

**DOMESTICATION OF *PIF* TRANSPOSABLE ELEMENTS
TRANSPOSASES AS REGULATORY PROTEINS IN
*DROSOPHILA MELANOGASTER***

by

DIWASH JANGAM,

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy at
The University of Texas at Arlington
August, 2018

Arlington, TX

Supervising Committee:

Esther Betrán, Supervising Professor

Cédric Feschotte

Michael Buszczak

Shawn Christensen

Jeff Demuth

Copyright © by Diwash Jangam 2018

All Rights Reserved

ACKNOWLEDGEMENTS

I want to thank my committee for their guidance, support, and encouragement. Dr. Esther Betrán for being such an amazing adviser and providing guidance and supporting me in every step. Dr. Cédric Feschotte for being very supportive and encouraging throughout my Ph.D. and also providing financial support for various experiments that I have conducted. Dr. Michael Buszczak for his advice and help with the technical aspects of fly genetics, allowing me to learn techniques in his lab over at UT Southwestern and for sharing fly stocks and plasmids available in his lab. Dr. Jeff Demuth and Dr. Shawn Christensen for their guidance and support.

I want to thank all the members of the Betrán lab. Susana Domingues for being a great mentor, an awesome friend and someone I could also look up to. Javier for keeping the lab running smoothly so that I could conduct my experiments without disruptions. Mehdi, Ayda and Fatema for their time in brainstorming ideas, valuable input in my experimental designs, and also for being great friends. Suzanne and Sophia for being very persistent and helping me out with some tedious experiments in my projects.

I want to specially thank Varsha Bhargava from Buszczak lab for teaching to generate mutants, taking incredible images, advice in experimental design, and also for being a great friend. I also want to thank Victor, Arnaldo, Courtney, and Mayu from the Buszczak lab for their valuable input in my project.

Lastly, I want to thank Rachel Cosby and Aurelie Kapusta for their help and valuable input in the bioinformatics aspects of my research.

This research was funded by grants from Phi Sigma Graduate honor society, UTA RIGS funds to Dr. Esther Betrán, and funds from Dr. Cedric Feschotte Lab.

Date: August 22, 2018

ABSTRACT

DOMESTICATION OF *PIF* TRANSPOSABLE ELEMENTS

TRANSPOSASES AS REGULATORY PROTEINS IN

DROSOPHILA MELANOGASTER

Diwash Jangam, Ph.D.

The University of Texas at Arlington, 2018

Supervising Professor: Esther Betrán

Transposable elements (TEs) are genetic units that are able to move and amplify within a host genome. As a result of their activities, TE insertions can cause disruptions of gene functions and ectopic recombination producing deleterious effects in the host and thus they are also referred to as selfish elements. The machineries that TEs harbor that facilitate their transposition have been shown to be co-opted by the host for their own benefit through a process called molecular domestication. In the first chapter, we review examples that show that TE proteins are domesticated as an adaptation to evolutionary conflicts. We provide evidence for TE proteins domestication through conflicts between host-pathogen, mother-embryo, host-TE, and potentially as a result of centromere drive. We also argue that as long as all the hallmarks of a TE are present, they remain opportunistic and could not be considered domesticated. In the two

other chapters, we focus on identifying functions of domesticated transposase from *PIF/Harbinger* DNA TE in *D. melanogaster*. These *PIF* domesticated genes are named *Drosophila PIF Like Genes (DPLGs)*. There are four *DPLGs* in *D. melanogaster* and all of these genes are old, under purifying selection, and arose through independent domestication events. We show that *DPLGs* are domesticated as regulatory proteins and a subset of these genes are involved in neuronal and gonadal functions, and affect the viability and survival of *D. melanogaster*. We also provide evidence for functional overlap of these independently domesticated *PIF* transposase providing support to the model that domestication of transposase promotes domestication of related transposases.

TABLE OF CONTENTS:

ACKNOWLEDGEMENTS iii

ABSTRACT iv

CHAPTER ONE: TRANSPOSABLE ELEMENT DOMESTICATION AS AN ADAPTATION TO
EVOLUTIONARY CONFLICTS 1

CHAPTER TWO: DOMESTICATION OF *PIF* TRANSPOSABLE ELEMENTS TRANSPOSASES AS
REGULATORY PROTEINS IN *DROSOPHILA MELANOGASTER* 17

CHAPTER THREE: CONCLUDING AND FUTURE DIRECTIONS CHAPTER 83

CHAPTER ONE

Transposable Element Domestication As an Adaptation to Evolutionary Conflicts

Diwash Jangam¹, Cédric Feschotte^{2,3,*} and Esther Betrán^{1,*}

¹ Department of Biology, University of Texas at Arlington, Arlington, TX, USA.

³ Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT, USA

² Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA

Special Issue: Transposable Elements

Review

Transposable Element
Domestication As an
Adaptation to Evolutionary
ConflictsDiwash Jangam,¹ Cédric Feschotte,^{2,3,*} and Esther Betrán^{1,*}

Transposable elements (TEs) are selfish genetic units that typically encode proteins that enable their proliferation in the genome and spread across individual hosts. Here we review a growing number of studies that suggest that TE proteins have often been co-opted or ‘domesticated’ by their host as adaptations to a variety of evolutionary conflicts. In particular, TE-derived proteins have been recurrently repurposed as part of defense systems that protect prokaryotes and eukaryotes against the proliferation of infectious or invasive agents, including viruses and TEs themselves. We argue that the domestication of TE proteins may often be the only evolutionary path toward the mitigation of the cost incurred by their own selfish activities.

Domestication of Transposable Element Proteins

Transposable elements (TEs) are **selfish genetic elements** (see [Glossary](#)) that are able to move and amplify within the genome of virtually all walks of life, including prokaryotes, unicellular and multicellular eukaryotes, and even large DNA viruses [1–3]. So-called autonomous TEs encode the enzymatic machinery to promote their own mobilization and propagation ([Figure 1](#), Key Figure) as well as those of related nonautonomous elements, and occasionally unrelated host sequences. The disruptive effects of TEs have been documented extensively, for instance, as they integrate into regulatory or coding regions of host genes or when they induce ectopic/nonallelic recombination events [4–6]. As a result, new TE insertions are often deleterious and removed from the population by **purifying selection** or they are effectively neutral and fixed through **genetic drift** [7]. Consistent with the idea that the bulk of TE sequences do not serve host function, the rate and pattern of sequence evolution of TEs that are fixed in a genome generally follows that of unconstrained, neutrally evolving DNA, leading to the accumulation of disabled and nonreplicative TE ‘skeletons’ throughout genomes [8–10].

These theoretical and empirical observations are in line with the notion that TEs owe their persistence and extraordinary diversification to their self-replicative and genetically invasive activities [11]. This selfish ‘raison d’être’ does not preclude that, on occasion, parts or whole TE sequences may be co-opted to serve cellular function beneficial to the host organism. This process of co-option or ‘molecular domestication’ [12] of TE sequences has become increasingly recognized in recent years as advances in genomics have facilitated the identification, in various organisms, of a growing number of instances of TE-derived sequences that have been repurposed to serve cellular functions. The most robust evidence for such domestication events has come from either (i) comparative genomics whereby particular TE sequences can be

Trends

Transposable elements are selfish DNA elements that are able to increase in copy number by exploiting host cellular functions.

Domestication of TE sequences by the host for cellular function is an evolutionary process that has been unexpectedly common.

Proteins encoded by TEs are often repurposed to perform host functions as part of novel protein-coding genes.

Domesticated TE proteins are frequently co-opted to mitigate evolutionary conflicts, especially in defense against pathogens and invasive genetic elements.

For certain TE conflicts, domestication might be an inevitable outcome.

¹Department of Biology, University of Texas at Arlington, Arlington, TX, USA

²Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT, USA

³Present address: Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA

*Correspondence: cf458@cornell.edu (C. Feschotte) and betran@uta.edu (E. Betrán).

inferred to have been immobilized and evolved under functional constraint acting at the level of the host organism for extended periods, and/or (ii) genetic evidence whereby mutation or experimental removal of TE sequences has advert effects on cell function and/or host fitness [1,13–16]. Many of these well-documented examples point to the frequent co-option of TE-derived sequences as noncoding elements modulating host gene expression at the DNA or RNA level [13,16–20]. By contrast, we herein focus on cases where the ‘coding’ regions of TEs have been recruited as proteins serving host cell function. Out of the well-studied examples of TE protein domestication (listed in the Supplemental Table online), a substantial fraction appears to have been driven by the necessity to cope with various evolutionary conflicts and these will be the focus of this review (Table 1). We will highlight specific cases that illustrate both recent and older domestication events and reveal that TE proteins are often domesticated in response to intraspecific and interspecific evolutionary conflicts, as part of an **arms race** characterized by ever-changing selective pressures [21]. Interestingly, some of these conflicts have played out repeatedly in multiple lineages and adaptation has occurred through independent TE domestication events, leading to convergent evolutionary innovations. Finally, we will argue that in certain conflicting situations TE domestication might be the sole, inevitable evolutionary resolution.

Conflicts between Hosts and Pathogens

A recurrent theme implicating TE protein domestication is the emergence of **adaptive immune systems**. V(D)J recombination is a conserved process of jawed vertebrates that creates a virtually infinite repertoire of antibodies in their B and T cells, which in turn allows the recognition and neutralization of a vast diversity of antigens expressed by pathogens [22,23]. The two crucial components of V(D)J recombination are (i) the recombination activating 1 (RAG1) and recombination activating 2 (RAG2) proteins, which catalyze the DNA rearrangement reaction, and (ii) their *cis*-acting DNA sequences, the recombination signal sequences (RSSs), which reside at the boundaries of the V (variable), D (diversity), and J (joining) segments that define the specific genomic sequences bound, cleaved, and joined to produce an essentially unlimited diversity of coding sequences [22,24]. It is now firmly established that the catalytic core of RAG1, the protein that is responsible for cleavage activity, is a domesticated transposase derived from an ancient lineage of DNA transposons dubbed *Transib* [23,25]. TE-encoded homologs of *RAG1* occur in various invertebrates such as sea urchin and oysters [25,26]. It is also likely that the RSS motif descends from the terminal inverted repeats (TIRs) of *Transib* elements, since these sequences and their arrangement are similar in *Transib* transposons [25]. In addition to *RAG1*, *RAG2* also was shown to have TE origins and several lines of evidence suggested that both *RAG1* and *RAG2* were domesticated from the same ancestral *Transib* element [26]. This evolutionary scenario has been solidified by a recent study that functionally characterized an active *Transib* element from the lancelet, a member of the cephalochordates, which lacks V(D)J recombination [27]. This transposon, coined *ProtoRAG*, encodes both *RAG1*- and *RAG2*-like genes arranged just like their domesticated vertebrate homologs and flanked by TIRs that resemble the RSS and is able to undergo TIR-dependent transposition through a mechanism strikingly similar to RAG1/2-mediated DNA rearrangement [27]. These results support the idea that not only *RAG1* derives from a transposase, but that *RAG2* and RSS also descend from an ancestral transposon related to *ProtoRAG*.

There is growing evidence that TE domestication was also instrumental to the emergence of another adaptive immune system, but this time of prokaryotes: the clustered regularly interspaced short palindromic repeats (CRISPR)-CRISPR-associated system (Cas) system. To defend against the continuous assault of invasive (and often deadly) genetic elements such as plasmids and phages, many bacteria and archaea have evolved a form of adaptive immunity that consists of two minimal components: (i) a core Cas protein complex, which has nucleic acid binding and endonuclease activities, and (ii) a guide RNA generated from CRISPR loci [28]. CRISPR loci are composed of noncontiguous direct repeats separated by variable spacer

Glossary

Adaptive immune systems:

antigen-specific response system that involves antigen-specific recognition and neutralization.

Arms race: in this publication, an evolutionary conflict that involves continuous competition between interacting species or genetic elements to adapt to the ever-changing interacting partner.

Genetic drift: evolutionary process leading to the chance change in allele frequencies due to the random effects caused by sampling in populations because of their finite population size.

Histone deacetylases: enzymes whose activity involves the removal of acetyl groups of histones. This histone modification most often leads to chromatin condensation.

Positive selection: evolutionary process leading to the increase in frequency of new beneficial alleles. When it occurs in the context of an arms race, it is recurrent and leaves behind a signature of fast protein evolution.

Purifying selection: evolutionary process leading to the decrease in frequency of new deleterious alleles.

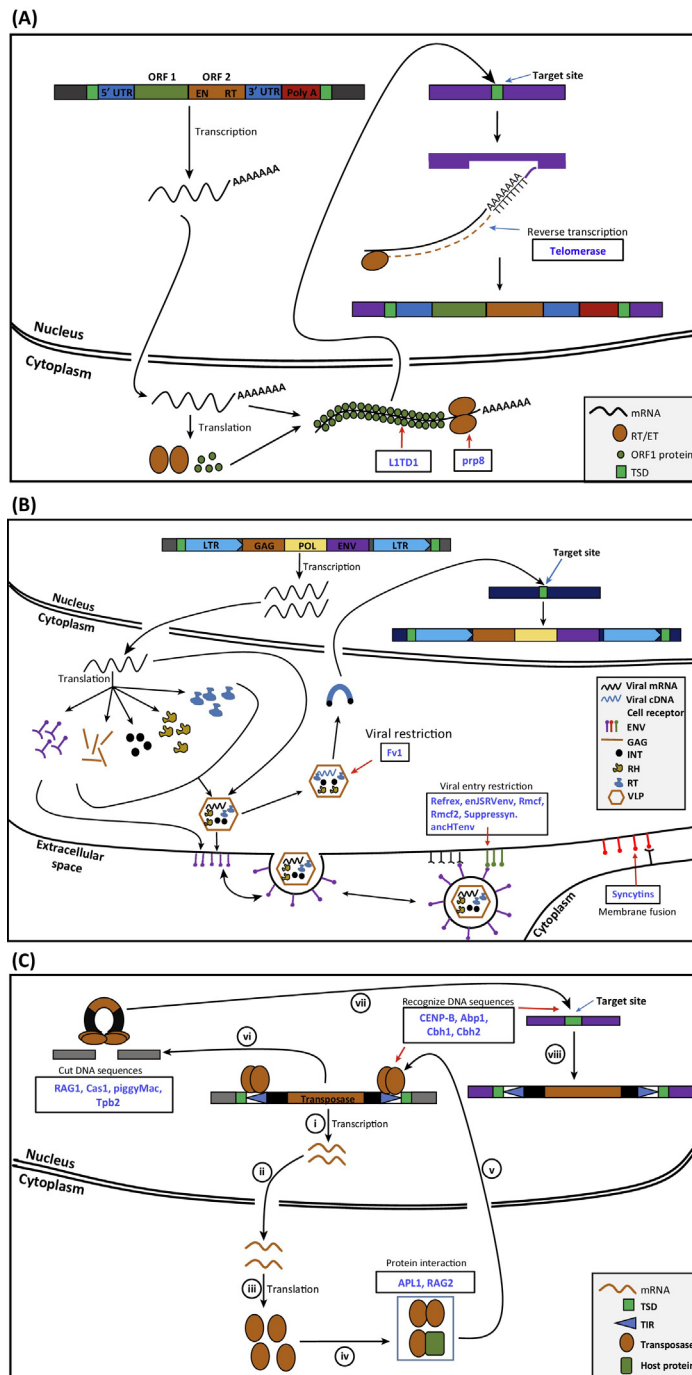
Reverse transcription: protein activity that leads to the synthesis of DNA using RNA as a template.

Selfish genetic element: DNA sequences that have evolved the ability to propagate at a cost to the host.

Small-interfering RNAs: in this publication, short cellular RNAs complementary to mRNAs that prompt targeted mRNA degradation and, often, targeted gene silencing through chromatin remodeling.

Key Figure

Transposition Mechanisms of Major Types of Transposable Elements Highlighting Genes and Functions that are Involved in Conflict.



Trends in Genetics

(See figure legend on the bottom of the next page.)

sequences that are formed by incorporating DNA bits of invasive genetic elements, thereby providing a heritable record of the cell's previous exposure to various parasitic elements [29,30]. If the invader in record enters the cell again, the CRISPR RNA (crRNA) will be used as a guide for the Cas protein complex to recognize and digest the invader's nucleic acids [31]. Recent studies have uncovered a wide diversity of CRISPR-Cas systems that can be divided into two major classes. Class 1 systems are the most widespread in bacteria and archaea and are defined by multisubunit crRNA-effector complexes, whereas Class 2 systems represent only about 10% of the CRISPR-Cas loci and are defined by a single subunit crRNA-effector module [32]. Cas1, the most prominent Cas protein found in CRISPR-Cas systems, shares sequence similarity to the transposases encoded by a group of TEs called *Casposons* [33]. In addition, Cas1 is able to integrate spacer sequences into the CRISPR locus through a biochemical mechanism that is strikingly similar to transposase [34,35]. Akin to RSS sequences in V(D)J, it has been proposed that the CRISPR repeats may be derived from the TIRs of *Casposon*-like elements based on the observation that the TIRs of these transposons are very similar to CRISPR repeats in terms of sequence, size, secondary structure, and their predicted ability to be bound and cleaved by Cas1-like transposases [36,37]. Thus, evidence is mounting that Class 1 CRISPR-Cas systems originated from the domestication of *Casposons* [33]. Other mobile genetic elements have also contributed to the further diversification of Class 1 CRISPR-Cas systems. Notably, reverse transcriptase sequences derived from mobile Group II introns have been co-opted repeatedly in bacterial evolution to form Cas1-reverse transcriptase fusion proteins, which enable the acquisition of spacer sequences from parasitic RNA agents [38].

The Class 2 CRISPR-Cas system, which is apparently less common, but still remarkably diverse in bacteria is thought to have an independent origin from Class 1 CRISPR-Cas system. There is growing support to the notion that Class 2 systems were also assembled from parts borrowed from transposons. Cas9-like proteins, which act as effectors in most Class 2

Figure 1. (A and B) Class I elements or retrotransposons transpose via RNA intermediates, and (C) Class II elements or DNA transposons mobilize directly as DNA molecules. We include retroviruses and endogenous retroviruses under long terminal repeat (LTR) retrotransposons as previously proposed [109]. (A) A typical non-LTR retroelement transposition (i.e., LINE element) is initiated by the transcription of the element. The transcript is translated into proteins and they associate with the mRNA and translocate into the nucleus. The reverse transcriptase (RT) protein has endonuclease activity and makes a nick on one of the strands and uses the 3' end to prime synthesis a cDNA copy and insert into the genome. This process is known as target-primed reverse transcription. The nicks generated on two DNA strands are generally staggered and this results in target site duplications (TSDs). (B) A typical LTR retrotransposon is characterized by LTRs and generally encodes for three major proteins (GAG, POL, and ENV). The transposition is initiated by the transcription of three encoded genes as a single mRNA. The transcript is translated into several protein products. The *POL* gene is translated to three proteins: integrase (INT), RNase H (RH), and RT. The GAG forms the capsid protein that encapsulated the LTR mRNA, int, RH, RT into a nucleocapsid virus-like particle (VLP). The ENV, which is a glycoprotein, can promote the escape of the VLP from the cell. In the extracellular space, the viral surface ENV glycoprotein can recognize susceptible cells through recognition of the cell receptors and fuse with the cell membrane. Once fused, the nucleocapsid can enter into the cell cytoplasm. Alternatively, the VLP, instead of escaping the cell, can continue the transposition process within a single host cell. In the VLP, the RT reverse transcribes the RNA into cDNA, which then associated with the INT. The INT guides the cDNA into the nucleus, where it finds a target site and integrates the element into the genome. Since the INT usually generates a staggered cut, the LTR elements are flanked by TSDs. (C) A typical cut-and-paste DNA TE is flanked by TSDs and terminal inverted repeats (TIRs), and encodes for at least a transposase. The transposition of the cut-and-paste element is initiated when the transposase is transcribed and translated. The transposase can either stay as monomer or form multimers. Alternatively, the transposase can also interact with other proteins (either encoded by the TE itself or host proteins). The transposase is then able to translocate into the nucleus where it recognizes and binds the TIRs. After binding to the TIRs, the transposase catalyzes the excision of the TE from the donor site. When the TE (bound to the transposase) finds a target site, it makes a staggered cut and inserts itself into the new site. When the staggered cuts are repaired, the TE remains flanked by TSDs. Abbreviations: ALP1, ANTAGONIST OF LIKE HETEROCHROMATIN PROTEIN 1; ENV, envelope protein; GAG, group antigens; LINE, long interspersed element; ORF, open reading frame; POL, reverse transcriptase; RAG2, recombination activating 2.

Table 1. Compilation of TE Protein Domestication Events Highlighted in the Text That Have Been Driven by Selection to Adapt to Evolutionary Conflicts

Gene names	Ancestral TE	TE protein (domain) ^a	Originally discovered	Taxonomic distribution	Possible conflict	Function related to the conflict	Refs
<i>abp1, cbh1, cbh2</i>	<i>Tc1/mariner/pogo</i>	Transposase (whole)	<i>Schizosaccharomyces pombe</i>	Schizosaccharomycetale	TE vs host/ Centromere drive	LTR retrotransposons repression; chromatin silencing at centromere	[44,45]
<i>ALP1</i>	<i>PIF/Harbinger</i>	Transposase (whole)	<i>Arabidopsis thaliana</i>	Land plants	TE vs host	Potential role in TE silencing through interaction with Polycomb Repressive Complex 2	[49]
<i>hsaHTenv</i>	<i>Gammaretroviruses</i>	Envelope	<i>Homo sapiens</i>	Hominids	Virus vs host	Restrict viral entry into the host cell	[42]
<i>cag</i>	<i>Tc1/mariner/pogo</i>	Transposase (DBD)	<i>Drosophila melanogaster</i>	Unknown	Centromere drive	Predicted to bind centromeric DNA sequence	[90]
<i>Cas1</i>	<i>Casposon</i>	Transposase (whole)	Bacteria	Archaea and bacteria	Virus/TE vs host	Defense response to virus and maintenance of CRISPR repeat elements	[33]
<i>Cas9</i>	<i>DNA transposon</i>	tnpB	Bacteria	Archaea and bacteria	Virus/TE vs host	Integration of invading viral DNA into CRISPR locus	[40]
<i>CENPB</i>	<i>Tc1/mariner/pogo</i>	Transposase (DBD)	<i>H. sapiens</i>	Mammals	Centromere drive	Facilitates mitotic centromere formation, recognizes and binds a 17-bp sequence in the centromeric alpha satellite DNA	[100,101]
<i>enJSRVenv</i>	<i>Betaretroviruses</i>	Envelope	Sheep	Ovine/unknown	Virus vs host	Restrict viral entry into the host cell	[41,102]
<i>Fv1</i>	<i>MuERV-L (gag) (Class III)</i>	Capsid Protein	<i>Mus musculus</i>	<i>Mus</i> subgenera	Virus vs host	Murine leukemia virus restriction	[103]
<i>L1TD1</i>	<i>L1</i>	ORF1	<i>H. sapiens</i>	Mammals	TE vs host	Potential role in TE control	[50]
<i>MAIL1</i>	<i>Ty3/gypsy LTR retrotransposon</i>	Plant mobile domain	<i>A. thaliana</i>	Unknown	TE vs host	Silencing of TEs and genes. Condensation of pericentromeric heterochromatin	[48]
<i>MAIN</i>	<i>Ty3/gypsy LTR retrotransposon</i>	Plant mobile domain	<i>A. thaliana</i>	Unknown	TE vs host	Silencing of TEs and genes. Condensation of pericentromeric heterochromatin	[48]
<i>PiggyMac</i>	<i>piggyBac</i>	Transposase (DBD? + core)	<i>Paramecium tetraurelia</i>	Unknown	TE vs host	Required for genome rearrangement in <i>P. tetraurelia</i>	[55]
<i>RAG1</i>	<i>Transib</i>	Transposase (core)	<i>H. sapiens</i>	Jawed vertebrates	Pathogen vs host	Important function in V(D)J combination and interacts with RAG2	[26]
<i>RAG2</i>	<i>Transib</i>	Transposase	<i>H. sapiens</i>	Jawed vertebrates	Pathogen vs host	Important function in V(D)J combination and interacts with RAG1	[26]
<i>Refrex-1</i>	<i>Gammaretroviruses</i>	Envelope	Domestic cats	Feline/unknown	Virus vs host	Restrict viral entry into the host cell	[41,104]
<i>Rmcf</i>	<i>Gammaretroviruses</i>	Envelope	<i>Mus castaneus</i>	<i>Mus</i> subgenera	Virus vs host	Defends against viral infection	[41,105]
<i>Rmcf2</i>	<i>Gammaretroviruses</i>	Envelope	<i>M. castaneus</i>	<i>Mus</i> subgenera	Virus vs host	Defends against viral infection	[41,106]

Table 1. (continued)

Gene names	Ancestral TE	TE protein (domain) ^a	Originally discovered	Taxonomic distribution	Possible conflict	Function related to the conflict	Refs
<i>Suppressyn</i>	<i>HERV-F</i>	Envelope	<i>H. sapiens</i>	Simians	Virus vs host	Regulates syncytins and potential viral restriction into host cells	[41,107]
<i>Syncytin A, Syncytin B</i>	<i>HERV-F or HERV-H</i>	Envelope	<i>M. musculus</i>	Murid rodents	Fetus vs mother	Placenta formation; fusogenic activities <i>ex vivo</i>	[69]
<i>Syncytin 1, Syncytin 2</i>	<i>HERV-W</i>	Envelope	<i>H. sapiens</i>	Humans, apes, Old World monkeys	Fetus vs mother	Cell fusion; placenta formation	[69]
<i>Syncytin-Car1</i>	<i>CarERV3 (Class I)</i>	Envelope	Carnivores	Carnivores	Fetus vs mother	Placenta-specific expression, fusogenic activities	[69]
<i>Syncytin-Ory1</i>	<i>Type D retroviruses</i>	Envelope	Rabbits and hares	Leporids: rabbits and hares	Fetus vs mother	Placenta-specific expression, fusogenic activities	[69]
<i>Syncytin-rum1</i>	<i>vertebrate retrovirus</i>	Envelope	<i>Bos taurus</i>	<i>B. taurus</i>	Fetus vs mother	Placenta-specific expression, fusogenic activities	[69]
<i>TERT</i>	<i>LINE-like retroelement?</i>	Reverse transcriptase	<i>H. sapiens</i>	Eukaryotes	TE vs host	RNA-directed DNA polymerase activity; telomerase RNA reverse transcriptase activity	[108]
<i>TPB1, TPB2, TPB6</i>	<i>piggyBac</i>	Transposase (DBD? + core)	<i>Tetrahymena thermophila</i>	Unknown	TE vs host	Required for genome rearrangement in <i>T. thermophila</i>	[59,60]
<i>Prp8</i>	<i>Retroelement</i>	Reverse transcriptase?	<i>H. sapiens</i>	Eukaryotes	Mobile introns vs host	Generation of catalytic spliceosome for second transesterification step; mRNA 3'-splice site recognition	[66]

^aAbbreviations: Core, catalytic core; DBD, DNA binding domain.

systems, show sequence similarity to proteins called TnpB, which are poorly characterized but are commonly found in both autonomous and nonautonomous TEs [39,40]. Phylogenetic analyses point to multiple domestication events of TnpB proteins giving rise to different lineages of Cas9-like effectors for several Class 2 CRISPR-Cas subtypes [39]. Taken together, these findings paint a remarkable picture whereby multiple CRISPR-Cas systems have been assembled independently from the co-option of various types of TE-derived proteins during prokaryotic evolution.

Host-TE Conflict Resolved through TE Domestication

In parallel to the arms race between host and pathogens plays another battle between cells and invasive genetic elements like TEs and retroviral relatives. Cells have evolved ways to overcome these conflicts through pathways and mechanisms that often rely on domesticated TE proteins. Several proteins derived from the *Envelope (Env)* gene of endogenous retroviruses are known to aid in the protection of the host cell by restricting infection of related retroviruses. Because *Env* is essential for entry of the virus into the host cell, endogenous *Env* expression can block viral entry through a competitive process called receptor interference [41]. There are at least six different *Env*-derived genes identified in species as diverse as mouse, cat, sheep, and primates that are capable of protecting against the infection of related retroviruses [41,42].

In addition to *Env* proteins, the *Gag* proteins encoded by endogenous retroviruses can also be co-opted for restricting retroviral infection. A classic example is the mouse *Fv1* gene, which is derived from the *Gag* gene of a member of the endogenous retrovirus (ERV) family ERV-L. *Fv1* was initially identified as a restriction factor for the murine leukemia virus (which is not directly related to ERV-L), but was subsequently shown to be capable of protecting against a wide variety of retroviruses [43]. Thus, *Fv1* may have acquired a broad antiviral function, though the molecular mechanisms by which restriction is achieved remain poorly understood. Interestingly, *Fv1* is a rapidly evolving gene with signature of **positive selection** diversifying the C-terminal region of the protein, which is known to be important for viral restriction [43], suggesting that this factor has been engaged in an arms race with retroviruses.

In the fission yeast *Schizosaccharomyces pombe*, three transposase-derived proteins (*Abp1*, *Cbh1* and *Cbh2*) that originated from *pogo* transposons have taken on partially overlapping function in controlling unrelated retrotransposons called Tf2 elements. These domesticated proteins have been reported to transcriptionally silence Tf2 retrotransposons by tethering **histone deacetylases** to the long terminal repeats of these elements, which also prevents the chromosomal integration of Tf2 elements via homologous recombination [44–46]. Thus, proteins derived from one TE class have acquired the ability to silence TEs from a completely different class. TE silencing may not be the sole cellular function of the *S. pombe* transposase-derived proteins as they are also required for proper chromosome segregation [47]. Therefore, it is unclear whether the TE silencing function of *Abp1*, *Cbh1*, and *Cbh2* evolved first or emerged secondarily through fortuitous recognition of Tf2 elements. Interestingly, *pogo*-like transposases have been domesticated in several additional lineages and also in part to serve centromeric function (Box 1). We speculate that these repeated episodes of transposase domestication might have been promoted to suppress another conflict: the so-called centromere drive (Box 1).

In *Arabidopsis*, two genes *MAIL1* and *MAIN* encode related proteins that appear to be evolutionarily derived from a subset of Ty3/gypsy retrotransposons found in angiosperms [48]. These proteins might have been initially captured from the host by these elements, but appear to have been reclaimed to partake in an epigenetic silencing pathway that transcriptionally represses a broad array of TEs [48]. Genetic loss of these genes resulted in impaired condensation of pericentromeric heterochromatin and upregulation of TE transcription,

Box 1. Is Centromere Drive Promoting TE Domestication?

Centromere binding protein B (CENP-B) is a conserved mammalian factor that is essential for the establishment of centromere identity and is derived from the transposase of a pogo-like DNA transposon [87]. Three CENP-B-like proteins (Abp1, Cbh1, and Cbh2) have also been identified in fission yeast by virtue of their sequence and functional similarities to mammalian CENP-B. However, the yeast genes are not orthologous to the mammalian CENP-B and were independently domesticated from a distinct lineage of pogo-like transposons, suggesting a form of convergent evolution [88]. In addition, there are two more reports of independent domestication events of pogo-like transposases: one in lepidopteran species with holocentric chromosomes [89] and one in *Drosophila* (called CAG), which may also be a centromere-associated protein [89,90].

The recurrent domestication of pogo-like transposases in evolution is intriguing and might be driven by the ability of these proteins to turn into TE silencers as described in the text for the CENP-B-like proteins of fission yeast [44]. However, the association of several CENP-B proteins with centromeric regions suggests that another genetic conflict might be another evolutionary force repeatedly underlying their co-option: the so-called centromere drive model [91]. Unlike mitosis, which produces two identical daughter cells or male meiosis, where all four gametes are produced, in female meiosis only one of the four meiotic products is passed to the next generation through the oocyte. This creates an opportunity for competition between homologous chromosomes to end up in the oocyte. The centromere is positioned in a way that it can effectively orient the chromosome during meiosis I through microtubule attachment and the proper orientation of a chromosome has been shown to be advantageous for its transmission to the next generation [11]. This phenomenon has led to a model dubbed centromere drive, where the centromere that positions the chromosome in the best orientation is selected, and is thought to be responsible for rapid evolution of centromeric DNA as its length and sequence can bias its transmission [92,93]. Centromere drive is believed not to cause direct negative fertility effects to females; however, if it reduces the fertility in males, the centromeric proteins would require to adapt and restore male fertility [94]. The conflict generated could have not only led to the rapid evolution of kinetochore proteins, but in addition, might have also led to the domestication of pogo-like transposase as centromere-binding proteins to adapt to the centromere drive.

suggesting that their protein products are acting as transcriptional repressors. The molecular mechanisms by which *MAIL 1* and *MAIN* promote the formation of silent chromatin remain to be characterized, but involve a molecular pathway independent of **small-interfering RNAs** and DNA methylation. It is intriguing that proteins related to *MAIL 1* and *MAIN* seem to have been acquired by TEs multiple times, including DNA transposons of the *Mutator*-like superfamily, which may reflect a counter-defense strategy deployed by these elements [48].

ANTAGONIST OF LIKE HETEROCHROMATIN PROTEIN 1 (ALP1) is another domesticated TE protein identified in *Arabidopsis* that is involved in yet another epigenetic silencing pathway [49]. ALP1 directly derives from a *PIF*-like transposase and antagonizes silencing through a direct interaction with the Polycomb Repressive Complex 2 (PRC2), which is known to contribute to TE silencing in *Arabidopsis* [49]. The authors hypothesized that ALP1's interaction with PRC2 could be an ancient property of *PIF*-like transposases that benefited the original transposon as a form of counter-repression [49]. Interestingly, in this case, the domesticated TE protein does not appear to play an effector role in TE repression but exerts a modulatory effect on a TE silencing pathway. The outcome must be beneficial to the host organism since the *ALP1* gene displays clear signature of evolutionary conservation and purifying selection across diverse land plant species.

Another example of domesticated TE protein that was potentially co-opted for TE control is LINE-1 type transposase domain-containing 1 (L1TD1 [50]). *L1TD1* was co-opted in the ancestor of placental mammals from the open reading frame 1 (ORF1) coding region of long interspersed nuclear elements LINE-1 (or L1s), one of the most abundant and persistent TE families in mammalian genomes. While the biochemical and cellular activities of L1TD1 remain to be defined, several observations suggest that it may be engaged in an arms race relationship with L1 elements. First, the evolution of the *L1TD1* gene is characterized by bouts of rapid diversification under positive selection in primates and mice, lineages where L1 elements have undergone particularly dramatic bursts of diversification and expansion. Second, the *L1TD1* gene has been lost multiple times during mammalian evolution, and at least in one lineage

(megabats), its loss correlates with the (otherwise rare) extinction of L1 elements, as if L1TD1 was no longer needed once L1 became extinct. Although human L1TD1 now appears to function as a regulatory protein to maintain embryonic stem cell pluripotency, the authors argue that L1TD1 was initially domesticated to defend against L1 or other TEs [50].

TE Proteins Co-opted to Eliminate TE-Derived Sequences

Ciliates are single-celled eukaryotes that are unique for harboring dimorphic nuclei [51]. The germ-line micronucleus (MIC) contains the genomic material that is passed down to the next generation while the somatic macronucleus (MAC) is not passed to the next generation but encodes all the proteins responsible for the organism's function [51]. Like other typical eukaryotic genomes, the MIC contains a large amount of repetitive sequences and TEs interspersed with DNA sequences essential for the host [52,53]. However, unlike any other genomes, the genes in the MIC are interrupted and sometimes even scrambled by a multitude of nongenic DNA segments, including repetitive elements called internal eliminated sequences (IESs) that must be excised at the DNA level for correct assembly of the MAC and proper gene expression [54]. To excise IESs, some ciliates have co-opted the cleavage activities of transposases. For instance, *Paramecium* uses *PiggyMac* (*Pgm* [55,56]), a domesticated transposase, while in *Tetrahymena* at least four other transposase-derived proteins [*Tetrahymena PiggyBac*-like (TPB) 2 (TPB2), TPB, TBP6, and LIA5] are required for IES removal [57–60]. While all of these proteins share sequence similarity to the *piggyBac* superfamily of transposases, their evolutionary relationship to one another remains obscure. For example, it is unclear whether the *Paramecium* and *Tetrahymena* genes encoding these proteins are orthologous or if each is derived from distinct transposons independently domesticated and/or the product of gene duplication events [54,58,61].

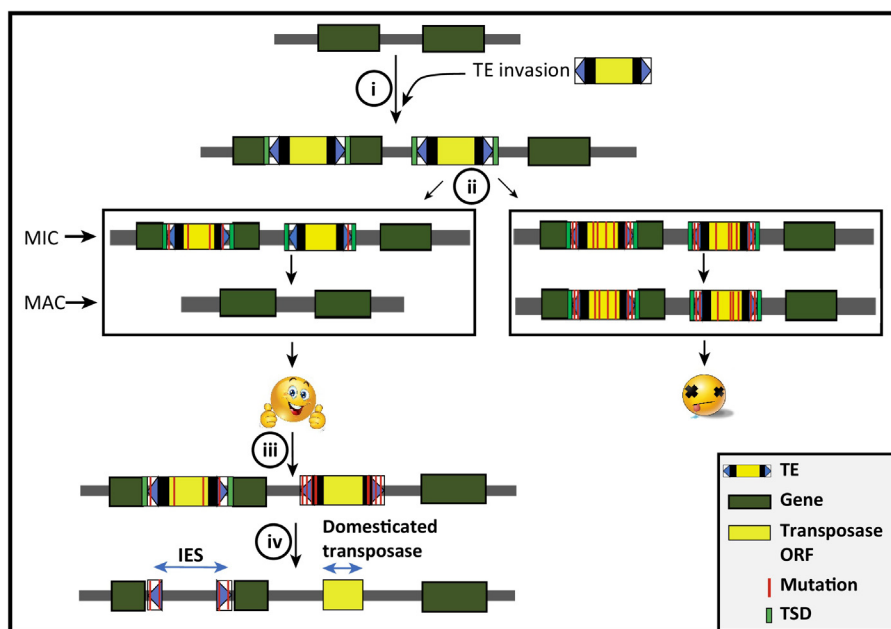
The functions of these transposase-derived proteins also appear to vary or to have diverged after they were domesticated. In *Paramecium*, DNA elimination involves the precise excision of IESs, which are located between and within genes, and requires the catalytically active PGM transposase as well as TA dinucleotides at the boundaries of IES. When excised in the MAC, only a single TA remains [61]. In *Tetrahymena*, IESs are primarily intergenic and while TPB2 has retained catalytic activities that are essential for IES removal [60], LIA5 appears to have lost its catalytic activity but remains essential for DNA elimination likely through its involvement in chromatin reorganization prior to IES excision [57,58]. Finally, TPB1 and TPB6 appear to function as catalytically active transposases, but differ from TPB2 in being dedicated to the removal of a small subset of IESs that resemble ancient *piggyBac* transposons [59].

There is growing evidence that IES themselves originated from TEs. In *Paramecium*, a fraction of IES appears to be derived from recent *Tc1/mariner*-like element invasions and resemble miniature elements or solitary TIRs, whose sequences appear to have converged to be excised by *Paramecium*'s *PiggyMac*, leaving no scar behind unlike typical *Tc1/mariner* transposition [61]. In *Tetrahymena*, there is also substantial overlap and terminal sequence similarities between IES and TEs, including *piggyBac*-like elements [53,59,62]. These observations and what is known about the biochemistry of IES excision [61] support the idea that the process of IES elimination in *Paramecium* and *Tetrahymena* closely resembles the excision of *piggyBac* transposons, and that this type of elements could have provided both the enzymatic machinery and at least some of the *cis*-acting sequences now required for the process.

Expanding upon this idea, we propose a model in which the dimorphic nuclei system of ciliates fosters the evolution of TEs that are specifically expressed during the transition from MIC to MAC as to be excised during that transition, which would minimize their deleterious impact and lead to their eventual domestication (Box 2). Another ciliate species, *Oxytricha*, provides an outstanding model to test the hypothesis. In this species, certain DNA transposons called TBE

Box 2. When TE Domestication May Be Inevitable

Here we argue that evolutionary conflicts drive TEs to evolve features that initially minimize their deleterious impact on the host, but eventually and inevitably lead to their domestication. We envision that the ciliate MIC/MAC binuclei transition, which involves a step of programmed DNA elimination when the MAC nucleus forms (see main text), might provide such an opportunity. Once this process is in place, natural selection would favor ciliate TEs that are expressed during the transition from MIC to MAC as this behavior would promote their propagation exclusively in the MIC and minimize deleterious effects on somatic development and function in the MAC (Figure I). This system further predicts that TEs capable of excision, such as cut-and-paste DNA transposons, would be favored over those that cannot like retrotransposons. Consistent with this, the TE landscape of *Tetrahymena* and *Oxytricha* MIC genomes appears to be dominated by DNA transposons [52,53] (but maybe less so for *Paramecium* [83]). Another prediction of this model is that if a transposon lands in a region where its excision becomes essential for proper MAC (somatic) function and fixes in the population, it would impose functional constraint on at least one active transposase gene as well as the *cis*-acting sequences (TIRs) required for excision of the transposon (Figure I). This situation may progress from a type of mutualism [54,64], where active TEs capable of excision are maintained by natural selection, to a complete domestication if the transposase becomes physically dissociated from its *cis*-acting sequences and can only function in *trans* without increasing the spread of the transposase. Interestingly, the transposases encoded by the TBE family of DNA transposons, which are currently active in *Oxytricha*, are known to bear the signature of purifying selection [64,84]. This signature of constraint suggests that TBE transposase activity has to be maintained to preserve host fitness and might reflect an intermediate step toward eventual domestication. A final prediction of the model is that transposons that insert within or close to genes but excise precisely, leaving no molecular ‘footprints’ of their insertion, would be more likely to succeed in colonizing the MIC and therefore be more prone to evolve toward domestication. In this regard, it is notable that all transposases identified to date as domesticated for DNA elimination in ciliates derive from *piggyBac* transposons, which are known to insert preferentially within or near genes [85] and to produce precise excision events [86].



Trends in Genetics

Figure I. Model for the Inevitability of Transposable Element (TE) Domestication Using Ciliates As Example. (i) A novel TE invades a naïve ciliate genome and expands in copy number. Upon integration, TEs can potentially disrupt genes and regulatory sequences. (ii) Because ciliates have dimorphic nuclei, micronucleus (MIC) and macronucleus (MAC), the organisms in which the TEs evolve to precisely excise during the transformation from the MIC to the MAC will have intact coding regions and regulatory sequences and will survive. These TEs will proliferate undetected by the host. Host with mutated copies of the TEs which cannot excise due to mutations in the terminal inverted repeats (TIRs) or transposases which cannot form intact open reading frame in the MAC do not survive. (iii) Thus, there is purifying selection and organisms that harbor TEs able to precisely excise have higher fitness provided there are enough active TE copies that provide a source of transposase for excision. These TEs will keep proliferating and the potential for deleterious mutations in the TIRs will increase. (iv) Conflict between TE and host is resolved by domesticating a TE protein to excise related TEs during the transition to the MAC nucleus. Overtime the TE-related sequences accumulate mutations beyond recognition. However, the regions of TE (generally parts of TIRs) are under purifying selection since these sequences are important for the excision of disruptive sequences and give rise to internal eliminated sequences (IESs). Abbreviation: TSD, target site duplication.

Box 3. Can Whole TEs Be Domesticated?

A recurrent theme in the examples of TE domestication summarized in this review is that the domesticated TE proteins act in *trans* even when they act on substrates that resemble TE ends or derive from the same cognate TE [e.g., RAG1/2 in V(D)J recombination]. However, can a whole TE be domesticated as a unit? In the following section, we argue that domestication is complete only when proteins are separated from rest of the TE sequences as genetic conflict still exists when the whole TE is still replicating even if it is potentially in the trajectory to being domesticated.

In *Oxytricha*, TEs belonging to Tc1/mariner superfamily excise themselves during the transition from MIC to MAC nucleus and their activity is needed for transition from MIC to MAC [63]. However, the conflict may still be ongoing in this case as the fitness of individuals is predictably lowered by new insertions of the active element that is likely still actively transposing [54], thereby increasing the chances of deleterious TE-excision-disabling mutations as offspring is produced (see Figure 1 in Box 2).

In *Drosophila*, the activity of non-LTR retrotransposons elongates telomeres. In these species, three non-LTR retrotransposons (Het-A, TART, and TAHRE) retrotranspose to the very ends of the chromosomes using the 3' end for **reverse transcription**, preventing the telomeres from shortening [95,96]. These elements preferentially target the end of the chromosomes and are rarely found in other genomic regions [97]. The presence of TEs at the ends of the chromosomes might be viewed as a domestication of whole TEs by the host. However, while the retroelements may have found a 'safe heaven' at the ends of the chromosomes, perhaps, after the demise of the telomerase in *Drosophila*, the genome might actually still be in conflict with the TEs. It should be emphasized that the three non-LTR retroelements that transpose to the ends of the chromosomes in *Drosophila* have evolved features that differ from their non-LTR relatives such as targeting of the 5' end of other telomeric TEs at the end of chromosomes combined to unusually long untranslated regions that appear to specialize them toward their genomic niche of telomere targeting [96]. However, they appear to remain selfish elements as given the opportunity, DNA double-strand breaks are recognized by these elements and they transpose into other genomic regions as well [96] and the opportunities for selfishness remain [98]. Thus, these TEs cannot be considered fully domesticated and although the organism presumably benefits from their insertion at the telomeres, there is evidence of ongoing conflict that might only be resolved by a complete domestication whereby the locus producing the template RNA is physically separated from that encoding the reverse transcriptase. It is tempting to speculate that telomerase, itself a reverse transcriptase, might have originated through this evolutionary path from a domesticated retroelement [99].

undergo self-excision during the MIC to MAC transition and TBE transposase expression is essential for IES excision and proper genome unscrambling [63]. Thus, TBE transposons might be at an early step toward domestication [54,63] (Boxes 2 and 3).

Parallels have been drawn between the ciliate IES and spliceosomal introns in eukaryotes [64]. Analogous to the TE domestication model for IES removal (Box 2), it is tempting to speculate that TE proteins might have been ancestrally co-opted to ensure spliceosomal intron splicing. There is solid evidence that spliceosomal introns are evolutionarily related and likely derive from sequences resembling Group II introns (i.e., a type of Class I mobile element). First, Group II introns have structural similarity with spliceosomal introns (e.g., several components of the Group II intron ribozyme including small RNAs are similar to the small nuclear RNAs of the spliceosome), and they also have a splicing mechanism that is strikingly reminiscent of the removal of spliceosomal introns [65]. In addition, the *prp8* protein, which is an integral component of the spliceosomal complex, displays significant sequence similarity with reverse transcriptases and as such has been proposed to derive from an anciently domesticated retroelement [66]. Since Group II introns typically encode a reverse transcriptase that is essential for their insertion (Figure 1), Group II and spliceosomal introns are not only similar in structure and excision mechanism, but also in parts of the enzymatic machinery that catalyzes their mobilization [65,66]. These observations bring credence to the idea that spliceosome and introns could have originated via domestication of TE-encoded proteins and their *cis*-acting sequences, respectively. This hypothesis is in line with the proposal that the invasion of an ancient Group II intron-like mobile element of an early eukaryotic ancestor might have led to the emergence of spliceosomal introns [64,67]. Insertions of these introns within genes would have been tolerated by natural selection as long as the introns would have been spliced out after transcription, restoring the reading frame, imposing functional constraint on the machinery that ensures their splicing, and leading to domestication of the machinery for proper cell function.

In summary, TEs evolve ways to replicate and spread in the genome while minimizing their deleterious effects on host fitness, for instance, by ensuring their excision at the DNA or RNA level. These processes create a dependence of the host on these enzymatic activities, which leads to the assimilation of TE-encoded proteins to the cellular machinery. Thus, the evolutionary dynamics of host–TE interactions create fertile ground for the domestication of TEs, which in turn add a layer of complexity to the organization and function of the genome.

TE Proteins Co-opted because of Conflict between Mother and Embryo

Syncytins are proteins derived from the *Env* gene of retroviruses that have been co-opted at least nine times during mammal evolution and are thought to play a role in placentation [68–71]. The placenta is a temporary organ that is formed by the fusion of fetal extraembryonic membrane and the maternal uterine tissue, which facilitates metabolic exchanges through the interface between the mother and the fetus [69]. The proposed function of Syncytins in the placenta is based on their restricted or high level of expression in that organ, their ability to mediate cell-to-cell fusion (which is required for the establishment of the syncytiotrophoblast layer at the fetal–maternal interface) and, in some cases, also immunosuppressive activities [68]. Knockout studies of two murine-specific Syncytins in the mouse firmly established their critical function in placenta development [72,73], but it remains unknown whether all Syncytins identified in other mammalian lineages are equally important for placentation. In fact, the evolution of Syncytins presents an evolutionary conundrum, because none are conserved across mammals, but instead each has emerged independently during mammal evolution through co-option events of distinct, lineage-specific retroviral *Env* sequences [68–71]. Some species such as mouse and human even harbor multiple Syncytins in their genomes that originated at different evolutionary time points and there is also evidence that some Syncytins have been lost during evolution [68].

Could the repeated co-option and turnover of Syncytins reflect convergent adaptation to a persistent evolutionary battle? It has been proposed that the interface between the mother and the fetus in the placenta sets the stage for a conflict whereby the fetus selects for the ability to maximize the transfer of nutrients to itself while the mother, in response, adapts to limit the nutrient transfer and maintain overall homeostasis maximizing her offspring number [74]. This conflict is predicted to result in an evolutionary arms race driving rapid placental evolution and potentially Syncytin evolution. The model is supported by several genetic observations, including certain patterns of gene expression (imprinting; that reveals that the conflict might even start as a mother–father conflict [75,76]) and evolution (positive selection) that are prevalent for placental-specific genes, and at an anatomical level by the remarkable diversification of this organ during mammalian evolution [69,74,77,78].

A placenta feature that might facilitate the recurrent Syncytin co-option for placenta function could be the low level of DNA methylation relative to other tissues, which tends to promote the expression of TEs and endogenous retroviruses in particular in this organ [76,79,80]. In addition to the Syncytins, that is, endogenous retrovirus gene domestication, numerous ERV and other TE sequences have been co-opted as *cis*-regulatory elements to coordinate placental or uterine gene expression during pregnancy [81,82]. Thus, placenta hypomethylation that might facilitate the recurrent recruitment of ERV proteins and placenta-specific regulatory sequences might influence how it adapts to the everlasting evolutionary arms race between the fetus and the mother [78].

Concluding Remarks

In this review, we highlight three major routes by which TE proteins have been domesticated in response to genetic conflicts. First, TE proteins from various classes of elements have been repeatedly co-opted to suppress TEs or retroviruses. The recurrence of this phenomenon may

be explained by the fact that these TE proteins had pre-existing interactions with cellular machinery and with TEs themselves, which can be readily repurposed for TE suppression. A second, unforeseen route invokes the transition from actively transposing elements toward domestication imposed by their own selfish invasive strategies. This route, which is best exemplified by the process of DNA elimination and the formation of IES in ciliates or possibly the transition from self-splicing to spliceosomal introns, likely contributed to increasing complexity in genome organization and function during evolution. Finally, the last route involves the co-option of TE proteins and sequences for seemingly unexpected novel biological processes such as adaptive immune systems of vertebrates [V(D)J recombination] and prokaryotes (CRISPR-Cas). In those examples where TE proteins are co-opted for seemingly completely new functions, the interactions of those proteins, the host cellular machinery, and the interactions of TEs they derive from need to be better characterized if we are to obtain a complete picture of how those novel functions evolved (see Outstanding Questions).

Lastly, we want to highlight the challenges in assigning function to domesticated TE proteins. These challenges are very similar to assigning functions to any other candidate protein and reverse genetics methods are often used. However, from an evolutionary point of view, there is an interest in exploring the function that initially facilitated the domestication of TE proteins. When a TE protein is domesticated to resolve a conflict, it will most likely be incorporated into host cellular pathways. These pathways may evolve over time to a point that may obscure our understanding of the activity or interaction that initially triggered the domestication event. Thus, the present function of a TE-derived protein may not always illuminate the initial process of domestication and, as a consequence, the role of genetic conflicts as the initial driver of TE domestication may remain underestimated.

Acknowledgments

E.B. would like to acknowledge the support from the National Institute of General Medical Sciences of the National Institutes of Health (GM071813). C.F. is supported by grants GM112972, GM059290, HG009391 from the National Institutes of Health. The content of this work is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Supplemental Information

Supplemental information associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.tig.2017.07.011>.

References

- Feschotte, C. and Pritham, E.J. (2007) DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.* 41, 331–368
- Curcio, M.J. and Derbyshire, K.M. (2003) The outs and ins of transposition: from mu to kangaroo. *Nat. Rev. Mol. Cell Biol.* 4, 865–877
- Sun, C. *et al.* (2015) DNA transposons have colonized the genome of the giant virus *Pandoravirus salinus*. *BMC Biol.* 13, 38
- Hancks, D.C. and Kazazian, H.H., Jr (2016) Roles for retrotransposon insertions in human disease. *Mob. DNA* 7, 9
- Beauregard, A. *et al.* (2008) The take and give between retrotransposable elements and their hosts. *Annu. Rev. Genet.* 42, 587–617
- Levin, H.L. and Moran, J.V. (2011) Dynamic interactions between transposable elements and their hosts. *Nat. Rev. Genet.* 12, 615–627
- Lynch, M. (2007) *The Origins of Genome Architecture*, Sinauer Associates
- Wacholder, A.C. *et al.* (2014) Inference of transposable element ancestry. *PLoS Genet.* 10, e1004482
- Carr, M. *et al.* (2012) Evolutionary genomics of transposable elements in *Saccharomyces cerevisiae*. *PLoS One* 7, e50978
- de Koning, A.P. *et al.* (2011) Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* 7, e1002384
- Burt, A. and Trivers, R. (2006) *Genes in Conflict: The Biology of Selfish Genetic Elements*, The Belknap Press of Harvard University Press
- Miller, W.J. *et al.* (1997) Molecular domestication of mobile elements. *Genetica* 100, 261–270
- Feschotte, C. (2008) Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* 9, 397–405
- Joly-Lopez, Z. *et al.* (2016) Phylogenetic and genomic analyses resolve the origin of important plant genes derived from transposable elements. *Mol. Biol. Evol.* 33, 1937–1956
- Joly-Lopez, Z. *et al.* (2012) A gene family derived from transposable elements during early angiosperm evolution has reproductive fitness benefits in *Arabidopsis thaliana*. *PLoS Genet.* 8, e1002931
- Chuong, E.B. *et al.* (2017) Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* 18, 71–86
- Kapusta, A. and Feschotte, C. (2014) Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends Genet.* 30, 439–452

Outstanding Questions

Why do DNA transposon proteins appear to be more prone to domestication than retroelement proteins?

What types of interaction interfaces between host and TEs facilitate TE domestication? Are some interfaces more likely to be preserved and co-opted?

How often does TE domestication lead to convergent molecular innovations?

How are TE genes put under the regulatory control of the host? Does it involve the replacement or modification of TE's ancestral regulatory properties? Does the local genomic environment near a TE's landing site play a major role?

To what extent is the population genetics of a species affecting the propensity and path toward TE domestication?

Are species accumulating more TEs in their genome more likely to co-opt TEs?

18. Bourque, G. (2009) Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Curr. Opin. Genet. Dev.* 19, 607–612
19. Rebollo, R. *et al.* (2012) Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu. Rev. Genet.* 46, 21–42
20. Elbarbary, R.A. *et al.* (2016) Retrotransposons as regulators of gene expression. *Science* 351, aac7247
21. Rice, W.R. (1998) Intergenomic conflict, interlocus antagonistic coevolution and evolution of reproductive isolation. In *Endless Forms: Species and Speciation* (Howard, D.J. and Berlocher, S. H., eds), pp. 261–270, Oxford University Press
22. Gellert, M. (2002) V(D)J recombination: RAG proteins, repair factors, and regulation. *Annu. Rev. Biochem.* 71, 101–132
23. Carmona, L.M. and Schatz, D.G. (2016) New insights into the evolutionary origins of the recombination-activating gene proteins and V(D)J recombination. *FEBS J.* 284, 1590–1605
24. Oettinger, M.A. *et al.* (1990) RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. *Science* 248, 1517–1523
25. Kapitonov, V.V. and Jurka, J. (2005) RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol.* 3, e181
26. Kapitonov, V.V. and Koonin, E.V. (2015) Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. *Biol. Direct* 10, 20
27. Huang, S. *et al.* (2016) Discovery of an active RAG transposon illuminates the origins of V(D)J recombination. *Cell* 166, 102–114
28. Horvath, P. and Barrangou, R. (2010) CRISPR/Cas, the immune system of bacteria and archaea. *Science* 327, 167–170
29. Barrangou, R. *et al.* (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315, 1709–1712
30. van der Ploeg, J.R. (2009) Analysis of CRISPR in *Streptococcus mutans* suggests frequent occurrence of acquired immunity against infection by M102-like bacteriophages. *Microbiology* 155, 1966–1976
31. Jiang, F. and Doudna, J.A. (2017) CRISPR-Cas9 structures and mechanisms. *Annu. Rev. Biophys.* 46, 505–529
32. Makarova, K.S. *et al.* (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* 13, 722–736
33. Krupovic, M. *et al.* (2014) Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. *BMC Biol.* 12, 36
34. Nunez, J.K. *et al.* (2015) Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* 519, 193–198
35. Hickman, A.B. and Dyda, F. (2015) The casposon-encoded Cas1 protein from *Aciduliprofundum boonei* is a DNA integrase that generates target site duplications. *Nucleic Acids Res.* 43, 10576–10587
36. Koonin, E.V. and Krupovic, M. (2015) Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat. Rev. Genet.* 16, 184–192
37. Beguin, P. *et al.* (2016) Casposon integration shows strong target site preference and recapitulates protospacer integration by CRISPR-Cas systems. *Nucleic Acids Res.* 44, 10367–10376
38. Silas, S. *et al.* (2016) Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science* 351, aad4234
39. Shmakov, S. *et al.* (2017) Diversity and evolution of class 2 CRISPR-Cas systems. *Nat. Rev. Microbiol.* 15, 169–182
40. Kapitonov, V.V. *et al.* (2015) ISC, a novel group of bacterial and archaeal DNA transposons that encode Cas9 homologs. *J. Bacteriol.* 198, 797–807
41. Malfavon-Borja, R. and Feschotte, C. (2015) Fighting fire with fire: endogenous retrovirus envelopes as restriction factors. *J. Virol.* 89, 4047–4050
42. Blanco-Melo, D. *et al.* (2017) Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors. *Elife* 6, e22519
43. Yap, M.W. *et al.* (2014) Evolution of the retroviral restriction gene Fv1: inhibition of non-MLV retroviruses. *PLoS Pathog.* 10, e1003968
44. Cam, H.P. *et al.* (2008) Host genome surveillance for retrotransposons by transposon-derived proteins. *Nature* 451, 431–436
45. Johansen, P. and Cam, H.P. (2015) Suppression of meiotic recombination by CENP-B homologs in *Schizosaccharomyces pombe*. *Genetics* 201, 897–904
46. Murton, H.E. *et al.* (2016) Restriction of retrotransposon mobilization in *Schizosaccharomyces pombe* by transcriptional silencing and higher-order chromatin organization. *Genetics* 203, 1669–1678
47. Baum, M. and Clarke, L. (2000) Fission yeast homologs of human CENP-B have redundant functions affecting cell growth and chromosome segregation. *Mol. Cell. Biol.* 20, 2852–2864
48. Ikeda, Y. *et al.* (2017) Arabidopsis proteins with a transposon-related domain act in gene silencing. *Nat. Commun.* 8, 15122
49. Liang, S.C. *et al.* (2015) Kicking against the PRCs – a domesticated transposase antagonises silencing mediated by Polycomb group proteins and is an accessory component of Polycomb repressive complex 2. *PLoS Genet.* 11, e1005660
50. McLaughlin, R.N., Jr *et al.* (2014) Positive selection and multiple losses of the LINE-1-derived L1TD1 gene in mammals suggest a dual role in genome defense and pluripotency. *PLoS Genet.* 10, e1004531
51. Prescott, D.M. (1994) The DNA of ciliated protozoa. *Microbiol. Rev.* 58, 233–267
52. Chen, X. *et al.* (2014) The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. *Cell* 158, 1187–1198
53. Hamilton, E.P. *et al.* (2016) Structure of the germline genome of *Tetrahymena thermophila* and relationship to the massively rearranged somatic genome. *Elife* 5, e19090
54. Vogt, A. *et al.* (2013) Transposon domestication versus mutualism in ciliate genome rearrangements. *PLoS Genet.* 9, e1003659
55. Baudry, C. *et al.* (2009) PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. *Genes Dev.* 23, 2478–2483
56. Dubois, E. *et al.* (2017) Multimerization properties of PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements. *Nucleic Acids Res.* 45, 3204–3216
57. Vogt, A. and Mochizuki, K. (2013) A domesticated PiggyBac transposase interacts with heterochromatin and catalyzes reproducible DNA elimination in *Tetrahymena*. *PLoS Genet.* 9, e1004032
58. Shieh, A.W. and Chalker, D.L. (2013) LIA5 is required for nuclear reorganization and programmed DNA rearrangements occurring during *Tetrahymena* macronuclear differentiation. *PLoS One* 8, e75337
59. Cheng, C.Y. *et al.* (2016) The piggyBac transposon-derived genes TPB1 and TPB6 mediate essential transposon-like excision during the developmental rearrangement of key genes in *Tetrahymena thermophila*. *Genes Dev.* 30, 2724–2736
60. Cheng, C.Y. *et al.* (2010) A domesticated piggyBac transposase plays key roles in heterochromatin dynamics and DNA cleavage during programmed DNA deletion in *Tetrahymena thermophila*. *Mol. Biol. Cell* 21, 1753–1762
61. Arnaiz, O. *et al.* (2012) The *Paramecium* germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. *PLoS Genet.* 8, e1002984
62. Fass, J.N. *et al.* (2011) Genome-scale analysis of programmed DNA elimination sites in *Tetrahymena thermophila*. *G3 (Bethesda)* 1, 515–522
63. Nowacki, M. *et al.* (2009) A functional role for transposases in a large eukaryotic genome. *Science* 324, 935–938
64. Witherspoon, D.J. *et al.* (1997) Selection on the protein-coding genes of the TBE1 family of transposable elements in the ciliates *Oxytricha fallax* and *O. trifallax*. *Mol. Biol. Evol.* 14, 696–706

65. Lambowitz, A.M. and Belfort, M. (2015) Mobile bacterial group II introns at the crux of eukaryotic evolution. *Microbiol. Spectr.* 3, MDNA3-0050-2014
66. Dlakic, M. and Mushegian, A. (2011) Prp8, the pivotal protein of the spliceosomal catalytic center, evolved from a retroelement-encoded reverse transcriptase. *RNA* 17, 799–808
67. Martin, W. and Koonin, E.V. (2006) Introns and the origin of nucleus-cytosol compartmentalization. *Nature* 440, 41–45
68. Esnault, C. *et al.* (2013) Differential evolutionary fate of an ancestral primate endogenous retrovirus envelope gene, the EnvV syncytin, captured for a function in placentation. *PLoS Genet.* 9, e1003400
69. Lavialle, C. *et al.* (2013) Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 368, 20120507
70. Cornelis, G. *et al.* (2015) Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. *Proc. Natl. Acad. Sci. U. S. A.* 112, E487–E496
71. Cornelis, G. *et al.* (2014) Retroviral envelope syncytin capture in an ancestrally diverged mammalian clade for placentation in the primitive Afrotherian tenrecs. *Proc. Natl. Acad. Sci. U. S. A.* 111, E4332–E4341
72. Dupressoir, A. *et al.* (2011) A pair of co-opted retroviral envelope syncytin genes is required for formation of the two-layered murine placental syncytiotrophoblast. *Proc. Natl. Acad. Sci. U. S. A.* 108, E1164–E1173
73. Dupressoir, A. *et al.* (2009) Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. *Proc. Natl. Acad. Sci. U. S. A.* 106, 12127–12132
74. Haig, D. (2008) Placental growth hormone-related proteins and prolactin-related proteins. *Placenta* 29, S36–S41
75. Moore, T. and Haig, D. (1991) Genomic imprinting in mammalian development: a parental tug-of-war. *Trends Genet.* 7, 45–49
76. Haig, D. (2016) Transposable elements: self-seekers of the germline, team-players of the soma. *Bioessays* 38, 1158–1166
77. Chuong, E.B. *et al.* (2010) Maternal-fetal conflict: rapidly evolving proteins in the rodent placenta. *Mol. Biol. Evol.* 27, 1221–1225
78. Chuong, E.B. (2013) Retroviruses facilitate the rapid evolution of the mammalian placenta. *Bioessays* 35, 853–861
79. Haig, D. (2015) Going retro: transposable elements, embryonic stem cells, and the mammalian placenta (retrospective on DOI 10.1002/bies.201300059). *Bioessays* 37, 1154
80. Schroeder, D.I. *et al.* (2015) Early developmental and evolutionary origins of gene body DNA methylation patterns in mammalian placentas. *PLoS Genet.* 11, e1005442
81. Lynch, V.J. *et al.* (2015) Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Rep.* 10, 551–561
82. Chuong, E.B. *et al.* (2013) Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat. Genet.* 45, 325–329
83. Guerin, F. *et al.* (2017) Flow cytometry sorting of nuclei enables the first global characterization of *Paramecium* germline DNA and transposable elements. *BMC Genomics* 18, 327
84. Chen, X. and Landweber, L.F. (2016) Phylogenomic analysis reveals genome-wide purifying selection on TBE transposons in the ciliate *Oxytricha*. *Mob. DNA* 7, 2
85. Gogoi-Doring, A. *et al.* (2016) Genome-wide profiling reveals remarkable parallels between insertion site selection properties of the MLV retrovirus and the piggyBac transposon in primary human CD4⁺ T cells. *Mol. Ther.* 24, 592–606
86. Mitra, R. *et al.* (2008) piggyBac can bypass DNA synthesis during cut and paste transposition. *EMBO J.* 27, 1097–1109
87. Smit, A.F. and Riggs, A.D. (1996) Tiggers and DNA transposon fossils in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* 93, 1443–1448
88. Casola, C. *et al.* (2008) Convergent domestication of pogo-like transposases into centromere-binding proteins in fission yeast and mammals. *Mol. Biol. Evol.* 25, 29–41
89. d'Alençon, E. *et al.* (2011) Characterization of a CENP-B homolog in the holocentric Lepidoptera *Spodoptera frugiperda*. *Gene* 485, 91–101
90. Mateo, L. and Gonzalez, J. (2014) Pogo-like transposases have been repeatedly domesticated into CENP-B-related proteins. *Genome Biol. Evol.* 6, 2008–2016
91. Henikoff, S. *et al.* (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* 293, 1098–1102
92. McLaughlin, R.N., Jr and Malik, H.S. (2017) Genetic conflicts: the usual suspects and beyond. *J. Exp. Biol.* 220, 6–17
93. Malik, H.S. (2005) *Mimulus* finds centromeres in the driver's seat. *Trends Ecol. Evol.* 20, 151–154
94. Roach, K.C. *et al.* (2012) Rapid evolution of centromeres and centromeric/kinetochore proteins. In *Rapidly Evolving Genes and Genetic Systems* (Singh, R.S., ed.), pp. 83–93, Oxford University Press
95. Pardue, M.L. and Debaryshe, P. (2011) Adapting to life at the end of the line: how *Drosophila* telomeric retrotransposons cope with their job. *Mob. Genet. Elem.* 1, 128–134
96. Pardue, M.L. and DeBaryshe, P.G. (2011) Retrotransposons that maintain chromosome ends. *Proc. Natl. Acad. Sci. U. S. A.* 108, 20317–20324
97. Servant, G. and Deininger, P.L. (2016) Insertion of retrotransposons at chromosome ends: adaptive response to chromosome maintenance. *Front. Genet.* 6, 358
98. Lee, Y.C.G. *et al.* (2016) Recurrent innovation at genes required for telomere integrity in *Drosophila*. *Mol. Biol. Evol.* 34, 467–482
99. de Lange, T. (2004) T-loops and the origin of telomeres. *Nat. Rev. Mol. Cell Biol.* 5, 323–329
100. Ohzeki, J. *et al.* (2002) CENP-B box is required for de novo centromere chromatin assembly on human alphoid DNA. *J. Cell. Biol.* 159, 765–775
101. Okada, T. *et al.* (2007) CENP-B controls centromere formation depending on the chromatin context. *Cell* 131, 1287–1300
102. Spencer, T.E. *et al.* (2003) Receptor usage and fetal expression of ovine endogenous betaretroviruses: implications for coevolution of endogenous and exogenous retroviruses. *J. Virol.* 77, 749–753
103. Best, S. *et al.* (1996) Positional cloning of the mouse retrovirus restriction gene Fv1. *Nature* 382, 826–829
104. Ito, J. *et al.* (2013) Refrex-1, a soluble restriction factor against feline endogenous and exogenous retroviruses. *J. Virol.* 87, 12029–12040
105. Kozak, C.A. (2014) Origins of the endogenous and infectious laboratory mouse gammaretroviruses. *Viruses* 7, 1–26
106. Wu, T. *et al.* (2005) Rmc2, a xenotropic provirus in the Asian mouse species *Mus castaneus*, blocks infection by polytropic mouse gammaretroviruses. *J. Virol.* 79, 9677–9684
107. Sugimoto, J. *et al.* (2013) A novel human endogenous retroviral protein inhibits cell-cell fusion. *Sci. Rep.* 3, 1462
108. Lingner, J. *et al.* (1997) Reverse transcriptase motifs in the catalytic subunit of telomerase. *Science* 27, 561–567
109. Wicker, T. *et al.* (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8, 973–982

CHAPTER TWO

Domestication of *PIF* transposable element transposases as regulatory proteins in *Drosophila*

Abstract

Proteins encoded by transposable elements (TEs) play a vital role in their proliferation. In recent years an increasing number of studies have shown that these TE proteins have contributed to the emergence of novel host proteins. Here we study the function of several of such proteins that were co-opted or 'domesticated' from the transposase of *PIF/Harbinger* elements in *Drosophila* called *Drosophila PIF Like Genes (DPLGs)*. There are four *DPLGs* in *D. melanogaster* and these genes are highly diverged, are under purifying selection, and likely arose through independent domestication events. *DPLGs* co-express with transcription factors across development, and *DPLG2-4* are able to localize with DNA in the nucleus of the ovaries suggesting that *DPLGs* were potentially domesticated as regulatory proteins. RNA *in situ* hybridization results show that *DPLG1-4* have strikingly similar pattern of transcript localization during embryogenesis and also show overlap of transcript localization in the gonads suggesting functional overlap between these genes. Protein localization studies of HA-tagged *DPLG2-4* in the ovaries also show overlap in protein localization further supporting their functional relatedness. Results from RNA-Seq analyses from the ovaries of *DPLG1* and *DPLG4* null mutants show that they contribute to mild but significant overlapping changes in gene expression. Further results from experiments in the mutants provide evidence for potential genetic interaction between *DPLG1* and *DPLG4*. Our data also provide evidence that *DPLG4* affects viability, fertility and survival in *D. melanogaster*. Together, we conclude that *PIF* transposases in *D. melanogaster* have been domesticated as regulatory proteins and affect multiple traits in flies. The genetic interaction and functional overlap of these independently domesticated

transposases from the same TE family support a stepping stone model in which domestication of a transposase might promote the domestication of related transposases.

Introduction

Transposable Elements (TEs) are genetic units that are able to move and amplify within host genomes, are found in diverse groups of organisms ranging from single-celled organisms to complex multicellular eukaryotes and have been shown to occupy a large fraction of various genomes (Lander et al. 2001; Feschotte and Pritham 2007; Schnable et al. 2009). In fact, at least half of the human genome is comprised of TEs, in contrast to less than 2% accounting for protein-coding genes (Cordaux and Batzer 2009; de Koning et al. 2011). According to the mechanism of transposition, TEs are divided into two classes: Class I elements or retrotransposons which transpose via RNA intermediate that is reverse transcribed, and Class II elements, or DNA transposons that move directly via DNA (Cordaux and Batzer 2009). All TEs have sequence features and encode proteins that aid in their mobilization and/or amplification (Curcio and Derbyshire 2003). As a result of their transposition activities, TEs can cause deleterious effects in the host if they integrate into regulatory or coding regions (Sinzelle et al. 2009). Moreover, due to sequence similarity between homologous TE copies at paralogous loci, they can induce ectopic recombination that can lead to translocations, inversions, and deletions causing detrimental effects to the host (Sinzelle et al. 2009). Thus, TEs are selfish elements and generally do not contribute positively towards the fitness of the host, and, as a consequence, TE insertions are either removed by purifying selection or disabled by random mutations at the mutation rates of the host genome (Kidwell and Lisch 2000).

Interestingly, TEs have also benefited the host. Genes that TEs harbor for transposition have been co-opted by hosts through a process known as molecular domestication (Casola et al. 2008; Sinzelle et al. 2009; Vogt et al. 2013; Mateo and González 2014; Jangam et al. 2017). In

principle, any protein of a TE can be domesticated, however, evidence accumulated up to now indicates that the transposases of DNA TEs are more prone to domestication than any other TE protein despite the general abundance of retrotransposons (Feschotte 2008; Jangam et al. 2017). Transposases encoded by the TEs have the ability to translocate into the nucleus, recognize and bind to the TE's TIRs (Terminal Inverted Repeats), and catalyze their transposition to new genomic sites (Curcio and Derbyshire 2003; Feschotte and Pritham 2007). The DNA binding ability of the transposase can potentially be recycled by the host to bind specific DNA sequences and regulate gene expression contributing towards several host cellular functions including TE silencing (Cam et al. 2008; Johansen and Cam 2015). The domestication of transposases has been described in a variety of organisms (Casola et al. 2008; Joly-Lopez et al. 2012; Pavelitz et al. 2013; Mateo and González 2014; Liang et al. 2015; Duan et al. 2017). Although a lot of these genes have good support for domestication, most of them still remain to be functionally characterized.

Several domesticated TEs from *PIF-Harbinger-IS5* superfamily of DNA transposons have been described in *Drosophila* (Casola et al. 2007). All *PIF* TEs described in plants and animals up to date are characterized by the presence of two independent open reading frames (ORFs) (Zhang et al. 2001; Jiang et al. 2003; Kapitonov and Jurka 2004; Zhang et al. 2004; Casola et al. 2007; Grzebelus et al. 2007). The first ORF encodes for a transposase protein that has an N-terminal HTH DNA binding domain and a C-terminal DDE catalytic domain, and the second ORF is a *PIFp2* gene which encodes for a protein that has an N-terminal SANT/Myb/MADF domain and a predicted C-terminal BESS domain (Casola et al. 2007; Sinzelle et al. 2008). In zebrafish, it has been shown that the proteins from the two ORFs interact with each other and the *PIFp2* protein is responsible for the translocation of the complex into the nucleus where it can recognize the

ends of the TE and catalyze its transposition (Sinzelle et al. 2008). In *Drosophila*, there have been seven events of domestication of the transposase of *PIF Like Transposons*; these genes were named *Drosophila PIF Like Genes 1 to 7 (DPLG1-7)*; (Casola et al. 2007). Sequence analyses show that all *DPLGs* have likely acquired disabling mutations in their DDE catalytic domain but have an intact HTH DNA binding domain (Casola et al. 2007). Thus authors proposed that *DPLGs* have lost the ability to cut DNA, however they have likely retained the ancestral DNA binding ability. All *DPLGs* are old genes and are under purifying selection. *DPLG1-4* are found in all *Drosophila* species, whereas *DPLG6* and *DPLG7* have been lost in some lineages of *Drosophila*, and *DPLG5* emerged in the *Sophophora* subgenus before the divergence of the *melanogaster* and *obscura* species groups but was later lost from the *melanogaster* subgroup. Interestingly, *DPLG1-7* likely originated from at least three independent domestication events (Casola et al. 2007).

To explore the function of these *PIF*-transposase-derived genes, we took advantage of the genetic tools available in *Drosophila melanogaster* and focused on the *DPLGs* present in this species (*DPLG1-4*). These are very old genes (domesticated at least 50 million years ago prior to *Sophophora/Drosophila* split), conserved but highly diverged from each other (Casola et al. 2007; Wiegmann et al. 2011; Misof et al. 2014). *DPLG1-4* are all transcribed in several tissues of the adult fly, however they tend to be relatively highly transcribed in the ovaries and the adult fly brain (Lovering et al. 2018). We provide data supporting that *PIF* transposases in *D. melanogaster* have been likely domesticated as regulatory proteins. Using a reverse genetics approach, we show that a subset of *DPLGs* contribute towards viability, fertility, and survival in flies. Further, we also provide support for a stepping stone model in which domestication of a transposase promotes domestication of related transposases.

Materials and Methods

Drosophila stocks

DPLG2-4 were tagged with HA in the Buszczak lab at UT Southwestern following a previously described protocol (Chan et al. 2012). *w¹¹¹⁸* stock for the knockout (KO) studies and stock expressing the Cre recombinase (BDSC stock #34516) were obtained from the Buszczak lab at UT Southwestern. *w¹¹¹⁸* for the knockdown (KD) studies was obtained from Vienna Drosophila Resource Center (VDRC). The *Act5C-Gal4* line was received from Bloomington Drosophila Stock Center (BDSC stock #4414), RNAi line for *DPLG4* was obtained from VDRC (GD library; VDRC ID: 40639; (Dietzl et al. 2007)). The line containing insertion of *P-element* in the first exon of *DPLG4* and the line used to mobilize it were obtained from Bloomington (BDSC stock #17472 and #1808, respectively).

Generating null mutants

Knockout (KO) lines for *DPLG1* and *DPLG4* were generated in collaboration with Varsha Bhargava in Buszczak lab at UT Southwestern using the CRISPR-Cas9 technology. Three constructs were generated for each KO, out of which two were the source of guide RNAs (gRNAs) that were responsible for guiding the double strand breaks on either side of the gene of interest and the third construct was a donor vector which was used during repair and enabled the replacement of the gene of interest with a DsRed cassette (Addgene plasmid # 51019). The guide RNAs were designed using CRISPR target finder available at <http://flycrispr.molbio.wisc.edu/tools>. Oligos

were ordered from IDT, Inc., and when annealed, formed small double stranded DNA fragments with sticky ends. The plasmid pU6-BbsI-chiRNA (Addgene plasmid # 45946) was cut using *BbsI* restriction enzyme (NEB, Inc.) and the annealed oligos (Supplementary table 1) were directionally cloned into this plasmid.

To produce the donor vector, ~1kb-long gene blocks (Supplementary table 1) corresponding to the 5' and 3' regions flanking the gene of interest and upstream and downstream of the sites targeted by the gRNAs were ordered from IDT, Inc. The pHD-DsRed-attP (provided by the Buszczak lab) plasmid was cut using *EcoRI* and *XhoI* restriction enzymes (Promega Corporation). These gene blocks contained ~30 bp overlapping sequence with the region flanking the restriction sites in pHD-DsRed-attP. The gene blocks were cloned into the cut plasmid using NEBuilder® HiFi DNA Assembly Master Mix (NEB, Inc.), following the protocol provided in the kit. This generated a plasmid that contained the DsRed cassette (which is under a regulatory region that drives expression in the eye) flanked by the 1 kb regions flanking the gene of interest in the genome. These constructs were put in a total volume of 200 µl with 250 ng/µl of donor plasmid, 20 ng/µl of each guide RNA and were sent to Rainbow Transgenic Flies, Inc. for injection. Plasmid concentrations appear to make a difference for the CRISPR-Cas9 gene replacement KO technology to work efficiently. Constructs for *DPLG1* and *DPLG4* were injected into nos-Cas9 attP2 and nos-Cas9 attP40 strains (Kondo and Ueda 2013), respectively. The flies that emerged from the injected embryos were crossed to the w^{1118} stock, the progeny were screened for red fluorescence in the eye and the lack of the gene of interest was confirmed by PCR. To control for the background effects, flies from a KO line were backcrossed with the w^{1118} for six generations as outlined in previous publications (Slawson et al. 2011; Chandler et al. 2013).

Three backcross replicates per KO were produced. Since the KOs have the dominant DsRed marker in place of the gene of interest, it was used to follow heterozygote mutants during backcrossing. After six generations, the heterozygote mutants were crossed among themselves, and individuals were crossed again to produce the three homozygote backcross mutant lines for the two KOs. PCR was used to screen for homozygote mutant fly pairs.

Generating *DPLG4* *P-element* excision mutant

We obtained a line from BDSC that contained a *P-element* insertion at the beginning of the first exon of *DPLG4* causing a disruption in its coding sequence, resulting in a potential null mutant of *DPLG4*. On close examination, we discovered that an alternative start site was present in the *P-element* what could restore the coding region of *DPLG4*, potentially resulting in the expression of a shorter but functional protein. Thus we decided to excise the *P-element* out to generate a frameshift mutant of *DPLG4*. Javier Rio generated the *P-element* excision frameshift mutant. This excision mutant was generated by crossing the *P-element* insertion line to a line that expressed a source of transposase (see the description of the fly stocks above) that would facilitate the excision of the *P-element*. Since the *P-element* incorporated a mini white gene, its excision would result in flies that lacked the red coloration in their eyes. These flies were made homozygotes and *DPLG4* was sequenced to screen for lines that did not express the intact protein. We generated a frameshift mutant line that contained a 31 bp insertion relative to wild type *DPLG4* which resulted in a frameshift mutation and introduction of multiple stop codons. A line that contained complete restoration of wildtype *DPLG4* as a result of perfect excision or

repair of the *P-element* excision was used as a control. These stocks were verified by sequencing and Supplementary figure 1 illustrates the 31 bp insertion.

Generating lines for antibody staining

DPLGs tagged with HA were used to study their protein localization. The HA tag has been used before to successfully study the localization of HARBI1, a related domesticated transposase, in mammalian cells and shown not to affect protein conformation (Sinzelle et al. 2008). *DPLG2-4*-HA tagged lines were generated using recombineering technology (Chan et al. 2012) and the RFP marker flanked by loxP sites was used to screen for successful integration of the construct containing the tagged gene. Before performing the protein localization studies, the RFP marker was removed from the tagged lines by crossing them to a stock expressing the Cre recombinase because removing the RFP marker enhances the antibody staining. In these stocks, the HA tagged genes are expressed from the attB site where they were initially inserted (i.e., they have not been moved to the endogenous gene site), but are flanked by a big part of the endogenous gene region and expected to express the gene in the wildtype pattern (Chan et al. 2012).

Generating probes for *in situ* hybridization

The protocol for generating RNA probes for *in situ* hybridization was adapted from Morris *et al.* (Morris et al. 2009). Exonic sequences of *DPLG1-4* were amplified using the primers specific to the respective *DPLGs* with the promoter for *in vitro* transcription added to one or the other sides to produce sense and antisense probes, respectively. Oligo sequences used are given in

Supplementary table 2. The PCR products were analyzed in 1% agarose gel and cleaned with PCR cleanup kit from Promega Corporation. *In vitro* transcription was performed on these DNA fragments to generate sense and antisense RNA probes using the DIG RNA Labelling kit (SP6/T7; Roche Ltd.).

Sample collection and fixation for *in situ* hybridization

The protocol for embryo collection and fixation was obtained from Tautz and Pfeifle (Tautz and Pfeifle 1989). Embryos were collected on agar plates within 24 hours and were dechorinated by exposing them to 50% bleach for 2 minutes. The embryos were then washed with DI water and put into a vial containing 10 ml of 4% paraformaldehyde (PFA) in PBS and 40 ml of heptane and were gently shaken for 30 minutes. PFA and heptane were then sucked out of the vial replaced with 5 ml of methanol and vigorously shaken for 1 minute. The embryos that had fallen to the bottom of the vial were transferred to a vial and washed with methanol three times for 10 minutes and stored in methanol for long term at -20°C.

The protocol used by Morris et al. was followed for the collection of testis and ovaries and their fixation (Morris et al. 2009). Testis from less-than-one-day-old males were obtained by dissection in 1X PBS within 20 minutes and fixed in 4% PFA in PBS for 30 minutes. Virgin females were collected and aged for three days. The ovaries from these females were dissected within 15 minutes and fixed in 4% PFA in PBS for 15 minutes.

RNA *in situ* hybridization

The protocol for RNA *in situ* hybridization was adopted from Morris et al. (Morris et al. 2009) with some modifications. Embryos were collected and permeabilized using proteinase K (2µg/ml in PBST) for ~5 minutes at RT or ~1 hour at 4°C and fixed. Testis from young males and ovaries did not need to be treated with proteinase K. The samples were then hybridized overnight with probes (1:100 concentration) in hybridization buffer at 55°C. Samples were thoroughly washed with PBST (PBST + 0.1% Tween-20) and incubated in 0.5 ml of anti-digoxigenin (1:2000 in PBST) overnight at 4°C. After thoroughly washing the sample with PBST, the samples were put in developing solution for color development. After desired development of the color, both the control and the experimental reactions were stopped at the same time by washing with PBST (4 times for 10 minutes). The samples were then put in 30% glycerol for 30 minutes, followed by 50% glycerol and then to 70% glycerol. The samples were then mounted on slides in Vectashield mounting media (Vector Laboratories, Inc.) for imaging. *In situ* hybridization showing transcript localization for *DPLG1* and *DPLG4* in the ovarioles were performed by Susana Domingues.

Antibody staining of *Drosophila* ovaries

Virgin females were collected and were aged for three days supplemented with yeast. The ovaries from these females were dissected within 15 minutes and fixed in 4% PFA in PBS for 10 minutes. The ovaries were then thoroughly washed with PBT (PBS + 0.3% Triton X). Primary antibody (Anti-HA; Cat # C29F4, Cell Signaling Technology, Inc.) was added to a concentration of 1:200 in PBT and stored at 4°C overnight. Next day, ovaries were thoroughly washed and then the secondary antibody (Anti-Rabbit; Cat # A11008, Invitrogen, Thermo Fisher Scientific) was added at a concentration of 1:300 in PBT and stored at RT for 6 hours. The ovaries were washed

and incubated in TO-PRO™-3 Iodide (Invitrogen, Thermo Fisher Scientific) at a concentration of 1:1000 for one hour and washed again thoroughly. The ovarioles were then mounted on slides in Vectashield mounting media (Vector Laboratories, Inc.) for imaging. Imaging of the localization of DPLG4-HA in the whole ovariole was performed in the confocal microscope at UT Arlington (Zeiss LSM 510 Confocal Microscope) while the rest of the images were taken by Varsha Bhargava using the confocal microscope in the Buszczak lab at UT Southwestern.

RNA-Seq experimental design and data analyses

Forty ovaries were dissected from five-day-old KOs and control (w^{1118}) female flies. There were three backcrossed lines of the KO line into w^{1118} for every gene and three replicates for the control (See details above). RNA was extracted using Direct-zol kit (Zymo Research) and stored at -70°C. RNA was sent to Genomics Core Facility at Cornell University for library preparation that selected for only poly-A enriched RNAs and were sequenced using Illumina NextSeq500 single-read platform. The reads from the sequencing were checked for their quality using FastQC program (Andrew 2010). For analyzing differential gene expression, STAR aligner (version 2.5; (Dobin et al. 2013)) was used to map the RNA-Seq reads to the *Drosophila* reference genome (genome assembly BDGP6.88), HTseq (Anders, et al. 2015) was used to count the reads that corresponded to genes and DESeq2 (Love et al. 2014) was used for differential expression (DE) analyses with a FDR of 5%. FlyMine was used for enrichment analyses of DE genes (Lyne et al. 2007). For analyzing differential TE expression, STAR aligner was used to map the RNA-Seq reads to the *Drosophila* reference genome with an additional flag --outFilterMultimapNmax 100.

TEtranscripts (Jin et al. 2015) was used to generate the counts of reads that mapped to TEs and genes. DESeq2 was used for differential expression analyses of TEs with a FDR of 5%.

Viability and fertility tests

To test the viability during embryogenesis, 100 embryos were lined up in an agar plate and the number of larvae that hatched after 32 hours were counted. For postembryonic viability tests, 50 larva were obtained from collected embryos after 24 hours and transferred into a vial with food, and the number of adults that emerged were counted after 15 days. All procedures had five replicates and were performed at 25°C.

To study viability in KD flies, reciprocal crosses of *Act5C-Gal4* (ubiquitous driver) were performed with *UAS-DPLG4-RNAi* line and the progeny was counted and compared to progeny resulting from *Act5c-Gal4* crossed with *w¹¹¹⁸* and *UAS-DPLG4-RNAi* crossed with *w¹¹¹⁸*.

Fertility was also studied in KD and KO flies. To test for male fertility, two males were crossed with two control females, and to test for female fertility two females were crossed with two control males. The parents were discarded after 8 days and the total number of offspring was counted after 15 days. All crosses had five replicates and were performed at 25°C.

To test for fertility changes as flies aged, adult flies were placed in chamber with plate containing agar mixed with corn syrup and yeast paste. The flies were let to lay eggs and the eggs were collected in a 15 ml conical tube after 24 hours. The eggs were then washed 2-3 times with 1X PBS and 32 μ l of eggs were transferred to a bottle containing media (three bottles per sample). The bottles were kept at 25°C and the adults were collected within two days. *DPLG4-KO* or *w¹¹¹⁸*

females were kept with w^{1118} males in presence of wet yeast. Every 10 days, older males were replaced with ~4 day old males so that there was no shortage of sperm supply and all flies were transferred into fresh media in presence of wet yeast every 3 days. From this pool of females and w^{1118} males, one female was randomly selected and put into a fresh vial in presence of wet yeast with 10 replicates for KO and control females. These females were let to lay eggs for three days and then discarded. This procedure was repeated after every three days up to day 30. The resulting number of progeny from these vials was counted after 15 days of the female being discarded. Two-way ANOVA was used to test if there was a significant difference between the KO and control with increased age, and t-test was used to test if there was a significant difference between the KO and control at a particular age.

Survival assay

The survival assay was adopted from (Linford et al. 2013). Flies were placed in chambers with plate containing agar mixed with corn syrup and yeast paste at the center of the plate. The eggs laid by the flies were collected in a 15 ml conical tube. The eggs were washed 2-3 times with 1X PBS and 32 μ l of eggs were transferred to a bottle containing media (three bottles per sample). Bottles were kept at 25°C and the adults were collected within two days. Following this, male and female flies were mated for 2 days and 30 flies were placed into a vial containing same gender with 10 replicates. The position of the vials was randomized to avoid any bias associated with the vial location in the incubator. For the first three weeks, flies were transferred into a new vial containing fresh food every two days and after 3 weeks flies were transferred to a new vial three times a week without anesthesia. During transfer the following data were recorded: 1) the age

of the flies, 2) dead flies in the old vial, 3) dead flies in the new vial, 4) total new deaths. If a fly escaped or died of unnatural cause, note of this event was recorded. This process was continued till all the flies died. Analysis of the data was performed in R using the “Survival” package (Therneau and Lumley 2011).

Quantitative RT-PCR

RNA was extracted from 5 males from each reciprocal KD cross, total of 10 males for control (UAS-*DPLG4*-RNAi crossed with *w¹¹¹⁸*) and KD (*Actin5c*-GAL4 crossed with UAS-*DPLG4*-RNAi) using Direct-zol kit (Zymo Research). cDNA synthesis, primer design and Quantitative PCR (qPCR) was done using the protocol outlined in Schmittgen and Livak (Schmittgen and Livak 2008). The *Rp49* gene was used as the internal control and the primers used followed the design by Lu and Clark (Lu and Clark 2010). Primers for amplifying *DPLG4* are provided in Supplementary table 2. GoScript™ Reverse Transcriptase (Promega Corporation) was used to generate cDNA, and qPCR was performed using Green-2-Go qPCR Mastermix (Bio Basic, Inc.).

Results

DPLGs show similar patterns of transcript localization

According to the RNA-Seq data depicted in FlyBase, all *DPLGs* show low to moderately high expression during embryogenesis, and tend to have relatively higher expression pattern in the brain and the gonads (specially in ovaries; (Brown et al. 2014; Gramates et al. 2017). Because of this, RNA *in situ* hybridization was used to explore the transcript localization of *DPLGs* during

embryogenesis and in the gonads. *piwi* was used as a positive control and sense probes were used as negative controls (Supplementary Figure 2 and 3). As expected, *piwi* localizes in the gonads of the embryos and larva (Supplementary Figure 2 and 3). Strikingly, all *DPLGs* show very similar patterns of transcript localization (Figure 1). In the first two hours of embryogenesis, all *DPLGs* show ubiquitous transcription localization in the embryos. At stage 1-3, all *DPLGs* are still broadly expressed in the embryos, however the transcripts of *DPLG2-4* start localizing towards the posterior part of the embryo where the pole cells reside around stage 4-5. In contrast, the posterior localization of *DPLG1* was not observed at this stage (Figure 1). In the later stages of embryogenesis, the transcripts of all *DPLGs* start to localize in the ventral nerve chord (Figure 1) suggesting that *DPLGs* may have a role during neurogenesis. In addition to the nervous system, transcripts of all *DPLGs* also seem to localize in the embryonic midgut (Figure 1).

During oogenesis, the transcripts of *DPLG1-3* are abundant in the nurse cells, whereas *DPLG4* transcripts are detected starting from the posterior part of the germarium all the way in nurse cells at the late stages of oogenesis (Figure 2). During spermatogenesis, *DPLG1* transcripts mostly localize in the primary spermatocytes and to a lesser extent in the anterior tip of the testis where the mitotic cells, and the stem cells reside (Figure 2). *DPLG2* and *DPLG3* transcripts are detected only after late primary spermatocytes up to round spermatids in the testis (Figure 2). Just like in the ovaries, *DPLG4* transcripts are detected very broadly in the testis. *DPLG4* transcripts are detected from early stages of spermatogenesis all the way to round spermatids (Figure 2). Taken together, these data point to differential regulation, but potentially redundant function of *DPLGs* during embryogenesis and gonadogenesis.

DPLGs localize in the nucleus like regulatory proteins

Ovaries were selected as the tissue to more precisely study the localization of the HA-tagged *DPLG* proteins as these genes showed abundant transcript localization in the ovarioles (Figure 2). Additionally, oogenesis in *D. melanogaster* is very well studied and is ideal for comparisons of protein localization studies. The results revealed that *DPLG2-4-HA* are translated into their respective protein product and show distinct localization during oogenesis.

DPLG2-HA localizes towards the anterior tip of the germarium where the stem cells and early differentiating cells reside (Figure 3B and 3C). During later stages of oogenesis, DPLG2-HA localizes exclusively in the nucleus of the germline cells (Figure 3D). In contrast, DPLG3-HA localizes throughout the germanium of the ovariole (Figure 3F) and is detected in the nucleus of the germline cells as well as the nucleus of somatic cells in subsequent stages (Figure 3G). In addition, DPLG3-HA is also detected in the nucleus of the oocyte (Figure 3G). DPLG4-HA is nearly absent in the anterior and shows some localization signals in the posterior part of the germarium (Figure 3I). In the later stages of oogenesis, DPLG4-HA shows a pattern of expression similar to DPLG3-HA, where DPLG4-HA localizes in the nucleus of the germline and the somatic cells in the ovariole (Figure 3J). However, unlike DPLG3-HA, for DPLG4-HA we observe more intense staining of somatic nuclei than germ cell nuclei (Figure 3J). This data also corroborates the *in situ* hybridization data and illustrates that DPLGs have overlapping expression patterns in the ovaries. All DPLGs tagged with HA seem to have not continuous but rather punctuated nuclear localization. Although, some localization signals do not overlay with DNA, there are extensive signals for DPLG2-4 that show overlay with DNA supporting the hypothesis that DPLGs are likely involved in chromatin or transcription regulation (Figure 2).

KO of *DPLG1* and *DPLG4* and their effects on viability and fertility

We produce three knockout lines for two *DPLGs*, *DPLG1* and *DPLG4*, (i.e., gene replacement with RFP using CRISPR-Cas9) by producing three backcrossing lines of the two KOs to w^{1118} (see details in Materials and Methods). These three KO replicates in the w^{1118} background were verified initially by PCR and confirmed by the absence of complete transcripts (based on RNA-Seq data) of both genes in the respective KO lines (Figure 4 A and B).

Viability during embryogenesis and post embryogenesis stages were studied in these null mutant backcrossed lines of *DPLG1* and *DPLG4*, and compared to the w^{1118} (control). *DPLG1-KO* flies showed a significant reduction in embryonic viability (Figure 4C), however they displayed an increase in post embryonic viability (more details in Materials and Methods; Figure 4D). Consistent with these results, knockdown (KD) of *DPLG1* ubiquitously using the UAS-GAL4 system did not show any effect in total viability compared to the controls (data not shown). In contrast, *DPLG4-KO* flies showed a lower viability compared to the control in both embryonic and post embryonic stages (Figure 4 E and F). This phenotype was corroborated by an independent KD experiment using UAS-GAL4 system, where depletion of *DPLG4* RNA ubiquitously caused reduction in the viability of flies compared to the controls (Figure 4G). Most embryos that did not survive as a result of loss of *DPLG4* had developed to late embryonic stages (data not shown) suggesting that this reduction in viability was due to defects in the embryos and not due to defects in the fertility of the *DPLG4-KO* females. Interestingly, this reduction in viability was not observed in an independent frameshift mutant of *DPLG4* (Supplementary figure 3) suggesting that the viability defect observed as a result of loss of *DPLG4* may depend on the genetic

background. Alternatively, *DPLG4 P-element* excision mutant may not represent a true null mutant. Taken together these data support that loss of *DPLG1* reduces embryonic viability while in contrast increases post-embryonic viability and loss of *DPLG4* results in reduction of viability that may be background dependent.

Both males and females of *DPLG1*-KO flies had a higher fertility compared to the control (Figure 4I) whereas *DPLG4*-KO flies did not show any significant difference in the fertility compared to the control flies (Figure 4J). These results showed that loss of *DPLG1* increases fertility in both males and females while loss of *DPLG4* does not have any fertility effects. For *DPLG4*, fertility was also tested under the stress of starvation and with increasing age. There was no significant effect in fertility under starvation (See Supplementary Results). Interestingly, loss of *DPLG4* resulted in increase in fertility of females with increase in age (two-way ANOVA, $p = 1.8e-05$). Using t-test we discovered that there was no significant effect in fertility of young *DPLG4*-KO females (up to day 15) (Figure 4K), however older females from day 21 showed a significantly higher fertility compared to the control (Figure 4K) suggesting that the effect of age in the fertility as a result of loss of *DPLG4* was restricted to older females. In fact the average fertility of the *DPLG4*-KO females stayed consistent throughout the experiment (up to day 30) while the fertility of the control females started to drastically decline towards day 21 and were sterile after day 27 (Figure 4K).

Since *DPLG1* and *DPLG4* showed differential effects compared to the control, we wanted to test if the effects would be exaggerated in the double mutant of *DPLG1* and *DPLG4*. We generated a double mutant of *DPLG1* and *DPLG4* in the w^{1118} background and Fatema B. Ruma tested for effects in embryonic viability, post embryonic viability, and female fertility.

Interestingly, there was no significant difference in embryonic viability despite reduction in the viability of both *DPLG1-KO* and *DPLG4-KO* embryos pointing to the existence of genetic interactions between the two genes. We are still in the process of obtaining the data for post embryonic viability test and fertility test for the double mutants.

Differentially expressed genes in ovaries of *DPLG1-KO* and *DPLG4-KO* show significant overlap

To begin exploring the molecular phenotype of *DPLG* mutants, we performed RNA-Seq of ovaries of *DPLG1-KO* and *DPLG4-KO* flies from each of the 3 independently backcrossed lines (i.e., triplicate) and 3 replicates for the control line. As we mentioned above, no reads from the respective *DPLGs* excised region were obtained for the KO backcrosses (Figure 4 A and B) supporting that we have obtained KOs for the genes under study. Results from FastQC showed that all raw data passed quality control test to be used for downstream analyses (data not shown). One of the control samples (*w¹¹¹⁸ #1*) showed an increased proportion of reads mapping to multiple loci (27.66% of total reads instead of ~2% observed for all other samples) out of which most were mapping to regions that produced 5SrRNA. Presence of this large proportion of reads mapping to multiple loci did not significantly affect the results (see Supplementary Results, Supplementary table 3 for the analyses excluding this sample) because reads mapping to multiple locations are routinely removed from the analyses. So, below we present the analyses including this sample.

Principal Components Analysis (PCA) of gene transcription for the samples was performed. The PCA plot of the first two principal components did not show very distinct

clustering of KOs from the control suggesting that the effects due to loss of *DPLG1* and *DPLG4* were not drastic (Supplementary Figure 4). *DPLG4*-KO ovaries did not contain a lot of DE genes. In contrast, *DPLG1*-KO ovaries contained a large number of DE genes (Figure 5A). There were only 131 genes DE genes in *DPLG4*-KO ovaries, where *DPLG4* was the most downregulated gene. There were 89 upregulated genes and 42 downregulated genes (Figure 5A) out of which there were no enriched GO terms or pathways for upregulated or downregulated genes. *DPLG1*-KO ovaries contained a total of 848 DE genes out of which 292 were upregulated and 556 were downregulated (Figure 5A). For upregulated genes, the enriched GO terms were cytoplasmic translation, and ribosome biogenesis. Enriched pathways were associated with non-sense mediated decay, translation, metabolism, and p53 independent DNA damage pathways (Supplementary table 3). For downregulated genes, most GO terms were associated with general development, regulation of metabolic processes, female gamete generation, and neuronal development (Supplementary table 4). Enriched pathways were signaling and generic transcription pathways (Supplementary table 4).

Interestingly, there was a significant overlap between the DE genes in the *DPLG1*-KO and *DPLG4*-KO ovaries (38 genes, $p < 8.261e-20$ using hypergeometric test) (Figure 5B). Further, the \log_2 foldchange of these overlapping DE genes in *DPLG1*-KO and *DPLG4*-KO ovaries were positively correlated ($r=0.7$, $p=8.6e-7$ using F-test; Figure 5C). In addition, *DPLG4* is downregulated (\log_2 foldchange -0.66) in *DPLG1*-KO ovaries and *DPLG1* is unaffected in *DPLG4*-KO ovaries. These overlapping genes are not enriched for any GO terms or pathways. Since we expected that regions nearby *DPLG1* and *DPLG4* would not introgress as readily into *w¹¹¹⁸* background, we excluded genes that did not show signs of introgression close to *DPLG1* and

DPLG4 in our GO and pathway analysis. There was no difference in the GO terms or pathways as a result of removal of these genes (data not shown). We also confirmed that the 38 overlapping DE genes between *DPLG1*-KO and *DPLG4*-KO ovaries showed signatures (same nucleotide variants as *w*¹¹¹⁸) of successful introgression into the *w*¹¹¹⁸ background ruling out that the 38 overlapping DE genes are due to lack of introgression of those regions into the KOs (Supplementary table 5). Together these data show that loss of *DPLG1* has drastic effect on ovarian gene expression in flies, and loss of *DPLG1* or *DPLG4* affect the ovarian gene expression of a set of genes in a similar manner.

In addition, we also decided to study the expression of TEs. Since *DPLGs* are TE derived and might still bind DNA, they could be involved in TE regulation. Our analyses revealed that several TE families were DE in the ovaries of both *DPLG1*-KO and *DPLG4*-KO flies, however the log2foldchanges of most DE TEs were rather modest (Log2foldchange range of DE TEs in *DPLG1*-KO ovaries: -6.8 to 3.2, Log2foldchange range of DE TEs in *DPLG4*-KO ovaries: -3.7 to 2.7) (Figure 6). Additionally, the direction of the DE of the TEs was variable. Many DE TEs were upregulated, but some were downregulated as well, with similar log2foldchange in both directions. Interestingly, most TEs that were upregulated were *Gypsy* elements and most TEs that were downregulated were telomeric elements in both KO ovaries. Although there was no enrichment of GO terms or pathways associated with TE control in either of the KOs (see above), there was an enrichment ($p < 0.011$; using hypergeometric test) for a few genes known to be in the piRNA pathways (89 manually curated by us; Supplementary table 6) and DE in *DPLG1* KO ovaries (Downregulated genes were all soma specific piRNA pathway genes except *BoYb*: *CG9821*, *CycT*, *wcy*, *omd*, *Droj2*, *IntS12*, *BoYb*. Upregulated genes were all germline and soma piRNA pathway

genes: *Gasz*, *piwi*, *Hen1*). An explanation to this observation could be a different representation of somatic and germline cells in the *DPLG1*-KO. To test this we analyzed the DE of germline-specific (*vasa*, *nanos*, *aub* and *AGO3*) and soma-specific (tj) genes and none of them were differentially expressed supporting that the overrepresentation of the piRNA genes is not due to different representation of ovarian somatic germline cells but rather due to potential effects on the piRNA pathway. Despite these observations, our RNA-Seq results provide some but not strong evidence that *DPLG1* and *DPLG4* are involved in regulating TE activities.

DPLG4 mutants show increased survival

We also tested if survival was affected in the *DPLG4* mutant flies compared to the control. In addition to the null mutant of *DPLG4* generated through CRISPR-Cas9 technology, we also tested the survival of an independent *DPLG4* mutant that contained a frameshift mutation. Analysis of both *P-element* excision mutant and CRISPR-Cas9 generated KO of *DPLG4* showed that loss of *DPLG4* results in a significant increase in the survival of flies compared to the control using log rank test (*DPLG4 P-element* excision male: $X^2 = 23.6$, $df = 1$, $p = 1.21e-06$; *DPLG4 P-element* excision female: $X^2 = 162$, $df = 1$, $p < 1e-16$; *DPLG4*-KO male: $X^2 = 31.1$, $df = 1$, $p = 2.44e-08$; *DPLG4*-KO female: $X^2 = 104$, $df = 1$, $p < 1e-16$) (Figure 7). The median survival time (MST) (the time at which the survival probability drops to 0.5) of the mutants increased by 7.6% in males and 15.8% females of *P-element* excision mutants, and the MST increased by 13% in males and 21.4% females of CRISPR KOs. The significant increase in the survival was observed in both males and females, although the increase in the survival was more substantial in the females of both the mutants compared to the males (Figure 7). One issue we faced in the protocol for *P-element*

excision mutant and its control was that during mating in the bottles, we observed food accumulation in the wings of the flies. Because of this reason we mated CRISPR generated mutants in multiple vials in small batches and the accumulation of the food in the wings was not observed for these flies. For the mutant and control of the *P-element* excision line we observed a spike in the death of flies in the first few days which we did not observe in the CRISPR mutants and there was almost no deaths in the first 20 days in experiments described in the literature (Buck et al. 2000; Linford et al. 2013; Wit, Sarup, et al. 2013; Galenza et al. 2016). We suspect that the food accumulation in the wings of the *P-element* excision controls and the mutants might be the reason for this initial spike of their death. Thus we discarded the data of the first 20 days in both the control and the KOs of the *P-element* excision lines.

Discussion

Domestication of *DPLGs* as regulatory proteins in *Drosophila*

Multiple sequence alignment of *DPLGs* with the ancestral *PIF* transposase revealed that they likely acquired disabling mutations in the catalytic domain but have retained the intact ancestral HTH putative DNA binding domain (Casola et al. 2007). Thus we hypothesized that *DPLGs* were domesticated as regulatory proteins. Interestingly, data from the ModENCODE consortium (The ModENCODE Consortium et al. 2010) revealed that genes that are in the coexpression cluster with *DPLGs* are enriched for transcription factor activities (Supplementary table 7) suggesting that *DPLGs* coexpress with transcription factors. Further, our protein

localization studies for *DPLG2-4* show that they localize with DNA in the ovarioles (Figure 3). These results provide supporting evidence that *DPLGs* are potential regulatory proteins.

Interestingly, evidence from two independent studies in *Arabidopsis* show that domesticated *PIF* transposases have been recruited as regulatory proteins to counteract epigenetic silencing. It has been shown that *ALP1*, a domesticated *PIF* transposase, physically interacts with and suppresses the activity of Polycomb Repressive complex 2 (*PRC2*), thereby opposing epigenetic silencing (Liang et al. 2015). *HDP1*, another protein domesticated from *PIF* transposase in *Arabidopsis*, interacts with histone acetyltransferase complexes and prevent DNA hypermethylation and epigenetic silencing (Duan et al. 2017). It is very intriguing that the two domesticated *PIF* transposase that have been functionally characterized up to date function to oppose the establishment of repressive chromatin. Several TEs have been described to have developed ways to evade host defense (Cui and Fedoroff 2002; Fu et al. 2013; Hosaka et al. 2017), thus it is possible that proteins encoded by *PIF* TEs might have also evolved to prevent their own sequences, which transposase normally bind to, from getting epigenetically repressed. Paradoxically, these selfish activities might have predisposed these transposases for cooption to serve a cellular function. The mechanism by which *DPLGs* and other *PIF* transposase-derived proteins modulate the establishment of repressive chromatin calls for future investigation.

DPLG1 is in head to head orientation with a gene called *piwi* which is only 402 bp upstream of it (Gramates et al. 2017). *Piwi* forms a riboprotein complex guided by piRNAs that is not only involved in development but also in silencing TEs at the DNA level through chromatin remodeling in the gonads of *D. melanogaster* (Klenov et al. 2011; Czech and Hannon 2016). Studies have shown that genes in head to head orientation often share common regulatory

regions (Kalitsis and Saffery 2009). Thus, we speculate that *DPLG1* and *piwi* may be under the control of some shared regulatory regions and may be functionally associated. Interestingly, according to ModENCODE data, *DPLG1* and *piwi* are in the same co-expression cluster (mE1_20_mRNA_expression_cluster_06) suggesting they have correlated expression profile across fly development (The ModENCODE Consortium et al. 2010; Gramates et al. 2017). The sharing of *piwi* regulatory regions might also have been the means by which *DPLG1* domestication was facilitated. Interestingly, there is an overrepresentation of DE piRNA pathway genes in *DPLG1*-KO ovaries supporting *DPLG1* could be involved in TE control as well.

Role of *DPLGs* in TE control

Data from RNA-Seq analyses of *DPLG1*-KO and *DPLG4*-KOs ovaries showed that several TEs are differentially expressed in these ovaries (Figure 6). Since the KOs were introgressed into the control background through six rounds of backcrossing, and although their genetic backgrounds are very similar, they are not identical. The signals of the DE TEs could be due to non introgressed loci and not due to actual derepression of TEs. Further, mutations in the genes known to be in TE control pathways show much higher derepression of TEs than observed in our results (Klenov et al. 2011; Handler et al. 2013; Huang et al. 2014; Wylie et al. 2016). There were, however, several interesting observations. For *DPLG4*-KO, there were a lot of *Gypsy* elements that were DE in ovaries in which most upregulated TEs were *Gypsy* elements (Figure 6). This observation is in line with a study in which out of a screen of over 7000 genes, *DPLG4* was one of the 368 genes that scored positive for upregulation of gypsy-lacZ reporter after KD in oocyte somatic cells. However further investigation with *DPLG4* was halted because only genes that

scored positive with two independent knockdowns were considered 'validated' and *DPLG4* had only one RNAi line available (Handler et al. 2013). *Gypsy* elements are mostly active in the follicle cells of the ovaries (Handler et al. 2013), and interestingly *DPLG4* shows strong signals of protein localization in the follicle cells as well (Figure 3). Several *Gypsy* elements are also DE in *DPLG1*-KO (Figure 6). Although, we do not observe any enrichment terms associated with control of TE activities, there is an overrepresentation of DE piRNA pathway genes in *DPLG1*-KO ovaries when we use our curated piRNA pathway gene set. In particular, some specific piRNA pathway genes are going down analogous to the downregulation of *DPLG4* in this mutant. So, additional analyses are needed to disentangle any role of *DPLGs* in TE control, including *DPLG2* and *DPLG3*. Double mutants should also be examined for TE derepression as some *DPLGs* might have overlapping functions.

Do domesticated transposases promote domestication of related transposases?

There are seven *DPLGs* across *Drosophila* and most of them predate *Drosophila* diversification. Interestingly, *DPLGs* were not derived from a single domestication event but rather from multiple independent domestications of *PIF* transposases. Phylogenetic analysis of *DPLGs* shows them clustering with *PIF* transposases of different *PIF* elements before clustering with each other and provides evidence for at least three independent domestication events that gave rise to the seven *DPLGs* (Casola et al. 2007), although the complete order of the domestication events of all *DPLGs* is still not clear. *In situ* hybridization studies show that all *DPLGs* have strikingly similar patterns of transcript localization during embryogenesis (Figure 1). Further, *DPLGs* also show overlapping transcript localization in the testis and the ovaries (Figure

2). Transcript localization results in the ovaries are further supported by protein localization studies that show *DPLG2-4* have overlap in protein localization during oogenesis (Figure 3). Together, these data provides support for the idea that *DPLGs* could have overlapping functions. RNA-Seq data revealed that there was a significant overlap of DE genes between *DPLG1-KO* and *DPLG4-KO* ovaries. Interestingly, these overlapping DE genes showed positive correlation in the log2foldchange and *DPLG4* was downregulated in *DPLG1-KO* ovaries, although *DPLG1* was not DE in *DPLG4-KO* ovaries (Figure 6C). So, *DPLG1* might upregulate *DPLG4* in addition to other regulatory roles. Loss of *DPLG1* or *DPLG4* affects ovarian gene expression of some genes in similar ways providing additional support that they may be working in similar pathways in the ovaries. However, the effects for female fertility for both mutants are not consistent (Figure 4I and 4J). Further, while independent mutations in either *DPLG1* or *DPLG4* have reduced viability in the embryos, double mutant of *DPLG1* and *DPLG4* have no effect in the embryonic viability. This observation suggests that *DPLG1* and *DPLG4* might be genetic interactors of each other in complex ways that might depend on the cell type and trait under study. Taken together, we have various lines of evidence that suggests *DPLGs* have functional relatedness, and additional support that *DPLG1* and *DPLG4* might work in some of the same pathways in ovaries and potentially interact. Evidence of functional relatedness between two independently domesticated transposases provides support to our stepping stone model where domestication of a transposase promotes domestication of related transposases.

It would be interesting to explore these initial inferences in more detail and understand if *DPLG2* and *DPLG3* have functions related to *DPLG1* and *DPLG4* given their overlap in transcript localization with *DPLG1* and *DPLG4*, and protein localization with *DPLG4*. Although *DPLG2* and

DPLG3 are not DE in the both *DPLG1*-KO and *DPLG4*-KO ovaries, they could be acting upstream of *DPLG1* and *DPLG4* and might be a reason we do not observe DE of *DPLG2* and *DPLG3* in the KOs that we analyzed. Further, all *DPLGs* have almost exactly the same pattern of transcript localization during embryogenesis. It would be interesting to perform RNA-Seq on various stages of embryogenesis in the KO of *DPLGs* to explore if they are performing similar functions during embryogenesis as well.

DPLG4 may be involved in the process of ageing and have neuronal function

Loss of *DPLG4* resulted in increased survival of both males and females, where the effect was more drastic in the females (Figure 7). A phenotype that has been repeatedly associated with long living flies is reduced viability early in their development (Buck et al. 2000). This is in line with what we observed with loss of *DPLG4* where the viability during embryogenesis and post embryogenesis stages is reduced compared to the control (Figure 4E and 4F). Additionally, our results showed that younger *DPLG4*-KO females do not show difference in fertility compared to the control, however as the females age the *DPLG4*-KO females have a significantly higher fertility compared to control females (Figure 4K). This pattern of no effect in fertility in younger females and significantly higher fertility in older females of long living flies has been reported in the past (Partridge and Fowler 1991) and the authors speculated that the long lived females may possess superior soma that could result in their increased fertility late in life compared to the controls. Taken together, loss of *DPLG4* is associated with increased survival in both males and females and inhibition of drastic decline in fertility of older females may be a result of reduced deterioration of the female ovaries.

Reduction of heterochromatin marks has been shown to be associated with increase in age suggesting that there is a tight relation between age of an organism and chromatin state (Wood et al. 2010; Wood and Helfand 2013). In fact, inhibition in the expression of a component of heterochromatin protein complex, HP1 (Heterochromatin protein 1), shortens lifespan of flies while its overexpression results in extension of lifespan (Larson et al. 2012). Since *DPLG4* negatively affects survival, it is possible that *DPLG4* is associated with prevention of heterochromatin formation. This result is in line with previous observations that *PIF* transposase are domesticated as regulatory proteins that prevent heterochromatin (Liang et al. 2015; Duan et al. 2017). Although this is not completely consistent with the postulated potential role of *DPLGs* in TE repression (see above), regulatory proteins might have different roles in different cell types or chromatin regions through interactions with different proteins.

We do not have convincing evidence that *DPLG4* is involved in preventing heterochromatin in young flies (preliminary studies of position effect variegation did not support this effect; data not shown), however *DPLG4* could be associated with accelerating the decrease of heterochromatin marks with increasing age and thus contributing negatively towards the survival in flies. Loss of heterochromatin marks with increase in age have been also associated with increase in TE activities in older flies (Wood and Helfand 2013; Orr 2016). If *DPLG4* is associated with preventing heterochromatin marks in older flies, this suggests that *DPLG4* could be promoting TE activities in older flies. Our results show that most telomeric elements are downregulated in both *DPLG1*-KO and *DPLG4*-KO ovaries (Figure 6). These KO flies were kept as homozygotes for close to two years before their RNA was extracted. During this time, these KO flies could have accumulated relatively less insertions compared to the control resulting in the

signal of their downregulation in the KO ovaries. This further supports that *DPLGs* may be involved in preventing heterochromatin and that opens doors to further investigation.

Given that *DPLG4*-KO flies have a higher survival and increased fertility, why would *DPLG4* be under purifying selection? One reason could be the existence of a trade-off. *DPLG4*-KO might be needed early in development but have a cost later in life. Additionally it has been shown that long living flies may do better in the lab conditions, however in the natural environment, these flies are not as efficient in finding food and this may be a surplus factor that reduces the fitness of *DPLG4*-KO flies in nature (Wit, Kristensen, et al. 2013). Further, the authors observe behavior defects in long lived flies which might have contributed towards their reduced fitness (Wit, Kristensen, et al. 2013). We did not observe any behavioral defects in the larva or the adults of *DPLG4* mutants, however these experiments were done in controlled lab setting and may not hold true in the natural environments. Despite lack of phenotypic data, other lines of evidence suggest that *DPLG4* might have neuronal function and may affect fly behavior. *DPLG4* is relatively highly expressed in adult central nervous system (Lovering et al. 2018) and *in situ* hybridization results reveal that *DPLG4* transcripts localize in the ventral nerve chord of the embryos (Figure 1). Additionally, top 5% genes that have co-evolved with *DPLG4* are enriched for GO terms relating to neuronal functions (table 1). Further, overexpression of *DPLG4* during the larval stage mildly rescues neuromuscular junction overgrowth phenotype caused by dominant negative of NSF2 (Laviolette et al. 2005). Thus *DPLG4* could be important during neurogenesis and may affect fly behavior resulting in their reduced fitness in nature. Taken together, our results show that loss of *DPLG4* affects survival in flies and inhibits drastic decline of fertility in females in lab conditions, thus supporting that *DPLG4* is likely involved in the process of ageing.

References

- Andrew S. 2010. A quality control tool for high throughput sequence data. Babraham Bioinforma. [Internet]. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, et al. 2014. Diversity and dynamics of the *Drosophila* transcriptome. *Nature* 512:393–399.
- Buck S, Vettraino J, Force AG, Arking R. 2000. Extended longevity in *Drosophila* is consistently associated with a decrease in developmental viability. *Journals Gerontol. - Ser. A Biol. Sci. Med. Sci.* 55.
- Cam HP, Noma KI, Ebina H, Levin HL, Grewal SIS. 2008. Host genome surveillance for retrotransposons by transposon-derived proteins. *Nature* 451:431–436.
- Casola C, Hucks D, Feschotte C. 2008. Convergent domestication of pogo-like transposases into centromere-binding proteins in fission yeast and mammals. *Mol. Biol. Evol.* 25:29–41.
- Casola C, Lawing AM, Betrán E, Feschotte C. 2007. PIF-like transposons are common in *Drosophila* and have been repeatedly domesticated to generate new host genes. *Mol. Biol. Evol.* 24:1872–1888.
- Chan C-C, Scoggin S, Hiesinger PR, Buszczak M. 2012. Combining recombineering and ends-out homologous recombination to systematically characterize *Drosophila* gene families: Rab GTPases as a case study. *Commun. Integr. Biol.* [Internet] 5:179–183. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3376058&tool=pmcentrez&rendertype=abstract>

- Chandler CH, Chari S, Dworkin I. 2013. Does your gene need a background check? How genetic background impacts the analysis of mutations, genes, and evolution. *Trends Genet.* 29:358–366.
- Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* 10:691–703.
- Cui H, Fedoroff N V. 2002. Inducible DNA Demethylation Mediated by the Maize Suppressor-mutator Transposon-Encoded TnpA Protein. *Plant Cell [Internet]* 14:2883–2899. Available from: <http://www.plantcell.org/lookup/doi/10.1105/tpc.006163>
- Curcio MJ, Derbyshire KM. 2003. The outs and ins of transposition: From MU to kangaroo. *Nat. Rev. Mol. Cell Biol.* 4:865–877.
- Czech B, Hannon GJ. 2016. One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. *Trends Biochem. Sci.* 41:324–337.
- Dietzl G, Chen D, Schnorrer F, Su KC, Barinova Y, Fellner M, Gasser B, Kinsey K, Oppel S, Scheiblauer S, et al. 2007. A genome-wide transgenic RNAi library for conditional gene inactivation in *Drosophila*. *Nature* 448:151–156.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21.
- Duan CG, Wang X, Xie S, Pan L, Miki D, Tang K, Hsu CC, Lei M, Zhong Y, Hou YJ, et al. 2017. A pair of transposon-derived proteins function in a histone acetyltransferase complex for active DNA demethylation. *Cell Res.* 27:226–240.

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

Feschotte C, Pritham EJ. 2007. DNA Transposons and the Evolution of Eukaryotic Genomes. *Annu. Rev. Genet.* [Internet] 41:331–368. Available from: <http://www.annualreviews.org/doi/10.1146/annurev.genet.40.110405.090448>

Fu Y, Kawabe A, Etcheverry M, Ito T, Toyoda A, Fujiyama A, Colot V, Tarutani Y, Kakutani T. 2013. Mobilization of a plant transposon by expression of the transposon-encoded anti-silencing factor. *EMBO J.* 32:2407–2417.

Galenza A, Hutchinson J, Campbell SD, Hazes B, Foley E. 2016. Glucose modulates *Drosophila* longevity and immunity independent of the microbiota. *Biol. Open* [Internet] 5:165–173. Available from: <http://bio.biologists.org/lookup/doi/10.1242/bio.015016>

Gramates LS, Marygold SJ, Dos Santos G, Urbano JM, Antonazzo G, Matthews BB, Rey AJ, Tabone CJ, Crosby MA, Emmert DB, et al. 2017. FlyBase at 25: Looking to the future. *Nucleic Acids Res.* 45:D663–D671.

Grzebelus D, Lasota S, Gambin T, Kucherov G, Gambin A. 2007. Diversity and structure of PIF/Harbinger-like elements in the genome of *Medicago truncatula*. *BMC Genomics* [Internet] 8:409. Available from: <http://bmcbgenomics.biomedcentral.com/articles/10.1186/1471-2164-8-409>

Handler D, Meixner K, Pizka M, Lauss K, Schmied C, Gruber FS, Brennecke J. 2013. The genetic makeup of the *drosophila* piRNA pathway. *Mol. Cell* 50:762–777.

- Hosaka A, Saito R, Takashima K, Sasaki T, Fu Y, Kawabe A, Ito T, Toyoda A, Fujiyama A, Tarutani Y, et al. 2017. Evolution of sequence-specific anti-silencing systems in Arabidopsis. *Nat. Commun.* 8.
- Huang H, Li Y, Szulwach KE, Zhang G, Jin P, Chen D. 2014. AGO3 Slicer activity regulates mitochondria-nuage localization of Armitage and piRNA amplification. *J. Cell Biol.* 206:217–230.
- Jangam D, Feschotte C, Betrán E. 2017. Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. *Trends Genet.* 33:817–831.
- Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, Wessler SR. 2003. An active DNA transposon family in rice. *Nature* 421:163–167.
- Jin Y, Tam OH, Paniagua E, Hammell M. 2015. Tetrascripts: A package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* 31:3593–3599.
- Johansen P, Cam HP. 2015. Suppression of meiotic recombination by CENP-B homologs in *Schizosaccharomyces pombe*. *Genetics* 201:897–904.
- Joly-Lopez Z, Forczek E, Hoen DR, Juretic N, Bureau TE. 2012. A Gene Family Derived from Transposable Elements during Early Angiosperm Evolution Has Reproductive Fitness Benefits in *Arabidopsis thaliana*. *PLoS Genet.* 8.
- Kalitsis P, Saffery R. 2009. Inherent promoter bidirectionality facilitates maintenance of sequence integrity and transcription of parasitic DNA in mammalian genomes. *BMC Genomics* 10:498.

- Kapitonov V V, Jurka J. 2004. Harbinger transposons and an ancient HARBI1 gene derived from a transposase. *DNA Cell Biol.* 23:311–324.
- Kidwell MG, Lisch DR. 2000. Transposable elements and host genome evolution. *Trends Ecol. Evol.* 15:95–99.
- Klenov MS, Sokolova OA, Yakushev EY, Stolyarenko AD, Mikhaleva EA, Lavrov SA, Gvozdev VA. 2011. Separation of stem cell maintenance and transposon silencing functions of Piwi protein. *Proc. Natl. Acad. Sci.* [Internet] 108:18760–18765. Available from: <http://www.pnas.org/cgi/doi/10.1073/pnas.1106676108>
- Kondo S, Ueda R. 2013. Highly Improved gene targeting by germline-specific Cas9 expression in *Drosophila*. *Genetics* 195:715–721.
- de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over Two-Thirds of the human genome. *PLoS Genet.* 7.
- 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.
- Larson K, Yan SJ, Tsurumi A, Liu J, Zhou J, Gaur K, Guo D, Eickbush TH, Li WX. 2012. Heterochromatin formation promotes longevity and represses ribosomal RNA synthesis. *PLoS Genet.* 8.
- Laviolette MJ, Nunes P, Peyre JB, Aigaki T, Stewart BA. 2005. A genetic screen for suppressors of *drosophila* NSF2 neuromuscular junction overgrowth. *Genetics* 170:779–792.

- Liang SC, Hartwig B, Perera P, Mora-García S, de Leau E, Thornton H, de Alves FL, Rapsilber J, Yang S, James GV, et al. 2015. Kicking against the PRCs – A Domesticated Transposase Antagonises Silencing Mediated by Polycomb Group Proteins and Is an Accessory Component of Polycomb Repressive Complex 2. *PLoS Genet.* 11.
- Linford NJ, Bilgir C, Ro J, Pletcher SD. 2013. Measurement of Lifespan in *Drosophila melanogaster*. *J. Vis. Exp.* [Internet]. Available from: <http://www.jove.com/video/50068/measurement-of-lifespan-in-drosophila-melanogaster>
- Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15.
- Lovering RC, Roncaglia P, Howe DG, Laulederkind SJF, Khodiyar VK, Berardini TZ, Tweedie S, Foulger RE, Osumi-Sutherland D, Campbell NH, et al. 2018. Improving Interpretation of Cardiac phenotypes and enhancing discovery with expanded knowledge in the gene ontology. *Circ. Cardiovasc. Genet.* 11.
- Lu J, Clark AG. 2010. Population dynamics of PIWI-interacting RNAs (piRNAs) and their targets in *Drosophila*. *Genome Res.* 20:212–227.
- Lyne R, Smith R, Rutherford K, Wakeling M, Varley A, Guillier F, Janssens H, Ji W, McLaren P, North P, et al. 2007. FlyMine: An integrated database for *Drosophila* and *Anopheles* genomics. *Genome Biol.* 8.
- Mateo L, González J. 2014. Pogo-Like transposases have been repeatedly domesticated into CENP-B-Related Proteins. *Genome Biol. Evol.* 6:2008–2016.

- Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* (80-.). 346:763–767.
- Morris CA, Benson E, White-Cooper H. 2009. Determination of gene expression patterns using in situ hybridization to *Drosophila* testes. *Nat. Protoc.* 4:1807–1819.
- Orr WC. 2016. Tightening the connection between transposable element mobilization and aging. *Proc. Natl. Acad. Sci.* [Internet] 113:11069–11070. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27663733>
- Partridge L, Fowler K. 1991. Direct and correlated responses to selection on age at reproduction in *Drosophila melanogaster*. *Evolution* (N. Y). 46:76–91.
- Pavelitz T, Gray LT, Padilla SL, Bailey AD, Weiner AM. 2013. PGBD5: A neural-specific intron-containing piggyBac transposase domesticated over 500 million years ago and conserved from cephalochordates to humans. *Mob. DNA* 4.
- Schmittgen TD, Livak KJ. 2008. Analyzing real-time PCR data by the comparative CT method. *Nat. Protoc.* [Internet] 3:1101–1108. Available from: <http://www.nature.com/doi/10.1038/nprot.2008.73>
- Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al. 2009. The B73 maize genome: Complexity, diversity, and dynamics. *Science* (80-.). 326:1112–1115.
- Sinzelle L, Izsvák Z, Ivics Z. 2009. Molecular domestication of transposable elements: From

detrimental parasites to useful host genes. *Cell. Mol. Life Sci.* 66:1073–1093.

Sinzelle L, Kapitonov V V, Grzela DP, Jursch T, Jurka J, Izsvák Z, Ivics Z. 2008. Transposition of a reconstructed Harbinger element in human cells and functional homology with two transposon-derived cellular genes. *Proc. Natl. Acad. Sci. U. S. A.* [Internet] 105:4715–4720. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18339812> <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2290759>

Slawson JB, Kuklin EA, Ejima A, Mukherjee K, Ostrovsky L, Griffith LC. 2011. Central regulation of locomotor behavior of *Drosophila melanogaster* depends on a CASK isoform containing CaMK-Like and L27 domains. *Genetics* 187:171–184.

Tautz D, Pfeifle C. 1989. A non-radioactive in situ hybridization method for the localization of specific RNAs in *Drosophila* embryos reveals translational control of the segmentation gene hunchback. *Chromosoma* [Internet] 98:81–85. Available from: <http://link.springer.com/10.1007/BF00291041>

The ModENCODE Consortium, Roy S, Ernst J, Kharchenko P V., Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, et al. 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* (80-.). 330:1787–1797.

Therneau T, Lumley T original S->R port. 2011. survival: Survival analysis including penalised likelihood. R Packag. version 2.36-5.

Vogt A, Goldman AD, Mochizuki K, Landweber LF. 2013. Transposon Domestication versus

Mutualism in Ciliate Genome Rearrangements. *PLoS Genet.* 9.

Wiegmann BM, Trautwein MD, Winkler IS, Barr NB, Kim J-W, Lambkin C, Bertone MA, Cassel BK, Bayless KM, Heimberg AM, et al. 2011. Episodic radiations in the fly tree of life. *Proc. Natl. Acad. Sci.* [Internet] 108:5690–5695. Available from: <http://www.pnas.org/cgi/doi/10.1073/pnas.1012675108>

Wit J, Kristensen TN, Sarup P, Frydenberg J, Loeschcke V. 2013. Laboratory selection for increased longevity in *Drosophila melanogaster* reduces field performance. *Exp. Gerontol.* 48:1189–1195.

Wit J, Sarup P, Lupsa N, Malte H, Frydenberg J, Loeschcke V. 2013. Longevity for free? Increased reproduction with limited trade-offs in *Drosophila melanogaster* selected for increased life span. *Exp. Gerontol.* 48:349–357.

Wood JG, Helfand SL. 2013. Chromatin structure and transposable elements in organismal aging. *Front. Genet.* 4.

Wood JG, Hillenmeyer S, Lawrence C, Chang C, Hosier S, Lightfoot W, Mukherjee E, Jiang N, Schorl C, Brodsky AS, et al. 2010. Chromatin remodeling in the aging genome of *Drosophila*. *Aging Cell* 9:971–978.

Wylie A, Jones AE, D’Brot A, Lu WJ, Kurtz P, Moran J V., Rakheja D, Chen KS, Hammer RE, Comerford SA, et al. 2016. p53 genes function to restrain mobile elements. *Genes Dev.* 30:64–77.

Zhang X, Feschotte C, Zhang Q, Jiang N, Eggleston WB, Wessler SR. 2001. P instability factor: An

active maize transposon system associated with the amplification of Tourist-like MITEs and a new superfamily of transposases. *Proc. Natl. Acad. Sci.* [Internet] 98:12572–12577. Available from: <http://www.pnas.org/cgi/doi/10.1073/pnas.211442198>

Zhang X, Jiang N, Feschotte C, Wessler SR. 2004. PIF- and Pong-Like Transposable Elements: Distribution, Evolution and Relationship with Tourist-Like Miniature Inverted-Repeat Transposable Elements. *Genetics* 166:971–986.

Figure legends:

Figure 1. RNA *in situ* hybridization showing transcript localization of *DPLG1-4* during embryogenesis where all *DPLGs* show very similar pattern. At **stage 1-3** (the first two hours of embryogenesis) all *DPLGs* localize broadly throughout the embryo. At **stage 4-5** *DPLG2-4* start to localize towards the posterior part of the embryo where the pole cells are found, posterior localization of *DPLG1* was not detected. At **stage 6** transcripts of all *DPLGs* start to localize in the ventral nerve chord (VNC) and continue to localize in the VNC in the subsequent stages. At **stages 8-11** signals of transcript localization for all *DPLGs* are also observed in the midgut.

Figure 2. RNA *in situ* hybridization showing transcript localization of *DPLG1-4* in the ovaries (A-D) and testis (E-H). *DPLG1-3* (A-C) show strong transcript localization signal in the nurse cells, while *DPLG4* (D) shows broad transcript localization throughout the ovariole except in the anterior part of the germarium. In the testis, *DPLG1* (E) mostly localizes in the primary spermatocytes, although there are some signals from the stem cells and the mitotic cells as well. *DPLG2* and *DPLG3* (F-G) are detected only in the late primary spermatocyte stage. *DPLG4* (H) transcripts are detected in the stem cells and in all subsequent stages of spermatogenesis till the round spermatids.

Figure 3. Protein localization of *DPLG2-4* tagged with HA in the ovaries. **(A-D)** Protein localization of *DPLG2-HA*. *DPLG2-HA* (B-C) localizes in the anterior part of the germarium where the stem cells and the early differentiating cells reside. In the later stages of oogenesis, *DPLG2-HA* (D) localizes in the nucleus of the germline cells. **(E-G)** Protein localization of *DPLG3-HA*. *DPLG3-HA*

(F) localizes broadly throughout the germarium. In the later stages of oogenesis, DPLG3-HA (G) localizes in the nucleus of the germline and the somatic cells. In addition, DPLG3-HA also localizes in the oocyte nucleus (shown by arrow). **(H-J)** Protein localization of DPLG4-HA. DPLG4-HA (I) does not show much localization signal from the anterior part of the germarium and shows some signals from the posterior part of the germarium. In the later stages of oogenesis, DPLG4-HA (J) localizes in the nucleus of the germline and the somatic cells. **(K)** Negative control.

Figure 4. CRISPR-Cas9 KO of *DPLG1* and *DPLG4* and their viability and fertility effects: **(A-B)** RNA-Seq data showing absence of complete transcripts of *DPLG1* and *DPLG4* in the respective KOs. Inverted red triangles indicate the guide RNA targets in the genome for the *DPLG1* and *DPLG4*. The coding sequence (CDS) of *DPLG1* is completely removed and there is no transcript detected (A). One of the guide targets for *DPLG4* was designed in the CDS and small amount of transcripts were detected for the region that was not removed from the genome (B). There are no transcripts detected from the region that was removed from the genome. **(C-H)** *DPLG1* and *DPLG4-KO* effects on viability. *DPLG1-KO* shows a decrease in viability during embryogenesis (C) and an increase in viability in the post embryogenesis stages (D). *DPLG4-KO* shows reduction in viability in both embryogenesis and post embryogenesis stages (E-F). *DPLG4-KD* flies also show reduction in viability compared to the controls supporting the KO results (G). Quantitative RT-PCR results showing relative mean expression of *DPLG4* is significantly reduced in the KD compared to the control (H). **(I-J)** *DPLG1* and *DPLG4-KO* effects on fertility. *DPLG1-KO* males and females show increased fertility (I). *DPLG4-KO* flies do not exhibit any fertility effects in both males and females (J). **(K)** *DPLG4-KO* female fertility at different ages compared to the control. Younger *DPLG4-KO*

females do not have significant difference in fertility, however, *DPLG4-KO* older females show increased fertility compared to the control. **(L-N)** Viability and fertility studied in double mutants of *DPLG1* and *DPLG4*. The double mutant does not show any difference in embryonic viability (L) but exhibit increased viability in the post embryonic stages (M) compared to the controls. Double mutant females do not show any difference in fertility compared to the control (N). Error bars indicate standard error and “star” indicates significantly different than the control ($p < 0.05$, t-test).

Figure 6 (A) Number of upregulated and downregulated genes in *DPLG4-KO* and *DPLG1-KO* ovaries. **(B)** Venn diagram showing overlap of DE genes in *DPLG1-KO* and *DPLG4-KO* ovaries. **(C)** Comparison of Log2foldchange of overlapping DE genes between *DPLG1-KO* and *DPLG4-KO* ovaries. Log2foldchange of overlapping DE genes in *DPLG1-KO* and *DPLG4-KO* ovaries shows positive correlation ($r = 0.7$, $p < 0.001$). **(D)** Venn diagram showing overlap of enriched pathways and GO terms between *DPLG1-KO* and *DPLG4-KO* ovaries.

Figure 7. Log2foldchange of DE TEs in *DPLG1-KO* and *DPLG4-KO* ovaries. Both KOs show some DE TEs, however, most log2foldchange are not very high.

Figure 8. Survival analyses of *DPLG4* mutants. Comparison of survival of CRIPSR generated *DPLG4* null mutant (A-B) and *P-element* excision generated *DPLG4* frameshift mutant (C-D). Both mutants show significant increase in survival of both males and females compared to their

respective controls using log rank test (*DPLG4 P-element* excision male: $X^2 = 23.6$, $df = 1$, $p = 1.21e-06$, MST of control = 53, MST of mutant = 68; *DPLG4 P-element* excision female: $X^2 = 162$, $df = 1$, $p < 1e-16$, MST control = 57, MST mutant = 66; *DPLG4-KO* male: $X^2 = 31.1$, $df = 1$, $p = 2.44e-08$, MST $w^{1118} = 54$, MST KO = 61; *DPLG4-KO* female: $X^2 = 104$, $df = 1$, $p < 1e-16$, MST $w^{1118} = 56$, MST KO = 68).

Tables and Figures

Table 1. GO term enrichment for top 5% genes that show signatures of co-evolution with *DPLG4*

GO term	P value
plasma membrane bounded cell projection organization	0.005
cell projection organization	0.007
neuron differentiation	0.015
generation of neurons	0.019
axonogenesis	0.028
cell morphogenesis involved in neuron differentiation	0.03
neuron development	0.045
cell morphogenesis involved in differentiation	0.047

Figure 1

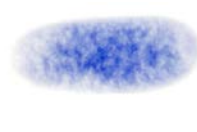




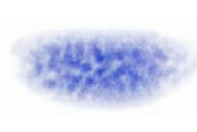

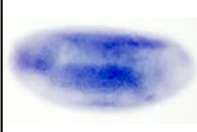
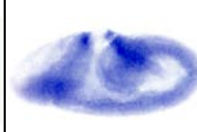

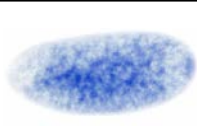

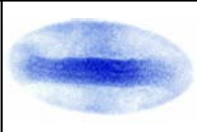


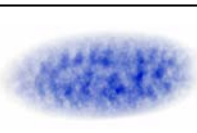
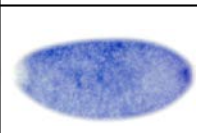
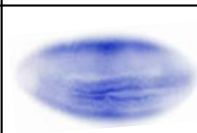
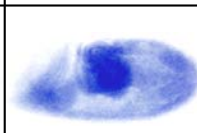
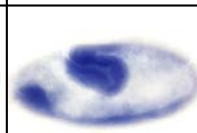
<i>DPLG1</i>					
<i>DPLG2</i>					
<i>DPLG3</i>					
<i>DPLG4</i>					
Stage	1 - 3	4 - 5	6 - 7	8 - 9	9 - 11

Figure 2

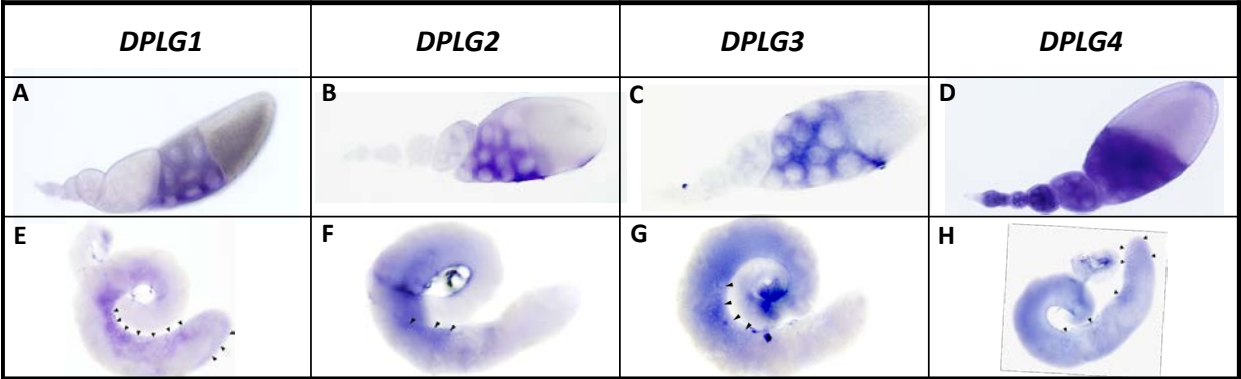


Figure 3

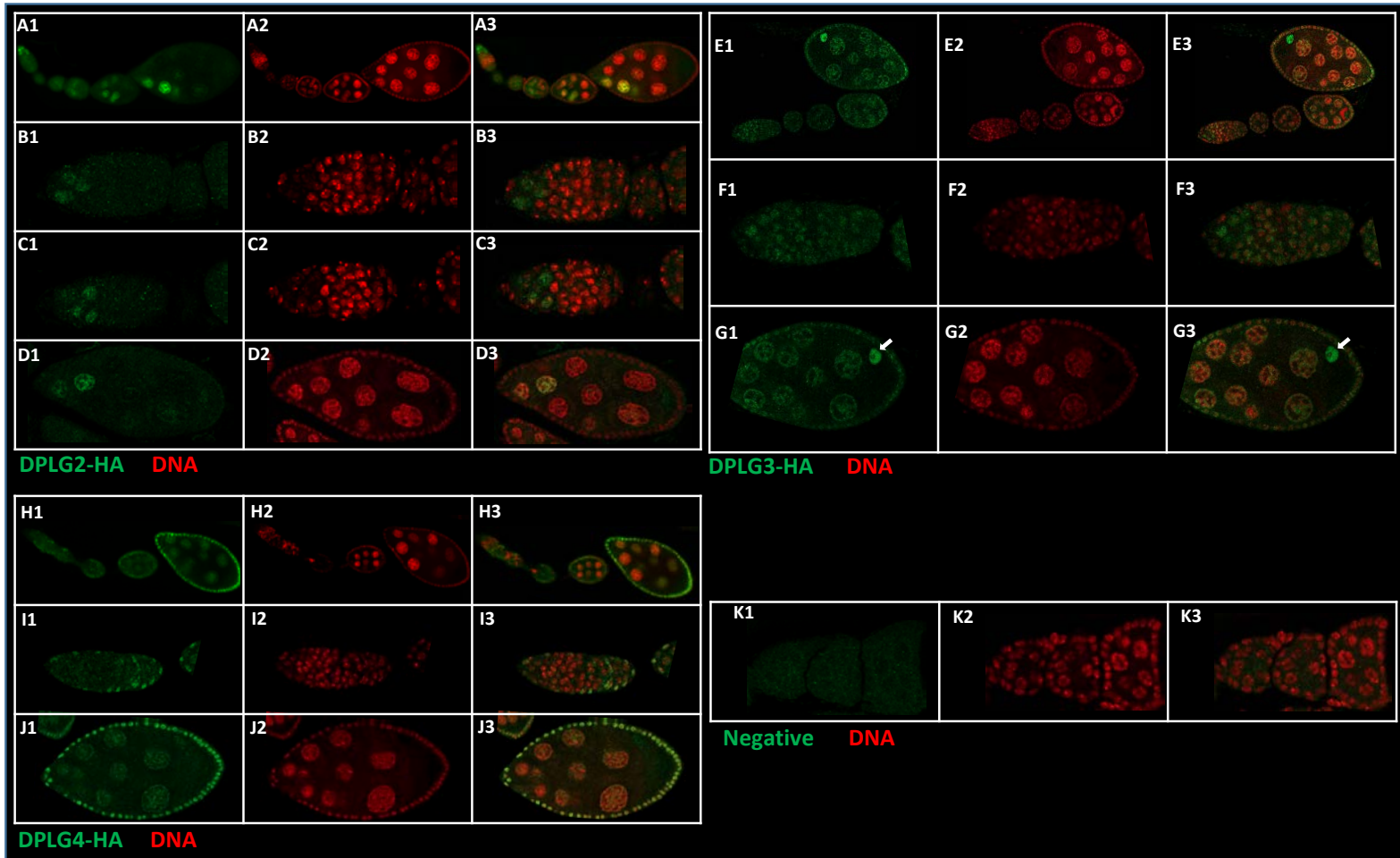


Figure 4

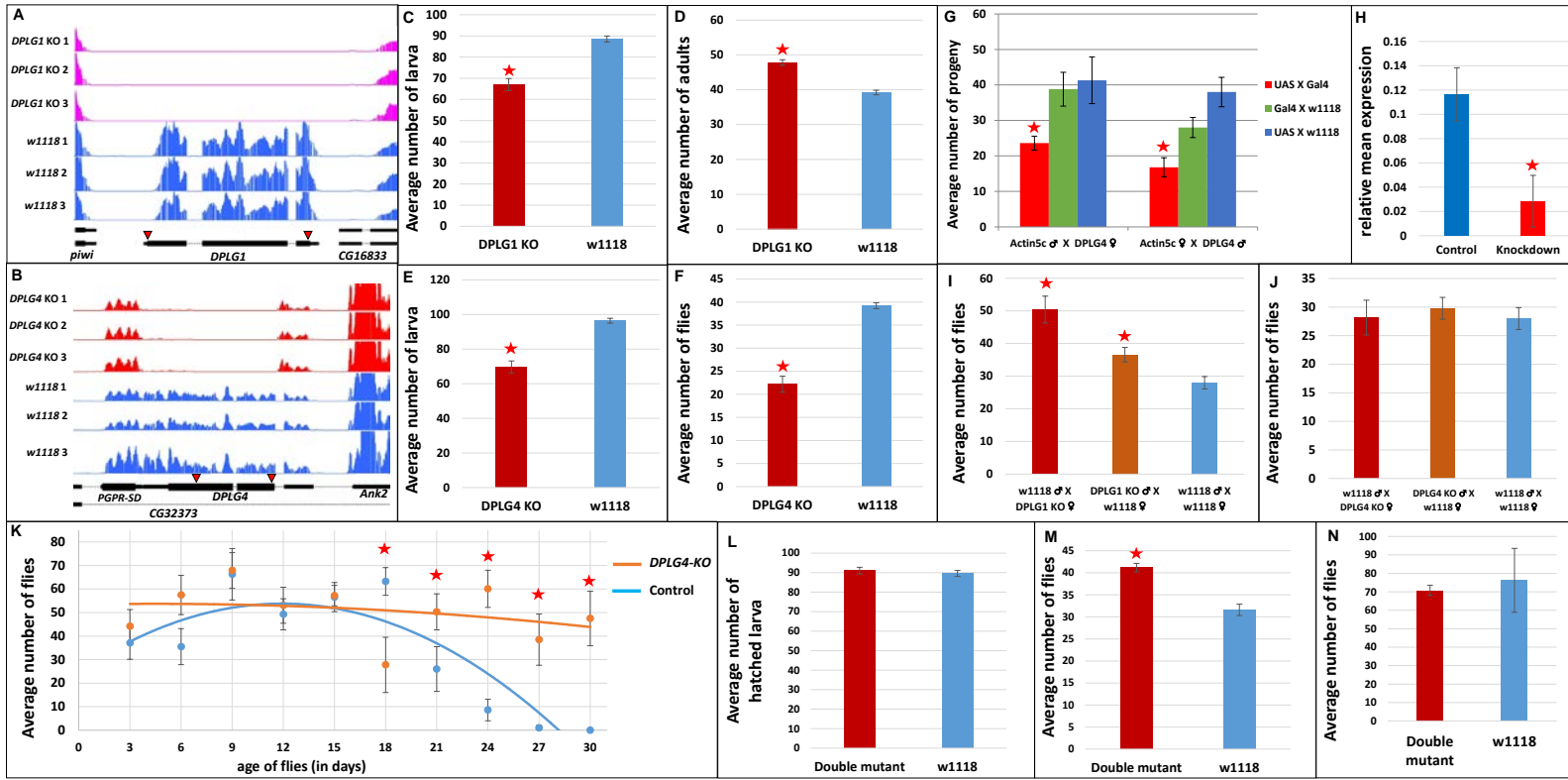


Figure 5

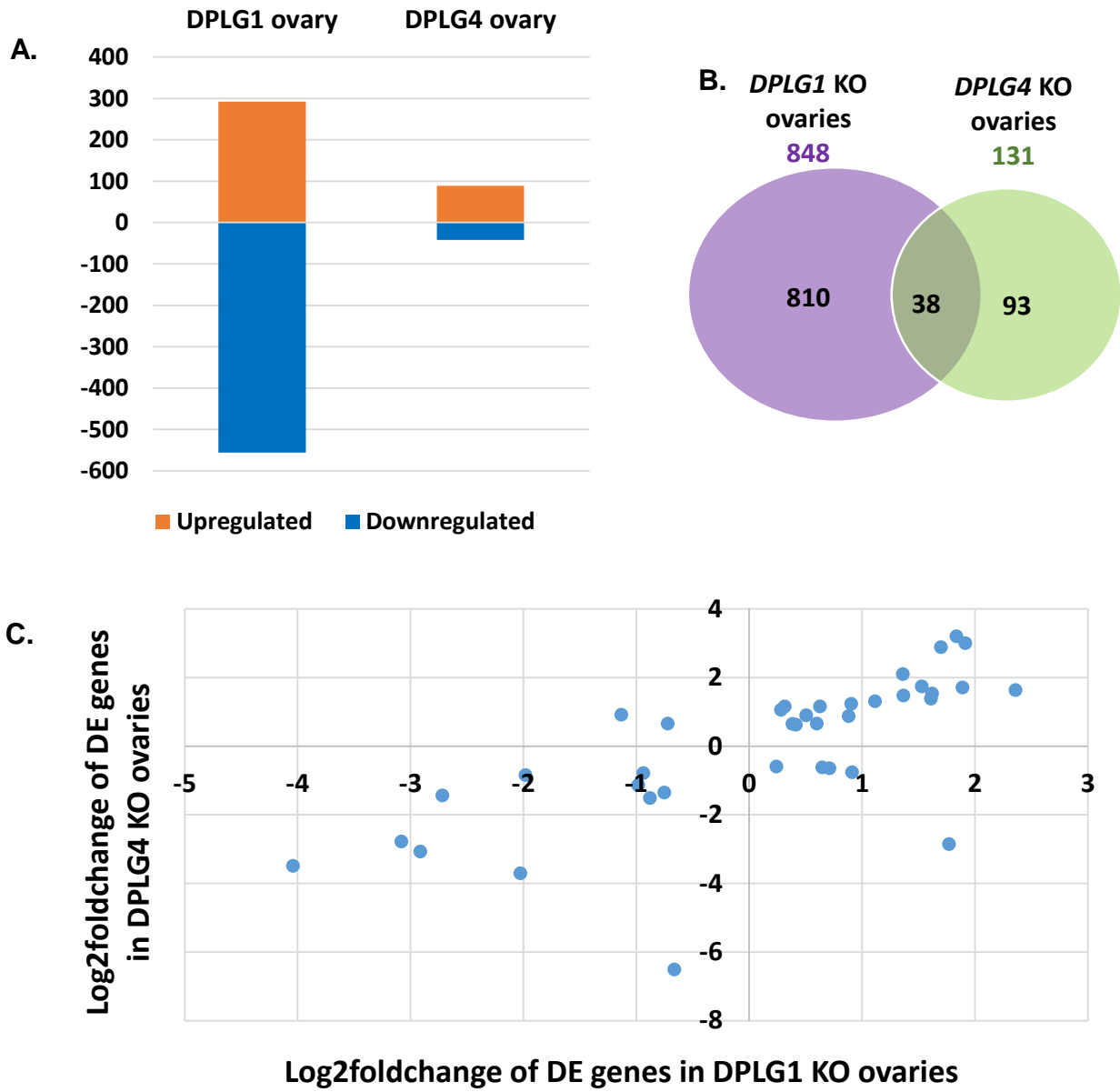


Figure 6

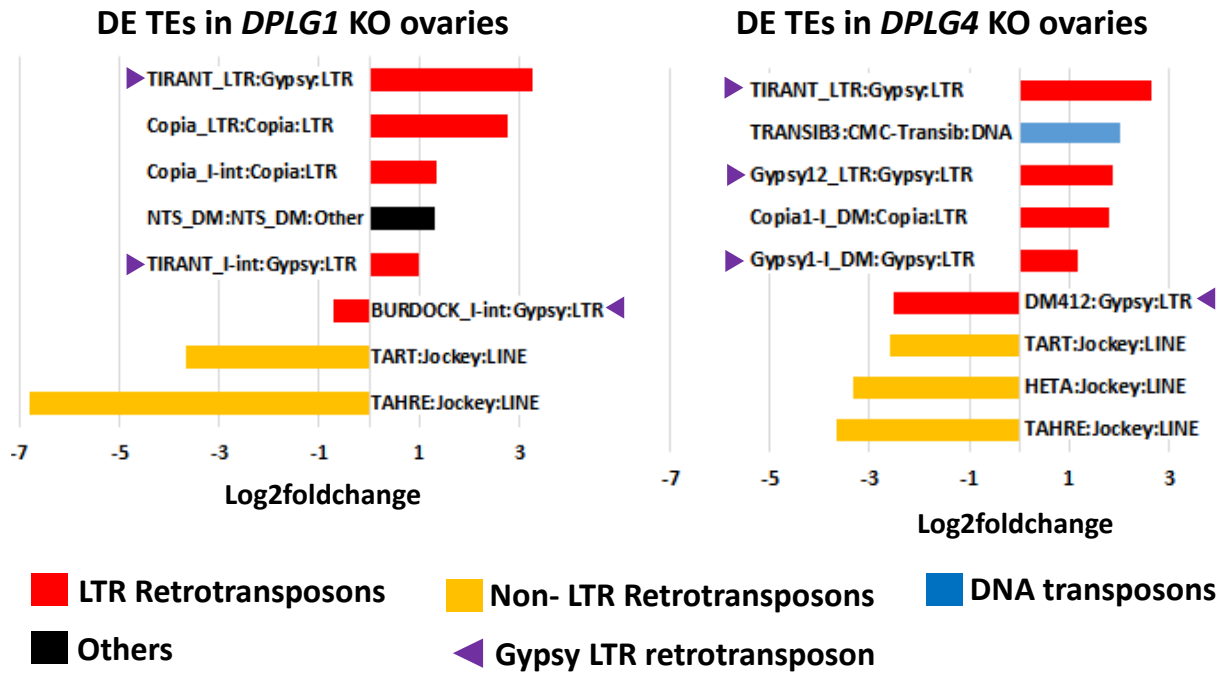
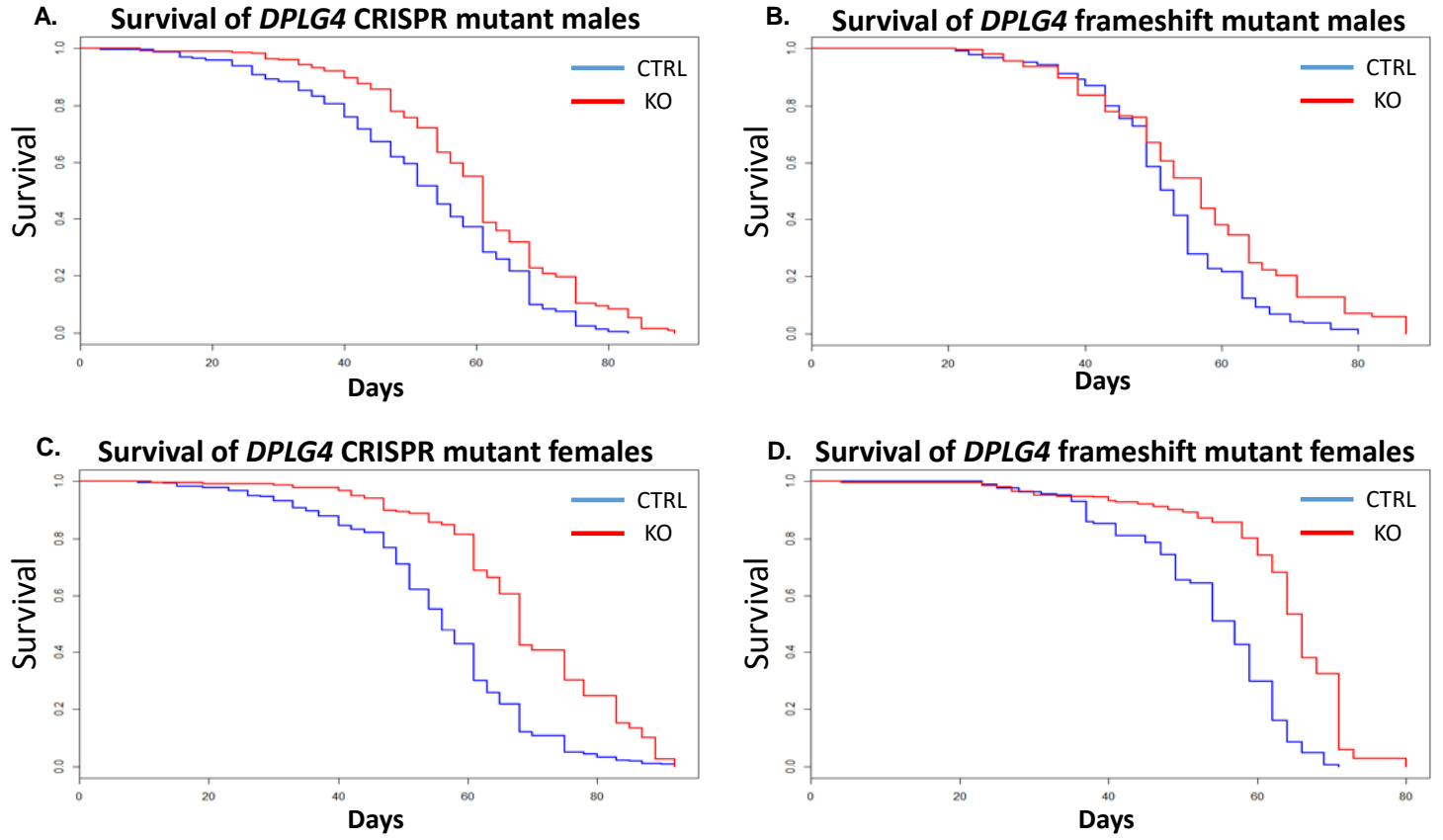


Figure 7



Supplementary Tables

Supplementary Table 1. Primers and gene blocks used for knocking out *DPLG1* and *DPLG4*

<i>DPLG1</i> guide 1 sense	CTTCGCACTTGTTTACTTGCATAGC
<i>DPLG1</i> guide 1 antisense	AAACGCTATGCAAGTAAACAAGTGC
<i>DPLG1</i> homologous arm 1 <i>EcoRI</i>	GGGTGTCGCCCTTCGCTGAAGCAGGTGGAATCCTGTCAGGTGATGGTTATTT AGCAGTTTTTATTAATAAAAAAAAAATAAATGAAACGAAACAGATAATGCTTTACAT TTCCTTGGAATTCAAATTGAAAGAAGCGTTTTTGGAAATATGTTAATGAATT CCCTTGAAAAAAGTGCTTGGTCAACAGAAATCAAGAAGATTTTATGCAGTC AAGCTCAAATTCATAGCAGATTTTTGTTTGTAGTTTTAGTGCTTCCTTTTGGT GTGCAATCCGAAGAAATGCGCGCATTACAATCTATGTACATGGCTGTGTTT GTGTTTGCACATAACGGAATTATAAGGAATTTGCTTACCCTATGATTTTTT TTCCTAGCTTGAATAAGAGTTTACATACCGGCCCATCACCCTACCTCGGGAA GTAGAGGAATCATCTTCGTTAAGTGGACGCTGCGTCCACGTCCCTGATCATC AGCCATTGTTACTTTTTGGCACTCGCGATCACTTAAAACCTCACAATGGAACT GAAAGAGTACTTTGAAAACGACGCTTTGGACTCAGTGTGACCAAATGGCCAG ATTTCAAAGATCGTTAAGCTTTTGGAAATGTGTAGATATTAGTCTTCACAACGA AAACAATCGTTTTGCAAATCTTCGCAATAGGTTCTATTTCCCTAGTATTTATTT TTTGGTACTTCGAGCTTTGATTTCTTTCTGTTCTGTTTTTTTTAAAAGTATAGAAA TACATAAAAAAAAAATTATTTAACTCCAAAAGTACATAACGTGCTTAAAAAAAA TAAAAGTATTAATTTACTTACTTACATTTTTTTAATTAACAACATTTAAAAGAT TATTTTTAAAGTGTAAGTTTCATCATAAAATTTCAATCGTTGCAAGAGCCTTA AATACCGCATGTACGTATGAGGTATTTCTAATATATACAAAACAGCCACCT GCTAATCTTGCATGCTAGCGGCCGCGGACATAT
<i>DPLG1</i> guide 2 sense	CTTCGATTCAAACCAGGCTAAAC
<i>DPLG1</i> guide 2 antisense	AAACGTTTAGCCTGGTTTTGAATC
<i>DPLG1</i> homologous arm 2 <i>XhoI</i>	TGCATAAGGCGCGCCTAGGCCTTCTGCAGCAACGGGTAAGTGAGCGATTAA AACGGTTTCTTCGGGAAGTCCCCTATCGGCCGAACAGCTGGTGCCTGCGCAA TCGGTATCTATCGCGAGGTGGCAGTTTGGAGTGCTCACGCTGCCATCTCCAAA CTGTGCACATGGATATTCTGTTCTTCAATAAAACAGTAGTTTATGTTTTTGTG TCATAGGCTGTGAAAAGTTTTAGTGAACGTTTAAGAATTTACGCTTAACTGG AAGCTACCAATAATCCAATATATTTATTAGACAAAATTGTGCATTTGCTATTA GTTATATTGTAGGTATATTGTGCCTCCGCAATCATTATGATTATAAAGGATGT GTGTTGATTTGCATTTTCAAGTTTTGCTCCATTTTACATACATTTACACCTGGC GCATTTTTCTGTGCTGGCAATGTAATCACACCGAGCTAGCTGACGACAAGTGCC AGTGACCTAGATTTTGGGAGTCTGCGCTTGGGTAGAGTAACGACACCACCA CCGCCACCGCAATCCTGCTCCGTGCCGTTAATCCTCCAGCTGCAGCAACAA AGTAAAGGTCACAAATCGCGAGGCAAGTAAGTTTTTGTACATATGACCTGGTT TACGGAGTGGCAGCTACGCTTGGTAGCCAGGGTGATGGCGGGTTAATCTA TATTGTATATTCACCGCCAGAGTGGCATACTATGCATTCTGGCTTCTCTATC TATCCACCAGATATGATTAATCCGCATCCGTAAACTTACGTCCAAATCCGATCA GAAATCATTACAGCATTTAAAACCTGTTTGTAGGCCTTTTTTGGGTTAGCGG AGTCAGAGATCACATCGAGCTTAGTACGTACCCCTCAACAGATATATGACTGG

	GTGACCAGGATCTGACATGCGTAACAACAGCGGCTATCATATCCCAGCTAGC AACAGTCTGAAGCTCGAGGCTCTCCGTCAATCGAGTTCAAG
<i>DPLG4</i> guide 1 sense	CTTCGCACAGCCGAGCACCCTTCG
<i>DPLG4</i> guide 1 antisense	AAACCGAAGTGGTGCTCGGCTGTGC
<i>DPLG4</i> homologous arm 1 <i>EcoRI</i>	GGGTGTCGCCCTTCGCTGAAGCAGGTGGAATGCCAGACCCGTTTAATAAGCA GGCCGCTCCCAGGAGCAGTGCAGTGGGCCACAATAGAGATTTGAGACATCA CATAGCATCATTACCGAACCGAACTTCCGCAAGAACTAAGCTAACTATAC AGAACCCTAGATGAAGCACGGAACGCAGATGGAACGAACGTGAACATAGTC GTAGTCGTAATGGTATGCAGTAGATAATGCGCCAAACGAGAGAGGTAAACAA AAGTCGGCAACGTGGCAACAGATGCTGTTACAGATTACAGATTTAGAAAATA TAAACAAGGAACTATACAAAATTCGTAAAGTACTTAGGTAAATTGCACTAAT ATAAACTCCCATTTGAACTGCCGCTCGCACTGTCTGTACTGCCGACAATGCTGC CATTTCCGCTCATATTACCATTGCCAATTCCTATGCCAATACCGCTTCCGTTTCC GCTCTGTTGGGCAAAGTCCAGCTGGAGCAGCCAGTTCCTCTTAGCCAGACCTT CGGGCGTCTCGCCACGCTGCGTACACAGCGGCTGGAAGCAGGTGGTCAGTCC AACGCTGGGCGTGAACCTCCGCCGTTTCGCGTAACTCCTCCTCGGTGGCCAGCA TGATGCAATCGTCGTAATTTTCATCGCTATCGACTGTAGGCCAGCTGCTGC TGCTCTGCCCTTTGCTGCACGATGCTGTGCTGGTGCGCCGAAATGAGTCGTC CGTCTTGCAGGAAATTGTACAGCGCCACACAGCCGAGCACCCTAATTCTT GCATGCTAGCGCCGCGGACATAT
<i>DPLG4</i> guide 2 sense	CTTCGAGATATTGACTATAATCCC
<i>DPLG4</i> guide 2 antisense	AAACGGGATTATAGTACAATATCTC
<i>DPLG4</i> homologous arm 2 <i>XhoI</i>	TGCATAAGGCGCGCCTAGGCCTTCTGCAGCCCCAGGTGCACCTAGAATCTATA GCTATGCCAGCTATATAGCCGCACCCCACTTTTCGGCCCCGATTTTCGCCAC TCGTCGTGTGTTTTGTGCAACAAATCGATTTTACTCACCTCCCAACCTCCCG TGTTCTTGCGATGTGTGTATAACCAGTTTCGGATCTTCGGGTGAAAATACGT TTTTGCTACGGGCTGCGTTGTATAATTAGAAATATCCGATAGATTGCTGTATCT ACTGCGTTTCTTTCATCAGACTCTCATCAGCAATCGGTTCTGGATTTTCATCCA CGGATTATATGTACGTGCAATGGTACCCAAGAAAAATCCGATGCCAATACGA AAATGCAGCCATCAGCAAGCAGTAGCAGTAGTAGCAGCAGCAGCAACAACAA CAACAAAACAGCCGAACGTCGACAAGTTTTTGATGGCAGAAATTGCATAGG GCCAGCCAAGGAGTTTTTGGCCAGTTGACGATTTTCACGGTTCGGTTATCC GTTTTTCTCGGCCAGCTCATCTGCGTTGCTGCTATCACTTGACCGAAAACCA CATTTTCCCAATTCTTCTTTTTTACCCTGATGTTGCCGATTCAGTGTGAC CGTGTGACGACGATAAAAAATGACATACGAACAGGCCGTTAATTGAATACC GTAATATACCACCAATGTCGTTTTCAAAAAATACCACTTTTATACCACTTTATTA CAAAAACAATTTATTTAAAATTATATTTAAATACTTAAGGCAAAAAGTATT TAAATTTAAAATTTAACCTAATACCTTTTAAAGTTAAAGATGTAATTTATCGCA GTAACCTAATTCGTTGGTCAATAATTTAGTTTAAAAAAGGCCAAATTACGATG TGCCCATATGGAATACTTTTATTGTCTATTCAATCTATATATCTACTTTATTTGC AATAGCTCGAGGCTCTCCGTCAATCGAGTTCAAG

Supplementary Table 2. Primers used to generate probes for *piwi* and *DPLG1-4*.

<i>piwi</i> Forward primer antisense	GTACTTCAGCACAGTCACGGAGTG
<i>piwi</i> Reverse primer antisense	AATACGACTCACTATAGGGACTCCTGACGAACTTGTTGCGAGACCAG
<i>piwi</i> Forward primer sense	TAATACGACTCACTATAGGGACTGTACTTCAGCACAGTCACGGAGTG
<i>piwi</i> Reverse primer sense	CCTGACGAACTTGTTGCGAGACCAG
<i>DPLG1</i> Forward primer antisense	GGCATCTGTCCTATCATCGAAAGC
<i>DPLG1</i> Reverse primer antisense	TAATACGACTCACTATAGGGACTCGAAGCGACTCATCAGCAGATTG
<i>DPLG1</i> Forward primer sense	TAATACGACTCACTATAGGGACTGGCATCTGTCCTATCATCGAAAGC
<i>DPLG1</i> Reverse primer sense	CGAAGCGACTCATCAGCAGATTG
<i>DPLG2</i> Forward primer antisense	CGATGTGCCTGTGGTGCTC
<i>DPLG2</i> Reverse primer antisense	TAATACGACTCACTATAGGGACTGATCCTCCAGTCCATCGTCCTC
<i>DPLG2</i> Forward primer sense	TAATACGACTCACTATAGGGACTCGATGTGCCTGTGGTGCTC
<i>DPLG2</i> Reverse primer sense	GATCCTCCAGTCCATCGTCCTC
<i>DPLG3</i> Forward primer antisense	GCGTCCTTGGCGTCTGCTC
<i>DPLG3</i> Reverse primer antisense	TAATACGACTCACTATAGGGACTCAAGCATGTGTGGTTCGCTCAG
<i>DPLG3</i> Forward primer sense	TAATACGACTCACTATAGGGACTGCGTCCTTGGCGTCTGCTC
<i>DPLG3</i> Reverse primer sense	CAAGCATGTGTGGTTCGCTCAG
<i>DPLG4</i> Forward primer antisense	ACGCTCTGGAGGAACAACAG
<i>DPLG4</i> Reverse primer antisense	TAATACGACTCACTATAGGGACTAGCAGCCAGTTCCTCTTAGC
<i>DPLG4</i> Forward primer sense	TAATACGACTCACTATAGGGACTACGCTCTGGAGGAACAACAG
<i>DPLG4</i> Reverse primer sense	AGCAGCCAGTTCCTCTTAGC
<i>DPLG4</i> qPCR primer forward	CGCAACGCAAGAAGAAGTC
<i>DPLG4</i> qPCR primer reverse	GAGGCACGGAAGCAGTAG
<i>Rp49</i> qPCR primer forward	CGTTGGGGTTGGTGAGG
<i>Rp49</i> qPCR primer reverse	CGTTTACTGCGGCGAGAT

Supplementary Figures

Supplementary Figure 1. Sequence alignment of *DPLG4* mutant with *DPLG4* reference

sequence showed that the mutant contains a 31 bp insertion (shown by red box) that creates a frameshift resulting in incorrect protein sequence and multiple stop codons (shown by *) after translation. Yellow box shows the canonical start codon. B) *DPLG4* mutant flies generate significantly more offspring in average compared to the control.

Supplementary Figure 2. Negative control for in situ hybridization of *DPLG1-4* during embryogenesis.

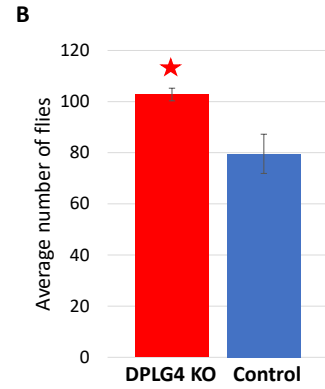
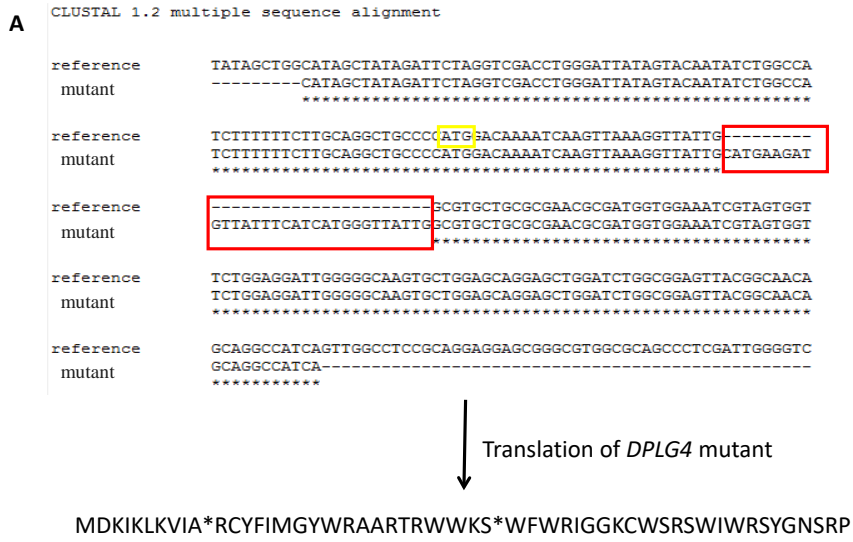
Supplementary Figure 3. Negative control for in situ hybridization of *DPLG1-4* in ovaries (A-D) and testis (E-H).

Supplementary Figure 4. PCA plot showing that *DPLG1-KO*, *DPLG4-KO* and control do not show very distinct clustering from each other suggesting the effects in the KOs are not very drastic compared to each other and the control.











Supplementary Figure 5 (A) Number of upregulated and downregulated genes in *DPLG4-KO* and *DPLG1-KO* ovaries. **(B)** Venn diagram showing overlap of DE genes in *DPLG1-KO* and *DPLG4-KO* ovaries. **(C)** Comparison of Log2foldchange of overlapping DE genes between *DPLG1-KO* and *DPLG4-KO* ovaries. Log2foldchange of overlapping DE genes in *DPLG1-KO* and *DPLG4-KO* ovaries shows positive correlation ($r= 0.7$, $p<0.001$). **(D)** Venn diagram showing overlap of enriched pathways and GO terms between *DPLG1-KO* and *DPLG4-KO* ovaries.

Supplementary Figure 6. Log₂foldchange of DE TEs in *DPLG1-KO* and *DPLG4-KO* ovaries. Both KOs show some DE TEs, however, their log₂foldchange are not very high.









Supplementary Figure 1



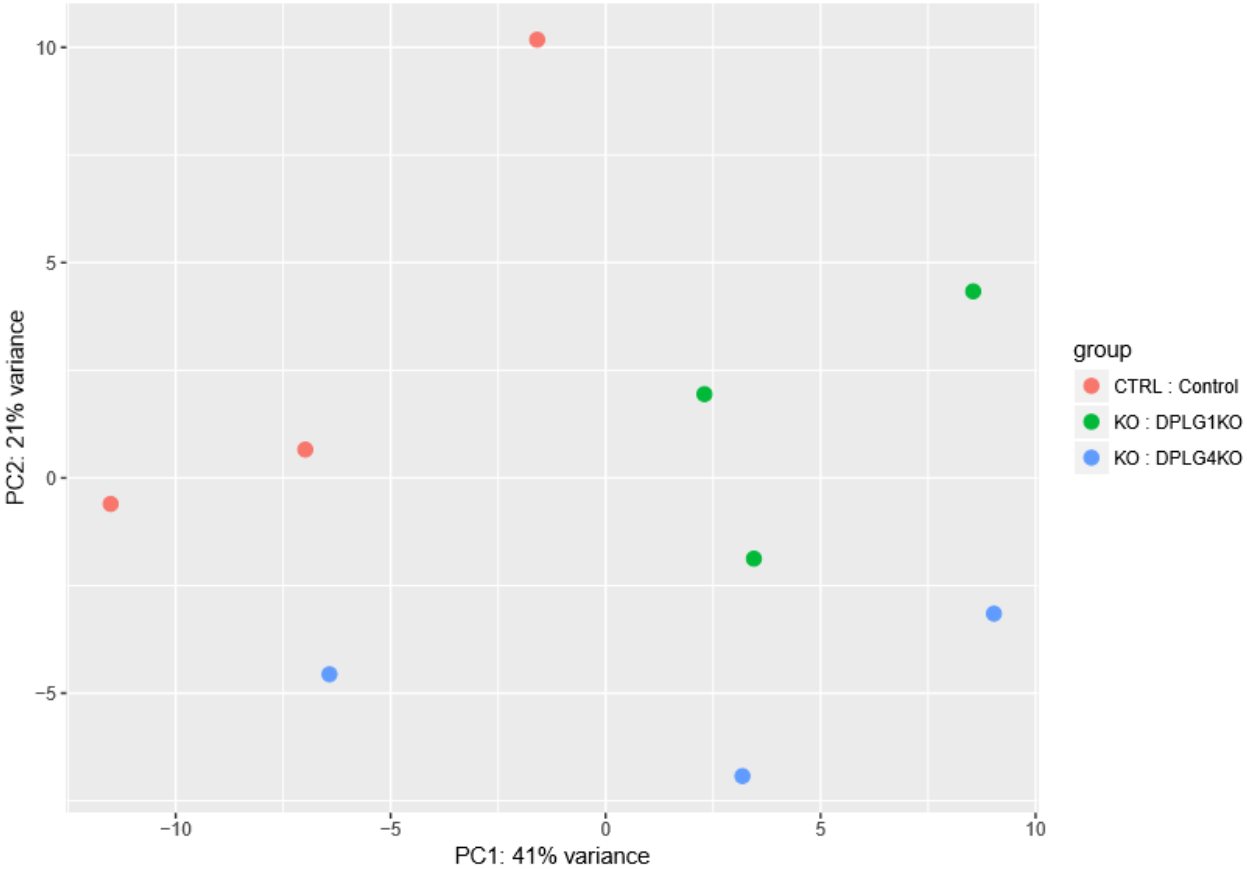
Supplementary Figure 2

<i>piwi</i> (Positive Control)		
<i>DPLG1</i> sense RNA (Negative control)		
<i>DPLG2</i> sense RNA (Negative control)		
<i>DPLG3</i> sense RNA (Negative control)		
<i>DPLG4</i> sense RNA (Negative control)		

Supplementary Figure 3

<i>DPLG1</i>	<i>DPLG2</i>	<i>DPLG3</i>	<i>DPLG4</i>
A 	B 	C 	D 
E 	F 	G 	H 

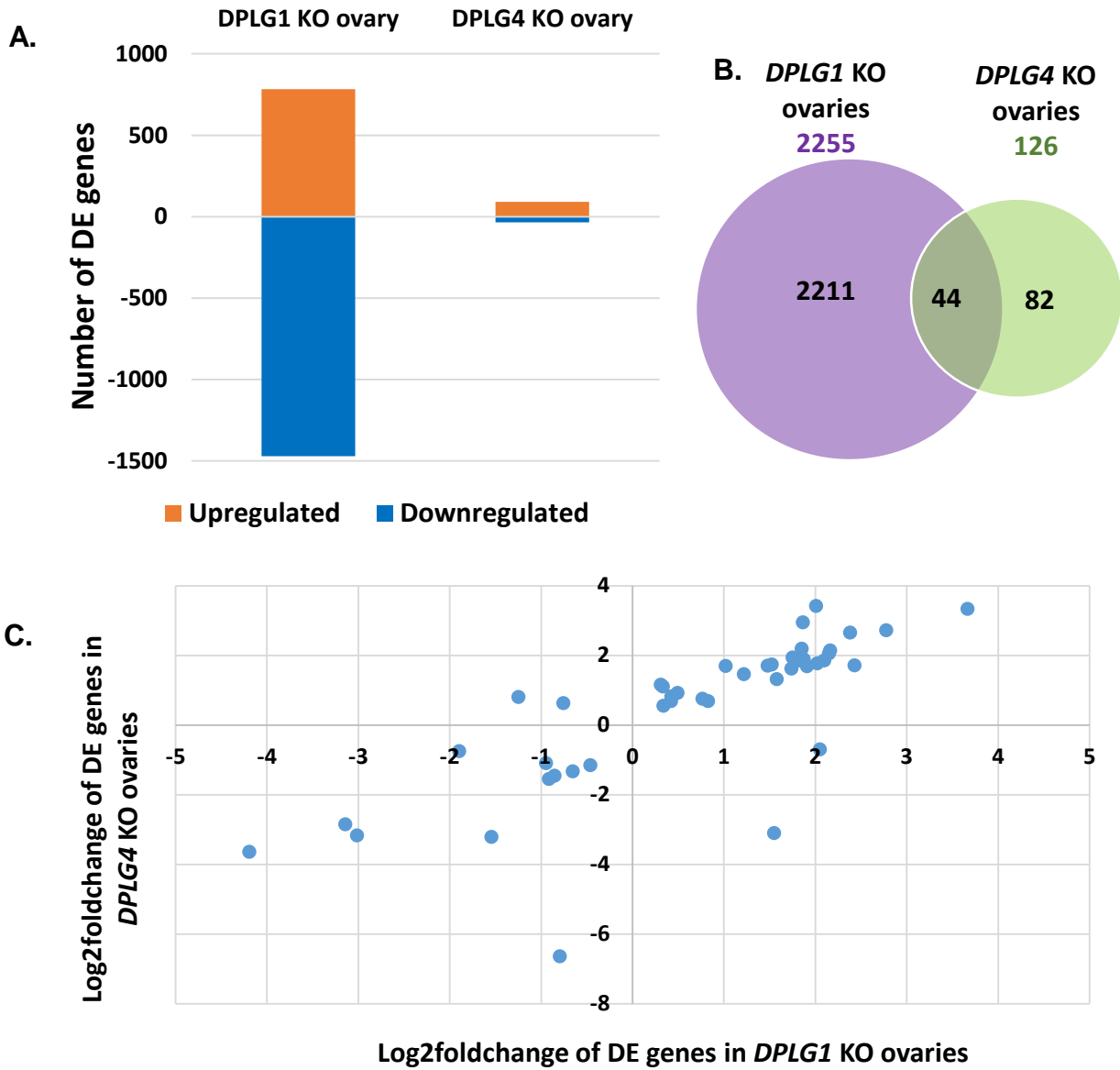
Supplementary Figure 4



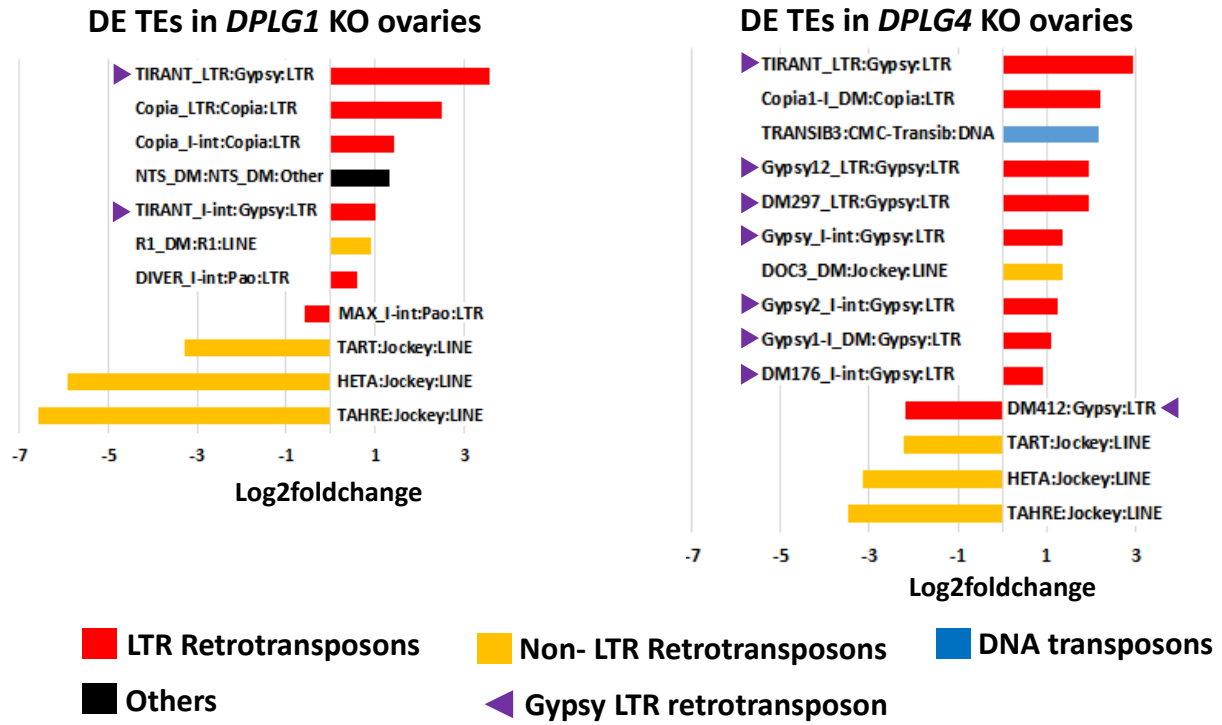
Supplementary Results

One of the control samples (*w¹¹¹⁸*) showed an increased proportion of reads mapping to multiple loci (27.66% of total reads instead of ~2% observed for all other samples) out of which most were mapping to regions that produced 5SrRNA. Although these are excluded from downstream analysis because they map to multiple loci, we wanted to test if there were major effects as a result of the smaller number of map reads for this sample. Exclusion of this sample from our analysis resulted in threefold increase in the number of DE genes (Supplementary Figure 5) and increase in only 3 DE TEs (2 upregulated and 1 downregulated) in *DPLG1*-KO ovaries (Supplementary Figure 6). Because of the increase in number of DE genes, the number of GO terms and pathways increased by more than two fold (Supplementary table 3). Despite this increase, most GO terms and pathways between the two analyses contained mostly related enriched terms. In contrast, there was not much change in the number of DE genes in *DPLG4*-KO ovaries (Supplementary Figure 5) and only 5 additional TEs upregulated (Supplementary Figure 6), however we observed 5 GO terms and 1 pathways enriched as a result of dropping this control sample (Supplementary table 3) in contrast to no GO terms and pathways enrichment when all three control samples were used, despite the number of DE genes remaining relatively similar. The overlap of DE genes between *DPLG1*-KO and *DPLG4*-KO ovaries from the initial analysis increased by 6 genes. Thus, although we did observe differences after dropping the control sample there were no major changes that we observed as a result of dropping one of our control samples suggesting that presence of these large proportion of reads mapping to multiple loci did not affect the overall analysis.

Supplementary Figure 5



Supplementary Figure 6



CHAPTER THREE

Concluding and Future Directions Chapter

Chapter one sheds light into the factors influencing domestication of TE proteins. After compiling examples of domestication events, we found evidence from various studies that TE proteins are domesticated as an adaptation to evolutionary conflicts. Conflicts between host and pathogen have resulted in domestication of TE proteins as part of adaptive immune system, conflicts between embryos and mother led to domestication of TE proteins involved in placentation, and conflicts between host and TEs themselves generated domesticated TE proteins that take part in suppressing their own activities and also potentially facilitate removal of TE related sequences as in the case of ciliates. We also discuss the possibility of centromere drive leading to TE protein domestication as several studies have described multiple independent domestication and convergent evolution of TE proteins in distinct organisms as centromere associated proteins. Lastly we discuss if whole TE could be domesticated and argue that as long as all the hallmarks of a TE are present, they remain opportunistic and would not be considered domesticated.

Chapter two addresses additional factors that influence TE protein domestication and provides insights into how these proteins are utilized by the host. We use *Drosophila melanogaster* as our model organism and domesticated transposase from *PIF/Harbinger* TE as domesticated genes of interest. We show that *DPLGs* co-express with transcription factors across *Drosophila* development and HA-tagged *DPLG 2-4* are able to localize with DNA in the nucleus of the ovaries and provide evidence that *DPLGs* likely are domesticated as regulatory proteins in *D. melanogaster*. We have some but not strong evidence that *DPLGs* are involved in controlling TE

activities. We also provide evidence that *DPLGs* show strikingly similar pattern of transcript localization during embryogenesis and show overlap in transcript localization in the gonads. Further results from HA-tagged *DPLG2-4* showed overlap in protein localization of these genes. Overlap of transcript and protein localization support that *DPLGs* might have related functions. RNA-Seq analysis of the ovaries of *DPLG1* and *DPLG4* null mutants showed positive correlation in the log2foldchange of DE genes and large overlap in the GO terms and enriched pathways. Further *DPLG1* and *DPLG4* show signatures of co-evolution. Together, these data provide strong support that at least *DPLG1* and *DPLG4* having related functions and provide support to the model that domestication of transposase promote domestication of related transposases. We also provide evidence that *DPLG4* has important functions in *D. melanogaster* including viability, fertility with increase in age, survival and neuronal development. Future extension of this work would be to follow up on functions of *DPLG2* and *DPLG3* and explore if they show overlap in function with other *DPLGs* as well.

An additional future direction would be to explore the domestication of the second ORF of *PIF* elements. We have generated some preliminary, but exciting results on this topic. All *PIF* TEs known in plants and animals distinguish themselves from traditional DNA transposons by the presence of two independent transcription units (Zhang, et al. 2001; Jiang, et al. 2003; Kapitonov and Jurka 2004; Zhang, et al. 2004; Casola, et al. 2007; Grzebelus, et al. 2007). One encodes a ~400-500-aa protein representing the catalytic transposase, while the other encodes a ~300-400 aa protein with a Myb-like domain (also known as SANT/trihelix or MADF domain). Studies in rice (Yang, et al. 2007; Hancock, et al. 2010) and zebrafish (Sinzelle, et al. 2008) indicate that both

proteins are required for transposition of *PIF* element. For zebrafish *Harbinger3_DR*, it is also known that the Myb-like protein interacts physically with the transposase and promotes localization of both proteins in the nucleus (Sinzelle, et al. 2008).

Interestingly, in addition to the domestication of *PIF* transposase, the Myb-like protein encoded by *PIF* TE has also been shown to be domesticated. In fact there have been several cases of co-domestication of both transposase and Myb-like proteins. In *D. pseudoobscura*, *D. persimilis* and *D. willistoni* a domesticated *PIF* transposase, *DPLG7*, is found immediately flanking a Myb-like (called MADF in *Drosophila*) gene *DPM7* (*Drosophila PIF MADF-like protein-encoding gene 7*). Since these genes are in proximity to each other it was proposed that these genes were co-domesticated from a single TE copy. The functions of these genes still remain to be elucidated. Additionally in vertebrates, HARBI1 (Harbinger Transposase Derived 1) is domesticated from *PIF* transposase and NAIF1 (Nuclear Apoptosis-Inducing Factor 1) is domesticated from the Myb-like protein (Sinzelle, et al. 2008). The functions of these genes have not been fully characterized, however, it is shown that HARBI1 physically interacts with NAIF1 and NAIF1 is responsible for the nuclear localization of HARBI1 (Sinzelle, et al. 2008). Co-domestication of yet another *PIF*-like transposase with a Myb-like protein was reported in *Arabidopsis*. In this study the domesticated transposase, HDP (Harbinger transposon-derived protein) 1, and the Myb-like protein, HDP2, were also shown to interact with each other physically. The HDP1 and HDP2 protein complex was shown to physically interact with histone acetyltransferase complex and prevents DNA hypermethylation and epigenetic silencing (Duan, et al. 2017). Together, these observations suggest that the domestication of *PIF*-derived transposase is frequently accompanied by the co-

domestication of their cognate Myb-like protein and in some cases have been shown to interact with each other just like the ancestral TE proteins.

In *Drosophila*, myb-like proteins or MADF proteins contains trihelix motif and bulky aromatic residues (Casola et al. 2007). They contain nuclear localization signals (NLS) and have the ability to bind DNA sequences. There are 48 MADF proteins in *Drosophila melanogaster*. These proteins are generally transcription factors, some with very important functions in the flies like *Stonewall*, *Adf-1*, and *Mes2*. We hypothesize that both DPLGs and MADF proteins were domesticated from *PIF* TEs and function through interaction with each other. Our Results show that DPLGs are able to localize with DNA in the nucleus of the ovaries. However, DPLGs lack putative nuclear localization signal (NLS) (Except *DPLG2* that shows presence of putative NLS) (Supplementary table 1) and can only translocate into the nucleus through interaction with proteins that contain NLS. We speculate that MADF proteins are the interactors that promote nuclear localization of DPLGs. If this holds true then MADF proteins could have facilitated the domestication of *DPLGs*.

Interestingly, a subset of MADF proteins associated with BESS domain show lineage specific expansion (Shukla, et al. 2013) in *Drosophila* that seems to coincide closely with domestication of *DPLGs*. It would be interesting to precisely date the rise of MADF proteins in *Drosophila* and explore if the domestication of MADF proteins coincided with the domestication of *DPLGs*. Like *DPLGs*, genes encoding for proteins with MADF domain also tend to be highly expressed in the CNS and the ovaries providing support for their functional relatedness (Figure 1). We further tested if *DPLGs* show signs of significant coexpression with MADF proteins across

developmental stages. We used data from modENCODE consortium (Graveley, et al. 2011), which groups genes according to their significant coexpression across fly development. All DPLGs are in separate clusters and strikingly, there are multiple MADFs that belong to coexpression clusters with *DPLG1*, *DPLG3* and *DPLG4*. In fact, the number of MADF proteins that coexpress with *DPLG1*, *DPLG3* and *DPLG4* are significantly higher than expected by chance from the proportion of MADFs in the genome and (Table 1 and Supplementary table 2).

We have also explored the Evolutionary Rate Covariation (ERC) of *DPLGs* and MADF proteins (Clark, et al. 2012; Findlay, et al. 2014). According to the calculated ERC values, several MADF proteins show signatures of coevolution with *DPLGs* (Supplementary table 3). *DPLG1* and *DPLG3* show again the highest number of coevolving MADFs. The MADF proteins that co-evolve and co-express with *DPLG1* are *az2*, *CG1602*, and *CG30403*, and with *DPLG3* are *CG11504*, *CG6683*, and *CG15601*. These genes would be the prime candidates to test for interaction with DPLGs opening doors to future research to answer if MADF proteins are domesticated from *PIF* TEs and interact with DPLGs and also elucidate if MADF proteins facilitated domestication of DPLGs. There are no MADF proteins that coevolve and coexpress with *DPLG2* and *DPLG4*.

References:

Casola C, Lawing AM, Betran E, Feschotte C. 2007. PIF-like transposons are common in drosophila and have been repeatedly domesticated to generate new host genes. *Mol Biol Evol* 24:1872-1888.

Clark NL, Alani E, Aquadro CF. 2012. Evolutionary rate covariation reveals shared functionality and coexpression of genes. *Genome Res* 22:714-720.

Duan CG, Wang X, Xie S, Pan L, Miki D, Tang K, Hsu CC, Lei M, Zhong Y, Hou YJ, et al. 2017. A pair of transposon-derived proteins function in a histone acetyltransferase complex for active DNA demethylation. *Cell Res* 27:226-240.

Findlay GD, Sitnik JL, Wang W, Aquadro CF, Clark NL, Wolfner MF. 2014. Evolutionary rate covariation identifies new members of a protein network required for *Drosophila melanogaster* female post-mating responses. *PLoS Genet* 10:e1004108.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al. 2011. The developmental transcriptome of *Drosophila melanogaster*. *Nature* 471:473-479.

Grzebelus D, Lasota S, Gambin T, Kucherov G, Gambin A. 2007. Diversity and structure of PIF/Harbinger-like elements in the genome of *Medicago truncatula*. *BMC Genomics* 8:409.

Hancock CN, Zhang F, Wessler SR. 2010. Transposition of the Tourist-MITE mPing in yeast: an assay that retains key features of catalysis by the class 2 PIF/Harbinger superfamily. *Mob DNA* 1:5.

Jiang N, Bao Z, Zhang X, McCouch SR, Eddy SR, Wessler SR. 2003. An active DNA transposon in rice. *Nature* 421:163-167.

Kapitonov VV, Jurka J. 2004. Harbinger transposons and an ancient HARBI1 gene derived from a transposase. *DNA Cell Biol* 23:311-324.

Sinzelle L, Kapitonov VV, Grzela DP, Jursch T, Jurka J, Izsvak Z, Ivics Z. 2008. Transposition of a reconstructed Harbinger element in human cells and functional homology with two transposon-derived cellular genes. *Proc Natl Acad Sci U S A* 105:4715-4720.

Yang G, Zhang F, Hancock CN, Wessler SR. 2007. Transposition of the rice miniature inverted repeat transposable element mPing in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A* 104:10962-10967.

Zhang X, Feschotte C, Zhang Q, Jiang N, Eggleston WB, Wessler SR. 2001. *P Instability Factor*: an active maize transposon system associated with the amplification of *Tourist*-like MITEs and a new superfamily of transposases. *Proc. Natl. Acad. Sci. USA* 98:12572-12577.

Zhang X, Jiang N, Feschotte C, Wessler SR. 2004. Distribution and evolution of *PIF*- and *Pong*-like transposons and their relationships with *Tourist*-like MITEs. *Genetics* 166:971-986.

Figure legends

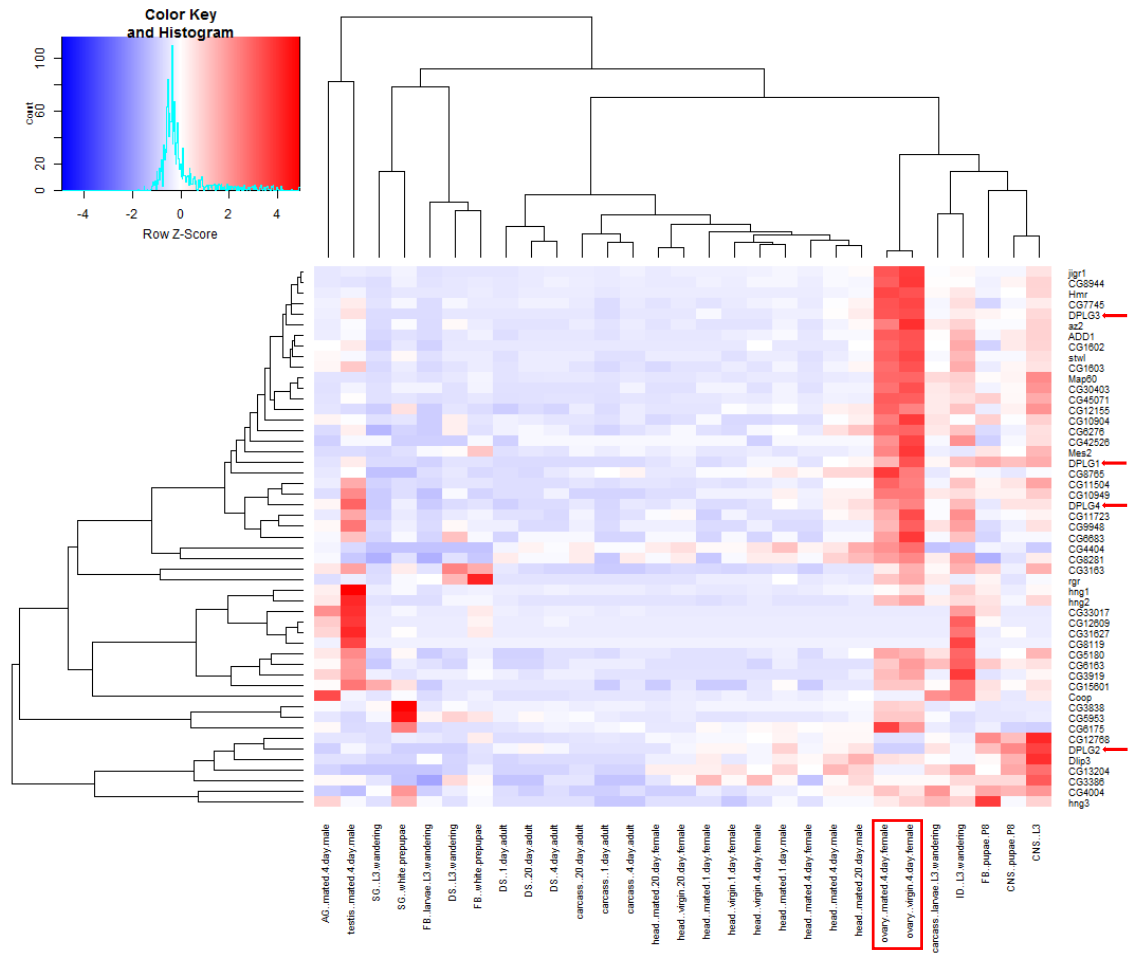
Figure 1. Heatmap showing transcription of genes encoding MADF domain and *DPLGs* across various tissues in *Drosophila*. *DPLGs* and MADF proteins show relatively higher expression in the ovaries and CNS. The ovaries are surrounded by red box in the X-axis and the *DPLGs* are pointed by red arrow in the Y-axis. FB: Fat body, CNS: Central nervous system, AG: Accessory gland, L3: Third instar larval stage, ID: Imaginal disc, SG: Salivary gland, DS: Digestive system.

Table 1. Number of MADF proteins that co-express with *DPLG1-4*.

Cluster Name	mE1_20_mRNA_expression_cluster_06	mE2_34_mRNA_expression_cluster_30	mE2_34_mRNA_expression_cluster_29	mE2_34_mRNA_expression_cluster_17	mE1_20_mRNA_expression_cluster_05	mE2_34_mRNA_expression_cluster_21	mE1_20_mRNA_expression_cluster_03
<i>DPLG</i> in the cluster	<i>DPLG1</i>	<i>DPLG1</i>	<i>DPLG2</i>	<i>DPLG3</i>	<i>DPLG3</i>	<i>DPLG4</i>	<i>DPLG4</i>
Total genes in the cluster	702	364	301	536	965	329	1036
MADF proteins	7	6	0	9	15	2	1
<i>P</i> -value	0.01	1.00E-03	0.35	7.58E-05	4.55E-07	3.14E-01	1.19E-01

The cluster name represents the cluster that incorporates the specific *DPLG* (See Supplementary Table 2 for all the details). *P*-values represent the probability of the number of MADF proteins being in a coexpression cluster of that size by chance (Fisher's exact test).

Figure 1



List of Supplementary tables

Supplementary table 1. Putative nuclear localization signals in *DPLGs*.

Supplementary table 2. List of MADF proteins that coexpress with *DPLGs*.

Supplementary table 3. Evolutionary rate covariation analysis highlighting MADF proteins that coevolve with *DPLGs*. Additionally list of MADF proteins that shows significant rate of coevolution and coexpress with respective *DPLGs* are also listed.