Evolution of *P*-element Repression in *Drosophila melanogaster* Through the Piwi-interacting RNA Pathway

---

A Dissertation Presented to

the Faculty of the Department of Biology and Biochemistry

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Shuo Zhang

August 2019

Evolution of *P*-element Repression in *Drosophila melanogaster* Through the Piwi-interacting RNA Pathway

Shuo Zhang

APPROVED:

Dr. Erin S. Kelleher, Chairman
Department of Biology and Biochemistry

Dr. Ricardo B. R. Azevedo
Department of Biology and Biochemistry

Dr. Alexander J. Stewart
Department of Biology and Biochemistry

Dr. Claudio Casola
Department of Ecosystem Science & Management
Texas A&M University

Dr. Dan E. Wells, Dean
College of Natural Sciences and Mathematics

Acknowledgements

First of all, I would like to express my deepest gratitude to my advisor Dr. Erin S. Kelleher. She inspires me when I got stuck, corrects me when I made mistakes, encourages me when I was down, and forgives me when I made no progress. Dr. Kelleher invested a lot in mentoring me how to address scientific questions, present findings to scientific communities, and write compelling papers. Studying abroad is challenging, especially in a place where people do not speak my native language. But her kindness to my family and me makes me feel I was at home. She offers suggestions on renting an apartment in Houston, points out some tips on how to take care of a newborn, and lends toys to my daughter. I feel extremely fortunate to have such a nice advisor and friend, who will continuously impact my academic career and life.

I also appreciate my present and former committee members: Dr. Ricardo B. R. Azevedo, Dr. Alexander J. Stewart, Dr. Claudio Casola, and Dr. Elizabeth Ostrowski. They provided constructive suggestions and helpful comments on my projects, which drive me to think more broadly and deeply. Without their kind support, this dissertation would have been impossible.

I would also like to thank Dr. Richard P. Meisel. I did my rotation under his supervision. I learnt wet lab and dry lab skills from him. In addition, he gave a lot of useful suggestions on my projects and presentation skills.

I have to thank my labmates: Luyang Wang, Jyoti Lama, Satyam Srivastav, Sadia Tasnim, Lily Ortega, Uche Akoma, Donald Hubbard. In addition, I need to thank members of Meisel lab: Dr. Jae Hak Son, Kiran Adhikari, Dr. Pablo Delclos. We had so many wonderful lab meetings together. Also, I benefited a lot from suggestions and feedback they provide. My special thanks go to high school teachers Beverly Pointer, Jamika Lasker, and Katherine W. Hartman, who helped me collect some phenotypic data.

I would also like to thank my friends in the Department of Biology and Biochemistry, including Haopeng Yang, Chenchu Lin, Fei Yuan, Hao Zhang, Jason Tarkington, Vrutant Shah, Afzal Ahrorov, etc. We either had discussion about sciences or shared experiences of living in Houston, which makes my life at the University of Houston (UH) memorable. In particular, Haopeng gave me a ride to supermarkets for my first year at UH and Vrutant took me in for several days during Hurricane Harvey.

I also thank staff in the Department of Biology and Biochemistry. In particular, Rosezelia E. Jackson helped me a lot, ranging from my application for UH graduate program to my application for graduation.

Lastly, but certainly not least, I would like to thank my parents and parent-in-law for their support. Especially, I thank my wife, Jiaqi, for taking care of our daughter Alice and keeping the family running. They and their countless support motivated me to move forward.

Evolution of *P*-element Repression in *Drosophila melanogaster* Through the Piwi-interacting RNA Pathway

---

An Abstract of a Dissertation

Presented to

the Faculty of the Department of Biology and Biochemistry

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Shuo Zhang

August 2019

Abstract

Transposable elements (TEs) are ubiquitous and selfish genetic entities whose mobilization poses a significant threat to their host. In the germline of metazoan, the Piwi-interacting RNAs (piRNAs) derived from TE-enriched loci (called piRNA clusters) regulate TE activity in a sequence specific manner. However, the emergence and dynamics of piRNA-mediated repressor alleles to an invading TE remain elusive. *P*-element, a DNA transposon that recently invaded the *D. melanogaster* genome around 1950, provides a unique opportunity to study the evolution of host repression. In this dissertation, I first adapted a targeted sequencing strategy and developed a computational pipeline to annotate *P*-element insertions in a sequenced *Drosophila melanogaster* genome. My approach precisely determined *P*-element insertion breakpoints and found new *P*-element insertions, which were undetected by previously methods. Next, I modified the pipeline to annotate *P*-element insertions in the *Drosophila melanogaster* genetic reference panel (DGRP), a panel composed of 205 fully sequenced inbred lines. I found over 90% of DGRP genomes have *P*-elements in ancestral piRNA clusters that are active prior to the *P*-element invasion. This indicates *de novo* mutation, in which *P*-elements transpose into pre-existing piRNA clusters, is the predominant mechanism for the origin of repressor alleles. Moreover, I detected no fewer than 84 independent *P*-element insertions in ancestral piRNA clusters. Finally, I observed that *P*-element insertions in piRNA clusters segregate at significant higher frequency

than *P*-elements outside of piRNA clusters, suggesting that cluster *P*-elements confer a selective advantage. Taken together, my results revealed a striking example of polygenic adaptation, in which a plethora of *de novo* beneficial *P*-element insertions into multiple piRNA clusters, fueled the evolution of a ubiquitous repressive phenotype in <60 years.

Table of Contents

Abbreviations x

Chapter 1 Introduction 1
1.1 What are transposable elements? 2
1.2 The impact of transposable elements on host genomes 3
1.3 The Piwi-interacting RNA pathway 5
1.4 Horizontal transfer of transposable elements 8
1.5 The origin of repressors: *de novo* mutation versus epigenetic mutation 9
1.6 Evolutionary dynamics of repressor alleles 12
1.7 The *P*-element invasion into the *D. melanogaster* genome 13
1.8 Identification of TEs from next-generation sequencing data 15
1.9 Dissertation outline 18

Chapter 2 Targeted identification of TE insertions in a *Drosophila* genome through hemi-specific PCR 20
2.1 Introduction 21
2.2 Results 24
2.2.1 Hemi-specific PCR amplifies abundant *P*-element insertions 24
2.2.2 Validation of novel insertions and identification of false positives 29
2.2.3 Sequence similarity to true insertion sites may produce false positives 30
2.2.4 The majority of sequencing reads are explained by annotated insertions 32
2.2.5 Improved insertion site identification and frequency estimation 33
2.3 Discussion 37
2.4 Conclusion 39
2.5 Materials and Methods 40
2.5.1 Genomic DNA samples 40
2.5.2 Primer design 40
2.5.3 Library construction by hemi-specific PCR 41
2.5.4 Identification of *P*-element-derived read-pairs and alignment to the reference genome 43
2.5.5 Annotation of *P*-element insertions based on uniquely mapping read pairs 44
2.5.6 Determining the number of *P*-element reads that arise from annotated insertions 45
2.5.7 Estimating the frequency of individual insertions from whole genome sequencing paired-end data 45
2.5.8 Site-specific PCR 46

Chapter 3 piRNA-mediated silencing of an invading transposable element evolves rapidly through abundant beneficial *de novo* mutations 48
3.1 Introduction 49
3.2 Results 52
3.2.1 Identification of ancestral piRNA clusters 52
3.2.2 Most North American genotypes have *P*-elements in ancestral piRNA clusters 54
3.2.3 Abundant repressors underpin repressive phenotype 59
3.2.4 Cluster *P*-element insertions are targets of positive selection 63
3.3 Discussion 65

3.4 Materials and Methods 70
3.4.1 DGRP stocks and genomes 70
3.4.2 piRNA cluster annotation 70
3.4.3 Detecting *P*-element insertions in DGRP genomes 70
3.4.4 Detecting *P*-element insertions in TAS 72
3.4.5 Localizing insertion sites of *P*-element insertions in TAS 73
3.4.6 PCR verification of insertion sites 74
3.4.7 Recombination rates 74
3.4.8 Data analysis 75

Chapter 4 Overall conclusions and discussion 76
4.1 Overall conclusions 77
4.2 Discussion 78
4.2.1 The application of hemi-specific PCR 78
4.2.2 Soft sweeps in evolution of host resistance to invading TEs 79
4.2.3 The long-term evolution of *P*-element insertions in piRNA clusters 80
4.2.4 The role of epigenetic mutation in the evolution of *P*-element repression 81
4.2.5 The evolution of *P*-element repression among *D. melanogaster* populations in other geographic regions 82
4.2.6 The invasion of *P*-elements into *Drosophila simulans* 83
4.2.7 The size of ancestral piRNA clusters 84

Appendix 85

References 140

Abbreviations

bp	base pairs
crRNA	CRISPR RNA
DGRP	*Drosophila melanogaster* genetic reference panel
DNA	deoxyribonucleic acid
H3K9me3	histone 3 lysine 9 trimethylation
HT	horizontal transfer
kb	kilobase pairs
LINE	long interspersed nuclear element
Mb	megabase pairs
nt	nucleotide
piRNA	Piwi-interacting RNA
PCR	polymerase chain reaction
RNA	ribonucleic acid
siRNA	small interfering RNA
TE	transposable element
WGS	whole genome re-sequencing
TAS	telomeric associated sequences

Chapter 1 Introduction

1.1 What are transposable elements?

Transposable elements (TEs), also known as "jumping genes", are selfish genetic entities that move and replicate within host genomes. TEs were first discovered around 1950 by Barbara McClintock in her now classic study of variegated maize kernel, which is induced by transposition of TEs (McClintock 1950; McClintock 1953). Later, TEs were found to be present in virtually all organisms and can make up a substantial fraction of host genomes (Figure 1.1; reviewed in Chénais et al. 2012). For example, TEs account for ~45% of human genome (Lander et al. 2001) and ~85% of the maize genome (Schnable et al. 2009).

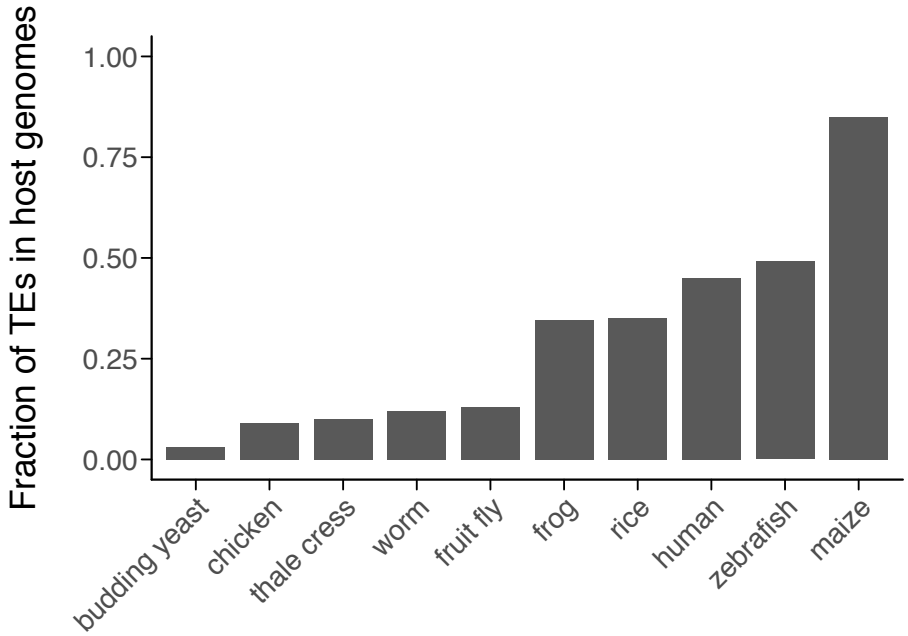

Figure 1.1 Fraction of TEs in some model organisms. TE: transposable element. Data sources: budding yeast, *Saccharomyces cerevisiae* (Kim et al. 1998); chicken, *Gallus gallus* (International Chicken Genome Sequencing Consortium, 2004); thale cress, *Arabidopsis thaliana* (Kaul et al. 2000); worm, *Caenorhabditis elegans* (Bessereau 2006); fruit fly, *Drosophila melanogaster* (Smith et al. 2007); frog, *Xenopus tropicalis* (Hellsten et al. 2010)*; rice, Oryza sativa* L. (International Rice Genome Sequencing Project. 2005); human, *Homo sapiens* (Lander et al. 2001); zebrafish, *Danio rerio* (Howe et al. 2013)*;* maize, *Zea mays*

(Schnable et al. 2009).

Based on the mode of transposition, TEs are categorized into two major classes: Class I (retrotransposons) and Class II (DNA transposons). Retrotransposons move through a "copy and paste" mechanism, in which a retrotransposon is first transcribed into a RNA intermediate and the RNA is reverse transcribed before inserting into a novel location. In contrast, DNA transposons move through a "cut and paste" mechanism, in which they are directly excised and integrated into a new location (Wicker et al. 2007). TEs are also classified into autonomous and non-autonomous elements. Autonomous TEs transpose using their own transposition machinery (transposase or reverse transcriptase), whereas transposition of non-autonomous TEs requires the presence of autonomous TE copies.

1.2 The impact of transposable elements on host genomes

Due to self-replication and mobility, TEs profoundly shape host genome architecture. First, TE proliferation contributes to genome expansion. For example, the LINE1 retrotransposons alone account for ~17% of the human genome (Lander et al. 2001) and proliferation of LTR retrotransposons contribute to the rapid expansion of plethodontid salamander genomes (Sun et al. 2012). Second, dispersed TE insertions can influence host gene expression by disrupting *cis*-regulatory elements, providing alternative transcriptional start sites, triggering the formation

of heterochromatin, and affecting splicing patterns (reviewed in Feschotte 2008). For instance, a *hAT* DNA transposon insertion results in hypermethylation of a sex determination locus and leads to the development of female flowers in melon (Martin et al. 2009).

TEs are occasionally domesticated to perform essential functions in the host genome. For example, a *Doc* element insertion into *CHKov1*, a putative gene involved in choline metabolism, results in a truncated allele which confers resistant to both organophosphate pesticides and the sigma virus in *D. melanogaster* (Aminetzach et al. 2005; Magwire et al. 2011). Moreover, *HeT-A* and *TART* retrotransposons preferably insert into telomeres of *Drosophila* to maintain telomere length (Pardue and DeBaryshe 2003).

However, TEs are generally considered genetic parasites as they impose a sizeable mutational burden on their hosts. First, TEs produce deleterious insertions that disrupt or interfere with functional sequences (Levis et al. 1984; McGinnis et al. 1983; Lee 2015). In addition, TEs cause double-strand DNA breaks through encoded endonucleases during transposition (Gasior et al. 2006), which would introduce mutations if DNA damage is not repaired correctly. Moreover, homologous TEs may undergo ectopic recombination, leading to chromosomal rearrangements, such as deletion, duplication or translocation (Lim 1988; Hedges and Deininger 2007). Finally, large portion of TEs burdens the host genome in terms of replication, transcription, and translation of TE insertions (Nuzhdin 1999). As a result, TEs can cause deleterious phenotypes, including human genetic diseases (reveiwed in

Chénais 2013; Jang et al. 2019; Burns 2017). For instance, a single insertion of LINE-1 element (one type of retrotransposon) into a clotting protein-coding gene can cause hemophilia A, an *X*-linked disorder in human (Kazazian et al. 1988).

1.3 The Piwi-interacting RNA pathway

Owing to mutagenic effects of TEs, their activity is strictly regulated, especially in the germline where TEs are exceptionally active and TE-induced mutations are transmitted to the next generation. Small RNA-mediated silencing is a common mechanism employed by both prokaryotes and eukaryotes to defend exogenous nucleic acids, including TEs (Malone and Hannon 2009; Girard and Hannon 2008; Kumar and Kevin 2012). In this silencing pathway, CRISPR RNAs (crRNAs) in prokaryotes, small interfering RNAs (siRNAs) in plants or Piwi-interacting RNAs (piRNAs) in animals guide proteins with nuclease activity to silence homologous DNA sequences transcriptionally or post-transcriptionally (Brennecke et al. 2007; Aravin et al. 2007b; Girard and Hannon 2008; Slotkin et al. 2009; Bhaya et al. 2011). This dissertation will focus on the emergence and dynamics of piRNA-based repressor alleles to invading TEs in animals.

In the germline of many metazoans, TEs are silenced by the small Piwi-interacting RNAs (piRNAs) pathway (Aravin et al. 2007b, 2006; Brennecke et al. 2007; Siomi et al. 2011; Houwing et al. 2007; Batista et al. 2008; Bagijn et al. 2012) piRNAs are 21-33 nucleotide (nt) single-stranded small RNAs. As the name implies, piRNAs are associated with Piwi proteins, a clade of the Argonaute protein family.

Argonaute proteins also play a central role siRNAs and miRNAs-mediated silencing (reviewed in Höck and Meister 2008). piRNA species are highly diverse, ranging from ~3000 in *Caenorhabditis elegans* to ~1, 000,000 in *D. melanogaster* (reviewed in Ozata et al. 2019), which enable hosts to silence diverse TE families.

piRNAs are derived from discrete specialized genomic loci known as piRNA clusters, which may make up ~3.5% of the host genome (Brennecke et al. 2007). piRNA clusters are very large complex loci, up to 240 kb in *Drosophila*, which contain many inactive TE copies from a variety of TE families (Brennecke et al. 2007). Many piRNA clusters are bi-directionally transcribed, producing sense and antisense piRNAs (Brennecke et al. 2007; Aravin et al. 2007b; Mohn et al. 2014). In the germline, piRNAs guide Piwi protein complexes to degrade complementary mRNA of TEs, or transcriptionally silence active TE insertions by establishing repressive chromatin marks, such as DNA methylation in mammals or H3K9me3 in *Drosophila* (Figure 1.2; Brennecke et al. 2007; Aravin et al. 2007, 2008; Le Thomas et al. 2013; Sienski et al. 2012; Huang et al. 2013).

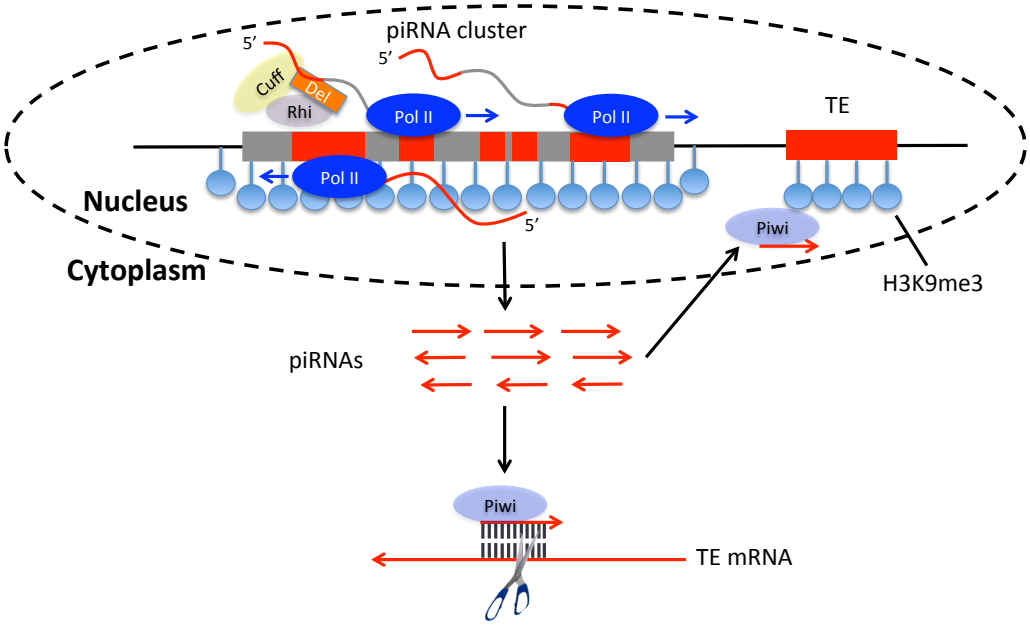

Figure 1.2 A simplified piRNA pathway in *Drosophila*. piRNA clusters (gray) contain multiple inactive TEs (red). piRNA clusters are enriched with H3K9me3, a repressive heterochromatin mark. Initialized by the Rhino-Cutoff-Deadlock complex, piRNA clusters can be bi-directionally transcribed (blue arrows) by Pol II, producing piRNA precursors. Precursors are processed into mature 23 – 29 nt piRNAs that can be sense (mapped to sense strand of TE mRNAs; red arrows pointing to the left) or antisense (mapped to antisense strand of TE mRNAs, red arrows pointing to the right). Mature piRNAs guide Piwi proteins to cleave complementary TE mRNAs in cytoplasm or establish repressive H3K9me3 marks around TE insertions in nucleus. Other proteins involved in piRNA biogenesis are not shown in this figure. TE: transposable element, piRNA: Piwi-interacting RNA, Pol II: RNA Polymerase II, H3K9me3: histone 3 lysine 9 trimethylation, mRNA: message RNA.

1.4 Horizontal transfer of transposable elements

Although TEs are transmitted vertically from parents to offspring, comparative genomics has revealed numerous examples of horizontal transfer (HT) of TEs between non-mating species (Dotto et al. 2015). Due to their inherent mobility, TEs are more likely to be horizontally transferred. In fact, HT is proposed to be the mechanism how TEs avoid extinction due to selection, genetic drift and/or accumulation of defective mutations (reviewed in Schaack et al. 2010). Although the mechanisms by which TEs move between non-mating species remain elusive, vectors such as viruses and parasites, may facilitate HT (Houck et al. 1991; Gilbert et al. 2010, 2016).

HT of TEs is widespread across tree of life (Baidouri et al. 2014; Thomas et al. 2010; Dotto et al. 2015; Peccoud et al. 2017). Among *Drosophila* genomes, it is estimated that nearly one-third of the TE families have experienced HT (Bartolomé et al. 2009). Among 40 sequenced plants belonging to different families, ~65% species contain at least one retrotransposon that was horizontally transferred to or from another genome (Baidouri et al. 2014). Moreover, HT of TEs occurs between more distantly related taxa, such as from animals to plants, even between prokaryotes and eukaryotes (Lin et al. 2016).

Widespread HT of TEs threatens the integrity of host genome. Without host repression, invading TEs can proliferate exponentially (Kofler et al. 2018) and the bursts of the TE invaders decrease the host fitness dramatically. For example, activity of *P*-elements is responsible for hybrid dysgenesis, a syndrome

characterized by sterility, elevated mutation rates, and high frequencies of chromosomal aberrations (Figure 1.3; Kidwell et al. 1977; Bingham et al. 1982).

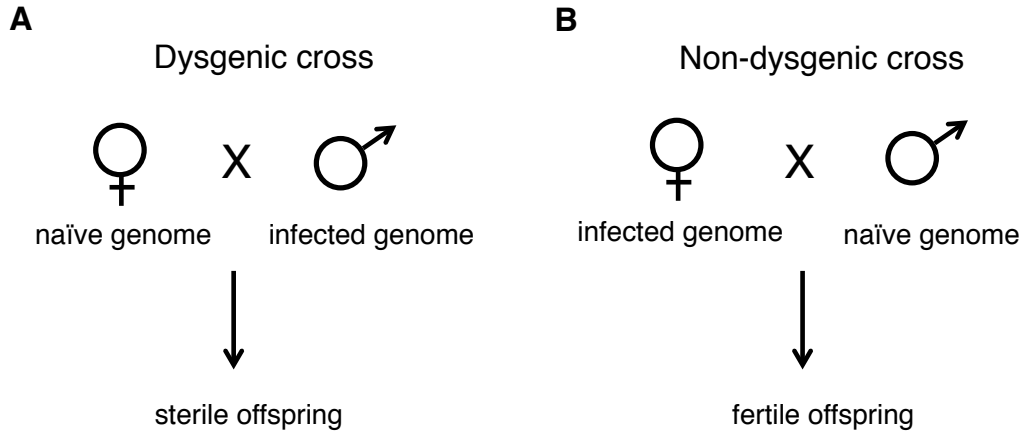

Figure 1.3 A schematic drawing of hybrid dysgenesis. (A) Hybrid dysgenesis occurs in the crosses between males infected with *P*-elements and naïve females devoid of *P*-elements. Because the offspring have no complementary piRNAs to silence the *P*-elements, *P*-elements are activated, resulting in gonadal atrophy. (B) Offspring from the reciprocal cross are fertile because *P*-element piRNAs are maternally transmitted from infected mother to the offspring (Brennecke et al. 2008). TE: transposable element.

1.5 The origin of repressors: *de novo* mutation versus epigenetic mutation

The widespread HT of TEs in animals challenges the host to silence those TE invaders by producing novel species of piRNAs. Novel species of piRNAs or piRNA-mediated repressors (*i.e.* TEs in piRNA clusters) can arise through two mechanisms. The prevailing proposed mechanism is *de novo* mutation (or "trap" model), in which novel TEs randomly transpose into pre-existing piRNA clusters producing repressor alleles (Bergman et al. 2006; Khurana et al. 2011; Girard and Hannon 2008). This

model is supported by the observations that DNA sequences inserted into an active piRNA cluster generate corresponding piRNAs (Khurana et al. 2011; Muerdter et al. 2012; Yamamoto et al. 2013; Han et al. 2015; Duc et al. 2019), and that TE insertion polymorphisms in piRNA clusters are associated with differences in piRNA production and silencing (Stuart et al. 2002; Brennecke et al. 2008; Zanni et al. 2013).

Alternatively, repressors may arise through epigenetic mutation, in which a non-piRNA producing locus is converted into a novel piRNA cluster by installment of H3K9me3 (Figure 1.4; De Vanssay et al. 2012). This process is mediated by piRNAs produced from homologous piRNA-generating loci called inducers (Le Thomas et al. 2014; Hermant et al. 2015). Although any TE insertions could form novel piRNA clusters (Shpiz et al. 2014), epimutated loci differ from those novel piRNA clusters in their ability to produce piRNAs over numerous generations in the absence of original inducers. Moreover, epimutated piRNA clusters are capable of converting other homologous regions inert for piRNA production into piRNA-producing loci (De Vanssay et al. 2012). Epigenetic mutation may play a vital role in the establishment of repression to invading TEs, by quickly creating repressor alleles and reinforcing repression capacity (Hermant et al. 2015; Kelleher 2016).

There are several challenges to elucidating the role of *de novo* and epigenetic mutation in evolution of repression of invading TEs. Because most piRNA clusters (>95% in *D. melanogaster*) are located in pericentromeric and telomeric heterochromatin with highly repetitive sequences (Brennecke et al. 2007), it is

challenging to detect polymorphic TE insertions in repeat-rich piRNA clusters due to an absence of unique surrounding sequence. Additionally, it is impossible to differentiate the *de novo* and epigenetic mutation models for most piRNA producing TE insertions invasions occurred in the distant evolutionary past, meaning it is unknown which arose first at a given locus, TE insertion or piRNA production.

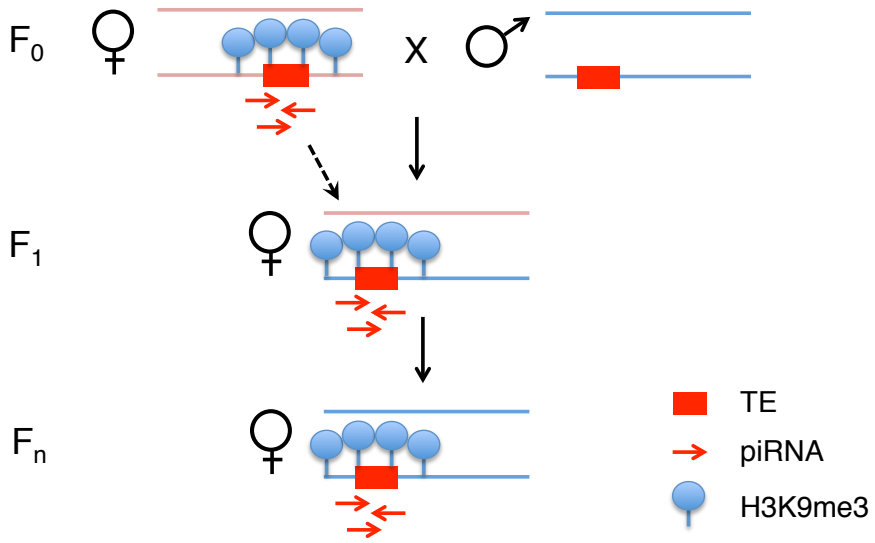

Figure 1.4 The origin of repressors via epigenetic mutation. piRNAs are produced from a piRNA cluster with repressive chromatin marks in the mother. These piRNAs are maternally transmitted to the offspring and interact with (black dashed arrow) a homologous TE, which is inherited from the father and is inert for piRNA production. The maternally inherited piRNAs mediate converting the non-piRNA producing TE insertion into a novel piRNA-producing locus by modifying chromatin marks. The epimutated piRNA cluster is stable over multiple successive generations (n > 100) if inherited maternally. In F_1 generation, chromosomes inherited from mother are indicated in pink and chromosomes from father are indicated in light blue. TE: transposable element, piRNA: Piwi-interacting RNA, H3K9me3: histone 3 lysine 9 trimethylation. The figure is adapted from Kelleher (2016).

1.6 Evolutionary dynamics of repressor alleles

Regardless of the mutational mechanism that produces small RNA-mediated repressor alleles, they increase host fitness by safeguarding host genome from deleterious TE-induced mutations. However, whether the fitness benefit associated with repressor alleles is strong enough for them to escape genetic drift is controversial.

On the one hand, forward simulation models suggest that piRNA-mediated repressor alleles are selectively more advantageous than non-repressor alleles (*i.e.* TEs outside of piRNA clusters), especially when TEs are highly deleterious (Lu and Clark 2010; Kelleher et al. 2018; Kofler 2019). Adaptive evolution of repressor alleles is further supported by population genomic data from *D. melanogaster* (Lu and Clark 2010).

On the other hand, Charlesworth and Langley found the selective advantage of a repressor allele is limited in sexually reproducing organisms with free recombination (Charlesworth and Langley 1986). This is because the advantage of a repressor is to reduce TE copy number in linked DNA, and recombination will weaken this advantage by reducing the extent of linkage disequilibrium (Charlesworth and Langley 1989). Nevertheless, they considered deleterious TE insertions as the only TE-induced fitness cost, ignoring the selective pressure imposed by hybrid dysgenesis, which greatly deceases host fitness by causing DNA damage during transposition. In addition, a simulation of dynamics of repressor alleles that reduce TE transposition rate indicates that selection pressure for host

repressors is strong only at early stage of a TE invasion so that the frequency increase of beneficial alleles is hard to detect (Lee and Langley 2012). Lee and Langley assumed repressors arise at the same time as a new TE invades a naïve genome. This assumption restrains the accumulation of the invading TE at the early invasion, reducing fitness cost associated with the TE invader and therefore decreasing selective pressure on repressor alleles. Moreover, a simulation by Kofler found that host repressor alleles go to fixation even when TE insertions were assumed to be neutral (Kofler 2019), making the role of positive selection on evolution of repressor alleles elusive.

1.7 The *P*-element invasion into the *D. melanogaster* genome

P-elements provide a unique opportunity to study not only the contributions of *de novo* mutation to the origin of repressors, but also evolutionary dynamics of repressors when selection is most strong. *P*-elements are DNA transposons that invaded the *D. melanogaster* genome around 1950 by horizontal transfer from *D. willistoni* (Daniels et al. 1990; Kidwell 1983; Anxolabéhère et al. 1988). Many natural populations of *D. melanogaster* evolved repression in less than 50 years (Figure 1.5A; Anxolabéhère et al. 1988; Kidwell 1983).

Multiple lines of evidence indicate that evolved *P*-element repression in natural populations reflects piRNA-mediated silencing. First, *P*-elements inserted into piRNA clusters, such as Telomeric Associated Sequences on the left arm of the X chromosome (*X*-TAS), are able to regulate *P*-element activity (Marin et al. 2000;

Stuart et al. 2002; Ronsseray et al. 1996). Additionally, repression in many natural populations exhibits a maternal effect (Figure 1.3; Stuart et al. 2002; Simmons et al. 2004), which is consistent with the maternal transmission of piRNAs. Moreover, mutations in genes involved in the piRNA pathway, such as *aubergine*, impair *P*-element repression (Reiss et al. 2004; Simmons et al. 2007). Finally, multiple strains that are able to repress *P*-elements have been demonstrated to produce abundant *P*-element piRNAs, whereas flies that are unable to suppress *P*-elements are devoid of *P*-element piRNAs (Figure 1.5B; Brennecke et al. 2008; Kofler et al. 2018; Khurana et al. 2011; Wakisaka et al. 2017).

Specifically, using *P*-elements has the following advantages. First, due to its recent invasion, numerous strains collected prior to *P*-element invasion are retained in laboratories and stock centers. This provides a historical record of ancestral piRNA clusters that were active before the *P*-element invasion, enabling us to examine the contribution of *de novo* mutation to the evolution of *P*-element repression. Furthermore, the recently sequenced *Drosophila melanogaster* genetic reference panel (DGRP), which contains 205 wild-derived inbred lines collected around 2003 (Mackay et al. 2012), makes it ideal to study *P*-element dynamics at the population level. Finally, *P*-element activity induces hybrid dysgenesis (Figure 1.3), which makes it is easy to quantify *P*-element repression ability in the lab conditions.

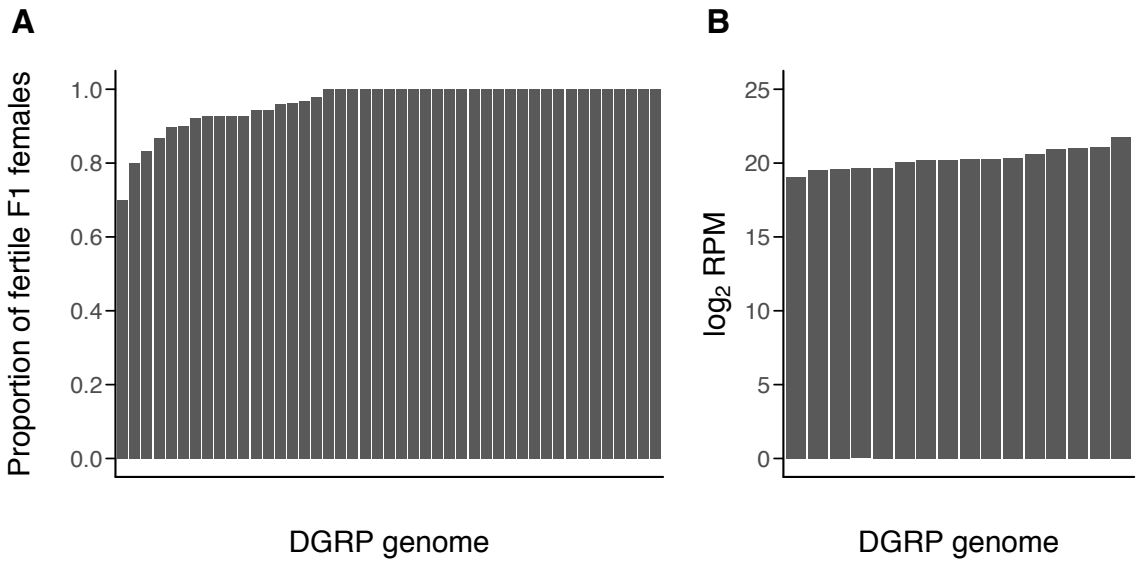

Figure 1.5 Natural populations evolved piRNA-mediated *P*-element repression. (A) DGRP genomes possess strong repression ability to *P*-elements. The *P*-element repression ability of a strain could be measured by crossing females of the test strain to a strain with the ability to induce hybrid dysgenesis (P strain), and examining the fertility in the progeny. In this experiment, the Harwich (a typical P strain) was used to test *P*-element repression in DGRP genomes. Female offspring were raised at 29 °C for 3 days and then were squashed under a slide to check if they could produce eggs, an indicator of ovarian development. The repression ability for each line was calculated as the percentage of F1 females that were able to lay eggs. Data were collected by Beverly Pointer. (B) The expression of *P*-element piRNAs from a subset of DGRP genomes. The expression level was measured in number of *P*-element piRNA reads per million mapped microRNA reads (RPM) and transformed to $\log_2$ scale ($\log_2$ RPM). Deep sequencing data came from Song et al. (2014). DGRP: *Drosophila melanogaster* genetics reference panel, piRNA: Piwi-interacting RNA.

1.8 Identification of TEs from next-generation sequencing data

To fully understand the emergence of piRNA-mediated repressor alleles and their evolutionary dynamics, the first step is to detect individual TE insertions in host genomes. However, TE insertions are highly polymorphic (*i.e.,* only present in certain genotypes within a single species) and tend to be rare (*i.e.,* segregate at low frequencies), making it necessary to annotate TEs in each genotype or population to

be studied. Illumina whole genome re-sequencing (WGS), one of next-generation sequencing (NGS) techniques, provides an unprecedented opportunity to annotate presence and absence of individual TEs in a given genome or pooled genomes. Multiple algorithms have been developed to annotate TEs from WGS data (Zhuang et al. 2014; Rahman et al. 2015; Adrion et al. 2017; Fiston-Lavier et al. 2011; Robb et al. 2013; Nakagome et al. 2014), most of which take advantage of discordant reads and split reads to infer the presence or absence of specific TE insertions (Figure 1.6; reviewed in Ewing 2015).

However, identifying TE insertions from WGS has some limitations. First, WGS is expensive and it is unnecessary to perform whole genome sequencing if a few TEs are investigated. In addition, reads supporting presence or absence of individual TE insertions are not always sampled from WGS or are not sampled with sufficient coverage to reliably infer an insertion or deletion event. This may be particularly true for detecting somatic TE mutations or TE insertions in pooled samples, where individual TE insertions tend to be rare. Moreover, precise insertion breakpoints could be not determined if there is a lack of split reads. Finally, due to the short length of Illumina reads (1 x 50 bp – 2 x 300 bp), detecting TE insertions in repetitive regions is challenging because can be uniquely assigned to a single location in the reference genome (reviewed in Barrón et al. 2014). This poses a serious challenge to the detection of TE insertions in piRNA clusters, as the majority of piRNA clusters are located in repetitive regions (Brennecke et al. 2007).

An alternative method to curate TE insertions is targeted genomic re-sequencing (TGS). The central feature of TGS is the enrichment of DNA fragment with TEs through biochemical marked probes and/or polymerase chain reaction (PCR) using TE-specific primers (Platt et al. 2015; Witherspoon et al. 2010; Ewing and Kazazian 2010). After sequencing the TE-enriched fragments, the insertion sites are determined by mapping reads to the reference genome (the same as WGS, Figure 1.6). One advantage of TGS is that the sequencing coverage is exceptionally high around the TE insertion sites, decreasing the likelihood that rare insertions go undetected. Moreover, precise insertion junctions can be pinpointed, as one read in a read pair always spans TE insertion junctions. However, compared to WGS, TGS is only suitable when a few TEs are investigated and TGS has not yet fundamentally solved the problem of detecting TEs in repetitive regions due to short read length.

The long-read sequencing techniques, including PacBio (https://www.pacb.com) and Oxford Nanopore (https://nanoporetech.com), provide a feasible way to annotate TE insertions in repetitive regions. Compared to short-read sequencing strategy, PacBio and Nanopore can produce longer reads (N50 > 10 kb, *i.e.*, 50% of total sequencing coverage is derived from reads greater than 10 kb), increasing unique alignment rate (Jain et al. 2018; Chakraborty et al. 2018). However, most current PacBio and Nanoproe platforms have high sequencing error rates (1% - 20%; Weirather et al. 2017). Moreover, PacBio sequencing is currently expensive. The advancement of long-read sequencing

techniques, cost-efficient and easy-to-use platforms of Nanopore in particular, promise greatly enhanced power to annotated TE insertions in repetitive regions.

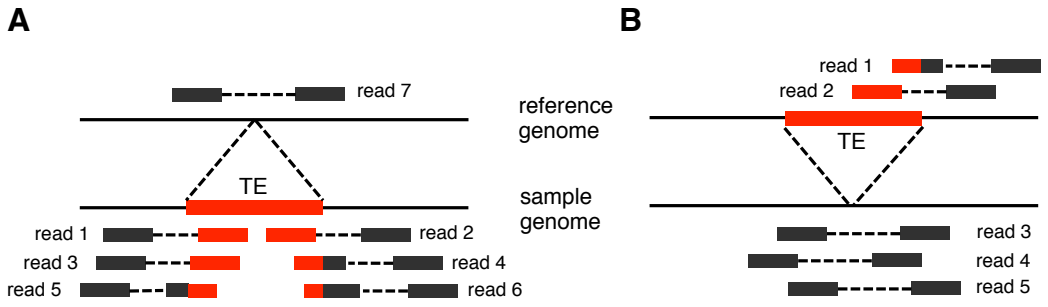

Figure 1.6 (A) Identification of TE insertions from paired-end WGS reads. Discordant reads are read pairs that one end of the pair is mapped to a TE and its mate is mapped to the reference genome (*e.g.,* read 1, 2, 3). Splits reads are read pairs where one segment of one read is mapped to a TE and the remaining is mapped to the reference genome (*e.g.,* read 4, 5, 6). Both discordant reads and split reads indicate the presence of novel TE insertions, but only split reads reveal precise insertion breakpoints. If the TE insertion is not fixed in the sample genome, reads spanning the junction (Read 7) may be present and the frequency of the insertion could be estimated as (discordant reads + split reads)/(discordant reads + split reads + reads spanning the junction). (B) Identification absence of TEs from paired-end WGS reads. In this case, discordant reads are reads spanning the insertion breakpoint (*e.g.,* read 3, 4, 5). If not all chromosomes lose the insertion, there would be reads aligned to the TE (*e.g.,* read 1, 2) and the frequency of the insertion could be estimated as reads aligned to the TE/(reads aligned to the TE + discordant reads). Reads aligned to the reference genome are black and reads aligned to TEs are red. TE: transposable element, WGS: whole genome re-sequencing.

1.9 Dissertation outline

This dissertation uses the invasion of the *D. melanogaster* genome by *P*-elements to study (1) the role of *de novo* mutation to the origin of piRNA-mediated repressor alleles (*P*-elements in piRNA clusters), (2) the role of positive selection on the population dynamics of repressor alleles.

In Chapter 2, I adapted a targeted sequencing strategy to PCR amplify *P*-elements and their flanking genomic DNA sequences in a DGRP genome using *P*-element specific primers and a set of degenerate primers. The PCR products were sequenced on the Illumina Miseq platform. To identify the *P*-element insertion sites, I developed a computational pipeline. I found that targeted sequencing is powerful for annotating the precise insertion sites of *P*-elements, even for *P*-elements in heterochromatic regions. Compared to whole genome re-sequencing (WGS), targeted sequencing provides a cost-efficient approach to identify TE insertions.

In Chapter 3, I modified the pipeline developed in Chapter 2 to annotate *P*-element insertions in all DGRP genomes from whole genome re-sequencing reads. I also annotated ancestral piRNA clusters that were active before *P*-element invasion from a large number of *P*-element-free *D. melanogaster* strains collected prior to *P*-element invasion. I found that over 90% of DGRP genomes have at least one *P*-element in ancestral piRNA clusters, suggesting *de novo* mutation is the predominant mechanism for the origin of piRNA-mediated repressors. Moreover, I found at least 84 different *de novo* insertions into ancestral piRNA clusters, which as allelic class is subject to positive selection. Therefore, the evolution of *P*-element repression presents a striking example of polygenic adaptation.

Chapter 2 Targeted identification of TE insertions in a *Drosophila* genome through hemi-specific PCR[1]

[1] This chapter has been previously published. Reference: Zhang, S. & Kelleher, E.S., 2017. Targeted identification of TE insertions in a *Drosophila* genome through hemi-specific PCR. *Mobile DNA*. 8:10. doi: 10.1186/s13100-017-0092-1.

2.1 Introduction

Transposable elements (TEs) are mobile genetic entities that are major contributors to the evolution of eukaryotic genomes. TE proliferation can drive dramatic changes in genome size (Vitte and Panaud 2005; Lee and Kim 2014; Sun et al. 2012; Sun and Mueller 2014) and gene regulation (Lee 2015; Hollister and Gaut 2009; Lynch et al. 2011; Jjingo et al. 2014). Additionally, ectopic recombination between TE insertions produces structural rearrangements within and between chromosomes (Startek et al. 2015; Robberecht et al. 2013; Zhang et al. 2013; Prada and Laissue 2014; Sarilar et al. 2014). Finally, transposition into novel genomic sites produces abundant intraspecific variation in the presence and absence of individual TE insertions (Robb et al. 2013; Kofler et al. 2015b; Rishishwar et al. 2015).

Despite their contribution to genetic variation, population genomic studies of TEs remain challenging. Like all repetitive elements, TEs are inherently problematic to assign to particular genomic locations. Furthermore, TEs are often found in heterochromatic regions, such that the genomic sequences that surround them may also be repetitive. Finally, TE insertions are often polymorphic within samples used for genome re-sequencing, meaning they are supported by few sequencing reads, and discerning between false positives and rare insertions can prove difficult (Kofler et al. 2012; Zhuang et al. 2014; Rahman et al. 2015; Fiston-Lavier et al. 2015).

Whole genome re-sequencing (WGS) is often employed to provide a comprehensive picture of genetic variation, including the presence and absence of

TE insertions. Numerous methodologies have been developed for annotation of polymorphic TE insertions from WGS data (Keane et al. 2013; Bergman 2012; Chen et al. 2017; Fiston-Lavier et al. 2015; Kofler et al. 2012; Zhuang et al. 2014; Rahman et al. 2015). However, WGS of a large population genomic sample remains expensive, and may be unnecessary for studies that focus on one or a few active TE families. Additionally because WGS provides variable sequence coverage across the genome, and the power to annotated particular TE insertions may be limited by stochastic low read-depth. Read depth may be critical for identification of a unique TE insertion site, particularly in heterochromatic repeat-rich regions that contain limited unique sequence.

Targeted genomic re-sequencing (TGS) of TE insertions allows for vastly increased sequencing depth at TE insertion sites in smaller sequencing libraries as compared to WGS (Ewing and Kazazian 2010; Platt et al. 2015; Witherspoon et al. 2010). TGS therefore offers combined potential for more robust identification of TE insertions that are rare or occur in repetitive regions, at a reduced sequencing cost. Here, we adapt a hemi-specific PCR approach for TGS of TE insertions on the Illumina platform (Ewing and Kazazian 2010) to *Drosophila* genomes. We further present a computational method for identification of precise TE insertion sites from TGS data. Although our approach is adaptable to any TE or genome, we piloted it by re-sequencing insertions of *P*-elements, DNA transposons that recently invaded the *D. melanogaster* genome and are highly polymorphic among strains (Kelleher 2016; Srivastav and Kelleher 2017; Anxolabéhère et al. 1988; Ronsseray et al. 1989;

Bingham et al. 1982; Rubin et al. 1982). To evaluate our approach, we compared our results to two TE annotation sets based on WGS data for the same strain (Mackay et al. 2012; Zhuang et al. 2014; Rahman et al. 2015).

We demonstrate that TGS by hemi-specific PCR is a powerful method for identification of polymorphic *P*-element TE insertions in *Drosophila*, identifying almost all known insertions (~94%), while also uncovering previously un-annotated insertions in repetitive genomic regions. False-positives in TGS data were easily differentiated from true insertions based on read support. We further demonstrate that TGS allows for identification of precise insertion sites for all annotated TEs, as compared to WGS, where the absence of reads spanning the TE insertion breakpoint often limits the resolution of the annotations to a genomic window. Finally, we describe a new method for estimating the polymorphic frequency of individual TE insertions from WGS data, which takes advantage of precise insertion sites provided by TGS. Overall, our results suggest that TGS based on hemi-specific PCR may be a more powerful and precise method for annotation of polymorphic TE insertions than WGS for the study of particular TE families, such as the *P*-element. However, the two approaches are complementary, and together provide the most complete picture of TE location and frequency.

2.2 Results

2.2.1 Hemi-specific PCR amplifies abundant *P*-element insertions

P-elements are absent from the *D. melanogaster* reference genome (y^1; cn^1 bw^1 sp^1)(Hoskins et al. 2015), but are ubiquitous among recently collected wild-type genomes (Zhuang et al. 2014; Rahman et al. 2015). We therefore chose to pilot our approach by examining *P*-elements in the wild-derived strain RAL-492, which was collected from Raleigh NC in 2003 (Mackay et al. 2012). Illumina paired-end whole-genome sequencing data was previously published for RAL-492, and genomic *P*-elements were previously annotated by the TEMP (33 insertions; Zhuang et al. 2014) and TIDAL (29 insertions; Rahman et al. 2015) TE annotation packages.

To amplify *P*-element insertions and adjacent sequence from the RAL-492 genome (Mackay et al. 2012), we employed a hemi-specific PCR approach, using a forward primer specific to a region at the 3' end of *P*-elements that is required for transposition (Mullins et al. 1989), and a series of 15 degenerate reverse primers (Figure 2.1). Each degenerate reverse primers contains a different common pentamer in the *D. melanogaster* genome followed by 5 four-fold degenerate nucleotides (N bases), allowing it to recognize a diversity of chromosomal sites (Table S2.1). To determine the optimal annealing temperature for hemi-specific PCR, and verify that our approach would amplify a range of DNA fragments corresponding to multiple *P*-element insertions, we examined the size distribution of amplicons for 4 degenerate primers at two different annealing temperatures (55 ˚C and 50 ˚C; Figure 2.2). Although a diversity of fragment sizes were observed

for both annealing temperatures, the range was broader and more evenly distributed among amplicons at 50 ˚C. We therefore separately conducted hemi-specific PCR for 15 degenerate primers at the annealing temperature of 50 ˚C to generate our sequencing libraries.

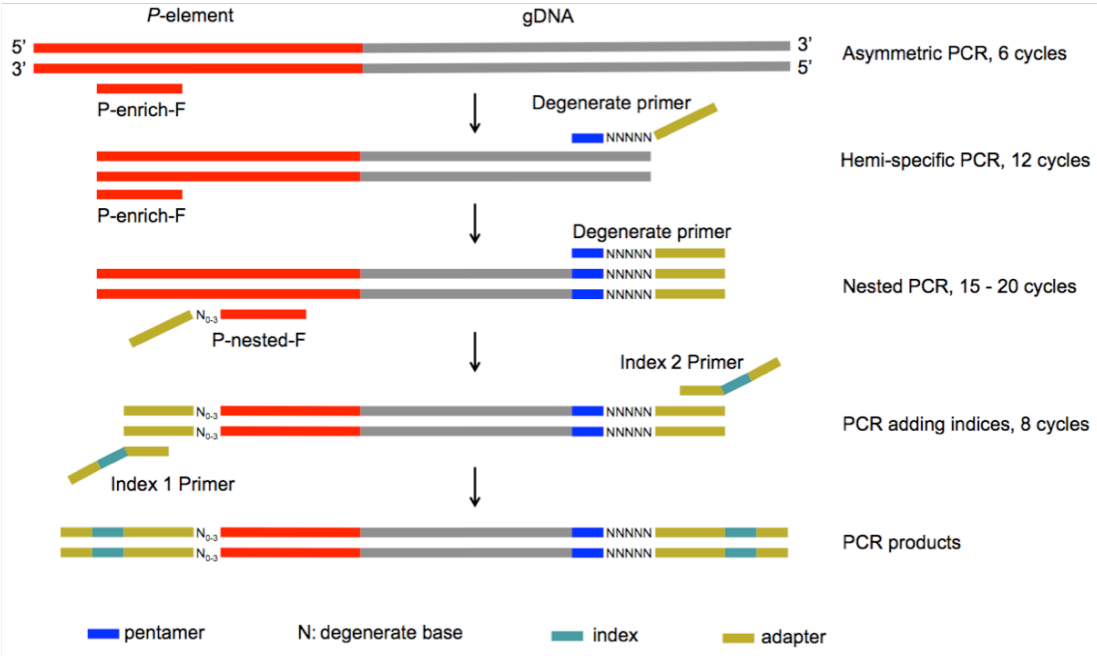

Figure 2.1 Hemi-specific PCR of *P*-element insertions. (A) Sequencing libraries were generated by nested hemi-specific PCR. First, asymmetric PCR enriches for *P*-element 3'ends using a *P*-element specific primer (P-enrich-F) that aligns to *P*-element from position 2752 to 2774 (out of 2907 total nucleotides). Next, a degenerate reverse primer is added recognize and amplify unknown sequences that are adjacent to *P*-element 3' ends. Third, nested PCR with the P-nested-F primer cocktail (positions 2856 to 2877) and the degenerate reverse primer enhances PCR specificity for *P*-elements and produces amplicons with 5' end read complexity, which is required for Illumina sequencing. Last, DNA fragments are amplified with indexing primers to allow for multiplexing. The resulting amplicons consist of adapters at each end, a *P*-element 3' end and its adjacent genomic sequences.

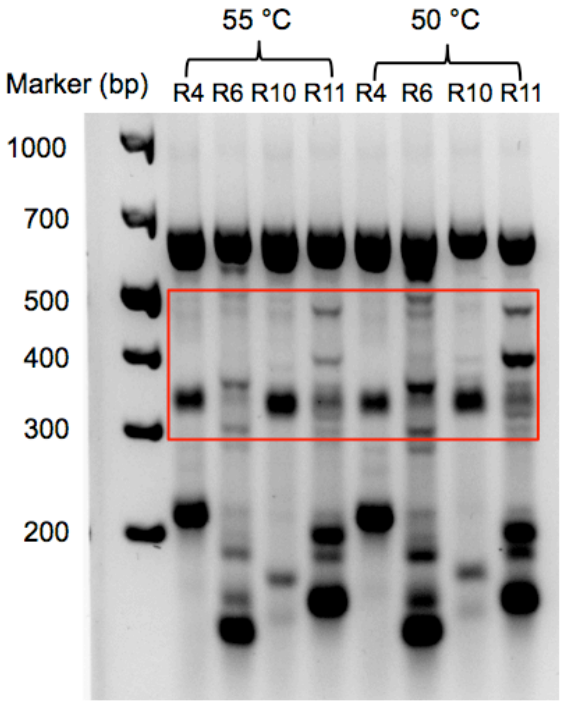

Figure 2.2 Hemi-specific PCR amplified multiple DNA fragments. PCR products from nested PCR with four degenerate primers (R4, R6, R10 and R11) are shown for two different annealing temperatures. DNA fragments between 300 – 500 bp (indicated in red box) were selected for sequencing.

We sequenced 0.43 – 1.31 million read pairs for each of 15 degenerate primers (Table S2.2). >93% of read pairs for all 15 degenerate primers contained 3' *P*-element sequences, indicating our PCR conditions were highly specific (Table S2.2). After trimming *P*-element sequence and low-quality ends, we aligned read pairs to release 6 of the *D. melanogaster* genome (dm6) (Hoskins et al. 2015), and the Telomere Associated Satellites of the *X*-chromosome (*X-TAS*) (Karpen and Spradling 1992). Although *X-TAS* is absent from the genome of the dm6 reference strain (y^1; cn^1 bw^1 sp^1) (Hoskins et al. 2015), these subtelomeric satellites are common among wild-derived genomes and often contain *P*-elements (Anderson et

al. 2008; Ronsseray et al. 1996; Marin et al. 2000; Stuart et al. 2002). Depending on the degenerate primer, 80.8% - 98.0% of read pairs were aligned to the reference, with 20.8% - 97.3% of read pairs aligning to the reference in unique genomic location (Table S2.2). Therefore, there is variation among the degenerate primers in the degree to which the insertions they amplify are surrounded by unique genome sequence.

To identify *P*-element insertions from our sequencing reads, we first considered read pairs that could be uniquely mapped to the reference genome (see Materials and Methods). In total, 53 independent *P*-element insertion sites were suggested in the RAL-492 genome, based on the unique and concordant alignment of >20 *P*-element-derived read pairs to the reference for each insertion (Table S2.3). Of these 53 insertions, 27 had previously been identified from whole genome re-sequencing data by both TIDAL (Zhuang et al. 2014) and TEMP packages (Rahman et al. 2015), and an additional six had been identified by TEMP only (Figure 2.3). By contrast, only two insertions found by TIDAL and TEMP were not detected by hemi-specific PCR. Hemi-specific PCR therefore identified almost all high-confidence *P*-element insertions detected in whole genome re-sequencing data while also suggesting up to 20 previously unknown insertions.

To determine why hemi-specific PCR may fail to detect a small number of insertions, we examined the insertion sites of the two *P*-elements annotated by both TIDAL and TEMP but not hemi-specific PCR. We discovered that in both cases, the annotated insertions were two tail-to-tail *P*-element insertions, meaning that

amplification from the 3' end of one element would produce sequence from the 3' end of the adjacent element, rather than genomic sequence corresponding to the insertion site. False negatives could therefore be avoided with this method in the future by placing *P*-element specific primers at both the 5' and 3' ends of the element.

We also did not detect 19 *P*-element insertions that were found only by TEMP (Figure 2.3). Notably, these insertions were excluded from the published TEMP annotations because they were not estimated to occur at more than 80% frequency in any inbred line, including RAL-492 (Zhuang et al. 2014). If these insertions are true positives that are segregating at a low frequency in RAL-492 (Figure S2.1A), they may not have been represented in the sample of genomic DNA that we used for Illumina library prep. Alternatively, these insertions may be false positives, as they are supported by fewer read-pairs in whole genome re-sequencing data than those that were also identified by TIDAL, hemi-specific PCR, or both (Figure S2.1B). Indeed, we attempted to amplify one of these insertions using standard PCR and were unable to do so (Table S2.4).

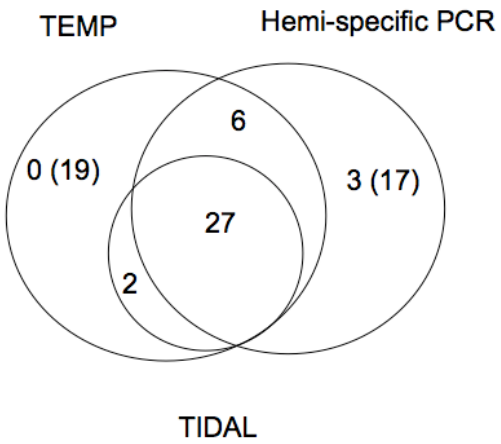

Figure 2.3 The number of *P*-element insertions found by Hemi-specific PCR, TEMP and TIDAL. The number of *P*-element insertions is indicated in each subset. The number in parentheses indicates the number of known or potential false positives.

2.2.2 Validation of novel insertions and identification of false positives

To validate the 20 candidate novel *P*-element insertions identified by hemi-specific PCR we performed site-specific PCR. Among the *P*-element insertions found only by hemi-specific PCR (Figure 2.3), three insertions (chr2L:20917521, chrX_TAS:4894 and chrY:768808) could be amplified from RAL-492 genomic DNA (Table S2.4). Insertions at chrX_TAS:4894 and chrY:768808 appear to be fixed in the RAL-492 strain, and we were able to identify read pairs (15 for chrX_TAS:4894 and 18 for chrY:768808) in the previous WGS data that support these two insertions. However, because these insertions are located in repetitive genomic regions, there were no read pairs in the WGS data that uniquely aligned to either insertion site, preventing their detection by TEMP and TIDAL. The read depth provided by TGS therefore offers greater power to identify TE insertions in heterochromatic regions.

The third insertion, chr2L:20917521 is polymorphic, as indicated by the presence of PCR amplicons corresponding to both inserted and un-inserted chromosomes (Figure S2.2). There were no read pairs supporting this polymorphic insertion in the previous WGS data, perhaps because the inserted chromosome was not sampled among individuals used for the sequencing library.

We could not validate the remaining 17 insertions that were uniquely identified by hemi-specific PCR, either through insertion-specific PCR or from previous whole-genome sequencing data (Table S2.4). We therefore believe these are false positives that result from PCR artifacts that occur during library prep. Fortunately, false positives are easily distinguished from true insertions by the low abundance of supporting reads among our sequencing libraries and their presence in sequencing libraries from only a few degenerate primers (Figure 2.4). If we require at least 100 read pairs and four degenerate primers to define a *P*-element insertion, we are able to exclude all but one of the false positives. Excluding false-positives, we detected 36 *P*-element insertions in the RAL-492 genome, three of which were previously unknown (Table S2.3).

2.2.3 Sequence similarity to true insertion sites may produce false positives

There is one outlier among the false positives: an insertion at chr3L:25797105 (Figure 2.4A) that is supported by 1478 read pairs and 13 degenerate primers. Notably, we found the sequence around this insertion site was 94% similar across 446 bp to sequence at a true insertion site (chr3L:26023661). Therefore, some false

positives may occur due to nucleotide substitutions introduced during PCR and sequencing, which cause a subset of reads derived from a true insertion to align better to highly similar sequences elsewhere in the genome. Consistent with this, the reads supporting the false positive were 0.17% as abundant in our data as compared to reads supporting the true insertion (Table S2.3), which is similar to what is expected based on the per-site mutation rate for Taq DNA polymerase (0.003%) (McInerney et al. 2014) and the Illumina MiSeq platform (0.8%) (Quail et al. 2012). Furthermore, reads supporting the true insertion site were separated by fewer mutations from the reference genome (mean 2.2 mutations per 100 bp) as compared to reads supporting the false positive insertion (mean 6.7 mutations per 100 bp).

To determine whether sequence identity might explain other potential false positives we observed in our data, we compared 0.8 Kb of the genomic region surrounding all insertion sites to each other via BLAST (Altschul et al. 1990). We found that the genomic sequence at two potential false positives chr3L:26834988 and chrUn_CP007074v1:15794 exhibited significant sequence similarity to the PCR-verified insertion chrX_TAS:4894 (87% across 83 bp for chr3L:26834988; 84% identity across 93 bp for chrUn_CP007074v1:15794). In both cases, reads supporting the potential false-positive insertions were <1% as abundant as reads supporting the true positive.

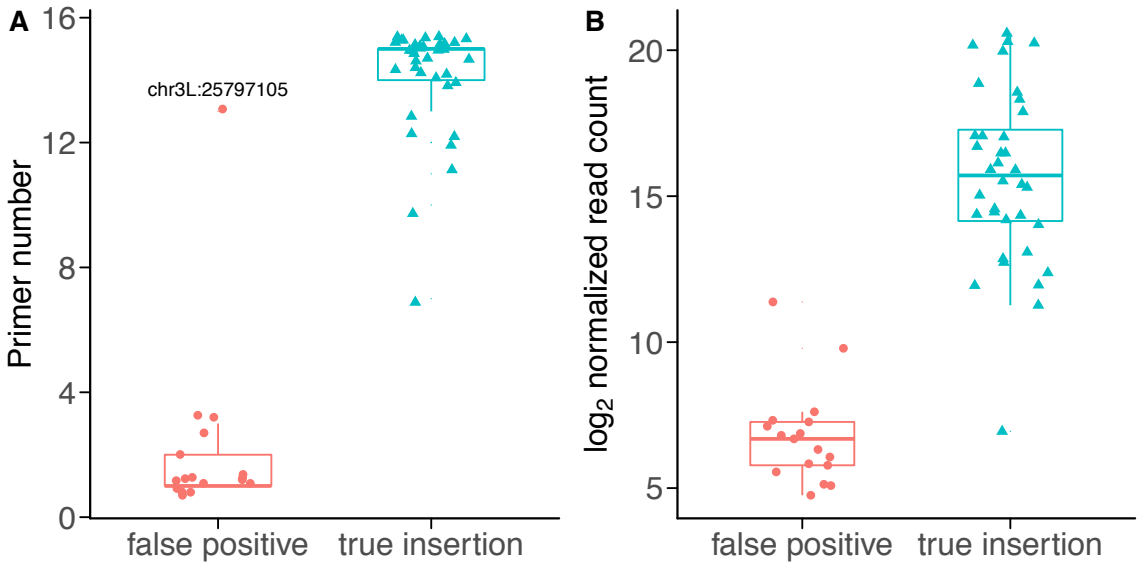

Figure 2.4 Read and primer support for true insertions and false positives detected by hemi-specific PCR. False-positives were detected by hemi-specific PCR but could not be validated by insertion-specific PCR or whole genome re-sequencing data, whereas true insertions were verified by one or both of these methods. (A) True insertions are sampled more sequencing libraries generated using different degenerate primers for hemi-specific PCR (Welch's t_{22} = 15.56, P = 2.91 x 10^{-13}). (B) True insertions are supported by larger number of uniquely mapping read pairs in hemi-specific PCR libraries (Welch's t_{50} = 13.78, P < 2.2 x 10^{-16}). The number of read pairs was normalized to reads per million based on total sequenced reads from each degenerate primer.

2.2.4 The majority of sequencing reads are explained by annotated insertions

For some degenerate primers, >50% of read pairs aligned to the reference genome in multiple locations (*i.e.* multiply mapping, Table S2.2). These read pairs might be derived from one of the 36 insertions that were annotated from unique alignments. Alternatively, they may indicate the presence of false negatives, which could not be annotated due to an absence of uniquely mapping reads. To differentiate between these alternatives, we constructed a putative contig for each of the 36 *P*-element

insertions, which was comprised of the full-length *P*-element consensus flanked by 500 nucleotides of adjacent genomic sequence (see Materials and Methods). Multiply mapping reads that support annotated insertions were then identified based on their alignment to the 36 putative insertion contigs.

For all but one of the degenerate primers, >95% of multiply mapping reads could be aligned to at least one of the 36 putative insertion contigs (Table S2.2). Furthermore, most multiply mapping reads were aligned to insertions in repetitive genomic regions, such as chrX_TAS:4894. Therefore, with the exception of the tail-to-tail elements, our analysis pipeline likely detects most or all of the *P*-elements present in hemi-specific Illumina libraries.

2.2.5 Improved insertion site identification and frequency estimation

Read-pairs generated by hemi-specific PCR include at least one "split-read" which is comprised of both TE and adjacent genomic sequences. Split reads are invaluable for TE annotation, because they allow for precise identification of the breakpoint that characterizes each insertion (Figure 2.5), but are often absent from annotations based on WGS data due to lower read depth at individual insertion sites. For example, although the precise insertion site of all 36 insertions detected in the RAL-492 genome by hemi-specific PCR were identified, five of these insertion sites were absent from TEMP annotations based on WGS data, due to a lack of split reads (Zhuang et al. 2014). An additional five insertions had slightly different insertion

sites inferred by hemi-specific and WGS, suggesting potential inaccuracy in annotation of the insertion site.

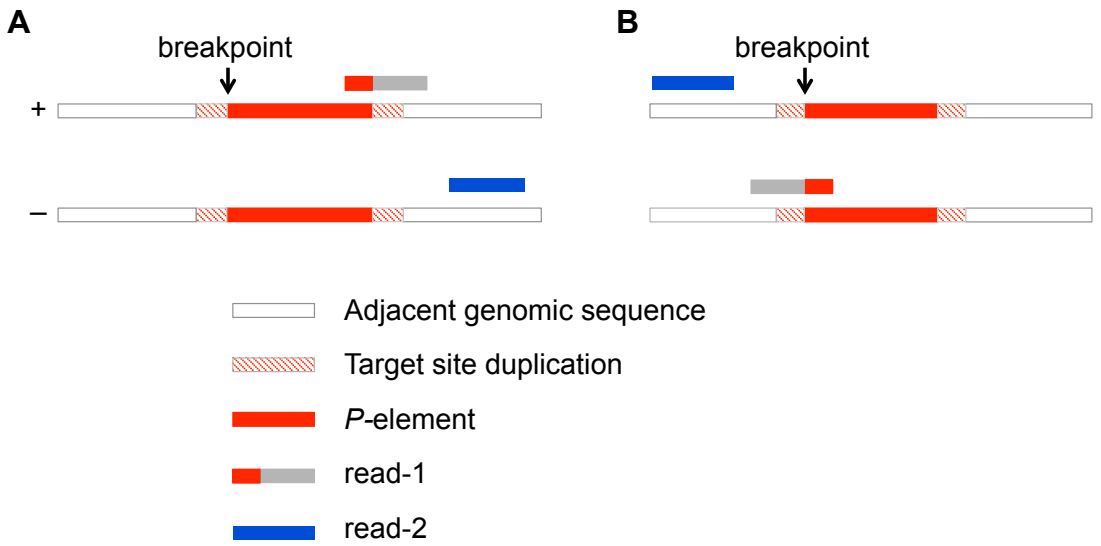

Figure 2.5 Insertion site identification and putative insertion contig structure. Read-1 of each pair generated by hemi-specific PCR is a split read that contains both *P*-element and adjacent genomic sequence. Breakpoints are determined based on the alignment of read-1 (red) to the plus (A) or minus genomic strand (B). Contigs are constructed through insertion of the *P*-element consensus at the insertion site, which is flanked by an 8 bp target site duplication on either side.

Precision and accuracy of insertions site annotation could be of particular value in facilitating the estimation of polymorphic TE insertion frequencies from WGS data. TE annotation packages such as TEMP and TIDAL estimate the frequency of an individual TE insertion among sequenced genomes as the proportion of read pairs aligning to the insertion site that support the insertion allele. However, because precise insertion sites are not always known, reads supporting each chromosome cannot be identified by concurrent alignment to the reference genome and a putative insertion allele. Rather, reads are aligned to the reference genome

only, and read-pairs supporting the insertion allele are identified by a minimal number of nucleotides (7 nt for TEMP and 22 nt for TIDAL) that align to the TE consensus. Such an approach likely underestimates the number of reads supporting the insertion chromosome by excluding read-pairs that include very little TE sequence.

Taking advantage of the precise breakpoints that are provided by hemi-specific PCR, we developed a new method for estimating the frequency of polymorphic TE insertions in WGS data. Unfortunately, the frequency of the insertion allele cannot be estimated from TGS data, because reads supported the reference allele (lacking a TE insertion) are not represented in the sequencing library. We aligned WGS reads concurrently to the reference genome as well as putative contigs for each of the 36 annotated insertions. We then estimated the frequency of each *P*-element insertion based on the number of read-pairs in WGS data that exhibit a significantly better alignment to the putative insertion contig than to the corresponding window in the reference genome.

Based on this approach, we estimate that 97.2% (35 out of 36) of the *P*-element insertions identified by both TEMP and hemi-specific PCR are completely fixed in RAL-492, as expected in a highly inbred line. By contrast, using the same WGS data as we employed, TIDAL and TEMP estimated that many insertions remained polymorphic after inbreeding (Figure 2.6A). Specifically, for the 27 insertions found by TEMP, TIDAL and hemi-specific PCR (Figure 2.3), the median frequency estimated from concurrent alignment to the reference and putative

insertion contig was 0.31 higher than the TIDAL estimate ($P < 1 \times 10^{-6}$, based on 10^6 permutations of the observed data) and 0.11 higher than the TEMP estimate ($P = 5.1 \times 10^{-4}$, based on 10^6 permutations of the observed data). The higher estimated TE insertion frequencies generated by concurrent mapping resulted from a larger number of identified read pairs that support the insertion chromosome, as compared to the TIDAL and TEMP approaches (Figure 2.6B; linear contrast $F_{1,54} = 564.54$, $P < 2 \times 10^{-16}$). Furthermore, TIDAL generated the lowest estimated frequencies and the fewest reads supporting the inserted chromosome, which is consistent with the most stringent requirements for identification of reads supporting the insertion (22 nt overlap with the consensus).

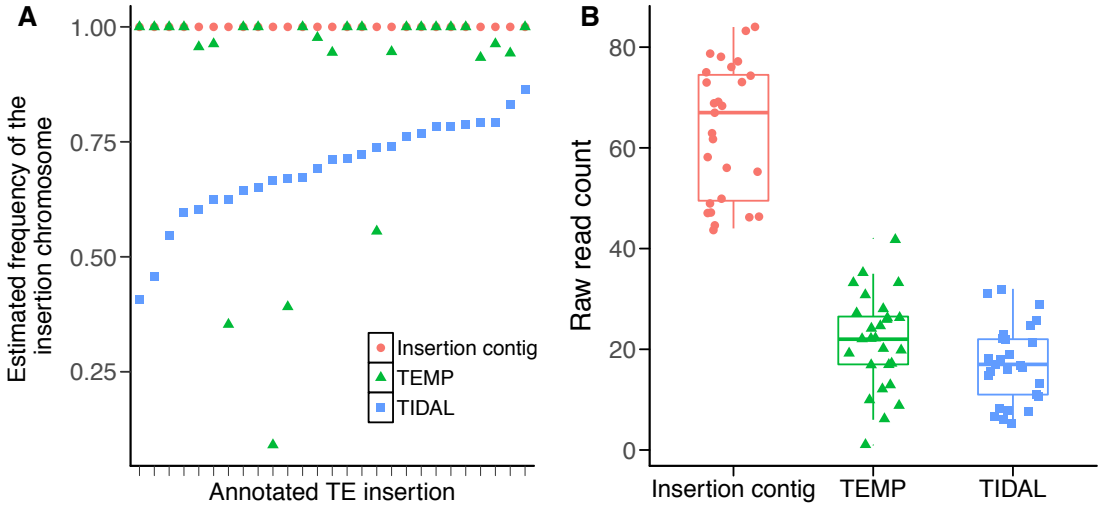

Figure 2.6 Estimation of TE insertion frequency. (A) Estimated frequencies for 27 TE insertions in RAL-492 generated by TEMP, TIDAL, and our concurrent alignment approach (contig). All three frequency estimates are based on previously published WGS data from RAL-492 (Mackay et al. 2012). (B) The number of WGS read pairs supporting each *P*-element insertion identified by TIDAL, TEMP and concurrent alignment (contig).

For six insertions, we validated that the insertion was fixed in our RAL-492 sample by performing PCR with primers on either side of the insertion site, such that both the insertion allele and the reference (un-inserted) allele would amplify if present. Only the insertion allele amplified, suggesting that the reference allele was absent. Collectively, our observations suggest a systematic bias towards low TE insertion frequency estimates when reads are not aligned to a putative insertion contig that is defined by precise breakpoints.

2.3 Discussion

Our results validate hemi-specific PCR as a powerful method for TGS of particular TE families. Of 38 true insertions in the RAL-492 genome, which were either independently validated by site-specific PCR (Table S2.4), or were found in multiple annotation sets (Table S2.3), 36 could be identified from sequencing reads generated by hemi-specific PCR. By contrast, TEMP detected 35 true insertions (Zhuang et al. 2014) while TIDAL detected 29 (Figure 2.3)(Rahman et al. 2015). Hemi-specific PCR therefore exhibited marginally to significantly improved power to detect true insertions when compared to previous analyses of WGS data, based on ~50% fewer sequencing reads (Table S2.2)(Mackay et al. 2012). Furthermore, given that all but one true insertion was supported by >1000 uniquely mapping reads in our data (Table S2.3), hemi-specific PCR libraries could be highly multiplexed while still retaining power to discover the vast majority of insertions. Importantly, we

were able to avoid almost all false positives by excluding insertions that were supported by few reads or degenerate primers (Figure 3), revealing that the enhanced power of TGS for genome annotation does not come at the expense of accuracy. By contrast, TEMP annotation of WGS data detected almost all true insertions but also exhibited a high false-positive rate, while TIDAL avoided false positives but missed many true insertions (Figure 3; Table S2.4).

Annotating TE insertions in heterochromatic regions based on WGS data remains challenging, as individual insertions are often supported by only few read pairs, which may not yield a unique alignment in repeat rich sequence. Annotation of polymorphic TE insertions in heterochromatic regions is of particular interest due to the known role of heterochromatic piRNA clusters in regulating germline TE activity in both mammals and insects (Aravin et al. 2007a; Brennecke et al. 2007). TGS by hemi-specific PCR offered improved annotation in heterochromatic regions, as two of the three previously un-annotated insertions we discovered here were in heterochromatin. Indeed, one of the previously unknown insertions we annotated is in the *X-TAS*, a prolific piRNA cluster (Aravin et al. 2007a) that plays an important role in *P*-element regulation (Ronsseray et al. 1991; Brennecke et al. 2008; Ronsseray et al. 1996; Marin et al. 2000; Stuart et al. 2002). TGS by hemi-specific PCR may therefore provide an opportunity to examine polymorphic TE insertions that determine differences in TE regulation (Zanni et al. 2013).

Our TGS and analysis method, based on hemi-specific PCR, also provided precise insertions sites for all annotated TEs, which are often lacking from

annotations based on WGS data. Precise insertion sites provide more information about the potential functional impact of a TE insertion. Additionally, as we demonstrated, they allow for more accurate estimates of the polymorphic frequency of TE insertions from WGS data. Estimating TE insertion site frequencies is critical for examining the selective forces that act on TE insertions (Petrov et al. 2011; Kofler et al. 2012, 2015b). They are also important to consider when evaluating associations between particular TE insertions and phenotypes of interest in genome-wide association studies.

2.4 Conclusion

Our results indicate that hemi-specific PCR offers an attractive alternative approach to WGS for identification of polymorphic TE insertions of particular TE families in *Drosophila* genomes. As expected for a targeted approach focused on a single TE family, TGS was more powerful for annotating true positive *P*-element insertions than WGS, and also offered enhanced precision and accuracy in determining the exact location of those insertions. Furthermore, this performance was achieved at a lower read depth and therefore reduced sequencing cost.

TGS is easily adapted to other host genomes or TE families through development of new nested and degenerate primer sets. Indeed our method is modeled after that of Ewing and Kazazian (2010), which curated LINE-1 elements in human genomes. Additionally, TGS could be expanded to identify polymorphic insertions for many TE families in the same library by incorporating multiple nested

primer pairs. Such an approach would be invaluable for population genomic studies that focus on the dynamics of particular active TE families.

2.5 Materials and Methods

2.5.1 Genomic DNA samples

RAL-492 and RAL-802 strains were obtained from the Bloomington *Drosophila* Stock Center. Genomic DNA was extracted using the Qiagen DNeasy Blood and Tissue kit.

2.5.2 Primer design

Our library preparation method is modeled after the approach described by Ewing and Kazazian (2010), which amplifies LINE-1 elements and adjacent genomic sequences in human genomes (Figure 2.1). By combining nested forward primers that are specific to 3' end of *P*-element with degenerate reverse primers, we preferentially amplified *P*-elements and their adjacent genomic sequences. The first *P*-element specific primer (P-enrich-F) enriches 3' *P*-element ends, while the second (P-nested-F) contains Illumina nextera adapter sequences to allow for sequencing of amplicons. The nested forward primers used for PCR bind to sequences that are required for *P*-element mobilization, and therefore are expected to a conserved among genomic *P*-elements (Mullins et al. 1989). In addition, the forward nested primer was an equimolar cocktail of 4 different primers, which are complementary

to the same stretch of the *P*-element 3' end (position 2856 to 2877), but have spacers of 0 - 3 "N" nucleotides from the Illumina adaptor sequence (Figure 2.1). The spacers ensure sequence complexity at the start of the sequencing read, which is critical to the success of the sequencing reaction.

To design degenerate reverse primers for hemi-specific PCR, we first identified common pentamers in the *D. melanogaster* genome with jellyfish (Marçais and Kingsford 2011). We selected a set of 15 pentamers that are common, but also diverse in their sequence composition, to maximize the breadth of genomic sequences that could be recognized by the degenerate primers. Each degenerate primer was comprised of an Illumina adapter for nextera sequencing, followed by 5 degenerate nucleotides, followed by a common pentamer from 5' to 3'. Primers used in library construction are listed in Table S2.1.

2.5.3 Library construction by hemi-specific PCR

The first six cycles of PCR were asymmetric, and enriched for the 3' end of *P*-elements. The PCR was conducted in a 46 μL reaction volume with 10 μL of 5X GoTaq Flexi Buffer (Promega), 6 μL of 25 mM $MgCl_2$, 2 μL of 20 μM P-enrich-F primer, 0.5 μL of 100% DMSO, 0.5 μL of Flexi GoTaq, 1 μL of 10 mM dNTPs, and ~500 ng template DNA. The PCR conditions were 2:30 min at 95 ˚C, followed by 6 cycles of 30 sec at 95 ˚C, 1 min at 62 ˚C and 2 min at 72 ˚C.

The second PCR was hemi-specific, and allowed for 12 cycles of amplification of *P*-element 3' ends and adjacent genomic sequences. 4 μL of each degenerate

primer (5 μM) was added to a separate asymmetric PCR reaction mix. The reaction conditions were 2 min at 95 ˚C, followed by 12 cycles of 30 sec at 95 ˚C, 30 sec at 50 ˚C and 2 min at 72 ˚C, followed by 10 min at 72 ˚C. The PCR product was purified using the QIAquick PCR Purification Kit (Qiagen), yielding 20 μL DNA.

The third PCR (15-20 cycles) was nested, and provides enhanced specificity for *P*-element targets. Purified PCR products from PCRs 1 and 2 were used as templates, and amplification was targeted by an Illumina-tagged forward nested *P*-element primer, and the same degenerate reverse primer employed PCR 2. The PCR was conducted in 50 μL reaction volume with 10 μL of 5X GoTaq Flexi Buffer, 6 μL of 25 mM $MgCl_2$, 4 μL of 5 μM equimolar forward primer, 4 μL of degenerate primer, 0.5 μL of 100% DMSO, 0.5 μL of Flexi GoTaq, 1 μL of 10 mM dNTPs, and 10 μL template DNA from last step. The PCR condition is: 2 min at 95 ˚C, followed by 15 - 20 cycles of 30 sec at 95 ˚C, 30 sec at 55 ˚C and 30 sec at 72 ˚C, followed by 10 min at 72 ˚C. For degenerate primers R4, R6, R8, R9, R11, R12, R13, R15, PCR 3 was performed for 15 cycles. Because the remaining degenerate primers yielded weak bands or no bands after 15 cycles, we increased the number of cycles to 20 for these primers. For all 15 libraries, 300 – 500 bp PCR products were isolated from agarose gels and purified using the QIAquick Gel Extraction Kit (Qiagen), and 22. 5 μL purified DNA were eluted.

The fourth PCR (8 cycles) incorporated indices for multiplexing on the Illumina platform using the Illumina Nextera XT Index Kit. The PCR was conducted in a 50 μL reaction volume with 10 μL of 5X GoTaq Flexi Buffer, 6 μL of 25 mM

$MgCl_2$, 5 μL of index 1, 5 μL of index 2, 0.5 μL of Flexi GoTaq, 1 μL of 10 mM dNTPs, and 22.5 μL template DNA from last step. The PCR conditions were: 3 min at 95 ˚C, followed by 8 cycles of 30 sec at 95 ˚C, 30 sec at 55 ˚C and 30 sec at 72 ˚C, followed by 5 min at 72 ˚C. PCR products between 300 – 500 bp were isolated from an agarose gel, and purified using the QIAquick Gel Extraction Kit. The resulting sequencing libraries were paired-end sequenced (2X150 nt reads) on the MiSeq platform by the Weill Cornell Epigenomics Core. Sequencing libraries are available in the NCBI sequence read archive (SRR5712353 to SRR5712367).

2.5.4 Identification of *P*-element-derived read-pairs and alignment to the reference genome

Based on the placement of the P-nested-F primer, read-1 from every read pair should begin with 52 nt at the 3' terminus of the *P*-element consensus (Fig. 1A). The first 22 nt are included in the P-nested-F primer, while the remaining 30 will occur only in amplicons that arise from true *P*-element 3' ends. We therefore locally aligned all read-1 sequences to the full-length *P*-element consensus sequence (O'Hare and Rubin 1983) using bowtie2 (v2.1.0) (Langmead and Salzberg 2012) and selected read pairs where the alignment of read-1 to 3' end of *P*-element was longer than 20 nt using a custom Perl script (1 mismatch and 1 gap allowed; Supplementary script2). Any remaining Illumina sequencing adapters and *P*-element sequences, as well as low-quality ends, were removed from our selected

read pairs using cutadapt (v1.9.1) (Martin 2011). The *P*-element derived and trimmed read pairs were used for all down-stream analyses (Table S2.2).

2.5.5 Annotation of *P*-element insertions based on uniquely mapping read pairs

To pinpoint *P*-element insertions in the RAL-492 genome, read pairs were globally aligned to dm6 as well as *X-TAS* using bowtie2 with default options. The results of alignments to the reference genome are reported in Table S2.2. For read pairs that concordantly (*i.e.* aligned with expected orientation and the distance between mates is within 500 bp) and uniquely aligned to the reference genome, we determined the breakpoints of *P*-element insertions based on the reported alignments using a custom shell and Perl scripts (Supplementary script1, 2, 3 and 4). As *P*-element transposition will generate 8-bp target site duplications (Beall and Rio 1997), we defined breakpoints as the 3' end of the 8-bp target site duplication on the plus genomic strand. If the *P*-element insertion is in the same orientation as the plus genomic strand, the breakpoint is equal to the location where left-most nucleotide was aligned in read-1 plus 7 bp (Figure 2.5A). In contrast, the breakpoint is equal to location where the right-most nucleotide was aligned in read-1 if the inserted *P*-element is in the same orientation as the minus genomic strand (Figure 2.5B). We required 20 concordant, uniquely mapping read pairs to annotate a single insertion. *P*-element insertions found by uniquely mapping read pairs are reported in Table S2.2.

2.5.6 Determining the number of *P*-element reads that arise from annotated insertions

To determine how many multiply mapping reads could be derived from one of the 36 insertions, we annotated based on unique and concordant alignment to the reference genome, we aligned multiply mapping reads to putative insertion contigs that we generated for each annotated insertion. Each of the ~300-500 bp PCR products that were sequenced contain 52 bp of *P*-element sequence and 77 bp of Illumina adapter sequence, with the remaining sequence (up to ~371 bp) deriving from the genomic region adjacent to each insertion. We therefore constructed putative insertion contigs that contained the *P*-element consensus and 500 bp adjacent genomic sequences at 5' and 3' end, including the inferred 8 bp target site duplication (Figure 2.5). Multiply-mapping read pairs were aligned to the putative insertion contigs using bowtie2, allowing for up to 5 mismatches and 2 gaps. The number of multiply mapping read pairs that could be aligned to at least one annotated insertion are listed in Table S2.2.

2.5.7 Estimating the frequency of individual insertions from whole genome sequencing paired-end data

To estimate frequency of each annotated TE insertion, we used previously published whole genome re-sequencing data for RAL-492 (Mackay et al. 2012) to compare the abundance of read pairs supporting the insertion allele and reference genome. Read

pairs were globally aligned to a hybrid assembly that combined the putative insertion contig for each of our insertions, as well as the dm6 assembly, using bowtie2. Only alignments with a mapping quality score (MAPQ) greater than 10, indicating high confidence that they are the correct alignment for a particular read-pair, were retained. A read pair was considered to support the insertion if it aligned to the putative insertion contig and its alignment spanned the breakpoint. Similarly, a read pair was considered to support the reference genome if it aligned to dm6 and the alignment spanned the breakpoint. The frequency of the TE insertion was estimated the proportion of the number of read pairs supporting the insertion out of total number of read pairs supporting either the inserted or un-inserted chromosomes.

2.5.8 Site-specific PCR

To verify the existence of *P*-element insertions found by hemi-specific PCR and other approaches, we designed two different types of PCR assays. Insertion site assays combined forward and reverse primers on either side of each insertion site, such that potential PCR products would include both the reference and the insertion allele. Breakpoint-specific assays combined a *P*-element specific primer and a primer in the adjacent genomic sequence, and were specific to the insertion allele. PCR products were Sanger sequenced to further verify the presence or absence of *P*-element insertions. The primers for each insertion site we examined, as well as the PCR and sequencing results, are summarized in Table S2.4.

With the exception on the *X-TAS* insertion, primers for site-specific PCR amplify a unique location in the reference genome. Even repetitive genomic regions often carry distinct combinations of adjacent repeats that allow for site-specific PCR. For the *X-TAS* insertion, we used a break point specific assay combining a primer anneals to a satellite sequence that is unique to *X-TAS* array (Asif-Laidin et al. 2017) with a *P*-element specific primer. A positive result is diagnostic of a *P*-element insertion in a particular orientation in the *X-TAS* locus.

Chapter 3 piRNA-mediated silencing of an invading transposable element evolves rapidly through abundant beneficial *de novo* mutations[2]

[2] This chapter has been submitted as "piRNA-mediated silencing of an invading transposable element evolves rapidly through abundant beneficial *de novo* mutations" to *Genome Research* with a preprint at *bioRxiv*
doi: https://doi.org/10.1101/611350 (Zhang and Kelleher 2019)

3.1 Introduction

Transposable elements (TEs) are widespread genomic parasites that increase their copy number by mobilizing and self-replicating within their host genomes. TEs impose a severe mutational load on their hosts by producing deleterious insertions that disrupt functional sequences (Levis et al. 1984; McGinnis et al. 1983), causing DNA damage through encoded endonucleases (Gasior et al. 2006), and mediating ectopic recombination leading to structural rearrangements (Lim 1988). TE expression and proliferation are therefore strictly regulated, particularly in germline cells where TEs are exceptionally active and resulting mutations are transmitted to offspring. In the germline of most metazoans, TEs are controlled by a conserved small RNA-mediated pathway, in which Piwi-interacting RNAs (piRNAs), in complex with Argonaute proteins, silence TEs in a sequence-specific manner (Houwing et al. 2007; Brennecke et al. 2007; Aravin et al. 2007; Girard and Hannon 2008).

On evolutionary time scales, TEs are frequently horizontally transferred between non-hybridizing species, allowing TE families to colonize new host genomes (Thomas et al. 2010; Dotto et al. 2015; Peccoud et al. 2017). Although host regulation of endogenous TEs by piRNAs is ubiquitous, how the host evolves repression to novel TEs invading the genome remains poorly understood. After invasion, repressor alleles are proposed to arise through *de novo* mutation, when an invading TE copy inserts into a piRNA producing locus referred to as a piRNA cluster (Khurana et al. 2011; Girard and Hannon 2008). The existence of numerous

alternative piRNA clusters (e.g., 142 loci or ~3.5% of assembled *D. melanogaster* genome based on Brennecke et al. 2007) may facilitate the evolution of repression by increasing the mutation rate to generate repressors (Kelleher 2016; Kelleher et al. 2018; Kofler 2019). However, the technical challenge of annotating polymorphic TE insertions in repeat-rich piRNA clusters has limited the identification and study of these repressor alleles. Furthermore, for most TE families it is impossible to distinguish repressor alleles that arose via *de novo* insertion into existing piRNA clusters from the reciprocal: *de novo* piRNA clusters that arose at existing TE insertions. In particular, recent studies suggest that novel piRNA clusters may emerge frequently via epigenetic mutation, when a change in chromatin state triggers bi-directional transcription and piRNA production (De Vanssay et al. 2012; Le Thomas et al. 2014; Hermant et al. 2015).

The role of selection in the evolution of host TE repression, through piRNA mediated silencing or otherwise, also remains controversial. In sexually reproducing organisms, the selective advantage of a repressor allele is limited by recombination, which separates the repressor from the DNA it has protected from deleterious mutation (Charlesworth and Langley 1986). Additionally, while selection for repression may be strong when the genome is invaded by a new TE family, it is unclear whether it is sustained for a sufficient number of generations to enact meaningful changes in repressor allele frequency (Lee and Langley 2012). On the other hand, forward simulation models suggest that piRNA-mediated repressor alleles are targets of positive selection, especially when transposition rates are high

and TEs are highly deleterious (Lu and Clark 2010; Kelleher et al. 2018; Kofler 2019). Moreover, an early population genomic analysis of *D. melanogaster* suggests that TE insertions in piRNA clusters may segregate at higher frequency than non-cluster insertions, although this is based on modest sample size and read depth (Lu and Clark 2010).

The recent invasion of both *Drosophila melanogaster* and *Drosophila simulans* by *P*-element DNA transposons (Kidwell 1983; Anxolabéhère et al. 1988; Kofler et al. 2015a) provides a unique opportunity to study not only the contributions of *de novo* mutation to the evolution of piRNA-mediated silencing by resolving the location of piRNA clusters both before and after an invasion event, but also evolutionary dynamics of repressors when selection is most strong. Unlike most TEs that have colonized host genome for a long evolutionary time, *P*-elements invaded the *D. melanogaster* genome around 1950 by horizontal transfer from *D. willistoni* (Daniels et al. 1990; Kidwell 1983; Anxolabéhère et al. 1988). Similarly, *D. simulans* acquired *P*-elements from *D. melanogaster* around 2010 (Kofler et al. 2015a). In response, many natural populations of *D. melanogaster* evolved piRNA-mediated repression in less than 50 years (Jensen et al. 2008; Brennecke et al. 2008; Kidwell 1983). However, numerous strains collected prior to both invasions are retained in laboratories and stock centers, providing a historical record of ancestral piRNA clusters that were active before the *P*-element invasion.

Here, we take advantage of ~200 fully sequenced *D. melanogaster* genomes, comprising the *Drosophila* Genetic Reference Panel (DGRP) (Mackay et al. 2012;

Huang et al. 2014), to study the emergence and evolutionary dynamics of piRNA-mediated repressor alleles after the *P*-element invasion into *D. melanogaster* populations. To differentiate *de novo* insertions into ancestral piRNA clusters from novel piRNA clusters, we identified piRNA clusters in *D. melanogaster* from 9 *P*-element free strains of *D. melanogaster* collected before invasion. Furthermore, to empower the identification of repressor alleles, we developed a novel approach to identify TE insertions in repetitive DNA. We show that more than 90% of DGRP lines have at least one *P*-element in an ancestral piRNA cluster, indicating *P*-element repressors are widespread in natural populations. Moreover, we detected no fewer than 84 independent *P*-element insertions in ancestral piRNA clusters, suggesting an exceptionally high *de novo* mutation rate for the formation of piRNA-mediated repressor alleles. Finally, we observed that *P*-element insertions in piRNA clusters segregate at higher frequency than putatively neutral insertions in similar genomic compartments, indicating they are targets of positive selection. Together, our results reveal a striking example of adaptation in a polygenic system, in which a plethora of *de novo* beneficial *P*-element insertions into piRNA clusters fueled the evolution of a ubiquitous repressive phenotype in <60 years.

3.2 Results

3.2.1 Identification of ancestral piRNA clusters

We first sought to annotate ancestral piRNA clusters in the *D. melanogaster* genome. We took advantage of 27 small RNA sequencing libraries from 9 wild-type derived

strains (Table S3.1), which were isolated from nature prior to the *P*-element invasion and are therefore devoid of genomic *P*-elements. Using proTRAC (Rosenkranz and Zischler 2012), we annotated piRNA clusters based on the density of mapped piRNAs. By varying the density of reads required to annotate a piRNA cluster (pdens = 0.01, 0.05, and 0.10), we generated three sets of annotations, which contained 32, 159 and 497 piRNA clusters, and comprised 0.30%, 1.27 %, 3.68% of the assembled *D. melanogaster* genome, respectively (Figure 3.1A).

We identified some genomic loci that differ in their status as a piRNA cluster between genotypes, producing abundant piRNAs in some strains while remaining quiescent in others (Figure 3.1A-B). We therefore defined ancestral piRNA clusters as genomic regions that were annotated from at least one small-RNA library. Notably, major known piRNA clusters such as *flamenco* and *42AB* (Robert et al. 2001; Malone et al. 2009; Li et al. 2009; Brennecke et al. 2007) produced abundant piRNAs in all genotypes, and were annotated as piRNA clusters regardless of stringency (Figure 3.1A). In light of clear examples of polymorphism, our annotations should not be considered a comprehensive list of the piRNA clusters segregating in ancestral populations, but rather a representative sample that includes most clusters segregating at high frequency or fixed at the time of invasion.

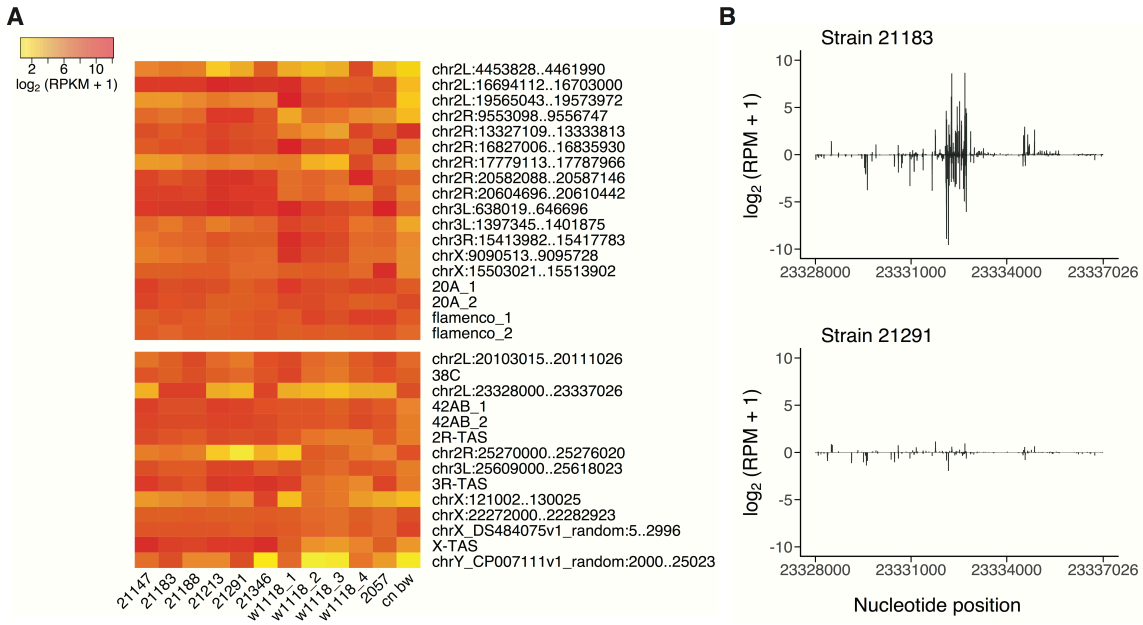

Figure 3.1 (A) Activity of 32 piRNA clusters in ancestral (*P*-element free) strains of *D. melanogaster*. Each column represents a small RNA sequencing library (biological replicates are combined) and each row represents a piRNA cluster annotated in at least one of these libraries. Coordinates of piRNA clusters are based on the *D. melanogaster* release 6 assembly (dm6: Hoskins et al. 2015). piRNA cluster expression levels are estimated by Reads Per Kilo base per Million mapped reads (RPKM) and transformed to $\log_2$ scale ($\log_2$ (RPKM + 1)). Clusters above the white line are uni-strand piRNA clusters and ones below the white line are dual-strand piRNA clusters. Details on small RNA library prep, which may be related to differences in annotated piRNA clusters between libraries from *w1118*, are provided in Table S3.1. (B) An example of a polymorphism in piRNA cluster activity in an annotated cluster on chromosome 2L (23328000..23337026). Abundant piRNAs are detected from strain 21183, whereas strain 21291 produces few piRNAs. Only uniquely mapping piRNAs are considered. piRNA density is measured in Reads Per Million mapped reads (RPM) and transformed to $\log_2$ scale ($\log_2$ (RPM + 1)). Positive value represent piRNAs mapped to sense strand of the reference genome, and negative value represent piRNAs from antisense strand.

3.2.2 Most North American genotypes have *P*-elements in ancestral piRNA clusters

To identify *P*-element insertions in extant populations, we took advantage of the published genomes from the DGRP, which includes more than 200 fully sequenced inbred lines collected from North America in 2003 (Mackay et al. 2012). All DGRP

genomes are known to harbor *P*-elements (Zhuang et al. 2014; Rahman et al. 2015) (Figure S3.1) and the majority of them are expected to repress *P*-element activity (Kidwell et al. 1983; Anxolabéhère et al. 1988; Kidwell 1983). Although previous annotations suggest that less than 20% DGRP genomes have *P*-elements in ancestral piRNA clusters (based on 32 annotated piRNA clusters) (Zhuang et al. 2014; Rahman et al. 2015), we suspected that this was a gross underestimate, because the common requirement for unique read alignment to the reference genome prohibits the identification TE insertions in repeat-rich piRNA clusters. This is particularly problematic for identifying TE insertions in telomeric associated sequence (TAS), piRNA clusters comprised of subtelomeric satellite repeats (Yin and Lin 2007; Asif-Laidin et al. 2017). Indeed, ~50% of wild-derived genomes are believed to harbor a *P*-element insertion in *X*-TAS (Ajioka and Eanes 1989; Ronsseray et al. 1989; Biémont et al. 1990), making the identification of *P*-element insertions in this piRNA cluster of particular significance. We therefore developed an alternative approach to annotate *P*-elements among DGRP genomes.

First, we annotated *P*-element insertion sites based on high-quality alignments of split reads (mapping quality score, MAPQ ≥ 20), which are not necessarily unique, yet still support a particular genomic location with high confidence. Including high-quality non-unique alignments increases the number of annotated *P*-elements by 44% and 24% increase when compared to TEMP and TIDAL, respectively, two approaches that rely on unique alignments (Zhuang et al. 2014; Rahman et al. 2015; Figure 3.2A). While we did not validate these new

insertions, the two additional insertions we identified in DGRP492 were also detected by previous study using hemi-specific PCR, indicating they are true insertions (Zhang and Kelleher 2017).

Despite relaxing the requirement for unique alignment, we still identified only 9 DGRP genomes (4.5%) with *P*-elements in *X*-TAS. High-quality alignments likely fail to provide a unique insertion site in TAS repeats because highly similar tandem satellite sequences provide multiple equivalent alignments (Figure 3.3A). Therefore, we first sought to detect TAS insertions by identifying *P*-derived reads that aligned only to TAS repeats. We found that the majority of DGRP genomic libraries contain *P*-derived read pairs that align to *X*, *2R* or *3R*-TAS (Table S3.2), while only 3 DGRP genomes contained *P*-derived reads aligning to *2L* and *3L*-TAS.

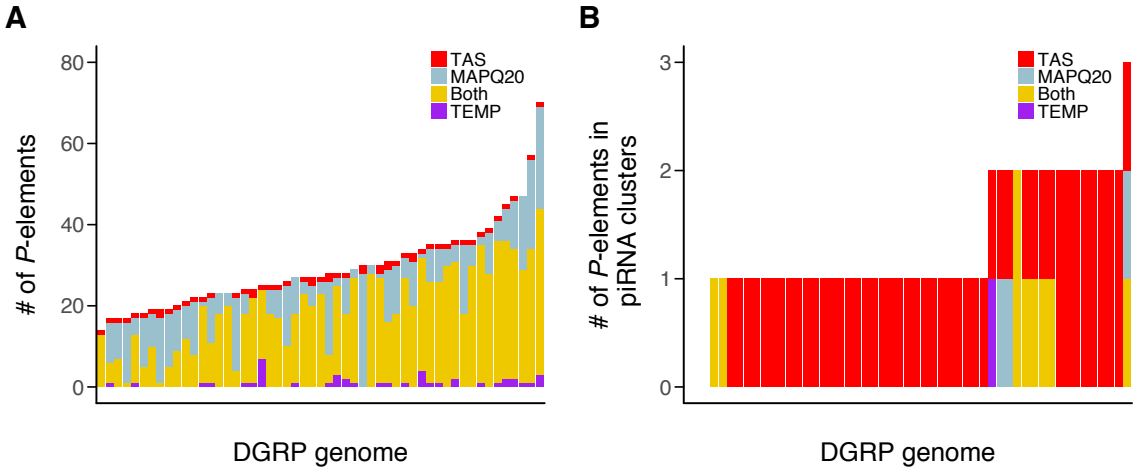

Figure 3.2 (A) Total number of *P*-elements and (B) number of *P*-elements in ancestral piRNA clusters annotated by different approaches. 53 DGRP genomes that were previously annotated by TEMP are compared. To identifying piRNA cluster insertions, the 32 cluster high confidence set was used. TEMP: insertions only found by TEMP; MAPQ20: insertions only found based on high-quality mapping; Both: insertions found by TEMP and MAPQ20; TAS: insertions only found when homologous TAS sequences were treated as a single locus.

To estimate the number (0, 1, >1) of *P*-elements in *X*, *2R* and *3R*-TAS for each DGRP line, we took advantage of the distribution of the number of read pairs supporting individual insertions outside of TAS from the same genome. We then calculated a *Z*-score for the number of *P*-derived reads mapped to TAS. Using this approach we identified 11 DGRP genomes that harbor no *P*-element insertions (6%, $Z < -1.96$), 137 DGRP genomes that harbor one *P*-element insertion (70%, $-1.96 < Z < 1.96$), and 47 genomes that carry two or more insertions into TAS arrays (24%, $1.96 < Z$) (Figure 3.3B; Table S3.2). Given that TAS arrays are ancestral piRNA clusters that are active in *P*-element free strains (Figure 3.1A; Brennecke et al. 2007; Yin and Lin 2007), our observations reveal that the majority of DGRP genomes carry repressor alleles that arose by *de novo* insertion into existing piRNA clusters (Figure 3.2B).

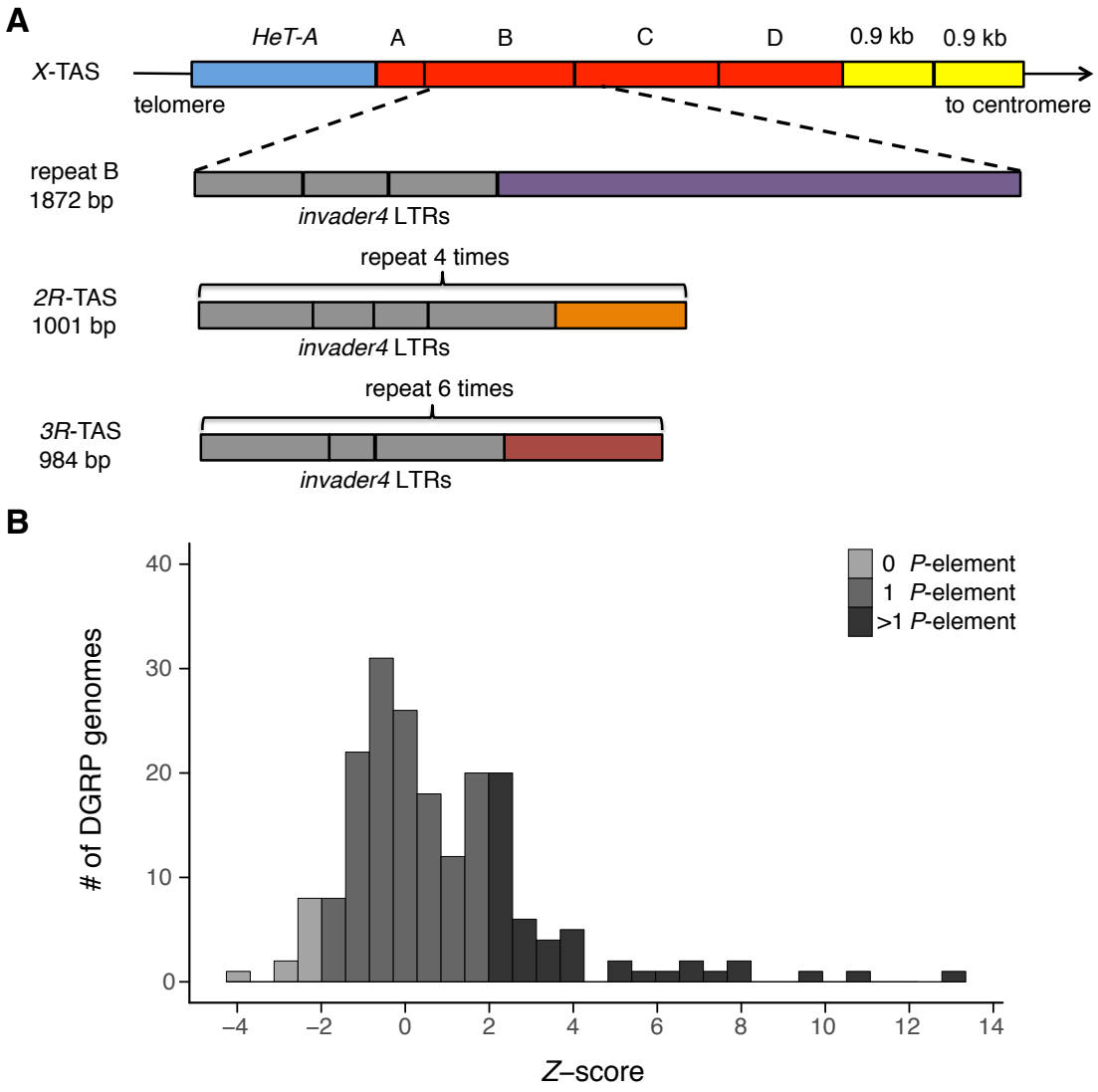

Figure 3.3 (A) The structure of TAS arrays (Asif-Laidin et al. 2017). *X*-TAS contains four tandem repeats (A, B, C and D in red) located between a *HeT-A* retrotransposon and two 0.9 kb repeats (Karpen and Spradling 1992). Repeat A is degenerated. Repeat B, C and D are ~1.8 kb in length and highly similar to each other (>95% identity). Repeat B (enlarged below) is compared to *2R*-chromosome TAS (*2R*-TAS) and *3R*-chromosome TAS (*3R*-TAS), which are represented by 4 and 6 copies in the assembled *D. melanogaster* genome, respectively. Each repeat of *2R*, *3R* and *X*-TAS also contains several subrepeats that are composed of *invader4* retrotransposon long terminal repeats (LTRs)(gray; Bergman et al. 2006). In addition to *invader4* LTRs, the *X, 2R* and *3R*-TAS repeats share other short homologous fragments (41-131 bp) with 909 bp being unique to the *X*-TAS repeat (not shown in the figure; Asif-Laidin et al. 2017). (B) The distribution of *Z*-scores among DGRP genomes. DGRP genomes with $Z < -1.96$, $-1.96 < Z < 1.96$ and $Z > 1.96$ were estimated to have 0, 1 and >1 *P*-element, respectively.

3.2.3 Abundant repressors underpin repressive phenotype

We next sought to isolate individual repressor alleles that arose via *de novo* insertion into TAS arrays. First, we identified the candidate TAS array(s) containing *P*-element insertions in each DGRP genome based on proportion of *P*-derived read pairs whose best alignment supported an insertion in *X*, *2R*, or *3R*-TAS (see methods). For the 82% of DGRP genomes for which we identified at least one candidate TAS array harboring a *P*-element insertion, we further identified the insertion site that was supported by the most read pairs (see methods). In addition, based on alternative breakpoints identified by alignments to TAS sequences, we also determined which of multiple alternate pseudo-genomes, containing *P*-element insertions into different sites, was supported by the most reads (see methods). Due to sequence homology among repeats within the same TAS array (>95% identity; Figure 3.3A), we assumed all homologous insertion sites among tandem repeats corresponded to a single insertion event for these analyses.

Among 92 DGRP genomes, we found 102 *P*-element insertions into TAS where the best insertion site identified by reference genome and pseudo-genome alignments agreed, suggesting well-supported insertion sites. These corresponded to 43 unique insertion sites, 32 of which we were able to verify by site-specific PCR (74.4%). For the remaining 11 insertions, PCR revealed that two were located at different sites, PCR failed for seven sites, and PCR was not attempted for two sites. We further attempted PCR to determine the insertion sites in 14 DGRP genomes

where the two computational methods did not agree, and 73 DGRP genomes where *P*-elements could not be assigned to a particular TAS or breakpoints could not be determined due to an absence of split reads. These PCRs determined an additional 40 *P*-element insertion sites in 71 DGRP genomes.

In total, we identified 89 independent insertions of *P*-elements into TAS sequences (*2R*, *3R* or *X*-TAS), 84 of which were verified by PCR in at least one DGRP genome (Table 3.1; Table S3.3). Consistent with previous studies (Ajioka and Eanes 1989; Ronsseray et al. 1989; Biémont et al. 1990), we found that >50% of DGRP genomes had *P*-element insertions in *X*-TAS and ~20% DGRP genomes had *P*-elements in *2R* and *3R*-TAS (Table 3.1; Table S3.3). Moreover, we discovered a multitude of insertion alleles in each TAS array: 20 in *2R*-TAS, 19 in *3R*-TAS and 50 in *X*-TAS (Table 3.1; Figure 3.4A, B; Table S3.3).

Table 3.1 *P*-element insertions in TAS

TAS array	# of genomes with insertions	# of alleles
2R-TAS	33	20 (20)
3R-TAS	38	19 (18)
X-TAS	133	50 (46)
2L and *3L*-TAS	1	1 (0)

Numbers in parentheses indicate PCR verified alleles.

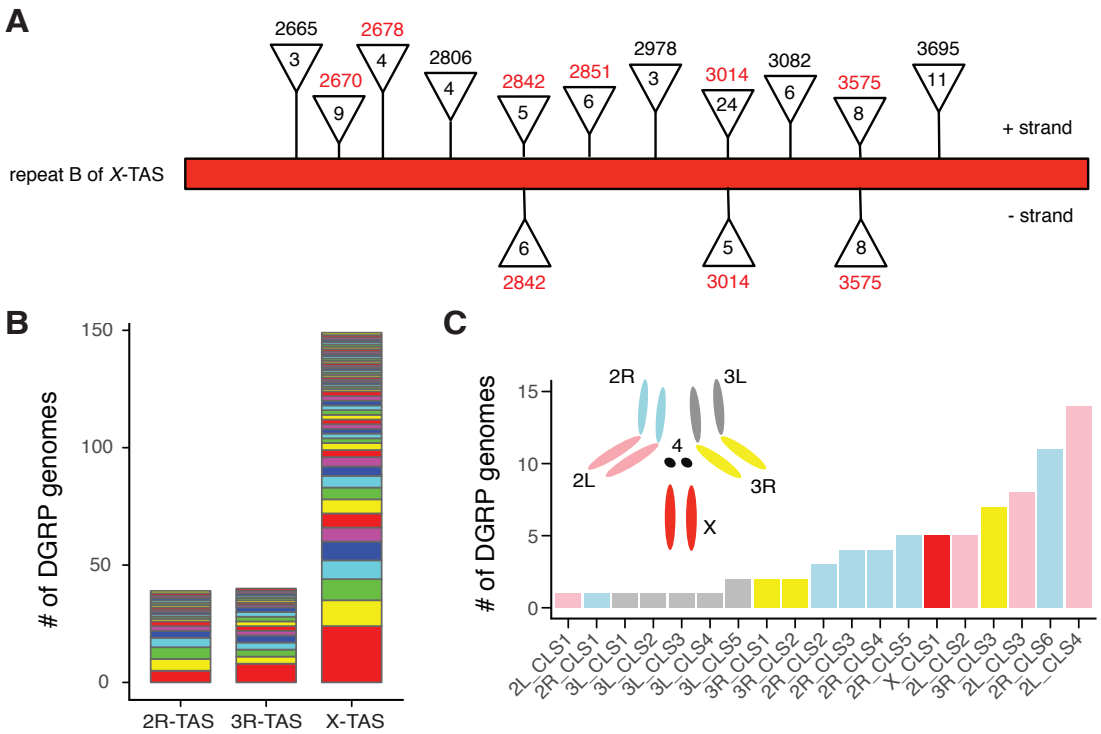

Figure 3.4 (A) Multiple *P*-element insertion sites were detected in *X*-TAS. Due to the high sequence similarity between repeats B, C and D of *X*-TAS, we arbitrarily assigned all *P*-element insertions in these repeats to repeat B. Only insertions present in more than 2 DGRP genomes are depicted. For each insertion, the breakpoint is labeled above the triangle and the number of DGRP genomes containing the insertion is indicated inside the triangle. Breakpoints positions in red correspond to insertion hotspots (Karpen and Spradling 1992). Sense *P*-elements, whose orientation is same as *X*-TAS were drawn above repeat B, whereas antisense *P*-elements were located below repeat B. (B) Multiple *P*-element insertion sites were detected in *2R*, *3R*, and *X*-TAS. Each color represents a unique *P*-element insertion. (C) *P*-elements are present in non-TAS ancestral piRNA clusters across all major chromosome arms, based on our annotation set of 159 piRNA clusters.

P-elements preferentially insert into sequence-specific sites in *X*-TAS (Karpen and Spradling 1992), which are also found in *2R* and *3R*-TAS. We therefore wondered whether *P*-element insertions into these hotspots were unusually common in among TAS insertion alleles from natural populations. Indeed, we found that these sites were greatly enriched for *P*-element insertion alleles: 88.2% (15 out of 17) of hotspots had a *P*-element insertion allele, as compared to only 1.5% (58

out of 3840) of non-preferred sites (Pearson's Chi-square test, $\chi^2 = 639.65$, P-value < 2.20×10^{-16}). Additionally, hotspots were more likely to have two distinguishable insertion alleles, one in each strand, when compared to non-preferred sites (Figure 3.5A; Pearson's Chi-square test, $\chi^2 = 18.92$, P-value = 1.36×10^{-5}). Finally, individual *P*-element insertions in hotspots were more common among DGRP chromosomes than those occurring at non-preferred sites (Figure 3.5B; Wilcoxon rank sum test, $W_{25, 64} = 1259$, P-value = 5.88×10^{-6}), suggesting that recurrent insertion into these positions elevates the frequency of these insertion alleles. Taken together, these results suggest that the exceptional abundance of *P*-elements insertions in TAS arrays is at least partially explained by an insertion site preference.

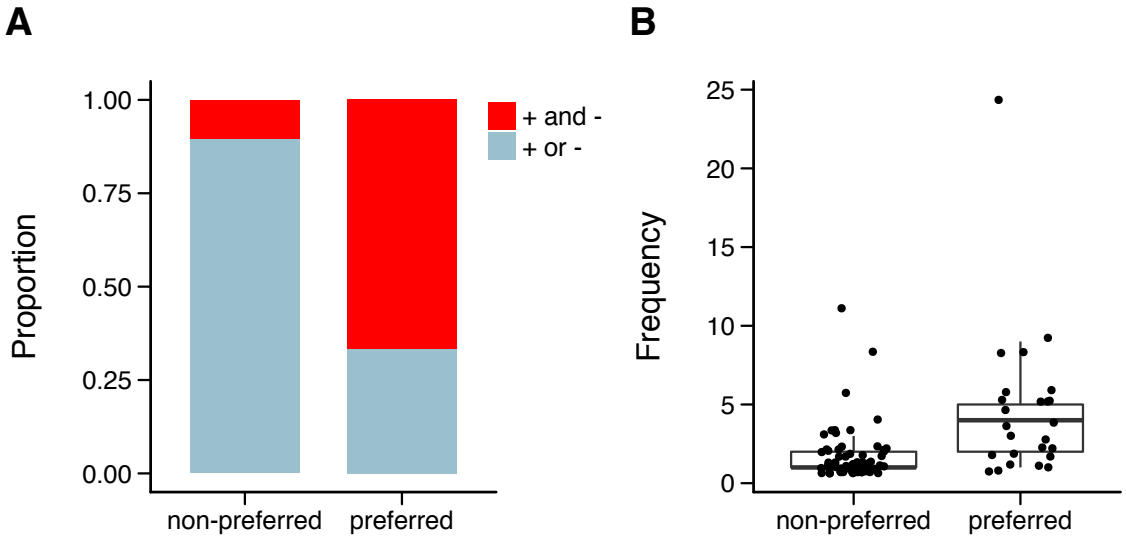

Figure 3.5 (A) The proportion of *P*-element insertions that insert into both sense and antisense strand of TAS (red), or single strand (blue) for non-preferred and preferred sites. (B) The comparison between frequencies of *P*-elements at non-preferred and preferred sites. *P*-elements inserted at same site but in opposite orientations were considered different insertions.

3.2.4 Cluster *P*-element insertions are targets of positive selection

Combining the TAS insertion alleles with those identified in non-TAS piRNA clusters, we detected up to 193 *P*-element insertion events into at least 15 (up to 37) different ancestral piRNA clusters, which are located on all of the major chromosome arms of the *Drosophila* genome (Figure 3.4C). To determine whether *P*-element insertions in ancestral piRNA clusters are targets of positive selection, we considered their site frequency spectrum. Positive selection is expected to increase the frequency of beneficial alleles in natural populations when compared to neutral alleles (Nielsen 2005). However, this observation is potentially confounded by the recurrent insertion of *P*-elements into known insertion hotspots in TAS sequences, which elevates their frequencies (Figure 3.5B). We therefore excluded *P*-element insertions in hotspots from our analysis of the site frequency spectrum. Consistent with positive selection, we found that *P*-element insertions in ancestral piRNA clusters segregate at higher frequencies than those outside of piRNA clusters (Figure 3.6A).

piRNA clusters occur predominantly in heterochromatic regions of low recombination (Brennecke et al. 2007), meaning that differences in the site-frequency spectra of cluster and non-cluster *P*-element insertions are also potentially confounded by differences in nature and efficacy of selection in different genomic compartments (Hill and Robertson 1966; Haddrill et al. 2007). In particular, heterochromatic insertions are less likely to disrupt functional sequences or participate in ectopic recombination, making them less deleterious than those in

euchromatin (Bartolomé and Maside 2004; Petrov et al. 2011; Kofler et al. 2012). However, when we restrict our comparison of cluster and non-cluster insertions to regions of low recombination where heterochromatin resides (≤1 cM/Mb) we observe that the elevated frequency of cluster insertions becomes more pronounced (Figure 3.6B; Figure S3.3B). This suggests that the relatively higher frequency of *P*-element insertions in ancestral piRNA clusters reflects positive selection, rather than reduced purifying selection against TE insertions in heterochromatin. Furthermore, the difference in frequency spectra between cluster and non-cluster *P* element insertions decreases when we include insertions in lower confidence piRNA clusters (Figure S3.3, S3.4), suggesting that the inclusion of false-positives (*i.e.,* insertions into incorrectly annotated piRNA clusters) dampens the signature of positive selection. Alternatively, this pattern could result from stronger positive selection for highly-expressed piRNA clusters, which are over-represented among high-confidence piRNA clusters.

A second important distinction between selection in heterochromatin and euchromatin lies in the efficacy of positive selection, which is reduced in regions of low recombination by linked deleterious variation ('Hill-Robertson effects': Hill and Robertson 1966). Although the sample size of piRNA cluster insertion in regions of high recombination (>1 cM/Mb) is small (n = 20 for high confidence piRNA clusters), we did not observe that they segregate at higher frequencies than non cluster insertions (Figure 3.6C; Figure S3.3C, S3.4C). While these observations are not consistent with Hill-Robertson effects, they are consistent with the theoretical

observation that reduced recombination increases the efficiency of selection on TE repressors by maintaining linkage between repressor alleles and the chromosomal sites they've protected from mutational load (Charlesworth and Langley 1986).

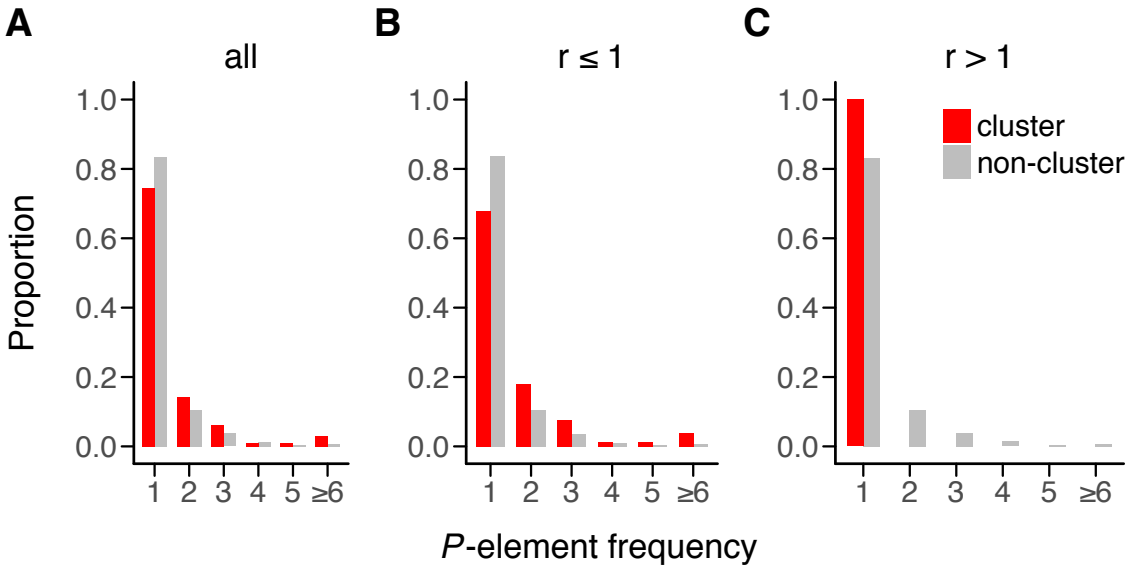

Figure 3.6. The frequency of *P*-elements in piRNA clusters (red) and the frequency of *P*-elements outside of clusters (gray) are compared for the high confidence set of 32 annotated piRNA clusters. The equivalent comparisons for lower stringency annotation sets are provided in figures S3 and S4. *P*-elements at hotspots were excluded from these comparisons. Insertions are compared in (A) all genomic regions ($W_{98, 4850}$ = 215720, P-value = 0.016), (B) regions of low recombination (≤1 cM/Mb, $W_{78, 1501}$ = 48966, P-value = 0.00019) and (C) high recombination (>1 cM/Mb, $W_{20, 3349}$ = 39130, P-value = 0.045).

3.3 Discussion

In this study, we took advantage of the recent invasion of the *Drosophila* melanogaster genome by *P*-element DNA transposons to chronicle the evolution of piRNA-mediated repression. We found that ~94% *D. melanogaster* genomes have at least one *P*-element in an ancestral piRNA cluster, suggesting *de novo* mutation, in which *P*-elements transpose into pre-existing piRNA clusters, is the predominant

mutational mechanism giving rise to piRNA-mediated silencing. Furthermore, we uncovered no fewer than 84 repressor alleles, which are targets of positive selection. Taken together, our results reveal that the common phenotype of *P*-element repression exhibited by North American *D. melanogaster* (Ogura et al. 2007; Kidwell et al. 1983; Kidwell 1983) is underpinned by an unprecedented number of beneficial repressor alleles, which have arisen since the *P*-element invasion in the mid 20th century.

The existence of numerous segregating repressor alleles indicates that *P*-element repression didn't evolve through a classical "hard sweep", in which a single beneficial mutation arises and then goes into fixation (Maynard Smith and Haigh 1974). Rather, the evolution of *P*-element repression in *Drosophila* melanogaster occurred through a plethora of "soft sweeps" (Pennings and Hermisson 2006), in which numerous repressor alleles arose and increased in frequency simultaneously. Indeed, to our knowledge the evolution of *P*-element repression represents one of the most striking examples of soft-sweeps in a eukaryotic genome. By comparison, other well-known examples such as insecticide resistance in *D. melanogaster* (Menozzi et al. 2004; Karasov et al. 2010) and lactose tolerance in human populations (Enattah et al. 2002; Tishkoff et al. 2007), include only 4 and 5 adaptive mutations, respectively. The extreme soft-sweeps we observe in the evolution of *P*-element repression are at least partially a consequence of the unique genetic architecture of piRNA mediated silencing. The presence of multiple, functionally redundant piRNA clusters, which will enact repression when occupied by *P*-

elements, provides an exceptionally large mutational target comprising at least 0.3% of the genome. Polygenic traits with large mutational targets are predicted to evolve via soft sweeps, because the overall beneficial mutation rate is increased (Pritchard et al. 2010; Messer and Petrov 2013; Pennings and Hermisson 2006; Karasov et al. 2010). Similarly, the per-site beneficial mutation rate within each piRNA cluster is also high, owing to a very high genome-wide transposition rate of *P*-elements (~0.1 new insertions per element multiplied by genomic copy number (Eggleston et al. 1988; Berg and Spradling 1991; Kimura and Kidwell 1994)).

In addition to documenting an abundance of putatively beneficial alleles, we discovered that positive selection predominantly acts on *P*-element insertions in heterochromatic piRNA clusters. This is consistent with the theoretical prediction that reduced recombination enhances positive selection on TE repressors by maintaining linkage to the genomic regions they've protected from deleterious insertions (Charlesworth and Langley 1986). Indeed, enhanced selection on repressors in low recombination regions might explain why piRNA clusters are predominantly located in heterochromatic regions (Blumenstiel 2011). Interestingly, however, our results differ from those of Lu and Clark, who show that piRNA cluster insertions of resident TE families that have occupied the genome for a long time segregate at higher frequency than those outside of piRNA clusters only in regions of high recombination (Lu and Clark 2010). The nature of the discrepancy is unclear, however, recent simulation models suggest that dynamics of piRNA

mediated repressor alleles can differ significantly between recently invaded TEs and those that are at copy number equilibrium (Kelleher et al. 2018; Kofler 2019).

P-elements are not randomly distributed among piRNA clusters in the *D. melanogaster* genome. Rather, 72.4% of them (based on 32 annotated piRNA clusters) occurred in TAS regions, consistent with previous studies that detected *P*-elements insertions using hybridization-based approaches (Ronsseray et al. 1991; Marin et al. 2000; Stuart et al. 2002; Brennecke et al. 2008). Our observations mirror those of a recent study of the evolution *P*-element repression in laboratory populations of *D. simulans,* which is established by multiple independent *P*-element insertions in piRNA clusters, particularly in the *3R*-TAS (Kofler et al. 2018). Furthermore, we discovered that *P*-elements are most commonly observed in previously identified insertion hotspots (Karpen and Spradling 1992), thereby demonstrating that this mutation bias shapes the distribution of *P*-element insertions in natural populations. TE insertions in TAS were likely not detected among DGRP genomes previously because the reliance on unique alignments excludes read pairs supporting insertions in satellite arrays (Linheiro and Bergman 2012; Zhuang et al. 2014; Rahman et al. 2015). Therefore, allowing for multiple mapping within highly homologous satellite repeats represents a powerful method for annotating TEs in these regions from short paired-end reads.

In summary, *P*-element repression in *Drosophila melanogaster* evolved rapidly though abundant *de novo* mutations that arise from the transposition of *P*-elements into pre-existing piRNA clusters. These concurrent beneficial alleles are

targets of positive selection, resulting a striking example of polygenic adaption. As piRNA-mediated silencing of TEs is conserved across animals, the model in which rapid adaptation to *P*-element invasion evolves through multiple beneficial *de novo* mutations applies to other TEs. Our observations reveal how the unique genetic architecture of piRNA-mediated silencing, in which insertion into multiple functionally redundant piRNA clusters results in a repressor allele, facilitates the evolution of repression of an invading TE.

3.4 Materials and Methods

3.4.1 DGRP stocks and genomes

All DGRP lines were ordered from the Bloomington Drosophila stock center.

3.4.2 piRNA cluster annotation

Ovarian small RNA sequencing libraries were downloaded from NCBI or were generated by our lab for another project (Lama and Kelleher unpublished, Table S3.1). The latter libraries are available from SRA archive (SRP160954). For each library, adapters were trimmed using cutadapt (version 1.9.1) (Martin 2011). Trimmed reads with 23 – 29 nt (typical size of piRNAs in *Drosophila*) were kept for piRNA cluster annotation. Then, piRNA clusters were predicted separately in each library using proTRAC (Rosenkranz and Zischler 2012), which identifies genomic loci corresponding to piRNA clusters based on the density of mapped piRNAs. We considered different values of the proTRAC pdens parameter (0.01, 0.05, 0.1), with lower pdens values corresponding to annotation sets that include a smaller number of higher confidence piRNA clusters. Annotated piRNA clusters detected less than 5 kb apart were considered a single cluster.

3.4.3 Detecting *P*-element insertions in DGRP genomes

DGRP whole genome sequencing reads were downloaded from the NCBI Sequence Read Archive (Mackay et al. 2012; Huang et al. 2014). 12 DGRP genomes were excluded from our analysis because 45 bp paired-end reads (DGRP357, DGRP379,

DGRP427, DGRP486, DGRP786), or 75 bp single-end reads (DGRP153, DGRP237, DGRP28, DGRP325, DGRP386, DGRP41, DGRP730) were too short to allow for identification of *P*-element insertion sites. To identify read pairs that include *P*-element sequence in the remaining genomes, individual reads were separately and locally aligned to full-length *P*-element consensus (O'Hare and Rubin 1983) using bowtie2 (v2.1.0) (Langmead and Salzberg 2012) with default parameters. *P*-element sequences were then trimmed from mapped reads using a custom Perl script. Trimmed reads longer than 30 bp were kept and used for down-stream analyses.

For each DGRP genome, the *P*-derived trimmed reads were first aligned to the *D. melanogaster* release 6 reference genome (dm6: Hoskins et al. 2015) as well as *X*-TAS (Karpen and Spradling 1992) using bowtie2. Reported alignments with mapping quality score greater than 20 and a mutational distance (sum of mismatches and gaps required to convert the read sequence to the reference) less than four were kept. Only the number of gaps was considered: we ignored gap extensions assuming that two adjacent nucleotide insertions or deletions were generated by one mutational event. To isolate breakpoints corresponding to *P*-element insertion sites, we took advantage of split reads, in which one segment aligned to the *P*-element consensus and the remainder aligned to the reference genome. After breakpoints were located, all non-split *P*-derived read pairs (*i.e.* one read aligns to *P*-element, its mate to the reference genome) within 500 bp were identified. At least 6 supporting read pairs (split or non-split) were required to annotate a single *P*-element insertion in non-TAS regions.

3.4.4 Detecting *P*-element insertions in TAS

We divided the dm6 reference genome into two parts: TAS and non-TAS regions. TAS regions included full-length of *X*-TAS (9872 bp, L03284) (Karpen and Spradling 1992), 2*R*-TAS (chr*2R*:25258060..25261551, 3492 bp) and *3R*-TAS (chr*3R*:32073015..32079331, 6317 bp) (Yin and Lin 2007), *2L*-TAS (chr*2L*:1..5041, 5041 bp) and *3L*-TAS (chr*3L*:1..19608, 19608 bp)(Walter et al. 1995). The other genomic regions were categories as non-TAS.

To determine if *P*-derived reads that did not map to the non-TAS regions corresponded to insertions in TAS, they were aligned to the TAS reference using bowtie2 outputting all valid alignments (-a). A read pair was considered mapped to TAS if the mutational distance was fewer than six for paired-end reads and four for single-end reads. For each DGRP genome, we calculated a *Z*-score for TAS-aligned reads according to the formula: $Z = (x - \mu) / \sigma$, where x is the number read pairs aligned to *X*, *2R* or *3R*-TAS, μ is the average number of reads supporting individual non-TAS *P*-element insertions in a given genome, and σ is the standard deviation for reads supporting non-TAS insertions. A significance level $\alpha = 0.05$ ($Z = \pm 1.96$) was used to estimate the number of *P*-elements in TAS in each DGRP genome (Table S3.2).

To determine which TAS arrays (*X*, *2R* or *3R*-TAS) contained a *P*-element insertion in each DGRP genome, we first calculated mutational distance for all reported alignments of each read pair in that genome. We then assigned each read

pair to the TAS array that it aligned to with lowest mutational distance. For DGRP genomes with one *P*-element in TAS ($-1.96 < Z < 1.96$), the insertion was predicted to occur in the TAS array whose supporting reads were at least 2 times greater than the reads supporting the other two TAS arrays. For DGRP genomes with more than one *P*-element in TAS ($1.96 < Z$), we sought to determine the locations of two *P*-elements. The first insertion was predicted to occur in the TAS assay supported by the highest number of reads. Then, we subtracted the average number of reads supporting a non-TAS *P*-element insertion in the given DGRP genome from the reads supporting the first TAS insertion. The second insertion was predicted the same way as DGRP genomes with one *P*-element.

3.4.5 Localizing insertion sites of *P*-element insertions in TAS

A read pair may be equally-well aligned to several homologous satellite repeats within a TAS array. Therefore, for *2R* and *3R*-TAS, we assigned *P*-elements to consensus sequences, as their repeats are indistinguishable from each other. Similarly for *X*-TAS, we were unable to determine whether a given insertion occurred in repeat B, C, or D, so we arbitrarily assigned all insertions to repeat B. We then identified the insertion breakpoint supported by the most split reads.

As an alternative approach, we also constructed pseudo genomes for each alternative TAS insertion site in a given DGRP genome, which included the *P*-element consensus sequence flanked at each end by an 8 bp target site duplication and 500 nt of adjacent TAS sequence. Paired-end reads were aligned to the

constructed pseudo genomes (MAPQ > 10), and the breakpoint corresponding to the pseudo genome with the most reads aligned was identified.

3.4.6 PCR verification of insertion sites

Genomic DNA was extracted using the QIAGEN DNeasy Blood & Tissue Kit (Cat. No. 69506) or a squish prep (Srivastav and Kelleher 2017). To determine the *P*-element insertion sites, a *P*-element specific and a TAS specific primer were used (Table S3.4). As multiple bands were generally produced, owing to alternative annealing of the TAS primer to multiple repeats, the main band was purified by gel extraction using the QIAGEN MinElute Gel Extraction Kit (Cat. No. 28606), and sequenced to determine the breakpoint.

3.4.7 Recombination rates

Recombination rates at *P*-element insertions sites were identified from the genome-wide map provided by Comeron *et al.* (Comeron et al. 2012). Because these rates were based on the release 5 of *D. melanogaster* reference genome, we converted our annotated *P*-element insertions in release 6 coordinates to release 5 on the Flybase (http://flybase.org). The recombination rate of insertions that didn't have release 5 counterparts was assumed to 0, because the major improvement of release 6 relative to release 5 is the assembly of heterochromatin regions (Dos Santos et al. 2015; Hoskins et al. 2015).

3.4.8 Data analysis

Annotating piRNA clusters and identifying *P*-element insertions were powered by the high performance computing resources from the Center for Advanced Computing and Data Science (CACDS) at the University of Houston (http://www.uh.edu/cacds/resources/hpc/). All statistical analyses were performed in R (version 3.3.1)(R Core Team 2016). Graphs were made in RStudio (RStudio Team 2015) with R packages ggplot2 (version 2.2.1)(Wickham 2017b), gplots (version 3.0.1)(G. W. Warnes, B. Bolker 2016), reshape2 (version 1.4.3)(Wickham 2017a), and cowplot (version 0.7.0)(Wilke 2017).

Chapter 4 Overall conclusions and discussion

4.1 Overall conclusions

TEs are generally considered genetic parasites because the majority of TEs are deleterious to hosts, producing deleterious mutations and DNA damage. The deleterious nature of TE insertions is evidenced by their scarcity in gene-rich genomic regions and their low frequency in natural populations (Kofler et al. 2012; Cridland et al. 2013). In animals, TEs are silenced by the piRNA pathway, in which small piRNAs derived from TE-enriched loci (piRNA clusters) suppress complementary TEs by base pairing (Brennecke et al. 2007; Aravin et al. 2007b; Girard and Hannon 2008; Slotkin et al. 2009; Bhaya et al. 2011). In this dissertation, I took advantage of the recent *P*-element invasion into the *D. melanogaster* genome to study the emergence and dynamics of piRNA-mediated repressor alleles in natural populations.

As *P*-element repressor alleles are *P*-element insertions located in piRNA clusters, I first adapted a targeted re-sequencing strategy using a combination of *P*-element specific primers and a set of degenerate primers (hemi-specific PCR) and developed a computational pipeline to identify *P*-element insertions in the *D. melanogaster* genome. I found that the targeted re-sequencing approach provides an efficient way to annotate *P*-element insertions, even in repetitive genomic regions where piRNA clusters reside (Brennecke et al. 2007). In addition, I demonstrated that hemi-specific PCR accurately determine *P*-element insertion breakpoints so that *P*-element insertion frequencies could be estimated precisely from the

sequencing data of pooled samples, which is valuable for studying the population dynamics of TEs.

Although I decided not to use hemi-specific PCR to detect *P*-element insertions in wild-derived DGRP genomes (Mackay et al. 2012), the computational pipeline I developed empowered me to re-annotate *P*-elements from existing whole genome re-sequence data of those strains. In particular, I focused on annotating *P*-element insertions in TAS, an important piRNA cluster in *P*-element regulation (Brennecke et al. 2007; Yin and Lin 2007; Stuart et al. 2002; Marin et al. 2000). I found that >90% of DGRP genomes have at least one *P*-element in TAS and therefore the majority of DGRP genomes have *P*-element insertions in ancestral piRNA clusters that were active before the *P*-element invasion. This suggests that *de novo* mutation, in which *P*-elements transpose into pre-existing piRNA clusters, is the predominant mechanism for the origin of *P*-element repressor alleles. Moreover, I found more than 84 independent *P*-element repressor alleles that are under positive selection, indicating the evolution of *P*-element repression is a process of polygenic adaptation.

4.2 Discussion

4.2.1 The application of hemi-specific PCR

The targeted re-sequencing strategy can be applied to broad TE-related studies. Revealing the genetic variation that underlies phenotypic diversity is one of central goals of biology. TEs, as a mutation inducer, greatly contribute to genetic variation,

ranging from large chromosomal rearrangements to small insertions and deletions. For example, TE insertions can account for ~70% of structural variants in inbred mouse strains (Quinlan et al. 2010). In addition, human populations are estimated to have ~2000 TE insertion polymorphisms (Bennett et al. 2004) and approximately 10% of insertions and deletions in human (> 100 bp) are caused by TEs (Xing et al. 2009). These TE-induced variants have broad phenotypic effects on host, including contributing to human diseases (Kazazian et al. 1988; reviewed in Chénais 2013), fueling adaptation to changing environments (Aminetzach et al. 2005; Daborn et al. 2002; Hof et al. 2016; Ishikawa et al. 2019; Niu et al. 2019), and facilitating speciation (reviewed in Serrato-Capuchina and Matute 2018). Therefore, the use of hemi-specific PCR to detect TE insertions lays foundation for uncovering TE-induced variants, their biological consequences on the host, and the underlying molecular mechanisms.

4.2.2 Soft sweeps in evolution of host resistance to invading TEs

The rapid adaptation to *P*-element invasion occurs through soft selective sweeps. In the process, multiple adaptive mutations (*i.e.*, *P*-elements in piRNA cluster) originate in the population concurrently and sweep to higher frequencies compared to neutral *P*-element insertions. This contrasts to hard selective sweeps, in which a single beneficial mutation rises and reaches fixation before another beneficial mutation occurs due to low adaptive mutation rates (Maynard Smith and Haigh 1974). Soft sweeps are more likely to occur when adaptive mutation rate is high

(Karasov et al. 2010; Messer and Petrov 2013). Therefore, *P*-element repression evolved through soft sweeps can due to a high mutation rate of piRNA-mediated repression, which is ultimately attributed to the following factors (reviewed in Kelleher 2016): 1) the genetic redundancy of piRNA clusters (~3.5% of the *D. melanogaster* genome based on Brennecke et al. 2007); 2) higher transposition rate of *P*-elements (10^{-3} – 10^{-1} new insertions per element; Eggleston et al. 1988; Berg and Spradling 1991; Kimura and Kidwell 1994) compared to per site mutation rate of *D. melanogaster* (~10^{-10}; Haag-Liautard et al. 2007); 3) the insertion bias of *P*-elements into TAS piRNA clusters.

Soft sweeps might be the dominant mode for the evolution of resistance to invading TEs due to elevated adaptive mutation rate. In animals, piRNAs are derived from piRNAs clusters in order to silence TEs (Brennecke et al. 2007; Houwing et al. 2007; Aravin et al. 2007b; Jehn et al. 2018). Similarly, 24-nt siRNAs are proposed to be generated from pericentromeric and TE-enriched islands in order to regulate TEs (reviewed in Sigman and Slotkin 2016). The existence of functionally redundant small RNA-producing loci increases the mutation rate to repressor alleles. Moreover, the high transposition rates of TEs (10^{-5} to 10^{-3}; Nuzhdin and Mackay 1995) similarly accelerates the production of beneficial alleles.

4.2.3 The long-term evolution of *P*-element insertions in piRNA clusters

We observed that the most frequent *P*-element in piRNA clusters occurred at a frequency of 12%, indicating none of *P*-element insertions in piRNA clusters that I

discovered are fixed. Although ~94% of DGRP genomes have at least one *P*-elements in ancestral piRNA clusters, sexual reproduction can still produce individuals without *P*-elements in piRNA clusters due to recombination between cluster *P*-element insertions. Individuals without *P*-elements in piRNA clusters have a lower fitness because they are susceptible to *P*-element activity. Therefore, positive selection will increase the number of *P*-elements in piRNA clusters in each individual to decrease the rate of producing susceptible offspring. It would be interesting to know the average number of *P*-element insertions in piRNA clusters after a long evolutionary time. A forward simulation shows that on average, four cluster *P*-element insertions are required in each individual to stop *P*-element invasion (Kofler 2019). However, little is known for the long-term evolution of cluster *P*-elements in natural populations. Most TEs have colonized host genomes for a long period. Studying those old TE invasions could shed light on the long-term evolution of *P*-element insertions in piRNA clusters.

4.2.4 The role of epigenetic mutation in the evolution of *P*-element repression

Although *de novo* mutation is the predominant mechanism for the evolution of *P*-element repression, I cannot rule out the role of epigenetic mutation in the process. First, for the >90% of DGRP genomes with *P*-elements in ancestral piRNA clusters, there might be epimutated *P*-element insertions in those genomes as our approach is biased to detect *de novo* mutations. Therefore, *de novo* and epigenetic mutations may act synergistically to regulate *P*-element activity. Moreover, I found 6 DGRP

genomes have no *P*-elements in ancestral piRNA clusters. Although there may be cluster *P*-elements failed to be detected in those genomes, it is possible that some *P*-elements were converted to novel piRNA-producing loci by add repressive heterochromatin marks. One future direction is to map the *P*-element insertions that are converted to novel piRNA clusters in DGRP genomes without *P*-element in ancestral piRNA clusters.

4.2.5 The evolution of *P*-element repression among *D. melanogaster* populations in other geographic regions

P-elements invaded *D. melanogaster* genome in Florida, and then spread into America and other continents (Kidwell 1983). Therefore, the DGRP genomes, which were collected in North American around 2003 (Mackay et al. 2012), represent later stage of *P*-element invasion into *D. melanogaster* genome compared to populations in other geographic regions. It would be interesting to know if *de novo* mutation is also the predominant mechanism for the evolution of *P*-element repression in populations on other continents, which represent the earlier stage of *P*-element invasion. Moreover, *P*-element-induced hybrid dysgenesis is severe at high temperature (*e.g.,* 29 ˚C), whereas hybrid dysgenesis is inhibited at low developmental temperature (*e.g.,* 18 ˚C)(Kidwell et al. 1977). Hence, selective pressure for host repressors may differs between *D. melanogaster* populations in different latitudes. It remains unknown whether repressors also evolved adaptively in all natural populations. The abundant sequenced wild-derived strains across

diverse geographic regions, including Africa, Europe, Australia, and Aisa (Grenier et al. 2015; Lack et al. 2015; Pool et al. 2012; Kofler et al. 2012), provide a unique opportunity to address those questions.

4.2.6 The invasion of *P*-elements into *Drosophila simulans*

P-elements recently invaded into the *D. simulans* genome by horizontal transfer from *D. melanogaster* around 2006 (Kofler et al. 2015a; Hill et al. 2016). In addition, *P*-elements cause hybrid dysgenesis in *D. simulans* and some strains collected after *P*-element invasion have acquired repression (Hill et al. 2016; Yoshitake et al. 2018). Therefore, the recent invasion of *P*-element into *D. simulans* genome also offers an opportunity to study the evolution of *P*-element repression at an early stage of *P*-element invasion. In fact, a study from experimentally evolving *D. simulans* populations shows the establishment of *P*-element repression through *de novo* mutation (Kofler et al. 2018). Moreover, most of the *de novo P*-element insertions in piRNA clusters are located in *3R*-TAS, consistent with the special role of TAS in *P*-element regulation in *D. melanogaster* (Kofler et al. 2018). However, they cannot disentangle *de novo* from epigenetic mutation as ~84% of piRNA-producing *P*-element insertions were not found in their study and those insertions could originate via *de novo* or epigenetic mutation (Kofler et al. 2018).

4.2.7 The size of ancestral piRNA clusters

The actual size of piRNA clusters in *D. melanogaster* remains unknown. Current annotations of piRNA clusters are based on mapping piRNAs to the reference genome. However, little is known about how piRNA clusters are formed, what nucleotide composition or chromatin marks are unique to piRNA clusters. More studies are needed to investigate what defines a piRNA cluster.

The conclusion that *de novo* mutation is the predominant mechanism for the evolution of *P*-element repress is robust regardless of the size of annotated ancestral piRNA clusters. Although we annotated ancestral piRNA clusters from diverse wild-derived *P*-element-free strains (27 small RNA sequencing libraries of 9 strains, Table S3.1), some piRNA clusters may be not captured in our analysis. Moreover, certain piRNA clusters may be false positives, particuarly when we applied a low stringency. However, TAS sequences are always annotated as piRNA clusters no matter how stringency we defined piRNA clusters. Because >90% of DGRP genomes have *P*-elements in TAS, our conclusion is not affected by the size of annotated piRNA clusters.

Appendix

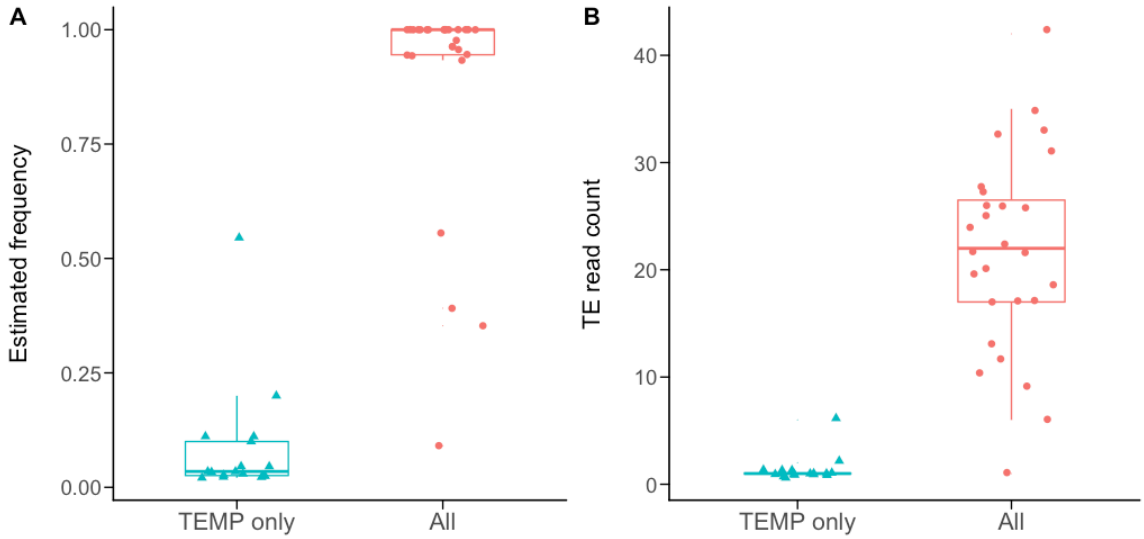

Figure S2.1 The *P*-element insertions found only by TEMP have a low frequency (A) and are supported by few reads (B). *P*-element insertion detected by TEMP (two tail-to-tail insertions are excluded) were divided into two groups: the "TEMP only" group which contains insertions found only by TEMP, and the "All" group which contains insertions found by TEMP, TIDAL and hemi-specific PCR. Frequency (called penetrance by TEMP) was estimated from whole genome sequencing of the RAL-492 line. TE read count indicates the number of paired-end reads that support the inserted chromosome. Insertions annotated only by TEMP have significantly lower frequencies than insertions also detected by TIDAL and hemi-specific PCR (Welch's t_{41} = 14.44, P-value < 2.2 x 10^{-16}). In addition, the number of TE reads supporting *P*-element insertions detected only by TEMP is significantly small (Welch's t_{27} = 11.05, df = 27, P-value = 1.33 x 10^{-11}).

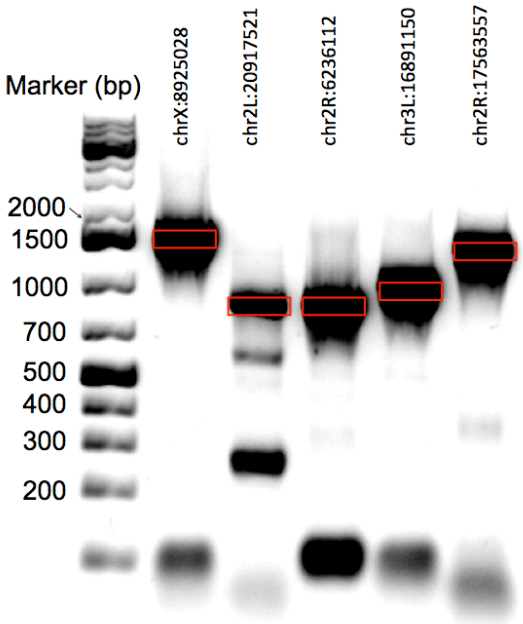

Figure S2. PCR of annotated insertion sites. To verify the existence of some *P*-element insertions, we performed insertion-specific PCR (Table S2.4). In the absence of *P*-element insertion, the length of all PCR products was expected to be less than 500 bp. However, the main bands of PCR products at all locations were more than 500 bp, indicating the existence of insertion events. To determine if they are *P*-elements, we sequenced the PCR products (indicated by red rectangles). Sequencing results indicated the existence of *P*-element at each location. Notably, chr2L:20917521 a polymorphic insertion which is absent from the TEMP and TIDAL annotation.

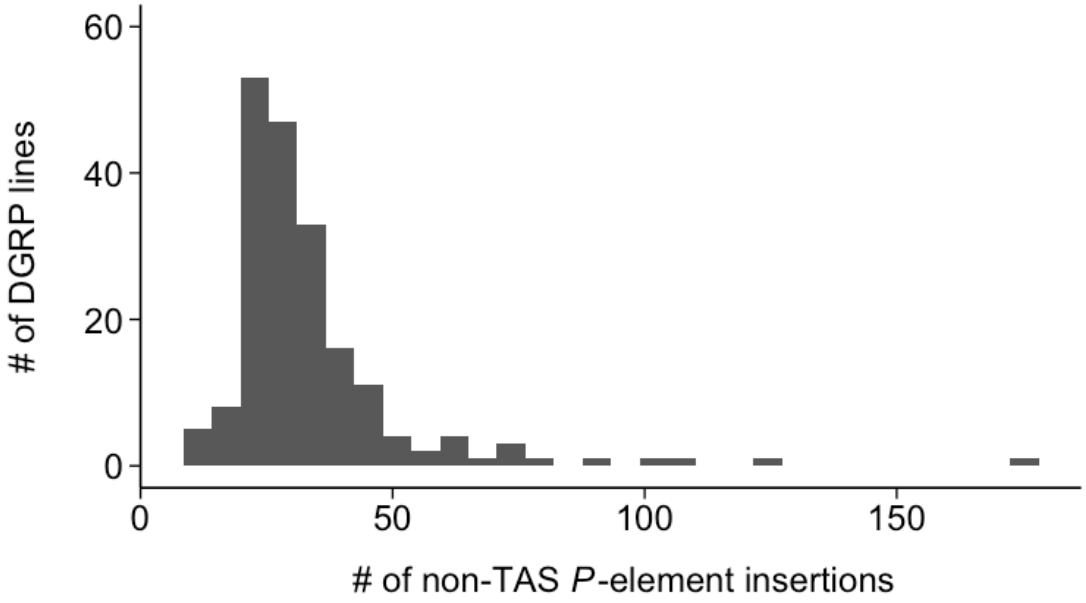

Figure S3.1 All DGRP genomes analyzed harbored *P*-elements in non-TAS regions. 193 DGRP genomes were analyzed.

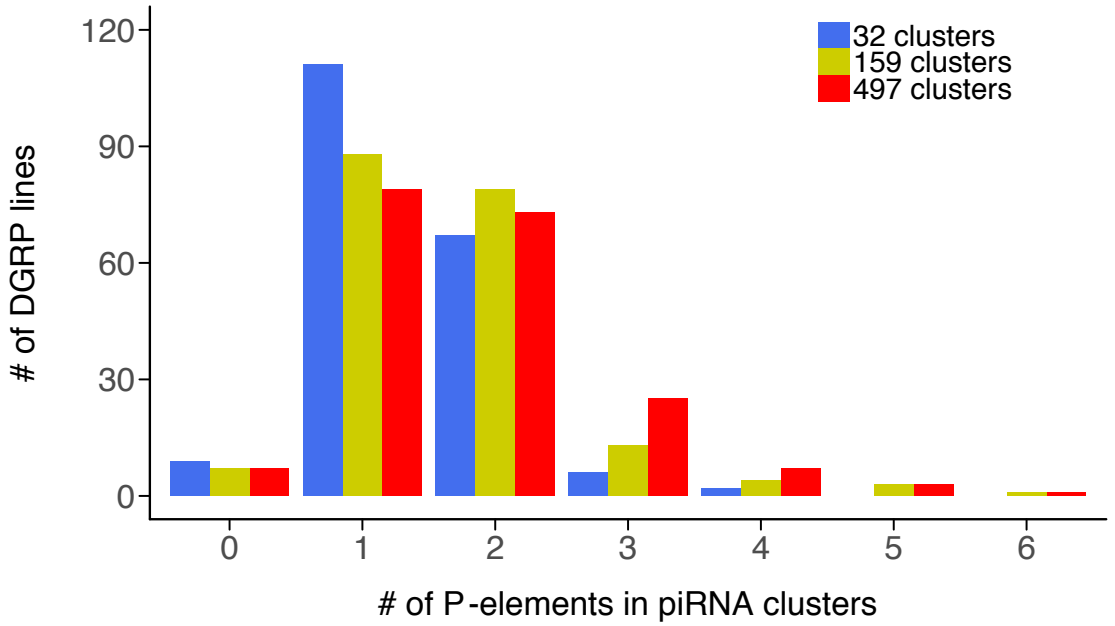

Figure S3.2 No matter how many annotated ancestral piRNA clusters were annotated, more than 90% DGRP lines had at least one *P*-element insertion in ancestral piRNA clusters.

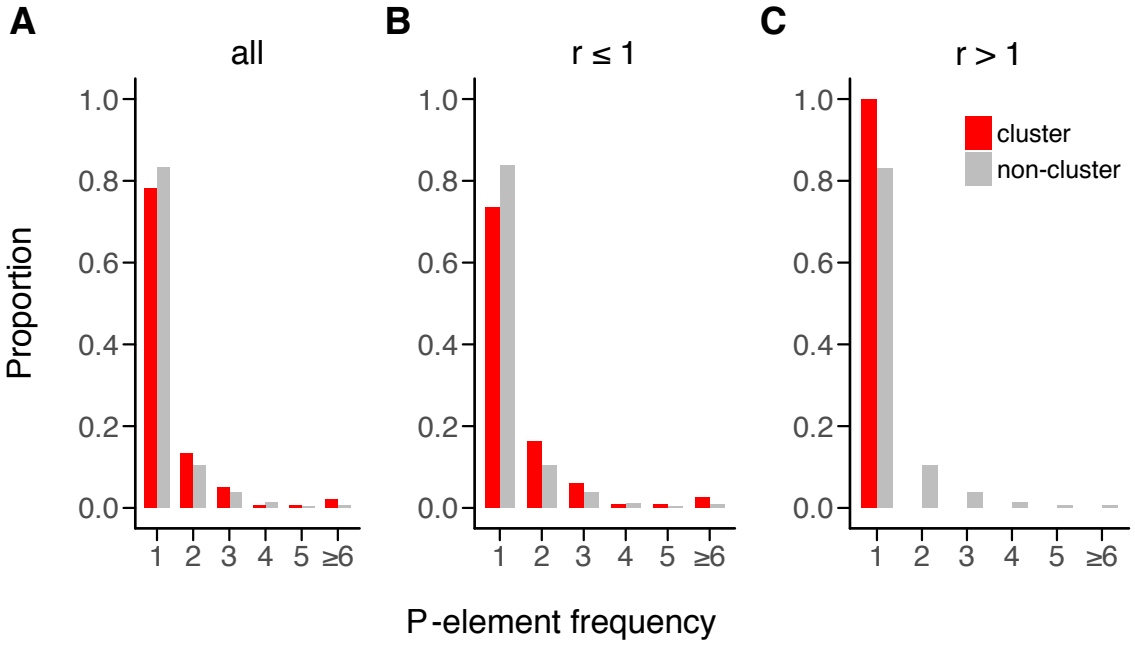

Figure S3.3 The frequency of *P*-elements in piRNA clusters and the frequency of *P*-elements outside of clusters were compared. Annotated 159 clusters were used. *P*-elements at hotspots were excluded from these comparisons. Insertions are compared in (A) all genomic regions ($W_{142, 4806}$ = 323360, P-value = 0.10), (B) regions of low recombination (≤1 cM/Mb, $W_{117, 1462}$ = 76641, P-value = 0.0042), and (C) high recombination (>1 cM/Mb, $W_{25, 3344}$ = 48850, P-value = 0.025).

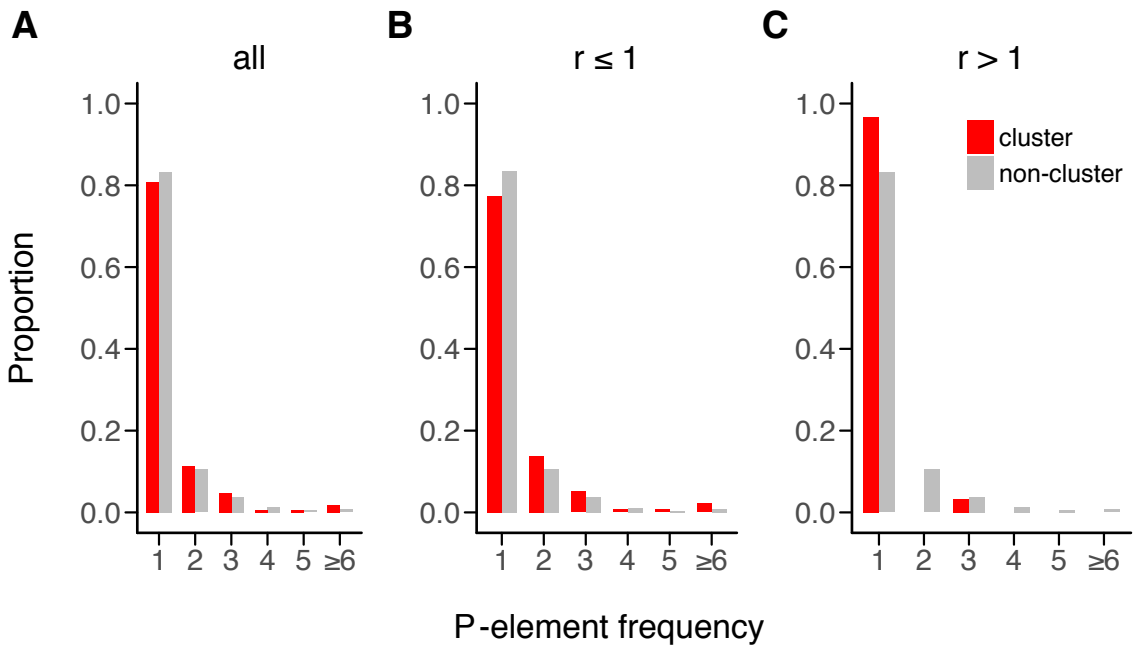

Figure S3.4 The frequency of *P*-elements in piRNA clusters and the frequency of *P*-elements outside of clusters were compared. Annotated 497 clusters were used. *P*-elements at hotspots were excluded from these comparisons. Insertions are compared in (A) all genomic regions ($W_{167, 4781}$ = 389090, P-value = 0.39), (B) regions of low recombination (≤1 cM/Mb, $W_{137, 1442}$ = 92581, P-value = 0.063), and (C) high recombination (>1 cM/Mb, $W_{30, 3349}$ = 56730, P-value = 0.054).

Table S2.1 Primers used in the library construction (Figure 2.1). P-enrich-F is the forward primer used in the asymmetric PCR. P-nest-F_1 – P-next-F_4 are forward primers used in the nested PCR. R1 – R15 are degenerate primers used to amplify adjacent genomic sequence.

Primer name	Sequence from 5' to 3'
P-enrich-F	CACGGACATGCTAAGGGTTAATC
P-nest-F_1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGTTAAGTGGATGTCTCTTGCC
P-nest-F_2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNCGTTAAGTGGATGTCTCTTGCC
P-nest-F_3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNCGTTAAGTGGATGTCTCTTGCC
P-nest-F_4	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNCGTTAAGTGGATGTCTCTTGCC
R1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNAAATG
R2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNAATTG
R3	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNAAATC
R4	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNAAACT
R5	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNAATGT
R6	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNGCCAA
R7	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNACAAT
R8	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNGCAGC
R9	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNCAGCA
R10	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNAATCA
R11	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNTGCCA
R12	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNTGGAA
R13	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNAGCAA
R14	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNATTCA
R15	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNAATGG

Table S2.2 The summary of sequenced paired-end reads. Percentage of *P*-derived read pairs (column 3) is the ratio of *P*-element derived read pairs to the number of sequenced read pairs (column 2). Unique alignment rate (column 4) is the percentage of read pairs aligned concordantly exactly one time, while overall alignment rate (column 5) is the percentage of read pairs aligned to the reference genome (one or more than one time). Un-annotated rate (column 6) is the percentage of multiply mapping read pairs that cannot be explained by insertions annotated by uniquely mapping reads.

Primer name	Read pairs	Percentage of P-derived read pairs	Unique alignment rate	Overall alignment rate	Un-annotated rate
R1	887994	94.50%	20.80%	85.73%	2.44%
R2	669016	96.13%	74.49%	97.45%	2.45%
R3	466546	95.23%	78.86%	85.82%	1.07%
R4	545050	95.59%	97.31%	97.97%	0.16%
R5	1309000	93.72%	24.58%	88.03%	2.39%
R6	791301	94.77%	89.14%	90.34%	0.36%
R7	448724	95.62%	31.86%	85.82%	14.41%
R8	716995	94.37%	75.46%	96.79%	1.10%
R9	798586	93.01%	57.36%	97.65%	1.25%
R10	822027	93.46%	60.29%	80.78%	2.85%
R11	584570	95.54%	78.95%	95.52%	2.10%
R12	425970	96.14%	67.14%	97.01%	2.27%
R13	479120	95.75%	89.57%	93.80%	0.84%
R14	629038	96.11%	38.88%	91.10%	4.60%
R15	523804	96.20%	72.15%	91.16%	1.78%

Table S2.3 Insertions found by concordantly and uniquely mapping read pairs. The inserted chromosomes and breakpoints are reported. The number in each cell indicates the number of read pairs that support each insertion. Sense means *P*-element is in the same orientation as the plus genomic strand. Antisense means *P*-element is in the same orientation as the minus genomic strand. Category indicates an insertion exists in the RAL-492 genome or it is a false positive. False positive insertions are marked in light grey. Insertions marked as yellow are two opposite *P*-element insertions at the same site, which are excluded from the down-stream analysis. The insertion chr3L:25797105 (light red) is similar to chr3L:26023661 (dark red). The insertion chr3L:26834988 and chrUn_CP007074v1:15794 (light blue) are similar to chrX_TAS:4894 (dark blue).

Insertion	R1	R2	R3	R4	R5	R6	R7	R8	R9
chr2L:220699	5694	1461	14998	128	12	2652	958	9	11
chr2L:3707098	0	69	0	0	0	0	0	0	0
chr2L:7573697	6348	1244	828	12	2660	114	439	1046	8
chr2L:8951631	0	0	26	0	0	0	0	0	0
chr2L:8951631	42	42	258	0	22	15	21	4	0
chr2L:12507829	187	10834	146182	839	383	475444	3567	1702	72
chr2L:15262593	0	37	0	0	0	0	0	0	0
chr2L:15264721	40	17	3661	0	626	46	45	29	197
chr2L:15762783	1213	30	678	36	13029	38190	884	1288	16
chr2L:20483955	0	75	0	0	0	0	0	0	0
chr2L:20917521	10	128	446	0	0	0	209	3	7
chr2L:21221219	209	273	11833	4	68	92	2709	48	5
chr2L:21222025	0	0	0	0	0	2	0	0	0
chr2R:6236112	10	3016	1324	19	361	36050	5872	387	26
chr2R:6236112	0	0	0	0	138	0	0	1	1
chr2R:6648629	282	163	420	18	1316	182	85	1717	24654
chr2R:17563557	92	96	10503	1	14	142	120	12	2

Table S2.3 continued

chr2R:20869396	0	0	0	0	0	28	0	0	0
chr2R:24292403	0	0	0	0	0	22	0	0	0
chr3L:3149449	11906	251129	26329	46	12033	108	480	466	1629
chr3L:6218052	0	6	93	0	4	7	0	20	10
chr3L:11710100	0	0	0	0	0	0	0	0	0
chr3L:11819921	70690	180982	122	24	3453	1487	2033	336	2686
chr3L:11820022	58	1	11	0	265	0	347	1850	3
chr3L:13265851	0	0	0	0	0	0	26	0	0
chr3L:13521910	160	8268	17	60	119	28	3	0	3
chr3L:14786218	0	0	0	0	0	0	0	0	0
chr3L:16891150	100	141	75842	12	55	33414	115	2	3
chr3L:17049414	266	1	73	0	31	2673	49	1838	5
chr3L:20104865	72	35	15	18	68	130	10420	8	2
chr3L:21240203	15	316	476	2	4	104	15	7899	1304
chr3L:22315456	0	0	0	0	0	0	0	0	0
chr3L:22408013	22	7	3606	17	161	709	41383	14	29
chr3L:22871816	25802	6805	336	45	205024	463	3700	356	7892
chr3L:22901871	11	530	35774	3	98	2781	150	47	27
chr3L:25797105	2	8	62	0	2	5	113	16	10
chr3L:26023661	7212	2287	522	168	1387	613	949	358782	382877
chr3L:26834988	50	0	0	0	94	0	15	0	0
chr3R:562644	2	14	1	0	3	3	0	0	0
chr3R:2210034	8	601	157	3	31	8	8	4	5
chr3R:4806575	292	988	13	1	692	214	52318	24	21
chr3R:4806575	34	0	59	0	0	0	0	1	0

Table S2.3 continued

chr3R:9533751	0	0	0	0	0	0	0	0	0
chr3R:14749333	0	0	72	0	0	0	0	0	0
chr3R:18399261	5933	1857	1106	1	120	1303	0	142	405
chr3R:31154703	0	0	0	0	0	0	0	0	0
chr3R:31713302	53	164	315	0	1	79	130	148	17
chr3R:31738027	27909	849	7	11	19931	5	4154	21	14
chrUn_CP007074v1:15794	30	0	0	0	211	0	1	0	0
chrX:2469894	69	44	2572	1	6	11768	175	7	1
chrX:4324864	0	0	0	0	0	0	0	0	0
chrX:4324864	166	414	54	74	7	632	103	161	2191
chrX:6679501	0	0	0	0	0	0	0	0	0
chrX:8925028	6372	69	6	1	59	43593	24	400	37
chrX:17314576	706	4534	346	503821	304	11427	666	127678	140
chrX_TAS:4894	1150	8	1	1	25194	7	85	2	3
chrY:768808	7	1	10	0	1	0	0	2	0

Table S2.3 continued

Insertion	R10	R11	R12	R13	R14	R15	Strand	Category
chr2L:220699	7547	246	60	2505	659	23076	antisense	TRUE
chr2L:3707098	0	0	0	0	0	0	antisense	FALSE
chr2L:7573697	964	1036	66	19	18	58	antisense	TRUE
chr2L:8951631	0	20	0	2	0	0	sense	TRUE
chr2L:8951631	28224	14	3351	64	0	57	antisense	
chr2L:12507829	24165	299	400	5478	23685	67	antisense	TRUE

Table S2.3 continued

chr2L:15262593	0	0	0	0	0	0	antisense	FALSE
chr2L:15264721	814	848	4	2524	22	44	antisense	TRUE
chr2L:15762783	1822	3665	27	8481	71	6	antisense	TRUE
chr2L:20483955	0	0	0	0	0	0	antisense	FALSE
chr2L:20917521	184	132	655	4	1061	0	sense	TRUE
chr2L:21221219	164	82	89	220	115	26	sense	TRUE
chr2L:21222025	0	37	1	0	0	0	antisense	FALSE
chr2R:6236112	2592	36947	116	1841	20	185	sense	TRUE
chr2R:6236112	0	0	8	0	0	0	antisense	
chr2R:6648629	1101	163	17	137842	540	56	sense	TRUE
chr2R:17563557	1315	14214	468	75	5511	1238	sense	TRUE
chr2R:20869396	0	0	35	0	0	0	antisense	FALSE
chr2R:24292403	0	0	0	0	0	0	antisense	FALSE
chr3L:3149449	449	668	7835	47157	102519	264150	antisense	TRUE
chr3L:6218052	195	237	1	5	967	5	antisense	TRUE
chr3L:11710100	0	28	0	0	0	0	sense	FALSE
chr3L:11819921	5869	292926	132910	2193	280	51490	antisense	TRUE
chr3L:11820022	101	11	1	79	0	10	sense	TRUE
chr3L:13265851	0	0	0	0	0	0	antisense	FALSE
chr3L:13521910	1768	97	19	3065	2	478	sense	TRUE
chr3L:14786218	0	47	0	0	0	0	sense	FALSE
chr3L:16891150	988	20115	49	615	516	81	sense	TRUE
chr3L:17049414	214	64	157	64	791	1	sense	TRUE
chr3L:20104865	7028	57	2460	7	5623	18	antisense	TRUE
chr3L:21240203	202	43	12	12201	152	8	antisense	TRUE

Table S2.3 continued

chr3L:22315456	0	0	0	67	0	0	antisense	FALSE
chr3L:22408013	44	916	1168	10937	5558	27	antisense	TRUE
chr3L:22871816	53181	310	13039	1553	42665	1804	sense	TRUE
chr3L:22901871	301848	7223	12	3951	3478	407	sense	TRUE
chr3L:25797105	994	28	34	0	557	3	antisense	FALSE
chr3L:26023661	5465	259	102391	140151	11912	5975	sense	TRUE
chr3L:26834988	0	0	0	0	0	0	antisense	FALSE
chr3R:562644	0	54	0	1	0	0	antisense	TRUE
chr3R:2210034	0	2	7	433	1039	24	sense	TRUE
chr3R:4806575	17	7766	89	236	180	130	sense	TRUE
chr3R:4806575	0	0	1	0	0	0	antisense	
chr3R:9533751	0	20	0	0	0	0	antisense	FALSE
chr3R:14749333	0	0	0	0	0	0	sense	FALSE
chr3R:18399261	1	3728	1	18	333	29	antisense	TRUE
chr3R:31154703	0	21	0	0	0	0	antisense	FALSE
chr3R:31713302	772	128	1123	466	270	3	antisense	TRUE
chr3R:31738027	3	89	159	10	9010	1	sense	TRUE
chrUn_CP007074v1:15794	0	0	0	0	0	0	antisense	FALSE
chrX:2469894	0	1404	40	5	130	325	antisense	TRUE
chrX:4324864	0	90	5	0	0	1	antisense	TRUE
chrX:4324864	1289	198	606	19936	6961	32	sense	
chrX:6679501	0	0	0	0	557	0	sense	FALSE
chrX:8925028	6	459	2205	146	1	11567	antisense	TRUE
chrX:17314576	3708	19230	3473	1032	3747	698	antisense	TRUE

Table S2.3 continued

chrX_TAS:4894	2	1	2	2	0	2	sense	TRUE
chrY:768808	246	8	3	2	4443	1	sense	TRUE

Table S2.4 PCR verification of insertions found by hemi-specific PCR.

Insertion	Forward primer	Reverse primer	Annotation sets	PCR condition	PCR result	Sequencing results
chr2L:20917521	TTTTCGCCCTACATCCATTT	TTCATGCATACCACTGCACA	Hemi-specific PCR	94 °C 2 min; 35 cycles of 94 °C 30 sec, 60 °C 30 sec and 72 °C 3:30 min; 72 °C 5 min	Both inserted and un-inserted chromosmes are amplified	The PCR product contains an internal deleted *P*-element and ajdecent genomic sequence around chr2L:20917521.
chr2L:8951631	TGTTTCTACCTGTGCAATCAACA	CCCTTAAGATAGTTTCCAACAGC	Hemi-specific PCR, TEMP and TIDAL	94 °C 2 min; 35 cycles of 94 °C 30 sec, 60 °C 30 sec and 72 °C 3:30 min; 72 °C 5 min	Inserted chromosome is amplified	The PCR product contains an internal deleted *P*-element and ajdecent genomic sequence around chr2L:8951631.
chr2L:20483955	ACGAGTTGAGAGAGTGCGAA	TCTTCTTGTCTTGCCTCCGT	Hemi-specific PCR	95 °C 3 min; 35 cycles of 95 °C 30 sec, 60 °C 30 sec and 72 °C 4:30 min; 72 °C 10 min	Un-inserted chromosome is amplied	NA
	AGTTTTCACCAAGGCTGCAC	TCTTCTTGTCTTGCCTCCGT	Hemi-specifc PCR	95 °C 3 min; 35 cycles of 95 °C 30 sec, 60 °C 30 sec and 72 °C 1:30 min; 72 °C 5 min	The PCR product is ~800 bp	Part of the PCR product belongs to *P*-element, and the other part is mapped to the genome region around chr3L: 26023660.

Table S2.4 continued

chr2L:21222025	ATCTGCAGGAGTGGTTGACA	TTTCAGGTTCCCTCCCACTC	Hemi-specific PCR	95 °C 3 min; 35 cycles of 95 °C 30 sec, 60 °C 30 sec and 72 °C 4:30 min; 72 °C 10 min	Un-inserted chromosome is amplied	NA
	AGTTTTCACCAAGGCTGCAC	TTTCAGGTTCCCTCCCACTC	Hemi-specific PCR	95 °C 3 min; 35 cycles of 95 °C 30 sec, 60 °C 30 sec and 72 °C 1:30 min; 72 °C 5 min	main band is not as expected	NA
chr2R:6236112	GAGATGCTCCTTAGCTGTTCG	GGAACGTAAGAGCTGCCAAA	Hemi-specific PCR, TEMP and TIDAL	94 °C 2 min; 35 cycles of 94 °C 30 sec, 60 °C 30 sec and 72 °C 3:30 min; 72 °C 5 min	Inserted chromosome is amplified	The PCR product contains an internal deleted *P*-element and adjacent genomic sequence around chr2R:6236112.
chr2R:17563557	GCAGTCGACAGAGTTTATGCC	CCTATGGTGGTTGCTGGGAT	Hemi-specific PCR, TEMP and TIDAL	94 °C 2 min; 35 cycles of 94 °C 30 sec, 60 °C 30 sec and 72 °C 3:30 min; 72 °C 5 min	Inserted chromosome is amplified	The PCR product contains an internal deleted *P*-element and adjacent genomic sequence around chr2R:17563557.
chr2R:23862575	AGTTTTCACCAAGGCTGCAC	TCCGGTACAGCGGTTACATT	TEMP and TIDAL	95 °C 3 min; 35 cycles of 95 °C 30 sec, 55 - 65 °C 30 sec and 72 °C 2 min; 72 °C 5 min	Inserted chromosome is amplified and the PCR product is ~1000 bp	Part of the PCR product belongs to *P*-element, and the other part is mapped to the genome region around chr2R: 23862575.
chr2R:20869396	ACGGTTCCAGCTGCGATATA	GCGCCCAGTTCCAGTTATTC	Hemi-specific PCR	95 °C 3 min; 35 cycles of 95 °C 30 sec, 60 °C 30 sec and 72 °C 4:30 min; 72 °C 10 min	Un-inserted chromosome is amplified.	NA

Table S2.4 continued

chr2R:20869396	AGTTTTCACCAAGGCTGCAC	GCGCCCAGTTCCAGTTATTC	Hemi-specific PCR	95 °C 3 min; 35 cycles of 95 °C 30 sec, 50 - 65 °C 30 sec and 72 °C 1:30 min; 72 °C 5 min	No clear main band.	NA
chr3L:16891150	GGTTCTTGCGGCTCTCTTG	ATTCCACAACACTGCCAACC	Hemi-specific PCR and TEMP	94 °C 2 min; 35 cycles of 94 °C 30 sec, 60 °C 30 sec and 72 °C 3:30 min; 72 °C 5 min	Inserted chromosome is amplified.	The PCR product contains an internal deleted P-element and adjacent genomic sequence around chr3L:16891150.
chr3L:25797105	GAGGACAGAGGAGAACGGAG	GGATGTTATGCTGCAGGGTG	Hemi-specific PCR	95 °C 2 min; 35 cycles of 95 °C 30 sec, 57 - 63 °C 30 sec and 72 °C 4 min; 72 °C 5 min	Un-inserted chromosome is amplified.	The PCR product doesn't have P-element sequence, and it is mapped to the region around chr3L:25797105.
chr3L:26023661	CACGGACATGCTAAGGGTTAATC	GAGGACAGAGGAGAACGGAG	Hemi-specific PCR and TEMP	94 °C 3 min; 35 cycles of 94 °C 30 sec, 60 °C 30 sec and 72 °C 3 min; 72 °C 10 min	The PCR product is ~300 bp.	Part of the PCR product belongs to P-element, and the other part is mapped to the genome region around chr3L:26023661.
chr3L:22408013	CACGGACATGCTAAGGGTTAATC	ACGTTAAGGCAATGCAAAACA	Hemi-specific PCR, TEMP and TIDAL	94 °C 3 min; 35 cycles of 94 °C 30 sec, 60 °C 30 sec and 72 °C 3 min; 72 °C 10 min	The main band is ~750 bp.	Part of the PCR product belongs to P-element, and the other part is mapped to the genome region around chr3L:22408013.

Table S2.4 continued

chr3L:3149449	CACGGACATG CTAAGGGTTA ATC	GTTCACGGCG TCTTCTTTGT	Hemi-specific PCR, TEMP and TIDAL	94 °C 3 min; 35 cycles of 94 °C 30 sec, 60 °C 30 sec and 72 °C 3 min; 72 °C 10 min	Multiple bands are amplified. The band ~500 bp is sequenced.	Part of the PCR product belongs to *P*-element, and the other part is mapped to the genome region around chr3L:3149449.
chr3L:22315456	ATGCGCGGCC ATTAAAGTTT	TGGCGATTTC ACACACTCAC	Hemi-specific PCR	95 °C 3 min; 35 cycles of 95 °C 30 sec, 60 °C 30 sec and 72 °C 4:30 min; 72 °C 10 min	Un-inserted chromosome is amplified.	NA
	AGTTTTCACC AAGGCTGCAC	TGGCGATTTC ACACACTCAC	Hemi-specific PCR	95 °C 3 min; 35 cycles of 95 °C 30 sec, 50 - 65 °C 30 sec and 72 °C 1:30 min; 72 °C 5 min	Multiple bands are amplified and the main band is not expected.	NA
chr3R:4806575	CACGGACATG CTAAGGGTTA ATC	GGATGTTATG CTGCAGGGTG	Hemi-specific PCR, TEMP and TIDAL	94 °C 3 min; 35 cycles of 94 °C 30 sec, 60 °C 30 sec and 72 °C 3 min; 72 °C 10 min	Multiple bands are amplified. The band ~350 bp is sequenced.	Part of the PCR product belongs to P-element, and the other part is mapped to the genome region around chr3R:4806575.
chr3R:2210034	CACGGACATG CTAAGGGTTA ATC	AACGGCACAT TAGTTGACCA	Hemi-specific PCR and TEMP	94 °C 3 min; 35 cycles of 94 °C 30 sec, 60 °C 30 sec and 72 °C 3 min; 72 °C 10 min	As expected, the length of PCR product is ~350 bp.	Part of the PCR product belongs to P-element, and the other part is mapped to the genome region around chr3R:2210034.

Table S2.4 continued

chr3R:8698090	ATGCCAGAAGTGCAACAACA	TCAATAGTGAAGCGGAGATGC	TEMP	95 °C 2 min; 35 cycles of 95 °C 30 sec, 57 - 63 °C 30 sec and 72 °C 4 min; 72 °C 5 min	Un-inserted chromosome is amplified.	The sequence of PCR product is mapped to the genome region around chr3R:8698090 and span the breakpoint estimated by TEMP.
chrX:8925028	GCACTTTCTCCGCAGTTCTC	AGAGGGCGCTATTGAGCATA	Hemi-specific PCR, TEMP and TIDAL	94 °C 2 min; 35 cycles of 94 °C 30 sec, 60 °C 30 sec and 72 °C 3:30 min; 72 °C 5 min	Inserted chromosome is amplified.	The PCR product contains an internal deleted *P*-element and ajdecent genomic sequence around chrX:8925028
chrX:8574278	AGTTTTCACCAAGGCTGCAC	TCAGTGCACCTTAGTCATCTGA	TEMP and TIDAL	95 °C 3 min; 35 cycles of 95 °C 30 sec, 60 °C 30 sec and 72 °C 1:40 min; 72 °C 10 min	Inserted chromosome is amplified, and the length of PCR product is expected.	Part of the PCR product is P-element, and the other part is mapped to the genome region around chrX: 8574278.
chrX:6679501	TCCACCCATAAATCCCGAGG	AAGTAAACACAGGGGCAGAC	Hemi-specifc PCR	95 °C 3 min; 35 cycles of 95 °C 30 sec, 60 °C 30 sec and 72 °C 4:30 min; 72 °C 10 min	Un-inserted chromosome is amplified.	NA
	TCCACCCATAAATCCCGAGG	AGTTTTCACCAAGGCTGCAC	Hemi-specifc PCR	95 °C 3 min; 35 cycles of 95 °C 30 sec, 50 - 65 °C 30 sec and 72 °C 1:30 min; 72 °C 5 min	The main band is not expected.	NA

Table S2.4 continued

chrX_TAS:4894	CACCGAAATGGATGAGTTGACG	CACACTTACCATAGAGCAAGGG	Hemi-specifc PCR	95 ˚C 3 min; 35 cycles of 95 ˚C 30 sec, 64 ˚C 30 sec and 72 ˚C 1:40 min; 72 ˚C 10 min	Inserted chromosome is amplified and the PCR product is ~1500 bp	Part of the PCR product belongs to *P*-element, and the other part is mapped to *X*-TAS
chrY:768808	CTTCTGATGTTGTGGACGCC	GTGATGGCTCTCTTTGCTTCT	Hemi-specifc PCR	95 ˚C 3 min; 35 cycles of 95 ˚C 30 sec, 60 ˚C 30 sec and 72 ˚C 4:30 min; 72 ˚C 10 min	Inserted chromosome is amplified and the PCR product is ~1800 bp	NA

Table S3.1 Small RNA sequencing libraries used for ancestral piRNA cluster annotation. Biological replicates are combined in the heatmap showing piRNA clusters are differentially expressed (Figure 3.1).

ID	SRR	Strain	Library preparation method	2S rRNA depletion	oxidated	Heatmap ID	Reference
1	SRR7814378	21147-R1	NEBNext Small RNA Library Prep for Illumina	Terminator oligo blocking	Not oxidated	21147	Libraries from 1 to 18 come from Jyoti & Kelleher unpublished paper
2	SRR7814377	21147-R2					
3	SRR7814385	21147-R3					
4	SRR7814381	21183-R1				21183	
5	SRR7814382	21183-R2					
6	SRR7814390	21183-R3					
7	SRR7814387	21188-R1				21188	
8	SRR7814388	21188-R2					
9	SRR7814384	21188-R3					
10	SRR7814383	21213-R1				21213	
11	SRR7814380	21213-R2					
12	SRR7814379	21213-R3					
13	SRR7814375	21291-R1				21291	
14	SRR7814389	21291-R2					
15	SRR7814376	21291-R3					
16	SRR7814386	21346-R1				21346	
17	SRR7814374	21346-R2					
18	SRR7814373	21346-R3					

Table S3.1 continued

19	SRR014280	w1118 3-5-day-old ovary	3' and 5' adapter were added to extracted RNA. Ligated products were reverse transcribed and PCR amplified before sequencing	without 2S depletion	Not oxidated	w1118_1	(Brennecke et al. 2008)
20	SRR3715417	w1118 ovaries	Based on Illumina TrueSeq Small RNA sample prep kit. Four random nucleotides were added at 3' end of 5' adaptor and 5' end of 3' adaptor.	without 2S depletion	oxidated using $NaIO_4$	w1118_2	(Hayashi et al. 2016)
21	SRR3715418	w1118 total ovaries		hybridization to immobilized oligos anchored to magnetic beads		w1118_3	
22	SRR1785658	w1118 2-day-old ovaries	Illumina TruSeq Small RNA sample prep kit	Terminator oligo blocking	Not oxidated	w1118_4	(Lo et al. 2016)
23	SRR1785659	w1118 2-day-old ovaries					
24	SRR1785660	w1118 2-day-old ovaries					

Table S3.1 continued

25	SRR827770	y[1]; cn[1] bw[1] sp[1]	19-28 nt small RNA were selected using polyacrylamide gels. Then, 3' and 5' adapter were added. Ligated products were reverse transcribed and PCR amplified before sequencing	without 2S depletion	Not oxidated	2057	(Shpiz et al. 2014)
26	SRR891254	cn bw	Illumina Small RNA Library Prep Kit	without 2S depletion	Not oxidated	cn bw	(Pane et al. 2011)
27	SRR891255						

Table S3.2 Estimation of the number of *P*-elements in TAS. "mean" is the average number of read pairs supporting individual non-TAS *P*-element insertions in a given genome. "standard deviation" is the standard deviation of reads supporting non-TAS insertions. "TAS reads" is the number of *P*-derived reads aligned to *2R*, *3R*, and *X*-TAS. If *Z* score is less than -1.96, we assigned 0 *P*-element to TAS. If *Z* score is greater than 1.96, we assigned >1 *P*-elements to TAS. Otherwise, 1 *P*-element was assigned.

ID	Mean	Standard deviation	TAS reads	*Z* score	Number of *P* in TAS
DGRP100	48.81	20.23	211.00	8.02	>1
DGRP101	22.23	9.99	28.00	0.58	1
DGRP105	29.50	7.92	23.00	-0.82	1
DGRP109	44.93	29.56	50.00	0.17	1
DGRP129	20.44	6.59	15.00	-0.83	1
DGRP136	22.62	10.12	18.00	-0.46	1
DGRP138	46.87	41.02	2.00	-1.09	1
DGRP142	25.50	11.70	4.00	-1.84	1
DGRP149	21.94	7.69	17.00	-0.64	1
DGRP158	12.41	5.63	8.00	-0.78	1
DGRP161	22.04	9.75	23.00	0.10	1
DGRP176	37.59	10.98	91.00	4.86	>1
DGRP177	23.88	8.49	19.00	-0.57	1
DGRP181	47.83	15.25	172.00	8.14	>1
DGRP189	30.96	12.84	167.00	10.60	>1
DGRP195	28.21	13.22	30.00	0.14	1
DGRP208	37.82	13.18	108.00	5.32	>1
DGRP21	20.89	8.89	42.00	2.37	>1
DGRP217	25.24	11.32	46.00	1.83	1
DGRP223	25.55	13.76	0.00	-1.86	1
DGRP227	24.79	9.13	26.00	0.13	1
DGRP228	26.00	9.45	20.00	-0.63	1
DGRP229	62.52	32.08	66.00	0.11	1
DGRP233	22.60	12.92	34.00	0.88	1
DGRP235	15.61	6.63	16.00	0.06	1
DGRP239	27.78	16.33	14.00	-0.84	1
DGRP256	39.40	13.04	18.00	-1.64	1
DGRP26	44.82	11.92	18.00	-2.25	0
DGRP272	18.27	8.80	0.00	-2.08	0
DGRP280	29.39	7.52	1.00	-3.77	0

Table S3.2 continued

DGRP287	14.21	4.82	0.00	-2.95	0
DGRP301	59.82	31.48	76.00	0.51	1
DGRP303	65.56	58.89	185.00	2.03	>1
DGRP304	52.83	24.13	53.00	0.01	1
DGRP306	59.78	31.44	62.00	0.07	1
DGRP307	76.45	26.73	116.00	1.48	1
DGRP309	18.81	7.64	25.00	0.81	1
DGRP31	38.03	21.38	60.00	1.03	1
DGRP310	26.63	8.63	41.00	1.67	1
DGRP315	95.83	43.28	148.00	1.21	1
DGRP317	19.04	6.25	19.00	-0.01	1
DGRP318	30.85	12.71	17.00	-1.09	1
DGRP319	37.45	23.27	92.00	2.34	>1
DGRP32	43.03	28.19	51.00	0.28	1
DGRP320	20.96	9.89	26.00	0.51	1
DGRP321	33.95	18.29	9.00	-1.36	1
DGRP324	133.53	74.92	430.00	3.96	>1
DGRP332	36.00	17.11	31.00	-0.29	1
DGRP336	14.58	8.21	9.00	-0.68	1
DGRP338	18.10	8.52	15.00	-0.36	1
DGRP340	36.78	25.20	100.00	2.51	>1
DGRP348	31.74	13.46	63.00	2.32	>1
DGRP350	18.13	9.52	22.00	0.41	1
DGRP352	21.55	9.76	21.00	-0.06	1
DGRP354	47.93	19.91	47.00	-0.05	1
DGRP355	44.30	17.94	54.00	0.54	1
DGRP356	17.56	6.58	31.00	2.04	>1
DGRP358	10.48	4.08	4.00	-1.59	1
DGRP359	18.24	7.27	26.00	1.07	1
DGRP360	22.59	6.03	40.00	2.89	>1
DGRP361	27.93	13.14	116.00	6.70	>1
DGRP362	20.96	10.90	44.00	2.11	>1
DGRP367	21.00	10.04	19.00	-0.20	1
DGRP370	25.26	9.85	14.00	-1.14	1
DGRP371	75.08	25.42	131.00	2.20	>1
DGRP373	36.83	18.05	151.00	6.33	>1
DGRP374	9.38	4.65	71.00	13.26	>1
DGRP375	43.33	25.92	19.00	-0.94	1

Table S3.2 continued

DGRP377	17.15	7.62	0.00	-2.25	0
DGRP378	15.98	8.94	21.00	0.56	1
DGRP38	21.91	10.80	42.00	1.86	1
DGRP380	49.15	13.23	34.00	-1.14	1
DGRP381	31.25	15.26	63.00	2.08	>1
DGRP382	26.19	9.98	31.00	0.48	1
DGRP383	21.95	9.51	25.00	0.32	1
DGRP385	8.92	2.87	6.00	-1.01	1
DGRP390	22.46	12.69	54.00	2.49	>1
DGRP391	25.54	12.37	20.00	-0.45	1
DGRP392	16.06	8.36	30.00	1.67	1
DGRP393	25.38	9.80	55.00	3.02	>1
DGRP395	33.85	21.01	87.00	2.53	>1
DGRP397	20.87	12.15	21.00	0.01	1
DGRP398	15.54	5.33	3.00	-2.35	0
DGRP399	30.00	9.31	25.00	-0.54	1
DGRP40	36.56	15.93	27.00	-0.60	1
DGRP405	13.68	6.20	17.00	0.54	1
DGRP406	16.63	8.08	28.00	1.41	1
DGRP409	27.65	12.47	11.00	-1.33	1
DGRP42	13.41	3.69	14.00	0.16	1
DGRP426	16.52	8.15	7.00	-1.17	1
DGRP437	54.14	19.17	55.00	0.05	1
DGRP439	11.71	6.94	27.00	2.21	>1
DGRP440	14.57	5.98	26.00	1.91	1
DGRP441	21.64	7.79	33.00	1.46	1
DGRP443	20.21	9.74	52.00	3.26	>1
DGRP45	35.18	11.15	78.00	3.84	>1
DGRP461	17.19	7.56	30.00	1.69	1
DGRP48	27.36	14.71	51.00	1.61	1
DGRP49	29.64	14.13	33.00	0.24	1
DGRP491	32.64	10.86	18.00	-1.35	1
DGRP492	28.81	12.22	21.00	-0.64	1
DGRP502	20.32	10.11	58.00	3.73	>1
DGRP505	56.82	26.97	48.00	-0.33	1
DGRP508	21.83	11.08	15.00	-0.62	1
DGRP509	35.52	14.94	0.00	-2.38	0
DGRP513	17.26	7.41	25.00	1.04	1

Table S3.2 continued

DGRP517	86.33	17.67	128.00	2.36	>1
DGRP528	30.65	14.28	4.00	-1.87	1
DGRP530	21.41	7.92	33.00	1.46	1
DGRP531	21.81	10.36	13.00	-0.85	1
DGRP535	28.91	8.84	28.00	-0.10	1
DGRP551	23.07	10.04	52.00	2.88	>1
DGRP554	41.68	10.76	55.00	1.24	1
DGRP555	20.27	5.97	33.00	2.13	>1
DGRP559	20.93	11.49	38.00	1.49	1
DGRP563	34.88	16.76	93.00	3.47	>1
DGRP566	23.91	15.69	42.00	1.15	1
DGRP57	50.76	28.25	19.00	-1.12	1
DGRP584	38.19	20.09	228.00	9.45	>1
DGRP589	37.28	21.14	53.00	0.74	1
DGRP59	32.53	11.68	48.00	1.33	1
DGRP591	30.56	6.95	23.00	-1.09	1
DGRP595	31.63	14.45	14.00	-1.22	1
DGRP596	44.86	27.80	8.00	-1.33	1
DGRP627	25.38	12.84	14.00	-0.89	1
DGRP630	15.40	8.02	21.00	0.70	1
DGRP634	25.16	7.34	34.00	1.20	1
DGRP639	26.38	8.70	18.00	-0.96	1
DGRP642	28.75	13.77	24.00	-0.35	1
DGRP646	21.47	9.45	17.00	-0.47	1
DGRP69	29.48	10.07	61.00	3.13	>1
DGRP703	16.90	6.45	35.00	2.81	>1
DGRP705	18.22	7.03	20.00	0.25	1
DGRP707	37.22	13.82	23.00	-1.03	1
DGRP712	34.59	12.27	81.00	3.78	>1
DGRP714	26.93	12.35	19.00	-0.64	1
DGRP716	23.17	11.79	47.00	2.02	>1
DGRP721	26.24	7.91	38.00	1.49	1
DGRP727	69.86	27.92	110.00	1.44	1
DGRP73	35.07	14.38	24.00	-0.77	1
DGRP732	29.08	9.45	47.00	1.90	1
DGRP737	22.64	12.71	23.00	0.03	1
DGRP738	29.11	15.84	45.00	1.00	1
DGRP748	41.35	17.62	87.00	2.59	>1

Table S3.2 continued

DGRP75	26.09	15.61	9.00	-1.10	1
DGRP757	33.38	14.25	24.00	-0.66	1
DGRP761	35.13	14.84	40.00	0.33	1
DGRP765	44.15	11.52	42.00	-0.19	1
DGRP774	69.59	14.27	74.00	0.31	1
DGRP776	36.13	12.23	26.00	-0.83	1
DGRP783	32.85	12.02	60.00	2.26	>1
DGRP787	31.97	12.55	23.00	-0.71	1
DGRP790	17.27	7.30	39.00	2.98	>1
DGRP796	35.69	19.61	35.00	-0.04	1
DGRP799	31.33	12.88	20.00	-0.88	1
DGRP801	27.59	16.33	1.00	-1.63	1
DGRP802	38.12	19.50	52.00	0.71	1
DGRP804	29.61	16.90	18.00	-0.69	1
DGRP805	29.33	12.08	2.00	-2.26	0
DGRP808	35.62	17.47	0.00	-2.04	0
DGRP810	34.15	10.86	58.00	2.20	>1
DGRP812	38.90	14.81	34.00	-0.33	1
DGRP818	32.62	11.47	60.00	2.39	>1
DGRP819	77.00	23.98	244.00	6.96	>1
DGRP820	20.43	9.92	22.00	0.16	1
DGRP821	115.84	60.11	258.00	2.36	>1
DGRP822	9.19	4.13	11.00	0.44	1
DGRP83	36.00	12.34	55.00	1.54	1
DGRP832	10.00	3.39	7.00	-0.88	1
DGRP837	14.31	4.44	21.00	1.51	1
DGRP843	43.63	22.98	24.00	-0.85	1
DGRP849	40.11	20.68	122.00	3.96	>1
DGRP85	27.31	13.65	0.00	-2.00	0
DGRP850	58.09	32.08	63.00	0.15	1
DGRP853	44.29	28.73	73.00	1.00	1
DGRP855	12.16	5.95	1.00	-1.88	1
DGRP857	13.00	5.78	32.00	3.29	>1
DGRP859	22.97	7.51	37.00	1.87	1
DGRP861	18.85	3.57	45.00	7.32	>1
DGRP879	14.64	4.95	8.00	-1.34	1
DGRP88	19.87	11.31	41.00	1.87	1
DGRP882	29.35	12.48	11.00	-1.47	1

Table S3.2 continued

DGRP884	20.11	7.19	18.00	-0.29	1
DGRP887	18.38	9.80	12.00	-0.65	1
DGRP890	22.12	6.58	38.00	2.41	>1
DGRP892	23.32	11.27	32.00	0.77	1
DGRP894	17.10	9.11	16.00	-0.12	1
DGRP897	48.00	19.56	56.00	0.41	1
DGRP900	53.41	26.44	208.00	5.85	>1
DGRP907	12.75	5.39	11.00	-0.32	1
DGRP908	23.70	10.49	40.00	1.55	1
DGRP91	37.92	14.22	0.00	-2.67	0
DGRP911	14.75	7.98	8.00	-0.85	1
DGRP913	41.63	25.52	79.00	1.46	1
DGRP93	32.63	15.60	32.00	-0.04	1

Table S3.3. Identification of TAS breakpoints. PCR verified *X*-TAS insertions were assigned to repeat B of *X*-TAS. PCR verified *2R* and *3R*-TAS insertions were assigned to the first repeat of *2R* and *3R*-TAS, respectively. "computationally inferred" represents insertions that were obtained from computational result but were not verified by PCR in a given genome.

ID	PCR verified	PCR condition	Computationally inferred
DGRP100	chrX_TAS:2678:+;chrX_TAS:2842:-	chrX_TAS:2678:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chrX_TAS:2842:- P-element5-R1 and TAS-XL-R1 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP101	chrX_TAS:3735:+	chrX_TAS:3735:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP105	chrX_TAS:2665:+;chr3R:32073857:-	chrX_TAS:2665:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr3R:32073857:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP109	chr2R:25258587:-	chr2R:25258587:- P-element5-R1 and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55-65°C gradient 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP129	chr2R:25258751:-	chr2R:25258751:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP136	chrX_TAS:3014:+	chrX_TAS:3014:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP138	untested		NA
DGRP142	chrX_TAS:3702:+	chrX_TAS:3702:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP149	chrX_TAS:2609:+	chrX_TAS:2609:+ TAS-XL-F2 and P-element5-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP153	chrX_TAS:3575:+	chrX_TAS:3575:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP158	chr2R:25258818:+	chr2R:25258818:+ P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes.	NA

Table S3.3 continued

DGRP161	chrX_TAS:3014:-;chr2R:25258635:-	chrX_TAS:3014:- P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55-65°C gradient 30 seconds, 72°C 2 minutes; 72°C 10 minutes. chr2R:25258635:- P-element5-R1 and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP176	chrX_TAS:2798:+;chr2R:25258680:+	chrX_TAS:2798:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes. chr2R:25258680:+ P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP177	failed		NA
DGRP181	chrX_TAS:3014:+;chr3R:32073805:-;chrX_TAS:3575:-;chrX_TAS:3014:-	chrX_TAS:3014:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chrX_TAS:3575:- P-element5-R1 and TAS-XL-R1 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr3R:32073805:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chrX_TAS:3014:- P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP189	chr3R:32073709:-;chrX_TAS:3695:-;chrX_TAS:4022:+	chrX_TAS:3695:- P-element5-R1 and TAS-XL-R1 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr3R:32073709:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chrX_TAS:4022:+ P-enrich and TAS-XL-R2; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP195	chr2R:25258751:+	chr2R:25258751:+ P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55-65°C gradient 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP208	chr3R:32073819:-;chrX_TAS:2842:+	chr3R:32073819:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds 65°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chrX_TAS:2842:+ P-enrich andTAS-XL-R1 ; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP21	chrX_TAS:2842:+	chrX_TAS:2842:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA

Table S3.3 continued

DGRP217	chr3R:32073634:-	chr3R:32073634:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP223	untested		NA
DGRP227	chrX_TAS:3575:+	chrX_TAS:3575:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP228	failed		NA
DGRP229	chrX_TAS:2670:+	chrX_TAS:2670:+ TAS-XL-F2 and P-element5-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP233	untested		NA
DGRP235	chr2R:25258587:-	chr2R:25258587:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP237	chrX_TAS:2670:+	chrX_TAS:2670:+: P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP239	chr2R:25258587:-	chr2R:25258587:- P-element5-R1 and chr2R_25259172; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 63°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP256	chrX_TAS:3695:-	chrX_TAS:3695:- P-element5-R1 and TAS-XL-R1 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP26	chrX_TAS:2670:+	chrX_TAS:2670:+ TAS-XL-F2 and P-element5-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP272	untested		NA
DGRP280	failed		NA
DGRP287	untested		NA
DGRP301	chrX_TAS:2670:+	chrX_TAS:2670:+ P-enrich and TAS-XL-R ; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP303	chrX_TAS:2670:+;chrX_TAS:2670:-	chrX_TAS:2670:+ P-enrich andTAS-XL-R1 ; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chrX_TAS:2670:- TAS-XL-F2 and P-enrich; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA

Table S3.3 continued

DGRP304	untested		chrX_TAS:3014:+
DGRP306	chrX_TAS:3014:+	chrX_TAS:3014:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP307	untested		chrX_TAS:3082:+;chrX_TAS:3575:-
DGRP309	chrX_TAS:3094:+	chrX_TAS:3094:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 50-65°C gradient 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP31	failed		chr3R:32073773:+
DGRP310	chrX_TAS:3014:-;chr2R:25258836:-	chrX_TAS:3014:- P-enrich and chr2R_25259172; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr2R:25258836:- P-element5-R1 and chr2R_25259172; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP315	chr2R:25258760:+	chr2R:25258760:+ P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes.	NA
DGRP317	failed		NA
DGRP318	chrX_TAS:3014:+	chrX_TAS:3014:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP319	chrX_TAS:2842:+;chrX_TAS:2842:-	chrX_TAS:2842:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chrX_TAS:2842:- TAS-XL-F2 and P-enrich; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP32	chr3R:32073784:-	chr3R:32073784:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55-65°C gradient 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP320	failed		chrX_TAS:2670:+

Table S3.3 continued

DGRP321	failed		chrX_TAS:3695:+
DGRP324	chrX_TAS:3021:+;chr3R:32073784:-	chrX_TAS:3021:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr3R:32073784:- P-element5-R1 and TAS-3R-QPCR-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP332	chrX_TAS:2678:+	chrX_TAS:2678:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP336	chrX_TAS:3014:+	chrX_TAS:3014:+ P-element5-R1 and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP338	failed		chrX_TAS:3014:+
DGRP340	failed		chrX_TAS:3595:-
DGRP348	chr3R:32073784:-	chr3R:32073784:- P-element5-R1 and TAS-3R-QPCR-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	chrX_TAS:3105:+
DGRP350	chr3R:32073784:-	chr3R:32073784:- P-element5-R1 and TAS-3R-QPCR-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP352	chr3R:32073871:-	chr3R:32073871:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP354	chrX_TAS:3695:+	chrX_TAS:3695:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP355	chrX_TAS:3695:+	chrX_TAS:3695:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP356	chr2R:25258728:+; chr2R:25258728:-	chr2R:25258728:+ P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes. chr2R:25258728:+ P-element5-R1 and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes.	NA

Table S3.3 continued

DGRP357	chr3R:32073927:-;chrX_TAS:2742..2752:+	chr3R:32073927:- P-element5-R1 and TAS-3R-QPCR-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chrX_TAS:2742..2752:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP358	untested	NA	NA
DGRP359	chrX_TAS:3014:+;chr2R:25258751:+	chrX_TAS:3014:+ P-enrich and TAS-XL-R ; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr2R:25258751:+ P-enrich and chr2R_25259172; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP360	chrX_TAS:2806:+	chrX_TAS:2806:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP361	chrX_TAS:2663:+;chr2R:25258760:+	chrX_TAS:2663:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr2R:25258760:+ P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes.	NA
DGRP362	untested		chrX_TAS:3014:-
DGRP367	chr2R:25258760:+	chr2R:25258760:+ P-enrich and chr2R_25259172; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP370	chrX_TAS:3014:+	chrX_TAS:3014:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP371	chr3R:32073586:-;chr2R:25258578:-;chr2R:25258578:+;	chr3R:32073586:- P-element5-R1 and TAS-3R-QPCR-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr2R:25258578:- P-element5-R1 and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes. chr2R:25258578:+ P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes.	NA
DGRP373	chrX_TAS:3714:+	chrX_TAS:3714:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA

Table S3.3 continued

DGRP374	chrX_TAS:2842:-;chr3R:32073595:+;chr3R:32073857:-	chrX_TAS:2842:- P-element5-R1 and TAS-XL-R1 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr3R:32073857:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr3R:32073595:+ P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes.	NA
DGRP375	failed		NA
DGRP377	untested		NA
DGRP378	untested		NA
DGRP379	chrX_TAS:3014:+	chrX_TAS:3014:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP38	chrX_TAS:3014:+	chrX_TAS:3014:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP380	untested		NA
DGRP381	chr2R:25258760:+	chr2R:25258760:+ P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP382	failed		chrX_TAS:3575:+
DGRP383	chrX_TAS:3570:+	chrX_TAS:3570:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP385	chrX_TAS:2978:+	chrX_TAS:2978:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP390	chrX_TAS:3014:+;chr3R:32073709:+	chrX_TAS:3014:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes. chr3R:32073709:+ P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP391	chrX_TAS:3735:+	chrX_TAS:3735:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA

Table S3.3 continued

DGRP392	failed		NA
DGRP393	untested		chrX_TAS:2842:-
DGRP395	chrX_TAS:2978:+;chr2R:25258751:-;chr2R:25258751:+	chrX_TAS:2978:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr2R:25258751:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr2R:25258751:+ P-enrich and chr2R_25259172; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP397	chrX_TAS:3596:+	chrX_TAS:3596:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP398	untested		NA
DGRP399	chrX_TAS:2978:+	chrX_TAS:2978:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP40	chr3R:32073855:-	chr3R:32073855:- P-element5-R1 and TAS-3R-QPCR-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 50-65°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP405	failed		chrX_TAS:3206:-
DGRP406	chrX_TAS:3575:-	chrX_TAS:3575:- P-element5-R1 and TAS-XL-R1 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP409	failed		NA
DGRP42	chrX_TAS:3007:+;chr2R:25258760:-	chrX_TAS:3007:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr2R:25258760:- P-element5-R1 and chr2R_25259172; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP426	failed		NA
DGRP427	chrX_TAS:3023:+	chrX_TAS:3023:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA

Table S3.3 continued

DGRP437	chrX_TAS:3695:+	chrX_TAS:3695:+ P-enrich andTAS-XL-R1 ; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP439	chr2R:25258680:-;chr2R:25258765:-	chr2R:25258680:- P-element5-R1 and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes. chr2R:25258765:- P-element5-R1 and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes.	NA
DGRP440	chrX_TAS:3014:+	chrX_TAS:3014:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP441	chrX_TAS:3016:+;chr2R:25258696:+	chrX_TAS:3016:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr2R:25258696:+ P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes.	NA
DGRP443	chr3R:32073855:-	chr3R:32073855:- P-element5-R1 and TAS-3R-QPCR-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP45	chr3R:32073819:-;chrX_TAS:3206:-	chr3R:32073819:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chrX_TAS:3206:- P-element5-R1 and TAS-XL-R1 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 50-65°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP461	failed		NA
DGRP48	chrX_TAS:3014:+	chrX_TAS:3014:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP49	untested		NA
DGRP491	chrX_TAS:3014:+	chrX_TAS:3014:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP492	chrX_TAS:2842:+	chrX_TAS:2842:+ P-enrich and TAS-XL-R ; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP502	chr3R:32073864:-;chrX_TAS:3014:+	chr3R:32073864:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chrX_TAS:3014:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA

Table S3.3 continued

DGRP505	failed		chrX_TAS:2670:+
DGRP508	chrX_TAS:3714:+	chrX_TAS:3714:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP509	failed		NA
DGRP513	chr3R:32073634:-	chr3R:32073634:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP517	chrX_TAS:2934:+;chr3R:32073634:-	chrX_TAS:2934:+ P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr3R:32073634:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP528	chrX_TAS:3695:+	chrX_TAS:3695:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP530	chr2R:25258751:-;chr2R:25258751:+	chr2R:25258751:- P-element5-R1 and chr2R_25259172; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 63°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr2R:25258751:+ P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes.	NA
DGRP531	untested		chrX_TAS:2823:+
DGRP535	chrX_TAS:3695:+	chrX_TAS:3695:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP551	chrX_TAS:2677:+;chr2R:25258827:+;chr3R:32073871:-	chrX_TAS:2677:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr2R:25258827:+ P-enrich and chr2R_25259172; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr3R:32073871:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP554	untested		chrX_TAS:3014:+

Table S3.3 continued

DGRP555	chrX_TAS:3014:+;chr2R:25258760:+	chrX_TAS:3014:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr2R:25258760:+ P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55-65°C gradient 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP559	chrX_TAS:2806:+;chr3R:32073871:-	chrX_TAS:2806:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr3R:32073871:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP563	failed		chrX_TAS:3575:-;chrX_TAS:3596:-
DGRP566	chr2R:25258760:-	chr2R:25258760:- P-element5-R1 and chr2R_25259172; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP57	chr3R:32073864:+	chr3R:32073864:+ P-enrich and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP584	chrX_TAS:3014:+;chr3R:32073595:+;chr3R:32073709:+	chrX_TAS:3014:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes. chr3R:32073595:+ P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes. chr3R:32073595:+ P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP589	chrX_TAS:3082:+	chrX_TAS:3082:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP59	chr3R:32073784:-	chr3R:32073784:- P-element5-R1 and TAS-3R-QPCR-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP591	untested		chrX_TAS:3583:+
DGRP595	failed		chrX_TAS:2670:+
DGRP596	failed		NA

Table S3.3 continued

DGRP627	untested		chrX_TAS:3575:+
DGRP630	chrX_TAS:2810:+	chrX_TAS:2810:+ TAS-XL-F2 and P-element5-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP634	failed		NA
DGRP639	chr2R:25258760:-	chr2R:25258760:- P-element5-R1 and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes.	NA
DGRP642	untested		NA
DGRP646	failed		NA
DGRP69	chrX_TAS:2944:+	chrX_TAS:2944:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	chr3R:32073784:-
DGRP703	chrX_TAS:2798:+;chr2R:25258751:+	chrX_TAS:2798:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr2R:25258751:+ P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes.	NA
DGRP705	chrX_TAS:3595:+;chrX_TAS:3595:-	chrX_TAS:3595:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chrX_TAS:3595:- P-element5-R1 and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP707	chrX_TAS:3073:+	chrX_TAS:3073:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP712	chrX_TAS:3082:+;chrX_TAS:3575:-;chr3R:32073784:-	chrX_TAS:3082:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chrX_TAS:3575:- P-element5-R1 and TAS-XL-R1 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr3R:32073784:- P-element5-R1 and TAS-3R-QPCR-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP714	chrX_TAS:2605:+	chrX_TAS:2605:+ TAS-XL-F2 and P-element5-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA

Table S3.3 continued

DGRP716	chrX_TAS:2678:+;chr3R:32073856:-	chrX_TAS:2678:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr3R:32073856:- P-element5-R1 and TAS-3R-QPCR-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP721	chrX_TAS:3704:+;chr3R:32073515:-	chrX_TAS:3704:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr3R:32073515:- P-element5-R1 and TAS-3R-QPCR-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP727	untested		NA
DGRP73	chrX_TAS:3704:+	chrX_TAS:3704:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP732	chr2R:25258594:-;chr3R:32073864:-	chr2R:25258594:- P-element5-R1 and chr2R_25259172; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr3R:32073864:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP737	chrX_TAS:3695:+	chrX_TAS:3695:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP738	chr3R:32073864:-	chr3R:32073864:- P-element5-R1 and TAS-3R-QPCR-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 50-65°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP748	chrX_TAS:3575:+;chrX_TAS:2670:+	chrX_TAS:3575:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chrX_TAS:2670:+ TAS-XL-F2 and P-element5-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP75	failed		NA
DGRP757	failed		NA
DGRP761	chrX_TAS:2761:+	chrX_TAS:2761:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP765	chrX_TAS:3695:+	chrX_TAS:3695:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA

Table S3.3 continued

DGRP774	chrX_TAS:3548:+	chrX_TAS:3548:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP776	chrX_TAS:3760:-	chrX_TAS:3760:- P-element5-R1 and TAS-XL-R1 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP783	chrX_TAS:3575:-	chrX_TAS:3575:- P-element5-R1 and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	chrX_TAS:2851:+
DGRP786	chr2R:25258756:-;chr3R:32073688:+	chr2R:25258756:- P-element5-R1 and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes. chr3R:32073688:+ P-enrich and TAS-3R-QPCR-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes.	NA
DGRP787	chrX_TAS:2678:+	chrX_TAS:2678:+ TAS-XL-F2 and P-element5-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP790	chrX_TAS:3082:+;chr3R:32073856:-	chrX_TAS:3082:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes. chr3R:32073856:- P-element5-R1 and TAS-3R-QPCR-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP796	chrX_TAS:3575:+	chrX_TAS:3575:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP799	failed		chrX_TAS:2806:+
DGRP801	untested		NA
DGRP802	chrX_TAS:2851:+	chrX_TAS:2851:+ P-enrich and TAS-XL-R ; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP804	failed		NA
DGRP805	untested		NA
DGRP808	failed		NA
DGRP810	chr3R:32073857:-	chr3R:32073857:- P-element5-R1 and TAS-3R-QPCR-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 50-65°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP812	failed		NA

Table S3.3 continued

DGRP818	chrX_TAS:3575:-;chrX_TAS:2842:-	chrX_TAS:3575:- P-element5-R1 and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes. chrX_TAS:2842:- TAS-XL-F2 and P-enrich; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP819	untested		chrX_TAS:3082:+;chrX_TAS:3575:-
DGRP820	chrX_TAS:3695:+;chr2R:25258587:+;chr2R:25258587:-	chrX_TAS:3695:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr2R:25258587:+ P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes. chr2R:25258587:- P-element5-R1 and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes.	NA
DGRP821	chrX_TAS:4004:+	chrX_TAS:4004:+ P-enrich and TAS-XL-R2; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	chrX_TAS:2842:+
DGRP822	chrX_TAS:3014:+	chrX_TAS:3014:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP83	chrX_TAS:3023:+;chr2R:25258587:-	chrX_TAS:3023:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr2R:25258587:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP832	chrX_TAS:2665:+	chrX_TAS:2665:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP837	chrX_TAS:2842:-	chrX_TAS:2842:- TAS-XL-F2 and P-enrich; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP843	failed		chrX_TAS:3014:+
DGRP849	untested		chrX_TAS:2743:+
DGRP85	untested		NA

Table S3.3 continued

DGRP850	failed		chrX_TAS:3014:+
DGRP853	failed		chrX_TAS:3007:+
DGRP855	failed		NA
DGRP857	chr2R:25258680:+	chr2R:25258680:+ P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55-65°C gradient 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP859	chrX_TAS:3014:+;chrX_TAS:3014:-	chrX_TAS:3014:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chrX_TAS:3014:- P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes.	NA
DGRP861	chrX_TAS:2665:+;chr3R:32073784:-	chrX_TAS:2665:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr3R:32073784:- P-element5-R1 and chr3R_32073398; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 55°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP879	chrX_TAS:3014:+	chrX_TAS:3014:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP88	chrX_TAS:3695:+;chr3R:32073595:-	chrX_TAS:3695:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr3R:32073595:- P-element5-R1 and TAS-3R-QPCR-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 50-65°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP882	chrX_TAS:2851:+	chrX_TAS:2851:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP884	untested		NA
DGRP887	chrX_TAS:2851:+	chrX_TAS:2851:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP890	chrX_TAS:3575:+	chrX_TAS:3575:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP892	chrX_TAS:2851:+	chrX_TAS:2851:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA

Table S3.3 continued

DGRP894	chrX_TAS:3073:+	chrX_TAS:3073:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP897	chrX_TAS:2851:+;chr2R:25258587:+	chrX_TAS:2851:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr2R:25258587:+ P-enrich and chr2R_25258834; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 5 minutes.	NA
DGRP900	chrX_TAS:3082:+	chrX_TAS:3082:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 2 minutes; 72°C 10 minutes.	NA
DGRP907	chrX_TAS:2939:+	chrX_TAS:2939:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP908	chr2R:25258751:-;chr3R:32073586:-	chr2R:25258751:- P-element5-R1 and chr2R_25259172; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes. chr3R:32073586:- P-element5-R1 and TAS-3R-QPCR-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 60°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP91	failed		NA
DGRP911	chrX_TAS:3695:+	chrX_TAS:3695:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA
DGRP913	failed		chrX_TAS:2806:+
DGRP93	chrX_TAS:3575:+	chrX_TAS:3575:+ P-enrich and TAS-XL-R1; 95°C 3 minutes; 35 cycles of 95°C 30 seconds, 64°C 30 seconds, 72°C 1.5 minutes; 72°C 10 minutes.	NA

Table S3.4 PCR primers used to determine *P*-element insertion breakpoints

Primer name	Sequence from 5' to 3'
P-element5-F1	TTGCTGCAAAGCTGTGACTG
P-enrich	CACGGACATGCTAAGGGTTAATC
TAS-XL-QPCR-R1	CACACTTACCATAGAGCAAGGG
TAS-XL-QPCR-F2	GACTCATATGTACAACTTTGCCAA
TAS-XL-QPCR-R2	CTTCTGGAAATCTTTTTTCAATTTTTATCG
chr2R_25259172	GCAAACAGTCACGTTGCGAA
chr2R_25258834	ACCGATCAACACGACCATTCA
chr3R_32073398	CGACATATGTGTACGCGTCG
chr3R_32073418	GAATGCGAACAGTCACGCTG
TAS-3R-QPCR-R1	TTTTTGTTAGTTGCTTTTGTAAGATT

Supplemental script1: A shell script used to detect *P*-element insertions from hemi-specific PCR

```
#!/usr/bin/sh

#######################################################################
#################################
#The shell script explains how to detect P-element insertions from
paired-end reads of hemi-specific PCR
#Shuo Zhang (shuozhang23@gmail.com)
#Erin Kelleher lab
#Program in Ecology and Evolution, Departement of Biology and
Biochemistry, University of Houston
#######################################################################
#################################

#To run this script, put your paired-end reads from each degenerate
primer into seperate directories.
#Then, create a directory and move your direacotories with sequencing
data into this created directory, run the shell script after you change
to this directory
module add bowtie2/2.1.0
module add samtools/0.1.19


#The 1st step: align read1 from the paired-end sequencing data to the
P-element consensus reference
for i in {1..15}
do
    cd Sample_lib_R$i  #change to directory containing paired-end reads
of a degenerate primer

    #align read1 from the paired-end sequencing data to the P-element
consensus reference
    #p_element.ref is the basename of the index for the P-element
consensus. For more details, please refer bowtie2 manual
    # read1.fastq is the read1 of paired-end reads
    bowtie2 --local -q --no-unal -x p_element.ref -U read1.fastq -S
read1.sam

    #select reads that align to plus strand of P-element
    samtools view -SXF 0x10 read1.sam > plus_strand_mapped_read1.sam

    #pick read pairs that aligned to 3' end of P-element
    perl pick_p_reads.pl plus_strand_mapped_read1.sam read1.fq r1.fq
lib*R2_001.fastq read2.fq r2.fq
    cat read1.fq r1.fq > p_read1.fq
    cat read2.fq r2.fq > p_read2.fq
    rm read1.fq r1.fq read2.fq r2.fq

    #cut adapter at 3' end and remove 3' end sequence of P element and
filter low quality nucleotides
```

```
    cutadapt  -a CTGTCTCTTATACACATCTCCGAGCCCAC -A
CTGTCTCTTATACACATCTGACGCTGCC  -o cut_p_read1.fq -p cut_p_read2.fq
p_read1.fq p_read2.fq
    cutadapt  -q 20,20 -g CGACGGGACCACCTTATGTTATTTCATCATG  -A
CATGATGAAATAACATAAGGTGGTCCCGTCG --minimum-length 20 -o
trimmed_p_read1.fq -p trimmed_p_read2.fq cut_p_read1.fq cut_p_read2.fq
    rm cut_p_read1.fq cut_p_read2.fq

    #align trimmed read paired to dm6
    bowtie2 -q -a --no-unal -x dm6.ref -1 trimmed_p_read1.fq
trimmed_p_read2.fq -S trimmed_read_all.sam
    samtools view -XSf 0x2  trimmed_read_all.sam  >
proper_trimmed_read_all.sam

    #pick uniquely mapping reads and their alignments
    cut -f 1 proper_trimmed_read_all.sam | uniq -c > alignment.txt
    perl pickUniqMulti.pl alignment.txt proper_trimmed_read_all.sam
uniq.sam multi.sam
    rm alignment.txt

    #detect the P-element insertions according to uniquely mapping
reads
    perl get_insertion_unique.pl -read1_len 30 -read2_len 30 -i
uniq.sam -o insertion.out
    cd ..
done
```

Supplemental script2: A Perl script used to detect *P*-element derived read pairs

```
#!/usr/bin/perl -w
use strict;
use List::Util qw(sum);
###
#This perl script is used to extract reads mapped to 2877-2907 of P-
element( one mismatch and one gap open are allowed)
#Usage: perl $0 out.sam read1.fq r1.fq raw_read2.fq read2.fq r2.fq
#History: 2015-09-08
###
my %read1;
my %read2;
open IN,'<', $ARGV[0] or die "Cannot open the input file:$!";
open OUT1,'>',$ARGV[1] or die "Cannot open the output file:$!";
open OUT2,'>',$ARGV[2] or die "Cannot open the output file:$!";
while (my $line = <IN>){
        chomp($line);
    next if $line=~/^\@/; ###skip header lines;
    ###parse CIGAR
    my @sam=split /\t/,$line,12;
    my (@cigar_m)=$sam[5]=~/(\d+)M/g;
        my (@cigar_d)=$sam[5]=~/(\d+)D/g;
        my (@cigar_s)=$sam[5]=~/(\d+)S/g;
    my (@cigar_i)=$sam[5]=~/(\d+)I/g;
        my $aln_ln=sum(@cigar_m,@cigar_d);
    next if $sam[3]<2850;
    my $end=$sam[3]+$aln_ln-1;
    my ($mismatch)= $sam[11]=~/XM:i:(\d+)/; #get the number of mismatch
    my ($gap) = $sam[11]=~/XO:i:(\d+)/; ###get the number of gap open
    if ($end==2907 and $sam[3] <= 2877 and $mismatch <=1 and $gap <=1
){ #first p element set
      $read1{$sam[0]}=1;
      print OUT1 "\@$sam[0]\n$sam[9]\n";
      print OUT1 "+\n";
      print OUT1 "$sam[10]\n";
      #print "$end\n";
      #print "+\n";
      next;
    }
    if ($end>=2896 and $aln_ln >= 20 and $mismatch <=1 and $gap <=1 ){
#second p elelement set
      $read2{$sam[0]}=1;
      print OUT2 "\@$sam[0]\n$sam[9]\n";
      print OUT2 "+\n";
      print OUT2 "$sam[10]\n";
      #print "$end\n";
                #print "+\n";
                next;
    }
}
```

```
close IN;
close OUT1;
close OUT2;

#reformat Read2 and output Read2
my $temp_file = $ARGV[3].".temp";
open IN,'<',$ARGV[3] or die "Cannot open the input file:$!";
open TEMP, '>',$temp_file or die "Cannot open the temporatory file:
$!\n";
while(my $line = <IN>){
        chomp($line);
        if ( $line =~ /^\@HWI-M/ ){
                print TEMP "\n$line\t";
        }
        else {
                print TEMP "$line\t";
        }
}
close TEMP;
close IN;

open IN, '<', $temp_file or die "Cannot open the temporatory file:
$!\n";
open OUT1,'>',$ARGV[4] or die "Cannot open the read1.fa file:$!\n";
open OUT2,'>',$ARGV[5] or die "Cannot open the read1.fa file:$!\n";
while (my $line=<IN>){
        chomp $line;
        my @items = split /\t/,$line;
        if (!($line =~ /^\@HWI-M/)){ next;} ##skip the empty line
    #$items[0] =~ s/\@HWI-M/>HWI-M/;
        my @names = split /\s/,$items[0];
    $names[0] =~ s/\@HWI-M/HWI-M/;
    if (exists $read1{$names[0]}) {
      print OUT1 "\@$names[0]\n";
      print OUT1 "$items[1]\n";
      print OUT1 "$items[2]\n";
      print OUT1 "$items[3]\n";
      next;
    }
    if (exists $read2{$names[0]}) {
      print OUT2 "\@$names[0]\n";
      print OUT2 "$items[1]\n";
      print OUT2 "$items[2]\n";
      print OUT2 "$items[3]\n";
      next;
    }
}
system("rm *.temp");
close IN;
close OUT1;
close OUT2;
print "The end!\n";
```

Supplementary script3: A Perl script used to pick up concordantly and uniquely mapping read pairs

```
#!/usr/bin/perl -w

###
#Description: This perl script is used to extract alignments with one
or more than one concordantly alignment from bowtie -a;
#Usage: perl $0 [options] -i in.sam --uniq uniq.sam --multi multi.sam
#Author: Shuo Zhang
#History: 2016-05-04
###

my %count;
die "wrong in and out file number: $!" unless @ARGV==4;
open IN,'<',$ARGV[0] or die "Cannot open the alignment count file: $!";
while (my $line=<IN>){
     chomp $line;
     my ($num) = $line=~ /(\d+) HWI/;
     my ($id) = $line=~ /(HWI.*)/;
     die if $num%2 != 0; #the number of alignment should be mutliple of
2
     $count{$id}=$num;
}
close IN;

open SAM,'<',$ARGV[1] or die "Cannot open the input .sam file";
open UNIQ,'>',$ARGV[2] or die "Cannot open the uniq.sam file";
open MULTI,'>',$ARGV[3] or die "Cannot open the multi.sam file";
while (my $line=<SAM>) {
     chomp $line;
     my $id=(split /\t/,$line)[0];
     if ($count{$id}==2){
       print UNIQ "$line\n";
     }
     else {
       print MULTI "$line\n";
     }
}
close SAM;
close UNIQ;
```

Supplementary script4: A Perl script used to determine the breaks of *P*-element insertions

```
#!/usr/bin/perl -w
use strict;
use List::Util qw(sum);
use Getopt::Long;

###
#This script is used to locate P-element insertions according to
alignments in SAM file.
#Usage: perl $0 [options] -i in.sam -o insertion.out
#Shuo Zhang
#History: 2015-11-10; 2015-11-12; 2016-05-17
###

my $read1_len=30; #default mapped read2 lenght: 30bp
my $read2_len=30;  # default mapped read2 length: 30bp
my $infile;
my $outfile;
my $mapped_read;
GetOptions ( 'read1_len:i' => \$read1_len, 'read2_len:i' =>
\$read2_len, 'infile=s' => \$infile, 'outfile=s' => \$outfile,
'mapped_read:s' => \$mapped_read) or die ("Error in command line
arguments\n");
my %insertion;
my %read;
#requirement for read1
open IN,'<', $infile or die "Cannot open the input file:$!";
while(my $line = <IN>){
     chomp($line);
     my @sam=split /\t/,$line;
     my (@cigar_m)=$sam[5]=~/(\d+)M/g;
     my $aln_ln=sum(@cigar_m);
     if ($sam[1]=~/1$/ and $aln_ln >=$read1_len){     # require mapped
length of read1 to be at least 90bp and filter low mapping quality
alignment
     $read{$sam[0]}=1;
     }
}
close IN;

#requirement for read2
open IN,'<', $infile or die "Cannot open the input file:$!";
while(my $line = <IN>){
        chomp($line);
        my @sam=split /\t/,$line;
     next unless exists $read{$sam[0]};
     if ($sam[1]=~/1$/){next;}
     if ($sam[1]=~/2$/){
       my (@cigar_m)=$sam[5]=~/(\d+)M/g;
```

```
                my $aln_ln=sum(@cigar_m);
      #my $rate=$aln_ln/($seq_len-54);
      delete $read{$sam[0]} unless (defined $aln_ln and $aln_ln >=
$read2_len);
     }
}
close IN;

if (defined $mapped_read){
     open READ,'>',$mapped_read or die "Cannot open the mapped_read
output file\n";
     foreach my $key (sort keys %read){
      print READ "$key\n";
     }
     close READ;
}


open IN,'<', $infile or die "Cannot open the input file:$!";
while(my $line = <IN>){
     chomp($line);
     my @sam=split /\t/,$line;
     next if $sam[1]=~/2$/;
     if (exists $read{$sam[0]}){
      my $strand="sense";
      if ($sam[1] =~/R/){
            my $coor = $sam[3] + 7; #breakpoint
            ${${$insertion{$sam[2]}}{$coor}}[0]++;
            ${${$insertion{$sam[2]}}{$coor}}[1]++;
            }
      elsif ($sam[1] =~/r/){
            $strand="antisense";
                        my (@cigar_m)=$sam[5]=~/(\d+)M/g;
                        my (@cigar_d)=$sam[5]=~/(\d+)D/g;
                        #my (@cigar_s)=$sam[5]=~/(\d+)S/g;
                        #my (@cigar_i)=$sam[5]=~/(\d+)I/g;
                        my $aln_ln=sum(@cigar_m,@cigar_d);
                        my $coor=$sam[3]+$aln_ln-1;
            ${${$insertion{$sam[2]}}{$coor}}[0]++;
            ${${$insertion{$sam[2]}}{$coor}}[2]++;
      }
      else { die "wrong sam file";}
     }
}
close IN;

###formatted print
open OUT,'>',$outfile or die "Cannot open the output file:$!";
print OUT "chromosome\tlocation\treads_support\tstrand\n";
foreach my $ele (sort keys %insertion){
     my @ele= keys %{$insertion{$ele}};
     my @ele_sorted = sort {$a <=> $b} @ele;
```

```
    foreach my $site (@ele_sorted ){
      my $num = ${${$insertion{$ele}}{$site}}[0];
      my $strand;
      if (defined ${${$insertion{$ele}}{$site}}[1] and
${${$insertion{$ele}}{$site}}[0] == ${${$insertion{$ele}}{$site}}[1]){
$strand = "sense";print OUT "$ele\t$site\t$num\t$strand\n"}
      elsif (defined ${${$insertion{$ele}}{$site}}[2] and
${${$insertion{$ele}}{$site}}[0] == ${${$insertion{$ele}}{$site}}[2]){
$strand = "antisense";print OUT "$ele\t$site\t$num\t$strand\n"}
      else {print OUT
"$ele\t$site\t${${$insertion{$ele}}{$site}}[1]\tsense\n"; print OUT
"$ele\t$site\t${${$insertion{$ele}}{$site}}[2]\tantisense\n";}
    }
}
close OUT;
close IN;
```

References

Adrion JR, Song MJ, Schrider DR, Hahn MW, Schaack S. 2017. Genome-wide estimates of transposable element insertion and deletion rates in *Drosophila melanogaster*. *Genome Biology and Evolution* **9**: 1329-1340.

Ajioka JW, Eanes WF. 1989. The accumulation of P-elements on the tip of the X chromosome in populations of *Drosophila melanogaster*. *Genet Res* **53**: 1–6.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Aminetzach YT, Macpherson JM, Petrov DA. 2005. Evolution: Pesticide resistance via transposition-mediated adaptive gene truncation in *Drosophila*. *Science* **309**: 764-767.

Anderson JA, Song YS, Langley CH. 2008. Molecular population genetics of *Drosophila* subtelomeric DNA. *Genetics* **178**: 477–487.

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

Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, Iovino N, Morris P, Brownstein MJ, Kuramochi-Miyagawa S, Nakano T et al. 2006. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442**: 203-207.

Aravin AA, Hannon GJ, Brennecke J. 2007a. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318**: 761–764.

Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, Bestor T, Hannon GJ. 2008. A piRNA pathway primed by individual transposons is linked to *de novo* DNA methylation in mice. *Molecular Cell* **31**: 785-799.

Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ. 2007b. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* **316**: 744-747.

Asif-Laidin A, Delmarre V, Laurentie J, Miller WJ, Ronsseray S, Teysset L. 2017. Short and long-term evolutionary dynamics of subtelomeric piRNA clusters in *Drosophila*. *DNA Research* **24**: 459-472.

Bagijn MP, Goldstein LD, Sapetschnig A, Weick E-M, Bouasker S, Lehrbach NJ, Simard MJ, Miska EA. 2012. Function, targets, and evolution of *Caenorhabditis elegans* piRNAs. *Science* **337**: 574-578.

Baidouri M El, Carpentier MC, Cooke R, Gao D, Lasserre E, Llauro C, Mirouze M, Picault N, Jackson SA, Panaud O. 2014. Widespread and frequent horizontal transfers of transposable elements in plants. *Genome Res* **24**: 831-838.

Barrón MG, Fiston-Lavier A-S, Petrov DA, González J. 2014. Population genomics of transposable elements in *Drosophila*. *Annu Rev Genet* **48**: 561-581.

Bartolomé C, Bello X, Maside X. 2009. Widespread evidence for horizontal transfer of transposable elements across *Drosophila* genomes. *Genome Biol* **10**: R22.

Bartolomé C, Maside X. 2004. The lack of recombination drives the fixation transposable elements on the fourth chromosome of *Drosophila melanogaster*. *Genet Res* **83**: 91-100.

Batista PJ, Ruby JG, Claycomb JM, Chiang R, Fahlgren N, Kasschau KD, Chaves DA, Gu W, Vasale JJ, Duan S et al. 2008. PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in *C. elegans*. *Mol Cell* **31**: 67-78.

Beall EL, Rio DC. 1997. *Drosophila* P-element transposase is a novel site-specific endonuclease. *Genes Dev* **11**: 2137–2151.

Bennett EA, Coleman LE, Tsui C, Pittard WS, Devine SE. 2004. Natural genetic variation caused by transposable elements in humans. *Genetics* **168**: 933-951.

Berg CA, Spradling AC. 1991. Studies on the rate and site-specificity of P element transposition. *Genetics* **127**: 515-524.

Bergman CM. 2012. A proposal for the reference-based annotation of *de novo* transposable element insertions. *Mob Genet Elements* **2**: 51-54.

Bergman CM, Quesneville H, Anxolabéhère D, Ashburner M. 2006. Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. *Genome Biol* **7**: R112.

Bessereau J-L. 2006. Transposons in *C. elegans*. *WormBook*, ed. The *C. elegans* Research Community, WormBook, doi/10.1895/wormbook.1.70.1, http://www.wormbook.org.

Bhaya D, Davison M, Barrangou R. 2011. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annu Rev Genet* **45**: 273-297.

Biémont C, Ronsseray S, Anxolabéhère D, Izaabel H, Gautier C. 1990. Localization of P elements, copy number regulation, and cytotype determination in *Drosophila melanogaster*. *Genet Res* **56**: 3-14.

Bingham PM, Kidwell MG, Rubin GM. 1982. The molecular basis of P-M hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family. *Cell* **29**: 995–1004.

Blumenstiel JP. 2011. Evolutionary dynamics of transposable elements in a small RNA world. *Trends Genet* **27**: 23-31.

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

Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, Hannon GJ. 2008. An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science* **322**: 1387–1392.

Bucheton A, Paro R, Sang HM, Pelisson A, Finnegan DJ. 1984. The molecular basis of I-R hybrid Dysgenesis in *Drosophila melanogaster*: identification, cloning, and properties of the I factor. *Cell* **38**: 153–163.

Burns KH. 2017. Transposable elements in cancer. *Nat Rev Cancer* **17**: 415-424.

Chakraborty M, VanKuren NW, Zhao R, Zhang X, Kalsow S, Emerson JJ. 2018. Hidden genetic variation shapes the structure of functional elements in *Drosophila*. *Nat Genet* **50**: 20-25.

Charlesworth B, Langley CH. 1989. The population genetics of *Drosophila* transposable elements. *Annu Rev Genet* **23**: 251-287.

Charlesworth B, Langley CH. 1986. The evolution of self-regulated transposition of transposable elements. *Genetics* **112**: 359-383.

Chen J, Wrightsman TR, Wessler SR, Stajich JE. 2017. RelocaTE2: a high resolution transposable element insertion site mapping tool for population resequencing. *PeerJ* **5**: e2942.

Chénais B. 2013. Transposable elements and human cancer: A causal relationship? *Biochim Biophys Acta - Rev Cancer* **1835**: 28–35.

Chénais B, Caruso A, Hiard S, Casse N. 2012. The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene* **509**: 7–15.

Comeron JM, Ratnappan R, Bailin S. 2012. The many landscapes of recombination in *Drosophila melanogaster*. *PLoS Genet* **8**: e1002905.

Cridland JM, Macdonald SJ, Long AD, Thornton KR. 2013. Abundance and distribution of transposable elements in two *Drosophila* QTL mapping resources. *Mol Biol Evol* **30**: 2311-2327.

Daborn PJ, Yen JL, Bogwitz MR, Le Goff G, Feil E, Jeffers S, Tijet N, Perry T, Heckel D, Batterham P et al. 2002. A single p450 allele associated with insecticide resistance in *Drosophila*. *Science* **297**: 2253-2256.

Daniels SB, Peterson KR, Strausbaugh LD, Kidwell MG, Chovnik A. 1990. Evidence for horizontal transmission of the P transposable element between *Drosophila* species. *Genetics* **124**: 339–355.

De Vanssay A, Bougé AL, Boivin A, Hermant C, Teysset L, Delmarre V, Antoniewski C, Ronsseray S. 2012. Paramutation in *Drosophila* linked to emergence of a piRNA-producing locus. *Nature* **490**: 112–115.

Dos Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, Emmert DB, Gelbart WM, FlyBase C. 2015. FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res* **43**: D690-697.

Dotto BR, Carvalho EL, Silva AF, Duarte Silva LF, Pinto PM, Ortiz MF, Wallau GL. 2015. HTT-DB: horizontally transferred transposable elements database. *Bioinformatics* **31**: 2915-2917.

Duc C, Yoth M, Jensen S, Mouniée N, Bergman CM, Vaury C, Brasset E. 2019. Trapping a somatic endogenous retrovirus into a germline piRNA cluster immunizes the germline against further invasion. *Genome Biol* **20**: 127.

Eggleston WB, Schlitz DMJ, Engels WR. 1988. P-M hybrid dysgenesis does not mobilize other transposable element families in *D. melanogaster*. *Nature* **331**: 368-370.

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I. 2002. Identification of a variant associated with adult-type hypolactasia. *Nature Genetics* **30**: 233-237.

Evgen'ev MB, Zelentsova H, Shostak N, Kozitsina M, Barskyi V, Lankenau DH, Corces VG. 1997. Penelope, a new family of transposable elements and its possible role in hybrid dysgenesis in *Drosophila virilis*. *Proc Natl Acad Sci U S A* **94**: 196–201.

Ewing AD. 2015. Transposable element detection from whole genome sequence data. *Mob DNA* **6**: 24.

Ewing AD, Kazazian HH. 2010. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res* **20**: 1262–1270.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**: 397-405.

Fiston-Lavier A-S, Barrón MG, Petrov DA, González J. 2015. T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. *Nucleic Acids Res* **43**: e22.

Fiston-Lavier AS, Carrigan M, Petrov DA, Gonzalez J. 2011. T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. *Nucleic Acids Res* **39**: e36.

Gasior SL, Wakeman TP, Xu B, Deininger PL. 2006. The human LINE-1 retrotransposon creates DNA double-strand breaks. *J Mol Biol* **357**: 1383–1393.

Gilbert C, Peccoud J, Chateigner A, Moumen B, Cordaux R, Herniou EA. 2016. Continuous influx of genetic material from host to virus populations. *PLoS Genet* **12**: e1005838.

Gilbert C, Schaack S, Pace JK, Brindley PJ, Feschotte C. 2010. A role for host-parasite interactions in the horizontal transfer of transposons across phyla. *Nature* **464**: 1347-1350.

Girard A, Hannon GJ. 2008. Conserved themes in small-RNA-mediated transposon control. *Trends Cell Biol* **18**: 136-148.

Grenier JK, Arguello JR, Moreira MC, Gottipati S, Mohammed J, Hackett SR, Boughton R, Greenberg AJ, Clark AG. 2015. Global diversity lines-a five-continent reference panel of sequenced *Drosophila melanogaster* Strains. *G3-Genes Genom Genet* **5**: 593-603.

Haag-Liautard C, Dorris M, Maside X, Macaskill S, Halligan DL, Houle D, Charlesworth B, Keightley PD. 2007. Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature* **455**: 82-85.

Haddrill PR, Halligan DL, Tomaras D, Charlesworth B. 2007. Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. *Genome Biol* **8:** R18.

Han BW, Wang W, Li CJ, Weng ZP, Zamore PD. 2015. piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. *Science* **348**: 817-821.

Hayashi R, Schnabl J, Handler D, Mohn F, Ameres SL, Brennecke J. 2016. Genetic and mechanistic diversity of piRNA 3′-end formation. *Nature* **539**: 588-592.

Hedges DJ, Deininger PL. 2007. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutat Res* **616**: 46-59.

Hellsten U, Harland RM, Gilchrist MJ, Hendrix D, Jurka J, Kapitonov V, Ovcharenko I, Putnam NH, Shu S, Taher L et al. 2010. The genome of the Western clawed frog *Xenopus tropicalis*. *Science* **328**: 633-636.

Hermant C, Boivin A, Teysset L, Delmarre V, Asif-Laidin A, van den Beek M, Antoniewski C, Ronsseray S. 2015. Paramutation in drosophila requires both nuclear and cytoplasmic actors of the piRNA pathway and induces cis-spreading of piRNA production. *Genetics* **201**: 1381-1396.

Hill T, Schlötterer C, Betancourt AJ. 2016. Hybrid Dysgenesis in *Drosophila simulans* Associated with a Rapid Invasion of the P-Element. *PLoS Genet* **12**: e1006058.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet Res* **8**: 269–294.

Höck J, Meister G. 2008. The Argonaute protein family. *Genome Biol* **9**: 210.

Hof AEvt, Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, Hall N, Darby AC, Saccheri IJ. 2016. The industrial melanism mutation in British peppered moths is a transposable element. *Nature* **534**: 102-105.

Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res* **19**: 1419–1428.

Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, Booth BW, Pfeiffer BD, George RA, Svirskas R, et al. 2015. The Release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome Res* **25**: 445–458.

Houck M, Clark J, Peterson K, Kidwell M. 1991. Possible horizontal transfer of Drosophila genes by the mite *Proctolaelaps regalis*. *Science* **253**: 1125-1128.

Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, Filippov DV, Blaser H, Raz E, Moens CB et al. 2007. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in zebrafish. *Cell* **129**: 69-82.

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**: 498-503.

Huang W, Massouras A, Inoue Y, Peiffer J, Ràmia M, Tarone AM, Turlapati L, Zichner T, Zhu D, Lyman RF, et al. 2014. Natural variation in genome architecture among 205 *Drosophila melanogaster* genetic reference panel lines. *Genome Res* **24**: 1193-1208.

Huang XA, Yin H, Sweeney S, Raha D, Snyder M, Lin H. 2013. A major epigenetic programming mechanism guided by piRNAs. *Dev Cell* **24**: 502-516.

International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. *Nature* **436**: 793-800.

Ishikawa A, Kabeya N, Ikeya K, Kakioka R, Cech JN, Osada N, Leal MC, Inoue J, Kume M, Toyoda A, et al. 2019. A key metabolic gene for recurrent freshwater colonization and radiation in fishes. *Science* **364**: 886—889.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**: 338-345.

Jang HS, Shah NM, Du AY, Dailey ZZ, Pehrsson EC, Godoy PM, Zhang D, Li D, Xing X, Kim S, et al. 2019. Transposable elements drive widespread expression of oncogenes in human cancers. *Nat Genet* **51**: 611–617.

Jehn J, Gebert D, Pipilescu F, Stern S, Kiefer JST, Hewel C, Rosenkranz D. 2018. PIWI genes and piRNAs are ubiquitously expressed in mollusks and show patterns of lineage-specific adaptation. *Commun Biol* **1**: 137-137.

Jensen PA, Stuart JR, Goodpaster MP, Goodman JW, Simmons MJ. 2008. Cytotype regulation of P transposable elements in *Drosophila melanogaster*: Repressor polypeptides or piRNAs? *Genetics* **179**: 1785–1793.

Jjingo D, Conley AB, Wang J, Mariño-Ramírez L, Lunyak V V, Jordan IK. 2014. Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. *Mob DNA* **5**: 14.

Karasov T, Messer PW, Petrov DA. 2010. Evidence that adaptation in *Drosophila* is not limited by mutation at single sites. *PLoS Genet* **6**: e1000924.

Karpen GH, Spradling AC. 1992. Analysis of subtelomeric heterochromatin in the *Drosophila* minichromosome Dp1187 by single P element insertional mutagenesis. *Genetics* **132**: 737–753.

The Arabidopsis Genome I. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796-815.

Kazazian HH, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. 1988. Haemophilia A resulting from *de novo* insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**: 164-166.

Keane TM, Wong K, Adams DJ. 2013. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics* **29**: 389–90.

Kelleher ES. 2016. Reexamining the P-element invasion of *Drosophila melanogaster* Through the Lens of piRNA Silencing. *Genetics* **203**: 1513–31.

Kelleher ES, Azevedo RBR, Zheng Y. 2018. The evolution of small-RNA-mediated silencing of an invading transposable element. *Genome Biol Evol* **10**: 3038–3057.

Khurana JS, Wang J, Xu J, Koppetsch BS, Thomson TC, Nowosielska A, Li C, Zamore PD, Weng Z, Theurkauf WE. 2011. Adaptation to P element transposon invasion in *Drosophila melanogaster*. *Cell* **147**: 1551-1563.

Kidwell MG. 1983. Evolution of hybrid dysgenesis determinants in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A* **80**: 1655–1659.

Kidwell MG, Frydrk T, Novy J. 1983. The hybrid dysgenesis potential of *Drosophila melanogaster* from diverse temporal and geographic origins. *Drosoph Inf Serv* **59**: 63–69.

Kidwell MG, Kidwell JF, Sved JA. 1977. Hybrid dysgenesis in *Drosophila melanogaster*: a syndrome of aberrant traits including mutation, sterility and male recombination. *Genetics* **86**: 813–833.

Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF. 1998. Transposable elements and genome organization: A comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res* **8**: 464-478.

Kimura K, Kidwell MG. 2009. Differences in P element population dynamics between the sibling species *Drosophila melanogaster* and *Drosophila simulans*. *Genetical Research* **63**: 27-38.

Kofler R. 2019. Dynamics of transposable element invasions with piRNA clusters. *Molecular Biology and Evolution* **36**: 1457-1472.

Kofler R, Betancourt AJ, Schlötterer C. 2012. Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. *PLoS Genet* **8**: e1002487.

Kofler R, Hill T, Nolte V, Betancourt AJ, Schlötterer C. 2015a. The recent invasion of natural *Drosophila simulans* populations by the P-element. *Proc Natl Acad Sci* **112**: 6659-6663.

Kofler R, Nolte V, Schlötterer C. 2015b. Tempo and mode of transposable element activity in *Drosophila*. *PLoS Genet* **11**: e1005406.

Kofler R, Senti K-A, Nolte V, Tobler R, Schlötterer C. 2018. Molecular dissection of a natural transposable element invasion. *Genome Research* doi:10.1101/gr.228627.117.

Kumar MS, Chen KC. 2012. Evolution of animal Piwi-interacting RNAs and prokaryotic CRISPRs. *Briefings in Functional Genomics* **11**: 277-288.

Lack JB, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, Stevens KA, Langley CH, Pool JE. 2015. The drosophila genome nexus: A population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. *Genetics* **199**: 1229-1241.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Le Thomas A, Rogers AK, Webster A, Marinov GK, Liao SE, Perkins EM, Hur JK, Aravin AA, Tóth KF. 2013. Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. *Genes Dev* **27**: 390–399.

Le Thomas A, Stuwe E, Li S, Du J, Marinov G, Rozhkov N, Chen YCA, Luo Y, Sachidanandam R, Toth KF, et al. 2014. Transgenerationally inherited piRNAs trigger piRNA biogenesis by changing the chromatin of piRNA clusters and inducing precursor processing. *Genes Dev* **28**: 1667–1680.

Lee S-I, Kim N-S. 2014. Transposable elements and genome size variations in plants. *Genomics Inform* **12**: 87–97.

Lee YCG. 2015. The role of piRNA-mediated epigenetic silencing in the population dynamics of transposable elements in *Drosophila melanogaster*. *PLOS Genet* **11**: e1005269.

Lee YCG, Langley CH. 2012. Long-term and short-term evolutionary impacts of transposable elements on *Drosophila*. *Genetics* **192**: 1411–32.

Levis R, O'Hare K, Rubin GM. 1984. Effects of transposable element insertions on RNA encoded by the white gene of Drosophila. *Cell* **38**: 471–481.

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

Lim JK. 1988. Intrachromosomal rearrangements mediated by hobo transposons in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A* **85**: 9153–9157.

Lin X, Faridi N, Casola C. 2016. An ancient transkingdom horizontal transfer of Penelope -like retroelements from arthropods to conifers. *Genome Biology and Evolution* **8**: 1252-1266.

Linheiro RS, Bergman CM. 2012. Whole genome resequencing reveals natural target site preferences of transposable elements in *Drosophila melanogaster*. *PLoS One* **7**: e30008.

Lo P-K, Huang Y-C, Poulton JS, Leake N, Palmer WH, Vera D, Xie G, Klusza S, Deng W-M. 2016. RNA helicase Belle/DDX3 regulates transgene expression in *Drosophila*. *Developmental Biology* **412**: 57-70.

Lu J, Clark AG. 2010. Population dynamics of PIWI-interacting RNAs (piRNAs) and their targets in *Drosophila*. *Genome Research* **20**: 212-227.

Lynch VJ, Leclerc RD, May G, Wagner GP. 2011. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat Genet* **43**: 1154–9.

Mackay TFC, Richards S, Stone E a, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The *Drosophila melanogaster* genetic reference panel. *Nature* **482**: 173–8.

Magwire MM, Bayer F, Webster CL, Cao C, Jiggins FM. 2011. Successive increases in the resistance of *Drosophila* to viral infection through a transposon insertion followed by a duplication. *PLoS Genet* **7**: e1002337

Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R, Hannon GJ. 2009. Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. *Cell* **137**: 522-535.

Malone CD, Hannon GJ. 2009. Small RNAs as guardians of the genome. *Cell* **136**: 656-668.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel

counting of occurrences of k-mers. *Bioinformatics* **27**: 764–770.

Marin L, Lehmann M, Nouaud D, Izaabel H, Anxolabéhère D, Ronsseray S. 2000a. P-element repression in *Drosophila melanogaster* by a naturally occurring defective telomeric P copy. *Genetics* **155**: 1841–1854.

Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, Morin H, Pitrat M, Dogimont C, Bendahmane A. 2009. A transposon-induced epigenetic change leads to sex determination in melon. *Nature* **461**: 1135-1138.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10–12.

Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favorable gene. *Genet Res* **23**: 23–35.

McClintock B. 1953. Induction of instability at selected Loci in maize. *Genetics* **38**: 579-599.

McClintock B. 1950. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences* **36**: 344-355.

McGinnis W, Shermoen AW, Beckendorf SK. 1983. A transposable element inserted just 5′ to a Drosophila glue protein gene alters gene expression and chromatin structure. *Cell* **34**: 75–84.

McInerney P, Adams P, Hadi MZ. 2014. Error rate comparison during polymerase chain reaction by DNA polymerase. *Mol Biol Int* **2014**: 287430.

Menozzi P, Shi MA, Lougarre A, Tang ZH, Fournier D. 2004. Mutations of acetylcholinesterase which confer insecticide resistance in *Drosophila melanogaster* populations. *BMC Evol Biol* **4**: 4-4.

Messer PW, Petrov DA. 2013. Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution* **28**: 659-669.

Mohn F, Sienski G, Handler D, Brennecke J. 2014. The Rhino-Deadlock-Cutoff complex licenses noncanonical transcription of dual-strand piRNA clusters in *Drosophila*. *Cell* **157**: 1364-1379.

Muerdter F, Olovnikov I, Molaro A, Rozhkov N V., Czech B, Gordon A, Hannon GJ, Aravin AA. 2012. Production of artificial piRNAs in flies and mice. *RNA* **18**: 42–52.

Mullins MC, Rio DC, Rubin GM. 1989. cis-acting DNA sequence requirements for P-element transposition. *Genes Dev* **3**: 729–38.

Nakagome M, Solovieva E, Takahashi A, Yasue H, Hirochika H, Miyao A. 2014. Transposon Insertion Finder (TIF): a novel program for detection of *de novo* transpositions of transposable elements. *BMC Bioinformatics* **15**: 71-71.

Nielsen R. 2005. Molecular signatures of natural selection. *Annu Rev Genet* **39**: 197-218.

Niu XM, Xu YC, Li ZW, Bian YT, Hou XH, Chen JF, Zou YP, Jiang J, Wu Q, Ge S et al. 2019. Transposable elements drive rapid phenotypic variation in *Capsella rubella*. *Proc Natl Acad Sci U S A* **116**: 6908-6913.

Nuzhdin SV, Mackay TF. 1995. The genomic rate of transposable element movement in *Drosophila melanogaster*. *Mol Biol Evol* **12**: 180-181.

Nuzhdin S V. 1999. Sure facts, speculations, and open questions about the evolution of transposable element copy number. *Genetica* **107**:129.

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

Ogura K, Woodruff RC, Itoh M, Boussy IA. 2007. Long-term patterns of genomic P element content and P-M characteristics of *Drosophila melanogaster* in eastern Australia. *Genes Genet Syst* **82**: 479-487.

Ozata DM, Gainetdinov I, Zoch A, O'Carroll D, Zamore PD. 2019. PIWI-interacting RNAs: small RNAs with big functions. *Nat Rev Genet* **20**: 89-108.

Pane A, Jiang P, Zhao DY, Singh M, Schupbach T. 2011. The Cutoff protein regulates piRNA cluster expression and piRNA production in the *Drosophila* germline. *EMBO J* **30**: 4601-4615.

Pardue ML, DeBaryshe PG. 2003. Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. *Annu Rev Genet* **37**: 485-511.

Peccoud J, Loiseau V, Cordaux R, Gilbert C. 2017. Massive horizontal transfer of transposable elements in insects. *Proc Natl Acad Sci* **114**: 4721–4726.

Pennings PS, Hermisson J. 2006. Soft sweeps II - Molecular population genetics of adaptation from recurrent mutation or migration. *Mol Biol Evol* **23**: 1076–1084.

Petrov DA, Fiston-Lavier A-S, Lipatov M, Lenkov K, González J. 2011. Population genomics of transposable elements in *Drosophila melanogaster*. *Mol Biol Evol* **28**: 1633–1644.

Platt RN, Zhang Y, Witherspoon DJ, Xing J, Suh A, Keith MS, Jorde LB, Stevens RD, Ray

DA. 2015. Targeted capture of phylogenetically informative Ves SINE Insertions in genus myotis. *Genome Biol Evol* **7**: 1664–75.

Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, Duchen P, Emerson JJ, Saelao P, Begun DJ et al. 2012. Population Genomics of sub-saharan *Drosophila melanogaster*: African diversity and non-African admixture. *Plos Genet* **8**: e1003080.

Prada CF, Laissue P. 2014. A high resolution map of mammalian X chromosome fragile regions assessed by large-scale comparative genomics. *Mamm Genome* **25**: 618–635.

Quail M, Smith ME, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y, Rothberg J, et al. 2012. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics* **13**: 341.

Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, Hurles ME, Mell JC, Hall IM. 2010. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res* **20**: 623-635.

Rahman R, Chirn GW, Kanodia A, Sytnikova YA, Brembs B, Bergman CM, Lau NC. 2015. Unique transposon landscapes are pervasive across *Drosophila melanogaster* genomes. *Nucleic Acids Res* **43**: 10655–10672.

Reiss D, Josse T, Anxolabehere D, Ronsseray S. 2004. aubergine mutations in *Drosophila melanogaster* impair P cytotype determination by telomeric P elements inserted in heterochromatin. *Mol Genet Genomics* **272**: 336-343.

Rishishwar L, Tellez Villa CE, Jordan IK. 2015. Transposable element polymorphisms recapitulate human evolution. *Mob DNA* **6**: 21.

Robb SMC, Lu L, Valencia E, Burnette JM, Okumoto Y, Wessler SR, Stajich JE. 2013b. The use of RelocaTE and unassembled short reads to produce high-resolution snapshots of transposable element generated diversity in Rice. *G3-Genes Genom Genet* **3**: 949–957.

Robberecht C, Voet T, Zamani Esteki M, Nowakowska BA, Vermeesch JR. 2013. Nonallelic homologous recombination between retrotransposable elements is a driver of *de novo* unbalanced translocations. *Genome Res* **23**: 411–418.

Robert V, Prud'homme N, Kim A, Bucheton A, Pélisson A. 2001. Characterization of the flamenco region of the *Drosophila melanogaster* genome. *Genetics* **158**: 701–713.

Ronsseray S, Lehmann M, Anxolabehere D. 1991. The maternally inherited

regulation of P elements in *Drosophila melanogaster* can be elicited by two P copies at cytological site 1A on the X chromosome. *Genetics* **129**: 501–512.

Ronsseray S, Lehmann M, Anxolabéhère D. 1989. Copy number and distribution of P and I mobile elements in *Drosophila melanogaster* populations. *Chromosoma* **98**: 207–214.

Ronsseray S, Lehmann M, Nouaud D, Anxolabéhère D. 1996. The regulatory properties of autonomous subtelomeric P elements are sensitive to a Suppressor of variegation in *Drosophila melanogaster*. *Genetics* **143**: 1663–1674.

Rosenkranz D, Zischler H. 2012. proTRAC - a software for probabilistic piRNA cluster detection, visualization and analysis. *BMC Bioinformatics* **13**: 5.

Rubin GM, Kidwell MG, Bingham PM. 1982. The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. *Cell* **29**: 987–994.

Sarilar V, Bleykasten-Grosshans C, Neuvéglise C. 2014. Evolutionary dynamics of hAT DNA transposon families in *Saccharomycetaceae*. *Genome Biol Evol* **7**: 172–190.

Schaack S, Gilbert C, Feschotte C. 2010. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol Evol* **25**: 537-546.

Schnable P, Ware D, Fulton R, Stein J. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science (80-)* **326**: 1112–1115.

Serrato-Capuchina A, Matute DR. 2018. The role of transposable elements in speciation. *Genes (Basel)* **9:** 254.

Shpiz S, Ryazansky S, Olovnikov I, Abramov Y, Kalmykova A. 2014. Euchromatic transposon insertions trigger production of novel pi- and endo-siRNAs at the target sites in the *Drosophila* Germline. *PLoS Genet* **10**: e1004138.

Sienski G, Donertas D, Brennecke J. 2012. Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. *Cell* **151**: 964-980.

Sigman MJ, Slotkin RK. 2016. The first rule of plant transposable element silencing: location, location, location. *Plant Cell* **28**: 304-313.

Simmons MJ, Raymond JD, Niemi JB, Stuart JR, Merriman PJ. 2004. The P cytotype in *Drosophila melanogaster*: a maternally transmitted regulatory state of the germ line associated with telomeric P elements. *Genetics* **166**: 243-254.

Simmons MJ, Ryzek DF, Lamour C, Goodman JW, Kummer NE, Merriman PJ. 2007. Cytotype regulation by telomeric P elements in *Drosophila melanogaster*: evidence for involvement of an RNA interference gene. *Genetics* **176**: 1945-1955.

Siomi MC, Sato K, Pezic D, Aravin AA. 2011. PIWI-interacting small RNAs: The vanguard of genome defence. *Nat Rev Mol Cell Biol* **12**: 246–258.

Slotkin RK, Vaughn M, Borges F, Tanurdžić M, Becker JD, Feijó JA, Martienssen RA. 2009. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell* **136**: 461-472.

Smith CD, Shu S, Mungall CJ, Karpen GH. 2007. The Release 5.1 annotation of *Drosophila melanogaster* heterochromatin. *Science* **316**: 1586-1591.

Song J, Liu J, Schnakenberg SL, Ha H, Xing J, Chen KC. 2014. Variation in piRNA and transposable element content in strains of *Drosophila melanogaster*. *Genome Biol Evol* **6**: 2786-2798.

Srivastav SP, Kelleher ES. 2017. Paternal induction of hybrid dysgenesis in *Drosophila melanogaster* is weakly correlated with both P-element and hobo element dosage. *G3-Genes Genom Genet* **7**: 1487-1497.

Stamatis N, Monastirioti M, Yannopoulos G, Louis C. 1989. The P-M and the 23.5 MRF (hobo) systems of hybrid dysgenesis in *Drosophila melanogaster* are independent of each other. *Genetics* **123**: 379-387.

Startek M, Szafranski P, Gambin T, Campbell IM, Hixson P, Shaw CA, Stankiewicz P, Gambin A. 2015. Genome-wide analyses of LINE-LINE-mediated nonallelic homologous recombination. *Nucleic Acids Res* **43**: 2188–98.

Stuart JR, Haley KJ, Swedzinski D, Lockner S, Kocian PE, Merriman PJ, Simmons MJ. 2002. Telomeric P elements associated with cytotype regulation of the P transposon family in *Drosophila melanogaster*. *Genetics* **162**: 1641–1654.

Sun C, Mueller RL. 2014. Hellbender genome sequences shed light on genomic expansion at the base of crown salamanders. *Genome Biol Evol* **6**: 1818–1829.

Sun C, Shepard DB, Chong RA, López Arriaza J, Hall K, Castoe TA, Feschotte C, Pollock DD, Mueller RL. 2012. LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. *Genome Biol Evol* **4**: 168–183.

Team RDC, R Development Core Team R. 2016. R: A language and environment for statistical computing. *R Found Stat Comput*.

Team Rs. 2015. RStudio: Integrated development for R. *RStudio, Inc, Boston, MA URL*

http//www rstudio com.

Thomas J, Schaack S, Pritham EJ. 2010. Pervasive horizontal transfer of rolling-circle transposons among animals. *Genome Biology and Evolution* **2**: 656-664.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M et al. 2006. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics* **39**: 31-40.

Vitte C, Panaud O. 2005. LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenet Genome Res* **110**: 91–107.

Wakisaka KT, Ichiyanagi K, Ohno S, Itoh M. 2017. Diversity of P-element piRNA production among M' and Q strains and its association with P-M hybrid dysgenesis in *Drosophila melanogaster*. *Mobile DNA* **8**: 13.

Walter MF, Jang C, Kasravi B, Donath J, Mechler BM, Mason JM, Biessmann H. 1995. DNA organization and polymorphism of a wild-type *Drosophila* telomere region. *Chromosoma* 104: 229-241.

Warnes GR, Bolker B, Bonebakker L, et al. 2016. gplots: various R programming tools for plotting data. *R Package version 3.0.1*. https://cran.r-project.org/web/packages/gplots/

Weirather J, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang X, Buck D, Au K. 2017. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore technologies and their applications to transcriptome analysis. *F1000Research* **6**.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**: 973-982.

Wickham H. 2017a. Flexibility reshape data: a reboot of the reshape package. *R CRAN Proj*.

Wickham H. 2017b. ggplot2: Elegant graphics for data analysis. *Journeal Stat Softw*.

Wilke CO. 2017. Cowplot: streamlined plot theme and plot annotations for ggplot2.

Witherspoon DJ, Xing J, Zhang Y, Watkins WS, Batzer MA, Jorde LB. 2010. Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. *BMC Genomics* **11**: 410.

Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer MA et al. 2009. Mobile elements create structural variation: analysis of a

complete human genome. *Genome Res* **19**: 1516-1526.

Yamamoto Y, Watanabe T, Hoki Y, Shirane K, Li Y, Ichiiyanagi K, Kuramochi-Miyagawa S, Toyoda A, Fujiyama A, Oginuma M et al. 2013. Targeted gene silencing in mouse germ cells by insertion of a homologous DNA into a piRNA generating locus. *Genome Research* **23**: 292-299.

Yannopoulos G, Stamatis N, Monastirioti M, Hatzopoulos P, Louis C. 1987. hobo is responsible for the induction of hybrid dysgenesis by strains of *Drosophila melanogaster* bearing the male recombination factor 23.5MRF. *Cell* **49**: 487-495.

Yin H, Lin H. 2007. An epigenetic activation role of Piwi and a Piwi-associated piRNA in *Drosophila melanogaster*. *Nature* **450**: 304–308.

Yoshitake Y, Inomata N, Sano M, Kato Y, Itoh M. 2018. The P element invaded rapidly and caused hybrid dysgenesis in natural populations of *Drosophila simulans* in Japan. *Ecol Evol* **8**: 9590-9599.

Zanni V, Eymery A, Coiffet M, Zytnicki M, Luyten I, Quesneville H, Vaury C, Jensen S. 2013. Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. *Proc Natl Acad Sci U S A* **110**: 19842–19847.

Zhang J, Zuo T, Peterson T. 2013. Generation of tandem direct duplications by reversed-ends transposition of maize Ac elements. *PLoS Genet* **9**: e1003691.

Zhang S, Kelleher E. 2019. piRNA-mediated silencing of an invading transposable element evolves rapidly through abundant beneficial *de novo* mutations. *bioRxiv* doi:10.1101/611350: 611350.

Zhang S, Kelleher ES. 2017. Targeted identification of TE insertions in a *Drosophila* genome through hemi-specific PCR. *Mobile DNA* **8**: 10.

Zhuang J, Wang J, Theurkauf W, Weng Z. 2014. TEMP: A computational method for analyzing transposable element polymorphism in populations. *Nucleic Acids Res* **42**: 6826–6838.

International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**: 695-716.