the geneticssociety

www.nature.com/hdy

ARTICLE

Check for updates

The evolution of tandem repeat sequences under partial selfing and different modes of selection

Vitor Sudbrack[1]✉ and Charles Mullon[1]

Tandem repeat (TR) sequences occur when short DNA motifs are repeated head-to-tail along chromosomes and are a major source of genetic variation. Population genetic models of TR evolution have focused on large, randomly mating, haploid populations. Yet many organisms reproduce partially through self-fertilisation ('selfing'), which increases homozygosity and thus may alter the evolutionary processes shaping TR sequences. Here we use mathematical modelling and simulations to study the evolution of homologous TR sequences in partially selfing, diploid populations under four different selective regimes that may be relevant to TRs: (i) additive purifying selection, (ii) truncation-like purifying selection, (iii) selection against heterozygotes due to misalignment costs, and (iv) stabilising selection favouring an intermediate TR sequence length. We show that selfing influences TR evolution primarily by increasing homozygosity, with two main consequences: (1) it enhances the variation produced by unequal recombination within individuals, and (2) it increases variation between individuals. Consequently, selection on TRs becomes more effective under partial selfing across all modes of selection considered, resulting in lower genetic load, despite higher genetic drift. Overall, our results suggest that mating systems and inbreeding are important factors shaping variation in TR sequences.

Heredity (2026) 135:99–112; https://doi.org/10.1038/s41437-026-00820-1

INTRODUCTION

Tandem repeat (TR) sequences consist of short genomic motifs, varying between 2 and 2000 base pairs, that are repeated head-to-tail (e.g. GAAGAAGAA...) from a dozen to thousands of times, spanning up to millions of base pairs in some cases (Depienne and Mandel 2021; Miklos and Gill 1982). TRs are ubiquitous across the tree of life (Verbiest et al. 2023), ranging from microsatellites (1–6 base pair motifs) and minisatellites (larger motifs) to very long satellite arrays (Buschiazzo and Gemmell 2006; Jeffreys et al. 1985; Richard et al. 2008). Short TR sequences, or 'simple sequence repeats', are a major source of genetic variation because individuals often differ markedly in repeat number. This polymorphism ('TR copy number', 'variable number of TRs', or 'TR sequence length') underlies their widespread use as genetic markers (e.g. in population genetics, paternity testing and linkage mapping; Ellegren 2004; Hammond et al. 1994; Slatkin 1995; Verbiest et al. 2023). For a long time, TR sequences have been considered part of 'junk DNA', moulded either by mutational processes alone or in combination with purifying selection (Charlesworth et al. 1994; Kruglyak et al. 1998, 2000; McGinty et al. 2025). Recent studies, however, have revealed that at least some of these sequences may have functional significance and therefore potentially non-straightforward fitness consequences. Understanding the processes that shape variation in TR sequences is therefore useful for both practical and fundamental reasons (Gymrek and Goren 2021).

Population genetic models have been useful to disentangle the effects of selection, mutation, recombination and genetic drift on the evolution of TRs (Ohta 1983b; Stephan 1986, 1987, 1989; Walsh 1987; see also Crow and Kimura 1970, p. 294–296; Krüger and Vogel 1975; Ohta 1983a; Takahata 1981 for models of chromosome size or gene content evolution that involve similar processes). This body of work has highlighted how the particular mechanisms of mutation and recombination affecting TRs contribute to variation in TR sequence length. Specifically, the repeated structure of these sequences makes them subject to replication slippage (a mutational process by which TR sequences gain or lose multiple motifs at once, in contrast to indels; Fan and Chu 2007) and unequal recombination (redistribution of motifs among gametes due to the misalignment between homologous TR sequences during crossover; Smith 1976). While replication slippage is key to the maintenance of TRs within populations (Ohta 1983b; Stephan 1987), unequal recombination tends to increase variation in copy number (Stephan 1986). In turn, greater variation leads to more efficient selection against TR copies in the face of genetic drift (Stephan 1987). The balance between these processes depends on motif length and genomic location, with unequal recombination particularly common in longer arrays and in regions of high recombination (Ellegren 2004; Stephan 1987).

Here, we extend current theory in two main directions, incorporating factors that existing TR models have not included. First, whereas existing models assume that individuals mate at random, we consider the common form of non-random mating that is partial selfing. Selfing leads to excess homozygosity, thereby affecting selection, drift and recombination (Burgarella and Glémin 2017). Because unequal recombination depends on the variance between homologous copies, increased homozygosity is expected to modify TR dynamics. Empirical patterns are

[1]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland. Associate editor: Lindsey Compton. ✉email: vitorsudbrack@gmail.com

Received: 4 July 2025 Revised: 5 January 2026 Accepted: 5 January 2026
Published online: 29 January 2026

consistent with this possibility: in *Arabidopsis thaliana*, where selfing rates regularly exceed 75–95% (Abbott and Gomes 1989), microsatellite content is lower than in closely related self-incompatible species (Clauss et al. 2002). Similarly, in nematodes, the partially selfing *Caenorhabditis briggsae* shows reduced TR content compared with its outcrossing relative *C. nigoni* (Subirana and Messeguer 2017). However, current models cannot account for differences that depend on the mating system.

Second, we move beyond the standard assumption that TR sequences experience only purifying selection. This matters because TRs are increasingly recognised for their potential roles in complex traits and phenotypic variation, as highlighted by recent genome-wide association studies (Gymrek and Goren 2021; reviewed in Depienne and Mandel 2021 for humans and in Verbiest et al. 2023 across various taxa; see Sureshkumar et al. 2025 for examples in plants). TRs can also influence chromosomal stability, as some TR sequences protect gene ends, prevent chromosomal erosion and delay cell senescence (Biscotti et al. 2015; Greider 1990; Ide et al. 2010). It remains unclear, however, how these various forms of function for TRs translate into selection pressures and influence their evolution.

Motivated by these observations, we consider several modes of selection acting at the diploid stage rather than assuming that selection on TR sequence length acts at the haploid gametic stage and is purifying with additive effects (i.e. each additional repeat has the same fitness effect, but see Stephan 1987 for a model of truncation-selection): (i) non-additive effects in TR sequence length, in line with the observation that in some cases the onset of disease is determined by an excess of repeats beyond a threshold (e.g. Fragile X syndrome, which is related to CGG expansions, Depienne and Mandel 2021; Usdin et al. 2015); (ii) interactions between homologous TR sequences within individuals, motivated by studies showing that the lack of TR homology can compromise chromosomal stability (for instance by inducing chromosomal loops or supercoiling, John and Miklos 1979; Usdin et al. 2015; Verbiest et al. 2023 or by increasing recombinational instability, Jarne and Lagoda 1996); and (iii) stabilising selection for an optimal TR sequence length, based on eQTL and mQTL studies that have demonstrated that some TRs can influence gene expression and regulation, and ultimately contribute to phenotypic variation (Fotsing et al. 2019; Gymrek et al. 2016; Quilez et al. 2016; Reinar et al. 2021; Sureshkumar et al. 2025; Verbiest et al. 2023; Zhang et al. 2025). In such cases, it is conceivable that an optimal level of gene expression may lead to an optimal sequence length, placing TRs under stabilising selection around that length.

Our aim is to determine how partial selfing interacts with these different modes of selection to shape TR variation at the mutation-selection equilibrium, considering the unique forms of mutation and recombination experienced by TR sequences.

MODEL

Life cycle, trait and its distribution

We consider a population of diploid hermaphrodites of constant size N with the following life cycle (Fig. 1A; Table 1 for a list of symbols): (1) Each adult produces a large number of gametes according to its fecundity and then dies. (2) Gametes fuse together to form zygotes. With probability α, a zygote is produced by combining two gametes of the same individual, or with complementary probability $1-\alpha$, of two different individuals, so that the parameter α is the selfing rate. (3) Zygotes compete randomly to form the N adults of the next generation.

Each individual $i \in \{1, \ldots, N\}$ is characterised by two positive integers, $z_{i1} \geq 1$ and $z_{i2} \geq 1$, which are the lengths of a focal TR sequence on the paternally and maternally-inherited chromosomes, respectively (Fig. 1A for illustration). Our goal is to investigate the effect of selfing rate α on the evolution and polymorphism of this TR sequence. To do so, we will track the evolutionary dynamics of the population mean,

$$\bar{z} = \frac{1}{2N}\sum_{i=1}^{N}(z_{i1} + z_{i2}), \tag{1}$$

and variance,

$$\begin{aligned}\sigma_{\mathrm{T}}^2 &= \frac{1}{2N}\sum_{i=1}^{N}\left[(z_{i1})^2 + (z_{i2})^2\right] - \bar{z}^2 \\ &= \underbrace{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{z_{i1} - z_{i2}}{2}\right)^2}_{\sigma_{\mathrm{W}}^2} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{z_{i1} + z_{i2}}{2} - \bar{z}\right)^2}_{\sigma_{\mathrm{B}}^2},\end{aligned} \tag{2}$$

which we decomposed as the sum between the variance within (σ_{W}^2) and between (σ_{B}^2) individuals (SI Text A.1 for details). This decomposition of variance allows us to define the *F*-statistic (also referred to as the *R*-statistic in the context of microsatellites, see Eq. (13) in Slatkin 1995) as

$$F_{\mathrm{IS}} = \frac{\sigma_{\mathrm{T}}^2 - 2\sigma_{\mathrm{W}}^2}{\sigma_{\mathrm{T}}^2}, \tag{3}$$

such that

$$\sigma_{\mathrm{B}}^2 = \left(\frac{1 + F_{\mathrm{IS}}}{2}\right)\sigma_{\mathrm{T}}^2 \text{ and } \sigma_{\mathrm{W}}^2 = \left(\frac{1 - F_{\mathrm{IS}}}{2}\right)\sigma_{\mathrm{T}}^2. \tag{4}$$

F_{IS} is a measure of excess homozygosity relative to random mating. When $F_{\mathrm{IS}} = 0$, this indicates random mating (i.e. full outcrossing) where the variance within individuals is equal to the variance between individuals. In contrast, any form of non-random mating that increases homozygosity, such as partial selfing, will lead to ($\sigma_{\mathrm{W}}^2 < \sigma_{\mathrm{B}}^2$) and thus $F_{\mathrm{IS}} > 0$.

In general, the expected change $\Delta\bar{z}$ in mean TR sequence length $\bar{z}$ over one generation can be expressed as

$$\Delta\bar{z} = \frac{1}{2}\mathrm{E}\left[w_i\left(z_i^{\mathrm{G}} - \bar{z}\right)\right], \tag{5}$$

where z_i^{G} denotes the expected TR sequence length in a gamete of individual i, and w_i denotes the expected number of gametes of individual i that are recruited in the next generation. Similarly, the expected change in total variance $\Delta\sigma_{\mathrm{T}}^2$ reads as

$$\Delta\sigma_{\mathrm{T}}^2 = \frac{1}{2}\mathrm{E}\left[w_i\left(\sigma_i^{2\mathrm{G}} - \sigma_{\mathrm{T}}^2\right)\right] + \frac{1}{2}\mathrm{E}\left[w_i\left(z_i^{\mathrm{G}}\right)^2\right] - \frac{1}{4}\mathrm{E}\left[w_i z_i^{\mathrm{G}}\right]^2, \tag{6}$$

where $\sigma_i^{2\mathrm{G}}$ is the variance in sequence length among gametes of individual i. In both Eqs. (5) and (6), the expectation $\mathrm{E}[\cdot]$ is taken over all individuals i and all events that occur in a full iteration of the life cycle (SI Texts A.2 and A.3 for derivations of Eqs. (5) and (6), respectively).

We assume that the fecundity of an individual depends on the TRs it carries, so that the fecundity f_i of individual i can be written as $f_i = f(z_{i1}, z_{i2})$. Given our assumptions on the life cycle, the (gene) fitness w_i of this individual is given by

$$w_i = 2N\frac{f_i}{\sum_{k=1}^{N} f_k}. \tag{7}$$

We will consider different forms for f_i to reflect different types of selection on TRs, and quantify the impact of TRs on the population by the genetic load

$$L = 1 - \frac{\overline{f_i}}{f_{\max}}, \tag{8}$$

where $\overline{f_i}$ is the average fecundity in the population and $f_{\max}$ is the maximum fecundity calculated when extra repeats are absent.

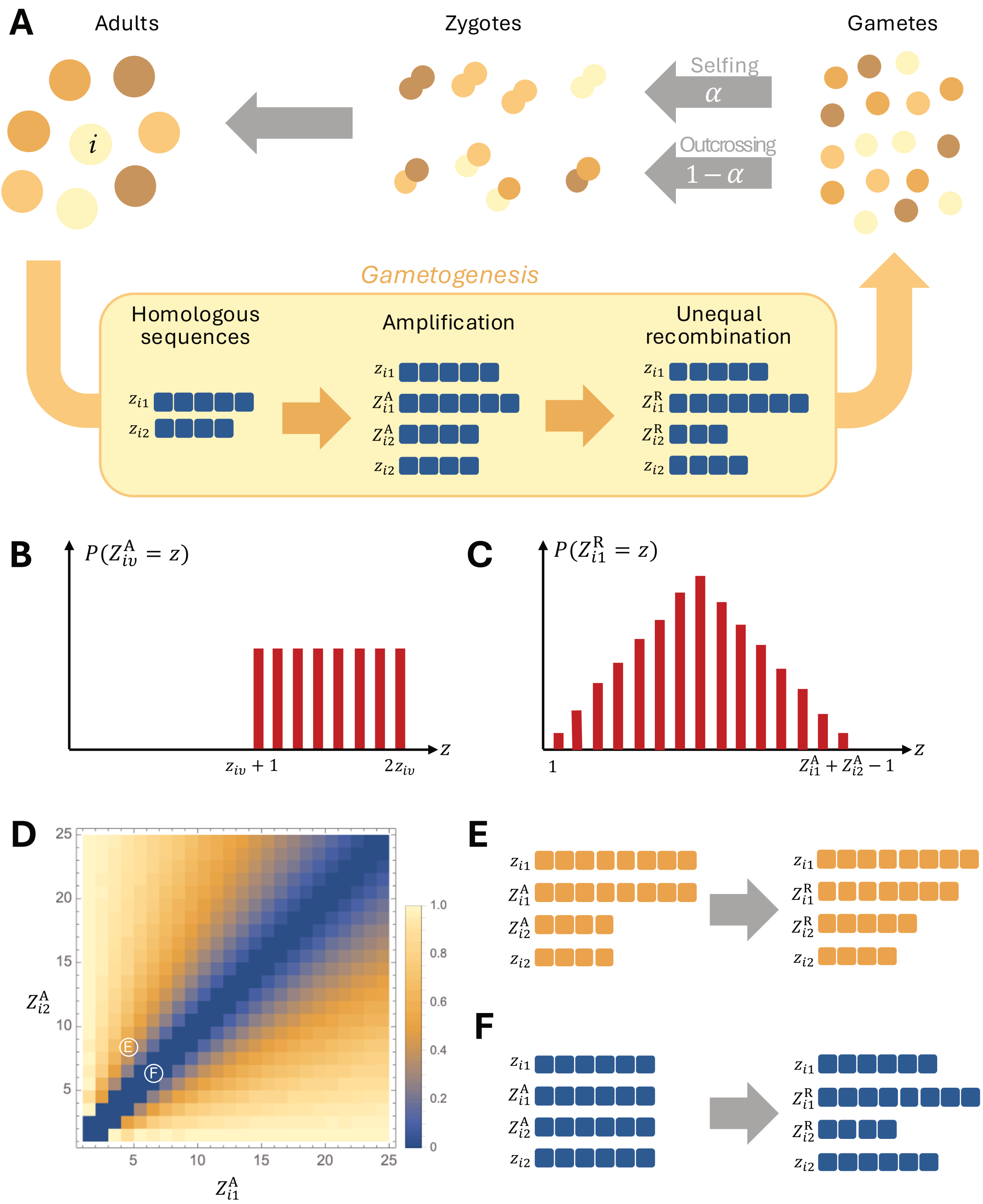


Fig. 1 Life cycle and gametogenesis. A Life cycle described in the section 'Life cycle, trait and its distribution', with a highlighted example of the production of gametes of an adult *i*, as detailed in the section 'Meiosis and gametogenesis'. In this example, $z_{i1}=5$ and $z_{i2}=4$, Z_{i1}^{A} is amplified while Z_{i2}^{A} is not, and unequal recombination between Z_{i1}^{A} and Z_{i2}^{A} takes place. **B** Probability distribution for the TR sequence length after amplification Z_{iv}^{A} given template z_{iv} (Eq. 9). **C** Probability distribution for the TR sequence length after unequal recombination Z_{i1}^{R} given amplified sequences Z_{i1}^{A} and Z_{i2}^{A} (Eq. 10). **D** Probability that unequal recombination decreases differences between gametes, i.e. probability that $\mathrm{P}[\min(Z_{i1}^{A},Z_{i2}^{A})<Z_{i1}^{R}<\max(Z_{i1}^{A},Z_{i2}^{A})|Z_{i1}^{A},Z_{i2}^{A}]$. Examples of both cases, when unequal recombination **E** decreases or **F** increases variance between gametes.

In the next section, we characterise the moments z_{i}^{G} and σ_{i}^{2G} that appear in Eqs. (5) and (6) as a result of meiosis and gametogenesis.

Meiosis and gametogenesis

We assume that during meiosis and gametogenesis, three processes can influence the distribution of TRs among the recruited gametes of an individual: amplification, unequal recombination and Mendelian segregation.

Amplification due to replication slippage. First, the length of a TR sequence may be amplified during interphase due to replication slippage (e.g. Fig. 1A; Buschiazzo and Gemmell 2006; Charlesworth et al. 1994; Levinson and Gutman 1987; see Khristich and Mirkin

Table 1. Summary of variables and parameters of the model.

Variable	Description
At the population level	
N	Population size
α	Selfing rate
F_{IS}	Inbreeding coefficient (Eq. 3)
$\bar{z}$	Mean TR sequence length (Eq. 1)
σ_{T}^2	Variance in TR sequence length (Eq. 2)
σ_{B}^2	Variance in mean TR sequence length between individuals (Eq. 2)
σ_{W}^2	Mean variance in TR sequence length between homologous sequences (within individuals; Eq. 2)
L	Genetic load caused by TRs (Eq. 8)
At the individual level	
z_{i1}, z_{i2}	TR sequence length of the two homologous chromosomes of individual i
f_i	Fecundity of individual i (Eqs. 12, 17, 18 or 19)
w_i	Fitness of individual i (Eq. 7)
z_i^{G}	Mean TR sequence length among gametes of individual i (Eq. 11a)
$\sigma_i^{2\mathrm{G}}$	Variance in TR sequence length among gametes of individual i (Eq. 11b)
At the TR sequence level	
γ	Probability per meiosis of an unequal recombination event between homologous sequences
μ	Probability per chromatid of a slippage replication event
s_{a}	Additive cost of each TR in the sequence (Eqs. 12 and 18)
s_{ϵ}	Strength of non-additive effects in selection (Eq. 17)
θ	Threshold TR sequence length after which fecundity quickly decays (Eq. 17)
s_{d}	Fecundity costs due to misalignment between homologous TR sequences (Eq. 18)
s_{b}	Strength of stabilising selection (Eq. 19)
Θ	Optimal TR sequence length per chromosome under stabilising selection (Eq. 19)

2020 for a deeper discussion on mutational processes in TRs). To model this process, let $Z_{i\nu}^{\mathrm{A}}$ be a random variable for the length of the TR sequence in a focal germ cell after replication on the sister (inner) chromatid whose template is $z_{i\nu}$ (with $\nu \in \{1, 2\}$) in individual i. We assume that amplification takes place independently with probability μ during the replication of each new chromatid, in which case saltatory amplification occurs following the model of Stephan (1987). This model, which is based on the evidence that TR sequence length tends to grow due to mutation in a way that increases with the length of the parental sequence (Buschiazzo and Gemmell 2006; Slatkin 1995), assumes that the TR sequence length in the new inner chromatid $Z_{i\nu}^{\mathrm{A}}$ increases by a random number of TRs that is uniformly distributed between 1 and the template size $z_{i\nu}$ (Fig. 1B). That is, with probability μ, we have

$$Z_{i\nu}^{\mathrm{A}} \overset{\mathrm{law}}{\sim} \mathrm{Unif}(z_{i\nu}+1,\ 2z_{i\nu})\,. \tag{9}$$

With complementary probability $1-\mu$, no amplification happens, so that the new sister chromatid is identical to its template, i.e. $\Pr\left[Z_{i\nu}^{\mathrm{A}} = z_{i\nu}\right] = 1$. We assume that the probability of undergoing replication slippage is independent of the TR sequence length.

Unequal recombination. Following replication, the newly formed inner chromatids may recombine. When chromosome pairing during synapsis is correct, recombination does not affect the length of TR sequences. When, however, pairing is incorrect, which readily occurs owing to the repetitive nature of TRs, unequal crossover causes the number of TRs to be redistributed between chromatids during recombination (Krüger and Vogel 1975; Ohta 1981; Ohta and Kimura 1981; Perelson and Bell 1977; Stephan 1986). To model this process, let us denote by Z_{i1}^{R} and Z_{i2}^{R} the random variables for the TR copy numbers produced by the inner chromatids after recombination. We assume that unequal recombination occurs with probability γ, in which case we have

$$\begin{aligned} &Z_{i1}^{\mathrm{R}} \,|\, Z_{i1}^{\mathrm{A}}, Z_{i2}^{\mathrm{A}} \overset{\mathrm{law}}{\sim} \mathrm{Tri}\left(1, \tfrac{Z_{i1}^{\mathrm{A}}+Z_{i2}^{\mathrm{A}}}{2}, Z_{i1}^{\mathrm{A}}+Z_{i2}^{\mathrm{A}}-1\right) \text{ and} \\ &Z_{i2}^{\mathrm{R}} = Z_{i1}^{\mathrm{A}}+Z_{i2}^{\mathrm{A}}-Z_{i1}^{\mathrm{R}}\,, \end{aligned} \tag{10}$$

where Tri denotes the Triangular distribution (Fig. 1C; Stephan 1986; Takahata 1981). This model assumes that any of the Z_{i1}^{A} sites of the chromatid 1 is equally likely to cross over any of the Z_{i2}^{A} sites of its homologous chromatid during synapsis and that the total number of copies is conserved, i.e. $Z_{i1}^{\mathrm{R}} + Z_{i2}^{\mathrm{R}} = Z_{i1}^{\mathrm{A}} + Z_{i2}^{\mathrm{A}}$(Stephan 1986; Takahata 1981).

Mendelian segregation. Finally, random Mendelian segregation distributes the products of meiosis fairly among the gametes of an individual. To see these effects, let us denote by Z_i^{G} the random TR sequence length in a gamete sampled among all gametes produced by individual i, whose TR sequence lengths are z_{i1} and z_{i2}. The moments of Z_i^{G} that are necessary to our analysis (i.e. that appear in Eqs. 5 and 6) are then given by

$$\begin{aligned} z_i^{\mathrm{G}} &= \mathrm{E}\left[Z_i^{\mathrm{G}} | z_{i1}, z_{i2}\right] \\ &= \tfrac{1}{4}z_{i1} + \tfrac{1}{4}z_{i2} + \tfrac{1}{4}\mathrm{E}\left[Z_{i1}^{\mathrm{R}} | z_{i1}, z_{i2}\right] + \tfrac{1}{4}\mathrm{E}\left[Z_{i2}^{\mathrm{R}} | z_{i1}, z_{i2}\right] \end{aligned} \tag{11a}$$

$$\begin{aligned} \sigma_i^{2\mathrm{G}} &= V\left[Z_i^{\mathrm{G}} | z_{i1}, z_{i2}\right] \\ &= \mathrm{E}\left[\left(Z_i^{\mathrm{G}}\right)^2 | z_{i1}, z_{i2}\right] - \left(z_i^{\mathrm{G}}\right)^2 \\ &= \tfrac{1}{4}(z_{i1})^2 + \tfrac{1}{4}(z_{i2})^2 + \tfrac{1}{4}\mathrm{E}\left[{Z_{i1}^{\mathrm{R}}}^2 | z_{i1}, z_{i2}\right] + \tfrac{1}{4}\mathrm{E}\left[{Z_{i2}^{\mathrm{R}}}^2 | z_{i1}, z_{i2}\right] - \left(z_i^{\mathrm{G}}\right)^2. \end{aligned} \tag{11b}$$

In both equations above, the first two terms correspond to the outer chromatids, while the next two terms represent the inner chromatids, which undergo amplification and unequal recombination.

Analyses

We investigate the evolution of TRs under different selection regimes with two complementary approaches. First, we analyse the change in mean and variance within the population across generations mathematically. We do this by computing the gametic moments in Eqs. (11a) and (11b) using the distributions given in Eqs. (9) and (10) (details in SI Text A.4), and then substituting these moments into Eqs. (5) and (6). Second, to validate these analyses and extend them, we perform individual-based simulations using SLiM v.4.3, implementing the life cycle and gametogenesis described above, with further details of the implementation available in SI Text B (Haller and Messer 2023).

Using these approaches, we vary the selfing rate α to quantify how mating systems influence both the evolutionary dynamics and the standing variation of TR sequence length. We also examine how different modes of selection shape TR sequence evolution by varying the fecundity function f_i and assessing how these effects interact with partial selfing.

In addition, we explore how the relative rates of amplification μ and unequal recombination γ influence TR sequence evolution. These rates allow us to model different classes of short-motif TR arrays: replication slippage is thought to dominate the evolution of short arrays such as microsatellites (Ellegren 2004; Levinson and Gutman 1987; Strand et al. 1993), whereas unequal recombination becomes increasingly important in larger arrays such as minisatellites or short-motif satellites (Ellegren 2004; Subirana and Messeguer 2017). Varying the parameter γ also enables us to contrast genomic regions with low recombination (e.g. heterochromatin) and regions with high recombination (e.g. euchromatin, as discussed in Stephan 1987).

RESULTS

Shorter TR sequences under purifying selection in selfing populations

As a baseline, we assume that TR sequences are under purifying selection with additive effects of repeats, e.g. owing to the time and energy cells invest in replicating TRs in the genome (Buschiazzo and Gemmell 2006; Charlesworth et al. 1994; Stephan 1986 1987; Verbiest et al. 2023). Several lines of evidence are consistent with purifying selection acting on TR sequences. In *Daphnia* for example, TR copy numbers are higher in mutation-accumulation lines than in isolated natural populations, suggesting that selection opposes repeat accumulation (Flynn et al. 2017). We assume that the fecundity f_i of a focal individual i carrying sequences of lengths z_{i1} and z_{i2} is

$$f_i = 1 - s_a(z_{i1} + z_{i2}), \tag{12}$$

where s_a tunes the strength of purifying selection.

Because the expressions of the population mean and variance (Eqs. 5 and 6) are too complicated to have an analytical form in general, we first assume that selection and amplification are weak ($s_a \sim \mathcal{O}(\delta)$ and $\mu \sim \mathcal{O}(\delta)$ where δ is small parameter) and that the population is large ($N \sim \mathcal{O}(1/\delta)$), obtaining

$$\Delta\bar{z} = \underbrace{\frac{1}{4}\mu(1+\bar{z})}_{\text{amplification}} - \underbrace{2s_a\sigma_B^2}_{\substack{\text{purifying}\\ \text{selection}}} + \mathcal{O}(\delta^2) \tag{13a}$$

$$\Delta\sigma_T^2 = \underbrace{-\frac{1}{2}\gamma\sigma_W^2}_{\substack{\text{homogenising}\\ \text{effect}}} + \underbrace{\frac{1}{12}\gamma\left[\bar{z}^2 + \sigma_B^2 - 1\right]}_{\substack{\text{reshuffling}\\ \text{effect}}} + \mathcal{O}(\delta) \tag{13b}$$

for the dynamics of the mean and variance in TR sequence length (SI Text A.6 for derivation). The two terms of Eq. (13a) highlight how the change in TR sequence length depends on a balance between: (i) amplification, whose effects are proportional to the mean $\bar{z}$ due to its saltatory nature (i.e. larger sequences gain on average more TR copies); and (ii) purifying selection, whose effects are proportional to the variance between individuals σ_B^2 as selection takes place among adults. The two terms of Eq. (13b), meanwhile, reflect how unequal recombination can either decrease or increase the variance in TR sequence length, depending on how the variance in TR sequence length is distributed within and between individuals (Fig. 1D). If the variance within individuals σ_W^2 is large (compared to σ_B^2), then the first term of Eq. (13b), which is negative, tends to dominate, indicating that recombination tends to reduce total variance. This is because when recombination takes place among homologous TR sequences that have sufficiently different lengths, recombination tends to homogenise these (e.g. Fig. 1E). If, however, the variance within individuals σ_W^2 is small (compared to σ_B^2), recombination in this case makes these sequences more different by reshuffling TR copies, so that the variance increases (e.g. Fig. 1F). This is captured by the second term of Eq. (13b).

Comparing Eqs. (13a) and (13b) shows that there is a separation of timescales between the dynamics of the mean and variance: changes in mean are of order δ while changes in the variance are of order 1 (under our assumption that $s_a \sim \mathcal{O}(\delta)$ and $\mu \sim \mathcal{O}(\delta)$ while $\gamma \sim \mathcal{O}(1)$). This entails that the dynamics of the variance should stabilise to an equilibrium σ_{eq}^2 before the mean when δ is small. Solving $\Delta\sigma_T^2 = 0$ with Eq. (4) for σ_{eq}^2 with a given $\bar{z}$, we obtain that this equilibrium is

$$\sigma_{eq}^2 = \frac{2}{5 - 7F_{IS}}\left(\bar{z}^2 - 1\right) + \mathcal{O}(\delta)\,, \tag{14}$$

where

$$F_{IS} = \frac{\alpha\left(1 - \frac{\gamma}{2}\right)^2}{2 - \alpha\left(1 - \frac{\gamma}{2}\right)^2} + \mathcal{O}(\delta) \tag{15}$$

(SI Text A.5 for derivation). Equations (14) and (15) show that for a given population mean $\bar{z}$, a greater selfing rate α leads to a greater equilibrium variance. This is because selfing increases the reshuffling effect relative to the homogenising effect via a decrease in σ_W^2 compared to σ_B^2 in Eq. (13b). The effect of selfing on the inbreeding coefficient F_{IS} is reduced by unequal recombination because unequal recombination decreases homozygosity (this effect of γ is weighted by 1/2 in Eq. (15) as only the inner chromatids can undergo unequal recombination).

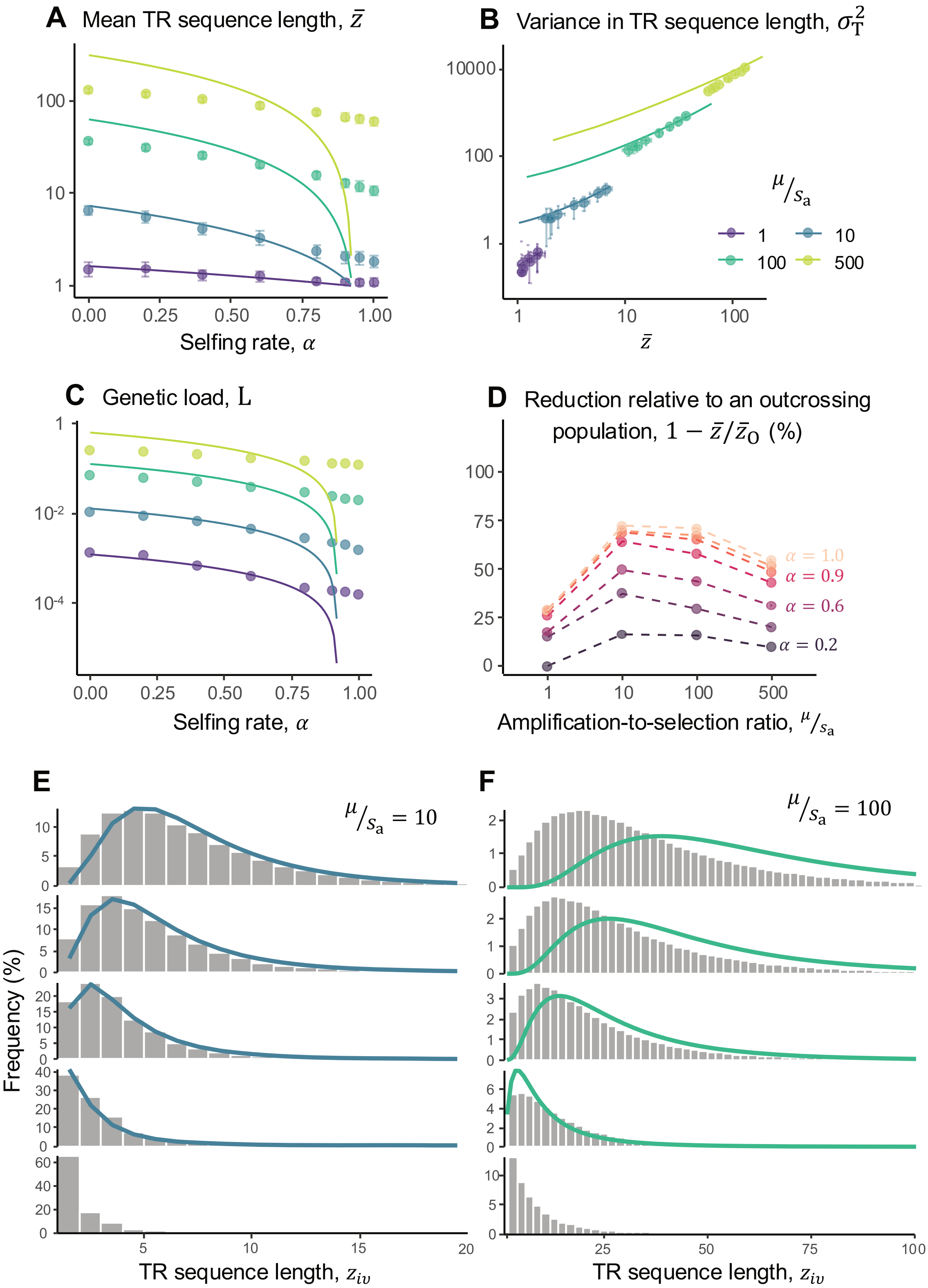
Solving $\Delta\bar{z} = 0$ for $\bar{z} = \bar{z}_{eq}$ with $\sigma_T^2 = \sigma_{eq}^2$ given by Eq. (14) (and using Eq. 4), we obtain that the mean of the equilibrium distribution of TR sequence length is

$$\bar{z}_{eq} = 1 + \left(\frac{5 - 7F_{IS}}{1 + F_{IS}}\right)\frac{\mu}{8s_a} + \mathcal{O}(\delta)\,, \tag{16}$$

which shows that selfing reduces the mean TR copy number in the population (Fig. 2A). This reduction is due to a greater proportion of the total variance in TRs that is between individuals, which increases the efficacy of selection (recall σ_B^2 in Eq. 13a). Plugging the mean in Eq. (16) into Eq. (14), we find that selfing also decreases the total variance among sequences in the population (Fig. 2B). Because selfing leads to fewer TR copies, it also reduces the load associated with them (Fig. 2C). Indeed, using Eq. (8) with Eqs. (12) and (16) we obtain the genetic load $L = \left(\frac{5-7F_{IS}}{1+F_{IS}}\right)\frac{\mu}{4} + \mathcal{O}(\delta^2)$. Therefore, genetic load decreases with selfing owing to shorter TR sequences (Fig. 2C). How much selfing reduces the mean and variance in TR sequence length depends on the amplification-to-selection ratio μ/s_a (curves in Fig. 2A).

We performed individual-based simulations to investigate the case $\mu/s_a \gg 1$ (points in Fig. 2A). As predicted, our analytical approach for $\bar{z}$ matches simulations quite well when amplification and selection are of similar order, and both are small relative to unequal recombination; with the exception when $F_{IS} > 5/7$ ($\alpha > 0.8$). In this latter case, our analytical model predicts the complete loss of repeated motifs ($\bar{z} = 1$ and $\sigma_T^2 = 0$) while simulations still show the maintenance of small TR sequences (mostly due to amplification events). This is because the separation of timescales breaks down when $\sigma_B^2 \gg \sigma_W^2$ (see Eq. 13b). When $\mu/s_a \gtrsim 100$, our simulations show shorter TR sequences in outcrossing populations than predicted by Eq. (16).

To see how the effects of selfing interact with other factors, it is useful to look at the relative reduction in copy number in a selfing population compared to an outcrossing one, all else being equal (i.e. we measured $(\bar{z}_O - \bar{z})/\bar{z}_O$ where $\bar{z}_O$ is the mean TR sequence length in outcrossing populations, $\alpha = 0$). This shows that the relative reduction due to selfing is smaller in the regime where $\mu \gg s_a$ (Fig. 2D). This is because: (i) amplification increases TR sequence length similarly in outcrossing and selfing populations thereby reducing the differences they show (recall Eq. 13a), and (ii)



amplification decreases the excess homozygosity caused by selfing and thus mitigates its effect.

Simulations also reveal that the skewness and kurtosis of the TR copy number distribution in the population are greater when selfing is more frequent (Fig. 2E and F for example; SI Fig. S1). Altogether, this means that under partial selfing, we expect more outliers with long sequences in a particular sample compared to if the TR sequence length were normally distributed. In fact, the distribution of TR sequence length shows a good fit to a lognormal distribution whose mean and variance are given by Eqs. (14) and (16). This fit is particularly good when $s_a < \mu < \gamma$ (see solid lines in Fig. 2E), but less good when $s_a < \mu \approx \gamma$ (Fig. 2F).

Fig. 2 TR sequence variation under purifying selection. **A** Mean TR sequence length $\bar{z}$ (in log scale) for various selfing rates α, with dots as averages from simulations (error bars show standard deviation; SI Text B for details on simulations). Lines represent the theoretical prediction obtained with Eqs. (14) and (16). Parameters: $s_a = 0.001$, $\gamma = 0.1$, $N = 2000$. **B** Relationship between the total variance σ^2_T and the mean TR sequence length $\bar{z}$ (both on log scale) for different selfing rates. Dots represent averages from simulations, while lines show theoretical predictions based on Eqs. (14) and (16). Notably, Eq. (14) accurately captures the relationship between the moments, even in cases where Eq. (16) does not accurately predict the mean. Parameters: same as (**A**). **C** Genetic load L (in log scale) as a function of selfing rate α, with dots as averages from simulations using Eq. (16). Lines represent the theoretical prediction. Parameters: same as (**A**). **D** Reduction in the mean TR sequence length $\bar{z}$ relative to outcrossing populations $\bar{z}_O$ for different selfing rates (from bottom to top: $\alpha = 0.2, 0.4, 0.6, 0.8, 0.9, 0.95$ and 1.0). **E**, **F** Distribution of TR sequence length in populations with different selfing rates (from top to bottom: $\alpha = 0, 0.25, 0.5, 0.75$ and 1.0). Histograms are based on simulations, and lines represent the theoretical prediction obtained with a lognormal distribution with mean given by Eq. (16) and variance given by Eq. (14). Parameters: $\mu/s_a = 10$ in (**E**) and $\mu/s_a = 100$ in (**F**), other parameters as (**A**).

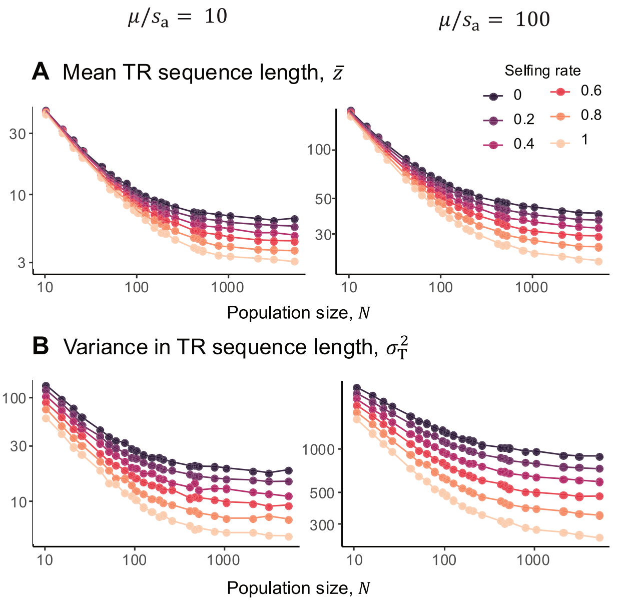


Fig. 3 TR sequence variation in small populations. **A** The mean TR sequence length $\bar{z}$ and **B** total variance σ^2_T (both in log scales) in simulations with small and intermediate population sizes N (between 10 and 5000 individuals, in log scale) and for populations with various selfing rates (see legend). The number of replicates varies between population sizes, with at least 10 simulations for each parameter combination. Parameters: amplification-to-selection ratio $\mu/s_a = 10$ (left) and $\mu/s_a = 100$ (right column), other parameters: $\gamma = 0.01$, $s_a = 0.001$.

The effect of genetic drift

We used simulations to investigate the effect of genetic drift in small populations, in particular when $2N \lesssim 1/\gamma$. These show that for a given selfing rate, smaller populations tend to have longer TR sequences (compare points of the same colour in Fig. 3A). This is due to less efficient purifying selection in smaller populations, which leads to an accumulation of more repeats. In fact, the variances between and within individuals are both greater in smaller populations (Fig. 3B for total variance σ^2_T; see SI Fig. S2 for each component). Additionally, the reduction in TR sequence length due to selfing is proportionally smaller in small populations (e.g. compare points when $N = 10$ in Fig. 3A). This is presumably because small outcrossing populations show some level of inbreeding due to sampling effects and thus greater homozygosity.

We also performed simulations with different population sizes N such that the effective population size N_e is held constant for different selfing rates α (with $N_e = (1-\alpha/2)N$, as in Eq. 6 of Pollak 1987; SI Fig. S3). As expected, these simulations show that the reduction due to selfing occurs when controlling for effective population size. More broadly, this shows that the effects of selfing via N_e are always weaker than those described in the section 'Shorter TR sequences under purifying selection in selfing populations', i.e. via selection (by boosting variance between individuals) and unequal recombination.

Truncation-like selection increases variance in TR copy number

We now consider the case where the TRs have non-additive effects such that the fitness of an individual rapidly declines after the number of repeats it carries goes beyond a certain threshold. Specifically, we assume that fecundity is

$$f_i = \frac{\pi - 2\arctan\left(s_\epsilon(z_{i1} + z_{i2} - 2\theta)\right)}{\pi + 2\arctan\left(2s_\epsilon\theta\right)} \tag{17}$$

where 2θ is a threshold for the total number of TRs, after which fecundity decreases towards zero at a rate that depends on the parameter $s_\epsilon > 0$. When $s_\epsilon \ll 1$, Eq. (17) approaches additive effects, and we recover Eq. (12) with a cost $s_a \to 2s_\epsilon/\pi$ per TR copy, regardless of θ. When $s_\epsilon \gg 1$, fecundity behaves similarly to a step function (in the limit $s_\epsilon \to \infty$): $f_i = 1$ for $z_{i1} + z_{i2} < 2\theta$ and $f_i = 0$ for $z_{i1} + z_{i2} > 2\theta$ (this limiting case would be equivalent to truncation selection, which was studied in haploids under random mating in Section 4 of Stephan 1987).

Using individual-based simulations, we find that as the fecundity in Eq. (17) approaches a truncation-selection curve (i.e. large s_ϵ), the mean and variance in TRs both increase (Fig. 4A). This is because selection against repeats when TR sequences are below the threshold (when $z_{i1} + z_{i2} < 2\theta$) is weaker when s_ϵ is large. In fact, as s_ϵ increases, the variance in TR sequence length per chromosome approaches $(\theta - 1)^2/12$, which is the variance of a random variable following a uniform distribution between 1 and θ (this is also true when amplification increases, i.e. as μ gets large, in line with the results of Stephan 1987, Fig. 4B, SI Fig. S4 for more details on the TR sequence length distribution).

Selfing, meanwhile, has similar effects to those found under additivity: it reduces the mean and variance in TR sequence length (compare dark and light lines in Fig. 4A, B). This reduction, in turn, lowers the genetic load in the population (Fig. 4C). The decrease in load is lower when selection is truncation-like (large s_ϵ) since the deleteriousness of each repeat (as long as their total remains below 2θ) is lower than when effects are additive. The effects of amplification and unequal recombination on TR sequence length are also similar to those under additivity across all selfing rates: amplification tends to increase TRs while recombination tends to reduce them (as in haploid populations; Stephan 1987).

Misalignment costs exacerbate the differences in the TR sequence length of selfing and outcrossing populations

We now consider potential costs arising from the misalignment of homologous genes surrounding a TR sequence, e.g. when physical distortions during synapsis due to different TR sequence lengths create DNA secondary structures, unstable recombination and loops that can lead to non-functional gametes after recombination (Balzano et al. 2021; Jarne and Lagoda 1996; Verbiest et al. 2023).

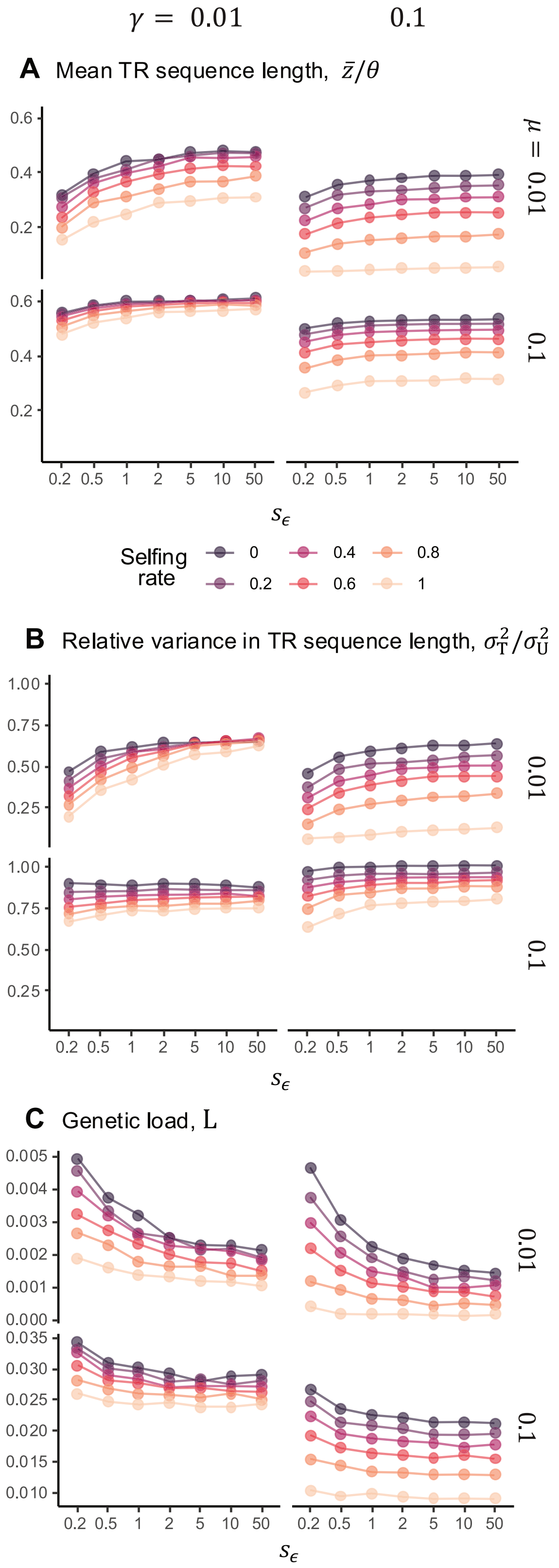


Fig. 4 TR sequence variation under truncation-like selection. **A** Mean TR sequence length $\bar{z}$ under non-additive fitness effects in Eq. (17). Mean is represented as a fraction of the threshold length θ. Unequal recombination rate is $\gamma = 0.01$ in the left column and $\gamma = 0.1$ in the right column, while amplification rates are indicated in the right margin. Parameters: $N = 2000$, $\theta = 100$. **B** Total variance in TR sequence length across various selfing rates (see legend in **A**) under non-additive fitness effects. As a reference, the variance is represented as a fraction of that of a uniform distribution where all sequences are equally present between 1 and the threshold θ, that is $\sigma_U^2 = (\theta - 1)^2/12$. Parameters: same as (**A**). **C** Genetic load under non-additive fitness effects between repeats, calculated from the simulation with Eq. (8), where $f_{\max}$ is given by Eq. (17) with $z_{i1} = z_{i2} = 1$. Parameters: same as (**A**).

To model this, we assume that the fecundity of an individual is

$$f_i = \underbrace{[1 - s_a(z_{i1} + z_{i2})]}_{\text{additive effects}} \times \underbrace{\left[1 - s_d \frac{|z_{i1} - z_{i2}|}{z_{i1} + z_{i2}}\right]}_{\text{misalignment effects}}, \qquad (18)$$

where the parameter $s_d \geq 0$ tunes the cost of the misalignment. When $s_d = 0$, we recover the additive model (Eq. 12), where each TR copy has effectively independent effects on fitness. When s_d is large, however, a difference in length between homologous TR sequences within individuals is also costly, reducing fitness. This can be seen as a form of under-dominance, causing heterozygote disadvantage.

Our simulations show that in outcrossing populations, misalignment costs cause an increase in mean TR sequence length (dark points, especially $s_d > 0.1$, in Fig. 5A). This is because any gamete carrying a TR sequence that is shorter than average will on average suffer a misalignment cost under random mating as it will likely fuse with a gamete carrying a longer TR sequence. Selection can therefore favour amplified sequences if it brings them closer in length to the average sequence, as long as s_a is not too large (Fig. 5A with $s_a = 0.001$ fixed). Contrastingly, misalignment costs have weak to no effects in selfing populations since these show a deficit of heterozygosity (light points in Fig. 5A). As a result, the difference in mean TR sequence length $\bar{z}$ between outcrossing and selfing populations is greater when s_d is large (Fig. 5A). The variance of TR sequence length is also affected by misalignment costs, but mostly in outcrossing populations where the variance within individuals σ_W^2 is reduced but between individuals σ_B^2 is increased. This is because misalignment costs simultaneously increase homozygosity and lead to longer sequences, inflating the total variance.

The effects of misalignment costs on the load L are similar to those on the mean TR sequence length, with the load increasing with misalignment costs in outcrossing but remaining similar in selfing populations (compare dark and light points in Fig. 5B). To disentangle the contribution to the load of misalignment effects from an increase in $\bar{z}$, we computed $2s_a(\bar{z} - 1)/L$ to measure the proportion of the load that is due to additive costs (i.e. arising from the first term in Eq. 18). When $s_d = 0$, the entire load is due to additive effects, but when s_d is larger, the costs of misalignment represent a larger fraction of the load despite longer sequences (Fig. 5C). In other words, misalignment costs grow faster than additive costs as s_d increases, and this is especially true in outcrossing populations (due to a deficit of heterozygotes in selfing populations).

Partial selfing reduces load when the TR sequence length is under stabilising selection

Finally, we examine the case where TR sequences are under stabilising selection for an optimal length, which aligns with recent evidence suggesting that some TRs can play functional roles in different biological processes (Balzano et al. 2021, for example) and contribute to phenotypic variation (Fotsing et al. 2019; Gymrek et al. 2016; Quilez et al. 2016; Verbiest et al. 2023). We assume fecundity is given by

$$f_i = \left[1 - s_b(z_{i1} - \Theta)^2\right] \times \left[1 - s_b(z_{i2} - \Theta)^2\right], \qquad (19)$$

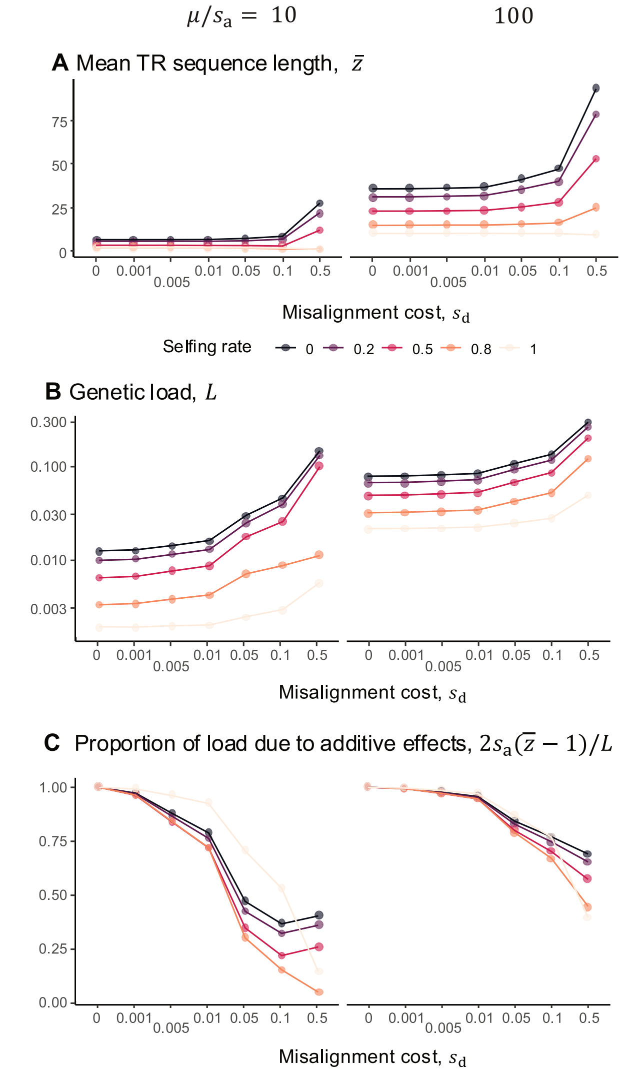


Fig. 5 TR sequence variation when sequence misalignment is costly. A Mean TR sequence length $\bar z$ across various selfing rates (see legend) for different costs of misalignment (s_d in Eq. (18) with $s_a = 0.001$ fixed). Parameters: $N = 2000$, $\gamma = 0.1$. **B** Genetic load L in the population (in log scale). Parameters: same as (**A**). **C** Proxy for the proportion of genetic load L due to additive effects. Parameters: same as (**A**).

where the parameter $s_b > 0$ tunes the strength of stabilising selection for the optimum Θ. We show in SI Text A.6.3 that by plugging Eq. (19) into Eq. (5), we obtain that the mean at mutation-selection balance in a large population can be expressed as

$$\bar z_{eq} = \frac{\Theta+1}{2} + \sqrt{\left(\frac{\Theta-1}{2}\right)^2 + \left(\frac{5-7F_{IS}}{1+F_{IS}}\right)\frac{\mu}{16s_b}} + \mathcal{O}(\delta), \qquad (20)$$

where we assumed that the distribution of TR sequence length is approximately Normal (which here is justified by stabilising selection; Walsh and Lynch 2018). With weak selection, the population variance in TR sequence length is still given by Eq. (14). The load, meanwhile, can be approximated as $\bar L = 2s_b\left((\bar z - \Theta)^2 + \sigma_T^2\right) + \mathcal{O}(\delta^2)$ using Eq. (19) in Eq. (8) with $f_{max} = 1$. We also performed individual-based simulations. These, together with Eq. (20), reveal that the effects of selfing depend on how large the optimal length Θ is compared to the amplification-to-selection ratio μ/s_b (Fig. 6).

When $\Theta \gg \mu/s_b$, the mean TR sequence length $\bar z$ is close to Θ, regardless of selfing rate. This can be seen from Eq. (20), putting μ/s_b to zero (see also Fig. 6A for simulations). This is because stabilising selection here is strong and amplification events are rare, so that rare mutations away from the optimum are efficiently purged in both outcrossing and selfing populations. In fact, the variance around Θ in the population is always small, especially so when unequal recombination is infrequent (compare top and bottom plots in left column, $\mu/s_b = 0.01$, in Fig. 6B). As a consequence of the absence of additional repeats and little variation, the load in the population is negligible (purple line in SI Fig. S5).

When $\Theta \lesssim \mu/s_b$, the mean TR sequence length in the population is greater than the optimum Θ as here stabilising selection is weak and amplification events frequent (Fig. 6A). The increase in $\bar z$ is smaller under selfing (compare points of different colours in Fig. 6A) as selfing increases the proportion of variance between individuals, thereby increasing the efficacy of stabilising selection (light grey area in Fig. 6B). Note that Eq. (20) tends to underestimate the $\bar z$ observed in simulations (Fig. 6A). This is because Eq. (20) uses the variance expected under neutrality (Eq. 14) and therefore neglects the effect of stabilising selection on variance. In simulations, stabilising selection reduces variance in TR sequence length close to the optimum, which weakens selection and leads to a higher equilibrium mean. The reduction in both the mean and total variance in TR sequence length under selfing leads to a corresponding reduction in genetic load (SI Fig. S5).

DISCUSSION

Our analyses indicate two main pathways through which selfing influences TRs via excess homozygosity: (i) by increasing the effect that unequal recombination has on generating variation among sequences within individuals, and (ii) by increasing the variance in TR copy number between individuals. As a result of these effects, selection (here assumed to take place at the diploid stage) tends to be more efficient under partial selfing, leading to shorter average TR sequence length and reduced polymorphism. In turn, this means that selfing populations show lower genetic load due to TRs. These effects of selfing on TR abundance and genetic load are especially strong in large populations and when selection is purifying.

Empirical implications

One implication of our results is that, all else being equal, TR sequence length should negatively correlate with selfing rates across species or populations. More broadly, since the effects of selfing operate through increased homozygosity, any mechanism increasing homozygosity is expected to similarly reduce mean TR sequence length. Thus, the TR sequence length should generally show a negative correlation with homozygosity, irrespective of the specific cause. The robustness of our results across amplification and recombination regimes indicates that they apply to both microsatellites and larger arrays such as minisatellites and short-motif satellites (Balzano et al. 2021; Ellegren 2004).

Genomic data that directly quantify TR sequence lengths within populations remain sparse, largely because TR arrays are often filtered out early in genome-processing pipelines. Nevertheless, available comparisons of microsatellite and minisatellite abundances across closely related species differing in mating systems are broadly consistent with our predictions. In nematodes, comparative genomic analyses between the obligate outcrossing species *Caenorhabditis nigoni* and its closely related, partially selfing hermaphroditic relative *C. briggsae* reveal a marked

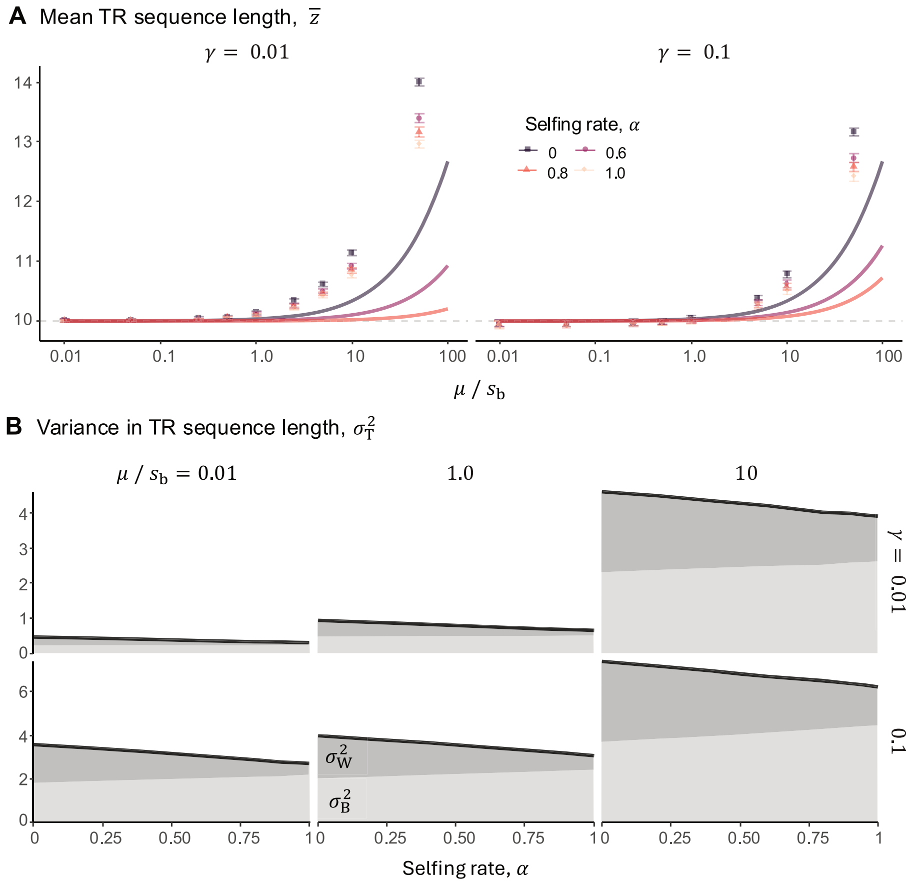


Fig. 6 TR sequence variation under stabilising selection. A Mean TR sequence length $\bar{z}$ across different values of amplification-to-selection ratio (μ/s_{b}) for various selfing rates (see legend). Unequal recombination rate is $\gamma = 0.01$ (left) and $\gamma = 0.1$ (right column). Curves present the analytical results of Eq. (20). The horizontal grey line indicates $\Theta = 10$. Parameters: $s_{\mathrm{b}} = 0.005$, $N = 5000$. **B** Variance in TR sequence length across various selfing rates under stabilising selection (from left to right, $\mu/s_{\mathrm{b}} = 0.01$, 1 and 10). The top row has $\gamma = 0.01$, while $\gamma = 0.1$ in the bottom row. Shades of grey under the curve represent the proportion of variation between (σ^2_{B}, light grey) and within individuals (σ^2_{W}, dark grey). Parameters: same as (**A**).

reduction in TR content in the latter: *C. briggsae* carries roughly 25% fewer TRs, and those present tend to be shorter and less stable (Subirana and Messeguer 2017). The authors attribute this to unequal recombination being more mutagenic in heterozygotes, leading to new TR arrays arising more frequently in outcrossers. By contrast, in our model, similar differences arise even when unequal recombination is not mutagenic, suggesting that mating-system effects on homozygosity and the efficacy of selection could, in principle, be sufficient to generate this pattern.

Parallel trends are seen in plants. Within the *Arabidopsis* genus, self-compatible populations harbour less microsatellite polymorphism than self-incompatible ones, a pattern consistent with repeat loss in highly inbred lineages (Clauss et al. 2002; Mable and Adam 2007). Across lineages of the monkeyflower *Mimulus guttatus*, selfing is associated with a reduction of microsatellite variation (Awadalla and Ritland 1997). This reduction was interpreted as evidence for recent bottlenecks in selfing lineages, yet our results show that substantial losses of TR variation can arise in selfing populations even when effective population size is held constant (SI Fig. S3). Finally, across *Plantago* species, minisatellite variation is reduced in the highly selfing *P. major* compared to the outcrossing *P. lanceolata* (Wolff et al. 1994), which is again consistent with an effect of mating system and population structure on TR variation as seen in our model.

Indirect evidence may also come from broader cross-taxon comparisons, although these are necessarily harder to interpret. For example, humans carry substantially longer TR sequences than other primates (Rubinsztein et al. 1995). One possible explanation is that humans tend to have lower levels of homozygosity than many other primate species (Kuderna et al. 2023), although differences in demographic history can also influence homozygosity and complicate this comparison (Prado-Martinez et al. 2013). Similarly, genome-size surveys generally report smaller genomes in selfing species than in their outcrossing relatives (e.g. across 14 plant species pairs in Wright et al. 2008, or within the genus *Veronica* in Albach and Greilhuber 2004, where higher selfing rates were associated with smaller genomes and less repetitive DNA). Because short TRs typically constitute a relatively uniform proportion (roughly 1–5%, with few exceptions) of eukaryotic genomes (Srivastava et al. 2019), such correlations could be consistent with our prediction that increased homozygosity reduces repeat copy number. However, genome size and TR abundance are shaped by many factors, so these broad comparisons must be interpreted with caution. More direct tests will require explicit quantification of TR copy-number variation across species or populations that differ in mating system or inbreeding level.

Our model further predicts that TR sequences should, on average, be longer in genomic regions with lower recombination or in smaller populations, where unequal recombination and selection against expansions are less effective; an effect similarly seen in haploid populations (Charlesworth et al. 1986; Stephan 1987). Low recombination regions in the genome are indeed often associated with higher TR densities (e.g. centromeres of most eukaryotes are largely composed of satellite DNA, with yeast as a notable exception, Charlesworth et al. 1986; Garrido-Ramos 2015; and repetitive DNA is abundant on heterogametic sex chromosomes, Cooke 1976; Kejnovsky et al. 2009; Kubat et al. 2008). Comparisons regarding population size, however, are more difficult to interpret due to the effects of recent bottlenecks or population structure.

The nature of selection on TRs

Previous theoretical work on TR sequence evolution has primarily considered purifying selection in haploid populations (Stephan 1987). Our model complements these analyses by explicitly incorporating diploidy, partial selfing and by introducing two additional selective scenarios: (i) selection against misalignment costs (heterozygote disadvantage), based on studies showing increased chromosomal or recombinational instabilities in heterozygous TR sequences (Jarne and Lagoda 1996; John and Miklos 1979; Usdin et al. 2015; Verbiest et al. 2023); and (ii) stabilising selection favouring an intermediate TR sequence length, motivated by evidence that some TRs influence the expression of nearby genes and thereby phenotypes that may themselves be under stabilising selection (Fotsing et al. 2019; Gymrek et al. 2016; Quilez et al. 2016; Reinar et al. 2021; Sureshkumar et al. 2025; Verbiest et al. 2023; Zhang et al. 2025).

Across all regimes considered, selfing consistently reduces the genetic load associated with TRs. The robustness of this effect suggests that correlations between mean TR sequence length and inbreeding coefficients (F_{IS}) may not, on their own, allow strong inferences about the underlying mode of selection. A lack of correlation might indicate either strong stabilising selection or a mutation-driven equilibrium that is insensitive to the mating system. Such equilibria arise when mutational processes such as slippage, point mutation, expansion and contraction maintain a stable mean TR sequence length (Kruglyak et al. 1998, 2000). Conversely, a very strong correlation could be indicative of selection against misalignment costs, as this scenario produced the largest differences in TR sequence length between selfing and outcrossing populations in our analyses.

Under purifying selection, the distribution of TR sequence lengths is heavily skewed and characterised by frequent long outliers, particularly when mean lengths are short (e.g. under low amplification or high selfing rates). Our results indicate that these distributions are well-described by a lognormal distribution, with many short sequences alongside occasional but readily sampled long ones. This matches with empirical descriptions of TRs as hypervariable (Jeffreys et al. 1985; Lareu et al. 1998; Legendre et al. 2007; Lundström et al. 2023; Verbiest et al. 2023; Wei et al. 2014). Additionally, our model predicts greater variation in TR sequence length within populations as mean TR sequence length increases, consistent with human data (Duitama et al. 2014; Legendre et al. 2007).

Under strong truncation-like selection, the average TR sequence length in the population remains well below the threshold above which fitness declines ($\bar{z} \lesssim 0.6\theta$ in Fig. 4A). Selection thus minimises the risk of deleterious expansions through replication slippage or recombination within a lineage. Some empirical patterns are consistent with this: for example, Fragile X syndrome manifests beyond 200 repeats of a CGG motif, whereas most humans carry only 5 to 40 repeats (Depienne and Mandel 2021; Lundström et al. 2023).

Contrasts and limitations

TR sequences belong to the broader category of repetitive DNA, which also includes transposable elements, but the two groups are expected to respond differently to selfing because their underlying mechanisms differ. Although both can amplify and be affected by recombination, they differ in their genomic organisation and in the recombinational processes they experience. TRs form tandem arrays in which unequal recombination is frequent and local, whereas transposable element copies are dispersed across the genome and primarily experience ectopic recombination, which is rarer and often more deleterious (Wicker et al. 2007). Because of these differences, theoretical predictions for how selfing affects transposable element abundance are less clear-cut than for TRs. In established models, selfing can either increase or decrease transposable element copy number depending on the interplay between recombination, amplification and non-additive fitness effects (Boutin et al. 2012; Charlesworth and Charlesworth 1995; Morgan 2001; Roze 2023; Wright and Schoen 1999). Empirical studies likewise show mixed patterns, with some selfing species carrying fewer transposable element copies than outcrossing relatives (Ågren et al. 2014; Albach and Greilhuber 2004; De La Chaux et al. 2012) and others showing the opposite trend (Dolgin et al. 2008; Lockton and Gaut 2010; Wright et al. 2001). These contrasts suggest that repetitive elements differ in how they respond to selfing, depending on the mechanisms that generate and remove copy-number variation.

Because selfing reduces effective population size and decreases effective recombination, it is typically associated with greater genetic load, except when load is driven by recessive deleterious mutations (Abu Awad and Roze 2018; Glémin 2007; Hartfield and Glémin 2014; Sianta et al. 2023; see Crow and Kimura 1970, p. 299 for an overview on mutation load, and Burgarella and Glémin 2017 for a review on effects of selfing). In such cases, selfing can reduce load by exposing recessive mutations to selection. In contrast, our analyses show that, because mutation and recombination shape TR variation through saltatory amplification and unequal recombination, selfing leads to shorter TR sequences and lower genetic load, even when TR copies have additive fitness effects within each sequence and between homologous sequences (Eq. 12).

Our model relies on several simplifying assumptions that point towards useful directions for future work. First, we considered a single well-mixed population. However, since TRs often contribute to genetic differentiation between populations (Goldstein et al. 1995; Slatkin 1995), investigating TR sequence evolution in subdivided populations connected by limited dispersal could be relevant, especially as kin selection and inbreeding may interact there (Rousset 2004).

Second, we considered selection acting on a single TR sequence, whereas in reality, each TR sequence exists within a genetic context that can also be affected by selfing (e.g. linkage with recessive deleterious mutations). Linked recessive deleterious mutations, for example, could interfere with selection on TRs, particularly in selfing species where effective recombination is reduced (Burgarella and Glémin 2017). If interference reduces the efficacy of purifying selection, we expect longer sequences than predicted in our model. We also did not consider selection against frameshifts in coding regions, which is most relevant for TRs composed of repeat units not divisible by three (Ellegren 2004).

Third, we assumed that replication slippage always increases TR sequence length at a constant rate. However, slippage can either increase or decrease TR sequence length depending on which DNA strand loops out, though there is a bias towards increases due to the flexibility of the newly synthesised strand (Ellegren 2004; Knox et al. 2024; Seyfert et al. 2008). We also neglected point mutations as well as their possible interactions with slippage (Ellegren 2002; Kruglyak et al. 1998). This omission is partly justified because slippage events are estimated to occur ten to one hundred times more frequently than point mutations in TRs (e.g. in primates; Pumpernik et al. 2008). In addition, the probability of slippage can depend on TR length, becoming less likely below a critical threshold (Pumpernik et al. 2008; Verbiest et al. 2023). Including this effect would likely amplify the differences we predict between outcrossing and selfing populations, because longer TRs would undergo more amplification events. Slippage is also less relevant for large-motif TRs, where amplification may instead be mediated by transposable elements (Balzano et al. 2021; Meštrović et al. 2015).

Fourth, we assumed that the two homologous TR sequences carried by a selfed offspring come from independent gametes. This is appropriate where offspring arise from two gametes produced by separate meioses, but other forms of reproduction, such as parthenogenesis, involve offspring deriving from gametes produced by the same meiosis. In these cases, homologous TR sequence lengths within offspring could be correlated.

Finally, we focused on autosomal TRs, although TRs are also abundant in sex chromosomes (e.g. TRs constitute about half of the Y chromosome in humans; Cooke 1976). Under our assumptions, the non-recombining heterogametic chromosome (Y or W) would presumably accumulate repeats, due to its inability to purge repeats combined with biased replication slippage towards increased lengths. In contrast, evolution on the recombining homogametic sex chromosome (X or Z) would mirror that of an autosome, but with appropriately rescaled effective population size and unequal recombination rates.

Conclusions

Our results show that mating systems, specifically partial selfing and the associated increase in homozygosity, influence the evolutionary dynamics of TRs. Selfing consistently reduces the genetic load associated with TRs across multiple selection scenarios. This reduction is primarily due to the way unequal recombination generates TR variation within individuals and how this interacts with selfing. Existing genomic data on microsatellites in partially selfing plants and nematodes are broadly consistent with these predictions, but more explicit comparisons of TR variation across populations or species that differ in mating system or inbreeding level will be needed to test them rigorously.

DATA AVAILABILITY

The code used for our simulations is available here: https://vsudbrack.github.io/TandemRepeatsSM. The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures and tables.

REFERENCES

Abbott RJ, Gomes MF (1989) Population genetic structure and outcrossing rate of Arabidopsis thaliana (L.) Heynh. Heredity 62(3):411–418.

Abu Awad D, Roze D (2018) Effects of partial selfing on the equilibrium genetic variance, mutation load, and inbreeding depression under stabilizing selection. Evolution 72(4):751–769.

Ågren JA, Wang W, Koenig D, Neuffer B, Weigel D, Wright SI (2014) Mating system shifts and transposable element evolution in the plant genus Capsella. BMC Genom 15(1):602.

Albach DC, Greilhuber J (2004) Genome size variation and evolution in Veronica. Ann Bot 94(6):897–911.

Awadalla P, Ritland K (1997) Microsatellite variation and evolution in the Mimulus guttatus species complex with contrasting mating systems. Mol Biol Evol 14(10):1023–1034.

Balzano E, Pelliccia F, Giunta S (2021) Genome (in) stability at tandem repeats. Semin Cell Dev Biol 113: 97–112.

Biscotti MA, Canapa A, Forconi M, Olmo E, Barucca M (2015) Transcription of tandemly repetitive DNA: functional roles. Chromosome Res 23(3):463–477.

Boutin TS, Le Rouzic A, Capy P (2012) How does selfing affect the dynamics of selfish transposable elements?. Mob DNA 3(1):5.

Burgarella C, Glémin S (2017) Population Genetics and Genome Evolution of Selfing Species. In eLS, John Wiley & Sons, Ltd (Ed.) https://doi.org/10.1002/9780470015902.a0026804.

Buschiazzo E, Gemmell NJ (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. Bioessays 28(10):1040–1050.

Charlesworth B, Langley CH, Stephan W (1986) The evolution of restricted recombination and the accumulation of repeated DNA sequences. Genetics 112(4):947–962.

Charlesworth B, Sniegowski P, Stephan W (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371(6494):215–220.

Charlesworth D, Charlesworth B (1995) Transposable elements in inbreeding and outbreeding populations. Genetics 140(1):415.

Clauss M, Cobban H, Mitchell-Olds T (2002) Cross-species microsatellite markers for elucidating population genetic structure in Arabidopsis and Arabis (Brassicaceae). Mol Ecol 11(3):591–601.

Cooke H (1976) Repeated sequence specific to human males. Nature 262(5565):182–186.

Crow JF, Kimura M (1970) An introduction to population genetics theory. Scientific Publishers. New Jersey (USA).

De La Chaux N, Tsuchimatsu T, Shimizu KK, Wagner A (2012) The predominantly selfing plant Arabidopsis thaliana experienced a recent reduction in transposable element abundance compared to its outcrossing relative Arabidopsis lyrata. Mob DNA 3(1):2.

Depienne C, Mandel J-L (2021) 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges?. Am J Hum Genet 108(5):764–785.

Dolgin ES, Charlesworth B, Cutter AD (2008) Population frequencies of transposable elements in selfing and outcrossing caenorhabditis nematodes. Genet Res 90(4):317–329.

Duitama J, Zablotskaya A, Gemayel R, Jansen A, Belet S, Vermeesch JR (2014) Large-scale analysis of tandem repeat variability in the human genome. Nucleic Acids Res 42(9):5728–5741.

Ellegren H (2002) Microsatellite evolution: a battle between replication slippage and point mutation. Trends Genet 18(2):70.

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5(6):435–445.

Fan H, Chu J-Y (2007) A brief review of short tandem repeat mutation. Genom Proteom Bioinform 5(1):7–14.

Flynn JM, Caldas I, Cristescu ME, Clark AG (2017) Selection constrains high rates of tandem repetitive DNA mutation in Daphnia pulex. Genetics 207(2):697–710.

Fotsing SF, Margoliash J, Wang C, Saini S, Yanicky R, Shleizer-Burko S (2019) The impact of short tandem repeat variation on gene expression. Nat Genet 51(11):1652–1659.

Garrido-Ramos MA (2015) Satellite DNA in plants: more than just rubbish. Cytogenet Genome Res 146(2):153–170.

Glémin S (2007) Mating systems and the efficacy of selection at the molecular level. Genetics 177(2):905–916.

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW (1995) An evaluation of genetic distances for use with microsatellite loci. Genetics 139(1):463–471.

Greider CW (1990) Telomeres, telomerase and senescence. Bioessays 12(8):363–369.

Gymrek M, Goren A (2021) Missing heritability may be hiding in repeats. Science 373(6562):1440–1441.

Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S (2016) Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet 48(1):22–29.

Haller BC, Messer PW (2023) SLiM 4: multispecies eco-evolutionary modeling. Am Nat 201(5):E127–E139.

Hammond HA, Jin L, Zhong Y, Caskey CT, Chakraborty R (1994) Evaluation of 13 short tandem repeat loci for use in personal identification applications. Am J Hum Genet 55(1):175.

Hartfield M, Glémin S (2014) Hitchhiking of deleterious alleles and the cost of adaptation in partially selfing species. Genetics 196(1):281–293.

Ide S, Miyazaki T, Maki H, Kobayashi T (2010) Abundance of ribosomal RNA gene copies maintains genome integrity. Science 327(5966):693–696.

Jarne P, Lagoda PJ (1996) Microsatellites, from molecules to populations and back. Trends Ecol Evol 11(10):424–429.

Jeffreys AJ, Wilson V, Thein SL (1985) Hypervariable 'minisatellite' regions in human DNA. Nature 314(6006):67–73.

John B, Miklos GLG (1979) Functional aspects of satellite DNA and heterochromatin. Int Rev Cytol 58:1–114.

Kejnovsky E, Hobza R, Cermak T, Kubat Z, Vyskot B (2009) The role of repetitive DNA in structure and evolution of sex chromosomes in plants. Heredity 102(6):533–541.

Khristich AN, Mirkin SM (2020) On the wrong DNA track: molecular mechanisms of repeat-mediated genome instability. J Biol Chem 295(13):4134–4170.

Knox MA, Biggs PJ, Garcia-R JC, Hayman DT (2024) Quantifying replication slippage error in cryptosporidium metabarcoding studies. J Infect Dis 230:e144–e148.

Krüger J, Vogel F (1975) Population genetics of unequal crossing over. J Mol Evol 4:201–247.

Kruglyak S, Durrett RT, Schug MD, Aquadro CF (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci USA 95(18):10774–10778.

Kruglyak S, Durrett R, Schug MD, Aquadro CF (2000) Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations. Mol Biol Evol 17(8):1210–1219.

Kubat Z, Hobza R, Vyskot B, Kejnovsky E (2008) Microsatellite accumulation on the Y chromosome in Silene latifolia. Genome 51(5):350–356.

Kuderna LF, Gao H, Janiak MC, Kuhlwilm M, Orkin JD, Bataillon T (2023) A global catalog of whole-genome diversity from 233 primate species. Science 380(6648):906–913.

Lareu M, Barral S, Salas A, Pestoni C, Carracedo A (1998) Sequence variation of a hypervariable short tandem repeat at the D1S1656 locus. Int J Leg Med 111(5):244–247.

Legendre M, Pochet N, Pak T, Verstrepen KJ (2007) Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Res 17(12):1787–1796.

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4(3):203–221.

Lockton S, Gaut BS (2010) The evolution of transposable elements in natural populations of self-fertilizing Arabidopsis thaliana and its outcrossing relative Arabidopsis lyrata. BMC Evol Biol 10(1):10.

Lundström OS, Verbiest MA, Xia F, Jam HZ, Zlobec I, Anisimova M (2023) Webstr: a population-wide database of short tandem repeat variation in humans. J Mol Biol 435(20):168260.

Mable BK, Adam A (2007) Patterns of genetic diversity in outcrossing and selfing populations of Arabidopsis lyrata. Mol Ecol 16(17):3565–3580.

McGinty RJ, Balick DJ, Mirkin SM, Sunyaev SR (2025) Inherent instability of simple DNA repeats shapes an evolutionarily stable distribution of repeat lengths. bioRxiv. https://doi.org/10.1101/2025.01.09.631797.

Meštrović N, Mravinac B, Pavlek M, Vojvoda-Zeljko T, Šatović E, Plohl M (2015) Structural and functional liaisons between transposable elements and satellite DNAs. Chromosome Res 23(3):583–596.

Miklos GLG, Gill AC (1982) Nucleotide sequences of highly repeated DNAs; compilation and comments. Genet Res 39(1):1–30.

Morgan MT (2001) Transposable element number in mixed mating populations. Genet Res 77(3):261–275.

Ohta T (1981) Population genetics of selfish DNA. Nature 292(5824):648–649.

Ohta T (1983a) On the evolution of multigene families. Theor Popul Biol 23(2):216–240.

Ohta T (1983b) Theoretical study on the accumulation of selfish DNA. Genet Res 41(1):1–15.

Ohta T, Kimura M (1981) Some calculations on the amount of selfish DNA. Proc Natl Acad Sci USA 78(2):1129–1132.

Perelson AS, Bell GI (1977) Mathematical models for the evolution of multigene families by unequal crossing over. Nature 265(5592):304–310.

Pollak E (1987) On the theory of partially inbreeding finite populations. I. Partial selfing. Genetics 117(2):353.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B (2013) Great ape genetic diversity and population history. Nature 499(7459):471–475.

Pumpernik D, Oblak B, Borštnik B (2008) Replication slippage versus point mutation rates in short tandem repeats of the human genome. Mol Genet Genom 279:53–61.

Quilez J, Guilmatre A, Garg P, Highnam G, Gymrek M, Erlich Y (2016) Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. Nucleic acids Res 44(8):3750–3762.

Reinar WB, Lalun VO, Reitan T, Jakobsen KS, Butenko MA (2021) Length variation in short tandem repeats affects gene expression in natural populations of Arabidopsis thaliana. Plant Cell 33(7):2221–2234.

Richard G-F, Kerrest A, Dujon B (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev 72(4):686–727.

Rousset F (2004) Genetic structure and selection in subdivided populations, vol 40. Princeton University Press New Jersey (USA).

Roze D (2023) Causes and consequences of linkage disequilibrium among transposable elements within eukaryotic genomes. Genetics 224(2):iyad058.

Rubinsztein DC, Amos W, Leggo J, Goodburn S, Jain S, Li S-H (1995) Microsatellite evolution-evidence for directionality and variation in rate between species. Nat Genet 10(3):337–343.

Seyfert AL, Cristescu ME, Frisse L, Schaack S, Thomas WK, Lynch M (2008) The rate and spectrum of microsatellite mutation in Caenorhabditis elegans and Daphnia pulex. Genetics 178(4):2113–2121.

Sianta SA, Peischl S, Moeller DA, Brandvain Y (2023) The efficacy of selection may increase or decrease with selfing depending upon the recombination environment. Evolution 77(2):394–408.

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139(1):457–462.

Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover: DNA whose sequence is not maintained by selection will develop periodicities as a result of random crossover. Science 191(4227):528–535.

Srivastava S, Avvaru AK, Sowpati DT, Mishra RK (2019) Patterns of microsatellite distribution across eukaryotic genomes. BMC Genom 20:1–14.

Stephan W (1986) Recombination and the evolution of satellite DNA. Genet Res 47(3):167–174.

Stephan W (1987) Quantitative variation and chromosomal location of satellite DNAs. Genet Res 50(1):41–52.

Stephan W (1989) Tandem-repetitive noncoding DNA: forms and forces. Mol Biol Evol 6(2):198–212.

Strand M, Prolla TA, Liskay RM, Petes TD (1993) Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature 365(6443):274–276.

Subirana JA, Messeguer X (2017) Evolution of tandem repeat satellite sequences in two closely related Caenorhabditis species. diminution of satellites in hermaphrodites. Genes 8(12):351.

Sureshkumar S, Chhabra A, Guo Y-L, Balasubramanian S (2025) Simple sequence repeats and their expansions: role in plant development, environmental response and adaptation. New Phytol 247(2):504–517.

Takahata N (1981) A mathematical study on the distribution of the number of repeated genes per chromosome. Genet Res 38(1):97–102.

Usdin K, House NC, Freudenreich CH (2015) Repeat instability during DNA repair: Insights from model systems. Crit Rev Biochem Mol Biol 50(2):142–167.

Verbiest M, Maksimov M, Jin Y, Anisimova M, Gymrek M, Bilgin Sonay T (2023) Mutation and selection processes regulating short tandem repeats give rise to genetic and phenotypic diversity across species. J Evol Biol 36(2):321–336.

Walsh B, Lynch M (2018) Evolution and selection of quantitative traits. Oxford University Press Walsh. New York (USA).

Walsh JB (1987) Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. Genetics 115(3):553–567.

Wei KH-C, Grenier JK, Barbash DA, Clark AG (2014) Correlated variation and population differentiation in satellite DNA abundance among lines of Drosophila melanogaster. Proc Natl Acad Sci USA 111(52):18793–18798.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8(12):973–982.

Wolff K, Rogstad S, Schaal B (1994) Population and species variation of minisatellite DNA in Plantago. Theor Appl Genet 87(6):733–740.

Wright SI, Schoen DJ (1999) Transposon dynamics and the breeding system. Genetica 107(1):139–148.

Wright SI, Le QH, Schoen DJ, Bureau TE (2001) Population dynamics of an ac-like transposable element in self-and cross-pollinating Arabidopsis. Genetics 158(3):1279–1288.

Wright SI, Ness RW, Foxe JP, Barrett SC (2008) Genomic consequences of outcrossing and selfing in plants. Int J Plant Sci 169(1):105–118.

Zhang Z-Q, Jiang J, Xu Y-C, Dent C, Sureshkumar S, Balasubramanian S (2025) Mutations of short tandem repeats explain abundant trait heritability in Arabidopsis. Genome Biol 26(1):242.

ACKNOWLEDGEMENTS

The authors thank Tanja Schwander for motivating this study, Ehouarn Le Faou, Denis Roze, Diala Abu Awad and William Toubiana for useful feedback, the editor and three anonymous reviewers for useful comments. This work was funded by the Swiss National Science Foundation (PCEFP3181243 to CM).

AUTHOR CONTRIBUTIONS

VS performed the analytical work with the supervision of CM, conducted the simulations, produced the figures and co-wrote the manuscript. CM co-wrote the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41437-026-00820-1.

Correspondence and requests for materials should be addressed to Vitor Sudbrack.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.