Theoretical Population Biology 164 (2025) 12–22

Contents lists available at ScienceDirect

Theoretical Population Biology

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

Muller's ratchet and gene duplication

Fabian Freund [a], Johannes Wirtz [b,c], Yichen Zheng [c,d], Yannick Schäfer [c], Thomas Wiehe [c,*]

[a] University of Leicester, University Street, Leicester, LE1 7RH, United Kingdom
[b] CEFE, Route de Mende 1819, Montpellier, France[1]
[c] Institut für Genetik, Universität zu Köln, Zülpicher Straße 47a, 50674 Köln, Germany
[d] Institut für Nutztiergenetik, Friedrich-Loeffler-Institut, Höltystrasse 10, 31535 Neustadt am Rübenberge, Germany[1]

ARTICLE INFO

Dataset link: https://github.com/y-zheng/Runaway-Duplication-Simulations/, https://github.com/fabfreund/genedup

Keywords:
Gene duplication
Multiplicative fitness
Synergistic epistasis
Muller's ratchet
Danio rerio
Homo sapiens

ABSTRACT

Copy number of genes in gene families can be highly variable among individuals and may continue to change across generations. Here, we study a model of duplication–selection interaction, which is related to Haigh's mutation–selection model of Muller's ratchet. New gene copies are generated by duplication but fitness of individuals decreases as copy number increases. Our model comes in two flavors: duplicates are copied either from a single template or from any existing copy. A duplication–selection equilibrium exists in both cases for infinite size populations and is given by a shifted Poisson or a negative binomial distribution. Unless counteracted by synergistic epistasis, finite populations suffer from loss of low copy-number haplotypes by drift, forcing them into a regime called 'run-away evolution' in which new copies accumulate without bound nor equilibrium. We discuss a few empirical examples and interpret them in the light of our models. Generally, large gene families appear too over-dispersed to fit the single template model suggesting a dynamic, and potentially accelerating, duplication process.

1. Introduction

In his classical paper (Muller, 1964), Muller studied "the relation of recombination to mutational advance" noting "a kind of irreversible ratchet mechanism in non-recombining species". The term "Muller's ratchet" appeared only several years later in the title of Haigh's paper from 1978, where he studied the dynamics of deleterious mutations in infinite and finite populations (Haigh, 1978). While a stable mutation–selection equilibrium is reached in infinite populations void of drift, finite populations continue to accumulate deleterious mutations if not counteracted by recombination. Together, the two papers triggered broad interest in this fundamental evolutionary mechanism among both theoreticians and experimentalists entailing a large volume of follow-up studies (see Desai, 2019 for a recent overview). One of the focal questions of theoretical interest concerned the speed of the ratchet. Clearly, this depends on the relative magnitude of the strength of selection and drift and the mutation rate (Stephan et al., 1993; Gordo and Charlesworth, 2000; Neher and Shraiman, 2012; Casanova et al., 2023). Importantly, the ratchet in finite populations can be slowed down and even be halted by appropriate recombination of mutation-loaded chromosomes. This effect keeps the mutation load of natural populations in check and also constitutes the most popular explanation for the evolution and maintenance of sexual reproduction (Felsenstein, 1974; Smith, 1978; Kondrashov, 1988; Charlesworth, 1990; Gerrish et al., 2021). Another mechanism capable of preventing build-up of mutation load is a steady influx of beneficial mutations, which may drive the population towards a mutation–selection equilibrium (Goyal et al., 2012). A third way to slow down the speed of the ratchet is by synergistic epistasis: it amplifies the deleterious fitness effect of additional mutations in an already mutation-loaded chromosome (Kondrashov, 1994; Wiehe, 1997; Jain, 2008) and poses a selection threshold to unlimited accumulation of deleterious mutations. In this sense, epistasis may be considered a substitute for recombination which should also work under forms of reproduction which are not obligately sexual or dioecious.

In the model considered here, we replace mutation by gene duplication and study gene copy number variation (gCNV) in a population. We consider a hypothetical gene family where new copies are produced by duplication of an existing gene, leading to variable family sizes among individuals. We study two versions of the duplication process: one in which only a template gene may duplicate, called 'single template model' (STM), and one in which any copy may duplicate, called 'compound copy model' (CCM). While both represent theoretical extremes, duplication in nature may be somewhere in between and depend on other features of the gene family under consideration. Our focus is

* Corresponding author.
E-mail address: twiehe@uni-koeln.de (T. Wiehe).
[1] Current address

https://doi.org/10.1016/j.tpb.2025.04.002
Received 23 December 2024
Available online 13 May 2025

on the overall count of gene copies in a given individual's genome and we disregard the effect of intra-chromosomal recombination on this count. For a multiplicative fitness function we derive formulae for the duplication–selection equilibrium in infinite populations. We show that this equilibrium is close to a family of standard distributions, and develop ways of estimating the relevant parameters from data. In simulations we study properties of the two model versions in finite populations and under the effect of epistasis. In particular, we show that the 'duplication ratchet' is slowed down, and can come to an effective halt, if synergistic epistasis is sufficiently strong. Finally, we interpret in the light of our model experimentally observed copy number counts of several gene families in human (Brahmachary et al., 2014) and of the NLR gene family of the vertebrate model species *D. rerio* (Suurväli et al., 2021; Schäfer et al., 2024).

2. Model

Let a population of constant size $N \leq \infty$ be propagated in discrete generations $t = 0, 1, 2, \ldots$. Consider a gene family where each individual of the population has $i_o + i$, $i \geq 0$, gene copies and $i_o \geq 1$ is the smallest copy number found in this population (i.e., i can differ between individuals, but i_o is the same for all individuals). In general, i_o and i are realizations of time-dependent discrete random variables $I_o(t)$ and $I(t)$.

Let $y_i(t)$, $i \geq 0$, be the relative frequency of (the equivalence class of) individuals with $i_o + i$ copies (i.e. those, which have i 'extra' copies) at time t. Clearly, for all $t \geq 0$ holds $\sum_{i\geq 0} y_i(t) = 1$. Over time, allele frequencies may change due to duplication, selection or – in the case of finite populations – due to drift. In this latter case, also $I_o(t)$ may change: it increases once all individuals with exactly i_o copies (i.e., the 'zero' class with $i = 0$) are lost.

Further, we assume a fitness function governed by the selection coefficient $0 < s < 1$ and the epistasis parameter $0 < \beta$ (cf. Kondrashov, 1994, 1988; Wiehe, 1997). Fitness $v_{i_o,i}$ of an individual with $i_o + i$ copies is

$$v_{i_o,i} = (1-s)^{(i_o+i)^\beta} / (1-s)^{i_o^\beta}\,,\ i \geq 0\,. \tag{1}$$

This definition implies that individuals which carry only the minimum number of genes ($i = 0$), have fitness 1. Mean and variance of fitness in generation t are

$$\begin{aligned}\bar{v}_{i_o}(t) &= \sum_{i\geq 0} v_{i_o,i}\ y_i(t)\,,\\ \mathbb{V}(v_{i_o}(t)) &= \sum_{i\geq 0} v_{i_o,i}^2\ y_i(t) - \bar{v}_{i_o}^2(t)\,.\end{aligned}$$

When $\beta = 1$ there is no epistasis. Fitness is multiplicative and depends only on i, but not on i_o. In this case

$$v_{i_o,i} = v_i = (1-s)^i\,,\ i \geq 0\,,$$

which is identical to the fitness function in Haigh's model (Haigh, 1978; Crow and Kimura, 1970). Although our definition of fitness refers to haplotypes, we also use it for genotypes of diploid individuals. Ploidy should be of no concern when assuming random mating and multiplicativity of maternal (i copies) and paternal (j copies) fitnesses, i.e., $v_{i_o,i,j} = v_{i_o,i}\ v_{i_o,j}$. Strictly, this holds only when non-linear fitness effects are neglected and in the absence of epistasis. While the former is unproblematic for small s, the effect of epistasis may not easily be hand-waved away and shall be considered in more detail elsewhere. Without epistasis, we have for copy class $i_o + i$ after random mating a total frequency $y_i(t)\sum_{j=0}^{\infty} y_j(t)$ for all genotypes having a class $i + i_o$ maternal chromosome. The selection step for these genotypes results in a total frequency

$$\sum_{j=0}^{\infty} y_i(t) y_j(t)\, v_{i_o,i+j} = y_i(t) v_{i_o,i} \sum_{j=0}^{\infty} y_j(t) v_{i_o,j}$$

before normalization. Thus, the selection strength for each maternal class $i + i_o$ genotype is proportional to the selection coefficient for the same class in the haploid case. The analogous argument holds for the paternal classes. Consequently, while the normalization factor may change, haplotype frequencies after selection are the same as in the haploid case. An analogous argument holds if we do consider multiple chromosomes: each chromosome is described by our model, if fitness is multiplicative across the copies on each chromosome.

Consider now the duplication process described by the matrix

$$\mathcal{D}_{i_o} = \left(d_{i_o+i,i_o+i+j}\right)_{i\geq 0, j\geq 0}\,.$$

It contains the probabilities d_{i_o+i,i_o+i+j} that an individual with $i_o + i$ copies has a descendant with $i_o + i + j$ copies in the next generation. Under duplication, copies are not lost. Hence, $j \geq 0$. To specify the entries of $\mathcal{D}$, we distinguish the single template (STM) and the compound copy (CCM) models. In STM only one gene, the template, can duplicate. Assuming this can happen once per generation, we have

$$\begin{aligned} d_{i_o+i,i_o+i} &= 1-d\\ d_{i_o+i,i_o+i+1} &= d\\ d_{i_o+i,i_o+i+j} &= 0,\ \forall j > 1\end{aligned} \tag{2}$$

for all $i_o > 0$ and $i \geq 0$. We call this the Bernoulli model.

Allowing that the template may produce more than one copy per generation leads to the Poisson model:

$$d_{i_o+i,i_o+i+j} = \exp(-d)\, d^j / j!\,. \tag{3}$$

For sufficiently small d both Eqs. (2) and (3) are nearly identical. On average d new copies are introduced per generation. The variance of the number of new copies is $(1-d)\,d$ under the Bernoulli model, and d under the Poisson model.

In CCM any copy may duplicate, leading to the binomial probabilities

$$d_{i_o+i,i_o+i+j} = \binom{i_o+i}{j} d^j (1-d)^{i_o+i-j}\,, \tag{4}$$

with parameters $i_o + i$ and d. For $j > i_o + i$, the binomial coefficient vanishes, so that $d_{i_o+i,i_o+i+j} = 0$, reflecting the assumption that any copy may duplicate at most once per generation. Similarly to STM, this restriction is not imposed under the Poisson approximation, where

$$d_{i_o+i,i_o+i+j} = \exp\left(-\left(i_o+i\right)\, d\right)\, \left(\left(i_o+i\right)\, d\right)^j / j!\,. \tag{5}$$

Again, both Eqs. (4) and (5) are nearly identical for small d. Both have mean $(i_o + i)\,d$ and the variance is $(i_o + i)(1-d)\,d$ under the Binomial, and $(i_o + i)\,d$ under the Poisson model.

Summarizing, for $N = \infty$ one generation cycle consists of

1. Selection/Normalization
$$\tilde{y}_i(t) = \frac{y_i(t)\, v_{i_o,i}}{\bar{v}_{i_o}(t)}\,,$$
2. Duplication
$$\tilde{\tilde{y}}_i(t) = \sum_{0\leq j\leq i} \tilde{y}_j(t) d_{i_o+j,i_o+i} = \sum_{0\leq j\leq i} \frac{y_j(t)\, v_{i_o,j}\, d_{i_o+j,i_o+i}}{\bar{v}_{i_o}(t)}\,,$$
3. Replication
$$y_i(t+1) = \tilde{\tilde{y}}_i(t)\,. \tag{6}$$

As in classical Wright–Fisher theory, random drift in finite populations ($N < \infty$) is realized by multinomial sampling with parameters $\tilde{\tilde{y}}_i(t)$ and N in step 3. This produces counts, denoted by the random variables $Y_i(t+1)$, of class i individuals in generation $t+1$. The counts are scaled back to relative frequencies by

$$y_i(t+1) = \frac{Y_i(t+1)}{N}\,. \tag{7}$$

Note, that any class $i \geq 0$ may be lost by drift if $N < \infty$. Once the class with the currently smallest copy number ($i = 0$) is lost, it cannot be re-surrected by any of the evolutionary forces considered here. Such a loss leads to an increase of I_o from i_o to $i_{\tilde{o}} > i_o$. The indices i of all classes are updated to $i - (i_{\tilde{o}} - i_o)$, such that the new class with 0 extra copies is again labeled with $i = 0$. Each such jump of the random variable $I_o(t)$ represents a click of the 'duplication ratchet'. Thus, there is always a non-empty class $i = 0$.

3. Results

Duplication–selection equilibrium. In the absence of drift ($N = \infty$) the fittest class ($i = 0$) will not be lost and the count $i_o > 0$ remains constant over time. Thus, $y_0(t)$ is strictly positive for all t. Further, without epistasis, a non-trivial duplication–selection equilibrium exists. We present the result for the haploid model with a single chromosome.

Proposition 1. *Let $0 < s, d < 1$. The frequencies $\mathbf{y}(t) = (y_0(t), y_1(t), \ldots)$ converge in ℓ^1 to an equilibrium $\mathbf{y}^* = (y_0^*, y_1^*, \ldots)$, with recursively determined components*

$$\forall i > 0: \quad y_i^* = \frac{\sum_{j=0}^{i-1} y_j^* d_{i_o+j,i_o+i}\, v_j}{d_{i_o,i_o} - d_{i_o+i,i_o+i}\, v_i} \tag{8}$$

such that $\sum_{i\geq 0} y_i^* = 1$.

A proof of this proposition is given in the supplementary material. As we have established that copy number evolution follows the same laws for frequency changes for each haplotype and chromosome, the haploid equilibria and convergence results hold also true in the diploid case, for each haplotype of each chromosome. Thus, to assess the total counts across a diploid genome with l chromosomes, we need to sum $2l$ independent random variables, with distribution given by the equilibrium frequencies above, and the minimum copy number i_o among the $2l$ chromosomes.

STM. For STM, the recursion Eq. (8) can be solved. With $d_{i_o+j,i_o+j} = 1 - d$, $d_{i_o+j-1,i_o+j} = d$, this simplifies to

$$\begin{aligned} y_j^* &= \frac{d}{(1-d)}(1-s)^{j-1}(1-(1-s)^j)^{-1} y_{j-1}^* \\ &= \frac{d}{(1-d)(1-s)}((1-s)^{-j}-1)^{-1} y_{j-1}^* \\ &= \underbrace{\left(\frac{d}{(1-s)(1-d)}\right)^j \prod_{i=1}^{j}((1-s)^{-i}-1)^{-1}}_{:=c_j} y_0^*. \end{aligned} \tag{9}$$

Since $1 = \sum_{j\in\mathbb{N}} y_j^* = y_0^* \sum_{j=0}^{\infty} c_j$, we have $y_0^* = (\sum_{j=0}^{\infty} c_j)^{-1}$. Starting from Eq. (9), we can derive by Taylor approximation an approximate equilibrium distribution for STM:

$$\begin{aligned} y_j^* &= \frac{d}{(1-s)(1-d)}((1-s)^{-j}-1)^{-1} y_{j-1}^* \\ &= \frac{d}{(1-s)(1-d)} \frac{(1-s)^j}{1-(1-s)^j} y_{j-1}^* \\ &= \frac{d}{(1-s)(1-d)s} \frac{(1-s)^j}{j+\mathcal{O}(s)} y_{j-1}^* \end{aligned}$$

For small s and d, this becomes approximately

$$y_j^* \approx \frac{d}{sj} y_{j-1}^*,$$

whose unique solution is the Poisson distribution with parameter d/s. Thus, the duplication model STM can be approximated by Haigh's model (Haigh, 1978) for the mutation–selection equilibrium for small s and d. In more detail, at equilibrium, the number of extra copies i is Poisson-distributed with parameter d/s, i.e.,

$$y_i^* = e^{-d/s}(d/s)^i/i! \quad \forall i \geq 0\,. \tag{10}$$

The approximation is very similar to the exact solution, see Table S1. Note the slight difference to Haigh's model: class 0 refers to those individuals which have no extra copies beyond the minimal number i_o. Thus, mean, variance and skewness of total copy number are

$$\begin{aligned} \mathbb{E}(i_o + I) &= i_o + \mathbb{E}(I) = i_o + d/s\,, \\ \mathbb{V}(i_o + I) &= \mathbb{V}(I) = d/s\,, \\ \mathrm{sk}(i_o + I) &= \mathrm{sk}(I) = \sqrt{s/d}\,, \end{aligned}$$

and the *Fano index* (Fano, 1947) (variance to mean ratio) is

$$\mathbb{F}(i_o + I) = \frac{\mathbb{V}(i_o + I)}{\mathbb{E}(i_o + I)} = d/(i_o \cdot s + d)\,.$$

Note that only the mean depends on i_o, but neither the variance nor the skewness, because they are invariant under shift by a constant.

An easy calculation shows that equilibrium mean fitness is

$$\bar{v} = \sum_{i\geq 0} v_i\, y_i = d_{i_o,i_o} = e^{-d} \approx 1 - d\,,$$

independent of the strength of selection. Fitness variance at equilibrium is

$$\mathbb{V}(v) = e^{-2d}\,(e^{d/s} - 1) \approx d/s\,(1 - 2d)\,.$$

If d is small compared to s, the fitness variance is approximately $d/s(1-2d) \approx d/s$ and is at the same scale as copy number variance. At equilibrium the effect of selection is balanced by duplication in each generation: after selection and standardization allele frequencies are

$$\tilde{y}_i = \frac{e^{-d/s}(d/s)^i/i!\,(1-s)^i}{e^{-d}} = e^{-d(1-s)/s}(d(1-s)/s)^i/i!\,.$$

They still follow a Poisson law, but with parameter $d(1-s)/s$. I.e., selection reduces mean and variance of copy number by the factor $1-s$ per generation. This reduction is balanced by duplication. Applying Eq. (3), one has

$$\tilde{\tilde{y}}_i = \sum_{0\leq j\leq i} \tilde{y}_j(t) d_{i_o+j,i_o+i} = e^{-d/s}(d/s)^i/i!\,,$$

returning allele frequencies to their values before selection and mean copy number to d/s. Thus, duplication shifts mean copy number by the factor $1/(1-s) \approx 1 + s$.

CCM. Consider equilibrium mean fitness first. As above, it suffices to satisfy the equilibrium condition for y_0^*

$$\bar{v} = d_{i_o,i_o} = (1-d)^{i_o} \approx e^{-d\,i_o}\,,$$

according to Eqs. (4) and (5). The equilibrium condition for class frequencies y_i^* is

$$y_i^* = \sum_{0\leq j\leq i} \frac{y_j^*\, v_j\, d_{i_o+j,i_o+i}}{d_{i_o,i_o}} = \sum_{0\leq j\leq i-1} \frac{y_j^*\, v_j\, d_{i_o+j,i_o+i}}{d_{i_o,i_o} - v_i d_{i_o+i,i_o+i}}\,. \tag{11}$$

To proceed, consider the ratio y_1^*/y_0^*. With duplication according to Eq. (5) this becomes

$$\begin{aligned} \frac{y_1^*}{y_0^*} &= \frac{d_{i_o,i_o+1}}{\bar{v} - v_1\, d_{i_o+1,i_o+1}} = \frac{e^{-d\,i_o} d\, i_o}{e^{-d\,i_o} - (1-s)\, e^{-d(i_o+1)}} \\ &\approx \frac{d\, i_o}{1-(1-s)(1-d)} \approx \frac{d\, i_o}{d+s}\,, \end{aligned} \tag{12}$$

where the last approximation holds if both s and d are sufficiently small. Similarly, we compute

$$\frac{y_2^*}{y_1^*} = \frac{\binom{i_o}{2} d^2 \frac{y_{i_o}^*}{y_{i_o+1}^*} + (i_o+1)d(1-d)^2(1-s)}{(1-d)^2\,(1-(1-d)^2(1-s)^2)}\,.$$

Still, letting s and d be sufficiently small, and retaining only terms of at most linear order in s and d in numerator and denominator, we get

$$\frac{y_2^*}{y_1^*} \approx \frac{d\,(i_o+1)}{2(d+s)}\,,$$

Table 1
Mean $\mathbb{E}$, Variance $\mathbb{V}$ and dispersion $\mathbb{F} = \mathbb{V}/\mathbb{E}$ for STM and CCM of random variables I (extra copy number) and J (total copy number) in terms of parameters i_o, d and s. Note that J is expected to be under-dispersed ($F < 1$) in STM.

	STM		CCM	
	I	J	I	J
$\mathbb{E}$	d/s	$i_o + d/s$	$i_o\ d/s$	$i_o\ (1 + d/s)$
$\mathbb{V}$	d/s		$i_o\ (d/s)\ (1 + d/s)$	
sk	$\sqrt{s/d}$		$\frac{1+d/(d+s)}{\sqrt{i_o\, d/(d+s)}}$	
$\mathbb{F}$	1	$\frac{d/s}{i_o+d/s}$	$1 + d/s$	d/s

and generally,

$$\frac{y_i^*}{y_{i-1}^*} \approx \frac{d\,(i_o + i - 1)}{i\,(d+s)}\,.$$

Upon iteration, one finds

$$\frac{y_i^*}{y_0^*} = \prod_{j=0}^{i-1} \frac{y_{j+1}^*}{y_j^*} \approx \prod_{j=0}^{i-1} \frac{d\,(i_o + j)}{(j+1)\,(d+s)} = \left(\frac{d}{d+s}\right)^i \binom{i_o + i - 1}{i}\,.$$

Summation yields

$$\sum_{i\geq 0} \frac{y_i^*}{y_0^*} = \left(\frac{s}{d+s}\right)^{-i_o}$$

and therefore

$$y_0^* = \left(\frac{s}{d+s}\right)^{i_o}\,.$$

Thus, the equilibrium frequencies of CCM are well approximated by a negative binomial law NB with parameters i_o and $s/(d+s)$ (Table S1). In short,

$$\begin{aligned} y_i^* &= \left(\frac{d}{d+s}\right)^i \left(\frac{s}{d+s}\right)^{i_o} \binom{i_o + i - 1}{i} \\ &= \left(\frac{d}{d+s}\right)^i \left(\frac{s}{d+s}\right)^{i_o} \binom{i_o + i - 1}{i_o - 1} \end{aligned} \tag{13}$$

for all $i \geq 0$. In terminology of the negative binomial distribution y_i^* is the probability to draw i_o copies which all individuals have ('failures') before one draws the ith extra copy ('success'), given a success probability of $d/(d+s)$. As expected intuitively, the number of extra copies is large compared to i_o, if d is large compared to s, and vice versa.

Mean extra copy number is $i_o\, d/s$. Mean, variance and skewness of the total copy number at equilibrium are

$$\begin{aligned} \mathbb{E}(i_o + I) &= \sum_{i=0}^{\infty} (i_o + i) y_i^* = i_o\, \frac{d+s}{s} \\ \mathbb{V}(i_o + I) &= i_o\, \frac{d}{s}\, \frac{d+s}{s} \\ \mathrm{sk}(i_o + I) &= \frac{1 + d/(d+s)}{\sqrt{i_o\, d/(d+s)}}\,. \end{aligned}$$

The variance to mean ratio is

$$\mathbb{F}(i_o + I) = \mathbb{V}(i_o + I)/\mathbb{E}(i_o + I) = \frac{d}{s}\,,$$

independent of i_o, unlike the moments above (see Table 1).

It is worth noting that the random variable $J = i_o + I$ with $I \sim \mathrm{NB}(i_o, p)$, $p = s/(s+d)$ and p small, is well approximated by a Gamma-distributed random variable $K \sim \Gamma(a, b)$ with $a = i_o/(1-p) = i_o(d+s)/d$ and $b = p/(1-p) = s/d$. In fact, K has the added beauty that it counts the total number of gene copies (which can be directly observed in data) and it is not shifted by a constant. Mean copy number is $\bar{K} = \bar{J} = i_o(d+s)/s$ and the Fano index is $\mathbb{F}(K) = \mathbb{F}(J) = d/s$. Thus, mean and variance of J and K are identical, but they differ in their skewness: they are $\mathrm{sk}(J) = (2-p)/\sqrt{i_o(1-p)}$ and, slightly smaller, $\mathrm{sk}(K) = 2\sqrt{(1-p)/i_o} = 2(1-p)/\sqrt{i_o(1-p)}$. The Gamma-approximation is inaccurate if p is not close to 0 or, equivalently, if s is large compared to d. As the sum of independent Poisson random variables, and also of independent negative binomial distributions and Gamma distributions with equal success probabilities/second parameter are again Poisson/negative binomial/Gamma-distributed random variables, these approximation also hold for the diploid and multi-chromosome case (by adding up the per-chromosome, per-haplotype parameters accordingly).

Epistasis. Presence of epistasis ($\beta \neq 1$) renders Eqs. (8) and (11) considerably more complicated. Duplication rate and fitness both depend on both variables i_o and i (see Eq. (1)). A selection–duplication equilibrium is not readily deduced analytically. Therefore, we resort to simulations (see below) and a qualitative description. Consider the fitness ratio of classes $i+1$ and i

$$v_{i_o,i+1}/v_{i_o,i} = (1-s)^{(i+i_o+1)^\beta - (i+i_o)^\beta}\,.$$

It holds for all $i_o \geq 1$

$$\lim_{i\to\infty} \frac{v_{i_o,i+1}}{v_{i_o,i}} = \begin{cases} 1 & \text{if } 0 < \beta < 1 \quad \text{(case 1)} \\ 1 - s & \text{if } \beta = 1 \quad \text{(case 2)} \\ 0 & \text{if } 1 < \beta \quad \text{(case 3)} \end{cases} \tag{14}$$

With synergistic epistasis ($\beta > 1$) and increasing i, fitness of class $i+1$ becomes vanishingly small compared to fitness of class i, and thus negative selection against further duplicates becomes increasingly strong. For $\beta = 1$ the ratio remains constant at $0 < 1 - s < 1$, independent of i_o and i. Hence, at equilibrium, we expect $\mathbb{E}_{\beta>1}(J) < \mathbb{E}_{\beta=1}(J)$. Since $\beta = 1$ admits a finite duplication–selection equilibrium in both models STM and CCM, such an equilibrium must also exist for $\beta > 1$. In contrast, under diminishing epistasis ($\beta < 1$) the fitness ratio between classes $i+1$ and i tends towards 1 with increasing i. Hence, with increasing i selection becomes less and less effective to counteract duplication, eventually driving mean copy number to infinity.

The effect of epistasis is more drastic in finite populations, since random drift induces a ratchet effect leading to eventual loss of the least loaded class, i.e. of the class with fewest gene copies, and to a steady increase of i_o. Without epistasis ($\beta = 1$) this happens at constant speed in the STM model. In contrast, in the CCM model the ratchet accelerates with increasing copy number due to the increasing duplication rate. This acceleration can – to some extent – be counteracted by synergistic epistasis ($\beta > 1$). On the other hand, diminishing returns ($\beta < 1$) always exacerbates the acceleration.

Finite population size. The discrete Markov process $\boldsymbol{Y}(t) = \left(Y_i(t)\right)_{i\geq 0}$ is governed by a multinomial transition law

$$\boldsymbol{Y}(t) \sim \mathrm{Mult}\left(N, (\tilde{\bar{y}}_i(t))_i\right) \tag{15}$$

with parameters $N < \infty$ and $(\tilde{\bar{y}}_i(t))_i$ defined in Eq. (6). In the course of one generation counts $(Y_i(t))_i$ change by duplication, selection and multinomial sampling. In every generation t there is a non-zero probability that the least loaded class, i.e. the class I_o with the fewest number of gene copies, is lost by drift. In other words, the random variable $I_o(t)$ increases over time in a ratchet-like process. Let T_i be a hitting time: the first time t_i^* when $I_o(t) = i$. This means that from t_i^* onwards all individuals carry at least i copies . We call $R_i = \mathbb{E}(T_i - T_{i-1})$ the (average) *interarrival time* between consecutive 'clicks'. The *speed* of the ratchet is the inverse of the interarrival time.

In Haigh's model (Haigh, 1978) this speed is constant and does not depend on i (except for the time to the first click, which depends on initial conditions). Haigh had derived an approximation and pointed out that the size of the currently fittest class is the major determinant of the interarrival time. However, outside of a narrow parameter range, his approximation is quite imprecise. Essentially, the problem

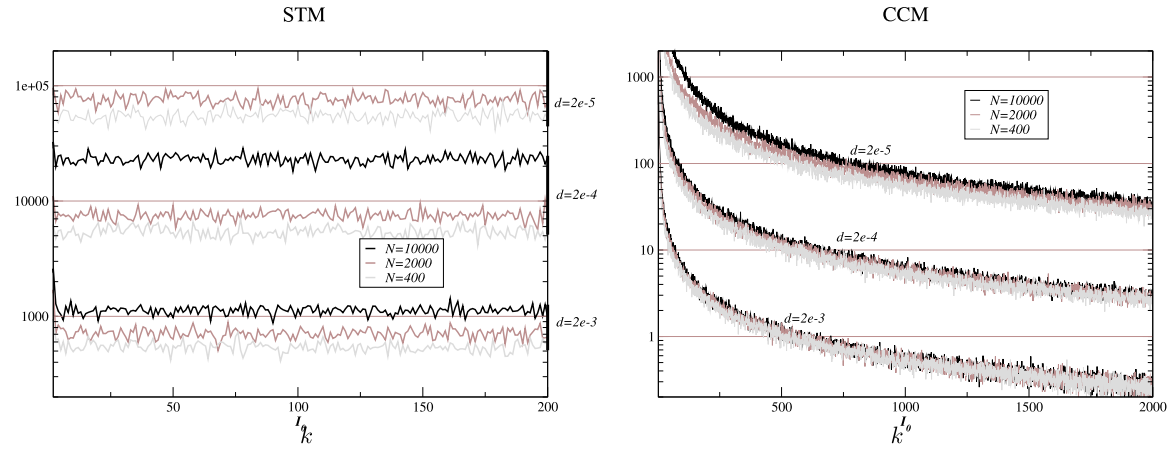

Fig. 1. Mean interarrival times R_k (across 100 iterations) *vs.* ratchet 'clicks' k. Without epistasis ($\beta = 1$). Simulation results for STM (left) and CCM-models (right). For all cases: $s = 2 \cdot 10^{-4}$. Duplication rate from $d = 2 \cdot 10^{-5}$ to $d = 2 \cdot 10^{-3}$, population size from $N = 400$ to $N = 10{,}000$ (see legend in panels). Case STM and $N = 10{,}000$ is not shown because interarrival times were exceedingly long in our simulation.

remains still open despite several attempts of improved approximations (e.g., Stephan et al., 1993; Neher and Shraiman, 2012).

Here, we investigated interarrival times with computer simulations and compared the effect of epistasis in STM and CCM models. We traced the stochastic process $(\mathbf{Y}(t))_t$ starting from the initial distribution $\mathbf{Y}(0)$ with $Y_0(0) = N$ and $Y_i(0) = 0 \; \forall \; i > 0$.

We terminate a simulation run at T_{2000}, or when $t = 50$ million generations are reached, whatever comes first. Given the equivalence of STM (without epistasis) with Haigh's model, it is not surprising that the interarrival times are constant in this case (Fig. 1A). In contrast, under CCM R_i quickly decrease with increasing i (Fig. 1B), reflecting the increasing effective duplication rate when the number of copies grows. Besides their strong dependence on the relative magnitudes of d and s, interarrival times increase with increasing population size.

In both STM and CCM the ratchet is slowed down under synergistic ($\beta > 1$) epistasis. If sufficiently strong, the ratchet can even come to a stop. Operationally, this means that mean copy number $\mathbb{E}(J(t))$ remains bounded on a time scale of $O(10^6)$. To be conservative, we ran extra simulations for 10M generations, recorded the population status every 20 generations, and explored the effect of epistasis ($\beta \in [1.5, 2.1]$) on slowing the ratchet (Figure S1). For instance, for $\beta = 2$ the ratchet came to a stop on the time scale of our simulation with values of $\mathbb{E}(J)$ from 1.12 (for $s = 0.001$, $d = 0.0001$, strong selection weak duplication, $d/s = 0.1$) to 49.71 (for $s = 0.0002$, $d = 0.0005$, weak selection strong duplication, $d/s = 2.5$). In contrast, epistasis with $\beta = 1.5$ was not sufficient to halt the ratchet. Even with $\beta = 1.7$ duplications accumulate and T_{1000} is quickly reached. However, with $\beta \geq 1.8$ a slow-down is observed at about 0.3 – 0.5M generations, and the ratchet comes effectively to a halt at about 2–4M (see Fig. 2). The threshold value of β between runaway expansion and stopped ratchet generally lies between 1.7 and 1.8 but depends on the value of s, d and N. The relationship between these parameters is potentially highly complex and non-linear, and is thus beyond the scope of this study.

Interestingly, a regime of synergistic epistasis may lead in CCM to a transient behavior of the ratchet: initial slow down, but ultimate acceleration (Fig. 2). This suggests the existence of a copy number threshold beyond which the dynamics enters irreversibly a regime of what can be called 'run-away evolution'.

Parameter estimation and model selection. Given a sample of n counts $\boldsymbol{J} = (J_1, J_2, \ldots, J_n) = (I_o + I_1, I_o + I_2, \ldots, I_o + I_n)$ one can define the following estimators for the parameters of the Poisson distribution associated to STM (with parameter λ) and of the negative binomial distribution associated to CCM (with parameters i_o and $p = s/(d + s)$). Let $\bar{\boldsymbol{J}}$, $\mathbb{V}(\boldsymbol{J})$ and $\mathbb{F}(\boldsymbol{J}) = \mathbb{V}(\boldsymbol{J})/\bar{\boldsymbol{J}}$ be sample mean, variance and dispersion, respectively, and

$$\begin{aligned} \hat{\lambda}_1 &= \bar{\boldsymbol{J}}\,, \\ \hat{\lambda}_2 &= \mathbb{V}(\boldsymbol{J})\,, \\ \hat{p} &= 1/(1 + \mathbb{F}(\boldsymbol{J})) \\ \hat{i_o} &= \bar{\boldsymbol{J}}\;\hat{p}\,. \end{aligned}$$

Note that the relationship $\mathbb{F}(\boldsymbol{J}) = (1 - p)/p = d/s$ yields an estimate of the ratio d/s for CCM.

Plausibility of STM may easily be checked by asking whether $\hat{\lambda}_1 \approx \hat{\lambda}_2$. Also, it holds that an under-dispersed negative binomial (i.e., $p < 1$) converges to a Poisson distribution with parameter $\lambda = i_o\, d/s$ when letting $p \uparrow 1$ and keeping the mean $i_o\, d/s$ of the negative binomial constant. If $\boldsymbol{J}$ were (nearly) Poisson-distributed with parameter $i_o\, d/s$, one still would expect its dispersion to be at most one. Thus, observing a dispersion larger than one is incompatible with a STM model.

As mentioned above, the (compound) random variable $J = I_o + I$ with $I_o = i_o$ and $I \sim \mathrm{NB}(i_o, p)$ is well approximated by a Gamma-distributed variable K with $K \sim \Gamma(a, b) = \Gamma(i_o(1 + s/d), s/d)$. It is straightforward to estimate parameters a, b from sample $\boldsymbol{J}$. We have

$$\hat{a} = \bar{\boldsymbol{J}}^2/\mathbb{V}(\boldsymbol{J}) = \bar{\boldsymbol{J}}/\mathbb{F}(\boldsymbol{J})\,, \tag{16}$$

$$\hat{b} = \bar{\boldsymbol{J}}/\mathbb{V}(\boldsymbol{J}) = 1/\mathbb{F}(\boldsymbol{J})\,. \tag{17}$$

By definition, $\bar{K} = \bar{\boldsymbol{J}}$ and $\mathbb{F}(K) = \mathbb{F}(\boldsymbol{J})$.

An alternative way to both distinguish between CCM and STM models and estimate parameters d/s and i_o is by maximum-likelihood grid search. Given a sample of gene copy counts, we compute the log-likelihood of the sample for $i_o \in \{1, 2, \ldots, m\}$, where m is the minimum of the counts in the sample and with $d/s \in \{1/100, 1/99.75, 1/99.5, \ldots, 1/1.25, 1, 1.25, 1.5, 1.75, \ldots, 99.75, 100\}$. For distinguishing between CCM and STM, we adopt the Kass–Raftery scale (Kass and Raftery, 1995) interpretation of likelihood quotients (as these are technically Bayes factors). To be precise, if the log-likelihood is > 1 (< -1), we have positive evidence for a better fit of CCM (STM). A log-likelihood above 3 (below -3), will be interpreted as strong evidence for the favored model. Seen as Bayesian model selection with equal priors for STM and CCM, this means that the posterior probability of the other hypothesis is at most 0.05. In that same line of thought, we can also control that the 'family-wise error', i.e. the summed probabilities of the 'other' hypotheses, across multiple comparisons to be below some (significance) threshold (which we will set at 0.05). See Section Suppl 4 for details.

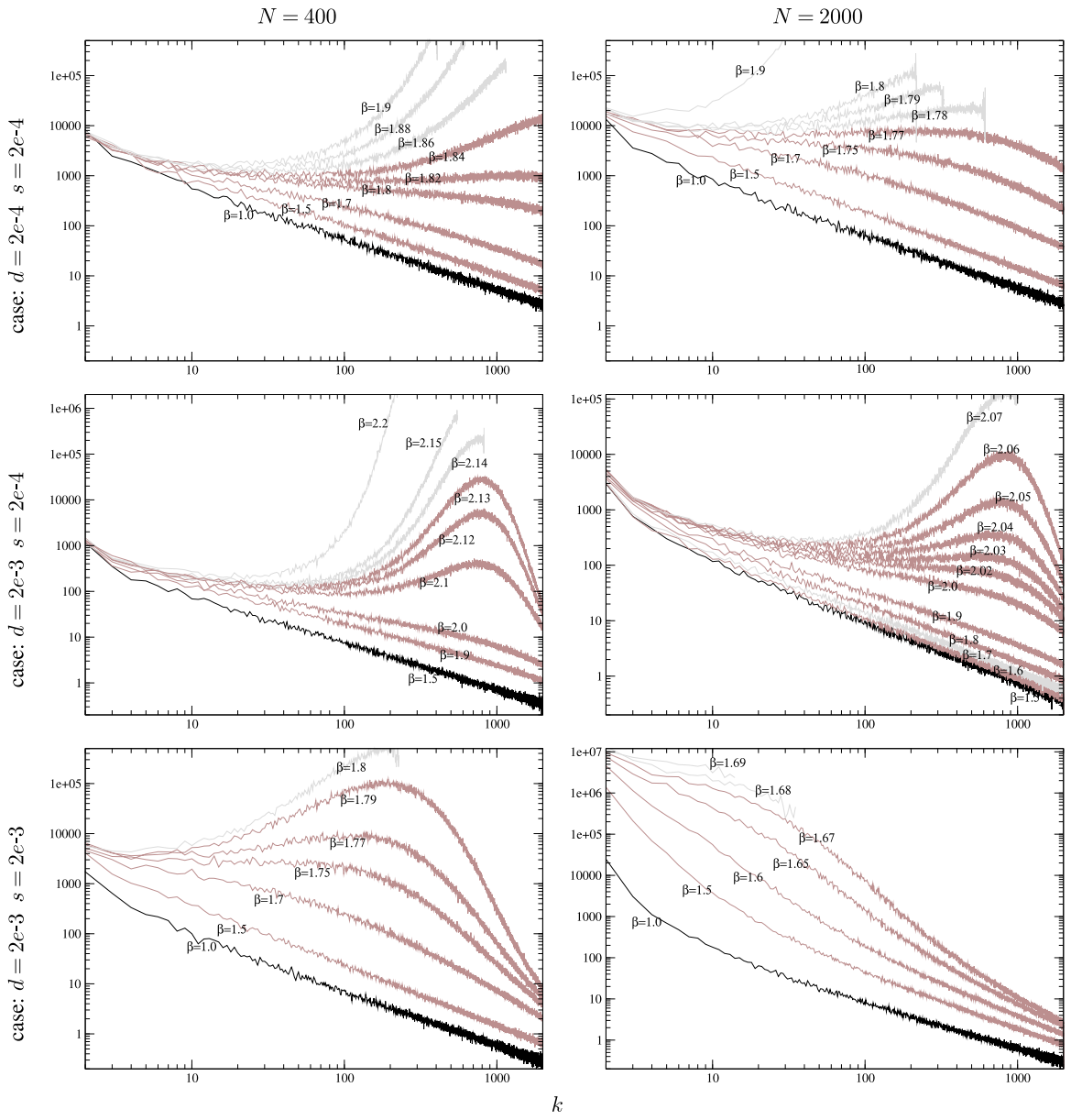

Fig. 2. Mean interarrival times R_k *vs.* ratchet 'clicks' k. CCM model with epistasis ($\beta \geq 1$). Mean interarrival times calculated across 100 iterations unless the max number of generations ($2 \cdot 10^7$) was exceeded. However, in all cases a minimum number of 10 iterations was taken to calculate the mean. This is why some trajectories terminate before 2000 clicks were reached. Left: $N = 400$. Right: $N = 2000$. Top: $d = s = 2e{-}4$, middle: $d = 2e{-}3 \quad s = 2e{-}4$, bottom: $d = s = 2e{-}3$. Initial configuration for all simulation runs: $Y_1(0) = N, \; Y_i(0) = 0 \; \forall i > 0$.

Application I: NLR genes in Danio rerio. Schäfer et al. (2024) determined copy numbers of the huge NLR gene family in two lab and four wild populations of zebrafish (*Danio rerio*). NLR genes are divided into groups, characterized by certain protein domain signatures. The so-called 'Fisnacht' domain is characteristic of all zebrafish NLR genes, while the 'B30.2' domain occurs only in some of the groups. We reproduce the gene counts, based on capture of either of the domains, in Supplementary Table S2. Summary statistics of mean, variance and dispersion, together with the estimates of d/s and i_o, both from moment estimates and from maximum likelihood estimation, are shown in Table 2 below. The parameters a and b are estimated from the data as described above.

Except for population CGN, we observe over-dispersed NLR copy counts. In most populations the Fano index is substantially larger than one. This observation is compatible with CCM, but not with STM. This is in agreement with the maximum likelihood grid search, where we find no evidence for STM (all log-likelihood ratios larger than −1) for either gene family, but at least positive evidence for all populations but one (log-likelihood ratio larger than 1) and strong evidence for all but two (log-likelihood ratio larger than 3) populations across both gene families (Table 2), even when controlling across all tests simultaneously (log-likelihood threshold < 5.64, see Section Suppl 4). Moreover, moment and maximum likelihood estimates agree very well (Table 2) and there is an overall good fit of the CCM model to the data (Fig. 3) yielding reliable parameter estimates. The discrepancy in dispersion between the lab population CGN and the other populations is most likely to be attributed to population history. CGN is a heavily inbred population of small effective size, established more recently than, and derived from, the other lab population in our sample (TU). This could have led to a strong founder effect due to a small number of founding individuals with a particularly large copy number.

Application II: Gene families in three populations of human. Brahmachary et al. (2014) reported a survey of copy number variation in about

Table 2
Summary statistics calculated from counts of NLR genes in four populations of *Danio rerio*. Lab strains: CGN and TU; wild populations: CHT, DP, KG and SN. Column n refers to the number of individuals which were typed. Further columns: mean, variance and minimum copy number observed. Estimates of parameters a and b of a Gamma distribution as given in Eqs. (16) and (17). Minimum copy number $\hat{i}_o$ is estimated as $\hat{a}/(1+\hat{b})$. Note that $1/\hat{b}$ is an estimator of the ratio d/s. Moreover, we report the maximum likelihood estimates of d/s and i_o for CCM, as well as the log likelihood of CCM over STM.

Pop.	n	Fisnacht								
		$\bar{J}$	$\mathbb{V}(J)$	$\min(J)$	$\hat{a}$	$1/\hat{b}$	$\hat{i}_o$	$(\widehat{d/s})_{\text{ml}}$	$(\hat{i}_o)_{\text{ml}}$	log L
CGN	8	271.9	214.6	245	344.5	0.789	151.962	0.8	151	1.54
TU	8	176.6	4828.7	78	6.459	27.343	6.231	28.5	6	174
CHT	18	394.1	6064.4	167	25.611	15.388	24.048	19.75	19	1364
DP	20	273.9	4468.8	61	16.788	16.315	15.818	26.5	10	1216
KG	20	400.8	7179.7	223	22.374	17.913	21.191	21.25	18	1005
SN	19	441.6	4842.2	335	40.273	10.965	36.907	11.25	36	404
wp[a]	77	376	9647.0	61	14.655	25.657	14.105	33.25	11	12 491
		B30.2								
		$\bar{J}$	$\mathbb{V}(J)$	$\min(J)$	$\hat{a}$	$1/\hat{b}$	$\hat{i}_o$	$(\widehat{d/s})_{\text{ml}}$	$(\hat{i}_o)_{\text{ml}}$	log L
CGN	8	83.5	34.8	73	668	0.417	58.937	0.44	58	−0.12
TU	8	54.1	360.1	28	433	6.656	7.066	8	6	16.36
CHT	18	129.9	770.9	66	2338	5.935	18.732	7	16	45.4
DP	20	62.9	457.3	12	1257	7.270	7.606	9.5	6	42.8
KG	20	135.1	1382.8	60	2702	10.235	12.025	14	9	120.8
SN	19	135.3	489.5	102	2571	3.618	29.299	3.5	30	21
wp[a]	77	115.2	1743.9	12	8868	15.138	7.138	22	5	565.6

[a] wp = all wild populations (CHT, DP, KG, SN) pooled.

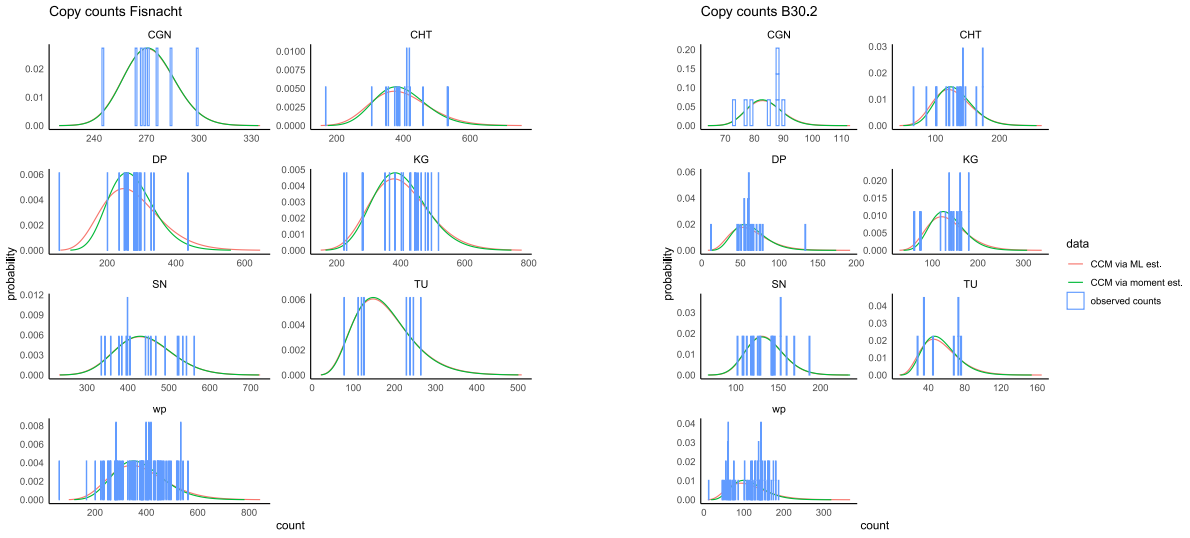

Fig. 3. Empirical gene copy count distribution (blue bars) and lineplots of the probabilities of counts for the fitted CCM distribution (using maximum likelihood or moment-based parameter estimators), for two gene domains (Fisnacht, B30.2) from samples of lab (CGN,TU) and wild populations (CHT,DP,KG,SN) of *D. rerio*. wp: wild populations pooled. The observed counts are listed in Table S2.

150 gene families (Suppl Table S3) in the three human populations YRI, CEU and CHB. Excluding loci with little variation (less than five copies in YRI, less than two copies in any of the populations), genes on sex chromosomes, and microsatellites, we selected a subset of 52 gene families (Table 4) for which we determined sample statistics and used our maximum likelihood approach to perform model selection and to estimate minimum copy number i_o and the ratio d/s under the better fitting model from CCM and STM (Fig. 4, as explained above). The gene families tend to fit better to the STM model than the CCM model in all three populations, albeit 19%–38% of gene families not clearly supporting one model over the other (Table 3). While gene families do not always fit to the same model in each population, the model preference, measured as the log likelihood ratio, is strongly correlated between families (Spearman/Pearson correlations of 0.59/0.98 for YRI:CEU, 0.64/1 for YRI:CHB, 0.79/0.99 for CEU:CHB; we report additionally Spearman rank correlation as the log likelihood ratios spread unevenly across their range). Most human gene families we analyzed tend to average below 50 copies, thus clearly fewer copies than the NLR genes in *D. rerio*. There is a positive correlation between average copy number and log likelihood ratios (Fig. 5): Across populations, Pearson correlations are at least 0.89, while rank (Spearman) correlations are at least 0.67. This indicates a tendency of larger families to fit CCM better. In particular, the largest gene families with more than 100 copies (DUX4 (Leidenroth and Hewitt, 2010), USP17, NBPF10, REXO1L1) all have strong support for the CCM model (log BF > 8.05, the threshold for strong evidence corrected for multiple testing), as shown in Fig. 5.

In general, there is a positive correlation between estimated d/s ratios and gene copy number both for gene families fitting better to STM and CCM across all populations (CCM: >0.89 (Pearson)/>0.56 (Spearman); STM: 0.71/>0.73 (Spearman)), even though not for all large gene families a high $\frac{d}{s}$ is estimated. In contrast, there is a clearly weaker correlation of minimum copy number (i_o) estimates with mean copy numbers in CCM (Pearson: 0.016–0.54, Spearman: 0.35–0.73), compared to STM (Pearson: > 0.95, Spearman: > 0.77).

When we consider the variance to mean ratio, the maximum likelihood ratios do coincide well with the maximum likelihood estimates

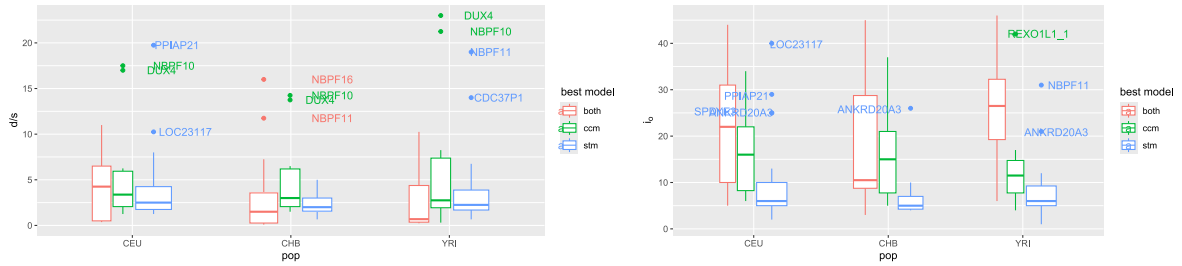

Fig. 4. Left: boxplot of d/s estimates (log-scaled ordinate). Right: boxplot of i_o estimates in 52 human gene families and three populations, listed in Table 4.

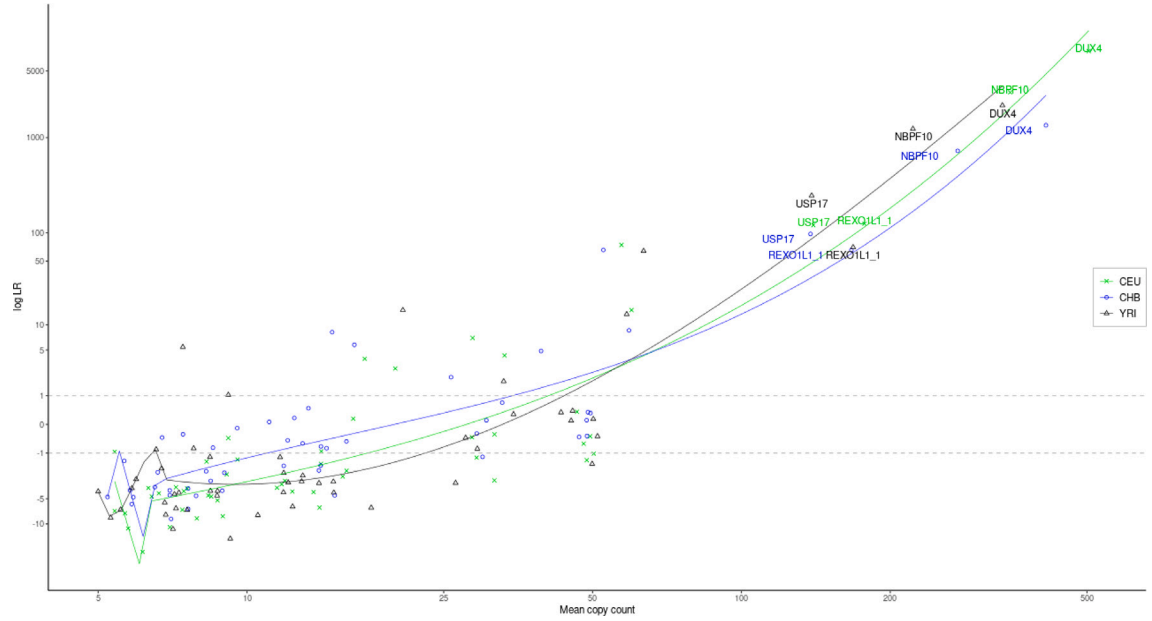

Fig. 5. Comparison of average copy number with log likelihood ratios between CCM and STM fit of gene families observed in three human populations (YRI, CEU, CHB). We labeled all families with average copy numbers above 100. The fitted lines are spline interpolations (degree 3, knots at quantiles 0,0.05,0.1, ...,1). The gray dotted lines highlight our cutoffs for positive evidence for either model (STM < -1, CCM > 1).

Table 3
Proportion of selected model (CCM, STM or no positive evidence for either) of gene families in three human populations (YRI, CEU and CHB; rounded to two digits). We consider a Bayes factor with $|\log(BF)| > 1$ positive evidence.

Population	CCM	STM	No evidence
YRI	0.19	0.62	0.19
CEU	0.19	0.63	0.17
CHB	0.19	0.42	0.38

of d/s, i.e. the maximum likelihood differ at most by 24% from the variance to mean ratio (in CEU; ≤28%, ≤18% in CHB, YRI). This adds to the plausibility of CCM for these gene families. On the contrary, gene families with better fit to STM do not show a variance to mean ratio of approximately 1, but often much less (median/95%-quantile for variance to mean ratio for CEU: 0.14/0.48, for CHB: 0.17/0.36, for YRI: 0.14/0.48). Thus, despite fitting the data better than CCM, it remains unclear if STM can truly explain the observed copy numbers in these cases.

4. Discussion

We considered an evolutionary model where new copies of a gene arise by duplication and where selection favors chromosomes with low copy number. In the so-called single template model (STM) there is only one gene which can act as template for new copies. In the compound copy model (CCM) any copy may duplicate again. Re-interpreting mutations as new copies, STM is almost identical to Haigh's (Haigh, 1978) model of mutation–selection balance (when population size is infinite) and of Muller's ratchet (when it is finite). We have shown that STM admits a finite equilibrium distribution with mean copy number i_o+d/s, where i_o is the minimum number of copies – i.e., the number present in all individuals – and d/s is the average number of extra copies — i.e. those which are copy-number-variable in the population. Under the assumptions of infinite population size and absence of epistatic interactions this number is determined by the ratio of duplication rate d and strength of (purifying) selection s.

For CCM we have shown that the limiting equilibrium distribution is essentially a negative binomial with parameters i_o and $s/(d+s)$ and mean copy number $i_o(1+d/s)$. The average number of extra copies is given by the product $i_o\ d/s$. For $i_o = 1$ both STM and CCM have on average the same number of extra copies. While remaining constant in STM, it linearly increases in i_o for CCM. For gene copy counts observed in a sample, one can assess the plausibility of the two models by several arguments. From the equilibrium distributions follows that overdispersion ($\mathbb{F}(\boldsymbol{J}) > 1$) is counterindicative of STM, while compatible with CCM. The parameters d/s and i_o can be estimated from moments of the hypothesized equilibrium distributions. One can also directly use a likelihood ratio approach for model selection and a maximum likelihood approach for parameter estimation.

We illustrated this on two data sets of gene copy number variation (gCNV): (a) the NLR genes in *D. rerio* (Schäfer et al., 2024) and (b)

Table 4
Gene families in three human populations (CEU, CHB and YRI). Data from Brahmachary et al. (2014) and adapted by Otto et al. (2024). Columns: mean $\bar{J}$ and variance $\mathbb{V}(J)$ of observed copy number. Log likelihood ratios ($\log(LR)$) between CCM and STM, and maximum likelihood estimates of i_o and d/s. Underlined $\log(LR)$ denote gene families with strong support after correction for all (3×52) comparisons in this table (see Sect. Suppl 4, threshold $\pm$ 8.05).

Gene name	CEU					CHB					YRI				
	$\bar{J}$	$\mathbb{V}(J)$	log(LR)	$\hat{i_o}$	$\hat{d/s}$	$\bar{J}$	$\mathbb{V}(J)$	log(LR)	$\hat{i_o}$	$\hat{d/s}$	$\bar{J}$	$\mathbb{V}(J)$	log(LR)	$\hat{i_o}$	$\hat{d/s}$
AMY1A	9.10	7.24	−2.34	3	6.00	11.09	11.45	0.06	5	1.25	8.70	3.57	−4.57	5	3.75
ANKRD20A3	29.13	3.20	−1.23	25	4.25	29.96	2.73	−1.19	26	4.00	26.42	3.40	−3.10	21	5.50
BOLA2B	6.63	0.88	−4.27	5	1.75	6.51	0.71	−3.53	5	1.50	7.18	1.03	−6.58	5	2.25
CBWD3	12.35	0.91	−4.04	10	2.25	12.96	3.13	−0.58	10	3.00	12.12	1.02	−3.06	10	2.00
CDC37P1	19.95	31.74	2.84	7	1.75	16.49	31.48	5.80	5	2.25	14.98	16.83	−4.18	1	14.00
CLEC18A	8.35	1.49	−4.54	6	2.25	7.89	2.06	−4.59	5	3.00	7.80	1.89	−0.79	6	1.75
CSH	7.18	0.42	−3.55	6	1.25	7.42	0.34	−0.27	7	0.44	6.72	0.31	−1.90	6	0.67
DEFA1	7.92	2.99	−8.65	4	4.00	6.98	2.75	−3.91	4	3.00	7.42	7.16	5.46	4	0.80
DEFB130	5.40	0.35	−7.05	4	1.50	5.22	0.31	−4.74	4	1.25	5.00	0.44	−4.04	4	1.00
DUX4	504.17	7703.90	8055.19	28	17.00	413.07	5560.65	1348.64	28	13.75	337.23	8608.96	2178.12	14	23.00
FAM72A	7.57	0.83	−3.68	6	1.50	7.60	0.52	−3.69	6	1.50	6.85	0.37	−7.83	5	1.75
FAM75A1	11.93	2.20	−2.90	9	3.00	13.31	4.22	0.48	10	0.33	11.87	2.39	−2.22	9	2.75
FAM75A5	11.50	1.47	−3.60	9	2.50	12.47	2.94	0.17	10	0.25	11.67	1.31	−1.21	10	1.75
FAM90A	57.20	288.50	74.37	8	6.25	52.58	282.07	65.80	7	6.50	63.42	295.94	64.19	11	4.75
FCGBP	5.67	0.56	−7.52	4	1.75	5.80	1.12	−3.90	4	1.75	5.30	1.74	−8.49	3	2.25
FOXD4L2	13.63	1.02	−4.11	11	2.75	14.49	3.71	−0.78	11	3.50	12.97	1.08	−2.43	11	2.00
GOLGA6L9	28.52	6.63	−0.36	22	6.50	29.16	6.95	−0.25	22	7.25	27.68	7.07	−0.38	20	7.75
GOLGA8G	31.67	7.55	−0.27	24	7.75	30.49	5.44	0.10	26	0.17	29.23	9.57	−0.80	19	10.25
GUSBP1	15.90	6.84	−2.06	8	8.00	13.98	4.98	−2.03	9	5.00	12.90	5.04	−2.96	8	5.00
HIST2	8.72	0.41	−5.24	7	1.75	8.91	0.58	−3.94	7	2.00	8.43	0.45	−3.96	7	1.50
KIR2DL2	9.57	2.52	−1.34	7	2.50	9.56	1.71	−0.09	8	1.50	9.17	2.82	1.04	7	0.31
KIR2DL5A	8.47	1.61	−4.68	6	2.50	8.53	1.44	−0.76	7	1.50	8.42	1.20	−1.20	7	1.50
KIR2DS2	9.17	2.34	−0.39	7	2.25	9.00	1.45	−2.21	7	2.00	8.72	2.14	−3.98	6	2.75
LIMS3	5.40	0.24	−0.94	5	0.40	5.64	0.33	−1.41	5	0.67	5.85	0.20	−3.66	5	0.80
LOC23117	50.28	8.92	−1.03	40	10.25	48.64	7.96	0.11	42	0.16	50.17	13.09	0.14	40	0.25
LOC653606	6.42	0.59	−4.68	5	1.50	5.89	1.01	−4.78	4	2.00	6.55	0.39	−0.83	6	0.57
MUC12	14.12	4.10	−1.59	10	4.00	12.09	3.40	−0.47	9	3.00	11.85	6.71	−4.14	5	6.75
NBPF10	349.57	5854.49	2980.85	19	17.50	273.84	3967.45	726.95	18	14.25	222.33	5606.36	1236.93	10	21.25
NBPF11	47.97	10.03	−0.59	37	11.00	48.71	11.30	−0.32	37	11.75	49.90	17.55	−1.59	31	19.00
NBPF16	46.50	26.56	0.36	31	0.50	46.96	16.27	−0.35	31	16.00	45.25	22.63	0.09	30	0.50
NPIP	49.45	5.30	−0.33	44	5.50	48.93	4.93	0.34	45	0.09	51.17	4.89	−0.33	46	5.25
PGA3	6.15	2.64	−20.59	2	4.25	8.44	1.84	−2.91	6	2.50	7.08	1.37	−11.47	4	3.00
PPIAP21	48.67	18.46	−1.37	29	19.75	49.51	15.03	0.31	39	0.27	43.17	14.01	0.34	33	0.31
PRAMEF14	11.75	5.24	−3.23	7	4.75	11.87	3.89	−1.72	8	3.75	10.52	1.54	−7.90	7	3.50
PRAMEF20	7.45	0.62	−4.00	6	1.50	7.60	0.93	−6.68	5	2.50	7.28	0.41	−4.18	6	1.25
PRAMEF5	16.40	5.77	0.15	12	0.36	15.89	6.83	−0.51	9	7.00	17.83	3.19	−6.45	12	5.75
PRAMEF8	5.75	1.75	−11.28	3	2.75	5.84	0.73	−5.83	4	1.75	5.97	0.51	−2.75	5	1.00
PRR11	8.28	0.99	−1.45	7	1.25	8.27	0.61	−2.09	7	1.25	6.82	0.86	−5.58	5	1.75
PRR20A_1	17.30	28.59	3.85	6	2.00	14.87	30.89	8.24	6	1.50	20.67	47.07	14.70	6	2.50
PSG3	15.62	1.49	−2.51	13	2.50	15.04	3.32	−4.53	10	5.00	14.95	1.81	−2.98	12	3.00
REXO1L1_1	177.30	691.54	124.65	34	4.25	166.91	580.72	65.56	37	3.50	168.25	497.89	70.46	42	3.00
RGPD1	14.02	1.75	−6.42	10	4.00	14.11	1.78	−0.70	12	2.00	13.98	0.69	−3.12	12	2.00
SLX1B	6.98	1.44	−10.91	4	3.00	7.02	1.29	−8.77	4	3.00	7.13	1.47	−4.39	5	2.25
SNAR-A1	59.95	141.13	14.70	17	2.50	59.24	129.55	8.64	17	2.50	58.62	132.99	13.22	17	2.50
SPDYE3	31.65	4.43	−2.86	25	6.75	32.82	7.01	0.69	28	0.17	34.60	7.94	0.28	28	0.24
SULT1A3	7.40	1.19	−6.84	5	2.50	6.98	0.93	−4.50	5	2.00	7.57	1.40	−6.86	5	2.50
TBC1D3_1	33.18	46.76	4.26	15	1.25	39.33	71.68	4.86	13	2.00	45.57	40.76	0.39	25	0.80
TCEB3C	28.60	54.82	7.02	9	2.25	25.87	40.21	2.12	10	1.50	33.05	49.17	1.82	12	1.75
TP53TG3	8.93	4.74	−8.16	4	5.00	6.73	3.29	−0.37	3	3.75	9.25	3.28	−14.72	4	5.25
TRIM49L1	14.13	4.46	−0.92	10	4.25	14.09	3.81	−1.70	10	4.00	12.37	2.81	−6.26	8	4.25
USP17	139.80	705.11	120.40	23	5.00	137.98	731.89	97.37	22	5.25	138.72	1239.19	246.86	15	8.25
ZNF658B	6.32	0.73	−3.63	5	1.25	6.60	1.20	−2.17	5	1.50	5.55	0.59	−6.80	4	1.50

several gene families in *H. sapiens* (Brahmachary et al., 2014) (Fig. 3, Tables 2, 4).

Counting zebrafish NLRs by their Fisnacht protein domains, and disregarding the lab population CGN, our estimates of the d/s ratio are roughly between 10 and 30 and the minimum number (i_o) between 6 (for TU) and 151 (for CGN). Both moment and maximum likelihood estimators agree very well in both estimates (Table 2). The CGN population, however, strongly differs from the other populations surveyed. Interestingly, the other lab population in our sample (TU) marks the opposite: d/s is estimated largest and i_o is estimated smallest among all populations. Again, very low effective population size, absence of random mating, and potentially a large copy number variance in the original founders could explain this observation. As noted before (Schäfer et al., 2024), lab populations are all but suitable to infer evolutionary properties of a species. Our results here confirm strong discrepancies in the immune gene repertoire between lab and wild populations of a model organism which is widely used in biomedical contexts to extrapolate research results to other vertebrates, including human.

As a second application, we analyzed in three populations of human a quite diverse set of gene families. They strongly differ in family size and range from very small to extremely large. Not surprisingly, copy number variation of small gene families is better fitted by STM, followed by a twilight zone where both models appear to be viable explanations of the observed gCNV, while large to super-large families can clearly be explained only by CCM. This mirrors our findings from large NLR gene copy numbers in *D. rerio*. In general, CCM appears to lead to better fits than STM. In small families, observed dispersion tends to be smaller than predicted under STM.

Cases of large families, where STM may be closer to biological reality than CCM, are short interspersed repeats (SINEs) of eukaryotic genomes. Being non-autonomous they depend on the replication

machinery and the presence of other elements capable of duplication (Singer, 1982; Kramerov and Vassetzky, 2011).

Despite the good fit of CCM, in particular in large families, our model may be criticized for ignoring mechanisms which can play an important role in the evolution of gene copy numbers. Copy number can change by unequal recombination and duplicates can also arise by segmental or whole genome duplication (WGD). The latter are extremely rare events compared to those acting on a population genetic timescale of $O(N_{\text{eff}})$ generations, and are neglected here. Our focus is on the copy numbers of individual gene families evolving within populations. As for the action of recombination we have previously analyzed a model where copy number change is realized by unequal cross-over (Otto and Wiehe, 2023; Otto et al., 2024), which can both increase and decrease copy number. We showed that the process approaches a recombination–selection equilibrium which is also very well approximated by a Gamma law. Therefore, while different mechanisms may prevent run-away evolution, we do not expect fundamentally different equilibria if they exist. However, to improve the accuracy of parameter estimation, it certainly remains desirable to integrate this mechanism into the model. The drawback are the additional parameters which need to be estimated or be determined by other means.

Another model assumption, which may be questioned, concerns the fitness function. The one considered here assumes a decrease in fitness with every additional gene copy. Insofar it is analogous to the fitness function of Haigh's model, where any newly arising mutation is assumed to have a fitness cost. The question which fitness model is, or is not, appropriate has been discussed at length before (Zhang et al., 2009; Kondrashov, 2012; Lynch and Conery, 2003). A one-fits-all answer does not exist. Rather, appropriateness depends not only on the size of the gene family but also on the genomic localization of the genes (arrayed in tandem or dispersed) and, above all, on gene function. Occasionally, a new duplicate may confer a fitness benefit (see, e.g., Lynch and Conery (2000)), a case which is not explicitly covered by our model. With sufficiently strong positive selection, fixation of the new copy would be almost instantaneous compared to the subsequent, and slower, drift–selection–duplication regime. Thus, while a 'selective sweep' would increase minimum copy number i_o, it is unlikely to affect the qualitative behavior of our model, in particular when such adaptive fixations are rare. The model from Lynch and Conery (2003) treats selection on duplicates independently for each copy. This may be warranted for (very) small gene families, where different copies may have different functions, perhaps because they occupy different positions in a gene regulatory network or because they may be tuned to different environmental cues, for instance color vision genes in mammals (Yokoyama and Bernhard Radlwimmer, 1998; Gai et al., 2023), post-WGD paralogs involved in fermentative metabolism in yeast (Escalera-Fanjul et al., 2019) or many other examples.

Large gene families, such as the NLRs considered as an example here, may comprise functionally active as well as pseudogenes. It certainly is conceivable that pseudogenes come with a fitness cost leading to decreasing overall fitness with increasing copy number. A unimodal fitness function may appear more appropriate when optimal fitness is associated with a certain fixed number of copies. However, similarly to Muller's ratchet, low copy genotypes will eventually be lost under the action of drift, thereby transforming the fitness function from unimodal to monotonically decreasing. In any case, even if simplistic, the multiplicative fitness function, together with a tunable component of epistasis, covers a wide spectrum of ways how natural selection may shape gene copy number variation.

We explored by computer simulations the dynamics under finite population size, such that genetic drift induces a recurrent ratchet-like loss of the currently fittest class. In particular, we focused on the interarrival times between consecutive clicks, or the speed, of the ratchet. In CCM duplication is a self-accelerating process leading to ever decreasing, and ultimately vanishing, interarrival times. Such a process may be called run-away evolution. With very different model assumptions the generation of large families by fast spread of copies in a population has also been the subject of a recent study by Wang et al. (2024).

By a very natural generalization of the fitness function, our models provide for including the effect of epistasis on the evolutionary dynamics of multi-copy gene families. One of the most striking properties of synergistic epistasis is its capacity to slow down the speed of the ratchet. Synergistic epistasis ($\beta > 1$) induces strong negative selection against individuals with high copy numbers and acts like a barrier against additional copies. Only once surmounted, the ratchet moves with accelerating speed and without upper bound on average copy number. Thus, depending on the strength of epistasis β and the relative magnitudes of d and s, interarrival times may show a transient pattern: increasing while copy number i is small, but then decreasing and eventually vanishing when i is sufficiently large (Fig. 2). Under an extreme form of synergistic epistasis – truncation selection – where individuals with more than a (finite) threshold number of copies have fitness 0, the ratchet ceases to click when all individuals have reached the threshold.

An ever increasing copy number does not seem plausible in practice. Even in large gene families, such as the NLRs in zebrafish, gene copy number variation appears to be smaller than expected under a regime of pure run-away evolution. Our simulations show that epistasis can substantially slow down the ratchet already in very small (here $N_{\text{eff}} = 400$) to moderately large (here $N_{\text{eff}} = 2000$) populations while other parameters are kept in a plausible range (e.g., $s = 2e{-}4$, $d = 2e{-}3$, corresponding to $d/s = 10$, see Fig. 2). In both cases we observe an increase of the interarrival times between ratchet clicks in the interval roughly between $k = 100$ and $k = 1000$. Using different arguments, and extrapolating from observed copy counts, Schäfer et al. (2024) estimated upper bounds on the total number of NLR genes in zebrafish populations and arrived at values of a few thousand for wild populations and of a few hundred for the lab strains (see their Table 1 in Schäfer et al. (2024)). Thus, even if natural populations may not (yet) be at a stable equilibrium, they may have entered a metastable state with interarrival times way beyond a population genetic time scale of $O(N_{\text{eff}})$ generations (Fig. 2).

At large, the same arguments can be extended to interpreting copy number variation of large gene families in human populations in the light of our model. However, as noted before, CCM may not apply to small gene families. If variable at all, their dynamics may well be explained by STM. Here, run-away evolution is not an issue.

CRediT authorship contribution statement

Fabian Freund: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Funding acquisition, Conceptualization. **Johannes Wirtz:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis. **Yichen Zheng:** Writing – review & editing, Software, Methodology, Investigation. **Yannick Schäfer:** Writing – review & editing, Resources, Data curation. **Thomas Wiehe:** Writing – review & editing, Writing – original draft, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

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

Acknowledgments

This work was supported by (separate) grants to TW and FF from the German Research Foundation (DFG) in the framework of SPP-1590 and SPP-1819. We thank Dr. J. Suurväli for sharing insights in the evolution of the zebrafish NLR gene family.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.tpb.2025.04.002.

Data availability

Simulation data are available at https://github.com/y-zheng/Runaway-Duplication-Simulations/. R-code used for numerical analyses is available at https://github.com/fabfreund/genedup. Experimental data from *D. rerio* and human are available in the cited references.

References

Brahmachary, Manisha, Guilmatre, Audrey, Quilez, Javier, Hasson, Dan, Borel, Christelle, Warburton, Peter, Sharp, Andrew J., 2014. Digital genotyping of macrosatellites and multicopy genes reveals novel biological functions associated with copy number variation of large tandem repeats. PLoS Genet. 10 (6), e1004418.

Casanova, Adrian González, Smadi, Charline, Wakolbinger, Anton, 2023. Quasi-equilibria and click times for a variant of Muller's ratchet. Electron. J. Probab. 28, 1–27.

Charlesworth, Brian, 1990. Mutation-selection balance and the evolutionary advantage of sex and recombination. Genet. Res. 55 (3), 199–221.

Crow, John F., Kimura, Motoo, 1970. An Introduction To Population Genetics Theory. Harper & Row, London.

Desai, Michael M., 2019. Haigh (1978) and Muller's ratchet. Theor. Popul. Biol. 133, 19–20.

Escalera-Fanjul, Ximena, Quezada, Héctor, Riego-Ruiz, Lina, González, Alicia, 2019. Whole-genome duplication and yeast's fruitful way of life. TIG 35 (1), 42–54.

Fano, Ugo, 1947. Ionization yield of radiations. II. The fluctuations of the number of ions. Phys. Rev. 72, 26–29.

Felsenstein, Joseph, 1974. The evolutionary advantage of recombination. Genetics 78 (2), 737–756.

Gai, Yulin, Tian, Ran, Liu, Fangnan, Mu, Yuan, Shan, Lei, Irwin, David M., Liu, Yang, Xu, Shixia, Yang, Guang, 2023. Diversified mammalian visual adaptations to bright- or dim-light environments. Mol. Biol. Evol. 40 (4), http://dx.doi.org/10.1093/molbev/msad063.

Gerrish, Philip J, Galeota-Sprung, Benjamin, Cordero, Fernando, Sniegowski, Paul, Colato, Alexandre, Hengartner, Nicholas, Vejalla, Varun, Chevallier, Julien, Ycart, Bernard, 2021. Natural selection and the advantage of recombination. http://dx.doi.org/10.1101/2021.06.07.447324, bioRxiv.

Gordo, Isabel, Charlesworth, Brian, 2000. The degeneration of asexual haploid populations and the speed of Muller's ratchet. Genetics 154 (3), 1379–1387.

Goyal, Sidhartha, Balick, Daniel J, Jerison, Elizabeth R, Neher, Richard A, Shraiman, Boris I, Desai, Michael M, 2012. Dynamic mutation–selection balance as an evolutionary attractor. Genetics 191 (4), 1309–1319.

Haigh, John, 1978. The accumulation of deleterious genes in a population - Muller's ratchet. Theor. Popul. Biol. 14 (2), 251–267.

Jain, Kavita, 2008. Loss of least-loaded class in asexual populations due to drift and epistasis. Genetics 179 (4), 2125–2134.

Kass, Robert E., Raftery, Adrian E., 1995. Bayes factors. J. Amer. Statist. Assoc. 90 (430), 773–795. http://dx.doi.org/10.1080/01621459.1995.10476572.

Kondrashov, Alexey S, 1988. Deleterious mutations and the evolution of sexual reproduction. Nature 336, 435–440.

Kondrashov, Alexey S, 1994. Muller's ratchet under epistatic selection. Genetics 136 (4), 1469–1473.

Kondrashov, Fjodor A., 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc. Biol. Sci. 279, 5048–5057.

Kramerov, D.A., Vassetzky, N.S., 2011. Origin and evolution of sines in eukaryotic genomes. Heredity 107 (6), 487–495.

Leidenroth, Andreas, Hewitt, Jane E., 2010. A family history of dux4: phylogenetic analysis of duxa, b, c and duxbl reveals the ancestral duxgene. BMC Evol. Biol. 10, 364.

Lynch, M., Conery, J.S., 2000. The evolutionary fate and consequences of duplicate genes. Science 290 (5494), 1151–1155. http://dx.doi.org/10.1126/science.290.5494.1151.

Lynch, Michael, Conery, John S., 2003. The evolutionary demography of duplicate genes. In: Genome Evolution. Springer, pp. 35–44.

Muller, Hermann J., 1964. The relation of recombination to mutational advance. Mutat. Res. 1 (1), 2–9.

Neher, Richard A, Shraiman, Boris I, 2012. Fluctuations of fitness distributions and the rate of Muller's ratchet. Genetics 191 (4), 1283–1293.

Otto, M., Wiehe, T., 2023. The structured coalescent in the context of gene copy number variation. Theor. Popul. Biol. 154, 67–78. http://dx.doi.org/10.1016/j.tpb.2023.08.001.

Otto, Moritz, Zheng, Yichen, Grablowitz, Paul, Wiehe, Thomas, 2024. Detecting adaptive changes in gene copy number distribution accompanying the human out-of-africa expansion. Hum. Genome Var. 11 (37), http://dx.doi.org/10.1038/s41439-024-00293-w.

Schäfer, Yannick, Palitzsch, Katja, Leptin, Maria, Whiteley, Andrew.R, Wiehe, Thomas, Suurväli, Jaanus, 2024. Copy number variation and population-specific immune genes in the model vertebrate zebrafish. ELife 13, e98058.

Singer, Maxinem F, 1982. Sines and lines: highly repeated short and long interspersed sequences in mammalian genomes. Cell 28 (3), 433–434.

Smith, John Maynard, 1978. The Evolution of Sex. Cambridge University Press Cambridge.

Stephan, Wolfgang, Chao, Lin, Smale, Joanne G., 1993. The advance of Muller's ratchet in a haploid asexual population: Approximate solutions based on diffusion theory. Genet. Res. 61 (3), 225–231.

Suurväli, Jaanus, Garroway, Colin J, Boudinot, Pierre, 2021. Recurrent expansions of b30. 2-associated immune receptor families in fish. Immunogenetics 1–19.

Wang, Xiaopei, Ruan, Yongsen, Zhang, Lingjie, Chen, Xiangnyu, Shi, Zongkun, Wang, Haiyu, Chen, Bingjie, Tracy, Miles.E, Wu, Chung-I, Wen, Haijun, 2024. The paradox of extremely fast evolution driven by genetic drift in multi-copy gene systems. ELife 13.

Wiehe, Thomas, 1997. Model dependency of error thresholds: the role of fitness functions and contrasts between the finite and infinite sites models. Genet. Res. 69 (2), 127–136.

Yokoyama, Shozo, Bernhard Radlwimmer, F., 1998. The five-sites' rule and the evolution of red and green color vision in mammals. Mol. Biol. Evol. 15 (5), 560–567.

Zhang, Feng, Gu, Wenlin, Hurles, Matthew E., Lupski, James R., 2009. Copy number variation in human health, disease, and evolution. Annu. Rev. Genom. Hum. Genet. 10, 451–481.