OPEN ACCESS Freely available online

PLOS ONE

Genetic Hitchhiking under Heterogeneous Spatial Selection Pressures

Kristan A. Schneider[1,2]*, Yuseob Kim[3]

1 Department Fakultät Mathematik/Naturwissenschaften/Informatik, University of Applied Sciences Mittweida, Mittweida, Germany, **2** Department of Mathematics, University of Vienna, Vienna, Austria, **3** Department of Life Science and Division of EcoScience, Ewha Womans University, Seoul, Korea

Abstract

During adaptive evolutionary processes substantial heterogeneity in selective pressure might act across local habitats in sympatry. Examples are selection for drug resistance in malaria or herbicide resistance in weeds. In such setups standard population-genetic assumptions (homogeneous constant selection pressures, random mating etc.) are likely to be violated. To avoid misinferences on the strength and pattern of natural selection it is therefore necessary to adjust population-genetic theory to meet the specifics driving adaptive processes in particular organisms. We introduce a deterministic model in which selection acts heterogeneously on a population of haploid individuals across different patches over which the population randomly disperses every generation. A fixed proportion of individuals mates exclusively within patches, whereas the rest mates randomly across all patches. We study how the allele frequencies at neutral markers are affected by the spread of a beneficial mutation at a closely linked locus (genetic hitchhiking). We provide an analytical solution for the frequency change and the expected heterozygosity at the neutral locus after a single copy of a beneficial mutation became fixed. We furthermore provide approximations of these solutions which allow for more obvious interpretations. In addition, we validate the results by stochastic simulations. Our results show that the application of standard population-genetic theory is accurate as long as differences across selective environments are moderate. However, if selective differences are substantial, as for drug resistance in malaria, herbicide resistance in weeds, or insecticide resistance in agriculture, it is necessary to adapt available theory to the specifics of particular organisms.

Citation: Schneider KA, Kim Y (2013) Genetic Hitchhiking under Heterogeneous Spatial Selection Pressures. PLoS ONE 8(4): e61742. doi:10.1371/journal.pone.0061742

Editor: Stephen R. Proulx, UC Santa Barbara, United States of America

Received October 19, 2012; **Accepted** March 17, 2013; **Published** April 24, 2013

Funding: This study was supported by the National Research Foundation of Korea grants funded by the Korean government (MEST) (Grant number: 2011-0001575 and 2012R1A1A2004932) to YK, the WTZ project KR 05/2011 to KS and YK, and partly by the grant R01GM084320 from the U.S. National Institutes of Health to Ananias Escalante. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: kristan.schneider@hs-mittweida.de

Introduction

When an advantageous mutation arises and rapidly increases to high frequency under strong positive selection, neutral variants on the same chromosome (initially linked to the mutation) "hitchhike" along the mutation to high frequency. This rapid change in neutral allele frequencies generates a characteristic pattern of polymorphism, commonly referred to as a "selective sweep". Meiotic recombination breaks the association between the advantageous and the neutral allele (the "hitchhiker"). Therefore, the pattern of a selective sweep is contained within a small map distance from the locus under selection. Signatures of selective sweeps include the local reduction of polymorphism (expected heterozygosity), skew of site frequency spectrum, and a unique spatial pattern of linkage disequilibrium. As vast amounts of genome-wide data becomes available, the characteristic patterns of genetic hitchhiking provide a powerful tool to identify candidate regions in the genome that were recently (or still are) under positive directional selection. Moreover, as the quality of genetic data improves, it might be possible to develop methods aiming to reconstruct the underlying evolutionary dynamics by "reverse engineering" hitchhiking patterns. This however requires to extend classical theory to situations that reflect organism-specific characteristics regarding particularities in e.g. the selective environment, demography, or mating structure.

[1] first provided a comprehensive mathematical analysis of this evolutionary process. Since then, remarkable advancements in the mathematical theory of selective sweeps were made [2–5]. Theories focused on the stochastic patterns of variation, mainly achieved through coalescent and diffusion approximations, in order to detect and interpret selective sweeps from DNA sequence polymorphism. Consequently, as more genomic data became available, clear cases of selective sweeps that confirm such theoretical predictions rapidly accumulated (reviewed in [6–8]).

Recent theoretical studies focus on the expansion of theory beyond the "standard model" of genetic hitchhiking. The standard model assumes that an advantageous mutation, arising as a single copy on a random chromosome, increases to high frequency under constant and homogeneous selective pressure in an ideal random-mating population of constant size. This model, the basic scenario of adaptive evolution that [1] considered, is simple enough to allow the application of diffusion and coalescent approximations and thus the prediction of stochastic patterns. However, a selective sweep in a real population must occur under a very complex demographic structure and under various modes of positive selection. Application of the standard model of genetic hitchhiking to the interpretation of actual data may thus lead to a serious

problem. Recent studies addressed this problem by modeling selective sweeps that occur from standing genetic variation or recurrent beneficial mutations [9,10], under arbitrary dominance of the beneficial allele [11], under selection on a quantitative trait [12], in newly derived populations [13–15], in geographically structured populations [16], and under the complex life cycle of malaria parasites [17,18].

Homogeneity of selective pressure driving the beneficial mutation to a high frequency is an important assumption in the standard model of selective sweeps. Typically this is well justified even for a population that is distributed over multiple "patches" with different selective environments, if individuals move rapidly over different patches and also mate with other individuals from other patches. In that case, the population might be modeled to be panmictic under homogeneous selective pressure, which is given by the selective advantage of the beneficial allele averaged over all patches. However, as will be argued below, if mating is restricted to between individuals within the same patch, it can alter the effective rate of meiotic recombination and thus the strength of genetic hitchhiking. This may be important for many species in which mating occurs between individuals within a restricted range (under the same selective environment) but the young offspring (or seeds) are dispersed over a much wider range. Such species include plants that reproduce frequently by self-fertilization and animals that lay eggs in common breeding sites from which the young disperses into random habitats, or agents causing vector-borne parasitic diseases. Particularly plasmodium species, parasites that cause human malaria, are further important examples: male and female gametocytes produced inside a single human host enter a mosquito's gut during the blood meal and release gametocytes, which immediately fuse and undergo meiosis, producing sporozoites that are transferred to different hosts. Therefore, given that hosts constitute heterogeneous selective environments ("patches") for parasites, an allele under selection experiences random switches of patches over malaria transmission cycles while mating always occurs between gametocytes from the same patch. This is an important consideration for malaria parasites in which strong patterns of selective sweeps due to the evolution of drug resistance were discovered [19–22].

This study investigates the hitchhiking effect of a mutant allele spreading over a heterogeneous environment, which is composed of patches with different selective pressures. Averaged over all patches, the mutant allele is advantageous over the wild type. This model of selection with random dispersal over patches between generations is known as the hard-selection Levene model (cf. [23], Ch. 6). [24] originally formulated this model assuming soft selection. While typically the Levene model is considered to study the maintenance of multiple alleles at the balance of selective pressures in different patches, we consider an overall advantage of the mutant allele that ensures the rapid increase of its frequency by directional selection. Our model also differs from the Levene model in which mating between individuals occur within patches.

After formulating the deterministic model and deriving the corresponding recursion equations, we study the effect of a single locus under positive directional selection on a neutral multialelic locus. We propose an analytical solution for the equilibrium frequencies and the expected heterozygosity at the neutral locus after the sweep is complete. We further derive approximations for the equilibrium heterozygosity that are easier to interpret. In particular we want to contrast within-patch mating and mating after random dispersion over the whole population. Even further, we present stochastic simulations in comparison to the analytic results of the deterministic model.

Methods

Overview of the Model

Assume that a haploid sexual population disperses randomly in finitely many patches $\mathcal{P}_1,\ldots,\mathcal{P}_K$. Offspring is born in a common breeding site and then migrates randomly into the K patches. Let the α_k denote the proportion of individuals that migrate into patch $\mathcal{P}_k$. Viability selection acts differently across patches. After reaching the reproductive age adults migrate to the common breeding site for reproduction. A proportion β_k of individuals of patch $\mathcal{P}_k$ mates randomly with individuals from the same patch, whereas the remaining individuals mate randomly in the common breeding site. The haploid offspring in the next generation migrates again into the different patches from the common breeding site. The proportion β_k of individuals mating with other individuals of the same patch has various interpretations. It might reflect that individuals from different patches arrive at different times at the common breeding site, and hence they have a higher chance to mate with individuals from their own patch. Alternatively, it might be interpreted as matings that occur on the way to the breeding site. It might also reflect that some matings occur within the patches before migrating to the breeding site. For simplicity, we will refer to the proportion β_k of matings, as within-patch and to the proportion $1-\beta_k$ as breeding-site matings.

Suppose that the size of the population is sufficiently large to treat the evolutionary changes deterministically. Then, the population in a given generation is represented by a vector $\boldsymbol{p}$ of haplotype frequencies, which is counted after sexual reproduction in the common breeding site. The single-generation change of $\boldsymbol{p}$ is determined by the reproductive success within the different patches. Mating and meiotic recombination is as described above.

This model superficially appears to be the hard-selection Levene model (cf. [23] Chapter 6, for a diploid version), which is equivalent to the standard haploid selection model without migration. However, there is a crucial difference. Namely, the Levene model assumes that mating occurs randomly within the common breeding site, while we assume that only a proportion of individuals of each patch mates within this site. Clearly, our model reduces to the hard selection Levene model if $\beta_k=0$ for $k\in\{1,\ldots,K\}$. On the contrary, if $\beta_k=1$ all matings occur within the patches. We will discuss the differences of our model and the hard-selection Levene model in more detail in the following sections. (The Levene model was introduced originally by [24] for soft-selection.).

Change of Haplotype Frequencies

Assume L multi-allelic loci in a genome of haploid individuals, and let n_i be the number of alleles segregating at locus i, yielding to $n=\prod_{i=1}^{L} n_i$ haplotypes in total. These are labelled $1,\ldots,n$ in the usual order. Their respective relative frequencies in the overall population are $1,\ldots,p_n$, which are summarized by the haplotype-frequency vector $p=(p_1,\ldots,p_n)$.

Let $\alpha_k p_k$ denote the frequency of haplotype i in patch k. Then the (absolute) frequency of haplotype i in patch k after selection is $\alpha_k p_i W_i^{(k)}$. Hence, $\alpha_k\beta_k p_i W_i^{(k)}$, and $\alpha_k(1-\beta_k)p_i W_i^{(k)}$ are the absolute numbers of individuals in patch $\mathcal{P}_k$ that mate randomly within the patch and at the breeding site, respectively. Moreover, $\alpha_k\bar{W}^{(k)}=\alpha_k\sum_{i=1}^{n} p_i W_i^{(k)}$ denotes the number of haplotypes in patch k after selection. The probability that a mating between an i- and a j-haplotype occurs in patch $\mathcal{P}_k$ is then given by

$$\frac{\alpha_k\beta_k p_i W_i^{(k)}\alpha_k\beta_k p_j W_j^{(k)}}{\alpha_k^2\beta_k^2\overline{W}^{(k)2}}=\frac{p_i W_i^{(k)}p_j W_j^{(k)}}{\overline{W}^{(k)2}}. \tag{1}$$

Let the probability that mating between an i- and a j-haplotype gives rise to a l-haplotype be $R(i,j\to l)$. Therefore, the number of l-haplotypes that are produced in patch $\mathcal{P}_k$ is given by

$$p_l^{*(k)}=\frac{\alpha_k\beta_k}{\overline{W}^{(k)}}\sum_{i,j=1}^{n}p_i W_i^{(k)}p_j W_j^{(k)}R(i,j\to l). \tag{2}$$

The number of l-haplotypes that arrive unmated in the common breeding site is

$$\sum_{k=1}^{K}p_i W_i^{(k)}\alpha_k(1-\beta_k)=p_i W_i, \tag{3}$$

where $W_i=\sum_{k=1}^{K}(1-\beta_k)\alpha_k W_i^{(k)}$, and the total number of unmated individuals at the breeding site is

$$\overline{W}=\sum_{i=1}^{n}p_i W_i. \tag{4}$$

Hence, the number of l-haplotypes produced in the breeding site is

$$p_l^{*}=\frac{1}{\overline{W}}\sum_{i,j=1}^{n}p_i W_i p_j W_j R(i,j\to l). \tag{5}$$

From (2) and (5), the relative frequency of haplotype l in the whole population is calculated to be

$$p'_l=\frac{p_l^{*}+\sum_{k=1}^{K}p_l^{*(k)}}{\sum_{j=1}^{n}(p_l^{*}+\sum_{k=1}^{K}p_l^{*(k)})}. \tag{6}$$

We shall briefly summarize the classical hard-selection Levene model:

Remark 1. *In the case of the hard selection Levene model, all individuals mate in a common pool. The relative frequency of i-haplotypes in the mating pool is* $p_i^{*}=\dfrac{\sum_{k=1}^{K}p_i W_i^{(k)}\alpha_k}{\sum_{j=1}^{n}\sum_{k=1}^{K}p_j W_j^{(k)}\alpha_k}=\dfrac{p_i W_i\alpha_k}{\bar{W}}$, *where* $W_j=\sum_{k=1}^{K}W_j^{(k)}\alpha_k$ *and* $\bar{W}=\sum_{j=1}^{n}p_j W_j$. *Hence, the frequency of l-haplotypes in the next generation is given by*

$$p'_l=\frac{1}{\overline{W}^2}\sum_{i,j=1}^{n}p_i W_i p_j W_j R(i,j\to l). \tag{7}$$

Results

Now we want to study genetic hitchhiking, i.e., the influence of selection at a single locus on a linked neutral locus. For this purpose we assume that the first locus is selected with two alleles $\mathcal{A}_1$ and $\mathcal{A}_2$, and that the second locus is selectively neutral with finitely many alleles $\mathcal{B}_1,\ldots,\mathcal{B}_M$. We number the haplotypes such that l stands for $\mathcal{A}_1\mathcal{B}_l$ and $l+M$ stands for $\mathcal{A}_2\mathcal{B}_l$ $(1\le l\le M)$. Moreover, we denote the recombination rate between the two loci by r.

Dynamics at the Selected Locus

Let us denote the frequency of $\mathcal{A}_1$ by p and that of $\mathcal{A}_2$ by $1-p$. The fitnesses of a haplotype carrying the allele $\mathcal{A}_1$ in patch $\mathcal{P}_k$ is $w_1^{(k)}$, whereas that of haplotypes carrying $\mathcal{A}_2$ is $w_2^{(k)}$. Moreover, let $w_i:=\sum_{k=1}^{K}(1-\beta_k)\alpha_k w_i^{(k)}$, so that we obtain $w_1=W_i$ and $w_2=W_{i+M}$ for $i\in\{1,\ldots,M\}$.

By marginalization of the above dynamics it is straightforward to derive the dynamics for p. In subsection 1 of Analysis we show

$$p'=\frac{p\lambda}{p\lambda+(1-p)\mu}, \tag{8a}$$

where

$$\lambda=\sum_{k=1}^{K}\alpha_k w_1^{(k)} \tag{8b}$$

is mean fitness of $\mathcal{A}_1$ among all patches and

$$\mu=\sum_{k=1}^{K}\alpha_k w_2^{(k)} \tag{8c}$$

is the mean fitness of $\mathcal{A}_2$ among all patches.

Note that the dynamics (8) are independent of the β_k's. In particular, the dynamics (8) at the selected locus are that of the standard haploid selection model, which is identical to the hard-selection Levene model.

Summarizing, we obtain:

Result 1. *The allele $\mathcal{A}_1$ will become fixed in the population if and only if*

$$\lambda>\mu.$$

.

Moreover, by iterating (8a) the frequency of $\mathcal{A}_1$ in generation t, with initial condition $p(0)=p_0$, is calculated to be

$$p(t)=\frac{p_0\lambda^t}{p_0\lambda^t+(1-p_0)\mu^t}. \tag{9}$$

Furthermore, we have

$$\lim_{t\to\infty} p(t) = \lim_{t\to\infty} \frac{p_0}{p_0+(1-p_0)(\frac{\mu}{\lambda})^t} = \begin{cases} 1 & \text{if } \lambda>\mu, \\ p_0 & \text{if } \lambda=\mu, \\ 0 & \text{if } \lambda<\mu. \end{cases} \quad (10)$$

Dynamics at the Neutral Locus

Now we want to study the hitchhiking effect of the spread of an resistant allele at a single locus on neutral variation. As before p denotes the frequency of the resistant allele $\mathcal{A}_1$. We have $p=\sum_{j=1}^{M} p_j$. Moreover, we denote the frequencies of the neutral allele $\mathcal{B}_l$ with an $\mathcal{A}_1$-background and $\mathcal{A}_2$-background by $Q_l = \frac{p_l}{\sum_{j=1}^{M} p_j}$, and $R_l = \frac{p_{M+l}}{\sum_{j=1}^{M} p_{M+j}}$, respectively.

In Analysis, subsection 2, we derive Q_l in generation T to be

$$Q_l(T) = Q_l(0) - \frac{r}{\lambda}(Q_l(0)-R_l(0)) \sum_{t=0}^{T-1} \Theta_t \prod_{\tau=0}^{t-1} \Lambda_{p_\tau}, \quad (11)$$

where

$$\Lambda_{p_t} = 1 - \frac{r\Theta_t}{\lambda}\left(\frac{p_0}{1-p_0}\left(\frac{\lambda}{\mu}\right)^{t+1}+1\right) \quad (12)$$

and

$$\Theta_t = \sum_{k=1}^{K} \alpha_k w_1^{(k)} \left(\frac{(1-\beta_k)}{\frac{p_0}{1-p_0}\left(\frac{\lambda}{\mu}\right)^t \frac{w_1}{w_2}+1} + \frac{\beta_k}{\frac{p_0}{1-p_0}\left(\frac{\lambda}{\mu}\right)^t \frac{w_1^{(k)}}{w_2^{(k)}}+1}\right) \quad (13)$$

$$\begin{aligned} &= \frac{w_1}{\frac{p_0}{1-p_0}\left(\frac{\lambda}{\mu}\right)^t \frac{w_1}{w_2}+1} + \sum_{k=1}^{K} \frac{\alpha_k\beta_k w_1^{(k)}}{\frac{p_0}{1-p_0}\left(\frac{\lambda}{\mu}\right)^t \frac{w_1^{(k)}}{w_2^{(k)}}+1} \\ &= \sum_{k=0}^{K} \frac{\gamma_k w_1^{(k)}}{\frac{p_0}{1-p_0}\left(\frac{\lambda}{\mu}\right)^t \frac{w_1^{(k)}}{w_2^{(k)}}+1}. \end{aligned} \quad (14)$$

In the last step we set $\gamma_k := \alpha_k\beta_k$ for $k\in\{1,\ldots,K\}$, $\gamma_0 := \sum_{k=1}^{K} \alpha_k(1-\beta_k)$, and $w_i^{(0)} = w_i$. Hence, we defined patch $\mathcal{P}_0$ as the breeding site, and γ_k is the proportion of the population mating in path $\mathcal{P}_k$.

Although, we could in principle derive $R_l(T)$ analogously, we refrain from doing so. We are only interested in the case in which the allele $\mathcal{A}_1$ sweeps through the population. Hence, at equilibrium $\mathcal{A}_2$ vanishes, and all neutral alleles are linked to the allele $\mathcal{A}_1$. Hence, the equilibrium frequency of $\mathcal{B}_l$ is given by $\lim_{t\to\infty} Q_l(t)$. Deriving these frequencies allows to study genetic hitchhiking. In particular, we have

$$\hat{Q}_l = \lim_{T\to\infty} Q_l(T+1) = Q_l(0) - (Q_l(0)-R_l(0))A_r, \quad (15a)$$

where

$$\begin{aligned} A_r &:= \frac{r}{\lambda}\sum_{t=0}^{T-1} \Theta_t \Pi_{\tau=0}^{t-1} \Lambda_{p_\tau} \\ &= \frac{r}{\lambda}\sum_{k=0}^{K} \alpha_k\beta_k w_1^{(k)} \sum_{t=0}^{\infty} \frac{\Pi_{\tau=0}^{t-1}\Lambda_{p_\tau}}{\frac{p_0}{1-p_0}\left(\frac{\lambda}{\mu}\right)^t \frac{w_1^{(k)}}{w_2^{(k)}}+1}. \end{aligned} \quad (15b)$$

From the above it is straightforward to calculate the equilibrium heterozygosity defined by

$$\hat{H} = \frac{M}{M-1}\sum_{l=1}^{M} \hat{Q}_l(1-\hat{Q}_l). \quad (16)$$

The equilibrium heterozygosity depends on the initial allele-frequency distribution at the neutral locus, because $\hat{Q}_l$ does. However, as shown in Analysis (subsection 3) the relative expected heterozygosity defined by

$$\mathcal{H} := \frac{E(\hat{H})}{E(H_0)} = \frac{E(\hat{H}|H_0)}{H_0}$$

is independent of the initial distribution of allele-frequency distribution. Here, E denotes the expectation (over the initial distribution of allele-frequencies), and

$$H(0) := \frac{M}{M-1}\sum_{l=1}^{M} Q_l(0)(1-Q_l(0))$$

is the initial heterozygosity. We summarize

Result 2. *The equilibrium frequency of the neutral allele $\mathcal{B}_l$ is given by (15). The expected relative heterozygosity is calculated to be*

$$\mathcal{H} := 2A_r - A_r^2 \quad (17)$$

where A_r is defined in (15b).

Remark 2. *For the hard-selection Levene model, we need to set $\beta_k = 0$ for all k, which gives*

$$\hat{Q}_l = Q_l(0) - r(Q_l(0)-R_l(0)) \sum_{t=0}^{\infty} \frac{(1-r)^t}{\frac{p_0}{1-p_0}\left(\frac{w_1}{w_2}\right)^{t+1}+1}, \quad (18)$$

which clearly is exactly the solution for standard hitchhiking.

The differences between our model and the hard-selection Levene model become obvious from the above remark. Whereas the dynamics at the selected model coincide for both models, differences occur at linked neutral loci. Not surprisingly, the hard-selection Levene model is equivalent to the standard haploid selection model. In particular, the relative heterozgosity which measures the hitchhiking effect (see section 3) does not coincide for the two models. Figures 1, 2, and 3 illustrate these differences.

The analytic solution (17) is insofar not satisfying as it is iterative and difficult to interpret. We will therefore derive approximations that have a simpler form and are easier to interpret in terms of the involved parameters.

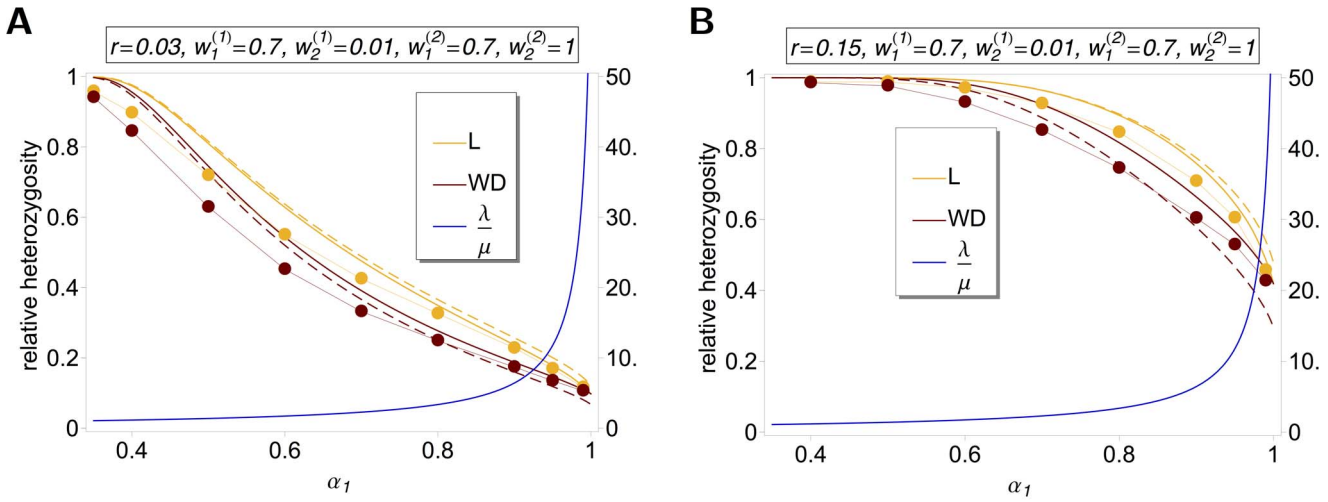

Figure 1. Heterozygosity as a function of α_1. Average relative heterozygosity $\mathcal{H}(r)$ (left y-axis) and $\frac{\lambda}{\mu}$ (right y-axis) as a function of α_1 assuming two patches ($\alpha_2 = 1-\alpha$). We assume either complete within-patch mating and dispersion (WD; $\beta_1 = \beta_2 = 1$) according to the model introduced here, or the hard-selection Levene model (L; $\beta_1 = \beta_2 = 0$). Solid lines correspond to exact solutions according to equations (15) and (18), respectively. Dashed lines show approximate solutions according to equation (25a) combined with equations (25b) and (25c), respectively. Dots represent the values obtained from stochastic simulations. Fitness values are shown in the boxes above the plot panels in (A) and (B). Stochastic simulations are based on 1000 repetitions for each parameter combination and $N = 10{,}000$. For the exact and approximate solutions we assumed $p_0 = 0.0001$ to compensate for the deterministic solution's overestimation of heterozygosity due to the prolonged initial spread of the beneficial mutation in the deterministic model.
doi:10.1371/journal.pone.0061742.g001

Approximations

By writing p_t for p, and using (8), Λ_{p_t} becomes

$$\Lambda_p = 1 - \frac{r(p\lambda + (1-p)\mu)}{\lambda\mu}\left(\frac{w_1}{p\frac{w_1}{w_2} + 1 - p} + \sum_{k=1}^{K} \frac{\alpha_k \beta_k w_1^{(k)}}{p\frac{w_1^{(k)}}{w_2^{(k)}} + 1 - p}\right) \quad (19)$$

(cf. 35, 28). Hence,

$$\lim_{p\to 0} \Lambda_p = 1 - \frac{r\mu}{\lambda\mu}\left(w_1 + \sum_{k=1}^{K} \alpha_k \beta_k w_1^{(k)}\right)$$
$$= 1 - \frac{r}{\lambda}\sum_{k=1}^{K} \alpha_k (1 - \beta_k + \beta_k) w_1^{(k)} = 1 - r.$$

Moreover,

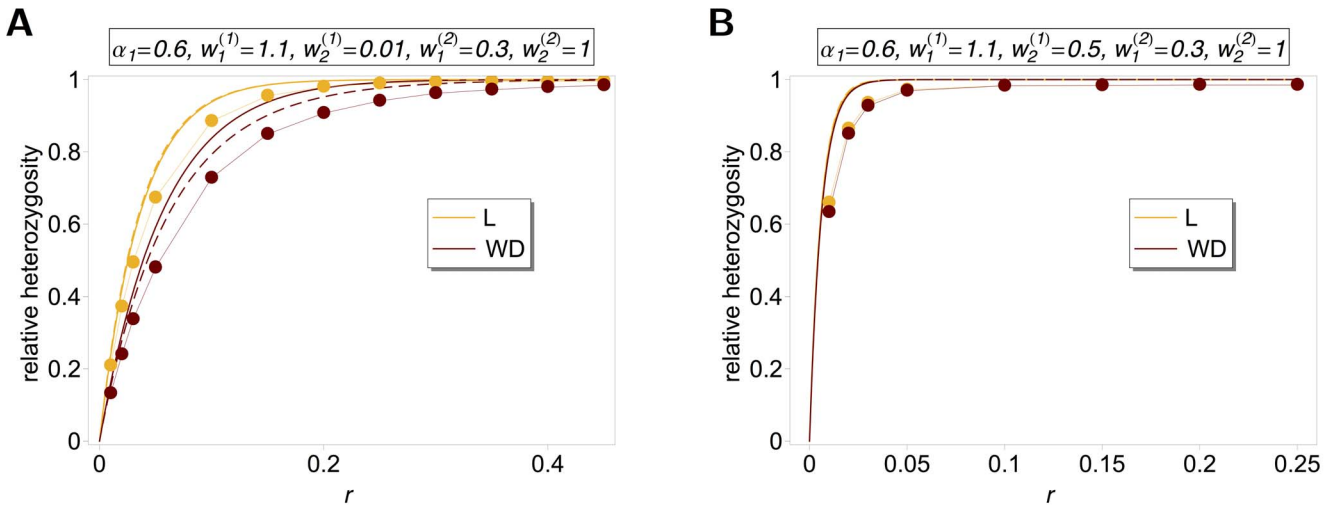

Figure 2. Heterozygosity as a function of r. Average relative heterozygosity $\mathcal{H}(r)$ as a function of r. See legend of Figure 1 for more details.
doi:10.1371/journal.pone.0061742.g002

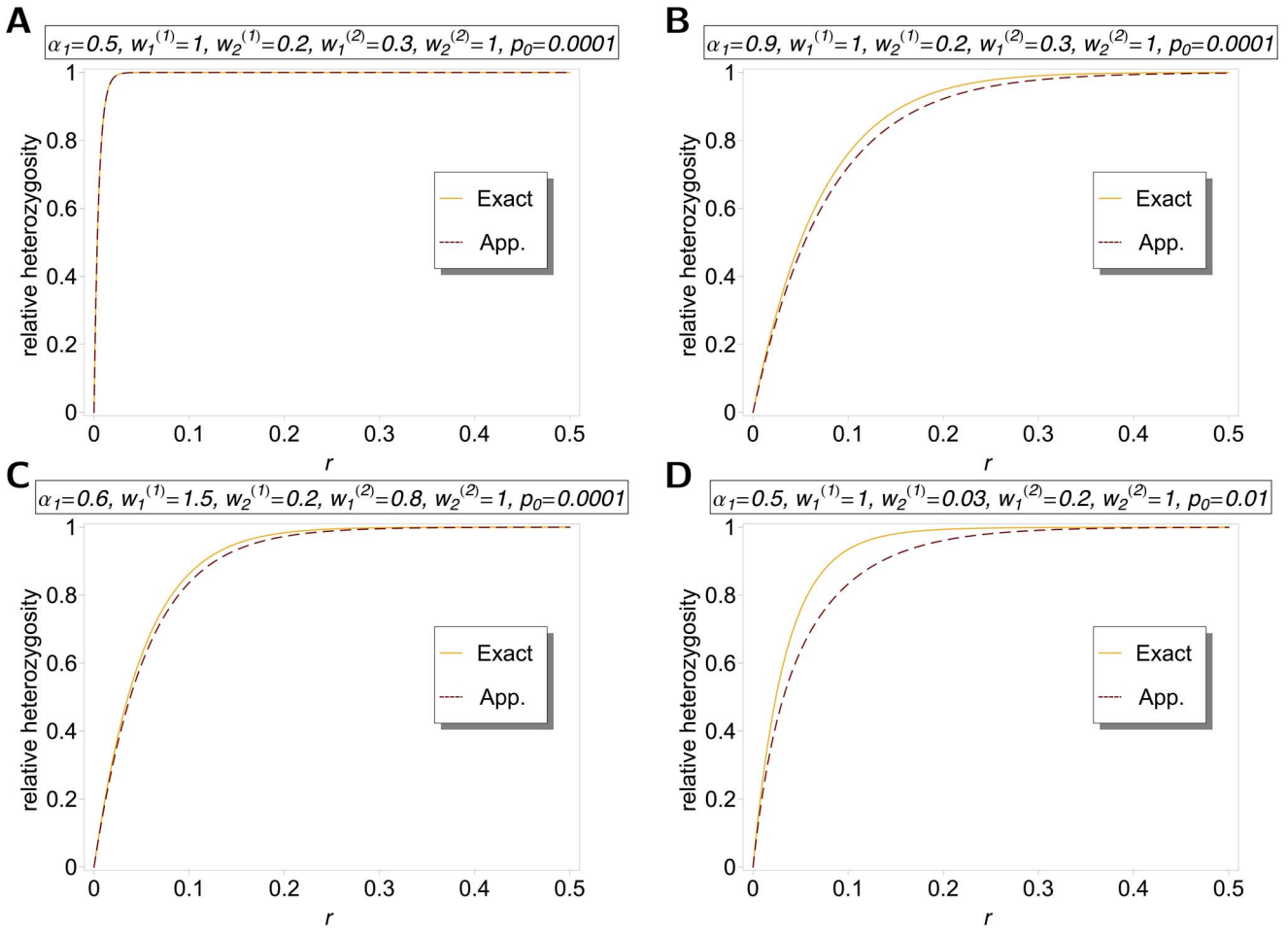

Figure 3. Exact *vs.* approximate average relative heterozygosity. Average relative heterozygosity $\mathcal{H}(r)$ as a function of r as given by (15) and (18), and equation (25a) combined with equation (25b). Two patches with $\beta_1=\beta_2=1$ were assumed. Moreover, fitness parameters and initial frequencies are shown in the boxes above the plot panels in (A), (B), (C), and (D).
doi:10.1371/journal.pone.0061742.g003

$$\lim_{p\to 1}\Lambda_p = 1-\frac{r\lambda}{\lambda\mu}\left(\frac{w_1}{\frac{w_1}{w_2}}+\sum_{k=1}^{K}\frac{\alpha_k\beta_k w_1^{(k)}}{\frac{w_1^{(k)}}{w_2^{(k)}}}\right)=1-\frac{r}{\mu}\sum_{k=1}^{K}\alpha_k w_2^{(k)}=1-r.$$

In the section Analysis we even show that $\Lambda_p\geq 1-r$, always holds. Hence, we can appoximately set $\Lambda_{p_t}\approx 1-r$. Therefore, we can approximate $\hat{Q}_l$ by

$$\hat{Q}_l \approx Q_l(0)-\frac{r}{\lambda}(Q_l(0)-R_l(0))\sum_{k=0}^{K}\alpha_k\beta_k w_1^{(k)}\sum_{t=0}^{\infty}\frac{(1-r)^t}{\frac{p_0}{1-p_0}\left(\frac{\lambda}{\mu}\right)^t\frac{w_1^{(k)}}{w_2^{(k)}}+1}. \quad (20)$$

Note that (20) has the same structure as comparable quantities in [18]. Hence, (20) can be further approximated with exactly the same methods as in [18]. This leads to

Result 3. *The equilibrium frequency $\hat{Q}_l$ of the allele $\mathcal{B}_l$ is given by (18). If $p_0\approx 0$, the frequency is approximately*

$$\hat{Q}_l\approx\tilde{Q}_l := Q_l(0)-(R_l(0)-Q_l(0))\begin{bmatrix}\frac{1}{\log(1-r)}\left(1+\frac{r}{2(1-r)}\right)-\frac{p_0^{\frac{\log(1-r)}{\log\mu-\log\lambda}}}{\lambda}\left(\frac{1}{\log(1-r)}+\frac{1}{\log(1-r)-\log\lambda+\log\mu}\right)\\ \sum_{k=0}^{K}\alpha_k w_1^{(k)}\left(\frac{w_1^{(k)}}{w_2^{(k)}}\right)^{\frac{\log(1-r)}{\log\mu-\log\lambda}}\end{bmatrix}. \quad (21)$$

If additionally $r\approx 0$, we approximately obtain

$$\hat{Q}_l\approx\tilde{Q}_l : = R_l(0)+(Q_l(0)-R_l(0))\sum_{k=0}^{K}\frac{\gamma_k w_1^{(k)}}{\lambda}\left(\frac{w_1^{(k)}}{w_2^{(k)}}p_0\right)^{\frac{r}{\log\lambda-\log\mu}} \quad (22)$$

$$= R_l(0) + (Q_l(0) - R_l(0)) p_0^{\frac{r}{\log\lambda - \log\mu}} \sum_{k=0}^{K} \frac{\gamma_k w_1^{(k)}}{\lambda} e^{r\frac{\log w_1^{(k)} - \log w_2^{(k)}}{\log\lambda - \log\mu}}. \quad (23)$$

A scratch of the proof based on the results of [18] is presented in the section Analysis (subsection 4).

The above results allows for a simple interpretation. The neutral allele's frequency is a weighted average over the respective frequencies resulting from each patch (including the breeding site $\mathcal{P}_0$). The weights are the proportion of individuals mating in each patch, γ_k, times the relative size of the patches, i.e., the relative frequency of individuals in the patch, $w_1^{(k)}/\lambda$. Moreover, within-patch mating leads to an adjustment factor $e^{r\frac{\log w_1^{(k)} - \log w_2^{(k)}}{\log\lambda - \log\mu}}$ for the neutral allele's frequency within each patch as compared to standard hitchhiking. This adjustment measures how deviations of the selective regime in patch $\mathcal{P}_k$ from the overall selection regime affects recombination. In particular, if $\frac{w_1^{(k)}}{w_2^{(k)}} \approx \frac{\lambda}{\mu}$, we have $e^{r\frac{\log w_1^{(k)} - \log w_2^{(k)}}{\log\lambda - \log\mu}} \approx e^r \approx 1$ for $r \approx 0$. This implies that patches that reflect the population average selective pressures can be subsumed within the common breeding site. However, in patches characterized by 'extreme' selective regimes, deviations might be substantial. We can summarize:

Result 4. *Let* $\Omega := \{k \in \{0,\ldots,K\} \mid |\frac{w_1^{(k)}}{w_2^{(k)}} - \frac{\lambda}{\mu}| < \varepsilon\}$ *($\varepsilon > 0$), be the set of patches that reflect the overall selective regime. If* $p_0, r \approx 0$, *the equilibrium frequency* $\hat{Q}_l$ *of the allele* $\mathcal{B}_l$ *is given by*

$$\hat{Q}_l \approx R_l(0) + \frac{Q_l(0) - R_l(0)}{\lambda} p_0^{\frac{r}{\log\lambda - \log\mu}} \left(\sum_{k\in\Omega} \gamma_k w_1^{(k)} + \sum_{k\in\Omega^c} \gamma_k w_1^{(k)} \left(\frac{w_1^{(k)}}{w_2^{(k)}} p_0\right)^{\frac{r}{\log\lambda - \log\mu}}\right), \quad (24)$$

where $\Omega^c = \{0,\ldots,K\} \backslash \Omega$.

The equilibrium heterozygosity is obtained by combining an adaptation of Result 2 with Result 3 and 4.

Result 5. *If* $p_0 \approx 0$ *and* $r \approx 0$ *we have*

$$\mathcal{H} = 1 - \phi^2 p_0^{\frac{2r}{\log\lambda - \log\mu}}, \quad (25a)$$

where

$$\phi = \sum_{k=0}^{K} \frac{\gamma_k w_1^{(k)}}{\lambda} e^{r\frac{\log w_1^{(k)} - \log w_2^{(k)}}{\log\lambda - \log\mu}} \quad (25b)$$

or

$$\phi = \sum_{k\in\Omega} \gamma_k w_1^{(k)} + \sum_{k\in\Omega^c} \gamma_k w_1^{(k)} \left(\frac{w_1^{(k)}}{w_2^{(k)}} p_0\right)^{\frac{r}{\log\lambda - \log\mu}} \quad (25c)$$

with Ω *defined as in Result 4. The factor* ϕ *is an adjustment due to increased inbreeding within patches caused by different survival rates resulting from different selective regimes.*

Note, that setting $\Omega = \{1,\ldots,K\}$ yields the approximate herterozygosity for the hard-selection Levene model. Figures 1, 2, and 3 illustrate the above result.

Stochastic Simulations

The stochastic behavior of our two-locus model is explored by computer simulation in which the population contains a finite number (N) of haploid individuals. We restrict our attention to contrast the two extreme situations of complete intra-patch mating ($\beta_k = 1$ for all k) and to the hard-selection Levene model ($\beta_k = 0$ for all k). Furthermore, we will assume only two patches for most of the simulations.

Given N individuals in generation t, sampling of individuals (offspring) for generation $t+1$ is performed in the following manner. First, a copy of a randomly-picked individual in generation t is sent to patch $\mathcal{P}_k$ with probability α_k. Then, a number x is drawn from uniform distribution between 0 and $\max_{i,k} w_i^{(k)}$. This copy is accepted (i.e. sampled) into generation $t+1$ if $x < w_i^{(k)}$, where $i = 1$ (2) if it carries the mutant (wildtype) allele. Otherwise this copy is discarded. This procedure is repeated until all N haploids are sampled. Next, to perform recombination, $Nr/2$ pairs of individuals are chosen and cross-overs occur. For each pair, the first individual is chosen randomly from the entire population. If $\beta_k = 0$, the second individual is also chosen over the entire population (Levene model). If $\beta_k = 1$, the second individual is chosen from the same patch. This completes reproduction for generation $t+1$. Simulations start ($t = 0$) with one mutant and $N-1$ wildtype alleles. If the mutant allele is lost, the simulation starts again from the initial condition. The simulation stops when the mutant allele reaches fixation in the entire population ($t = \tau$). We use the method of quantifying the short-term coalescent rate from the individual-based simulation, as described in [25], to determine the expected heterozygosity at a neutral locus. Briefly, at the beginning of the simulation, all N individuals carry distinct neutral alleles, as the neural allele of the ith individual is represented by the "ancestral number" i ($= 1,\ldots,N$). Then, let $q_i(t)$ be the frequency of ancestral number i at time t during simulation. As described above, $q_i(0) = 1/N$ for all i. As a result of the selective sweep, $q_i(\tau) = 0$ for many i, while $\sum_{i=1}^{N} q_i(\tau) = 1$. Assuming that new neutral mutations between time 0 and τ can be ignored, the expected heterozygosity at $t = \tau$ is given by

$$H = H_0\left(1 - \sum_{i=1}^{N} q_i^2\right),$$

(cf. [25]).

The results of the simulation model are presented in Figures 1 and 2. As expected, the heterozygosity is lower than predicted by the deterministic model. This can be adjusted by adjusting the initial frequency in the deterministic model (i.e., by shortening the length of the trajectory).

Discussion

While adaptive evolution in reality follows complex patterns (demography, heterogeneous selection pressures, spatial structure, mating behavior, etc.), such processes can often be accurately described within the idealized framework formed by standard population-genetic assumptions (constant homogeneous selection pressures, constant population size, random mating). Deviations from standard assumptions - particularly heterogeneities in selective pressures - are obviously important in allopatry and

parapatry. However, even individuals living in sympatry might experience substantial differences in selective pressures. Examples include selection for herbicide resistance in weeds [26–28], stress tolerance in insects and weeds in agriculture, insecticide resistance in bed bugs [29–32], drug resistance in vector borne diseases (see below). Whereas in these examples candidate regions under selection might be inferred with population-genetic methods that build up on standard theory, substantial errors could result when attempting to reconstruct the underlying evolutionary dynamics (e.g., estimating selection coefficients, speed of evolution, recombination rates, etc.) from the selective sweep patterns. To avoid misinferences under such scenarios, it is therefore necessary to validate the applicability of standard population-genetic theory, and - if appropriate - adapt existing theory, particularly since many of the mentioned examples are matters of economic relevance and/or global health interest.

For instance, Plasmodium parasites causing human malaria typically experience different 'environmental conditions' depending on characteristics of human hosts determining selective regimes (drug treatment, drug dosage, immune response, levels of host-acquired or natural immunity, etc.). Parasites conferring resistance to antimalarial drugs are advantageous only in hosts treated with the respective drugs, whereas they are slightly deleterious in untreated hosts due to metabolic costs. In parallel, sexual reproduction occurs inside the mosquito vector, randomly but exclusively between parasites that were extracted from the same host, manifesting another deviation from standard assumption. Heterogeneous selection pressures act also on a spatial scale because drug-deployment policies and control interventions are country specific. This is particularly relevant along the borders of Cambodia, Laos, Myanmar, and Thailand where the containment of emerging artemisinin resistance is of fundamental importance to sustain successful malaria control [33]. Inferences based on standard population-genetic assumptions might be misleading as parasites experience highly varying selective environments and severe inbreeding is immanent to the specifics of malaria transmission.

More generally, parasites or pathogens that sexually reproduce within hosts might experience radically heterogeneous selection pressures, as immune responses may occur differently across organs or within specific tissues. Sexual reproduction might be common even in fungal pathogens [34]. In agriculture patches of contrasting selective regimes are created in sympatry by human interventions (cf. [35,36]). The use of fertilizer, manure, herbicides, pesticides along with interventions such as plowing and irrigation varies across farmed land. Therefore, insects or weeds might experience radically different selective conditions across nearby acres. A striking example of a rapid evolutionary change under such a setting is the fast progression of glyphosate ("roundup") resistance in many species of weeds, economically challenging US agriculture. Genetic understanding of glyphosate resistance will require the detection and analyses of selective sweeps in the plants, including those reproducing by self-pollination and long-distance seed dispersal.

In this study we introduced a model for heterogeneous selection in sympatry within a haploid population that randomly disperses across patches in every generation. Viability selection acts differently within the patches and mating occurs randomly within or between patches. In the limiting case that mating occurs randomly between all patches, the model reduced to the hard-selection Levene model (cf. [23], Ch. 6), which is identical to the standard selection model. However, if mating occurs exclusively within demes, the deviations from the standard model can be substantial. We showed that the dynamics at a single selected locus are independent of the dispersal pattern. Namely, they are solely determined by the average selection intensities across patches. However, as soon as two or more linked loci are considered deviations from standard-population-genetic assumptions become apparent. Particularly, we studied how the genetic variation at a neutral locus is affected as a beneficial mutation sweeps at a nearby linked locus.

We were able to derive an analytic solution for the allele frequencies at a neutral locus after the beneficial mutation became fixed. As the analytic solution is complicated we also derived approximations, which allow for clear and simple interpretations. Namely they reflect the frequency change driven by the selective pressure averaged over patches, however adjusted by a factor determining the relative importance of the patches. As long as differences in selection pressures are moderate the hitchhiking effect is accurately described by standard population-genetic theory. However, if selection pressures are extreme as it might be the case in the above mentioned examples, heterogeneities in selection pressures in combination with intra-patch mating leads to stronger reductions in genetic variation than predicted by the standard model. The reason is as follows. Radically reversing directional selection across patches leads to mating only between individuals carrying the allele that allows survival within the respective selective environments, thus greatly increasing the effect of inbreeding. Hence, meiotic recombination is less efficient to restore genetic variation. This effect however cannot be just summarized by an adjustment of the recombination rate. In fact the unique mating scheme leads to a process for which selection and recombination cannot be decoupled.

We also performed stochastic simulations to verify the results of the deterministic model's analytic prediction. As expected the deterministic solutions were underestimating the reduction of genetic variation at neutral loci. However, as usual this can be compensated by adjusting the effective initial frequency of the advantageous allele, which reflects the shorter allele frequency trajectory of the advantageous allele conditional on its escape from extinction by random genetic drift.

In general our results are informative to properly interpret selection coefficients when these are attempted to be measures from the patterns of selective sweeps. Unfortunately, appropriate data is unavailable for the mentioned examples to which our model would apply (pesticide and herbicide resistance). Nevertheless, as the examples are of great economic interest, and as population genetic theory continues to advance such data hopefully become available soon. Anyhow, the model is applicable to malaria where attempts have been made to link estimates of selection to the hitchhiking pattering (e.g. [19,22]).

The hitchhiking effect revealed in this study might be compared to that of another study assuming the subdivision of population into many small demes or patches [16,37]. They predicted the reduced strength of hitchhiking (higher heterozygosity), in contrast to our current result, due to population subdivision. Their model however assumes homogeneous selective pressure over demes and limited migration of individuals between demes. In such a case, the delay in the propagation of advantageous allele into the entire population provides more opportunities for recombination that breaks the hitchhiking. Most populations in nature would violate the assumptions of both studies (instantaneous dispersal among demes of the current study and homogeneous selective pressure in [16]). Further investigation is needed for the joint effect of the two forces.

Analysis

1 Single-locus Dynamics

Here, we derive the marginal dynamics at a single locus. Let p denote the frequency of allele $\mathcal{A}_1$, i.e., $p=\sum_{i=1}^{M} p_i$. The fitness of allele $\mathcal{A}_i$ in patch $\mathcal{P}_k$ is denoted by $w_i^{(k)}$. Hence, we have $W_i^{(k)}=w_1^{(k)}$ and $W_{i+m}(k)=w_2^{(k)}$ for $i\in\{1,\ldots,M\}$. Moreover, let $w_i := \sum_{k=1}^{K}(1-\beta_k)\alpha_k w_i^{(k)}$, so that we obtain $w_1=W_i$ and $w_2=W_{i+M}$ for $i\in\{1,\ldots,M\}$.

With the above notation we can derive $\bar{W}^{(k)}=pw_1^{(k)}+(1-p)w_2^{(k)}$. Thus,

$$\begin{aligned}p_l^{*(k)} &= \frac{\alpha_k\beta_k}{\overline{W}^{(k)}}\sum_{i,j=1}^{2M} p_i W_i^{(k)} p_j W_j^{(k)} R(i,j\to l)\\ &= \frac{\alpha_k\beta_k}{\overline{W}^{(k)}}\Big[\sum_{i,j=1}^{M} w_1^{(k)} 2p_i p_j R(i,j\to l)\\ &+2\sum_{i=1}^{M}\sum_{j=M+1}^{2M} w_1^{(k)} w_2^{(k)} p_i p_j R(i,j\to l)\\ &+\sum_{i,j=M+1}^{2M} w_2^{(k)} 2p_i p_j R(i,j\to l)\Big].\end{aligned}$$

Assume $l\in\{1,\ldots,M\}$. Then, by denoting the the Kronecker-δ by $\delta_{i,j}$ we obtain

$$\begin{aligned}p_l^{*(k)} &= \frac{\alpha_k\beta_k}{\overline{W}^{(k)}}\Big[\sum_{i,j=1}^{M} w_1^{(k)} 2p_i p_j \frac{\delta_{i,l}+\delta_{j,l}}{2}\\ &+2\sum_{i=1}^{M}\sum_{j=M+1}^{2M} w_1^{(k)} w_2^{(k)} p_i p_j \frac{\delta_{i,l}(1-r)+r\delta_{j,l+M}}{2}\\ &+\sum_{i,j=M+1}^{2M} w_2^{(k)} 2p_i p_j 0\Big]\\ &= \frac{\alpha_k\beta_k w_1^{(k)}}{\overline{W}^{(k)}}\Big[w_1^{(k)} p_l\sum_{i=1}^{M} p_i + w_2^{(k)}\big((1-r)p_l\sum_{j=M+1}^{2M} p_j + r p_{l+M}\sum_{i=1}^{M} p_i\big)\Big]\\ &= \frac{\alpha_k\beta_k w_1^{(k)}}{\overline{W}^{(k)}}\big(w_1^{(k)} p_l p + w_2^{(k)}(p_l(1-p)(1-r)+p_{l+M}pr)\big).\end{aligned}$$

Similarly, for $l\in\{1,\ldots,M\}$, we obtain

$$\begin{aligned}&p_{l+M}^{*(k)} =\\ &\frac{\alpha_k\beta_k w_2^{(k)}}{\overline{W}^{(k)}}\big(w_1^{(k)}(p_l(1-p)r+p_{l+M}p(1-r))+w_2^{(k)}p_{l+M}(1-p)\big).\end{aligned}$$

Therefore,

$$\begin{aligned}&\sum_{l=1}^{M} p_l^{*(k)}\\ &= \frac{\alpha_k\beta_k w_1^{(k)}}{\overline{W}^{(k)}}\big(w_1^{(k)}p^2+w_2^{(k)}(p(1-p)(1-r)+(1-p)pr)\big)\backslash\\ &= \alpha_k\beta_k w_1^{(k)} p\end{aligned}$$

and

$$\begin{aligned}&\sum_{l=M+1}^{2M} p_l^{*(k)}\\ &= \frac{\alpha_k\beta_k w_2^{(k)}}{\overline{W}^{(k)}}\big(w_1^{(k)}(p(1-p)r+(1-p)p(1-r))+w_2^{(k)}(1-p)^2\big)\\ &= \alpha_k\beta_k w_2^{(k)}(1-p).\end{aligned}$$

Using $\bar{W}=pw_1+(1-p)w_2$ a similar calculation as above gives,

$$\begin{aligned}p_l^{*} &= \frac{1}{\overline{W}}\sum_{i,j=1}^{2M} p_i W_i p_j W_j R(i,j\to l)\\ &= \frac{1}{\overline{W}}\Big(\sum_{i,j=1}^{M} p_i w_1 p_j w_1 R(i,j\to l)+2\sum_{i=1}^{M}\sum_{j=M+1}^{2M} p_i w_1 p_j w_2 R(i,j\to l)\\ &+\sum_{i,j=M+1}^{2M} p_i w_2 p_j w_2 R(i,j\to l)\Big).\end{aligned}$$

Hence, for $l\in\{1,\ldots,M\}$, we have

$$\begin{aligned}p_l^{*} &= \frac{1}{\overline{W}}\sum_{i,j=1}^{2M} p_i W_i p_j W_j R(i,j\to l)\\ &= \frac{1}{\overline{W}}\big(p_l w_1^2\sum_{j=1}^{M} p_j + p_l w_1 w_2(1-r)\sum_{j=1}^{M} p_{j+M} + p_{l+M} w_1 w_2 r\sum_{i=1}^{M} p_i + 0\big)\\ &= \frac{1}{\overline{W}}\big(p p_l w_1^2 + (1-p)p_l w_1 w_2(1-r) + p p_{l+M} w_1 w_2 r\big).\end{aligned}$$

Similarly, for $l\in\{1,\ldots,M\}$, we have

$$p_{l+M}^{*} = \frac{1}{\overline{W}}\big((1-p)p_l w_1 w_2 r + p p_{l+M} w_1 w_2(1-r) + (1-p)p_{l+M} w_2^2\big)$$

Hence,

$$\sum_{l=1}^{M} p_l^* = pw_1 \quad \text{and} \quad \sum_{l=1}^{M} p_{l+M}^* = (1-p)w_2 .$$

Therefore, (6) simplifies to

$$p_l' = \frac{p_l^* + \sum_{k=1}^{K} p_l^{*(k)}}{\sum_{j=1}^{2M} (p_l^* + \sum_{k=1}^{K} p_l^{*(k)})} = \frac{p_l^* + \sum_{k=1}^{K} p_l^{*(k)}}{pw_1 + (1-p)w_2 + \sum_{k=1}^{K} \alpha_k \beta_k (w_1^{(k)} p + w_2^{(k)}(1-p))}$$

$$= \frac{p_l^* + \sum_{k=1}^{K} p_l^{*(k)}}{p \sum_{k=1}^{K} \alpha_k w_1^{(k)} + (1-p) \sum_{k=1}^{K} \alpha_k w_2^{(k)}} = \frac{p_l^* + \sum_{k=1}^{K} p_l^{*(k)}}{p\lambda + (1-p)\mu} .$$

Hence, it is easily seen that

$$p' = \sum_{l=1}^{M} p_l' = \frac{p\lambda}{p\lambda + (1-p)\mu} .$$

2 Two-locus Dynamics

Here we derive $Q_{l'}$ and $R_{l'}$. First, we need to drive $p_l^{*(k)}$ and p_l^* from (2) and (5), respectively. For $l \in \{1, \ldots, M\}$, straightforward calculation (similar as in Analysis, subsection 1) yields.

$$p_l^{*(k)} = \frac{\alpha_k \beta_k}{\overline{W}^{(k)}} \sum_{i,j=1}^{2M} p_i W_i^{(k)} p_j W_j^{(k)} R(i,j \to l)$$

$$= \frac{\alpha_k \beta_k}{\overline{W}^{(k)}} \Big(\sum_{j=1}^{M} p_l W_l^{(k)} p_j W_j^{(k)} + r \sum_{j=1}^{M} p_{l+M} W_{l+M}^{(k)} p_j W_j^{(k)} + (1-r) \sum_{j=1}^{M} p_l W_l^{(k)} p_{M+j} W_{M+j}^{(k)} \Big)$$

$$= \frac{p_l W_l^{(k)} \alpha_k \beta_k}{\overline{W}^{(k)}} \Big(\sum_{j=1}^{2M} p_j W_j^{(k)} + r \sum_{j=1}^{M} p_{l+M} (W_{l+M}^{(k)} p_j W_j^{(k)} - p_l W_l^{(k)} p_{M+j} W_{M+j}^{(k)}) \Big)$$

$$= p_l W_l^{(k)} \alpha_k \beta_k + \frac{r \alpha_k \beta_k}{\overline{W}^{(k)}} \sum_{j=1}^{M} (p_{l+M} W_{l+M}^{(k)} p_j W_j^{(k)} - p_l W_l^{(k)} p_{M+j} W_{M+j}^{(k)})$$

$$= \alpha_k \beta_k p_l w_1^{(k)} + \frac{\alpha_k \beta_k r w_1^{(k)} w_2^{(k)}}{\overline{W}^{(k)}} \sum_{j=1}^{M} (p_{l+M} p_j - p_l p_{M+j})$$

$$= \alpha_k \beta_k p_l w_1^{(k)} - \frac{\alpha_k \beta_k r w_1^{(k)} w_2^{(k)}}{\overline{W}^{(k)}} p(1-p)(Q_l - R_l) .$$

Clearly, we have

$$\overline{W}^{(k)} = w_1^{(k)} p + w_2^{(k)} (1-p) .$$

For $l \in \{M+1, \ldots, 2M\}$ the calculation is similar. Summarizing we obtain

$$p_l^{*(k)} = \begin{cases} p_l \alpha_k \beta_k w_1^{(k)} - \frac{r \alpha_k \beta_k w_1^{(k)} w_2^{(k)}}{\overline{W}^{(k)}} p(1-p)(Q_l - R_l) & \text{for } l \in \{1, \ldots, M\}, \\ p_l \alpha_k \beta_k w_2^{(k)} + \frac{r \alpha_k \beta_k w_1^{(k)} w_2^{(k)}}{\overline{W}^{(k)}} p(1-p)(Q_{l-M} - R_{l-M}) & \text{for } l \in \{M+1, \ldots, 2M\}. \end{cases}$$

Exactly the same calculation as above yields

$$p_l^* = \frac{1}{\overline{W}} \sum_{i,j=1}^{2M} p_i W_i p_j W_j R(i,j \to l) = p_l w_1 - \frac{r w_1 w_2}{\overline{W}} p(1-p)(Q_l - R_l) .$$

Hence,

$$p_l^* = \begin{cases} p_l w_1 - \frac{r w_1 w_2}{\overline{W}} p(1-p)(Q_l - R_l) & \text{for } l \in \{1, \ldots, M\}, \\ p_l w_2 - \frac{r w_1 w_2}{\overline{W}} p(1-p)(Q_{l-M} - R_{l-M}) & \text{for } l \in \{M+1, \ldots, 2M\}. \end{cases} \quad (26)$$

Therefore, for $l \in \{1, \ldots, M\}$, we have

$$p_l^* + \sum_{k=1}^{K} p_l^{*(k)} = p_l \sum_{k=1}^{K} \alpha_k (1-\beta_k) w_1 + \alpha_k \beta_k w_1 - rp(1-p)(Q_l - R_l) \left(\frac{w_1 w_2}{\overline{W}} + \sum_{k=1}^{K} \frac{\alpha_k \beta_k w_1^{(k)} w_2^{(k)}}{\overline{W}^{(k)}} \right) = p_l \lambda - rp(1-p)(Q_l - R_l)\vartheta_p \quad (27)$$

where

$$\vartheta_p = \frac{w_1 w_2}{\overline{W}} + \sum_{k=1}^{K} \frac{\alpha_k \beta_k w_1^{(k)} w_2^{(k)}}{\overline{W}^{(k)}} . \quad (28)$$

Similarly, for $l \in \{1, \ldots, M\}$, we have

$$p_{l+M}^* + \sum_{k=1}^{K} p_{l+M}^{*(k)} = p_{l+M} \mu - rp(1-p)(Q_l - R_l)\vartheta_p . \quad (29)$$

Consequently, because $1 = \sum_{l=1}^{M} R_l = \sum_{l=1}^{M} Q_l$, we have

$$\sum_{l=1}^{M} (p_l^* + \sum_{k=1}^{K} p_l^{*(k)}) = p\lambda ,$$

$$\sum_{l=M+1}^{2M}(p_l^*+\sum_{k=1}^{K}p_l^{*(k)})=(1-p)\mu,$$

$$\sum_{l=1}^{2M}(p_l^*+\sum_{k=1}^{K}p_l^{*(k)})=p\lambda+(1-p)\mu=:\overline{w}. \quad (30)$$

In particular, by combining (6) with (29) or (27), and (30) we obtain.

$$p_l^{'}=\begin{pmatrix} p_l\frac{\lambda}{\overline{w}}-(Q_l-R_l)\frac{rp(1-p)\vartheta_p}{\overline{w}} & \text{for } l\in\{1,\ldots,M\},\\ p_l\frac{\mu}{\overline{w}}-(Q_{l-M}-R_{l-M})\frac{rp(1-p)\vartheta_p}{\overline{w}} & \text{for } l\in\{M+1,\ldots,2M\}.\end{pmatrix} \quad (31)$$

Therefore, we deduce from (26) and (31).

$$Q_l^{'}=\frac{p_l\lambda-rp(1-p)(Q_l-R_l)\vartheta_p}{\lambda\sum_{j=1}^{M}p_j}=Q_l-\frac{r\vartheta_p}{\lambda}(1-p)(Q_l-R_l). \quad (32)$$

Similarly, we obtain

$$R_l^{'}=R_l+\frac{r\vartheta_p}{\mu}p(Q_l-R_l). \quad (33)$$

We have

$$Q_l^{'}-R_l^{'}=(Q_l-R_l)\Lambda_p, \quad (34)$$

where

$$\Lambda_p=1-\frac{r\vartheta_p(p\lambda+(1-p)\mu)}{\lambda\mu}. \quad (35)$$

Iteration of (34) yields.

$$Q_l(t)-R_l(t)=(Q_l(0)-R_l(0))\prod_{\tau=0}^{t-1}\Lambda_{p_\tau} \quad (36)$$

Hence, from iterating (32) using first (36) and then (28) and (9) we obtain.

$$Q_l(T+1)=Q_l(T)-\frac{r\vartheta_{p_T}}{\lambda}(1-p_T)(Q_l(T)-R_l(T))$$

$$=Q_l(0)-\sum_{t=0}^{T}\frac{r\vartheta_{p_t}}{\lambda}(1-p_t)(Q_l(t)-R_l(t))$$

$$=Q_l(0)-(Q_l(0)-R_l(0))\sum_{t=0}^{T}\frac{r\vartheta_{p_t}}{\lambda}(1-p_t)\prod_{\tau=0}^{t-1}\Lambda_{p_\tau}$$

$$=Q_l(0)-(Q_l(0)-R_l(0))\frac{r}{\lambda}$$

$$\times\sum_{t=0}^{T}\Big(\frac{w_1w_2}{p_tw_1+(1-p_t)w_2}+\sum_{k=1}^{K}\frac{\alpha_k\beta_kw_1^{(k)}w_2^{(k)}}{p_tw_1^{(k)}+(1-p_t)w_2^{(k)}}\Big)\frac{(1-p_0)\mu^t}{p_0\lambda^t+(1-p_0)\mu^t}\prod_{\tau=0}^{t-1}\Lambda_{p_\tau}$$

$$=Q_l(0)-(Q_l(0)-R_l(0))\frac{r}{\lambda}$$

.

$$\times\sum_{t=0}^{T}\Big(\frac{w_1w_2}{p_0\lambda^tw_1+(1-p_0)\mu^tw_2}+\sum_{k=1}^{K}\frac{\alpha_k\beta_kw_1^{(k)}w_2^{(k)}}{p_0\lambda^tw_1^{(k)}+(1-p_0)\mu^tw_2^{(k)}}\Big)(1-p_0)\mu^t\prod_{\tau=0}^{t-1}\Lambda_{p_\tau}$$

$$=Q_l(0)-\frac{r}{\lambda}(Q_l(0)-R_l(0))\sum_{t=0}^{T}\Big(\frac{w_1w_2}{\frac{p_0}{1-p_0}(\frac{\lambda}{\mu})^tw_1+w_2}+\sum_{k=1}^{K}\frac{\alpha_k\beta_kw_1^{(k)}w_2^{(k)}}{\frac{p_0}{1-p_0}(\frac{\lambda}{\mu})^tw_1^{(k)}+w_2^{(k)}}\Big)\prod_{\tau=0}^{t-1}\Lambda_{p_\tau}$$

$$=Q_l(0)-\frac{r}{\lambda}(Q_l(0)-R_l(0))\sum_{t=0}^{T}\Big(\frac{w_1}{\frac{p_0}{1-p_0}(\frac{\lambda}{\mu})^t\frac{w_1}{w_2}+1}+\sum_{k=1}^{K}\frac{w_1^{(k)}\alpha_k\beta_k}{\frac{p_0}{1-p_0}(\frac{\lambda}{\mu})^t\frac{w_1^{(k)}}{w_2^{(k)}}+1}\Big)\prod_{\tau=0}^{t-1}\Lambda_{p_\tau}$$

$$=Q_l(0)-\frac{r}{\lambda}(Q_l(0)-R_l(0))\Big(w_1\sum_{t=0}^{T}\frac{\prod_{\tau=0}^{t-1}\Lambda_{p_\tau}}{\frac{p_0}{1-p_0}(\frac{\lambda}{\mu})^t\frac{w_1}{w_2}+1}+\sum_{k=1}^{K}\alpha_k\beta_kw_1^{(k)}\sum_{t=0}^{T}\frac{\prod_{\tau=0}^{t-1}\Lambda_{p_\tau}}{\frac{p_0}{1-p_0}(\frac{\lambda}{\mu})^t\frac{w_1^{(k)}}{w_2^{(k)}}+1}\Big).$$

By combining (35), (28), and (9) we see that Λ_{pt} is given by (12). Hence, the above yields (11), by substitution T by $T-1$.

3 Equilibrium Heterozygosity

Obviously, (15) has the form.

$$\hat{Q}_l=Q_l(0)-(Q_l(0)-R_l(0))A_r,$$

where A_r does not depend on $Q_l(0)$ or $R_l(0)$. The equilibrium heterozygosity is given by

$$\hat{H} = \frac{M}{M-1}\sum_{l=1}^{M}\hat{Q}_l(1-\hat{Q}_l),$$

because the beneficial allele becomes fixed at equilibrium.

Now, assume that initially only a single copy of the beneficial mutation arises. Hence, we have $Q_l(0)=1$ with probability $R_l(0)$ and $Q_l(0)=0$ with probability $1-R_l(0)$, i.e., $P(Q_l(0)=1|R_1(0),\ldots,R_M(0))=R_l(0)$ and $P(Q_l(0)=0|R_1(0),\ldots,R_M)=1-R_l(0)$. Therefore, we have.

$$\begin{aligned}
&E(\hat{H}|R_1(0),\ldots,R_M)=\\
&E(\tfrac{M}{M-1}\sum_{l=1}^{M}(Q_l(0)-(Q_l(0)-R_l(0))A_r)\\
&(1-Q_l(0)+(Q_l(0)-R_l(0))A_r))|R_1(0),\ldots,R_M(0)),\\
&=\frac{M}{M-1}\sum_{l=1}^{M}E((Q_l(0)-(Q_l(0)-R_l(0))A_r)\\
&(1-Q_l(0)+(Q_l(0)-R_l(0))A_r))|R_1(0),\ldots,R_M(0))\\
&=\frac{M}{M-1}\sum_{l=1}^{M}R_l(0)((1-R_l(0))A_r(1-(1-R_l(0))A_r)\\
&+\sum_{\substack{j=1\\ j\neq l}}^{M}R_l(0)A_r(1-R_l(0)A_r))\\
&=\frac{M}{M-1}\sum_{l=1}^{M}R_l(0)(1-R_l(0))A_r(1-(1-R_l(0))A_r)\\
&+\frac{M}{M-1}\sum_{l=1}^{M}R_l(0)A_r(1-R_l(0)A_r)\sum_{\substack{j=1\\ j\neq l}}^{M}R_j(0)\\
&=\frac{M}{M-1}\sum_{l=1}^{M}R_l(0)(1-R_l(0))A_r(1-(1-R_l(0))A_r)\\
&+\frac{M}{M-1}\sum_{l=1}^{M}R_l(0)(1-R_l(0))A_r(1-R_l(0)A_r)\\
&=\frac{M}{M-1}\sum_{l=1}^{M}R_l(0)(1-R_l(0))(2A_r-A_r^2)\\
&=\frac{M}{M-1}(2A_r-A_r^2)\sum_{l=1}^{M}R_l(0)(1-R_l(0)).
\end{aligned}$$

Since $H_0=\frac{M}{M-1}\sum_{l=1}^{M}R_l(0)(1-R_l(0))$, we have

$$E(\hat{H}|R_1(0),\ldots,R_M)=E(\hat{H}|H_0)=(2A_r-A_r^2)H_0.$$

Hence we have

$$\begin{aligned}
E(\hat{H})&=E(E(\hat{H}|R_1(0),\ldots,R_M))\\
&=\frac{M}{M-1}(2A_r-A_r^2)E(\sum_{l=1}^{M}R_l(0)(1-R_l(0))).
\end{aligned}$$

Since

$$E(H_0)=E(\sum_{l=1}^{M}R_l(0)(1-R_l(0)))$$

is the heterozygosity before the sweep, we see that the relative heterozygosity

$$\mathcal{H}:=\frac{E(\hat{H})}{E(H_0)}=\frac{E(\hat{H}|H_0)}{H_0}=2A_r-A_r^2$$

is independent of the initial allele-frequency distribution before the sweep

4 Approximations

Let $f(x,y):=\frac{1}{\frac{p}{x}+\frac{1-p}{y}}$ for $x,y\in\mathbb{R}^+$. Its Hessian matrix is calculated to be

$$H=\begin{pmatrix}\frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial y\partial x}\\ \frac{\partial^2 f}{\partial x\partial y} & \frac{\partial^2 f}{\partial y^2}\end{pmatrix}=\frac{2p(1-p)}{((1-p)x+py)^3}\begin{pmatrix}-y^2 & xy\\ xy & -x^2\end{pmatrix}.$$

Clearly, we have $-\frac{2p(1-p)y^2}{((1-p)x+py)^3}<0$ and $\det H=0$, i.e., the leading minors of H are non-positive. Hence, f is concave but not strictly concave (note that $f(x,x)=x/2$). Hence, for positive random variables X and Y the Jensen's inequality for higher dimensions yields.

$$E[f(X,Y)]\leq f(E[X],E[Y]).$$

Now, choose $X(k)=w_1^{(k)}$ and $Y(k)=w_2^{(k)}$ with probability $\alpha_k\beta_k$ for $k=1,\ldots,K$, $X(0)=w_1$ and $Y(0)=w_2$ with probability $\sum_{k=1}^{K}\alpha_k(1-\beta_k)$. Then the Jensen's, inequality gives.

$$\vartheta_p=\frac{\sum_{k=1}^{K}\alpha_k(1-\beta_k)w_1w_2}{pw_1+(1-p)w_2}+\sum_{k=1}^{K}\frac{\alpha_k\beta_k w_1^{(k)}w_2^{(k)}}{pw_1^{(k)}+(1-p)w_2^{(k)}}$$

$$= \frac{\sum_{k=1}^{K} \alpha_k(1-\beta_k)}{\frac{p}{w_2} + \frac{1-p}{w_1}} + \sum_{k=1}^{K} \frac{\alpha_k \beta_k}{\frac{p}{w_2^{(k)}} + \frac{1-p}{w_1^{(k)}}} \leq \frac{1}{\frac{p}{\mu} + \frac{1-p}{\lambda}} = \frac{\lambda\mu}{p\lambda + (1-p)\mu}.$$

Using this inequality yields

$$\Lambda_p = 1 - \frac{r\vartheta_p(p\lambda + (1-p)\mu)}{\lambda\mu} > 1 - r.$$

Proof of Result 3. First, we approximate Λ_{p_t} by $1-r$. Therefore, we obtain the approximation (20), which we can rewrite as.

$$\hat{Q}_l \approx Q_l(0) - \frac{r}{\lambda}(Q_l(0) - R_l(0)) \sum_{k=0}^{K} \alpha_k w_1^{(k)} \sum_{t=0}^{\infty} g_k(t), \quad (37)$$

with

$$g_k(t) := \frac{a^t}{c_k b^t + 1}, \quad (38)$$

where

$$a := 1-r, \quad b := \frac{\lambda}{\mu}, \quad \text{and} \quad c_k := \frac{w_1^{(k)} p_0}{w_2^{(k)}(1-p_0)}. \quad (39)$$

As in eq. 45 in [18], we can approximate.

$$\sum_{t=0}^{\infty} g_k(t) = \frac{1}{2}\int_0^{\infty} g_k(x)dx + \frac{1}{2}\int_1^{\infty} g_k(x)dx =: B_k. \quad (40)$$

The integrals can be expressed in terms of the hypergeometric function. Exactly the same derivations as in the proofs of Theorem 1 and, Remarks 1 and 2 in [18] yield the desired expressions. The hypergeometric function can be further approximated as in the proofs of Theorem 2 and Remark 3 in [18]. Finally, assuming $p_0 \approx 0$, the same approximations as in Theorems 3 and 4 in [18] can be applied. According to eq. 95 in [18] we obtain.

$$\begin{aligned} B_k \approx{}& \frac{1}{\log a}(c_k^{-\frac{\log a}{\log b}} - 1) - \frac{1-a}{2a\log a} \\ &- \frac{1}{\log \frac{a}{b}} c_k^{-\frac{\log a}{\log b}} - \frac{c_k}{\log ab}(c_k^{-\frac{\log ab}{\log b}} - 1) \\ &+ \frac{c_k}{2\log ab}\left(\frac{1}{ab} - 1\right). \end{aligned} \quad (41)$$

If $p_0 \approx 0$, we have $c_k \approx 0$. We have to keep in mind that the expressions c_k^x are delicate for $x \approx 0$. However, since $(1-p_0)^x \approx 1$, we have $c_k^x \approx d_k^x$, with $d_k = p_0 \frac{w_1^{(k)}}{w_2^{(k)}}$. We obtain.

$$B_k \approx \frac{1}{\log a}(d_k^{-\frac{\log a}{\log b}} - 1) - \frac{1-a}{2a\log a} - \frac{1}{\log \frac{a}{b}} d_k^{-\frac{\log a}{\log b}}. \quad (42)$$

By combining (37), (40), and (43) and a little rearrangement, we obtain (21).

Furthermore, for small r we additionally have $\log a \approx -r$, such that.

$$B_k \approx \frac{1}{r}(1 - d_k^{\frac{r}{\log b}}) + \frac{1}{2(1-r)} + \frac{1}{r + \log b} d_k^{\frac{r}{\log b}}. \quad (43)$$

Clearly, for $r \approx 0$, the first second and third term in the above expression are negligible compared with the first term since $\frac{1}{r}$ is large. Hence, we have

$$B_k \approx \frac{1}{r}(1 - d_k^{\frac{r}{\log b}}). \quad (44)$$

Now, (22) follows from combination of (37), (40), and (44).

Acknowledgments

The authors gratefully acknowledge the work of Hua Chen and an anonymous reviewer as well as their helpful comments on an earlier draft of this manuscript.

Author Contributions

Contributed reagents/materials/analysis tools: KS YK. Wrote the paper: KS YK.

References

1. Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genetics Research 23: 23–35.
2. Kaplan NL, Hudson RR, Langley CH (1989) The "hitchhiking effect" revisited. Genetics 123: 887–99.
3. Stephan W, Wiehe THE, Lenz MW (1992) The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. Theoretical Population Biology 41: 237–254.
4. Barton NH (2000) Genetic hitchhiking. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 355: 1553–1562.
5. Etheridge A, Pfaffelhuber P, Wakolbinger A (2006) An approximate sampling formula under genetic hitchhiking. The Annals of Applied Probability 16: 685–729.
6. Thornton KR, Jensen JD, Becquet C, Andolfatto P (2007) Progress and prospects in mapping recent selection in the genome. Heredity (Edinb) 98: 340–8.
7. Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19: 711–22.
8. Stephan W (2010) Detecting strong positive selection in the genome. Mol Ecol Resour 10: 863–72.
9. Hermisson J, Pennings PS (2005) Soft Sweeps: Molecular Population Genetics of Adaptation From Standing Genetic Variation. Genetics 169: 2335–2352.
10. Pennings PS, Hermisson J (2006) Soft sweeps iii: the signature of positive selection from recurrent mutation. PLoS Genet 2: e186.
11. Teshima KM, Przeworski M (2006) Directional positive selection on an allele of arbitrary dominance. Genetics 172: 713–718.
12. Chevin LM, Hospital F (2008) Selective sweep at a quantitative trait locus in the presence of background genetic variation. Genetics 180: 1645–1660.
13. Li H, Stephan W (2006) Inferring the demographic history and rate of adaptive substitution in drosophila. PLoS Genet 2: e166.

14. Innan H, Kim Y (2008) Detecting local adaptation using the joint sampling of polymorphism data in the parental and derived populations. Genetics 179: 1713–1720.
15. Kim Y, Gulisija D (2010) Signatures of recent directional selection under different models of population expansion during colonization of new selective environments. Genetics 184: 571–585.
16. Kim Y, Maruki T (2011) Hitchhiking effect of a beneficial mutation spreading in a subdivided population. Genetics 189: 213–226.
17. Schneider KA, Kim Y (2010) An analytical model for genetic hitchhiking in the evolution of antimalarial drug resistance. Theor Popul Biol 78: 93–108.
18. Schneider K, Kim Y (2011) Approximations for the hitchhiking effect caused by the evolution of antimalarial-drug resistance. Journal of Mathematical Biology 62: 789–832.
19. Nair S, Williams JT, Brockman A, Paiphun L, Mayxay M, et al. (2003) A Selective Sweep Driven by Pyrimethamine Treatment in Southeast Asian Malaria Parasites. Mol Biol Evol 20: 1526–1536.
20. Nash D, Nair S, Mayxay M, Newton PN, Guthmann JP, et al. (2005) Selection strength and hitchhiking around two anti-malarial resistance genes. Proceedings of the Royal Society B: Biological Sciences 272: 1153–1161.
21. McCollum AM, Mueller K, Villegas L, Udhayakumar V, Escalante AA (2007) Common origin and fixation of plasmodium falciparum dhfr and dhps mutations associated with sulfadoxinepyrimethamine resistance in a low-transmission area in south america. Antimicrob Agents Chemother 51: 2085–2091.
22. McCollum AM, Schneider KA, Grifing SM, Zhou Z, Kariuki S, et al. (2012) Differences in selective pressure on dhps and dhfr drug resistant mutations in western kenya. Malar J 11: 77.
23. Nagylaki T (1992) Introduction to theoretical population genetics, volume 21 of *Biomathematics*. Berlin: Springer-Verlag, xii+369 pp.
24. Levene H (1953) Genetic equilibrium when more than one ecological niche is available. Am Nat 87: 331–333.
25. Kim Y, Wiehe T (2009) Simulation of dna sequence evolution under models of recent directional selection. Brief Bioinform 10: 84–96.
26. Neve P (2008) Simulation modelling to understand the evolution and management of glyphosate resistance in weeds. Pest Manag Sci 64: 392–401.
27. Powles SB, Yu Q (2010) Evolution in action: plants resistant to herbicides. Annu Rev Plant Biol 61: 317–347.
28. Powles SB (2010) Gene amplification delivers glyphosate-resistant weed evolution. Proc Natl Acad Sci U S A 107: 955–956.
29. Bai X, Mamidala P, Rajarapu SP, Jones SC, Mittapalli O (2011) Transcriptomics of the bed bug (cimex lectularius). PLoS One 6: e16336.
30. Adelman ZN, Kilcullen KA, Koganemaru R, Anderson MAE, Anderson TD, et al. (2011) Deep sequencing of pyrethroid-resistant bed bugs reveals multiple mechanisms of resistance within a single population. PLoS ONE 6: e26228.
31. Jones SC, Bryant JL (2012) Ineffectiveness of over-the-counter total-release foggers against the bed bug (heteroptera: Cimicidae). J Econ Entomol 105: 957–963.
32. Booth W, Saenz VL, Santangelo RG, Wang C, Schal C, et al. (2012) Molecular markers reveal infestation dynamics of the bed bug (hemiptera: Cimicidae) within apartment buildings. J Med Entomol 49: 535–546.
33. Cheeseman IH, Miller BA, Nair S, Nkhoma S, Tan A, et al. (2012) A major genome region underlying artemisinin resistance in malaria. Science 336: 79–82.
34. Heitman J (2010) Evolution of eukaryotic microbial pathogens via covert sexual reproduction. Cell Host Microbe 8: 86–99.
35. Amos JN, Bennett AF, Mac Nally R, Newell G, Pavlova A, et al. (2012) Predicting landscapegenetic consequences of habitat loss, fragmentation and mobility for multiple species of woodland birds. PLoS One 7: e30888.
36. Bianchi FJJA, Booij CJH, Tscharntke T (2006) Sustainable pest regulation in agricultural landscapes: a review on landscape composition, biodiversity and natural pest control. Proc Biol Sci 273: 1715–1727.
37. Slatkin M, Wiehe T (1998) Genetic hitch-hiking in a subdivided population. Genet Res 71: 155–160.