Long-term and Short-term Evolutionary Impacts of Transposable Elements on Drosophila

Yuh Chwen G. Lee and Charles H. Langley

Center for Population Biology and Department of Evolution and Ecology, The University of California, Davis, Davis, California, 95616
Running title: Evolutionary Impacts of Transposable Elements

Key words: Transposable elements; Coevolution; Host-pathogen interaction; Selection; Population differentiation

Author for Correspondence: Grace Yuh Chwen Lee
Department of Evolution and Ecology, One Shields Ave. The University of California, Davis, Davis, California, 95616, USA
530-754-9551
grylee@ucdavis.edu
ABSTRACT

Transposable elements (TEs) are considered to be genomic parasites and their interactions with their hosts have been likened to the coevolution between host and other nongenomic, horizontally transferred pathogens. TE families, however, are vertically inherited as integral segments of the nuclear genome. This transmission strategy has been suggested to weaken the selective benefits of host alleles repressing the transposition of specific TE variants. On the other hand, the elevated rates of TE transposition and high incidences of deleterious mutations observed during the rare cases of horizontal transfers of TE families between species could create at least a transient process analogous to the influence of horizontally transmitted pathogens. Here, we formally address this analogy using empirical and theoretical analysis to specify the mechanism of how host-TE interactions may drive the evolution of host genes. We found that host TE-interacting genes actually have more pervasive evidence of adaptive evolution than immunity genes that interact with nongenomic pathogens in Drosophila. Yet, both our theoretical modeling, and empirical observations comparing D. melanogaster populations before and after the horizontal transfer of P elements, which invaded D. melanogaster early last century, demonstrated that horizontally transferred TEs have only a limited influence on host TE-interacting genes. We propose that the more prevalent and constant interaction with multiple vertically transmitted TE families may instead be the main force driving the fast evolution of TE-interacting genes, which is fundamentally different from the gene-for-gene interaction of host-pathogen coevolution.
INTRODUCTION

Host-pathogen interactions affect the population dynamics and the evolutionary trajectories of both species. In particular, coevolutionary dynamics will affect the pattern of polymorphism and divergence of genes underlying host-parasite interactions either through an arms race (VAN VALEN 1973; DAWKINS and KREBS 1979) or balancing selection (HALDANE 1949; HUGHES et al. 1990; TAKAHATA et al. 1992; HUGHES and YEAGER 1998; ROSE et al. 2004). In either case, accelerated rates of protein evolution and/or recurrent adaptive substitutions are expected in genes engaged in these interactions, which has been observed in both immunity genes (SCHLENKE and BEGUN 2003; JIGGINS and KIM 2007; SACKTON et al. 2007; OBBARD et al. 2009b) and anti-virus siRNA genes (OBBARD et al. 2006, 2009a; b, 2011) in Drosophila.

Transposable elements (TEs) are ubiquitous genomic constituents that increase their copy number (the number of TEs in a host genome) by replicative transposition (copying to new genomic locations). Like Drosophila, most host genomes are occupied by multiple TE families, which are defined by sequence similarity (homology) and by replication and transposition mechanisms. Even though incidences of potentially adaptive individual TE insertions with high population frequencies have been reported (DABORN et al. 2002; AMINETZACH et al. 2005; GONZÁLEZ et al. 2008; SCHMIDT et al. 2010), the mutagenic effects of TE insertions are typically deleterious because they disrupt gene structure and function (FINNEGAN 1992) and can lead to deleterious chromosomal rearrangement (MONTGOMERY et al. 1987, 1991; LANGLEY et al. 1988; BATZER and DEININGER 2002; MIECKOWSKI et al. 2006). Supporting this, TEs in natural populations of Drosophila are generally found in intergenic regions (AQUADRO et al. 1986;
Kaminker et al. 2002; Bergman et al. 2006) and present at low population frequencies (and reviewed in Charlesworth and Langley 1989; Lee and Langley 2010; Lee et al. 2007; Harshman and Decelle 2005). *D. melanogaster* strains with larger copy number of several TE families surveyed also have lower fitness (Mackay 1989; Pasyukova et al. 2004).

Because of these deleterious fitness impacts of TEs on their hosts, the interaction between host and TEs has been suggested to be analogous to an arms race between host and other more familiar pathogens (Kidwell and Lisch 2001; Aravin et al. 2007; Omura et al. 2008; Oubbard et al. 2009a; Blumenstiel 2011). Surprisingly, there has been no systematic comparison of the evolution of host genes interacting with TEs and those interacting with pathogens to examine this analogy. Furthermore, no specific models for the evolutionary impact of antagonistic interactions between host and TEs on protein evolution have been analyzed.

Due to the horizontal transmission of pathogens, there can be strong associations between specific host alleles and the level of pathogen load, leading to large selective benefits for host alleles that can effectively suppress pathogen infections. On the contrary, TE families are inherited vertically as part of the host parental genome. Vertical inheritance of TE insertions, random mating, and recombination lead to weak associations between selectively favored, TE-suppressive host variants and any net reduction in TE replication. Indeed, in organisms with random mating and weak linkage disequilibrium (e.g. *Drosophila*), a host variant suppressing a single TE family will enjoy only weak adaptive advantage (Charlesworth and Langley 1986). If the suppression of TE transposition involves specific associations between host alleles and
TE variants, this is unlikely to drive the fast evolution of host genes involved in TE suppression.

In rare cases, TEs are observed to have been horizontally transferred between host species (reviewed by Silva et al. 2004; Loreto et al. 2008; Schaack et al. 2010). In these cases, their evolutionary impacts on the host may be more analogous to those of nongenomic pathogens. During the spread of a newly invaded TE family, transposition rate can be exceptionally high (Kidwell et al. 1988; Good et al. 1989; reviewed in Silva et al. 2004). This is usually associated with a high incidence of deleterious insertions and sterilities, imposing a strong selective pressure on the host. However, the spread of horizontal transfer of TE families to all genomes of the new host species is found to happen on a short time scale, generally within thousands of generations (Daniels et al. 1990; Simmons 1992). After the initial horizontal transfer, copies of the new TE family appear to follow the typical TE mode of vertical transmission. A quantitative analysis of the hypothesis that horizontal transferred TE families are able to impose strong enough selection to elicit specific adaptive evolution in host genes remains unaddressed.

One of the best-studied examples of a TE horizontal transfer is the P element (reviewed in Engels 1997; Rio 2002), which invaded D. melanogaster from the distantly related D. willistoni less than a hundred years ago (Brookfield et al. 1984; Anxolabéhère et al. 1988; Daniels et al. 1990; Clark et al. 1994). While virtually all more recently surveyed D. melanogaster populations have P elements, D. melanogaster strains collected in the early last century, and maintained in the laboratory since that time, are still free of P elements (Kidwell et al. 1983; Anxolabéhère et al. 1988). Matings
between females lacking *P elements* (the M strains, said to have *M cytotype*) and males with *P elements* (the P strains, said to have *P cytotype*) result in progeny with “hybrid dysgenesis” syndrome, which consists of high rates of male recombination, sterility, mutation and chromosomal rearrangement (Hiraizumi 1971; Kidwell *et al.* 1973, 1977; Engels 1979). This dysgenic effects are attributed to the high rate of *P element* transposition (Bingham *et al.* 1982; Rubin *et al.* 1982). *P elements* of *D. melanogaster* thus provide a rare opportunity to investigate the evolutionary impacts of horizontally transferred TE families on the host genes interacting with them.

Our systematic comparisons demonstrate that the molecular evolution of TE-interacting genes exhibits comparable evidence of recurrent adaptive fixations to that of genes mediating the interactions between *Drosophila* and horizontally transferred pathogens. We took two approaches to investigate whether horizontally transferred TEs have discernable selective impacts and can contribute to the observed long-term adaptive evolution of TE-interacting genes. We developed and analyzed a model of TE invasion after horizontal transfer. And we used the recently invaded *P element* as a system to empirically contrast the genetic differentiation of candidate host genes between *D. melanogaster* populations before and after *P element* invasions. Both our analytical modeling and empirical observations suggest that the selective pressure imposed by horizontally transferred TE families is limited. We proposed a hypothesis other than the gene-for-gene host-TE coevolutionary model to explain our observations.
**MATERIALS and METHODS**

*D. melanogaster* variation of *post-P element* invasion population and *Drosophila* divergence data

We used the release 1.0 assembly of 44 *D. melanogaster* genome sequences generated by the *Drosophila* Population Genomic Project [DPGP, www.dpgp.org (Langley et al. 2012)]. The DPGP data consists of seven strains from Malawi and 37 strains from North Carolina, which were all collected after *P element* invasions and have *P elements*. Coding regions of each candidate gene were defined according to the *D. melanogaster* reference genome annotation (version 5.16) and parsed out from above genomic sequences using Perl scripts. Bases with quality scores lower than 30 and regions that appeared as identical by descent (IBD) or exhibit residual heterozygosity [see (Langley et al. 2012)] were treated as missing data. We also removed alleles when more than 50% of the bases were missing data. To compare the evolution of candidate genes with immunity genes, we used 236 immunity genes included in Sackton et al. (2007). Retrieval of coding sequences and population genetics analysis (see below) of these immunity genes were the same as those of candidate genes.

*D. simulans* (Begun et al. 2007) and *D. yakuba* (Clark et al. 2007) alleles were retrieved according to *D. melanogaster* coordinates from the DPGP multi-species alignment, which includes *D. melanogaster*, *D. simulans*, *D. yakuba* and *D. erecta* genomes (Langley et al. 2012). When a *D. simulans* allele was used as an outgroup in statistical inferences (see below), we chose the allele with the smallest proportion of missing data and alignment gaps (or highest base coverage) among the mosaic *D. simulans* genome (Clark et al. 2007) and six *D. simulans* genomes (Begun et al. 2007).
For Ago3, all D. simulans alleles from the above seven genomes had low coverage. We used D. melanogaster exon sequences of Ago3 to blast against the trace reads generated from D. simulans population genomics project (BEGUN et al. 2007) and assembled retrieved reads using the codoncode aligner (http://www.codoncode.com/aligner/). We removed reads whose alignment outside exons was incongruent with majority of other reads. Consensus sequence was called if there were at least three reads covering the region.

Population genetics and molecular evolution analysis of candidate genes

$\pi$ was estimated as average pairwise differences (NEI 1987). Lineage-specific divergences were estimated by maximum likelihood using PAML version 4 (YANG 2007) on the branch leading to D. melanogaster and D. simulans, using D. yakuba as the outgroup. Genes with fewer than 100 sites included in the PAML analysis, or with a $dS$ value smaller than 0.0001 were excluded. We used both D. melanogaster and D. simulans within-species polymorphism to carry out McDonald-Kreitman tests [two-species MK test, (McDONALD and KREITMAN 1991)]. Codons having more than two states within species were removed. Codons that are both polymorphic and divergent between species were counted as both polymorphism and divergence. We used the mutational path minimizing the number of nonsynonymous differences. $P$-values of MK tests were determined by Fisher’s Exact Tests (FET). For genes without D. simulans variation (Ago3 and mael), we carried out one-species MK tests using D. melanogaster polymorphism and the D. simulans allele with highest base coverage to count the number of fixations. For candidate genes with significant MK test results, we used Pfam (FINN et
al. 2009) with $E$-value cutoff $10^{-5}$ to annotate known domains and perform MK test on each annotated domain. We estimated average $\alpha$ (the proportion of amino acid fixations driven by positive selection) for different classes of genes using WELCH (2006) with default parameters. We also used the likelihood ratio test to investigate whether a single $\alpha$ or a two $\alpha$ model better fits the data when we included both candidate and immunity/all genes in the analysis, testing whether there are differences in $\alpha$ between classes of genes.

When comparing population genetic estimates or statistics of candidate genes with genome-wide distribution, we used a conservative gene set used by DPGP (LANGLEY et al. 2012), which consists of genes whose $D. melanogaster$ alleles of DPGP data and outgroup alleles all have the same gene model as the reference annotations (canonical initiation codon, splice junction and termination codon).

$D. melanogaster$ variation data before $P$ element invasions

Variation data from pre-$P$ element $D. melanogaster$ populations were collected by resequencing the coding regions of candidate genes from laboratory-maintained strains collected before the 1960s and previously identified as M strains (KIDWELL et al. 1983). PCR with primers amplifying the second exon of $P$ element transposase (O’HARE and RUBIN 1983; CLARK et al. 1994) was used to confirm the absence of $P$ elements. To have comparable sampling locations to the DPGP data, we first used four African strains (CA1, KSA2, KSA3 and KSA4) and four North and South Carolina strains (Wild 10E, Wild 11A, Wild 11C and Wild 11D) in the initial survey for unusual temporal differentiation. Five candidate genes ($Irbp$, $squ$, $Spn-E$, $Krimp$ and $Hen1$) showed significant differentiation between alleles from the above eight strains, and post-$P$
element alleles from DPGP data. Additional alleles were then collected on these five
genes using four Asian, five European, one South American and 11 Northern American
strains. Details of *D. melanogaster* strains used in this study can be found in
Supplementary Table 1. For control genes near Hen1, we only sequenced the 15 North
American M strains.

Despite exhaustive efforts to locate M strains collected before *P element*
horizontal transfer, the available *pre-P* element strains are far from ideal. Within North
America, where there is the largest set of *pre-P-element* strains, spatial locations are
disperse: strains were collected on the west and the east coast as well as the northern and
southern latitudes. Latitudinal clines for various loci have been observed in *D.

Unfortunately, this may increase the possibility of falsely concluding that there is
temporal genetic differentiation, while the actual difference would be a result of the
geographic heterogeneity of between-time samples. Accordingly, we also examined other
aspects of the data (heterozygosity and haplotypes) in addition to temporal differentiation
and included control genes near candidate genes showing strong temporal differentiation
before drawing conclusions (see below).

DNA samples were prepared with 30 flies from each *D. melanogaster* M strain.
PCR and sequencing primers for coding regions of candidate genes were designed using
the Primer3 program (Rozen and Skaltsky 2000) and the *D. melanogaster* reference
genome. PCR products were purified and sequenced directly. Most of the *D.
melanogaster* strains used in this study have been maintained in the laboratory for over
fifty years and are highly inbred. For targeted regions with residual heterozygosity within
lines, PCR products were cloned with TOPO-TA cloning (Invitrogen) and one clone of each PCR product was sequenced. Sequences generated are deposited to GenBank with accession number XXXX.

**Analysis of temporal differentiation between pre-P and post-P element populations**

Sequences of *pre-P* and *post-P element* populations were aligned using ClustalW (Chenna et al. 2003), followed by manual curation. We estimated $F_{st}$ according to Weir and Cockerham (1984) and used permutations to determine the $p$-values (Hudson et al. 1992). To further test for unusual haplotypic structures, we used methods based on the frequency of major haplotypes (Hudson et al. 1994), the number of haplotypes and the heterozygosity of haplotypes (Depaulis and Veuille 1998), and used coalescent simulation without recombination to determine the $p$-values (Hudson 2002). Although these three haplotype-based tests are related conceptually, their power to detect deviations from the same null hypothesis vary with the alternatives and thus are not fully redundant (Depaulis and Veuille 1998). It is worth noting that the significance of haplotype tests were based on coalescent simulations *without* recombination, which is especially conservative, and our observation of strong evidence for 20kb haplotypic structure around *Hen 1* is highly unusual (see results).

For analysis of the upstream and downstream regions of *Hen1*, we used sliding window of size 5kb incremented every 100bp to depict the divergence between *D. melanogaster* and *D. simulans*, the polymorphism within *D. simulans*, and the polymorphism of *D. melanogaster* African and North American populations separately.
Analytical model for the dynamics of host alleles that can reduce transposition during the spread of an invading transposable element such as the P element

We considered a panmictic population of diploid hosts with infinite population size and initially devoid of the invading TE. After invasion of the TE, each host genome carries a number ($n \geq 0$) of TEs and zero, one or two resistant alleles at the host locus of interest. We assumed that there is complete linkage equilibrium among the invading TEs, and the TEs and the host resistance locus. The low frequency of virtually all TE insertions in natural populations of D. melanogaster (Aquadro et al. 1986; Montgomer y et al. 1987; Charlesworth and Langley 1989; Lee and Langley 2010) coupled with the small scale of linkage disequilibrium between SNPs with more intermediate frequency in D. melanogaster (Miyashita and Langley 1988; Long et al. 1998; Langley et al. 2000, Langley et al. 2012) assure that the magnitude of linkage disequilibrium among elements is small and our assumption is reasonable. The assumption of no linkage disequilibrium and low TE frequencies motivate the further modeling of distribution of TE copy number as Poisson (see next section for details). The use of the Poisson distribution of TE copy number among individuals of a population has been developed as an approximation and successfully applied in theoretical analyses of TEs (Charlesworth and Charlesworth 1983; Langley et al. 1983). As mentioned above, it has an empirical basis in studies of specific TE families and surveys of genomic variation (reviewed in Charlesworth and Langley 1989). The transposition of P elements was also modeled following a Poisson process.

We considered two aspects of the deleterious effects of P elements on host fitness. According to previous theoretical analysis (Charlesworth and Charlesworth 1983),
in order to have stable containment of transposable elements, the logarithm of fitness
must decline more rapidly than linearly with average copy number. We considered a
synergistic epistasis for the deleterious effects of P elements insertions described
previously (Dolgin and Charlesworth 2006, 2008) and the fitness of an individual
with n copies of P element is:

\[ w(n) = e^{-an - bn^2/2}. \]

a and b were chosen as $10^{-5}$ and $10^{-6}$ respectively (see Appendix). The other deleterious
fitness effect is caused when the transposition of P elements generates more double-
stranded breaks than the host recombination repair machinery can efficiently repair,
leading to the reported reduced fertility (reviewed in Rio 2002). P element transposes
through a cut-and-paste mechanism and the process starts with a double-stranded break
generated at the original P element insertion (donor site). Approximately 85% of the
double-stranded breaks at the donor sites are repaired using the sister chromatid as the
template (Engels et al. 1990), resulting in regeneration of a P element at the donor site
and increase in copy number by one. With the assumption that every P element
transposition leads to a net gain of one P element copy, we used a truncation selection
model: an offspring with more than \( n_{HD} \) (the maximum number of new TE transposition a
host can tolerate before having hybrid dysgenic syndrome) new P element insertions is
sterile. The mean fitness of offspring of parents with average \( m \) P element copies is:

\[ \bar{w}(m, u) = \sum_{n=0}^{\infty} e^{-m} m^n \left( \sum_{i=0}^{n_{HD}} \frac{e^{-nu} (nu)^i}{i!} e^{-a(n+i) - b(n+i)^2/2} \right). \]

\( u \), the transposition rate per copy per generation, changes according to parental cytotypes,
with P element transposition rate in MxP dysgenesis cross (\( u_0 \)) being much higher than
that in other crosses ($u_i$). Individuals with $P$ elements are set to have $P$ cytotype and others without are set to have $M$ cytotype.

One of the two segregating alleles of the host locus is able to reduce $P$ element transposition rate by a proportion $d$ in homozygote (i.e. the transposition rate is then $u(1 - d)$ in homozygotes). The heterozygotic effect of this allele is $h$ of that of homozygotes and transposition rate in heterozygotes is thus $u(1 - hd)$. We were interested in three aspects of the host population that changed over generations: (1) the proportion of $P$ cytotype, $r$, (2) the allele frequency of host allele reducing $P$ element transposition, $l$, (3) the average copy number of $P$ elements among individuals with $P$ cytotype, $\mu$. With the assumption that there is linkage equilibrium between $P$ element insertions and the host locus, the reduction of $P$ element transposition rate due to the host locus can be considered as independent event from the type of cross. The mean fitness of offspring of a specific cross with $m$ parental $P$ element copies is:

$$w_{cross}(m,u,l) = l^2w(m,u(1-d)) + 2l(1-l)w(m,u(1-hd)) + (1-l)^2w(m,u).$$

Based on this formula, we derived equations for $r$, $l$ and $\mu$, which can be found in the Appendix. We used R to calculate $r$, $l$ and $\mu$ for 1,000 – 10,000 generations. For most cases, we reported the result for 1,000 generations, the approximate number of $D. melanogaster$ generations since the M strains were first collected. At generation zero, we assumed $r_0$ and $l_0$ equal to $10^{-3}$. Parameters without a significant impact on the dynamics of $l$ are set as constant values ($d = 0.5$, $h = 0.5$, $u_i = 10^{-4}$ and $\mu_0 = 10$, see Appendix for results of all parameters tested). $u_0$ was tested for $10^{-1}$ and 1. $n_{HD}$ was tested for 2, 3, 5, 7, and 10.
RESULTS

Candidate genes

We took a candidate gene approach and focused on two groups of genes (Table 1). The first group consists of genes known to be involved in the piwi-RNA (piRNA) biogenesis [hereafter termed piRNA genes (reviewed in Klattenhoff and Theurkauf 2008)]. piRNA is a class of small RNAs that has been implicated in TE transposition rate regulation (reviewed in Klattenhoff and Theurkauf 2008). Generation of piRNAs is disrupted in piRNA gene mutants (Aravin et al. 2004; Lim and Kai 2007; Pane et al. 2007; Li et al. 2009; Klattenhoff et al. 2009), leading to elevated expression levels of more than a dozen TE families (Vagin et al. 2004; Aravin et al. 2004; Lim and Kai 2007; Pane et al. 2007; Li et al. 2009; Klattenhoff et al. 2009; Lu and Clark 2010).

We also included vasa (vas), whose mutant phenotypes also include piRNA generation disruptions and elevated transcription of several TE families (Vagin et al. 2004; Lim and Kai 2007). piRNAs corresponding to P element sequences have been observed in P strains but not M strains, suggesting P elements are a common target of the piRNA pathway (Brennecke et al. 2008).

The second group of candidates contains genes known to interact with P elements via other pathways. The double-stranded breaks left after P element transpositions in the germline are repaired by host recombination–repair machinery of heterodimers formed by Irbp and Ku80 (Rio and Rubin 1988; Beall et al. 1994). Double-stranded breaks generated by other DNA-based TEs may be repaired using a similar mechanism. Splicing factor Psi is shown to specifically bind to the 5’ splice site of the P element third intron, suppressing the proper splicing of mRNA of P element transposase and thereby repressing

**Genes interacting with transposable elements show evidence of positive selection**

Recurrent directional selection can lead to an accelerated rate of protein divergence relative to synonymous site divergence. We used maximum likelihood methods (YANG 2007) to estimate the *dN/dS* ratios on both the *D. melanogaster* and *D. simulans* branches with *D. yakuba* as an outgroup and then rank estimates of candidate genes among other *D. melanogaster* annotated genes (Table 2 for *dN/dS* ratio and Supplementary Table 2 for separate *dN* and *dS* values). Consistent with a previous report (VERMAAK *et al.* 2005), *rhi* has a *dN/dS* ratio that is larger than one and is among the fastest evolving genes in *D. melanogaster* (*dN/dS* = 1.415, rank 0.31% genome-wide). Two nuage component genes, *krimp* and *mael*, both rank in the top 5% genome-wide while *aub* and *zuc* are among top 10%. Estimates of the *dN/dS* ratios on the *D. simulans* branch, except for that of *zuc*, gave similar results. Our inferences are generally in agreement with previous reports that used a branch-site model to detect recurrent amino acid substitutions on the phylogeny of 12 *Drosophila* species (HEGER and PONTING 2007; KOLACZKOWSKI *et al.* 2011). Even though *Spn-E* was identified as the *RNA interference* gene showing the most extensive signal of recurrent adaptation on the phylogenetic tree (HEGER and PONTING 2007; KOLACZKOWSKI *et al.* 2011), the branch leading to *D. melanogaster* was not significant (KOLACZKOWSKI *et al.* 2011). The differences between studies are not surprising given the fundamental differences in methodology (comparing
relative rates of amino acid substitutions of entire coding sequences among all genes

*versus* identification of a subset of sites or branches that recurrently substituted across the

phylogeny).

We used the McDonald-Kreitman test (*MK test*, MCDONALD and KREITMAN 1991) to detect genes whose evolution does not follow the neutral model of evolution.

Rejection of the null hypothesis due to the presence of more than the expected number of

amino acid fixations has been interpreted as evidence supporting adaptive protein

evolution. To have greater statistical power, we considered polymorphisms from both *D.

melanogaster* (LANGLEY et al. 2012) and *D. simulans* (BEGUN et al. 2007). We identified

*aub* and *armi* as significant while *Spn-E, krimp, vas* and *Ku80* were marginally

significant (Table 3). All of these genes rejected the null hypothesis of neutral evolution

in the direction of an excess of amino acid divergence (Table 3), consistent with a history

of positive selection acting on these genes. To further localize protein domains showing

signals of positive selection, we performed *MK tests* on Pfam-annotated domains. We

found the PAZ domain of *aub* (*p-value* = 0.011) and DEAD domain of *vas* (*p-value* =

0.0007) showed significant enrichment of amino acid substitutions. When only

considering within *D. melanogaster* polymorphism, we did not find evidence of adaptive

evolution for *Spn-E, vas* and *Ku80* (Table 3), perhaps due to a generally lower level of

variation in *D. melanogaster* than in *D. simulans* (AQUADRO et al. 1988; ANDOLFATTO

2001; ANDOLFATTO et al. 2011), and thus lower statistical power.


*Genes interacting with transposable elements show more prevalent evidence of*

*positive selection than immunity genes*
The fundamental differences in the mechanism of transmission between TE families and other nongenomic pathogens raised the question of what was the relative intensity of evolutionary impacts they imposed on hosts. To address this question, we compared the proportion of genes exhibiting evidence of positive selection between our candidate genes and immunity genes using the same population genetic and molecular evolution analysis.

We found five out of 12 piRNA genes (41.67%) have D. melanogaster dN/dS estimates among the top 10% genome-wide, which is significantly greater than that of immunity genes (24 out of 214 genes, 11.21%, \(FET p = 0.01\), Figure 1A and Supplementary Table 2). Studies have found that the rates of adaptive evolution vary with respect to the function of immunity genes (SACKTON et al. 2007; OBBARD et al. 2009b; and reviewed in LAZZARO 2008). We categorized immunity genes into “recognition”, “signaling” and “effector” (see methods) and still found piRNA genes have a higher proportion of fast evolving genes than all categories of immunity genes. However, we only found statistical significance when comparing piRNA genes to either “signaling” or “effector” categories (Figure 1A). Comparisons considering all candidate genes (both piRNA genes and genes known to interact with P elements; Figure 1A), or focusing on relative rates of amino acid evolution on the D. simulans lineage (data not shown), are both consistent with our findings using piRNA pathway gene evolution along the D. melanogaster lineage.

We observed even more dramatic enrichment in the proportion of genes showing evidence of recurrent adaptive evolution (rejection of MK tests with overabundant amino acid fixation) of piRNA genes (five out of 10 genes, 50%) than that of immunity genes.
(18 out of 180 genes, 9.52%, \textit{FET p-values} = 0.002; Figure 1B) and the genome-wide proportion (888 out of 8,085 genes, 10.98%, \textit{FET p-values} = 0.003). “Signaling” genes have the largest proportion of genes showing adaptive evolution among three immunity gene categories (13 out of 97 genes, 10.92%), which is still significantly lower than \textit{piRNA} genes or all candidate genes (\textit{FET p-values} = 0.012). We found a consistent pattern when including all candidate genes in the comparisons.

We used the maximum-likelihood method proposed by Welch (2006) to formally test whether averaged \( \alpha \) (the proportion of amino acid substitution fixed by positive selection) is different between classes of genes. When comparing either “\textit{piRNA v.s. immunity genes}” or “\textit{piRNA v.s. all genes}”, we found a two \( \alpha \) consistently fits the data better \( [A = 446.46 \text{ (v.s. immunity genes)} \text{ and } 613.82 \text{ (v.s. all genes)}; p < 0.001 \text{ for both comparisons}] \). Comparisons between \textit{piRNA} genes and a specific subset of immunity genes were also highly significant \( (p < 0.001) \). Permutation analysis also found the maximum-likelihood estimated \( \alpha \) of \textit{piRNA} genes \( (\alpha = 0.82) \) is significantly higher than that of immunity genes \( (\alpha = 0.40 \text{ for all immunity genes}) \text{ and all genes } (\alpha = 0.36), \text{ except for “recognition” immunity genes (Figure 1C). Again, comparisons considering all candidate genes gave consistent results. Another estimates of adaptive protein evolution \( \omega_\alpha \), the rate of adaptive substitution relative to the rate of neutral substitutions \( (\text{GOSSMANN et al. 2010}) \), supports the same conclusion based on \( \alpha \) (Figure 1D). Either \textit{piRNA} genes or all candidate genes have larger \( \omega_\alpha \) than immunity genes except for “recognition” immunity genes \( (\text{Mann-Whitney U test, } p < 0.05) \text{ and all genes } (\text{Mann-Whitney U test, } p < 0.001) \), suggesting that the larger \( \alpha \) of TE-interacting genes is not due to differences in proportion of effectively neutral mutations.
Horizontal transfer of TE families does not impose enduringly strong selection on host beneficial variants

We used a deterministic model to analyze the dynamics of host alleles that can reduce the transposition rate of a newly horizontally transferred TE family by a fixed proportion during the spread of that TE family in a panmictic host population (see methods and Appendix for model details). We specifically considered the well-documented horizontal transfer of P elements, which provided the biological context needed to specify details of the model. The model considered both epistatic selections against increases in P element copy number (Charlesworth and Charlesworth 1983; Dolgin and Charlesworth 2006, 2008) and host sterility caused by too many double-stranded chromosomal breaks generated through P element transposition. We set individuals with more than $n_{HD}$ new P element transpositions (and thus double-stranded breaks) to be completely sterile. This sterility effect would most likely occur in the M(female) x P(male) hybrid dysgenic cross, as the transposition rate of P elements in dysgenic cross ($u_0$) is several orders of magnitude higher than in nondysgenic cross ($u_1$) (Eggleston et al. 1988). The host beneficial allele will especially enjoy strong fitness advantages in the dysgenic crosses because hosts with this allele will be more likely to have fewer than $n_{HD}$ double-stranded breaks and therefore higher expected fertility. Of course, the degree of sterility elicited by P element transposition can be a continuous phenotype. The usage of the truncation selection model will maximize the selective benefit of a suppressive host allele, making our overall conclusion conservative.
The increase in frequency of the host beneficial allele ($l$) is dependent on how fast
**P elements** spread through the population and, during its spread, how likely hybrid
dysgenesis is to occur. We found that the spread of **P elements** is fast in most cases
(Figure 2A for $u_0 = 1$ and $n_{HD} = 5$, See Appendix for discussions of other cases), a finding
that is consistent with several caged experiments introducing **P elements** into M strain
populations (KIDWELL et al. 1988; GOOD et al. 1989). This quick spread leads to the host
beneficial allele having selective advantage only during a narrow period (Figure 2B). We
found that the largest selective advantage occurs when the proportion of **P cytotype**
individuals ($r$) is intermediate, and the probability of a hybrid dysgenesis cross is high
(Figure 2C). This is also reflected in the dynamics of $l$, which has a phase of rapid
increase followed by a longer phase of slow increase (Figure 2D).

The combined effects of $u_0$ and $n_{HD}$ had the most noticeable influence on the
dynamics of the host beneficial allele. With increased $n_{HD}$, it takes fewer generations until
the **P element** is found in nearly all genomes ($r > 0.99$) in the population ($t_{0.99}$) (Figure
3A) and $l_{1000}$ is smaller (host beneficial allele frequency at generation 1,000) because of
the decreased duration during which the probability of hybrid dysgenesis is high (Figure
3B). Generally, **P elements** spread faster when $u_0$ is larger (except for $n_{HD} = 2$, see
below). Yet, the probability of hybrid dysgenesis is also higher, leading to larger $l_{1000}$
(Figure 3A and 3B). For all cases examined except one (see Appendix for all parameters
tested), the difference between $l_0$ and $l_{1000}$ (< 2%) would hardly be detected with the
regular size of samples.

The only exception is when $u_0 = 1$ and, in which the temporal $F_{st}$ can be as large
as 0.15. However, it takes more than 10,000 generations for the **P element** to become
nearly fixed in the population due to the still high probability of hybrid dysgenesis even when the majority of the hosts have \textit{P cytotype}. This is much longer than the time scale of actual \textit{P element} spread. Furthermore, the tolerance of non-programmed double-stranded breaks in the germline can be much higher than this threshold ($n_{HD} = 2$; Orsi \textit{et al.} 2010).

It is worth noting that we modeled the distribution of \textit{P element} copy number as approximately \textit{Poisson}. The derivation and previous application of the \textit{Poisson} approximation were focused on situations where the TEs population is near equilibrium and the mean copy number of TEs is much larger than one (Charlesworth and Charlesworth 1983; Langley \textit{et al.} 1983). However, the critical aspect of our analytical modeling here is the initial phase of the invasion of a new TE family (the \textit{P element}), which may lead to a copy number distribution different from \textit{Poisson}. We thus used Monte Carlo simulations (see Supplementary text for details of simulations) to investigate how the distribution of TE copy number reaches a \textit{Poisson} distribution. We find that soon after the \textit{P cytotype} is common in the population, the TE copy number distribution is close to \textit{Poisson} (Figure S5 of Supplementary text). More importantly, these simulations show that the estimated change of host beneficial allele frequency from the analytical approximation is within 2\% of the simulated results (Figure S3 and S4 of Supplementary text), supporting our overall conclusions.

\textbf{Recent horizontal transfer of \textit{P element} does not have widespread evolutionary impacts on candidate genes}

To empirically investigate the short-term evolutionary impacts of TE horizontal transfer on hosts, we compared the genetic differentiation between \textit{pre-\textit{P element}}
invasion (pre-P) and current (post-P) populations (temporal differentiation). For five candidate genes that showed strong temporal differentiation in our initial survey with a smaller number of M strains (Irbp, krimp, Hen1, Spn-E and squ; see methods and Supplementary Table 3), only Hen-1 and squ still showed highly significant temporal differentiation with increased size of pre-P element samples (Table 4). However, we observed strong genetic differentiation between North American and African contemporary post-P populations for our candidate genes (geographic differentiation, Table 4). If the geographic differentiation of the current population was also present in the pre-P population, the wide geographic distribution of M strains used may lead to false conclusions of the temporal differentiation (see methods and Supplementary Table 1).

We further restricted our analysis to the North American population, which has the largest number of alleles for both pre-P (15) and post-P (37) samples and still found strong temporal differentiation of Hen1 (Fst = 0.173, p = 0.003; Table 4). Such differentiation may reflect in divergence in protein functions, as Fst estimated using amino acid sequences also showed significant temporal differentiation (Fst = 0.261, p < 0.001; Table 4).

**The strong temporal differentiation of Hen1 is likely the result of genetic hitchhiking from a nearby, strongly selected gene**

Given that the spread of P elements is fairly recent and fast, any associated directional selection on suppressive host variants is expected to result in reduced genomic polymorphism around the selected SNPs and increased linkage disequilibrium due to genetic hitchhiking (Maynard Smith and Haigh 1974; Kim and Nielsen 2004;
Consistent with the analysis of temporal differentiation, we observed marginally significant haplotypic structures for the Hen1 gene region for post-P North American populations [frequency of major haplotypes = 15, \( p = 0.042 \) (Hudson et al. 1994) and haplotype heterozygosity = 0.759, \( p = 0.031 \) (Depaulis and Veuille 1998)]. This strong haplotypic structure was extended at least 10kb upstream and downstream of Hen1 (frequency of major haplotypes = 10, \( p = 0.003 \) and haplotype heterozygosity = 0.290, \( p = 0.001 \)). We also observed a dramatic reduction of post-P element North American variation for a almost 100kb genomic segment around Hen1 when compared with polymorphism in post-P element African D. melanogaster, polymorphism in closely related D. simulans and divergence between D. melanogaster and D. simulans (Figure 4A).

However, the genomic region around Hen1 is highly gene-rich, with more than twenty genes, including Cyp6g1. Studies have identified a recent selective sweep associated with the Cyp6g1 allele that has an Accord transposable element inserted upstream (Daborn et al. 2002; Catania et al. 2004; Chung et al. 2007; Schmidt et al. 2010). Functional analysis confirmed that this Accord insertion confers insecticide resistance (Daborn et al. 2002; Catania et al. 2004; Schmidt et al. 2010), which is the most likely force driving the strong, recent selective sweep on Cyp6g1. The Accord inserted allele was found fixed in non-African populations, yet it was intermediate in African population (Catania et al. 2004), which may also explain why the reduction of heterozygosity around Hen1 was most apparent in the North American populations while P element is virtually fixed worldwide.
To further investigate whether the strong temporal differentiation observed on
Hen1 is the result of genetic hitchhiking from strongly selected Cyp6g1 or any other
genes in the nearby region, we estimated the temporal differentiation of coding regions
for another 11 genes that are within 30kb to either Hen1 or Cyp6g1 and have functions
unrelated to either TE suppression or insecticide resistant (“control genes”,
Supplementary Table 4 and Figure 4C). Compared with the observed strong temporal
differentiation of these control genes, Hen1 (Fst = 0.173) is no longer exceptional (Table
5 and Figure 4B). We also found that the closer a gene is to Cyp6g1, the stronger the
temporal differentiation was (Table 5 and Figure 4B). The geographic differentiation was
also strong for these 11 genes (Table 5), which is consistent with the scenario that
application of insecticide in the non-African regions is leading to strong genetic
hitchhiking on genes around Cyp6g1 in the post-P North American population we
studied.

DISCUSSION

TEs are selfish genetic elements in the genome. Their interactions with their hosts
are often analogized to the molecular arms race between hosts and other nongenomic
pathogens, such as bacteria, viruses, fungi and protozoa. This analogy deserves further
mechanistic specification and analysis because of the fundamental differences in
transmission mode between TEs and horizontally transmitted pathogens.

To address this analogy, we first systematically compared the long-term evolution
of host TE-interacting genes to that of immunity genes. We found the proportion of TE-
interacting genes with evidence of positive selection is at least as large as, if not greater
than, that of genes in pathways conferring immunity to pathogens. *aub*, which showed strong evidence of adaptive protein evolution, is a key component in *piRNA*-mediated TE silencing (Gunawardane *et al.* 2007; Brennecke *et al.* 2007) and its PAZ domain, known to mediate the binding of single-strand *RNAs* associated with Argonaute/Piwi proteins (Lingel *et al.* 2003; Yan *et al.* 2003; Song *et al.* 2003), also showed an excess of amino acid fixations. Both RNA helicases surveyed, *armi* and *Spn-E*, showed evidence of adaptive evolution, although the detailed mechanism of their involvement in *piRNA* biogenesis and TE suppression is still not clear. Most interestingly, all three nuage component genes (*mael, krimp* and *vas*) showed strong evidence of positive selection.

The nuage, where many proteins encoded by *piRNA* genes localize (reviewed in Klattenhoff and Theurkauf 2008), is considered the major battleground for host-TE arms race (Blumenstiel 2011). In addition, *vas* is involved in the formation of pole plasm (the future germline) and is essential for the proper localization and translational control of maternally deposited *mRNAs* at the rear of developing embryos (Lasko and Ashburner 1990; Stybler *et al.* 1998; Johnstone and Lasko 2004). RNA genome of TE* s* that can be preferentially incorporated into the pole plasm and have its RNA tertiary structure being detangled and translated will have large fitness benefits from increased transmission. DEAD box of *VAS*, the key domain mediating the unwinding of RNAs with self-annealed structure (reviewed in Linder 2006; Arkov and Ramos 2010), had a strong enrichment of amino acid fixations under McDonald-Kreitman framework.

Our analytical model found that host alleles suppressing *P element* transposition enjoy strong selective benefits only during a limited time frame due to the fast spread of *P elements* through strict vertical parent-offspring inheritance. This led to negligible host
allele frequency differences between pre and post-P element populations with all biologically reasonable parameters tested. Such theoretical prediction may be extended to other known cases of horizontally transferred TEs in Drosophila, such as I-element, hobo and mariner (reviewed in Silva et al. 2004). For DNA-based TE families that also transpose with the cut-and-paste mechanism (hobo and mariner), our model should be readily applicable. RNA-based TE families, such as I-element, transpose through a “copy-and-paste mechanism” (replicating through insertions of reversed transcribed cDNA into another position in the genome). The hybrid dysgenic syndrome observed for I-element is caused by the catastrophic meiosis of eggs, which includes failure to produce functional female pronucleus and developmental abnormalities after fertilization (Orsi et al. 2010). If such meiotic defect is initiated with a higher than tolerable threshold of I-elements transposition activity and is also responsible for the hybrid dysgenic syndromes of other unobserved horizontal transfer of RNA-based TEs, our model could be a reasonable generalization for all TE families.

Consistent with our quantitative analysis, we found no strong evidence supporting recent selection imposed by P elements on candidate genes. The significant temporal differentiation and haplotypic structures of Hen1, our only strong candidate, is likely to be the result of strong genetic hitchhiking effects of the nearby insecticide resistant gene, Cyp6g1. Certainly it is possible that other types of alleles and other loci may have responded to the P element invasion. We did not find a strong reductions in polymorphism of post-P element population 10kb upstream and downstream of each candidate gene, providing no support for the alternative that selection has acted on polymorphisms of local cis-acting elements (data not shown). Yet, the possibility that
strong selection on cis-regulatory elements outside the surveyed region or trans-acting regulatory variation cannot be ruled out. The possibility that there was instead strong selection for a P element variant that can reduce the deleterious impact on its host is not plausible because there is only one-base difference observed between D. willistoni and D. melanogaster canonical P elements (Daniels et al. 1990). Still another alternative is the prevalent D. melanogaster variants of host genes segregating in the pre-P element population were already similar to those of D. willistoni, which coevolved with P elements and can thus effectively reduce P element transposition and minimize their deleterious effects. We found that the amino acid sequence divergence between D. melanogaster and D. willistoni orthologs (Clark et al. 2007) for most candidate genes are greater than the genome-wide median (Figure 5). The only two exceptions [Hrb27C (1.7%) and Psi (16.4%)] both have other essential host functions (Hammond et al. 1997; Labourier et al. 2002; Dreyfuss et al. 2002; Goodrich et al. 2004; Huynh et al. 2004; Yano et al. 2004; Blanchette et al. 2005) and their evolution is expected to be strongly constrained. These observations therefore do not support the scenario of D. melanogaster pre-adaptation to P elements.

The selection coefficient for a host allele that can reduce TE transposition rate in an outbred population with limited linkage disequilibrium is (Charlesworth and Langley 1986):

\[ s = \frac{\delta u \left( \frac{\bar{n} u}{2 \tilde{H}} \right)}{2} \]

which depends on the change in transposition rate \( \delta u \) and the expected number of new TE copies (average copy number \( \bar{n} \) times transposition rate \( u \)). \( \tilde{H} \) is the approximated harmonic mean of recombination frequency between pairs of TE insertions, which
approaches half in species with free recombination. This theoretical prediction suggests that the selective benefit of a host allele increases with the number of TE copies whose transposition it can suppress. Given the TE transposition rate \(10^{-5} \sim 10^{-4}\) (NUZH DIN and MACKAY 1995; NUZH DIN et al. 1997; MASIDE et al. 2000, 2001; and reviewed in CHARLES WORTH and LANGLEY 1989; LE ROUZIC and DECE LIERE 2005)] and the number of active copies of individual TE families [below a hundred (KAMIN KER et al. 2002; QUES NEVILLE et al. 2005; BERGMAN et al. 2006)], the selective benefits for a host allele targeting a specific TE family is small. However, Drosophila genomes are occupied by more than a hundred TE families (KAMINKER et al. 2002; QUESNEVILLE et al. 2005; BERGMAN et al. 2006; Clark et al. 2007), and TE families can be further classified into clade or superfamilies according to encoded protein products and similarities in sequences [such as Ty1-copia-like or Ty3-gypsy like, (reviewed in WICKER et al. 2007)]. In this case, a host variant targeting attributes shared between multiple TE families can enjoy larger selective benefits than a variant targeting a single TE family and is likely to spread in the host population. For example, a host allele that can reduce half of the transposition rate of a subset of TE families with total one thousand copies can have selection coefficient as large as \(10^{-5}\), which would be strong enough to overcome the effect of genetic drift in D. melanogaster \([N_e \sim 10^6\), (LANGLEY et al. 1982; KREITMAN 1983)]. Accordingly, unlike the arms race between host and horizontally transferred pathogens, where strong selective benefit comes from the precise targeting of host allele to a specific pathogen variant, the antagonistic interaction between a host variant targeting the aggregated influence of multiple vertically inherited TE families could be the main force driving the fast evolution of host TE-interacting genes.
ACKNOWLEDGMENT

Thanks to Drosophila Bloomington stock center and M. Itoh for providing M strains. We appreciate D.J. Begun and M.T. Levine for extensive discussion of the project and to D.J. Begun, M.T. Levine, P.L. Ralph, M. Turelli and E.I. Dietrich for critical reading of the manuscript. We also thank D.J. Obbard for helpful discussion and kindly sharing unpublished results regarding MK tests for piRNA genes. We are grateful to C.M. Cardeno, P. Saelao, and T. Trejo-Cantwell for experimental assistance. Finally, we acknowledge the thoughtful, constructive and patient criticism of two anonymous reviewers.

REFERENCES


1 ANDOLFATTO P., WONG K. M., BACHTROG D., 2011 Effective Population Size and the
2 Efficacy of Selection on the X Chromosomes of Two Closely Related

4 ANXOLABÉHÈRE D., KIDWELL M. G., PERIQUET G., 1988 Molecular characteristics of
5 diverse populations are consistent with the hypothesis of a recent invasion of

7 AQUADRO C. F., DESSE S. F., BLAND M. M., LANGLEY C. H., LAURIE-AHLBERG C. C., 1986
8 MOLECULAR POPULATION GENETICS OF THE ALCOHOL DEHYDROGENASE
9 GENE REGION OF DROSOPHILA MELANOGASTER. Genetics 114: 1165 –
10 1190.

12 melanogaster and Drosophila simulans. I. Contrasting levels of naturally
13 occurring DNA restriction map variation and divergence. Genetics 119: 875–
14 888.


17 ARAVIN A. A., KLENOV M. S., VAGIN V. V., BANTIGNIES F., CAVALLI G., GVOZDEV V. A., 2004
18 Dissection of a Natural RNA Silencing Process in the Drosophila


CHENNA R., SUGAWARA H., KOIKE T., LOPEZ R., GIBSON T. J., HIGGINS D. G., THOMPSON J. D.,
2003  Multiple sequence alignment with the Clustal series of programs.
Nucleic Acids Res. 31: 3497–3500.

CHUNG H., BOGWITZ M. R., MCCART C., ANDRIANOPoulos A., FFRENCH-CONSTANT R. H.,


DABORn P. J., YEN J. L., BOGWITZ M. R., GOFF G. LE, FEIL E., JEFFERS S., TIJET N., PERRY T.,
HECKEL D., BATTERHAM P., FEYEREISEN R., WIlSON T. G., FFRENCH-CONSTANT R. H.,


1. Dolgin E. S., Charlesworth B., 2008  The effects of recombination rate on the
distribution and abundance of transposable elements. Genetics 178: 2169–
2177.

2. Dreyfuss G., Kim V. N., Kataoka N., 2002  Messenger-RNA-binding proteins and the

does not mobilize other transposable element families in D. melanogaster.
Nature 331: 368–370.

4. Engels W. R., 1979  Germ line aberrations associated with a case of hybrid


203–218.

8. Finn R. D., Mistry J., Tate J., Coggill P., Heger A., Pollington J. E., Gavin O. L.,


Research 38: D211–D222.

1 GONZÁLEZ J., LENKOV K., LIPATOV M., MACPHERSON J. M., PETROV D. A., 2008  High rate of


7 GOODRICH J. S., CLOUSE K. N., SCHÜPBACH T., 2004  Hrb27C, Sqd and Otu cooperatively
regulate gurken RNA localization and mediate nurse cell chromosome

10 GOSSMANN T. I., SONG B.-H., WINDSOR A. J., MITCHELL-OLDS T., DIXON C. J., KAPRALOV M. V.,
FILATOV D. A., EYRE-WALKER A., 2010  Genome Wide Analyses Reveal Little
Evidence for Adaptive Evolution in Many Plant Species. Mol Biol Evol 27:
1822–1832.

14 GUNAWARDANE L. S., SAITO K., NISHIDA K. M., MIYOSHI K., KAWAMURA Y., NAGAMI T., SIOMI
H., SIOMI M. C., 2007  A slicer-mediated mechanism for repeat-associated


18 HAMMOND L. E., RUDNER D. Z., KANAAR R., RIO D. C., 1997  Mutations in the hrp48 gene,
which encodes a Drosophila heterogeneous nuclear ribonucleoprotein
particle protein, cause lethality and developmental defects and affect P-

HEGER A., PONTING C. P., 2007 Evolutionary rate analyses of orthologs and paralogs

HIRAIZUMI Y., 1971 Spontaneous recombination in Drosophila melanogaster males.

HOFFMANN A. A., WEEKS A. R., 2007 Climatic selection on genes and traits after a 100
year-old invasion: a critical look at the temperate-tropical clines in
Drosophila melanogaster from eastern Australia. Genetica 129: 133–147.

HUDSON R. R., 2002 Generating samples under a Wright-Fisher neutral model of

positive selection in the superoxide dismutase (Sod) region of Drosophila


Drosophila hnRNPA/B homolog, Hrp48, is specifically required for a distinct


LANGLEY C. H., LAZZARO B. P., PHILLIPS W., HEIKINEN E., BRAVERMAN J. M., 2000  Linkage disequilibria and the site frequency spectra in the su(s) and su(w(a)) regions of the Drosophila melanogaster X chromosome. Genetics **156**: 1837–1852.


1 LEE Y. C. G., LANGLEY C. H., 2010 Transposable elements in natural populations of
3 1228.
4 Li C., VAGIN V. V., Lee S., Xu J., Ma S., Xi H., SEITZ H., HORWICH M. D., SYRZYCKA M., HONDA
5 B. M., KITTLER E. L. W., ZAPP M. L., KLATTENHOFF C., SCHULZ N., THEURKAUF W. E.,
6 WENG Z., ZAMORE P. D., 2009 Collapse of germline piRNAs in the absence of
8 Lim A. K., KAI T., 2007 Unique germ-line organelle, nuage, functions to repress
11 LINDER P., 2006 Dead-box proteins: a family affair—active and passive players in
16 region contribute to naturally occurring variation in bristle number in
18 Loreto E. L. S., CARARETO C. M. A., CAPY P., 2008 Revisiting horizontal transfer of


1 Montgomery E., Charlesworth B., Langley C. H., 1987  A test for the role of natural
2 selection in the stabilization of transposable element copy number in a

5 rearrangement by ectopic recombination in Drosophila melanogaster:

7 Nei M., 1987 Molecular evolutionary genetics.

8 Nuzhdin S. V., Mackay T. F., 1995  The genomic rate of transposable element

10 Nuzhdin S. V., Pasyukova E. G., Mackay T. F., 1997  Accumulation of transposable
11 elements in laboratory lines of Drosophila melanogaster. Genetica 100: 167–
12 175.

13 O'Hare K., Rubin G. M., 1983  Structures of P transposable elements and their sites of
14 insertion and excision in the Drosophila melanogaster genome. Cell 34: 25–
15 35.

16 Obbard D. J., Gordon K. H. J., Buck A. H., Jiggins F. M., 2009a  The evolution of RNAi as
17 a defence against viruses and transposable elements. Philos. Trans. R. Soc.

19 Obbard D. J., Welch J. J., Kim K.-W., Jiggins F. M., 2009b  Quantifying adaptive
dysgenesis is associated with catastrophic meiosis and abnormal zygote

PANE A., WEHR K., SCHÜPBACH T., 2007  zucchini and squash Encode Two Putative
Nucleases Required for rasiRNA Production in the Drosophila Germline.

PASYUKOVA E. G., NUZHDIN S. V., MOROZOVA T. V., MACKAY T. F. C., 2004  Accumulation of
transposable elements in the genome of Drosophila melanogaster is

QUESNEVILLE H., BERGMAN C. M., ANDRIEU O., AUTARD D., NOUAUD D., ASHBURNER M.,
ANXOLABÉHERE D., 2005  Combined Evidence Annotation of Transposable

RIO D. C., 2002 P transposable elements in Drosophila melanogaster. In: Mobile DNA

RIO D. C., RUBIN G. M., 1988  Identification and purification of a Drosophila protein
that binds to the terminal 31-base-pair inverted repeats of the P

ROUZIC A. LE, DECELIÈRE G., 2005  Models of the population genetics of transposable
1 Rozen S., Skaletsky H., 2000  Primer3 on the WWW for general users and for

2 Rubin G. M., Kidwell M. G., Bingham P. M., 1982  The molecular basis of P-M hybrid


4 Schaack S., Gilbert C., Feschotte C., 2010  Promiscuous DNA: horizontal transfer of
transposable elements and why it matters for eukaryotic evolution. Trends

5 Schmidt J. M., Good R. T., Appleton B., Sherrard J., Raymant G. C., Bogwitz M. R.,
Martin J., Daborn P. J., Goddard M. E., Batterham P., Robin C., 2010  Copy
number variation and transposable elements feature in recent, ongoing
adaptation at the Cyp6g1 locus. PLoS Genet. 6: e1000998.

6 Schmidt P. S., Matzkin L., Ippolito M., Eanes W. F., 2005  Geographic variation in
diapause incidence, life-history traits, and climatic adaptation in Drosophila

7 Siebel C. W., Admon A., Rio D. C., 1995  Soma-specific expression and cloning of PSI, a

SIEBEL C. W., KANAAR R., RIO D. C., 1994  Regulation of tissue-specific P-element pre-mRNA splicing requires the RNA-binding protein PSI. Genes Dev. 8: 1713–1725.


1 Stephan W., Song Y. S., Langley C. H., 2006 The Hitchhiking Effect on Linkage

2 Styhler S., Nakamura A., Swan A., Suter B., Lasko P., 1998 vasa is required for
   GURKEN accumulation in the oocyte, and is involved in oocyte differentiation

3 Vagin V. V., Klenov M. S., Kalmykova A. I., Stolyarenko A. D., Kotelnikov R. N.,
   Gvozdev V. A., 2004 The RNA interference proteins and vasa locus are
   involved in the silencing of retrotransposons in the female germline of

4 Vermaak D., Henikoff S., Malik H. S., 2005 Positive selection drives the evolution of
   rhino, a member of the heterochromatin protein 1 family in Drosophila. PLoS

5 Weir B. S., Cockerham C. Cl., 1984 Estimating F-Statistics for the Analysis of

6 Welch J. J., 2006 Estimating the Genomewide Rate of Adaptive Protein Evolution in

7 Wicker T., Sabot F., Hua-Van A., Bennetzen J. L., Capy P., Chalhoub B., Flavell A.,
   A unified classification system for eukaryotic transposable elements. Nat Rev
   Genet 8: 973–982.


FIGURE 1. Proportion of candidate and immunity genes showing evidence of positive selection. Proportion of candidate and immunity genes having *D. melanogaster* $dN/dS$ among top 10% genome-wide are shown in (A). The proportion of candidate gene, immunity gene and all gene having significant two-species *MK tests* ($p$-value $< 0.05$) and positive $\alpha$ are shown in (B). Dashed lines are the expectations assuming uniformity. Number of genes with *MK* test and PAML results in each category is shown in parenthesis. Figure (C) is the maximum-likelihood estimates of averaged $\alpha$ and Figure (D) is the boxplots (25, 50 and 75 percentile) of estimated $\omega_a$ for different class of genes. Error bars represent the 95% bootstrapping intervals around each estimate. Significant comparisons between *piRNA* genes and another class of genes are denoted * (Fisher-exact test, $p$-value $< 0.05$), ** ($p$-value $< 0.01$), and *** ($p$-value $< 0.001$). Comparisons of proportions (A, B) were based on Fisher-exact test, comparisons of maximum-likelihood estimated $\alpha$ (C) was based on permutations and comparison of $\omega_a$ (D) was based on Mann-Whitney U test.
FIGURE 2. The dynamics of the host population during the spread of P elements.

These figures show the change of proportion of P cytotype individuals, $r$ (A), the
selection coefficient against non-suppressive host allele, $s$ (B), the allele frequency of
host beneficial allele, the relationship between $s$ and $r$ (C) when $u_0 = 1$ and $n_{HD} = 5$, and $l$
(D) over 1,000 generations.
FIGURE 3. The influences of $u_0$ and $n_{HD}$ on the time for near-fixation of $P$ elements and host beneficial allele frequency. The generations until $P$ element is nearly fixed in the population ($r > 0.99$) for different $u_0$ and $n_{HD}$ are shown in (A). Green dots are when $u_0 = 1$ and blue dots are when $u_0 = 10^{-1}$. The dashed line denotes generation 1,000. The allele frequencies of host beneficial allele ($l_{1000}$) at generation 1,000 for different $u_0$ and $n_{HD}$ are shown in (B). The dashed line denotes 0.001, which is the value of $l_0$. 
FIGURE 4. Polymorphism, divergence and temporal differentiation around Hen1.

Divergence between D. melanogaster and D. simulans (red), polymorphism of D. simulans (orange), polymorphism of post-P element African D. melanogaster population (green) and polymorphism of post-P element North American D. melanogaster population (blue) of 100kb upstream and downstream of Hen1 (from 8,033,231 – 8,040,259) are shown on a log scale in (A). There is dramatic drop of polymorphism in the North American D. melanogaster populations around Hen1. Temporal differentiation between pre-P and post-P element North American populations of control genes are shown in (B), with their relative position shown in (C). Genes in (C) are Hen1 (red), Cyp6g1 (orange), jeb (1), CG8378 (2), CG13178 (3), CG8878 (4), CG8407 (5), Oda (6), wash (7), CG33964 (8), Cyp6t3 (9), RpS11 (10) and Sr-CII (11). Sequenced regions of each control gene are shown in blue while un-sequenced regions are in light blue. The coordinates of three figures are aligned.
FIGURE 5. Divergence between *D. willistoni* and *D. melanogaster* of candidate genes among other genes. Amino acid sequence divergences were estimated for genes with annotated *D. willistoni* ortholog (11 candidate genes and 10,029 other genes) and the genome-wide distribution of the divergence is shown. The divergences of the 11 candidate genes on the x-axis are shown in the lower part of the figure.
### TABLE1. Information of candidate genes

<table>
<thead>
<tr>
<th>Category</th>
<th>FBgn</th>
<th>symbol</th>
<th>gene name</th>
<th>chr</th>
<th>CDs length</th>
<th>gene functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>piRNA genes</td>
<td>FBgn0004872</td>
<td>piwi</td>
<td>Piwi</td>
<td>2L</td>
<td>2532</td>
<td>piRNA generation</td>
</tr>
<tr>
<td></td>
<td>FBgn0000146</td>
<td>aub</td>
<td>Aubergine</td>
<td>2L</td>
<td>2601</td>
<td>piRNA generation</td>
</tr>
<tr>
<td></td>
<td>FBgn0250816</td>
<td>AGO3</td>
<td>Ago3</td>
<td>3L</td>
<td>2604</td>
<td>piRNA generation</td>
</tr>
<tr>
<td></td>
<td>FBgn0041164</td>
<td>armi</td>
<td>Armitage</td>
<td>3L</td>
<td>3825</td>
<td>RNA helicase</td>
</tr>
<tr>
<td></td>
<td>FBgn0003483</td>
<td>Spn-E</td>
<td>Spindle-E</td>
<td>3R</td>
<td>4305</td>
<td>RNA helicase</td>
</tr>
<tr>
<td></td>
<td>FBgn0002652</td>
<td>squ</td>
<td>Squash</td>
<td>2L</td>
<td>726</td>
<td>nuclease</td>
</tr>
<tr>
<td></td>
<td>FBgn0261266</td>
<td>zuc</td>
<td>Zucchini</td>
<td>2L</td>
<td>762</td>
<td>nuclease</td>
</tr>
<tr>
<td></td>
<td>FBgn0033686</td>
<td>Hen1</td>
<td>Pimet/DmHen1</td>
<td>2R</td>
<td>1176</td>
<td>piRNA methyltransferase</td>
</tr>
<tr>
<td></td>
<td>FBgn0016034</td>
<td>mael</td>
<td>Maelstrom</td>
<td>3L</td>
<td>1389</td>
<td>nuage component</td>
</tr>
<tr>
<td></td>
<td>FBgn0034098</td>
<td>krimp</td>
<td>Krimper</td>
<td>2R</td>
<td>2241</td>
<td>nuage component</td>
</tr>
<tr>
<td></td>
<td>FBgn0003970</td>
<td>vas</td>
<td>Vasa</td>
<td>2L</td>
<td>661</td>
<td>nuage component</td>
</tr>
<tr>
<td></td>
<td>FBgn0004400</td>
<td>rhi</td>
<td>Rhino</td>
<td>2R</td>
<td>1257</td>
<td>HP1 paralog, heterochromatin binding</td>
</tr>
<tr>
<td>P element specific genes</td>
<td>FBgn0014870</td>
<td>Psi</td>
<td>PSI</td>
<td>2R</td>
<td>2394</td>
<td>P element somatic inhibitor; mRNA splicing factor</td>
</tr>
<tr>
<td></td>
<td>FBgn0004838</td>
<td>Hrb27C</td>
<td>Hrb27C</td>
<td>2L</td>
<td>1266</td>
<td>mRNA splicing factor</td>
</tr>
<tr>
<td></td>
<td>FBgn0011774</td>
<td>Irbp</td>
<td>Irbp</td>
<td>3R</td>
<td>1897</td>
<td>recombination repair protein</td>
</tr>
<tr>
<td></td>
<td>FBgn0041627</td>
<td>Ku80</td>
<td>Ku80</td>
<td>2L</td>
<td>2100</td>
<td>recombination repair protein</td>
</tr>
</tbody>
</table>
### TABLE 2. Linage specific relative rates of protein evolution

<table>
<thead>
<tr>
<th>gene</th>
<th>mel dN/dS</th>
<th>percentile&lt;sup&gt;a&lt;/sup&gt;</th>
<th>sim dN/dS</th>
<th>percentile&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGO3</td>
<td>0.261</td>
<td>12.67%</td>
<td>0.265</td>
<td>18.79%</td>
</tr>
<tr>
<td>armi</td>
<td>0.297</td>
<td>10.22%</td>
<td>0.391</td>
<td>10.88%</td>
</tr>
<tr>
<td>aub</td>
<td>0.382</td>
<td>6.60%</td>
<td>0.530</td>
<td>6.31%</td>
</tr>
<tr>
<td>Hen1</td>
<td>0.246</td>
<td>13.64%</td>
<td>0.266</td>
<td>18.66%</td>
</tr>
<tr>
<td>Hrb27C</td>
<td>0.176</td>
<td>21.99%</td>
<td>0.019</td>
<td>82.02%</td>
</tr>
<tr>
<td>Irbp</td>
<td>0.101</td>
<td>38.42%</td>
<td>0.164</td>
<td>32.83%</td>
</tr>
<tr>
<td>krimp</td>
<td><strong>0.471</strong></td>
<td><strong>4.71%</strong></td>
<td><strong>0.964</strong></td>
<td><strong>2.09%</strong></td>
</tr>
<tr>
<td>Ku80</td>
<td>0.181</td>
<td>21.28%</td>
<td>0.245</td>
<td>20.80%</td>
</tr>
<tr>
<td>mael</td>
<td><strong>0.902</strong></td>
<td><strong>1.05%</strong></td>
<td><strong>0.491</strong></td>
<td><strong>7.17%</strong></td>
</tr>
<tr>
<td>piwi</td>
<td>0.081</td>
<td>45.87%</td>
<td>0.194</td>
<td>27.71%</td>
</tr>
<tr>
<td>Psi</td>
<td>0.091</td>
<td>42.12%</td>
<td>0.027</td>
<td>78.55%</td>
</tr>
<tr>
<td>rhi</td>
<td><strong>1.415</strong></td>
<td><strong>0.31%</strong></td>
<td><strong>0.508</strong></td>
<td><strong>6.77%</strong></td>
</tr>
<tr>
<td>Spn-E</td>
<td>0.219</td>
<td>16.28%</td>
<td>0.281</td>
<td>17.42%</td>
</tr>
<tr>
<td>squ</td>
<td>0.240</td>
<td>14.25%</td>
<td>0.338</td>
<td>13.72%</td>
</tr>
<tr>
<td>vas</td>
<td>0.243</td>
<td>14.05%</td>
<td>0.306</td>
<td>15.63%</td>
</tr>
<tr>
<td>zuc</td>
<td><strong>0.359</strong></td>
<td><strong>7.48%</strong></td>
<td>0.207</td>
<td>25.73%</td>
</tr>
</tbody>
</table>

<sup>a</sup>dN/dS for each candidate gene was ranked among 9,172 (D. melanogaster) and 9,051 (D. simulans) genes that have PAML results. Candidate genes that ranked among the top 10% among all the genes with PAML results are in bold type.
<table>
<thead>
<tr>
<th>gene</th>
<th>two-species MK test&lt;sup&gt;a&lt;/sup&gt;</th>
<th>mel MK test&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no. codons</td>
<td>p-value</td>
</tr>
<tr>
<td>AGO3</td>
<td>NA</td>
<td>NA&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>armi</td>
<td>1186</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>aub</td>
<td>835</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Hen1</td>
<td>389</td>
<td>0.837</td>
</tr>
<tr>
<td>Hrb27C</td>
<td>421</td>
<td>0.095</td>
</tr>
<tr>
<td>Irbp</td>
<td>628</td>
<td>0.189</td>
</tr>
<tr>
<td>krimp</td>
<td>712</td>
<td>0.020</td>
</tr>
<tr>
<td>Ku80</td>
<td>694</td>
<td>0.012</td>
</tr>
<tr>
<td>mael</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>piwi</td>
<td>837</td>
<td>0.732</td>
</tr>
<tr>
<td>Psi</td>
<td>796</td>
<td>0.189</td>
</tr>
<tr>
<td>rhi</td>
<td>413</td>
<td>0.396</td>
</tr>
<tr>
<td>Spn-E</td>
<td>1427</td>
<td>0.021</td>
</tr>
<tr>
<td>squ</td>
<td>130</td>
<td>0.493</td>
</tr>
<tr>
<td>vas</td>
<td>634</td>
<td>0.022</td>
</tr>
<tr>
<td>zuc</td>
<td>253</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<sup>a</sup>MK tests using both <i>D. melanogaster</i> and <i>D. simulans</i> polymorphism (see methods)

<sup>b</sup>MK tests using only <i>D. melanogaster</i> polymorphism

<sup>c</sup>NA: not available due to lack of <i>D. simulans</i> polymorphism data

<i>MK tests</i> with significant <i>p-values</i> and positive α are in bold type
### TABLE 4. Temporal differentiation of a subset of candidate genes

<table>
<thead>
<tr>
<th></th>
<th>geographic differentiation&lt;sup&gt;a&lt;/sup&gt;</th>
<th>temporal differentiation&lt;sup&gt;b&lt;/sup&gt;</th>
<th>temporal differentiation&lt;sup&gt;c&lt;/sup&gt;</th>
<th>amino acid differentiation&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all samples</td>
<td>only North American samples</td>
<td>nucleotide differentiation</td>
<td>nucleotide differentiation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fst</td>
<td>p-value</td>
</tr>
<tr>
<td>Irbp</td>
<td>0.144</td>
<td>0.111</td>
<td></td>
<td></td>
</tr>
<tr>
<td>krimp</td>
<td>0.143</td>
<td>0.004</td>
<td>0.011</td>
<td>0.262</td>
</tr>
<tr>
<td>Hen1</td>
<td>0.632</td>
<td>&lt; 0.001</td>
<td>0.303</td>
<td>0.001</td>
</tr>
<tr>
<td>Spn-E</td>
<td>0.459</td>
<td>&lt; 0.001</td>
<td>0.053</td>
<td>0.08</td>
</tr>
<tr>
<td>squ</td>
<td>0.028</td>
<td>0.25</td>
<td>0.202</td>
<td>0.001</td>
</tr>
</tbody>
</table>

<sup>a</sup> Genetic differentiation between current (post-<i>P element</i>) African and North American populations  
<sup>b</sup> Genetic differentiation between pre-<i>P</i> and post-<i>P</i> element invasion populations  
<sup>c</sup> The temporal differentiation were estimated only considering North American samples of both populations before and after <i>P element</i> invasions.  
<sup>d</sup> Genetic differentiation estimated using nucleotide sequences  
<sup>e</sup> Genetic differentiation estimated using amino acid sequences  
All the significant results are in bold-type.
### TABLE 5. Temporal and geographic differentiation of site frequencies of control genes

<table>
<thead>
<tr>
<th>Gene</th>
<th>distance to Hen1</th>
<th>distance to Cyp6g1</th>
<th>M strain π</th>
<th>temporal differentiation</th>
<th>geographic differentiation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fst</td>
<td>p-value</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>nonsyn</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>syn</td>
<td></td>
<td></td>
</tr>
<tr>
<td>jeb</td>
<td>-31911</td>
<td>-69718</td>
<td>0.0021</td>
<td>0.089</td>
<td>0.037</td>
</tr>
<tr>
<td>CG8378</td>
<td>-7273</td>
<td>-45080</td>
<td>0.0000</td>
<td>0.239</td>
<td>0.001</td>
</tr>
<tr>
<td>CG13178</td>
<td>-5378</td>
<td>-43185</td>
<td>0.0007</td>
<td>0.271</td>
<td>0.002</td>
</tr>
<tr>
<td>CG8878</td>
<td>0</td>
<td>-37807</td>
<td>0.0001</td>
<td>0.195</td>
<td>0.001</td>
</tr>
<tr>
<td>CG8407</td>
<td>4426</td>
<td>-33381</td>
<td>0.0000</td>
<td>0.230</td>
<td>0.005</td>
</tr>
<tr>
<td>Oda</td>
<td>21177</td>
<td>-16630</td>
<td>0.0012</td>
<td>0.369</td>
<td>0.001</td>
</tr>
<tr>
<td>wash</td>
<td>32266</td>
<td>-5541</td>
<td>0.0010</td>
<td>0.393</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>CG33964</td>
<td>34832</td>
<td>-2975</td>
<td>0.0009</td>
<td>0.377</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Cyp6t3</td>
<td>42227</td>
<td>4420</td>
<td>0.0021</td>
<td>0.549</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>RpS11</td>
<td>50683</td>
<td>12876</td>
<td>0.0000</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Sr-CII</td>
<td>60799</td>
<td>22992</td>
<td>0.0025</td>
<td>0.238</td>
<td>0.001</td>
</tr>
</tbody>
</table>

*a* comparisons between pre- and post-P element invasion populations with samples from North American populations only.

*b* comparisons between North American and African post-P element invasion populations

*c* the distance in base pairs between the mid-points of the focused gene and Hen1. CG8878 has zero distance because it is nested within Hen1.

*d* the distance in base pairs between the mid-points of the focused gene and Cyp6g1

*e* only the coding region of first exon of jeb was sequenced.

NA: not available due to no SNPs differences between pre-P and post-P element North American populations

The differentiation were estimated using nucleotide sequence.