A Population Genetics Theory for piRNA-regulated Transposable Elements

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

1 Laboratoire Évolution: Génomes, Comportement, Écologie; IDEEV; Université Paris-Saclay, CNRS, IRD.

*** arnaud.le-rouzic@universite-paris-saclay.fr**

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 pi-RNAs, which require a specific mutational event (the insertion of a TE copy in a pi-RNA 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 differed substantially from the existing expectations. We proposed three sub-models, depending on whether or not genomic TE copies and pi-RNA 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 monotonous, 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 consensual "TE life cycle" model (Kidwell and Lisch 2001; Wallau et al. 2016), TE families do not maintain 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 less 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 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 of 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 small-RNA 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). Small-RNA regulation seems to concern a wide range of species, and defines a regulation scenario known as the "trap model" (Kofler 2019). In such a scenario, regulation is triggered by the insertion of a TE in specific "trap" regions of the genome, the pi-clusters. TE sequences inserted in pi-clusters are transcribed into small regulating pi-RNAs, that are able to silence homologous mRNAs from close TE families by recruiting proteins from the PIWI family.

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 are a mixture of permissive and repressive genetic backgrounds), and (iii) the regulation mechanism is independent from the 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 pi-RNA 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 that the major factor conditioning the TE success (in terms of copy number) is not the transposition rate, but rather the size of the pi-clusters. 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 approximations for the trap model equilibrium copy number and equilibrium cluster frequency in three scenarios: (i) neutral TEs, (ii) deleterious TEs and neutral clusters, and (iii) deleterious TEs and deleterious clusters. We show that these scenarios correspond to qualitatively distinct outcomes, and we validate the predicted equilibria based on 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 is constant, the copy number dynamics is exponential. If the transposition rate u_n is regulated by the copy number ($u_0 > v$, $du_n/dn < 0$), $\overline{u_n} \simeq u_{\bar{n}}$, and $\lim(u_n) < v$, with $u_{\hat{n}} = v$, 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 across

generations as:

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

where $s_{\bar{n}} = \partial \log w_n / \partial n|_{\bar{n}}$. 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). After transposition, the copy number rises to $\bar{n} = \bar{n}(1+u)$, while its variance becomes $\mathrm{V}(n) = \bar{n}(1+u)^2$. In the following, we used a correction $s' \simeq s(1+2u)$; this correction is subtle and was proposed solely to ensure a better match with numerical simulations, it does not affect the conclusions qualitatively.

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 Gra14. 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 $n = 10,000$ possible TE insertion sites. k pi-clusters of size $n\pi/k$ were distributed on different chromosomes, the parameter π standing for the proportion of the n loci corresponding to pi-clusters. Insertion sites were freely recombining, except within pi-clusters. Generations were non-overlapping; reproduction consisted in generating and pairing randomly $2N$ haploid gametes from $2N$ parents sampled with replacement proportionally to their fitness. Transposition occurred with a rate u_i computed for each individual as a function of its genotype at pi-clusters, with the same assumptions as described below. The location of the transposed copy was drawn uniformly in the diploid genome. Transposition events in occupied loci were cancelled, which happened rarely as TE genome contents were always far from saturation. Populations were initialized with 10 heterozygote insertions (in non-piRNA loci), randomly distributed in the population at frequency $p0 = 0.05$ each, resulting in $n0 = 1$ copy in average per diploid individual. For each parameter set, simulations were replicated 10 times, and the average number of diploid TE copies was reported. Average cluster frequencies were calculated by dividing the number of diploid TE copies in pi-clusters by $2k$; frequencies could theoretically be slightly > 1 due to rare events in which several TEs could insert simultaneously in the same cluster. 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 a cluster region 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 $1-p$). Allele frequencies at all clusters were considered to be the same (infinite population and identical cluster

properties). If the regulatory allele frequency at generation t is p_t, the average number of cluster insertions for a diploid individual is $2kp_t$. TE deletions were neglected ($v = 0$). The presence of a single regulatory allele at any cluster site was supposed to trigger complete regulation: the transposition rate per copy and per generation was u in absence of regulation, and 0 otherwise. Assuming random mating and no linkage disequilibrium (i.e. no correlation between n and the genotype at the regulatory cluster), we approximated the discrete generation model with a continuous process, and the neutral model (equation 2) was rewritten as a set of two differential equations on $\bar{n}$ (relabelled n for simplicity) and p:

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

Initially, there are n_0 copies per individual in the population, and $p_0 = 0$.

The system of equation 3 admits three equilibria (characterized by the equilibrium values $\hat{n}$ and $\hat{p}$): $E_1 : u = 0$ (no transposition, $\hat{n} = n_0$ and $\hat{p} = 0$), $E_2 : \hat{n} = 0$ (loss of the transposable element), and $E_3 : \hat{p} = 1$ (fixation of all regulatory clusters). Equilibria E_1 and E_2 do not need to be investigated further, as $u = 0$, $n_0 = 0$, or $p_0 = 1$ do trivially result in the absence of any 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:

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

Cluster fixation is asymptotic ($\lim_{t\to\infty} p = 1$), and the equilibrium is asymptotically stable ($\mathrm{d}n/\mathrm{d}t > 0$ and $\mathrm{d}p/\mathrm{d}t > 0$). Figure 1 illustrates the effect of u and k on the dynamics n_t and p_t. 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. The fact that the equilibrium copy number does not depend on the transposition rate u is one of the most counter-intuitive results of the trap model. This prediction was confirmed by simulations. It relies on the absence of linkage disequilibrium between regulatory clusters and genomic copies. 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 equations 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 $\pi \ll 1$, (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: (i) TE insertions

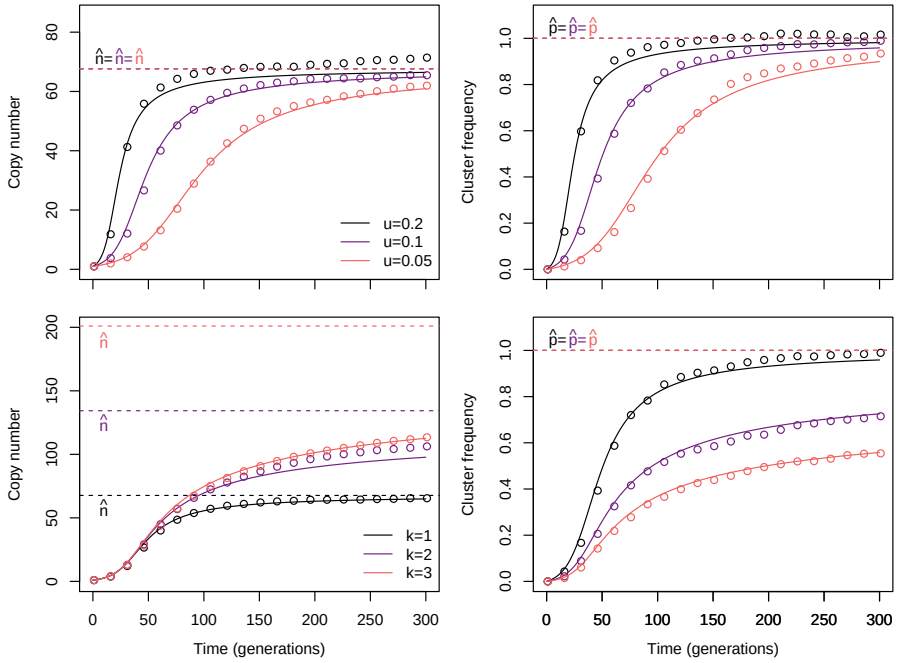

Figure 1: Dynamics in the neutral piRNA model, $n_0 = 1, \pi = 0.03$. The top panel illustrates the influence of the transposition rate, the bottom panel of the number of clusters. Left: number of copies n, right: frequency of the segregating clusters in the population (p). Open symbols: simulations, plain lines: difference equations, hyphenated lines: predicted equilibria. The copy number equilibrium $\hat{n}$ does not depend on the transposition rate, and the cluster frequency at equilibrium $\hat{p} = 1$ in all conditions.

in pi-clusters are neutral, and (ii) TE insertions in pi-clusters are as deleterious as the other insertions.

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

$$\begin{aligned} \frac{\mathrm{d}n}{\mathrm{d}t} &= nu(1-p)^{2k} - ns' \\ \frac{\mathrm{d}p}{\mathrm{d}t} &= \frac{\pi}{2k} nu(1-p)^{2k}. \end{aligned} \tag{5}$$

This equation only gives two equilibria, $E_2 : \hat{n} = 0$, and $E_3 : s' = 0$ and $\hat{p} = 1$, which is the same as for the neutral model (equation 3): no selection and fixation of all regulatory clusters. 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 ($\mathrm{d}n/\mathrm{d}t > 0$) up to a maximum n^*, which is achieved when $p = p^*$ (Figure 2). The maximum copy number can be

obtained analytically (Appendix A.1):

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

Once the maximum number of copies is achieved, cluster copies keep on accumulating, decreasing the transposition rate, which leads to a decrease in the copy number, up to the loss of the the element ($\hat{n} = 0$ at equilibrium). At that stage, clusters are not fixed, and the equilibrium cluster frequency $\hat{p}$ can be expressed as a function of copy number and cluster frequency at the maximum (p^* and n^*) (Appendix A.2):

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

from which an exact solution for $\hat{p}$ could not be calculated. The following approximation (from Appendix A.3):

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

happens to be acceptable for a wide range of transposition rates and for small selection coefficients ($s < 0.1$) (Figure 3).

Equation 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 cluster insertions are as deleterious as other TEs, selection acts on cluster frequency as predicted by population genetics (assuming no dominance):

$$\begin{aligned} \frac{\mathrm{d}n}{\mathrm{d}t} &= nu(1-p)^{2k} - ns' \\ \frac{\mathrm{d}p}{\mathrm{d}t} &= \frac{\pi}{2k}nu(1-p)^{2k} - sp\frac{1-p}{1-sp}. \end{aligned} \tag{10}$$

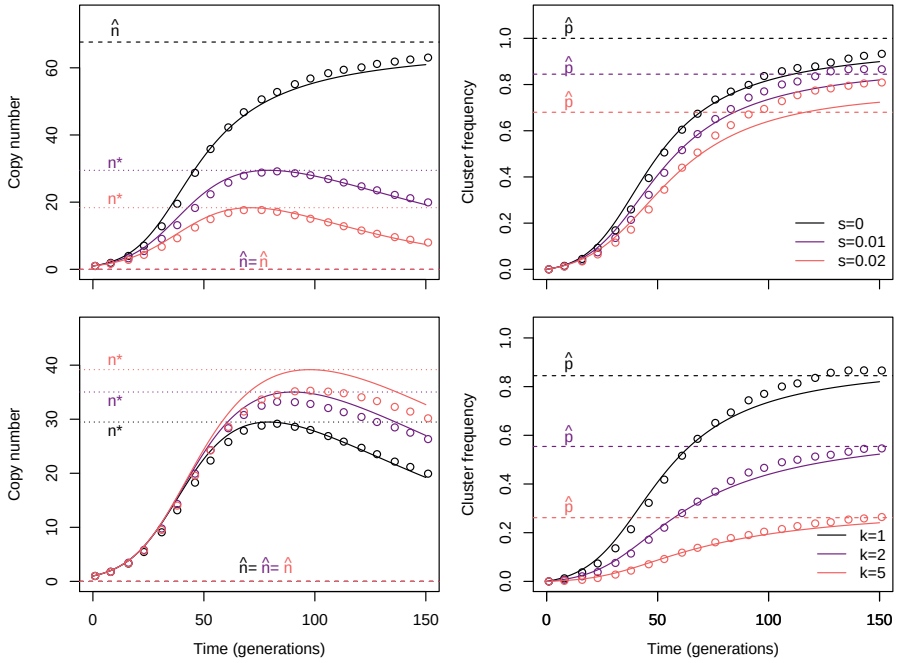

Figure 2: Dynamics of the deleterious TE - neutral cluster model. The top panel illustrates the influence of the selection coefficient s (with $u = 0.1, k = 1$), the bottom panel of the number of clusters k (with $s = 0.01, u = 0.1$). Left: number of copies n, right: frequency of the segregating clusters in the population p. Open symbols: simulations, plain lines: difference equations, hyphenated lines: predicted cluster frequency equilibrium $\hat{p}$, dotted lines: predicted copy number maximum n^*. Whenever $s > 0$, the copy number equilibrium $\hat{n}$ is 0.

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. The dynamics for n and p are illustrated in Figure 4, and the influence of model parameters (u, s, and π) on equilibrium values are depicted in Figure 5.

A linear stability analysis (Appendix A.4) 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 sizes, 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

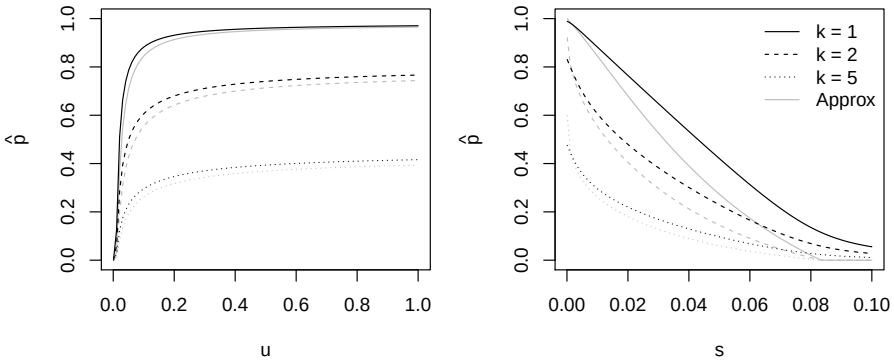

Figure 3: Equilibrium cluster frequency $\hat{p}$ for the deleterious TE - neutral cluster model as a function of the transposition rate u and the selection coefficient s. The number of clusters k is indicated with different line types. The approximation proposed in equation 8 is illustrated in gray.

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 deleterious TE - neutral cluster model does not allow for an equilibrium, comparisons had to be performed before the stabilization of the copy number (arbitrarily, at $T = 100$ 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 (Figure 6). 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 equlibrium 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, TE amplification 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 ($N < 100$), but this effect never exceeds 20% of the total copy number (Figure B1). Overall, drift has a very limited impact on the mean copy number when $N > 50$.

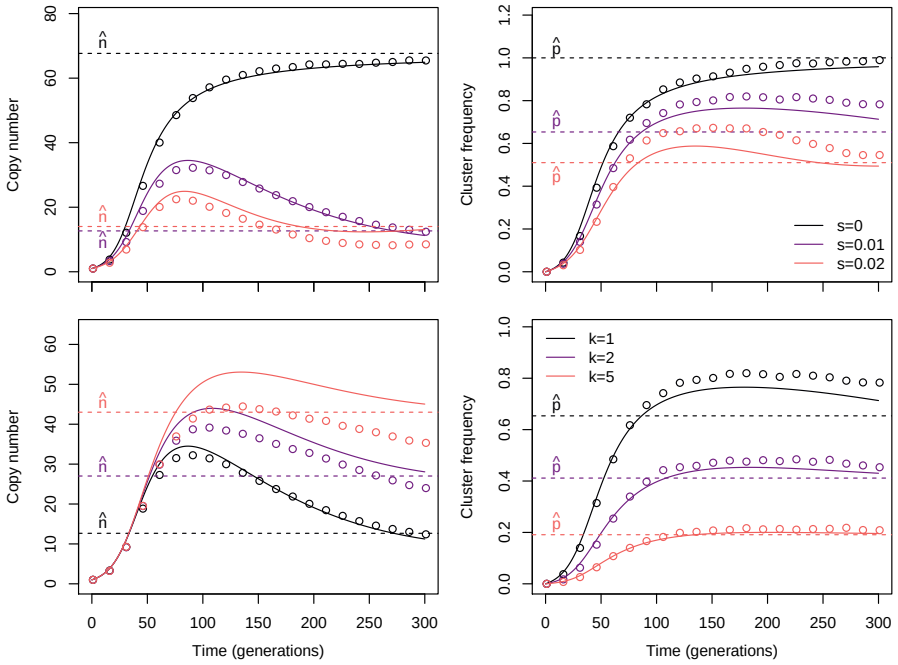

Figure 4: Dynamics of the copy number (n, left) and the cluster frequency (p, right) in the deleterious TE - deleterious cluster model. Top panels: influence of the selection coefficient, bottom panels: influence of the number of clusters. Plain lines: predicted dynamics from equation 10, hyphenated lines: predicted equilibrium (eq. 11), open circles: simulations.

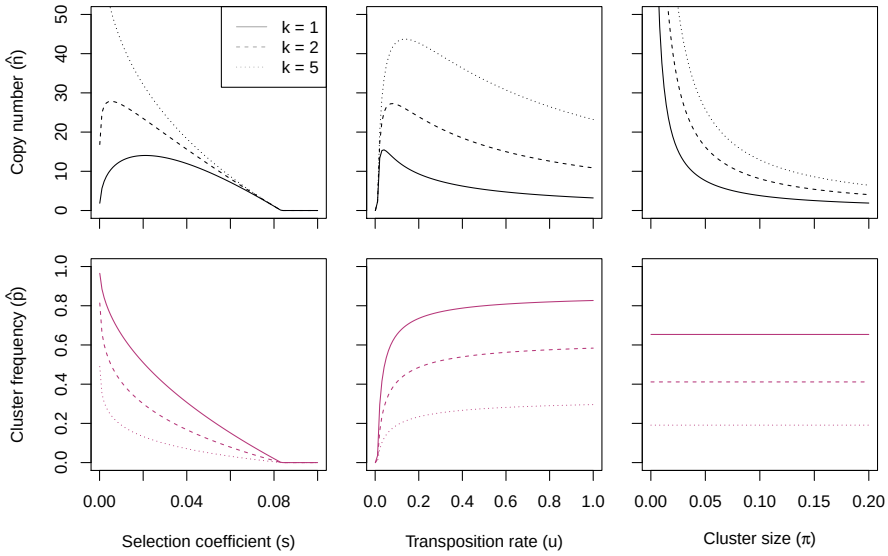

Figure 5: Equilibrium copy number ($\hat{n}$, black) and cluster frequency ($\hat{p}$, red) as a function of model parameters (u, s, and π) in the deleterious TEs - deleterious cluster model. Default parameter values were $u = 0.1$, $s = 0.01$, and $\pi = 0.03$. The number of clusters ($k = 1$, $k = 2$, and $k = 5$) is indicated by different line styles.

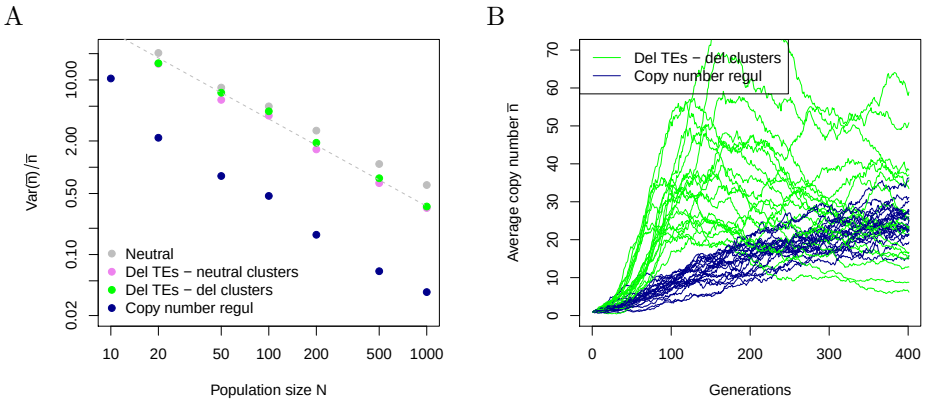

Figure 6: A: The effect of genetic drift is larger in the trap model than for copy-nuber regulation models. The figure displays the average copy number $\bar{n}$ in 20 independent replicates, $N = 100$ for both models. Parameters were $u = 0.1$, $s = 0.01$, $\pi = 0.03$, $k = 2$ for the trap model, and $u_n = 0.1/(1+0.3n)$, and $s = 0.01$, for the copy-number regulation model. Regulation strengh was set so that the expected equilibrium copy number $\hat{n} \simeq 25$ was the same for both models. B: 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 piRNA model (slope of -1 on the log-log plot).

4 Discussion

4.1 Population genetics of the trap model

The formalization of TE regulation by pi-RNA 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 efficiency of regulation 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 only if copies inserted in clusters are deleterious as well. When cluster copies 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. Pi-RNA regulation being a mutational process, the TE copy number is more stochastic and substiantially more sensitive to genetic drift than other regulation models.

These results confirm and formalizes 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 pi-clusters).

While the equilibrium for the neutral trap model can be expressed with a very simple formula (equation 3), the derivation of copy number and cluster frequencies is less straightforward when selection is accounted for (equations 10 and 11). In all cases, the TE copy number depends on the cluster size and distribution. The most effective configuration for TE regulation is a single, large pi-cluster. Dividing the cluster in smaller parts increases equilibrium TE copy numbers, and reducing the total cluster size as well. 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-monotonous 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 optimal rate in the trap model (about 0.1 to 0.2 transpositions per copy and per generation in unregulated genetic backgrounds, figure 5) seems convincingly 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 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 (plants, 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 pi-cluster regulation was also simplified. We considered that pi-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 cluster frequency at equilibrium but not its stability. In contrast, imperfect regulation (a residual transposition rate even when clusters 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 coeffcient 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.

Very recent data might suggest that pi-RNA regulation may not be sufficient to explain the early regulation of TE activity. For instance, Kofler et al. (2022) observed, in a lab experimental evolution context, that the transposition of the P element in Drosophila decreases before the first pi-cluster insertion appears in the population. 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 pi-RNA clusters and recombination

Kofler (2019) has already noticed that recombination among cluster loci reduces the efficiency of regulation. For a given proportion of the genome π occupied by piRNA clusters, regulation is 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). Indeed, when several clusters segregate, recombination decreases the heritability of the transposition regulation (recombination between clusters can generate permissive genotypes in the offspring of a cross between two transposition resistant individuals). In a similar way, regulation efficiency is expected to decrease with the within-cluster recombination rate (not modeled here). We confirmed here that the number of copies at equilibrium is indeed 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, pi-clusters seem to be located at several loci needs to be explained by other factors (such as functional constraints) than the regulation efficiency. The need to regulate independently different TE families might also play a role in the scattering of pi-clusters; the interactions between several TE families invading simultaneously may generate new constraints on the regulation system, which probably deserves further investigation.

An interesting hypothesis was raised by Kelleher et al. (2018) about the possibility that pi-cluster 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 regulatory clusters should sweep in the population as advantageous alleles. Our model, neglecting linkage disequilibrium between TEs and clusters, 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 pi-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 regulatory cluster alleles is of the order of magnitude of $n \times s$ (at least when few clusters are segregating), and selection may have an effect on cluster 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 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 artefacts 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 predictions from models is thus of utterly importance, even when models are necessarily naive and imperfect.

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 regulatory cluster TE 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 new kind of equilibrium was 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 regulatory clusters, which could be directly measured in populations to estimate the likelihood of the different regulation scenarios.

References

Adams, M. D., R. S. Tarng, and D. C. Rio (1997). The alternative splicing factor PSI regulates P-element third intron splicing in vivo. *Genes & Development* 11.1, pp. 129–138.

Arkhipova, I. and M. Meselson (2005). Deleterious transposable elements and the extinction of asexuals. *Bioessays* 27.1, pp. 76–85.

Brennecke, J., A. A. Aravin, A. Stark, M. Dus, M. Kellis, R. Sachidanandam, and G. J. Hannon (2007). Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* 128.6, pp. 1089–1103.

Brookfield, J. F. and R. M. Badge (1997). Population genetics models of transposable elements. *Genetica* 100.1, pp. 281–294.

Charlesworth, B. and D. Charlesworth (1983). The population dynamics of transposable elements. *Genetics Research* 42.1, pp. 1–27.

Deniz, Ö., J. M. Frost, and M. R. Branco (2019). Regulation of transposable elements by DNA modifications. *Nature Reviews Genetics* 20.7, pp. 417–431.

Dolgin, E. S. and B. Charlesworth (2006). The fate of transposable elements in asexual populations. *Genetics* 174.2, pp. 817–827.

Doolittle, W. F. and C. Sapienza (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284.5757, pp. 601–603.

Gilbert, C. and C. Feschotte (2018). Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. *Current Opinion in Genetics & Development* 49, pp. 15–24.

Gladyshev, E. (2017). Repeat-induced point mutation and other genome defense mechanisms in fungi. *Microbiology Spectrum* 5.4, pp. 5–4.

Goriaux, C., E. Théron, E. Brasset, and C. Vaury (2014). History of the discovery of a master locus producing piRNAs: the flamenco/COM locus in *Drosophila melanogaster*. *Frontiers in Genetics* 5, p. 257.

Harris, C. R., 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 (Sept. 2020). Array programming with NumPy. *Nature* 585.7825, pp. 357–362.

Hartl, D., E. Lozovskaya, and J. Lawrence (1992). Nonautonomous transposable elements in prokaryotes and eukaryotes. *Genetica* 86.1, pp. 47–53.

Kelleher, E. S., R. B. Azevedo, and Y. Zheng (2018). The evolution of small-RNA-mediated silencing of an invading transposable element. *Genome Biology and Evolution* 10.11, pp. 3038–3057.

Khan, A. I., D. M. Dinh, D. Schneider, R. E. Lenski, and T. F. Cooper (2011). Negative epistasis between beneficial mutations in an evolving bacterial population. *Science* 332.6034, pp. 1193–1196.

Kidwell, M. G. and D. R. Lisch (2001). Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution* 55.1, pp. 1–24.

Kofler, R. (2019). Dynamics of transposable element invasions with piRNA clusters. *Molecular Biology and Evolution* 36.7, pp. 1457–1472.

— (2020). piRNA clusters need a minimum size to control transposable element invasions. *Genome Biology and Evolution* 12.5, pp. 736–749.

Kofler, R., V. Nolte, and C. Schloetterer (2022). The transposition rate has little influence on equilibrium copy numbers of the P-element. *bioRxiv*.

Kouyos, R. D., O. K. Silander, and S. Bonhoeffer (2007). Epistasis between deleterious mutations and the evolution of recombination. *Trends in Ecology & Evolution* 22.6, pp. 308–315.

Le Rouzic, A. (2014). Estimating directional epistasis. *Frontiers in Genetics* 5, p. 198.

Le Rouzic, A. and P. Capy (2005). The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. *Genetics* 169.2, pp. 1033–1043.

— (2006). Population genetics models of competition between transposable element subfamilies. *Genetics* 174.2, pp. 785–793.

Lee, Y. C. G. (2022). Synergistic epistasis of the deleterious effects of transposable elements. *Genetics* 220.2, iyab211.

Lohe, A. R. and D. L. Hartl (1996). Autoregulation of mariner transposase activity by overproduction and dominant-negative complementation. *Molecular biology and evolution* 13.4, pp. 549–555.

Lu, J. and A. G. Clark (2010). Population dynamics of PIWI-interacting RNAs (piRNAs) and their targets in Drosophila. *Genome Research* 20.2, pp. 212–227.

Lynch, M. and J. S. Conery (2003). The origins of genome complexity. *Science* 302.5649, pp. 1401–1404.

Maisnier-Patin, S., J. R. Roth, Å. Fredriksson, T. Nyström, O. G. Berg, and D. I. Andersson (2005). Genomic buffering mitigates the effects of deleterious mutations in bacteria. *Nature Genetics* 37.12, pp. 1376–1379.

Malone, C. D. and G. J. Hannon (2009). Small RNAs as guardians of the genome. *Cell* 136.4, pp. 656–668.

Orgel, L. E. and F. H. Crick (1980). Selfish DNA: the ultimate parasite. *Nature* 284.5757, pp. 604–607.

Ozata, D. M., I. Gainetdinov, A. Zoch, D. O'Carroll, and P. D. Zamore (2019). PIWI-interacting RNAs: small RNAs with big functions. *Nature Reviews Genetics* 20.2, pp. 89–108.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Robillard, É., A. Le Rouzic, Z. Zhang, P. Capy, and A. Hua-Van (2016). Experimental evolution reveals hyperparasitic interactions among transposable elements. *Proceedings of the National Academy of Sciences* 113.51, pp. 14763–14768.

Roessler, K., A. Bousios, E. Meca, and B. S. Gaut (2018). Modeling interactions between transposable elements and the plant epigenetic response: a surprising reliance on element retention. *Genome biology and evolution* 10.3, pp. 803–815.

Saint-Leandre, B., P. Capy, A. Hua-Van, and J. Filée (2020). pi RNA and Transposon Dynamics in Drosophila: A Female Story. *Genome biology and evolution* 12.6, pp. 931–947.

Selker, E. U. and J. N. Stevens (1985). DNA methylation at asymmetric sites is associated with numerous transition mutations. *Proceedings of the National Academy of Sciences* 82.23, pp. 8114–8118.

Soetaert, K., T. Petzoldt, and R. W. Setzer (2010). Solving Differential Equations in R: Package deSolve. *Journal of Statistical Software* 33.9, pp. 1–25. ISSN: 1548-7660.

Sousa, A., C. Bourgard, L. M. Wahl, and I. Gordo (2013). Rates of transposition in Escherichia coli. *Biology letters* 9.6, p. 20130838.

Wallau, G. L., P. Capy, E. Loreto, A. Le Rouzic, and A. Hua-Van (2016). VHICA, a new method to discriminate between vertical and horizontal transposon transfer: Application to the mariner family within Drosophila. *Molecular Biology and Evolution* 33.4, pp. 1094–1109.

Zanni, V., A. Eymery, M. Coiffet, M. Zytnicki, I. Luyten, H. Quesneville, C. Vaury, and S. Jensen (2013). Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. *Proceedings of the National Academy of Sciences* 110.49, pp. 19842–19847.

Zhang, S., B. Pointer, and E. S. Kelleher (2020). Rapid evolution of piRNA-mediated silencing of an invading transposable element was driven by abundant de novo mutations. *Genome Research* 30.4, pp. 566–575.

Appendix A Mathematical details

A.1 Equation 6

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

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

A.2 Equation 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:

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

A.3 Equation 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 equation 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:

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

A.4 Equilibrium stability for equation 11

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

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

Eigenvalues are negative (i.e., the equilibrum 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.

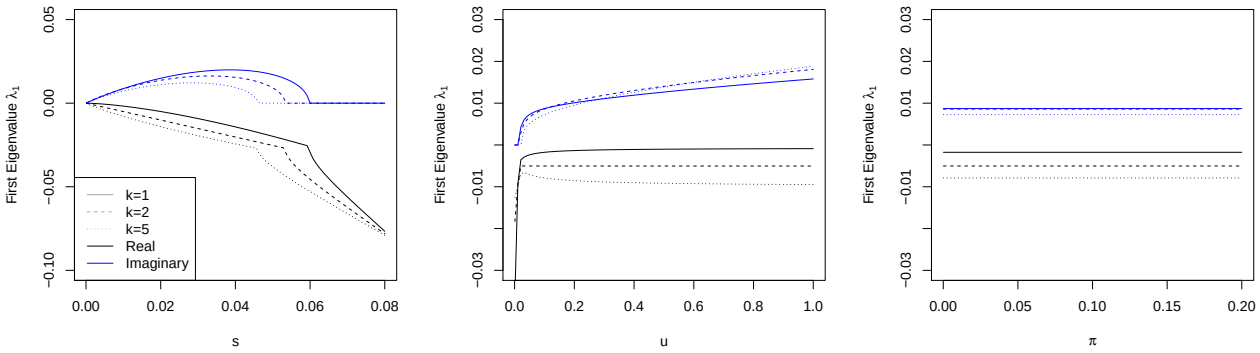

Figure A1: First Eigenvalue of the Jacobian matrix as a function of model parameters (u, s, and π) in the deleterious TEs - deleterious cluster model. Default parameter values were $u = 0.1$, $s = 0.01$, and $\pi = 0.03$. The number of clusters ($k = 1$, $k = 2$, and $k = 5$) is indicated by different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

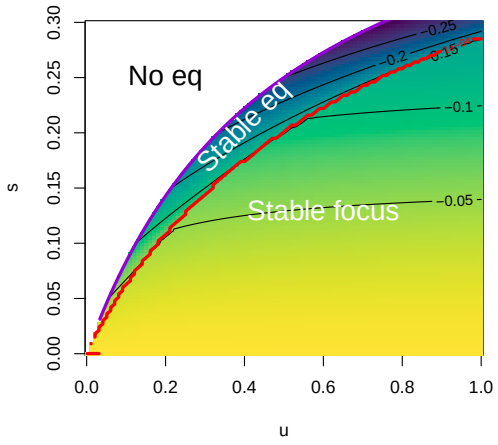

Figure A2: Equilibrium stability for the deleterious TES - deleterious cluster model. The figure represents the real part of the first eigenvalue of the Jacobian matrix for two major parameters (u and s), with $k = 1$ and $\pi = 0.03$. The eigenvalue is negative for the whole parameter range, and is 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).

Appendix B Supplementary results

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

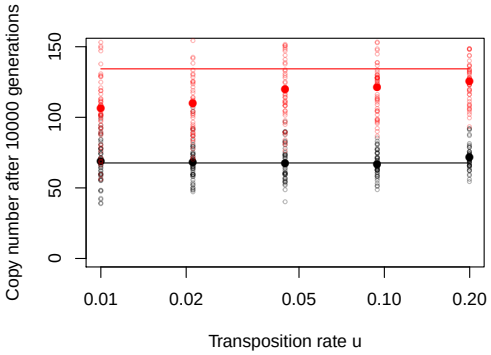

Figure B1: Effect of the transposition rate u on the simulated equilibrium copy number in the neutral model. Equation 4 predicts that, in the neutral model, the equilibrium does not depend on the transposition rate, which is at odds with previous TE regulation models, and does not seem to be observed in earlier numerical models (Kofler 2019). 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 the 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.

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

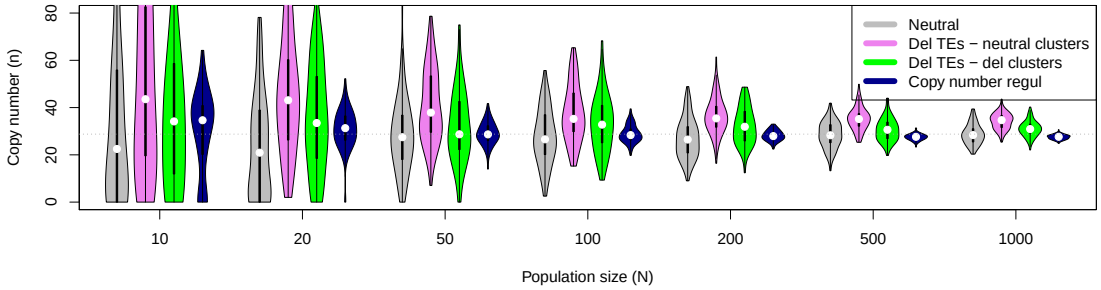

Figure B2: Distribution of the average copy number $\bar{n}$ among 1000 replicates in different models, with various population sizes. Models were parameterized so that they achieve very similar average copy numbers ($\bar{n} \sim 19$) in large populations (horizontal dotted line): $s = 0.01$ for all 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.