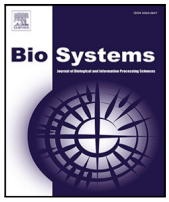
BioSystems 238 (2024) 105176

Contents lists available at ScienceDirect

BioSystems

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



The pseudoentropy of allele frequency trajectories, the persistence of variation, and the effective population size

Nikolas Vellnow [a], Toni I. Gossmann [a,*], David Waxman [b]

[a] TU Dortmund University, Computational Systems Biology, Faculty of Biochemical and Chemical Engineering, Emil-Figge-Str. 66, 44227 Dortmund, Germany
[b] Fudan University, Centre for Computational Systems Biology, ISTBI, 220 Handan Road, Shanghai 200433, People's Republic of China

ARTICLE INFO

Keywords:
Genetic variation
Time-series
Diffusion approximation
Segregation
Simulation
Genetic drift
Selection

ABSTRACT

To concisely describe how genetic variation, at individual loci or across whole genomes, changes over time, and to follow transitory allelic changes, we introduce a quantity related to entropy, that we term *pseudoentropy*. This quantity emerges in a diffusion analysis of the mean time a mutation segregates in a population. For a neutral locus with an arbitrary number of alleles, the mean time of segregation is generally proportional to the pseudoentropy of initial allele frequencies. After the initial time point, pseudoentropy generally decreases, but other behaviours are possible, depending on the genetic diversity and selective forces present.

For a biallelic locus, pseudoentropy and entropy coincide, but they are distinct quantities with more than two alleles. Thus for populations with multiple *biallelic* loci, the language of entropy suffices. Then entropy, combined across loci, serves as a concise description of genetic variation. We used individual based simulations to explore how this entropy behaves under different evolutionary scenarios. In agreement with predictions, the entropy associated with unlinked neutral loci decreases over time. However, deviations from free recombination and neutrality have clear and informative effects on the entropy's behaviour over time.

Analysis of publicly available data of a natural *D. melanogaster* population, that had been sampled over seven years, using a sliding-window approach, yielded considerable variation in entropy trajectories of different genomic regions. These mostly follow a pattern that suggests a substantial effective population size and a limited effect of positive selection on genome-wide diversity over short time scales.

1. Introduction

Individuals within a population vary genetically, with mutation the ultimate source of such variation. While the ultimate fate of any mutation is either fixation or loss, at a given time, the genetic variation of a population arises only from those mutations that are segregating in different individuals of the population. To understand this variation, population genetics focusses on the effects of the multiple evolutionary forces that act on mutations within a population, which include random genetic drift, selection and recombination. In this work we characterise the behaviour of the variation of a population using an approach based on theory and simulation. These have their origin in a *diffusion analysis*, in which the dynamics of a population are approximated in terms of continuous trajectories of allele frequencies. Such an approach was introduced into population genetics by Fisher (1922) (see also Charlesworth, 2022), Wright (1945), and then was substantially extended and developed by Kimura (1955). The development of diffusion methods has continued to the present, and a very considerable body of work, linked to population genetics, has been established — see e.g., the textbook by Ewens (2004).

Genetic variation can be measured at multiple levels, for example at the level of the individual, or the population, or across species. While allozyme variation has previously been used as an indirect measure for the genetic variation, we now are able to determine genetic variation directly from DNA sequences, often across whole genomes. The most frequent type of mutations are single nucleotide polymorphisms (SNPs). During the time SNPs segregate, a population will exhibit genetic variation. However, as SNPs are a very frequent type of mutation, they may occur within the genome in close proximity to each other. Such linked mutations do not have independent fates, with the rate of recombination between them being an important feature (Stapley et al., 2017). Because of this, genetic diversity, as measured at single sites or single loci, is of limited value. Instead, information that is combined over regions of the genome is far more informative (Gossmann et al., 2011).

* Corresponding author.
E-mail addresses: nikolas.vellnow@tu-dortmund.de (N. Vellnow), toni.gossmann@tu-dortmund.de (T.I. Gossmann), davidwaxman@fudan.edu.cn (D. Waxman).

https://doi.org/10.1016/j.biosystems.2024.105176
Received 10 November 2023; Received in revised form 1 March 2024; Accepted 1 March 2024
Available online 11 March 2024

Large scale sequencing technologies, applied, e.g., to whole genomes or full transcriptomes, have led to large quantities of sequencing data becoming available, that covers a wide range of taxonomic groups including non-model organisms (Ekblom and Galindo, 2011). State-of-the-art sequencing technologies provide the possibility of detecting and quantifying the amount of genetic diversity at various levels, including genes (Nadeau et al., 2007), pathways (Gossmann and Ziegler, 2014), tissues (Karr et al., 2019), genomes (Gossmann et al., 2011), in populations (Bosse et al., 2017) — both within and between entire species (Gossmann et al., 2012), and under different scenarios, such as natural or controlled environments (Joly and Faure, 2015) or even entire ecosystems (Garud and Pollard, 2020). This illustrates some of the powerful methods we have to infer selection, from genetic diversity (Casillas and Barbadilla, 2017).

Important summary statistics in evolutionary biology can be obtained from genome-wide diversity, such as Watterson's θ and π or Tajima's D (Watterson, 1975; Tajima, 1989). Principally such summary statistics are obtained from population wide samples obtained at a single time point. However, the way these summary statistic change over time is fundamental to our understanding of population genetics (Coop and Ralph, 2012). Due to advances in DNA sequencing, time-series data has become increasingly available (Malaspinas et al., 2012). While common population genetic methods estimate genetic diversity by assuming a sample scheme that is based on a *single* time point, it is now feasible to use sequencing data from *multiple* time points. As evolution is a dynamic process *by definition*, time-series data presents the opportunity for more accurate detection of causes of evolutionary change at the genetic level, due to the link between allele frequency trajectories and the strength of selection (Bollback et al., 2008). As a result of this, likelihood-based methods have been developed to co-estimate selection coefficients, s, and the effective population size, N_e (Bollback et al., 2008; Foll et al., 2015). The quantity N_e plays a fundamental role in population genetics and is crucial to how variable a population is, as well as distinguishing the strength of selection relative to that of drift (Charlesworth, 2009).

As noted above, we base our work on a diffusion analysis. One such diffusion-based result, that we are especially concerned with, is the time of segregation of a mutation in a population, which equivalently can be described as the time the population exhibits variation. In the case of, e.g., asexual individuals with a single neutral biallelic locus, the mean time of segregation can be expressed in terms of an *entropy* associated with the initial frequencies (Buss and Clote, 2004). In other words, in this case, an important attribute of allelic diversity can be expressed in terms of an entropy. There have been a number of studies in the literature on entropy and related subjects (Iwasa, 1988; Frank, 2009; Mustonen and Lässig, 2010; Peck and Waxman, 2010; Frank, 2012; Baez, 2021; Hledík et al., 2022; Peck and Waxman, 2023)

However, by extending the analysis to a multiallelic case, with more than two alleles, we gain a broader perspective that takes us away, in general, from considerations of entropy. In particular, we show that the fundamental quantity, underlying population variation, is *not* an entropy but a generally distinct quantity that we term *pseudoentropy*. This quantity captures key aspects of genetic diversity. It allows us to give an estimate of the time that a population exhibits variation, as well as allowing a concise characterisation of the behaviour of genetic variation over time. We analytically derive mathematical features of the pseudoentropy and show that under neutrality, its decay over time gives a clear signal for the rate of genetic drift, in terms of the effective population size. To go beyond unlinked neutral loci, we explore changes in pseudoentropy, when combined across loci, under different evolutionary scenarios, using extensive individual-based evolutionary simulations. We find, in agreement with our analytical predictions, that under neutrality, in the absence of linkage, pseudoentropy decreases over time. However, linkage and selection have distinct, informative influences on the behaviour of pseudoentropy over time.

2. Segregation time

We start with a basic model where we can study the persistence of variation. This is for an isolated population of haploid asexual organisms that have one locus and carry one gene, and have discrete generations. Later in this work, we shall consider populations with more complex genetics.

Within the model, time is measured in generations, which we label by t, and which takes the values $0, 1, 2, \ldots$.

Individuals can carry one of n different alleles, with $n \geq 2$. We shall use i to label the alleles, and i can take the values $1, 2, \ldots, n$.

We take there to be a finite number of N adults present in each generation, and we census the population at the adult stage of the life cycle. Each adult is taken to produce the same very large number of offspring and then die. For the population sizes of interest, we assume we can neglect mutation, so any offspring is identical to its parent. If there are no viability (fitness) differences of carriers of different alleles, then the N adults of the next generation are obtained by an unspecified ecological number-regulating mechanism, that is equivalent to picking N offspring at random from the large pool of offspring present in the population. If carriers of different alleles have fitness differences, then there will generally be selective deaths prior to number regulation, and the different alleles will have a modified representation, at number regulation.

What we have just described is a Wright–Fisher model (see e.g., the textbook by Ewens (2004)).

Let $X_i(t)$ denote the relative frequency of type i alleles within adults in the population at time t. Generally the $X_i(t)$ are non-negative and sum to unity ($X_i(t) \geq 0$ and $\sum_{i=1}^{n} X_i(t) = 1$). Henceforth, we shall refer to the relative frequency of an allele just as its *frequency*.

Since we neglect mutation, and have implicitly assumed no migration, no new alleles enter the population. Fixation or loss of the alleles that are initially present are then the only possible outcomes at long times.

Genetic variation persists in the population until all alleles except one have been lost from the population.

Starting at an initial time of 0, with a population where there are n (≥ 2) different alleles present, the population initially exhibits genetic variation. This variation persists up to a random time, that we denote by T, at which time (and beyond) only a single allele remains in the population. Genetic variation is thus absent for all times $\geq T$ and we call T the *time of segregation*.

3. Mean time of segregation for two alleles

We first consider the segregation time of the asexual model, described above, in the special case where there are just $n = 2$ alleles initially present in the population.

With two alleles, it is usual to focus attention on just one allele, and describe the state of the population by the frequency of this allele. For example, knowing the frequency of one allele at time t, say allele 1 (i.e., knowing $X_1(t)$), means the frequency of allele 2 has the value $1 - X_1(t)$ and hence is fully determined. From this viewpoint, where allele 1 is the focus of our attention, the time of segregation can be taken to be the time it takes for allele 1 to become either fixed or lost. For the case $n = 2$ we shall write the initial frequency of allele 1 carriers as y, i.e.,

$$X_1(0) = y. \tag{1}$$

When fixation of allele 1 is the ultimate outcome, we use T_{fix} to denote the random time it takes for fixation of this allele to occur. Similarly, we use T_{loss} to denote the corresponding time to loss of allele 1, given that its loss ultimately occurs. We can then write the segregation time as

$$T = \begin{cases} T_{\text{fix}} & \text{if fixation of allele 1 ultimately occurs} \\ T_{\text{loss}} & \text{if loss of allele 1 ultimately occurs.} \end{cases} \tag{2}$$

With y the frequency of allele 1 at time 0 (i.e., the *initial frequency* - see Eq. (1)), the mean (or expected) value of T, conditional on this frequency, is

$$E_y[T] = E_y[T_{\text{fix}}|\text{fix}] \times P_{\text{fix}}(y) + E_y[T_{\text{loss}}|\text{loss}] \times P_{\text{loss}}(y) \tag{3}$$

in which:

(i) $E_y[...]$ denotes an expected value, conditional on an initial frequency of allele 1 of y;
(ii) $E_y[...|\text{fix}]$ denotes an expected value, conditional on an initial frequency of allele 1 of y *and* the ultimate occurrence of fixation of this allele, while $E_y[...|\text{loss}]$ is the corresponding expectation, when loss takes the place of fixation;
(iii) $P_{\text{fix}}(y)$ is the probability that fixation of allele 1 ultimately occurs, when the initial frequency of allele 1 is y, while $P_{\text{loss}}(y)$ is the corresponding probability, when loss takes the place of fixation.

3.1. Neutral case

Considering still, a population with $n = 2$ alleles, we now assume they have no selective advantage over each other, i.e., we consider a neutral scenario.

The probabilities of fixation and loss of allele 1, when this allele begins with an initial frequency of y, are

$$P_{\text{fix}}(y) = y \quad \text{and} \quad P_{\text{loss}}(y) = 1 - y, \tag{4}$$

respectively (Kimura, 1962). On introducing the effective population size, N_e, into the model (Wright, 1931), the expected times to fixation and loss of allele 1, under the diffusion approximation, are given by

$$E_y[T_{\text{fix}}|\text{fix}] \simeq -\frac{2N_e(1-y)\ln(1-y)}{y} \quad \text{and} \quad E_y[T_{\text{loss}}|\text{loss}] \simeq -\frac{2N_e y \ln(y)}{1-y}, \tag{5}$$

respectively (Kimura and Ohta, 1969).

Using the results in Eqs. (4) and (5), we can write the expected segregation time of Eq. (3) as

$$E_y[T] \simeq 2N_e \times H(y) \tag{6}$$

where

$$H(y) = -(1-y)\ln(1-y) - y\ln(y). \tag{7}$$

The quantity $H(y)$, appearing in Eq. (7), has precisely the form of the Shannon entropy of a random variable which, with the probabilities y and $1-y$, takes two distinct values.[1]

The presence of the entropy in the result for the mean segregation time, under neutrality, has been previously reported in the literature (Buss and Clote, 2004).

We note that in other biological contexts, Shannon entropy has previously been used as a measure of the genetic diversity of a population, (Lewontin, 1972), however in this case entropy was adopted by deliberate choice, as a useful statistic. By contrast, the entropy that appears in the present work, in Eqs. (6) and (7), has a very different status, namely as a quantity that naturally emerges from diffusion-like dynamics.

[1] Generally, for a random variable which, with the probabilities $p_1, p_2, \dots, p_n$, takes n distinct values, the Shannon entropy is given by $-\sum_{i=1}^{n} p_i \ln(p_i)$ (Cover and Thomas, 2006). The Shannon entropy depends on the probabilities of occurrence of the n distinct values, but not on the n distinct values themselves. Throughout the paper, we employ natural logarithms within the entropy.

3.1.1. Dependence on the initial frequency

We have just seen that under the diffusion approximation, the mean time of segregation, for $n = 2$ neutral alleles, depends on the initial frequency of allele 1, namely y, in the guise of the entropy, $H(y)$ (Eq. (7)). Let us consider the character of this y dependence and hence the form of the entropy.

We start by noting that because both alleles are equivalent, the mean time of segregation will be unchanged if their initial frequencies are interchanged, i.e., if y is replaced by $1-y$. This behaviour is equivalent to saying $E_y[T]$, as a function of y, is *symmetric* around the value $y = \frac{1}{2}$. Additionally, it is highly plausible that the further y is from $\frac{1}{2}$, the smaller will be the mean segregation time. The entropy, $H(y)$, has precisely these properties: it is a symmetric around $y = \frac{1}{2}$, it has a maximum value at $y = \frac{1}{2}$, and it decreases with the distance of y from $\frac{1}{2}$, ultimately vanishing at $y = 0$ and $y = 1$ (see Fig. S1 of the Supplementary Material).

4. Mean time of segregation with more than two alleles

To gain a perspective on how fundamental the entropy is, in the present context, we shall determine the mean time of segregation when there are more than two alleles initially present in the population, i.e., when $n > 2$.

If, initially, there are more than two alleles present in a population, then the time over which the population exhibits variability has, at first sight, a more complicated behaviour than the case of two alleles. For example, when $n > 2$, loss of one allele, at any time, does not generally correspond to the disappearance of variability in the population at that time since the remaining alleles can be at intermediate frequencies. However, if we focus on *fixation*, then there is no ambiguity: the occurrence of fixation of one allele guarantees loss of all other alleles, and corresponds to the termination of variability.

Let us consider a population that starts, at the initial time of 0, with n alleles that are at the n initial frequencies

$$\mathbf{y} = (y_1, y_2, \dots, y_n). \tag{8}$$

These are non-negative and sum to unity ($y_i \geq 0$ and $\sum_{i=1}^{n} y_i = 1$).

Let $T_{\text{fix},i}$ denote the random time of fixation of allele i, conditional on allele i ultimately fixing. Then the analogue of Eq. (2), for the random time of segregation, T, that is expressed solely in terms of fixation times of the different alleles, is

$$T = \begin{cases} T_{\text{fix},1} & \text{if fixation of allele 1 ultimately occurs} \\ T_{\text{fix},2} & \text{if fixation of allele 2 ultimately occurs} \\ T_{\text{fix},3} & \text{if fixation of allele 3 ultimately occurs} \\ \vdots & \vdots \\ T_{\text{fix},n} & \text{if fixation of allele } n \text{ ultimately occurs.} \end{cases} \tag{9}$$

The analogue of Eq. (3), for the mean time of segregation, is

$$E_{\mathbf{y}}[T] = \sum_{i=1}^{n} E_{\mathbf{y}}[T_{\text{fix},i}|\text{fix}, i] \times P_{\text{fix},i}(\mathbf{y}) \tag{10}$$

where the notation parallels that of Eq. (3), i.e., $E_{\mathbf{y}}[...]$ corresponds to the expected value, conditional on the initial frequencies, $\mathbf{y}$ (which are given in Eq. (8)); $E_{\mathbf{y}}[...|\text{fix}, i]$ is the expected value, conditional on the initial frequencies of $\mathbf{y}$ *and* fixation of allele i ultimately occurring; $P_{\text{fix},i}(\mathbf{y})$ is the probability of the ultimate fixation of allele i, given the initial frequencies of $\mathbf{y}$.

It can be seen that the mean segregation time with $n = 2$ alleles, given in Eq. (3), is of the same form as Eq. (10), which depends solely on fixation properties, since with just two alleles in a population, loss of one allele is equivalent to fixation of the other allele.

4.1. Neutral case — and emergence of the pseudoentropy

To see what we have gained from considering a general number of n different alleles being present in the population, consider again the case of selective neutrality of all n alleles. Then, as follows from the $n = 2$ results, we have[2] $P_{\text{fix},i}(\mathbf{y}) = y_i$ and under the diffusion approximation, $E_{\mathbf{y}}[T_{\text{fix},i}|\text{fix}, i] \simeq -2N_e(1-y_i)\ln(1-y_i)/y_i$. Eq. (10) then yields

$$E_{\mathbf{y}}[T] \simeq 2N_e \times \mathcal{H}(\mathbf{y}) \tag{11}$$

where

$$\mathcal{H}(\mathbf{y}) = -\sum_{i=1}^{n}(1-y_i)\ln(1-y_i). \tag{12}$$

The quantity $\mathcal{H}(\mathbf{y})$ has a *similar* appearance to the entropy associated with a random variable that takes n distinct values[1]. However, $\mathcal{H}(\mathbf{y})$ depends on the set of n numbers

$$(1-y_1, 1-y_2, \ldots, 1-y_n) \tag{13}$$

which have the property

$$\sum_{i=1}^{n}(1-y_i) = n-1. \tag{14}$$

Hence for $n > 2$, the set of n numbers in Eq. (13) do not sum to unity, and cannot represent a probability distribution. Since entropy is defined for a set of numbers that constitute a probability distribution, the quantity $\mathcal{H}(\mathbf{y})$, which is defined in Eq. (12), is generally not an entropy. Rather, we call $\mathcal{H}(\mathbf{y})$ of Eq. (12) the *pseudoentropy* of $\mathbf{y}$. Only in the special case of $n = 2$ do the numbers in Eq. (13) sum to unity, and so can represent a probability distribution, and only then does $\mathcal{H}(\mathbf{y})$ coincide with an entropy. For all $n \geq 3$ the quantity $\mathcal{H}(\mathbf{y})$ is not naturally viewed as entropy.[3]

It would thus seem that under selective neutrality of all alleles, it is a mathematical accident or coincidence that for $n = 2$ a function with the form of an entropy appears in the expected segregation time, since for $n = 2$, entropy can also be viewed as a pseudoentropy, and for all other n (i.e., $n = 3, 4, 5, \ldots$) it is not an entropy that appears in the result but a pseudoentropy. The fundamental quantity, at the heart of the mean time of segregation, thus appears to be a pseudoentropy.

Let us return to the general n allele result, for the mean segregation time, $E_{\mathbf{y}}[T]$, given in Eq. (11), when there is selective neutrality. The mean segregation time generally depends on the pseudoentropy of $\mathbf{y}$, namely $\mathcal{H}(\mathbf{y})$, and two key properties of this function are as follows:

(i) $\mathcal{H}(\mathbf{y})$ has a minimum value of zero, which is achieved when one $y_i = 1$ and the remaining y_i are zero (the minimum value is obtained by taking a limit of the y_i);
(ii) $\mathcal{H}(\mathbf{y})$ has a maximum value of $(n-1)\ln\left(1+\frac{1}{n-1}\right)$ and this is achieved when all $y_i = 1/n$.

Thus generally, the pseudoentropy lies in the range

$$0 \leq \mathcal{H}(\mathbf{y}) \leq (n-1)\ln\left(1+\frac{1}{n-1}\right). \tag{15}$$

Note that the maximum value that $\mathcal{H}(\mathbf{y})$ can take, for a given n, namely $(n-1)\ln\left(1+\frac{1}{n-1}\right)$, has the property of increasing with n but being bounded from above. In particular, it takes the value $\ln(2) \simeq 0.69$ at $n = 2$ and asymptotically increases to unity as $n \to \infty$. Thus, under any circumstances, $\mathcal{H}(\mathbf{y}) \leq 1$ and by Eq. (11) we have $E_{\mathbf{y}}[T] \lesssim 2N_e$.

5. Pseudoentropy trajectories, under neutrality

So far we have considered the segregation time, T, for the neutral case when, initially, there are n different alleles present in the population. We have found that the mean value of T is proportional to a quantity we termed the pseudoentropy, evaluated at the initial frequencies, $\mathbf{y}$. That is, $E_{\mathbf{y}}[T] \propto \mathcal{H}(\mathbf{y})$. In the present section, we shall restrict all considerations to the case of *selective neutrality* of all alleles. We shall, however, investigate the behaviour of the pseudoentropy, when *evaluated at the set of frequencies of the different alleles*, at some arbitrary time, t. To this end, we write

$$\mathbf{X}(t) = (X_1(t), X_2(t), \ldots, X_n(t)). \tag{16}$$

for the set of frequencies of the n different alleles within a single population, at time t. Then the quantity whose behaviour we shall look at is the pseudoentropy of $\mathbf{X}(t)$:

$$\mathcal{H}(\mathbf{X}(t)) = -\sum_{i=1}^{n}(1-X_i(t))\ln(1-X_i(t)). \tag{17}$$

The behaviour of $\mathbf{X}(t)$ over time constitutes a set of n allele frequency trajectories, of the n different alleles in the population. The behaviour of $\mathcal{H}(\mathbf{X}(t))$, over time, is then very naturally described as a *pseudoentropy trajectory* of the population. It is, however, substantially simpler to consider, than the set of allele frequency trajectories, since at any time the pseudoentropy collapses all n frequencies of the n different alleles into a single number.

A key relation exists between the expected value of $\mathcal{H}(\mathbf{X}(t))$, and the probability distribution of the segregation time, T. Let $\mathcal{P}_{\mathbf{y}}(T > t)$ denote the probability that the segregation time, T, exceeds t, when the initial frequencies are $\mathbf{y}$ (Eq. (8)). Then assuming a constant effective population size, N_e, a diffusion approximation motivates the following result for the expected value of $\mathcal{H}(\mathbf{X}(t))$, conditioned on the initial frequencies $\mathbf{y}$:

$$E_{\mathbf{y}}[\mathcal{H}(\mathbf{X}(t))] - \mathcal{H}(\mathbf{y}) \simeq -\frac{1}{2N_e}\int_0^t \mathcal{P}_{\mathbf{y}}(T > t')dt' \tag{18}$$

- see the appendix for details.

Eq. (18), which has more general and alternative formulations,[4] relates the pseudoentropy of allele frequencies to statistical properties of the segregation time.

5.1. Long time properties

Eq. (18) leads, for $t \to \infty$, to the result for the mean time of segregation given in Eq. (11), as follows. In the long time limit, one allele will ultimately fix, so $\lim_{t\to\infty} E_{\mathbf{y}}[\mathcal{H}(\mathbf{X}(t))] = 0$, thereby allowing the left hand side of Eq. (18) to be simplified. Furthermore, we generally have the relation $\int_0^\infty \mathcal{P}_{\mathbf{y}}(T > t)dt = E_{\mathbf{y}}[T]$ (see e.g., Haigh, 2013 or the appendix). Thus the right hand side of Eq. (18) also simplifies, reducing this equation to $0 - \mathcal{H}(\mathbf{y}) \simeq -\frac{1}{2N_e}E_{\mathbf{y}}[T]$, which is equivalent to Eq. (11).

[2] With n equal fitness alleles, properties of one allele, say allele i, can be obtained by lumping together all other alleles, leading to an effectively two allele problem: allele i and the 'rest'. The probability of ultimate fixation of allele i is then given by the first result in Eq. (4) with y replaced by the frequency of allele i, namely y_i, and the mean time to fixation of allele i is given by the first result in Eq. (5), with y again replaced by y_i. This procedure does not generally work when fitnesses of different alleles are different.

[3] We could, of course, express $\mathcal{H}(\mathbf{y})$ in terms 'new probabilities' that we define as $p_i = (1-y_i)/(n-1)$, which are non negative and sum to unity. Then $\mathcal{H}(\mathbf{y})$ could be written as an expression involving the entropy of the p_i. It is not clear what is gained by this: the p_i do not have an obvious interpretation, and any p_i cannot take values in the full range $0 \leq p_i \leq 1$, that standard probabilities can, but rather are restricted to the range $0 \leq p_i \leq 1/(n-1)$.

[4] In Eq. (18) we have assumed that N_e is independent of time. More generally, N_e will depend on the time, in which case N_e will lie *under* the integral sign in Eq. (18). Additionally, we note that equivalent to Eq. (18) is the differential equation $dE_{\mathbf{y}}[\mathcal{H}(\mathbf{X}(t))]/dt \simeq -\mathcal{P}_{\mathbf{y}}(T > t)/(2N_e)$.

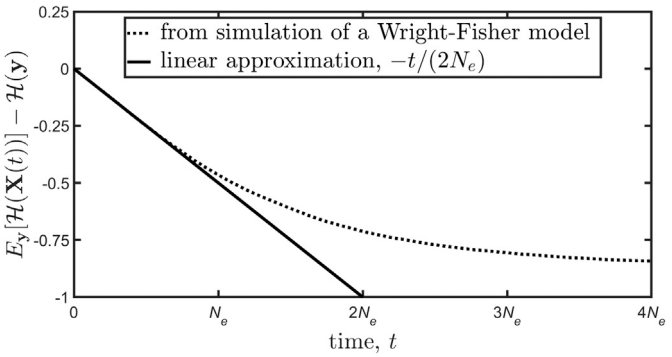


Fig. 1. A plot of the pseudoentropy difference, $E_{\mathbf{y}}[\mathcal{H}(\mathbf{X}(t))] - \mathcal{H}(\mathbf{y})$, to illustrate Eq. (18) and its small time approximation, Eq. (19). For the figure, the parameter values adopted were: $n = 4$ alleles, initial frequencies $\mathbf{y} = (1/4, 1/4, 1/4, 1/4)$, census population size $N = 400$, and effective population size $N_e = 100$. To produce the simulated curve, 10^4 replicate runs were made – see the appendix – using the method of Zhao et al. (2016).

5.2. Short time properties

We can also consider the short time properties of $E_{\mathbf{y}}[\mathcal{H}(\mathbf{X}(t))]$, which follow from Eq. (18), and which are potentially informative.

For t values which are small we have that $\mathcal{P}_{\mathbf{y}}(T > t)$ is close to unity.[5] It is reasonable to assume typical values of T are of order N_e, so we take a small time regime to correspond to $0 \leq t \ll N_e$ where $\mathcal{P}_{\mathbf{y}}(T > t) \simeq 1$, and from Eq. (18), this suggests that

$$E_{\mathbf{y}}[\mathcal{H}(\mathbf{X}(t))] - \mathcal{H}(\mathbf{y}) \simeq -\frac{t}{2N_e} \quad \text{for} \quad 0 \leq t \ll N_e. \tag{19}$$

Eq. (19) tells us that for small t, a plot of the statistic $E_{\mathbf{y}}[\mathcal{H}(\mathbf{X}(t))] - \mathcal{H}(\mathbf{y})$, against time, will linearly decrease and lead to an estimate of $1/(2N_e)$, independent of the initial frequencies of the different alleles, $\mathbf{y}$, and independent of the number of different alleles of the locus, n. Such a plot suggests a way of obtaining an estimate of the effective population size, N_e. Alternatively, if N_e is known, but the generation time is unknown, then a plot of $E_{\mathbf{y}}[\mathcal{H}(\mathbf{X}(t))] - \mathcal{H}(\mathbf{y})$ against the actual time (as opposed to time measured in generations) can lead to an estimate of the generation time.

In Fig. 1 we have plotted the pseudoentropy difference, $E_{\mathbf{y}}[\mathcal{H}(\mathbf{X}(t))] - \mathcal{H}(\mathbf{y})$ against time. In the figure there are two curves — one curve from simulations of a Wright–Fisher model (see the first part of the appendix for details), the other illustrating the linear behaviour in Eq. (19). The leading part of the simulated curve closely matches the predicted linear behaviour at small t, and has a slope that is close to $-1/(2N_e)$.

We note that because the right hand side of Eq. (19) is independent of the initial frequencies (the y_i) we can carry out an unrestricted average of Eq. (19), over the initial frequencies, and obtain the result $E[\mathcal{H}(\mathbf{X}(t))] - E[\mathcal{H}(\mathbf{X}(0))] \simeq -t/(2N_e)$.

6. Simulations incorporating different selection scenarios

So far we have presented analytically motivated results for the relatively simple genetics of haploid individuals. In the following, we present simulation results for more complex genetic architectures.

6.1. Multiple biallelic loci

We now consider multiple biallelic loci. Noting that for a biallelic locus, there is no distinction between its entropy and its pseudoentropy, we shall adopt the simpler language of *entropy* in what follows.

[5] The probability of T taking a positive value is unity, which corresponds to $\mathcal{P}_{\mathbf{y}}(T > 0) = 1$.

In order to summarise the allelic diversity over L biallelic loci we adopt the *average entropy*. This allows connection with the earlier results since, up to a numerical factor, the average entropy is simply a sum over entropies of individual loci.

With $X_i(t)$ and $1 - X_i(t)$ the relative frequencies of the two alleles at locus i at time t, we take the entropy associated with this locus to be given by Eq. (7) (see Fig. S1 of the Supplementary Material, with y replaced by $X_i(t)$). The entropy treats the two allele frequencies completely equivalently, and for locus i at time t it is given by

$$H(X_i(t)) = -X_i(t)\ln(X_i(t)) - (1 - X_i(t))\ln(1 - X_i(t)). \tag{20}$$

Then for a set of times labelled by t, starting with $t = 0$ (the t's do not need to be evenly spaced), we define the average entropy as

$$H_L(t) = \frac{1}{L}\sum_{i=1}^{L} H(X_i(t)), \tag{21}$$

which has an initial value of

$$H_L(0) = \frac{1}{L}\sum_{i=1}^{L} H(X_i(0)), \tag{22}$$

and the corresponding entropy difference is

$$\Delta H_L(t) = H_L(t) - H_L(0). \tag{23}$$

Since we consider the biallelic case we have $0 \leq H_L(t) \leq \ln(2)$ and $-\ln(2) \leq \Delta H_L(t) \leq \ln(2)$, where $\ln(2) \simeq 0.69$.

6.2. Aims and general setup of SLiM simulations

In order to explore how the entropy changes, for a set of polymorphic loci, under several different evolutionary scenarios, we ran extensive individual-based evolutionary simulations of a Wright–Fisher model implemented in the software package SLiM v4.0 (Haller and Messer, 2022). These simulations complement and generalise the analytic results of the previous sections for the haploid case. More specifically, the simulations explored the effects of: (i) sexual reproduction, including recombination between loci, (ii) positive and negative selection, and (iii) heterozygote advantage and negative frequency-dependent selection — as examples of balancing selection.

SLiM is an evolutionary simulation package for constructing genetically explicit individual-based evolutionary models (see also https://messerlab.org/slim/). We set up simulations of diploid hermaphroditic individuals, with a mutation rate of 10^{-7} per site per gamete. The recombination rate was also measured per site per gamete, but varied in different simulated scenarios. After a burn-in of 10,000 generations the mutations currently segregating in the population and their frequencies were tracked for every 5 of the next 400 generations (i.e., at 80 time points). Then $H_L(t)$ and $\Delta H_L(t)$ (Eqs. (21) and (23), respectively), for each sampled time point, were calculated. In particular, this means that only mutations that were present in the first generation after burn-in were used for the calculation of $H_L(t)$ and $\Delta H_L(t)$, but not mutations that emerged later in the simulations. The population size of $N = 400$ was kept constant from generation to generation, with each individual contributing to the next generation according to its relative fitness. The genotypic fitness of a locus, in SLiM, was calculated as follows:

- wildtype homozygote: $w_{wt,wt} = 1$
- mutant homozygote: $w_{mt,mt} = 1+s$, where s is the chosen selection coefficient
- heterozygote: $w_{wt,mt} = 1 + hs$, where h is the chosen dominance coefficient

Unless stated to the contrary, we set $h = 0.5$, so that the fitness of heterozygotes, at a locus, were precisely intermediate between that of the homozygotes. The genotypic fitnesses of all loci were multiplied by SLiM to calculate the relative fitness of an individual. Genotypic fitnesses were calculated differently in the scenario of negative frequency-dependent selection, as explained below.

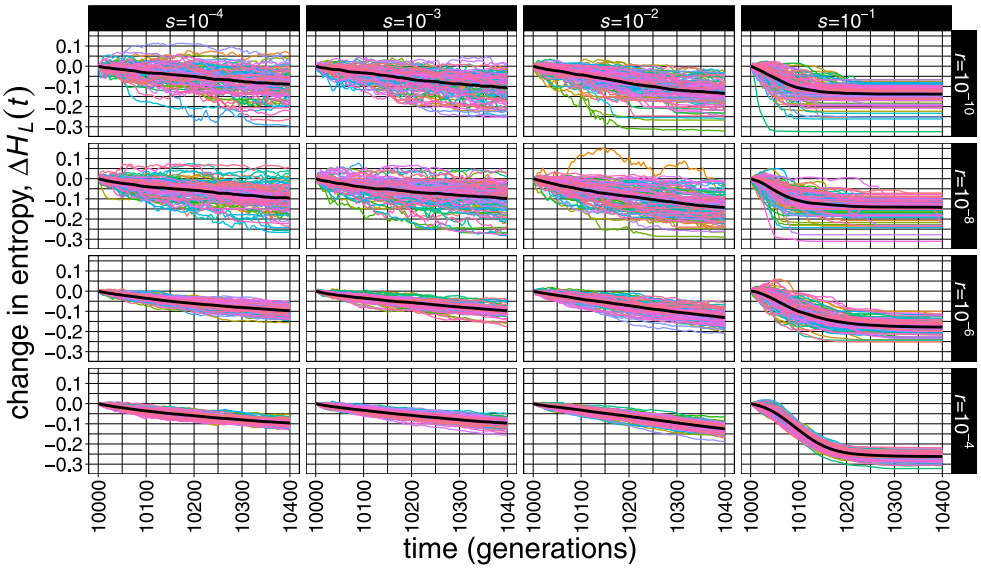


Fig. 2. Positive selection and varying recombination rate. The change in entropy, $\Delta H_L(t)$ over a period of 400 generations following a burn-in of 10,000 generations with positive selection. Each column is for a given selection coefficient, s, and each row is for a given rate of recombination, r. Mean changes of $\Delta H_L(t)$, over 100 replicate runs, are shown in black.

6.3. Neutral case

We considered a neutral scenario, which involves all loci evolving neutrally, by setting $s = 0$, so all genotypes have the same fitness, i.e. $w_{wt,wt} = w_{wt,mt} = w_{mt,mt} = 1$. We investigated how the entropy change, $\Delta H_L(t)$, behaves in 10 different cases, where the recombination rate between adjacent loci takes the values $0, 10^{-10}$, 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3} and 10^{-2}.

From the simulations, we find that the entropy change, $\Delta H_L(t)$, when averaged over replicate populations, has a rate of decrease over time that is the same for all recombination rates. However, the corresponding *variance* of $\Delta H_L(t)$ decreased with recombination rate (see Fig. S2 of the Supplementary Material).

For the simulations in the following sections, we investigated only the four representative recombination rates: 10^{-10}, 10^{-8}, 10^{-6} and 10^{-4}.

6.4. Positive selection

To explore the effect of positive selection on the change in entropy, $\Delta H_L(t)$, we simulated replicate populations in which all mutations were equally beneficial, for the selection coefficients 0.0001, 0.001, 0.01 and 0.1, and under four different recombination rates — see Fig. 2.

We observe that the rate of decay of the entropy with time increases with selection and recombination rate.

6.5. Negative selection

To explore the effect of negative selection on entropy, we simulated replicate populations in which all mutations were equally detrimental, for the selection coefficients of −0.1, −0.01, −0.001, −0.0001 and 0, under four different recombination rates — see Fig. 3.

When selection was strong, entropy decayed by a small amount during the initial 50 generations and then stayed constant, irrespective of recombination rate. In simulations with weaker selection the decay in entropy was observed over the whole observation period and resulted in a larger decrease at the end. Furthermore, under weak selection high recombination rates led to a decreased variance of $\Delta H_L(t)$ among replicate simulations.

6.6. Overdominance

As an example of the effect of balancing selection on entropy, we simulated scenarios with overdominance ($h = -0.5$) with different strengths of negative selection and different recombination rates. For example, with $s = -0.1$ this led to the following relative fitnesses: wildtype homozygotes $w_{wt,wt} = 1$, mutant homozygotes $w_{mt,mt} = 1 - 0.1 = 0.9$, and heterozygotes $w_{wt,mt} = 1 + (-0.5) \cdot (-0.1) = 1.05$ (see Fig. S3 of the Supplementary Material).

We observed little change in the entropy in simulations with large selection coefficients and low recombination rates but a continued decay of $\Delta H_L(t)$ when selection was weak and/or when recombination rate was high (see Fig. S3 of the Supplementary Material).

6.7. Soft sweep

To explore the change $\Delta H_L(t)$, during a soft sweep (Hermisson and Pennings, 2005) – arguably a more realistic scenario – we simulated a burn-in under neutral evolution after which we selected one mutation with a frequency between 0.2 and 0.8, at a random chromosomal position, and changed its selection coefficient to a positive value.

During the subsequent increase of this mutation, the change in entropy, $\Delta H_L(t)$, decreased at the fastest rate, and by the largest amount, at the highest levels of selection and the lowest recombination rates — see Fig. 4.

6.8. Negative frequency-dependent selection

We also simulated scenarios with negative frequency-dependent selection (NFDS). For this, we calculated the fitness of each genotype so that it decreases proportionally to the genotype's frequency with the proportionality constant c, where $0 \leq c < 1$ (Hartl and Clark 1997 p. 240). Consequently, c can serve as a measure for the strength of NFDS. However, to implement these calculations in SLiM we normalised fitness values so that the wildtype homozygote had a fitness of one. Letting p and q denote the frequency of the wildtype and the mutant allele, respectively, it follows:

Genotype	Hartl and Clark 1997	Normalized
$w_{wt/wt}$	$1 - cp^2$	$\frac{1-cp^2}{1-cp^2} = 1$
$w_{wt/mt}$	$1 - c2pq$	$\frac{1-c2pq}{1-cp^2}$
$w_{mt/mt}$	$1 - cq^2$	$\frac{1-cq^2}{1-cp^2}$

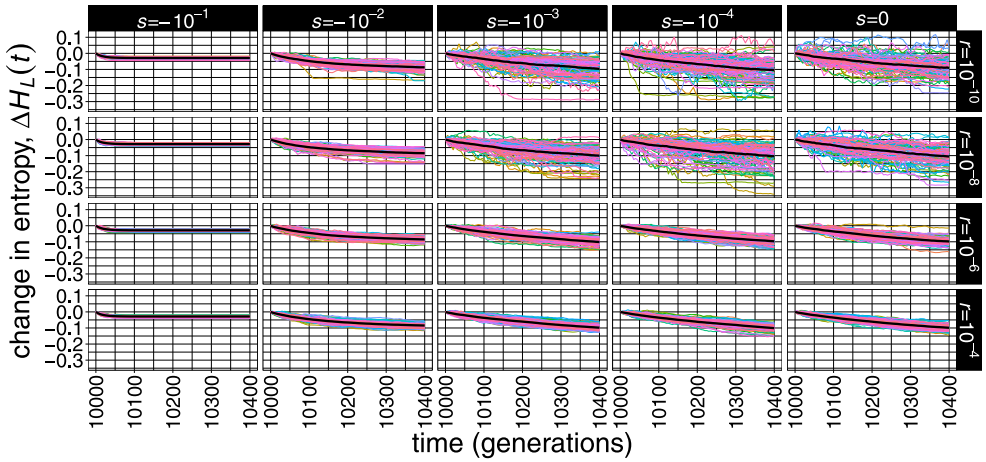


Fig. 3. Negative selection and varying recombination rate. The change in entropy, $\Delta H_L(t)$ over a period of 400 generations following a burn-in of 10,000 generations with negative selection. Each column is for a given selection coefficient, s, and each row is for a given rate of recombination, r. Mean changes of $\Delta H_L(t)$, over 100 replicate runs, are shown in black.

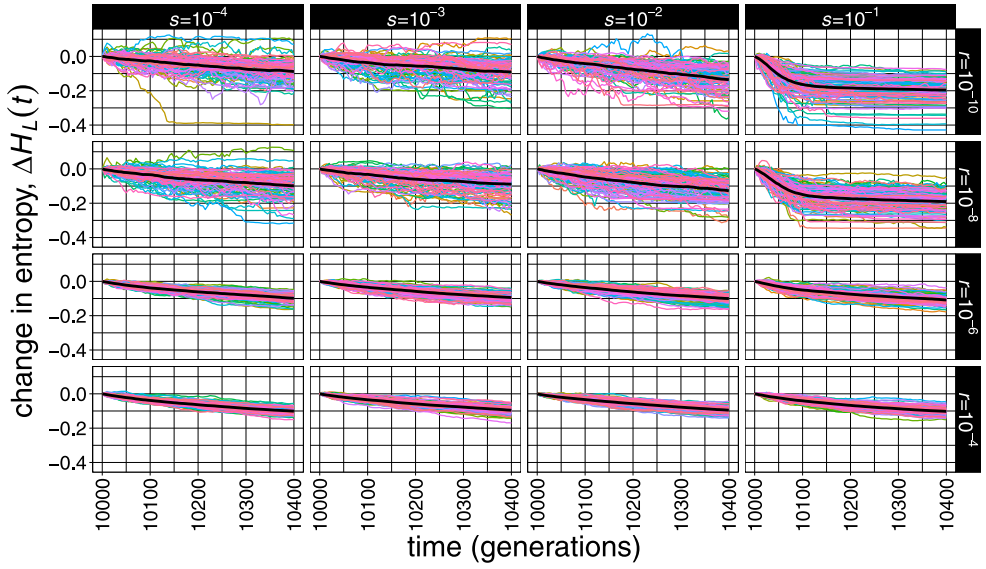


Fig. 4. Single locus soft sweep scenario and varying recombination rate. The change in entropy, $\Delta H_L(t)$, over a period of 400 generation, during a soft sweep scenario following a burn-in of 10,000 generations of neutral evolution. Each column is for a given selection coefficient, s, and each row is for a given rate of recombination, r. Mean changes of $\Delta H_L(t)$, over 100 replicate runs, are shown in black.

6.8.1. Single locus under negative frequency-dependent selection linked to neutral loci

We simulated a scenario similar to a soft sweep (see above) but with negative frequency-dependent selection (NFDS) acting on a single locus, after a burn-in under neutral evolution.

We observe the change in entropy, $\Delta H_L(t)$, decreasing over time. Additionally, in simulations with low recombination rates there is appreciable variation between different replicate runs and less decrease in entropy, irrespective of the strength of NFDS, i.e. the value of c (see Fig. S4 of the Supplementary Material).

6.8.2. Linked loci under frequency-dependent selection

Here, we simulated a scenario in which all loci were under neutral selection during the burn-in but, after this, all loci were subject to negative frequency-dependent selection (NFDS).

We observed a strong and fast decrease in entropy when NFDS was strong but an initial increase in entropy when NFDS was less strong and recombination rate was high — see Fig. 5.

7. Entropy in a natural population

7.1. Aims and data set

To test the usefulness of allelic entropy for studying genomic data from wild, naturally evolving populations, we used Eqs. (21) and (23) to calculate the average entropy for genomic data from a free-living *D. melanogaster* population. We used the publicly available genomic data ("Drosophila Evolution over Space and Time" — DEST) of a population from Linvilla, USA that was sampled at 15 time points over seven years (Kapun et al., 2021). The population was sampled twice (2009, 2010, 2014, 2015), three times (2011, 2012) or once (2013) per year. Assuming 15 generations/year (Pool, 2015), the sampling period spanned around 105 generations. On average 69.1 flies were sampled per time point (range: 33–116), which were then sequenced with Pool-Seq. We downloaded the already pre-processed bcf-file containing SNPs and their chromosome positions. We then excluded all SNPs that were not polymorphic at the first sampling time point ($t = 0$), and SNPs from both sex chromosomes as well as those from the very small non-recombining chromosome 4. Consequently, we included only SNPs that

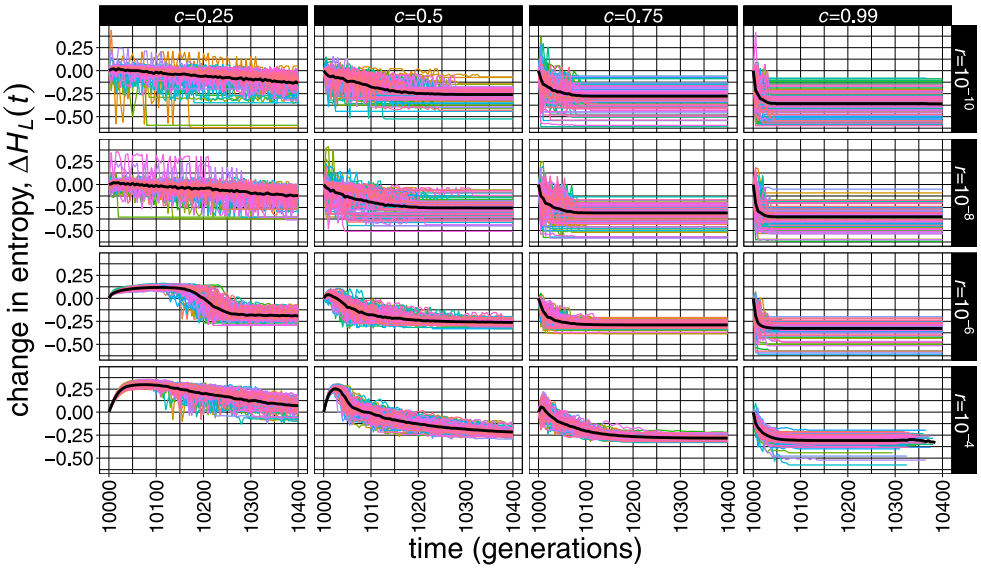


Fig. 5. Multi locus negative frequency dependent selection and varying recombination rate. The change in entropy, $\Delta H_L(t)$, during a scenario in which all loci are under negative frequency-dependent selection over a period of 400 generations following a burn-in of 10,000 generations of neutral evolution. Each column is for a given proportionality constant, c, and each row is for a given rate of recombination, r. Mean changes of $\Delta H_L(t)$, over 100 replicate runs, are shown in black.

Table 1
Summary statistics for sliding windows of length 5 kb for each chromosome including only windows with more than 10 SNPs.

Chromosome	No. SNPs/window		No. windows
	Mean	Range	
2L	73.6	11–255	4236
2R	57.6	11–198	3883
3L	63.9	11–188	4464
3R	55.0	11–188	5182

Table 2
Summary statistics for sliding windows of length 5 kb for each chromosome including only windows with more than 100 SNPs.

Chromosome	No. SNPs/window		No. windows
	Mean	Range	
2L	701.8	116–1420	445
2R	549.4	101–1261	406
3L	610.2	105–1490	466
3R	522.7	102–1217	546

were polymorphic at the first time point from chromosomes 2L, 2R, 3L and 3R in our analysis.

7.2. Method and data analysis

We used a sliding window approach where we partitioned the genome into non-overlapping windows of a fixed size of either 5 kb or 50 kb. We then calculated the average entropy (Eq. (21)) over all SNPs within that window for each of the 15 time points enabling us to track entropy trajectories for each window over time. We excluded 5 kb or 50 kb windows that contained less than 10 or 100 SNPs, respectively, to reduce variability between trajectories due to sampling error. However, results were qualitatively similar when no windows or 5 kb or 50 kb windows with less than 25 or 250 SNPs, respectively, were excluded, even though the number of windows changed slightly in those cases (see Supplementary Material Tables S1–S4).

7.3. Results

7.3.1. Sliding windows of 5 kb

Partitioning chromosomes into 5 kb windows resulted in mean numbers of SNPs per window between 55.0 and 73.6 — see Table 1.

The entropy trajectories for 5 kb windows of chromosome 2L were quite variable and exhibited substantial changes, from one time point to another — see panel A of Fig. 6. During the first observed time points, the entropy-values decreased slightly, on average, but reached a similar value to the starting value at time points 11 to 14 — see panel B of Fig. 6. The other chromosomes showed qualitatively similar patterns (see panels A–F of Fig. S5 of the Supplementary Material).

7.3.2. Sliding windows of 50 kb

Partitioning the chromosomes into 50 kb windows resulted in mean numbers of SNPs per window between 522.7 and 701.8 — see Table 2.

The entropy trajectories for 50 kb windows of chromosome 2L were less variable and exhibited weaker changes from one time point to another than the corresponding trajectories for the 5 kb windows — see panel C of Fig. 6. During the first observed time points the entropy-values decreased slightly on average but reached a similar value to the start value at time points 11 to 14 — see panel D of Fig. 6. The other chromosomes showed a similar behaviour — see panels A–F of Fig. S6 of the Supplementary Material.

8. Discussion

In this work, we used a quantity we termed *pseudoentropy* to describe changes of genetic variation in a population over time. We first developed a single locus model that allows multiple alleles. We showed that for more than two alleles the relevant quantity that relates to genetic variation is pseudoentropy rather than entropy. A key finding is that a small time after census of the population, the pseudoentropy change behaves linearly with time, and the negative slope is related to the effective population size, N_e. Importantly, while the pseudoentropy contains population variation, its change, at small times, is independent of the initial frequencies. This means that we can use all available trajectories, independent of their initial frequency, in a data set to obtain an estimate of the slope, while other models require restrictions on data usage (Gossmann and Waxman, 2022).

Going beyond our analytical results, we then explored changes in *entropy* (which is relevant to a biallelic model) under different evolutionary scenarios using extensive individual-based evolutionary

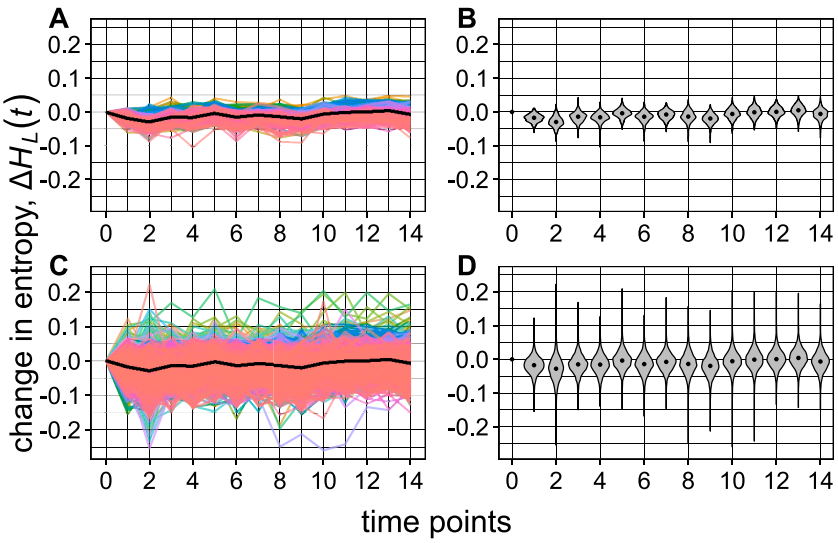


Fig. 6. The change in entropy, $\Delta H_L(t)$ along the 2L chromosome. (A) Each coloured line represents one of 4236 windows of fixed size 5 kb. (B) Violin plots indicating the distribution of 4236 windows of size 5 kb at each time point (medians as black dots). (C) Each coloured line represents one of 445 windows of fixed size 50 kb. (D) Violin plots indicating the distribution of 445 windows of size 50 kb at each time point (medians as black dots).

simulations. In particular, we simulated trajectories for multiple loci that are linked on the same haplotype and calculated the mean entropy for multiple loci as a summary statistic (Eq. (21)). For a neutral model we found, in agreement with our analytical predictions, that entropy decreases over time, but that recombination rate influences the variance and decay among entropy trajectories. We then incorporated directional selection in the model and showed that strong positive selection and high recombination rates lead to a faster decay of entropy. Hence estimates of N_e, based on the decay, would be lower than estimates made assuming a neutral scenario. In the case of negative selection and overdominance we also saw a decay in the slope, however under strong selection but low recombination, the slope is flat. This is because we integrated selection into the burn-in of the simulations, which means that allele frequencies are already low at the beginning of the observation period. Taken together we suggest our method can be used to infer N_e from time series data. We highlight the potential factors that may influence these estimates, in particular in regions of strong selection, or regions of low recombination, or both.

Our initial simulations were conducted on selective effects common to all loci that are linked on a haplotype. As most mutations have small fitness effects that are nearly neutral (Ohta, 1992), this model is likely to be unrealistic in most biological populations. Therefore, we modified our simulations by assuming an allele equilibrium frequency distribution that was generated under neutrality (by using a neutral burn-in) and choosing a single site to be under positive selection, and simulating this situation. Due to linkage it is expected that selection acts on nearby loci and has therefore also an effect on neutral variation. To understand the impact of selection on a single locus we applied a soft sweep model (Johri et al., 2022), where selection acts on a single locus that is already segregating at a certain frequency in the population (i.e., contrary to a new mutation). Under strong selection and weak recombination the effect for such a soft sweep model on the entropy is similar to that to a 'selection on all loci' model, with positive selection across all sites. This is somewhat expected as low recombination leads to linkage of nearby loci and neutral variation cannot segregate freely. Therefore, if there is sufficient recombination, pseudoentropy in regions under strong positive selection will decay faster than in regions not under strong selection. We therefore postulate that changes in the pseudoentropy may be used in a sliding window approach, across the genome, to identify loci under strong positive selection.

Motivated by the results from the single locus soft sweep model, we asked the question whether balancing selection (in our case frequency-dependent selection) can have an impact on the decay of the entropy over time. This seems likely because under balancing selection an increase or a less extreme decay in entropy may arise due to genetic diversity being maintained over longer periods of time (Koenig et al., 2019). Indeed, with low recombination rates, the impact of different strengths of balancing selection lead to less extreme decays of the entropy (cf. Fig. S4 and Fig. 4). Consequently, N_e, estimated from such a slope would be an overestimate. To further investigate our findings we extended our model of balancing selection to all mutations of the linked haplotype. In this model we found that for modest levels of balancing selection, and high rates of recombination, we can observe a temporary *increase* of the entropy (Fig. 5). Genomic regions under recent balancing selection may consequently show characteristic entropy trajectories. We therefore speculate that a sliding window approach may be used to identify the short-term action of balancing selection on the genome.

To test whether our simulation findings may be reflected in real biological data we determined entropy estimates from a wild Drosophila population collected at 15 time points over a period of 7 years (Kapun et al., 2021). We used a sliding window technique to calculate entropy for short genomic fragments. Using this approach we obtained local estimates of entropy across the genome that may identify potential regions in the genome that show signatures of strong selection or non-neutral evolution.

Generally, we see little entropy decay compared to the simulations. A potential explanation is that the population is very large and genetic drift therefore weak. The excess of low frequency alleles in the site frequency spectra of this population is consistent with a large population size (see Fig. S7 of the Supplementary Material). Furthermore, observing a decay in entropy may be difficult because our observation period was only about 105 generations, assuming 15 generations/year (Pool, 2015). Although we can observe the decay of entropy already in the first 100 generations of our simulations (see e.g., Fig. S2 of the Supplementary Material) a longer observation period might be needed in large natural populations. Another potential explanation for the lack of entropy decay could be strong negative selection (cf. Fig. 3). However, we do not think this is a likely explanation because most mutations have small effects on fitness, i.e. are nearly-neutral (Ohta, 1992; Eyre-Walker and Keightley, 2007). Therefore, it is unlikely that widespread, strongly negative selection is the major force driving allele frequency changes across the whole genome.

Furthermore, we observe that the entropy values vary from one time point to another. These changes might be attributed to sampling error, i.e. different number of individuals were sampled at different

time points, and sequencing depth and quality was different from time point to time point, which affects in particular alleles segregating at low frequency (Gossmann and Waxman, 2022).

We also identified potential outlier trajectories. Generally, we see little evidence for outlier trajectories, in particular at larger window sizes. That might suggest that most regions are not under strong selective forces and/or that the population is in a momentary equilibrium and no changes in selective pressures happened during the observation period. Another explanation might be that selection is strongly restricted to regions of high recombination. We identify one outlier region on chromosome 2L between 15,586,079 and 15,591,079bp, which reached $\Delta H_L(t)$-values of below −0.2 at time points 8 to 11 (see panel C of Fig. 6). This window contains only 13 SNPs and harbours an LTR, the closest gene to this region is *kek3*. *kek3* forms a protein network of 11 interaction partners (see Fig. S8 of the Supplementary Material) which are associated with the regulation of epidermal growth factor receptor signalling pathway (Bogdan and Klämbt, 2001) — which is a crucial pathway in early development and is also highly conserved across taxa. Based on the pseudoentropy trajectory we speculate that this region might have experienced strong positive selection during the observation period.

Our results suggest that (pseudo) entropy is an important measure of genetic diversity that we believe merits further attention. We think it may be used to obtain estimates of the effective population size at very high resolutions, and to identify regions in the genome undergoing non-neutral evolution. In particular it has the potential to identify regions under strong positive and balancing selection. However, to establish the practical use of pseudoentropy for high resolution estimation of the effective population size and the detection of ongoing non-neutral evolution (especially weak selection), a detailed comparison with alternative methods may be desirable.

9. Data deposition

All scripts are available at github https://github.com/NikolasVellnow/entropy_scripts, https://github.com/NikolasVellnow/trajectories_scripts and https://github.com/NikolasVellnow/dest_scripts.

CRediT authorship contribution statement

Nikolas Vellnow: Formal analysis, Writing – original draft, Writing – review & editing. **Toni I. Gossmann:** Conceptualization, Data curation, Funding acquisition, Supervision, Writing – original draft, Writing – review & editing. **David Waxman:** Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

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

Data availability

links have been provided.

Acknowledgements

This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A). One of us (DW) thanks Andrew Overall for helpful comments and suggestions.

Appendix A. Theoretical underpinnings

This appendix has three purposes:

(i) Providing details of the simulations carried out for finite population of haploid asexual individuals with one locus and n selectively neutral alleles.
(ii) Motivating the stochastic differential equation that is equivalent to the diffusion approximation of the Wright–Fisher model.
(iii) Establishing results for the behaviour of the pseudoentropy under the diffusion approximation.

We shall distinguish between the census size of the population, N, and the effective population size, N_e, and use a variant of the Wright–Fisher model that incorporates an effective population size (Zhao et al., 2016), yielding results in close agreement with a diffusion approximation.

To begin, we label generations by t with $t = 0, 1, 2, \ldots$. Initial frequencies of the n different alleles in generation 0 are given by[6]

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}. \tag{A.1}$$

Note that in accordance with Zhao et al. (2016), the numbers y_i appearing in Eq. (A.1) are the actual initial frequencies, i.e., frequencies appropriate to a population of size N. Thus each y_i in Eq. (A.1) is of the form of an integer divided by the census size,[7] N.

We use

$$\mathbf{X}_t = \begin{pmatrix} X_{1,t} \\ X_{2,t} \\ \vdots \\ X_{n,t} \end{pmatrix} \tag{A.2}$$

to denote the set of frequencies of the n different alleles within a population at time t. Thus we have

$$\mathbf{X}_0 = \mathbf{y}. \tag{A.3}$$

To define the dynamics of $\mathbf{X}_t$ we use $\mathbf{M}_t(k, \mathbf{x})$, for different t, to represent multinomial random variables that are independent for all t, irrespective of the parameters k and $\mathbf{x}$, where k represents the number of trials, while $\mathbf{x}$ is an n component column vector that contains the probabilities of falling into each of n different categories (Anderson et al., 2018).

A realisation of the multinomial random number $\mathbf{M}_t(k, \mathbf{x})$ is a column of n integers of the form

$$\mathbf{m} = \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_n \end{pmatrix} \tag{A.4}$$

with $0 \le m_i \le k$ and $\sum_{i=1}^{n} m_i = k$. The probability of the occurrence of $\mathbf{m}$ is $\frac{k!}{m_1!m_2!\ldots m_n!} x_1^{m_1} x_2^{m_2} \ldots x_n^{m_n}$.

For $t = 0, 1, 2, \ldots$ the equation that $\mathbf{X}_t$ obeys, following Zhao et al. (2016), is

$$\mathbf{X}_{t+1} = \frac{\mathbf{M}_t(N_e, \mathbf{X}_t)}{N_e}. \tag{A.5}$$

This equation, supplemented by Eq. (A.3) provides a simple and direct way to simulate the dynamics

The simulation procedure captures: (i) the initial frequencies, in terms of the census size, N, and (ii) statistical fluctuations on the scale

[6] For the purposes of this appendix we represent the set of allele frequencies by a *column* vector.

[7] For a diploid, we replace N by $2N$.

of N_e, and leads to results that are very close to results derived from the diffusion approximation (Zhao et al., 2016).

A.1. Motivating the stochastic differential equation

To motivate the stochastic differential equation that is equivalent to the diffusion approximation, we begin, by writing Eq. (A.5) as

$$\mathbf{X}_{t+1} - \mathbf{X}_t = \eta_t(\mathbf{X}_t) \tag{A.6}$$

where

$$\eta_t(\mathbf{x}) = \frac{\mathbf{M}_t(N_e, \mathbf{x})}{N_e} - \mathbf{x}. \tag{A.7}$$

The expected value of $\eta(\mathbf{x})$ is zero, and using properties of multinomial random variables, it may be verified that with $\delta_{i,j}$ a Kronecker delta, and $\mathbf{V}(\mathbf{x})$ an $n \times n$ matrix with elements

$$V_{i,j}(\mathbf{x}) = x_i\delta_{i,j} - x_i x_j \quad \text{for} \quad i, j = 1, 2, \ldots, n \tag{A.8}$$

the variance–covariance matrix of $\eta(\mathbf{x})$ is the matrix $\mathbf{V}(\mathbf{x})/N_e$. Thus $\mathbf{X}_{t+1} - \mathbf{X}_t$ is a random variable with mean zero, and, conditional on the value of $\mathbf{X}_t$, it a variance covariance matrix given by $\sqrt{\mathbf{V}(\mathbf{X}_t)}/\sqrt{N_e}$, where the square root of $\mathbf{V}(\mathbf{X}_t)$ denotes the *principal square root*.[8]

We next introduce a set of n independent Wiener processes

$$\boldsymbol{W}(t) = \begin{pmatrix} W_1(t) \\ W_2(t) \\ \vdots \\ W_n(t) \end{pmatrix} \tag{A.9}$$

(Tuckwell, 1995). The increments of $\boldsymbol{W}(t)$ are given by $d\mathbf{W}(t) = \mathbf{W}(t + dt) - \mathbf{W}(t)$ and have the properties

$$E\left[dW_i(t)\right] = 0 \qquad dW_i(t)dW_j(t) = \delta_{i,j}dt. \tag{A.10}$$

Then with $\mathbf{X}(t)$ a continuous time, continuous state analogue of $\mathbf{X}_t$, the diffusion approximation of Eq. (A.6) is the equation

$$d\mathbf{X}(t) = \frac{\sqrt{\mathbf{V}(\mathbf{X}(t))}}{\sqrt{N_e}} d\mathbf{W}(t). \tag{A.11}$$

This is an *Ito stochastic differential equation* (Tuckwell, 1995). It preserves key properties of $\mathbf{X}(t)$ like $X_i(t) \geq 0$ and

$$\sum_{i=1}^{n} X_i(t) = 1. \tag{A.12}$$

Using the properties of $d\mathbf{W}(\mathbf{t})$ in Eq. (A.10), it can be verified that the distribution of $\mathbf{X}(t)$ obeys

$$-\frac{\partial f(\mathbf{x}, t)}{\partial t} = -\frac{1}{2N_e} \sum_{i,j=1}^{n} \frac{\partial^2}{\partial x_i \partial x_j} \left[V_{i,j}(\mathbf{x}) f(\mathbf{x}, t)\right] \tag{A.13}$$

(Tuckwell, 1995).

An analysis based on Eq. (A.11) is fully equivalent to an analysis based on the diffusion equation, Eq. (A.13), but often at a much lower algebraic cost, and in a form that is closer to intuition.

A.2. Pseudoentropy trajectories

The pseudoentropy of frequency trajectories, written $\mathcal{H}(\mathbf{X}(t))$, is defined in Eq. (17) of the main text, which we reproduce here for convenience:

$$\mathcal{H}(\mathbf{X}(t)) = -\sum_{i=1}^{n} (1 - X_i(t)) \ln(1 - X_i(t)). \tag{A.14}$$

We shall often suppress t arguments.

[8] The principal square root of a real symmetric matrix with non-negative eigenvalues is also a real symmetric matrix with non-negative eigenvalues.

Applying the rules of stochastic (*Ito*) calculus (Tuckwell, 1995) to $\mathcal{H}(\mathbf{X}(t))$ leads to

$$\begin{aligned} d\mathcal{H}(\mathbf{X}) &= \sum_{i=1}^{n} \left.\frac{\partial \mathcal{H}(\mathbf{x})}{\partial x_i}\right|_{\mathbf{x}=\mathbf{X}} dX_a + \frac{1}{2} \sum_{i,j=1}^{n} \left.\frac{\partial^2 \mathcal{H}(\mathbf{x})}{\partial x_i \partial x_j}\right|_{\mathbf{x}=\mathbf{X}} dX_i dX_j \\ &= \sum_{i=1}^{n} \left[\ln(1 - X_i) + 1\right] dX_i - \frac{1}{2} \sum_{i,j=1}^{n} \frac{1}{1 - X_i} \delta_{i,j} dX_i dX_j. \end{aligned} \tag{A.15}$$

Using Eqs. (A.11) and (A.12), the term $\sum_{i=1}^{n} \left[\ln(1 - X_i) + 1\right] dX_i$ in Eq. (A.15) simplifies to

$$\sum_{i=1}^{n} \left[\ln(1 - X_i) + 1\right] dX_i = \sum_{i,j=1}^{n} \ln(1 - X_i) \frac{\left[\sqrt{\mathbf{V}(\mathbf{X})}\right]_{i,j} dW_j}{\sqrt{N_e}} \tag{A.16}$$

where $\left[\sqrt{\mathbf{V}(\mathbf{X})}\right]_{i,j}$ is the (i, j) element of the matrix $\sqrt{\mathbf{V}(\mathbf{X})}$.

The term $\sum_{i,j=1}^{n} \frac{1}{1-X_i} \delta_{i,j} dX_i dX_j$ that appears in Eq. (A.15) also simplifies. Using Eq. (A.11) we have

$$\begin{aligned} \sum_{i,j=1}^{n} \frac{1}{1 - X_i} \delta_{i,j} dX_i dX_j &= \frac{1}{N_e} \sum_{i,j,k,l=1}^{n} \frac{1}{1 - X_i} \delta_{i,j} \left[\sqrt{\mathbf{V}(\mathbf{X})}\right]_{i,k} dW_k \left[\sqrt{\mathbf{V}(\mathbf{X})}\right]_{j,l} dW_l \\ &= \frac{1}{N_e} \sum_{i,j,k,l=1}^{n} \frac{1}{1 - X_i} \delta_{i,j} \left[\sqrt{\mathbf{V}(\mathbf{X})}\right]_{i,k} \left[\sqrt{\mathbf{V}(\mathbf{X})}\right]_{j,l} \delta_{k,l} dt \\ &= \frac{1}{N_e} \sum_{i,k=1}^{n} \frac{1}{1 - X_i} \left[\sqrt{\mathbf{V}(\mathbf{X})}\right]_{i,k} \left[\sqrt{\mathbf{V}(\mathbf{X})}\right]_{i,k} dt \\ &= \frac{1}{N_e} \sum_{i=1}^{n} \frac{1}{1 - X_i} \left[\mathbf{V}(\mathbf{X})\right]_{i,i} dt = \frac{1}{N_e} \sum_{i=1}^{n} X_i dt = \frac{1}{N_e} dt. \end{aligned} \tag{A.17}$$

Using Eqs. (A.16) and (A.17), Eq. (A.15) becomes

$$d\mathcal{H}(\mathbf{X}(t)) = \frac{1}{\sqrt{N_e}} \sum_{i,j=1}^{n} \ln(1 - X_i(t)) \left[\sqrt{\mathbf{V}(\mathbf{X}(t))}\right]_{i,j} dW_j(t) - \frac{1}{2N_e} dt. \tag{A.18}$$

This equation was obtained from a formal application of Ito's rules but it cannot hold for all t. The pseudoentropy is non-negative, and while the first term on the right hand side is non problematic in the sense it does not drive $\mathcal{H}(\mathbf{X}(t))$ negative,[9] the final term on the right hand side, namely $-dt/(2N_e)$, can lead to $\mathcal{H}(\mathbf{X}(t))$ becoming negative at sufficiently large t. We give Eq. (A.18) the interpretation of applying only for times where there is variation in the population, and the pseudoentropy is non-zero. This interpretation is fully consistent with the relation between mean time of segregation and the pseudoentropy of initial frequencies in Eq. (11). When there is no variation in the population the pseudoentropy vanishes, and so must its rate of variation. Thus there is variation in the population only for times prior to the random time of segregation, T.

With $\Theta(t)$ a Heaviside step function ($\Theta(t)$ is 1 for $t > 0$ and 0 for $t \leq 0$) we interpret Eq. (A.18) as

$$d\mathcal{H}(\mathbf{X}(t)) = \frac{1}{\sqrt{N_e}} \sum_{i,j=1}^{n} \ln(1 - X_i(t)) \left[\sqrt{\mathbf{V}(\mathbf{X}(t))}\right]_{i,j} dW_j(t) - \frac{\Theta(T - t)}{2N_e} dt \tag{A.19}$$

where the Heaviside step function sets the $-dt/(2N_e)$ term to zero when variation has ceased in the population, and $\mathcal{H}(\mathbf{X}(t))$ has hit zero. As we shall shortly see, this interpretation leads to precisely the expected value for the mean time of segregation, assuming neutrality of all alleles, as given in Eqs. (11) and (12) of the main text.

Let us take the expected value of this equation, subject to

$$\mathbf{X}(0) = \mathbf{y}. \tag{A.20}$$

[9] We note that first term on the right hand side of Eq. (A.18), namely $\frac{1}{\sqrt{N_e}} \sum_{i,j=1}^{n} \ln(1 - X_i(t)) \left[\sqrt{\mathbf{V}(\mathbf{X}(t))}\right]_{i,j} dW_j(t)$, is non problematic: it vanishes as any $X_i \to 1$, and hence vanishes at the disappearance of variation.

Writing the corresponding expected value as $E_{\mathbf{y}}[...]$, we obtain $dE_{\mathbf{y}}[\mathcal{H}(\mathbf{X}(t))] = -\frac{1}{2N_e}E_{\mathbf{y}}[\Theta(T-t)]dt$. The quantity $E_{\mathbf{y}}[\Theta(T-t)]$ is the probability that $T > t$, and we write this probability as

$$E_{\mathbf{y}}[\Theta(T-t)] = \mathcal{P}_{\mathbf{y}}(T > t). \tag{A.21}$$

We thus obtain

$$\frac{dE_{\mathbf{y}}[\mathcal{H}(\mathbf{X}(t))]}{dt} = -\frac{1}{2N_e}\mathcal{P}_{\mathbf{y}}(T > t). \tag{A.22}$$

We can integrate this equation from time 0 to an arbitrary time t. For the case where N_e is independent of time we have

$$E_{\mathbf{y}}[\mathcal{H}(\mathbf{X}(t))] - \mathcal{H}(\mathbf{y}) = -\frac{1}{2N_e}\int_0^t \mathcal{P}_{\mathbf{y}}(T > t')dt'. \tag{A.23}$$

Note that integrating Eq. (A.21) from $t = 0$ to ∞ leads to $\int_0^\infty \mathcal{P}_{\mathbf{y}}(T > t)dt = \int_0^\infty E_{\mathbf{y}}[\Theta(T-t)]dt$. The right hand side of this equation is $\int_0^\infty E_{\mathbf{y}}[\Theta(T-t)]dt = E_{\mathbf{y}}[\int_0^\infty \Theta(T-t)dt] = E_{\mathbf{y}}[T]$ thus we have

$$\int_0^\infty \mathcal{P}_{\mathbf{y}}(T > t)dt = E_{\mathbf{y}}[T] \tag{A.24}$$

(see also Haigh, 2013). Taking the $t \to \infty$ limit of Eq. (A.23) then yields $\lim_{t\to\infty} E_{\mathbf{y}}[\mathcal{H}(\mathbf{X}(t))] - \mathcal{H}(\mathbf{y}) = -\frac{1}{2N_e}E_{\mathbf{y}}[T]$. Given that loss of variation ultimately occurs, we have $\lim_{t\to\infty}\mathcal{H}(\mathbf{X}(\infty)) = 0$ and hence obtain $\mathcal{H}(\mathbf{y}) = \frac{1}{2N_e}E_{\mathbf{y}}[T]$, which relates the initial value of the pseudoentropy to the expected time of segregation, and is derived in the main text, using elementary methods.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.biosystems.2024.105176.

References

Anderson, D.F., Seppäläinen, T., Valkó, B., 2018. Introduction To Probability. Cambridge University Press, Cambridge.

Baez, J.C., 2021. The fundamental theorem of natural selection. Entropy (ISSN: 1099-4300) 23 (11), 1436. http://dx.doi.org/10.3390/e23111436, https://www.mdpi.com/1099-4300/23/11/1436.

Bogdan, S., Klämbt, C., 2001. Epidermal growth factor receptor signaling. Curr. Biol. 11 (8), R292–R295. http://dx.doi.org/10.1016/s0960-9822(01)00167-1.

Bollback, J.P., York, T.L., Nielsen, R., 2008. Estimation of 2Nes from temporal allele frequency data. Genetics (ISSN: 0016-6731) 179 (1), 497–502. http://dx.doi.org/10.1534/genetics.107.085019.

Bosse, M., Spurgin, L.G., Laine, V.N., Cole, E.F., Firth, J.A., Gienapp, P., Gosler, A.G., McMahon, K., Poissant, J., Verhagen, I., Groenen, M.A.M., van Oers, K., Sheldon, B.C., Visser, M.E., Slate, J., 2017. Recent natural selection causes adaptive evolution of an avian polygenic trait. Science (ISSN: 0036-8075) 358 (6361), 365–368. http://dx.doi.org/10.1126/science.aal3298, https://www.sciencemag.org/lookup/doi/10.1126/science.aal3298.

Buss, S.R., Clote, P., 2004. Solving the fisher-wright and coalescence problems with a discrete markov chain analysis. Adv. Appl. Probab. 36, 1175–1197.

Casillas, S., Barbadilla, A., 2017. Molecular population genetics. Genetics 205.

Charlesworth, B., 2009. Effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet. 10 (3), 195–205. http://dx.doi.org/10.1038/nrg2526.

Charlesworth, B., 2022. The effects of weak selection on neutral diversity at linked sites. Genetics 221, iyac027. http://dx.doi.org/10.1093/genetics/.

Coop, G., Ralph, P., 2012. Patterns of neutral diversity under general models of selective sweeps. Genetics 192.

Cover, T.M., Thomas, J.A., 2006. Elements of Information Theory, second ed. John Wiley and Sons, New Jersey.

Ekblom, R., Galindo, J., 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. (ISSN: 0018067X).

Ewens, W., 2004. Mathematical Population Genetics I. Theoretical Introduction, second ed. Springer-Verlag, New York.

Eyre-Walker, A., Keightley, P.D., 2007. The distribution of fitness effects of new mutations. Nat. Rev. Genet. (ISSN: 1471-0064) 8 (8), 610–618. http://dx.doi.org/10.1038/nrg2146.

Fisher, R., 1922. On the dominance ratio. Proc. R. Soc. Edinb. 42, 321–431.

Foll, M., Shim, H., Jensen, J.D., 2015. WFABC: a Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. Mol. Ecol. Resour. (ISSN: 1755098X) 15 (1), 87–98. http://dx.doi.org/10.1111/1755-0998.12280, http://www.ncbi.nlm.nih.gov/pubmed/24834845.

Frank, S.A., 2009. Natural selection maximizes Fisher information. J. Evol. Biol. (ISSN: 1420-9101) 22 (2), 231–244. http://dx.doi.org/10.1111/j.1420-9101.2008.01647.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1420-9101.2008.01647.x, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1420-9101.2008.01647.x.

Frank, S.A., 2012. Natural selection. V. How to read the fundamental equations of evolutionary change in terms of information theory. J. Evol. Biol. (ISSN: 1420-9101) 25 (12), 2377–2396. http://dx.doi.org/10.1111/jeb.12010, https://onlinelibrary.wiley.com/doi/abs/10.1111/jeb.12010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jeb.12010.

Garud, N.R., Pollard, K.S., 2020. Population genetics in the human microbiome. Trends Genet. (ISSN: 01689525) 36 (1), 53–67. http://dx.doi.org/10.1016/j.tig.2019.10.010, https://linkinghub.elsevier.com/retrieve/pii/S0168952519302215.

Gossmann, T.I., Keightley, P.D., Eyre-Walker, A., 2012. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biol. Evol. (ISSN: 1759-6653) 4 (5), 658–667. http://dx.doi.org/10.1093/gbe/evs027, https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evs027.

Gossmann, T.I., Waxman, D., 2022. Correcting bias in allele frequency estimates due to an observation threshold: A markov chain analysis. Genome Biol. Evol. (ISSN: 01689525) 14 (4), http://dx.doi.org/10.1093/gbe/evac047, https://linkinghub.elsevier.com/retrieve/pii/S0168952519302215.

Gossmann, T.I., Woolfit, M., Eyre-Walker, A., 2011. Quantifying the variation in the effective population size within a genome. Genetics (ISSN: 0016-6731) 189 (4), 1389–1402. http://dx.doi.org/10.1534/genetics.111.132654, http://www.genetics.org/lookup/doi/10.1534/genetics.111.132654.

Gossmann, T.I., Ziegler, M., 2014. Sequence divergence and diversity suggests ongoing functional diversification of vertebrate NAD metabolism. DNA Repair (ISSN: 15687864) 23, 39–48. http://dx.doi.org/10.1016/j.dnarep.2014.07.005, https://linkinghub.elsevier.com/retrieve/pii/S1568786414001918.

Haigh, J., 2013. Probability Models. In: Springer Undergraduate Mathematics Series, Springer London, ISBN: 9781447153436, https://books.google.co.uk/books?id=6QhGAAAAQBAJ.

Haller, B.C., Messer, P.W., 2022. SLiM 4: Multispecies Eco-Evolutionary Modeling. The American Naturalist, (ISSN: 0003-0147) http://dx.doi.org/10.1086/723601, https://www.journals.uchicago.edu/doi/10.1086/723601, Publisher: The University of Chicago Press.

Hermisson, J., Pennings, P., 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169.

Hledík, M., Barton, N., Tkačik, G., 2022. Accumulation and maintenance of information in evolution. Proc. Natl. Acad. Sci. 119 (36), http://dx.doi.org/10.1073/pnas.2123152119, e2123152119. https://www.pnas.org/doi/abs/10.1073/pnas.2123152119.

Iwasa, Y., 1988. Free fitness that always increases in evolution. J. Theoret. Biol. (ISSN: 0022-5193) 135 (3), 265–281. http://dx.doi.org/10.1016/S0022-5193(88)80243-1, https://www.sciencedirect.com/science/article/pii/S0022519388802431.

Johri, P., Stephan, W., Jensen, J.D., 2022. Soft selective sweeps: Addressing new definitions, evaluating competing models, and interpreting empirical outliers. PLOS Genet. 18 (2), e1010022. http://dx.doi.org/10.1371/journal.pgen.1010022.

Joly, D., Faure, D., 2015. Next-generation sequencing propels environmental genomics to the front line of research. (ISSN: 13652540).

Kapun, M., Nunez, J.C.B., Bogaerts-Márquez, M., Murga-Moreno, J., Paris, M., Outten, J., Coronado-Zamora, M., Tern, C., Rota-Stabelli, O., Guerreiro, M.P.G., Casillas, S., Orengo, D.J., Puerma, E., Kankare, M., Ometto, L., Loeschcke, V., Onder, B.S., Abbott, J.K., Schaeffer, S.W., Rajpurohit, S., Behrman, E.L., Schou, M.F., Merritt, T.J.S., Lazzaro, B.P., Glaser-Schmitt, A., Argyridou, E., Staubach, F., Wang, Y., Tauber, E., Serga, S.V., Fabian, D.K., Dyer, K.A., Wheat, C.W., Parsch, J., Grath, S., Veselinovic, M.S., Stamenkovic-Radak, M., Jelic, M., Buendía-Ruíz, A.J., Gómez-Julián, M.J., Espinosa-Jimenez, M.L., Gallardo-Jiménez, F.D., Patenkovic, A., Eric, K., Tanaskovic, M., Ullastres, A., Guio, L., Merenciano, M., Guirao-Rico, S., Horváth, V., Obbard, D.J., Pasyukova, E., Alatortsev, V.E., Vieira, C.P., Vieira, J., Torres, J.R., Kozeretska, I., Maistrenko, O.M., Montchamp-Moreau, C., Mukha, D.V., Machado, H.E., Lamb, K., Paulo, T., Yusuf, L., Barbadilla, A., Petrov, D., Schmidt, P., Gonzalez, J., Flatt, T., Bergland, A.O., 2021. Drosophila evolution over space and time (DEST): A new population genomics resource. Mol. Biol. Evol. (ISSN: 1537-1719) 38 (12), 5782–5805. http://dx.doi.org/10.1093/molbev/msab259.

Karr, T.L., Southern, H., Rosenow, M.A., Gossmann, T.I., Snook, R.R., 2019. The old and the new: Discovery proteomics identifies putative novel seminal fluid proteins in drosophila. Mol. Cellular Proteom. (ISSN: 1535-9476) 18 (Supplement 1), S23–S33. http://dx.doi.org/10.1074/mcp.RA118.001098, http://www.mcponline.org/lookup/doi/10.1074/mcp.RA118.001098.

Kimura, M., 1955. Stochastic processes and distribution of gene frequencies under natural selection. Cold Spring Harbour Symp. Quant. Biol. 20.

Kimura, M., 1962. On the probability of fixation of mutant genes in a population. Genetics 47, 713–719.

Kimura, M., Ohta, T., 1969. The average number of generations until fixation of a mutant gene in a finite population. Genetics 61, 763–771.

Koenig, D., Hagmann, J., Li, R., Bemm, F., Slotte, T., Neuffer, B., Wright, S., Weigel, D., 2019. Long-term balancing selection drives evolution of immunity genes in capsella. elife 8, e43606.

Lewontin, R.C., 1972. The apportionment of human diversity. Evolutionary Biology 6, 381–398.

Malaspinas, A., Malaspinas, O., Evans, S., Slatkin, M., 2012. Estimating allele age and selection coefficient from time-serial data. Genetics (ISSN: 0016-6731) 192 (2), 599–607. http://dx.doi.org/10.1534/genetics.112.140939, http://www.ncbi.nlm.nih.gov/pubmed/22851647, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3454883, http://www.genetics.org/cgi/doi/10.1534/genetics.112.140939.

Mustonen, V., Lässig, M., 2010. Fitness flux and ubiquity of adaptive evolution. Proc. Natl. Acad. Sci. 107 (9), 4248–4253. http://dx.doi.org/10.1073/pnas.0907953107, https://www.pnas.org/doi/abs/10.1073/pnas.0907953107.

Nadeau, N.J., Burke, T., Mundy, N.I., 2007. Evolution of an avian pigmentation gene correlates with a measure of sexual selection. Proc. R. So. B (ISSN: 0962-8452) 274 (1620), 1807–1813. http://dx.doi.org/10.1098/rspb.2007.0174, https://royalsocietypublishing.org/doi/10.1098/rspb.2007.0174.

Ohta, T., 1992. The nearly neutral theory of molecular evolution. Ann. Rev. Ecol. Systemat. (ISSN: 0066-4162) 23 (1), 263–286. http://dx.doi.org/10.1146/annurev.es.23.110192.001403, http://www.annualreviews.org/doi/10.1146/annurev.es.23.110192.001403.

Peck, J.R., Waxman, D., 2010. Is life impossible? Information, sex, and the origin of complex organisms. Evolution (ISSN: 0014-3820) 64 (11), 3300–3309. http://dx.doi.org/10.1111/j.1558-5646.2010.01074.x.

Peck, J.R., Waxman, D., 2023. Homogenizing entropy across different environmental conditions: a universally applicable method for transforming continuous variables. IEEE Trans. Inf. Theory (ISSN: 1557-9654) 69 (3), 1394–1412. http://dx.doi.org/10.1109/TIT.2022.3217387, https://ieeexplore.ieee.org/abstract/document/9930795.

Pool, J.E., 2015. The mosaic ancestry of the drosophila genetic reference panel and the d. melanogaster reference genome reveals a network of epistatic fitness interactions. Mol. Biol. Evol. (ISSN: 0737-4038) 32 (12), 3236–3251. http://dx.doi.org/10.1093/molbev/msv194.

Stapley, J., Feulner, P.G.D., Johnston, S.E., Santure, A.W., Smadja, C.M., 2017. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. Phil. Trans. R. Soc. B 372 (1736), 20160455. http://dx.doi.org/10.1098/rstb.2016.0455.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. Genetics 123.

Tuckwell, H., 1995. Elementary Applications of Probability Theory, second ed. Chapman and Hall, London.

Watterson, G.A., 1975. On the number of segregating sites in genetical models without recombination. Theoret. Popul. Biol. (ISSN: 0040-5809) 7 (2), 256–276. http://dx.doi.org/10.1016/0040-5809(75)90020-9, https://www.sciencedirect.com/science/article/pii/0040580975900209.

Wright, S., 1931. Evolution in mendelian populations. Genetics 16, 97–159.

Wright, S., 1945. The differential equation of the distribution of gene frequencies. Proc. Natl Acad. Sci. USA 31, 382–389.

Zhao, L., Gossmann, T.I., Waxman, D., 2016. A modified wright–fisher model that incorporates ne: A variant of the standard model with increased biological realism and reduced computational complexity. J. Theoret. Biol. 393, 218–228.