Genetic demixing and evolution in linear stepping stone models

K. S. Korolev[*],

Department of Physics and FAS Center for Systems Biology, Harvard University, Cambridge, Massachusetts 02138, USA

Mikkel Avlund,

Department of Physics and FAS Center for Systems Biology, Harvard University, Cambridge, Massachusetts 02138, USA and Niels Bohr Institute, University of Copenhagen, Copenhagen 2100, Denmark

Oskar Hallatschek, and

Max Planck Research Group for Biological Physics and Evolutionary Dynamics, Max Planck Institute for Dynamics and Self-Organization (MPIDS), Göttingen 37073, Germany

David R. Nelson[†]

Department of Physics and FAS Center for Systems Biology, Harvard University, Cambridge, Massachusetts 02138, USA

Abstract

Results for mutation, selection, genetic drift, and migration in a one-dimensional continuous population are reviewed and extended. The population is described by a continuous limit of the stepping stone model, which leads to the stochastic Fisher-Kolmogorov-Petrovsky-Piscounov equation with additional terms describing mutations. Although the stepping stone model was first proposed for population genetics, it is closely related to "voter models" of interest in nonequilibrium statistical mechanics. The stepping stone model can also be regarded as an approximation to the dynamics of a thin layer of actively growing pioneers at the frontier of a colony of micro-organisms undergoing a range expansion on a Petri dish. The population tends to segregate into monoallelic domains. This segregation slows down genetic drift and selection because these two evolutionary forces can only act at the boundaries between the domains; the effects of mutation, however, are not significantly affected by the segregation. Although fixation in the neutral well-mixed (or "zero-dimensional") model occurs exponentially in time, it occurs only algebraically fast in the one-dimensional model. An unusual sublinear increase is also found in the variance of the spatially averaged allele frequency with time. If selection is weak, selective sweeps occur exponentially fast in both well-mixed and one-dimensional populations, but the time constants are different. The relatively unexplored problem of evolutionary dynamics at the edge of an expanding circular colony is studied as well. Also reviewed are how the observed patterns of genetic diversity can be used for statistical inference and the differences are highlighted between the well-mixed and one-dimensional models. Although the focus is on two alleles or variants, *q*-allele Potts-like models of gene segregation are considered as well. Most of the analytical results are checked with simulations and could be tested against recent spatial experiments on range expansions of inoculations of *Escherichia coli* and *Saccharomyces cerevisiae*.

I. INTRODUCTION

The quantitative theory of evolution is an important open problem. The theory is necessary to determine the history of species migrations, and it could shed light on the origin and development of life. Moreover, a better understanding of the evolutionary dynamics could help control epidemics (Murray, 2003), fight diseases with an evolutionary character such as cancer and acquired immune deficiency syndrome (Nowak, 2006), and guide the engineering of artificial evolution for practical applications (Bar-Yam, 2005; Poli *et al.*, 2008).

Most of the current understanding of evolutionary dynamics comes from population genetics, a scientific discipline that studies how evolutionary forces shape the genetic diversity of populations. The majority of theoretical models and experiments in population genetics study only one or a few well-mixed populations, i.e., populations without spatial structure, where every individual is equally likely to interact with any other individual inside the same population. Micro-organisms growing and evolving in a well-mixed liquid culture provide an important example. While nonspatial models are often easier to analyze than spatial ones, they do miss what can be essential features of natural populations.

In nature, organisms often occupy areas that are much larger than the square of the dispersal distance, i.e., the distance typically traveled by an individual in one generation. This causes two main problems for well-mixed-population models. First, well-mixed-population models underestimate the role of genetic drift (fluctuations due to the discreteness of the number of individuals). The difference arises because the organisms can only interact with their neighbors, and the number of neighbors within the dispersal distance is much smaller than the total number of organisms in the entire population. Second, well-mixed-population models neglect the spatial structure of the population that can be created by external factors or by internal dynamics. Such spatial structures often exist, and, as we show in this paper, they can significantly affect evolutionary processes in the population.

Well-mixed-population models are particularly inadequate when applied to expanding populations. Expansions are very common in biology. Species spread to new territories from the locations where they first evolved. Expansions also occur because of environmental changes such as the global warming and the glacial cycles or due to sudden long-distance migrations to new habitats. Even though well-mixed-population models can account for the growing number of individuals (population size), these models do not capture the fact that the newly settled areas are colonized by the offspring of only a small number of individuals at the expanding front. Since the ancestral population is small, the genetic drift is strong. As a result, neutral genetic diversity decreases with the distance from the origin of the expansion. This reduction in genetic diversity, which is often called "the founder effect" (Mayr, 1942), has been observed in humans (Templeton, 2002; Ramachandran *et al.*, 2005) and many other species. For example, the founder effect in the population waves following the receding glaciers is believed to be responsible for the reduced genetic diversity in high latitude regions compared to equatorial ones (Hewitt, 1996).

The spreading of *Escherichia coli* (*E. coli*) and *Saccharomyces cerevisiae* (*S. cerevisiae*) on Petri dishes has been investigated in recent experiments by Hallatschek *et al.* (2007). In these experiments, microbes grown in the dark carried one of two selectively neutral alleles, differing only in a gene encoding for proteins with two distinct fluorescence spectra. Figure 1 shows the expansion of an initially well-mixed 50:50 population of *E. coli* into two unoccupied half planes initiated by a razor blade inoculation with cells grown up in liquid culture. The distinctive feature illustrated by the typical experiment in Fig. 1 is that the population does not remain well mixed, instead, it segregates into well-defined domains. The segregation occurs because

the strong genetic drift associated with reduced population size facilitates fixation of one of the two alleles at the front.

Analogous phenomena should also occur in a nonexpanding one-dimensional population because its dynamics is similar to the dynamics of the front of a growing population. The front of a population wave and a literally one-dimensional habitat are not exactly equivalent because the contour of the front undergoes undulations while a one-dimensional habitat has a fixed linear shape. Nevertheless, both are effectively one-dimensional and should deviate from the predictions of well-mixed-population models in similar ways. The advantage of a flat one-dimensional habitat is that it is easier to analyze. In addition, although most species live in effectively two-dimensional habitats, a quasi-one-dimensional habitat could describe a bank of a river, a seacoast, and a slope of a linear mountain range.

To study the dynamics of a population analytically, we adopt the stepping stone model proposed by Kimura and Weiss (1964). This model considers many well-mixed populations, demes, located on a spatial lattice. Each deme is subject to mutation, selection, genetic drift, and short-range migration between neighboring demes. In the limit of weak evolutionary forces and large number of demes, the stepping stone model is equivalent to the continuous models proposed by Wright (1943) and Malécot (1955) and is described by the stochastic Fisher-Kolmogorov-Petrovsky-Piscounov equation (Fisher, 1937; Kolmogorov *et al.*, 1937) with additional terms representing mutation. On the other hand, when each deme contains only one organism, the model is analogous to the Eden model (Saito and Müller-Krumbhaar, 1995) used to describe the growth of interfaces and the voter model (Cox and Griffeath, 1986) discussed in Appendix F.

We also performed numerical simulations to better understand the relationship between the experiments in Hallatschek *et al.* (2007) and our analytical results. An illustrative simulation (with periodic boundary conditions) is shown in Fig. 2, which also shows the difference between a growing population front with undulations and a literally one-dimensional habitat advancing uniformly in time. Figure 3 shows qualitative agreement between the experiments and the simplified row-by-row growth model that we studied analytically.

In this paper, we first focus on the spatial segregation due to genetic drift and its effect on the dynamics of a linear one-dimensional population. We find that segregation of two neutral alleles has two stages. During the first stage, distinguishable domains emerge from the well-mixed population. During the second stage, domain boundaries diffuse and annihilate upon collision. As a result, some of the domains vanish whereas others grow. We show how our calculations might be used to extract the diffusion constant and the effective population size from experiments like those in Hallatschek *et al.* (2007) and discuss how well the model describes the behavior of microbes. A *detailed* comparison (beyond the qualitative agreement we find with the main features) would require more extensive and precise experiments; we hope such experiments will be carried out in the future. The spatial segregation dramatically changes the effects of genetic drift and selection on the population compared to the predictions of well-mixed-population models. For the neutral model without mutation, we find that local diversity or "heterozygosity" decays as $t^{-1/2}$, and the standard deviation of the global fraction of an allele grows subdiffusively as $t^{1/4}$. The evolutionary dynamics during a radial expansion (see Fig. 14) is studied as well. In this case, migration and genetic drift slowly weaken as the circumference grows. As a result, the domain boundaries eventually stop coalescing leading to a finite number of domains in the long-time limit. We find that this final number of domains grows as a square root of the initial radius of the colony. We also study the dynamics in the presence of weak selection and find that it differs markedly from that of a well-mixed population. Because of the spatial segregation into domains, selection acts only near domain boundaries, which constitute only a small fraction of the population. Hence, extinction of a

deleterious allele proceeds much more slowly in one-dimensional populations than in well-mixed populations. Unlike genetic drift and selection, the effects of mutation in the spatial model are essentially the same as in the well-mixed-population model, but the spatial model gives a more accurate description of the population and accounts for the spatial correlations. Finally, we discuss how one can estimate important model parameters by sampling and sequencing DNA from organisms in a natural population. The differences between spatial and nonspatial models used for genetic inference are highlighted.

A substantial fraction of our results for the neutral dynamics in a one-dimensional habitat has been derived previously in population genetics (Kimura and Weiss, 1964; Malécot, 1975; Barton *et al.*, 2002), ecology (Houchmandzadeh and Vallade, 2003), and nonequilibrium statistical mechanics (Cox and Griffeath, 1986; Bramson and Lebowitz, 1991). Here we present a single self-contained derivation of these earlier results in a novel context of expanding populations in two dimensions and in a language familiar to physicists, with future microbial tests of the theory in mind. Our new results are primarily confined to the analysis of radial expansions and natural selection.

This paper is organized as follows. First, we review classical results for well-mixed populations in Sec. II. We then introduce the one-dimensional stepping stone model in Sec. III and derive the equations of motion for spatial correlation functions. In Secs. IV and V we solve these equations for zero and nonzero mutation rates, respectively. While the neutral stepping stone model has been treated before, we derive some new results and use a different technique that can be easily extended to radially expanding populations. The effects of selection are considered in Sec. VI, and in Sec. VII we test our analytical results with simulations. In Sec. VIII, evolutionary dynamics during a radial range expansion is analyzed, and Sec. IX deals with genetic inference. Various details are relegated to Appendixes A–F. In Appendix E, we indicate how some of the two-state (i.e., "two-allele") results can be generalized for the Potts-model-like nonequilibrium dynamics of q-alleles with $q \geq 3$.

II. POPULATION GENETICS IN WELL-MIXED POPULATIONS

Well-mixed-population models are relevant to microorganisms vigorously shaken in a test tube, but they do not describe spatial phenomena. Indeed, if cells visit all parts of the test tube during a cell division time, they live in an effectively zero-dimensional habitat. Nevertheless, well-mixed-population models can serve as a useful reference point to which spatial models can be compared. Nonspatial populations also provide a simple context to introduce genetic drift, mutation, and selection, and the stepping stone model presented in Sec. III uses a well-mixed-population model to describe the dynamics of allele frequencies within the demes. This section summarizes the classical results of nonspatial population genetics, which are primarily due to Wright, Fisher, Haldane, and Kimura; Crow and Kimura (1970) and Hartl and Clark (1989) provided a good introduction to the subject and referred to the original literature, which is too extensive to be discussed here [see also Blythe and McKane (2007) for a recent review written for physicists].

To simplify the discussion and to make a direct connection with the experiments in Hallatschek *et al.* (2007), we consider two alleles in a population of N haploid organisms, i.e., organisms with a single set of chromosomes.[1] The two-allele approximation may seem very restrictive, but many of our results can be generalized to an arbitrary integer number of $q \geq 3$ alleles. In addition, a two-allele model can be used to describe the dynamics of an allele of interest (with

---

[1]The theory of haploid organisms also describes the dynamics of genes in cellular organelles such as mitochondria and chloroplast and on certain sex chromosomes such as Y chromosome in *Homo sapiens*. For N diploid organisms, the theory is essentially the same under certain assumptions, provided one focuses on the dynamics of $2N$ gene copies in each generation [see Hartl and Clark (1989)].

or without a selective advantage) when all other alleles have the same fitness. We assume that each of the individuals in the population can die, give birth (divide), and mutate. The details of this birth and death process are species dependent, but the dynamics on time scales larger than the generation time τ_g is believed to be universal provided N is large. This universal dynamics is often referred to as the diffusion or continuous approximation. Two simple models are commonly used to illustrate the continuous approximation: the Wright-Fisher model and the Moran model. Here we use the latter because it more closely resembles microbes with overlapping generations.

First, we consider the Moran model without selection and mutation. During a time step, two individuals are randomly selected with replacement from the population. The first individual is chosen to reproduce and the second one to die; thus, the total number of the organisms is conserved. If the "frequency" of allele one (i.e., the fractional number of individuals with genotype one) at time step $\tilde{t}$ is $f(\tilde{t})$, then at the next time step it is $f+1/N$ with probability $f(1-f)$, $f-1/N$ with probability $f(1-f)$, and f with probability $f^2+(1-f)^2$. The expectation value and variance of $f(t+1)$ are then given by

$$\langle f(\tilde{t}+1)\rangle = f(\tilde{t}), \tag{1}$$

$$\left\langle [f(\tilde{t}+1) - \langle f(\tilde{t}+1)\rangle]^2 \right\rangle = \frac{2f(\tilde{t})[1 - f(\tilde{t})]}{N^2}, \tag{2}$$

where angular brackets represent average with respect to the random choice of individuals for reproduction and death. Because only one of N organisms gives birth in a Moran time step, $\tilde{t}$ measures time in fractional generation time, τ_g/N.

Equations (1) and (2) imply that $f(\tilde{t})$ performs an unbiased random walk in the space of allele frequencies. In the continuum limit, this random walk can be described by the following Fokker-Planck equation with a frequency-dependent diffusion coefficient (Crow and Kimura, 1970; Hartl and Clark, 1989):

$$\frac{\partial P(t,f)}{\partial t} = \frac{\mathfrak{D}_g}{2}\frac{\partial^2}{\partial f^2}[f(1-f)P(t,f)], \tag{3}$$

where $P(t,f)$ is the probability density function for f at time t measured in generations and $\mathfrak{D}_g$ is the genetic diffusion constant. Here t is the time measured in generations; as discussed above, N Moran time steps constitute a generation time τ_g. Thus, in the Moran model, we have

$$\mathfrak{D}_g = \frac{2}{N\tau_g}. \tag{4}$$

Alternative reproduction schemes, such as the Wright-Fisher sampling (Crow and Kimura, 1970; Hartl and Clark, 1989), lead to an equation identical to Eq. (3) but with a different numerical coefficient in Eq. (4).

Equation (3) is subject to absorbing boundary conditions[2] at $f=0$ and 1 because if one of the alleles is lost, it cannot appear again in the absence of mutation. Therefore the population eventually becomes fixed at one of the absorbing states. We calculate the rate of the fixation by considering the average heterozygosity of the population,

$$H(t) \equiv \langle h(t) \rangle = \langle 2f(t)[1 - f(t)] \rangle,$$

(5)

which is the (averaged over realizations) probability that two randomly selected individuals have different alleles. When the population is close to the fixation ($f \approx 0$ or ≈ 1), the heterozygosity is close to zero. The equations of motion for $F(t) \equiv \langle f(t) \rangle$ and $H(t)$ follow from Eq. (3) by multiplying both sides with f or h, integrating over f, and eliminating the derivatives with respect to f via integration by parts. The results are

$$\frac{dF(t)}{dt} = 0,$$

(6)

$$\frac{dH(t)}{dt} = -\mathfrak{D}_g H(t).$$

(7)

Equations (6) and (7) imply that, while the average frequencies of these neutral alleles do not change $F = \langle f \rangle = f(t=0) \equiv F_0$, the population reaches fixation exponentially fast, $H(t) = H(0) e^{-\mathfrak{D}_g t} = F_0(1-F_0)e^{-\mathfrak{D}_g t}$.

The average heterozygosity is closely related to the variance of $f(t)$, the fraction of the first allele,

$$V(t) = \left\langle [f(t) - \langle f(t) \rangle]^2 \right\rangle = F(t)[1 - F(t)] - \frac{1}{2}H(t).$$

(8)

Thus, even if a population starts with zero variance, the fluctuations grow until the variance reaches its maximum value of $F_0(1-F_0)$, which corresponds to a population fixed to allele one with probability F_0 and to allele two with probability $1-F_0$. Note that, for small t, $V(t)$ grows linearly with time, but, at large times, the variance approaches its limiting value exponentially fast. The linear growth of variance at small times also follows from the Fokker-Planck equation because, at small times, Eq. (3) can be approximated by a diffusion equation with a constant diffusivity.

Next, we generalize Eq. (3) to account for mutations. In the Moran model, mutation is included at the end of a time step by allowing the offspring to mutate with probability $\tilde{\mu}_{12}$ from allele one to allele two and with probability $\tilde{\mu}_{21}$ from allele two to allele one. If the frequency of allele one at time step t is $f(t)$, then, at the next time step, the expectation value of $f(t+1)$ is given by

$$\langle f(\tilde{t}+1) \rangle = f(\tilde{t}) + \frac{\tilde{\mu}_{21}[1 - f(\tilde{t})] - \tilde{\mu}_{12}f(\tilde{t})}{N},$$

(9)

---

[2]Since Eq. (3) is singular at the boundaries, we require $\lim_{f \to 0,1} f(1-f)P(t,f)=0$. See Kimura (1955) and Risken (1989) for a more detailed discussion.

and the variance of $f(\tilde{t}+1)$ is given by Eq. (2) to the leading order in the mutation rates and the inverse population size.

Since the expectation value of $\tilde{f}(t)$ changes with time, mutation leads to an f-dependent drift term in the Fokker-Planck equation. Upon recalling that N Moran time steps equal one generation time, we have

$$\frac{\partial P(t,f)}{\partial t} = -\frac{\partial}{\partial f}\{[\mu_{21}-(\mu_{12}+\mu_{21})f]P(t,f)\}+\frac{\mathfrak{D}_g}{2}\frac{\partial^2}{\partial f^2}[f(1-f)P(t,f)],$$

(10)

where $\mu_{12} \equiv \tilde{\mu}_{12}\tau_g^{-1}$ and $\mu_{21} \equiv \tilde{\mu}_{21}\tau_g^{-1}$ are the mutation rates per generation.

Because the alleles can mutate into each other, the probability flux through the boundaries must be zero, so Eq. (10) has reflecting boundary conditions, and a nontrivial stationary solution for $P(t,f)$ exists. While the stationary distribution can be obtained easily [see Eq. (20), Fig. 4, and Crow and Kimura (1970)], it is sufficient to analyze the moments $F(t)$ and $H(t)$ introduced above. This will also allow us to make a direct comparison with the corresponding solutions of the one-dimensional stepping stone model in Sec. V. The equations of motion for $F(t)$ and $H(t)$ are obtained from Eq. (10) in the same way as for the absence of mutation. The results are

$$\frac{dF(t)}{dt}=\mu_{21}-(\mu_{12}+\mu_{21})F(t),$$

(11)

$$\frac{dH(t)}{dt}=-(\mathfrak{D}_g+2\mu_{12}+2\mu_{21})H(t)+2[\mu_{21}+(\mu_{12}-\mu_{21})F(t)].$$

(12)

Since these equations are linear differential equations with constant coefficients, the equilibrium is approached exponentially fast. The stationary solutions, which are obtained in the limit $t\rightarrow\infty$, are given by

$$F(\infty)=\frac{\mu_{21}}{\mu_{12}+\mu_{21}},$$

(13)

$$H(\infty)=\frac{2F(\infty)[1-F(\infty)]}{1+\frac{\mathfrak{D}_g}{2(\mu_{12}+\mu_{21})}}.$$

(14)

From Eqs. (14) and (8), we see that, when the population size is large enough, i.e., $\mathfrak{D}_g\ll(\mu_{12}+\mu_{21})$, $H(\infty)\approx2F(\infty)[1-F(\infty)]$, the stationary value of the heterozygosity is consistent with $f(t)\approx F(\infty)$. Thus $V(\infty)\approx0$, and the fluctuations of $f(t)$ are negligible. In the opposite limit, $H(\infty)\propto(\mu_{12}+\mu_{21})/\mathfrak{D}_g$ is significantly smaller, which suggests that most of the time the population is fixed to one of the alleles, and mutations lead to rare transitions between states with $f=0$ and 1. Consequently, the stationary distribution is dominated by the regions around $f=0$ and 1, as one can see in Fig. 4. Our interpretation of Eq. (14) is consistent with a more rigorous analytical and numerical analysis by Duty (2000).

Finally, we introduce the Darwinian natural selection, which is usually related to the difference in the reproduction or survival probability of the organisms. In the continuous time limit

considered here, both mechanisms of selection lead to the same dynamics; therefore, we only consider selection due to different growth rates. In the Moran model, a growth rate difference is embodied in modified probabilities of reproduction: the individual with allele one is chosen to reproduce not with probability f but with probability $w_1 f/[w_1 f + w_2(1-f)]$, where w_1 and w_2 are the fitnesses (i.e., growth rates) of alleles one and two, respectively. In the absence of mutations, this modification results in

$$\langle f(\tilde{t}+1) \rangle = f(\tilde{t}) + \frac{f(\tilde{t})[1 - f(\tilde{t})](w_1 - w_2)}{N\{w_1 f(\tilde{t}) + w_2[1 - f(\tilde{t})]\}}.$$

(15)

When selection is weak, that is $|w_1 - w_2| \ll w_1 + w_2$, Eq. (15) reduces to

$$\langle f(\tilde{t}+1) \rangle = f(\tilde{t}) + \frac{\tilde{s}}{N} f(\tilde{t})[1 - f(\tilde{t})],$$

(16)

where $\tilde{s} = 2(w_1 - w_2)/(w_1 + w_2)$ is the selective advantage of allele one, which has to be much smaller than 1 for the approximation to hold. When $\tilde{s} > 0$, allele one is advantageous; for $\tilde{s} < 0$, it is deleterious. In the following, we assume that allele one is advantageous because one can always relabel the alleles to satisfy this condition.

Similar to the case of mutations without selection, the variance of $f(\tilde{t}+1)$ is given by Eq. (2) to the leading order in $\tilde{s}$ and N^{-1}, and the corresponding Fokker-Planck equation acquires an f-dependent drift term due to selection,

$$\frac{\partial P(t,f)}{\partial t} = - s \frac{\partial}{\partial f}[f(1 - f)P(t,f)] + \frac{\mathfrak{D}_g}{2} \frac{\partial^2}{\partial f^2}[f(1 - f)P(t,f)],$$

(17)

where $s = \tilde{s}\tau_g^{-1}$. The equations for F and H are not as useful as before because the hierarchy of the moment equations does not close. Nevertheless, Eq. (17) can be easily analyzed in two limits. When the population size is large ($\mathfrak{D}_g \ll s$), fluctuations are not important, and $[dF(t)]/dt \approx sF(t)[1 - F(t)]$. Upon setting $F_0 \equiv F(0)$, we have

$$F(t) \approx \frac{1}{1 + [(1 - F_0)/F_0]e^{-st}},$$

(18)

so the selective sweep is exponentially fast. When the fluctuations dominate the dynamics, the selection slightly increases the odds of fixation of the advantageous allele but does not significantly affect the rate of fixation. For a detailed analysis of Eq. (17), see Crow and Kimura (1970).

In the continuous limit, the population genetics of a well-mixed effectively zero-dimensional population with genetic drift, selection, and mutation is summarized by the following Fokker-Planck (or forward Kolmogorov) equation:

$$\frac{\partial P(t,f)}{\partial t} = -s \frac{\partial}{\partial f}[f(1-f)P(t,f)] - \frac{\partial}{\partial f}\{[\mu_{21} - (\mu_{12} + \mu_{21})f] \times P(t,f)\} + \frac{\mathfrak{D}_g}{2} \frac{\partial^2}{\partial f^2}[f(1-f)P(t,f)].$$

(19)

The stationary distribution for Eq. (19) is reached exponentially fast and takes the following form (Crow and Kimura, 1970; Duty, 2000):

$$P(\infty, f) = Ce^{2sf/\mathfrak{D}_g} f^{2\mu_{21}/D_g - 1} (1 - f)^{2\mu_{12}/D_g - 1}, \tag{20}$$

where C is the normalization constant chosen to set $\int_0^1 P(\infty, f) df = 1$. This stationary distribution is plotted in Fig. 4 for both strong and weak genetic drift.

Although the formulation in terms of a Fokker-Plank equation is appropriate for nonspatial models, an alternative formulation via a stochastic differential equation can be generalized to spatial models more easily. Equation (19) is equivalent to

$$\frac{df(t)}{dt} = sf(t)[1 - f(t)] + \mu_{21} - (\mu_{12} + \mu_{21})f(t) + \sqrt{\mathfrak{D}_g f(t)[1 - f(t)]}\,\Gamma(t)\ (\text{Itô}), \tag{21}$$

$$\langle \Gamma(t_1)\Gamma(t_2)\rangle = \delta(t_1 - t_2), \tag{22}$$

where $\Gamma(t)$ is a white zero mean Gaussian noise and $\delta(t)$ is Dirac's delta function; to get the correct Fokker-Planck equation (19), one must use Itô's prescription to define how Eq. (21) steps the dynamics forward in time. This interpretation of the noise term ensures that $f(t)$ depends only on $\Gamma(t')$ with $t' < t$ as it is appropriate for population genetics. Itô's prescription is adopted throughout the paper and an introduction to the Itô calculus is given in Appendix A [see also Gardiner (1985), Risken (1989), and Duty (2000)]. In Sec. III, we use Eq. (21) to formulate the stepping stone model in one dimension.

Well-mixed-population models do not describe migration and subdivision of natural populations (Hartl and Clark, 1989). To remedy this deficiency, two common approaches exist: to assume a uniformly populated spatial habitat with free diffusion or to assume a patchy habitat with a prescribed pattern of limited migration between the patches. The former is the subject of this paper and can be regarded as the continuum limit of the stepping stone model (Kimura and Weiss, 1964); see Sec. III. The simplest variant of the latter approach is known as the island model (Wright, 1931). The island model assumes that all patches or islands have the same number of organisms and populations in every patch obey well-mixed-population dynamics. The migration occurs between any two patches with equal probability, so, in some sense, this is a mean-field or infinite-dimensional model. The island model successfully predicts that the organisms are more likely to be related locally than globally, but most of its predictions are similar to those of well-mixed-population models because the migration does not account for spatial structure. In the limit of an infinitely large number of islands, the effect of migration in and out of any patch is equivalent to an effective mutation rate; however, this is not the case in a one-dimensional model considered below.

III. ONE-DIMENSIONAL STEPPING STONE MODEL

In Sec. II, we formulated a model to describe genetic drift, mutation, and selection in an effectively zero-dimensional habitat. For a one-dimensional population considered in this section, we extend the model to account for short-range migrations during every generation. Migration is usually modeled either as exchange of individuals between neighboring island populations (demes) (Wright, 1931; Kimura and Weiss, 1964) or as dispersal of offspring or adults within a continuous population (Wright, 1943; Nagylaki, 1974; Malécot, 1975).

Although the first approach was developed to model patchy populations, it can be used to describe continuous populations if the deme sizes are much smaller than the whole population, and spatial variations are gradual. In this limit, both migration models should give essentially the same results. Here we adopt the first approach because it is conceptually simpler.

To specify the one-dimensional stepping stone model, we consider an infinite set of demes arranged on a line. Neighboring demes are separated by distance a and indexed by an integer $l=-\infty,\ldots,-1,0,1,\ldots,\infty$. Each deme has N organisms (but the total population size is infinite), and the frequency of allele one in deme l is $f_l(t)$. Migration occurs only between nearest neighbors, and, every generation, a deme exchanges $\tilde{m}N/2$ individuals with its right neighbor and $\tilde{m}N/2$ individuals with its left neighbor. We assume that the exchange fraction $\tilde{m}$ is much smaller than 1 and that the individuals of both allelic types are equally likely to be exchanged. Thus, in one generation, $\langle f_l \rangle$ changes by $\tilde{m}(f_{l-1}+f_{l+1}-2f_l)/2$ due to migration. The variance of f_l grows due to randomness in the exchange process, but this increase is negligible compared to the genetic drift within an island. In the continuous time limit, $f_l(t)$ obeys the following generalization of Eq. (21):

$$\frac{df_l}{dt}=\frac{m}{2}(f_{l-1}+f_{l+1}-2f_l)+sf_l(1-f_l)+\mu_{21}-(\mu_{12}+\mu_{21})f_l+\sqrt{\mathfrak{D}_g f_l(1-f_l)}\Gamma_l(\text{Itō}),$$

(23)

$$\langle\Gamma_{l_1}(t_1)\Gamma_{l_2}(t_2)\rangle=\delta_{l_1 l_2}\delta(t_1-t_2),$$

(24)

where $m=\tilde{m}\tau_g^{-1}$ and $\delta_{l_1 l_2}$ is Kronecker's delta. We can also write Eq. (23) in the continuous space limit by introducing a spatial coordinate $x=la$,

$$\frac{\partial f}{\partial t}=D_s\frac{\partial^2 f}{\partial x^2}+sf(1-f)+\mu_{21}-(\mu_{12}+\mu_{21})f+\sqrt{D_g f(1-f)}\Gamma\ (\text{Itō}),$$

(25)

$$\langle\Gamma(t_1,x_1)\Gamma(t_2,x_2)\rangle=\delta(t_1-t_2)\delta(x_1-x_2),$$

(26)

where the spatial and genetic diffusion constants are $D_s=ma^2/2$ and $D_g=a\mathfrak{D}_g=2a/\tau_g N$, respectively. Thus the continuous time and space limit of the stepping stone model is described by the stochastic Fisher-Kolmogorov-Petrovsky-Piscounov equation (Fisher, 1937;Kolmogorov *et al.*, 1937) with additional terms describing mutation.

Similar to the analysis of the well-mixed-population model discussed in Sec. II, we use equal-time-correlation functions of $f(t,x)$ to characterize the dynamics of the stepping stone model. The spatial versions of the average frequency and heterozygosity are defined as follows:

$$F(t,x)=\langle f(t,x)\rangle,$$

(27)

$$H(t,x_1,x_2)=\langle f(t,x_1)[1-f(t,x_2)]\rangle+\langle f(t,x_2)[1-f(t,x_1)]\rangle.$$

(28)

The equation of motion for $F(t,x)$ depends on $H(t,x,x)$ and is readily derived by averaging Eq. (25), which gives

$$\frac{\partial F}{\partial t} = D_s \frac{\partial^2 F}{\partial x^2} + \mu_{21} - (\mu_{12} + \mu_{21})F + \frac{s}{2}H(t, x, x). \tag{29}$$

The dynamics of $H(t, x_1, x_2)$ is obtained by differentiating Eq. (28) with respect to t and then eliminating $\partial f / \partial t$ with the help of Eq. (25). Note that Itô's formula (see Appendix A) must be used to differentiate Eq. (28) correctly. The result is

$$\frac{\partial}{\partial t} H(t, x_1, x_2) = D_s \left(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} \right) H(t,$$
$$x_1, x_2) - D_g H(t,$$
$$x_1, x_2) \times \delta(x_1$$
$$- x_2) - 2(\mu_{12}$$
$$+ \mu_{21})H(t, x_1, x_2) + 2\mu_{21} + (\mu_{12} - \mu_{21})[F(t,$$
$$x_1) + F(t,$$
$$x_2)] + \frac{s}{2}[H(t,$$
$$x_1, x_1) + H(t,$$
$$x_2, x_2)]$$
$$- s \langle 2f(t, x_1)[1 - f(t, x_1)]f(t, x_2) \rangle$$
$$- s \langle 2f(t, x_2)[1 - f(t, x_2)]f(t, x_1) \rangle . \tag{30}$$

Equations (29) and (30) agree with the ones derived by Nagylaki (1978) in the limit of no mutations considered there.

From Eq. (30), one can see that the hierarchy of the moment equations does not close unless selection is absent. Similar to the well-mixed case, the correlation functions for neutral models with and without mutations can be found analytically (see Secs. V and IV) but different methods are required to analyze the dynamics in the presence of selection (see Sec. VI). To simplify the analysis, we consider well-mixed spatially homogeneous initial conditions. Then F is only a function of t and H is a function of t and $x = x_1 - x_2$. With these simplifying assumptions, the equations of motion for $F(t)$ and $H(t, x)$ take the following form:

$$\frac{dF(t)}{dt} = \mu_{21} - (\mu_{12} + \mu_{21})F(t) + \frac{s}{2}H(t, 0), \tag{31}$$

$$\frac{\partial}{\partial t} H(t, x) = 2D_s \frac{\partial^2}{\partial x^2} H(t,$$
$$x) - D_g H(t,$$
$$0)\delta(x)$$
$$- 2(\mu_{12} + \mu_{21})H(t,$$
$$x) + 2\mu_{21} + 2(\mu_{12}$$
$$- \mu_{21})F(t,$$
$$x) + sH(t,$$
$$0) - 2s \langle 2f(t, 0)[1 - f(t, 0)]f(t, x) \rangle . \tag{32}$$

IV. NEUTRAL MODEL WITHOUT MUTATIONS

We start the analysis of the one-dimensional stepping stone model by considering neutral alleles that do not mutate. In practice, this means $N^2\tilde{\mu}_{12}$, $N^2\tilde{\mu}_{21} \ll 1$ and $N^2\tilde{s} \ll 1$ (as we show below). Although these assumptions are not always realistic, they help to clarify the role of genetic drift in a spatial context. In addition, neglecting mutations is a good approximation on time scales shorter than the waiting times for the mutations μ_{12}^{-1} and μ_{21}^{-1}. Under these assumptions, F does not change, $F(t)=F_0$, and Eq. (32) reads

$$\frac{\partial}{\partial t}H(t,x) = 2D_s\frac{\partial^2}{\partial x^2}H(t,x) - D_g H(t,0)\delta(x).$$

(33)

Equation (33) can also be derived by tracing the ancestral lineages of organisms backward in time. The average spatial heterozygosity $H(t,x)$ is the average probability of sampling two different individuals chosen at time t from demes separated by distance x. As we trace the lineages of the two sampled organisms backward in time, the lineages diffuse in space due to migration and, when they are at the same point, they have a chance to coalesce in which case the sampled organisms must be identical because they have a common ancestor. Such a coalescence event changes the probability of being different from $H(t,0)$ to 0 and acts like a sink at $x=0$. The first term on the right-hand side of Eq. (33) describes the diffusion and the second term describes the coalescence. Since this argument is valid for an arbitrary number of alleles, Eq. (33) is valid for an arbitrary number of spatially diffusing neutral alleles. See Appendix E for a more detailed discussion of the q-allele problem, with $q \geq 3$.

To better understand the microbiology experiments on neutral alleles by Hallatschek *et al.* (2007), we consider uncorrelated initial conditions $F(0)=F_0$ and $H(0,x)=H_0$, where F_0 is the fraction of allele one and $H_0=2F_0(1-F_0)$, which is the heterozygosity of a well-mixed population with the frequency of allele one equal to F_0. For these initial conditions, Eq. (33) is solved in Appendix B. The results are

$$H(t,x) = H_0 - D_g\int_0^t dt' e^{-x^2/8D_s(t-t')}\frac{H(t',0)}{\sqrt{8\pi D_s(t-t')}},$$

(34)

$$H(t,0) = H_0\,\text{erfc}\left(\sqrt{\frac{D_g^2 t}{8D_s}}\right)e^{D_g^2 t/8D_s},$$

(35)

where erfc(y) is the complementary error function.

The spatial heterozygosity at vanishing separation $H(t,0)$ is particularly interesting because it indicates the degree of spatial segregation: if $H(t,0) \ll 1$, then, locally, the demes are fixed to one of the two alleles. From Eq. (35), one can see that, for $t \geq 8D_s/D_g^2$,

$$H(t,0) = H_0\left(\frac{\pi D_g^2 t}{8D_s}\right)^{-1/2} + O(t^{-3/2}),$$

(36)

which means that at long times one of the alleles reaches fixation locally. Therefore we see that the spatial model we are considering is consistent with the experiments by Hallatschek *et al.* (2007) (see Fig. 1) because it predicts the formation of domains (regions of local fixation). Thus, similar to the well-mixed model considered in Sec. II, one of the alleles reaches fixation locally with fixation time $\tau_f=8D_s/\pi D_g^2 \sim N^2$. But not only is this fixation time proportional to N^2, instead of N, the functional form of heterozygosity decay is different: instead of a rapid exponential decay, the spatial model shows a slow algebraic decay of local heterozygosity. These results agree with the previous works on population genetics of Nagylaki (1974) and Malécot (1975). Local fixation and $t^{-1/2}$ decay of local heterozygosity have also been found in the voter model (Cox and Griffeath, 1986), which corresponds to the stepping stone model with $N=1$ (see Appendix F).

The characteristic demixing time can also be estimated by the following scaling argument: the characteristic population size at time t in the coarsening process is $N_{ch}(t) \sim n_0 \sqrt{tD_s}$, where the population density $n_0 \sim N/a$. Upon recalling that the fixation time in zero dimensions is $\tau_f \sim N_{ch}(\tau_f)\tau_g$ and solving self-consistently for τ_f, we have $\tau_f \sim D_s N^2 \tau_g^2 / a^2 \sim D_s/D_g^2 \sim N^2 \tau_g$.

Another important characteristic of $H(t,x)$ is the length scale over which $H(t,x)$ changes from its minimum to its maximum values. Figure 5 plots Eq. (34) and shows the spatial variation of $H(t,x)$ at different times. One can see that the spatial heterozygosity is reduced near the origin due to the local fixation, but $H(t,x)$ rises to its initial value H_0 at large x, where the alleles remain uncorrelated. After the domains form, i.e., for $t \geq 8D_s/D_g^2$, this change from $H(t,0)$ to $H(0,x)$ happens on a length scale that is set by the average size of the domains ℓ, which is proportional to the diffusion length $\sqrt{2D_s t}$, as follows from Eq. (34). Since this characteristic length scale changes with time, it is convenient to rescale distances: $\bar{x}=x/\sqrt{2D_s t}$. Upon using Eq. (36) to simplify Eq. (34), we see that $H(t,x)$ approaches a nontrivial limit in terms of $\bar{x}$ as time goes to infinity,

$$H(t,\bar{x}) \xrightarrow{t\to\infty} H_0\left(1 - \frac{1}{\pi}\int_0^1 \frac{d\zeta}{\sqrt{\zeta(1-\zeta)}}e^{-|\bar{x}|^2/4\zeta}\right),$$

(37)

which agrees with the known results for the voter model (Cox and Griffeath, 1986).

A more precise evaluation of the domain density and hence an average domain size $\ell(t)$ can be obtained from $H(t,x)$, as shown in Appendix C. From Eq. (C4), we know that $\ell(t)=4D_s/[D_g H(t,0)]$, so using Eq. (36) we see that

$$\ell=\frac{\sqrt{2\pi D_s t}}{2f_0(1-f_0)},$$

(38)

which is consistent with the analysis of Hallatschek and Nelson (2010). Note that the genetic diffusion constant $D_g \sim 1/N$ drops out because, at large times, the only dynamics left is the diffusive motion of the domain walls. With neutral alleles, these boundaries behave as annihilating random walks, and the average domain size can be easily calculated (Hallatschek and Nelson, 2010).

Equations (35) and (38) suggest that the processes driven by the genetic drift slow down with time because the logarithmic time derivatives of $H(t,0)$ and ℓ tend to zero as time goes to infinity. In the annihilating random-walk picture of Hallatschek and Nelson (2010),

annihilations become rarer and rarer as the coarsening progresses. A more direct measure of genetic drift, which is also interesting from the biological point of view, is the fluctuations of the total fraction of, say, the first allele $\mathfrak{f}(t)$ in a finite population of length L. We define $\mathfrak{f}(t)$ as

$$\mathfrak{f}(t) = \frac{1}{L} \int_0^L f(t, x) dx \tag{39}$$

and compute its variance $v(t)$ to characterize its fluctuations.

Upon integrating Eq. (25) over x with $s=\mu_{12}=\mu_{21}=0$, we obtain the equation of motion for $\mathfrak{f}$,

$$\frac{d\mathfrak{f}}{dt} = \frac{1}{L} \int_0^L \sqrt{D_g f(t, x)[1 - f(t, x)]} \Gamma(t, x) dx, \tag{40}$$

where the spatial diffusion term vanishes after integration by parts provided periodic or Newman boundary conditions are imposed. Upon noting that $\langle \mathfrak{f} \rangle = F = \text{const}$ [the Itô interpretation of the noise $\Gamma(t,x)$ is crucial here] and defining

$$v = \langle \mathfrak{f}^2 \rangle - \langle \mathfrak{f} \rangle^2, \tag{41}$$

one finds immediately that $dv/dt = d\langle \mathfrak{f}^2 \rangle/dt$. To evaluate the time derivative, we use the rules of the Itô calculus, sketched in Appendix A, and find

$$\frac{dv(t)}{dt} = \frac{1}{L^2} \int_0^L \int_0^L \sqrt{D_g f(t, x_1)[1 - f(t, x_1)]} \times \sqrt{D_g f(t, x_2)[1 - f(t, x_2)]} \delta(x_1 - x_2) dx_1 dx_2, \tag{42}$$

where the delta function comes from averaging over the noise and using Eq. (26). From Eq. (42), it follows that

$$v(t) = \frac{D_g}{2L} \int_0^t H(t', 0) dt', \tag{43}$$

where we assume $v(0)=0$. Hence, we know $v(t)$ exactly because $H(t,0)$ is given by Eq. (35).

For small times $t \ll 8D_s/D_g^2$, the variance grows linearly with time. For large times, we can use the asymptotic expansion of $H(t,0)$ given by Eq. (36) to calculate $v(t)$. The result is

$$v(t) = \frac{H_0 \sqrt{8D_s t}}{\sqrt{\pi} L} + O\left(\frac{D_s}{D_g L} \right). \tag{44}$$

Equation (44) is consistent with Bramson and Lebowitz (1991), and we immediately conclude that the standard deviation $\Delta(t) = \sqrt{v(t)}$ grows as $t^{1/4}$ for large times. This important result is generalized for the flat-front and undulating-front models with q-alleles in Appendix E by approximating the dynamics of the domain boundaries by annihilating random walks. Thus, $\mathfrak{f}$ performs a subdiffusive random walk, and genetic drift of the global frequency $\mathfrak{f}(t)$ becomes weaker with time. Equation (44) is valid only for $t \ll L^2/D_s$ because it relies on Eq. (36), which is valid for an infinite population, and should break down at times that are long enough for a

domain boundary to diffuse from one end of the population to the other. Using Eqs. (8) and (43), one can also calculate the behavior of the global heterozygosity $\mathscr{H}(t)=L^{-1}\int_0^L H(t, x)dx$, i.e., the probability to sample two different alleles from the population regardless of their spatial locations,

$$2F_0(1 - F_0) - \mathscr{H}(t) \begin{aligned} &= \frac{D_g}{L}\int_0^t H(t', 0)dt' \\ &= \frac{H_0\sqrt{32D_s t}}{\sqrt{\pi}L} + O\left(\frac{D_s}{D_g L}\right), \end{aligned}$$

(45)

where the second equality requires $8D_s/D_g^2 \ll t \ll L^2/D_s$ for the reasons mentioned above. In the opposite limit $t \gtrsim L^2/D_s$, the global heterozygosity $\mathscr{H}(t,x)$ obeys zero-dimensional dynamics of a well-mixed population with an effective $\mathfrak{D}_g = D_g/L$ as shown by Nagylaki (1974).

The local heterozygosity and average domain size can be obtained from experiments on microbial spreading like the one shown in Fig. 1. If the data are sufficiently precise, Eqs. (36) and (38) could be used to extract D_s and D_g from the experiments. Since $D_g \sim 1/N$, extracting D_g from experimental data determines the effective deme size for the equivalent stepping stone model. D_s can be obtained from the diffusion of individual domain boundaries or $v(t)$. These two parameters completely determine the neutral dynamics without mutation and play an important role when selection or mutation is present.

V. NEUTRAL MODEL WITH MUTATIONS

While on short time scales mutation can be neglected, it is the long time scales and the patterns of genetic diversity created by mutations that are of particular interest in population genetics. Noticeable mutations also arise in microbiology experiments like those in Fig. 1, especially if mutation rates are enhanced by DNA damaging chemicals or radiation. In this section, we extend the results of Sec. IV by allowing for nonzero mutation rates between the two alleles. We assume, as before, statistically homogeneous initial conditions and note that the dynamics of the one- and two-point correlation functions is then given by

$$\frac{dF(t)}{dt} = \mu_{21} - (\mu_{12} + \mu_{21})F(t),$$

(46)

$$\frac{\partial}{\partial t}H(t, x) = 2D_s\frac{\partial^2}{\partial x^2}H(t, x) - D_g H(t, x)\delta(x) - 2(\mu_{12} + \mu_{21})H(t, x) + 2\mu_{21} + 2(\mu_{12} - \mu_{21})F(t),$$

(47)

where $F(t) \equiv \langle f(t,x)\rangle$ is independent of x. The equation of motion for F in the spatial model is exactly the same as Eq. (11), which describes the well-mixed-population model. Therefore F relaxes to its equilibrium value $F(\infty) = \mu_{21}/(\mu_{12} + \mu_{21})$ [see Eq. (13)], exponentially fast with time constant $(\mu_{12} + \mu_{21})^{-1} \geq \tau_g$. The similarity to the nonspatial model is not surprising because neutral mutations occur equally likely at any point within the population regardless of its spatial structure. The dynamics of $H(t,x)$ is, however, more complicated because both mutation and genetic drift determine the behavior of the spatial heterozygosity.

The stationary solution of Eq. (47) reads

$$H(\infty, x) = \frac{2\mu_{12}\mu_{21}}{(\mu_{12}+\mu_{21})^2}\left(1 - \frac{e^{-\sqrt{[(\mu_{12}+\mu_{21})/D_s]}|x|}}{1+4\sqrt{D_s(\mu_{12}+\mu_{21})/D_g^2}}\right).$$

(48)

Equation (48) agrees with the solution of Kimura and Weiss (1964), which was obtained in the discrete space and time limit. One can see that, for $x \geq \sqrt{D_s/(\mu_{12}+\mu_{21})}$, the spatial heterozygosity approaches $2F(\infty)[1-F(\infty)]$. Thus mutations cause the frequencies of allele one to eventually become uncorrelated at large separations. At shorter distances, however, there are correlations, and $H(\infty,x) < H(\infty,\infty)$ for all $x < \infty$. Note, in particular, that

$$H(\infty, 0) = \frac{2F(\infty)[1-F(\infty)]}{1+\frac{1}{4}D_g/\sqrt{D_s(\mu_{12}+\mu_{21})}} < H(\infty, \infty).$$

(49)

Note also that, for small mutation rates, the heterozygosity is proportional to $\mu_{12}+\mu_{21}$ [see Eq. (14)] in a well-mixed population, but the local heterozygosity in a one-dimensional population is proportional to $\sqrt{\mu_{12}+\mu_{21}}$ whenever $\tau_g(\mu_{12}+\mu_{21}) \ll a^2/N^2 D_s\tau_g$, which is a consequence of weaker genetic drift in one dimension.[3]

When $H(\infty,0) \ll H(\infty,\infty)$, the population is segregated into domains of different allelic types. Upon invoking Eq. (C4), we obtain the following average domain size:

$$\ell = \frac{2D_s(\mu_{12}+\mu_{21})^2}{D_g\mu_{12}\mu_{21}}\left(1+\frac{1}{4}\frac{D_g}{\sqrt{D_s(\mu_{12}+\mu_{21})}}\right) \approx \frac{D_s^{1/2}(\mu_{12}+\mu_{21})^{3/2}}{2\mu_{12}\mu_{21}}.$$

(50)

This result, together with Eq. (13), can be used to extract the mutation rates from experimental data.

We can also determine how fast $H(t,x)$ reaches its stationary value. Since the heterozygosity cannot be in equilibrium unless the frequency of the alleles has equilibrated, we assume, for simplicity, that $F(0)$ equals its stationary value. Then the deviation of the spatial heterozygosity from its long-time equilibrium value $\tilde{H}(t,x) = H(t,x) - H(\infty,x)$ obeys the following equation:

$$\frac{\partial}{\partial t}\tilde{H} = 2D_s\frac{\partial^2}{\partial x^2}\tilde{H} - D_g\tilde{H}\delta(x) - 2(\mu_{12}+\mu_{21})\tilde{H}.$$

(51)

Equation (51) can be further simplified by the change of variables $\tilde{H} = e^{-2(\mu_{12}+\mu_{21})t}\widehat{H}$, which leads to

$$\frac{\partial}{\partial t}\widehat{H} = 2D_s\frac{\partial^2}{\partial x^2}\widehat{H} - D_g\widehat{H}\delta(x).$$

(52)

---

[3]In population genetics, population structure and spatial correlations are often reported via $F_{st} = H(\infty,0)/H(\infty,\infty)$, which can readily be obtained from Eq. (49).

Since Eq. (52) is identical to Eq. (33), we conclude that, at long times, the difference between

$H(t,x)$ and the stationary solution decays as $\widehat{C}\sqrt{D_s/[D_g^2 t]}e^{-2(\mu_{12}+\mu_{21})t}$, where $\hat{C}$ is a constant. Thus, apart from an algebraic prefactor (and a nontrivial spatial dependence), the dynamics of $H(t,x)$ is essentially the same as in the well-mixed case.

In this section, we considered a model with only two alleles; however, in many circumstances, an infinite allele model is more appropriate. The infinite allele model is discussed in Appendix D. Some results for q alleles, $2<q<\infty$, are discussed in Appendix E.

VI. SELECTION

Unlike the neutral models with spatial diffusion and mutation discussed above, the one-dimensional stepping stone model with selection is difficult to treat analytically because the hierarchy of moment equations does not close. We examined three closure schemes,

$$\langle 2f(t,x_1)[1-f(t,x_1)]f(t,x_2)\rangle \approx H(t,0)F(t),\tag{53}$$

$$\langle 2f(t,x_1)[1-f(t,x_1)]f(t,x_2)\rangle \approx 2F(t)[1-F(t)][1-2F(t)]-[1-2F(t)]H(t,x)+H(t,0)F(t),\tag{54}$$

and

$$\langle f(t,x_1)f(t,x_2)f(t,x_3)\rangle \approx \frac{\langle f(t,x_1)f(t,x_2)\rangle\langle f(t,x_2)f(t,x_3)\rangle}{F},\text{ assuming } x_1\le x_2\le x_3.\tag{55}$$

The first scheme is a simple factorization approximation; the second scheme, which assumes small fluctuations, is due to Nagylaki (1978); and the third scheme, which provides a good approximation for some diffusion-limited reactions, was proposed by Lin (1991). Unfortunately, none of the schemes describe the behavior of the system correctly. The progress can be made, however, for some initial conditions in two limiting cases of strong selection $sD_s/D_g^2\ge 1$ and weak selection $sD_s/D_g^2\ll 1$. Note that we now use the term weak selection in a different sense than in Sec. II. For the rest of this section, we include spatial diffusion and genetic drift but neglect mutations, which are justified on short time scales.

First, consider the initial condition $f(0,x)=1-\theta(x)$, where $\theta(x)$ is the Heaviside step function. This initial condition specifies just one domain boundary, which, for any positive s, undergoes Brownian motion with a drift to the right. This is a good description of an expansion of a new advantageous mutant spreading through the population. In the strong selection limit $(s\ge D_g^2/D_s)$, Fisher (1937) found that the sharp boundary above broadens to a width of order $\sqrt{D_s/s}$, and the velocity of the genetic wave is given by

$$\upsilon_s=2\sqrt{sD_s}.\tag{56}$$

When, in contrast, selection is weak compared to genetic drift, it was recently found that the velocity is given by Doering *et al.* (2003) and Hallatschek and Korolev (2009),

$$v_w = \frac{2sD_s}{D_g}.$$

(57)

When the population contains multiple domains, the domain walls bordering a favorable genetic variant ("allele one") expand to engulf the regions occupied by the more deleterious allele.

Another interesting initial condition is $f(0,x) = F_0 =$ const; i.e., the population is initially well mixed. This scenario, for example, describes the quasi-one-dimensional strip of pioneers advancing at the front of a two-dimensional population wave that originated from a well-mixed ancestral population and is propagating in the region where one of the alleles has higher fitness [see Hallatschek and Nelson (2010)].

If selection is strong enough, then allele one ("the preferred variant") takes over the population before spatial correlations have time to appear. To see this, note from Eq. (18) that allele one wins locally on the time scales of s^{-1}, but, from Eq. (35), the time for spatial correlations to appear is on the order of D_s/D_g^2, which is much larger than s^{-1} in the strong selection limit. Thus the behaviors of one-dimensional and well-mixed populations are similar when $s \geq D_g^2/D_s = a^2/\tau_g^2 N^2 D_s$.

In the limit of weak selection, however, spatial correlations appear before allele two is eliminated. Qualitatively, we can divide the selective sweep into two stages. During the first stage, the effects of selection are negligible and spatial segregation occurs as described in Sec. IV. During the second stage, the domains of allele two shrink at each end with wall velocity v_w given by Eq. (57), and the stochastic motion of the domain boundaries can be neglected. The crossover time between the stages occurs when the diffusive displacement of the walls is of the same order as their deterministic displacement, i.e., when $v_w t = \sqrt{2D_s t}$. Thus the crossover time τ^* is on the order of D_s/v_w^2, which can be expressed as $D_g^2/s^2 D_s$ with the help of Eq. (57). Then, from Eq. (38), the average domain size at the crossover ℓ^* is on the order of D_g/sH_0.

The dynamics during the second stage depends on the probability distribution of domains of size η, $P_d(\eta)$ at time τ^*. For annihilating random walkers, which are a good approximation to domain boundaries during the first stage, Bramson and Griffeath (1980) proved that $P_d(\eta)$ has exponential tail for large η of the form $e^{-\gamma' \eta/\ell^*}$, where γ' is a number of the order 1. Since each domain shrinks with velocity $2v_w$, the fraction of allele two can be expressed as

$$1 - F(t) \propto \int_0^\infty \eta e^{-\gamma(\eta + 2v_w t)/\ell^*} d\eta \propto e^{-\lambda s^2 D_s t/D_g^2},$$

(58)

where λ is a number of order 1. From Eq. (58) it follows that, as in the well-mixed case, the selective sweep is exponentially fast, but the time constant of this process is proportional to s^{-2} rather than s^{-1}.

The analysis leading to Eq. (58) can be generalized to an arbitrary initial probability distribution $P_d(\eta)$ provided the dynamics is dominated by selection and genetic drift. For example, if a population initially in equilibrium with respect to mutations and genetic drift (see Sec. V) is affected by an abrupt environmental change that makes allele one advantageous, then the shift to the new equilibrium occurs exponentially fast with a time constant proportional to

$(sD_s^{1/2}\mu_{12}\mu_{21})/[D_g(\mu_{12}+\mu_{21})^{3/2}]$, assuming $P_d(\eta)$ has exponential tail of the form $e^{-\gamma''\eta/\ell}$, where γ'' is a constant of order unity and ℓ is given by Eq. (50).

Finally, we address a slightly different but equally important question: What is the probability p_{surv} that a few copies of the advantageous allele survive and establish a growing domain? Doering *et al.* (2003) solved this problem exactly,

$$p_{\text{surv}}=1-\exp\left[-\frac{2s}{D_g}\int_{-\infty}^{+\infty}f(0,x)dx\right].$$

(59)

Surprisingly, the survival probability does not depend on the diffusion constant. For a small initial number of advantageous alleles, we can qualitatively explain this result in the limit of weak selection by the following argument. Initially, the dynamics is almost neutral, and the probability of survival within a small interval of length Δx is proportional to the relative fraction of the advantageous allele in this interval, $\int f(0,x)dx/\Delta x$, because every organism has approximately the same probability to reach fixation. Once a domain of size Δx is formed, its survival probability equals the probability that the two biased random walks performed by the domain boundaries never meet, which is $1-\exp(-\upsilon\Delta x/D_s)\approx\upsilon\Delta x/D_s$ for small Δx and s [see Redner (2001) and Hallatschek and Nelson (2010)]. Then, using Eq. (57), the survival probability is

$$(\upsilon\Delta x/D_s)\int f(0,x)dx/\Delta x=2s\int f(0,x)dx/D_g \approx 1-\exp\left[-2s\int f(0,x)dx/D_g\right].$$

Note that, even though p_{surv} does not depend on D_s, the expression for the survival in one dimension does not reduce to its analog in well-mixed populations (Crow and Kimura, 1970; Doering *et al.*, 2003),

$$p_{\text{surv}}=\frac{1-e^{-2sf(0)/\mathfrak{D}_g}}{1-e^{-2s/\mathfrak{D}_g}},$$

(60)

unless one assumes $s/\mathfrak{D}_g\gtrsim 1$.

We make two important observations based on the results of this section. First, the temporal dynamics of the one-dimensional stepping stone model with selection can depend strongly on the initial conditions. Second, the results in the weak selection limit are sometimes related to the results in the strong selection limit or in well-mixed-population models by a parameter substitution, e.g., $s \rightarrow s^2 D_s/D_g^2$, at least up to a numerical factor. The second observation suggests that, while data can be naively fitted to a well-mixed-population model, the fit in fact gives the "renormalized" value of s instead of the "bare" one.

Although the one-dimensional stepping stone model provides a reasonable approximation to neutral genetic demixing at a linear front of an expanding two-dimensional microbial colony (see Fig. 3), there are additional subtleties associated with the dimensional reduction from two to one dimensions when one of the alleles is more fit. Apart form the undulations of the front mentioned in the caption of Fig. 2, the front develops additional structure because favorable sectors bulge out ahead of their less fit neighbors. In the limit of very small genetic drift, there are kink singularities where favorable and unfavorable domains meet. Nevertheless, the basic picture of domain boundaries engulfing unfavorable sectors at a constant velocity is still valid; see Hallatschek and Nelson (2010) for further details.

VII. SIMULATIONS

In Secs. III–VI, we reviewed and extended the theoretical analysis of the one-dimensional stepping stone model. This model, while of great theoretical interest, relies on a restrictive set of assumptions including large deme sizes and slow diffusive migration. The recent experiments by Hallatschek *et al.* (2007), on the other hand, were carried out with bacterial fronts that were only a monolayer thick; therefore, demes consisted of only a few microbes. Moreover, depending on the micro-organism, nearby demes can exchange a significant fraction of cells in each generation. In this section, we discuss numerical simulations not subject to these restrictions and compare them with the theoretical predictions.

We simulate L organisms arranged on a line and labeled by an integer l, $l=1,2,\ldots,L$. Each organism can be of either allelic type 1 or allelic type 2. During even generations, the offspring at site l comes from an organism at either site $l-1$ or site l, whereas during odd generations it comes from either site l or $l+1$. The simulations embody the process illustrated in Fig. 2, laid out on a triangular lattice in space and time. Periodic boundary conditions are imposed at the left and right ends of the population. Let $12\rightarrow2$ refer to the event that the offspring has allelic type 2, while one of its possible parents has allelic type 1, the other has allelic type 2, etc. The transition probabilities, which depend on the states of the possible ancestors, are then given by

$$
\begin{aligned}
11 &\rightarrow 1 : 1 - \tilde{\mu}_{12}, \\
11 &\rightarrow 2 : \tilde{\mu}_{12}, \\
22 &\rightarrow 1 : \tilde{\mu}_{21}, \\
22 &\rightarrow 2 : 1 - \tilde{\mu}_{21}, \\
12 &\rightarrow 1 : \left(\tfrac{1}{2}+\tfrac{\tilde{s}}{4}\right)(1-\tilde{\mu}_{12}) + \left(\tfrac{1}{2}-\tfrac{\tilde{s}}{4}\right)\tilde{\mu}_{21}, \\
12 &\rightarrow 2 : \left(\tfrac{1}{2}+\tfrac{\tilde{s}}{4}\right)\tilde{\mu}_{12} + \left(\tfrac{1}{2}-\tfrac{\tilde{s}}{4}\right)(1-\tilde{\mu}_{21}), \\
21 &\rightarrow 1 : \left(\tfrac{1}{2}+\tfrac{\tilde{s}}{4}\right)(1-\tilde{\mu}_{12}) + \left(\tfrac{1}{2}-\tfrac{\tilde{s}}{4}\right)\tilde{\mu}_{21}, \\
21 &\rightarrow 2 : \left(\tfrac{1}{2}+\tfrac{\tilde{s}}{4}\right)\tilde{\mu}_{12} + \left(\tfrac{1}{2}-\tfrac{\tilde{s}}{4}\right)(1-\tilde{\mu}_{21}).
\end{aligned}
\tag{61}
$$

The event $12\rightarrow2$ can happen either if allele one was selected for reproduction (probability $1/2 +\tilde{s}/4$) and mutated (probability $\tilde{\mu}_{12}$) or if allele two was selected for reproduction (probability $1/2-\tilde{s}/4$) and did not mutate (probability $1-\tilde{\mu}_{21}$). Other transition probabilities are obtained analogously. Thus, the system we simulate is similar to the voter model (Cox and Griffeath, 1986), which equivalent to population genetics models with $N=1$ (see Appendix F). However, to make the calculation faster, we use discrete generations rather than exponentially distributed waiting times until reproduction. We found no significant differences in dynamics between the voter model and the model used here.

First, we simulate the neutral model without mutations. To illustrate the similarities and differences between the stepping stone model and the undulating-front model, we also simulate a linear population wave in a two-dimensional habitat; both models are displayed in Fig. 2. Our model with an undulating front is the same as in Saito and Müller-Krumbhaar (1995), but we use a triangular grid instead of a square one. Figures 6 and 7 show how the average number of domain boundaries decreases with time; the insets show the mean-square displacements of the boundaries as a function of time. The simulations confirm that the domains move diffusively for the one-dimensional model with a flat front and superdiffusively for the undulating-front (Eden) model, in agreement with Saito and Müller-Krumbhaar (1995). Therefore, the number of domain boundaries decays faster for the undulating-front model compared to the flat-front model, which, as we show below, most closely tracks the prediction of the one-dimensional stepping stone model.

A single run of a simulation is shown in Fig. 8, and the spatial heterozygosity averaged over many realizations is shown in Fig. 9. Figure 10 shows that $H(t,x)$ for large t is described well by the limiting shape of spatial heterozygosity given by Eq. (37).

To properly represent $H(t,0)$, we artificially define demes of a larger size by grouping M neighboring individuals into one deme. From a theoretical point of view, this procedure is similar to the formation of the Kadanoff block spins as in renormalization-group methods (Wilson and Kogut, 1974; Goldenfeld, 1992) whereas, from the point of view of population genetics, this procedure is similar to the methods of collecting data from a dispersed natural population. In field studies, scientists do not typically sample every single individual, instead, they often divide the habitat into patches and sample a representative number of individuals from those patches. To summarize, we keep the dynamics of the simulation exactly the same but define the spatial heterozygosity on the demes of size M rather than on single individuals (see Fig. 9). We found that the local heterozygosity has the form $H(t,0)=\beta(M)t^{-1/2}$, as predicted by our analysis of the stepping stone model, for all M studied ($1 \leq M \leq 64$). From Eq. (36), we expect that $\beta \propto M^{-2}$; this expectation is also confirmed by our simulations.

As discussed in Sec. IV, the *total* fraction of allele one $\mathfrak{f}(t)$ fluctuates in an unusual way with time. Figure 11(a) shows examples of these remarkable variable-step-length random walks. The fluctuations of $\mathfrak{f}(t)$ obey Eq. (44) and grow subdiffusively, as shown in Fig. 11(b). We also find good agreement between the theory and the simulations in the presence of mutation for all values of M studied. The stationary average heterozygosity for $M=1$ is shown in Fig. 12. Finally, we studied selective sweeps in an initially well-mixed population. Figure 13(a) confirms the prediction from Eq. (58) that $F \propto 1-e^{-\alpha t}$ and Fig. 13(b) confirms the result of Sec. VI that, for strong genetic drift, the effective extinction rate α is proportional to s^2. Numerical results for *three* neutral alleles are presented in Appendix E.

VIII. INFLATION

Throughout this review we have focused on the evolutionary forces acting at a linear (flat or undulating) front, whose total length (averaged over the undulations) does not change in time. In this section, we explore the changes in the evolutionary dynamics caused by a constant increase of the total front length, for example, at the edge of an expanding circular colony (see Fig. 14). We now show that this increase, which we term "inflation" in an analogy with cosmology (Guth, 1981), slows down genetic drift and natural selection at the front.

Models of both linear and circular fronts are relevant to biology. Linear fronts describe the essential features of the dynamics when the effects of curvature and changes in the front length are negligible or when the spreading is limited by some geographical barriers, say, a receding glacier between two parallel rivers. If one focuses on genetic markers in *Homo sapiens* (e.g., in mitochondria DNA), dynamics of a linear front also resembles the abrupt settlement by pioneers via a "land run" in 1889 across the border of Oklahoma (Gibson, 1965). Circular fronts are more appropriate for modeling an initial colonization by a small number of pioneers arriving in the interior of a large spatially homogeneous habitat. Semicircular fronts are relevant to colonization after landing on a coast line. The circular scenario is often realized in microbiological experiments when a Petri dish is inoculated with micro-organisms. A radial range expansion of *E. coli* is shown in Fig. 14, which highlights the effects of genetic drift at the front.

The growth of a circular colony lengthens the front, thereby increasing characteristic local length scales. This inflation is specified by the dependence of the radius of the colony R on time t. Here we assume $R(t)=R_0+\mathrm{v}t$, which corresponds to a colony expanding with a constant velocity v from an initial radius R_0. The velocity of the expansion has been found constant in

the experiments by Hallatschek *et al.* (2007) and in the theoretical studies of the two-dimensional Fisher equation (Murray, 2003), provided the width of the front is much smaller than its length, a condition necessary for a one-dimensional model to hold.

To highlight the effects of inflation, we consider the simplest version of the one-dimensional stepping stone model without mutations and natural selection. In a circular geometry, it is convenient to use the angle $\varphi = x/R(t)$ instead of x to reference positions along the front. Then, the equation of motion for $H(t,x)$ takes a form analogous to Eq. (33),

$$\frac{\partial}{\partial t}H(t,\varphi) = \frac{2D_s}{(R_0 + \mathrm{v}t)^2}\frac{\partial^2}{\partial \varphi^2}H(t,\varphi) - \frac{D_g}{R_0 + \mathrm{v}t}H(t,0)\delta(\varphi),$$

(62)

where the factors of $R_0 + \mathrm{v}t$ have been introduced to account for the inflation. Like Eq. (33), Eq. (62) is valid for an arbitrary number of neutral alleles and can be understood by tracing two lineages backward in time. The time dependence of the coefficients in front of the diffusion and coalescence terms accounts for the fact that, as the colony grows, the same sizes in the φ space correspond to different sizes in the x space, where the diffusion and coalescence terms have their familiar time-independent form as in Eq. (33). When re-expressed in terms of t and $x = (R_0 + \mathrm{v}t)\varphi$, Eq. (62) contains an advection term describing the deterministic decrease of the separation between the lineages as they go back to the initial radius R_0.

Equation (62) is defined on a bounded domain $\varphi \in [-\pi, \pi]$ with periodic boundary conditions. Nevertheless, we can approximate the problem well by considering an unbounded domain $\varphi \in (-\infty, \infty)$, provided two diametrically opposite lineages are sufficiently unlikely to coalesce. From Eq. (62), we see that diffusion effectively stops after a characteristic time R_0/v, so our approximation of an unbounded domain should be valid if the distance traveled by the lineages during this time is small compared to the radius of the colony: $\sqrt{R_0 D_s/\mathrm{v}} \ll R_0$ or $D_s/\mathrm{v} \ll R_0$; this corresponds to a regime with many sectors as we show later. One can also test the goodness of the approximation by evaluating $H(0,\pi) - H(t,\pi)$, which is expected to be small if the approximation is valid.

To simplify the analysis, we make Eq. (62) dimensionless in terms of the new variables t and ϕ such that $t = tD_s/D_g^2$ and $\varphi = \phi D_s/D_g R_0$. The equation of motion for $H(t,\phi)$ then reads

$$\frac{\partial}{\partial t}H(t,\phi) = \frac{2}{(1+\sigma t)^2}\frac{\partial^2}{\partial \phi^2}H(t,\phi) - \frac{1}{1+\sigma t}H(t,0)\delta(\phi),$$

(63)

where the dimensionless parameter $\sigma = \mathrm{v}D_s/R_0 D_g^2$ is proportional to the ratio of two characteristic time scales in the problem: the local fixation time $\tau_f \sim D_s/D_g^2$ in the model of a linear front and the time in which the colony doubles its initial radius.

Upon assuming $\phi \in (-\infty, \infty)$, we obtain the exact solution of Eq. (63) for the initial condition $H(0,\varphi) = H_0$ by a generalization of the method presented in Appendix B,

$$H(t,\phi) = H_0 - \int_0^t dt' \frac{H(t',0)}{1+\sigma t'}\sqrt{\frac{(1+\sigma t)(1+\sigma t')}{8\pi(t-t')}} \times \exp\left[-\frac{\phi 2(1+\sigma t)(1+\sigma t')}{8(t-t')}\right],$$

(64)

where

$$H(\mathrm{t},0)=H_0-\frac{H_0\sqrt{1+\sigma\mathrm{t}}}{\sqrt{2\pi\sigma}}\left[\frac{\pi}{2}-\arcsin\left(\frac{1}{\sqrt{1+\sigma\mathrm{t}}}\right)\right]+H_0\sqrt{\frac{\pi(1+\sigma\mathrm{t})}{8\sigma}}\exp\left(\frac{1+\sigma\mathrm{t}}{8\sigma}\right)\times\left[\mathrm{erf}\left(\sqrt{\frac{1+\sigma\mathrm{t}}{8\sigma}}\right)-\mathrm{erf}\left(\frac{1}{\sqrt{8\sigma}}\right)\right]-H_0\sqrt{\frac{\pi(1+\sigma\mathrm{t})}{8\sigma}}(e^{\mathrm{t}/8}$$

$$-1)+H_0\sqrt{\frac{1+\sigma\mathrm{t}}{128\pi\sigma}}$$

$$\times\ e^{\mathrm{t}/8}\int_0^{\mathrm{t}}dt' e^{-t'/8}\arcsin\left(\frac{1}{\sqrt{1+\sigma t'}}\right).$$

(65)

The behavior of $H(\mathrm{t},\phi)$ is shown in Fig. 15. Similar to a linear front, the local heterozygosity $H(t,0)$ vanishes for large times, and the characteristic angular length scale over which $H(t,\varphi)$ changes from 0 to H_0 increases. Thus, Eq. (62) predicts the formation and growth of the domains shown in Fig. 14. However, there are two important differences that distinguish radial expansions from linear ones. First, $H(t,0)$ tends to zero as t^{-1} rather than as $t^{-1/2}$. Second, the curve $H(t,\varphi)$ approaches a nontrivial limit shape, unlike the case with a linear front, where the analogous curve widens indefinitely.

Following Hallatschek *et al.* (2007), we can qualitatively understand this behavior by noticing that the diffusion and lineage coalescence effectively stop after the characteristic time R_0/v. After this point, the number of domain boundaries and the angular width of the domains remain approximately constant. Therefore, $H(t,0)$, which is proportional to the fraction of the circumference occupied by the boundaries between the domains, should decay as t^{-1}, and the shape of $H(t,\varphi)$ should approach a nontrivial limit.

From the exact solution [Eqs. (64) and (65)], we compute the average angular size of the domains $\ell_\varphi(t)$ and the average number of the domains $\mathcal{N}(t)=2\pi/\ell_\varphi(t)$. Similar quantities were calculated by Hallatschek and Nelson (2010) in the approximation of random walking domain boundaries, which appropriate when $t>\tau_f$. Although we cannot use Eq. (C4) because of the inflation, Eq. (C2) remains valid and takes the following form:

$$\ell_\varphi(t)=\left(\frac{\partial H(t,+0)}{\partial\varphi}\right)^{-1}.$$

(66)

By integrating Eq. (62) over φ in the neighborhood of 0, we express $\partial H(t,+0)/\partial\varphi$ in terms of $H(t,0)$ and obtain

$$\ell_\varphi(t)=\frac{4D_s}{D_g(R_0+\mathrm{v}t)H(t,0)},$$

(67)

which approaches a constant at long times. This limit can be computed analytically, with the results

$$\ell_\phi(\infty) = \left(\frac{H_0 \mathrm{v}}{D_g} + H_0 \sqrt{\frac{R_0 \mathrm{v}}{2\pi D_s}} \right)^{-1},$$

(68)

$$\mathcal{N}(\infty) = \frac{2\pi H_0 \mathrm{v}}{D_g} + H_0 \sqrt{\frac{2\pi R_0 \mathrm{v}}{D_s}}.$$

(69)

Equation (69) implies two things. First, by measuring $\mathcal{N}(\infty)$ as a function of the initial homeland radius R_0, one can estimate both D_s and D_g for a microbial population, which could potentially be easier than the experiments with linear fronts that we proposed in Sec. IV. Second, if all individuals in the founding population are distinguishable ($H_0=1$), then each of the final sectors must originate from a single ancestor. Hence, $\mathcal{N}(\infty)$ for $H_0=1$ gives the average number of ancestors of the genetically segregated population at the periphery, which contains most of the organisms. This number is remarkably small. Figure 14, where $H_0=1/2$, has about 20 domains, so, since $\mathcal{N}(\infty) \propto H_0$, the segregated part of the population descended from only about 40 ancestors, a tiny fraction of about 20 000 founding cells. Although some of these cells are trapped in the interior of the homeland, a large number of them are piled in a ring at the edge of the homeland within minutes of inoculation as the carrier fluid dries out (Hallatschek *et al.*, 2007).

We can further quantify the amount of genetic drift in the population by the variance $v(t)$ of the total fraction of allele one $\mathfrak{f}(t)$. For simplicity, we assume a population with only two alleles. For several alleles, the global heterozygosity $\mathcal{H}(t)$ is more appropriate and can be easily obtained from our expressions for $v(t)$ because $\mathcal{H}(t) = H_0 - 2v(t)$. We compute $\mathcal{H}(t)$ and thus $v(t)$ by integrating Eq. (62) over φ; the result is

$$v(t) = \frac{D_g}{4\pi} \int_0^t \frac{H(t',0)}{R_0 + \mathrm{v}t} dt'.$$

(70)

We are mostly interested in the long-time limit $v(\infty)$, which is approached asymptotically as t^{-1}. This limit can be expressed as

$$v(\infty) = \frac{D_s H_0}{4\pi R_0 D_g} K(\sigma),$$

(71)

where $K(\sigma) = \int_0^\infty H^*(t',0)/(1+\sigma t') dt'$; $H^*(t,\phi)$ is the solution of Eq. (63) for $H_0=1$. The dependence of K on σ is shown in Fig. 16. In the limit of large D_g (approximated by the voter model; see Appendix F), an analytical expression for $v(\infty)$ is given by Eq. (F7).

Even though inflation slows down lineage diffusion and coalescence, genetic drift can still cause large fluctuations in the relative frequency of the alleles. These fluctuations are particularly important for any organism that undergoes spatial colonizations followed by almost complete extinctions (such life cycles are common both in nature and in the laboratory). For such organisms, $v(\infty)$ or $\mathcal{H}(\infty)$ characterizes the effective genetic drift, which is much larger than the one predicted by well-mixed-population models. We also note that the effects of genetic drift could be more pronounced at a circular front because natural selection is less efficient in the presence of inflation: domains of deleterious alleles persist longer because the

contraction of the domain due to the Darwinian selection must also be able to overcome its natural expansion due to inflation.

IX. GENETIC INFERENCE

So far, we have focused on forward-in-time dynamics while trying to calculate the patterns of genetic diversity from simple models of evolutionary dynamics. However, it is often necessary to reverse the question: Given the observed genetic diversity, how do we infer the recent history of the population and estimate important parameters such as the mutation rates and the effective population size? This question is particularly important because the current state of genetic diversity is often the only clue to the past. Fortunately, genetic inference can be very powerful because the differences among the genomes of individuals contain valuable information about the evolution of the population, and these differences can now be easily measured via DNA sequencing. For example, genetic inference has been used to determine the time and origin of the recent expansion of *Homo sapiens* (Ramachandran *et al.*, 2005) and to test whether *Homo sapiens* and *Homo neanderthalensis* used to interbreed (Nordborg, 1998).

Genetic inference is a well-developed subject, which becomes rather technical when one wants to incorporate biological details and use advanced statistical tools. In this section, we address some of the basic questions in genetic inference and highlight the differences between the spatial and nonspatial models. The results for well-mixed-population models presented here are usually attributed to Kingman (1982); we refer the interested reader to the book of Wakeley (2008) for an introduction.

In a typical study, *n* organisms are sampled from the population and parts of their genomes are sequenced (see Fig. 17). A genetic sequence, say …*ACTGAA*…, is an ordered string of letters taken from a four-letter alphabet: *A*, *T*, *C*, and *G*, where the letter stand for the nucleotides adenine, thymine, cytosine, and guanine, respectively. For haploid organisms considered here, an offspring inherits its sequences from the parent with possibly a few mutations (but no recombination). While a wide range of mutations is possible, we consider only point mutations, i.e., substitutions of one letter for another. Moreover, we assume that every new point mutation occurs at a new site (position) in the genome (Kimura, 1969). Because the mutation rate per site μ is very small and the total number of sites $\mathcal{L}_g$ (i.e., the length of the sequenced section of the genome) is large, most of the mutations occur at different positions along the sequence, so this infinite site approximation is reasonable on time scales shorter than μ^{-1}. For simplicity, we neglect the dependence of the mutation rates on the position within the genome as well as on the type of substitution, i.e., all 12 possible substitutions are assumed to occur at the same rate μ. We further assume that all genetic variation is neutral (Kimura, 1983).

In principle, the complete data set of *n* sequences of length $\mathcal{L}_g$ can be and often is used to estimate parameters in the model. However, we can understand the basic principles of genetic inference by considering two simple summary statistics: the average number of pairwise differences Π and the number of segregating sites *S*, i.e., the number of sites that do not have identical nucleotides in at least two sequences in the sample. The former is intimately related to the average heterozygosity *H* and the latter illustrates the use of genealogical trees in genetic inference (see Fig. 17).

We first consider Π, which is defined as the expected number of different sites in two randomly selected sequences. In a finite population, any two sequences have a common ancestor, so, as we trace them backward in time, their lineages must coalesce. We denote the average time it takes two lineages to coalesce by T_2. Then the expected number of pairwise differences is given by

$$\Pi = 2\mu \mathscr{L}_g T_2, \tag{72}$$

where the factor of 2 accounts for the fact that mutations occur in both lineages.

Since genetic inference deals with backward-in-time dynamics, it is convenient to use reverse time $\tau = -t$. To calculate T_2, we introduce the persistence probability $U_2(\tau)$, the probability that two lineages sampled at $t=0$ have not coalesced between $t=0$ and $-\tau$. Because $U_2(\tau)$ is the cumulative probability distribution function for the coalescence times, the desired probability density function is $-dU_2(\tau)/d\tau$ and T_2 can be calculated from

$$T_2 = \int_0^\infty \tau \left[-\frac{dU_2(\tau)}{d\tau} \right] d\tau = \int_0^\infty U_2(\tau) d\tau. \tag{73}$$

For a one-dimensional population, we have to take into account the positions where the organisms are sampled. Therefore, we introduce $U_2(\tau, x_1, x_2)$ as the probability that two lineages have not coalesced and are at positions x_1 and x_2 at reverse time τ. Then, $U_2(\tau)$ is given by

$$U_2(\tau) = \int_0^L \int_0^L U_2(\tau, x_1, x_2) dx_1 dx_2, \tag{74}$$

where L is the length of the habitat.

The time evolution of the persistence probability and the average heterozygosity are intimately related in both well-mixed and spatial populations because both quantities describe the fate of two lineages traced backward in time. In fact, the equation of motion for $H(t)$ is identical to that of $U_2(\tau)$, and the same is true for $H(t, x_1, x_2)$ and $U_2(\tau, x_1, x_2)$. For example, in the well-mixed-population model considered in Sec. II, $H(t)$ and $U_2(\tau)$ change only due to coalescence events, and each coalescent event changes both quantities from their current values to 0. Thus, analogously to Eq. (7) the equation of motion for $U_2(\tau)$ reads

$$\frac{d}{d\tau} U_2(\tau) = -\mathfrak{D}_g U_2(\tau), \tag{75}$$

with the initial conditions $U_2(0) = 1$. For the one-dimensional stepping stone model, we obtain that

$$\frac{\partial}{\partial \tau} U_2(\tau, x_1, x_2) = D_s \left(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} \right) U_2(\tau, x_1, x_2) - D_g U_2(\tau, x_1, x_2) \delta(x_1 - x_2) \tag{76}$$

in analogy with Eq. (30) for $\mu_{12} = \mu_{21} = s = 0$. The initial condition is $U_2(0, x_1, x_2) = \delta(x_1 - x_1^0)\delta(x_2 - x_2^0)$, where x_1^0 and x_2^0 are the positions of the first and second lineages at the time of sampling.

For the well-mixed case, we integrate both sides of Eq. (75) with respect to τ from zero $[U_2(0)=1]$ to infinity $[U_2(\infty)=0]$ and use Eq. (73) to find that $T_2 = \mathfrak{D}_g^{-1}$. Then, from Eq. (72), we obtain the average number of pairwise differences,

$$\Pi_{\text{well mixed}} = \frac{2\mu\mathscr{L}_g}{\mathfrak{D}_g}.$$

(77)

The mutation rate μ can often be measured experimentally (Drake, 1991;Araten *et al.*, 2005), so Eq. (77) and the knowledge of $\Pi_{\text{well mixed}}$ can be used to estimate the effective population size encoded in $\mathfrak{D}_g$ [see Eq. (4)].

For the one-dimensional stepping stone model, one has to specify the spatial boundary conditions for Eq. (76). Since a lineage can neither go outside the habitat nor disappear at its edge, reflecting boundary conditions should be used. With these boundary conditions, Eq. (76) has been analyzed by Wilkins and Wakeley (2002). Here we assume periodic boundary conditions, which are appropriate for a population living on a coast line of an island. These boundary conditions are simpler because they ensure translational invariance: the average coalescence time for two lineages sampled at x_1^0 and x_2^0 can only depend on $|x_1^0 - x_2^0|$ but not on x_1^0 and x_2^0 separately.

Following Wilkinson-Herbots (1998), we solve Eq. (76) with periodic boundary conditions by the Fourier transform in the positions and the Laplace transform in reverse time. The result is

$$T_2(x_1^0, x_2^0) = \frac{L}{D_g} + \frac{|x_1^0 - x_2^0|(L - |x_1^0 - x_2^0|)}{4D_s},$$

(78)

where the first term on the right-hand side is the average coalescence time for two lineages sampled at the same point and the second term is the average time for two lineages to meet for the first time.[4] Note that $T_2(x_1,x_1)$ is identical to T_2 in a well-mixed population, provided we take the effective population size to be the total size of the spatially extended population: $L/D_g = \mathfrak{D}_g^{-1}L/a = (N\tau_g/2)L/a$, where L/a is the number of demes (a is the distance between neighboring demes). Note that the distribution of the coalescence times is highly skewed, and the average coalescence time does not characterize the distribution well: most of the time coalescence occurs very fast compared to $T_2(x_1,x_1)$, but in rare cases lineages persist for times much longer than $T_2(x_1,x_1)$ (Charlesworth *et al.*, 2003).

The average number of pairwise differences for the whole data set is obtained by averaging over the spatial positions of the samples x_j, $j=1,2,\dots,n$,

$$\Pi_{1d} = \frac{4\mu\mathscr{L}_g}{n(n-1)} \sum_{j_1=2}^{n} \sum_{j_2=1}^{j_1-1} T_2(x_{j_1}, x_{j_2}),$$

(79)

where the factor $n(n-1)/2$ accounts for the total number of different ways to pair up the sequences. Given the mutation rate μ and a sufficiently large sample size n, one can use Eqs. (79) and (78) to estimate D_g and D_s. Both parameters can be estimated because Π_{1d}, unlike $\Pi_{\text{well mixed}}$, depends on the spatial positions of the samples as well as on the properties of the population. Thus, one can generate independent equations to estimate D_g and D_s using different

---

[4]The average time to the first encounter of two random walks on an interval with periodic boundary conditions is equal to the average survival time of a single random walk with twice the diffusion constant on the same interval but with absorbing boundary conditions. This equivalent problem can be solved by a standard method [see, e.g., Redner (2001)].

subsets of the samples; for example, D_g can be estimated from the samples taken from the same point and D_s can be estimated from the remaining samples. Note, however, that Π_{1d} depends only on μ/D_g and μ/D_s, so at most two parameters can be estimated from the data; similar considerations hold for the well-mixed case as well.

The average number of pairwise differences is relatively easy to compute in both spatial and nonspatial models because it depends on the history of only two lineages. For the same reason, Π does not illuminate the underlying treelike genealogy of the sample (see Fig. 17), and a different statistic is needed for that purpose. Under the infinite site assumption, a given site is either monomorphic, i.e., all samples have the same nucleotide at this site, or polymorphic, i.e., two different nucleotides are found: one is ancestral and the other is due to a mutation. Only the polymorphic sites contain information about the underlying genealogy, and the frequencies of mutations at each site are often used for genetic inference. Here we consider a simpler summary statistic, the number of segregating sites S, i.e., the expected number of polymorphic sites in the sample.

As we go backward in time, the number of lineages decreases due to coalescence events from n to $n-1$, to $n-2$, etc., until it eventually reaches 1; we consider only pairwise coalescence events assuming the population size is sufficiently large so that the coalescence of more than two lineages at one time is unlikely. Let T_j be the average time when j lineages are present. Then the expected number of polymorphic sites is given by

$$S = \mu \mathscr{L}_g \sum_{j_1=2}^{n} j T_j,$$

(80)

where the factors of j account for the fact that mutations can occur in any of the j lineages during the time interval T_j.

For a well-mixed population, we compute T_j by noticing that any of $j(j-1)/2$ distinct pairs of lineages can coalesce next and, from Eq. (75), each pair has a constant coalescence rate of $\mathfrak{D}_g$. Hence,

$$T_j = \frac{2}{j(j-1)\mathfrak{D}_g},$$

(81)

and from Eq. (80)

$$S_{\text{well mixed}} = \frac{2\mu \mathscr{L}_g}{\mathfrak{D}_g} \sum_{j_1=1}^{n-1} \frac{1}{j} \approx \frac{2\mu \mathscr{L}_g}{\mathfrak{D}_g} \left[\ln(n) + \gamma - \frac{1}{2n} \right],$$

(82)

where the approximation is valid for large sample sizes n (Gradshteyn and Ryzhik, 1980) and γ is the Euler constant.

For a one-dimensional population, one could try to generalize the approach used to calculate T_2 for multiple lineages, but this method seems prohibitively difficult for large n. However, we can qualitatively understand the effects of spatial structure by considering n lineages sampled uniformly in x from the habitat. While analyzing a related problem of annihilating random walks (see Appendix F), Doering and ben-Avraham (1988) and Zhong and ben-Avraham (1995) showed that for a generic uniform spatial distribution of the samples the number of surviving lineages j at reverse time τ decays as

$$j(\tau) \sim \frac{1}{\sqrt{2\pi D_s \tau}}$$

(83)

for intermediate times, when, on the one hand, the time is sufficiently small for any lineage to diffuse across the whole habitat ($\tau \ll L^2/D_s$), but, on the other hand, the time is sufficiently large for neighboring lineages to coalesce ($\tau \ll Ln^{-1}D_g^{-1} + L^2 n^{-2} D_s^{-1}$).

Since T_j is the time during which the number of lineages changes from j to $j-1$, it follows from Eq. (83) that

$$T_j \approx \tau(j-1) - \tau(j) \approx -\frac{d\tau(j)}{dj} = \frac{1}{\pi D_s j^3},$$

(84)

where $\tau(j)$ is the inverse function of $j(\tau)$ used in Eq. (83). Equation (84) is only valid for intermediate j values: $1 \ll j \ll n$ because of the similar restrictions on Eq. (83). Upon comparing this one-dimensional result for T_j to Eq. (81), we see that the well-mixed model overestimates the contribution to the number of segregating sites from the recent part of genealogy with a large number of lineages. This should also be true for the initial stage $j \approx n$, where Eq. (83) is not valid, because faster coalescence results from the fact that a lineage has to travel only about L/n to meet its neighbor. Other statistics that rely on the relative duration of periods with j lineages should be affected in a similar way. This is particularly important when the deviations of the observed genealogical data from the predictions based on Eq. (81) are used to infer past evolutionary events, such as a selective pressure or geographic isolation (Wakeley, 2008), because some of these deviations could be due to the spatial structure of the population rather than external or internal perturbations.

In summary, the classic theory of genetic inference can be extended to spatial populations. These extensions are not only more accurate and realistic than assuming the well-mixed-population dynamics but also can be used to obtain information about the migration within the habitat (Wilkins and Wakeley, 2002). As spatially resolved genetic data sets become more readily available, better statistical tools based on spatial population genetic models will be needed.

X. CONCLUSIONS

Fluctuations due to sampling error during reproduction significantly affect the evolutionary dynamics of quasi-one-dimensional populations, e.g., two-dimensional populations undergoing range expansions. These fluctuations lead to the genetic demixing illustrated in Fig. 18, where an initially well-mixed population of alleles "phase separates" into monoallelic domains. The transition is somewhat analogous to spinodal decomposition in physics and materials science (Scheucher and Spohn, 1988) but is also markedly different. In particular, unlike conventional demixing phase transitions in finite-temperature statistical mechanics, genetic demixing occurs only in a *low* number of spatial dimensions d ($d \le 2$) (Scheucher and Spohn, 1988; Duty, 2000). The dependence of genetic demixing on the number of spatial dimensions d is illustrated by the decay of local heterozygosity in the absence of selection and mutation. For long times, the functional form of the decay is given by Duty (2000),

$$H(t,0) \sim \begin{cases} e^{-2t/(N\tau_g)}, & d=0 \\ (\tau_f/t)^{1/2}, & d=1 \\ 1/\ln(t), & d=2 \\ \text{const}, & d>2. \end{cases}$$

(85)

Note that $d=2$ is the critical dimension.

Here we have shown that the one-dimensional stepping stone model has very different dynamics compared to the standard well-mixed-population models used in population genetics. Most of the differences arise because, in the spatial model, populations segregate into monoallelic domains. As a result, genetic drift and selection can only act at the boundaries of the domains, which slows down the dynamics of the model. In particular, we found that, in the neutral model without mutation, fixation occurs exponentially fast in a well-mixed population, but the decay of heterozygosity is algebraic in the spatial model. Genetic drift in the population as a whole becomes weaker with time as spatial diffusion causes the effective population size to increase. For a linear one-dimensional model, we also found that the standard deviation of the total fraction of one of the alleles (in the absence of selection and mutation) increases subdiffusively as $t^{1/4}$. Selective sweeps also occur more slowly in the spatial model: for weak selection, $s \ll D_g^2/D_s$, we found that the time constant of the selective sweep is quadratic in s in the linear spatial model, but it is only linear in s in the well-mixed-population model. The effects of mutation do not differ as dramatically in spatial and nonspatial models, but the stepping stone model reveals nontrivial spatial correlations and predicts a different value for the local steady-state heterozygosity, proportional to $\sqrt{\mu_{12}+\mu_{21}}$ for small mutation rates, compared to $\mu_{12}+\mu_{21}$ in the well-mixed-population model. The evolutionary dynamics of spatial models also depends on the geometry of the expansion. For radial expansions, we found that the number of domains approaches a finite limit, which is, up to an additive constant, proportional to the square root of the initial radius of the colony R_0.

Our main conclusion is that the data from natural populations may not always conform to the predictions of the well-mixed-population model and, even when they do, the estimated parameters from the model may not be biologically meaningful. The spatial model contains an important additional parameter, the spatial diffusion constant parameter D_s, which enters into many of the predictions. For example, the time scale for local fixation is given by D_s/D_g^2 rather than $N\tau_g$ [see Eq. (36)] and, for small selective advantage, s is sometimes replaced by $s^2 D_s/D_g^2$ (see Sec. VI). Moreover, as we saw in Sec. VII, the time scale of fixation depends on the partitioning of the population by the experimenter into measurement sites. Thus care must be taken when interpreting the data from the natural populations. Finally, well-mixed-population models and experiments without spatial resolution do not account for spatial correlations, which contain important information about the population (see Secs. V and IX).

Acknowledgments

APPENDIX A: THE ITÔ CALCULUS

In this appendix, we discuss the Itô calculus. Our presentation relies on Gardiner (1985) and Risken (1989), which can be consulted for a more extensive presentation. For simplicity, we only consider nonspatial stochastic differential equations, but the results can be extended to spatial problems straightforwardly.

We analyze the following stochastic differential equation, which includes Eq. (21) as a special case:

$$\frac{d\psi}{dt} = \omega(\psi) + g(\psi)\Gamma(t), \tag{A1}$$

where $\Gamma(t)$ satisfies Eq. (22) and $\omega(\psi)$ and $g(\psi)$ are arbitrary continuously differentiable functions. From the point of view of ordinary calculus, Eq. (A1) is not well defined because $\Gamma(t)$ is discontinuous at every point. One way to circumvent this problem is to use discrete time steps of infinitesimal length δt rather than continuous time. Then, Eq. (A1) takes the following form:

$$\frac{\psi(t+\delta t) - \psi(t)}{\delta t} = \omega(\psi(t)) + g(\psi(t))\Gamma(t). \tag{A2}$$

However, this is not the only way to interpret Eq. (A1). For example, an alternative way to go from the continuous to a discrete description is to write Eq. (A1) as

$$\frac{\psi(t+\delta t) - \psi(t)}{\delta t} = \omega\left(\frac{\psi(t)+\psi(t+\delta t)}{2}\right) + g\left(\frac{\psi(t)+\psi(t+\delta t)}{2}\right)\Gamma(t). \tag{A3}$$

In fact, there is an infinite number of ways to interpret Eq. (A1), depending on the relative weight of $\psi(t)$ and $\psi(t+\delta t)$ inside the arguments of the functions on the right-hand side of the equation. The two most commonly used interpretations are Itô's and Stratonovich's prescriptions. The former corresponds to Eq. (A2) and the latter to Eq. (A3).

In physics, Stratonovich's prescription is commonly used because $\Gamma(t)$ is usually an approximation to a thermal noise with small but finite correlation time; therefore, the argument of $g(\cdot)$ should be an average value of ψ over the time that the correlations persist. In population genetics, on the other hand, Ito's prescription is appropriate because a random change of the allele frequencies depends only on the genetic composition of the population *prior* to the change.

Without the stochastic term, Eqs. (A2) and (A3) would yield the same results provided δt is sufficiently small, but the stochastic terms remain different even in the limit $\delta t \to 0$. An easy way to see this difference is to average Eqs. (A2) and (A3) with respect to the nondifferentiable noise function $\Gamma(t)$. Itô's prescription gives $\langle\psi(t+\delta t)\rangle - \langle\psi(t)\rangle = \langle\omega(\psi(t))\rangle\delta t$ because $\langle g(\psi(t))\Gamma(t)\rangle = \langle g(\psi(t))\rangle\langle\Gamma(t)\rangle = 0$ due to the independence of $\psi(t)$ and $\Gamma(t)$. A similar simplification, however, cannot be applied to Stratonovich's prescription because, generically, the stochastic term depends on $\psi(t+\delta t)$, which is not independent of $\Gamma(t)$.

Because of the aforementioned ambiguity in interpreting stochastic differential equations with multiplicative noise, care must be taken while differentiating stochastic variables. While the rules of ordinary calculus apply to Stratonovich's prescription, special rules of the Itô calculus

are required for Itô's prescription when tracking the evolution of a composite function $u(\psi(t))$ of the stochastic variable obeying Eq. (A1). In this paper, we use Itô's formula, namely (Gardiner, 1985;Risken, 1989),

$$\frac{d}{dt}u(\psi(t))=u'(\psi(t))\omega(\psi(t))+u'(\psi(t))g(\psi(t))\Gamma(t)+\frac{1}{2}u''(\psi(t))g^2(\psi(t)),$$

(A4)

where $u(\psi)$ is a twice continuously differentiable function and the primes now indicate differentiation with respect to ψ. The last term is the crucial addition due to the Itô calculus.

We conclude this discussion with an illustration of how Eq. (A4) can be used by deriving Eqs. (6) and (7) from Eq. (21) assuming $s=0$ and $\mu_{12}=\mu_{21}=0$. Thus, we start from the following equation of motion for $f(t)$:

$$\frac{df(t)}{dt}=\sqrt{\mathfrak{D}_g f(t)[1-f(t)]}\Gamma(t)(\text{Itô}).$$

(A5)

Thus, $\psi(t)=f(t)$, $\omega(\psi(t))=0$, and $g(\psi(t))=\sqrt{\mathfrak{D}_g \psi(t)[1-\psi(t)]}$. Since $F(t)=\langle f(t)\rangle$, we obtain Eq. (6) by averaging Eq. (A5). For $H(t)=\langle h(t)\rangle=\langle 2f(t)[1-f(t)]\rangle$, we use Eq. (A4) with $u(\psi(t))=2\psi(t)[1-\psi(t)]$ to obtain the equation of motion for $h(t)$,

$$\frac{dh(t)}{dt}=0+2[1-2f(t)]\sqrt{\mathfrak{D}_g f(t)[1-f(t)]}\Gamma(t)+\frac{1}{2}(-4)\mathfrak{D}_g f(t)[1-f(t)](\text{Itô}).$$

(A6)

Upon averaging Eq. (A6) with the rules described above, we obtain Eq. (7).

APPENDIX B: SOLUTION OF THE NEUTRAL MODEL WITHOUT MUTATIONS

In this appendix, we solve Eq. (33) subject to the initial condition $H(0,x)=H_0$. It is advantageous to first solve a simpler equation,

$$\frac{\partial}{\partial t}H=2D_s\frac{\partial^2}{\partial x^2}H-b(t)\delta(x),$$

(B1)

where $b(t)$ is an arbitrary function of time. Equation (B1) is a standard diffusion equation with a sink term, and it can be readily solved in the Fourier domain. The result is

$$H(t,x)=H_0-\int_0^t dt'\frac{b(t')e^{-x^2/8D_s(t-t')}}{\sqrt{8\pi D_s(t-t')}}.$$

(B2)

Note the convolution of $b(t')$ with the diffusion propagator. Now, we impose a self-consistency condition $b(t)=D_g H(t,0)$, which leads to

$$H(t,0)=H_0-D_g\int_0^t dt'\frac{H(t',0)}{\sqrt{8\pi D_s(t-t')}}.$$

(B3)

This is Abel's integral equation of the second kind, canonically written as

$$y(x)+\lambda\int_a^x\frac{y(t)dt}{\sqrt{x-t}}=g(x),$$

(B4)

where $g(x)$ is a known function. The general solution of Eq. (B4) given by Polyanin and Manzhirov (1998) reads as

$$y(x)=G(x)+\pi\lambda^2\int_a^x e^{\pi\lambda^2(x-t)}G(t)dt,$$

where

$$G(x)=g(x)-\lambda\int_a^x\frac{g(t)dt}{\sqrt{x-t}}.$$

(B5)

Equations (34) and (35) follow from Eqs. (B2)–(B5).

For radial expansions considered in Sec. VIII, one can solve the equation of motion for $H(t,\phi)$ by following the same set of steps.

APPENDIX C: AVERAGE DOMAIN DENSITY FROM THE SPATIAL HETEROZYGOSITY H(t,x)

In this appendix, we derive the relationship between the spatial heterozygosity $H(t,x)$ and the average domain density $n_d(t)$. From n_d, we can get a domain size by defining $\ell\equiv n_d^{-1}$. The result for the domain density is valid for an arbitrary number of alleles, so in this appendix we use a broader definition of $H(x,t)$ as the average probability of sampling at time t two different alleles from two demes distance x apart. We assume that the domains have formed, and they are on average much larger than the boundary regions.

Let $h(t,x_1,x_2)$ equal to 1 if both x_1 and x_2 are occupied by organisms in different allelic state and 0 otherwise. To compute ℓ, we use an alternative definition of $H(t,x)$ with ensemble average replaced by space average,

$$H(t,x)=\lim_{L\to\infty}\frac{1}{L}\int_0^L h(t,\xi,\xi+x)d\xi,$$

(C1)

where we assume periodic boundary conditions. We compute

$$H(t,x+\delta x)-H(t,x)=\lim_{L\to\infty}L^{-1}\int_0^L[h(t,\xi,\xi+\delta x)-h(t,\xi,\xi)]d\xi$$

for δx small compared to typical domain size but large compared to the deme spacing a. To do so, we expand both sides in δx. At the lowest order in δx, each domain boundary contributes δx to the right-hand side; therefore, $\partial H(t,+0)/\partial x$ equals the density of the domain boundaries. Upon defining the average domain size $\ell(t)$ as the inverse of the domain boundary density, we obtain the following relationship:

$$\ell(t) = \left(\frac{\partial H(t, +0)}{\partial x} \right)^{-1}.$$

(C2)

This relation is analogous to the one derived by ben-Avraham (1998).

We can further simplify Eq. (C2) by observing that

$$\frac{\partial H(t, +0)}{\partial x} = \frac{D_g H(t, 0)}{4 D_s},$$

(C3)

which follows from integrating Eq. (32) or (D1) with respect to x from $-\varepsilon$ to ε, $0 < \varepsilon \ll 1$, and noticing that $H(t,x)$ is an even function of x. The final result then reads

$$\ell(t) = \frac{4 D_s}{D_g H(t, 0)}.$$

(C4)

It should be emphasized that this result is only valid in the limit of very large domain sizes compared to the boundary regions, which means $H(t,0) \ll 1$. Therefore the leading term in $H(t,0)$ is sufficient at this level of approximation. Note that Eqs. (C2) and (C4) are valid in the presence of genetic drift, migration, selection, and mutation. For radially expanding populations subject to inflation, Eq. (C2) remains valid, but Eq. (C4) is replaced by Eq. (66).

APPENDIX D: INFINITE ALLELE MODEL

In this appendix, we extend the analysis of the stepping stone model with mutations presented in Sec. V to the infinite allele model. The infinite allele model assumes that every new mutation creates a new allele, which is a good approximation for genes encoded by a large number of nucleotides because the number of all possible mutations is much larger than the number of all possible back mutations (Hartl and Clark, 1989). The equation of motion for $H(t,x)$, which we interpret as the average probability of sampling two different alleles from demes x apart, can be derived by following two lineages backward in time, as done in Sec. IV. In the presence of mutation, the right-hand side of Eq. (33) should contain an additional term describing the rate of increase of $H(t,x)$ due to mutations in both of the lineages. Because, in the infinite allele model, a mutation changes the probability that the organisms have different alleles from H to 1, that is, by $1-H$, the new term is $2\mu(1-H)$, where μ is the mutation rate that is assumed to be the same for all types of mutations. Thus, Eq. (33) becomes

$$\frac{\partial}{\partial t} H = 2 D_s \frac{\partial^2}{\partial x^2} H + 2\mu(1 - H) - D_g H \delta(x)$$

(D1)

for the infinite allele model [compare Eq. (47)]. The stationary solution of Eq. (D1) is given by

$$H(\infty, x) = 1 - \frac{e^{-(\sqrt{\mu/D_s})|x|}}{1 + \frac{1}{4} D_g / \sqrt{D_s \mu}}.$$

(D2)

At large separations, $H(\infty,x)$ approaches 1, which is consistent with the infinite number of alleles. Locally, $H(\infty,0)=[1+(1/4)D_g/\sqrt{D_s\mu}]^{-1}$, and if $H(\infty,0)\ll 1$ the population is segregated into domains containing only one allelic type. The average size of such domains follows from Eq. (C4),

$$\ell=\frac{4D_s}{D_g}\left(1+\frac{1}{4}\frac{D_g}{\sqrt{D_2\mu}}\right)\approx\sqrt{\frac{D_s}{\mu}},$$

(D3)

where the last equality follows from the assumption that $H(\infty,0)\ll 1$. The approach to the stationary state can be obtained either by methods of Appendix B or by the change of variables $H(t,x)=H(\infty,t)+e^{-2\mu t}\hat{H}(t,x)$, which reduces Eq. (D1) to Eq. (33). The result is that the slowest decaying mode vanishes as $\tilde{C}t^{-1/2}e^{-2\mu t}$, where $\tilde{C}$ is a constant.

The infinite allele model and Eq. (D1) have been analyzed before by Nagylaki (1974) and Malécot (1975), who calculated the stationary solution and the long-time approach to the equilibrium. Our results are consistent with their findings.

APPENDIX E: A MODEL WITH SEVERAL NEUTRAL ALLELES

A model with q neutral alleles is an intermediate case between the two-allele model that we focus on in this paper and the infinite allele model discussed in Appendix D. The q-allele model is also analogous to nonequilibrium q-state Potts models. In this appendix, we outline how the q-allele model can be formulated and solved in the language of one- and two-point correlation functions, compare our analytical predictions to simulations, and extend Eq. (44) to the undulating-front model.

To specify the q-allele model, we let $f_i(t,x)$ be the frequency of allele i at time t and position x; these quantities satisfy $\Sigma_{i=1}^{q}f_i(t,x)=1$. The spatial diffusion and coalescence probability of two lineages are still characterized by D_s and D_g, respectively. Intra-allelic mutations are described by the mutation matrix μ_{ij}, which is the probability of allele i mutating into allele j if $i\neq j$. When $i=j$, we let $\mu_{ii}=-\Sigma_{j=1,j\neq i}^{q}\mu_{ij}$ to describe the outflow of alleles from allelic state i due to mutations.

The dynamics of the q-allele model can be analyzed by considering one-point correlation functions $F_i(t,x)=\langle f_i(t,x)\rangle$ and two-point correlation functions $F_{ij}(t,x_1,x_2)=\langle f_i(t,x_1)f_j(t,x_2)\rangle$. $F_i(t,x)$ is the probability to find allele i at position x at time t and $F_{ij}(t,x_1,x_2)$ is the probability to simultaneously find at time t allele i at position x_1 and allele j at position x_2. The evolution equations for these correlation functions are obtained by tracing one and two lineages backward in time; the results are

$$\frac{\partial F_i(t,x)}{\partial t}=D_s\frac{\partial^2 F_i(t,x)}{\partial x^2}+\sum_{j=1}^{q}\mu_{ji}F_j(t,x),$$

(E1)

$$\frac{\partial F_{ij}(t, x_1, x_2)}{\partial t}$$

$$= D_s\left(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}\right)F_{ij}(t,$$

$$x_1, x_2) + D_g\delta(x_1$$

$$- x_2)[\delta_{ij}F_i(t, x_1) - F_{ij}(t,$$

$$x_1, x_2)] + \sum_{i'=1}^{q}\sum_{j'=1}^{q}[\mu_{i'i}F_{i'j}(t, x_1, x_2) + \mu_{j'j}F_{ij'}(t, x_1, x_2)],$$

(E2)

where δ_{ij} is Kronecker's delta, which is 0 if $i \neq j$ and 1 otherwise. Thus, for a generic mutation matrix μ_{ij} one has to solve a system of coupled linear partial differential equations.

For simplicity and the ease of comparison with the other results in this paper, assume spatial homogeneity and identical mutation rates between any two alleles, $\mu_{i\neq j} = \mu/q$. Under these assumptions, Eq. (E2) can be simplified by introducing averaged spatial heterozygosity,

$$H(t, x) = \sum_{i=1}^{q}\sum_{\substack{j=1 \\ j \neq i}}^{q} F_{ij}(t, 0, x) = \sum_{i=1}^{q}\langle f_i(t, 0)[1 - f_i(t, x)]\rangle,$$

(E3)

which is the probability to sample two different alleles at time t distance x apart. The equation of motion for $H(t,x)$ can be derived both from Eq. (E2) and, more simply, by tracing two lineages backward in time,

$$\frac{\partial}{\partial t}H = 2D_s\frac{\partial^2}{\partial x^2}H + 2\mu\left(\frac{q-1}{q} - H\right) - D_g H\delta(x).$$

(E4)

Note that Eq. (E4) agrees with Eq. (D1) in the limit $q \to \infty$ and with Eq. (47) for $\mu_{12} = \mu_{21} = \mu/2$. Since Eq. (E4) has the same functional form as Eq. (D1), the methods of Appendix D can be used to solve for $H(t,x)$.

In the absence of mutations, Eq. (E4) is identical to Eq. (33), as mentioned in Sec. IV. However, q-allele models with different q may have slightly different dynamics due to q-dependent initial conditions: for example, an initially well-mixed population is represented by $H(0,x) = H_0 = 1 - 1/q$. Thus the results of Sec. IV apply to the q-allele model, provided appropriate initial conditions are used. In particular, we expect the standard deviation of $f_i(t)$, the total frequency of allele i, in a finite population to grow as $t^{1/4}$. This is indeed confirmed by our simulations shown in Fig. 19. Spatial correlations in the nonequilibrium q-state Potts model have recently been analyzed by Masser and ben-Avraham (2000), who also found that two-point correlation functions obey the same q-independent equation of motion.

Finally, one can obtain the behavior of the standard deviation of the total frequency of allele one $\Delta(t)$ in the undulating-front model by the following scaling argument. We consider $\Delta(t)$ at large times after monoallelic domains have formed. Let $N_d(t)$ be the number of domains consisting of allele one and $d_k(t)$, $k=1,2,\ldots,N_d(t)$, be lengths of these domains. Then, $\Delta(t)$ is given by

$$\Delta(t) = \sqrt{\frac{1}{L^2} \left\langle \left| \sum_{k=1}^{N_d(t)} d_k(t) - \left\langle \sum_{k=1}^{N_d(t)} d_k(t) \right\rangle \right|^2 \right\rangle}.$$

(E5)

We simplify Eq. (E5) by making an approximation that $N_d(t)$ and $d_k(t)$ for $k=1,2,\dots,N_d(t)$ are independent random variables, which gives

$$\Delta^2(t) \approx \frac{1}{L^2} \{ \langle N_d(t) \rangle \, [\langle d_1^2(t) \rangle - \langle d_1(t) \rangle^2] + \langle d_1(t) \rangle^2 [\langle N_d^2(t) \rangle - \langle N_d(t) \rangle^2] \},$$

(E6)

where we used the fact that $d_i(t)$ are identically distributed.

By using first passage time analysis discussed by Redner (2001), one can show that

$$\langle N_d^2(t) \rangle - \langle N_d(t) \rangle^2 \propto \langle N_d(t) \rangle \, [\langle d_1^2(t) \rangle - \langle d_1(t) \rangle^2] / \langle d_1(t) \rangle^2 \propto L [\langle d_1^2(t) \rangle - \langle d_1(t) \rangle^2] / \langle d_1(t) \rangle^3.$$

Thus

$$\Delta^2(t) \propto \frac{1}{L \langle d_1(t) \rangle} [\langle d_1^2(t) \rangle - \langle d_1(t) \rangle^2].$$

(E7)

Upon recalling that, in the undulating-front model, $\langle d_1(t) \rangle \propto t^\zeta$ and $\langle d_1(t) \rangle^2 \propto t^{2\zeta}$, we conclude that

$$\Delta(t) \propto \frac{t^{\zeta/2}}{\sqrt{L}} \propto \frac{t^{1/3}}{\sqrt{L}},$$

(E8)

where, in the last proportionality, we used $\zeta = 2/3$ from Saito and Müller-Krumbhaar (1995). Equation (E8) is in good agreement with the simulations of the undulating-front model shown in Fig. 20.

APPENDIX F: CONNECTION WITH THE VOTER MODEL AND ONE-DIMENSIONAL REACTION KINETICS

The stepping stone model with only one organism per island or "deme," $N=1$, has been extensively studied in probability theory (Durrett, 1988; Liggett, 2004) and nonequilibrium statistical mechanics (Ódor, 2004), where it is known as the voter model. The model typically considers a set of voters on a hypercubic lattice in d dimensions. Each voter holds one of the q possible opinions about an issue (corresponding to q alleles in population genetics), and, at a certain rate, each voter reconsiders the issue and adopts the opinion of a randomly chosen nearest neighbor. The voter model can be mapped onto the dynamics of the q-state Potts model at zero temperature. In one and two dimensions, opinions in the voter model coarsen spatially with time, and the model approaches one of the q absorbing states, in which all the voters have the same opinion (Cox and Griffeath, 1986; Duty, 2000). In higher dimensions, the voters still form cluster of opinions, but these clusters stop growing after reaching a certain limiting size. Selection and mutation are typically not considered in voter models.

The voter model can be solved exactly by tracing the history of opinion adoptions backward in time (Cox and Griffeath, 1986; Scheucher and Spohn, 1988; Duty, 2000). The opinion of a given individual performs a random walk as we follow the opinion from its current holder to its ultimate ancestor. With this observation, we can easily understand how the behavior of the voter model depends on the number of spatial dimensions. In one and two dimensions, a pair of random walks always meet (Redner, 2001), so the histories of opinion adoptions starting from two different voters will eventually converge to a single voter as we trace them backward in time. Therefore, any two voters should have the same opinion after a sufficiently long time has elapsed. In higher dimensions, however, there is a finite probability that two random walkers never meet (Redner, 2001); therefore, the voters never agree, and an absorbing state is never reached.

Another important property of the voter model is that the dynamics occurs only at the boundaries between the opinion clusters; inside a cluster the opinions cannot change because every voter has the same opinion as its nearest neighbors. This property is particularly useful in one spatial dimension, where it allows us to map the dynamics of the voter model to the one-dimensional diffusion-limited chemical kinetics of point particles. We identify each domain wall with a particle performing a random walk due to opinion changes at the boundary. When two particles meet, they react with two possible outcomes. They annihilate ($A+A \rightarrow 0$) if the flanking domains have the same opinion or coalesce ($A+A \rightarrow A$) otherwise. If the initial state is uncorrelated, the annihilation occurs with probability $1/(q-1)$, and the coalescence with probability $(q-2)/(q-1)$. In one dimension, this reaction-diffusion system has been analyzed by Masser and ben-Avraham (2000), who found that the density of the domain walls decays as $t^{-1/2}$ in agreement with Eq. (38). A related model of annihilating random walks for radial and linear range expansions was solved by Hallatschek and Nelson (2010).

It is not surprising that the voter model and the stepping stone model have the same long-time behavior in one dimension. At long times, most of the voters belong to large domains; therefore, we do not affect the system by combining neighboring sites into larger coarse-grained demes, as in Sec. VII. For these large demes, the equations of motion of the stepping stone model are valid, so the two models are equivalent in the long-time limit. The voter and stepping stone models are also equivalent in the small-D_s limit of very slow migration. In this case, each deme reaches fixation much faster than it sends out or accepts new migrants; hence, the stepping stone model reduces to the voter model with one voter representing an entire deme.

We can further illustrate the connection between the stepping stone model and the voter model by calculating the probability that two voters l sites apart have different opinions. This probability is analogous to the average spatial heterozygosity, so we call it $H(t,l)$. The equation of motion for $H(t,l)$ is obtained by following the histories of opinion adoptions backward in time. Since $H(t,l)$ changes only due to the diffusion of the history traces, the equation of motion reads

$$\frac{d}{dt}H(t,l)=[H(t,l-1)+H(t,l+1)-2H(t,l)],$$

(F1)

where we measure time in such units that the rate of opinion adoption is set to unity. While Eq. (F1) can be solved exactly (Houchmandzadeh and Vallade, 2003), it is more instructive to go to the continuum limit, in which the equation of motion for $H(t,x)$ takes the following form:

$$\frac{\partial}{\partial t}H(t,x)=2D_s\frac{\partial^2}{\partial x^2}H(t,x),$$

(F2)

where D_s denotes the spatial diffusion constant as in Eq. (33).

Upon comparing Eqs. (33) and (F2), one might naively conclude that the voter model corresponds to D_g=0 limit of the stepping stone model; in fact, the opposite is true: the voter model corresponds to the limit D_g=∞. Qualitatively, one can see this from the fact that $D_g \propto N^{-1}$, so, as the deme size N approaches its lowest value of 1, we expect D_g to increase. On more rigorous grounds, we should note that the role of the delta function in Eq. (33) is to enforce a boundary condition at x=0, provided one considers $H(t,x)$ only for x>0. This boundary condition is derived in Appendix C and is given by Eq. (C3). The corresponding boundary condition for Eq. (F2) is $H(t,0)$=0 because the probability of one voter having two different opinions is 0. We indeed recover $H(t,0)$=0 by letting $D_g \to \infty$ in Eq. (C3).

One can solve Eq. (F2) for the initial condition $H(0,x)=H_0$ by the Laplace transform in time or a self-similar ansatz; the solution reads

$$H(t, x) = H_0 \mathrm{erf}\left(\frac{|x|}{\sqrt{8D_s t}}\right).$$

(F3)

We can now compute the average size of the domains with the help of Eq. (C2). As we expect, the result is given by Eq. (38) because the long-time limits of the stepping stone model and the voter model agree.

The D_g=∞ approximation is particularly valuable for circular fronts undergoing inflation because the exact solution of the stepping stone model in this case [Eqs. (64) and (65)] is rather unwieldy. The equation of motion for $H(t,\varphi)$ in the voter model with inflation is given by

$$\frac{\partial}{\partial t}H(t,\varphi) = \frac{2D_s}{(R_0+vt)^2}\frac{\partial^2}{\partial \varphi^2}H(t,\varphi).$$

(F4)

One can solve Eq. (F4) in the Fourier domain and compute the nontrivial limit-shape as $t \to \infty$. The result reads

$$H(\infty, \varphi) = H_0 \mathrm{erf}\left(|\varphi|\sqrt{\frac{R_0 v}{8D_s}}\right).$$

(F5)

With the help of the angular version of Eq. (C2) [see Eq. (66)], we calculate the final number of sectors,

$$\mathcal{N}(\infty) = H_0 \sqrt{\frac{2\pi R_0 v}{D_s}},$$

(F6)

which agrees with Eq. (69) in the limit D_g=∞. In the same limit, we can also obtain an analytical expression for the long-time variance $v(\infty)$,

$$v(\infty) = \frac{1}{4\pi}\int_{-\pi}^{\pi}[H_0 - H(\infty,\varphi)]d\varphi = H_0\sqrt{\frac{2D_s}{\pi^3 R_0 v}},$$

(F7)

where we used the relationship between the variance $v(t)$ and the global heterozygosity $\mathcal{H}$ (t,φ) given by the spatial generalization of Eq. (8).

Finally, we note that the mapping to a one-dimensional reaction-diffusion system of particles could be generalized to account for superdiffusive boundaries in the undulating-front model, for example, by considering continuous time Lévy flights instead of random walks [see Hinrichsen and Howard (1999)]. In the chemical kinetics picture, one can also account for mutations by introducing a birth process $0 \rightarrow 2A$ and for natural selection by imposing an attraction between the particles flanking domains of the deleterious allele.

REFERENCES

Araten D, Golde D, Zhang R, Thaler H, Gargiulo L, Notaro R, Luzzatto L. Cancer Res 2005;65:8111. [PubMed: 16166284]

Barton N, Depaulis F, Etheridge A. Theor Popul. Biol 2002;61:31. [PubMed: 11895381]

Bar-Yam, Y. Making Things Work: Solving Complex Problems in a Complex World. Brookline, MA: Knowledge; 2005.

ben-Avraham D. Phys. Rev. Lett 1998;81:4756.

Blythe R, McKane A. J. Stat. Mech.: Theory Exp 2007;2007:P07018.

Bramson M, Griffeath D. Ann. Probab 1980;8:183.

Bramson M, Lebowitz J. J. Stat. Phys 1991;62:297.

Charlesworth B, Charlesworth D, Barton N. Annu. Rev. Ecol. Syst 2003;34:99.

Cox J, Griffeath D. Ann. Probab 1986;14:347.

Crow, J.; Kimura, M. An Introduction to Population Genetics Theory. New York: Harper & Row; 1970.

Doering C, ben-Avraham D. Phys. Rev. A 1988;38:3035. [PubMed: 9900720]

Doering C, Mueller C, Smereka P. Physica A 2003;325:243.

Drake J. Proc. Natl. Acad. Sci. U.S.A 1991;88:7160. [PubMed: 1831267]

Durrett, R. *Particle Systems and Percolation*, Lecture Notes on Particle Systems and Percolation. Pacific Grove, CA: Wadsworth; 1988.

Duty, TL. Ph.D. thesis. The University of British Columbia; 2000.

Fisher R. Ann. Eugenics 1937;7:353.

Gardiner, C. Handbook of Stochastic Methods. New York: Springer; 1985.

Gibson, A. Oklahoma: A History of Five Centuries. Harlow, Norman: 1965.

Goldenfeld, N. *Phase Transitions and the Renormalization Group*, Lectures on Phase Transitions and the Renormalization Group. Reading, MA: Addison-Wesley; 1992.

Gradshteyn, I.; Ryzhik, I. Table of Integrals, Series, and Products. New York: Academic; 1980.

Guth A. Phys. Rev. D 1981;23:347.

Hallatschek O, Hersen P, Ramanathan S, Nelson DR. Proc. Natl. Acad. Sci. U.S.A 2007;104:19926. [PubMed: 18056799]

Hallatschek O, Korolev KS. Phys. Rev. Lett 2009;103:108103. [PubMed: 19792344]

Hallatschek O, Nelson DR. Evolution 2010;64:193. [PubMed: 19682067]

Hartl, D.; Clark, A. Principles of Population Genetics. Sunderland: Sinauer; 1989.

Hewitt G. Biol. J. Linn. Soc 1996;58:247.

Hinrichsen H, Howard M. Eur. Phys. J. B 1999;7:635.

Houchmandzadeh B, Vallade M. Phys. Rev. E 2003;68:061912.

Kimura M. Proc. Natl. Acad. Sci. U.S.A 1955;41:144. [PubMed: 16589632]

Kimura M. Genetics 1969;61:893. [PubMed: 5364968]

Kimura, M. The Neutral Theory of Molecular Evolution. Cambridge: Cambridge University Press; 1983.

Kimura M, Weiss G. Genetics 1964;49:561. [PubMed: 17248204]

Kingman J. J. Appl. Probab 1982;27:235.

Kolmogorov A, Petrovsky N, Piscounov N. Moscow Univ. Math. Bull 1937;1:1.

Liggett, T. Interacting Particle Systems. New York: Springer; 2004.

Lin J. Phys. Rev. A 1991;44:6706. [PubMed: 9905799]

Malécot G. Cold Spring Harbor Symp. Quant. Biol 1955;20:52.

Malécot G. Theor Popul. Biol 1975;8:212. [PubMed: 1198353]

Masser T, ben-Avraham D. Phys. Lett. A 2000;275:382.

Mayr, E. Systematics and the Origin of Species from the Viewpoint of a Zoologist. New York: Columbia University; 1942.

Murray, J. Mathematical Biology. New York: Springer; 2003.

Nagylaki T. Proc. Natl. Acad. Sci. U.S.A 1974;71:2932. [PubMed: 4528770]

Nagylaki T. Proc. Natl. Acad. Sci. U.S.A 1978;75:423. [PubMed: 16592484]

Nordborg M. Am. J. Hum. Genet 1998;63:1237. [PubMed: 9758610]

Nowak, M. Evolutionary Dynamics: Exploring the Equations of Life. Cambridge, MA: Harvard University; 2006.

Ódor G. Rev. Mod. Phys 2004;76:663.

Poli, R.; Langdon, W.; McPhee, N. A field guide to genetic programming. 2008. Published via http://lulu.com and freely available at http://www.gp-field-guide.org.uk, 2008 (with contributions by J. R. Koza)

Polyanin, A.; Manzhirov, A. Handbook of Integral Equations. Boca Raton, FL: CRC; 1998.

Ramachandran S, Deshpande O, Roseman C, Rosenberg N, Feldman M, Cavalli-Sforza L. Proc. Natl. Acad. Sci. U.S.A 2005;102:15942. [PubMed: 16243969]

Redner, S. A Guide to First Passage Processes. Cambridge: Cambridge University Press; 2001.

Risken, H. The Fokker-Planck Equation: Methods of Solution and Applications. Berlin: Springer; 1989.

Saito Y, Müller-Krumbhaar H. Phys. Rev. Lett 1995;74:4325. [PubMed: 10058472]

Scheucher M, Spohn H. J. Stat. Phys 1988;53:279.

Templeton A. Nature (London) 2002;416:45. [PubMed: 11882887]

Wakeley, J. Coalescent Theory: An Introduction. Greenwood Village, CO: Roberts; 2008.

Wilkins J, Wakeley J. Genetics 2002;161:873. [PubMed: 12072481]

Wilkinson-Herbots H. J. Math. Biol 1998;37:535.

Wilson K, Kogut J. Phys. Rep 1974;12:75.

Wright S. Genetics 1931;16:97. [PubMed: 17246615]

Wright S. Genetics 1943;28:114. [PubMed: 17247074]

Zhong D, ben-Avraham D. Phys. Lett. A 1995;209:333.
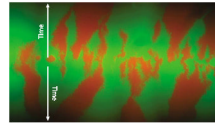
FIG. 1.
(Color online) Spatial segregation in an expanding microbial population. Different colors label different alleles. The Petri dish was inoculated with a well-mixed population occupying a narrow horizontal linear region between the arrows, which show the direction of the growth. As this population expands, it segregates into well-defined monoallelic domains. The colony is of the order 1 cm in height. Details of the experiment are presented in Hallatschek *et al.* (2007).
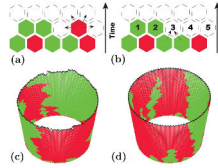
FIG. 2.

(Color online) An illustration of the two models of a growing front. (a) and (c) Illustration of the model with a rough undulating front, which is a natural result of an unconstrained two-dimensional growth. (b) and (d) Illustration of the model with a flat front, which is constrained to have no lateral undulations to simplify the analytical analysis. The blank hexagons represent empty sites, and different colors of the occupied hexagons represent different alleles. (a) The model of an undulating population front. The highlighted hexagon is a randomly chosen cell that can reproduce and deposit an identical offspring in any of its four empty nearest-neighbor sites (shown with arrows) with equal probability. (b) The model of a one-dimensional habitat, where each row represents a generation. Thus, each row is completed before moving on to the next one, so an empty site can be filled only by an offspring of one of its nearest neighbors in the previous generation (shown with arrows). Both (a) and (b) show the effects of genetic drift (sampling error) when, e.g., the second from the left cell in the bottom row leaves no offspring. Such events lead to coarsening seen in (c) and (d). (c) and (d) Single simulation runs for models in (a) and (b), respectively. A population of 100 cells was wrapped around a cylinder to illustrate periodic boundary conditions used in this paper. Note that in (d) the front is flat, whereas in (c) it is rough. This roughness affects some aspects of the shapes of the monoallelic domains shown in (c): a domain boundary followed from its lowest point to its highest point always goes up in (d) but, in (c), it sometimes goes down. As discussed by Hallatschek *et al.* (2007), domain walls are expected to wander more vigorously in (c) than in (d). Despite the apparent differences, both models exhibit the same qualitative behavior.
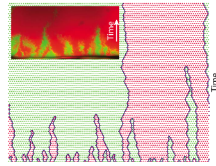
FIG. 3.
(Color online) Qualitative comparison of a gene segregation experiment from a linear inoculation (inset) and the simulation of a one-dimensional habitat. The experiment is analogous to the one depicted in Fig. 1.
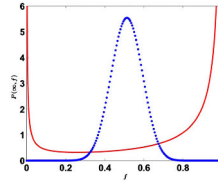
FIG. 4.

(Color online) The stationary distribution $P(\infty, f)$ in the presence of selection, mutation, and genetic drift [see Eqs. (19) and (20)]. The dotted line shows $P(\infty, f)$ for $s = \mathfrak{D}_g$ and $\mu_{12} = \mu_{21} = 10 \, \mathfrak{D}_g$, which corresponds to weak genetic drift, $\mu_{12}/D_g \gg 1$. The solid line shows $P(\infty, f)$ for $s = \mathfrak{D}_g$ and $\mu_{12} = \mu_{21} = 0.1 \mathfrak{D}_g$, which corresponds to strong genetic drift, $\mu_{12}/D_g \ll 1$. Note the difference in curvature between the two cases and the fact that the distribution is dominated by the central region in the weak genetic drift limit but by the tails in the opposite limit. The transition between these two regimes occurs when the mutation rates equal to $\mathfrak{D}_g/2$ and $P(\infty, f)$ diverges at $f=0$ and 1. Also note that the effect of natural selection is to bias the distribution toward $f=1$. As s increases, the maximum of the distribution shifts to the right for weak genetic drift, and the right tail of the distribution becomes more prominent for strong genetic drift.
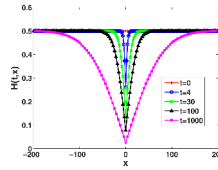
FIG. 5.

(Color online) Solutions of Eq. (33) at various times given $H(0, x) = \dfrac{1}{2}$. Time and distance are measured in units such that $D_s = 1$ and $D_g = 1$. Time increases from the top curves to the bottom curves. Note the statistical reflection symmetry, $H(t,x) = H(t,-x)$.
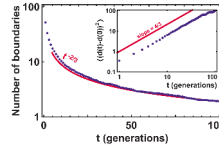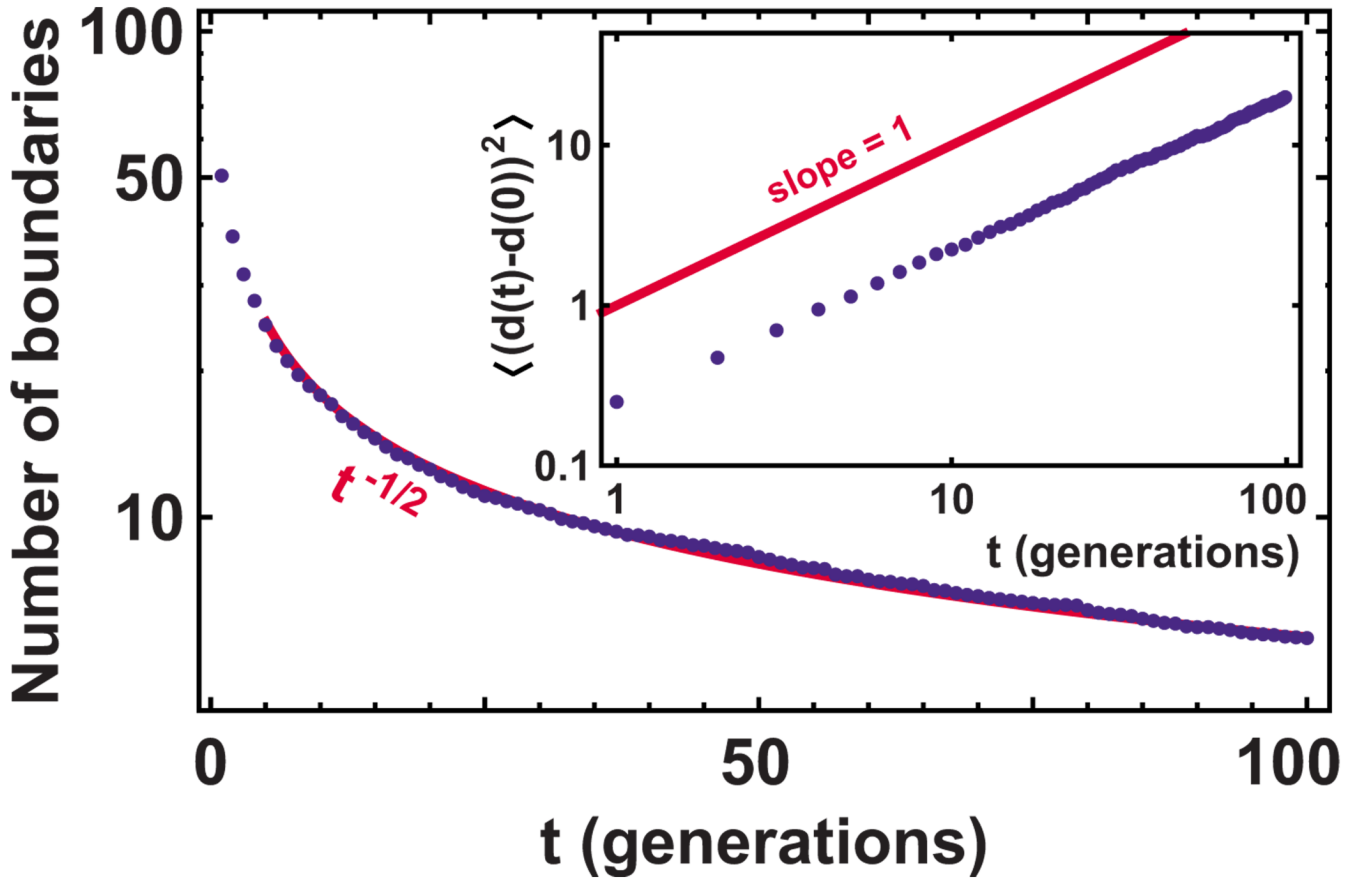
FIG. 6.

(Color online) The number of monoallelic domain boundaries as a function of time in the undulating-front model. The simulation of 100 demes averaged over 100 runs is plotted (dots), and the theoretically expected decay of the number of boundaries as $t^{-2/3}$ [see Saito and Müller-Krumbhaar (1995)] is plotted (solid line). Note that the agreement between the theory and the simulations is not expected during the transitory regime at early times. At $t=0$, each site is assigned either allele one or allele two with equal probability. Inset: Log-log plot of the mean-square displacement of the domain boundaries as a function of time in the same set of simulations as the main plot. The dots are the simulation data and the solid line is the expected slope according to Saito and Müller-Krumbhaar (1995).

FIG. 7.

(Color online) The number of monoallelic domain boundaries as a function of time in the flat-front model. The simulation of 100 demes averaged over 100 runs is plotted (dots), and the theoretically expected decay of the number of boundaries as $t^{-1/2}$ (see Hallatschek and Nelson, 2010) is plotted (solid line). Note that the agreement between the theory and the simulations is not expected during the transitory regime at early times. At t=0, each site is assigned either allele one or allele two with equal probability. Inset: Log-log plot of the mean-square displacement of the domain boundaries as a function of time in the same set of simulations as the main plot. The dots are the simulation data and the solid line is the expected slope according to Bramson and Lebowitz (1991) and Hallatschek and Nelson (2010) and Eq. (38).
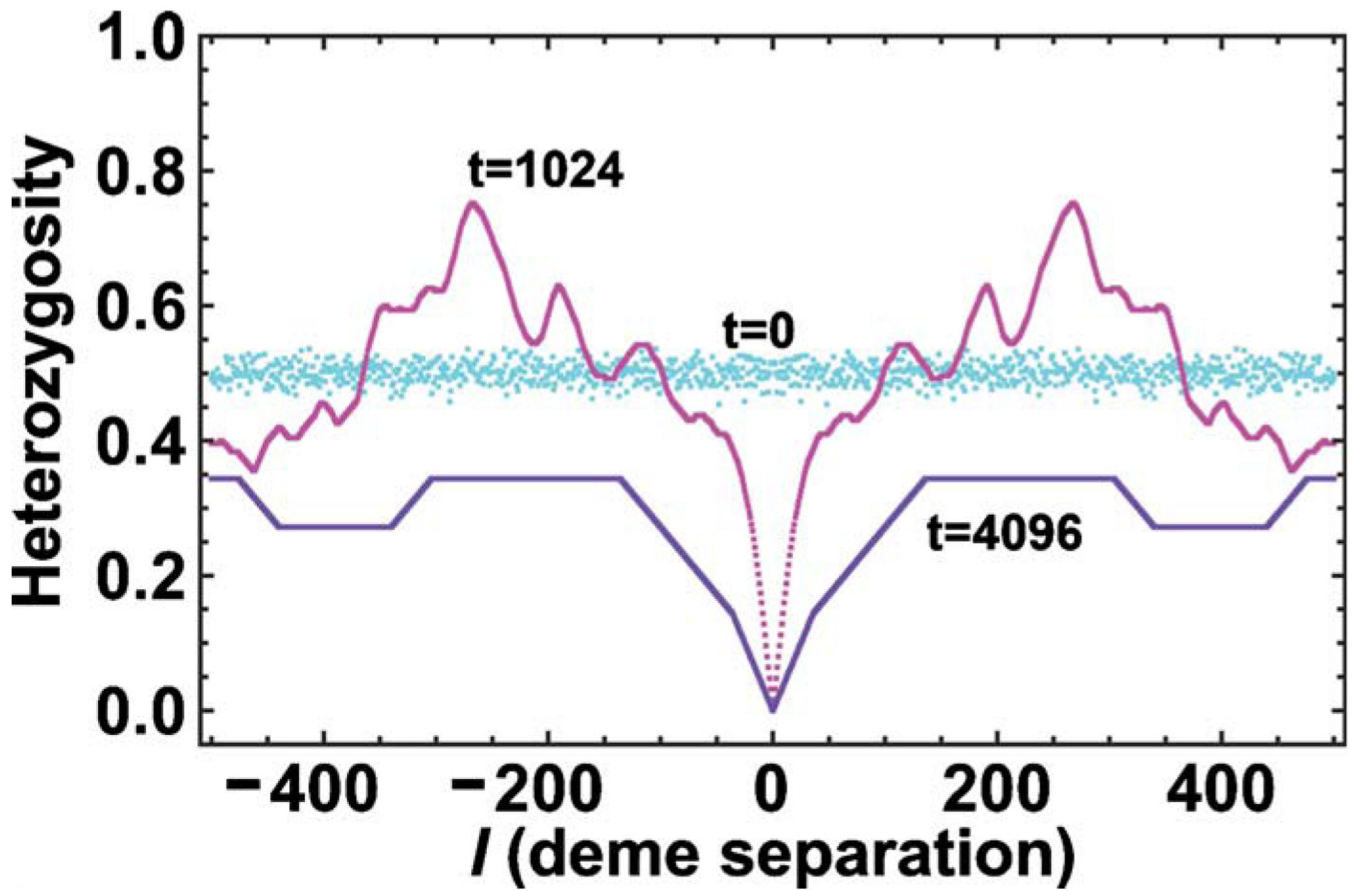
FIG. 8.

(Color online) A single run of the flat-front model with 1000 organisms. At t=0, each site is assigned either allele one or allele two with equal probability. The spatially averaged heterozygosity [defined in the sense of Eq. (C1) but without taking the limit $L\to\infty$] for three times measured in generations: t=0, 1024, and 4096. The separation l is the shortest distance between two points around the cylinder, and we take the clockwise direction to be positive. At inoculation, the heterozygosity fluctuates around 1/2 since the two alleles have equal probabilities of occupying any site. The only exception is the site l=0, where the heterozygosity is zero automatically because we only allow a single microorganism per site. After 1024 generations, short-range correlations are clearly visible and after 4096 generations, one can relate the abrupt changes in the slope of $H(t,l)$ to the sizes of the sectors in the population (not shown). The wiggles are eliminated when averaged over many realizations, as shown in Fig. 9. Note that the curve for t=4096τ_g lies completely below 1/2 because, at this time, the relative fraction of the alleles deviates significantly from the initial 50:50 ratio due to genetic drift.
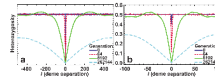
FIG. 9.

(Color online) The effects of coarse graining on the time evolution of the spatial heterozygosity $H(t,l)$ averaged over 100 realizations of the flat-front model with a 1000 organisms. At t=0, each site is assigned either allele one or allele two with equal probability. (a) Each deme hosts only one organism. Consequently, the heterozygosity at l=0 is zero at all times. (b) The same set of simulations, but the organisms have now been grouped into demes of size 5 for the purpose of calculating $H(t,l)$. (b) Qualitative agreement with the solution of stepping stone model displayed in Fig. 5, and (a) outside the region around l=0. Note that, unlike the calculation presented in Fig. 5, this simulation was run sufficiently long to show the effects of the boundary conditions.
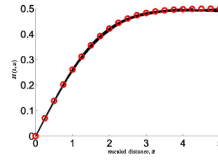
FIG. 10.
(Color online) Comparison between the analytical prediction for the limiting shape of $H(\infty,\bar{x})$ and the simulations of the flat-front model. The continuous curve (black) is formed by the data points representing $H(t,x)$ for several times between $t=2\times10^4$ and 4×10^5 plotted in the rescaled coordinates $\bar{x}$ and the circles represent the theoretical prediction of the limiting shape of the average spatial heterozygosity [see Eq. (37)]. The data are obtained in a simulation of 3200 individuals for 4×10^5 generations with averaging over 500 realizations. At $t=0$, each site is assigned either allele one or allele two with equal probability.
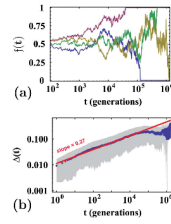
FIG. 11.

(Color online) Genetic drift in a finite population. At t=0, each site is assigned either allele one or allele two with equal probability. (a) The total fraction of allele one $f(t)$ vs time in four single runs of the neutral model with a flat front. Here L=1000, and there are no mutations. (b) The average standard deviation of the frequency of allele one $\Delta(t)$ is obtained from 200 realizations of the simulations described in (a). The solid line shows the best power-law fit, and the slope is close to the exponent expected from Eq. (44). The gray area encloses the points within one standard deviation from the mean.
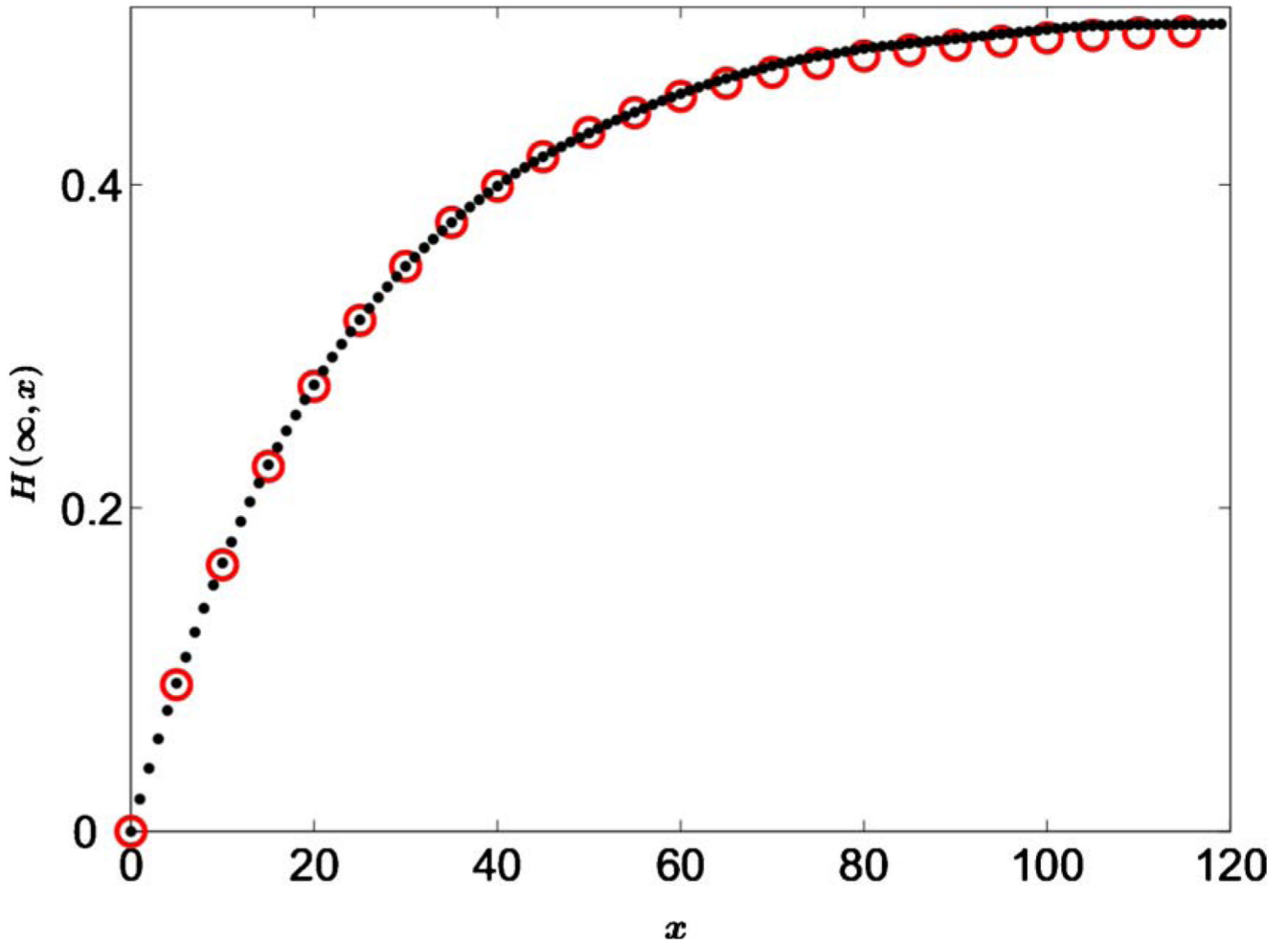
FIG. 12.
(Color online) Equilibrium between mutation and genetic drift in the absence of selection. Comparison between the analytical prediction for the steady-state heterozygosity $H(\infty,x)$ and the simulations with $\tilde{\mu}_{12}=\tilde{\mu}_{21}=10^{-4}$. The black dots represent the results of the simulation and the circles represent the best fit of theoretical result given by Eq. (48) to the data. Here only D_g is a fitting parameter; the values of D_s, μ_{12}, and μ_{21} follow from the correspondence between the discrete and continuum models. The data are obtained in a simulation of 3200 individuals for 2×10^5 generations with averaging over 100 realizations. At $t=0$, each site is assigned either allele one or allele two with equal probability.
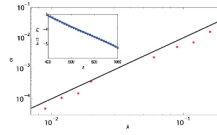
FIG. 13.

(Color online) The effective extinction rate α vs $\tilde{s}$ in the limit of weak selection. The dots are the data from the simulation and the black line has the slope equal to 2, which is the expected slope from Eq. (58). The data support $\alpha \propto \tilde{s}^2$. The values of α are obtained from graphs like the one shown in the inset. Inset: $\ln(1-f)$ vs t for $\tilde{s}=0.12$. The circles are the actual data points and the line is the best least-squares linear fit. The simulation confirms exponentially fast fixation. The data are obtained in a simulation of 1600 individuals for 6000 generations with averaging over 100 realizations. At $t=0$, each site is assigned either allele one or allele two with equal probability.
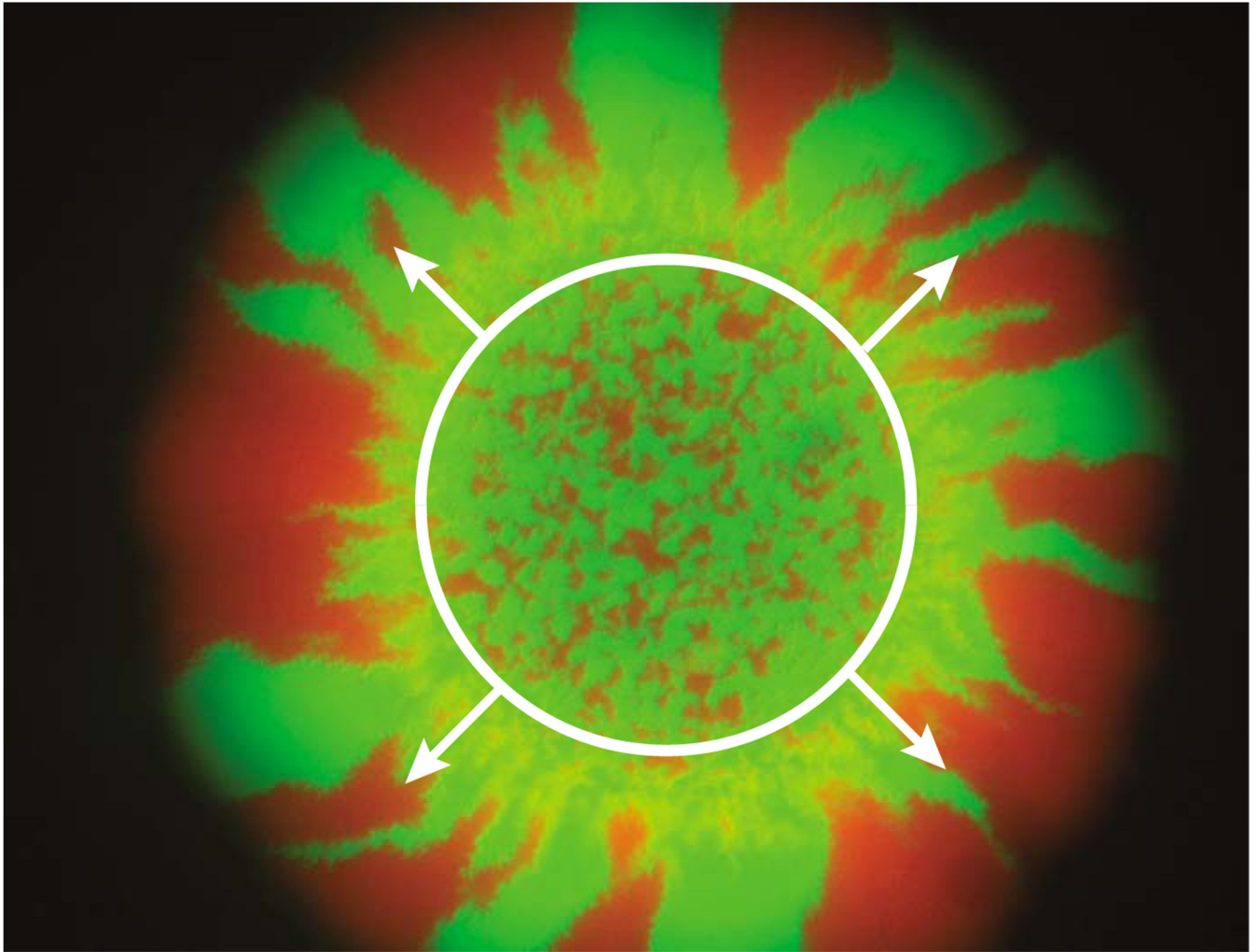
FIG. 14.
(Color online) Spatial segregation in an expanding circular bacterial colony of *E. coli*. Different colors label different alleles. The Petri dish was inoculated with a well-mixed population occupying the circled region of the colony, leading to many small domains in the central "homeland." As this population expands (shown with arrows), it segregates into well-defined monoallelic domains, which coalesce at early times but seem to stop coalescing in the final stages of the experiment. Note that the boundaries between the domains are biased to move away from each other due to inflation, in addition to their diffusive random-walk-like motion. Details of the experiment are presented by Hallatschek *et al.* (2007).
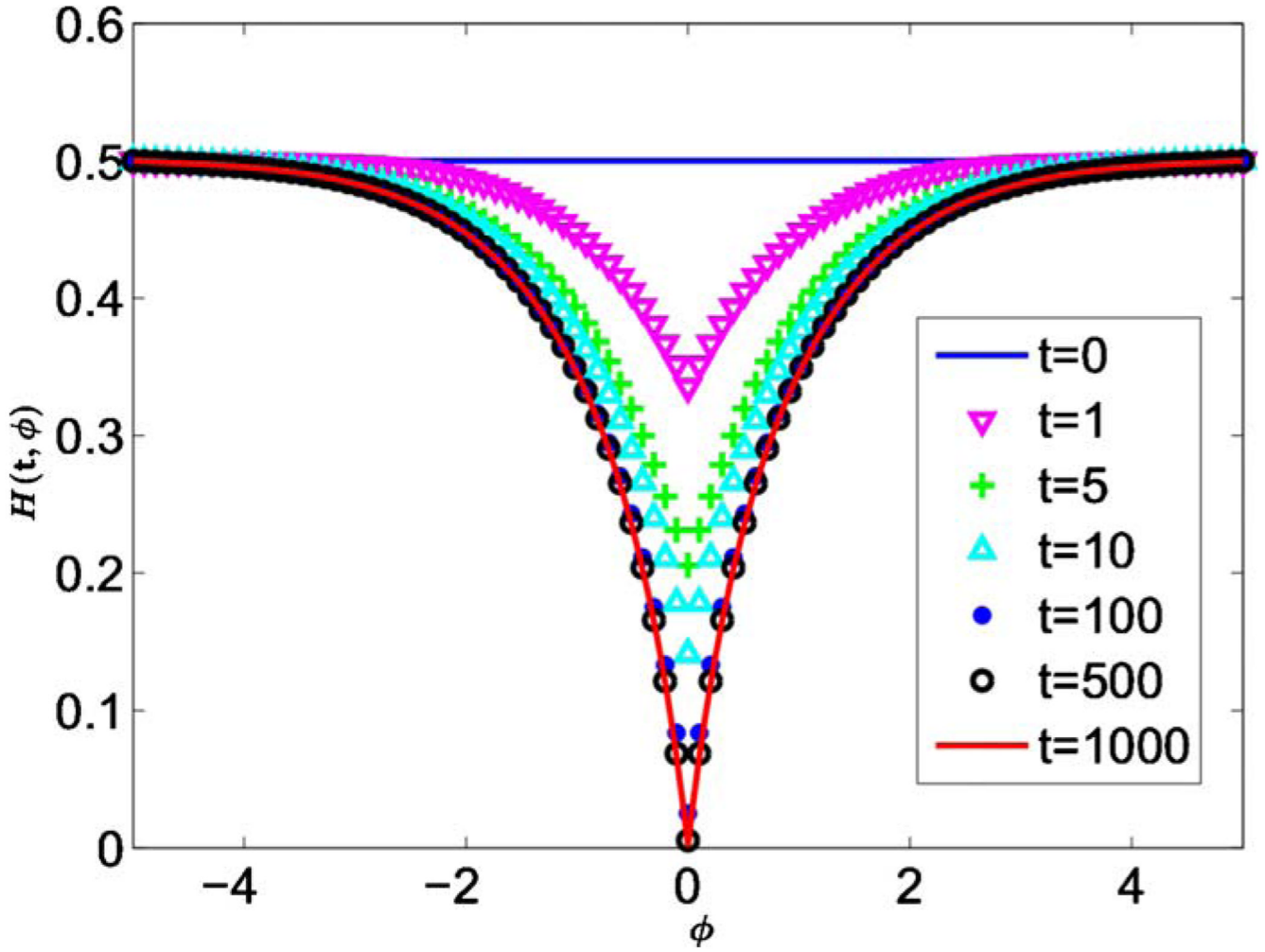
FIG. 15.
(Color online) Solutions of Eq. (63) with σ=1 at various rescaled times t given random initial

conditions on the circle bounding the homeland, $H(0,\phi)=\frac{1}{2}$. Time increases from the top curves to the bottom curves. Note that there is no observable difference between $H(500,\phi)$ and $H(1000,\phi)$ because $H(t,\varphi)$ reaches a nontrivial limit shape as t→∞.
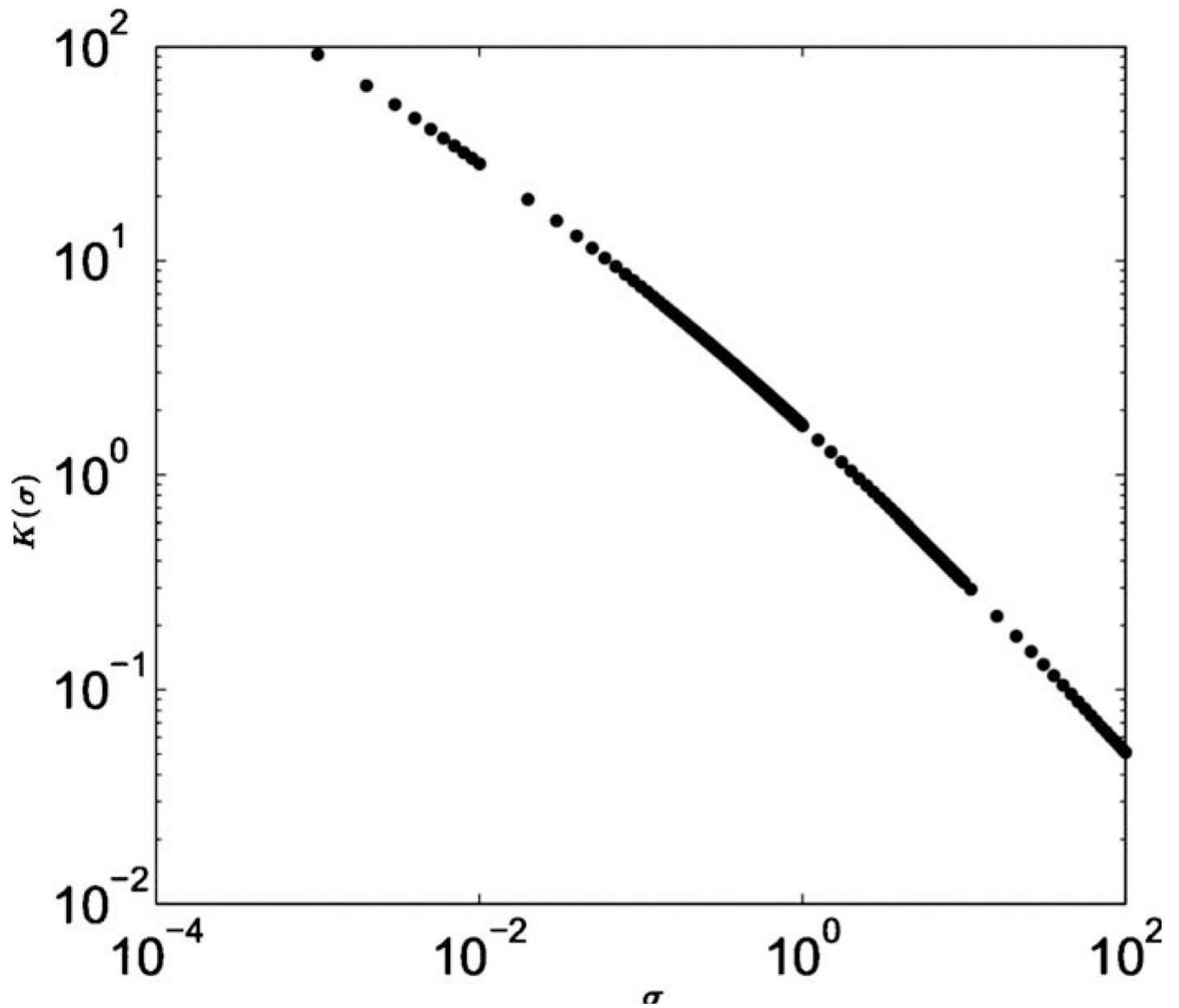
FIG. 16.
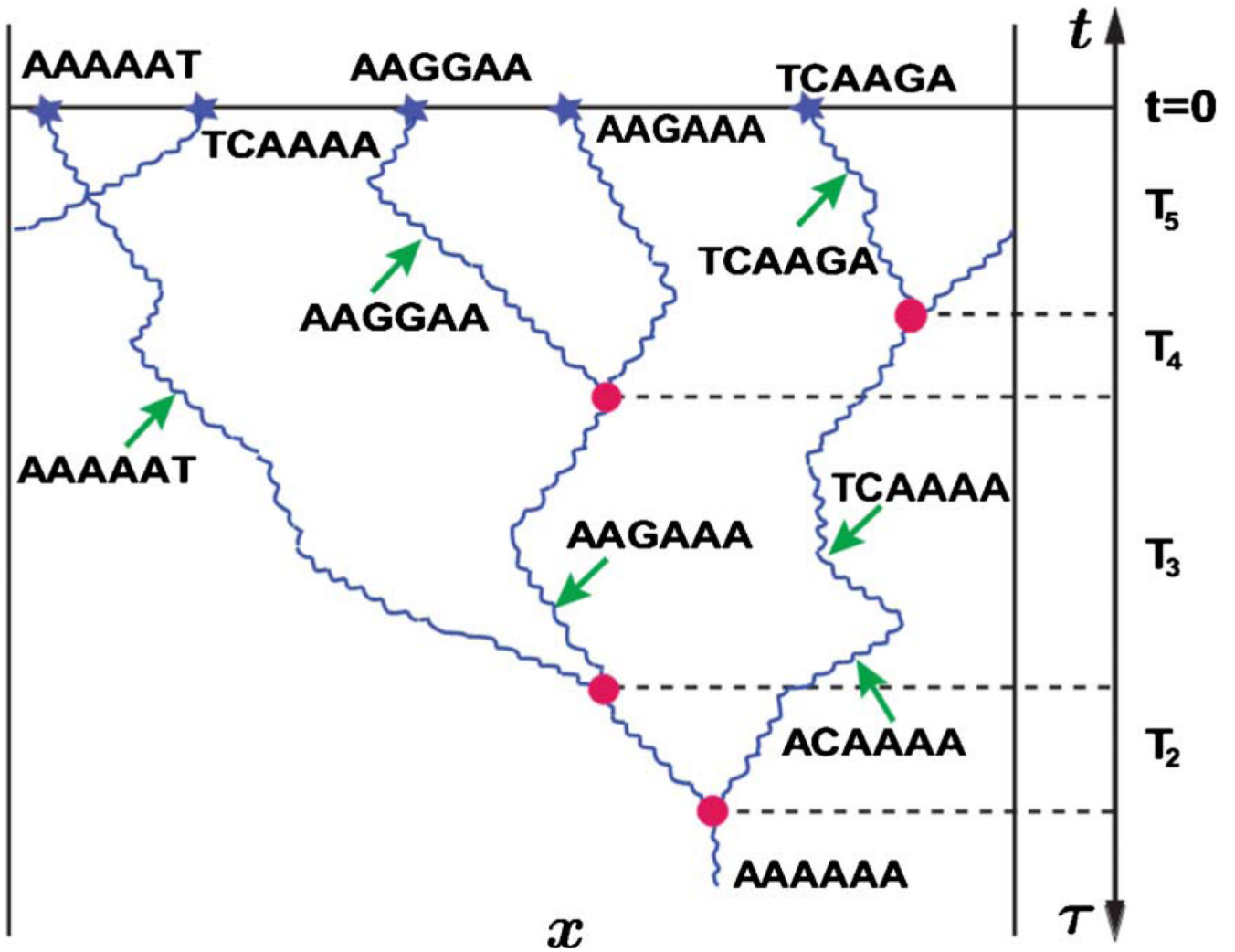A plot of $K(\sigma)$ for Eq. (71) from a numerical solution of Eq. (63).

FIG. 17.
(Color online) An illustration of the backward-in-time dynamics of the ancestral lineages in a one-dimensional habitat with periodic boundary conditions. Five organisms (i.e., *n*=5) are sampled from the much larger population at *t*=0 and their DNA is sequenced. We do not display the sites that are identical for all organisms, which are usually the majority of the sequenced sites, i.e., only the segregating sites are shown. For illustration purposes, we also assumed that all samples differ in at least one nucleotide, but in experiments one often finds organisms that have identical sequences. We trace the spatial diffusion and coalescence of the lineages backward in time until they merge into a single lineage of the common ancestor of the whole sample. The coalescence events are denoted by circles and the mutations are denoted by arrows and the resulting mutated sequences. Note that lineages may cross without coalescing as shown in the top left corner. The ancestral process shown here satisfies the infinite site model and illustrates the fact that the more genetically similar the lineages are the more likely they are to have a common ancestor in the recent past. Most genetic inference methods rely on this relationship as shown in the text.
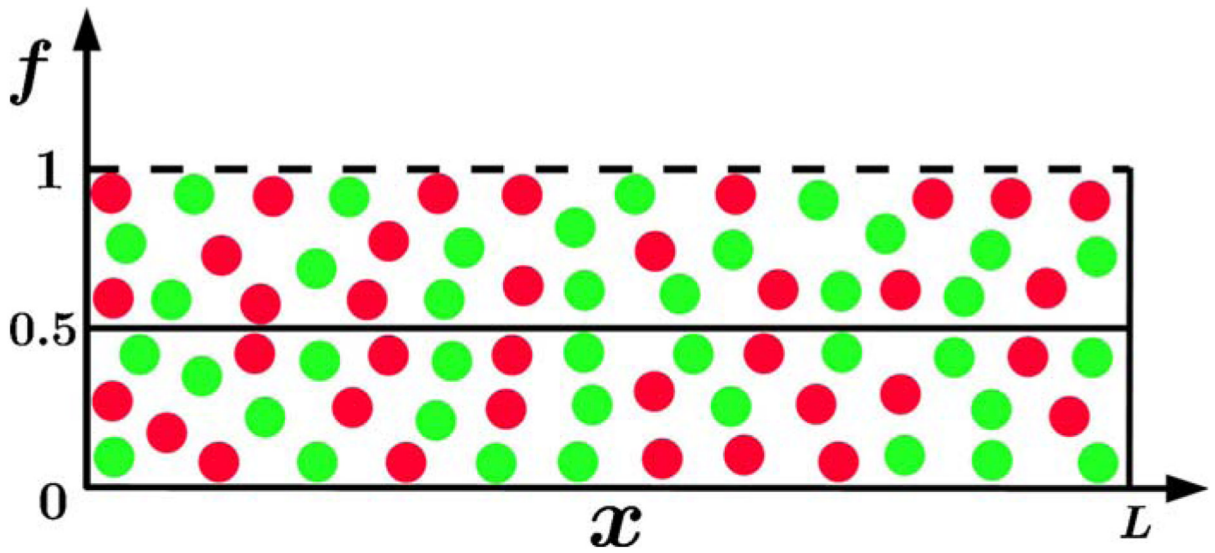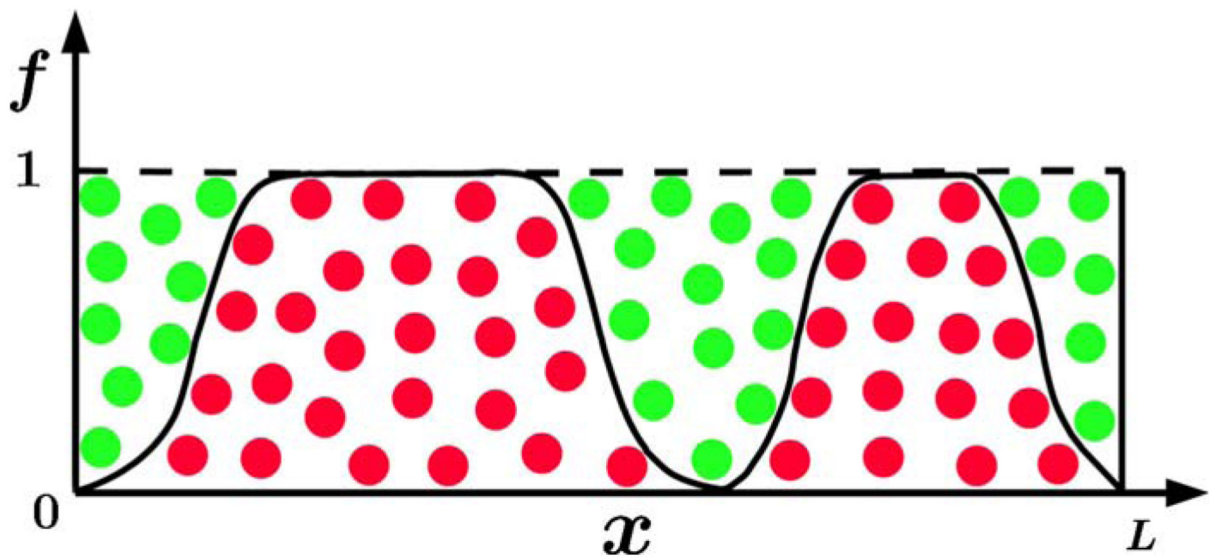
FIG. 18.
(Color online) Illustration of spinodal-decomposition-like genetic demixing in a one-dimensional population. (a) Initially well-mixed population with colors labeling different genotypes. (b) The same population, several generations later. The frequency of one of the alleles is now oscillating between 0 and 1 because the population segregates into monoallelic domains.
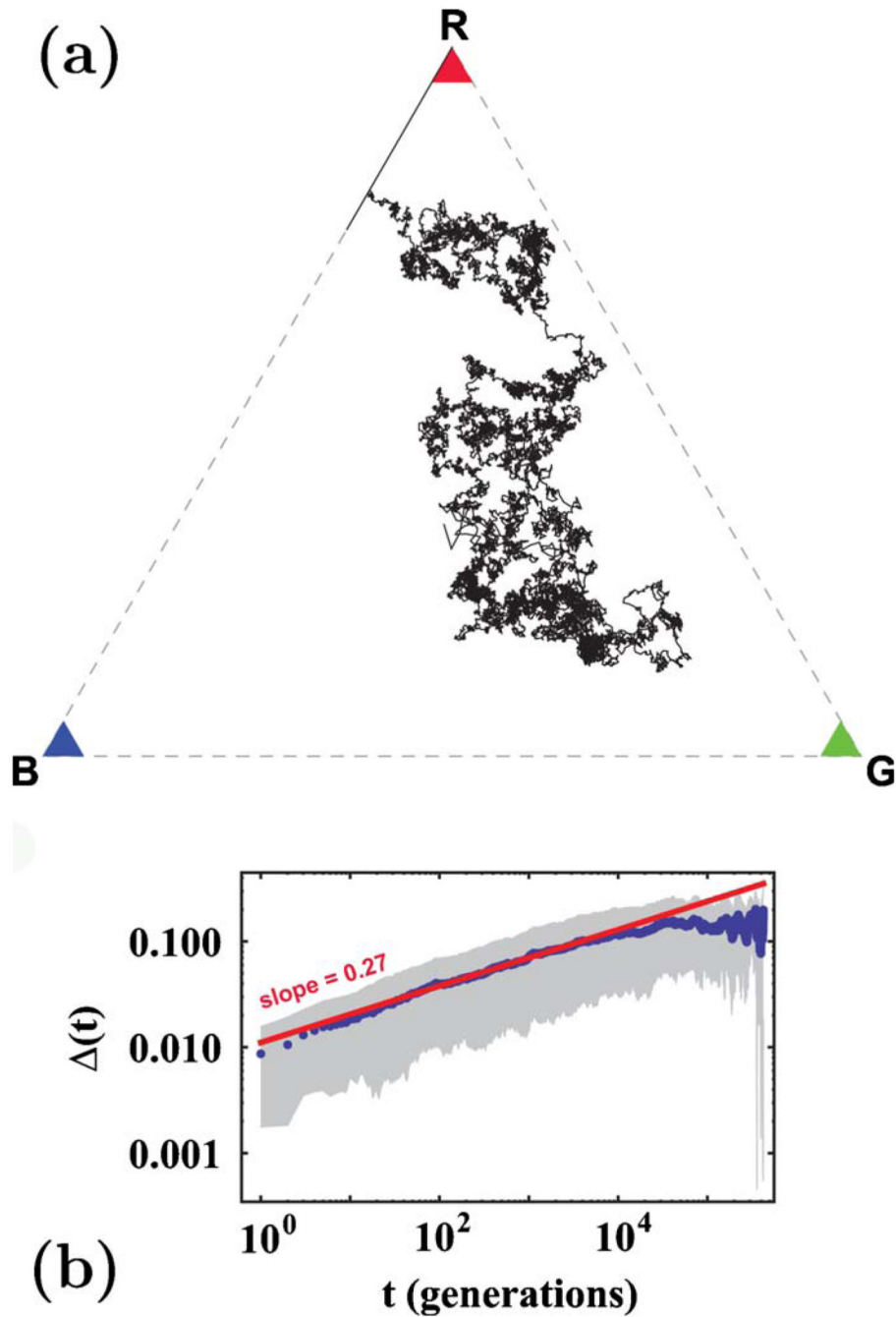
FIG. 19.

(Color online) Genetic drift during a linear range expansion in the flat-front model with three alleles. (a) The genetic composition of the population $[f_1(t), f_2(t), f_3(t)]$ projected on the plane $\Sigma_{i=1}^{3} f_i(t)=1$ in a single run of the neutral three-allele model with a flat front. The population is finite, $L=1000$, and there are no mutations. (b) The average standard deviation of the frequency of allele one $\Delta(t)$ is obtained from 200 realizations of the simulations described in (a). The solid line shows the best power-law fit, and the slope is close to the exponent expected from Eq. (44). The gray area encloses the points within one standard deviation from the mean. At

t=0, each site is assigned either allele one or allele two with equal probability, which corresponds to the center of the triangle in (a).
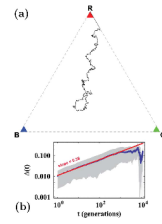
FIG. 20.

(Color online) Genetic drift during a linear range expansion in the undulating-front model with three alleles. (a) The genetic composition of the population $[f_1(t), f_2(t), f_3(t)]$ projected on the plane $\Sigma_{i=1}^{3} f_i(t)=1$ in a single run of the neutral three-allele model with an undulating front. The population is finite, $L=1000$, and there are no mutations. (b) The average standard deviation of the frequency of allele one $\Delta(t)$ is obtained from 200 realizations of the simulations described in (a). The solid line shows the best power-law fit and the slope is close to the exponent expected from Eq. (E8). The gray area encloses the points within one standard deviation from the mean. At $t=0$, each site is assigned either allele one or allele two with equal probability, which corresponds to the center of the triangle in (a).