("mi", "pi", "si", "vi"))2.Gene 0.0006833 8.664e-09
## at.level(Subclass, c("mi", "pi", "si", "vi"))3.Gene 0.0005370 2.340e-13
## at.level(Subclass, c("mi", "pi", "si", "vi"))4.Gene 0.1036248 2.972e-08
##                                                     u-95% CI eff.samp
## at.level(Subclass, c("mi", "pi", "si", "vi"))1.Gene 0.001419    649.3
## at.level(Subclass, c("mi", "pi", "si", "vi"))2.Gene 0.002115   1000.0
## at.level(Subclass, c("mi", "pi", "si", "vi"))3.Gene 0.002051    438.2
## at.level(Subclass, c("mi", "pi", "si", "vi"))4.Gene 0.325240   1000.0
## 
##                 ~Organism
## 
##          post.mean  l-95% CI  u-95% CI eff.samp
## Organism 6.993e-05 9.763e-11 0.0002299    514.5
## 
##  R-structure:  ~idh(Subclass):units
## 
##                       post.mean  l-95% CI  u-95% CI eff.samp
## SubclassControl.units 0.0003251 0.0001810 0.0004723    698.9
## Subclassmi.units      0.0005521 0.0001729 0.0010869    771.4
## Subclasspi.units      0.0060251 0.0033111 0.0094265    688.9
## Subclasssi.units      0.0012674 0.0003819 0.0025295   1000.0
## Subclassvi.units      0.0431069 0.0083074 0.0930210   1000.0
## 
##  Location effects: omega_A ~ Subclass 
## 
##             post.mean  l-95% CI  u-95% CI eff.samp  pMCMC    
## (Intercept)  0.009849  0.002151  0.018498     1000  0.024 *  
## Subclassmi   0.011261 -0.006067  0.033797     1000  0.220    
## Subclasspi   0.081436  0.058334  0.103126     1000 <0.001 ***
## Subclasssi   0.031000  0.005495  0.065159     1000  0.044 *  
## Subclassvi   0.188296 -0.077555  0.471909     1000  0.108    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The posterior distribution of the fixed and random effects of the model are stored in the object named model.2A$Sol, with columns pertaining to each distribution. For example:

```
colnames(data.frame(model.2A$Sol))[c(1:10, 44:50)] #Show columns of interest
```

```
##  [1] "X.Intercept."
##  [2] "Subclassmi"
##  [3] "Subclasspi"
##  [4] "Subclasssi"
##  [5] "Subclassvi"
##  [6] "at.level.Subclass..c..mi....pi....si....vi...1.Gene.ago1"
##  [7] "at.level.Subclass..c..mi....pi....si....vi...1.Gene.ars2"
##  [8] "at.level.Subclass..c..mi....pi....si....vi...1.Gene.dcr1"
##  [9] "at.level.Subclass..c..mi....pi....si....vi...1.Gene.drosha"
## [10] "at.level.Subclass..c..mi....pi....si....vi...1.Gene.loqs"
## [11] "at.level.Subclass..c..mi....pi....si....vi...3.Gene.tsn"
## [12] "at.level.Subclass..c..mi....pi....si....vi...3.Gene.vig"
## [13] "at.level.Subclass..c..mi....pi....si....vi...4.Gene.ago2"
## [14] "at.level.Subclass..c..mi....pi....si....vi...4.Gene.dcr2"
## [15] "at.level.Subclass..c..mi....pi....si....vi...4.Gene.r2d2"
## [16] "Organism.Anopheles"
## [17] "Organism.Apis"
```

We obtain the posterior distribution for ωA of an individual homologue by adding the posterior distributions for the intercept, subclass, and homologue. For example, the ωA posterior for Argonaute-2 is:

```
ago2.posterior <- model.2A$Sol[,"(Intercept)"] + model.2A$Sol[,"Subclassvi"] +
model.2A$Sol[,"at.level(Subclass, c(\"mi\", \"pi\", \"si\", \"vi\"))4.Gene.ago2
"]
```

We obtain 95% HPD confidence intervals using the command HPDinterval():

```
HPDinterval(ago2.posterior)
```

```
##           lower     upper
## var1 0.05833242 0.4022962
## attr(,"Probability")
## [1] 0.95
```

This show the lower 95% HPD interval (0.07) is greater than 0. We test whether this is greater than control genes by subtracting the posterior of ωA estimates of Argonaute-2 from the control gene class posterior, and see the proportion of MCMC intervals where it overlaps zero.

```
control.posterior <- model.2A$Sol[,"(Intercept)"]
control.minus.ago2.posterior <-  control.posterior - ago2.posterior

iterations.less.than.zero <- length(which(control.minus.ago2.posterior > 0))
total.chain.length <- length(control.minus.ago2.posterior)
pMCMC <- iterations.less.than.zero/total.chain.length
pMCMC*2 #Multiply by 2 to make a two tailed test
```

```
## [1] 0.012
```

We conclude Argonaute-2 has a greater adaptive rate than control genes (pMCMC = 0.012).

Model 2B: Comparison of RNAi genes

```
prior.2B=list(R=list(V=diag(1), nu=0.002),
             G=list(G1=list(V=diag(1),
                            nu=1,
                            alpha.mu=rep(0,1),
                            alpha.V=diag(1)),
                    G2=list(V=diag(1),
                            nu=1,
                            alpha.mu=rep(0,1),
                            alpha.V=diag(1))))
model.2B <-MCMCglmm(omega_A~1,
                   random=~idv(at.level(Class, c('RNAi'))):Gene + Organism,
                   data=dat, mev=dat$omega_se^2, pr=TRUE, prior = prior.2B, ver
bose=FALSE, nitt = 50000)
```

We also parameterise the second model without assigning genes to subpathways, and remove the subclass fixed effect and subclass-specific error-variances. The following is the output of the model:

```
summary(model.2B)

##
##  Iterations = 3001:49991
##  Thinning interval  = 10
##  Sample size  = 4700
##
##  DIC: -1749.396
##
##  G-structure:  ~idv(at.level(Class, c("RNAi"))):Gene
##
##                                post.mean l-95% CI u-95% CI eff.samp
## at.level(Class,c("RNAi")).Gene  0.003086 0.001022 0.005389      973
##
##                ~Organism
##
##          post.mean  l-95% CI u-95% CI eff.samp
## Organism 0.0003451 5.648e-06  0.00105    917.2
##
##  R-structure:  ~units
##
##       post.mean  l-95% CI  u-95% CI eff.samp
## units 0.0006735 0.0003879 0.0009833     1893
##
##  Location effects: omega_A ~ 1
##
##             post.mean l-95% CI u-95% CI eff.samp  pMCMC
## (Intercept)  0.019337 0.002746 0.035655     4089 0.0281 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Without a subclass effect, we assess whether a gene has a significantly elevated rate of adaptive amino acid evolution by comparing the posterior distributions of each gene effect to zero. For example, for Argonaute-2:

```
ago2.posterior <- model.2B$Sol[,"at.level(Class, c(\"RNAi\")).Gene.ago2"]
pMCMC <- length(which(ago2.posterior < 0))/length(ago2.posterior)
pMCMC*2

## [1] 0.04723404
```

We conclude that Argonaute-2 has significantly elevated ωA (pMCMC = 0.047).

SnIPRE-like analysis

Model 3A: SnIPRE-like analysis, with subpathway as a fixed effect

Set up data

```
library(MCMCglmm)
library(MASS)
dat<-read.csv("C:/Users/willi/Desktop/xspecies_rnaigene_counts_withpiRNA.csv",
sep = " ")
dat$nfac<-as.factor(paste(dat$divergence, dat$nonsynonymous))
dat$length<-dat$Ln
dat$length[which(dat$synonymous==1)]<-dat$Ls[which(dat$synonymous==1)]
dat$gene.id<-paste(dat$organism, dat$gene, dat$Duplicate)
dat<-dat[-which(dat$organism=="bombyx" & dat$gene=="vas"),] #This gene had a pr
oblematic alignment
missing<-table(dat$gene.id,dat$nfac)
missing<-missing[which(rowSums(missing)!=4),] #Remove genes with missing data

for(i in 1:nrow(missing)){
  combi<-strsplit(rownames(missing)[i], " ")[[1]]
  combj<-which(missing[i,]==0)

  for(j in 1:length(combj)){
     dat[nrow(dat)+1,]<-dat[which(dat$organism==combi[1] & dat$gene==combi[2] &
dat$Duplicate==combi[3])[1],]
     dat[nrow(dat),"nfac"]<-colnames(missing)[combj[j]]
     dat[nrow(dat),"count"]<-0
     dat[nrow(dat),"length"]<-1
     dat[nrow(dat),"divergence"]<-as.numeric(substr(colnames(missing)[combj[j]]
,1,1))
     dat[nrow(dat),"nonsynonymous"]<-as.numeric(substr(colnames(missing)[combj[
j]],3,3))
   }
}
rownames(dat)<-1:nrow(dat)
dat$organism <- factor(dat$organism, levels = c("dmel", "anopheles", "apis", "b
ombyx", "dpse", "heliconius"))
```

The data look like the following. Each gene in each species is represented by 4 rows of count data, one for each of the MK observations of polymorphism and divergence by synonymous and nonsynonymous mutation types.

```
head(dat[dat$RNAi==1,c(1,2,3,6,7,10,12:19)])

##     gene count divergence nonsynonymous  organism Duplicate piRNA siRNA
## 465 ago1     3          1             1 anopheles         A     0     0
## 466 ago1    10          1             0 anopheles         A     0     0
## 467 ago1     8          0             0 anopheles         A     0     0
## 468 ago1     2          0             1 anopheles         A     0     0
## 469 ago1     0          1             1      apis         A     0     0
## 470 ago1    39          1             0      apis         A     0     0
##     miRNA viRNA effector transcriptional biogenesis nfac
## 465     1     0        0               0          0  1 1
## 466     1     0        0               0          0  1 0
## 467     1     0        0               0          0  0 0
## 468     1     0        0               0          0  0 1
## 469     1     0        0               0          0  1 1
## 470     1     0        0               0          0  1 0

prior.3A<-list(R=list(V=diag(4), nu=0.002),
              G=list(G1=list(V=diag(4),
                             nu=4,
                             alpha.mu=rep(0,4),
                             alpha.V=diag(4)*1000)))
model.3A<-MCMCglmm(count~log(length) + nonsynonymous + divergence + nonsynonymo
us:divergence + organism +
                         (nonsynonymous + divergence + nonsynonymous:divergence
):(organism + piRNA + viRNA + miRNA + siRNA),
                       random=~us(1 + nonsynonymous + divergence + nonsynonymou
s:divergence):gene,
                       rcov=~us(nfac):organism:gene:Duplicate, family="poisson"
,
                       data=dat, pr = TRUE, pl=TRUE, prior = prior.3A, verbose
= FALSE)
```

For the SnIPRE-like analysis, we model the counts of each type of mutation (Pn, Ps, Dn, Ds) in each gene in each organism as a poisson response variable. Following Eilertson et al (2012), we set the fixed effects to the length of the gene, the type of mutation (by fitting either a nonsynonymous or divergence effect), and the interaction between nonsynonymous and divergence effects. Because we are interested in estimating effects of certain pathways across species, we also fit nonsynonymous, divergence and nonsynonymous-by-divergence effects separately for each gene class and organism with the term nonsynonymous+divergence+nonsynonymous:divergence):(organism+piRNA+siRNA+miRNA+viRNA). The fixed effects portion of the model looks as follows:

```
summary(model.3A)

## 
##  Iterations = 3001:12991
##  Thinning interval  = 10
##  Sample size  = 1000 
## 
##  DIC: 9979.245 
## 
##  G-structure:  ~us(1 + nonsynonymous + divergence + nonsynonymous:divergence
):gene
## 
##                                                          post.mean
## (Intercept):(Intercept).gene                            0.0300609
## nonsynonymous:(Intercept).gene                          0.0092904
## divergence:(Intercept).gene                            -0.0062428
## nonsynonymous:divergence:(Intercept).gene              -0.0028054
## (Intercept):nonsynonymous.gene                          0.0092904
## nonsynonymous:nonsynonymous.gene                        0.7257428
## divergence:nonsynonymous.gene                           0.0679281
## nonsynonymous:divergence:nonsynonymous.gene            -0.0110024
## (Intercept):divergence.gene                            -0.0062428
## nonsynonymous:divergence.gene                           0.0679281
## divergence:divergence.gene                              0.0397815
## nonsynonymous:divergence:divergence.gene                0.0006943
## (Intercept):nonsynonymous:divergence.gene              -0.0028054
## nonsynonymous:nonsynonymous:divergence.gene            -0.0110024
## divergence:nonsynonymous:divergence.gene                0.0006943
## nonsynonymous:divergence:nonsynonymous:divergence.gene  0.0263167
##                                                           l-95% CI u-95% CI
## (Intercept):(Intercept).gene                            2.216e-07  0.09298
## nonsynonymous:(Intercept).gene                         -9.516e-02  0.09389
## divergence:(Intercept).gene                            -4.354e-02  0.01857
## nonsynonymous:divergence:(Intercept).gene              -3.070e-02  0.01703
## (Intercept):nonsynonymous.gene                         -9.516e-02  0.09389
## nonsynonymous:nonsynonymous.gene                        5.079e-01  0.99459
## divergence:nonsynonymous.gene                          -1.419e-02  0.17282
## nonsynonymous:divergence:nonsynonymous.gene            -9.904e-02  0.06650
## (Intercept):divergence.gene                            -4.354e-02  0.01857
## nonsynonymous:divergence.gene                          -1.419e-02  0.17282
## divergence:divergence.gene                              8.076e-08  0.08962
## nonsynonymous:divergence:divergence.gene               -2.709e-02  0.02161
## (Intercept):nonsynonymous:divergence.gene              -3.070e-02  0.01703
## nonsynonymous:nonsynonymous:divergence.gene            -9.904e-02  0.06650
## divergence:nonsynonymous:divergence.gene               -2.709e-02  0.02161
## nonsynonymous:divergence:nonsynonymous:divergence.gene  3.772e-09  0.08581
##                                                         eff.samp
## (Intercept):(Intercept).gene                              75.05
## nonsynonymous:(Intercept).gene                            72.18
## divergence:(Intercept).gene                               56.54
## nonsynonymous:divergence:(Intercept).gene                 68.75
## (Intercept):nonsynonymous.gene                            72.18
## nonsynonymous:nonsynonymous.gene                         110.97
## divergence:nonsynonymous.gene                             80.44
## nonsynonymous:divergence:nonsynonymous.gene               20.63
## (Intercept):divergence.gene                               56.54
## nonsynonymous:divergence.gene                             80.44
## divergence:divergence.gene                               143.92
## nonsynonymous:divergence:divergence.gene                  75.55
## (Intercept):nonsynonymous:divergence.gene                 68.75
## nonsynonymous:nonsynonymous:divergence.gene               20.63
## divergence:nonsynonymous:divergence.gene                  75.55
## nonsynonymous:divergence:nonsynonymous:divergence.gene    12.38
##
##  R-structure:  ~us(nfac):organism:gene:Duplicate
##
##                                          post.mean  l-95% CI u-95% CI
## nfac0 0:nfac0 0.organism:gene:Duplicate   0.61205  0.494162  0.73623
## nfac0 1:nfac0 0.organism:gene:Duplicate   0.40907  0.300592  0.52785
## nfac1 0:nfac0 0.organism:gene:Duplicate   0.07070  0.027120  0.12159
## nfac1 1:nfac0 0.organism:gene:Duplicate  -0.04874 -0.134091  0.04111
## nfac0 0:nfac0 1.organism:gene:Duplicate   0.40907  0.300592  0.52785
## nfac0 1:nfac0 1.organism:gene:Duplicate   0.59946  0.457727  0.75343
## nfac1 0:nfac0 1.organism:gene:Duplicate   0.06481  0.008542  0.11479
## nfac1 1:nfac0 1.organism:gene:Duplicate   0.25187  0.146510  0.37272
## nfac0 0:nfac1 0.organism:gene:Duplicate   0.07070  0.027120  0.12159
## nfac0 1:nfac1 0.organism:gene:Duplicate   0.06481  0.008542  0.11479
## nfac1 0:nfac1 0.organism:gene:Duplicate   0.11693  0.079694  0.15503
## nfac1 1:nfac1 0.organism:gene:Duplicate   0.10049  0.050778  0.14859
## nfac0 0:nfac1 1.organism:gene:Duplicate  -0.04874 -0.134091  0.04111
## nfac0 1:nfac1 1.organism:gene:Duplicate   0.25187  0.146510  0.37272
## nfac1 0:nfac1 1.organism:gene:Duplicate   0.10049  0.050778  0.14859
## nfac1 1:nfac1 1.organism:gene:Duplicate   0.42194  0.308206  0.54007
##                                         eff.samp
## nfac0 0:nfac0 0.organism:gene:Duplicate    230.4
## nfac0 1:nfac0 0.organism:gene:Duplicate    170.7
## nfac1 0:nfac0 0.organism:gene:Duplicate    122.4
## nfac1 1:nfac0 0.organism:gene:Duplicate    153.7
## nfac0 0:nfac0 1.organism:gene:Duplicate    170.7
## nfac0 1:nfac0 1.organism:gene:Duplicate    178.2
## nfac1 0:nfac0 1.organism:gene:Duplicate    115.4
## nfac1 1:nfac0 1.organism:gene:Duplicate    184.9
## nfac0 0:nfac1 0.organism:gene:Duplicate    122.4
## nfac0 1:nfac1 0.organism:gene:Duplicate    115.4
## nfac1 0:nfac1 0.organism:gene:Duplicate    168.3
## nfac1 1:nfac1 0.organism:gene:Duplicate    321.7
## nfac0 0:nfac1 1.organism:gene:Duplicate    153.7
## nfac0 1:nfac1 1.organism:gene:Duplicate    184.9
## nfac1 0:nfac1 1.organism:gene:Duplicate    321.7
## nfac1 1:nfac1 1.organism:gene:Duplicate    242.8
## 
##  Location effects: count ~ log(length) + nonsynonymous + divergence + nonsyn
onymous:divergence + organism + (nonsynonymous + divergence + nonsynonymous:div
ergence):(organism + piRNA + viRNA + miRNA + siRNA) 
## 
##                                              post.mean l-95% CI u-95% CI
## (Intercept)                                   -4.46904 -4.97306 -3.94758
## log(length)                                    0.97623  0.91561  1.04389
## nonsynonymous                                 -1.49970 -1.77328 -1.24692
## divergence                                     1.21542  1.00947  1.42464
## organismanopheles                              0.21142 -0.08762  0.44189
## organismapis                                  -1.81890 -2.10749 -1.52871
## organismbombyx                                 0.40217  0.14216  0.69221
## organismdpse                                  -0.43007 -0.70159 -0.18840
## organismheliconius                             0.05955 -0.21181  0.34851
## nonsynonymous:divergence                       0.25602  0.01122  0.46658
## nonsynonymous:organismanopheles               -0.46541 -0.75293 -0.17766
## nonsynonymous:organismapis                    -0.40197 -0.77895 -0.04547
## nonsynonymous:organismbombyx                  -0.24661 -0.58073  0.06554
## nonsynonymous:organismdpse                     0.34119  0.03611  0.64421
## nonsynonymous:organismheliconius               0.01839 -0.27399  0.31943
## nonsynonymous:piRNA                            0.54011  0.15190  0.93668
## nonsynonymous:viRNA                            0.58109 -0.60559  1.64163
## nonsynonymous:miRNA                           -0.44753 -1.18277  0.25110
## nonsynonymous:siRNA                           -0.82456 -1.74869  0.04430
## divergence:organismanopheles                 -0.95889 -1.23915 -0.67998
## divergence:organismapis                       1.06272  0.77986  1.36073
## divergence:organismbombyx                     0.27959 -0.01646  0.53432
## divergence:organismdpse                      -0.76135 -1.03303 -0.50080
## divergence:organismheliconius                -1.06374 -1.31553 -0.76169
## divergence:piRNA                              0.08859 -0.04536  0.21826
## divergence:viRNA                              0.13431 -0.24487  0.45353
## divergence:miRNA                             -0.02860 -0.24507  0.18153
## divergence:siRNA                             -0.04164 -0.30738  0.20942
## nonsynonymous:divergence:organismanopheles    0.05467 -0.17126  0.27527
## nonsynonymous:divergence:organismapis        -0.35464 -0.63657  0.02153
## nonsynonymous:divergence:organismbombyx       0.06181 -0.22513  0.33049
## nonsynonymous:divergence:organismdpse         0.39663  0.17378  0.65118
## nonsynonymous:divergence:organismheliconius  -0.55927 -0.78868 -0.34668
## nonsynonymous:divergence:piRNA                0.36223  0.13452  0.59720
## nonsynonymous:divergence:viRNA                0.90780  0.50794  1.33614
## nonsynonymous:divergence:miRNA                0.26052 -0.06620  0.51761
## nonsynonymous:divergence:siRNA                0.78372  0.18102  1.45044
##                                             eff.samp  pMCMC
## (Intercept)                                  101.086 <0.001 ***
## log(length)                                   82.580 <0.001 ***
## nonsynonymous                                 56.027 <0.001 ***
## divergence                                   403.392 <0.001 ***
## organismanopheles                            723.666  0.126
## organismapis                                 264.677 <0.001 ***
## organismbombyx                               514.172  0.004 **
## organismdpse                                 311.035  0.002 **
## organismheliconius                           706.023  0.640
## nonsynonymous:divergence                      11.749  0.038 *
## nonsynonymous:organismanopheles              150.165 <0.001 ***
## nonsynonymous:organismapis                    20.804  0.020 *
## nonsynonymous:organismbombyx                  60.972  0.140
## nonsynonymous:organismdpse                    36.646  0.038 *
## nonsynonymous:organismheliconius             147.820  0.888
## nonsynonymous:piRNA                          211.155  0.008 **
## nonsynonymous:viRNA                          294.525  0.310
## nonsynonymous:miRNA                          326.843  0.236
## nonsynonymous:siRNA                           16.762  0.082 .
## divergence:organismanopheles                 529.568 <0.001 ***
## divergence:organismapis                      254.734 <0.001 ***
## divergence:organismbombyx                    345.638  0.052 .
## divergence:organismdpse                      291.953 <0.001 ***
## divergence:organismheliconius                654.453 <0.001 ***
## divergence:piRNA                             437.990  0.178
## divergence:viRNA                             321.484  0.422
## divergence:miRNA                             689.476  0.764
## divergence:siRNA                             574.703  0.752
## nonsynonymous:divergence:organismanopheles    39.134  0.680
## nonsynonymous:divergence:organismapis          8.341  0.080 .
## nonsynonymous:divergence:organismbombyx       19.681  0.644
## nonsynonymous:divergence:organismdpse         18.507 <0.001 ***
## nonsynonymous:divergence:organismheliconius   42.853 <0.001 ***
## nonsynonymous:divergence:piRNA                13.101  0.004 **
## nonsynonymous:divergence:viRNA                36.579 <0.001 ***
## nonsynonymous:divergence:miRNA                17.685  0.092 .
## nonsynonymous:divergence:siRNA                 5.187  0.016 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model output shows organisms (e.g. nonsynonymous:divergence:organismapis effect) and subpathways (e.g. nonsynonymous:divergence:viRNA effect) differ in their genome-wide level of positive selection. We test whether a homologue has an increased selection effect (e.g. Figure 4, Figure S4) by comparing the posterior distributions of the selection effect for control genes and the homologues. For example:

```
ago2.selection.effect <- model.3A$Sol[,"nonsynonymous:divergence"] +
  model.3A$Sol[,"nonsynonymous:divergence:viRNA"] +
  model.3A$Sol[,"nonsynonymous:divergence.gene.ago2"]

control.selection.effect <- model.3A$Sol[,"nonsynonymous:divergence"]

control.minus.ago2.posterior <-  control.selection.effect - ago2.selection.effe
ct

iterations.less.than.zero <- length(which(control.minus.ago2.posterior > 0))
total.chain.length <- length(control.minus.ago2.posterior)
pMCMC <- iterations.less.than.zero/total.chain.length
pMCMC*2 #Multiply by 2 to make a two tailed test

## [1] 0
```

We conclude that the Ago2 selection effect is greater than control genes (MCMCp < 0.001).

In addition to homologue-specific random effects, we include gene-specific random effects in the SnIPRE model. The homologue-specific random effects are coded as "random=~us(1 + nonsynonymous+divergence+nonsynonymous:divergence):gene", very similar to the random effect structure of Eilertson et al (2012), except we have an average of 24 observations per gene (6 species by 4 types of mutation). Instead of using the per-gene random effect structure from Eilertson et al (2012), we reparameterise the model so that residuals for each class of mutation would be estimated for each gene, and (by specifying pl=TRUE), the posterior distribution of each of these residuals stored. We code the different classes of mutation in the nfac column, where "0 0" denotes synonymous polymorphism, "0 1" denotes nonsynonymous polymorphism, and so on. The unstructured covariance matrix between mutation classes (us(nfac)) is estimated for each gene (organism:gene:Duplicate), from which we extract the gene-level selection effect. The posterior distributions of these residuals are saved in the model.3A$Liab data structure, with columns in the same order as the rows in our original data table:

```
nrow(dat) == ncol(data.frame(model.3A$Liab))

## [1] TRUE
```

To extract the gene specific selection effect from each residual, we create the mapping matrix (X), the model matrix for all nonsynonymous:divergence effects (X.matrix, a template for which fixed effects link with the rows of data), the columns of the model.3A fixed effects which match X.matrix (X.model.hit), and a list of unique genes (unique.genes):

```
X<-rbind(c(1,0,0,0), c(0,0,1,0), c(0,1,0,0), c(0,1,1,1))
X.model<-model.matrix(~nonsynonymous:divergence-1+(nonsynonymous:divergence):(p
iRNA+siRNA+miRNA+viRNA+gene), data=dat)
colnames(X.model) <- gsub(colnames(X.model), pattern= ":gene", replacement = "\
\.gene\\.")
```

```
X.model.hit<-match(colnames(X.model), colnames(model.3A$Sol))
unique.genes<-unique(dat$gene.id)
residuals.transform<-model.3A$Liab
X.fixed.random.effects<-model.3A$Liab
```

We loop through rows of the stored latent variables (model.3A$Liab), with each row being an iteration of the MCMC chain, and subtract the fixed and random effects corresponding to each data point ((model.3A$X%*%model.3A$Sol[i,1:ncol(model.3A$X)])) from the ωA estimate (model.3A$Z%*%model.3A$Sol[i,(ncol(model.3A$X)+1):ncol(model.3A$Sol)])), resulting in residuals for each gene in each species. Then, for each gene (gene.id - a combination of gene name, organism, and duplicate) in each iteration, we map the residuals onto the design matrix X to solve for the random effects of each observation. (residuals.transform). Finally, we obtain the posterior distribution of the selection effect for a particular gene (selection.effects) by adding the random nonsynonymous:divergence:gene effect (residuals.transform) to the fixed and random nonsynonymous:divergence effects (X.fixed.random.effects).

```
for(i in 1:nrow(model.3A$Liab)){
   residuals<-model.3A$Liab[i,]-(model.3A$X%*%model.3A$Sol[i,1:ncol(model.3A$X)
])@x -
     (model.3A$Z%*%model.3A$Sol[i,(ncol(model.3A$X)+1):ncol(model.3A$Sol)])@x #
Subtract the fixed pathway effects and random homologue effects from the latent
variables
   # residuals for each observation at iteration i
    for(j in 1:length(unique.genes)){
       hits<-which(dat$gene.id==unique.genes[j])
       # find positions of residuals gene j
       if(length(hits)==4){
        beta<-solve(X,residuals[hits]) # Map the residuals from the difference
between latent variables and fixed effects onto design matrix
        residuals.transform[i,hits]<-beta
        # solve for the (random) b effects for each observation
        X.fixed.random.effects[i,hits]<-c(X.model[hits,]%*%model.3A$Sol[i,X.mod
el.hit])
        # get nonsynonymous+divergence predictions for each obseravtion
      }
    }
}
selection.effects<-data.frame(X.fixed.random.effects+residuals.transform)
selection.effects <- selection.effects[,(dat$nfac == "1 1") & (dat$RNAi == 1)]
colnames(selection.effects) <- dat$gene.id[(dat$nfac == "1 1") & (dat$RNAi == 1
)]
head(selection.effects)[,1:5]

##   anopheles ago1 A apis ago1 A bombyx ago1 A dmel ago1 A dpse ago1 A
## 1        1.4193757   0.7722761     0.5842877  0.08680539   0.3625202
## 2        0.8455055   0.6456496     0.4544315  0.23100181   0.1364505
## 3        0.4975981   0.5436095     0.2534048 -0.33675612  -0.6107161
## 4        1.4048516   0.6310795     0.1799840  0.83226188  -0.5621510
## 5        1.2629835   0.3232525    -0.3350957  0.34226755  -0.1326866
## 6        0.7548817   0.3769485    -0.4292804  0.56985419   0.1456308
```

We calculate a "species-corrected" selection effect where nonsynonymous:divergence:organism effects are excluded in order to visualise differences between subpathways (e.g. Figure 3). We add these effects later when assessing positive selection in individual genes within a species. Also, in this example, we have only solved for the gene-level nonsynonymous:divergence random effects, however, the other gene-level effects could be obtained in a similar way.

SnIPRE was originally intended to identify genes in a single organism which shows signs of elevated positive selection. To do this, we add the organism specific selection effect to the "species-corrected" selection effect we have already obtained, and ask whether the selection effect overlaps zero. For example, to estimate the selection effect for the genes in Apis mellifera, we add the nonsynonymous:divergence:apis posterior to the columns of selection.effects which belong to Apis.

```
apis <- model.3A$Sol[,"nonsynonymous:divergence:organismapis"]
selection.effects.apis <- selection.effects[,grep("apis",colnames(selection.eff
ects))] + apis
```

Then, using HPDinterval() and colMeans(), we can get the selection effect for each gene, along with the upper and lower 95% highest posterior density intervals.

```
selection.effects.apis.summary <- data.frame(HPDinterval(as.mcmc(selection.effe
cts.apis)))
selection.effects.apis.summary <- cbind(selection.effects.apis.summary,data.fra
me(selectioneffect=c(as.vector(colMeans(selection.effects.apis)))))
head(selection.effects.apis.summary)
```

```
##                   lower     upper selectioneffect
## apis ago1 A -1.0316050 1.2296024      0.16360626
## apis ago2 A  0.8680883 2.3909126      1.64137753
## apis ago3 A -0.4233092 1.2884881      0.35796870
## apis armi A -0.6455877 0.7797752      0.06855869
## apis ars2 A -1.0879748 0.6938276     -0.13471034
## apis arx A  -1.0893511 1.0259998     -0.06135972
```

Finally, we identify genes with significantly positive selection effects.

```
selection.effects.apis.significant <- selection.effects.apis.summary[selection.
effects.apis.summary$lower > 0,]
print(selection.effects.apis.significant)
```

```
##                    lower    upper selectioneffect
## apis ago2 A 0.868088349 2.390913       1.6413775
## apis piwi A 0.350946926 1.734872       1.0338514
## apis dcr2 A 0.382093697 1.785551       1.0686899
## apis hen1 A 0.125590039 1.452687       0.7121331
## apis r2d2 A 0.007973288 2.385391       1.2971375
## apis tud A  0.280620893 1.627606       0.9377881
## apis vas A  0.615037761 2.525881       1.5089618
```

Model 3B: SnIPRE-like analysis, with piRNA split into biogenesis factors, effectors, and transcriptional silencing

```
prior.3B <-list(R=list(V=diag(4), nu=0.002),
                G=list(G1=list(V=diag(4), nu=4, alpha.mu=rep(0,4), alpha.V=diag
(4)*1000)))
model.3B <- MCMCglmm(count~log(length)+nonsynonymous + divergence+nonsynonymous
:divergence+organism+
                                 (nonsynonymous+divergence+nonsynonymous:diverg
ence):(organism + effector + biogenesis + transcriptional + viRNA + miRNA + siR
NA),
                               random=~us(1 + nonsynonymous + divergence + nons
ynonymous:divergence):gene,
                               rcov=~us(nfac):organism:gene:Duplicate, family="
poisson",
                               data=dat, pr = TRUE, pl=TRUE, prior = prior.3B,
verbose = FALSE)
```

We also fit the SnIPRE model (Model 3A) with the piRNA pathway split into different functional categories, akin to the Model 1B and 1C. We only used this model to estimate the selection effects associated with the piRNA categories (transcriptional silencing, effectors, and biogenesis machinery).

Model 3C: SnIPRE-like analysis, without subpathway as a fixed effect

```
prior.3C <- list(R=list(V=diag(4), nu=0.002), G=list(G1=list(V=diag(4), nu=4, a
lpha.mu=rep(0,4), alpha.V=diag(4)*1000)))
model.3C <- MCMCglmm(count~log(length)+nonsynonymous + divergence+nonsynonymous
:divergence+organism+
                         (nonsynonymous+divergence+nonsynonymous:divergence):(o
rganism),
                       random=~us(1 + nonsynonymous + divergence + nonsynonymou
s:divergence):gene,
                       rcov=~us(nfac):organism:gene:Duplicate, family="poisson"
,
                       data=dat, pr = TRUE, pl=TRUE, prior = prior.3C, verbose
= FALSE)
```

Finally, we fit the SniPRE model (Model 5A) without assuming genes belong to any particular subpathway,similar to the difference between Model 2A and 2B. Selection effects were then calculated in the same way, excluding the addition of a subpathway fixed effect.

S2 Text: Supplementary R code for models

To assess significance in the SweeD analyses, we used ms (Hudson, 2002) to perform 1000 coalescent simulations for each gene region of interest in each species, given the observed number of segregating sites, reported recombination rate, and a previously published estimate of the demographic history of that species. When population scaled recombination rate estimates were not available, we used estimates of N_e to scale per-base rate estimates. Although the details of the demographic scenarios we modelled are unlikely to impact substantially upon our qualitative comparisons of between sweep frequency in different types of gene, we attempted to use null models consistent with the published literature. The demographic scenarios modelled for each species are illustrated in Figure S1. For *D.*

melanogaster, recombination rates from the *Drosophila* recombination rate calculator were used with a constant N_e for African populations of 1.15x10^6 (Charlesworth, 2009). Some genes (*ael, AGO3, pasha,* and *Rm62*) are reported to lie in areas with zero recombination (Fiston-Lavier, et al., 2010), so we set the recombination rate in these genes at the lowest non-zero rate observed. For *D. pseudoobscura,* we simulated a population expansion (Haddrill, et al., 2010; Larracuente & Clark, 2014), and used the population scaled rates of recombination and gene conversion from Larracuente and Clark (2014). For *Anopheles gambiae,* we used demographic history parameters from Crawford and Lazzaro et al (2010) for the Cameroon population, and the recombination rates for each individual chromosome arm (1 cM/Mb for the X, 1.3 cM/Mb for 3L and 2R, 1.6 cM/Mb for 3R, and 2 cM/Mb for 2L) from Pombi et al (2006) and Stump et al (2007). Effective population size (N_e) was set to 2.4x10^6 estimated using the *D. melanogaster* mutation rate of Keightley et al (2014) and the Watterson's theta (θ_W) estimate in Crawford and Lazzaro (2010). For *H. melpomene,* we simulated three Costa Rican populations corresponding to *H. melpomene, H. cydno, and H. pachinus,* using the migration rates provided in Table 2 of Kronforst et al (2006). We used a constant recombination rate of 7.51 cM/Mb across the entire genome with an N_e of 2.1x10^6 for *H. melpomene*, 3.3x10^6 for cydno, and 2.7x10^6 for *H. pachinus*. For *B. mandarina,* we modelled the "gene-flow at bottleneck" scenario (Yang et al, 2014), with an N_e of 500,000 for *B. mandarina* and 73,000 for *B. mori*, and a recombination rate of 2.97 cM/Mb (Yamamoto et al, 2008; Yang et al, 2014). For *A. mellifera,* four subpopulations were modelled using N_e values in Table 1 of Wallberg et al (2014), following Figure 1F in Wallberg et al (2014) when modelling past subpopulation size changes. These subpopulations share migrants, and migration rates were estimated based on F_{ST} values between subpopulations reported in Whitfield et al (2006). A recombination rate of 19 cM/Mb is assumed to be constant across the genome (Beye et al, 2006). For *C. briggsae*, coalescent simulations and SweeD analyses were carried out on the 25 "tropical" samples in order to avoid modelling complicated demographic scenarios. These are expected to have an effective population size of 60,000, and to have undergone a recent bottleneck 0.916 N_e generations in the past (Cutter et al, 2006; Denver et al, 2009), assuming a 60-day generation time (Barrière & Félix, 2005). We used recombination rates for *C. briggsae* from Ross et al (2011), which are estimated to be 9.97 x 10^{-8} per bp per generation in autosomes and 4.6 x 10^{-8} per bp per generation on the X chromosome (Ross et al, 2011). Finally, for *P. pacificus,* four subpopulations were modelled corresponding to clade A1, A2, C, and 9 individuals whose clade was unknown (Rödelsperger, et al., 2014) which coalesced 0.849 N_e generations in the past (McGaughran et al, 2013). N_e was estimated by calculating θ_W for each contig and assuming a mutation rate of2x10^{-9} (Weller et al, 2014).To minimise differences between the real data and simulations, sites were randomly chosen to be folded, ancestrally invariant, or fixed for a derived substitution, in each case matching the numbers observed in the real data before the SweeD analysis.

Supplementary Materials References

Barrière, Antoine, and Marie-Anne Félix. 2005. "High local genetic diversity and low outcrossing rate in Caenorhabditis elegans natural populations." *Current biology : CB* 15 (13): 1176-84.

Beye, Martin, Irene Gattermeier, Martin Hasselmann, Tanja Gempe, Morten Schioett, John F Baines, David Schlipalius, et al. 2006. "Exceptionally high levels of recombination across the honey bee genome." *Genome research* 16 (11): 1339-44.

Brennecke, Julius, Alexei A Aravin, Alexander Stark, Monica Dus, Manolis Kellis, Ravi Sachidanandam, and Gregory J Hannon. 2007. "Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila." *Cell* 128 (6): 1089-103.

Caudy, Amy A, Mike Myers, Gregory J Hannon, and Scott M Hammond. 2002. "Fragile X-related protein and VIG associate with the RNA interference machinery." *Genes & development* (Cold Spring Harbor Laboratory Press) 16 (19): 2491-6.

Caudy, Amy A., René F. Ketting, Scott M. Hammond, Ahmet M. Denli, Anja M. P. Bathoorn, Bastiaan B. J. Tops, Jose M. Silva, Mike M. Myers, Gregory J. Hannon, and Ronald H. A. Plasterk. 2003. "A micrococcal nuclease homologue in RNAi effector complexes." *Nature* (Nature Publishing Group) 425 (6956): 411-414.

Charlesworth, Brian. 2009. "Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation." *Nature reviews. Genetics* (Nature Publishing Group) 10 (3): 195-205.

Crawford, Jacob E, and Brian P Lazzaro. 2010. "The demographic histories of the M and S molecular forms of Anopheles gambiae s.s." *Molecular biology and evolution* 27 (8): 1739-44.

Csink, A K, R Linsk, and J A Birchler. 1994. "The Lighten up (Lip) gene of Drosophila melanogaster, a modifier of retroelement expression, position effect variegation and white locus insertion alleles." *Genetics* 138 (1): 153-63.

Cutter, Asher D, Marie-Anne Félix, Antoine Barrière, and Deborah Charlesworth. 2006. "Patterns of nucleotide polymorphism distinguish temperate and tropical wild isolates of Caenorhabditis briggsae." *Genetics* 173 (4): 2021-31.

Czech, Benjamin, Jonathan B Preall, Jon McGinn, and Gregory J Hannon. 2013. "A transcriptome-wide RNAi screen in the Drosophila ovary reveals factors of the germline piRNA pathway." *Molecular cell* 50 (5): 749-61.

Denver, Dee R, Peter C Dolan, Larry J Wilhelm, Way Sung, J Ignacio Lucas-Lledó, Dana K Howe, Samantha C Lewis, et al. 2009. "A genome-wide view of Caenorhabditis elegans base-substitution mutation processes." *Proceedings of the National Academy of Sciences of the United States of America* 106 (38): 16310-4.

Dönertas, Derya, Grzegorz Sienski, and Julius Brennecke. 2013. "Drosophila Gtsf1 is an essential component of the Piwi-mediated transcriptional silencing complex." *Genes & development* (Cold Spring Harbor Laboratory Press) 27 (15): 1693-705.

Eilertson, Kirsten E, James G Booth, and Carlos D Bustamante. 2012. "SnIPRE: selection inference using a Poisson random effects model." *PLoS computational biology* (Public Library of Science) 8 (12): e1002806.

Fiston-Lavier, Anna-Sophie, Nadia D Singh, Mikhail Lipatov, and Dmitri A Petrov. 2010. "Drosophila melanogaster recombination rate calculator." *Gene* 463 (1-2): 18-20.

Gruber, Joshua J., D. Steven Zatechka, Leah R. Sabin, Jeongsik Yong, Julian J. Lum, Mei Kong, Wei-Xing Zong, et al. 2009. "Ars2 Links the Nuclear Cap-Binding Complex to RNA Interference and Cell Proliferation." *Cell* 138 (2): 328-339.

Gunawardane, Lalith S., Kuniaki Saito, Kazumichi M. Nishida, Keita Miyoshi, Yoshinori Kawamura, Tomoko Nagami, Haruhiko Siomi, and Mikiko C. Siomi. 2007. "A Slicer-Mediated Mechanism for Repeat-Associated siRNA 5' End Formation in Drosophila." *Science* 315 (5818).

Haase, Astrid D, Silvia Fenoglio, Felix Muerdter, Paloma M Guzzardo, Benjamin Czech, Darryl J Pappin, Caifu Chen, Assaf Gordon, and Gregory J Hannon. 2010. "Probing the initiation and effector phases of the somatic piRNA pathway in Drosophila." *Genes & development* (Cold Spring Harbor Laboratory Press) 24 (22): 2499-504.

Haddrill, Penelope R, Laurence Loewe, and Brian Charlesworth. 2010. "Estimating the parameters of selection on nonsynonymous mutations in Drosophila pseudoobscura and D. miranda." *Genetics* 185 (4): 1381-96.

Han, Bo W, Wei Wang, Chengjian Li, Zhiping Weng, and Phillip D Zamore. 2015. "piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production." *Science (New York, N.Y.)* 348 (6236): 817-21.

Handler, Dominik, Katharina Meixner, Manfred Pizka, Kathrin Lauss, Christopher Schmied, Franz Sebastian Gruber, and Julius Brennecke. 2013. "The genetic makeup of the Drosophila piRNA pathway." *Molecular cell* 50 (5): 762-77.

Horwich, Michael D., Chengjian Li, Christian Matranga, Vasily Vagin, Gwen Farley, Peng Wang, and Phillip D. Zamore. 2007. "The Drosophila RNA Methyltransferase, DmHen1, Modifies Germline piRNAs and Single-Stranded siRNAs in RISC." 1265-1272.

Hudson, R. R. 2002. "Generating samples under a Wright-Fisher neutral model of genetic variation." *Bioinformatics* (Oxford University Press) 18 (2): 337-338.

Ishizuka, Akira, Mikiko C Siomi, and Haruhiko Siomi. 2002. "A Drosophila fragile X protein interacts with components of RNAi and ribosomal proteins." *Genes & development* (Cold Spring Harbor Laboratory Press) 16 (19): 2497-508.

Jiang, Feng, Xuecheng Ye, Xiang Liu, Lauren Fincher, Dennis McKearin, and Qinghua Liu. 2005. "Dicer-1 and R3D1-L catalyze microRNA maturation in Drosophila." *Genes & development* (Cold Spring Harbor Laboratory Press) 19 (14): 1674-9.

Keightley, Peter D, Rob W Ness, Daniel L Halligan, and Penelope R Haddrill. 2014. "Estimation of the spontaneous mutation rate per nucleotide site in a Drosophila melanogaster full-sib family." *Genetics* 196 (1): 313-20.

Kirino, Yohei, Namwoo Kim, Mariàngels de Planell-Saguer, Eugene Khandros, Stephanie Chiorean, Peter S. Klein, Isidore Rigoutsos, Thomas A. Jongens, and Zissimos Mourelatos. 2009. "Arginine methylation of Piwi proteins catalysed by dPRMT5 is required for Ago3 and Aub stability." *Nature Cell Biology* (Nature Publishing Group) 11 (5): 652-658.

Klattenhoff, Carla, Hualin Xi, Chengjian Li, Soohyun Lee, Jia Xu, Jaspreet S. Khurana, Fan Zhang, et al. 2009. "The Drosophila HP1 Homolog Rhino Is Required for Transposon Silencing and piRNA Production by Dual-Strand Clusters." *Cell* 138 (6): 1137-1149.

Koch, Carmen M., Mona Honemann-Capito, Diane Egger-Adam, and Andreas Wodarz. 2009. "Windei, the Drosophila Homolog of mAM/MCAF1, Is an Essential Cofactor of the H3K9 Methyl Transferase dSETDB1/Eggless in Germ Line Development." Edited by Asifa Akhtar. *PLoS Genetics* 5 (9): e1000644.

Kronforst, Marcus R., Laura G. Young, Lauren M. Blume, and Lawrence E. Gilbert. 2006. "Multilocus analyses of admixture and introgression among hybridizing Heliconius butterflies." *Evolution* 60 (6): 1254-1268.

Larracuente, Amanda M, and Andrew G Clark. 2014. "Recent selection on the Y-to-dot translocation in Drosophila pseudoobscura." *Molecular biology and evolution* 31 (4): 846-56.

Lee, Yoontae, Chiyoung Ahn, Jinju Han, Hyounjeong Choi, Jaekwang Kim, Jeongbin Yim, Junho Lee, et al. 2003. "The nuclear RNase III Drosha initiates microRNA processing." *Nature* (Nature Publishing Group) 425 (6956): 415-419.

Lewis, Samuel H, Heli Salmela, and Darren J Obbard. 2016. "Duplication and Diversification of Dipteran Argonaute Genes, and the Evolutionary Divergence of Piwi and Aubergine." *Genome biology and evolution* 8 (3): 507-18.

Lewis, Samuel H., Claire L. Webster, Heli Salmela, and Darren J. Obbard. 2016. "Repeated Duplication of Argonaute2 Is Associated with Strong Selection and Testis Specialization in Drosophila." *Genetics* 204 (2).

Liu, Qinghua, Tim A. Rand, Savitha Kalidas, Fenghe Du, Hyun-Eui Kim, Dean P. Smith, and Xiaodong Wang. 2003. "R2D2, a Bridge Between the Initiation and Effector Steps of the Drosophila RNAi Pathway." *Science* 301 (5641).

Liu, Y., X. Ye, F. Jiang, C. Liang, D. Chen, J. Peng, L. N. Kinch, N. V. Grishin, and Q. Liu. 2009. "C3PO, an Endoribonuclease That Promotes RNAi by Facilitating RISC Activation." *Science* 325 (5941): 750-753.

Lo, Pang-Kuo, Yi-Chun Huang, John S. Poulton, Nicholas Leake, William H. Palmer, Daniel Vera, Gengqiang Xie, Stephen Klusza, and Wu-Min Deng. 2016. "RNA helicase Belle/DDX3 regulates transgene expression in Drosophila." *Developmental Biology* 412 (1): 57-70.

Malone, Colin D., Julius Brennecke, Monica Dus, Alexander Stark, W. Richard McCombie, Ravi Sachidanandam, and Gregory J. Hannon. 2009. "Specialized piRNA Pathways Act in Germline and Somatic Tissues of the Drosophila Ovary." *Cell* 137 (3): 522-535.

McGaughran, Angela, Katy Morgan, and Ralf J Sommer. 2013. "Unraveling the evolutionary history of the nematode Pristionchus pacificus: from lineage diversification to island colonization." *Ecology and evolution* 3 (3): 667-75.

Mohn, Fabio, Dominik Handler, and Julius Brennecke. 2015. "piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis." *Science (New York, N.Y.)* 348 (6236): 812-7.

Mohn, Fabio, Grzegorz Sienski, Dominik Handler, and Julius Brennecke. 2014. "The Rhino-Deadlock-Cutoff Complex Licenses Noncanonical Transcription of Dual-Strand piRNA Clusters in Drosophila." *Cell* 157 (6): 1364-1379.

Muerdter, Felix, Paloma M Guzzardo, Jesse Gillis, Yicheng Luo, Yang Yu, Caifu Chen, Richard Fekete, and Gregory J Hannon. 2013. "A genome-wide RNAi screen draws a genetic framework for transposon control and primary piRNA biogenesis in Drosophila." *Molecular cell* 50 (5): 736-48.

Neumüller, Ralph A., Joerg Betschinger, Anja Fischer, Natascha Bushati, Ingrid Poernbacher, Karl Mechtler, Stephen M. Cohen, and Juergen A. Knoblich. 2008. "Mei-P26 regulates microRNAs and cell growth in the Drosophila ovarian stem cell lineage." *Nature* (Nature Publishing Group) 454 (7201): 241-245.

Nishida, Kazumichi M, Tomoko N Okada, Takeshi Kawamura, Toutai Mituyama, Yoshinori Kawamura, Sachi Inagaki, Haidong Huang, et al. 2009. "Functional involvement of Tudor and dPRMT5 in the piRNA processing pathway in Drosophila germlines." *The EMBO journal* (European Molecular Biology Organization) 28 (24): 3820-31.

Ohtani, Hitoshi, Yuka W Iwasaki, Aoi Shibuya, Haruhiko Siomi, Mikiko C Siomi, and Kuniaki Saito. 2013. "DmGTSF1 is necessary for Piwi-piRISC-mediated transcriptional transposon silencing in the Drosophila ovary." *Genes & development* 27 (15): 1656-61.

Okamura, Katsutomo, Akira Ishizuka, Haruhiko Siomi, and Mikiko C Siomi. 2004. "Distinct roles for Argonaute proteins in small RNA-directed RNA cleavage pathways." *Genes & development* (Cold Spring Harbor Laboratory Press) 18 (14): 1655-66.

Olivieri, Daniel, Kirsten-André Senti, Sailakshmi Subramanian, Ravi Sachidanandam, and Julius Brennecke. 2012. "The Cochaperone Shutdown Defines a Group of Biogenesis Factors Essential for All piRNA Populations in Drosophila." *Molecular Cell* 47 (6): 954-969.

Pane, Attilio, Peng Jiang, Dorothy Yanling Zhao, Mona Singh, and Trudi Schüpbach. 2011. "The Cutoff protein regulates piRNA cluster expression and piRNA production in the Drosophila germline." *The EMBO journal* 30 (22): 4601-15.

Patil, Veena S, Amit Anand, Alisha Chakrabarti, and Toshie Kai. 2014. "The Tudor domain protein Tapas, a homolog of the vertebrate Tdrd7, functions in the piRNA pathway to regulate retrotransposons in germline of Drosophila melanogaster." *BMC Biology* 12 (1): 61.

Patil, Veena S., and Toshie Kai. 2010. "Repression of Retroelements in Drosophila Germline via piRNA Pathway by the Tudor Domain Protein Tejas." 724-730.

Pombi, Marco, Aram D Stump, Alessandra Della Torre, and Nora J Besansky. 2006. "Variation in recombination rate across the X chromosome of Anopheles gambiae." *The American journal of tropical medicine and hygiene* 75 (5): 901-3.

Preall, Jonathan B, Benjamin Czech, Paloma M Guzzardo, Felix Muerdter, and Gregory J Hannon. 2012. "shutdown is a component of the Drosophila piRNA biogenesis machinery." *RNA (New York, N.Y.)* (Cold Spring Harbor Laboratory Press) 18 (8): 1446-57.

Rangan, Prashanth, Colin D. Malone, Caryn Navarro, Sam P. Newbold, Patrick S. Hayes, Ravi Sachidanandam, Gregory J. Hannon, and Ruth Lehmann. 2011. "piRNA Production Requires Heterochromatin Formation in Drosophila." 1373-1379.

Rödelsperger, Christian, Richard A Neher, Andreas M Weller, Gabi Eberhardt, Hanh Witte, Werner E Mayer, Christoph Dieterich, and Ralf J Sommer. 2014. "Characterization of genetic diversity in the nematode Pristionchus pacificus from population-scale resequencing data." *Genetics* 196 (4): 1153-65.

Ross, Joseph A, Daniel C Koboldt, Julia E Staisch, Helen M Chamberlin, Bhagwati P Gupta, Raymond D Miller, Scott E Baird, and Eric S Haag. 2011. "Caenorhabditis briggsae recombinant inbred line genotypes reveal inter-strain incompatibility and the evolution of recombination." *PLoS genetics* 7 (7): e1002174.

Saito, K., Y. Sakaguchi, T. Suzuki, T. Suzuki, H. Siomi, and M. C. Siomi. 2007. "Pimet, the Drosophila homolog of HEN1, mediates 2'-O-methylation of Piwi- interacting RNAs at their 3' ends." *Genes & Development* 21 (13): 1603-1608.

Saito, Kuniaki, Akira Ishizuka, Haruhiko Siomi, Mikiko C Siomi, EJ Sontheimer, and T Tuschl. 2005. "Processing of Pre-microRNAs by the Dicer-1–Loquacious Complex in Drosophila Cells." Edited by James C. Carrington. *PLoS Biology* (Public Library of Science) 3 (7): e235.

Saito, Kuniaki, Hirotsugu Ishizu, Miharu Komai, Hazuki Kotani, Yoshinori Kawamura, Kazumichi M Nishida, Haruhiko Siomi, and Mikiko C Siomi. 2010. "Roles for the Yb body components Armitage and Yb in primary piRNA biogenesis in Drosophila." *Genes & development* (Cold Spring Harbor Laboratory Press) 24 (22): 2493-8.

Sato, Kaoru, Yuka W. Iwasaki, Aoi Shibuya, Piero Carninci, Yuuta Tsuchizawa, Hirotsugu Ishizu, Mikiko C. Siomi, and Haruhiko Siomi. 2015. "Krimper Enforces an Antisense Bias on piRNA Pools by Binding AGO3 in the Drosophila Germline." *Molecular Cell* 59 (4): 553-563.

Saxe, Jonathan P, Mengjie Chen, Hongyu Zhao, and Haifan Lin. 2013. "Tdrkh is essential for spermatogenesis and participates in primary piRNA biogenesis in the germline." *The EMBO Journal* 32 (13): 1869-1885.

Sienski, Grzegorz, Derya Dönertas, Julius Brennecke, S.I. Grewal, V.L. Trudeau, M. Savitsky, A. Kalmykova, et al. 2012. "Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression." *Cell* (Elsevier) 151 (5): 964-80.

Stump, A D, M Pombi, L Goeddel, J M C Ribeiro, J A Wilder, A della Torre, and N J Besansky. 2007. "Genetic exchange in 2La inversion heterokaryotypes of Anopheles gambiae." *Insect molecular biology* 16 (6): 703-9.

Van Rij, Ronald P., Maria Carla Saleh, Bassam Berry, Catherine Foo, Andrew Houk, Christophe Antoniewski, and Raul Andino. 2006. "The RNA silencing endonuclease Argonaute 2 mediates specific antiviral immunity in Drosophila melanogaster." *Genes and Development* 20 (21): 2985-2995.

Wallberg, Andreas, Fan Han, Gustaf Wellhagen, Bjørn Dahle, Masakado Kawata, Nizar Haddad, Zilá Luz Paulino Simões, et al. 2014. "A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee Apis mellifera." *Nature genetics* (Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.) 46 (10): 1081-8.

Wang, Xiao-Hong, Roghiyh Aliyari, Wan-Xiang Li, Hong-Wei Li, Kevin Kim, Richard Carthew, Peter Atkinson, and Shou-Wei Ding. 2006. "RNA Interference Directs Innate Immunity Against Viruses in Adult Drosophila." *Science* 312 (5772).

Weller, Andreas M, Christian Rödelsperger, Gabi Eberhardt, Ruxandra I Molnar, and Ralf J Sommer. 2014. "Opposing forces of A/T-biased mutations and G/C-biased gene conversions shape the genome of the nematode Pristionchus pacificus." *Genetics* 196 (4): 1145-52.

Whitfield, Charles W, Susanta K Behura, Stewart H Berlocher, Andrew G Clark, J Spencer Johnston, Walter S Sheppard, Deborah R Smith, Andrew V Suarez, Daniel Weaver, and Neil D Tsutsui. 2006. "Thrice out of Africa: ancient and recent expansions of the honey bee, Apis mellifera." *Science (New York, N.Y.)* 314 (5799): 642-5.

Xiol, Jordi, Pietro Spinelli, Maike A. Laussmann, David Homolka, Zhaolin Yang, Elisa Cora, Yohann Couté, et al. 2014. "RNA Clamping by Vasa Assembles a piRNA Amplifier Complex on Transposon Transcripts." *Cell* 157 (7): 1698-1711.

Yamamoto, Kimiko, Junko Nohata, Keiko Kadono-Okuda, Junko Narukawa, Motoe Sasanuma, Shun-Ichi Sasanuma, Hiroshi Minami, et al. 2008. "A BAC-based integrated linkage map of the silkworm Bombyx mori." *Genome biology* (BioMed Central Ltd) 9 (1): R21.

Yang, Shao-Yu, Min-Jin Han, Li-Fang Kang, Zi-Wen Li, Yi-Hong Shen, and Ze Zhang. 2014. "Demographic history and gene flow during silkworm domestication." *BMC evolutionary biology* (BioMed Central Ltd) 14 (1): 185.

Yeom, Kyu-Hyeon, Yoontae Lee, Jinju Han, Mi Ra Suh, V. Narry Kim, Li Y., Hao Y.L., Ooi C.E., Godwin B., and Vitols E. 2006. "Characterization of DGCR8/Pasha, the essential cofactor for Drosha in primary miRNA processing." *Nucleic Acids Research* (Oxford University Press) 34 (16): 4622-4629.

Zamparini, Andrea L., Marie Y. Davis, Colin D. Malone, Eric Vieira, Jiri Zavadil, Ravi Sachidanandam, Gregory J. Hannon, and Ruth Lehmann. 2011. "Vreteno, a gonad-specific protein, is essential for germline development and primary piRNA biogenesis in Drosophila." *Development* 138 (18).

Zhang, Fan, Jie Wang, Jia Xu, Zhao Zhang, Birgit S. Koppetsch, Nadine Schultz, Thom Vreven, et al. 2012. "UAP56 Couples piRNA Clusters to the Perinuclear Transposon Silencing Machinery." *Cell* 151 (4): 871-884.

Zhang, Zhao, Jia Xu, Birgit S. Koppetsch, Jie Wang, Cindy Tipping, Shengmei Ma, Zhiping Weng, William E. Theurkauf, and Phillip D. Zamore. 2011. "Heterotypic piRNA Ping-Pong Requires Qin, a Protein with Both E3 Ligase and Tudor Domains." *Molecular Cell* 44 (4): 572-584.