Population genetics of transposable element load: a mechanistic account of observed overdispersion

Ronald D. Smith[1,‡], Joshua R. Puzey[2,§,⋆] and Gregory D. Conradi Smith[1,†,⋆]

[1]Department of Applied Science and

[2]Department of Biology, William & Mary, Williamsburg, VA 23187

[⋆]These authors contributed equally

[†]Corresponding author: greg@wm.edu ORCID: 0000-0002-1054-6790

[§]ORCID: 0000-0001-8019-9993

[‡]ORCID: 0000-0003-2736-0092

Draft manuscript compiled on November 14, 2020

Abstract

In an empirical analysis of transposable element (TE) abundance within natural populations of *Mimulus guttatus* and *Drosophila melanogaster*, we found a surprisingly high variance of TE count (e.g., variance-to-mean ratio on the order of 10 to 100). To obtain insight regarding those evolutionary genetic mechanisms that are may underlie the overdispersed population distributions of TE abundance, we developed a mathematical model of TE population genetics that includes the dynamics of element proliferation and purifying selection on TE load. The modeling approach begins with a master equation for a birth-death process and it extends the predictions of the classical theory of TE dynamics in several ways. In particular, moment-based analysis of stationary population distributions of TE load reveal that overdispersion is most likely to arise via copy-and-paste (as opposed to cut-and-paste) dynamics. Parameter studies suggest that overdispersed population distributions of TE abundance are probably not a consequence of purifying selection on total element load.

1 INTRODUCTION

The genomics revolution has revealed that a significant portion of eukaryotic genomes are comprised of transposable elements (TEs, also called mobile DNA elements or transposons). Notable examples include the human and maize genomes, 44% and 85% of which are TE sequences (Mills et al., 2007; Springer et al., 2009). TEs are capable of moving throughout a genome via copy-and-paste or cut-and-paste mechanisms. Their effect can range from having little to no consequence on phenotype to being powerful mutagens (Bourque et al., 2018). In addition to the innate tendency of TEs to proliferate, factors such as recombination, epigenetics, and selection contribute to the complex genomic distribution of these elements (Kent et al., 2017). While it is clear that TEs have been an integral part of the long-term evolution of genome architecture, much about the role of TEs in evolution remains unknown. Knowledge of the dynamics of TE abundance in natural populations is an important step toward an increasing understanding of how genomes evolve.

The seminal and most commonly cited population genetic theory of TEs was developed in 1983 via a combination of mathematical analysis, computer simulation and a limited amount of experimental data (Charlesworth and Charlesworth, 1983). This modeling considered a single class of TEs with a drift-diffusion representation of TE proliferation, with either no selection or weak selection acting on total TE copy number. While such models have informed our understanding of the population genetics of TEs for several decades, the classical theory does not reproduce experimentally observed within-population variances that often greatly exceed the population mean. The cause of this discrepancy is that the classical model assumes a binomial distribution of within-population TE loads, which constrains the population variance to be no greater than the population mean.

This paper begins with a brief review of classical TE population genetics. This is followed by an examination of genome-sequence data from two natural populations (*Mimulus guttatus* and *Drosophila melanogaster*). Notably, in both cases, we observe that the within-population variance of TE load is highly overdispersed. Because these empirical results violate the expectation of classical TE modeling, we developed a master equation formulation of the population distribution of TE loads in a large randomly mating population. This master equation formulation extends the predictions of the classical theory of TE dynamics in several ways. In particular, our calculations and moment-based analysis of stationary distributions of TE load reveal that overdispersion is most likely to arise via copy-and-paste (as opposed to cut-and-paste) dynamics. Parameter studies further suggest that overdispersed population distributions of TE abundance are probably not a consequence of purifying selection on total element load.

1.1 Classical population genetics of TEs

A good starting point for discussing the evolutionary dynamics of TEs is the seminal paper by Charlesworth and Charlesworth (1983) and subsequent work (Brookfield and Badge, 1997; Charlesworth and Charlesworth, 2010; Deceliere, 2004; Le Rouzic and Deceliere, 2005). This classical theory represents a chromosome as a finite set of m available insertion sites (loci) per haploid genome, each of which can either be occupied by a transposable element (or not). For a single family of TEs, the state of an infinite diploid population at a given chromosomal site i, for $i = 1, 2, \ldots, m$, is described by its frequency, x_i, where $0 \leq x_i \leq 1$. Assuming insertion sites exhibit no linkage disequilibrium, the set of frequencies, $\{x_i\}_{i=1}^m$, describes the state of the population. The mean copy number of TEs per individual is $\bar{n} = 2\sum_{i=1}^m x_i$, where the factor of 2 accounts for diploidy.

The evolutionarily neutral version of the classical theory includes two processes affecting TE load (gain and loss). Gain of TEs is represented by a proliferation rate (per individual per element per generation) in the germ line of an individual with n elements. This proliferation rate, denoted u_n, is typically assumed to be a decreasing function of TE load ($du_n/dn < 0$). Loss of TEs is represented by a first-order excision rate constant (per individual per element per generation) denoted by ν. The change (per generation) in the mean TE copy number per individual is thus

$$\Delta\bar{n} = \mathsf{E}[\mathbf{n}u_{\mathbf{n}}] - \nu\bar{n}\,, \tag{1}$$

where $\mathbf{n}$ is the diploid TE load of a randomly sampled individual, the expected value is taken over individuals in the population, and $\bar{n} = \mathsf{E}[\mathbf{n}]$ is the population mean of TE copy number. Expanding Eq. 1 around the mean TE load gives the following second-order approximation,

$$\Delta\bar{n} \approx \bar{n}(u_{\bar{n}} - \nu) + \frac{V_n}{2}\left(2\frac{du_{\bar{n}}}{d\bar{n}} + \bar{n}\frac{d^2u_{\bar{n}}}{d\bar{n}^2}\right)\,, \tag{2}$$

where V_n denotes the population variance in TE copy number (Charlesworth and Charlesworth, 1983). If the higher order terms that scale the population variance are negligible, the change in mean TE copy number per generation is $\Delta\bar{n} \approx \bar{n}(u_{\bar{n}} - \nu)$. For this neutral model of TE population dynamics, one concludes that $\bar{n}$ will approach an (stable) equilibrium value satisfying $u_{\bar{n}} \approx \nu$ provided $du_{\bar{n}}/d\bar{n} < 0$.

To extend this model of TE population genetics to include the effect of natural selection, it is customary to assume an individual viability function, w_n, that is a decreasing function of total genome-wide TE load ($dw_n/dn < 0$) . Approximating the mean fitness of the population ($\mathsf{E}[w_{\mathbf{n}}]$) by the fitness of an individual with an average number of copies ($w_{\bar{n}}$), Eq. 2 can be extended to include the effect of selection on TE load (Charlesworth and Charlesworth, 2010),

$$\Delta\bar{n} \approx V_n\frac{d\ln w_{\bar{n}}}{d\bar{n}} + \bar{n}(u_{\bar{n}} - \nu) + \frac{V_n}{2}\left(2\frac{du_{\bar{n}}}{d\bar{n}} + \bar{n}\frac{d^2u_{\bar{n}}}{d\bar{n}^2}\right)\,. \tag{3}$$

As a specific example, consider the proliferation rate function $u_n = \xi_0/n$ with $\xi_0 > 0$ and the selection function $w_n = e^{-\gamma n}$ for $\gamma > 0$ (viability is a decreasing function of TE copy number). Because $du_n/dn = -\xi_0/n^2$ and $d^2u_n/dn^2 = 2\xi_0/n^3$, the higher order terms involving derivatives of u_n evaluate to zero. Consequently, Eq. 3 becomes

$$\Delta\bar{n} \approx V_n \frac{d\ln w_{\bar{n}}}{d\bar{n}} + \xi_0 - \nu\bar{n}\,.$$

Substituting $d\ln w_{\bar{n}}/d\bar{n} = -\gamma$ and setting $\Delta\bar{n} = 0$ gives $0 = -\gamma V_n + \xi_0 - \nu\bar{n}$. Solving for the equilibrium mean TE load gives,

$$\bar{n} = \frac{\xi_0 - \gamma V_n}{\nu}\,. \tag{4}$$

This result is biologically meaningful for $\xi_0 > \gamma V_n$. As expected, the equilibrium TE load is an increasing function of the proliferation rate constant, ξ_0, and a decreasing function of the excision rate constant, ν. Furthermore, stronger selection against TE load (greater γ) decreases the mean value of the equilibrium TE load in the population.

1.2 Population variance in the classical model

Analysis of the classical model of TE population genetics often proceeds by making further assumptions regarding the population variance V_n. For example, Charlesworth and Charlesworth (1983) assume the population variance takes the form

$$V_n = \bar{n}\left(1 - \frac{\bar{n}}{2m}\right) - 2m\sigma_x^2 + 4\sum_{i<j} D_{ij}\,, \tag{5}$$

where D is a matrix of linkage disequilibrium coefficients (Bulmer, 1980), and $\sigma_x^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})^2$ is the variance in element frequency across loci (see Supplemental Material, Section S1). If one further assumes that linkage effects are small enough to be ignored, then

$$V_n \approx \bar{n}\left(1 - \frac{\bar{n}}{2m}\right) - 2m\sigma_x^2\,. \tag{6}$$

Charlesworth and Charlesworth (1983) argue that for a large enough population, one expects the variance in element frequency across loci to be eventually become negligable, $\sigma_x^2 \to 0$ and, consequently, the equilibrium population variance of TE load should approach that of a binomial distribution,

$$V_n \approx \bar{n}\left(1 - \frac{\bar{n}}{2m}\right)\,. \tag{7}$$

In that case, assuming occupiable loci are not limiting ($\bar{n} << 2m$), the population variance will be well-approximately by the mean ($V_n \approx \bar{n}$). Substituting this value into Eq. 4, the classical model indicates that the equilibrium TE load will be

$$\bar{n} = \frac{\xi_0}{\gamma + \nu}\,. \tag{8}$$

As in Eq. 4, the equilibrium TE load is an increasing function of the copy-and-paste rate (ξ_0), and a decreasing function of both the excision rate constant (ν) and the strength of selection against TE load (γ).

The classical model (Eqs. 3–8) has informed expectations regarding the population genetics of TEs for several decades. For example, an extension of this classical theory predicts that in a finite population of effective size N_e, the the stationary distribution of TE frequency (x) will take the form $\rho(x) \propto x^{a-1}(1-x)^{b-1}$ where $a = 4N_e\bar{n}u_{\bar{n}}/(2m-\bar{n})$ and $b = 4N_e(\nu + |d \ln w_{\bar{n}}/d\bar{n}|)$ (Le Rouzic and Deceliere, 2005). For $u_n = \xi_0/n$, $w_n = e^{-\gamma n}$, and $\bar{n} << 2m$, this gives $a = 4N_e\xi_0$ and $b = 4N_e(\nu+\gamma)$. On the other hand, the classical approach to modeling TE population genetics has obvious limitations. For one thing, the derivation and analysis of the classical model makes assumptions about the population variance that may not be consistent with experimental observations (see Results). Furthermore, the population variance of TE load ought to be an emergent property of the model used to understand the population genetics of TEs, rather than a modeling assumption imposed upon a preexisting framework (Eq. 7).

The remainder of this paper summarizes recent work that addresses these two issues in detail. We begin by presenting empirical evidence that population variance of TEs is neither binomial nor well-approximated by the mean. This motivates the presentation of an alternative population genetic framework that may be used to predict both the population variance as well as the mean TE load. This model of TE population genetics is then interrogated in order to elucidate those evolutionary genetic mechanisms that have the greatest influence on the population variance of TE load.

2 RESULTS

2.1 Dispersion of TE loads in the classical model

In the classical modeling of TE population genetics discussed above, analytical results are obtained by assuming a randomly mating population with a binomial distribution of TE loads,

$$\mathbf{n} \sim \mathrm{Binomial}(2m, \bar{n}/2m)\,, \tag{9}$$

with mean $\mathsf{E}[\mathbf{n}] = \bar{n}$ and variance $\mathsf{Var}[\mathbf{n}] = \bar{n}(1-\bar{n}/2m)$ (Eqs. 5–7). A simple measure of the variability of TE load within a population is the *index of dispersion* (Fano factor) given by

$$\mathsf{Fano}[\mathbf{n}] = \frac{\mathsf{Var}[\mathbf{n}]}{\mathsf{E}[\mathbf{n}]}\,. \tag{10}$$

Substituting the mean and variance of the binomial distribution into Eq. 10, it is apparent that the classical model of TE population genetics predicts (i.e., assumes) a Fano factor that is less than one,

$$\mathsf{Fano}[\mathbf{n}] = 1 - \frac{\bar{n}}{2m} < 1\,. \tag{11}$$

In fact, when the number of sites occupied by TEs is small compared to the total number of occupiable loci ($m \to \infty$ with $\bar{n}$ fixed), the Fano factor approaches one from below ($\mathsf{Fano}[\mathbf{n}] \to 1$). In this limit, the binomial distribution of Eq. 9 is well-approximated by $\mathbf{n} \sim \text{Poisson}(\bar{n})$. If it were the case that the TE load within a population were Poisson distributed, then the mean and variance of TE load would be equal ($\mathsf{E}[\mathbf{n}] = \mathsf{Var}[\mathbf{n}] = \bar{n}$) and the index of dispersion would be $\mathsf{Fano}[\mathbf{n}] = 1$. With our expectations set by the classical model of TE population genetics discussed above, empirical observations of a Fano factor greater than one ($\mathsf{Fano}[\mathbf{n}] > 1$) would indicate *overdispersion* of TE load within a population.

2.2 Overdispersion of empirical TE counts

Fig. 1 presents analysis of two data sets, both of which indicate that the variance of TE load in experimentally studied populations is far greater than would be predicted by classical models of TE population genetics. The first data set consists of whole-genome sequence data from 164 lines of *Mimulus guttatus* derived from a naturally occurring population in Iron Mountain, Oregon, USA (estimated population size is about 300,000). The second data set comes from an analysis of 131 lines of *Drosophila melanogaster* from the Drosophila Genetic Reference Panel (DGRP) (Cridland et al., 2013). Their analysis identified over 17,000 TE insertions across individual lines that were derived from a large population in Raleigh, NC, USA. See Section **??** of the Supplementary Material for a description of experimental methods and data analysis.

Comparison of the marker locations in Fig. 1A with the dashed line (labelled Poisson) shows that in both species, *Mimulus guttatus* and *Drosophila melanogaster*, the population distribution of TE load is *overdispersed* (the variance of TE load is greater than the mean TE load). In *D. melanogaster*, this overdispersion is greater for so called cut-and-paste TEs with a DNA intermediate as opposed to copy-and-paste TEs with an RNA intermediate (compare open triangle to open circle). The corresponding Fano factors (TE load variance relative to mean, as defined in Eq. 10) are 16 and 2.7, respectively, values that indicate overdispersion (see Table 1). Overdispersion of TE load is even more pronounced in *M. guttatus*. In this case, the Fano factors are 61 for copy-and-paste elements LINE and LTR (filled red symbols), and 646 for cut-and-paste elements including DNA and Helitron (filled blue symbols).

Fig. 1B (left) shows the estimated number of copy-and-paste and cut-and-paste TEs in each of the 164 lines of *M. guttatus* (horizontal bar graph). In both cases, the variance (illustrated by the width of red and blue histograms) is far greater than the variance that would be consistent with classical model of TE population genetics (gray curves). Fig. 1B (right) shows the corresponding analysis for the 131 lines of *D. melanogaster*. From these analyses we conclude that in both species, *Mimulus guttatus* and *Drosophila melanogaster*, and for both classes of TEs, copy-and-paste and cut-and-paste, the distribution of TE load within the studied population is highly *overdispersed*.

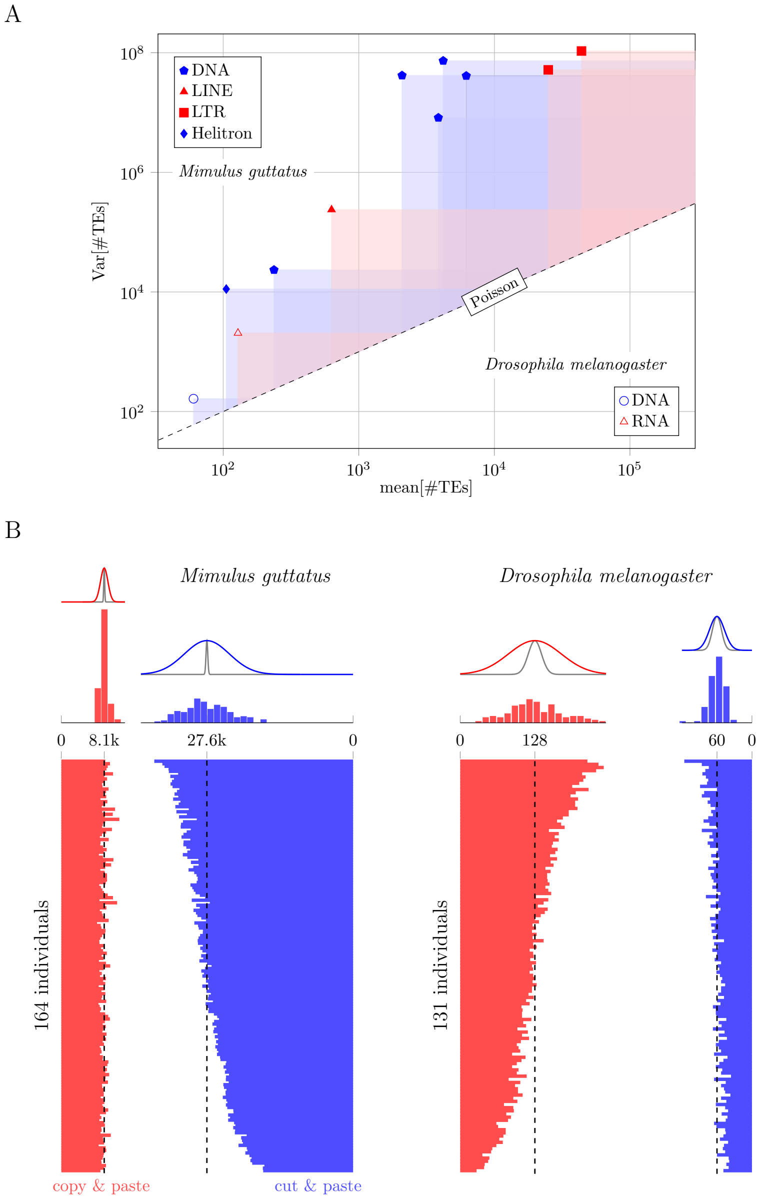


Figure 1: A: Mean-variance plot of TE copy number in *M. guttatus* and *D. melanogaster* populations compared to theoretical expectation (Poisson line). B: Estimated TE copy number for 164 *M. guttatus* individuals (left panels) and 131 *D. melanogaster* individuals (right). TE counts are separated by class (red, copy-and-paste; blue, cut-and-paste). The variability in TE load can be observed in the counts from individuals (bottom) as well as histograms (top). The overdispersion in TE load is apparent in the deviation of the observed counts (red and blue histograms) from the corresponding Poisson distributions (gray lines). The black dotted lines show the population mean of TE load.

Species	TE class	$\mathsf{E}[\mathbf{n}]$	$\mathsf{Var}[\mathbf{n}]$	$\mathsf{Fano}[\mathbf{n}]$
M. guttatus	copy-and-paste	8,082	4.9×10^5	61
	cut-and-paste	27,559	1.8×10^7	646
D. melanogaster	copy-and-paste	128	2,053	16
	cut-and-paste	60	164	2.7

Table 1: Empirically observed mean, variance, and index of dispersion (Fano factor) of the population distribution of TE load in 164 *M. guttatus* and 131 *D. melanogaster* individuals (cf. Fig. 1).

2.3 Overdispersion is not explained by distinct TE families

The overdispersion documented in Fig. 1 cannot be explained away as a consequence of the heterogeneity of the properties of distinct TE types. That is, if the population variances of two different families of TEs follow the classical model (i.e., $\mathbf{n}_1$ and $\mathbf{n}_2$ are binomially distributed according to Eq. 9), then $\mathsf{E}[\mathbf{n}_i] = \bar{n}_i$, $\mathsf{Var}[\mathbf{n}_i] = \bar{n}_i(1-\bar{n}_i/2m)$, and $\mathsf{Fano}[\mathbf{n}_i] \leq 1$ for $i = 1, 2$. If these families of TEs were independently distributed in the population, but not distinguished, the composite mean, $\mathsf{E}[\mathbf{n}] = \mathsf{E}[\mathbf{n}_1] + \mathsf{E}[\mathbf{n}_2] = \bar{n}_1 + \bar{n}_2$, and variance, $\mathsf{Var}[\mathbf{n}] = \mathsf{Var}[\mathbf{n}_1] + \mathsf{Var}[\mathbf{n}_2] = \bar{n}_1(1-\bar{n}_1/2m) + \bar{n}_2(1-\bar{n}_2/2m)$, would yield

$$\mathsf{Fano}[\mathbf{n}] = \frac{\bar{n}_1(1-\bar{n}_1/2m) + \bar{n}_2(1-\bar{n}_2/2m)}{\bar{n}_1 + \bar{n}_2} \leq 1\,.$$

This composite Fano factor is less than one, indicating that the presence of different varieties of TEs does not explain observed overdispersion in the classical model.

In fact, a stronger statement can be made; one that does not depend on the variance of each TE family being underdispersed ($\mathsf{Fano}[\mathbf{n}_i] \leq 1$). Consider two families of TEs with mean TE loads $\bar{n}_1$ and $\bar{n}_2$ and Fano factors F_1 and F_2. In that case, the variances of TE load are $F_1\bar{n}_1$ and $F_2\bar{n}_2$, respectively. If these two families were not distinguished, the observed composite mean load, $\bar{n}_1 + \bar{n}_2$, and variance, $F_1\bar{n}_1 + F_2\bar{n}_2$, yield the following index of dispersion,

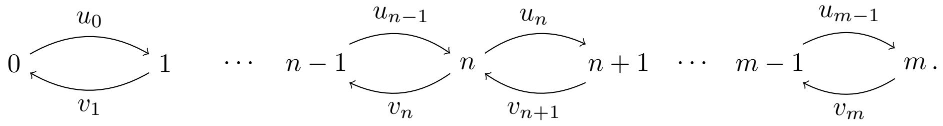
$$F = \frac{F_1\bar{n}_1 + F_2\bar{n}_2}{\bar{n}_1 + \bar{n}_2}\,.$$

Because F is a weighted average of Fano factors for each family, the composite overdispersion is bounded by $\min(F_1, F_2) \leq F \leq \max(F_1, F_2)$. The dispersion of TE load that results when families are not distinguished is always *less* than the overdispersion of at least one of the TE families.

2.4 Master equation for TE population dynamics

Our modeling aims to clarify the observed overdispersion of TE load in *Mimulus guttatus* and *Drosophila melanogaster*, following classical TE population genetics, but with a few important modifications. Because the variance in TE load is not the result of heterogeneity in TE types (see above), our analysis will focus on a single TE family.

Let $p_n(t)$ denote the probability that a randomly sampled haploid genome (gamete) has a TE count of n at time t. Prior to considerations of selection, our neutral model of TE population dynamics will take the form of a skip-free birth-death process with gain and loss rates denoted u_n and v_n. The state space for haploid TE load is $n \in \{0, 1, 2, \ldots, m\}$ and the state-transition diagram of the stochastic process is



The master equation for this stochastic process is the following system of $m+1$ differential equations,

$$\frac{dp_0}{dt} = -u_0 p_0 + v_1 p_1 \tag{12}$$

$$\frac{dp_n}{dt} = -(u_n + v_n)p_n + u_{n-1}p_{n-1} + v_{n+1}p_{n+1} \qquad 1 \le n \le m-1 \tag{13}$$

$$\frac{dp_m}{dt} = -v_m p_m + u_{m-1}p_{m-1}\,. \tag{14}$$

The expected value of TE load of a randomly sampled diploid genotype is

$$\bar{n} = \mathsf{E}[\mathbf{n}] = 2\sum_{n=0}^{m} np_n = 2\mu_1\,. \tag{15}$$

where $\mu_1 = \sum_{n=0}^{m} np_n$ is the mean TE load of a randomly sampled haploid gamete. By differentiating Eq. 15 to obtain

$$\frac{d\bar{n}}{dt} = 2\sum_{n=0}^{m} n\frac{dp_n}{dt} \tag{16}$$

and substituting Eqs. 12–14, the master equation formulation is found to be consistent with the classical model (Section S2). For example, if we assume that TE excision occurs with first order rate constant

$$v_n = \nu n\,, \tag{17}$$

one may derive from Eqs. 12–16 the following differential equation for the mean diploid TE load,

$$\frac{d\bar{n}}{dt} = \mathsf{E}[\mathbf{n}(u_{\mathbf{n}} - v_{\mathbf{n}})] = \mathsf{E}[\mathbf{n}u_{\mathbf{n}}] - \nu\bar{n}\,, \tag{18}$$

which is a continuous-time version of Eq. 1.

Limit	$\mathsf{E}[\mathbf{n}] = \bar{n}$	$\mathsf{Var}[\mathbf{n}] = \sigma_n^2$	$\mathsf{Fano}[\mathbf{n}] = \mathsf{Var}[\mathbf{n}]/\mathsf{E}[\mathbf{n}]$
$\eta = 0$	$\dfrac{2m\,\eta_0/\nu}{m+\eta_0/\nu}$	$\dfrac{2m^2\,\eta_0/\nu}{(m+\eta_0/\nu)^2}$	$\dfrac{m}{m+\eta_0/\nu}$
$\nu > \eta,\ m \to \infty$	$\dfrac{2\eta_0}{\nu-\eta}$	$\dfrac{2\eta_0\nu}{(\nu-\eta)^2}$	$\dfrac{\nu}{\nu-\eta}$

Table 2: The evolutionarily neutral moment equations (Eqs. 23 and 24) for the mean and variance of TE load make predictions in various limits (see Sections S2.2 and S2.3).

2.5 Master equation predicts the variance of TE load

One feature of the master equation formulation (Eqs. 12–14) is that the dynamics of the population variance of TE load are an emergent property of the model. To illustrate, let us assume that the insertion rate for a single family of TEs is

$$u_n = (\eta_0 + \eta n)(1 - n/m)\,, \tag{19}$$

where η is the copy-and-paste rate per transposon (a first-order rate constant), η_0 is the rate at which transposons are arriving from other sources (a zeroth order rate constant), n is the TE copy number, and m is the number of occupiable loci (in a haploid gamete). Substituting this constitutive relation for u_n, as well as $v_n = \nu n$, into Eqs. 12–14 gives

$$\begin{aligned}
\frac{dp_0}{dt} &= -\eta_0 p_0 + \nu p_1 && (20)\\
\frac{dp_n}{dt} &= -\left[(\eta_0+\eta n)(1-n/m) + \nu n\right] p_n \\
&+ \left[\eta_0 + \eta(n-1)\right]\left[1-(n-1)/m\right] p_{n-1} + \nu(n+1)p_{n+1} \qquad 1 \le n \le m-1 && (21)\\
\frac{dp_m}{dt} &= -\nu m p_m + \left[\eta_0 + \eta(m-1)\right]\left[1-(m-1)/m\right] p_{m-1}\,. && (22)
\end{aligned}$$

Fig. 2 shows representative numerical solutions of this master equation for the population dynamics of TE load. When the copy-and-paste rate constant is zero ($\eta = 0$) and occupiable loci are not limiting ($\bar{n} << 2m$), the stationary probability distribution is well-approximated by a Poisson distribution with $\mathsf{Var}[\mathbf{n}] \approx \bar{n}$ and $\mathsf{Fano}[\mathbf{n}] \approx 1$ (blue histograms). For both *Mimulus*- and *Drosophila*-like parameters, no overdispersion is observed provided that the copy-and-paste rate constant is zero ($\eta = 0$). These results should be compared to the green and red histograms, for which the copy-and-pate rate is nonzero (see caption for parameters). Notably, an increase in the copy-and-paste rate leads to significant overdispersion of the TE load for both simulated populations ($\mathsf{Fano}[\mathbf{n}]$ ranging from 7 to 100).

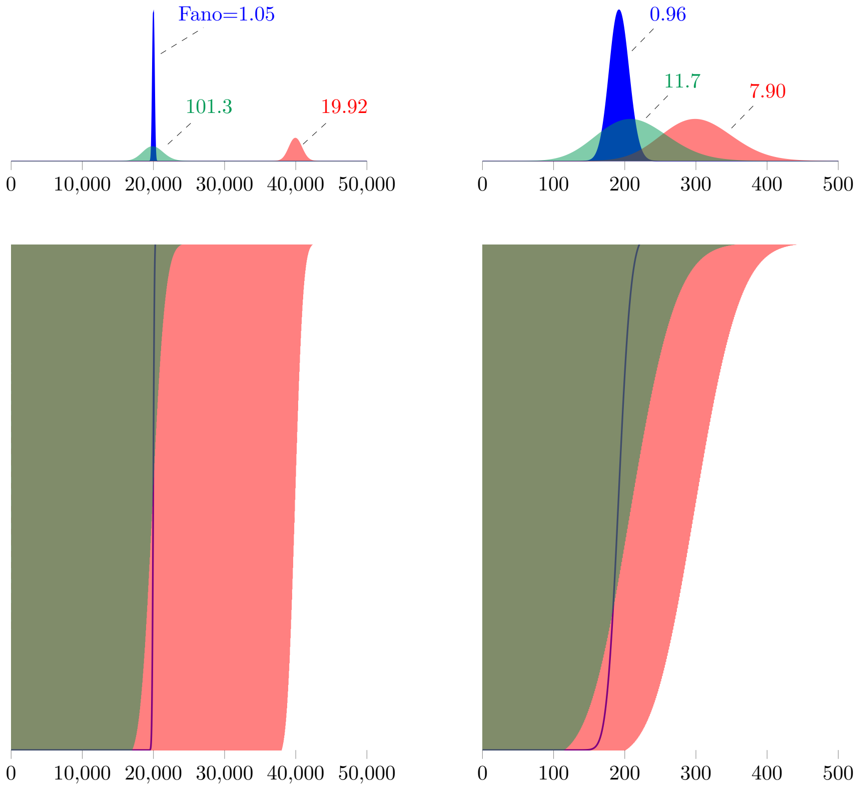


Figure 2: Top: Stationary population distributions of TE load in haploid genomes (gametes) calculated using the evolutionarily neutral master equation model (Eqs. 20–22). For mean loads similar to *Mimulus* (left) and *Drosophila* (right), no overdispersion is observed in simulations absent copy-and-paste transposition ($\eta = 0$, $\mathsf{Fano}[\mathbf{n}] \approx 1$). Green and red histograms show overdispersed population distributions of TE load that are obtained when copy-and-paste transposition is included. *Mimulus* parameters: $\nu = 0.1$, $m = 10^9$; η_0, η = 2000, 0 (blue), 200, 0.095 (red), 20, 0.099 (green). *Drosophila* parameters: $\nu = 0.1$, $m = 5000$; η_0, η = 20, 0 (blue), 1, 0.1 (green), 2, 0.1 (red). See Section S5 For details of numerical methods.

2.6 Moment equations for mean and variance of TE load

The previous section showed that the evolutionarily neutral master equation model provides information about the population variance of TE load that is unavailable in classical theory. Because this realism comes at the expense of a more complex model formulation (Eqs. 20–22 compared to Eq. 2), we derived ordinary differential equations (ODEs) that summarize the dynamics of the mean and variance of the population distribution of diploid TE loads predicted by the master equation. Section S2 of the Supplementary Material shows that the mean and variance of TE load solve the following ODEs,

$$\begin{aligned} \frac{d\bar{n}}{dt} &= 2\eta_0 - \left(\nu - \eta + \frac{\eta_0}{m}\right)\bar{n} - \frac{\eta}{m}\left(\sigma_n^2 + \frac{\bar{n}^2}{2}\right) && (23) \\ \frac{d\sigma_n^2}{dt} &= 2\eta_0 + \left(\nu + \eta - \frac{\eta_0}{m}\right)\bar{n} - 2\left(\nu - \eta + \frac{\eta_0 + \eta/2}{m}\right)\sigma_n^2 \\ &- \frac{2\eta}{m}\left(\bar{n}\sigma_n^2 + \frac{\bar{n}^2}{4} + \mathsf{E}[(\mathbf{n} - \bar{n})^3]\right). && (24) \end{aligned}$$

The term $\mathsf{E}[(\mathbf{n} - \bar{n})^3]$ that appears in Eq. 24 is the third central moment of the within-population diploid TE load. Analysis of this system of ODEs and the third central moment is provided below (Section 2.9).

If number of occupiable loci are not limiting ($\bar{n} << 2m$), we may take the limit of Eqs. 23 and 24 as $m \to \infty$ to obtain simpler equations for the mean and variance,

$$\begin{aligned} \frac{d\bar{n}}{dt} &= 2\eta_0 - (\nu - \eta)\bar{n} && (25) \\ \frac{d\sigma_n^2}{dt} &= 2\eta_0 + (\nu + \eta)\bar{n} - 2(\nu - \eta)\sigma_n^2 \,. && (26) \end{aligned}$$

This reduced system of ODEs is linear and in this limit the equation for the variance (Eq. 26) does not depend on the third central moment. The steady-state solution of Eqs. 25 and 26 given by

$$\begin{aligned} \bar{n} &= \frac{2\eta_0}{\nu - \eta} && (27) \\ \sigma_n^2 &= \frac{2\eta_0\nu}{(\nu - \eta)^2} = \frac{\nu\bar{n}}{\nu - \eta} && (28) \end{aligned}$$

is physical provided $\nu > \eta$, that is, when m is large, the rate of excision ν must be greater than the copy-and-past rate constant η for positive mean TE load ($\bar{n} > 0$). This physical steady state is stable because the Jacobian of Eqs. 25 and 26, given by the 2×2 matrix with entries $J_{11} = -(\nu - \eta)$, $J_{12} = 0$, $J_{21} = \nu + \eta$, $J_{22} = -2(\nu - \eta)$, has real valued eigenvalues $\lambda = -(\nu - \eta) < 0$ and $2\lambda < 0$.

The values for the steady-state mean and variance of TE load given by Eqs. 27–28 correspond to the following index of dispersion,

$$\mathsf{Fano}[\mathbf{n}] = \frac{\sigma_n^2}{\bar{n}} = \frac{\nu}{\nu - \eta} \,. \qquad (29)$$

Notably, that the condition for a stable steady state ($\nu > \eta$) implies an index of dispersion greater than unity ($\mathsf{Fano}[\mathbf{n}] > 1$) for any nonzero copy-and-paste rate constant ($\eta > 0$). For this reason, we conclude that *a steady state within-population distribution of TE loads will be overdispersed whenever the number of occupiable loci are not limiting* ($\bar{n} << 2m$). Further analysis of the moment equations (Eqs. 23 and 24) shows that overdispersion will not occur in the absence of copy-and-paste dynamics (see $\eta = 0$ case in Table 2).

This preliminary analysis of an evolutionarily neutral master equation for TE proliferation (Eqs. 20–22) indicates that *a nonzero copy-and-paste rate may lead to overdispersed population distributions of*

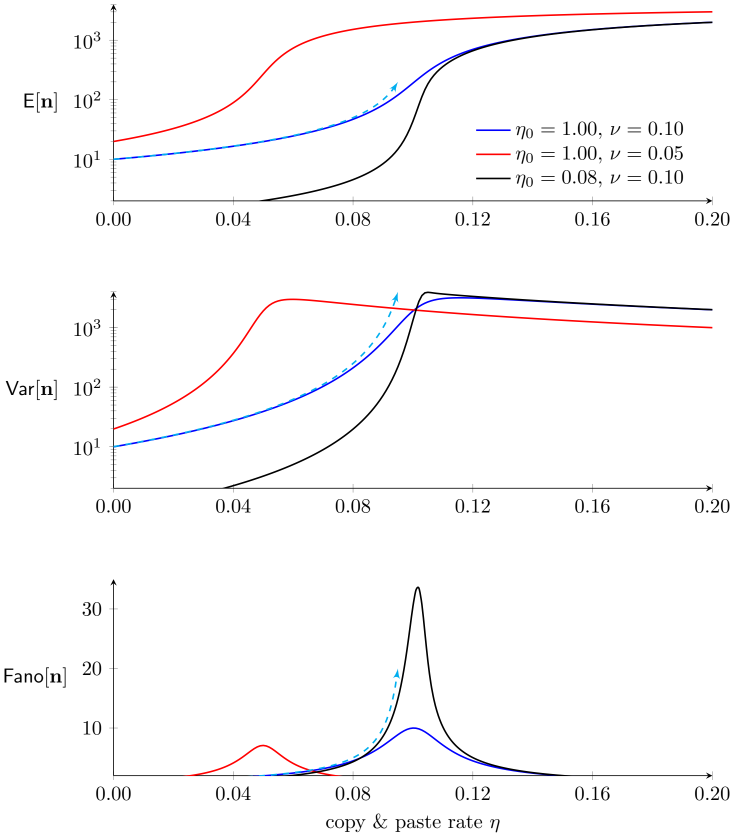


Figure 3: Parameter studies of the neutral master equation model showing the mean ($\bar{n}$), variance (σ_n^2), and index of dispersion (Fano[**n**]) of within-population TE load as a function of the copy-and-paste rate constant (η). Other parameters: $m = 4 \times 10^3$ and as in legend. Cyan curves indicate analytical approximations in the limit as $m \to \infty$ (see Table 2). These calculations were accelerated using a Fokker-Planck approximation to Eqs. 20–22 (see Section S5).

TE load (Eq. 29). That is, copy-and-paste TE dynamics is one possible explanation for our empirical observations of overdispersed TE counts (Fig. 1). Furthermore, this analysis predicts that a large index of dispersion may be a consequence of balanced dynamics of TE gain and loss (i.e., Fano[**n**] $\to \infty$ as ν decreases to η in Eq. 29). While the divergence in the analytical result is an artifact of taking the $m \to \infty$ limit, the parameter study shown in Fig. 3 confirms that blowup of Fano[**n**] occurs in master equation simulations when m is large and the dynamics of TE gain and loss are balanced ($\eta \approx \nu$).

2.7 Influence of selection on overdispersion

To investigate the effect of purifying selection on the population variance of TE load, we assume a selection coefficient (w_n) that depends on total diploid TE load (n) with $dw_n/dn < 0$ (higher load is

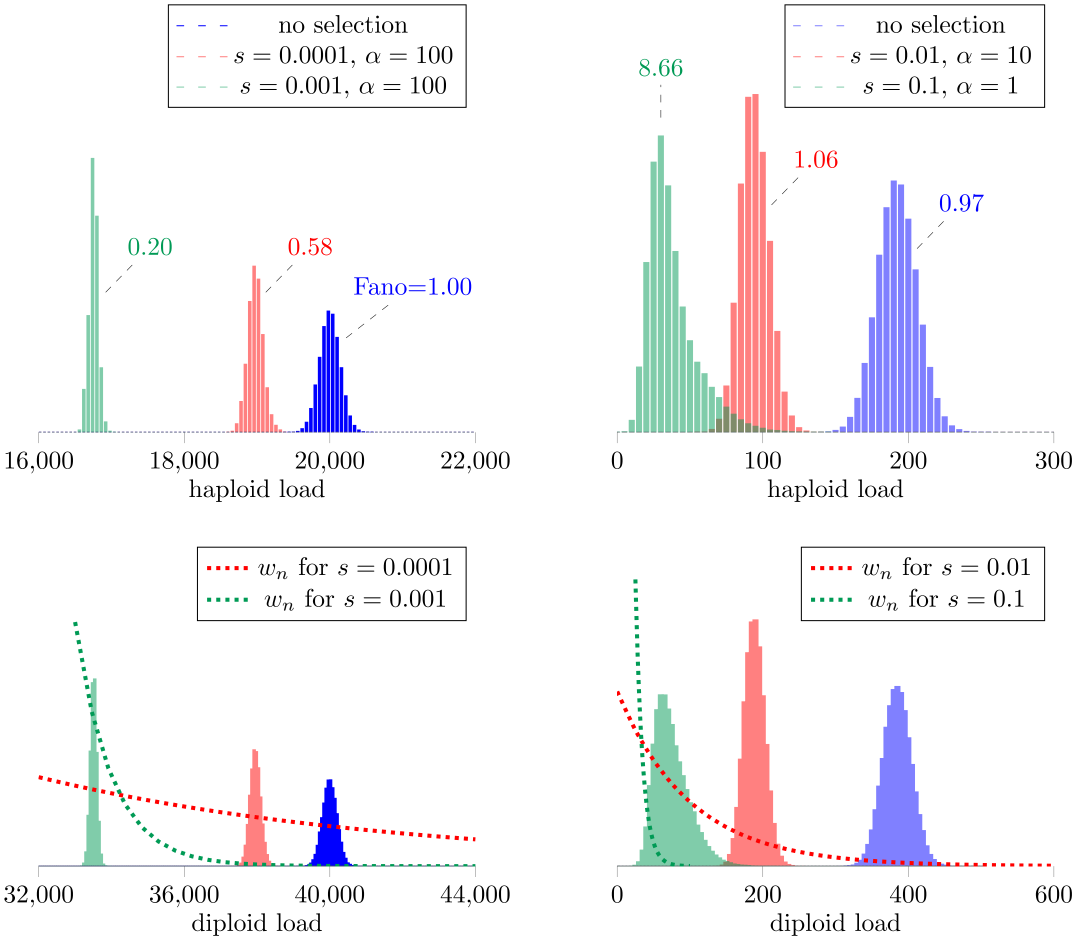


Figure 4: Stationary population distributions of TE abundance with and without selection given by numerical solution of Fokker-Planck equation associated to the master equations (Eqs. 31 and 32). Parameters as in Fig. 2 and legends.

less viable). For concreteness, let

$$w_n = (1-s)^{\ell} \quad \text{for} \quad 0 \le s << 1\,, \tag{30}$$

where s is the strength of selection against TE load. When the neutral model (Eqs. 20–22) is modified to include selection, the master equation becomes

$$\begin{aligned}\frac{dp_n}{dt} &= \alpha(p'_n - p_n) - [(\eta_0 + \eta n)(1 - n/m) + \nu n]\, p_n \\ &+ [\eta_0 + \eta(n-1)]\,[1-(n-1)/m]\, p_{n-1} + \nu(n+1)p_{n+1}\,. \end{aligned} \tag{31}$$

for $1 \leq n \leq m$. The first term in this expression represents each load probability p_n relaxing to a target probability p'_n given by

$$p'_n = \frac{p_n \sum_j w_{n+j} p_j}{\sum_i p_i \sum_j w_{i+j} p_j} \qquad 0 \leq i, j \leq m \tag{32}$$

where $w_{i+j} = (1-s)^{i+j}$. The equations for for dp_0/dt and dp_m/dt have fewer gain/loss terms than Eq. 31, but are analogous (cf. Eqs. 20 and 22). The parameter α that occurs in Eq. 31 is the inverse of the generation time. The quantity $\bar{w} = \sum_i p_i \sum_j w_{i+j} p_j$ is the mean fitness under the assumption of random mating (Gillespie, 2004).

Fig. 4 shows steady-state distributiosn of haploid (top row) and diploid (bottom) TE loads calculated using Eq. 31 both with and without of selection on diploid load. As expected, for both *Mimulus*- and *Drosophila*-like mean loads, the effect of weak selection (red and green histograms) is to decrease the TE load in the population as compared to the neutral model (blue histograms). This decrease in mean TE load occurs for a wide range of generation times ($1/\alpha$) and selection coefficients (s).

More important (and less obvious) is the impact of selection on the variance of TE load and overdispersion. Using *Drosophila* parameters, Fig. 4 (top right) shows an example simulation (green histogram) in which selection leads to increased dispersion (the Fano factor increases from 1 to 8.66). However, in a second case (red histogram), selection increases the index of dispersion only slightly (to a Fano factor of 1.06). Notably, in three representative simulations using *Mimulus* parameters, selection does not increase the dispersion of TE load (Fig. 4, left). This observation is consistent with the moment-based analysis presented in the following section.

2.8 Moment equations with selection

For a deeper understanding of the impact of selection on the distribution of TE load in a population, one may begin with Eqs. 31 and 32 and derive the dynamics of the mean and variance of TE load under the action of simple selection functions. For example, in the limit of weak selection $0 < s << 1$, Eq. 30 is well-approximated by $w_n = 1 - sn$. In this case, as derived in Section S3, the dynamics of the mean and variance of TE load solve

$$\begin{aligned} \frac{d\bar{n}}{dt} &= -\frac{\alpha s}{1 - s\bar{n}} \cdot \sigma_n^2 + 2\eta_0 - \left(\nu - \eta + \frac{\eta_0}{m}\right)\bar{n} - \frac{\eta}{m}\left(\sigma_n^2 + \frac{\bar{n}^2}{2}\right) && (33) \\ \frac{d\sigma_n^2}{dt} &= -\frac{\alpha s}{1 - s\bar{n}} \cdot \mathsf{E}[(\mathbf{n} - \bar{n})^3] + 2\eta_0 + \left(\nu + \eta - \frac{\eta_0}{m}\right)\bar{n} - 2\left(\nu - \eta + \frac{\eta_0 + \eta/2}{m}\right)\sigma_n^2 \\ &- \frac{2\eta}{m}\left(\bar{n}\sigma_n^2 + \frac{\bar{n}^2}{4} + \mathsf{E}[(\mathbf{n} - \bar{n})^3]\right) . && (34) \end{aligned}$$

These ODEs may be compared to the moment equations for the neutral model (Eqs. 25 and 26). As expected, the influence of selection on the mean TE load is proportional to the population variance

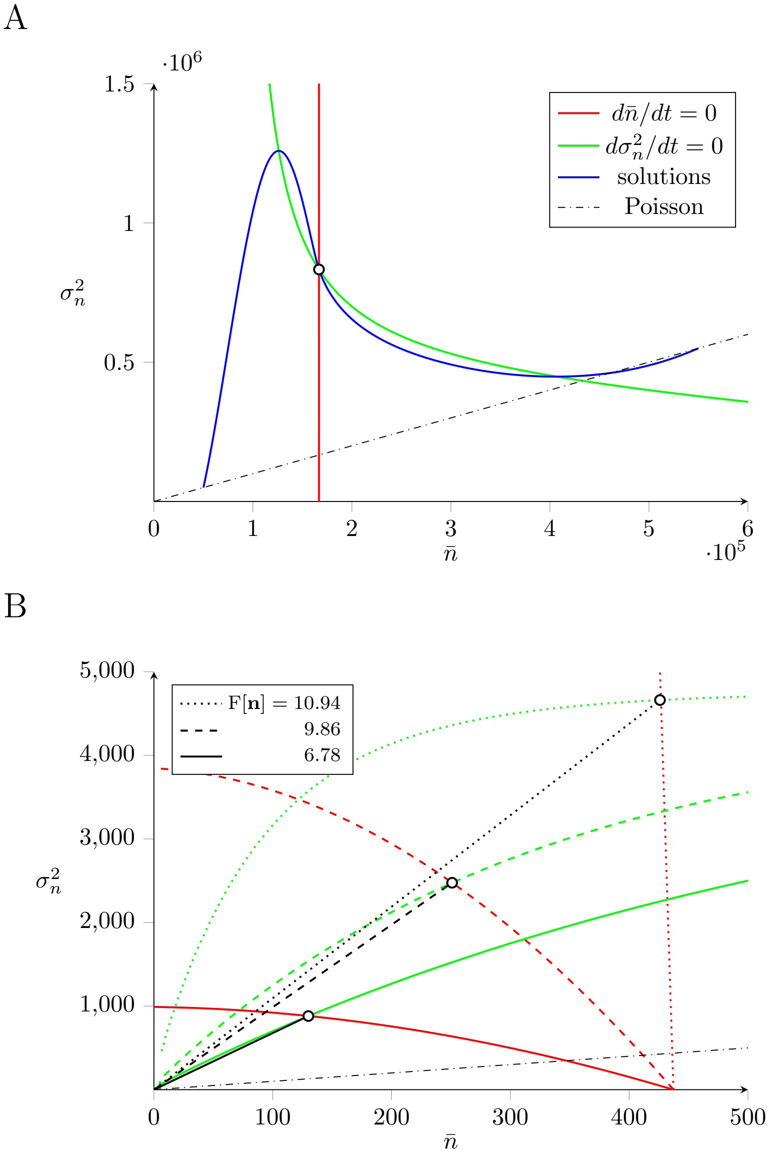


Figure 5: The phase plane for the dynamics of mean (n) and variance (σ_n^2) of TE load (Eqs. 42 and 43). The red and green curves are the nullclines for the mean and variance, respectively, with intersection corresponding to the steady state (open circle). (A) Mean loads similar to *Mimulus*. The blue trajectories show the dynamics of equilibration. (B) Mean loads similar to *Drosophila*. Increased selection decreases the index of dispersion (Fano[**n**]).

(through the factor $-\alpha s\sigma_n^2/(1-s\bar{n})$ in Eq. 35). Similarly, the influence of selection on the population variance is proportional to the third central moment of the diploid load (through the factor $-\alpha s\mathsf{E}[(\mathbf{n}-\bar{n})^3]/(1-s\bar{n})$ in Eq. 34). In both cases, the quantity $1-s\bar{n}$ is the mean fitness of the population, i.e., $\bar{w}=\mathsf{E}[w_{\mathbf{n}}]=\mathsf{E}[1-s\mathbf{n}]=1-s\bar{n}$.

Under the assumption that mean TE load is much smaller than the number of loci ($\bar{n} << 2m$), we may simplify the moment equations with selection (Eqs. 33 and 34) by taking the limit $m\to\infty$ to obtain

$$\frac{d\bar{n}}{dt} = -\frac{\alpha s}{1-s\bar{n}}\cdot\sigma_n^2 + 2\eta_0 - (\nu-\eta)\bar{n} \tag{35}$$

$$\frac{d\sigma_n^2}{dt} = -\frac{\alpha s}{1-s\bar{n}}\cdot\mathsf{E}[(\mathbf{n}-\bar{n})^3] + 2\eta_0 + (\nu+\eta)\bar{n} - 2(\nu-\eta)\sigma_n^2\,. \tag{36}$$

Setting the left side of Eq. 35 to zero, we observe that the steady-state mean and variance are related as follows,

$$\bar{n} = \frac{2\eta_0}{\nu-\eta} - \frac{\alpha s}{1-s\bar{n}}\cdot\frac{\sigma_n^2}{\nu-\eta} = \frac{2\eta_0}{\nu-\eta}\left[1-\frac{\alpha s}{1-s\bar{n}}\cdot\frac{\sigma_n^2}{2\eta_0}\right]\,. \tag{37}$$

Comparing this expression to Eq. 27, noting that the variance is nonnegative ($\sigma_n^2\geq 0$), we see that the effect of weak selection ($0<s<<1$) is to decrease the mean TE load in the population as compared to the neutral model (as expected). Similar analysis of Eq. 36 shows how selection may impact on the variance of of TE load and, consequently, the index of overdispersion. Setting the left side of Eq. 36 to zero and solving for the steady-state variance, gives

$$\sigma_n^2\left[1+\frac{\alpha s}{1-s\bar{n}}\cdot\frac{\nu+\eta}{2(\nu-\eta)^2}\right] = \frac{2\eta_0\nu}{(\nu-\eta)^2} - \frac{\alpha s}{1-s\bar{n}}\cdot\frac{\mathsf{E}[(\mathbf{n}-\bar{n})^3]}{2(\nu-\eta)}\,, \tag{38}$$

where the first term on the right side, $2\eta_0\nu/(\nu-\eta)^2$, is the variance of TE load in the absence of selection. Consistent with the master equation simulations shown in Fig. 4, Eq. 38 shows that the effect of selection is to either decrease or increase the population variance of TE load, depending on the sign of the third central moment ($\mathsf{E}[(\mathbf{n}-\bar{n})^3]$).

2.9 Moment closure and the ($\bar{n}$, σ_n^2) phase plane

In their current form, the moment equations (Eqs. 33 and 34) are an open system of ODEs, because the equation for the variance (σ_n^2) depends on $\mathsf{E}[(\mathbf{n}-\bar{n})^3]$, the unknown third central moment. As discussed in Section S4, a moment closure technique that is applicable in this situation assumes the third central moment of the diploid load is algebraic function of the mean and variance,

$$\mathsf{E}[(\mathbf{n}-\bar{n})^3] = \psi(\bar{n},\sigma_n^2)\,. \tag{39}$$

We investigated two possibilities for this function based on the properties of the beta-binomial and negative binomial distributions. The beta-binomial moment closure, derived in Section S4.3, is a

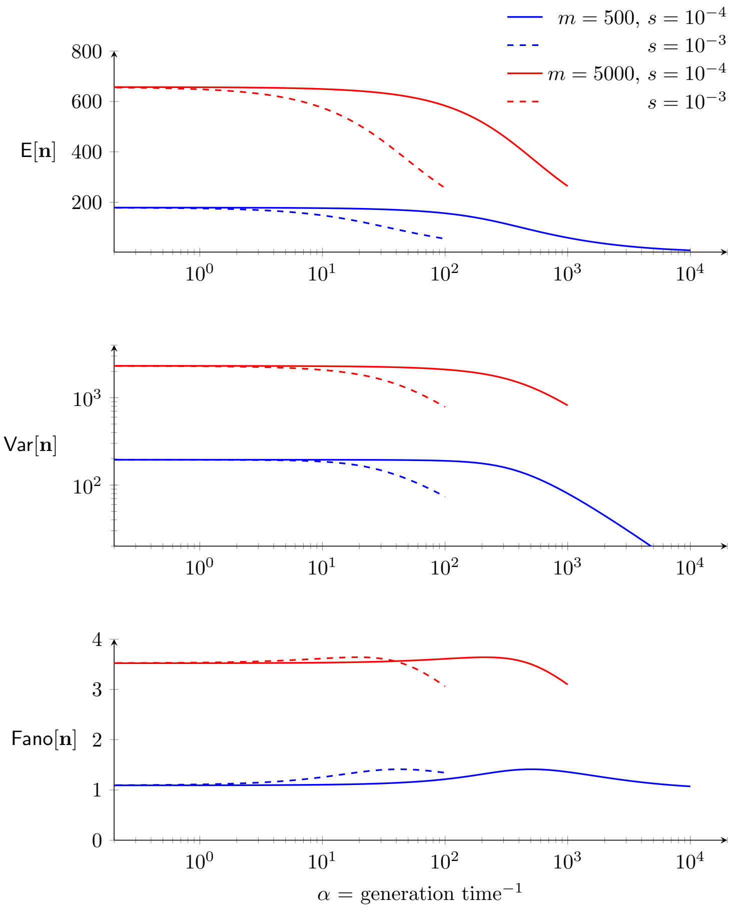


Figure 6: The moment equations derived under the assumption of weak selection (Eqs. 33 and 34) with beta-binomial moment closure (Eq. 40) enabled these parameter studies of the mean, variance, and dispersion of TE load as a function of generation time ($1/\alpha$). Parameters: $\nu = 0.1$, $\eta_0 = 10$, $\eta = 0.1$, and as in legend.

complicated expression involving the mean, variance, and number of loci m,

$$\psi_{BB}(\bar{n}, \sigma_n^2) = \sigma^2 \frac{(m-\bar{n})(\bar{n}^2 - 2m\bar{n} - 2\sigma^2 + 4m\sigma^2)}{m\bar{n}(2m - \bar{n} - 4) + 2m\sigma^2 + 2\bar{n}^2} \,. \tag{40}$$

Moment closure motivated by the properties of the negative binomial distribution results in a simpler expression that does not involve the number of loci m,

$$\psi_{NB}(\bar{n}, \sigma_n^2) \quad = \quad \sigma_n^2 \left(\frac{2\sigma_n^2 - \bar{n}}{\bar{n}} \right) \,. \tag{41}$$

Although the beta-binomial closure (Eq. 40) is arguably a better approximation, in our experience it does not perform markedly better than the negative binomial closure (Eq. 41), as assessed through comparison of moment ODE and master equation simulations. In the analysis that follows, we use the negative binomial closure, motivated by its simplicity and the fact the two expressions coincide the number of loci are not limiting (to see this, observe that $\psi_{BB} \to \psi_{NB}$ as $m \to \infty$). When the algebraic relationship representing the negative binomial closure (Eq. 41) is substituted into Eqs. 23–24, we obtain the following system of ODEs for the mean and variance of diploid load under the influence of selection:

$$\begin{aligned} \frac{d\bar{n}}{dt} &= -\frac{\alpha s}{1 - s\bar{n}} \cdot \sigma_n^2 + 2\eta_0 - \left(\nu - \eta + \frac{\eta_0}{m}\right)\bar{n} - \frac{\eta}{m}\left(\sigma_n^2 + \frac{\bar{n}^2}{2}\right) \qquad (42) \\ \frac{d\sigma_n^2}{dt} &= -\frac{\alpha s}{1 - s\bar{n}} \cdot \sigma_n^2 \left(\frac{2\sigma_n^2 - \bar{n}}{\bar{n}}\right) + 2\eta_0 + \left(\nu + \eta - \frac{\eta_0}{m}\right)\bar{n} - 2\left(\nu - \eta + \frac{\eta_0 + \eta/2}{m}\right)\sigma_n^2 \\ &- \frac{2\eta}{m}\left[\bar{n}\sigma_n^2 + \frac{\bar{n}^2}{4} + \sigma_n^2\left(\frac{2\sigma_n^2 - \bar{n}}{\bar{n}}\right)\right]. \qquad (43) \end{aligned}$$

Fig. 5A presents a representative (n, σ_n^2) phase plane for the dynamics of the mean and variance of TE load predicted by Eqs. 42 and 43. The red and green lines are the nullclines for the mean and variance, respectively, with intersection corresponding to the steady state. This calculation uses parameters resulting in a steady-state TE load similar to our empirical observations of *Mimulus guttatus* (counts on the order of 10^5). This steady state predicted by the moment equations is located far above the broken black line denoting $\sigma_n^2 = n$ and Fano factor of 1. The blue curves show two numerically integrated solutions using initial conditions for which the population variance is equal to the mean. Interestingly, these solutions show that dynamics of TE load can include a transient phase in which the index of dispersion is far greater or less than the steady-state value.

Fig. 5B shows how the nullclines for the mean and variance of TE load depend on the strength of selection in three cases with parameters corresopnding to TE loads similar to *Drosophila melanogaster* (counts on the order of 100). As the strength of selection increases, both the mean and variance of TE load decrease, in such a manner that the index of dispersion decreases (compare slopes of broken black lines).

Although the model obtained by moment closure and the phase plane analysis of Fig. 5 does not assume $\bar{n} << 2m$, we may consider Eqs. 42 and 43 in the limit as $m \to \infty$,

$$\begin{aligned} \frac{d\bar{n}}{dt} &= -\frac{\alpha s}{1 - s\bar{n}} \cdot \sigma_n^2 + 2\eta_0 - (\nu - \eta)\bar{n} \qquad (44) \\ \frac{d\sigma_n^2}{dt} &= -\frac{\alpha s}{1 - s\bar{n}} \cdot \sigma_n^2 \left(\frac{2\sigma_n^2 - \bar{n}}{\bar{n}}\right) + 2\eta_0 + (\nu + \eta)\bar{n} - 2(\nu - \eta)\sigma_n^2 \,. \qquad (45) \end{aligned}$$

Setting the left sides of Eqs. 42–43 to zero, and assuming weak selection ($0 \leq s << 1$), we can derive first-order accurate asymptotic expressions for the steady-state mean and variance,

$$\begin{aligned} \bar{n} &\approx \frac{2\eta_0}{\nu - \eta}\left[1 - \alpha s \frac{\nu}{(\nu - \eta)^2}\right] \qquad (46) \\ \sigma_n^2 &\approx \frac{2\nu\eta_0}{(\nu - \eta)^2}\left[1 - \alpha s \frac{(\nu + \eta)}{(\nu - \eta)^2}\right] . \qquad (47) \end{aligned}$$

Because $v/(\nu - \eta)^2 > 0$, this expression indicates that weak selection decreases the mean TE load, consistent with our intuition. Similarly, the factor $(\nu + \eta)/(\nu - \eta)^2$ is positive, so we conclude that weak selection decreases the population variance when m is large. As for the index of dispersion, this analysis indicates that under weak selection the Fano factor is

$$\frac{\sigma_n^2}{\bar{n}} \approx \frac{\nu}{\nu - \eta}\left[1 - \alpha s \frac{\eta}{(\nu - \eta)^2}\right] . \tag{48}$$

Because $\eta/(\nu - \eta)^2$ is positive any nonzero copy-and-paste rate ($\eta > 0$), we conclude that the Fano factor is also expected to decrease, because weak selection causes the within-population variance of TE load to decrease more than the mean. This conclusion that selection on diploid TE load is unlikely to be responsible for overdispersion is consistent with numerical parameter studies summarized in Fig. 6 that were enabled by the moment equations with selection (Eqs. 33 and 34) and beta-binomial moment closure (Eq. 40).

3 DISCUSSION

Although mathematical modeling has informed our understanding of the population genetics of transposable elements (TEs) for several decades, classical theory has emphasized analytical results that assume a binomial distribution of TE loads in a randomly mating population (Sections 1.1 and 1.2). Because the variance of a binomial distribution is less than or equal to its mean, the classical theory effectively assumes that the population distribution of TE loads are underdispersed ($\mathsf{Fano}[\mathbf{n}] \leq 1$).

In an empirical analysis of TE copy number in two natural populations (*M. guttatus* and *D. melanogaster*), we found (in both cases) that the population distribution of TE loads was dramatically overdispersed (Fig. 1 and Table 1). Because the classical theory of TE population genetics is not applicable to this situation, we extended this theory and explored mechanisms that may be responsible for observed overdispersion. The model presented here predicts the entire distribution function of TE loads, and from this distribution we calculate the mean, variance, and index of dispersion as a function of model parameters.

Prior to considerations of selection, the parameters of neutral model encode assumptions regarding the dynamics of TE proliferation (cut-and-paste, copy-and-paste, and excision rate constants) as well as an estimate of the number of loci that may be occupied by TEs. Using parameter sets that yield TE counts in the observed ranges (tens of thousands for *M. guttatus*, hundreds for *D. melanogaster*), we found (in both cases) that copy-and-paste TE proliferation dynamics often resulted in an overdispersed TE loads (Fig. 2). Moment-based analysis of the neutral model suggests that overdispersed population distributions are to be expected when the copy-and-paste transposition rate constant (η) and excision rate constant (ν) are comparable in magnitude (Fig. 3 and Table 1).

We next extended the master equation model to include purifying selection on TE load. For a parameter set corresponding to *M. guttatus*, selection decreased the mean and variance of TE load; however, because the variance decreased more than the mean, purifying selection had the effect of decreasing the index of dispersion (Fig. 4, left). For a parameter set corresponding to *D. melanogaster*, we found that purifying selection, when sufficiently strong, may lead to an increased index of dispersion of TE load (Fig. 4, right). Most importantly, in both parameter regimes, our simulations (Fig. 6) and analysis (Eqs. 46–48) agree that weak purifying selection decreases both the mean and variance of TE load in such a way that the index of dispersion is unchanged or slightly increases. Moment-based analysis of the master equation confirmed that weak selection has the effect of decreasing the index of dispersion (Section 2.9).

3.1 Comparison of *M. guttatus* and *D. melanogaster*

Parameter studies using the master equation model indicate that the mechanism of copy-and-paste transposition may lead to overdispersed population distributions of TE load, whereas cut-and-paste transition is less likely to do so (Fig. 2). When our empirical analysis of TE load in *D. melanogaster* was refined to consider these two broad classes of TEs, we found that copy-and-paste TEs were 6-fold more highly overdispersed than cut-and-paste TEs (see Table 1), consistent with the model prediction. On the other hand, our empirical analysis of TE load in *M. guttatus* shows that in this natural population copy-and-paste TEs are far less dispersed than cut-and-paste TEs.

3.2 Limitations of the model

The master equation for the dynamics of TE proliferation presented here extends the classical theory in several ways. Most importantly, in the master equation simulations, the relationship between the population variance and mean is a prediction of the model (as opposed to a modeling assumption, as in the classical theory). This feature of the model enables parameter studies exploring how the dynamics of TE proliferation and purifying selection influence the dispersion in TE load.

One limitation of our model is the harsh (but common) assumption that selection acts on overall TE load (Brookfield and Badge, 1997; Charlesworth and Charlesworth, 1983, 2010; Deceliere, 2004; Le Rouzic and Deceliere, 2005). This choice is consistent with the finding that most TE insertions have negative fitness consequences and are located outside of genes (Bartolomé et al., 2002; Duret et al., 2000; Mackay, 1989; Pasyukova et al., 2004). On the other hand, many TEs are located in heterochromatic regions of the genome. It is unlikely that these large masses of TEs have fitness consequences comparable to TEs that are proximal to genes. In future work, our model could be extended to include variability in the selective cost of TE insertions.

The most significant limitation of the master equation model is that the dynamics of recombination are not represented. Indeed, the population distribution of TE load is modeled without any representation of the location of TEs within the genome. To the extent that recombination promotes linkage equilibrium, one expects that recombination will decrease the dispersion of TE load and, consequently, is not a likely explanation for observed overdispersion. We recommend interpreting the master equation model as a representation of the dynamics of a single linkage class of TEs, with the tacit understanding that the index of dispersion for a genome composed of multiple linkages classes will be less than the model prediction. Admittedly, this viewpoint does not account for the fact that recombination is less frequent in regions of the genome that have a high density of TEs. Studying the influence of such density-dependent recombination on the dispersion of TE load is beyond the scope of this paper, as it would require a modeling framework that is explicitly spatial.

We note that events involving the loss or gain of multiple TEs (as could occur via ectopic recombination or other mechanisms) are expected to contribute to overdispersion. To see this, consider a master equation simulation in which the gain and loss of TEs occurs in blocks of size b. If there is no other change to the model, we may reinterpret the random variable $\mathbf{n}$ as the number of blocks of TEs in a randomly sampled diploid genome. In that case, the mean and variance of TE count are increased by a factor of b and b^2, respectively. The Fano factor, given by the ratio of variance to mean, increases by a factor of b,

$$\mathsf{Fano}[b\mathbf{n}] = \frac{\mathsf{Var}[b\mathbf{n}]}{\mathsf{E}[b\mathbf{n}]} = \frac{b^2\mathsf{Var}[\mathbf{n}]}{b\mathsf{E}[\mathbf{n}]} = b\mathsf{Fano}[\mathbf{n}]\,.$$

This scaling implies that block-wise inheritance of TEs is expected to increase the index of dispersion by a factor proportional to the representative block size. This intriguing and relatively simple explanation for empirically observed overdispersion could be studied using an explicitly spatial model of TE population genetics, preferably one that includes a mechanistic account of ectopic recombination and perhaps other genome rearrangements.

ACKNOWLEDGEMENTS

The work was supported in part by National Science Foundation Grant DMS 1951646 to GDCS and National Insitutes of Health (NCCIH) Grant R01 AT010816.

References

Carolina Bartolomé, Xulio Maside, and Brian Charlesworth. On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. *Molecular Biology and Evolution*, 19(6):926–937, 2002.

Guillaume Bourque, Kathleen H Burns, Mary Gehring, Vera Gorbunova, Andrei Seluanov, Molly Hammell, Michaël Imbeault, Zsuzsanna Izsvák, Henry L Levin, Todd S Macfarlan, et al. Ten things you should know about transposable elements. *Genome Biology*, 19(1):199, 2018.

John F.Y. Brookfield and Richard M. Badge. Population genetics models of transposable elements. *Genetica*, 100(1-3):281–294, 1997.

Michael George Bulmer. *The Mathematical Theory of Quantitative Genetics.* Clarendon Press, 1980.

Brian Charlesworth and Deborah Charlesworth. The population dynamics of transposable elements. *Genetics Research*, 42(1):1–27, 1983.

Brian Charlesworth and Deborah Charlesworth. *Elements of Evolutionary Genetics.* W. H. Freeman, January 2010.

Julie M Cridland, Stuart J Macdonald, Anthony D Long, and Kevin R Thornton. Abundance and distribution of transposable elements in two *Drosophila* QTL mapping resources. *Molecular Biology and Evolution*, 30(10):2311–2327, 2013.

Grégory Deceliere. The dynamics of transposable elements in structured populations. *Genetics*, 169 (1):467–474, 2004.

Laurent Duret, Gabriel Marais, and Christian Biémont. Transposons but not retrotransposons are located preferentially in regions of high recombination rate in *Caenorhabditis elegans*. *Genetics*, 156 (4):1661–1669, 2000.

Crispin Gardiner. *Stochastic Methods: A Handbook for the Natural and Social Sciences.* Springer, 4th edition, 2009.

John H. Gillespie. *Population Genetics: A Concise Guide.* The Johns Hopkins University Press, 2004.

Tyler V Kent, Jasmina Uzunović, and Stephen I Wright. Coevolution between transposable elements and recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372 (1736):20160458, 2017.

Arnaud Le Rouzic and Grégory Deceliere. Models of the population genetics of transposable elements. *Genetics Research*, 85(3):171–181, 2005.

Trudy FC Mackay. Transposable elements and fitness in *Drosophila melanogaster*. *Genome*, 31(1): 284–295, 1989.

Ryan E Mills, E Andrew Bennett, Rebecca C Iskow, and Scott E Devine. Which transposable elements are active in the human genome? *Trends in Genetics*, 23(4):183–191, 2007.

EG Pasyukova, SV Nuzhdin, TV Morozova, and TFC Mackay. Accumulation of transposable elements in the genome of *Drosophila melanogaster* is associated with a decrease in fitness. *Journal of Heredity*, 95(4):284–290, 2004.

Nathan M Springer, Kai Ying, Yan Fu, Tieming Ji, Cheng-Ting Yeh, Yi Jia, Wei Wu, Todd Richmond, Jacob Kitzman, Heidi Rosenbaum, A. Leonardo Iniguez, W. Brad Barbazuk, Jeffrey A. Jeddeloh, Dan Nettleton, and Patrick S. Schnable. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genetics*, 5(11):e1000734, 2009.

Nicolaas Godfried Van Kampen. *Stochastic Processes in Physics and Chemistry*. North Holland, 3rd edition, 2007.

Supplemental Materials

Population genetics of transposable element load: a mechanistic account of observed overdispersion

Ronald D. Smith, Joshua R. Puzey, Gregory D. Conradi Smith

S1 Comments on the classical model

In the classical model of TE population genetics, the state of an infinite diploid population at a given chromosomal site i, for $0 \le i \le m$, is described by its frequency, x_i, where $0 \le x_i \le 1$. Assuming insertion sites exhibit no linkage disequilibrium, the set of frequencies, $\{x_i\}_{i=1}^m$, describes the state of the population. The TE load of a randomly sampled diploid individual is a random number $\mathbf{n}$ given by

$$\mathbf{n} = \sum_{i=1}^{m} \mathbf{x}_i + \sum_{i=1}^{m} \mathbf{y}_i \,. \tag{S1}$$

where $\mathbf{x}_i$ and $\mathbf{y}_i$ are pairs of i.i.d. Bernoulli random variables with parameters x_i. Thus, $\mathsf{E}[\mathbf{x}_i] = \mathsf{E}[\mathbf{y}_i] = x_i$ and the mean copy number of TEs per individual is

$$\bar{n} = \mathsf{E}[\mathbf{n}] = \sum_{i=1}^{m} \mathsf{E}[\mathbf{x}_i] + \sum_{i=1}^{m} \mathsf{E}[\mathbf{y}_i] = 2\sum_{i=1}^{m} x_i = 2m\bar{x}$$

where in the last equality we have written $\bar{n}$ in terms of the number of loci and the mean frequency, $\bar{x} = (1/m)\sum_{i=1}^m x_i$. The variance of a Bernoulli random variable with parameter x_i is $x_i(1 - x_i)$. As a consequence, the variance of TE load is

$$\mathsf{Var}[\mathbf{n}] = \sum_{i=1}^{m} \mathsf{Var}[\mathbf{x}_i + \mathbf{y}_i] = 2\sum_{i=1}^{m} x_i(1 - x_i)\,.$$

The diploid load given by Eq. S1 is the sum of two i.i.d. Poisson-binomial random variables, that is, $\mathbf{n} = \mathbf{X} + \mathbf{Y}$ where $\mathbf{X} = \sum_{i=1}^m \mathbf{x}_i$ is the sum of independent Bernoulli random variables that are not necessarily independent (and similarly for $\mathbf{Y} = \sum_{i=1}^m \mathbf{y}_i$). It is well-known that

$$\mathsf{Var}[\mathbf{X}] = m\bar{x}(1 - \bar{x}) - m\sigma_x^2\,. \tag{S2}$$

where $\sigma_x^2 = (1/m)\sum_{i=1}^m (x_i - \bar{x})^2$ is the “variance” among the parameters of the Poisson-binomial distribution, $\{x_i\}_{i=1}^m$ (i.e., the variability of frequencies of occupation of the TE loci). Using $\mathsf{Var}[\mathbf{Y}] = \mathsf{Var}[\mathbf{X}]$ and $\mathsf{Var}[\mathbf{n}] = 2\mathsf{Var}[\mathbf{X}]$, and Eq. S2, we see that the variance of TE load is

$$\mathsf{Var}[\mathbf{n}] = 2m\bar{x}(1 - \bar{x}) - 2m\sigma_x^2\,.$$

Substituting $\bar{x} = \bar{n}/2m$ in the above expression gives Eq. 6.

To extend this model of TE population genetics to include the effect of natural selection, Charlesworth and Charlesworth (1983) assume a viability function, w_n, that is a decreasing function of total genome-wide TE load ($dw_n/dn < 0$). The effect of selection is to decrease the occupation frequency at each loci in a manner that is proportional to $\frac{1}{2}x_i(1-x_i)$, which is the variance of a Bernoulli random variable with parameter x_i, and also proportional the derivative, with respect to x_i, of the mean fitness of the population, $\bar{w} = \mathsf{E}[w_{\mathbf{n}}]$,

$$\Delta x_i = \frac{x_i(1-x_i)}{2\bar{w}} \frac{\partial \bar{w}}{\partial x_i} = x_i(1-x_i) \frac{\partial \ln \bar{w}}{\partial \bar{n}} \, .$$

The second equality is obtained using $(1/x_i)\partial\bar{w}/\partial x_i = \partial \ln \bar{w}/\partial x_i$ and noting that $\bar{n} = 2\sum_{i=1}^m x_i$ implies $\partial \bar{n}/\partial x_i = 2$. Summing over all sites gives

$$\Delta \bar{n} = 2\sum_{i=1}^{m} \Delta x_i = \bar{n}\left(1 - \frac{\bar{n}}{2m}\right) \frac{\partial \ln \bar{w}}{\partial \bar{n}} \, .$$

Using $V_n = \bar{n}(1 - \bar{n}/2m)$ and approximating the mean fitness of the population ($\bar{w} = \mathsf{E}[w_{\mathbf{n}}]$) by the fitness of an individual with an average number of copies ($w_{\bar{n}}$) gives the first term of Eq. 3.

S2 Derivation of moment equations with gain and loss terms

Let $\mathbf{n}$ be a random variable representing TE load of a randomly sampled diploid individual, and $\mathbf{x}$ and $\mathbf{y}$ be random variables representing the TE load of randomly sampled haploid genomes (gametes). Define the moments of the probability distribution of haploid TE loads (given by p_i for $0 \le i \le m$) as

$$\mu_q = \mathsf{E}[\mathbf{x}^q] = \sum_{n=0}^{m} n^q p_n \quad \text{for} \quad q = 0, 1, 2, \cdots \tag{S3}$$

where $\mu_0 = 1$ (conservation of probability), $\mu_1 = \mathsf{E}[\mathbf{x}]$ and $\mathsf{Var}[\mathbf{x}] = \mathsf{E}[\mathbf{x}^2] - \mathsf{E}[\mathbf{x}]^2 = \mu_2 - \mu_1^2$. We assume random mating for which the diploid TE load is the sum of two i.i.d. gametic loads ($\mathbf{n} = \mathbf{x} + \mathbf{y}$). In the case, the mean and variance of the within-population diploid TE load are related to μ_2 and μ_1 as follows,

$$\mathsf{E}[\mathbf{n}] = 2\mathsf{E}[\mathbf{x}] = 2\mu_1 \tag{S4}$$

$$\mathsf{Var}[\mathbf{n}] = 2\mathsf{Var}[\mathbf{x}] = 2(\mu_2 - \mu_1^2) \, . \tag{S5}$$

Using Eqs. S4–S5 we see that the Fano factor of the diploid load is

$$\mathsf{Fano}[\mathbf{n}] = \frac{\mathsf{Var}[\mathbf{n}]}{\mathsf{E}[\mathbf{n}]} = \frac{2(\mu_2 - \mu_1^2)}{2\mu_1} = \frac{\mu_2}{\mu_1} - \mu_1 \, .$$

Because the factor of two occurs in both the numerator and denominator of the above expression, the dispersion of the diploid load is equal to that of the haploid load, i.e. $\mathsf{Fano}[\mathbf{n}] = \mathsf{Fano}[\mathbf{x}]$.

S2.1 ODEs for the dynamics of μ_1 and μ_2

The moment equations for the haploid TE load are derived by differentiating Eq. S3 to obtain

$$\frac{d\mu_q}{dt} = \sum_{n=0}^{m} n^q \frac{dp_n}{dt} \, .$$

Substituting Eqs. 20–22 into the above expression gives

$$\begin{aligned}
\frac{d\mu_q}{dt} &= \sum_{n=0}^{m} n^q [-(u_n + v_n)p_n + u_{n-1}p_{n-1} + v_{n+1}p_{n+1}] \\
&= \underbrace{-\sum_{n=0}^{m} n^q v_n p_n + \sum_{n=0}^{m} n^q v_{n+1} p_{n+1}}_{V_q} \underbrace{- \sum_{n=0}^{m} n^q u_n p_n + \sum_{n=0}^{m} n^q u_{n-1} p_{n-1}}_{U_q} \, .
\end{aligned}$$

The terms V_q are evaluated as follows,

$$\begin{aligned}
V_q &= \sum_{n=0}^{m} n^q v_{n+1} p_{n+1} - \sum_{n=0}^{m} n^q v_n p_n = \sum_{n=1}^{m} (n-1)^q v_n p_n - \sum_{n=1}^{m} n^q v_n p_n \\
&= \sum_{n=1}^{m} [(n-1)^q - n^q] v_n p_n = \sum_{n=0}^{m} [(n-1)^q - n^q] v_n p_n \, ,
\end{aligned}$$

where we use $v_0 = 0$. For $q = 1$, $[(n-1)^q - n^q] = -1$; thus, V_1 is given by

$$V_1 = -\sum_{n=0}^{m} v_n p_n = -\sum_{n=0}^{m} \nu n p_n = -\nu \sum_{n=0}^{m} n p_n = -\nu\mu_1 \, , \tag{S6}$$

where we use $v_n = \nu n$. Similarly,

$$\begin{aligned}
U_q &= \sum_{n=0}^{m} n^q u_{n-1} p_{n-1} - \sum_{n=0}^{m} n^q u_n p_n = \sum_{n=0}^{m-1} (n-1)^q u_n p_n - \sum_{n=0}^{m-1} n^q u_n p_n \\
&= \sum_{n=0}^{m-1} [(n+1)^q - n^q] u_n p_n = \sum_{n=0}^{m} [(n+1)^q - n^q] u_n p_n \, .
\end{aligned}$$

Using $q = 1$, $u_m = 0$, $[(n+1)^q - n^q] = +1$, and $u_n = (\eta_0 + \eta n)(1 - n/m)$, we obtain

$$U_1 = \sum_{n=0}^{m} u_n p_n = \sum_{n=0}^{m} (\eta_0 + \eta n)(1 - n/m) p_n = \eta_0 \sum_{n=0}^{m} (1 - n/m) p_n + \eta \sum_{n=0}^{m} n(1 - n/m) p_n \, .$$

Thus, U_1 is given by

$$U_1 = \eta_0 - \eta_0 \mu_1/m + \eta\mu_1 - \eta\mu_2/m \, . \tag{S7}$$

Combining the expression for U_1 and V_1 gives the ODE for μ_1, the first moment of haploid TE load,

$$\frac{d\mu_1}{dt} = \eta_0[1 - \mu_1/m] - \nu\mu_1 + \eta[\mu_1 - \mu_2/m] = V_1 + U_1 \, . \tag{S8}$$

Similar calculations give $d\mu_0/dt = 0$ (conservation of probability) and

$$V_2 = \nu\mu_1 - 2\nu\mu_2 \tag{S9}$$

$$U_2 = \eta_0[1 - \mu_1/m] + 2\eta_0[\mu_1 - \mu_2/m] + \eta[\mu_1 - \mu_2/m] + 2\eta[\mu_2 - \mu_3/m] \, . \tag{S10}$$

Using these expressions, the ODE for the second moment, $d\mu_2/dt = V_2 + U_2$, is found to be

$$\frac{d\mu_2}{dt} = \eta_0[1 - \mu_1/m] + 2\eta_0[\mu_1 - \mu_2/m] + \nu\mu_1 - 2\nu\mu_2 + \eta[\mu_1 - \mu_2/m] + 2\eta[\mu_2 - \mu_3/m]\,. \tag{S11}$$

Eqs. S8–S11 are the first two ODEs in a sequence for which $d\mu_1/dt$ depends on μ_1 and μ_2, $d\mu_2/dt$ depends on μ_1, μ_2 and μ_3, and so on, as follows,

$$\begin{aligned} \frac{d\mu_1}{dt} &= f_1(\mu_1, \mu_2) && \text{(S12)}\\ \frac{d\mu_2}{dt} &= f_2(\mu_1, \mu_2, \mu_3) && \text{(S13)}\\ &\vdots \\ \frac{d\mu_q}{dt} &= f_q(\mu_{q-1}, \mu_q, \mu_{q+1})\,. && \text{(S14)} \end{aligned}$$

Section S4 shows how this open system of ODEs can be closed by assuming an algebraic relationship between the third moment and those of lower order. The influence of selection on the both the master equation and moment equation models is discussed in Section S3. The following two sections (Sections S2.2 and S2.3) explore parameter regimes for which the moment equations decouple and it is possible to derive analytical steady states for the population mean and variance of TE load.

S2.2 Moment ODEs in absence of copy-and-paste transposition

When there is no copy-and-paste transposition ($\eta = 0$), Eqs. S8–S11 simplify as follows:

$$\begin{aligned} \frac{d\mu_1}{dt} &= \eta_0(1 - \mu_1/m) - \nu\mu_1 \\ \frac{d\mu_2}{dt} &= \eta_0(1 - \mu_1/m) + 2\eta_0(\mu_1 - \mu_2/m) + \nu\mu_1 - 2\nu\mu_2\,. \end{aligned}$$

Notice that the dependence of $d\mu_1/dt$ on μ_2, and $d\mu_2/dt$ on μ_3, vanishes when $\eta = 0$. Regrouping terms gives

$$\begin{aligned} \frac{d\mu_1}{dt} &= \eta_0 - (\nu + \eta_0/m)\mu_1 \\ \frac{d\mu_2}{dt} &= \eta_0 + (v + 2\eta_0 - \eta_0/m)\mu_1 - 2(\nu + \eta_0/m)\mu_2\,. \end{aligned}$$

This system has steady state given by

$$\begin{aligned} \mu_1 &= \frac{\eta_0}{\nu + \eta_0/m} = \frac{m\,\eta_0/\nu}{m + \eta_0/\nu} \\ \mu_2 &= \frac{\eta_0 + (\nu + 2\eta_0 - \eta_0/m)\mu_1}{2(\nu + \eta_0/m)} = \mu_1^2 + \frac{\nu}{\nu + \eta_0/m}\mu_1\,. \end{aligned}$$

The central moment $\hat{\mu}_2 = \mu_2 - \mu_1^2$, the variance in haploid load, is thus

$$\hat{\mu}_2 = \frac{\nu}{\nu + \eta_0/m}\mu_1 = \frac{\eta_0\nu}{(\nu + \eta_0/m)^2} = \frac{m^2\,\eta_0/\nu}{(m + \eta_0/\nu)^2}\,.$$

Noting that $\mathsf{Var}[\mathbf{n}] = \sigma_n^2 = 2\hat{\mu}_2$ and $\mathsf{E}[\mathbf{n}] = \bar{n} = 2\mu_1$, we find

$$\begin{aligned}\bar{n} &= \frac{2\eta_0}{\nu + \eta_0/m} = \frac{2m\,\eta_0/\nu}{m + \eta_0/\nu} \\ \sigma_n^2 &= \frac{2\eta_0\nu}{(\nu + \eta_0/m)^2} = \frac{2m^2\,\eta_0/\nu}{(m + \eta_0/\nu)^2} = \frac{\nu\,\bar{n}}{\nu + \eta_0/m} = \frac{m\,\bar{n}}{m + \eta_0/\nu}\,.\end{aligned}$$

The index of dispersion for the diploid load is thus

$$\mathsf{Fano}[\mathbf{n}] = \frac{\nu}{\nu + \eta_0/m} = \frac{m}{m + \eta_0/\nu}\,.$$

S2.3 Moment ODEs when occupiable loci are not limiting

When occupiable loci are not limiting ($\mu_1 << m$), we may consider Eqs. S8–S11 in the limit as $m \to \infty$,

$$\begin{aligned}\frac{d\mu_1}{dt} &= \eta_0 - \nu\mu_1 + \eta\mu_1 \\ \frac{d\mu_2}{dt} &= \eta_0 + 2\eta_0\mu_1 + \nu\mu_1 - 2\nu\mu_2 + \eta\mu_1 + 2\eta\mu_2\,.\end{aligned}$$

Note that the large m limit uncouples the moment ODEs. Regrouping terms, we see that

$$\begin{aligned}\frac{d\mu_1}{dt} &= \eta_0 - (\nu - \eta)\mu_1 \\ \frac{d\mu_2}{dt} &= \eta_0 + (2\eta_0 + \eta + \nu)\mu_1 - 2(\nu - \eta)\mu_2\,.\end{aligned}$$

Provided $\nu > \eta$, this system has the stable steady state given by

$$\begin{aligned}\mu_1 &= \frac{\eta_0}{\nu - \eta} \\ \mu_2 &= \frac{\eta_0 + (2\eta_0 + \eta + \nu)\mu_1}{2(\nu - \eta)} = \mu_1^2 + \frac{\nu}{\nu - \eta}\mu_1\,.\end{aligned}$$

The central moment $\hat{\mu}_2 = \mu_2 - \mu_1^2$, the variance in haploid load, is thus

$$\hat{\mu}_2 = \frac{\nu}{\nu - \eta}\mu_1 = \frac{\eta_0\nu}{(\nu - \eta)^2}\,.$$

Noting that $\mathsf{Var}[\mathbf{n}] = \sigma_n^2 = 2\hat{\mu}_2$ and $\mathsf{E}[\mathbf{n}] = \bar{n} = 2\mu_1$ gives Eqs. 27–28, namely,

$$\begin{aligned}\bar{n} &= \frac{2\eta_0}{\nu - \eta} \\ \sigma_n^2 &= \frac{2\eta_0\nu}{(\nu - \eta)^2} = \frac{\nu\bar{n}}{\nu - \eta}\,.\end{aligned}$$

The index of dispersion for the diploid load is thus

$$\mathsf{Fano}[\mathbf{n}] = \frac{\nu}{\nu - \eta}\,.$$

S2.4 Central moment equations

Recall that the open system of moment equations for the probability distribution of TE load take the form

$$\begin{aligned} \frac{d\mu_1}{dt} &= \eta_0(1-\mu_1/m) - \nu\mu_1 + \eta(\mu_1 - \mu_2/m) \\ \frac{d\mu_2}{dt} &= \eta_0(1-\mu_1/m) + 2\eta_0(\mu_1 - \mu_2/m) + \nu\mu_1 - 2\nu\mu_2 + \eta(\mu_1 - \mu_2/m) + 2\eta(\mu_2 - \mu_3/m)\,, \end{aligned}$$

where $d\mu_3/dt = f_3(\mu_2, \mu_3, \mu_4)$, and so on (cf. Eqs. S12–S14). Rearranging terms in the equations for the first two moments gives

$$\frac{d\mu_1}{dt} = \eta_0 - \left(\nu - \eta + \frac{\eta_0}{m}\right)\mu_1 - \frac{\eta}{m}\mu_2 \tag{S15}$$

$$\frac{d\mu_2}{dt} = \eta_0 + \left(\nu + \eta + 2\eta_0 - \frac{\eta_0}{m}\right)\mu_1 - 2\left(\nu - \eta + \frac{\eta_0 + \eta/2}{m}\right)\mu_2 - \frac{2\eta}{m}\mu_3 \tag{S16}$$

It is convenient to express Eqs. S15–S16 in terms of the central moments. The first central moment of the haploid TE load is the mean, $\mu_1 = \mathsf{E}[\mathbf{x}]$. The second central moment is the variance

$$\hat{\mu}_2 = \mathsf{Var}[\mathbf{x}] = \mathsf{E}[(\mathbf{x} - \mathsf{E}[\mathbf{x}])^2] = \mathsf{E}[\mathbf{x}^2] - \mathsf{E}[\mathbf{x}]^2 = \mu_2 - \mu_1^2\,.$$

The third central moment is

$$\hat{\mu}_3 = \mathsf{E}[(\mathbf{x} - \mathsf{E}[x])^3] = \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3 = \mu_3 - 3\mu_1\hat{\mu}_2 - \mu_1^3\,. \tag{S17}$$

To find an ODE for the dynamics of the variance, we differentiate $\hat{\mu}_2 = \mu_2 - \mu_1^2$ to obtain

$$\frac{d\hat{\mu}_2}{dt} = \frac{d\mu_2}{dt} - 2\mu_1\frac{d\mu_1}{dt}\,. \tag{S18}$$

Substituting Eqs. S15–S16 into this expression we obtain,

$$\frac{d\mu_1}{dt} = \eta_0 - \left(\nu - \eta + \frac{\eta_0}{m}\right)\mu_1 - \frac{\eta}{m}\left(\hat{\mu}_2 + \mu_1^2\right) \tag{S19}$$

$$\frac{d\hat{\mu}_2}{dt} = \eta_0 + \left(\nu + \eta - \frac{\eta_0}{m}\right)\mu_1 - 2\left(\nu - \eta + \frac{\eta_0 + \eta/2}{m}\right)\hat{\mu}_2 - \frac{\eta}{m}(4\mu_1\hat{\mu}_2 + \mu_1^2 + 2\hat{\mu}_3)\,. \tag{S20}$$

Using $\mathsf{E}[\mathbf{n}] = \bar{n} = 2\mu_1$ and $\mathsf{Var}[\mathbf{n}] = \sigma_n^2 = 2\hat{\mu}_2$, and $\mathsf{E}[(\mathbf{n} - \bar{n})^3] = 2\hat{\mu}_3$, Eqs. S19–S20 may be transformed into equations for the mean and variance of diploid load. To see this, write $\mu_1 = \bar{n}/2$ and $\hat{\mu}_2 = \sigma_n^2/2$ and differentiate to obtain

$$\frac{1}{2}\frac{d\bar{n}}{dt} = \frac{d\mu_1}{dt} \quad \text{and} \quad \frac{1}{2}\frac{d\sigma_n^2}{dt} = \frac{d\hat{\mu}_2}{dt}\,.$$

Substitution gives

$$\begin{aligned} \frac{1}{2}\frac{d\bar{n}}{dt} &= \eta_0 - \frac{1}{2}\left(\nu - \eta + \frac{\eta_0}{m}\right)\bar{n} - \frac{\eta}{2m}\left(\sigma_n^2 + \frac{\bar{n}^2}{2}\right) \\ \frac{1}{2}\frac{d\sigma_n^2}{dt} &= \eta_0 + \frac{1}{2}\left(\nu + \eta - \frac{\eta_0}{m}\right)\bar{n} - \left(\nu - \eta + \frac{\eta_0 + \eta/2}{m}\right)\sigma_n^2 - \frac{\eta}{m}\left(\bar{n}\sigma_n^2 + \frac{\bar{n}^2}{4} + \mathsf{E}[(\mathbf{n} - \bar{n})^3]\right)\,, \end{aligned}$$

where we gave used Eq. S17. After simplifying, these equations become

$$\frac{d\bar{n}}{dt} = 2\eta_0 - \left(\nu - \eta + \frac{\eta_0}{m}\right)\bar{n} - \frac{\eta}{m}\left(\sigma_n^2 + \frac{\bar{n}^2}{2}\right) \tag{S21}$$

$$\begin{aligned}\frac{d\sigma_n^2}{dt} &= 2\eta_0 + \left(\nu + \eta - \frac{\eta_0}{m}\right)\bar{n} - 2\left(\nu - \eta + \frac{\eta_0 + \eta/2}{m}\right)\sigma_n^2 \\ &- \frac{2\eta}{m}\left(\bar{n}\sigma_n^2 + \frac{\bar{n}^2}{4} + \mathsf{E}[(\mathbf{n} - \bar{n})^3]\right) .\end{aligned} \tag{S22}$$

Taking the limit as $m \to \infty$ gives Eqs. 25–26.

S3 Selection in the master equation and moment equation models

In the master equation formulation, p_n is the probability of randomly sampling a gamete with a TE load of n. Under the assumption of random mating, selection leads to the following probabilities for each load in the next generation,

$$p_i' = \frac{p_i \sum_j w_{i+j} p_j}{\sum_i p_i \sum_j w_{i+j} p_j} = \frac{p_i \bar{w}_i}{\sum_i p_i \bar{w}_i} = \frac{p_i \bar{w}_i}{\bar{w}} \qquad 0 \le i, j \le m$$

where

$$\bar{w} = \mathsf{E}[w_{\mathbf{n}}] = \sum_i p_i \sum_j p_j w_{i+j} = \sum_{i,j} p_i p_j w_{i+j}$$

is the mean fitness of the diploid population. Selection may be included in the master equations for TE load (Eqs. 20–22) as follows,

$$\frac{dp_n}{dt} = \alpha(p_n' - p_n) + \cdots = \alpha\left(\frac{p_n \bar{w}_n}{\bar{w}} - p_n\right) + \cdots = \alpha\frac{p_n(\bar{w}_n - \bar{w})}{\bar{w}} + \cdots$$

where for typographical convenience we do not write the reaction terms involving u_n and v_n (these are indicated by $\cdots$). In the weak selection limit, $w_n = (1-s)^n \approx 1 - sn$ and the mean fitness $\bar{w}$ becomes

$$\bar{w} = \sum_n p_n \bar{w}_n \approx \sum_n p_n[1 - sn - s\mu_1] = 1 - s\mu_1 - s\sum_n np_n = 1 - 2s\mu_1 .$$

Thus, weak selection can be included in the moment equations as follows

$$\frac{dp_n}{dt} = \alpha p_n \frac{1 - sn - s\mu_1 - (1 - 2s\mu_1)}{1 - 2s\mu_1} + \cdots = \frac{\alpha s}{1 - 2s\mu_1} p_n(\mu_1 - n) + \cdots .$$

This expression leads to the following differential equation for the first moment in the weak selection limit,

$$\frac{d\mu_1}{dt} = \sum_n n\frac{dp_n}{dt} = \frac{\alpha s}{1 - 2s\mu_1}\left(\mu_1 \sum_n np_n - \sum_n n^2 p_n\right) + V_1 + U_1 \tag{S23}$$

$$= -\frac{\alpha s}{1 - 2s\mu_1}\left[\mu_2 - \mu_1^2\right] + V_1 + U_1 , \tag{S24}$$

where the quantity in brackets is the variance in haploid load ($\hat{\mu}_2 = \mu_2 - \mu_1^2$) and V_1 and U_1 are given by Eqs. S6 and S7. A similar calculation gives the dynamics of the second moment of the haploid load, moment

$$\begin{aligned}\frac{d\mu_2}{dt} &= \sum_n n^2 \frac{dp_n}{dt} = \frac{\alpha s}{1-2s\mu_1}\left(\mu_1 \sum_n n^2 p_n - \sum_n n^3 p_n\right) + V_2 + U_2 \\ &= -\frac{\alpha s}{1-2s\mu_1}\left[\mu_3 - \mu_1\mu_2\right] + V_2 + U_2\,.\end{aligned}$$

where V_2 and U_2 are given by Eqs. S9 and S10. Using Eq. S18, an ODE for the variance in haploid load is found,

$$\begin{aligned}\frac{d\hat{\mu}_2}{dt} &= \frac{d\mu_2}{dt} - 2\mu_1\frac{d\mu_1}{dt} \\ &= -\frac{\alpha s}{1-2s\mu_1}\left[(\mu_3 - \mu_1\mu_2) - 2\mu_1(\mu_2 - \mu_1^2)\right] + \cdots \\ &= -\frac{\alpha s}{1-2s\mu_1}\left[\mu_3 - 3\mu_1\mu_2 + 2\mu_1^3\right] + \cdots \\ &= -\frac{\alpha s}{1-2s\mu_1}\hat{\mu}_3 + \cdots\,.\end{aligned}$$

Using $\bar{n} = 2\mu_1$ and $d\bar{n}/dt = 2d\mu_1/dt$, and $\sigma_n^2/2 = \mu_2 - \mu_1^2$, we see that Eq. S24 is equivalent to

$$\frac{d\bar{n}}{dt} = -\frac{\alpha s}{1-s\bar{n}}\sigma_n^2 + \cdots\,.$$

Using $\mathsf{E}[(\mathbf{n}-\bar{n})^3] = 2\hat{\mu}_3$ and $\sigma_n^2/2 = \hat{\mu}_2$, we obtain

$$\frac{d\sigma_n^2}{dt} = -\frac{\alpha s}{1-s\bar{n}}\mathsf{E}[(\mathbf{n}-\bar{n})^3] + \cdots\,.$$

Combining these results for the effect of selection with the reaction terms of the neutral model (Eqs. S21 and S22), we obtain the following equations for the mean and variance of diploid load under the influence of selection:

$$\begin{aligned}\frac{d\bar{n}}{dt} &= -\frac{\alpha s}{1-s\bar{n}}\cdot\sigma_n^2 + 2\eta_0 - \left(\nu - \eta + \frac{\eta_0}{m}\right)\bar{n} - \frac{\eta}{m}\left(\sigma_n^2 + \frac{\bar{n}^2}{2}\right) && \text{(S25)} \\ \frac{d\sigma_n^2}{dt} &= -\frac{\alpha s}{1-s\bar{n}}\cdot\mathsf{E}[(\mathbf{n}-\bar{n})^3] + 2\eta_0 + \left(\nu + \eta - \frac{\eta_0}{m}\right)\bar{n} - 2\left(\nu - \eta + \frac{\eta_0 + \eta/2}{m}\right)\sigma_n^2 \\ &- \frac{2\eta}{m}\left(\bar{n}\sigma_n^2 + \frac{\bar{n}^2}{4} + \mathsf{E}[(\mathbf{n}-\bar{n})^3]\right)\,. && \text{(S26)}\end{aligned}$$

Taking the limit of Eqs. S25–S26 as $m \to \infty$ gives Eqs. 35–36.

S4 Moment closure

To analyze solutions of Eqs. S25–S26 without assuming that m is large or η is zero, the dependence of $d\sigma_n^2/dt$ on $\mathsf{E}[(\mathbf{n}-\bar{n})^3]$ (equivalently, the dependence of $d\hat{\mu}_2/dt$ on $\hat{\mu}_3$) must be accounted for. This is accomplished using the technique of moment closure, whereby we assume an algebraic relationship of the form $\mu_3 = \psi(\mu_2, \mu_1)$ or

$$\hat{\mu}_3 = \psi(\hat{\mu}_2, \mu_1)\,. \tag{S27}$$

One way to motivate a particular choice of algebraic relationship ψ is to select a distribution with properties similar to those exhibited by the master equation simulations. Next, one derives the relation Eq. S27 that would be exact if the model truly exhibited the selected distribution.

S4.1 Negative binomial closure

One possibility we have investigated is the negative binomial distribution. This choice is motivated by a few key properties. First, the negative binomial distribution is supported on the non-negative integers. The probability mass function for a negative binomial random variable, $\mathbf{X} \sim \mathrm{NB}(r,p)$ for $r > 0$ and $p \in [0,1]$, is

$$\mathsf{P}[\mathbf{X} = k] = \binom{k+r-1}{k} p^r (1-p)^k\,, \quad k \in \{0,1,2,3,\ldots\}\,.$$

Second, overdispersion is a propery of the the negative binomial distribution. The mean $\mu_1 = E[\mathbf{X}]$, variance $\hat{\mu}_2 = \mu_2 - \mu_1^2$, and index of dispersion $\mathsf{Fano}[\mathbf{X}] = \hat{\mu}_2/\mu_1$ are given by

$$\mu_1 = \frac{r(1-p)}{p}\,, \quad \hat{\mu}_2 = \frac{r(1-p)}{p^2}\,, \quad \frac{\hat{\mu}_2}{\mu_1} = \frac{1}{p} \geq 1\,.$$

The third central moment is

$$\hat{\mu}_3 = \frac{r(2-3p+p^2)}{p^3} = \frac{r(p-1)(p-2)}{p^3}\,. \tag{S28}$$

Inverting the above expressions to give

$$p = \frac{\mu_1}{\hat{\mu}_2} \qquad r = \frac{\mu_1^2}{\hat{\mu}_2 - \mu_1}\,.$$

Substituting into Eq. S28 gives

$$\hat{\mu}_3 = \frac{2(\hat{\mu}_2)^2}{\mu_1} - \hat{\mu}_2 = \hat{\mu}_2\left(\frac{2\hat{\mu}_2 - \mu_1}{\mu_1}\right) =: \psi_{NB}(\hat{\mu}_2, \mu_1)\,. \tag{S29}$$

The corresponding expression for the third central moment of the diploid load is

$$\mathsf{E}[(\mathbf{n} - \bar{n})^3] = \sigma_n^2\left(\frac{2\sigma_n^2 - \bar{n}}{\bar{n}}\right)\,. \tag{S30}$$

S4.2 Negative binomial closure: analysis of the weak selection limit

When Eqs. 35 and 36 are modified consistent with the negative binomial moment closure (Section S4.1), we obtain the closed system,

$$\begin{aligned}
\frac{d\bar{n}}{dt} &= -\frac{\alpha s}{1-s\bar{n}} \cdot \sigma_n^2 + 2\eta_0 - (\nu - \eta)\bar{n} \\
\frac{d\sigma_n^2}{dt} &= -\frac{\alpha s}{1-s\bar{n}} \cdot \sigma_n^2\left(\frac{2\sigma_n^2 - \bar{n}}{\bar{n}}\right) + 2\eta_0 + (\nu + \eta)\bar{n} - 2(\nu - \eta)\sigma_n^2\,.
\end{aligned}$$

Setting the left sides to zero and clearing the denominators gives

$$\begin{aligned} 0 &= -\alpha s\sigma_n^2 + \left[2\eta_0 - (\nu-\eta)\bar{n}\right]\left[1 - s\bar{n}\right] \\ 0 &= -\alpha s\sigma_n^2\left(2\sigma_n^2 - \bar{n}\right) + \left[2\eta_0 + (\nu+\eta)\bar{n} - 2(\nu-\eta)\sigma_n^2\right]\bar{n}\left[1 - s\bar{n}\right] . \end{aligned}$$

Assuming asymptotic expansions of the form $\bar{n} = \bar{n}_0 + s\bar{n}_1 + \cdots$ and $\sigma_n^2 = \sigma_0^2 + s\sigma_1^2 + \cdots$, we obtain

$$\begin{aligned} 0 &= -\alpha s(\sigma_0^2 + s\sigma_1^2 + \cdots) + [2\eta_0 - (\nu-\eta)(\bar{n}_0 + s\bar{n}_1 + \cdots)][1 - s(\bar{n}_0 + s\bar{n}_1 + \cdots)] \\ 0 &= -\alpha s(\sigma_0^2 + s\sigma_1^2 + \cdots)[2(\sigma_0^2 + s\sigma_1^2 + \cdots) - (\bar{n}_0 + s\bar{n}_1 + \cdots)] \\ &+ [2\eta_0 + (\nu+\eta)(\bar{n}_0 + s\bar{n}_1 + \cdots) - 2(\nu-\eta)(\sigma_0^2 + s\sigma_1^2 + \cdots)][\bar{n}_0 + s\bar{n}_1 + \cdots][1 - s(\bar{n}_0 + s\bar{n}_1 + \cdots)] . \end{aligned}$$

The zeroth order equations are

$$\begin{aligned} 0 &= -\left[2\eta_0 - (\nu-\eta)\bar{n}_0\right] \\ 0 &= \left[2\eta_0 + (\nu+\eta)\bar{n}_0 - 2(\nu-\eta)\sigma_0^2\right]\bar{n}_0 . \end{aligned}$$

Assuming $\bar{n}_0 > 0$, we find

$$\begin{aligned} \bar{n}_0 &= 2\eta_0/(\nu-\eta) \\ \sigma_0^2 &= \left[2\eta_0 + (\nu+\eta)\bar{n}_0\right] / \left[2(\nu-\eta)\right] = 2\eta_0\nu/(\nu-\eta)^2 , \end{aligned}$$

consistent with the neutral model (Eqs. 27 and 28). The first-order equations are

$$\begin{aligned} 0 &= -\alpha\sigma_0^2 - (\nu-\eta)\bar{n}_1 - [2\eta_0 - (\nu-\eta)\bar{n}_0]\bar{n}_0 \\ 0 &= -\alpha\sigma_0^2[2\sigma_0^2 - \bar{n}_0] + [(\nu+\eta)\bar{n}_1 - 2(\nu-\eta)\sigma_1^2]\bar{n}_0 + [2\eta_0 + (\nu+\eta)\bar{n}_0 - 2(\nu-\eta)\sigma_0^2][\bar{n}_1 - \bar{n}_0^2] . \end{aligned}$$

In the first equation, the expression in brackets evaluates to zero, so

$$\bar{n}_1 = -\alpha\sigma_0^2/(\nu-\eta) = -2\alpha\eta_0\nu/(\nu-\eta)^3$$

After some algebra, we find that the second equation yields,

$$\sigma_1^2 = -\frac{2\alpha\nu\eta_0\,(\nu+\eta)}{(\nu-\eta)^4} .$$

The above expressions may be combined to form the following two-term approximation for the mean and variance of TE load,

$$\bar{n} \approx \frac{2\eta_0}{\nu-\eta} - \frac{2\alpha s\nu\eta_0}{(\nu-\eta)^3} = \frac{2\eta_0}{\nu-\eta}\left[1 - \alpha s\frac{\nu}{(\nu-\eta)^2}\right] .$$

$$\sigma_n^2 \approx \frac{2\nu\eta_0}{(\nu-\eta)^2} - \frac{2\alpha s\nu\eta_0\,(\nu+\eta)}{(\nu-\eta)^4} = \frac{2\nu\eta_0}{(\nu-\eta)^2}\left[1 - \alpha s\frac{(\nu+\eta)}{(\nu-\eta)^2}\right] .$$

Because $v/(\nu-\eta)^2 > 0$, the equation for $\bar{n}$ indicates that weak selection decreases the mean TE load, consistent with our intuition. In the equation for σ_n^2, the factor $(\nu+\eta)/(\nu-\eta)^2$ is positive, so we conclude that weak selection also decreases the population variance. As for the index of dispersion, this analysis indicates that under weak selection the Fano factor is well-approximated by

$$\frac{\sigma_n^2}{\bar{n}} \approx \frac{\nu}{\nu-\eta} \cdot \frac{1-\alpha s(\nu+\eta)/(\nu-\eta)^2}{1-\alpha s\nu/(\nu-\eta)^2} = \frac{\nu}{\nu-\eta}\left[1-\alpha s\frac{\eta}{(\nu-\eta)^2}\right] .$$

That is, under weak selection, the Fano factor is expected to decrease, because weak selection causes variance to decrease more than the mean.

S4.3 Beta-binomial closure

For comparison to the negative binomial closure, we have worked through the possibility of choosing ψ to be the function that would be correct if the actual distribution of TE loads were beta-binomial distributed. The probability mass function for a beta-binomial random variable, $\mathbf{X} \sim \mathrm{BB}(\alpha,\beta)$ for $\alpha > 0$ and $\beta > 0$ on the interval 0 to m is

$$\mathsf{P}[\mathbf{X}=k] = \binom{m}{k}\frac{B(k+\alpha, m-k+\beta)}{B(\alpha,\beta)} , \quad k \in \{0,1,\ldots,m\} . \tag{S31}$$

where $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt$ is the beta function. The first three raw moments of a beta-binomial random variable are

$$\begin{aligned}
\mu_1 &= \frac{m\alpha}{\alpha+\beta} \\
\mu_2 &= \frac{m\alpha[m(1+\alpha)+\beta]}{(\alpha+\beta)(1+\alpha+\beta)} \\
\mu_3 &= \frac{m\alpha[m^2(1+\alpha)(2+\alpha)+3m(1+\alpha)\beta+\beta(\beta-\alpha)]}{(\alpha+\beta)(1+\alpha+\beta)(2+\alpha+\beta)} ,
\end{aligned}$$

while the variance is

$$\hat{\mu}_2 = \mu_2 - \mu_1^2 = \frac{m\alpha\beta(\alpha+\beta+m)}{(\alpha+\beta)^2(\alpha+\beta+1)} .$$

Using Eq. S17, it can be shown that the third central moment of a beta-binomial random variable is

$$\hat{\mu}_3 = \hat{\mu}_2\frac{(\alpha+\beta+2m)(\beta-\alpha)}{(\alpha+\beta)(\alpha+\beta+2)} .$$

Inverting Eqs. S32 and S32 gives

$$\begin{aligned}
\alpha &= \frac{m\mu_1-\mu_2}{m(\mu_2/\mu_1-\mu_1-1)+\mu_1} \\
\beta &= \frac{(m-\mu_1)(m-\mu_2/\mu_1)}{m(\mu_2/\mu_1-\mu_1-1)+\mu_1} .
\end{aligned}$$

From these values we calculate the following algebraic relationship for the third central moment in terms of the mean and variance of the haploid TE load,

$$\hat{\mu}_3 = \hat{\mu}_2 \frac{(m - 2\mu_1)(\mu_1^2 - m\mu_1 - \hat{\mu}_2 + 2m\hat{\mu}_2)}{m\mu_1(m - \mu_1 - 2) + \hat{\mu}_2 m + 2\mu_1^2} =: \psi_{BB}(\hat{\mu}_2, \mu_1)\,. \tag{S32}$$

For the mean and variance of the diploid TE load, the corresponding expression is

$$\mathsf{E}[(\mathbf{n} - \bar{n})^3] = 2\hat{\mu}_3 = \sigma^2 \frac{(m - \bar{n})(\bar{n}^2 - 2m\bar{n} - 2\sigma^2 + 4m\sigma^2)}{m\bar{n}(2m - \bar{n} - 4) + 2m\sigma^2 + 2\bar{n}^2}\,. \tag{S33}$$

This beta-binomial closure represented by this expression for the third central moment of diploid TE load is arguably preferable to the negative binomial closure discussed in Section S4.1, because a beta-binomial random variable has finite support (values between 0 and m as in Eq. S31). On the other hand, the negative binomial closure results in a simpler expression that often gives approximately the same nullclines and solution trajectories (Fig. 5). It is notable that the expression for $\mathsf{E}[(\mathbf{n} - \bar{n})^3]$ obtained using the beta-binomial distribution (Eq. S33) is well-approximated by the negative binomial result (Eq. S30) when the number of loci are not limiting ($\bar{n} << m$). To see this, one may compare Eqs. S29 and S32 and show that $\psi_{BB}(\hat{\mu}_2, \mu_1) \to \psi_{NB}(\hat{\mu}_2, \mu_1)$ as $m \to \infty$.

S5 Numerical methods

The master equation model given by Eq. 31 is a system of $m + 1$ ordinary differential equations. When m is sufficiently small, it is straightforward to use a relaxation method to calculate the limiting probability distributions for the master equation. Because the number of ODEs in the master equation grows with the number of occupiable loci, it can be more efficient, especially when m is large, to numerically solve for the limiting probability distribution of the associated Fokker-Planck equation. Writing $\rho(n,t)\,dn = \mathsf{Pr}[n \le \mathbf{n} \le n + dn]$ for the time-dependent probability density function for TE load, $\rho(n,t)$ solves the following Fokker-Planck equation (Gardiner, 2009; Van Kampen, 2007),

$$\frac{\partial \rho}{\partial t} = -\frac{\partial}{\partial n}\left[a(n)\rho\right] + \frac{1}{2}\frac{\partial^2}{\partial n^2}\left[b(n)\rho\right]\,. \tag{S34}$$

In these expressions, $\mathbf{n}$ is the random variable (TE load of a randomly sampled haploid genome) and n is the independent variable of the probability density $\rho(n,t)$. The drift and diffusion terms of the Fokker-Planck equation are

$$a(n) \quad = \quad -v(n) + u(n) = \eta_0 - \left(\nu - \eta + \frac{\eta_0}{m}\right) n - \frac{\eta}{m} n^2 \tag{S35}$$

$$b(n) \quad = \quad v(n) + u(n) = \eta_0 + \left(\nu + \eta - \frac{\eta_0}{m}\right) n - \frac{\eta}{m} n^2 \tag{S36}$$

where we have used the gain and loss terms given by $u(n) = (\eta_0 + \eta n)(1 - n/m)$ and $v(n) = \nu n$ (Eqs. 17 and 19). Writing Eq. S34 in conservative form as $\partial\rho/\partial t = -\partial\phi/\partial n$, where $\phi(n)$ is the probability flux, we find

$$\phi(n) = a(n)\rho - \frac{1}{2}\frac{\partial}{\partial n}[b(n)\rho]\,. \tag{S37}$$

For a steady-state solution $\hat{\rho}(n)$ with no flux boundary conditions, setting $\phi(n) = 0$ leads to the analytical solution

$$\hat{\rho}(n) = \frac{\theta}{b}\exp{(2U)}\ , \tag{S38}$$

where $b(n)$ is given by Eq. S36, and θ is a normalization constant such that $\int \hat{\rho}(n)\,dn = 1$ and

$$U(n) = \int_0^n \frac{a(n')}{b(n')}\,dn'\,.$$

In fact, $U(n)$ may be any antiderivative satisfying $U' = a/b$, because the normalization constant θ absorbs the arbitrary constant of integration.

Several numerical methods were used to simulate the models of TE population dynamics defined by the master equation for the neutral model (Eqs. 20–22) and the master equation that accounts for selection (Eq. 31). Fig. 2 used a flux-limiting numerical scheme and the method of lines to integrate the Fokker-Planck equation (Eq. S34) until a limiting value was reached. Fig. 3 was obtained using the analytical steady state of the Fokker-Planck equation (Eq. S38). Fig. 4 used Monte Carlo simulation of drift-diffusion process associated to the Fokker-Planck equation (Eq. S34). Comparing these results to the flux-limiting numerical scheme revealed that, for some parameter sets that included strong selection, the infinite population model exhibits periodic solutions that are rarely observed in large finite populations. Fig. 5 was calculated using the moment equations with selection and negative binomial moment closure (Eqs. 42 and 43). Beta-binomial moment closure leads to very similar results unless m is quite small (on the order of 100). Fig. 6 used moment equations with selection and Beta-binomial closure (Eqs. 33 and 34 with Eq. 40).