The CLSW files shown here were generated in a stepwise process as follows:

1. fasta files were downloaded from DPGP, with ambiguous nucleotides substituted with the most common nucleotide within the population at a given position, generating polymorphism_files_5-25-12
2. Because these fasta files corresponded to sequences with gaps, affecting recombination rate estimates, a custom script was used to convert the coordinates (from flybase v. 5.29) into genomic distances, resulting in ms output (empirical_ms_files)
3. These ms files were analyzed by clsw (Kim and Stephan 2002) using the -p option to supply parameter files (empirical_clsw_inp) producing results (empirical_clsw_out)
4. The parameters from ms and clsw were used to generate parameters (ssw_inp_neutral) which were run through ssw (Kim and Stephan 2002) to produce 1000 simulated neutral datasets for every population sample (ssw_out_neutral)
5. These simulated neutral ssw files were analyzed with clsw with parameters from “clsw_inp_simulations”, producing results (neutral_ssw_analyzed_clsw)
6. ssw was run again, this time with parameters assuming recent positive selection (ssw_inp_selective), to produce 1000 simulated selected datasets for every sample (ssw_out_selective)
7. These simulated selected ssw files were again analyzed as in 5., producing results (selective_ssw_analyzed_clsw)

selective_ssw_analyzed_clsw and neutral_ssw_analyzed_clsw can be used to generate empirical distributions which the GOF and LR values from empirical_clsw_out can be compared against to generate an empirical p value.

Note that Ago3 has no segregating sites in the Malawi population and no simulans population representatives. Ago3 North Carolina did not complete initial clsw analysis, and therefore could not be analyzed or simulated with positive selection. Therefore Ago3 could not be analyzed by ssw and clsw.

The PAML files shown here were generated as follows:

Annotated orthologs of piRNA proteins were retrieved from flybase (version 5.29). In some cases the annotations were missing or known to be incorrect. Using reciprocal blast and annotations from GenBank (Vermaak et al. 2005), fasta files were obtained and aligned with PRANK, with the -codon and -F options enabled, and reformatted into .nuc format, producing the files zipped as “PAML piRNA divergence alignments.”
Using these alignments, control files (codeml.ctl files), and trees (trees), the branch vs. M0, branch-sites neutral vs. branch-sites, M1a vs. M2a, M7 vs. M8, and M8a vs. M8 lnL scores were compared to each other with a chi-square test. Raw result files excluding branch-sites findings are included (most PAML piRNA results) as well as a spreadsheet summarizing branch-sites results (branch-sites results final)

Annotated 1 to 1 orthologs across all six Drosophila species analyzed in this study were downloaded from flybase and aligned using PRANK with the -F and -codon options. The resulting alignments were analyzed with all of the above methods and control files. The results are summarized in the spreadsheet “genomic results summaries.” The "b-s" tab gives branch-sites results for each foreground lineage. "b-s sums" counts the number of significant lineages for each gene from branch-sites results. "b" estimates dN/dS values for each lineage of each gene in the branch model. "s" gives "sites" results including the M0 neutral model. "summary" integrates information from the previous sheets. "pairwise dN-dS" gives an estimate of genomic pairwise dN/dS scores.

McDonald Kreitman tests were run with the MK test from the “analysis” package of the “libsequence” software (http://www.molpopgen.org/software/lseqsoftware.html) using fasta data from polymorphism_files_5-25-12.
Fay and Wu's H tests were analyzed on the 1,000 neutral and 1,000 selected datasets (ssw_out_neutral and ssw_out_selective) using the “msstats” package from libsequence.