Dynamics of Transposable Elements Under Regulation By piRNA

Siddharth Singh Tomar

université PARIS-SACLAY

Dynamics of Transposable Elements Under Regulation By piRNA

Dynamique des éléments transposables sous régulation par piARN

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°577 : structure et dynamique des systèmes vivants (SDSV)
Spécialité de doctorat : Évolution
Graduate School : Life Sciences and Health, Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'**UMR EGCE** (Université Paris-Saclay, CNRS, IRD), sous la direction d'**Aurélie HUA-VAN**, Professeure à l'Université Paris-Saclay, et la co-direction d'**Arnaud LE ROUZIC**, Chargé de Recherche CNRS

Thèse soutenue à Paris-Saclay, le le 19 octobre 2023, par

Siddharth Singh TOMAR

Composition du jury

Membres du jury avec voix délibérative

Sébastien BLOYER Professeur, Université Paris-Saclay	Président
Cristina VIEIRA-HEDDI Professeure, Université Claude-Bernard Lyon 1	Rapporteure & Examinatrice
Anna-Sophie FISTON-LAVIER Maîtresse de conférences (HDR), Université de Montpellier	Rapporteure & Examinatrice
Silke JENSEN Chargé de Recherche, CNRS & Université Clermont Auvergne	Examinatrice

THESE DE DOCTORAT

NNT : 2023UPASL088

université PARIS-SACLAY
ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Dynamique des éléments transposables sous régulation par piARN
Mots clés : Éléments transposables, Évolution, piARN, Modélisation, Régulation, Transfert horizontal

Résumé : Les éléments transposables (ET) sont des séquences d'ADN mobiles trouvées dans les génomes de presque tous les organismes. L'activité incontrôlée des ET peut poser des risques importants pour l'intégrité et la stabilité du génome d'une espèce et sa capacité de survie, et presque tous les organismes ont développé des mécanismes pour se défendre contre les invasions de ces éléments. Dans la plupart des métazoaires, une classe de petits ARN connus sous le nom d'ARN interagissant avec PIWI (piARN) cible et empêche l'expression des ET. Les piARN proviennent de locus génomiques spécifiques appelés clusters de piARN. Les copies d'ET qui ont transposé dans un de ces clusters peuvent agir comme des allèles régulateurs pour toutes les autres copies de la même famille. De cette manière, les clusters de piRNA agissent comme des pièges à éléments transposables, formant une théorie de la régulation des ET connu sous de nom de "modèle piège". Des études antérieures ont étudié ce "modèle piège" et confirmé sa capacité à réguler les élements envahissants le génome d'une population. Cependant, nous manquons d'un traitement analytique du modèle piège et de la façon dont il diffère des modèles antérieurs décrivant la régulation des éléments transposables. Cette thèse propose un traitement mathématique et numérique du modèle piège, sous diverses hypothèses. Il montre des différences significatives entre la dynamique des ET par rapport aux modèles traditionnels.

Les ET sont également impliqués dans les transferts horizontaux entre espèces, car plusieurs études ont trouvé des ET partagés entre des espèces non apparentées, suggérant que non seulement les ET peuvent transposer et s'amplifier dans le génome, mais aussi traverser la barrière des espèces pour sauter entre espèces éloignées. La thèse examine la possibilité que des clusters de piARN agissent comme un système immunitaire adaptatif pour le génome contre les ET. En particulier, l'hypothèse selon laquelle une expositions passée à des familles d'ET pourrait fournir une immunité contre de nouvelles invasions sera examinée. En simulant une invasion simultanée de deux familles d'ET apparentés dans la même population, j'ai pu montrer que ce n'est que dans des conditions très spécifiques que les piARN d'une famille peuvent réguler et contrôler l'invasion de l'autre famille. J'ai également étudié l'influence de la régulation croisée des piARN dans les ET partagés entre la famille des *Drosophilidae*. J'ai montré que les piARN générés par *Drosophila melanogaster* ne pouvaient pas réguler la plupart des ET des autres espèces de *Drosophilidae*, à l'exception de ceux des espèces proches. J'ai également montré que les insertions d'ET dans les piARN n'empêchent pas l'activité ultérieure chez *Drosophila melanogaster*. Cette thèse ajoute à notre compréhension du modèle piège en tant que modèle de régulation complexe qui peut contrôler les invasions des éléments transposables. Cependant, il remet également en question l'efficacité de la machinerie piRNA en tant que gardien infaillible du génome, suggérant que le modèle piège pourrait n'être qu'une petite partie d'une machinerie complexe entourant les ET et les défenses contre leur activité.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Title : Dynamics of Transposable Elements Under Regulation By piRNA
Keywords : Transposable elements, Evolution, piRNA, Modelling, Regulation, Horizontal transfer

Abstract : Transposable elements (TEs) are mobile DNA sequences found in the genomes of almost all organisms. Uncontrolled activity of TEs can pose significant risks to the integrity and stability of an organism's genome and its survivability, and nearly all organisms have developed mechanisms to defend themselves against TE invasions. In most metazoans, a class of small RNA known as PIWI-interacting RNAs (piRNAs) targets and silences TEs. piRNAs originate from specific genomic loci known as piRNA clusters. TEs which transpose into piRNA clusters can act as regulatory alleles for all other TE copies of the same family. In this manner, piRNA clusters can be considered as traps for invasive elements, defining a regulation theory known as the "trap" mode. Past studies have investigated the trap model and confirmed its capacity to explain the dynamics of TEs in populations. However, we lack an analytical treatment of the trap model, and how it differs from past models describing TE regulation remains unknown. This thesis delivers a mathematical and numerical treatment of the trap model under various assumptions. It shows significant differences between TE dynamics under traditional models vs. the trap model.

TEs are also implicated in horizontal transfers between species, as multiple reports have found shared TEs between unrelated species, suggesting that not only TEs can mobilize in genomes but also cross the species barrier to jump between distant species. The thesis considers the possibility that the exposure to TEs from various families may confer to piRNA clusters a role in an adaptive genomic immune system against new horizontally-transfered elements. By simulating the simultaneous invasion of two related TE families in the same population, I show that the conditions in which piRNA from one family can cross-regulate efficiently other TEs are quite restrictive. I also investigated the presence of piRNA cross-regulation in TEs shared between the *Drosophilidae* family, and showed that piRNAs generated by *Drosophila melanogaster* could not regulate most TEs from the *Drosophilida*, beyond those from very close species. I have also shown that TE insertions in piRNA clusters do not prevent subsequent TE bursts in *Drosophila melanogaster*, dismissing the "genome immunity" hypothesis as a general and widespread control system against TE horizontal transfers. This thesis adds to our understanding of the trap model as a complex, stochastic regulation model that can control TE invasions. However, It also questions the efficacy of piRNA machinery as an infallible guardian of the genome, suggesting that the trap model might just be a small part of a complex machinery surrounding TEs and TE defenses.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my thesis supervisors, Dr. Arnaud Le Rouzic and Prof. Aurélie Hua-Van, for their guidance and invaluable support. Without them, I would not get to work on an exceptional topic and an impressive project. Thank you for instilling the scientific method in me and encouraging me throughout the pandemic.

I am also grateful to the jury members, Prof. Sébastien Bloyer, Prof. Cristina Vieira-Heddi, Dr. Anna-Sophie Fiston-Lavier and Dr. Silke Jensen for evaluating my work and the incredible scientific discussion and suggestions during my disputation. I would like to extend my sincere thanks to my thesis committee, Dr. Anne Genissel, Dr. Matthieu Falque, Dr. Vincent Castric and Dr. Florian Maumus, for the helpful suggestions throughout the course of the thesis.

The work in my thesis was performed during an unprecedented (hopefully) once-in-a-lifetime event, and I cannot thank enough all the people in EGCE for their support and valuable help. Many thanks to all the members of the team Evolution and Genomes for the incredible discussions about science and life in France. I would also like to acknowledge all the PhD students in EGCE for their support. Zhen, Apolline and Helloïse - I am immensely fortunate to have you as colleagues. Thank you for all the moral support and help.

I would like to thank my family, especially my brother Jyotiraditya, for keeping my spirits high during the tough times. This thesis would be impossible if not for the continuous support of my parents.

Lastly, I would be remiss in not mentioning the beautiful CNRS campus at Gif-Sur-Yvette. If I ever have to study something as challenging as the dynamics of transposable elements, I wish it would be in a place as charming and full of character as Gif-Sur-Yvette.

Contents

1 Introduction

1.1 A very brief introduction to transposable elements

Transposons, or Transposable Elements (TEs), are genomic elements that are present in all branches of the tree of life. They are repeated sequences in a genome and their ability to move within the genome of host organism, giving them the colloquial name - "jumping genes".

They were first discovered in maize by Barbara McClintock, who observed phenotypic changes in maize kernel based on a gene (now a known TE) that moved within the maize genome (McClintock, 1950; Ravindran, 2012). Soon after, TEs were discovered in bacteria in the form of a phage-like sequence that impacted the expression of nearby genes (Taylor, 1963). Subsequent studies revealed their presence in other organisms like *Drosophila melanogaster* (Kidwell et al., 1977). Eventually, the realization was made that TEs exist in nearly all organisms, including vertebrates such as primates (Britten and Kohne, 1968; Pace and Feschotte, 2007).

The aforementioned studies revealed that TEs influence the host genome directly by disrupting genes while jumping (transposing) into them. TEs can also disrupt the expression of genes by influencing the promoters or enhancer regions (Feschotte, 2008). Moreover, TE activity generates insertion polymorphism in the population where each individual could have a varying number of the same TE family in their genome (Rishishwar et al., 2015).

Since TEs can potentially harm the host, most organisms have developed mechanisms to contain their spread within the genome. Nevertheless, even with these defenses, one of the most striking features of TEs is their ability to transfer between different species (i.e., horizontal transfer) in addition to the traditional vertical transfer of genetic information (Schaack et al., 2010).

To summarise, TEs can:

- Increase their copies and move in the host genome.
- Cause changes to the host genome with detrimental and deleterious consequences.
- Introduce genetic polymorphism into a population or species.
- Force the organism to develop defenses to counter their spread.
- Cross the species barrier.

These characteristics of TEs make them exciting candidates for further studies on their ability to invade and persist in the genome of nearly all studied organisms. This thesis aims to answer questions regarding the ability of TE to infect new organisms or the same organism repeatedly, even in the presence of host defenses.

1.1.1 TEs are diverse

Detailed study of TEs led to the discovery of many different superfamilies, both in Eukaryotes and Prokaryotes. These superfamilies use different mechanisms for transposition and have different structural features. Following is the classification of TEs found in eukaryotes (Wicker et al., 2007):

- Class-I: These elements use an RNA intermediate to spread in the genome. This replication mode is made possible using enzymes that perform complementary DNA (cDNA) synthesis using TE RNA as a template. Class-I elements are further divided into following orders:

  - Long Terminal Repeat (LTR) retrotransposon: LTR retrotransposons are characterized by the presence of long terminal repeats flanking edges of the TE and two genes, gag, and pol. The gag gene encodes a viral particle-like protein which is required for replication (LTR retrotransposons require a virus-like intermediate stage.). The pol gene encodes for multiple proteins, including *reverse transcriptase* (RT), responsible for cDNA synthesis, and *integrase* (IN), which integrates the cDNA into the host genome.

  - Long interspersed nuclear elements (LINEs): Unlike LTR retrotransposons, LINEs do not require a virus-like intermediate stage but instead use the gene ORF2 to encode *reverse transcriptase* (RT) and *endonuclease* (EN). The *endonuclease* is responsible for identifying insertion sites in the host genome and creating a nick for the insertion

of RNA, which is followed by reverse transcription of RNA into cDNA. LINE elements also contain an additional gene called ORF1, which encodes for a gag-like protein.

- Penelope-like elements (PLEs): PLEs contain LTR on either one or both ends[1] (Arkhipova, 2006). One of the hallmarks of PLEs is the presence of multiple truncated copies of the TE on the 5'prime end of the "parent" TE. These truncated copies are generated due to an aborted or incomplete replication process (Gladyshev and Arkhipova, 2007). Moreover, they contain protein with *GIY-YIG endonuclease* and *reverse transcriptase* domain.

- *Dictyostelium* Intermediate Repeat (DIR): DIRs contain *tyrosine recombinase* (YR) instead of *integrase* found in LTR retrotransposons. Their hallmark is a lack of target site duplications and terminal repeats.

- Short interspersed nuclear elements (SINEs): SINEs are non-autonomous; hence they require the transposition machinery from other TEs like LINEs. SINEs also share some sequence features with LINEs, allowing them to utilize LINE proteins for replication (Sun et al., 2007).

• Class-II: Conversely, Class-II elements do not use any RNA intermediate and instead excise themselves from the existing locus and move to a new locus. Class-II elements are further divided into multiple orders:

- TIR elements: These TEs are flanked by terminal inverted repeats, which are recognized by transposase for excision. TIRs are ubiquitous in both prokaryotes and eukaryotes.

- Helitron: These Class-II elements use a rolling circle amplification method for replication. They do not require displacement of existing double-stranded DNA, which differentiates them from TIR transposons[2].

[1] Due to the amplification method employed by PLEs, they generate so-called "pseudo-LTRs", which usually result in flanking LTRs, even if the TE amplification process generates only one LTR (Gladyshev and Arkhipova, 2007).

[2] Helitrons also lack the characteristic terminal inverted repeats of TIR elements.

- Polintons: Some of the largest TEs, Polintons or Mavericks, are elements that encode their own *DNA polymerase* and other proteins required for transposition. Like LTR retrotransposons, they also contain *integrase* for integration into the host genome (Kapitonov and Jurka, 2006).

1.1.2 TEs are omnipresent

Not only are TEs present in nearly all organisms, but they also make up a significant fraction of the genome in many of them. They contribute to almost 40% to 45% of the *Mus musculus* and *Homo sapiens* genome. Similarly, maize has a TE content of more than 80%. Other model organisms like *Drosophila melanogaster* and *Arabidopsis thaliana* have TE content of 20% and 12%, respectively (Özgen Deniz et al., 2019; Stitzer et al., 2021).

The TE family and superfamily content varies among species. While the *M. musculus* genome still contains various active LTR retrotransposons, *H. sapiens* only have a few intact LTR retrotransposons families left in their genome in the form of HERV (Human Endogenous RetroVirus, mostly subtype K) (Hughes and Coffin, 2004; Bannert and Kurth, 2006). Some species, like *Caenorhabditis elegans*, have more class-II TEs than class-I TEs, whereas most *D. melanogaster* TEs are class-I (Laricchia et al., 2017; Mérel et al., 2020). These differences in TE order and family distribution illustrate the vast amount of genetic diversity introduced by TEs in the tree of life.

1.1.3 TEs have a significant effect on the host species

TE activity has a detrimental effect on the host organism as it replicates or transpose within the host genome, disregarding how this will affect the host. Thus the interaction between TE and the host genome can fall anywhere between commensalism and parasitism. TEs can influence the host organism and the species in multiple ways:

- Ectopic recombination: Since the TE copies are spread throughout the genome and chromosome, they can cause recombination events based on shared homology between the copies of same TE family present in different

loci on chromatids. These events can potentially be highly deleterious[3], i.e., they harm the host genome as the newly recombined region might have disrupted or missing genes (Sasaki et al., 2010; Lim and Simmons, 1994; Kent et al., 2017).

- Disruptions in essential genes: As TEs spread within the genome, they can insert into existing essential genes or disrupt their structure and/or expression by altering promoters or exon/intron boundaries (Hancks and Kazazian, 2016; Chuong et al., 2016).

- Disruption in genomic architecture: TEs can potentially disrupt the three-dimensional conformation of the genome by disrupting topologically associating domain (TADs). This disruption can be potentially disastrous if the TAD contains essential genes; this also influences the developmental gene expression dynamics (Zhang et al., 2019).

- Cost of TE activity: As TEs require materials from the host for their transposition, they compete for the available resources inside the cell. Moreover, they also compete for the transcription and translation machinery (polymerase and ribosomes). Some TEs also require DNA repair machinery to repair the transposition sites. The repair of double-stranded breaks can cause mutagenesis, and the double-stranded break itself is detrimental to DNA stability (Wicker et al., 2016; Galla et al., 2011).

1.1.4 Hosts can develop defenses against an invading TE

Due to the adverse effects of TEs on the host genome, the host needs to develop methods to prevent TEs from causing havoc. In most eukaryotic organisms, this is accomplished by using small RNAs and the associated argonaute protein family (Swarts et al., 2014). The small RNA transcripts work together with argonaute proteins to create RISC (RNA-Induced Silencing Complex). This complex then induces RNAi (RNA interference) pre- and post-transcriptionally, silencing the target (Iwakawa and Tomari, 2022). These small RNAs can be divided into three families based on their length and mechanism of action:

- miRNA (microRNA): A 21-nucleotide RNA transcript that can regulate protein-coding genes and are part of the genome regulatory machinery. They are conserved across multiple species across the tree of life. They

[3] Petrov et al. (2010) also observed purifying selection against TE families which could be a potential source of ectopic recombination events.

target transcribed genes (mRNA) using partial sequence complementarity (post-transcriptional regulation) (He and Hannon, 2004; Fabian and Sonenberg, 2012).

- siRNA (small interfering RNA): A 20-24 nucleotide RNA transcript, similar to miRNA, is involved in post-transcriptional regulation and has a similar mechanism of action. However, in addition, they can also induce heterochromatin formation and are mainly involved in genomic defense against foreign DNA, such as viruses (Volpe and Martienssen, 2011; Carthew and Sontheimer, 2009).

- piRNA (Piwi-interacting RNA): A 18-29 nucleotide RNA transcript that primarily targets TEs and silences them. They were initially discovered in *D. melanogaster* (Aravin et al., 2007). Like miRNA and siRNA, piRNA employs members of the argonaute protein family and can regulate TEs at the epigenetic level by silencing the TE locus[4]. piRNAs work with Histone/DNA methyltransferases (HMTs/DNMTs) and methylate the histone/DNA to prevent further transcription of TE (Ozata et al., 2018). Both miRNAs and siRNAs are conserved across multiple species and clades, but piRNAs are rarely conserved at the sequence level (Özata et al., 2019; Parhad and Theurkauf, 2019).

  A typical piRNA cluster sequence is mainly composed of remnants of TEs, which are the source of sequence complementarity required to target active TE families. The piRNA-generating loci are referred to as piRNA clusters[5] (Brennecke et al., 2007). The nascent long piRNA precursor transcript is processed into smaller fragments, i.e., the piRNA themselves. This process is further accelerated by the germline and piRNA-specific feedback pathway known as the ping-pong cycle. The Ping-pong cycle involves partial complementarity between piRNAs and their targets (piRNA precursor transcript or TE transcript). piRNA-loaded argonaut protein then processes the targets and dices them into smaller pieces, generating more piRNAs. By using sequencing complementarity, these processed piRNAs find the TE sequence in the genome and recruit the silencing component of piRNA machinery (figure 1.1). This silencing component (RISC) varies between organisms but usually silences the TE locus by either performing histone modifications or DNA methylation. In *D. melanogaster*, this process is accomplished by histone methylation.

[4]While siRNA mainly silences loci by heterochromatin formation, piRNA can utilize DNA methylation machinery in addition to heterochromatin formation.

[5]Due to observed piRNA read aggregation when mapped during their discovery

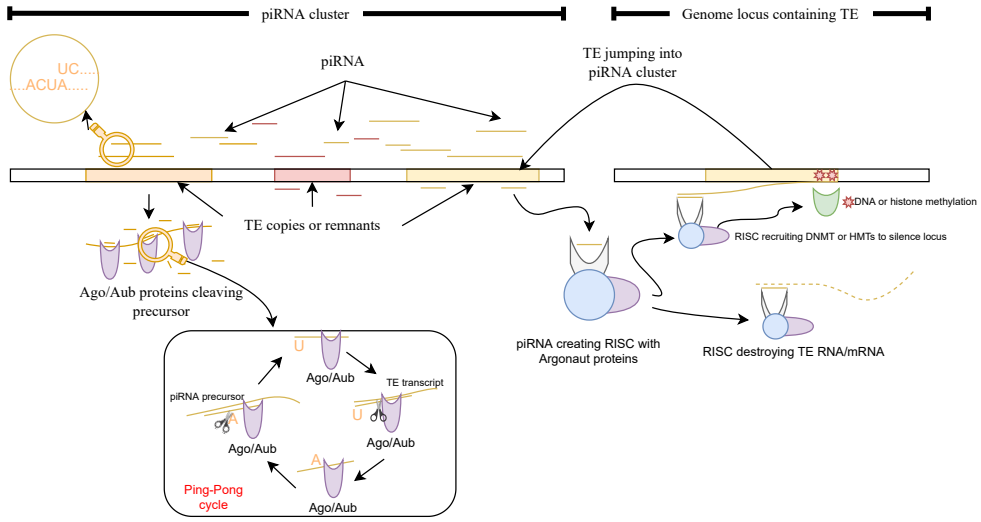


Figure 1.1: A piRNA cluster is a locus containing either intact or fragmented TE copies. The piRNA cluster generates a precursor long non-coding RNA transcript, which is then processed by argonaute proteins (PIWI in *D. melanogaster*) to generate the piRNA transcripts. These transcripts then associate with downstream Ago and other proteins to create an RNA interference complex. The complex can degrade the TE messenger RNA or recruit methylation machinery to silence TE locus by DNA/histone methylation.

There are some significant differences between the mechanism of piRNA and miRNA/siRNA:

- While miRNA and siRNA can target TEs, piRNA machinery is the primary method of genomic defense against TEs in most metazoans. piRNA uses the TE sequence to target the respective TEs, so they confer adaptive immunity at the genome level[6].

- miRNA and siRNA transcripts can target multiple loci or genes simultaneously, whereas piRNA are specific to the TE sequence they are derived from. However, the nascent primary transcript, which contains the piRNA, can contain remnants of multiple TE families.

- piRNA can degrade transposon mRNA, similar to siRNA and miRNA, using RISC-mediated degradation. However, unlike miRNA, piRNA can also silence the locus containing TE, thereby preventing it from producing any mRNA copy.

[6]Similar to how an adaptive immune system would use antigens derived from a pathogen to target it.

- piRNA-induced silencing can be maternally transfered. This inheritance is in the form of piRNAs present in the gametes (oocytes) (Brennecke et al., 2008; Thomas et al., 2014; Roovers et al., 2015).

- Processed piRNA transcripts have unique features, including enrichment for 5'prime uridine and 10^{th}-position adenine. This pattern is attributed to the activity of the Aubergine protein involved in ping-pong amplification, which cleaves longer piRNA transcripts based on the position of adenine (Wang et al., 2014). Furthermore, they have 2'-O-methylation modification at their 3'prime end (Lin, 2007).

piRNA and semantics: piRNA derive their name from the *PIWI* proteins which process them. *PIWI* proteins themselves derive their name from the P-element: P-element Induced WImpy. Their (*PIWI*) discovery was due to the mutations caused by P-element insertion in the locus containing *PIWI* genes (Lin and Spradling, 1997). These genes, when silenced, caused impaired development of *D. melanogaster* ovaries and testes. Homologs of *PIWI* proteins in other animals share an informal naming scheme, e.g., in mice, they are known as *MIWI*, and in humans, they are known as *HIWI* (Aravin et al., 2008; Sasaki et al., 2003). However, the small RNA these proteins process is always referred to as piRNA, regardless of the species they are found in.

Small-RNA are not always required for TE silencing: It is essential to point out that there are other methods for TE defense that do not require small RNA. One example is Repeat-Induced Point Mutation employed by fungi where TE sequences are enriched with mutations that inactivate the TE. The machinery targets methylated GC regions and mutate them to AT. This type of defence directly modifies genetic information instead of silencing it (Hood et al., 2005).

1.1.5 TEs can cross species boundaries

The standard mechanism of transfer of genetic information is vertical transfer, i.e., transfer of genetic information from parents to offspring. However, the "horizontal" transfer (HT) of genetic information between species is also observed. HT is prominent in prokaryotes, which are used to gain adaptive advantages such as antibiotic resistance. However, it is also observed in eukaryotes (Soucy et al., 2015). An example of HT in eukaryotes is the significant presence of bacterial genes in bdelloid rotifers (Nowell et al., 2021). Similarly, the whitefly *Bemisia tabaci* received multiple genes from plants that allows it to mitigate

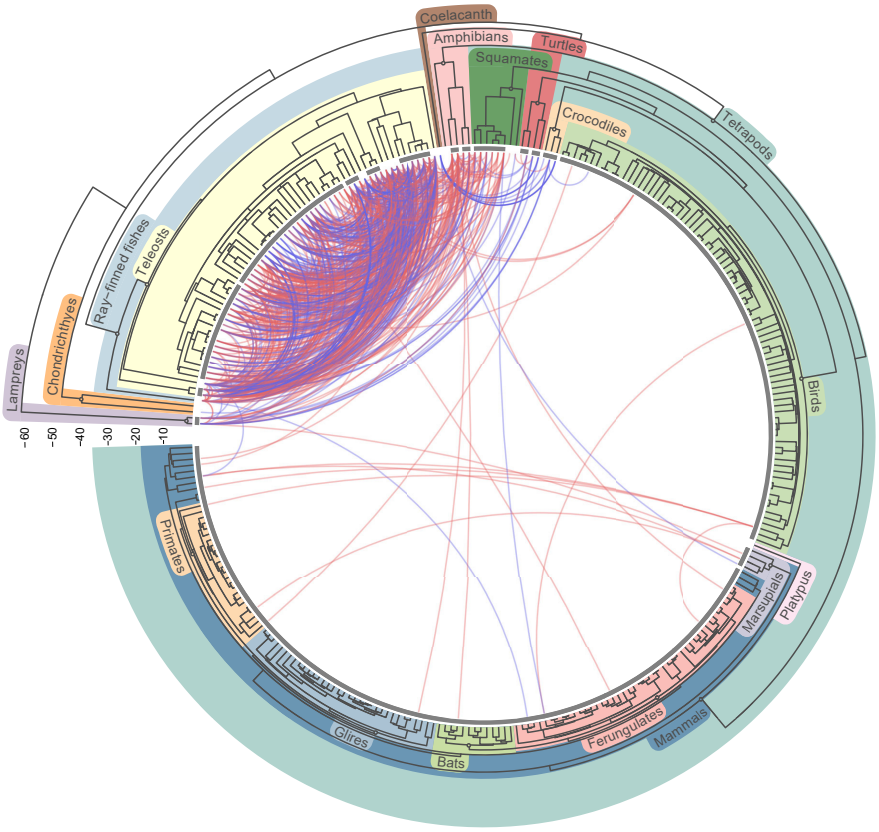


Figure 1.2: From Zhang et al. (2020). Blue and red arcs indicate the horizontal transfer of Class-I TEs and Class-II TEs, respectively, in vertebrates. Notice the increase in observed TE transfer events in teleost fishes and amphibians.

plant toxins (Gilbert and Maumus, 2022). Nevertheless, compared to bacteria, eukaryotic horizontal gene transfers are sporadic and hard to identify (Stanhope et al., 2001; Salzberg, 2017).

However, multiple studies support extensive TE horizontal transfer in insects and vertebrates compared to genes. In insects, more than 2200 instances of TE horizontal transfer have been recorded (Peccoud et al., 2017). Similarly, more than 900 TE horizontal transfers have been recorded in vertebrates (Zhang et al., 2020). Observation made by Zhang et al. (2020) indicated the presence of extensive TE transfer in aquatic vertebrates and amphibians (figure 1.2). The vectors that enable TEs to jump species boundaries are still unknown. However, it has been suggested that viruses, parasites, and endosymbiosis play a role in transferring genetic material between species (Keeling and Palmer, 2008).

A TE not only needs to evade the genomic defense of the host species (piRNA), but they need to be able to invade the germ cells of the target species. TE invasions in somatic cells are inconsequential to the population, even if they damage or destroy the cell. However, if a TE integrates in the germ cell, it can enable vertical transmission of TE in the population.

A TE silenced by the piRNAs can reinvade the same species. This phenomenon is observed with multiple invasions of I-element in the *D. melanogaster* genome, indicating that TEs can survive the host defenses (Blumenstiel, 2019). The TE[7] can:

- Evade the piRNA machinery; a TE can potentially accumulate non-deleterious mutations allowing it to remain active, so piRNAs can no longer target it using sequence complementarity.
- Invade a sub-population where the genomic defenses have not been established and reinvade the original population, which has accumulated mutations in the TE insertions in the piRNA cluster.
- Transpose during gametogenesis or early embryonic development. In higher-order metazoans, epigenetic regulation(DNA/histone methylation) is erased briefly during early embryo development and gamete formation, which gives TE a window to invade (Messerschmidt et al., 2014).

TE and host defenses are in a constant arms race, but it is undeniable that TEs have shaped the eukaryotic genome. Therefore, it is essential to understand the dynamics of TE invasion in a population and how the population contains the invasion. The following section will detail past studies on TE dynamics.

1.2 Introduction to TE dynamics

Doolittle and Sapienza (1980), along with Orgel and Crick (1980), considered TEs as selfish genomic components separate from the organism's phenotype. This conjecture meant that TEs undergo non-phenotypic selection or selection at the molecular scale. They stipulated that the only purpose of TEs is to multiply in

[7] When we consider that a single TE can potentially have hundreds or thousands of copies evolving independently in a genome, it is possible that one of those copies might be able to escape the piRNA machinery.

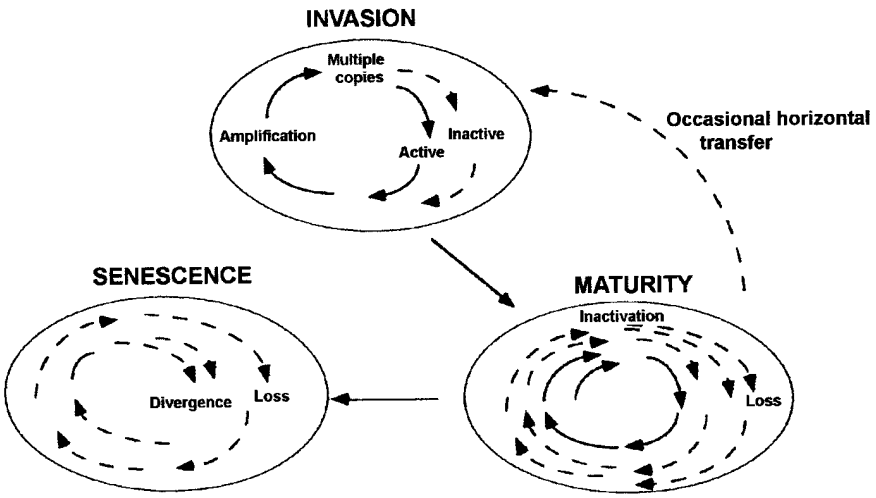


Figure 1.3: Figure from Kidwell and Lisch (2001). The TE family will initially increase the copy number in the genome, followed by a decline in its ability to transpose. During this time, it will incur mutations that will inactivate TE copies. Subsequently, the TE family can reinvade another population or become extinct as a genomic relic in the host organism.

the DNA[8], and the DNA itself needs to survive within the ecosystem of the cell. Hence TEs can replicate independently without affecting selection on the whole organism. However, as discussed in the previous section, TEs have deleterious effects on the cell's survivability. In multicellular organisms, this effect extends to the organism containing the cell. Most studies reflected on the overall negative effect of TEs and considered their effect either neutral or deleterious (Cordaux and Batzer, 2009; Chuong et al., 2016).

TEs, behaving very much like a parasite of the genome, undergo multiple stages, which could be considered as a life cycle independent of that of a host (figure 1.3):

1.2.1 TEs have a life cycle

- Invasion: The first step includes the introduction of TE family in the population by horizontal transfer. Functionally, each TE copy in this stage has the maximum probability of jumping again, i.e., a high transposition rate. Earlier models (Le Rouzic and Deceliere, 2005; Charlesworth and

[8] and thus, used the term "selfish-DNA", due to its ability to multiply in the genome without regard for the host's fitness. This behavior contrasts genes that increase their copy in the population by increasing the host's fitness.

Charlesworth, 1983; Le Rouzic et al., 2007) have characterized this step in detail, allowing us to predict the long-term trajectory of TE invasion in the population.

- Maturity and regulation: This stage includes the regulation[9] of TE family by the host and the persistence of a stable copy number[10]. At this step, the TE copies can become inactive due to mutations or suffer incomplete transposition events, leading to the loss of some copies. For a TE family to survive, it must find a new population or an individual without regulation against the TE family. This transfer can allow the TE family to sustain itself in another population during senescence in the original population.

- Senescence: Most TE copies are degraded or cannot transpose at this stage. The accumulation of deleterious mutations will lead to the eventual cessation of all transposition activity.

1.2.2 Early population genetics models

Studying TE invasion in natural populations is possible, as with P-element in *D. melanogaster*, but finding species where groups with and without invading TE exists is difficult. Therefore, an alternative is to model the invasion of TEs using analytical and individual-based models, which either mathematically resolve or simulate TE invasions in-silico, respectively. Presented below are early models which investigated the spread of TEs. [11]

- Brookfield (1982); Kaplan and Brookfield (1983); Langley et al. (1983): J. Brookfield postulated that recombination would assist the increase in TE copy number in the population. Recombination will segregate the TE copies, dilute them in the population, and counter a rapid increase in TE copies in a lineage, hence the potentially extreme effects of negative selection as each individual will have fewer TE copies compared to a (hypothetical) non-recombining individual of the same population. His conclusions suggested that a higher recombination rate can segregate TEs faster. Nevertheless, more importantly, to answer if these TEs are

[9] A decrease in transposition rate.

[10] Charlesworth and Charlesworth (1983) considered TEs to self-regulate, i.e., their ability to transpose is inversely proportional to the TE copies present in the genome; however, we now know that targeted defenses against TE exist in most organisms.

[11] The first complete metazoan genomes were sequenced in the late 1990s. Thus these studies had to compare the results from their models and estimate the extent of TE activity using incomplete datasets.

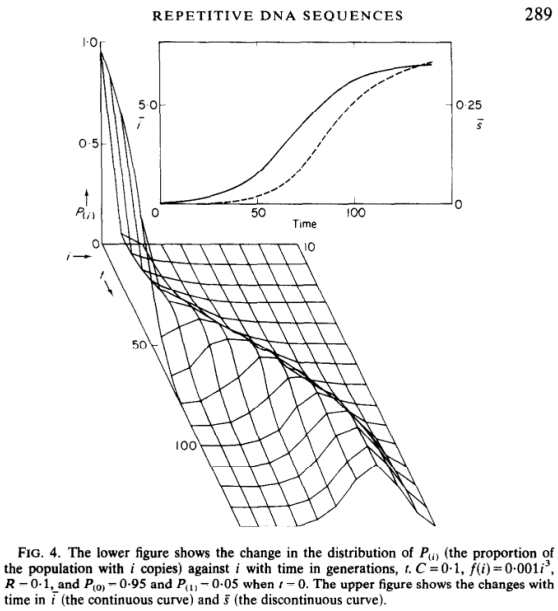


FIG. 4. The lower figure shows the change in the distribution of $P_{(i)}$ (the proportion of the population with i copies) against i with time in generations, t. $C = 0{\cdot}1$, $f(i) = 0{\cdot}001i^3$, $R = 0{\cdot}1$, and $P_{(0)} = 0{\cdot}95$ and $P_{(1)} = 0{\cdot}05$ when $t = 0$. The upper figure shows the changes with time in $\bar{i}$ (the continuous curve) and $\bar{s}$ (the discontinuous curve).

Figure 1.4: From Brookfield (1982). Figure describing dynamics of TE copy number in a simulated population. ī and s̄ are the mean TE copy number and mean selection respectivly. R is the recombination rate, $f(i)$ is the fitness cost associated with TE, and C is the transposition rate.

under stable equilibrium and are actively transposing, he designed simulations(figure 1.4) where the various parameters like transposition rate, fitness cost, carrier frequency, and recombination rate were explored.

His conclusions stated that a TE would continue to increase its copy in the population until population is close to extinction: "...the per copy duplication rate by transposition does not decrease with mean copy number per individual, the mean copy number will increase until it is very close to a number causing lethality". His model was based on the assumption that only selection acts on eliminating TE from the host. Later theoretical studies showed the presence of a stable TE copy number equilibrium if the transposition rate decreases with an increase in copy number or if the deleterious effects of TE are not additive (Charlesworth and Charlesworth, 1983; Le Rouzic et al., 2007).

His simulation design was further expanded in subsequent studies of TE dynamics. In these subsequent studies, Kaplan and Brookfield (1983); Langley et al. (1983) used individual-based simulations to describe the dynamics of TE spread in a simulated population in the presence of selection and

transposition inactivating mutations. These studies further established a framework for studying TE dynamics in-silico and juxtaposing the results with observed TE copy numbers in natural populations.

- Charlesworth and Charlesworth (1983): Perhaps one of the most influential studies in the field of TE dynamics, they described the dynamics of TE invasion using individual-based simulations under two assumptions,

  - Self-regulation: The TE can self-regulate the transposition rate (u_n) as it invades the population, with u_n inversely related to the TE copy number n. With the deletion rate as v and no effect of TE on the fitness of the host (i.e., neutral insertions), the change in TE copy number between generations can be approximated using the equation (Le Rouzic and Deceliere, 2005) :

$$\Delta\overline{n} \approx \overline{n}(u_{\overline{n}} - v) \tag{1.1}$$

  - Regulation by selection: The TE carries a fitness penalty, and therefore selection acts on the population to check the growth of TE. Given the average population fitness w_n, it is possible to approximate the change in copy TE copy number using the equation (Le Rouzic and Deceliere, 2005) :

$$\Delta\overline{n} \approx \overline{n}(\frac{\partial \ln w_{\overline{n}}}{\partial \overline{n}}) + \overline{n}(u_{\overline{n}} - v) \tag{1.2}$$

Both equations, 1.1 and 1.2 can yield[12] an equilibrium state for TE to exist in the population (figure 1.5), thus contrasting with the conclusion from Brookfield (1982) [13]. The study demonstrated that TE family could exist in the genome without reducing the host fitness considerably. It thus also gave a plausible explanation for the observed prevalence of TEs in natural populations.

[12] Depending on the shape of $u_{\overline{n}}$ and $w_{\overline{n}}$

[13] Under specific constraints: (1) Self-regulation model: u_n is smooth and decreasing, i.e., $\frac{\partial u}{\partial n} < 0$, and $\lim_{(u)_n \to \infty} < v$. (2) Selection model: $\frac{\partial log(w)}{\partial n} < 0$

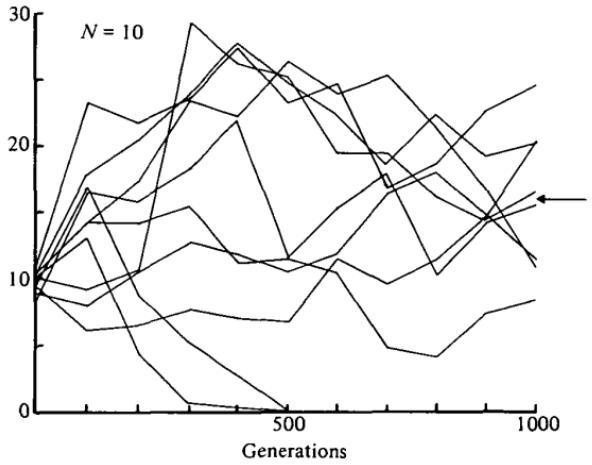


Figure 1.5: From Charlesworth and Charlesworth (1983). Figure describing dynamics of TE copy number in a simulated population of 10 individuals. x axis is the mean TE copy number per simulation. The plots are noisy due to the low population size and a low number of simulation replicates; however it is possible to observe the equilibrium as the average TE copy number hovers around 20.

1.2.3 Species/TE specific models

The early models were generalized to explain the dynamics of TEs in a freely recombining population (haploid in Brookfield (1982) and diploid individuals in Charlesworth and Charlesworth (1983)); natural populations sometimes have unique features that are not accounted for in such models. Following are some models or systems which differ from the standard assumptions (diploid genome, random mating, free recombination, and so on):

- Species biology: The species' biology, including attributes like the mode of reproduction, can affect the efficiency of TE invasion.
  - Mode of reproduction: Asexually and sexually reproducing organisms have significant differences in how they produce and process the transfer of genetic material to progeny. Sexually reproducing organisms have the process of meiosis (Lenormand et al., 2016), which also entails the process of crossing over (or chromosomal recombination), asexual organisms may dispense such mechanisms, resulting in near perfect replica of genetic material between generations. Such differences can significantly impact the dynamics of TE invasion in a population (Arkhipova, 2005). For example, prokaryotes, most unicellular eukaryotes, and even some metazoans (and plants) can

reproduce clonally. While this is the only reproduction route for prokaryotes , it is rare in complex[14] metazoans (Neaves and Baumann, 2011). As TEs are detrimental to organisms, their increase in clonal or asexual organisms would be highly deleterious to the host due to the host's (potential) inability to purge TE via recombination (Muller, 1964; Arkhipova and Meselson, 2004). An unchecked TE expansion in the genome would lead to the extinction of the lineage which contains the TE family. However, we still observe TEs in asexual organisms, indicating the presence of selective sweeps that might carry deleterious TE insertions along with rare beneficial TE insertions (Charlesworth et al., 1992; Barton, 2010). Simulations by Dolgin and Charlesworth (2006) in asexual populations observed the elimination of TEs from a large population in the presence of excision and an equilibrium-like state. However, the simulations also found that the TEs always overwhelm the host in smaller populations and lead to extinction[15].

  - Development: An organism might undergo multiple developmental stages, and some stages might repress the regulation of TEs. For mammals, it is typical for gametes and early stages of embryogenesis to be devoid of any DNA/histone methylation, which presents a brief window of opportunity for TEs to escape regulation (Surani and Hajkova, 2010). Such cases create an exception to the models where regulation is fixed and absolute.

- TE-specific molecular mechanisms: TEs can have different replication methods and interact with the host molecular machinery for replication. All TEs require RNA-polymerase from their host for transcribing the transposase and other proteins. However, for the success of TE invasions, one of the essential molecular aspects is the generation of their transposition machinery. TEs are usually autonomous, i.e., they produce their own functioning transposition machinery. However, it is possible to find TE copies which are non-autonomous and are dependent on transposition machinery from their autonomous counterparts. There are multiple examples of such TEs present in metazoans, including in primates in the form of the Alu SINE element. SINE elements use the host RNA polymerase-III proteins to transcribe but depend on LINE reverse-transcriptase for reverse-transcription and integration back into the host genome (Bennett et al., 2008) - making SINE a parasitic TE order in relation to LINE. Such relationship is usually detrimental to the replicative success of au-

[14] i.e., in metazoans with a significantly different cell types (Valentine et al., 1994)
[15] Dolgin and Charlesworth (2006) also found an equilibrium state for TEs in infinite population.

tonomous TE. Such differences require additional constraints on the TE invasion model and disregard the assumption that each existing TE copy can independently create more genomic copies.

- Hybrid dysgenesis and regulation: Certain TE families can introduce hybrid dysgenesis, i.e., the mating direction among TE carriers can influence progeny phenotype. This phenomenon was first observed in *D. melanogaster*, where P-element (a class-II TE) was present. The crosses between P-element-lacking females (*M*-strain) and P-element-positive males (*P*-strain) resulted in sterile progeny, whereas a cross in the opposite direction resulted in viable progeny (Bingham et al., 1982). Moreover, the sterile progeny exhibits many characteristics of deleterious effects of unchecked TE activity, including aberrant recombination, high mutation rates, and chromosomal rearrangements. These anomalies are mainly restricted to the germ cells since P-element is mainly active in the gonads of the fly (Ghanim et al., 2020).

  Such behaviors were modeled with success and reflected the distribution of P-element in the wild (Brookfield, 1991)[16]. The P-M system is one of many examples illustrating the role of TE in hybrid dysgenesis (Bucheton et al., 1992; Yannopoulos et al., 1987). Later studies discovered that piRNA confers defense against P-element in female oocytes and thus checked the spread of P-element in the progeny. In *D. melanogaster*, the piRNAs are maternally inherited and found in the oocyte's cytoplasm. Since the P-strain males lack any piRNA defenses in their germ cells, they cannot control the proliferation of P-element in progeny when mated with M-strain females who have not encountered P-element yet and have no piRNA-based defenses against them (Simmons et al., 2014). With the discovery of piRNA regulation, it was essential to incorporate the now-known regulatory element into existing TE invasion models and thus create a new model to better explain the proliferation of TEs even under active molecular regulation by the host. This new iteration of models is referred to as the "Trap model."

[16]The model by Brookfield (1991) predicted the inheritance of mutated TE proteins or some cytoplasmic defense before the discovery of piRNA. I want to point out that multiple studies during this period postulated the defenses against TEs based on observations in natural populations.

1.2.4 The trap model

Models discussed above inferred the presence of a regulating factor that allows a TE family to maintain a stable copy number[17]. As described in section 1.1.4, we now know that piRNAs are primarily responsible for the defense against TE in the population. The addition of piRNA requires further modification to the models describing TE dynamics, as piRNA clusters introduce a non-trivial layer of regulation. Bergman et al. (2006) hypothesized that specific genomic regions act as "traps" for TEs, which after capturing a moving TE, would regulate all other genomic copies of the same TE family through a co-suppression mechanism that may involve small RNA. This study was published one year before the demonstration that regulating piRNA derived from these high TE density regions acts as a trap and works by using nested TE sequences against TE mRNA (Aravin et al., 2007). Thus, the model incorporating piRNA regulation of TE is colloquially known as the "Trap model" (Bergman et al., 2006). Following are three relevant studies which used the trap model to describe TE dynamics:

- Lu and Clark (2009): Lu and Clark studied the effect of piRNA regulation on class-I TEs in the context of an individual-based model. Specifically, their model altered the transposition rate in the presence and absence of TE insertion in piRNA clusters. Their model referred to TEs, which insert into piRNA cluster and generate piRNA as piRTs (piRNA retrotransposons), and insertions targeted by piRNA generated by piRTs as targetRTs. Furthermore, they also incorporated recombination in their simulations and used the following function for determining the fitness cost of TE insertions:

$$w = e^{-an-\frac{1}{2}bn^2} \tag{1.3}$$

  Where a and b represent are constants representing an RT's insertion cost.

  Their simulations suggested a reduction in the fitness cost of targetRTs in the presence of piRTs. This apparent fitness benefit (relative to no regulation) could be explained by decreased TE copies generated post piRT insertion. Furthermore, another key observation was the positive correlation between the piRT's ability to reduce transposition rate and their frequency in the population, i.e., piRTs, which can regulate targetRTs more aggressively, were selected for in the population(figure 1.6).

[17] in the presence of selection and deleterious mutations in TE

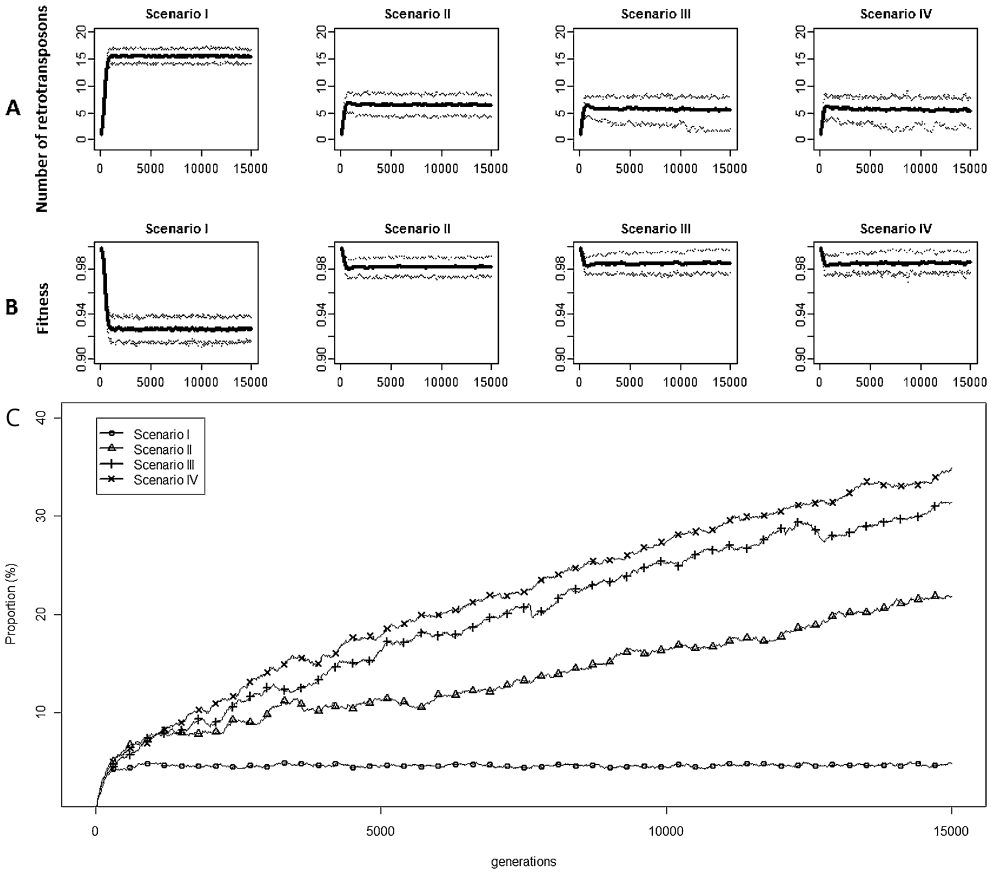


Figure 1.6: From Lu and Clark (2009). x-axis is generation. Panel A: Scenario I corresponds to no effect of piRNA on TE transposition. Scenarios II, III, and IV reduce the TE transposition rate to 10%, 1%, and 0.1%, respectively. Lu and Clark observed a similar effect on RT copy numbers in the population in scenarios II-IV, suggesting that piRNA regulation yielded roughly 60% less transposition than no regulation. Panel B: They also observed an increase in the host's fitness relative to no piRT insertion, suggesting that piRNA reduces the fitness cost of RT invasion. Panel C: Lu and Clark suggested that piRTs were selected for in the simulated population, i.e., they were enriched in the progeny because they conferred an advantage. Their analysis suggested that the piRNA cluster's ability to decrease the transposition of TE is directly proportional to their probability of being selected for in the population.

Lastly, they also observed the fixation of targetRTs, suggesting that this phenomenon resulted from reduced fitness cost due to piRTs. They concluded by stating that regulation by piRNA clusters can be a "trojan horse" due to their ability to shield the harmful effect of TEs, thus, allowing TEs to retain active copies in the genome[18].

[18] Unlike previous models, which incorporated deleterious mutations to TE, thereby reducing the amount of TEs which can actively transpose, Lu and Clark (2009) did not use any parameter analogous to deletion rate. Thus, their conclusion is probably not valid on the time scales

- Kelleher et al. (2018): Kelleher et al. used an extended model to study TE dynamics with an increased parameter space, including additional parameters which were not considered for past TE dynamics models[19]. The model included discrete male and female individuals to study dysgenic sterility and the possibility of ectopic recombination based on TE content. Furthermore, they incorporated insertion sites that reflected an actual genome with more granularity, such as the heterochromatin region (called pseudo-small RNA sites in the study). Their model also reflected a dosage-dependent response to piRNA regulation, i.e., it is imperfect and could require multiple insertions to work effectively. They argued that the model used by Lu and Clark (2009) did not accurately reflect ectopic recombination or hybrid dysgenesis, which are observed in natural populations.

  Their observations reflected a positive trend in the transposition and invasion rates, i.e., a greater $u > 0.1$ yielded a quicker invasion. They also argued that the transposition rate is more important than selection in determining the invasion's success; however, the average TE copy number in the population was still a function of both μ and the fitness cost of the TE (figure 1.7). The TE invasion was relatively unperturbed by the difference in population sizes[20]. However, they observed positive selection for repressor alleles (piRNA sites with TE insertions) in larger population sizes, arguing that the effect of linkage disequilibrium and drift is reduced in such cases. Furthermore, when they accounted for hybrid dysgenesis, they observed a robust selection against TE copies and failure of TE invasion. Only when the effect of hybrid dysgenesis was attenuated were TEs able to invade, albeit with a lower copy number and high frequency of TE insertions in piRNA sites.

  Finally, they concluded that selection for individual piRNA clusters is weaker when there are many of them[21] compared to when there are only a few regulatory sites. They also did not observe the fixation of regulatory alleles in the population, regardless of the population size[22](figure 1.8).

they discuss because their approximations for the dynamic/equilibrium stage cannot hold at the evolutionary time scale. Given enough time, the regulated TEs will mutate enough to either escape regulation or lose their ability to transpose.

[19] i.e., in previous studies the perimeters were explored individually

[20] They simulated two population sizes consisting of 200 and 20000 individuals each

[21] piRNA sites - as a % of the genome - for reference, nearly 3% of the *D. melanogaster* genome is composed of piRNA

[22] Nevertheless, they commented on the repressor alleles arising nearly simultaneously in smaller and larger populations.

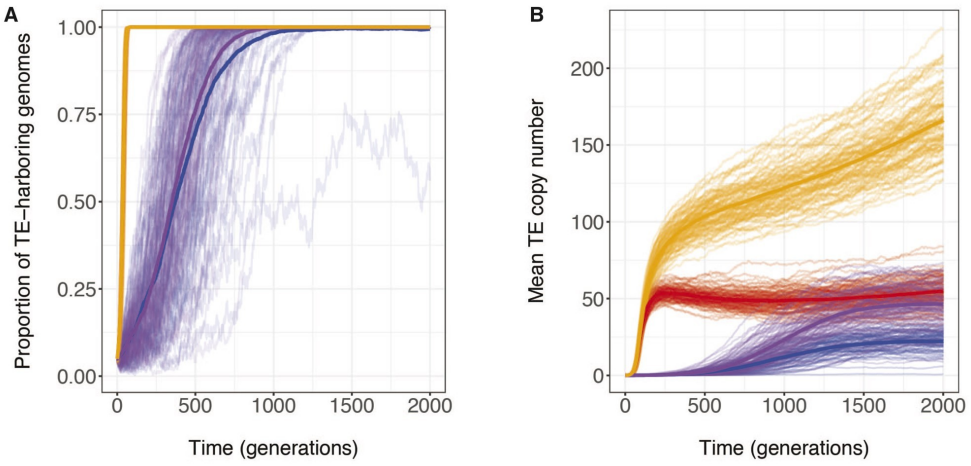


Figure 1.7: From Kelleher et al. (2018). The different colors indicate: $\mu = 0.01, \omega_e$: blue, $\mu = 0.01, \omega_r$: purple, $\mu = 0.1, \omega_e$: red, $\mu = 0.1, \omega_r$: orange. Here ω_r represents fitness model derived from empirical estimates in *D. melanogaster* and ω_e represents a fitness model with 10-fold reduced effect with respect to ω_r.

- Kofler (2019): This study is one of the most relevant studies concerning the course of this thesis. He comprehensively explored the dynamics of TEs in the presence of piRNA in a generalized population consisting of individuals containing diploid genomes with free recombination. Kofler (2019)'s model considered the presence of piRNA clusters, constituting multiple consecutive insertion sites. This abstraction fits well with the biological observation of a piRNA cluster structure, as described in section 1.1.4. Moreover, Kofler (2019)'s model simplified the genomic structure, including the presence of only two types of insertion sites, regulatory (piRNA cluster) and normal site, with no distinctions made for heterochromatin or other genetic features like genes. This simplification allows for a much more flexible exploration of sample space which is not constrained to a single model organism (figure 1.9). However, he did use *D. melanogaster* as a reference organism to set the parameters of the model[23]. Similarly, he simulated a single cluster system to represent flamenco-like piRNA clusters. *flamenco* is a large piRNA cluster in *D. melanogaster*, responsible for piRNA production in somatic follicle cells.

[23] Number of chromosomes, recombination rate, genome size, the proportion of genome as piRNA cluster.

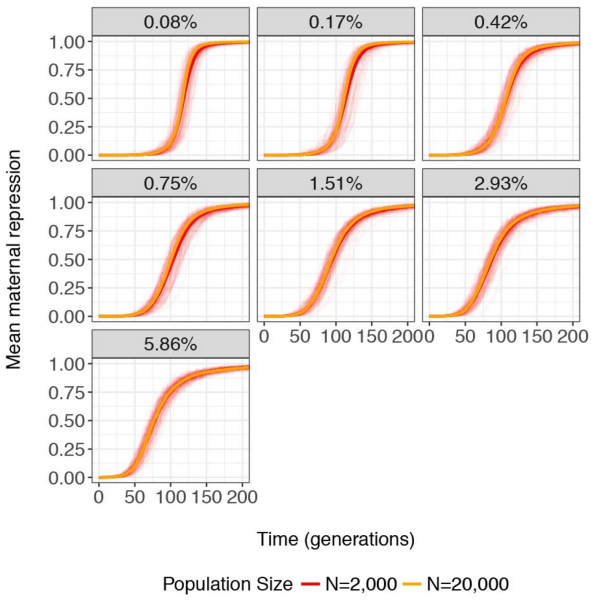


Figure 1.8: From Kelleher et al. (2018). Each panel represents the portion of piRNA sites in the genome. Kelleher et al. (2018) observed a near-simultaneous emergence of piRNA regulation, regardless of piRNA site % in the genome

Kofler (2019)'s results suggested that the regulatory insertions do get fixed, given enough generations. He also noticed that in most simulations, if the population has archived enough cluster insertions, this was enough to restrict the TE invasion, even if there were no fixed clusters. He classified the TE invasion process into three phases to explain this observation:

- Rapid invasion: Not many regulatory TE insertions. The transposition rate determines the duration of this stage.

- Shotgun phase: Cluster insertions are not fixed but found in high proportion (i.e., at least each individual has a single cluster insertion). These segregating clusters then confer immunity to TE invasion. The length of this phase is influenced by population size.

- Inactive: Fixed cluster insertions entirely deactivate the TE in the population.

Kofler (2019) also discussed the selection for piRNA insertions. The study could not find any difference between cluster and non-cluster TE insertions in the neutral model(i.e., TEs do not have negative fitness costs). However, he noted that roughly 4-6 cluster insertions were required in most simulations to arrest a TE invasion. The study also found significant

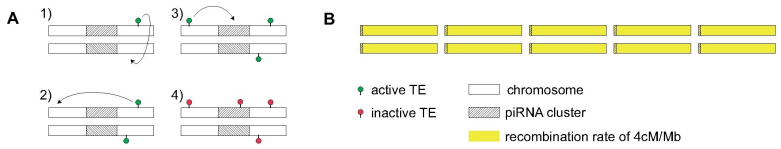


Figure 1.9: From Kofler (2019). The trap model, as desribed by Kofler (2019).

differences between the somatic and germline regulation model, i.e., a single large non-recombining cluster (*flamenco*) against multiple small recombining clusters (in germline). A single large somatic piRNA cluster was more efficient in containing the TE expansion, with fewer TE insertions required. He suggested that recombination might play a role in reducing the effectiveness of piRNA defenses (after observing similar results with recombination enabled in the somatic model).

One of the more significant results of the study was the exploration of invasion space, i.e., identifying which parameter combinations allow a TE to invade. Kofler (2019) observed a stark difference in the ability of a TE to invade a population with just negative selection acting on the TE[24] compared to negative selection acting in conjunction with piRNA defense (figure 1.10). TEs have a very narrow parameter space to invade the population without piRNA regulation; a high u will cause the population to go extinct, and a high negative selection against TE coupled with a low u will cause TE to go extinct. In contrast, piRNA defenses allow much broader flexibility in terms of TE u and negative selection against the TE.

Kofler (2019) termed the TE invasion permissive parameter space "Transposition-Selection-Cluster" (TSC) balance which involves piRNA and negative selection on TEs. He observed that this balance is only possible when both cluster and non-cluster insertions are negatively selected and there is a high transposition rate and sufficient negative selection against TE[25]. However, in the absence of fitness cost on cluster insertions, the TSC balance is not observed; most TEs become inactive due to the spread of regulating clusters in the populations (figure 1.11).

Finally, the study measured the theoretical prediction of TE distribution and observed TE distribution in *D. melanogaster*. While the TE copy number predictions matched with TE insertions in germline piRNA clusters, they were discordant with TE insertions found in somatic piRNA cluster(TEs found in *flamenco* piRNA cluster). Kofler (2019) observed

[24] $w = 1 - xn$, where x is the fitness cost
[25] $N \times u > 1$

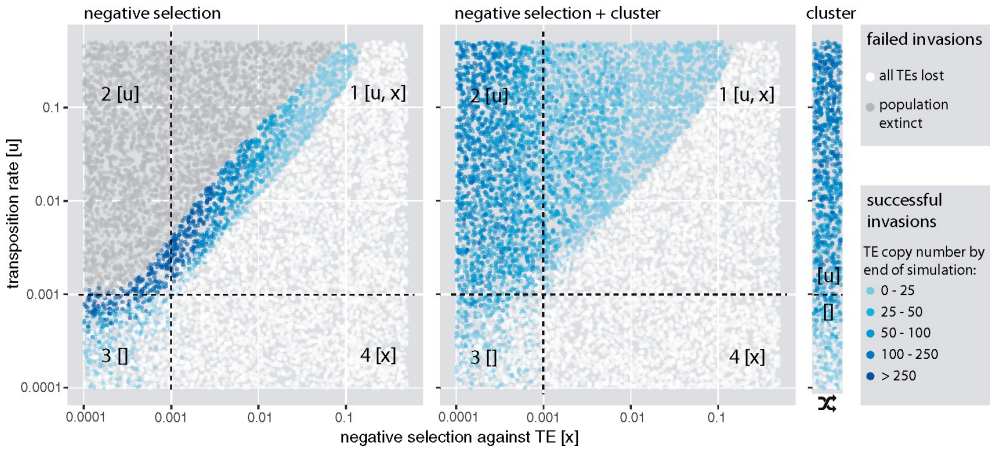


Figure 1.10: From Kofler (2019). Robert Kofler analysed the parameter space for the invasion, in this case the interaction between negative selection against TE and transposition rate. piRNA mechanism greatly expands the TE ability to invade the population, even under low negative selection against TE, thus rescuing the population from extinction.

substantially fewer somatic TEs in the wild population compared to the expected TE copy number derived from simulations(<100 vs. >500). Kofler (2019) hypothesized that this might be due to the preferential insertion of TEs active in somatic tissue in *flamenco* piRNA cluster or due to cryptic recombination.

Robert Kofler's model is an important landmark in the dynamics of TE invasion in the presence of piRNA regulation; many more questions and contradictions need to be answered, including questions regarding the distribution of piRNA clusters in genomes, the question of the possibility of horizontal TE transfer in the presence of piRNA defences and more.

1.3 Objectives of the thesis

While Kofler's simulations have shed light on the working of the trap model and the parameter space that enables a TE invasion, we still need a mathematical analysis of the trap model that can explain TE invasion's dynamics under regulation by piRNA clusters. Therefore, the thesis aims to understand what kind of TE copy number equilibrium exists when genomic TE and cluster TE insertions are either neutral or deleterious and how the distribution of piRNA clusters in a genome can affect TE dynamics.

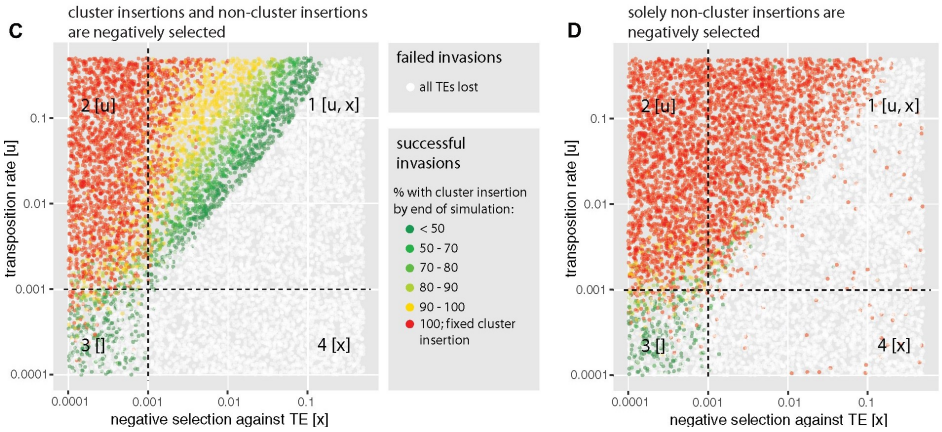


Figure 1.11: From Kofler (2019). Figure describing the Transposition-Selection-Cluster (TSC) balance: The yellow and green dots represent populations where TEs can still invade. D: In the absence of negative selection against cluster insertions, no TSC balance is observed, and the cluster insertions get fixed.

Another significant aspect of the study of TEs is their frequent horizontal transfers between species and multiple bursts of transposition in the same species. While we have some insights into the dynamics of a single TE family invasion in a population, the literature lacks the study of multiple simultaneous TE family invasions contained by piRNA machinery. These simultaneous invasions can represent repeated horizontal transfer in the same species by similar or completely different TE families or re-activation of the same TE family in the population after diverging from the parent copy. We expect the trap model to provide a memory of past invasions and possible protection against new invasions from the same or similar TE families. However, this hypothesis requires further investigation. Thus the thesis aims to analyse the phenomenon of horizontal transfer and TE re-activation and further study TE (re)invasion dynamics while under regulation by trap model, and asses the possibility of the presence of genomic immunity, i.e., the capability of the genome to prevent a previously encountered TE from invading again.

Model and simulation framework 2

The primary function of the TE is to amplify within the genome, and by extension, in the population. However, depending on the host and the population characteristics, factors like selection, drift, mutations, and regulation can influence the ability of TE to invade the population. The purpose of our individual based model is to describe the spread of TE under constraints: defense mechanisms based on trap model (piRNA), selection, and the presence of other TE families. The software tracks and traces the path taken by the TE in the population, indicating the ability of a TE to invade a population successfully.

2.1 The individual-based model

2.1.1 Parameters and mechanisms

Fitness defines an individual's success in spreading its genotype to the next generation (Crow and Kimura, 2009). In our model, any increase in the copies of a particular TE family decreases the fitness of the individual carrying those TE copies. Considering that all TEs have an equivalent cost to the host, which allows us to calculate the fitness (w) of the host as:

$$w_i = e^{\Sigma_j s_j} \tag{2.1}$$

where s_j is the deleterious effect ($s_j < 0$) of TE insertion j on fitness (w) of individual i. If all TEs have the same effect on fitness, then:

$$w_i = e^{(s \cdot n_i)} \tag{2.2}$$

where n_i is the number of TEs present in the host i genome and s is the fitness cost ($s < 0$) per TE insertion. Natural selection will favor individuals with higher fitness, i.e., a higher chance of passing their genotype to the next generation. We can rephrase this as the probability, based on their fitness, that an individual from the current population will be "selected" for reproduction. For a population undergoing a TE invasion, individuals with fewer TE copies will be selected over individuals with high TE content.

Epistasis is the mechanism where the effect of an allele can affect another allele. In the simulation model, we added the possibility to include a coefficient ε quantifying directional epistasis on fitness (equation 2.3). When $\varepsilon = 0$, the fitness function collapses to equation 2.2; the situation in which $\varepsilon < 0$ (negative epistasis on fitness) is expected to decrease the spread of TEs as each new insertion is more deleterious than the previous one.

The fitness calculation under epistasis is done by modifying the equation 2.2:

$$w_i = e^{(s \cdot n_i) + (\frac{1}{2} \cdot n_i^2 \cdot s^2 \cdot \epsilon)} \tag{2.3}$$

Genetic drift is the random shift in the frequency of an allele in the population. Thus, drift will cause an allele to become fixed or be eliminated from the population depending on the characteristics of the population, like its effective population size (N_e); in our model populations are finite-sized and we thus account for genetic drift

TE regulation is the mechanism where TE insertions lose their ability to transpose due to host defences. In the piRNA "trap model", specific genomic loci that are known as piRNA clusters act as traps for TE sequence to jump in and, as a result, generate complementary sequences which target TE transcripts and attenuate TE activity. In the past exploration of the model, a single insertion in one of the piRNA clusters can render a TE family inactive, i.e., reduce its transposition rate (u) to 0 (Kofler, 2019). In our model, the influence of each TE insertion in the piRNA cluster is determined by the parameter τ. In our model, the transposition rate (u_i) for a TE family in individual i with a copy in a piRNA cluster (cluster insertion) is calculated using the following equation:

$$u_i = \begin{cases} u(1 - \tau \cdot m) & \text{if } u(1 - \tau \cdot m) \geq 0 \\ 0, & \textit{otherwise} \end{cases} \tag{2.4}$$

where m is the number of TE insertions in all piRNA clusters and u is the transposition rate of TE family without any cluster insertions.

The model is designed to work with multiple TE families in a population, each with a different transposition rate. Furthermore, it implements cross-regulation, where cluster insertion of one TE family regulates other TE families. To study TE dynamics under cross-regulation, the model incorporates the parameter η. In the presence of two TE families, α and β, the effective transposition rate of TE family α under the influence of cluster insertions of TE family β can be calculated using the following equation:

$$u^{\alpha}_{\pi_i} = \begin{cases} u^{\alpha}_i - (\tau \cdot m_\beta \cdot \eta) & \text{if } u^{\alpha} - (\tau \cdot m_\beta \cdot \eta) \geq 0 \\ 0 & otherwise \end{cases} \tag{2.5}$$

The equation is symmetrical for calculating $u^{\beta}_{\pi_i}$.

Recombination is the mechanism that involves the reshuffling of genetic information. The model can be parameterized with a recombination rate r_i between each consecutive loci; however, to simplify the simulations, we considered the recombination rate constant for all loci (insertion sites), except for loci which delimit chromosomes ($r_i = 0.5$).

2.1.2 The genome and population structure

The model simulates a diploid genome with finite and fixed discrete insertion sites (T). Each insertion site represents an empty region where a TE can jump. A prespecified fraction (π) of the genome is designated as the piRNA cluster(s). Any TE insertion into the piRNA cluster will attenuate the transposition rate of all copies from the same family and all TE copies if cross-regulation is enabled ($\eta > 0$). Genomes are split into k chromosomes of equal length and are demarcated by setting the recombination rate to 0.5 between two loci belonging to different chromosomes. Figure 2.1 illustrates the regulation mechanism and genome structure.

The population consists of N hermaphrodite diploid individuals. Each generation is reproductively isolated, and the individuals do not overlap between generations. Each individual carries information about the TE insertion in its

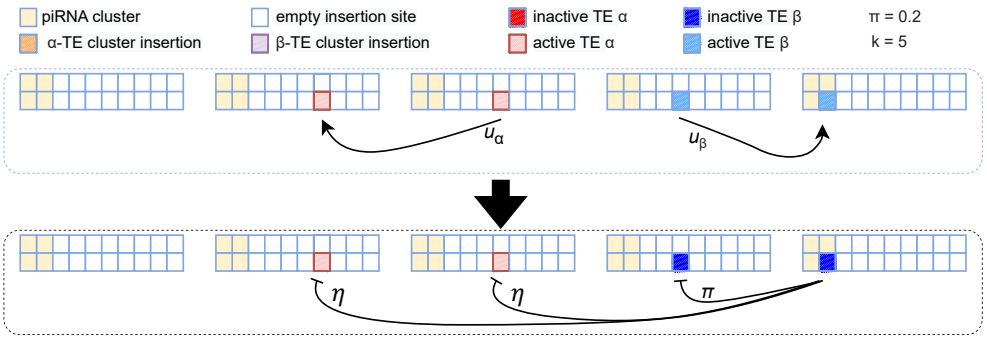


Figure 2.1: Representation of an individual's genome in the model and the mechanism of regulation and co-regulation.

genome. To minimize TE loss due to drift in the small population size, TEs are distributed to a fraction of the population, ensuring that some TEs make it into the next generation. Figure 2.2 illustrates the data-structures used for storing TE, population, and genome information.

2.1.3 Simulation steps and design

- Initialization is done by generating a population of N individuals. The parameter $n_{\alpha 0}$ defines the average number of TE copies (for TE family α) per individual at the start, and these copies are distributed into randomly selected insertion sites in the genome. This step is performed only at the start of the simulation.

- Horizontal TE transfer - The second TE family is introduced into the population at generation H. The parameter n_{β}^{0} defines the average number of TE copies (for TE family β) per individual during horizontal TE transfer. Similar to TE family α, the TE copies of β are distributed randomly in the genome.

- Reproduction - Two individuals are picked with a probability proportional to their fitness (described by equations 2.2 and 2.3). These two individuals produce a haploid gamete after meiosis (recombination). The resulting gametes are fused to form a new individual for the next generation. This process is repeated N times so that the population size N is constant.

- Transposition - The TEs in the newly formed individual undergo transposition. Each copy of the TE family α (and β) has a probability of transposing based on their transposition rates (u^{α} and u^{β}). In the presence of regulation and co-regulation, the transposition rate is calculated using equations 2.4 and 2.5, respectively. For each TE copy, transposition will occur if $u_{\pi_i}^{x} < X \sim Uniform[0,1]$. Due to the implementation, each TE copy can replicate only once per generation per individual.

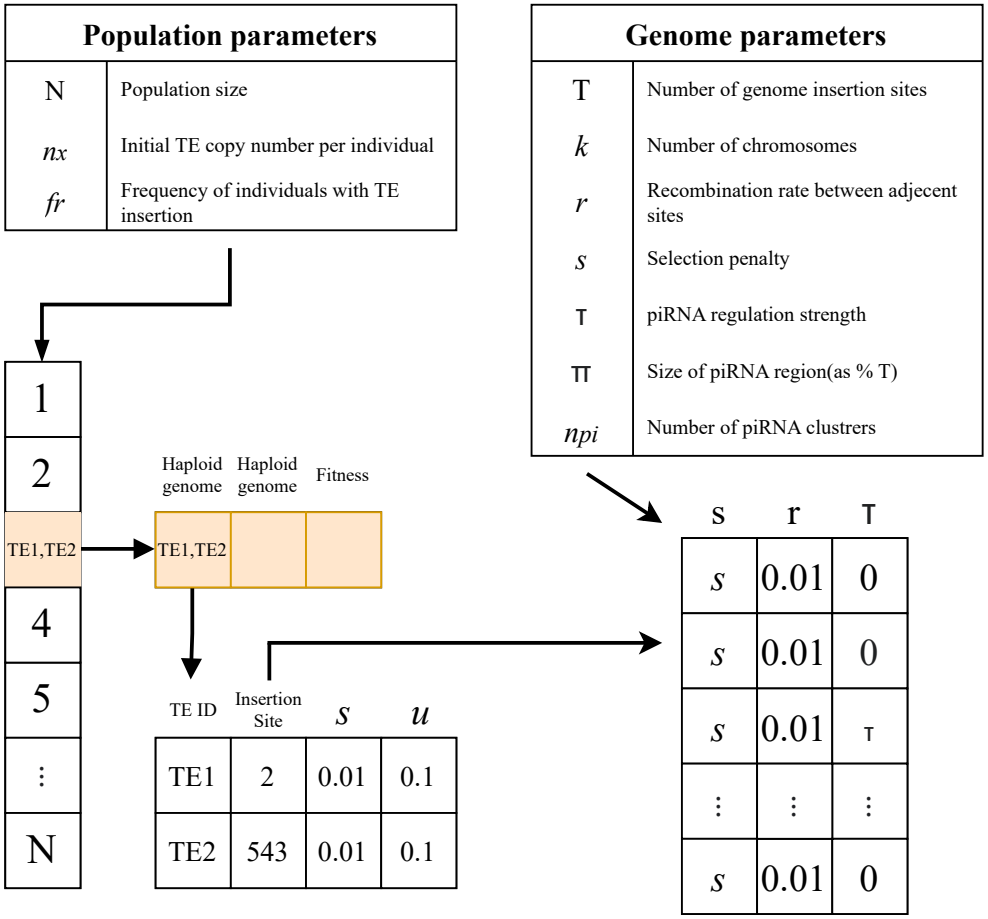


Figure 2.2: Each parameter file initializes a data structure based on the supplied arguments. Matrices are used to store information about TEs, genomes, and individuals. Each individual carries information about the TE content present in its genome. This information is then cross-referenced with the matrix, which stores information about that TE type, including insertion sites. Additional details on insertion sites, such as selection penalty and regulation strength, are cross-referenced from the genome matrix.

- Calculation of summary statistics - Summary statistics are calculated for each generation; however, these are reported at the end of the simulation run. These statistics represent the state of the TE invasion and are calculated using information from the diploid genome of each individual. Following are all the statistics generated by a simulation run:

  - Average copy number of all TEs present in the population per generation
  - Variance in copy number of all TEs present in the population per generation
  - Average copy number of specific TE family per generation
  - Variance number of specific TE family per generation
  - Average piRNA regulation on specific TE family per generation
  - Average fitness of the population per generation

  Figure 2.3 illustrates all the steps taken during a simulation run in form of a flowchart. Table 2.1 references all the parameters and their default values for the model (modeled after *D. melanogaster*).

Table 2.1: Parameters for the model

Parameter	Meaning	Default value
N	Population Size	2000
T	Number of genomic insertion sites	1000
k	Number of chromosomes	5
r	Recombination rate between adjacent insertion sites	0.1%
H	Horizontal transfer generation	
π	size of piRNA region (as % genome)	3%
τ	piRNA Regulation strength	1.0
η	Cross-regulation coefficient	0
n_{x0}	Initial number of copies per individual for TE family x	0.2
u^x	Maximum transposition rate for TE family x	0.05
s	Selection penalty	-0.005
ϵ	Epistasis coefficient on fitness	0

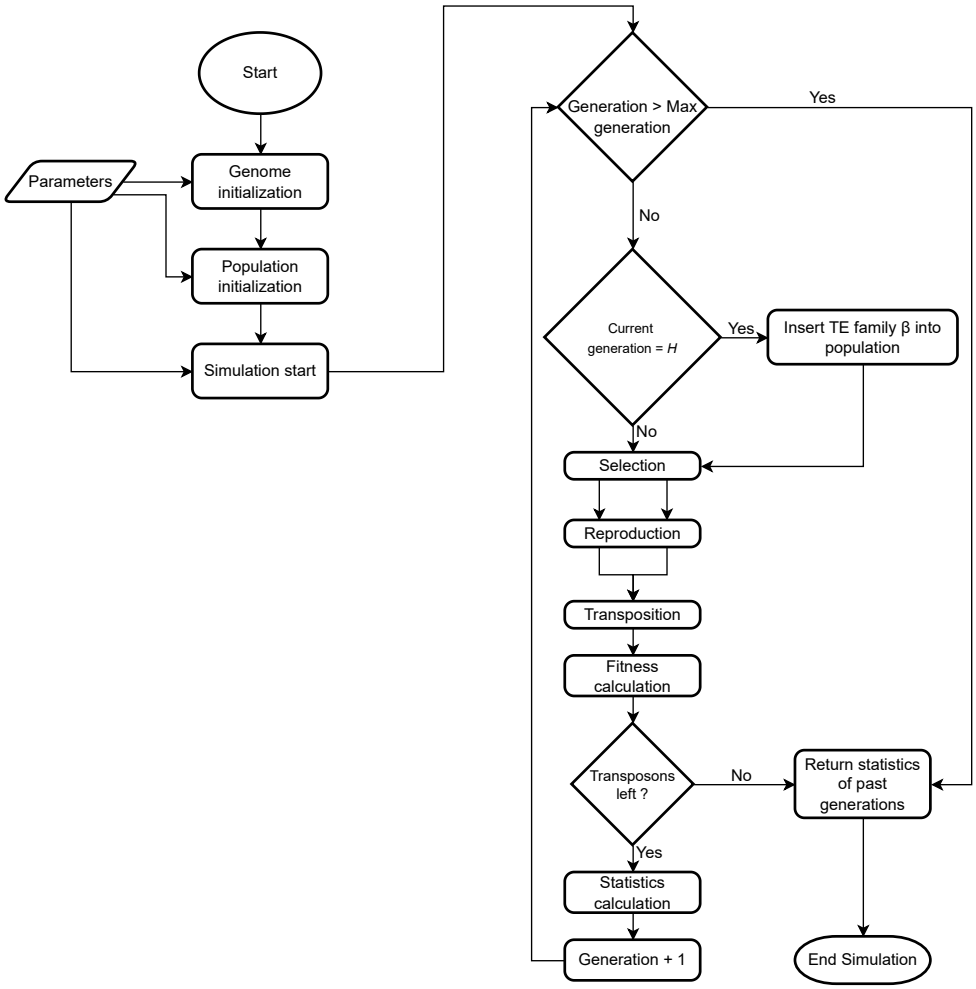


Figure 2.3: Flowchart explaining simulation steps.

2.1.4 Implementation

Language The model is implemented entirely in Python 3.6 as a library. Numpy (version 1.23.5) (Harris et al., 2020) matrices are used as a data structure for storing information about the population and genome. All other data structures (lists, sets, and dictionaries) are native to standard Python installation. Using native Python and Numpy creates a minimal requirement, with only Python and Numpy libraries required to run the simulations.

Numpy The decision to use Numpy was influenced by the speed and ease of using Numpy matrix containers. Not only do they allow mixed datatypes in a matrix, but it is also trivial to use selection and slicing operations in a matrix. Numpy also offers functions for random number generation and selection based on weights. These tools are essential as they simplify multiple steps in the matrix-based implementation of mode, including recombination, selection, and transposition, all requiring slicing, random number generation, and matrix manipulations. However, it is entirely possible to have an object-oriented implementation of the model where a class represents each individual.

Pickle The simulation results are stored as a pickle file containing a dictionary with various reporting parameters. Pickle files are serialized Python objects - and for the simulation results, they include a dictionary with keys corresponding to the required statistics (e.g., average TE copy number). This format simplifies the exploration of simulation results in Python itself, and the files can be loaded and worked on interactively in Jupyter Notebook.

Implementation The simulation steps are implemented as library submodules (functions). Since each function is isolated from the other, users can modify them easily as long as they return the required output. The main library (`popSim`) glues everything together and acts as a driver of the simulation. Users will primarily interact with the primary function (`runSim`) to run the simulations.

Running a simulation Initialisation of simulations and running them are two different processes. The function `generateGenome` generates the immutable genome, which acts as a template for mutable genomes between all individuals in a population. The output of this function is then passed to the `generatePopulation` function, which initializes the population array. After this step, the output from `generateGenome` and `generatePopulation` are passed to `runSim` to execute the primary sim-

ulation process. If users choose to, the user can modify `generateGenome` and `generatePopulation` to create custom populations using their functions as long as the returning matrices respect the data structure defined in the model. The code is executed in a Python script, and parameters are passed using a parameter file - the simulation library is not distributed as a binary.

Multithreading Each simulation is independent of other simulations. Users can implement multiprocessing by either using the inbuilt python multiprocessing library or using tools like GNU Parallel (Tange, 2018) to run the python script.

2.1.5 Code availability and license

The simulation software (Simulicron) code is available on a GitHub public repository and can be accessed via the following URL: https://github.com/siddharth-st/Simulicron

As a proponent of open science, I believe that not only the software code and the scripts used to generate results should be openly available but also that the code should be reusable and modifiable without restrictions; hence I have decided to use the CeCILL license version 2.1, designed by CNRS, INRIA and CEA. This license is approved by OSI (Open Souce Initiative) as an open-source license and is compatible with GNU General Public License. Furthermore, since not all journals maintain open-access policies, all studies conducted during this thesis are either available or will be made available as preprints on bioRxiv.

3 Article 1

The first study of this thesis deals with population genetics models and population scale simulations of TE dynamics under the trap model and how different selective pressure influences the TE equilibrium state. We further contrast observations with past population genetics models and highlight the stochastic nature of TE dynamics under the trap model. This study is available as a preprint on bioRxiv and can be accessed using the following URL: https://-doi.org/10.1101/2022.07.05.498868

The study has been published in the journal "Theoretical Population Biology", and can be accessed using the following URL: https://-doi.org/10.1016/j.tpb.2023.02.001

Publication and journal information:

Siddharth S. Tomar, Aurélie Hua-Van, Arnaud Le Rouzic, A population genetics theory for piRNA-regulated transposable elements, Theoretical Population Biology, Volume 150, 2023, Pages 1-13, ISSN 0040-5809

Theoretical Population Biology 150 (2023) 1–13

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

A population genetics theory for piRNA-regulated transposable elements

Siddharth S. Tomar, Aurélie Hua-Van, Arnaud Le Rouzic *

Université Paris-Saclay, CNRS, IRD, UMR EGCE, 12 Route 128, Gif-sur-Yvette, 91190, France

ARTICLE INFO

Article history:
Received 2 September 2022
Available online 1 March 2023

Keywords:
Transposition regulation
Model
Simulations
piRNA clusters
Trap model

ABSTRACT

Transposable elements (TEs) are self-reproducing selfish DNA sequences that can invade the genome of virtually all living species. Population genetics models have shown that TE copy numbers generally reach a limit, either because the transposition rate decreases with the number of copies (transposition regulation) or because TE copies are deleterious, and thus purged by natural selection. Yet, recent empirical discoveries suggest that TE regulation may mostly rely on piRNAs, which require a specific mutational event (the insertion of a TE copy in a piRNA cluster) to be activated — the so-called TE regulation "trap model". We derived new population genetics models accounting for this trap mechanism, and showed that the resulting equilibria differ substantially from previous expectations based on a transposition–selection equilibrium. We proposed three sub-models, depending on whether or not genomic TE copies and piRNA cluster TE copies are selectively neutral or deleterious, and we provide analytical expressions for maximum and equilibrium copy numbers, as well as cluster frequencies for all of them. In the full neutral model, the equilibrium is achieved when transposition is completely silenced, and this equilibrium does not depend on the transposition rate. When genomic TE copies are deleterious but not cluster TE copies, no long-term equilibrium is possible, and active TEs are eventually eliminated after an active incomplete invasion stage. When all TE copies are deleterious, a transposition–selection equilibrium exists, but the invasion dynamics is not monotonic, and the copy number peaks before decreasing. Mathematical predictions were in good agreement with numerical simulations, except when genetic drift and/or linkage disequilibrium dominates. Overall, the trap-model dynamics appeared to be substantially more stochastic and less repeatable than traditional regulation models.

1. Introduction

Transposable elements (TEs) are repeated sequences that accumulate in genomes and often constitute a substantial part of eukaryotic DNA. According to the canonical "TE life cycle" model (Kidwell and Lisch, 2001; Wallau et al., 2016), TE families are not maintained actively for a long time in genomes. TEs are the most active upon their arrival in a new genome (often involving a horizontal transfer, Gilbert and Feschotte, 2018); their copy number increases up to a maximum, at which point transposition slows down. TE sequences are then progressively degraded and fragmented, accumulate substitutions, insertions, and deletions, up to being undetectable and not identifiable as such. The reasons why the total TE content, the TE families, and the number of copies per family vary substantially in the tree of life, even among close species, are far from being well-understood, which raises interesting challenges in comparative genomics.

TEs spread in genomes by replicative transposition, which ensures both the genomic increase in copy number and the invasion of populations across generations of sexual reproduction. They are often cited as a typical example of selfish DNA sequences, as they can spread without bringing any selective advantage to the host species, and could even be deleterious (Orgel and Crick, 1980; Doolittle and Sapienza, 1980). Even if an exponential amplification of a TE family could, in theory, lead to species extinction (Brookfield and Badge, 1997; Arkhipova and Meselson, 2005), empirical evidence rather suggests that TE invasion generally stops due to several (non-exclusive) physiological or evolutionary mechanisms, including selection, mutation, and regulation. Selection limits the TE spread whenever TE sequences are deleterious for the host species: individuals carrying fewer TE copies will be favored by natural selection, and will thus reproduce preferentially, which tends to decrease the number of TE copies at the next generation (Charlesworth and Charlesworth, 1983; Lee, 2022). The effect of mutations relies on the degradation of the

* Corresponding author.
E-mail addresses: siddharth.tomar@universite-paris-saclay.fr (S.S. Tomar), aurelie.hua-van@universite-paris-saclay.fr (A. Hua-Van), arnaud.le-rouzic@universite-paris-saclay.fr (A. Le Rouzic).

https://doi.org/10.1016/j.tpb.2023.02.001
0040-5809/

protein-coding sequence of TEs, which decreases the amount of functional transposition machinery, and thus the transposition rate (Le Rouzic and Capy, 2006). Even though TEs can be inactivated by regular genomic mutations, as any other DNA sequences, there exist documented mutational mechanisms that specifically target repeated sequences, such as repeat induced point mutations in fungi (Selker and Stevens, 1985; Gladyshev, 2017). Alternatively, substitutions or internal deletions in TEs could generate non-autonomous elements, able to use the transposition machinery without producing it, decreasing the transposition rate of autonomous copies (Hartl et al., 1992; Robillard et al., 2016).

Transposition regulation refers to any mechanism involved in the control of the transposition rate by the TE itself or by the host. There is a wide diversity of known transposition regulation mechanisms; some prevent epigenetically the transcription of the TE genes (Deniz et al., 2019), others target the TE transcripts (Adams et al., 1997), or act at the protein level (Lohe and Hartl, 1996). Recently, the discovery of piRNA regulation systems have considerably improved and clarified our understanding of TE regulation (Brennecke et al., 2007; Malone and Hannon, 2009; Zanni et al., 2013; Ozata et al., 2019). piRNA regulation seems to concern a wide range of metazoan species (Huang et al., 2021), and acts on TE expression through a series of complex mechanisms, which can be summarized by a simplified regulation scenario known as the "trap model" (Bergman et al., 2006; Kofler, 2019). In such a scenario, regulation is triggered by the insertion of a TE in specific "trap" regions of the genome, the piRNA clusters. TE sequences inserted in piRNA clusters (thereafter TE cluster insertions) are transcribed into small regulating piRNAs, that are able to silence homologous mRNAs from the same TE family by recruiting proteins from the PIWI pathway.

Early models, starting from Charlesworth and Charlesworth (1983), assumed that the strength of regulation increases with the copy number. The transposition rate is then expected to drop progressively in the course of the TE invasion up to the point where transposition stops. In contrast, the PIWI regulation pathway displays unique features that may affect substantially the evolutionary dynamics of TE families: (i) it relies on a mutation-based mechanism, involving regulatory loci that may need several generations to appear (ii) the regulatory loci in the host genome segregate independently from the TE families and have their own evolutionary dynamics (the TE amplifies in a genetically-variable population, which is a mixture of permissive and repressive genetic backgrounds), and (iii) the regulation mechanism may not be strongly dependent on genomic copy number. The consequences of these unique features on the TE invasion dynamics are not totally clear yet. Individual-based stochastic simulations have shown that piRNA regulation is indeed capable of allowing a limited spread of TEs, compatible with the TE content of real genomes (Lu and Clark, 2010; Kelleher et al., 2018). Kofler (2019) has shown, by simulation, that a major factor conditioning the TE success (in terms of copy number) was the size of the piRNA clusters, while the influence of the transposition rate was reduced. The dynamics of transposable elements when regulated by a trap model thus appear to differ substantially from the predictions of the traditional population genetics models.

With this paper, we extend the existing corpus of TE population genetics models by proposing a series of mathematical approximations for the dynamics of TE copy number when regulated by piRNA in a "trap model" setting. We studied three scenarios, differing by whether or not TE copies induce a fitness cost when inserted in piRNA clusters and/or in other genomic locations: Scenario (i) neutral TEs, (ii) deleterious TEs and neutral TE cluster insertions, and (iii) deleterious TEs and deleterious TE cluster insertions. We showed that an equilibrium copy number could be achieved in scenarios (i) and (iii), while the TE family decays (after having invaded) in scenario (ii). We confirmed that the transposition rate does not condition the equilibrium number of copies in the neutral model (i), but it does when TEs are deleterious. Model (iii), in which both genomic and TE cluster insertions are deleterious, lead to a complex transposition–selection–regulation equilibrium (the "transposition–selection cluster" balance in Kofler (2019)), in which regulatory alleles do not reach fixation and stabilize at an intermediate frequency. The robustness of these predictions was validated by numerical simulations.

2. Models and methods

2.1. Population genetic framework

Model setting and notation traces back to Charlesworth and Charlesworth (1983), who proposed to track the mean TE copy number $\bar{n}$ in a population through the difference equation:

$$\bar{n}_{t+1} = \bar{n}_t + \bar{n}_t(u - v), \tag{1}$$

where u is the transposition rate (more exactly, the amplification rate per copy and per generation), and v the deletion rate. In this neutral model, if u and v are constant, the copy number dynamics is exponential. If the transposition rate u_n is regulated by the copy number ($u_0 > v$, $\mathrm{d}u_n/\mathrm{d}n < 0$, and $\lim(u_n) < v$), then a stable equilibrium copy number $\hat{n}$ can be reached.

However, in most organisms, TEs are probably not neutral. If TEs are deleterious, fitness w decreases with the copy number ($w_n < w_0$). As a consequence, individuals carrying more copies reproduce less, which decreases the average copy number every generation. The effect of selection can be accounted for using traditional quantitative genetics, considering the number of copies n as a quantitative trait: $\Delta\bar{n} \simeq \mathrm{Var}(n)\partial \log(w_n)/\partial n$, where $\mathrm{Var}(n)$ is the variance in copy number in the population, and $\partial \log(w_n)/\partial n$ approximates the selection gradient on n. The approximation is better when the fitness function w_n is smooth and the copy number n is not close to 0. Assuming random mating and no linkage disequilibrium, n is approximately Poisson-distributed in the population, and $\mathrm{Var}(n) \simeq \bar{n}$.

Charlesworth and Charlesworth (1983) proposed to combine the effects of transposition and selection to approximate the variation in copy number among generations t and $t+1$ as:

$$\bar{n}_{t+1} \simeq \bar{n}_t + \bar{n}_t(u_{\bar{n}_t} - v) + \bar{n}_t s_{\bar{n}_t}, \tag{2}$$

where $s_{\bar{n}_t} = \partial \log w_n/\partial n|_{\bar{n}_t}$.

When the transposition rate is high, the Poisson approximation does not hold and $\mathrm{Var}(n) > \bar{n}$: transposition overdisperses the copies in the population, as new TEs tend to appear in TE-rich genomes; random mating only halves this bias every generation but does not cancel it if transposition persists over generations. After transposition, the copy number rises to $\bar{n} = \bar{n}(1+u)$, while its variance becomes $\mathrm{V}(n) = \bar{n}(1+u)^2$; the drop in copy number due to selection thus becomes $\bar{n}(1+u)^2 s_{\bar{n}}$ instead if $\bar{n}s_{\bar{n}}$. In order to match our simulation algorithm described below, in which selection takes place after transposition, we accounted for linkage disequilibrium and replaced the selection coefficient $s_{\bar{n}}$ by $s'_{\bar{n}} = s_{\bar{n}}(1+2u)$ (neglecting u^2 in $(1+u)^2$). This correction remains an approximation, as the effect of linkage disequilibrium of TE copies is more subtle and complex (Roze, 2022). Overall, linkage disequilibrium due to transposition (slightly) amplifies the selection penalty for high transposition rates, and tends to decrease the genomic copy number.

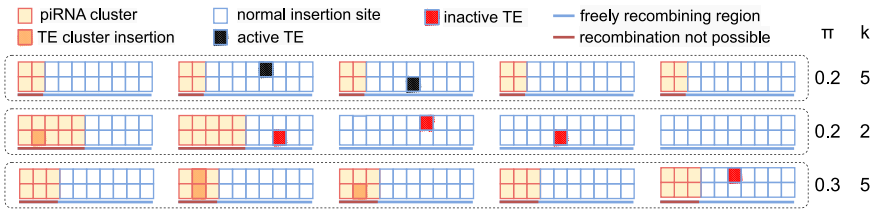


Fig. 1. Schematic representation of the genomic model in the simulations. The K possible insertion sites are equally spread on chromosomes; the k piRNA clusters are distributed at one of the chromosome tips, representing a proportion π of the genome. Recombination is suppressed within clusters, but is possible in non-cluster genomic regions. In transposition-permissive genomes (no TE insertion in piRNA clusters), TEs located in normal insertion sites can transpose in any random location with a transposition rate u. TEs located in clusters cannot transpose, and prevent the transposition of all other elements in the genome. In order to ensure the generality of the results, simulations were set up to minimize genetic linkage: recombination between normal sites was free (recombination rate $r = 0.5$), and the genome had 30 chromosomes (always larger than the number of clusters). In the default conditions, 300 out of $K = 10{,}000$ insertion sites are piRNA clusters ($\pi = 0.03$, close to the estimated proportion in the genome of *Drosophila melanogaster*).

2.2. Numerical methods

Data analysis was performed with R version 4.0 (R. Core Team, 2020). Mathematical model analysis involved packages deSolve (Soetaert et al., 2010) and phaseR (Grayling, 2014). All figures and analyses can be reproduced from the scripts available at https://github.com/lerouzic/amodelTE.

Mathematical predictions were validated by individual-based simulations. Populations consisted in $N = 1000$ hermaphroditic diploid individuals, with an explicit genome of 30 chromosomes and a total of $K = 10{,}000$ possible TE insertion sites (Fig. 1). k piRNA clusters of size $K\pi/k$ were distributed on different chromosomes, the parameter π standing for the proportion of the K loci corresponding to piRNA clusters. Insertion sites were freely recombining, except within piRNA clusters. Generations were non-overlapping; reproduction consisted in generating and pairing randomly $2N$ haploid gametes from $2N$ parents sampled with replacement, with a probability proportional to their fitness. Transposition occurred with a rate u_i computed for each individual as a function of its genotype at piRNA clusters ($u_i = u$ if no TE insertion in clusters, $u_i = 0$ otherwise). The location of the transposed copy was drawn uniformly in the diploid genome. Transposition events in occupied loci were canceled, which happened rarely as TE genome contents were always far from saturation ($K \gg n$). Populations were initialized with 10 heterozygote insertions (in non-piRNA loci), randomly distributed in the population at frequency 0.05 each, resulting in $n_0 = 1$ copy on average per diploid individual. For each parameter set, simulations were replicated 10 times, and the average number of diploid TE copies was reported. Average allele frequencies in clusters were calculated by dividing the number of diploid TE cluster insertions by $2k$. The possibility to have two TE insertions in the same cluster was discarded, as such a situation requires two simultaneous transpositions in the same cluster: transposition is repressed in presence of TE cluster insertion, and there is no within-cluster recombination. We did not distinguish different regulatory alleles at the same cluster (i.e., TEs inserted in different sites within the cluster), as those are functionally equivalent. The simulation software was implemented in python (version 3.8.10 for Linux), with data structures from the numpy library (Harris et al., 2020). The code is available at https://github.com/siddharthst/Simulicron/tree/amodel.

3. Results

3.1. Neutral trap model

The model assumes k identical piRNA clusters in the genome, and the total probability to transpose in cluster regions is π. Each cluster locus can harbor two alleles: a regulatory allele (i.e., the cluster carries a TE insertion), which segregates at frequency p, and an "empty" allele (frequency $1 - p$). When all clusters are identical, and neglecting genetic drift (infinite population size), regulatory (TE cluster insertion) allele frequencies at all clusters are expected to be the same (p); at generation t, the average number of TE cluster insertions for a diploid individual is $m_t = 2kp_t$. TE deletions were neglected ($v = 0$). The model posits that the presence of a single regulatory allele at any cluster locus triggers complete regulation: the transposition rate per copy and per generation was u in "permissive" genotypes (frequency $(1-p)^{2k}$ in the population), and 0 in regulated genotypes (frequency $1 - (1-p)^{2k}$). New regulatory alleles appear when a TE transposes in a piRNA cluster (with probability π), which is only possible in permissive genotypes. Assuming random mating and no linkage disequilibrium between clusters and the rest of the genome ($\mathrm{Cov}(n_t, p_t) = 0$, i.e. no correlation between n and the genotype at the regulatory clusters), we approximated the discrete generation model with a continuous process, and the neutral model (Eq. (2)) was rewritten as a set of two differential equations on $\bar{n}$ (relabeled n for readability) and p:

$$\begin{aligned} \frac{\mathrm{d}n}{\mathrm{d}t} &= nu(1-p)^{2k} \\ \frac{\mathrm{d}p}{\mathrm{d}t} &= \frac{\pi}{2k} nu(1-p)^{2k}. \end{aligned} \tag{3}$$

Here, n stands for non-cluster TE copies; for simplicity, Eq. (3) assumes that $\pi \ll 1$, so that $1 - \pi \simeq 1$ (a more precise version of Eq. (3) for large values of π is provided in Appendix A.1). Expressing the variation of the number of TE cluster insertions $m = 2kp$ as $\mathrm{d}m/\mathrm{d}t = \pi \mathrm{d}n/\mathrm{d}t$ would be more straightforward here, we kept the $\mathrm{d}p/\mathrm{d}t$ formulation solely for consistency with more complex models described below.

The initial state of the system is $n_0 > 0$ copies per individual in the population, and no piRNA cluster insertions ($p_0 = 0$). The system of Eq. (3) admits three equilibria (characterized by the equilibrium values $\hat{n}$ and $\hat{p}$): $E_1 : \hat{n} = n_0$ and $\hat{p} = 0$ (no transposition, achieved when $u = 0$), $E_2 : \hat{n} = 0$ and $\hat{p} = 0$ (no transposable element, $n_0 = 0$), and $E_3 : \hat{p} = 1$ (fixation of all TE cluster insertions). Equilibria E_1 and E_2 do not need to be investigated further, as $u = 0$ or $n_0 = 0$ do trivially result in the absence of TE invasion. Equilibrium E_3 is analytically tractable, as $\mathrm{d}n/\mathrm{d}p = 2k/\pi$, and $n = n_0 + 2pk/\pi$ at any point of time:

$$\begin{cases} \hat{n} &= n_0 + 2k/\pi \\ \hat{p} &= 1. \end{cases} \tag{4}$$

The fixation of regulatory alleles in all clusters is asymptotic ($\lim_{t\to\infty} p = 1$), and the equilibrium is asymptotically stable ($\mathrm{d}n/\mathrm{d}t > 0$ and $\mathrm{d}p/\mathrm{d}t > 0$). Fig. 2 illustrates the effect of u and k on the dynamics n_t and p_t. The increase in regulatory

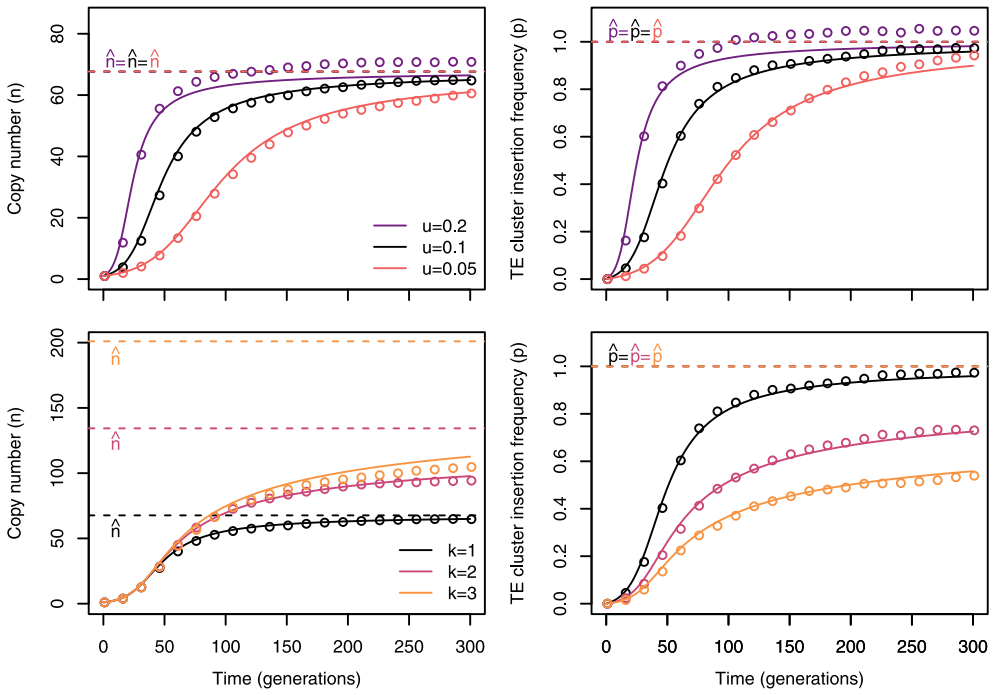


Fig. 2. Dynamics in the neutral model. Unless indicated otherwise, default parameters were set to $n_0 = 1$, $\pi = 0.03$, $k = 1$ cluster, and the transposition rate was $u = 0.1$. The top panel illustrates the influence of the transposition rate, the bottom panel of the number of clusters. Left: number of copies n, right: frequency of the TE cluster insertions in the population (p). Open symbols: simulations, plain lines: difference equations, hyphenated lines: predicted equilibria. The copy number equilibrium $\hat{n}$ does not depend on the transposition rate, and the TE cluster insertion frequency at equilibrium $\hat{p} = 1$ in all conditions. The time necessary to reach the equilibrium (both for $\hat{n}$ and $\hat{p}$) increases with k. In simulations, frequencies could be slightly > 1 due to rare events in which several TEs could insert simultaneously in the same cluster.

allele frequency p is due to new transposition events in clusters (there would be multiple alleles at the sequence level, all being functionally equivalent). Assuming that n_0 is small, the number of copies at equilibrium is proportional to the number of clusters k, and inversely proportional to the cluster size π. With several clusters ($k > 1$), the increase in copy number is slow, and the system may take a very long time to reach equilibrium. The copy number dynamics in simulations with different number of clusters remains very similar for the first hundreds of generations. The equilibrium copy number did not depend on the transposition rate u. This result relies on the absence of linkage disequilibrium between regulatory clusters and genomic copies; simulations showed that this assumption does not hold when the number of clusters increases, or when the transposition rate is very high (Appendix B.1).

3.2. Selection

Natural selection, by favoring the reproduction of genotypes with fewer TE copies, generally acts in the same direction as regulation. A piRNA regulation model implementing selection could be derived by combining Eqs. (2) and (3). In order to simplify the analysis, we derived the results assuming that the deleterious effects of TE copies were independent, i.e. $w_n = \exp(-ns)$, where n is the copy number and s the coefficient of selection (deleterious effect per insertion), so that $\partial \log w_n / \partial n = -s$.

The following calculation relies on the additional assumption that piRNA clusters do not represent a large fraction of genomes ($\pi \ll 1$, leading to $n \gg 2kp$, i.e. that the number of TE copies in the clusters is never large enough to make a difference in the total TE count). We will describe two selection scenarios that happened to lead to qualitatively different outcome: TE insertions in piRNA clusters are neutral, and TE cluster insertions are as deleterious as the other insertions.

Deleterious TEs and neutral clusters. If TE cluster insertions are neutral, the model becomes:

$$\begin{aligned} \frac{dn}{dt} &= nu(1-p)^{2k} - ns' \\ \frac{dp}{dt} &= \frac{\pi}{2k} nu(1-p)^{2k}. \end{aligned} \tag{5}$$

This equation only gives two equilibria, $E_2 : \hat{n} = 0$ (loss of all copies), and $E_3 : \hat{p} = 1$ (when $s' = 0$), which is the same as for the neutral model (Eq. (3)): no selection and fixation of all TE cluster insertions. At the beginning of the dynamics, assuming $p_0 = 0$, the TE invades if $u > s'$ (otherwise the system converges immediately to equilibrium E_2 and the TE is lost). The copy number increases ($dn/dt > 0$) up to a maximum n^*, which is achieved when $p = p^*$ (Fig. 3). The maximum copy number can be obtained analytically (Appendix A.2):

$$\begin{aligned} p^* &= 1 - \left(\frac{s'}{u}\right)^{1/2k} \\ n^* &= n_0 + \frac{2k}{\pi}\left[1 - \frac{1}{2k-1}\left(2k\left(\frac{s'}{u}\right)^{1/2k} - \frac{s'}{u}\right)\right]. \end{aligned} \tag{6}$$

The match between simulations (featuring a finite number of insertion sites, linkage disequilibrium, and non-overlapping generations) and this theoretical result was very good for a single cluster ($k = 1$), but degraded with larger number of clusters; n^* was overestimated by $\sim 10\%$ compared to simulations for $k = 5$ (Fig. 3). Once the maximum number of copies is achieved, TE cluster insertions keep on accumulating, decreasing the transposition

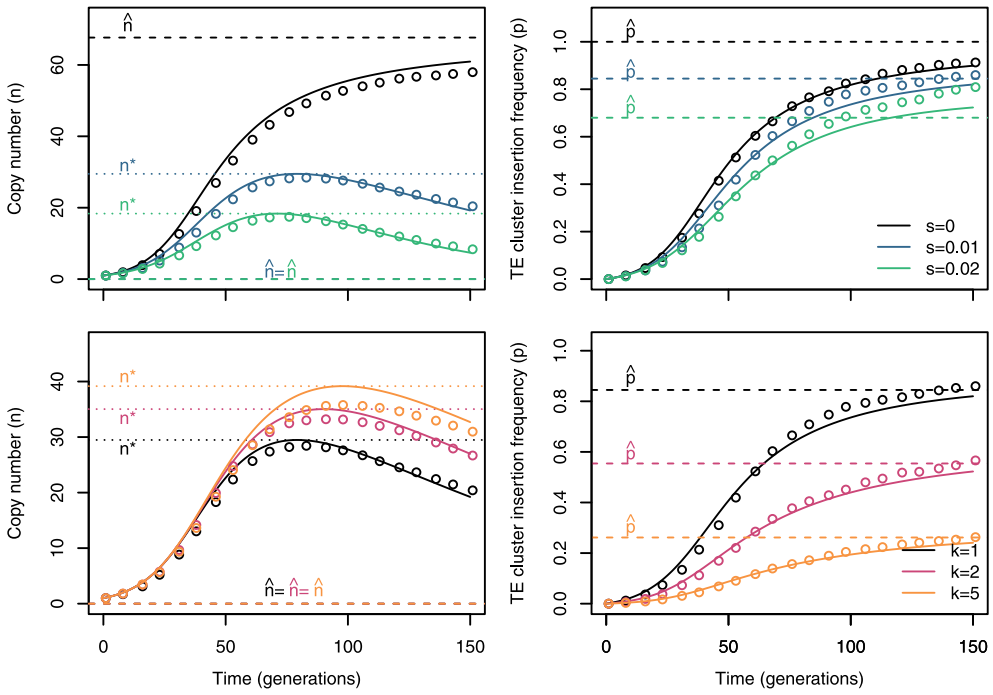


Fig. 3. Dynamics of the deleterious TE–neutral cluster model. The top panel illustrates the influence of the selection coefficient s (with $u = 0.1, k = 1$), the bottom panel of the number of clusters k (with $s = 0.01, u = 0.1$). Left: number of copies n, right: frequency of the segregating TE cluster insertions in the population p. Open symbols: simulations, plain lines: difference equations, hyphenated lines: predicted equilibrium ($\hat{n}$ on the left, $\hat{p}$ on the right), dotted lines: predicted copy number maximum n^*. Whenever $s > 0$, the copy number equilibrium $\hat{n}$ is 0.

rate, which leads to a decrease in the copy number, up to the loss of the element ($\hat{n} = 0$ at equilibrium). At that stage, TE cluster insertions are not fixed, and the equilibrium TE cluster insertion frequency $\hat{p}$ can be expressed as a function of copy number and TE cluster insertion frequency at the maximum (p^* and n^*) (Appendix A.3):

$$\hat{p} - \frac{s'}{u(2k-1)}\frac{1}{(1-\hat{p})^{2k-1}} = p^* - \frac{s'}{u(2k-1)}\frac{1}{(1-p^*)^{2k-1}} - \frac{\pi n^*}{2k}, \tag{7}$$

from which an exact solution for $\hat{p}$ could not be calculated. The following approximation (from Appendix A.4):

$$\hat{p} \simeq 1 - \left[\frac{u}{s'}(2k-1)p^* + 1\right]^{\frac{1}{1-2k}} \tag{8}$$

happens to be acceptable for a wide range of transposition rates and for small selection coefficients ($s < 0.1$) (Fig. 4).

The drop in $\hat{p}$ when the number of clusters k increases suggests that there might be a number of TE cluster insertions above which the transposition is effectively canceled. From Eq. (5), the copy number stabilizes when $u(1-p)^{2k} = s'$, i.e. when $m = 2k(1 - (s'/u)^{1/2k})$, where $m = 2kp$ is the total number of TE cluster insertions per diploid individual. When TE copies are deleterious ($s = 0.01$), this expression tends to $m = -\log(s'/u)$ when $k \to \infty$, which is about $m = 2.1$ copies per individual with our default parameters. In absence of selection, Eq. (4) predicts that transposition should occur up to the total fixation (i.e., $m = 2k$). However, the effective transposition rate will drop to very low levels much before fixation and its effects may rapidly become overwhelmed by genetic drift in finite-size populations. The variance of the change in copy number due to genetic drift between consecutive generations is about $\bar{n}/N$; in a population of size $N = 1000$ with $\bar{n} = 10$ copies per individual, drift changes the average copy number by $\pm 1\%$, while transposition will generate $u\exp(-m)$ new insertions in average (1% of $\bar{n}$ for $m = 2$ TE cluster insertions, and 0.2% of $\bar{n}$ for $m = 4$ TE cluster insertions with our default parameter set $u = 0.1$). Based on simulations, Kofler (2019) estimated that transposition effectively stops above 4 TE cluster insertions per individual in average, which matches this prediction.

Eq. (6) can be reorganized to address the problem of population extinction, as formulated in Kofler (2020). Numerical simulations have indeed shown that even if the final equilibrium state involves the loss of all TE copies, populations need to go through a stage where up to n^* deleterious copies are present in the genome. This makes it possible to approximate mathematically the critical cluster size π_c, from which the population fitness drops below an arbitrary threshold w_c and is at risk of extinction:

$$\pi_c > \frac{2k}{-(\log w_c)/s - n_0}\left[1 - \frac{1}{2k-1}\left(2k\left(\frac{s'}{u}\right)^{\frac{1}{2k}} - \frac{s'}{u}\right)\right]. \tag{9}$$

Setting $s = 0.01$, $u = 0.1$, and $n_0 = 1$, as in the other examples, and taking $w_c = 0.1$ gives $\pi_c > 0.0036$ for $k = 1$ and $\pi_c > 0.005$ for $k = 5$, these values being of the same order of magnitude than the interval 0.1% to 0.2% determined numerically by Kofler (2020).

Deleterious TEs and deleterious clusters. If the TE cluster insertions are as deleterious as other TEs, selection acts on TE cluster insertion frequency as predicted by population genetics (assuming no dominance):

$$\begin{aligned}\frac{dn}{dt} &= nu(1-p)^{2k} - ns' \\ \frac{dp}{dt} &= \frac{\pi}{2k}nu(1-p)^{2k} - sp\frac{1-p}{1-sp}.\end{aligned} \tag{10}$$

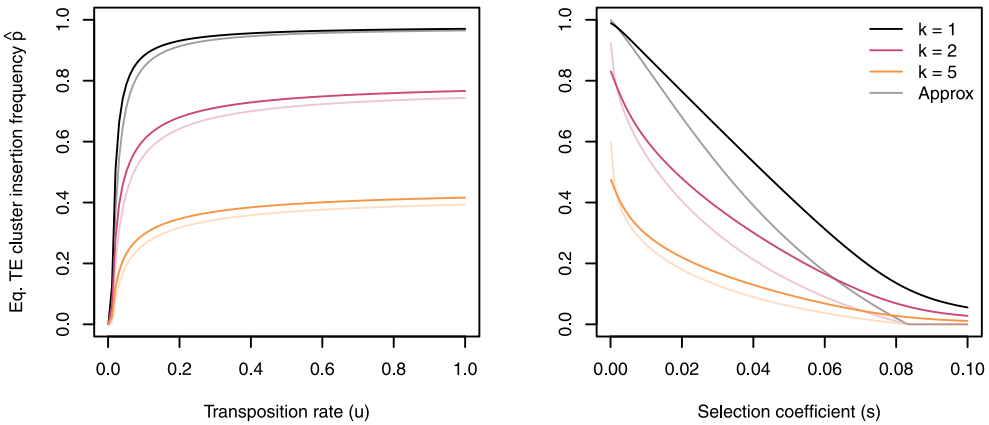


Fig. 4. Influence of model parameters on the equilibrium TE cluster insertion frequency in the deleterious TE–neutral cluster model. The number of clusters k is indicated with different line colors. In this model, the TE is finally eliminated from the genome ($\hat{n} = 0$), the equilibrium TE cluster insertion frequency $\hat{p}$ depends on the duration of the stay of the TE in the genome. Deleterious TEs are eliminated more rapidly, and have thus less opportunity to transpose into piRNA clusters, thus the lower $\hat{p}$. The approximation proposed in Eq. (8) is indicated in light colors.

This allows for a new equilibrium E_4:

$$\begin{cases} \hat{n} = \dfrac{2k}{\pi}\dfrac{s}{s'}\left(\dfrac{s'}{u}\right)^{1/2k}\dfrac{\hat{p}}{1-s\hat{p}} \\ \hat{p} = 1-\left(\dfrac{s'}{u}\right)^{1/2k}. \end{cases} \tag{11}$$

The equilibrium exists ($\hat{n} > 0$ and $\hat{p} > 0$) whenever $s < u(1 + 2u)$, i.e. the transposition rate must be substantially larger than the selection coefficient. It corresponds to the "Transposition–selection cluster" equilibrium described from numerical simulations in Kofler (2019). The dynamics for n and p are illustrated in Fig. 5; the convergence to a non-zero equilibrium $\hat{n}$ and an intermediate equilibrium for $\hat{p}$ (no fixation) was confirmed by simulations. As for the neutral-cluster model, the match between simulations and mathematical predictions tends to degrade for large values of k. The influence of model parameters (u, s, and π) on equilibrium values is depicted in Fig. 6; both the transposition rate u and the selection coefficient s show a non-monotonic relationship with the equilibrium copy number $\hat{n}$ (maximum number of copies for an intermediate value of u and s).

A linear stability analysis (Appendix A.5) shows that for the whole range of u, π, and k, as well as for most of the reasonable values of s, the equilibrium is a stable focus, i.e. the system converges to the equilibrium while oscillating around it.

3.3. Genetic drift

The models described above assume infinite population size, which may not hold for low-census species and for laboratory (experimental evolution) populations. We assessed the influence of population size on the copy number with numerical simulations, comparing the neutral model, the deleterious TE–neutral cluster model, and the deleterious TE–deleterious cluster model with a "classical" copy-number regulation model in which $u_n = u_0/(1 + bn)$ (Charlesworth and Charlesworth, 1983). Since the only equilibrium in the deleterious TE–neutral cluster model is the loss of all TE copies ($\hat{n} = 0$), comparisons had to be performed at an intermediate time point of the dynamics (arbitrarily, after $T = 100$ generations). Models were parameterized such that the copy number n was approximately the same after 100 generations. Drift affects piRNA models substantially more than copy number regulation, the variance of all trap models being approximately one order of magnitude larger (Fig. 7). Consistently with population genetics theory, the variance across simulation replicates decreased with $1/N$ for all models.

The standard population genetics theory predicts that selection in small populations is less effective at eliminating slightly deleterious alleles. Assuming that TE copies are deleterious, they should be eliminated faster in large populations compared to small ones. Although this mechanism has been proposed to explain the accumulation of junk DNA (including TE copies) in multicellular eukaryotes (Lynch and Conery, 2003), little is known about how the equilibrium copy number of an active TE family is expected to be affected by drift even in the simplest scenarios (Charlesworth and Charlesworth, 1983). Yet, informal models suggest that drift may have a limited effect, as copy number equilibria rely on the assumption that evolutionary forces that limit TE amplification (regulation and/or selection) increase in intensity when the copy number increases. Thus, when drift pushes the average copy number up or down, transposition is expected to be less or more effective respectively, which compensates the random deviation. Simulations show that, whatever the model, the copy number is indeed slightly higher in small populations ($N < 100$) when TEs are deleterious, but this effect never exceeds 20% of the total copy number (Supplementary Fig. B2). Overall, drift has a very limited impact on the mean copy number when $N > 50$.

4. Discussion

4.1. Population genetics of the trap model

The formalization of TE regulation by piRNA clusters (the "trap model") made it possible to derive a series of non-intuitive results, and evidence how trap regulation differs from traditional (copy-number based) regulation models. Among the most striking results: (i) in absence of selection (neutral trap model), the equilibrium copy number does not depend on the transposition rate, (ii) the proportion of genotypes able to regulate TEs increases with the size of clusters and decreases with the number of clusters, (iii) deleterious TEs can always invade when the transposition rate is larger than the selection coefficient, but the TE family can persist on the long-term only if TE cluster insertions are deleterious as well. When TE cluster insertions are neutral, they can increase in frequency up to fixation, which leads to the loss of all non-cluster TE copies. Equilibria are always stable. piRNA regulation being a mutational process, the TE copy number

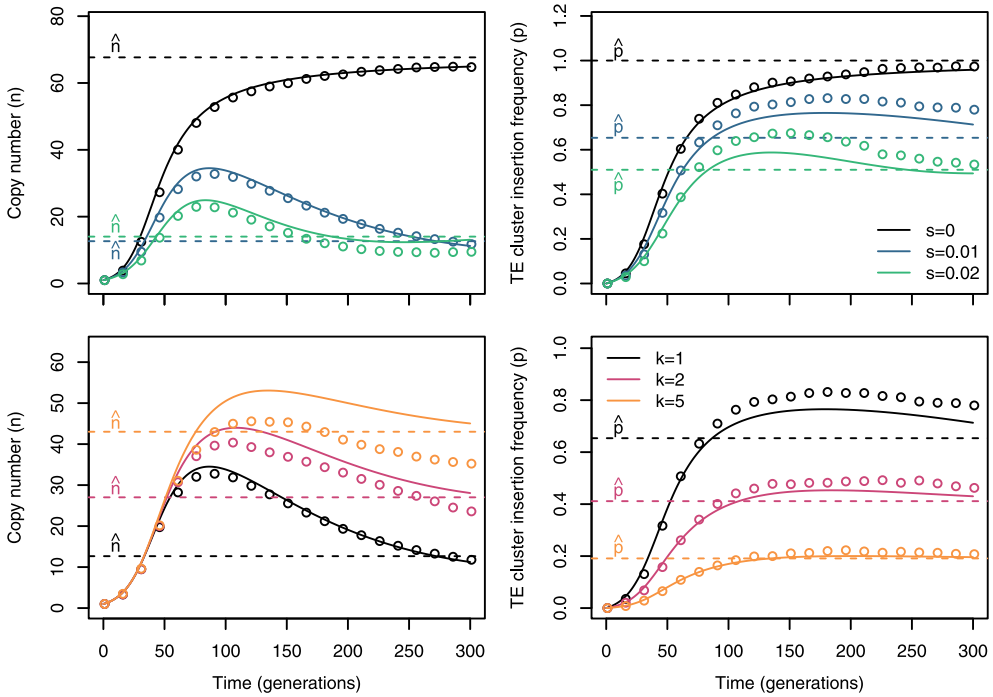


Fig. 5. Dynamics of the deleterious TE–deleterious cluster model. Default parameters were $\pi = 0.03$, $k = 1$, $u = 0.1$, $s = 0.01$. Top panels: influence of the selection coefficient, bottom panels: influence of the number of clusters. Plain lines: predicted dynamics from Eq. (10), hyphenated lines: predicted equilibrium (Eq. (11)), open circles: simulations.

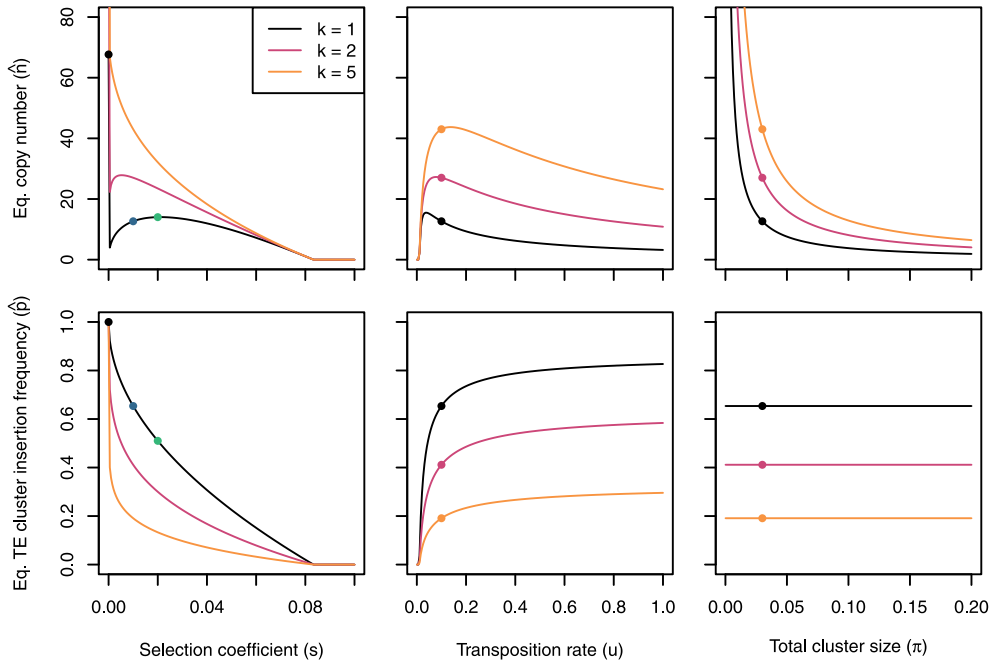


Fig. 6. Influence of model parameters on the equilibrium TE copy number and the TE cluster insertion frequency in the deleterious TE–deleterious cluster model. Default parameter values were $u = 0.1$, $s = 0.01$, and $\pi = 0.03$. The number of clusters ($k = 1$, $k = 2$, and $k = 5$) is indicated by different colors. Colored dots indicate the equilibria illustrated in the panels of Fig. 5 (same color code as in the figure).

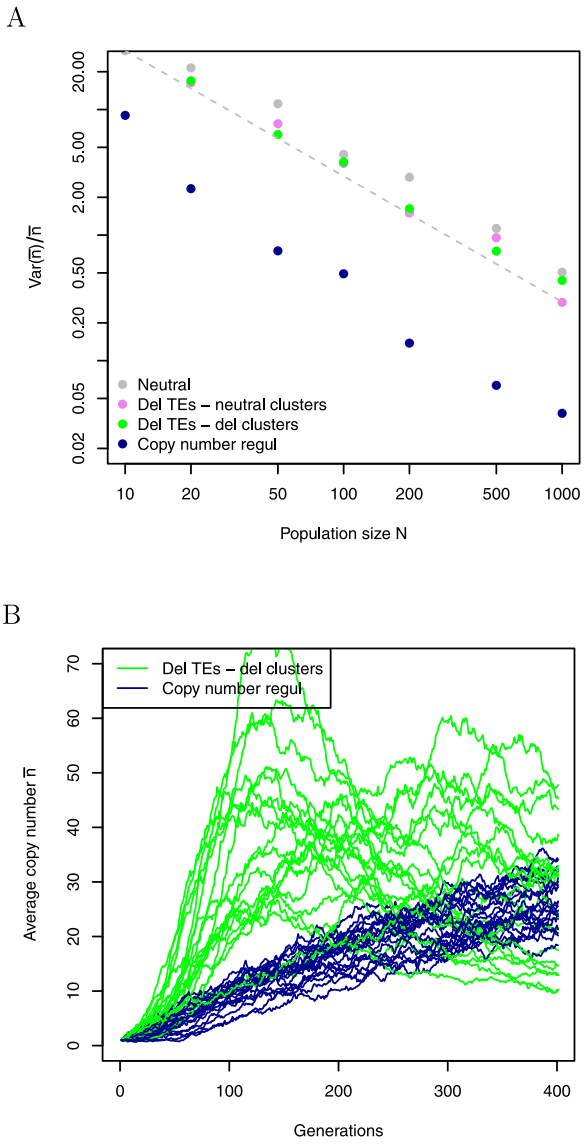


Fig. 7. Effect of genetic drift on the TE copy number. A: Variance in the average copy number (relative to the average copy number) at generation 100 among replicated simulations for various population sizes. Four models are displayed: neutral trap model, deleterious TE–neutral clusters, deleterious TE–deleterious cluster, and copy number regulation. Models were parameterized so that they have very similar copy numbers (about 18) at generation 100; Neutral trap model: $u = 0.045$, $\pi = 0.03$, $k = 2$); Deleterious TE–neutral clusters: $u = 0.13$, $\pi = 0.03$, $s = 0.01$, $k = 2$; Deleterious TE–deleterious clusters: $u = 0.07$, $\pi = 0.03$, $s = 0.01$, $k = 2$; Copy number regulation: $u_n = 0.07/(1 + 0.3n)$, $s = 0.01$. The theoretically-expected decrease in variance (in $1/N$) is illustrated for the neutral model (slope of -1 on the log–log plot). B: The effect of genetic drift is larger in the trap model than for copy-number regulation models. The figure displays the average copy number $\bar{n}$ in 20 independent replicates, $N = 100$ for both models. Parameters were $u = 0.1$, $s = 0.01$, $\pi = 0.03$, $k = 2$ for the trap model, and $u_n = 0.1/(1 + 0.3n)$, and $s = 0.01$, for the copy-number regulation model. Regulation strength was set so that the expected equilibrium copy number $\hat{n} \simeq 25$ was the same for both models.

is more stochastic and substantially more sensitive to genetic drift than other regulation models.

These results confirm and formalize previous work based on numerical simulations, in particular from Kofler (2019) who has already pointed out the small effect of transposition rate on the final state of the population and the inverse relationship between the number of clusters and the number of TE copies. The characterization of the equilibria demonstrate how the neutral trap model differs from the transposition–selection balance model proposed by Charlesworth and Charlesworth (1983); while the transposition–selection balance mostly depends on the transposition rate, the trap model equilibrium is determined by the mutational target (the size and the number of piRNA clusters).

While the equilibrium for the neutral trap model can be expressed with a very simple formula (Eq. (3)), the derivation of copy number and TE cluster insertion frequencies is less straightforward when selection is accounted for (Eqs. (10) and (11)). When TEs are deleterious even when inserted in the clusters, the equilibrium copy number depends on the transposition rate u and the selection coefficient s in a non-monotonic way (less copies when u or s are either very low or very large). The fact that there exists an optimal transposition rate when TE insertions are deleterious have been proposed previously, in a different theoretical framework (Le Rouzic and Capy, 2005). The elimination of high-transposition rate TEs from the genome is due to linkage disequilibrium; when transposition is very active, new TE copies are not randomly spread in the population but rather gathered into high-copy number (and thus, low fitness) genotypes. The optimal rate in the trap model (about 0.1 to 0.2 transpositions per copy and per generation in unregulated genetic backgrounds, Fig. 6) is compatible with empirical estimates (Robillard et al., 2016; Kofler et al., 2022).

4.2. Model approximations

The mathematical formulation of the trap model relies on a series of approximations. The general framework is strongly inspired from (Charlesworth and Charlesworth, 1983), and is based on the same general assumptions, such as a uniform transposition rates and selection coefficients among TE copies, diploid, random mating populations, and no linkage disequilibrium. This framework fits better some model species, including Drosophila or humans, than others (such as self-fertilizing plants or nematodes) for which the population genetics setup needs to be adapted. In general, individual-based (non-overlapping generations) simulations fit convincingly the predictions, but errors are cumulative in the trap model: small biases in the differential equations could add up over time and generate a visible discrepancy after several dozens generations.

The biology of the piRNA cluster regulation was also simplified. We considered that piRNA clusters were completely dominant and epistatic, i.e. a single genomic insertion drives the transposition rate to zero. Relaxing slightly this assumption is unlikely to modify qualitatively the model output, e.g. considering that regulatory insertions are recessive would change the frequency of permissive genotypes from $(1-p)^{2k}$ to $(1-p^2)^k$, which would increase the TE cluster insertion frequency at equilibrium but not its stability. In a similar way, if the strength of regulation was increasing with the number of TE cluster insertions (or the number of genomic TEs), equilibrium would still be achieved for a higher number of copies. In contrast, imperfect regulation (a residual transposition rate even when all TE cluster insertions are fixed, such as in Lu and Clark (2010)) would break the equilibrium in the neutral case, and copy number would raise indefinitely. This only affects the neutral model though, as imperfect regulation would have a much more limited effect when TEs are deleterious.

In order to compute the equilibria, we assumed no epistasis on fitness, i.e. constant $\partial \log w/\partial n = -s$. Deriving the model

with a different fitness function is possible, although solving the differential equations could be more challenging. Instead of our fitness function $w_n = e^{-ns}$, Charlesworth and Charlesworth (1983) proposed $w_n = 1 - sn^c$ (c being a coefficient quantifying the amount of epistasis on fitness), while Dolgin and Charlesworth (2006) later used $w_n = e^{-sn-cn^2}$ (different parameterizations for directional epistasis are discussed in e.g. Le Rouzic (2014)). Considering negative epistasis on fitness (i.e. the cost of additional deleterious mutations increases) in TE population genetic models has two major consequences: (i) the strength of selection increasing with the copy number, it ensures and stabilizes the equilibrium even in absence of regulation (Charlesworth and Charlesworth, 1983), and (ii) the model is more realistic, as epistasis on fitness for deleterious mutations has been measured repeatedly on many organisms (Maisnier-Patin et al., 2005; Kouyos et al., 2007; Khan et al., 2011). Interestingly, there is little evidence of negative epistasis for fitness among TE insertions (Lee, 2022), suggesting that epistasis is probably not a major explanation for the stabilization of the copy number. In the trap model, regulation itself is strong enough to achieve an equilibrium in absence of selection, so epistasis on fitness is expected to modify the equilibrium copy number and the range of parameters for which a reasonable copy number can be maintained (Kofler, 2019), but not the presence of a theoretical equilibrium.

Recent data might suggest that piRNA regulation may not be the only mechanism involved in early regulation of TE activity. For instance, Kofler et al. (2022) observed, in a lab experimental evolution context, that the number of TE cluster insertions was lower than expected in the trap model. Combining a copy-number regulation component and the trap model framework is theoretically possible and does not invalidate our approach, at the cost of introducing a new regulation parameter in the equations. In a more general way, the diversity of transposition regulation mechanisms in animals (Lu and Clark, 2010; Saint-Leandre et al., 2020), plants (Roessler et al., 2018), and micro-organisms (Sousa et al., 2013), makes it impossible to derive models that are both accurate and universal.

4.3. piRNA clusters, selection, and recombination

The most effective configuration for TE regulation is a single, large piRNA cluster. Dividing the cluster in smaller parts increases equilibrium TE copy numbers, and reducing the total cluster size as well. Kofler (2019) has already noticed that recombination among cluster loci reduces the efficiency of regulation, and that regulation was more efficient with one large, non-recombining cluster than with many small clusters spread on several chromosomes (the single-cluster model was called the "flamenco" model in Kofler (2019), inspired from the *flamenco* regulatory locus in Drosophila, Goriaux et al., 2014). In our setting, the proportion of transposition-permissive genotypes in the population is $(1 - m/2k)^{2k}$ when $m = 2kp$ TE cluster insertions are present in the genome and evenly distributed among clusters. This proportion increases as a function of k, meaning that the most efficient regulation is achieved when $k = 1$. Furthermore, the number of copies at equilibrium is expected to be proportional to the number of clusters k. Selection for TE regulation should thus minimize recombination within and across clusters; the fact that, in most organisms, piRNA clusters seem to be located at several loci needs to be explained by other factors (such as functional constraints) than the regulation efficiency. For instance, spreading piRNA clusters at several genomic locations may prevent TEs to evolve cluster avoidance strategies. The need to regulate independently different TE families might also play a role in the scattering of piRNA clusters; the interactions between several TE families invading simultaneously may generate new constraints on the regulation system, which probably deserves further investigation.

Our theoretical analysis displays contrasting results depending on the selection pressure on TE insertions located in piRNA clusters. The nature and the size of TE-induced deleterious effects is a long-standing issue (Nuzhdin, 2000; Lee, 2022) that remains poorly understood. TEs can affect the host fitness through various mechanisms, including the interruption of functional genomic sequences, chromosomal rearrangements due to ectopic recombinations between TEs inserted at different loci, or a direct poisoning effect of the transposition process. The properties of piRNA clusters seem to limit most of these potential effects: the density of functional regions in probably low; clusters tend to be located in low-recombination regions, and the transcription into mRNA is likely to be repressed. It is thus tempting to speculate that the effect of TE cluster insertions on the host fitness should be lower than TE copies inserted in random genomic positions, and that the deleterious TE–neutral cluster model is more realistic than the deleterious TE–deleterious cluster model. Solid empirical evidence is necessary to confirm this speculation, which cannot be addressed solely based on theoretical considerations.

An interesting hypothesis was raised by Kelleher et al. (2018) about the possibility that TE cluster insertion frequency could be influenced by positive selection. Assuming deleterious TEs, genotypes able to control TE spread are indeed expected to display a selective advantage over those in which transposition is unregulated, suggesting that TE cluster insertions should sweep in the population as advantageous alleles. Our model, neglecting linkage disequilibrium between TEs and TE cluster insertions, would then underestimate the increase in frequency of regulatory alleles (and thus overestimate the copy number). Although the reasoning is theoretically valid, the actual strength of positive selection on piRNA clusters is probably limited in general. Assuming that TE insertions have a local deleterious effect (because they disrupt genes or gene regulation), the selective advantage of a regulatory locus is weak and indirect (of the same order of magnitude as $n \times u \times s$, the deleterious effect of the few insertions arising in a single generation). In contrast, if active transposition is deleterious (such as in the hybrid dysgenesis scenario explored by Kelleher et al., 2018), the selective advantage of TE cluster insertions is of the order of magnitude of $n \times s$, and selection may have an effect on TE cluster insertion frequencies. Although it is experimentally difficult to determine how selection acts on TEs, both scenarios are expected to leave different genomic footprints, as the positive selection hypothesis posits that regulatory alleles should be shared among many individuals of the population, while the neutral hypothesis expects that various individuals are regulated by independent TE cluster insertions. Empirical evidence is scarce, but seems to favor the neutral hypothesis (Zhang et al., 2020).

4.4. Concluding remarks

Comparative genomics applied to transposable elements is hard. Notwithstanding the countless potential artifacts associated with sequencing, assembly, and annotation biases, understanding the evolutionary history of genomes is limited by the small number of evolutionary replicates, and the number of TE families and TE copy numbers accumulated in a single species is frequently dominated by stochastic and contingent factors. Being able to compare observed patterns with model predictions is thus of utter importance.

By extending the existing theory of transposable elements population genetics, we were able to demonstrate that the trap regulation model was affecting deeply the dynamics of TE invasion. In particular, when TE cluster insertions are neutral, the possibility to maintain a stable copy number equilibrium disappears, and all active TE copies are expected to be lost eventually.

When regulatory insertions are slightly deleterious, a hypothetical transposition–selection equilibrium can be achieved, in which genomic TEs maintain as selfish DNA sequences, while regulatory insertions maintain as a result of a selection–mutation balance. This situation prevents the fixation of TE cluster insertions, and thus maintains a low frequency of transposition-permissive genotypes (without piRNA against active TE families), which could be measured empirically in populations to estimate the likelihood of the alternative regulation scenarios.

Funding

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) (project TRANSPHORIZON, ANR-18-CE02-0021). The funding agency had no role in the data analysis, the writing of the report, and in the decision to submit the article for publication.

CRediT authorship contribution statement

Siddharth S. Tomar: Software, Writing – review & editing. **Aurélie Hua-Van:** Funding acquisition, Project administration, Supervision, Writing – review & editing. **Arnaud Le Rouzic:** Conceptualization, Methodology, Software, Formal analysis, Supervision, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

The authors thank three anonymous reviewers for their positive and constructive suggestions.

Appendix A. Mathematical details

A.1. Eq. (3) for large π.

Eq. (3) assumes for simplicity that $1-\pi \simeq 1$, which may not hold in all species. If only a proportion $1-\pi$ of new insertions fall in non-cluster regions, the model can be re-written as:

$$\frac{\mathrm{d}n}{\mathrm{d}t} = (1-\pi)nu(1-p)^{2k}$$
$$\frac{\mathrm{d}p}{\mathrm{d}t} = \frac{\pi}{2k}nu(1-p)^{2k}.$$

The equilibrium then becomes:

$$\begin{cases} \hat{n} &= n_0 + 2k\frac{1-\pi}{\pi}, \\ \hat{p} &= 1. \end{cases}$$

A.2. Eq. (6)

When the copy number n achieves its maximum n^*, $\mathrm{d}n/\mathrm{d}t = 0$. This happens when the TE cluster insertion frequency p^* is:

$$\frac{\mathrm{d}n}{\mathrm{d}t} = n^*u(1-p^*)^{2k} - n^*s' = 0$$
$$p^* = 1 - \left(\frac{s'}{u}\right)^{\frac{1}{2k}}.$$

The number of copies cumulated while p is rising from p_0 to p^* can be calculated by integrating both sides:

$$\frac{\mathrm{d}n}{\mathrm{d}p} = \frac{2k}{\pi}\left(1 - \frac{s'}{u(1-p)^{2k}}\right)$$
$$\int_{n_0}^{n^*}\mathrm{d}n = \frac{2k}{\pi}\left[\int_{p_0}^{p^*}\mathrm{d}p - \frac{s'}{u}\int_{p_0}^{p^*}(1-p)^{-2k}\mathrm{d}p\right]$$
$$n^* - n_0 = \frac{2k}{\pi}\left[p^* - p_0 - \frac{s'}{u(2k-1)}((1-p^*)^{1-2k} - 1)\right]$$
$$n^* = n_0 + \frac{2k}{\pi}\left[p^* + \frac{s'}{u(2k-1)}(1-(1-p^*)^{1-2k})\right]$$
$$n^* = n_0 + \frac{2k}{\pi}\left[1 - \frac{1}{2k-1}\left(2k\left(\frac{s'}{u}\right)^{\frac{1}{2k}} - \frac{s'}{u}\right)\right].$$

A.3. Eq. (7)

The strategy was very similar than for obtaining n^*, with $\mathrm{d}p/\mathrm{d}n$ integrated both sides from the maximum to the equilibrium:

$$\int_{n^*}^{\hat{n}=0}\mathrm{d}n = \frac{2k}{\pi}\left[\int_{p^*}^{\hat{p}}\mathrm{d}p - \frac{s'}{u}\int_{p^*}^{\hat{p}}(1-p)^{-2k}\mathrm{d}p\right]$$
$$-n^* = \frac{2k}{\pi}\left[(\hat{p}-p^*) - \frac{s'}{u}\left(\frac{(1-\hat{p})^{1-2k} - (1-p^*)^{1-2k}}{2k-1}\right)\right].$$

A.4. Eq. (8)

Rewriting the previous equation with $\delta p = \hat{p} - p^*$ and $1-p^* = q^*$ gives:

$$-n^* = \frac{2k}{\pi}\left[\delta p - \frac{s'}{u(1-2k)}\frac{1}{(q^*-\delta p)^{2k-1}} - \frac{s'}{u(1-2k)}{q^*}^{1-2k}\right],$$

which turns out to be dominated by the second term ($1/(q^* - \delta p)^{2k-1} \gg \delta p$ when δp increases) for most parameter values. As a consequence, neglecting $\hat{p} - p^*$ leads to:

$$n^* \simeq \frac{2k}{\pi}\left[\frac{s'}{u}\left(\frac{(1-\hat{p})^{1-2k} - (1-p^*)^{1-2k}}{2k-1}\right)\right]$$
$$\iff \hat{p} \simeq 1 - \left[(1-p^*)^{1-2k} + \frac{\pi u(2k-1)}{2s'k}n^*\right]^{\frac{1}{1-2k}}.$$

Replacing p^* and n^* with their expressions from Eq. (6) and reorganizing gives:

$$\hat{p} \simeq 1 - \left[\frac{u}{s'}(2k-1)(1 + \frac{n_0\pi}{2k} - \left(\frac{s'}{u}\right)^{\frac{1}{2k}}) + 1\right]^{\frac{1}{1-2k}}.$$

Assuming that n_0 is reasonably small and $\pi \ll 1$, the term $n_0\pi/2k$ can be further neglected, and:

$$\hat{p} \simeq 1 - \left[\frac{u}{s'}(2k-1)(1 - \left(\frac{s'}{u}\right)^{\frac{1}{2k}}) + 1\right]^{\frac{1}{1-2k}}.$$

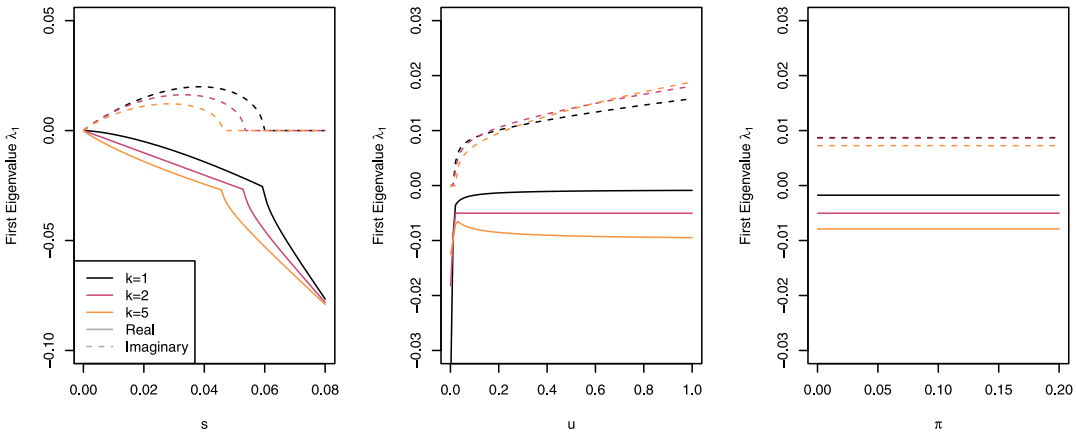


Fig. A1. First Eigenvalue of the Jacobian matrix as a function of model parameters (u, s, and π) in the deleterious TEs–deleterious cluster model. Default parameter values were $u = 0.1$, $s = 0.01$, and $\pi = 0.03$. The number of clusters ($k = 1$, $k = 2$, and $k = 5$) is indicated by different line colors. Eigenvalues are complex for most of the range of the parameters, real part: plain lines, imaginary part: dotted lines.

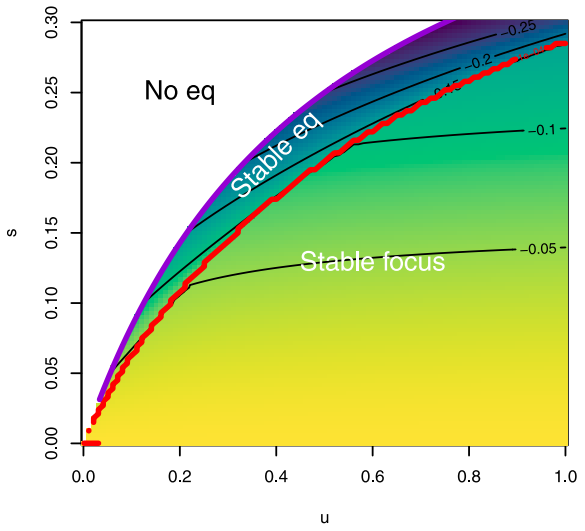


Fig. A2. Equilibrium stability for the deleterious TEs - deleterious cluster model. The figure represents the real part of the first eigenvalue of the Jacobian matrix for two major parameters (u and s), with $k = 1$ and $\pi = 0.03$. The eigenvalue is negative for the whole parameter range, and is a complex number for most of the range (below the red line). The purple line delineates $s = u/(1 + 2u)$, beyond which selection is too strong to let the TE invade (white area).

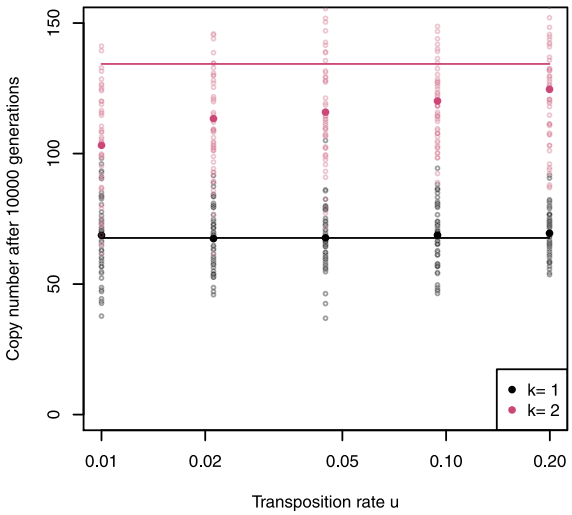


Fig. B1. Effect of the transposition rate u on the simulated equilibrium copy number in the neutral model. Eq. (4) predicts that, in the neutral model, the equilibrium does not depend on the transposition rate. Simulations were run for $k = 1$ and $k = 2$ clusters in populations of size $N = 5,000$; simulations were stopped after 10,000 generations. The figure displays the final copy number in each simulation (open symbols), their average (filled symbols), and the theoretical prediction (plain lines). Simulations display a slight increase in the equilibrium copy number for large transposition rates, due to linkage disequilibrium. This effect increases with the number of clusters. Conversely, when the transposition rate is low, the invasion dynamics is slower, and all TEs might not be fixed by the end of the simulations. Overall, theoretical predictions fit well for a single cluster, but simulations featuring several clusters are slower, and the final copy number remains below the theoretical expectation in finite populations from $k = 2$ clusters.

A.5. Equilibrium stability for Eq. (11)

The Jacobian matrix corresponding to the equilibrium $(\hat{n}, \hat{p})$ from Eq. (11) is:

$$\boldsymbol{J} = \begin{bmatrix} 0 & -2k\hat{n}u\left(\frac{s'}{u}\right)^{\frac{2k-1}{2k}} \\ \frac{\pi s'}{2k} & \frac{1-s}{(1-s\hat{p})^2} - \hat{n}u\pi\left(\frac{s'}{u}\right)^{\frac{2k-1}{2k}} - 1 \end{bmatrix}.$$

(see Figs. A1 and A2).

Eigenvalues are negative (i.e., the equilibrium is stable) for all tested parameter combinations. Eigenvalues happen to be complex for all parameter combinations (except for very large values of s), the equilibrium is thus a stable focus, reached asymptotically by oscillating around it.

Appendix B. Supplementary results

The analytical predictions assume that linkage disequilibrium and genetic drift can be ignored when computing the dynamics of TE copy number. This appendix compares the predictions to

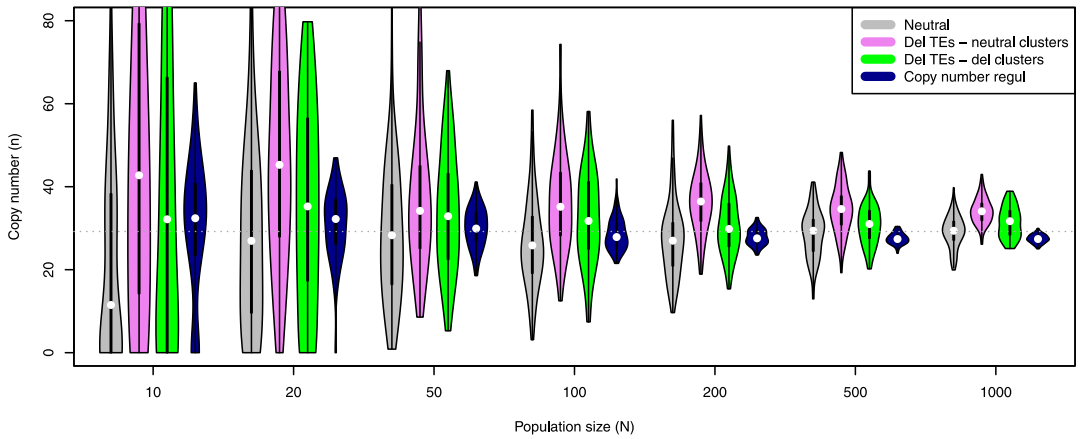


Fig. B2. Distribution of the average copy number $\bar{n}$ among 1000 replicates in different models, with various population sizes. Models were parameterized so that they achieve similar average copy numbers ($\bar{n} \sim 28$) in large populations (horizontal dotted line): $s = 0.01$ for all models (except the neutral model), $k = 1$ cluster and $\pi = 0.03$ in all trap models. Transposition rates were: $u = 0.045$ for the neutral model, $u = 0.05$ for the Deleterious TE–neutral cluster model, $u = 0.15$ for the Deleterious TE–Deleterious cluster model, and $u_n = 0.17/(1 + 0.45n)$ for the regulation model.

numerical simulations (linkage disequilibrium in Appendix B.1, genetic drift in Appendix B.2).

B.1. Sensitivity of the neutral equilibrium (Eq. (4)) to model assumptions.

See Fig. B1.

B.2. Effect of genetic drift on the average and variance of the copy number

See Fig. B2.

References

Adams, M.D., Tarng, R.S., Rio, D.C., 1997. The alternative splicing factor PSI regulates P-element third intron splicing *in vivo*. Genes Dev. 11 (1), 129–138. http://dx.doi.org/10.1101/gad.11.1.129.

Arkhipova, I., Meselson, M., 2005. Deleterious transposable elements and the extinction of asexuals. Bioessays 27 (1), 76–85. http://dx.doi.org/10.1002/bies.20159.

Bergman, C.M., Quesneville, H., Anxolabéhère, D., Ashburner, M., 2006. Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. Genome Biol. 7 (11), 1–21. http://dx.doi.org/10.1186/gb-2006-7-11-r112.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., Hannon, G.J., 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell 128 (6), 1089–1103. http://dx.doi.org/10.1016/j.cell.2007.01.043.

Brookfield, J.F., Badge, R.M., 1997. Population genetics models of transposable elements. Genetica 100 (1), 281–294. http://dx.doi.org/10.1007/978-94-011-4898-6_28.

Charlesworth, B., Charlesworth, D., 1983. The population dynamics of transposable elements. Genet. Res. 42 (1), 1–27. http://dx.doi.org/10.1017/S0016672300021455.

Deniz, Ö., Frost, J.M., Branco, M.R., 2019. Regulation of transposable elements by DNA modifications. Nature Rev. Genet. 20 (7), 417–431. http://dx.doi.org/10.1038/s41576-019-0106-6.

Dolgin, E.S., Charlesworth, B., 2006. The fate of transposable elements in asexual populations. Genetics 174 (2), 817–827. http://dx.doi.org/10.1534/genetics.106.060434.

Doolittle, W.F., Sapienza, C., 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature 284 (5757), 601–603. http://dx.doi.org/10.1038/284601a0.

Gilbert, C., Feschotte, C., 2018. Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. Curr. Opin. Genet. Dev. 49, 15–24. http://dx.doi.org/10.1016/j.gde.2018.02.007.

Gladyshev, E., 2017. Repeat-induced point mutation and other genome defense mechanisms in fungi. In: Heitman, J., Howlett, B.J., Crous, P.W., Stukenbrock, E.H., James, T.Y., Gow, N.A.R. (Eds.), The fungal kingdom. Wiley Online Library, pp. 687–699. http://dx.doi.org/10.1128/9781555819583.ch33.

Goriaux, C., Théron, E., Brasset, E., Vaury, C., 2014. History of the discovery of a master locus producing piRNAs: the flamenco/COM locus in *Drosophila melanogaster*. Front. Genet. 5, 257. http://dx.doi.org/10.3389/fgene.2014.00257.

Grayling, M.J., 2014. phaseR: An R package for phase plane analysis of autonomous ODE systems. R J. 6 (2), 43–51. http://dx.doi.org/10.32614/RJ-2014-023.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585 (7825), 357–362. http://dx.doi.org/10.1038/s41586-020-2649-2.

Hartl, D., Lozovskaya, E., Lawrence, J., 1992. Nonautonomous transposable elements in prokaryotes and eukaryotes. Genetica 86 (1), 47–53. http://dx.doi.org/10.1007/BF00133710.

Huang, S., Yoshitake, K., Asakawa, S., 2021. A review of discovery profiling of PIWI-interacting RNAs and their diverse functions in Metazoans. Int. J. Mol. Sci. 22 (20), 11166. http://dx.doi.org/10.3390/ijms222011166.

Kelleher, E.S., Azevedo, R.B., Zheng, Y., 2018. The evolution of small-RNA-mediated silencing of an invading transposable element. Genome Biol. Evol. 10 (11), 3038–3057. http://dx.doi.org/10.1093/gbe/evy218.

Khan, A.I., Dinh, D.M., Schneider, D., Lenski, R.E., Cooper, T.F., 2011. Negative epistasis between beneficial mutations in an evolving bacterial population. Science 332 (6034), 1193–1196. http://dx.doi.org/10.1126/science.1203801.

Kidwell, M.G., Lisch, D.R., 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution 55 (1), 1–24. http://dx.doi.org/10.1111/j.0014-3820.2001.tb01268.x.

Kofler, R., 2019. Dynamics of transposable element invasions with piRNA clusters. Mol. Biol. Evol. 36 (7), 1457–1472. http://dx.doi.org/10.1093/molbev/msz079.

Kofler, R., 2020. piRNA clusters need a minimum size to control transposable element invasions. Genome Biol. Evol. 12 (5), 736–749. http://dx.doi.org/10.1093/gbe/evaa064.

Kofler, R., Nolte, V., Schlötterer, C., 2022. The transposition rate has little influence on the plateauing level of the P-element. Mol. Biol. Evol. 39 (7), msac141. http://dx.doi.org/10.1093/molbev/msac141.

Kouyos, R.D., Silander, O.K., Bonhoeffer, S., 2007. Epistasis between deleterious mutations and the evolution of recombination. Trends Ecol. Evol. 22 (6), 308–315. http://dx.doi.org/10.1016/j.tree.2007.02.014.

Le Rouzic, A., 2014. Estimating directional epistasis. Front. Genet. 5, 198. http://dx.doi.org/10.3389/fgene.2014.00198.

Le Rouzic, A., Capy, P., 2005. The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. Genetics 169 (2), 1033–1043. http://dx.doi.org/10.1534/genetics.104.031211.

Le Rouzic, A., Capy, P., 2006. Population genetics models of competition between transposable element subfamilies. Genetics 174 (2), 785–793. http://dx.doi.org/10.1534/genetics.105.052241.

Lee, Y.C.G., 2022. Synergistic epistasis of the deleterious effects of transposable elements. Genetics 220 (2), iyab211. http://dx.doi.org/10.1093/genetics/iyab211.

Lohe, A.R., Hartl, D.L., 1996. Autoregulation of *mariner* transposase activity by overproduction and dominant-negative complementation. Mol. Biol. Evol. 13 (4), 549–555. http://dx.doi.org/10.1093/oxfordjournals.molbev.a025615.

Lu, J., Clark, A.G., 2010. Population dynamics of PIWI-interacting RNAs (piRNAs) and their targets in Drosophila. Genome Res. 20 (2), 212–227. http://dx.doi.org/10.1101/gr.095406.109.

Lynch, M., Conery, J.S., 2003. The origins of genome complexity. Science 302 (5649), 1401–1404. http://dx.doi.org/10.1126/science.1089370.

Maisnier-Patin, S., Roth, J.R., Fredriksson, A.s., Nyström, T., Berg, O.G., Andersson, D.I., 2005. Genomic buffering mitigates the effects of deleterious mutations in bacteria. Nature Genet. 37 (12), 1376–1379. http://dx.doi.org/10.1038/ng1676.

Malone, C.D., Hannon, G.J., 2009. Small RNAs as guardians of the genome. Cell 136 (4), 656–668. http://dx.doi.org/10.1016/j.cell.2009.01.045.

Nuzhdin, S.V., 2000. Sure facts, speculations, and open questions about the evolution of transposable element copy number. In: Transposable Elements and Genome Evolution. Springer, pp. 129–137. http://dx.doi.org/10.1007/978-94-011-4156-7_15.

Orgel, L.E., Crick, F.H., 1980. Selfish DNA: the ultimate parasite. Nature 284 (5757), 604–607. http://dx.doi.org/10.1038/284604a0.

Ozata, D.M., Gainetdinov, I., Zoch, A., O'Carroll, D., Zamore, P.D., 2019. PIWI-interacting RNAs: small RNAs with big functions. Nature Rev. Genet. 20 (2), 89–108. http://dx.doi.org/10.1038/s41576-018-0073-3.

R. Core Team, 2020. R: A language and environment for statistical computing. URL https://www.R-project.org/.

Robillard, É., Le Rouzic, A., Zhang, Z., Capy, P., Hua-Van, A., 2016. Experimental evolution reveals hyperparasitic interactions among transposable elements. Proc. Natl. Acad. Sci. 113 (51), 14763–14768. http://dx.doi.org/10.1073/pnas.1524143113.

Roessler, K., Bousios, A., Meca, E., Gaut, B.S., 2018. Modeling interactions between transposable elements and the plant epigenetic response: a surprising reliance on element retention. Genome Biol. Evol. 10 (3), 803–815. http://dx.doi.org/10.1093/gbe/evy043.

Roze, D., 2022. Causes and consequences of linkage disequilibrium among transposable elements within eukaryotic genomes. BioRxiv http://dx.doi.org/10.1101/2022.10.24.513493.

Saint-Leandre, B., Capy, P., Hua-Van, A., Filée, J., 2020. PiRNA and transposon dynamics in drosophila: A female story. Genome Biol. Evol. 12 (6), 931–947. http://dx.doi.org/10.1093/gbe/evaa094.

Selker, E.U., Stevens, J.N., 1985. DNA methylation at asymmetric sites is associated with numerous transition mutations. Proc. Natl. Acad. Sci. 82 (23), 8114–8118. http://dx.doi.org/10.1073/pnas.82.23.8114.

Soetaert, K., Petzoldt, T., Setzer, R.W., 2010. Solving differential equations in R: Package desolve. J. Stat. Softw. 33 (9), 1–25. http://dx.doi.org/10.18637/jss.v033.i09, URL http://www.jstatsoft.org/v33/i09.

Sousa, A., Bourgard, C., Wahl, L.M., Gordo, I., 2013. Rates of transposition in *Escherichia coli*. Biol. Lett. 9 (6), 20130838. http://dx.doi.org/10.1098/rsbl.2013.0838.

Wallau, G.L., Capy, P., Loreto, E., Le Rouzic, A., Hua-Van, A., 2016. VHICA, a new method to discriminate between vertical and horizontal transposon transfer: Application to the *mariner* family within drosophila. Mol. Biol. Evol. 33 (4), 1094–1109. http://dx.doi.org/10.1093/molbev/msv341.

Zanni, V., Eymery, A., Coiffet, M., Zytnicki, M., Luyten, I., Quesneville, H., Vaury, C., Jensen, S., 2013. Distribution, evolution, and diversity of retrotransposons at the *flamenco* locus reflect the regulatory properties of piRNA clusters. Proc. Natl. Acad. Sci. 110 (49), 19842–19847. http://dx.doi.org/10.1073/pnas.1313677110.

Zhang, S., Pointer, B., Kelleher, E.S., 2020. Rapid evolution of piRNA-mediated silencing of an invading transposable element was driven by abundant *de novo* mutations. Genome Res. 30 (4), 566–575. http://dx.doi.org/10.1101/gr.251546.119.

4 Article 2

Building on the analytical model in Article 1 and the simulation framework, we wanted to study the dynamics of concurrent invasions by two TE families and answer if piRNA can provide genomic immunity against repeated TE invasions. These simultaneous invasions can represent horizontal transfers in the same species by similar or completely different TE families or reactivation of the same TE family in the population. Article 2 investigates the dynamics of TE invasion by two families under the trap model and the cross-mobilization of piRNA between genomic copies of the TE families in *D. melanogaster* and between TE families shared between *Drosophilidae* species.

Article 2 is currently in preparation.

Does cross-regulation between transposable element families affect their invasion dynamics?

Siddharth S. Tomar[1], Arnaud Le Rouzic[1], Aurélie Hua-Van[1]

1: Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, 91190, Gif-sur-Yvette, France

Abstract

Piwi-interacting RNAs (piRNAs) regulate the expression of Transposable Elements (TE) in metazoans: mobile sequences that happen to insert into preexisting genomic regions (the piRNA clusters) trigger a specific silencing mechanism. This "trap model" predicts that piRNAs may not only stop active TE proliferation, but also maintain a temporary genomic immunity against TEs from the same family. To what extent piRNA immunity affects the turnover of transposable elements in genomes is virtually unknown. We designed an individual-based population genetics model featuring two related, cross-regulating TE lineages, and simulated horizontal transfers in naive vs immune populations. Our results indicate that the amplification dynamics of the newcomer TE is affected only when cross-regulation is very high, i.e. when the invading TE is essentially identical to the one that triggered immunity. The TE distribution in Drosophilidae confirmed that cross-regulation by piRNA is insufficient by itself to provide genomic immunity against horizontal transfers. Within-species genomic observations in *Drosophila melanogaster* suggest that close successive invasions by similar TEs are possible, with no strong evidence of immunity from past TE invasions.

Introduction

Transposable elements (TEs) are mobile elements, able to invade genomes and, widely distributed in nearly all branches of life, from prokaryotes to metazoans. TEs are diverse, with over 19,000 curated TE families in the Dfam database alone [26], many TEs can be broadly categorised into large orders based on similarities in their replicative mechanisms and transposition machinery [58]. Transposable elements tend to accumulate in genomes, up to representing the majority of the total DNA in many species. Transposition activity is mutagenic and generates insertions, deletions, translocations, and ectopic recombinations [46]. Even if some insertions can be occasionally recruited by natural selection and participate to adaptation [15], TEs are in average deleterious [5, 39], and they are a prominent example of "selfish DNA", able to invade genomes and populations in spite of a fitness cost [47, 19].

At the evolutionary scale, the TE content is dynamic, and the degradation of old, inactive TE families is compensated by novel TEs, often transferred horizontally from other species [29]. Among the most documented examples, the P element in *Drosophila melanogaster* has rapidly proliferated after a transfer from another Drosophila species, *D. willistoni* [18, 53]; the

P element was later transmitted from *D. melanogaster* to its close relative *D. simulans* with a similar success [33]. Similarly, the I element has been implicated in successive re-invasions in populations of *D. melanogaster*, which originally lacks functional I element. More generally, a countless number of horizontal transfers involving a wide diversity of TE families have been documented in Drosophila [56], in other insects [49], in vertebrates [61], in plants [20], and in various eukaryotic and prokaryotic microorganisms [24]. After having being horizontally-transferred, TEs are vertically inherited within populations from parent to offspring, as any other gene. Theoretical models predict that TEs should rapidly multiply at the start of the invasion, before reaching an equilibrium [13, 37] (see [14, 38] for review). This dynamic pattern is compatible with empirical observations, including experimental evolution in fast-reproducing organisms like *Drosophila* [25, 6, 52].

The origin of transposition regulation was rather unclear for a long time: evidence often pointed out the role of TEs and TE-related sequences in their own regulation [e.g., 11], while theoretical models predicted that TEs had little advantage at self-regulating [12, 46]. Mechanisms of transposition regulation are diversified, but it is now established that TE regulation by small RNA sequences, produced by the host from TE-derived DNA sequences, is universal in eukaryotes. For instance, metazoans regulate TE activity with small RNA (24-31nt) sequences called Piwi-interacting RNA(piRNA)[3]. They were discovered in *Drosophila* where they interact with the members of Argonaut protein family in germ cells to control the expression of TEs [42]. The piRNA-loaded protein is then guided either to complementary mRNAs (TE transcripts), that will be sliced into small pieces, hence triggering post-transcriptional silencing. Alternatively, the piRNA-protein complex can enter the nucleus and target and silence euchromatic TE copies through deposition of epigenetic marks. piRNAs originate from long non-coding RNA precursors, transcribed from genomic regions known as piRNA clusters. These piRNA clusters are spread throughout genomes, and contain sequences complementary to their target TEs [60, 48]. They behave as genomic regions able to trap and regulate the activity of mobile sequences.

The regulation of a TE family is expected to start with the random insertion of an active TE sequence into a piRNA cluster. Numerical simulations have shown that such a "trap model" [31] was convincingly explaining TE dynamics, at least qualitatively [41, 28, 32], and mathematical approaches confirmed that the expected equilibrium TE copy number depends mostly on the number of clusters, and not on TE activity [55]. This expectation differs qualitatively from earlier models, which assumed a physiological decrease of the transposition rate with the number of copies [13, 37], and justifies to ground population genetic models into this new regulation paradigm.

The trap model expects that transposition is robustly silenced at equilibrium, where a large proportion of the population genotypes carry regulatory alleles (TE insertions) at one or several clusters. As a consequence, the genome of the species becomes "immune" to the invasion by the TE family. Conciliating genome immunity and recurrent horizontal transfers requires that TEs can escape immunity, either because pi-RNA immunity fades out with time, or because TE sequences from different TE lineages can diverge enough to escape recognition by related pi-RNA from a previous TE invasion. This last possibility has rarely been addressed quantitatively, and will be the focus of this paper, combining a theoretical approach and an empirical analysis. In the theoretical part, we performed individual-based simulations featuring two TE families invading the genome of a species, and compared their amplification dynamics

with and without cross-regulation. We backed up these simulations with two bioinformic analyses: at the intra-specific scale, we scanned the genome of *Drosophila melanogaster* for large copy-number TE families susceptible to display several waves of invasion; and at the interspecific scale, we studied the pattern of possible TE horizontal transfers between different species of *Drosophila*.

Results

Theoretical simulations

First, we explored theoretically the consequences of TE cross-regulation through individual-based simulations. TE regulation follows a pi-RNA "trap model", in which transposition regulation is triggered by the insertion of a TE copy in a pi-RNA cluster. We simulated the successive invasion of two related TE families α and β; both families have identical features (same transposition rate: $u_\alpha = u_\beta = 0.05$ replicative transpositions per copy and per generation, and same deleterious effect on the host fitness: $s = 0.005$ per copy). TE families only differ by the order in which they are introduced in the genome (α at generation 0, β at generation $H \geq 0$). The strength of cross-regulation η stands for the relative effect of pi-RNA insertions of one family on the other family. When $\eta > 0$, any TE insertion from either TE family (α and β) into a piRNA cluster will reduce the transposition rate of the other family ($\eta = 0$: no cross regulation, $\eta = 1$: α and β cross-regulated as if they belong to the exact same TE family).

When $\eta = 0$ both TEs are regulated independently, and their invasion dynamics are identical (Figure 1b), with a shift in time when TE β is transferred H generations after TE α (Figure 1a). The dynamics of copy number confirms earlier "trap model" simulations [31] as well as mathematical expectations [55]; the copy number increases exponentially during the very first generations, and transposition slows down with the accumulation of insertions in pi-RNA clusters. There was a transient maximum copy number, which is theoretically expected when TE insertions in piRNA are deleterious [55]. The number of copies eventually stabilizes in a state called "transposition-selection cluster balance" in Kofler [31], reaching this equilibrium (when the vast majority of genotypes in the population carries at least one regulatory allele) can take several hundred generations. In contrast, when cross-regulation is total between both families ($\eta = 1$), TE copies from α and β had the same transposition rate at a given generation. When α was introduced 1000 generations before β (Figure 1d), the dynamics of α was unaltered, but β could not invade the genome (the α pi-cluster insertions prevented β to transpose). When both TEs were introduced at the same time, they directly competed with each other and their total copy number stabilized at the expected equilibrium value, equally shared among α and β (Figure 1e). The ratio of of α vs. β copies at equilibrium for intermediate values of H and η is displayed in figure 1c. Only very strong cross-regulation $\eta > 0.9$ prevented TE β to invade. When $0 < \eta < 0.9$, β could invade and reach copy numbers of the same order of magnitude than α. Therefore, in the trap model setting, partial cross-regulation is unlikely to have a drastic effect on the long-term genome content.

The non-linear effect of the cross-regulation coefficient on the invasion of new-coming TE families is illustrated in Figure 2. When TE β invaded 300 generations after α, the maximum copy number of α was barely affected, as it was already settled in the genome at the arrival

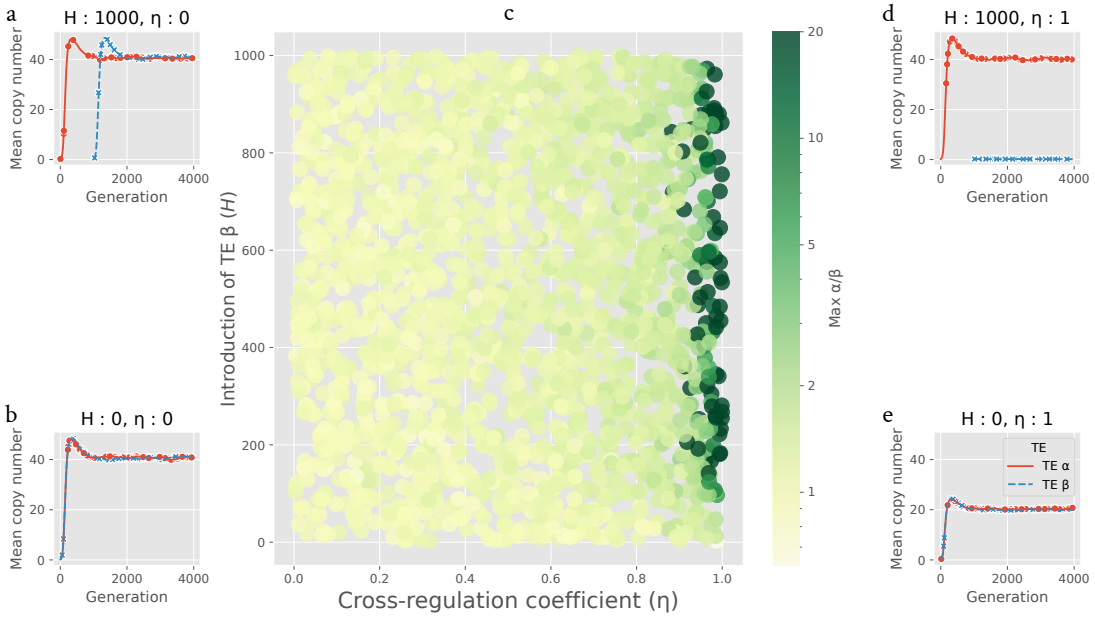


Figure 1: Influence of cross-regulation on the dynamics of related TE families. Small panels (a),(b),(d),(e) show the simulated evolution of average copy number for TEs α (red) an β (blue) for extreme parameter combinations. $\eta = 0$: no cross-regulation, TE dynamics are independent; $\eta = 1$ full cross-regulation, TEs α and β are regulated by the same pool of piRNAs; $H = 0$: α and β were introduced simultaneously, $H = 1000$: β was introduced 1000 generations after α. Central panel (c): ratio between the maximum number of TE α and β as a function of η and H; the color scale ranges from about 1 (as many α as β) to > 20 (20 times more α than β. Each dot stands for a single simulation, values for η and H were sampled in uniform distributions for each simulation.

of β. In contrast, after 300 generations, α was already strongly regulated and produced a substantial amount of piRNA, which reduced the effective transposition rate of cross-regulated β.

Genomics approach in Drosophila

In order to back-up these theoretical predictions, we performed a bioinformatics study in the model species Drosophila and related species. Two related questions were addressed: (i) given the piRNA produced in *Drosophila melanogaster*, would a TE from another Drosophila species be able to actively transpose if horizontally-transfered?, and (ii) is it possible to find in the genome of *Drosophila melanogaster* two or more closely-related TE groups that are differentially regulated? Based on the simulation model, our two main expectations were the following: (E1) piRNA immunity should prevent TE horizontal transfers from very

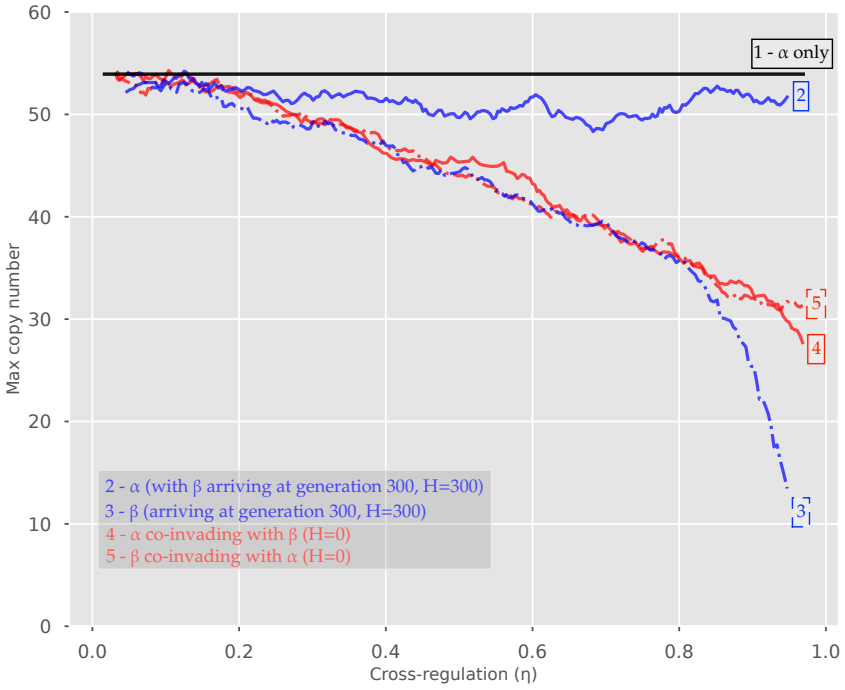


Figure 2: (a) Effect of the cross-regulation parameter η on the maximum number of copies for TEs α and β, when introduced at the same time (red lines 4 and 5) or when β invades $H = 300$ generations after α (blue lines 2 and 3). The figure displays the average copy number for a sliding window of 25 simulations (250 simulations in total, with η sampled in a uniform distribution). The copy number for α alone is provided as a reference (1 - black line).

closely related species only, as divergent TE sequences may easily escape regulation, and (E2) successive invasions from slightly divergent TE lineages from the same family should be observed.

Regulation of incoming TEs from other species To find evidence of cross-regulation, we investigated to what extent piRNA complement was shared between different *Drosophila* species. We used a publicly available dataset containing piRNA information sequenced from *Drosophila melanogaster* ovaries [23]. These sequenced reads were then aligned to TE copies in different *Drosophila* species to ascertain the level of *D. melanogaster* piRNA aligning and potentially regulating TEs from other species. The TEs in other species were identified using BLAST+ [10] utilizing the *D. melanogaster* TE library [4]. Our expectation was that TEs from *Drosophila* species will share less piRNA as the phylogenetic distance between them increases, but we also wanted to figure out at what phylogenetic scale the piRNA matching vanished, allowing potentially-unregulated horizontal transfers.

We indeed observed that the best match of *D. melanogaster* piRNAs was against *D. melanogaster* TEs and TEs from very closely related species (from the "melanogaster" subgroup) (Figure 3). Species outside of the melanogaster subgroup generally displayed no hit against *D. melanogaster* piRNAs, even when the TE family was present. Interestingly, many (but not all) TEs in

species from the distantly-related genus *Zaprionus* shared higher than expected similarity with *D. melanogaster* piRNA. Specifically, 12 TE families - *Copia, Copia, DM412, Doc, FW, G2, Gypsy4, Gypsy, Helena, Micropia* and *Stalker4* exhibited strong cross-mobilisation of piRNA from *D. melanogaster* to members of *Zaprionus* group, indicating possible horizontal transfers. We reconstructed the TE sequence phylogenetic tree for two of these families (Supplementary figure 1), and we could confirm that these spikes in potential piRNA cross-regulation was a signature of independent TE horizontal transfers. For TE families *Copia* and *Gypsy*, the TE copies from the *Zaprionus* genus were either nested or sister groups to to their *D. melanogaster* counterparts, which cannot be explained if vertically inherited. These horizontal transfers are plausible as *Drosophila* species from the *melanogaster* subgroup and flies from *Zaprionus* genus share similar habitats in the same geographical area in Africa [44]. In other TE families, such as R2 and 1731, TE sequence trees were close to the species phylogeny and thus compatible with vertical transmission.

Differential regulation TE lineages from the same family TE sequences evolve within the genome, and we expect some differences in piRNA shared between different lineages of the same TE family. This would lead to a partial cross-regulation situation, similar to our previous simulations with intermediate η values. We identified 65 TE families from *Drosophila melanogaster* displaying several full-length copies (at least 70% of the canonical sequence) with a maximum divergence between 5% and 20%.

Figure 4 illustrates the case of element Gypsy1 (a LTR class I TE), which displays a typical double-burst pattern, similar to our α and β lineages in the simulations. Assuming a regular molecular clock (phylogenetic trees were mid-point rooted), the following scenario can be reconstructed. A putative ancestral transposition burst left no surviving full-length copies, but could be detected based on piRNA cluster sequences (tagged 63, 43, and 85 in the figure). These very old piRNA clusters still produce piRNAs, as a small amount of small RNAs match their sequences. Another group of ancient piRNA clusters (88, 50, 54, 47, 49, and 40) may be the remnant of another ancestral TE burst, but they could also belong to the same group as piRNA 63 due to the uncertain tree rooting. These active clusters did not prevent another transposition burst, from which three full-length copies remain (labelled as 34, 17, and 10). Four piRNA cluster sequences were associated with this burst (piRNA 87, 86, 83, and 23); these piRNA cluster sequences are phylogenetically related to the full-length TEs, and the piRNAs they produce match the TE copies. Finally, a set of 9 full-length copies (78, 66, and related TE sequences) correspond to a more recent transposition burst, associated with a very active piRNA cluster (84).

Generalizing to the 65 TE families in *Drosophila melanogaster*, we observed diverse scenarios featuring one to four transposition bursts, with a large diversity in terms of sequence divergence, number of TE copies, and number of piRNA clusters. Even if the evolutionary scenario was not always easy to reconstruct (possibly because of the approximate rooting of phylogenetic trees), the following trends were retrieved consistently: (i) TE bursts in spite of pre-existing, active piRNA clusters were very frequent (at least 40 out of 65 TE families); (ii) Multiple successive bursts featuring 2 to 4 identifyable TE lineages were common (13/65 TE families); (iii) In more than half the multiple-burst cases (8 TE families /13), piRNA regulation was burst-specific, suggesting that the new transposition wave was associated with a TE sequence change allowing to escape piRNA regulation. This also means that the opposite situation (burst in spite of potentially-active piRNAs) was possible (supplementary table 1).

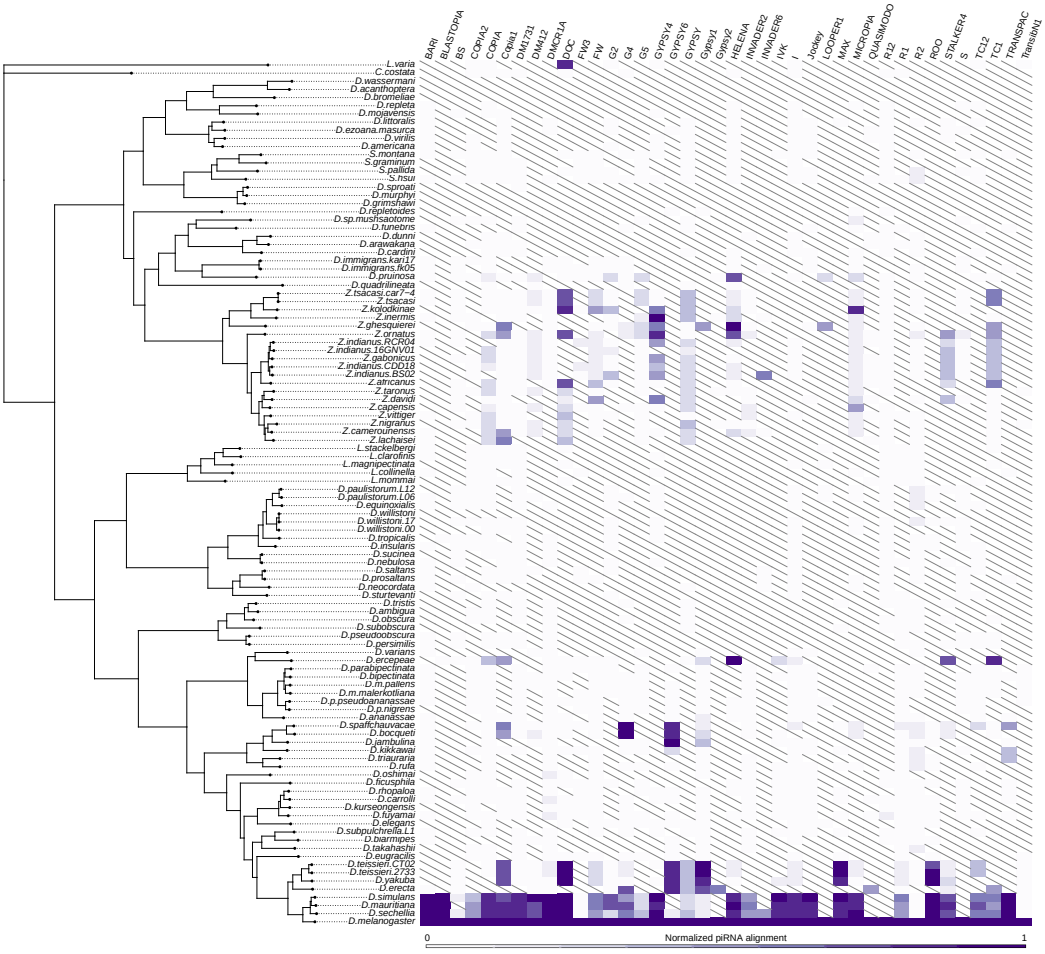


Figure 3: Analysis of *Drosophila melanogaster* piRNA aligning to TEs found in *Drosophila* and *Drosophila*-related species. Blanks correspond to TEs present in the species, but without any match to *Drosophila melanogaster* piRNAs.

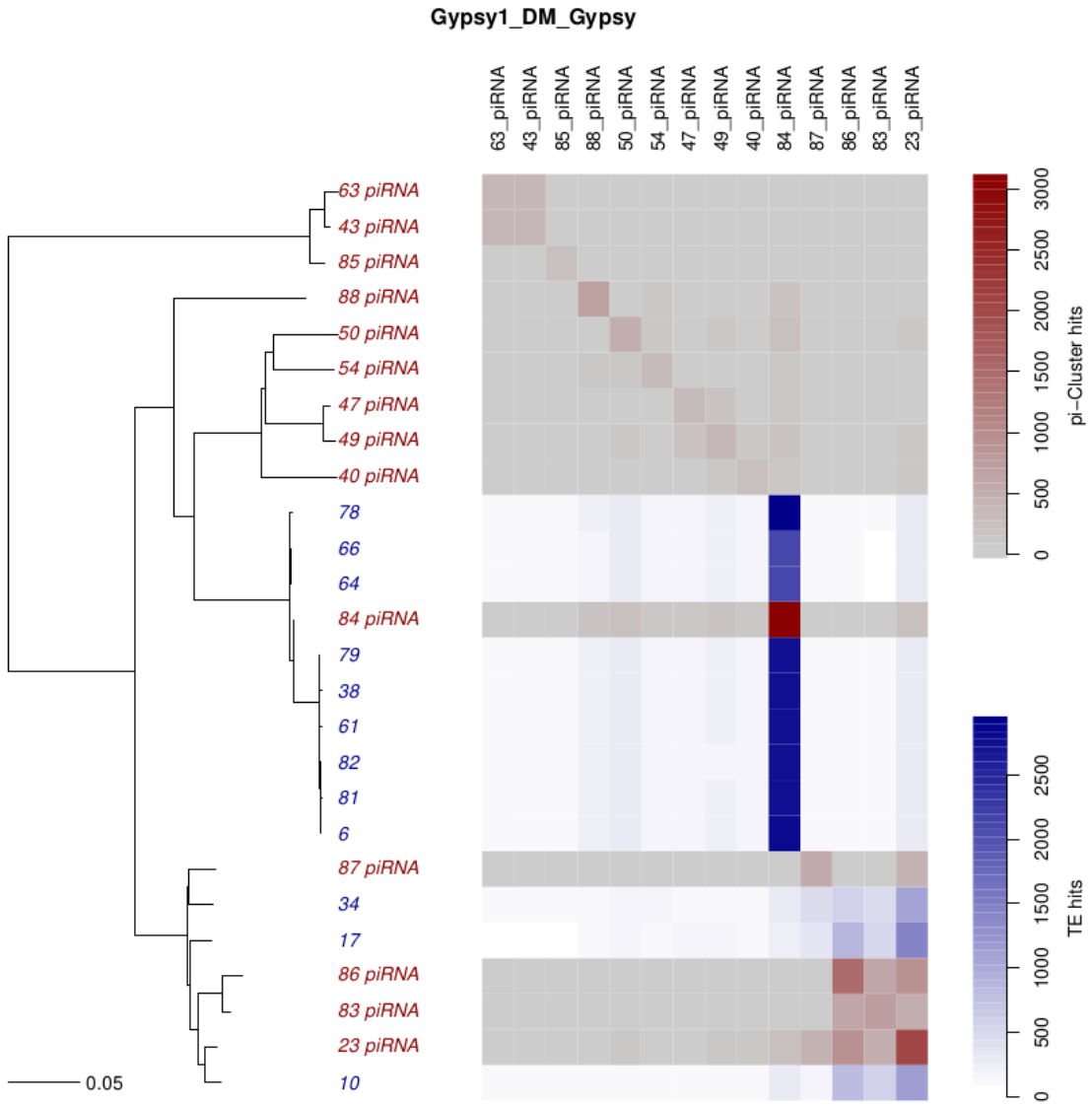


Figure 4: Analysis of shared piRNA aligning to Gypsy 1 copies within *Drosophila melanogaster*. Full-length TE copies (at least 70% of the canonical element) are in blue, sequences located in piRNA clusters are in red. Color intensity indicates the number of piRNA hits between each sequence (rows) and piRNA clusters (columns), which is a proxy for cluster activity (in red) and regulation strength (in blue). Sequence labels are arbitrary.

Discussion

Cross regulation in the piRNA trap model

Both simulations [31] and mathematical analysis [55] established that piRNA regulation generates some specific population genetics features, compared to copy-number regulation models [13]. In particular, the accumulation of TE copies does not depend a lot on the transposition rate: faster transposition creates regulatory piRNA earlier, and the final TE copy number tends to stabilize at around the same value. From our simulations, considering multiple TE families at once does not contradict these general conclusions on piRNA regulation. Our implementation of the piRNA trap model follows e.g. Kofler [31], and remains an abstraction of the complexity of the underlying regulation mechanisms (see recent discussions in e.g. [32, 22]). For instance, we considered that a single active TE-carrying cluster was enough to trigger total regulation, which is obviously an approximation. We also considered that piRNA clusters are the only contributors to TE regulation, which has been recently challenged [59]. Replacing the strict trap model by a quantitative regulation mechanism would increase the number of active clusters needed for TE silencing, and extend the persistence of piRNA regulation in the genome.

Here, we focused on close TE families, which can share partly or totally the pool of piRNA regulatory molecules, and thus cross-regulate each other. Simulations showed that in most cases, a moderate amount of cross-regulation is unlikely to prevent a new TE family to invade the genome, except if the rate of cross-regulation is very high (more than 90%). This observation is compatible with existing population genetics models. TE invasion is possible when the effective transposition rate is positive, i.e. new TE copies arising by transposition are more numerous than copies lost by deletion or eliminated by natural selection. Since the effective transposition rate of an incoming family is of the order of $u_e = u(1-\eta)$, invasion remains possible when $u_e > s$, which happens when $\eta < 1 - s/u$, i.e. $\eta < 0.9$ in our default parameter set (Table 1). From earlier analysis of the trap model [31, 55] the transposition rate is expected to affect the speed of the invasion, but barely changes the equilibrium copy number. This explains why the ratio α/β was close to 1 for most of the parameter space of the simulations (cross-regulation affects the speed, but not the outcome, of TE invasion). This scenario was also compatible with our within-species analysis in *Drosophila melanogaster*, as the presence of ancient TE lineages from the same family and their associated active piRNA clusters does not preclude new transposition bursts. As a consequence, our results confirmed that piRNA regulation can control the spread of TEs, but active "genome immunity" through piRNA regulation appears as a rather inefficient mechanism to control for even slightly divergent TEs.

Genomics analyses across Drosophila species confirmed that the "immunity" to horizontally-transfered TEs remains restricted to close species only. When the donor species was outside of the *Drosophila melanogaster* subgroup, there was virtually no match between *D. melanogaster* piRNA and transferable TE sequences – and thus, no possibility for piRNA regulation. The phylogenetic distance beyond which cross-regulation becomes unlikely can be roughly estimated: assuming a random distribution of substitutions, 93% of 26-nt piRNAs are expected to display at least one mismatch with the target TE sequence at 10% sequence divergence, and 99.5% at 20% sequence divergence. Obviously, in reality, piRNA regulation is likely to be more complex than a match/no match binary observation, as it could allow a limited amount of mismatches

between sequences [2], but our theoretical analysis showed how partial cross-regulation was unlikely to be efficient. Partial matching is thus likely to provide little protection against TE proliferation.

Interactions between TE families

Most species host several active TE families. While there has been a substantial effort in trying to understand TE invasion and regulation dynamics, both theoretically and empirically, much less is known about how different TE families interact, and whether these interactions facilitate or inhibit TE activity.

The diversity among transposition mechanisms, TE structure, and the specificity of TE regulation limits the opportunities of biochemical interactions among TEs from different families. A notable exception is the parasitic relationship between non-autonomous TEs (which have lost the capacity to produce some functional transposition machinery) and their autonomous counterparts. Non-autonomous TE lineages have been documented in all major TE groups, such as MITEs for class II transposons, or SINEs for non-LTR class I elements. Many non-autonomous TEs originate from a deletion within an autonomous copy, but there exists interesting examples (including *Alu* elements, the most common TE in the human genome) of phylogenetically-independent non-autonomous families. The presence of non-autonomous copies, which can titrate the transposition-related proteins without producing them, is theoretically predicted to affect substantially the dynamics of the corresponding full-length TEs [8, 36]. This interaction is also supported empirically; for instance Robillard et al. [52] showed that autonomous *mos1* copies (belonging to the *mariner* family) were almost completely silenced in presence of non-autonomous *peach* copies (see also [1]). In this last case, the sequence of *peach* elements was very close to the autonomous *mos1* [45], suggesting that undetectable, cryptic pairs of autonomous – non-autonomous TEs may be widespread.

More general would be a genome-scale interaction among TEs due to constraints related to the genetic load. By inserting randomly in potentially-important genomic regions and promoting rearrangements, TEs are mutagenic, and their activity is believed to decrease, in average, the host fitness [46]. If deleterious mutations were not independent, the presence of other active TE families would change the fitness effect of TEs, thus modifying their dynamics [13]. Even if measuring directly or indirectly the effects of TEs on fitness remains challenging, and that measurement of epistasis among TEs is equivocal [40], it is generally believed that mutations are expected to display negative epistasis (i.e., diminishing returns) on fitness [43, 16]. Negative epistasis on fitness implies that accumulating deleterious mutations have a stronger effect, thereby limiting the total amount of mutagenic factors (such as TEs) in the genome. However, there are so many TE families in the genome of most species that competition among TE families, if it exists, does not seem to have major evolutionary consequences. The reason why different TE families display highly variable TE copy numbers remains an open question.

Horizontal transfers and interspecies dynamics of TEs

The conflict between the host genome and TEs is ancient, and a form of small RNA-based regulation exists in most higher-order metazoans. The intra-genomic dynamics of TEs can hardly be dissociated from their interspecific distribution patterns, as the maintenance of active TE families heavily relies on frequent horizontal transfers [29]. The piRNA trap model fits

convincingly in this theoretical framework, as piRNA regulation specifically targets active TE families, and prevents temporarily its reinvasion in the same genome. When piRNA clusters and/or TE sequences have diverged sufficiently, a different TE clade from the same family can invade again, before jumping again to another species. Such a regular TE turnover fueled by horizontal transfers has been documented in most living organisms, including insects [49], vertebrates [61], or plants [21]. Here, we have reported strong evidence of multiple horizontal transfers, involving at least 12 TE families, between distant groups of African Drosophilidae (and in particular, between the *Zaprionus* genus and the *Drosophila menanogaster* subgroup). Given previous reports of frequent TE horizontal transfers among Drosophila [56], this pattern is not surprising, and supports the view that sympatric species may share a community of TEs that are frequently exchanged, although limited by the inertia of piRNA regulation.

Model and methods

Simulation model

We designed simulation model to explore the evolutionary dynamics of two piRNA-regulated transposable element families (α and β) in the genome of a random-mating population.

Population model The host species (a hermahrodite, diploid population of constant size N) evolved according to a traditional Wright-Fisher model; generations were non-overlapping. The model was individual-based, each individual had its own genotype, featuring explicitly the location of TEs. The genome consisted in $T = 1,000$ potential TE insertion sites regularly distributed on $k = 5$ chromosomes. A fraction π of these insertion sites was considered as belonging to pi-clusters and involved in TE regulation. The recombination rate between consecutive sites on the same chromosome was $r = 0.001$ (so, 20cM for each chromosome), recombination was free among chromosomes. Every generation, N new offspring were generated by sampling $2N$ parents in the population, with a probability proportional to their fitness. Each parent gave a haploid gamete, recombination probabilities being determined from the genetic distance between heterozygous sites, and both gametes were merged to form a new individual. A generation consisted in the succession of transposition and reproduction stages; fitness was calculated after transposition.

Transposition TEs from lineages α and β (n_α and n_β copies, respectively) were featured by maximum replicative transposition rates of u_α and u_β, respectively (transposition rates were expressed as the number of new copies per copy and per generation). Transposition rates were affected by the number of TE copies m_α and m_β inserted in regulatory piRNA clusters, the effective transposition rates being $u^e_\alpha = u_\alpha(1 - R_\alpha - \eta R_\beta)$ and $u^e_\beta = u_\beta(1 - R_\beta - \eta R_\alpha)$, respectively, where $R_\alpha = \min(1, m_\alpha)$ and $R_\beta = \min(1, m_\beta)$. R_α and R_β can be interpreted as the strength of regulation for TEs α and β, respectively. Here we considered for simplicity that the presence of a single TE copy in a piRNA cluster was enough to silence transposition. Transposition rates could not be negative and were set to 0 whenever necessary. The cross-regulation coefficient η quantified how α piRNAs regulate β, and vice versa.

In each individual, the actual number of new transpositions for TEs α (resp. β) was drawn in Poisson distributions of means $n_\alpha u^e_\alpha$ (resp. $n_\beta u^e_\beta$). The new insertion sites were drawn

uniformly among the $2T$ sites of the diploid genome (including $2T\pi$ piRNA cluster sites); TE copies at occupied sites were replaced.

Selection The presence of TEs induced a fitness cost of s per copy. In presence of $n = n_\alpha + n_\beta$ copies, the fitness value in the default model was $w = \exp(-ns)$, i.e., TEs had independent insertion effects (constant effect on log fitness).

Simulation scenario At generation 0, TE α was introduced as $n_{\alpha 0} = 0.2$ copies on average per individual (enough to ensure that the TE cannot be lost by genetic drift), randomly distributed in the genome, excluding piRNA cluster sites. After H generations, TE β was introduced as $n_{\beta H} = 0.2$ copies on average per individual, using the same procedure. A realistic horizontal transfer scenario would suggest $n_{\beta H} = 1/2N$, but the loss rate of β would be very high and require too many replicated horizontal transfer attempts to be computationally tractable.

Table 1: Model parameters and symbols

Parameter	Meaning	Default value
N	Population Size	2000
T	Number of genomic insertion sites	1000
k	Number of chromosomes	5
r	Recombination rate between adjacent insertion sites	0.1%
H	Horizontal transfer generation	
π	size of piRNA region(as % genome)	3%
η	Cross-regulation coefficient	
$n_{\alpha 0}$	Initial number of copies per individual for TE lineage α	0.2
$n_{\beta 0}$	Number of copies per individual for TE lineage β after horizontal transfer	0.2
u_α	Maximum transposition rate for TE lineage α	0.05
u_β	Maximum transposition rate for TE lineage β	0.05
s	Selection coefficient	0.005

Implementation The model was implemented using a Python-based library (Simulicron). Summary statistics, including average TE copy numbers and average transposition rates, were calculated and stored at regular intervals. Figures were generated with Python graphical library Seaborn[57]. The scripts used for simulation are available at https://github.com/siddharthst/Simulicron.

Bioinformatics analysis

Identification of genomic TE copies We designed a species-agnostic bioinformatics pipeline to discover potential evidence of piRNA cross-regulation between TEs insertions of the same family in the *D. melanogaster*. The pipeline uses an unmasked *D. melanogaster* genome downloaded from the Ensembl project[Release 105] [17]. For the discovery of TE copies, NCBI BLAST+ [10] was used with parameters `-outfmt 6 -evalue 20`. The search queries for nucleotide BLAST+ comprised of in-lab curated *D. melanogaster* TE family database derived from RepBase [4]. Some TEs might contain insertions, which will be reported as partial

matches by BLAST+. To account for small insertions (<250nt), we use Bedtools Merge [51] to merge all fragments of the same family if they are 250nt apart and on the same strand.

piRNA-cluster identification and mapping The piRNA-cluster coordinates were taken from a previous study[7]. Since these clusters were annotated on an earlier version of *D. melanogaster* genome assembly (R5), BLASTn was used to identify the respective cluster loci in the current release of *D. melanogaster* genome (R6). We used a publicly available small RNA dataset [24] sequenced from the ovaries of *D. melanogaster* strain w1118(SRR14569563). The reads were then processed using TrimGalore [34] with default parameters to remove any sequencing adapters. Furthermore, additional filtering was performed to remove any tRNA and rRNA sequences using BBMap [9]. Alignment pf the piRNAs to TE copies was performed using Bowtie [35]. Specifically, the parameter $-v$ 0 was used to disallow any mismatches, and the $-a$ option was used to report all valid alignments for each sequencing read.

piRNAs shared between piCluster and copies Written in R, the pipeline identifies the piRNA shared between piRNA clusters and TEs. Each TE copy of a TE family is aligned to the consensus sequence, using mafft [27], and all the insertions in the TE copy that are not present in the consensus were removed. The resulting alignment was then used to create a phylogenetic tree using fasttree (options -nt -gtr)[50]. The script isolates TE insertions in the annotated piRNA clusters if their coordinates overlap. The script then assigns the piRNA from the piRNA cluster insertions to the potential target non-cluster TE insertions using the alignment information. Only near full-length TE copies (at least 70% of the consensus length) and piRNA insertions with more than 200 aligned piRNA were kept; TE families with less than 6 full-length copies or less than 2 cluster insertions were discarded. This filtering procedure is detailed in supplementary table 2.

piRNAs shared between *D. melanogaster* and *Drosophilidae* A second pipeline was used to detect signatures of cross-regulation between the piRNA complement of *D. melanogaster* and other members of *Drosophilidae*. We used the same public dataset for small RNA reads, and acquired the 101 Drosophilid genomes from a previously published dataset [30]. We extracted the genomic TE copies from each species (if found) using RepeatMasker[54] (options `-s -gff -no_is -nolow -norna -div 40`) and kept the full-length copy with the most aligned piRNA. This second pipeline thus reports the amount of *D. melanogaster* piRNA reads matching the most targeted TE copy in various Drosophilidae. The complete results can be found in Supplementary Table 3 which can be accessed online at https://github.com/siddharthst/Simulicron/blob/master/Supplementary_data/Supplementary_Table_3.csv. Supplementary Figure 2 describes the steps taken by the bioinformatics pipeline. The scripts to reproduce the results are available at https://github.com/siddharthst/Simulicron/tree/master/Bioinformatics.

References

[1] Daniel De Anguiar and Daniel L Hartl. "Regulatory potential of nonautonomous mariner elements and subfamily crosstalk". In: *Genetica* 107.1 (1999), pp. 79–85.

[2] Todd A. Anzelon et al. “Structural basis for piRNA targeting”. In: *Nature* 597.7875 (Sept. 2021), pp. 285–289. DOI: 10.1038/s41586-021-03856-x. URL: https://doi.org/10.1038/s41586-021-03856-x.

[3] Alexei A. Aravin, Gregory J. Hannon, and Julius Brennecke. “The Piwi-piRNA Pathway Provides an Adaptive Defense in the Transposon Arms Race”. In: *Science* 318.5851 (Nov. 2007), pp. 761–764. DOI: 10.1126/science.1146484. URL: https://doi.org/10.1126/science.1146484.

[4] Weidong Bao, Kenji K. Kojima, and Oleksiy Kohany. “Repbase Update, a database of repetitive elements in eukaryotic genomes”. In: *Mobile DNA* 6.1 (June 2015). DOI: 10.1186/s13100-015-0041-9. URL: https://doi.org/10.1186/s13100-015-0041-9.

[5] Maite G Barrón et al. “Population genomics of transposable elements in Drosophila”. In: *Annual review of genetics* 48 (2014), pp. 561–581.

[6] C Biémont. “Dynamic equilibrium between insertion and excision of P elements in highly inbred lines from an M strain of *Drosophila melanogaster*”. In: *Journal of Molecular Evolution* 39.5 (1994), pp. 466–472.

[7] Julius Brennecke et al. “Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila”. In: *Cell* 128.6 (Mar. 2007), pp. 1089–1103. DOI: 10.1016/j.cell.2007.01.043.

[8] JFY Brookfield. “Models of the spread of non-autonomous selfish transposable elements when transposition and fitness are coupled”. In: *Genetics Research* 67.3 (1996), pp. 199–209.

[9] Brian Bushnell. “BBMap: A Fast, Accurate, Splice-Aware Aligner”. In: (Mar. 2014). URL: https://www.osti.gov/biblio/1241166.

[10] Christiam Camacho et al. “BLAST+: architecture and applications”. en. In: *BMC Bioinformatics* 10.1 (Dec. 2009), p. 421.

[11] Juliana P Castro and Claudia Carareto. “Drosophila melanogaster P transposable elements: mechanisms of transposition and regulation”. In: *Genetica* 121.2 (2004), pp. 107–118.

[12] B Charlesworth and CH Langley. “The evolution of self-regulated transposition of transposable elements”. In: *Genetics* 112.2 (1986), pp. 359–383.

[13] Brian Charlesworth and Deborah Charlesworth. “The population dynamics of transposable elements”. In: *Genetics Research* 42.1 (1983), pp. 1–27.

[14] Brian Charlesworth, Paul Sniegowski, and Wolfgang Stephan. “The evolutionary dynamics of repetitive DNA in eukaryotes”. In: *Nature* 371.6494 (1994), pp. 215–220.

[15] Rachel L Cosby, Ni-Chen Chang, and Cédric Feschotte. “Host–transposon interactions: conflict, cooperation, and cooption”. In: *Genes & development* 33.17-18 (2019), pp. 1098–1116.

[16] Alejandro Couce and Olivier A Tenaillon. “The rule of declining adaptability in microbial evolution experiments”. In: *Frontiers in genetics* 6 (2015), p. 99.

[17] Fiona Cunningham et al. “Ensembl 2022”. In: *Nucleic Acids Research* 50.D1 (Nov. 2021), pp. D988–D995. DOI: 10.1093/nar/gkab1049. URL: https://doi.org/10.1093/nar/gkab1049.

[18] Stephen B Daniels et al. “Evidence for horizontal transmission of the P transposable element between Drosophila species.” In: *Genetics* 124.2 (1990), pp. 339–355.

[19] W Ford Doolittle and Carmen Sapienza. “Selfish genes, the phenotype paradigm and genome evolution”. In: *Nature* 284.5757 (1980), pp. 601–603.

[20] Moaine El Baidouri et al. “Widespread and frequent horizontal transfers of transposable elements in plants”. In: *Genome research* 24.5 (2014), pp. 831–838.

[21] Moaine El Baidouri et al. “Widespread and frequent horizontal transfers of transposable elements in plants”. In: *Genome research* 24.5 (2014), pp. 831–838.

[22] D. Gebert et al. “Large Drosophila germline piRNA clusters are evolutionarily labile and dispensable for transposon regulation”. In: *Mol Cell* 81.19 (Oct. 2021), pp. 3965–3978.

[23] Daniel Gebert et al. “Large Drosophila germline piRNA clusters are evolutionarily labile and dispensable for transposon regulation”. In: *Molecular Cell* 81.19 (Oct. 2021), 3965–3978.e5. DOI: 10.1016/j.molcel.2021.07.011. URL: https://doi.org/10.1016/j.molcel.2021.07.011.

[24] Clement Gilbert and Richard Cordaux. “Horizontal transfer and evolution of prokaryote transposable elements in eukaryotes”. In: *Genome biology and evolution* 5.5 (2013), pp. 822–832.

[25] Allen G Good et al. “Rapid spread of transposable p elements in experimental populations of Drosophila melanogaster.” In: *Genetics* 122.2 (1989), pp. 387–396.

[26] Robert Hubley et al. “The Dfam database of repetitive DNA families”. In: *Nucleic Acids Research* 44.D1 (Nov. 2015), pp. D81–D89. DOI: 10.1093/nar/gkv1272. URL: https://doi.org/10.1093/nar/gkv1272.

[27] K. Katoh. “MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform”. In: *Nucleic Acids Research* 30.14 (July 2002), pp. 3059–3066. DOI: 10.1093/nar/gkf436. URL: https://doi.org/10.1093/nar/gkf436.

[28] Erin S Kelleher, Ricardo BR Azevedo, and Yichen Zheng. “The evolution of small-RNA-mediated silencing of an invading transposable element”. In: *Genome biology and evolution* 10.11 (2018), pp. 3038–3057.

[29] Margaret G Kidwell and Damon R Lisch. “Perspective: transposable elements, parasitic DNA, and genome evolution”. In: *Evolution* 55.1 (2001), pp. 1–24.

[30] Bernard Y Kim et al. “Highly contiguous assemblies of 101 drosophilid genomes”. In: *eLife* 10 (July 2021). DOI: 10.7554/elife.66405. URL: https://doi.org/10.7554/elife.66405.

[31] Robert Kofler. “Dynamics of Transposable Element Invasions with piRNA Clusters”. In: *Mol Biol Evol* 36.7 (July 2019), pp. 1457–1472. DOI: 10.1093/molbev/msz079.

[32] Robert Kofler, Viola Nolte, and Christian Schloetterer. “The transposition rate has little influence on equilibrium copy numbers of the P-element”. In: *bioRxiv* (2021). DOI: 10.1101/2021.09.20.461050.

[33] Robert Kofler et al. “The recent invasion of natural Drosophila simulans populations by the P-element”. In: *Proceedings of the National Academy of Sciences* 112.21 (2015), pp. 6659–6663.

[34] Felix Krueger et al. *TrimGalore: v0.6.10*. 2023. DOI: 10.5281/ZENODO.7598955. URL: https://zenodo.org/record/7598955.

[35] Ben Langmead et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome". In: *Genome Biology* 10.3 (2009), R25. DOI: 10.1186/gb-2009-10-3-r25. URL: https://doi.org/10.1186/gb-2009-10-3-r25.

[36] Arnaud Le Rouzic and Pierre Capy. "Population genetics models of competition between transposable element subfamilies". In: *Genetics* 174.2 (2006), pp. 785–793.

[37] Arnaud Le Rouzic and Pierre Capy. "The first steps of transposable elements invasion: parasitic strategy vs. genetic drift". In: *Genetics* 169.2 (2005), pp. 1033–1043.

[38] Arnaud Le Rouzic and Grégory Deceliere. "Models of the population genetics of transposable elements". In: *Genet Res* 85.3 (June 2005), pp. 171–81. DOI: 10.1017/S0016672305007585.

[39] Yuh Chwen G Lee. "Synergistic epistasis of the deleterious effects of transposable elements". In: *Genetics* 220.2 (2022), iyab211.

[40] Yuh Chwen G Lee. "Synergistic epistasis of the deleterious effects of transposable elements". In: *Genetics* 220.2 (2022), iyab211.

[41] Jian Lu and Andrew G Clark. "Population dynamics of PIWI-interacting RNAs (piRNAs) and their targets in Drosophila". In: *Genome research* 20.2 (2010), pp. 212–227.

[42] Maartje J Luteijn and René F Ketting. "PIWI-interacting RNAs: from generation to transgenerational epigenetics". In: *Nature Reviews Genetics* 14.8 (2013), pp. 523–534.

[43] Guillaume Martin, Santiago F Elena, and Thomas Lenormand. "Distributions of epistasis in microbes fit predictions from a fitness landscape model". In: *Nature genetics* 39.4 (2007), pp. 555–560.

[44] K. Maruyama and D. L. Hartl. "Evidence for interspecific transfer of the transposable element mariner between Drosophila and Zaprionus". In: *J Mol Evol* 33.6 (Dec. 1991), pp. 514–524.

[45] M Medhora, K Maruyama, and D L Hartl. "Molecular and functional analysis of the mariner mutator element Mos1 in Drosophila." In: *Genetics* 128.2 (June 1991), pp. 311–318. DOI: 10.1093/genetics/128.2.311. URL: https://doi.org/10.1093/genetics/128.2.311.

[46] Sergey V Nuzhdin. "Sure facts, speculations, and open questions about the evolution of transposable element copy number". In: *Transposable Elements and Genome Evolution*. Springer, 2000, pp. 129–137.

[47] Leslie E Orgel and Francis HC Crick. "Selfish DNA: the ultimate parasite". In: *Nature* 284.5757 (1980), pp. 604–607.

[48] Deniz M Ozata et al. "PIWI-interacting RNAs: small RNAs with big functions". In: *Nat Rev Genet* 20.2 (Feb. 2019), pp. 89–108. DOI: 10.1038/s41576-018-0073-3.

[49] Jean Peccoud et al. "Massive horizontal transfer of transposable elements in insects". In: *Proceedings of the National Academy of Sciences* 114.18 (2017), pp. 4721–4726.

[50] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments". In: *PLoS ONE* 5.3 (Mar. 2010). Ed. by Art F. Y. Poon, e9490. DOI: 10.1371/journal.pone.0009490. URL: https://doi.org/10.1371/journal.pone.0009490.

[51] Aaron R. Quinlan and Ira M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6 (Jan. 2010), pp. 841–842. DOI: 10.1093/bioinformatics/btq033. URL: https://doi.org/10.1093/bioinformatics/btq033.

[52] Émilie Robillard et al. "Experimental evolution reveals hyperparasitic interactions among transposable elements". In: *Proc Natl Acad Sci U S A* 113.51 (Dec. 2016), pp. 14763–14768. DOI: 10.1073/pnas.1524143113.

[53] Joana C Silva and Margaret G Kidwell. "Horizontal transfer and selection in the evolution of P elements". In: *Molecular Biology and Evolution* 17.10 (2000), pp. 1542–1557.

[54] A.F.A Smit, R. Hubley, and P. Green. *Repeatmasker*. URL: https://www.repeatmasker.org/ (visited on 04/05/2023).

[55] Siddharth S. Tomar, Aurélie Hua-Van, and Arnaud Le Rouzic. "A population genetics theory for piRNA-regulated transposable elements". In: *Theoretical Population Biology* 150 (Apr. 2023), pp. 1–13. DOI: 10.1016/j.tpb.2023.02.001. URL: https://doi.org/10.1016/j.tpb.2023.02.001.

[56] Gabriel Luz Wallau et al. "VHICA, a new method to discriminate between vertical and horizontal transposon transfer: Application to the mariner family within Drosophila". In: *Molecular biology and evolution* 33.4 (2016), pp. 1094–1109.

[57] Michael L. Waskom. "seaborn: statistical data visualization". In: *Journal of Open Source Software* 6.60 (2021), p. 3021. DOI: 10.21105/joss.03021. URL: https://doi.org/10.21105/joss.03021.

[58] Thomas Wicker et al. "A unified classification system for eukaryotic transposable elements". In: *Nature Reviews Genetics* 8.12 (Dec. 2007), pp. 973–982. DOI: 10.1038/nrg2165. URL: https://doi.org/10.1038/nrg2165.

[59] Filip Wierzbicki and Robert Kofler. "The composition of piRNA clusters in *Drosophila melanogaster* deviates from expectations under the trap model". In: *bioRxiv* (2023), pp. 2023–02.

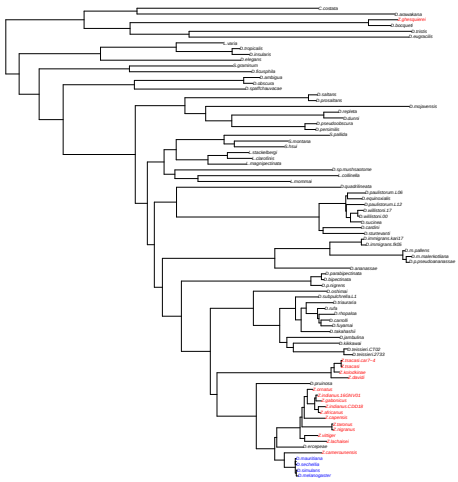
[60] Vanessa Zanni et al. "Distribution, evolution, and diversity of retrotransposons at the *flamenco* locus reflect the regulatory properties of piRNA clusters". In: *Proceedings of the National Academy of Sciences* 110.49 (2013), pp. 19842–19847. DOI: 10.1073/pnas.1313677110. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1313677110. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1313677110.

[61] Hua-Hao Zhang et al. "Horizontal transfer and evolution of transposable elements in vertebrates". In: *Nature communications* 11.1 (2020), pp. 1–10.

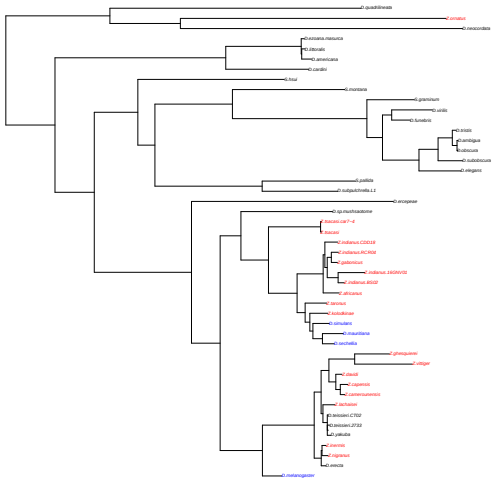
Supplementary material

Supplementary figure 1

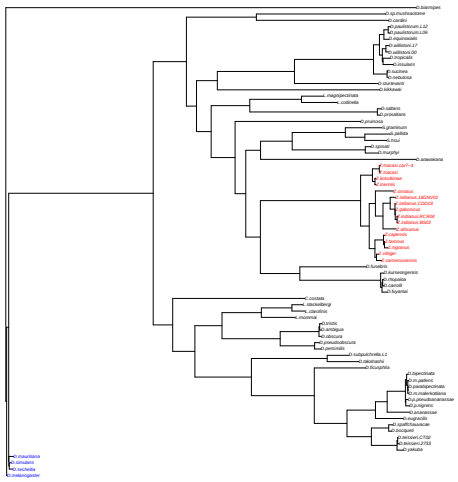
Copia



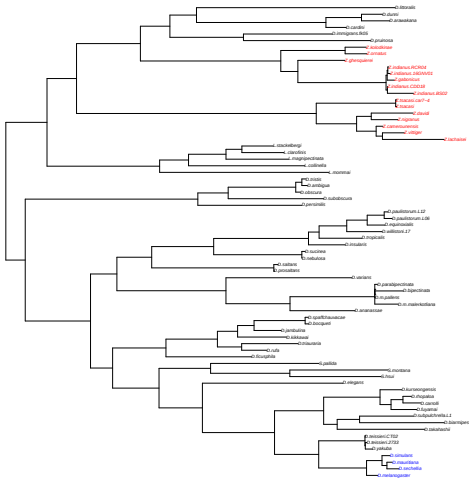
Gypsy



1731

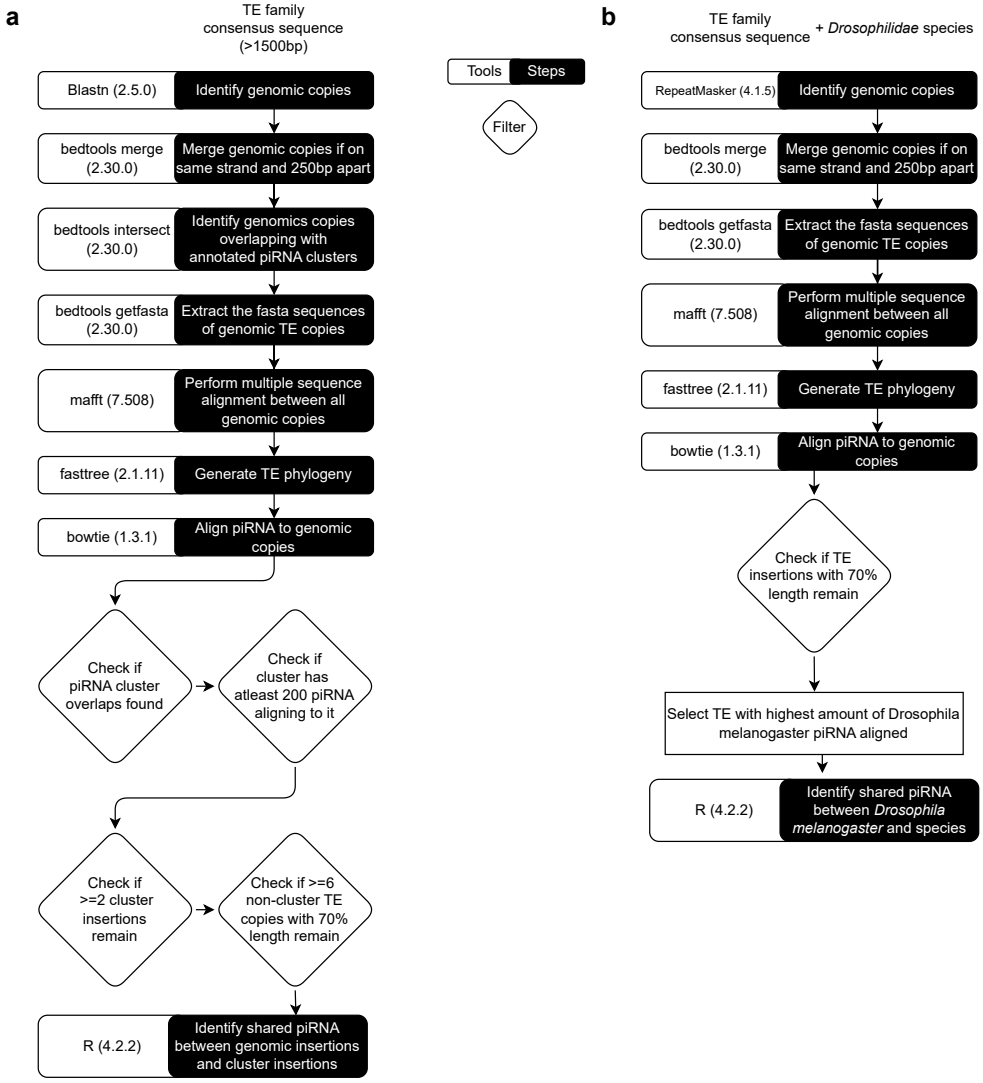


R2



Phylogeny of TE sequences (one TE sequence per species in which the TE family is present) for four typical families (Copia and R2 are non-LTR class I TEs, Gypsy and 1731 are LTR elements). Species belonging to the *melanogaster* subgroup are in blue, species from the *Zaprionus* genus are in red. The top row (Copia and Gypsy) display some horizontal transfer signal (TE copies from *D. melanogaster* and *Zaprionus* are clustered in the same clade), while the bottom row phylogenies (1731 and R2) are compatible with vertical transmission (the *melanogaster* subgroup and *Zaprionus* are far apart). Trees were midpoint-rooted (except 1731).

Supplementary figure 2



The pipelines used for bioinformatics analysis. a) Pipeline used for detecting signatures of piRNA cross-regulation in the TE complement of single species. b) Pipeline used to identify signatures of piRNA cross-regulation/genomic immunity between multiple species with a single species as a focus(the source of piRNA).

Supplementary table 1

Table describing the observed invasion history of 65 TE families with piRNA cluster insertions in *Drosophila melanogaster*

TE family	Invasion in spite of piRNA regulation	Multiple waves	Independent regulation
BAGGINS1_Loa	Yes	Perhaps	
Tc1mariner	No	No	
Batumi	Yes	Yes	No
Bel	No	No	
Bica_Gypsy	Yes	No	
Blood_gypsy	Yes	No	
BS2_Jockey	Yes	No	
BS3_Jockey	No	Yes	Perhaps
BS_Jockey	No	No	
Burdock	No	Yes	No
Chimpo	No	No	
Copia1	No	No	
Copia2	No	No	
Copia	No	No	
Diver2_bel	Yes	No	
Diver_bel	No	No	
1731_Copia	Yes	No	
176_Gypsy	Yes	No	
297_Gypsy	Yes	No	
412_Gypsy	Yes	No	
CR1A	Yes	No	
RT1A	Yes	No	
RT1B	Yes	Perhaps	Perhaps
Doc2	No	No	
Doc3	Yes	Yes	Yes
Doc6	No	No	
Doc	Yes	No	
FW	Yes	No	
G2_Jockey	Yes	Yes	Yes
G6_Jockey	Yes	No	
G_Jockey	No	Yes	Yes
GTWIN	Yes	No	
Gypsy1	Yes	Yes	Yes
Gypsy2	No	Yes	No
Gypsy3	Yes	Yes	Yes
Gypsy4	Yes	No	
Gypsy6	No	Yes	Yes
Gypsy	Yes	No	

Heta_Jockey	No	Yes	Yes
HSMBeagle	No	Yes	Yes
Idefix	No	No	
I	Perhaps	Yes	No
Invader2	Yes	Perhaps	No
Invader3	Yes	No	
Invader4	Yes	No	
IVK	Yes	Perhaps	Perhaps
Max_bel	Yes	Perhaps	No
McClintock	Yes	No	
MDG1	Yes	No	No
MDG3	Yes	No	
Micropia	Yes	No	
Nomad	Yes	No	
Pogo	No		
Protop	Yes	No	
Quasimodo2	Yes	No	
Quasimodo	Yes	No	
R1	Perhaps	No	
Roo	Yes	Perhaps	No
Rover	No	Perhaps	No
S	No	No	
Stalker2	Yes	No	
Syalker4	Yes	No	
Tabor	Yes	No	
TC12	No	No	
Transib2	No	Perhaps	Perhaps

Supplementary table 2

Table describing all the filtering steps taken during bioinformatics pipeline for processing TEs in *Drosophila melanogaster*

TE	TE Length	# TE copies	# TE copies post-filter	#Full length copies in piCluster	Is removed ?	# piClusters	# piClusters post filtering
McClintock_Gypsy	6450	93	8	1	No	10	7
ACCORD_Gypsy	7404	31	6	0	Yes/MinimumTEFailed	0	0
ACCORD2_Gypsy	7642	132	7	0	Yes/MinPiCLuster	18	1
BAGGINS1_Loa	5453	222	17	7	No	85	68
BARI_DM_Tc1mariner	1728	13	10	2	No	2	2
BATUMI_BEL	8507	388	18	5	No	70	68
BEL_BEL	6126	21	9	3	No	5	5
BLASTOPIA_LTR_Retrotransposon	0	0	0	0	Yes/NPi	0	0
BLOOD_Gypsy	7411	203	30	4	No	34	27
BS_Jockey	5126	55	9	1	No	7	5
BS2_Jockey	4721	147	13	6	No	31	20
BS3_DM_Jockey	1790	99	25	8	No	21	5
BURDOCK_Gypsy	6410	55	13	1	No	6	2
Bica_Gypsy	5107	40	14	2	No	6	4
CIRCE_LTR_Retrotransposon	6356	136	6	0	Yes/MinimumTEFailed	0	0
COPIA2_DM_Copia	4558	95	11	3	No	12	12
COPIA_DM_Copia	5146	79	43	3	No	5	5
Chimpo_Gypsy	4473	48	15	3	No	6	3
Chouto_Gypsy	6951	25	6	0	Yes/MinimumTEFailed	0	0

Copia1_DM_Copia	4530	29	7	1	No	2	2
DIVER_BEL	6091	33	17	0	No	2	2
DIVER2_BEL	4917	135	7	2	No	28	24
DM1731_Copia	4649	63	12	2	No	7	6
DM176_Gypsy	7439	163	30	2	No	15	5
DM297_Gypsy	6995	250	37	1	No	41	19
DM412_Gypsy	7439	218	32	3	No	27	18
DMCR1A_CR1	4470	654	36	5	No	150	87
DMRT1A_R1	5183	96	18	11	No	31	27
DMRT1B_R1	5183	232	40	14	No	72	51
DMRT1C_R1	5443	85	0	0	Yes/MinimumTEFailed	0	0
DOC_Jockey	4725	204	91	12	No	18	15
DOC2_DM_Jockey	4793	70	15	5	No	17	11
DOC3_DM_Jockey	4740	165	13	2	No	49	24
DOC4_DM_Jockey	2791	50	1	0	Yes/MinimumTEFailed	0	0
DOC5_DM_Jockey	4682	86	2	0	Yes/MinimumTEFailed	0	0
DOC6_DM_Jockey	4222	20	12	3	No	3	3
FB4_DM_Tc1mariner	4089	187	3	0	Yes/MinimumTEFailed	0	0
FROGGER_Copia	0	0	0	0	Yes/NPi	0	0
FW_Jockey	4692	275	105	31	No	61	55
FW2_DM_Jockey	3961	26	2	0	Yes/MinimumTEFailed	0	0
FW3_DM_Jockey	3132	19	1	0	Yes/MinimumTEFailed	0	0
G2_DM_Jockey	3102	152	17	5	No	16	15
G3_DM_Jockey	4605	51	1	0	Yes/MinimumTEFailed	0	0
G4_DM_Jockey	3856	61	5	0	Yes/MinimumTEFailed	0	0

G5A_DM_Jockey	2841	78	9	1	Yes/NoClusterAfterFilter	13	0
G5_DM_Jockey	4856	78	5	0	Yes/MinimumTEFailed	0	0
G6_DM_Jockey	2042	81	14	6	No	12	12
GTWIN_Gypsy	7411	74	7	2	No	21	5
GYPSY_Gypsy	7471	136	7	1	No	20	8
GYPSY10_Gypsy	6006	50	1	0	Yes/MinimumTEFailed	0	0
GYPSY11_Gypsy	4428	56	0	0	Yes/MinimumTEFailed	0	0
GYPSY12_Gypsy	10218	286	5	0	Yes/MinimumTEFailed	0	0
GYPSY12A_LTR_Gypsy	2280	250	2	0	Yes/MinimumTEFailed	0	0
GYPSY2_Gypsy	6841	98	1	0	Yes/MinimumTEFailed	0	0
GYPSY3_Gypsy	6973	90	11	2	No	29	11
GYPSY4_Gypsy	6862	50	17	7	No	24	23
GYPSY5_Gypsy	7369	40	4	0	Yes/MinimumTEFailed	0	0
GYPSY6_Gypsy	7826	65	10	2	No	11	7
GYPSY7_Gypsy	5486	58	5	0	Yes/MinimumTEFailed	0	0
GYPSY8_Gypsy	4955	71	1	0	Yes/MinimumTEFailed	0	0
GYPSY9_Gypsy	5349	35	1	0	Yes/MinimumTEFailed	0	0
G_DM_Jockey	4346	93	10	3	No	16	9
Gypsy1_DM_Gypsy	7543	99	14	2	No	31	14
Gypsy2_DM_Gypsy	7221	152	11	1	No	39	10
HETA_Jockey	6081	224	17	6	No	17	11
HMSBEAGLE_Gypsy	7061	186	15	1	No	36	9

HOBO_hAT	3016	48	1	0	Yes/MinimumTEFailed	0	0
IDEFIX_Gypsy	7410	323	26	1	No	23	8
INVADER1_Gypsy	4032	125	2	0	Yes/MinimumTEFailed	0	0
INVADER2_Gypsy	5124	142	31	4	No	19	8
INVADER3_Gypsy	5484	59	12	2	No	11	8
INVADER4_Gypsy	3105	68	12	2	No	25	15
INVADER5_Gypsy	4038	23	1	0	Yes/MinimumTEFailed	0	0
INVADER6_Gypsy	4885	20	6	0	Yes/MinimumTEFailed	0	0
IVK_DM_Jockey	5402	51	18	5	No	11	9
I_DM_I	5374	116	19	0	No	23	12
JOCKEY2_Jockey	3428	66	0	0	Yes/MinimumTEFailed	0	0
Jockey_Jockey	0	0	0	0	Yes/NPi	0	0
LOOPER1_DM_piggyBac	1881	13	2	0	Yes/MinimumTEFailed	0	0
MAX_BEL	8556	161	12	6	No	36	23
MDG1_Gypsy	7480	255	30	3	No	35	24
MDG3_DM_Gypsy	5520	42	16	3	No	5	5
MICROPIA_Gypsy	5428	79	13	0	No	10	2
NOMAD_Gypsy	7608	69	30	1	No	10	3
POGO_Tc1mariner	2121	53	7	1	No	4	4
PROTOP_P	4480	1676	8	1	No	291	144
QUASIMODO_Gypsy	7376	229	28	4	No	39	25
QUASIMODO2_DM_Gypsy	6452	93	8	1	No	10	7
R12_DM_NonLTR_Retrotransposon	3216	24	1	0	Yes/MinimumTEFailed	0	0

R1_DM_NonLTR_Retrotransposon	5356	198	19	7	No	27	27
R2_DM_R2	3607	38	5	0	Yes/MinimumTEFailed	0	0
ROO_BEL	9112	259	117	9	No	39	39
ROOA_BEL	7621	86	1	0	Yes/MinimumTEFailed	0	0
ROVER_DM_Gypsy	7318	134	13	3	No	15	6
S2_DM_Tc1mariner	1735	65	4	0	Yes/MinimumTEFailed	0	0
STALKER2_Gypsy	8250	497	9	2	No	76	42
STALKER4_Gypsy	7334	503	44	7	No	89	52
S_DM_Tc1mariner	1736	148	49	8	No	27	10
TABOR_Gypsy	7348	163	7	2	No	31	18
TAHRE_Jockey	10463	212	2	0	Yes/MinimumTEFailed	0	0
TART_Jockey	14477	322	0	0	Yes/MinimumTEFailed	0	0
TART_B1_Jockey	10654	64	1	0	Yes/MinimumTEFailed	0	0
TC12_DM_Tc1mariner	1644	44	13	2	No	9	2
TC1_DM_Tc1mariner	1666	52	20	3	Yes/NoClusterAfterFilter	5	0
TIRANT_LTR	8526	29	20	1	Yes/NoClusterAfterFilter	1	0
TRANSIB2_Transib	2844	73	12	2	No	5	2
TRANSIB3_Transib	2883	41	1	0	Yes/MinimumTEFailed	0	0
TRANSPAC_LTR_Retrotransposon	0	0	0	0	Yes/NPi	0	0

Transib1_Transib	3014	16	0	0	Yes/MinimumTEFailed	0	0
Transib4_Transib	2610	22	0	0	Yes/MinimumTEFailed	0	0
Transib5_Transib	3001	26	3	0	Yes/MinimumTEFailed	0	0
ZAM_Gypsy	8434	44	3	0	Yes/MinimumTEFailed	0	0

5 Discussion

This thesis has dealt with two questions in the field of transposable elements and their regulation:

- How TE invasion dynamics are influenced by piRNA regulation (trap model) and how they differ from TE regulation models based on selection.
- If piRNA regulation can act as a viable defence against TE reinvasion and horizontal TE transfers, i.e. can it provide "adaptive genomic immunity"?

The study of transposable elements is complex and paradoxical. They constitute a significant fraction of genomes in most complex organisms. They yet are not as well understood as protein-coding genes, which only represent a minor fraction of the genome. TEs actively spread between organisms, yet we know relatively less about how they make such jumps than viruses. The genomic defences of TE, i.e., the piRNA, are less understood than miRNA. All these gaps in knowledge present an opportunity for greater exploration in the field of TEs and their dynamics.

The result chapters of the thesis have extended the analytical framework associated with the trap model, and we have investigated the potential of cross-regulation(genomic immunity from past invasions) within the genome of *D. melanogaster* and between the species of the family *Drosophilidae*. Our simulations suggest that genomic immunity is possible but under very restrictive conditions. Furthermore, our bioinformatic analysis suggests that cross-immunity might not be prevalent between *Drosophilidae* species. Similarly, we noticed multiple TE families with successive transposition bursts despite having cluster insertions from previous invasions in *D. melanogaster*, indicating that piRNA might not act as a reliable genomic immune system against recurrent TE invasions. The subsequent sections will dwell on the possibilities enabled by the frameworks developed and the analysis done during this thesis and discuss the recent expansion of our knowledge in the field of TE dynamics and regulation.

5.1 Future directions

5.1.1 Identification of primary transcripts

The processed piRNA transcripts can be divided into two major categories, primary and secondary piRNA. Primary transcripts initiate the ping-pong cycle cascade and enable the amplification of piRNA. The piRNA produced downstream of this process, either by the processing of TE transcript or the long non-coding piRNA precursor, is referred to as secondary piRNA. So far, there is no consensus on how primary piRNA transcripts are produced. However, since they kick-start the silencing mechanism, they represent an essential piece of piRNA machinery.

One possible hypothesis I wanted to investigate was the relationship between the age of a pi-cluster and the amount of primary and secondary piRNA generated by the cluster. The study by Saint-Leandre et al. (2020) sheds some light on how prolific TE families are targeted by relatively high amounts of piRNA compared to older TE families in the *D. melanogaster* genome. However, determining if these piRNAs targeting successful TE families originate from existing or newer piRNA clusters can help us understand more about the evolution of the piRNA cluster loci itself. Do older clusters generate a specific species of piRNA (primary or secondary) more than newer clusters? Alternatively, are newer clusters self-sufficient to generate both subspecies of piRNA? The answer to this question can shed some insights into the prominence of old clusters and whether they are absolutely required for silencing a TE family or dispensable. It will also help us understand if there are redundancies in piRNA machinery and if multiple loci can provide primary piRNA simultaneously, and perhaps shed some light on the abundance of piRNA clusters in many genomes.

Based on the sequencing dataset described in Chapter 4, I attempted to measure the potential contribution of pi-clusters in terms of primary or secondary piRNA aligning to them. Disappointingly, I could not observe any discernable difference between the piRNA composition of pi-clusters. That is not to say there is no such difference, but the algorithm I used could not find it. The repetitive nature of TE sequences is an issue; thus, reliably assigning the source of piRNA is difficult. However, isolating long piRNA precursor transcripts from ovarian tissues is possible (Li et al., 2013; Murota et al., 2014). A total-RNA sequencing library will contain these long precursor transcripts. These precursors can be assigned to piRNA clusters more robustly, as they stretch for more than 200bp, containing substantially more information than the small mature piRNA. It is then possible to align a small RNA library containing piRNA to these long precursors to identify

the potential source piRNA cluster. The only pre-requisite for this method is the availability of both small RNA and total RNA datasets from the same pool of organisms, which is quite tricky for *D. melanogaster* due to the limited availability of biological material from each individual. Nevertheless, if we ever have a robust pipeline to identify the origin of primary and secondary piRNA, we can gain much more insight into the redundancy of piRNA machinery and the significance of old piRNA clusters.

5.1.2 Moving beyond *Drosophilidae*

The bioinformatic frameworks described in Chapter 4 are largely species-agnostic. It is possible to move beyond *Drosophilidae* to investigate the extent of piRNA regulation shared between different species of the same family or even between different strains of the same species. The only requirement to do so is the availability of high-quality genome assembly and small RNA dataset pairs for the species under investigation. Many insect species fulfil this criterion and thus can be investigated for piRNA cross-regulation within their genome. Furthermore, with resources like the Hymenoptera genome database (Walsh et al., 2021), it is possible to investigate the presence of genomic immunity against horizontal transfers within closely related *Hymenoptera* and other insect species.

Another opportunity is to adapt the framework to work on the *Muridae* family (which includes the genus *Mus*). Lilue et al. (2018) assembled the genome of 16 standard mice strains. These strains show great diversity in genotype, with divergences ranging from 1 Mya to less than 0.5 Mya. They even have different TE complements, suggesting active TE dynamics and in/outflow of TEs (Ferraj et al., 2022). For an organism-agnostic TE regulatory model, comparing results between *Drosophilidae* and *Mus musculus* would be intriguing. *Mus musculus* strain C57BL/6 has a well-annotated genome, similar to *D. melanogaster*. Multiple small RNA libraries and even PIWI RIP-seq libraries are available for *Mus musculus*, allowing the identification of cross-regulating piRNA between different mice strains.

Various factors change as we move from *D. melanogaster* to mammals. Mammals have larger genomes compared to *D. melanogaster* and piRNA machinery with more components (Ophinni et al., 2019). Mice and primates use DNA methylation in conjunction with histone modifications to silence the TE loci, whereas *D. melanogaster* only uses histone modifications (Aravin et al., 2008). Similarly, based on their expression during different developmental stages, mammalian piRNA insertions can be distinctly classified into two, pre-pachytene and

pachytene piRNA (Yu et al., 2021). They also differ in their TE content drastically, with mice and most primates having > 40% of their genome consisting of repeatable elements comapred to 20% in *D. melanogaster*. Terrestrial vertebrates also exhibit a distinct lack of horizontal transfers between them, starkly contrasting insects (Peccoud et al., 2017; Zhang et al., 2020). Hence, while the bioinformatic frameworks can be applied to these species and orders without modification, the results will reflect (sometimes) drastically different biological systems and require more scrutiny.

It is also possible to study species that share the same ecological habitat in specific geographical locations using the same bioinformatics pipeline. We observed potential horizontal transfers between the genus *Zaprionus* and *D. melanogaster*. However, it would be interesting to study the effectiveness of the piRNA pathway as a defence against horizontal transfer between sympatric species beyond *Drosophilidae*.

5.1.3 Implementing migration

Most species are not homogenous; they can be separated spatially (e.g. in different geographical areas) or temporally (e.g. different mating seasons) and could be divided into several random mating populations. Nevertheless, there exists a possibility of genetic exchange between populations due to gene flow. In this thesis, I have only considered the spread of TE and TE regulatory alleles within a single population. However, as we know from evidence gathered from multiple TE families, TEs often move between populations. Hence, gene flow among partially isolated populations is essential to understand the dynamics of TEs at the species level.

Quesneville and Anxolabéhère (1998) used the stepping stone migration model, transferring members from the originally invaded population to the second population and from the second to the third population, simulating recurrent TE flow between different populations. However, they had two concepts for P-element repression, one based on the insertion site[1], which influenced the transposition rate of the TE copy present in that insertion site and another based on non-autonomous repressor copies[2] and neutral copies. These repressor copies moved along with migrant individuals into new populations, thereby simultaneously

[1] These regulatory insertion sites affect TE's ability to transpose, resembling the now-known piRNA clusters!

[2] We now know that non-autonomous copies can either be part of piRNA and repress the TE family or "steal" the transposase from autonomous copies.

introducing both the active P-element and its regulatory allele. They studied the dynamics of the P-element and observed that migration profoundly affected the P-element activity. Not only they observed a reduction in P-element copies but also an increase in defective copies and regulatory cytotype. Similarly, Deceliere et al. (2005) used simulations to demonstrate how migration rate affects the occupancy of insertion sites and the mean TE copy number in the populations undergoing migration. They also observed a decrease in the mean TE copy number in the population[3], and an equal mean TE copy number shared between populations experiencing migration.

The models from Quesneville and Anxolabéhère (1998) and Deceliere et al. (2005) demonstrate that migration significantly impacts TE dynamics, and the model from Quesneville and Anxolabéhère (1998) even incorporates piRNA-like regulation; augmenting them with trap model can lead to a better understanding of TE dynamics in the wild population, where species can experience gene flow.

5.1.4 Fitting Simulations with Experimental Evolution

Simulations and models can provide us with approximations of TE dynamics under various parameters. However, these approximations are only as accurate as the model is correct in its ability to abstract an organism undergoing TE invasion and the characteristics of the population. It is essential to compare the results of the simulations with the experiments (in-vivo) along with the sampling of natural populations to understand TE dynamics. Experimental evolution provides a controlled environment where parameters like population size can be controlled. During these experiments, we can introduce a novel TE into the species of interest (*D. melanogaster*) and measure the passage of TE invasion as the population evolves. Robillard et al. (2016) introduced Mos1 TE into the *D. melanogaster* genome and measured the copy number of Mos1 elements in the experimental *D. melanogaster* population using RT-PCR in fixed intervals. Such datasets give us a reliable measure of TE activity in near real-time.

Similarly, capturing the state of populations evolving in the wild is essential. However, sampling from wild populations has many confounding factors that can influence the interpretation, such as the location or if the population is under stress due to environmental factors. If we can account for these factors in our sampling methodology, we can get more robust or "real" insights into the

[3]Compared to the population without migration.

dynamics of TE[4]. A prerequisite for studying TE invasions in natural populations is the detection or presence of an invading TE. In the past, we have identified the recent expansion of TEs in *Drosophilidae*, like the P-element. However, it is debatable how often we can detect signatures of invading TEs in the wild.

If the datasets about TE activity are available from multiple sampling of wild populations in regular intervals or through experimental evolution, we can use them to estimate the model parameters; that is to say, we can fit the data to the model. However, deriving a likelihood function that maps well with summary statistics associated with TE activity is challenging. One workaround is using likelihood-free inference methods, like Approximate Bayesian computation (ABC). ABC allows us to find the posterior distribution of the parameters associated with TE invasion, like the TE transposition rate and works well when we have a simulation framework that accepts parameters and generates results that can be compared with the observation (Sunnåker et al., 2013).

Our simulation software Simulicron is well-suited to generate the data required for ABC approximations. It thus can be used in conjunction with biological data generated from experimental evolution studies conducted in the lab. However, to use such methods, it is crucial to identify relevant summary statistics, such as the TE copy number at a specific generation or the site frequency spectrum. Identifying a correct distance function and tolerances that compares the summary statistics generated from the simulation and observed data and reject the simulation is also essential - i.e. for the ABC rejection algorithm (Tavaré et al., 1997). Even in a relatively simple model like Simulicron, more than five parameters can be explored simultaneously. If we randomly choose each parameter, the resulting exploration space can be vast and computationally prohibitive (depending on our tolerances). Employing Markov chain Monte Carlo (MCMC) methods with ABC can reduce the computational burden significantly by sampling a limited set of parameter combinations (Marjoram et al., 2003). If we can optimise an ABC MCMC framework to guide simulations and compare the results with experimental evolution, we may be able to identify the factors which enable a TE to invade efficiently (or otherwise).

[4]For example, organisms experiencing deleterious effects of TE might have different chances of survival in a wild and experimental population.

5.1.5 Variable piRNA strength

piRNA clusters can have a different level of expression and activity within the same genome/organism. Under a classical trap model, we assume that a single insertion of TE silences all the other TE copies of the same family. In contrast, if we have clusters with variable expression levels or "strength", some clusters would inhibit TE families by a single insertion, and others might require multiple insertions of the same family within a single locus. Akulenko et al. (2018) found such "strong" and "weak" clusters that differed in their ability to recruit *Rhino* protein, a crucial component of piRNA production machinery. The strong clusters associated with *Rhino* binding have a stronger effector piRNA expression than weak clusters. Although the trap model can be modified to incorporate such characteristics, and Simulicron even has provisions to do the same, it has yet to be studied. Similarly, Fablet et al. (2014) found variation between the expression and sequence of piRNA machinery in different *Drosophila* species. Such variations can also be studied using the model, and their effect is global, i.e., if a single cluster insertion reduces the transposition rate of the TE family by 1, we can modify it to 0.5 to reflect weaker piRNA machinery.

5.2 Looking beyond the trap model

The trap model explains the working of piRNA-based TE regulation quite comprehensively by describing the host defences and the ability of the TE to co-exist with them (Kofler, 2019), explaining the ubiquitous nature of repeated elements in the genomes of most organisms. Itou et al. (2015) experimentally validated the trap model by inserting novel sequences into piRNA clusters and silencing the reporter region, so we have proof that the model itself is valid and capable of targeting homologous sequences outside piRNA clusters. Moreover, our brief exploration of the trap model in both result chapters seems capable of stopping a TE invasion under different conditions, including multiple TEs invading simultaneously or differences in piRNA cluster composition and number.

However, growing evidence suggests that the trap model is not the only way to stop a TE invasion.

5.2.1 The alternatives ?

A particular species of bdelloid rotifer *Adineta vaga* lacks functional piRNA machinery, with a complete lack of ping-pong cycle (Rodriguez and Arkhipova, 2016). Nevertheless, it exhibits a similar level of TE dynamics as its close relatives, which have the complete canonical piRNA pathway (Nowell et al., 2021). This observation suggests some unknown TE silencing mechanism is in effect in *A. vaga*, or perhaps the RNA interference pathway has alternative mechanisms of TE control which are yet to be discovered.

Looking into plants, Piriyapongsa and Jordan (2008) discussed a model where TEs generate siRNA and eventually TE regulating miRNA. Creasey et al. (2014) discovered that miRNA in plants initiates the secondary siRNA pathway targeting TEs, acting analogous to the primary piRNA transcript in animals. Guo et al. (2022) discovered >2000 miRNA, which originates from TEs. If TE sequences indeed convert into miRNA[5] or are responsible for initiating the siRNA generation cascade, this would yield a drastically different regulatory system due to the ability of miRNA and siRNA to target sequences based on partial complementarity and their post-transcriptional silencing activity (Anzelon et al., 2021; Agarwal et al., 2015). Another possibility is that piRNA clusters alone do not check the TE expansion and invasion in the population and instead work in conjunction with non-cluster-based small RNA defense or a completely undiscovered mechanism. The studies mentioned above present several models of TE defence, which still require TE sequences but are independent of specific genomic loci like piRNA clusters.

It is also possible that the regulation of TE happens at the protein level. Lohe and Hartl (1996) describe overproduction inhibition(OPI) in the Mariner family of TEs. They found that the increased concentration of transposases causes a decline in transposition. Heinlem et al. (1994) observed non-functional Activator (Ac) transposase aggregates in conjunction with hyperactive promoter regions adjacent to the transposase protein. These observations allude to the negative dosage effect; it is possible that hyper-concentration of transposase can cause them to form aggregates and, as a result, reduce their activity or inhibit them completely. Multiple examples of proteins of non-TE origin show this behavior, including amyloid-B proteins, which are implicated in neurological disorders (Koo et al., 1999). Lohe et al. (1997) also describes mutated transposases, which even reduce

[5]There is some evidence for it (Shalgi et al., 2010; Petri et al., 2019) - however, the discovery of bonafide miRNA derived from TEs in animals is still an emerging topic.

germline TE excision in mariner elements and compete with wild-type transposes, reducing their activity in heteroallelic mutant/nonmutant systems, acting as a robust TE regulation system (termed as dominant-negative complementation).

Simmons and Bucholz (1985) describes another method of TE self-regulation in the form of titration of transposase by non-functional TE copies. They discovered non-functional P-elements in *D. melanogaster*, which could titrate functional P-element transposase. Indeed, Robillard et al. (2016) observed the same with the Mos1 element, where the non-functional copy of Mos1, named Peach, amplified more than the autonomous Mos1 copies by sequestering Mos1 transposase, thus diluting autonomous Mos1 copies in subsequent generations. Indeed, it is possible that "regulation by titeration" is possible for TE families with autonomous and non-autonomous counterparts. Given time, some TE copies for a TE family would have mutations that render their transposase non-functional and make them non-autonomous, thus acting as a sponge for functioning transposases. However, can we call them regulatory alleles? These observations bring us to the self-regulating TE model introduced by Charlesworth and Charlesworth (1983), where all TEs eventually self-regulate, and the transposition rate decreases with the increase in copy number[6]. TE self-regulation is independent of piRNA-based TE regulation and has some evidence; perhaps the eventual mutation in the transposase catalyzes this self-regulation. However, the timescale of these mutations might be too slow to save an organism from the adverse effects of TE activity.

5.2.2 The contradiction

Recent studies have also questioned the necessity of piRNA clusters or if they are at all required to defend against the TEs. The study by Gebert et al. (2021) investigated the activity of TE families in *D. melanogaster* by deliberately deleting three significant piRNA cluster loci hypothesized to regulate those respective TE families. They found no increased TE activity post cluster deletion, which contradicts the trap model. Gebert et al. (2021) suggested that piRNA clusters are dispensible, i.e., clusters are unnecessary after the TE invasion has subsided. We can consider the possibility of TE sequences evolving and collecting inactivating mutations, which can be deleterious for the TE (and can also give rise to TE killing non-autonomous copies). After this inactivation, we could still find remnants of the TE family, which have lost their ability to transpose. We could also find

[6] even though their model did not explicitly consider mutated or non-autonomous TEs as the source of this self-regulation

their cluster insertions that are not required anymore. Genomes often contain the history of TE invasions instead of an ongoing invasion, and these derelict piRNA clusters might reflect those past invasions.

The findings of Wierzbicki and Kofler (2023) question the validity of the trap model. Their finding indicates a discordance between simulations and observations from multiple *D. melanogaster* strains. If we consider piRNA clusters are indeed special regions, and the cluster insertion is sufficient to silence TEs. In that case, they should have a different distribution of TEs within them than the rest of the genome. Indeed, their simulations suggested no correlation between TE copies in piRNA clusters and the rest of the genome. In contrast, they found no distinction in the distribution of TEs in the piRNA cluster and the rest of the genome in the investigated *D. melanogaster* strains. This observation suggests that piRNA clusters are not special regions of the genome, or at least they run counter to the observations made using the trap model. Moreover, they found multiple TE families not present in annotated[7] piRNA clusters. Instead, they found multiple discrete loci non-cluster that generated piRNA for those TE families[8].

Indeed, there are multiple studies indicating the presence of piRNA of non-cluster origin. Shpiz et al. (2014) identified non-cluster TE insertions that generated piRNA and siRNA, which silenced the corresponding TE family. They also observed the production of small RNA from the region flanking the TE insertions. This observation suggests that these insertions might act as proto-piRNA clusters. They also noticed the production of piRNA from TE insertions in the UTRs (specifically, 3' untranslated region) of genes. Mohn et al. (2014) have also described Rhino-Deadlock-Cutoff (RDC) complex initiating piRNA formation from non-cluster TE insertions. Their observation suggests that the small RNA itself guides the RDC complex and defines the piRNA locus. In mice, the A-MYB promoter is closely associated with piRNA clusters[9], and there are more than 3500 A-MYB binding sites, thus indicating many more potential piRNA loci than the current 300 or so discovered piRNA clusters (Yamanaka et al., 2014). These observations suggest that not only can piRNA clusters form ex-nihilo but also that piRNA-based TE regulation is not limited to canonical piRNA clusters.

Scarpa and Kofler (2023) studied the effect of induced paramutations (i.e. by maternally inherited piRNA) in the silencing of TE families and observed that they significantly contribute to TE regulation by using non-cluster TE insertions

[7] piRNA cluster annotation can differ with different tools and the quality of sequenced piRNA libraries.
[8] Referred to as dispersed source loci.
[9] with respect to upstream/downstream vicinity of the gene

as the source of piRNA. Similarly, Hermant et al. (2015) observed the activation of dormant P element-derived transgene cluster *BX2* when crossed with females containing *T-1* locus producing complementary piRNA sequences against *BX2*. The activated *BX2** cluster produced piRNA even without the *T-1* locus, highlighting the importance of paramutations in piRNA machinery. While these observations explain the observed discrepancies between the distribution of TEs in piRNA and the genome, it raises an even bigger question regarding the genetic basis of the inheritance of information. Is maternal piRNA an absolute requirement to initiate the regulatory cascade (and thus act as the primary piRNA), and if so, does that mean that multiple piRNA clusters are supernumerary, perhaps much more than the three clusters deleted by Gebert et al. (2021)? If true, this would mean that TE regulation is more dependent on the efficiency of epigenetic information transfer than simple TE insertions in piRNA clusters. Scarpa and Kofler (2023) also modified the paramutation model to include siRNA generating sites, which, when occupied, could trigger paramutations yielding piRNA from non-cluster TE insertions. Their simulations suggested that siRNA-generating sites alone were sufficient to stop a TE invasion, yielding TE dynamics comparable to silencing by piRNA clusters. This observation, yet again, raises questions about the necessity of piRNA clusters.

5.3 Closing thoughts

During the course of this thesis, we saw multiple studies investigating the trap model and piRNA, and some even challenging the importance of piRNA clusters. Some of these studies have incorporated alternative strategies for TE defence and even highlighted the vital role of epigenetics and other small RNA species, potentially overshadowing the trap model. Moreover, we are rapidly expanding our repertoire of genomes, some from enigmatic species, and perhaps we may find novel ways of TE defences in those. We need a new simulation framework and an analytical model that considers the recent discoveries of alternative TE regulatory mechanisms to understand TE dynamics better. Such a model would be very complex, but so are the TEs and their regulatory systems. I end this thesis in the hope that I have contributed to a better understanding of TE dynamics; and I optimistically anticipate future studies that will address the perplexing systems regulating Transposable Elements.

Résumé en français

Les éléments transposables (ET) constituent un aspect fascinant de la génétique, présents chez presque tous les organismes eucaryotes. Ces éléments génétiques mobiles peuvent s'amplifier dans le génome, entraînant des changements importants dans la constitution génétique de l'organisme. Les ET sont classés en deux catégories principales : les ET de classe I et les ET de classe II.

Les ET de classe I se multiplient en utilisant un intermédiaire ARN au cours de la transposition, contrairement aux familles de classe II qui utilisent quant à elles un intermédiaire ADN et un mode couper-coller ou copier-coller. Au sein de ces grandes classes, il existe de nombreuses sous catégories (ordre, superfamilles, familles, sous-familles), qui partagent ou non des similitudes en termes de mécanismes de transposition, de structure ou des similitudes de séquences protéiques ou nucléotidiques. Le niveau de classification le plus bas est la famille, qui peut être divisée en groupes plus spécifiques en fonction de leurs caractéristiques uniques.

La diversité des éléments transposables à travers l'arbre du vivant est particulièrement vaste, chaque espèce possédant des centaines voire des milliers de familles d'ET différentes dont les nombreuses copies sont dispersées dans le génome. Alors que les ET sont considérés généralement comme délétères pour l'organisme, ils jouent un rôle important dans la structure, le fonctionnement et l'évolution du génome. Par exemple, en s'insérant dans où à proximité des gènes, ils peuvent les inactiver ou modifier leur régulation en perturbant les promoteurs et les activateurs, entraînant ainsi des modifications dans l'expression spatiotemporelle des gènes. De plus, les ET étant répétés dans le génome, ils peuvent être à l'origine de recombinaisons dites ectopiques, qui sont à l'origine de délétions à grande échelle, d'inversions chromosomiques, de duplications ou de translocation, ayant un impact significatif sur l'architecture génétique de l'organisme, et souvent délétères. Cependant les ETs, en transposant activement, sont une source majeure de diversité génétique au sein des populations. Certaines copies confèrent parfois un effet bénéfique à la cellule ou l'organisme, qui se traduit par une augmentation de la fréquence de cette insertion dans la population, et finalement par sa fixation. Cela s'accompagne de la perte de mobilité de la copie. On parle de domestication moléculaire. Plusieurs exemples d'innovations évolutives majeures résultent de la domestication de gènes issus d'éléments transposables.

Les éléments transposables restent néanmoins à la base des parasites du génome dont l'amplification massive peut s'avérer néfaste pour l'hôte. Pour contrer la prolifération des ET, la plupart des organismes ont ainsi développé des défenses. Chez les métazoaires, la principale voie permettant de contrôler et d'empêcher la transposition des ET est réalisée par des mécanismes faisant appel à de l'interférence ARN et la déposition de marques épigénétiques. En particulier la voie des piARN (PIWI interacting) est spécialisée dans le contrôle des ET. Cette voie de régulation, qui implique le mécanisme d'interférence ARN (ARNi), permet de réguler les ET.

Les piARN sont une classe spécifique de molécules d'ARNi qui ciblent les ET, ils proviennent de loci génomiques appelés clusters de piARN. Ces petites molécules d'ARN mesurent entre 24 et 32 nucléotides et sont prises en charge par des protéines spécifiques, appartenant à la famille des protéines PIWI. Grâce à la complémentarité des séquences, les piARN peuvent réguler les ET en induisant leur méthylation ou une autre modification épigénétique de leur séquence ou en détruisant les ARN messagers transcrits à partir des éléments.

Lorsqu'un ET d'une famille particulière transpose dans un cluster piARN, le cluster génère des piARN contre toutes les copies génomiques de cette famille et régule la transposition de toutes les copies. Ce mécanisme constitue la base du modèle "piège" (trap model), qui postule qu'une seule insertion d'ET dans un cluster piARN suffit à stopper l'activité de transposition de l'ensemble d'une famille d'éléments.

Les ET ont un cycle de vie spécifique dans les espèces, qui peut durer plusieurs millions d'années. Ce cycle commence par l'invasion d'une population naïve, via l'introduction d'une copie unique dans l'un des individus d'une population (souvent, par un transfert horizontal). Le nombre de copies augmente alors rapidement dans le génome de la population hôte. Simultanément, des génotypes régulateurs émergent dans la population, en particulier du fait de la transposition dans les clusters piARN. Quand la plupart des individus de la population possèdent des piARN contre la famille d'ET qui l'envahit, l'amplification des ET s'arrête. À ce stade, les ET peuvent persister longtemps au sein de la population, en gardant un certain nombre de copies dans le génome. Cependant, au fil du temps, les copies d'ET accumulent progressivement des mutations inactivatrices, qui les rendent incapables de transposer, même en absence de régulation. Finalement, la plupart des copies d'ET se dégradent, et deviennent des "reliques" génomiques, iden-

tifiables par les algorithmes bio-informatiques, mais sans activité biologique.

Cette thèse vise à répondre à quatre questions spécifiques sur la capacité des ET à envahir les populations :

1. Effets de la sélection : Comment différentes hypothèses sur le coût des copies d'ET influencent-elles la dynamique d'invasion ?

2. Architecture des clusters piARN : les clusters de piARN distribués sur plusieurs chromosomes et plus petits sont-ils plus efficaces qu des clusters plus grands mais moins nombreux ?

3. Comprendre le concept d'immunité génomique : les clusters de piARN peuvent-ils fournir une immunité génomique à long terme contre les familles d'ET après l'invasion initiale ?

4. Trouver des preuves d'immunité génomique : existe-t-il des preuves d'immunité génomique conférée par les piARN chez l'espèce modèle *Drosophila melanogaster* et les autres espèces de *Drosophilidae* ?

Un nouveau logiciel de simulation a été développé pour répondre à ces questions. Les simulations individu-centrées peuvent reproduire la dynamique de l'invasion des copies dans une population. Chaque individu a une capacité reproductive qui peut être altérée par les ET insérés dans son génome. Certaines régions du génome sont définies comme des clusters de piARN, et tout ET qui saute dans un cluster de piARN verra son taux de transposition réduit à zéro. Le cadre de simulation peut être modifié via divers paramètres, tels que la taille de la population, ou le taux de recombinaison. La simulation fournit plusieurs indicateurs statistiques, tels que le nombre moyen de copies, et la répartition des allèles

régulateurs dans la population. Le programme de simulation est implémenté en Python, ce qui garantit la portabilité et l'extensibilité du logiciel.

Effets de la sélection

Si l'insertion de copies d'ET n'est pas associées à un coût vis-à-vis de la sélection, quelle que soit leur position dans le génome, les modèles traditionnels de génétique des populations prédisent que le nombre de copies dépend du taux de transposition. Cependant, les simulations réalisées au cours de la thèse, basées sur l'hypothèse du modèle de piège (trap model), suggèrent que le taux de transposition a un effet minimal sur le nombre de copies à l'équilibre. En effet, la probabilité d'insertion dans un cluster de piARN, responsable de la régulation, est proportionnelle au taux de transposition : la régulation arrivera plus rapidement chez les familles d'ET très actives, et moins rapidement pour les éléments transposant peu. Ces résultats de simulation ont ensuite été confirmés par des dérivations mathématiques.

Cependant, cette hypothèse de neutralité sélective est peu réaliste, puisque les ET sont en général délétères pour l'organisme. Cependant, toutes les copies n'ont pas nécessairement le même effet. En particulier, les copies d'ET insérées dans l'hétérochromatine et les clusters de piARN sont potentiellement moins délétères, tandis que les copies d'ET situées dans l'euchromatine (à proxilité de nombreux gènes) peuvent avoir des conséquences plus sévères. Deux modèles ont été analysés :

1. ET génomiques délétères et insertions dans les clusters neutres : les simulations prédisent que le nombre de copies à l'équilibre devrait être proche de zéro, et les prédictions mathématiques le confirment.

2. ET génomiques et insertions dans les clusters délétères : les simulations prédisent un équilibre stable, permettant aux ET de persister dans la population via un équilibre de transposition-sélection. Dans ces conditions, la plupart des individus auront des insertions de piARN, mais certains génotypes permissifs permettront aux ETs de rester actifs dans la population.

Architecture de cluster optimale

La question suivante abordée dans la thèse est de déterminer la distribution optimale des clusters de piARN. Dans la plupart des organismes modèles, environ 3% du génome est dédié aux clusters piARN. Cependant, deux possibilités théoriques existent pour distribuer ces 3% du génome sous forme de clusters de piARN : un seul grand cluster de piARN, ou plusieurs petits clusters recombinants répartis dans tout le génome. Quelles que soient les modalités de sélection, les simulations ont montré qu'un seul grand cluster était toujours plus efficace pour contrôler l'invasion des ET que plusieurs clusters plus petits. Cependant, dans les organismes réels, ces deux types de structures coexistent. Un exemple de grand cluster de piARN est le locus flamenco chez Drosophila melanogaster, dédié à la régulation des ET dans les cellules somatiques des gonades.. Par contraste, les cellules de la lignée germinale sont régulées dans cette espèce par plusieurs petits clusters de piARN.

Comprendre l'immunité génomique

Dans le cycle de vie des ET évoqué précédemment, l'invasion des familles actives est arrêtée par la régulation (via les piARN) et par l'accumulation de mutations dans les copies. Cependant, des mutations inverses peuvent théoriquement restaurer la capacité des ET à transposer, et la modification des séquences d'ET peuvent aider les copies à échapper aux mécanismes de défenses.

Cette thèse soulève la question de savoir si les piARN peuvent empêcher de telles réinvasions. Dans le cadre de simulations deux familles d'ET apparentés sont définies : α et β. La famille α est la famille résidente, présente au début de la simulation, et la famille β est la famille ET envahissante, qui arrive dans le génome dans un second temps. Le modèle de simulation introduit un paramètre, η, qui définit le coefficient de régulation croisée, ou la capacité des piARN d'une famille ET à cibler l'autre famille TE. L'analyse des simulations pour différents jeux de paramètres révèle que :

— Le nombre de copies à l'équilibre pour α reste constant à moins que le β n'envahisse presque simultanément et que la co-régulation soit quasiment totale.

— Au contraire, β ne peut plus envahir une fois que l'invasion α est terminée, à moins que la co-régulation soit faible.

Ces résultats suggèrent que les piARN pourraient uniquement être capables de fournir une immunité génomique aux familles ET dans le cas où les copies de TE sont très proches, et susceptibles d'être corégulées systématiquement. Pour les familles d'ET avec des copies divergentes, la voie piARN pourrait ne pas être en mesure d'empêcher la réinvasion, puisque les piARN produits contre la première famille serait incapables de reconnaître la deuxième.

Trouver des preuves d'immunité génomique

Cette thèse examine également l'histoire de l'invasion des familles d'éléments transposables chez *Drosophila melanogaster* afin de trouver des preuves d'une immunité génomique contre les nouvelles invasions d'ET. Les résultats des simulations suggèrent que les piARN ne peuvent pas empêcher la réinvasion si les copies de TE ont divergé. Pour vérifier ce résultat, le génome de la drosophile a été analysé pour identifier les copies d'ET appartenant à des familles apparentées. Toutes les copies complètes ou presque

complètes de chaque famille ont été extraites du génome, et les petits ARN extraits de *D. melanogaster* ont été alignés sur celles-ci. De plus, les copies d'ET insérées dans des clusters de piARN annotés ont également été identifiées. Environ 20 familles d'ET parmi les 65 familles ET analysées présentaient plusieurs vagues d'invasion, chacune avec son propre ensemble de clusters de piARN régulateurs. Ce résultat confirme les résultats de simulation : la présence de copies issues d'une invasion précédente, ainsi que la production de piARN contre ces copies anciennes, n'empêchent pas une nouvelle invasion par un élément quelque peu différent. Cependant, toutes les familles de TE ne présentent pas de vagues d'invasion multiples, ce qui peut s'expliquer par l'efficacité de la régulation, ou par une moindre implication de ces familles dans des transferts horizontaux issus d'autres espèces.

Le transfert horizontal d'ET est répandu dans l'arbre du vivant, et largement concentré chez les insectes et les vertébrés. L'analyse décrite précédemment a été étendue à la famille des *Drosophilidae* pour identifier si les piARN peuvent conférer une immunité contre les transferts horizontaux, en analysant plus de 100 génomes de *Drosophilidae* pour la régulation croisée de TE similaires. Plus précisément, les familles d'ET de *D. melanogaster* ont été recherchées chez d'autres espèces de Drosophiles. Par ailleurs, les piARN de *D. melanogaster* ont été alignés sur les ET d'autres *Drosophilidae*. Il est clairement apparu que les ET des espèces du sous-groupe "melanogaster" (notamment *D. simulans*, *D. mauritiana*, *D. sechellia*) présentaient un degré élevé de régulation croisée et les transferts horizontaux efficaces entre ces groupes sont peu probables. Cependant, les piARN de *D. melanogaster* pourraient ne pas cibler les ET trop divergents d'espèces distantes, auquel cas il existe une possibilité de transferts horizontaux entre ces espèces et *D. melanogaster*.

Les espèces apparentées au genre *Zaprionus* (un genre très proche de Drosophila, dont la monophylie est très controversée)

ont par ailleurs montré des propriétés particulières en terme d'ET. En effet, les piARN de *D. melanogaster* ont le potentiel de cibler plusieurs familles d'ET de *Zaprionus*, reflétant une proximité inattendue entre les ET de ces groupes d'espèces assez divergentes. Cette observation suggère de multiples transferts horizontaux entre *Zaprionus* et *D. melanogaster* ou ses ancêtres récents. Après avoir analysé les arbres phylogénétiques des ET, les ET de *Zaprionus* ciblés par les piARN de *D. melanogaster* étaient trop proches pour être compatibles avec l'hypothèse d'une hérédité classique, par ailleurs, leur phylogénie était discordante par rapport à l'arbre des espèces. La plupart des *Zaprionus* et *D. melanogaster* sont des espèces sympatriques en Afrique, ; il existe donc des suspicions de transferts horizontaux récurrents d'ET entre les deux groupes. Dans cette situation particulière, les piARN de *D. melanogaster* sont donc susceptibles d'empêcher les transferts horizontaux à partir d'espèces sympatriques.

Conclusion

Cette thèse présente un nouveau modèle de dynamique des ET, répondant aux quatre questions soulevées initialement :

1. Toutes les copies d'ET (euchromatiques et hétérochromatiques, y compris celles insérées dans les clusters de piARN) doivent être délétères pour qu'une famille d'ET persiste dans le génome à l'équilibre.

2. Un seul grand cluster non recombinant est théoriquement plus efficace pour inhiber l'activité des ET et réduire le nombre de copies à l'équilibre, par comparaison avec les systèmes impliquant plusieurs clusters plus petits. Néanmoins, la régulation par de nombreux petits clusters, bien que théoriquement peu efficace, reste la norme dans les espèces connues.

3. L'action des piARN ne fonctionne que sur des séquences très similaires. Les piARN ne peuvent pas réguler les copies TE divergentes au sein d'une famille TE.

4. Les piARN ne peuvent pas empêcher le transfert horizontal d'ET similaires provenant d'espèces éloignées.

Quoi qu'il en soit, de multiples contradictions subsistent contre le modèle de piège et la machinerie piARN :

— Si les clusters de piARN sont des régions régulatrices spéciales du génome, ils devraient avoir une composition différente et termes d'ET. Pourtant, la littérature récente montre que la distribution des ET est fortement corrélée entre le génome et les clusters, ce qui indique que les clusters de piARN accumulent les ET de la même manière que le reste du génome. Il a également été rapporté que que le nombre observé de copies d'ET insérées dans des clusters s'écarte des simulations.

— Par ailleurs, il existe des preuves expérimentales indirectes que l'émergence d'une régulation contre l'élément P arrive avant que le piARN contre l'élément P puisse être détecté. Ces observations indiquent que les hypothèses du modèle de piège (trap model) pourraient nécessiter des révisions supplémentaires.Elles suggèrent que d'autres mécanismes de régulations pourraient exister, en parallèle de la régulation par les piARN, ou que des piARNs pourraient être produits à partir de régions non décrites comme des clusters de piARN.

En conclusion, le modèle de piège explique la plupart des aspects de la dynamique d'invasion des ET. Cependant, il s'agit d'une simple abstraction d'un mécanisme biologique complexe, et certaines contradictions nécessitent des modèles révisés pour ex-

pliquer la dynamique des ET avec plus de précision et de détails. Peut-être qu'un modèle plus complet incluant l'autorégulation des ET, la voie des siARN combinée avec cette des piARN pourrait être en mesure de répondre aux questions sur la distribution des ET dans les génomes et dans l'arbre du vivant.

Bibliography

V. Agarwal, G. W. Bell, J.-W. Nam, and D. P. Bartel. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4, Aug. 2015. doi: 10.7554/elife.05005. URL https://doi.org/10.7554/elife.05005.

N. Akulenko, S. Ryazansky, V. Morgunova, P. A. Komarov, I. Olovnikov, C. Vaury, S. Jensen, and A. Kalmykova. Transcriptional and chromatin changes accompanying de novo formation of transgenic piRNA clusters. *RNA*, 24(4):574–584, Jan. 2018. doi: 10.1261/rna.062851.117. URL https://doi.org/10.1261/rna.062851.117.

T. A. Anzelon, S. Chowdhury, S. M. Hughes, Y. Xiao, G. C. Lander, and I. J. MacRae. Structural basis for piRNA targeting. *Nature*, 597(7875):285–289, Sept. 2021. doi: 10.1038/s41586-021-03856-x. URL https://doi.org/10.1038/s41586-021-03856-x.

A. A. Aravin, G. J. Hannon, and J. Brennecke. The piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science*, 318 (5851):761–764, Nov. 2007. doi: 10.1126/science.1146484. URL https://doi.org/10.1126/science.1146484.

A. A. Aravin, R. Sachidanandam, D. Bourc'his, C. Schaefer, D. Pezic, K. F. Toth, T. Bestor, and G. J. Hannon. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Molecular Cell*, 31(6):785–799, Sept. 2008. doi: 10.1016/j.molcel.2008.09.003. URL https://doi.org/10.1016/j.molcel.2008.09.003.

I. Arkhipova. Mobile genetic elements and sexual reproduction. *Cytogenetic and Genome Research*, 110(1-4):372–382, 2005. doi: 10.1159/000084969. URL https://doi.org/10.1159/000084969.

I. Arkhipova and M. Meselson. Deleterious transposable elements and the extinction of asexuals. *BioEssays*, 27(1):76–85, 2004. doi: 10.1002/bies.20159. URL https://doi.org/10.1002/bies.20159.

I. R. Arkhipova. Distribution and phylogeny of penelope-like elements in eukaryotes. *Systematic Biology*, 55(6):875–885, Dec. 2006. doi: 10.1080/10635150601077683. URL https://doi.org/10.1080/10635150601077683.

N. Bannert and R. Kurth. The evolutionary dynamics of human endogenous retroviral families. *Annual Review of Genomics and Human Genetics*, 7(1):149–173, Sept. 2006. doi: 10.1146/annurev.genom.7.080505.115700. URL https://doi.org/10.1146/annurev.genom.7.080505.115700.

N. H. Barton. Mutation and the evolution of recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1544):1281–1294, Apr. 2010. doi: 10.1098/rstb.2009.0320. URL https://doi.org/10.1098/rstb.2009.0320.

E. A. Bennett, H. Keller, R. E. Mills, S. Schmidt, J. V. Moran, O. Weichenrieder, and S. E. Devine. Active *Alu* retrotransposons in the human genome. *Genome Research*, 18(12):1875–1883, Oct. 2008. doi: 10.1101/gr.081737.108. URL https://doi.org/10.1101/gr.081737.108.

C. M. Bergman, H. Quesneville, D. Anxolabéhère, and M. Ashburner. "recurrent insertion and duplication generate networks of transposable element sequences in the drosophila melanogaster genome". *Genome Biology*, 7(11): R112, 2006. doi: 10.1186/gb-2006-7-11-r112. URL https://doi.org/10.1186/gb-2006-7-11-r112.

P. M. Bingham, M. G. Kidwell, and G. M. Rubin. The molecular basis of p-m hybrid dysgenesis: The role of the p element, a p-strain-specific transposon family. *Cell*, 29(3):995–1004, July 1982. doi: 10.1016/0092-8674(82)90463-9. URL https://doi.org/10.1016/0092-8674(82)90463-9.

J. P. Blumenstiel. Birth, school, work, death, and resurrection: The life stages and dynamics of transposable element proliferation. *Genes*, 10(5):336, May 2019. doi: 10.3390/genes10050336. URL https://doi.org/10.3390/genes10050336.

J. Brennecke, A. A. Aravin, A. Stark, M. Dus, M. Kellis, R. Sachidanandam, and G. J. Hannon. Discrete small RNA-generating loci as master regulators of transposon activity in drosophila. *Cell*, 128(6):1089–1103, Mar. 2007. doi: 10.1016/j.cell.2007.01.043. URL https://doi.org/10.1016/j.cell.2007.01.043.

J. Brennecke, C. D. Malone, A. A. Aravin, R. Sachidanandam, A. Stark, and G. J. Hannon. An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science*, 322(5906):1387–1392, Nov. 2008. doi: 10.1126/science.1165171. URL https://doi.org/10.1126/science.1165171.

R. J. Britten and D. E. Kohne. Repeated sequences in DNA. *Science*, 161(3841): 529–540, Aug. 1968. doi: 10.1126/science.161.3841.529. URL https://doi.org/10.1126/science.161.3841.529.

J. Brookfield. Interspersed repetitive DNA sequences are unlikely to be parasitic. *Journal of Theoretical Biology*, 94(2):281–299, Jan. 1982. doi: 10.1016/0022-5193(82)90313-7. URL https://doi.org/10.1016/0022-5193(82)90313-7.

J. F. Brookfield. Models of repression of transposition in p-m hybrid dysgenesis by p cytotype and by zygotically encoded repressor proteins. *Genetics*, 128(2): 471–486, June 1991. doi: 10.1093/genetics/128.2.471. URL https://doi.org/10.1093/genetics/128.2.471.

A. Bucheton, C. Vaury, M. C. Chaboissier, P. Abad, A. Pélisson, and M. Simonelig. I elements and the drosophila genome. *Genetica*, 86(1-3):175–190, 1992. doi: 10.1007/bf00133719. URL https://doi.org/10.1007/bf00133719.

R. W. Carthew and E. J. Sontheimer. Origins and mechanisms of miRNAs and siRNAs. *Cell*, 136(4):642–655, Feb. 2009. doi: 10.1016/j.cell.2009.01.035. URL https://doi.org/10.1016/j.cell.2009.01.035.

B. Charlesworth and D. Charlesworth. The population dynamics of transposable elements. *Genetical Research*, 42(1):1–27, Aug. 1983. doi: 10.1017/s0016672300021455. URL https://doi.org/10.1017/s0016672300021455.

B. Charlesworth, A. Lapid, and D. Canada. The distribution of transposable elements within and between chromosomes in a population of *Drosophila melanogaster*. II. inferences on the nature of selection against elements. *Genetical Research*, 60(2):115–130, Oct. 1992. doi: 10.1017/s0016672300030809. URL https://doi.org/10.1017/s0016672300030809.

E. B. Chuong, N. C. Elde, and C. Feschotte. Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics*, 18(2):71–86, Nov. 2016. doi: 10.1038/nrg.2016.139. URL https://doi.org/10.1038/nrg.2016.139.

R. Cordaux and M. A. Batzer. The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10(10):691–703, Oct. 2009. doi: 10.1038/nrg2640. URL https://doi.org/10.1038/nrg2640.

K. M. Creasey, J. Zhai, F. Borges, F. V. Ex, M. Regulski, B. C. Meyers, and R. A. Martienssen. miRNAs trigger widespread epigenetically activated siRNAs from transposons in arabidopsis. *Nature*, 508(7496):411–415, Mar. 2014. doi: 10.1038/nature13069. URL https://doi.org/10.1038/nature13069.

J. F. Crow and M. Kimura. *An introduction to population genetics theory*. Blackburn Press, West Caldwell, NJ, Jan. 2009. ISBN 1932846123.

G. Deceliere, S. Charles, and C. Biémont. The dynamics of transposable elements in structured populations. *Genetics*, 169(1):467–474, Jan. 2005. doi: 10.1534/genetics.104.032243. URL https://doi.org/10.1534/genetics.104.032243.

E. S. Dolgin and B. Charlesworth. The fate of transposable elements in asexual populations. *Genetics*, 174(2):817–827, Oct. 2006. doi: 10.1534/genetics.106.060434. URL https://doi.org/10.1534/genetics.106.060434.

W. F. Doolittle and C. Sapienza. Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284(5757):601–603, Apr. 1980. doi: 10.1038/284601a0. URL https://doi.org/10.1038/284601a0.

M. R. Fabian and N. Sonenberg. The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. *Nature Structural & Molecular Biology*, 19(6):586–593, June 2012. doi: 10.1038/nsmb.2296. URL https://doi.org/10.1038/nsmb.2296.

M. Fablet, A. Akkouche, V. Braman, and C. Vieira. Variable expression levels detected in the drosophila effectors of piRNA biogenesis. *Gene*, 537(1):149–153, Mar. 2014. doi: 10.1016/j.gene.2013.11.095. URL https://doi.org/10.1016/j.gene.2013.11.095.

A. Ferraj, P. A. Audano, P. Balachandran, A. Czechanski, J. I. Flores, A. A. Radecki, V. Mosur, D. S. Gordon, I. A. Walawalkar, E. E. Eichler, L. G. Reinholdt, and C. R. Beck. Resolution of structural variation in diverse mouse genomes reveals chromatin remodeling due to transposable elements. Sept. 2022. doi: 10.1101/2022.09.26.509577. URL https://doi.org/10.1101/2022.09.26.509577.

C. Feschotte. Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, 9(5):397–405, May 2008. doi: 10.1038/nrg2337. URL https://doi.org/10.1038/nrg2337.

M. Galla, A. Schambach, C. S. Falk, T. Maetzig, J. Kuehle, K. Lange, D. Zychlinski, N. Heinz, M. H. Brugman, G. Göhring, Z. Izsvák, Z. Ivics, and C. Baum. Avoiding cytotoxicity of transposases by dose-controlled mRNA delivery. *Nucleic Acids Research*, 39(16):7147–7160, May 2011. doi: 10.1093/nar/gkr384. URL https://doi.org/10.1093/nar/gkr384.

D. Gebert, L. K. Neubert, C. Lloyd, J. Gui, R. Lehmann, and F. K. Teixeira. Large drosophila germline piRNA clusters are evolutionarily labile and dispensable for transposon regulation. *Molecular Cell*, 81(19):3965–3978.e5, Oct. 2021. doi: 10.1016/j.molcel.2021.07.011. URL https://doi.org/10.1016/j.molcel.2021.07.011.

G. E. Ghanim, D. C. Rio, and F. K. Teixeira. Mechanism and regulation of p element transposition. *Open Biology*, 10(12), Dec. 2020. doi: 10.1098/rsob.200244. URL https://doi.org/10.1098/rsob.200244.

C. Gilbert and F. Maumus. Multiple horizontal acquisitions of plant genes in the whitefly *Bemisia tabaci*. *Genome Biology and Evolution*, 14(10), Sept. 2022. doi: 10.1093/gbe/evac141. URL https://doi.org/10.1093/gbe/evac141.

E. A. Gladyshev and I. R. Arkhipova. Telomere-associated endonuclease-deficient *Penelope*-like retroelements in diverse eukaryotes. *Proceedings of the National Academy of Sciences*, 104(22):9352–9357, May 2007. doi: 10.1073/pnas.0702741104. URL https://doi.org/10.1073/pnas.0702741104.

Z. Guo, Z. Kuang, Y. Tao, H. Wang, M. Wan, C. Hao, F. Shen, X. Yang, and L. Li. Miniature inverted-repeat transposable elements drive rapid MicroRNA diversification in angiosperms. *Molecular Biology and Evolution*, 39(11), Oct. 2022. doi: 10.1093/molbev/msac224. URL https://doi.org/10.1093/molbev/msac224.

D. C. Hancks and H. H. Kazazian. Roles for retrotransposon insertions in human disease. *Mobile DNA*, 7(1), May 2016. doi: 10.1186/s13100-016-0065-9.

C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

L. He and G. J. Hannon. MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews Genetics*, 5(7):522–531, July 2004. doi: 10.1038/nrg1379. URL https://doi.org/10.1038/nrg1379.

M. Heinlem, T. Brattig, and R. Kunze. In vivo aggregation of maize activator (ac) transposase in nuclei of maize endosperm and petunia protoplasts. *The Plant Journal*, 5(5):705–714, May 1994. doi: 10.1111/j.1365-313x.1994.00705.x. URL https://doi.org/10.1111/j.1365-313x.1994.00705.x.

C. Hermant, A. Boivin, L. Teysset, V. Delmarre, A. Asif-Laidin, M. van den Beek, C. Antoniewski, and S. Ronsseray. Paramutation in *Drosophila* requires both nuclear and cytoplasmic actors of the piRNA pathway and induces *Cis*-spreading of piRNA production. *Genetics*, 201(4):1381–1396, Oct. 2015. doi: 10.1534/genetics.115.180307. URL https://doi.org/10.1534/genetics.115.180307.

M. E. Hood, M. Katawczik, and T. Giraud. Repeat-induced point mutation and the population structure of transposable elements in *Microbotryum violaceum*. *Genetics*, 170(3):1081–1089, July 2005. doi: 10.1534/genetics.105.042564. URL https://doi.org/10.1534/genetics.105.042564.

J. F. Hughes and J. M. Coffin. Human endogenous retrovirus k solo-LTR formation and insertional polymorphisms: Implications for human and viral evolution. *Proceedings of the National Academy of Sciences*, 101(6):1668–1672, Feb. 2004. doi: 10.1073/pnas.0307885100. URL https://doi.org/10.1073/pnas.0307885100.

D. Itou, Y. Shiromoto, Y. Shin-ya, C. Ishii, T. Nishimura, N. Ogonuki, A. Ogura, H. Hasuwa, Y. Fujihara, S. Kuramochi-Miyagawa, and T. Nakano. Induction of DNA methylation by artificial piRNA production in male germ cells. *Current Biology*, 25(7):901–906, mar 2015. doi: 10.1016/j.cub.2015.01.060. URL https://doi.org/10.1016%2Fj.cub.2015.01.060.

H.-O. Iwakawa and Y. Tomari. Life of RISC: Formation, action, and degradation of RNA-induced silencing complex. *Mol. Cell*, 82(1):30–43, Jan. 2022.

V. V. Kapitonov and J. Jurka. Self-synthesizing DNA transposons in eukaryotes. *Proceedings of the National Academy of Sciences*, 103(12):4540–4545, Mar. 2006. doi: 10.1073/pnas.0600833103. URL https://doi.org/10.1073/pnas.0600833103.

N. L. Kaplan and J. F. Y. Brookfield. TRANSPOSABLE ELEMENTS IN MENDELIAN POPULATIONS. III. STATISTICAL RESULTS. *Genetics*, 104(3):485–495, July 1983. doi: 10.1093/genetics/104.3.485. URL https://doi.org/10.1093/genetics/104.3.485.

P. J. Keeling and J. D. Palmer. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8):605–618, Aug. 2008. doi: 10.1038/nrg2386. URL https://doi.org/10.1038/nrg2386.

E. S. Kelleher, R. B. R. Azevedo, and Y. Zheng. The evolution of small-RNA-mediated silencing of an invading transposable element. *Genome Biology and Evolution*, 10(11):3038–3057, Sept. 2018. doi: 10.1093/gbe/evy218. URL https://doi.org/10.1093/gbe/evy218.

T. V. Kent, J. Uzunović, and S. I. Wright. Coevolution between transposable elements and recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1736):20160458, Nov. 2017. doi: 10.1098/rstb.2016.0458. URL https://doi.org/10.1098/rstb.2016.0458.

M. G. Kidwell and D. R. Lisch. PERSPECTIVE: TRANSPOSABLE ELEMENTS, PARASITIC DNA, AND GENOME EVOLUTION. *Evolution*, 55(1):1–24, Jan. 2001. doi: 10.1111/j.0014-3820.2001.tb01268.x. URL https://doi.org/10.1111/j.0014-3820.2001.tb01268.x.

M. G. Kidwell, J. F. Kidwell, and J. A. Sved. Hybrid dysgenesis in DROSOPHILA MELANOGASTER: A syndrome of aberrant traits including mutation, sterility and male recombination. *Genetics*, 86(4):813–833, Aug. 1977. doi: 10.1093/genetics/86.4.813. URL https://doi.org/10.1093/genetics/86.4.813.

R. Kofler. Dynamics of transposable element invasions with piRNA clusters. *Molecular Biology and Evolution*, 36(7):1457–1472, Apr. 2019. doi: 10.1093/molbev/msz079. URL https://doi.org/10.1093/molbev/msz079.

E. H. Koo, P. T. Lansbury, and J. W. Kelly. Amyloid diseases: Abnormal protein aggregation in neurodegeneration. *Proceedings of the National Academy of Sciences*, 96(18):9989–9990, Aug. 1999. doi: 10.1073/pnas.96.18.9989. URL https://doi.org/10.1073/pnas.96.18.9989.

C. H. Langley, J. F. Y. Brookfield, and N. Kaplan. TRANSPOSABLE ELEMENTS IN MENDELIAN POPULATIONS. I. A THEORY. *Genetics*, 104(3):457–471, July 1983. doi: 10.1093/genetics/104.3.457. URL https://doi.org/10.1093/genetics/104.3.457.

K. Laricchia, S. Zdraljevic, D. Cook, and E. Andersen. Natural variation in the distribution and abundance of transposable elements across the *Caenorhabditis elegans* species. *Molecular Biology and Evolution*, 34(9):2187–2202, May 2017. doi: 10.1093/molbev/msx155. URL https://doi.org/10.1093/molbev/msx155.

A. Le Rouzic and G. Deceliere. Models of the population genetics of transposable elements. *Genetical Research*, 85(3):171–181, June 2005. doi: 10.1017/s0016672305007585. URL https://doi.org/10.1017/s0016672305007585.

A. Le Rouzic, T. S. Boutin, and P. Capy. Long-term evolution of transposable elements. *Proceedings of the National Academy of Sciences*, 104(49):19375–19380, Dec. 2007. doi: 10.1073/pnas.0705238104. URL https://doi.org/10.1073/pnas.0705238104.

T. Lenormand, J. Engelstädter, S. E. Johnston, E. Wijnker, and C. R. Haag. Evolutionary mysteries in meiosis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1706):20160001, Oct. 2016. doi: 10.1098/rstb.2016.0001. URL https://doi.org/10.1098/rstb.2016.0001.

X. Z. Li, C. K. Roy, M. J. Moore, and P. D. Zamore. Defining pirna primary transcripts. *Cell Cycle*, 12(11):1657–1658, 2013. doi: 10.4161/cc.24989. URL https://doi.org/10.4161/cc.24989. PMID: 23673320.

J. Lilue, A. G. Doran, I. T. Fiddes, M. Abrudan, J. Armstrong, R. Bennett, W. Chow, J. Collins, S. Collins, A. Czechanski, P. Danecek, M. Diekhans, D.-D. Dolle, M. Dunn, R. Durbin, D. Earl, A. Ferguson-Smith, P. Flicek, J. Flint, A. Frankish, B. Fu, M. Gerstein, J. Gilbert, L. Goodstadt, J. Harrow, K. Howe, X. Ibarra-Soria, M. Kolmogorov, C. J. Lelliott, D. W. Logan, J. Loveland, C. E. Mathews, R. Mott, P. Muir, S. Nachtweide, F. C. P. Navarro, D. T. Odom, N. Park, S. Pelan, S. K. Pham, M. Quail, L. Reinholdt, L. Romoth, L. Shirley, C. Sisu, M. Sjoberg-Herrera, M. Stanke, C. Steward, M. Thomas, G. Threadgold, D. Thybert, J. Torrance, K. Wong, J. Wood, B. Yalcin, F. Yang, D. J. Adams, B. Paten, and T. M. Keane. Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nature Genetics*, 50(11):1574–1583, Oct. 2018. doi: 10.1038/s41588-018-0223-8. URL https://doi.org/10.1038/s41588-018-0223-8.

J. K. Lim and M. J. Simmons. Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. *BioEssays*, 16(4):269–275, Apr. 1994. doi: 10.1002/bies.950160410. URL https://doi.org/10.1002/bies.950160410.

H. Lin. piRNAs in the germ line. *Science*, 316(5823):397–397, Apr. 2007. doi: 10.1126/science.1137543. URL https://doi.org/10.1126/science.1137543.

H. Lin and A. Spradling. A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the *Drosophila* ovary. *Development*, 124(12):2463–2476, June 1997. doi: 10.1242/dev.124.12.2463. URL https://doi.org/10.1242/dev.124.12.2463.

A. Lohe, D. De Aguiar, and D. Hartl. Mutations in the *mariner* transposase: The d, d(35)e consensus sequence is nonfunctional. *Proceedings of the National Academy of Sciences*, 94(4):1293–1297, Feb. 1997. doi: 10.1073/pnas.94.4.1293. URL https://doi.org/10.1073/pnas.94.4.1293.

A. R. Lohe and D. L. Hartl. Autoregulation of mariner transposase activity by overproduction and dominant-negative complementation. *Molecular Biology and Evolution*, 13(4):549–555, Apr. 1996. doi: 10.1093/oxfordjournals.molbev.a025615. URL https://doi.org/10.1093/oxfordjournals.molbev.a025615.

J. Lu and A. G. Clark. Population dynamics of PIWI-interacting RNAs (piRNAs) and their targets in *Drosophila*. *Genome Research*, 20(2):212–227, Nov. 2009. doi: 10.1101/gr.095406.109. URL https://doi.org/10.1101/gr.095406.109.

P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26): 15324–15328, Dec. 2003. doi: 10.1073/pnas.0306899100. URL https://doi.org/10.1073/pnas.0306899100.

B. McClintock. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, 36(6):344–355, June 1950. doi: 10.1073/pnas.36.6.344. URL https://doi.org/10.1073/pnas.36.6.344.

V. Mérel, M. Boulesteix, M. Fablet, and C. Vieira. Transposable elements in drosophila. *Mobile DNA*, 11(1), July 2020. doi: 10.1186/s13100-020-00213-z. URL https://doi.org/10.1186/s13100-020-00213-z.

D. M. Messerschmidt, B. B. Knowles, and D. Solter. DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes & Development*, 28(8):812–828, Apr. 2014. doi: 10.1101/gad.234294.113. URL https://doi.org/10.1101/gad.234294.113.

F. Mohn, G. Sienski, D. Handler, and J. Brennecke. The rhino-deadlock-cutoff complex licenses noncanonical transcription of dual-strand piRNA clusters in drosophila. *Cell*, 157(6):1364–1379, June 2014. doi: 10.1016/j.cell.2014.04.031. URL https://doi.org/10.1016/j.cell.2014.04.031.

H. Muller. The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 1(1):2–9, May 1964. doi: 10.1016/0027-5107(64)90047-8. URL https://doi.org/10.1016/0027-5107(64)90047-8.

Y. Murota, H. Ishizu, S. Nakagawa, Y. W. Iwasaki, S. Shibata, M. K. Kamatani, K. Saito, H. Okano, H. Siomi, and M. C. Siomi. Yb integrates piRNA intermediates and processing factors into perinuclear bodies to enhance piRISC assembly. *Cell Rep.*, 8(1):103–113, July 2014.

W. B. Neaves and P. Baumann. Unisexual reproduction among vertebrates. *Trends in Genetics*, 27(3):81–88, Mar. 2011. doi: 10.1016/j.tig.2010.12.002. URL https://doi.org/10.1016/j.tig.2010.12.002.

R. W. Nowell, C. G. Wilson, P. Almeida, P. H. Schiffer, D. Fontaneto, L. Becks, F. Rodriguez, I. R. Arkhipova, and T. G. Barraclough. Evolutionary dynamics of transposable elements in bdelloid rotifers. *eLife*, 10, Feb. 2021. doi: 10.7554/elife.63194. URL https://doi.org/10.7554/elife.63194.

Y. Ophinni, U. Palatini, Y. Hayashi, and N. F. Parrish. PiRNA-guided CRISPR-like immunity in eukaryotes. *Trends Immunol.*, 40(11):998–1010, Nov. 2019.

L. E. Orgel and F. H. Crick. Selfish DNA: the ultimate parasite. *Nature*, 284(5757): 604–607, Apr. 1980.

D. M. Ozata, I. Gainetdinov, A. Zoch, D. O'Carroll, and P. D. Zamore. PIWI-interacting RNAs: small RNAs with big functions. *Nature Reviews Genetics*, 20(2):89–108, Nov. 2018. doi: 10.1038/s41576-018-0073-3. URL https://doi.org/10.1038/s41576-018-0073-3.

D. M. Özata, T. Yu, H. Mou, I. Gainetdinov, C. Colpan, K. Cecchini, Y. Kaymaz, P.-H. Wu, K. Fan, A. Kucukural, Z. Weng, and P. D. Zamore. Evolutionarily conserved pachytene piRNA loci are highly divergent among modern humans. *Nature Ecology & Evolution*, 4(1):156–168, Dec. 2019. doi: 10.1038/s41559-019-1065-1. URL https://doi.org/10.1038/s41559-019-1065-1.

Özgen Deniz, J. M. Frost, and M. R. Branco. Regulation of transposable elements by DNA modifications. *Nature Reviews Genetics*, 20(7):417–431, Mar. 2019. doi: 10.1038/s41576-019-0106-6. URL https://doi.org/10.1038/s41576-019-0106-6.

J. K. Pace and C. Feschotte. The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. *Genome Research*, 17(4): 422–432, Mar. 2007. doi: 10.1101/gr.5826307. URL https://doi.org/10.1101/gr.5826307.

S. S. Parhad and W. E. Theurkauf. Rapid evolution and conserved function of the piRNA pathway. *Open Biology*, 9(1), Jan. 2019. doi: 10.1098/rsob.180181. URL https://doi.org/10.1098/rsob.180181.

J. Peccoud, V. Loiseau, R. Cordaux, and C. Gilbert. Massive horizontal transfer of transposable elements in insects. *Proceedings of the National Academy of Sciences*, 114(18):4721–4726, Apr. 2017. doi: 10.1073/pnas.1621178114. URL https://doi.org/10.1073/pnas.1621178114.

R. Petri, P. L. Brattås, Y. Sharma, M. E. Jönsson, K. Pircs, J. Bengzon, and J. Jakobsson. LINE-2 transposable elements are a source of functional human microRNAs and target sites. *PLOS Genetics*, 15(3):e1008036, Mar. 2019. doi: 10.1371/journal.pgen.1008036. URL https://doi.org/10.1371/journal.pgen.1008036.

D. A. Petrov, A.-S. Fiston-Lavier, M. Lipatov, K. Lenkov, and J. Gonzalez. Population genomics of transposable elements in drosophila melanogaster. *Molecular Biology and Evolution*, 28(5):1633–1644, Dec. 2010. doi: 10.1093/molbev/msq337. URL https://doi.org/10.1093/molbev/msq337.

J. Piriyapongsa and I. K. Jordan. Dual coding of siRNAs and miRNAs by plant transposable elements. *RNA*, 14(5):814–821, Mar. 2008. doi: 10.1261/rna.916708. URL https://doi.org/10.1261/rna.916708.

H. Quesneville and D. Anxolabéhère. Dynamics of transposable elements in metapopulations: A model of p element invasion in *Drosophila*. *Theoretical Population Biology*, 54(2):175–193, Oct. 1998. doi: 10.1006/tpbi.1997.1353. URL https://doi.org/10.1006/tpbi.1997.1353.

S. Ravindran. Barbara McClintock and the discovery of jumping genes. *Proceedings of the National Academy of Sciences*, 109(50):20198–20199, Dec. 2012. doi: 10.1073/pnas.1219372109. URL https://doi.org/10.1073/pnas.1219372109.

L. Rishishwar, C. E. T. Villa, and I. K. Jordan. Transposable element polymorphisms recapitulate human evolution. *Mobile DNA*, 6(1), Nov. 2015. doi: 10.1186/s13100-015-0052-6. URL https://doi.org/10.1186/s13100-015-0052-6.

É. Robillard, A. Le Rouzic, Z. Zhang, P. Capy, and A. Hua-Van. Experimental evolution reveals hyperparasitic interactions among transposable elements. *Proceedings of the National Academy of Sciences*, 113(51):14763–14768, Dec. 2016. doi: 10.1073/pnas.1524143113. URL https://doi.org/10.1073/pnas.1524143113.

F. Rodriguez and I. R. Arkhipova. Multitasking of the piRNA silencing machinery: Targeting transposable elements and foreign genes in the bdelloid rotifer *Adineta vaga*. *Genetics*, 203(1):255–268, May 2016. doi: 10.1534/genetics.116.186734. URL https://doi.org/10.1534/genetics.116.186734.

E. F. Roovers, D. Rosenkranz, M. Mahdipour, C.-T. Han, N. He, S. M. C. de Sousa Lopes, L. A. van der Westerlaken, H. Zischler, F. Butter, B. A. Roelen, and R. F. Ketting. Piwi proteins and piRNAs in mammalian oocytes and early embryos. *Cell Reports*, 10(12):2069–2082, Mar. 2015. doi: 10.1016/j.celrep.2015.02.062. URL https://doi.org/10.1016/j.celrep.2015.02.062.

B. Saint-Leandre, P. Capy, A. Hua-Van, and J. Filée. piRNA and transposon dynamics in *Drosophila*: A female story. *Genome Biology and Evolution*, 12(6):931–947, May 2020. doi: 10.1093/gbe/evaa094. URL https://doi.org/10.1093/gbe/evaa094.

S. L. Salzberg. Horizontal gene transfer is not a hallmark of the human genome. *Genome Biology*, 18(1), May 2017. doi: 10.1186/s13059-017-1214-2. URL https://doi.org/10.1186/s13059-017-1214-2.

M. Sasaki, J. Lange, and S. Keeney. Genome destabilization by homologous recombination in the germ line. *Nature Reviews Molecular Cell Biology*, 11(3): 182–195, Feb. 2010. doi: 10.1038/nrm2849. URL https://doi.org/10.1038/nrm2849.

T. Sasaki, A. Shiohama, S. Minoshima, and N. Shimizu. Identification of eight members of the argonaute family in the human genome. *Genomics*, 82(3): 323–330, Sept. 2003. doi: 10.1016/s0888-7543(03)00129-0. URL https://doi.org/10.1016/s0888-7543(03)00129-0.

A. Scarpa and R. Kofler. The impact of paramutations on the invasion dynamics of transposable elements. Mar. 2023. doi: 10.1101/2023.03.14.532580. URL https://doi.org/10.1101/2023.03.14.532580.

S. Schaack, C. Gilbert, and C. Feschotte. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in Ecology & Evolution*, 25(9):537–546, Sept. 2010. doi: 10.1016/j.tree.2010.06.001. URL https://doi.org/10.1016/j.tree.2010.06.001.

R. Shalgi, Y. Pilpel, and M. Oren. Repression of transposable-elements – a microRNA anti-cancer defense mechanism? *Trends in Genetics*, 26(6):253–259, June 2010. doi: 10.1016/j.tig.2010.03.006. URL https://doi.org/10.1016/j.tig.2010.03.006.

S. Shpiz, S. Ryazansky, I. Olovnikov, Y. Abramov, and A. Kalmykova. Euchromatic transposon insertions trigger production of novel pi- and endo-siRNAs at the target sites in the drosophila germline. *PLoS Genetics*, 10 (2):e1004138, Feb. 2014. doi: 10.1371/journal.pgen.1004138. URL https://doi.org/10.1371/journal.pgen.1004138.

M. J. Simmons and L. M. Bucholz. Transposase titration in *Drosophila melanogaster*: a model of cytotype in the p-m system of hybrid dysgenesis. *Proceedings of the National Academy of Sciences*, 82(23):8119–8123, Dec. 1985. doi: 10.1073/pnas.82.23.8119. URL https://doi.org/10.1073/pnas.82.23.8119.

M. J. Simmons, M. W. Thorp, J. T. Buschette, and J. R. Becker. Transposon regulation in *Drosophila*: piRNA-producing p elements facilitate repression of hybrid dysgenesis by a p element that encodes a repressor polypeptide. *Molecular Genetics and Genomics*, 290(1):127–140, Aug. 2014. doi: 10.1007/s00438-014-0902-9. URL https://doi.org/10.1007/s00438-014-0902-9.

S. M. Soucy, J. Huang, and J. P. Gogarten. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16(8):472–482, July 2015. doi: 10.1038/nrg3962. URL https://doi.org/10.1038/nrg3962.

M. J. Stanhope, A. Lupas, M. J. Italia, K. K. Koretke, C. Volker, and J. R. Brown. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature*, 411(6840):940–944, June 2001. doi: 10.1038/35082058. URL https://doi.org/10.1038/35082058.

M. C. Stitzer, S. N. Anderson, N. M. Springer, and J. Ross-Ibarra. The genomic ecosystem of transposable elements in maize. *PLOS Genetics*, 17(10):e1009768, Oct. 2021. doi: 10.1371/journal.pgen.1009768. URL https://doi.org/10.1371/journal.pgen.1009768.

F.-J. Sun, S. Fleurdépine, C. Bousquet-Antonelli, G. Caetano-Anollés, and J.-M. Deragon. Common evolutionary trends for SINE RNA structures. *Trends in Genetics*, 23(1):26–33, Jan. 2007. doi: 10.1016/j.tig.2006.11.005. URL https://doi.org/10.1016/j.tig.2006.11.005.

M. Sunnåker, A. G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz. Approximate bayesian computation. *PLoS Computational Biology*, 9(1):e1002803, Jan. 2013. doi: 10.1371/journal.pcbi.1002803. URL https://doi.org/10.1371/journal.pcbi.1002803.

M. A. Surani and P. Hajkova. Epigenetic reprogramming of mouse germ cells toward totipotency. *Cold Spring Harbor Symposia on Quantitative Biology*, 75(0):211–218, Jan. 2010. doi: 10.1101/sqb.2010.75.010. URL https://doi.org/10.1101/sqb.2010.75.010.

D. C. Swarts, K. Makarova, Y. Wang, K. Nakanishi, R. F. Ketting, E. V. Koonin, D. J. Patel, and J. van der Oost. The evolutionary journey of argonaute proteins. *Nature Structural & Molecular Biology*, 21(9):743–753, Sept. 2014. doi: 10.1038/nsmb.2879. URL https://doi.org/10.1038/nsmb.2879.

O. Tange. *Gnu Parallel 2018*. Zenodo, 2018. doi: 10.5281/ZENODO.1146014. URL https://zenodo.org/record/1146014.

S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, Feb. 1997. doi: 10.1093/genetics/145.2.505. URL https://doi.org/10.1093/genetics/145.2.505.

A. L. Taylor. BACTERIOPHAGE-INDUCED MUTATION IN ESCHERICHIA COLI. *Proceedings of the National Academy of Sciences*, 50(6):1043–1051, Dec. 1963. doi: 10.1073/pnas.50.6.1043. URL https://doi.org/10.1073/pnas.50.6.1043.

A. L. Thomas, E. Stuwe, S. Li, J. Du, G. Marinov, N. Rozhkov, Y.-C. A. Chen, Y. Luo, R. Sachidanandam, K. F. Toth, D. Patel, and A. A. Aravin. Transgenerationally inherited piRNAs trigger piRNA biogenesis by changing the chromatin of piRNA clusters and inducing precursor processing. *Genes & Development*, 28(15):1667–1680, Aug. 2014. doi: 10.1101/gad.245514.114. URL https://doi.org/10.1101/gad.245514.114.

J. W. Valentine, A. G. Collins, and C. P. Meyer. Morphological complexity increase in metazoans. *Paleobiology*, 20(2):131–142, 1994. ISSN 00948373, 19385331. URL http://www.jstor.org/stable/2401015.

T. Volpe and R. A. Martienssen. RNA interference and heterochromatin assembly. *Cold Spring Harbor Perspectives in Biology*, 3(9):a003731–a003731, Jan. 2011. doi: 10.1101/cshperspect.a003731. URL https://doi.org/10.1101/cshperspect.a003731.

A. T. Walsh, D. A. Triant, J. J. L. Tourneau, M. Shamimuzzaman, and C. G. Elsik. Hymenoptera genome database: new genomes and annotation datasets for improved go enrichment and orthologue analyses. *Nucleic Acids Research*, 50(D1):D1032–D1039, Nov. 2021. doi: 10.1093/nar/gkab1018. URL https://doi.org/10.1093/nar/gkab1018.

W. Wang, M. Yoshikawa, B. W. Han, N. Izumi, Y. Tomari, Z. Weng, and P. D. Zamore. The initial uridine of primary piRNAs does not create the tenth adenine that is the hallmark of secondary piRNAs. *Molecular Cell*, 56 (5):708–716, Dec. 2014. doi: 10.1016/j.molcel.2014.10.016. URL https://doi.org/10.1016/j.molcel.2014.10.016.

T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, and A. H. Schulman. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8(12):973–982, Dec. 2007. doi: 10.1038/nrg2165. URL https://doi.org/10.1038/nrg2165.

T. Wicker, Y. Yu, G. Haberer, K. F. X. Mayer, P. R. Marri, S. Rounsley, M. Chen, A. Zuccolo, O. Panaud, R. A. Wing, and S. Roffler. DNA transposon activity is associated with increased mutation rates in genes of rice and other grasses. *Nature Communications*, 7(1), Sept. 2016. doi: 10.1038/ncomms12790. URL https://doi.org/10.1038/ncomms12790.

F. Wierzbicki and R. Kofler. The composition of piRNA clusters in *Drosophila melanogaster* deviates from expectations under the trap model. Feb. 2023. doi: 10.1101/2023.02.14.528490. URL https://doi.org/10.1101/2023.02.14.528490.

S. Yamanaka, M. C. Siomi, and H. Siomi. piRNA clusters and open chromatin structure. *Mobile DNA*, 5(1), Aug. 2014. doi: 10.1186/1759-8753-5-22. URL https://doi.org/10.1186/1759-8753-5-22.

G. Yannopoulos, N. Stamatis, M. Monastirioti, P. Hatzopoulos, and C. Louis. hobo is responsible for the induction of hybrid dysgenesis by strains of drosophila melanogaster bearing the male recombination factor 23.5mrf. *Cell*, 49(4):487–495, May 1987. doi: 10.1016/0092-8674(87)90451-x. URL https://doi.org/10.1016/0092-8674(87)90451-x.

T. Yu, K. Fan, D. M. Özata, G. Zhang, Y. Fu, W. E. Theurkauf, P. D. Zamore, and Z. Weng. Long first exons and epigenetic marks distinguish conserved pachytene piRNA clusters from other mammalian genes. *Nature Communications*, 12(1), Jan. 2021. doi: 10.1038/s41467-020-20345-3. URL https://doi.org/10.1038/s41467-020-20345-3.

H.-H. Zhang, J. Peccoud, M.-R.-X. Xu, X.-G. Zhang, and C. Gilbert. Horizontal transfer and evolution of transposable elements in vertebrates. *Nature Communications*, 11(1), Mar. 2020. doi: 10.1038/s41467-020-15149-4. URL https://doi.org/10.1038/s41467-020-15149-4.

Y. Zhang, T. Li, S. Preissl, M. L. Amaral, J. D. Grinstein, E. N. Farah, E. Destici, Y. Qiu, R. Hu, A. Y. Lee, S. Chee, K. Ma, Z. Ye, Q. Zhu, H. Huang, R. Fang, L. Yu, J. C. I. Belmonte, J. Wu, S. M. Evans, N. C. Chi, and B. Ren. Transcriptionally active HERV-h retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nature Genetics*, 51(9):1380–1388, Aug. 2019. doi: 10.1038/s41588-019-0479-7. URL https://doi.org/10.1038/s41588-019-0479-7.