Spatial soft sweeps: patterns of adaptation in populations with long-range dispersal

Jayson Paulose,[1] Joachim Hermisson,[2] and Oskar Hallatschek[1]

[1]*Departments of Physics and Integrative Biology,*
University of California, Berkeley,
USA

[2]*Department of Mathematics and Max F. Perutz Laboratories,*
University of Vienna,
Austria

Adaptation in extended populations often occurs through multiple independent mutations responding in parallel to a common selection pressure. As the mutations spread concurrently through the population, they leave behind characteristic patterns of polymorphism near selected loci—so-called soft sweeps—which remain visible after adaptation is complete. These patterns are well-understood in two limits of the spreading dynamics of beneficial mutations: the panmictic case with complete absence of spatial structure, and spreading via short-ranged or diffusive dispersal events, which tessellates space into distinct compact regions each descended from a unique mutation. However, spreading behaviour in most natural populations is not exclusively panmictic or diffusive, but incorporates both short-range and long-range dispersal events. Here, we characterize the spatial patterns of soft sweeps driven by dispersal events whose jump distances are broadly distributed, using lattice-based simulations and scaling arguments. We find that mutant clones adopt a distinctive structure consisting of compact cores surrounded by fragmented "haloes" which mingle with haloes from other clones. As long-range dispersal becomes more prominent, the progression from diffusive to panmictic behaviour is marked by two transitions separating regimes with differing relative sizes of halo to core. We analyze the implications of the core-halo structure for the statistics of soft sweep detection in small genomic samples from the population, and find opposing effects of long-range dispersal on the expected diversity in global samples compared to local samples from geographic subregions of the range. We also discuss consequences of the standing genetic variation induced by the soft sweep on future adaptation and mixing.

I. INTRODUCTION

Rare beneficial alleles can rapidly increase their frequency in a population in response to a new selective pressure. When adaptation is limited by the availability of mutations, a single beneficial mutation may sweep through the entire population in the classical scenario of a "hard sweep". However, populations may exploit a high availability of beneficial mutations due to standing variation, recurrent new mutation, or recurrent migration [1, 2, 3, 4, 5] to respond quickly to new selection pressures. As a result, multiple adaptive alleles may sweep through the population concurrently, leaving genealogical signatures that distinguish them from hard sweeps. Such events are are termed *soft sweeps*. Soft sweeps are now known to be frequent and perhaps dominant in many species [6, 7]. Well-studied examples in humans include multiple origins for the sickle cell trait which confers resistance to malaria [8], and of lactose tolerance within and among geographically separated human populations [9, 10].

Soft sweeps rely on a supply of beneficial mutations on distinct genetic backgrounds, which has two main origins. One is when selection acts on an allele which has multiple copies in the population due to standing genetic variation—a likely source of soft sweeps when the potentially beneficial alleles were neutral or only mildly deleterious before the appearance of the selective pressure [3]. In this work, we consider the other important scenario of soft sweeps due to recurrent new mutations which arise after the onset of the selection pressure. Soft sweeps become likely when the time taken for an established mutation to fix in the entire population is long compared to the expected time for additional new mutations to arise and establish. In a panmictic population, the relative rate of the two processes is set primarily by the rate at which new mutations enter the population as a whole [5].

Most examples of soft sweeps in nature, however, show patterns consistent with arising in a geographically structured rather than a panmictic population [7]. Spatial structure promotes soft sweeps [11]: when lineages spread diffusively (i.e. when offspring travel a restricted distance between local fixation events), a beneficial mutation advances as a constant-speed wave expanding outward from the point of origin, much slower than the logistic growth expected in a well-mixed population. Therefore, fixation is slowed down by the time taken for genetic information to spread through the range, making multi-origin sweeps more likely. However, the detection of such a *spatial soft sweep* crucially depends on the sampling strategy: the wavelike advance of distinct alleles divides up the range into regions within which a single allele is predominant.

If genetic samples are only taken from a small region within the species' range, the sweep may appear hard in the local sample even if it was soft in the global range.

Between the two limits of wavelike spreading and panmictic adaptation lies a broad range of spreading behaviour driven by dispersal events that are neither local nor global. Many organisms spread through long-range jumps drawn from a probability distribution of dispersal distances (dispersal kernel) that does not have a hard cutoff in distance but instead allows large, albeit rare, dispersal events that may span a significant fraction of the population range [12, 13]. A recent compilation of plant dispersal studies showed that such so-called "fat-tailed" kernels provided a good statistical description for a majority of data sets surveyed [14]. Fat-tailed dispersal kernels accelerate the growth of mutant clones, whose sizes grow faster-than-linearly with time and ultimately overtake growth driven by a constant-speed wave [12, 15]. Besides changing the rate at which beneficial alleles take over the population, long-range dispersal also breaks up the wave of advance [16]: the original clone produces geographically separated satellites which strongly influence the spatial structure of regions taken over by distinct alleles.

Despite its prominence in empirically measured dispersal behaviour and its strong effects on mutant clone structure and dynamics, the impact of long-range dispersal on soft sweeps is poorly understood. Past work incorporating fat-tailed dispersal kernels in spatial soft sweeps [11] relied on deterministic approximations of the jump-driven spreading behaviour of a single beneficial allele [12]. However, recent analysis has shown that deterministic approaches are accurate only in the two extreme limits of local (i.e. wavelike) and global (i.e. panmictic) spreading, and break down over the entire regime of intermediate long-range dispersal [17]. Away from the limiting cases, the correct long-time spreading dynamics is obtained only by explicitly including rare stochastic events which drive the population growth. Deterministic approaches also do not account for the disconnected satellite structure, which has consequences for soft sweep detection in local samples.

Here, we study soft sweeps driven by the stochastic spreading of alleles via long-range dispersal. We perform simulations of spatial soft sweeps in which beneficial alleles spread via fat-tailed dispersal kernels which fall off as a power law with distance, focusing on the regime in which multiple alleles arise concurrently. We find that long-range dispersal gives rise to distinctive spatial patterns in the distribution of mutant clones. In particular, when dispersal is sufficiently long-ranged, mutant clones are discontiguous in space, in contrast to the compact clones expected from wavelike spreading models. We identify qualitatively different regimes for spatial soft sweep patterns depending on the tail of the jump distribution. We show that analytical results for the stochas-

tic jump-driven growth of a *single* allele [17], combined with a mutation-growth balance relevant for spatial soft sweeps [11], allow us to predict the range sizes beyond which soft sweeps become likely. We also analyze how stochastic aspects of growth of independent alleles, particularly the establishment of satellites disconnected from the initial expanding clone, influence the statistics of observing soft sweeps in a small sample from the large population. We find that long-range dispersal has contrasting effects on the likelihood of soft sweep detection, depending on whether the population is sampled locally or globally.

II. MODEL

We consider a haploid population that lives in a d-dimensional habitat consisting of demes that are arranged on a square lattice. Local resource limitation constrains the deme population to a fixed size $\hat{n}$, assumed to be the same for all demes. Denoting the linear dimension of the lattice as L, the total population size is $N = L^d \hat{n}$. The population is panmictic within each deme. With a rate m per generation, individuals migrate from one deme to another. For each dispersal event, the distance r to the target deme is chosen from a probability distribution with weight $J(r)$, appropriately discretized, with the normalization $\int_1^\infty J(r) \, d^d r = 1$. The function $J(r)$ is called the jump kernel. New mutations arise in all demes at a constant rate u per individual per generation. Each new mutation is distinguishable from previous mutations (e.g. due to different genomic backgrounds), but all mutations confer the same selective advantage s. Back mutations are ignored.

To focus on the effects of long-range dispersal over local dynamics, we now impose a set of bounds on the individual-based parameters following Ref. 11. In particular, we consider only situations where $s\hat{n} \gg 1$; $u\hat{n} \ll 1$; $m\hat{n} \ll 1$ (strong selection, and low mutation and migration rates at the deme level). Mutations are also assumed to be fully redundant, i.e. a second mutation confers no additional advantage. The strong selection condition implies that genetic drift within a deme is irrelevant relative to selection: a new mutation, upon surviving stochastic drift and fixing within a deme (which happens with probability $2s$) cannot be subsequently lost due to genetic drift. The bounds on mutation and migration rates meanwhile imply that the fixation dynamics of a beneficial mutation within a deme is fast compared to the dynamics of mutation within a deme or of migration among demes. The time to fixation of a beneficial allele from a single mutant individual in the deme, $\log(\hat{n}s)/s$, is a few times $1/s$. When $u\hat{n} \ll 1$ and $m\hat{n} \ll 1$, the fixation time scale is much shorter than the establishment time scales of new alleles arising due to mutation or migration, which are $(2sm\hat{n})^{-1}$ and $(2su\hat{n})^{-1}$ respectively. Therefore, the
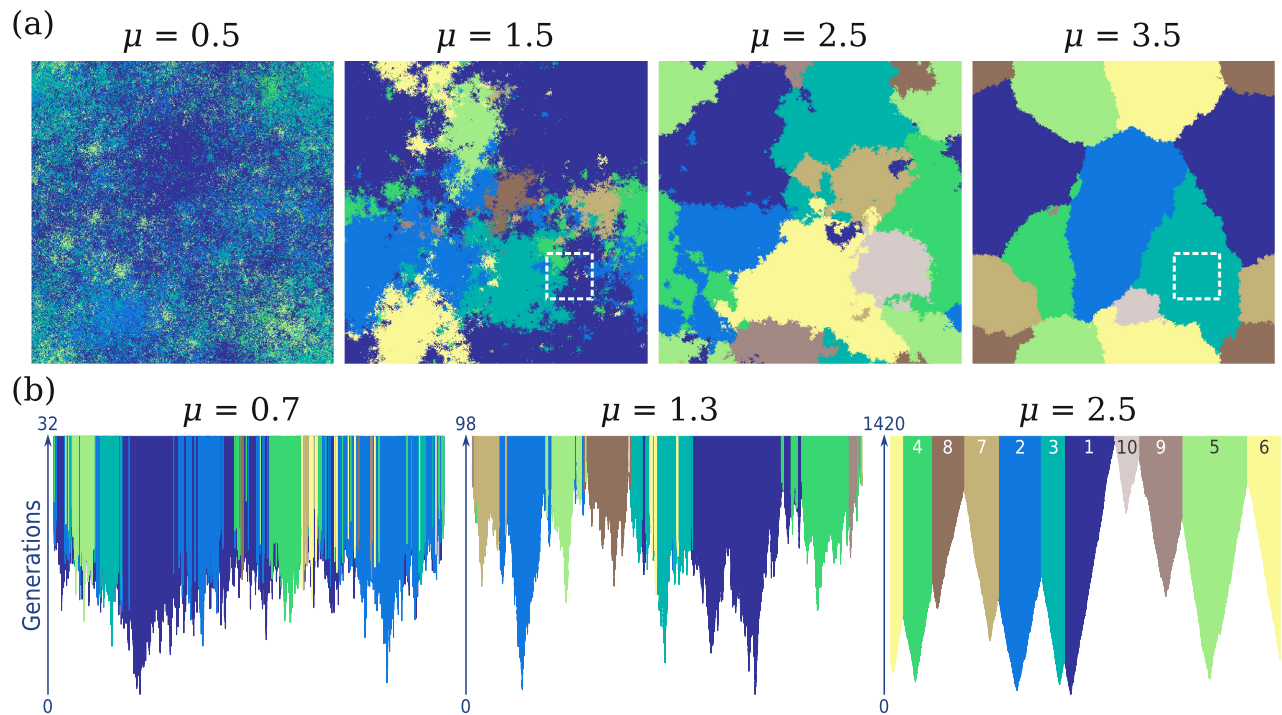
FIG. 1 **Spatial soft sweep patterns with the same number of distinct alleles vary strongly with jump kernel.** **a,** Final states of 2D simulations on a lattice of size 512×512, for a range of values of the kernel exponent μ. Each pixel corresponds to a deme, and is coloured according to the identity of the allele occupying that deme; demes belonging to the same mutant clone share the same colour. Simulations were chosen which have ten unique alleles in the final state; colours reflect the temporal order of the originating mutations as labeled in the lower right panel. Rescaled mutation rates are 3×10^{-6}, 10^{-6}, 10^{-6}, and 10^{-7} for kernel exponents 0.4, 1.5, 2.5, and 3.5 respectively. The subrange highlighted by a dashed line contains six distinct alleles for $\mu = 1.5$ but only one allele for $\mu = 3.5$. **b,** Full time-evolution of 1D simulations with $L = 16384$ for three kernel exponents, chosen so that the final state has ten unique alleles. Each vertical slice displays the lattice state at a particular time (measured in generations), starting from an empty lattice (white) and continuing until all sites are filled and the sweep is completed. The rescaled mutation rates are 3×10^{-5}, 7×10^{-6}, and 7×10^{-7} respectively from left to right. In the last panel, the colours are labelled according to the order of appearance of the originating mutation; the same order is shared among all panels in the figure.

first beneficial allele that establishes in a deme, whether through mutation or migration, fixes in that deme without interference from other alleles. Furthermore, the assumption of mutual redundancy means that subsequent mutations that arrive after the first fixation event also have no effect. As a result, the first beneficial allele that establishes in a deme excludes any subsequent ones—a situation termed allelic exclusion [11].

Taken together, these assumptions lead to a simplified model that ignores the microscopic dynamics of mutations within demes. For each deme, we keep track of a single quantity: the allelic identity (whether wildtype or one of the unique mutants that has arisen) that has fixed in the deme. At the deme level, new mutations fix within wildtype demes at the rate $2s\hat{n}u$, and each mutated deme sends out migrants at rate $2s\hat{n}m$ with the target deme selected according to the dispersal kernel $J(r)$ (the rates explicitly include the fixation probability $2s$ of a single mutant in a wildtype deme). The first successful mutant to arrive at a wildtype deme, whether through mutation

or migration, immediately fixes within that deme. The state of the deme thereafter is left unchanged by mutation or migration events, because of allelic excusion.

When time is measured in units of the expected interval $(2s\hat{n}m)^{-1}$ between successive dispersal events per deme, the reduced model is characterized by just three quantities: L; $J(r)$; and the per-deme rate of mutations per dispersal attempt $\tilde{u} \equiv 2s\hat{n}u/(2s\hat{n}m) = u/m$, which we call the rescaled mutation rate of our model. Simulations are begun with a lattice of demes of size L^d all occupied by the wildtype. Each discrete simulation step is either a mutation or an attempted migration event, with the relative rates determined by $\tilde{u}$ and the fraction of wildtype sites at that step. Mutation events flip a randomly-selected wildtype deme into a new allelic identity. Migration events first pick a mutated origin and then pick a target deme according to the jump kernel. If the target site is wildtype, it acquires the allelic identity of the origin; otherwise the migration is unsuccessful. Simulations are run until all demes have been taken over

by mutants.

The fat-tailed jump kernels we use are of the form $J(r) = \mu r^{-(d+\mu)}$. The exponent μ characterizes the "heaviness" of the tail of the distribution. We have chosen power-law kernels because they span a dramatic range of outcomes that connect the limiting cases of well-mixed and wavelike growth upon varying a single parameter. The growth dynamics of more general fat-tailed kernels in the stochastic regime of interest (i.e. driven by rare long jumps) are largely determined by the power-law falloff of the tail, and details of the dispersal kernel at shorter length scales are less consequential. Therefore, our qualitative results should extend to kernels sharing the same power law behaviour of the tail, provided the typical clones are large enough so that rare jumps picked from the tail of the distribution become relevant. The underlying analysis leading to the results is even more general, and can be applied to any jump kernel that leads to faster-than-linear growth in the extent of an individual clone with time.

The output of a simulation at a given set of L, μ and $\tilde{u}$ values is the final configuration of mutants, which can be grouped into distinct clones of the same allelic identity. Note that we have ignored the post-sweep mixing of alleles which are now relatively neutral to each other due to migration; this is justified by the separation of time scales between fast fixation and slow neutral migration [11]. In addition, although we restrict ourselves to weak mutation and migration at the deme level, the population-level mutation and migration rates Nu, Nm are typically large which allows for soft sweeps with strong migration effects.

III. RESULTS

A. Jump-driven growth and the core-halo structure of mutant clones

Some typical outcomes of the simulation model are shown in Fig. 1 for both two-dimensional (2D) and one-dimensional (1D) ranges. To emphasize variations in the spatial patterns for the same average clone size, simulations were chosen in which the final state has exactly ten unique alleles; this required varying the rescaled mutation rate as μ was increased. This feature, which is tied to the slower growth of individual clones apparent in the space-time plots of Fig. 1(b), is explored in depth in Section III.B.

In both 2D and 1D, the spatial soft sweep patterns of Fig. 1 display systematic differences as the kernel exponent is varied. Clones are increasingly fragmented as the kernel exponent is reduced; i.e. as long-range dispersal becomes more prominent. At the highest value of μ in each dimension, the range is divided into compact, essentially contiguous domains each of which shares a unique mutational origin. As the kernel exponent μ is reduced,

the contiguous structure of clones is lost as they break up into disconnected clusters of demes. For most clones, however, a compact region can still be identified in the range which is dominated by that clone (i.e. the particular allele reaches a high occupancy that is roughly uniform within the region but begins to fall with distance outside it) and in turn contains a significant fraction of the clone. We call this region the *core* of the clone. The remainder of the clone is distributed among many satellite clusters which produce local regions of high occupancy for a particular clone. The satellites become increasingly sparse and smaller in size as we move away from the core. For the broadest kernels ($\mu = 0.5$ in 2D and $\mu = 0.7$ in 1D), the clone also includes isolated demes which do not form a cluster but are embedded within cores and satellite clusters of a different allele. We term the collection of satellites and isolated demes the *halo* region surroundng the core of the clone. The spatial extent of the clone including the halo (i.e. the diameter of the region which encompasses the entire clone) can be many times the extent of the core alone, and increases relative to the core extent as μ is reduced. (We will use "extent" to refer to linear dimensions, and "mass" or "size" to refer to the number of demes.)

The space-time evolution displayed in 1D reveals the role of jump-driven growth in producing the observed spatial structures. At $\mu = 2.5$, the growth of clones appears nearly deterministic, with fronts separating mutant from wildtype advancing outwards from the originating mutations at near-constant velocity. These fronts are arrested when they encounter advancing fronts of other clones, leaving behind a tessellation of the range into contiguous clones. By contrast, at the lower values $\mu = 1.3$ and 0.7, the stochastic nature of jump-driven growth becomes apparent. Clones advance through long-distance dispersal events, which seed satellite clusters that may merge with each other before the sweep is complete. For all except the smallest clones, the originating mutation is surrounded by a region which is dominated by that particular allele—these form the core regions defined above. Satellites are seeded by stochastic jumps that extend over regions which either were occupied by a different allele already, or get filled in by a different allele before the satellite has a chance to merge with the core. For $\mu = 1.3$, haloes extend only a short distance out from the core, whereas at $\mu = 0.7$ the haloes often extend over a distance many times the core extent.

The increased fragmentation of clones with broader dispersal kernels has a marked impact on local diversity in sub-regions of the range. Haloes belonging to different alleles overlap to produce regions of high diversity, as exemplified by the dashed box in Fig. 1(a) for $\mu = 1.5$, which contains demes belonging to six of the 10 unique alleles despite being a small fraction of the total range area. By contrast, the same region contains only one allele at $\mu = 3.5$ for which clones form contiguous domains.
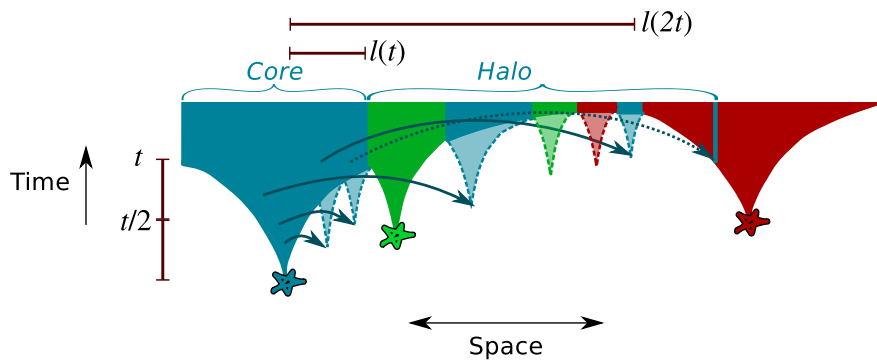
FIG. 2 **Marginal dynamics of jump-driven soft sweeps.** When long-range dispersal is significant, clones of different allelic identity (distinguished by colors) grow out of their originating mutations (stars) by accumulating satellite clusters (translucent cones). If only a single mutation were present, the extent $\ell(t)$ of the high-occupancy core (consisting of satellite clusters which have merged with the cluster growing out of the originating mutation) would, on average, follow a faster-than-linear growth rule, depicted schematically by the border of the opaque regions. For a wide range of kernels in the vicinity of $\mu = d$, the growth rule arises from a hierarchy of length and time scales related by doublings in time: the satellite that merges with the core at time t would typically originate from a key jump out of the core at time $t/2$, that extended over a length $\ell(t)$ (solid arrows out of blue region). Looking forward in time, a clone whose core has grown for time t without being obstructed will have likely seeded satellites out to a distance $\ell(2t)$. In the presence of recurrent mutations, these satellites may be obstructed from merging with the core due to intervening cores and satellites of different mutational origin (green and red regions). For very broad dispersal kernels, the halo also includes rarer jumps out of the core (dashed arrow) that land in regions that are being taken over by other alleles, but establish themselves in stochastic gaps in those regions. The disconnected satellite clusters and isolated demes comprise the halo region of the mutant clone.

Other effects of broadening the dispersal kernel are also visible in Fig. 1: the spread in clone sizes becomes larger, and individual clones take many more generations to attain a given size.

To build a quantitative understanding of these variations, we begin by noting that at early times in Fig. 1(b), single clones grow largely unencumbered by other clones. We can therefore gain insight from existing results on the jump-driven growth of a *single* advantageous clone expanding into a wildtype background [17]. The key features are summarized here and illustrated for the blue clone in Fig. 2. Consider a clone that grows from a mutation that originated at time $t_0 = 0$ at the origin. At times longer than a short transient, the clone fills most sites out to some distance from the origin. In line with the terminology established above, we call this region of high occupancy the *core* of the growing clone. Its typical extent over time (i.e. the average radius of a core that has grown for time t) is quantified by a function $\ell(t)$ which itself depends on the dispersal kernel. As sites in the core get filled, they send out offspring through long-range dispersal events drawn from the specified kernel, which then grow into independent satellite clusters. As a result, at any time t there are also demes outside the core which are occupied by the mutant. However, the occupancy of sites outside the core decays as $r^{-(d+\mu)}$ with distance r from the originating mutation [17], fast enough that the total mass of the clone at time t is proportional to $\ell^d(t)$.

As sketched in Fig. 2, the core grows through the merger of satellite clusters that grew out of rare but consequential "key jumps" out of the core at earlier times

(solid arrows in Fig. 2). Ref. 17 identified qualitative differences in the behaviour of key jumps and the resulting functional forms of $\ell(t)$ as the kernel exponent is varied. When $\mu > d+1$, the extent of typical key jumps remains constant over time, which implies that they must originate and land within a fixed distance from the boundary of the high-occupancy region at all times. As a result, clones advance via a constant-speed front similar to the case of wavelike growth; i.e. $\ell(t) \propto t$. Furthermore, the separation between the core and satellites is insignificant at long times, giving rise to essentially contiguous clones. By contrast, for $\mu < d+1$, growth is increasingly driven by jumps that originated in the interior of the core at earlier times, and key jumps become longer with time. The resulting growth of $\ell(t)$ is faster-than-linear with time. The value $\mu = d$ is an important marginal case which separates two distinct types of long-time asymptotic behaviour for $\ell(t)$: power-law growth for $d < \mu < d+1$ and stretched-exponential growth for $0 < \mu < d$ (see the second column of Table I for the asymptotic growth forms in all regimes). As $\mu \to 0$, spatial structure becomes increasingly irrelevant and the growth dynamics approaches the exponential growth of a well-mixed population.

These features of single-clone growth can be directly connected to the spatial patterns in Fig. 1 when recurrent mutations are allowed. The tessellation of the range into contiguous domains for the highest values of μ is exactly as expected from the wavelike growth situation when $\mu > d + 1$. When $\mu < d+1$, by contrast, each clone consists of a growing core and well-separated satel-

Kernel exponent	Asymptotic $\ell(t)$	χ
$\mu < d$	$\exp(B_\mu t^\eta)$	$[\xi/W(\xi)]^{\frac{1}{\eta}}, \quad \xi \equiv B_\mu \eta d(\tilde{u}\omega_d)^{-\eta}$
$\mu = d$	$\exp(c\log^2 t)$	$(\tilde{u}\omega_d)^{-1} e^{2\left(\log 2 - \sqrt{\log 2 \log(2/\tilde{u}\omega_d)}\right)}$
$d < \mu < d+1$	$t^{1/(\mu-d)}$	$\tilde{u}^{-1/\mu}$
$\mu = d+1$	$t\log t$ (1D)	$\frac{1}{2}\sqrt{\frac{W(1/\tilde{u})}{\tilde{u}}}$
$\mu > d+1$	t	$\tilde{u}^{-1/(d+1)}$

TABLE I **Core growth asymptotics and characteristic core sizes.** The table catalogues the asymptotic behaviour of $\ell(t)$ (from Ref. 17) along with the expected scaling of the characteristic clone size χ, omitting distance and time scales for ℓ and t respectively. W is the Lambert W-function, $B_\mu \approx 2d\log(2)/(\mu-d)^2$, and $\eta = \log[2d/(d+\mu)]/\log 2$. Note that for values of μ near d, the asymptotic growth forms are of limited value since the time befor the asymptotics is reached becomes very large. In this situation, more accurate values of χ are obtained using a scaling form for $\ell(t)$, see Appendix B for more details.

lite clusters at any time. Unlike the single-mutant case, satellites belonging to a particular clone are no longer guaranteed to merge with the core or with each other at later times: due to allelic exclusion, mergers are obstructed by cores and satellites with a different allelic identity, as shown schematically in Fig. 2. The final pattern of frozen-in satellite clusters comprises the previously identified halo structure around each core when $\mu < d+1$.

1. Marginal dynamics and the relative sizes of core and halo

We can quantify the expected spatial extent of entire clones (including haloes) relative to cores by considering the dynamics in the vicinity of the marginal value $\mu = d$. Although the long-time asymptotic dynamics are qualitatively different above and below this value (power-law in t for $d < \mu < d+1$, and stretched-exponential for $0 < \mu < d$), the approach to asymptotic behaviour is extremely slow for values of μ close to d, with the intermediate-time evolution controlled by the marginal dynamics at $\mu = d$. As a result, the marginal dynamics is important for a wide range of values of μ at biologically-relevant time scales [17].

In the marginal regime, the scaling behaviour of key jumps follows a particularly simple pattern, illustrated schematically in Fig. 2: satellite clusters which merge with the core at time t are seeded by key jumps that typically happened around time $t/2$ and covered a distance of order $\ell(t) \ll \ell(t/2)$. Therefore, a core that has grown up to some extent $\ell(t)$ has likely already seeded satellites out to a distance of order $\ell(2t)$, some of which will have reached an appreciable size as illustrated in Fig. 2. If the core has grown to some linear size X', we then expect satellites that have reached a significant size to extend to a distance $Y' \equiv \ell(2\ell^{-1}(X'))$, which may be considered to be a lower bound on the expected extent of the halo. When isolated demes embedded within cores and satellites belonging to other clones are included, the full spatial extent of the clone is even larger, because there remains a finite probability of rare jumps from the core

out to distances farther than Y' (dotted arrow in Fig. 2).

The above estimate for Y', when approximated using the long-time asymptotic growth rules for different jump kernels, reveals qualitatively different scaling behaviours for the clone extent on either side of the critical point $\mu = d$. For power-law growth, $\ell(t) \sim t^{1/(\mu-d)}$, we find $Y'/X' \sim 2^{1/(\mu-d)}$; i.e. the ratio of halo size to core size is a constant that grows as $\mu \to d$ but is independent of the size of the clone. By contrast, in the stretched-exponential regime, with $\ell(t) \sim \exp(Bt^\eta)$ where B and η depend on μ, we find $Y'/X' \sim X'^{(2^\eta-1)}$; i.e. the ratio of halo to core depends on the core size as well as on the kernel exponent. Since $\eta = \log[2d/(d+\mu)]/\log 2 > 1$, the halo becomes increasingly prominent as $\mu \to 0$. These scaling estimates break down as μ approaches d—for instance, the ratio Y'/X' diverges in the power-law growth regime—mirroring the limited utility of the approximate asymptotic forms for $\ell(t)$ near $\mu = d$. Instead, we must use the more accurate forms for $\ell(t)$ (Appendix B) to evaluate X' and Y'. However, upon using these forms with fitted magnitude and time scales for 1D clones, the qualitative picture is largely unchanged: the ratio Y'/X' becomes very weakly dependent on the core size as $\mu \searrow d = 1$, but the dependence is much stronger when μ drops below 1. For instance, at $\mu = 1.2$, the predicted ratio Y'/X' merely doubles from 4 to 8 as X' spans four orders of magnitude from $X' = 10$ to $X' = 10^5$; by contrast, at $\mu = 0.8$ the ratio changes by an order of magnitude (from roughly 5 to 70) over the same range of core sizes.

In summary, the growth dynamics for single clones at different kernel exponents predicts the following structure for large mutant clones: contiguous, compact clones for $\mu > d+1$; a high-occupancy core with a halo of well-developed satellite clusters that extends out to a size-independent (but kernel-dependent) multiple of the core radius for $d < \mu < d+1$; and a sparse halo which is significantly larger in extent than the core and becomes more prominent with increasing clone size for $\mu < d$. We now assume that these conclusions, and the associated scaling relations for the linear extent of the core and halo, also apply to the spatial structure of mutant clones that have

grown during soft sweeps and have been frozen in due to interference with clones of differing mutational origin. This key assumption is tested in the following section.

2. Occupancy profiles

To verify the structural features outlined above, we measured average occupancy profiles of distinct clones in 1D soft sweep simulations. The occupancy profile of a particular clone is defined as the probability $\rho(r)$ that a deme at distance r from its point of origin is occupied by that clone. Occupancy profiles from clones of different sizes are averaged by scaling the distance coordinate of each profile by the total mass X of that clone, to obtain an averaged profile $\bar{\rho}(r/X)$ for each kernel and rescaled mutation rate. This choice of distance scale is justified by the observation that averaged occupancy profiles for a given kernel with vastly different average clone sizes collapse onto a single curve when the distance coordinate is rescaled by the size, consistent with the core radius being proportional to X, see Supplementary Fig. A1.

Averaged occupancy profiles for different jump kernels are shown in Fig. 3(a) with $\tilde{u} = 10^{-5}$. We observe that when $\mu > d + 1$, the averaged occupancy is negligible for $r/X > 1$, and has a profile symmetric about $r/X = 1/2$ for lower values. This form is consistent with the entire clone being contained in a single domain, which grows to different lengths on either side of the originating mutation, as illustrated in the inset.

The predicted breakup of clones due to long-range dispersal is reflected in the overall broadening of occupancy profiles as the kernel exponent μ is reduced below the critical value $d + 1$. In this range, an appreciable portion of the clone lies outside the maximal distance from the origin ($r/X = 1$) that would be allowed for a contiguous domain. The dropoff in occupancy becomes increasingly steep at low values of r/X as μ is reduced, but more gradual at larger distances, consistent with a narrowing high-occupancy core balanced by a halo of increasing prominence. At large distances, the falloff in occupancy is consistent with the power-law behaviour expected from the single-clone growth dynamics, $\bar{\rho} \propto r^{-(\mu+d)}$ [dashed lines in Fig. 3(b)], which supports our assumption that the final structure of a mutant clone in a spatial soft sweep is similar to that of a single clone expanding without interference.

To quantify the relative prominence of the halo to the core across all growth regimes, we define the rescaled average core radius, $\tilde{r}_c$, as the value of r/X at which the averaged occupancy drops below $1/2$. We also define the core occupancy as the fraction of the total occupancy contained within $0 < r < 2\tilde{r}_c$. Although the precise values of these thresholds are not important, our definitions are chosen so that the core radius for contiguous domains is half the clone size, and the entire clone

mass is contained within twice the core radius, matching our intuition for clones with no halo. Deviations from the values for contiguous domains are only seen when $\mu < d + 1$ [Fig. 3(c)], consistent with the expectation from the growth dynamics of single clones. We find that both quantities approach zero as $\mu \to 0$, confirming that the core becomes irrelevant in the panmictic limit. However, the core contains an appreciable fraction of the total occupancy for all values of μ that we have simulated. This observation, together with the fact that the absolute core size scales with the clone mass, implies that insights about the geometric extent of cores can be connected to the total mass of their respective clones, as we do in the following section.

B. Characteristic scales via mutation-expansion balance

So far, we have focused on the spatial structure of individual clones within a soft sweep, and have shown that many aspects of this structure can be understood from the theory of growth of a single clone under the same dispersal kernel. To address questions of global and local allelic diversity, however, we need to explicitly consider the concurrent growth of multiple clones. We now show how the balance between jump-driven growth and the dynamics of introduction of new mutations sets the typical size and spatial extent of clones.

1. Size of a "typical" clone

In an infinitely large range, a single clone could grow without bound, but in the presence of recurrent mutations, the growth of any one clone is obstructed by other clones. Balancing mutation and growth gives rise to a characteristic time scale t^*, and associated characteristic linear extent χ, for mutant clones in multi-origin spatial sweeps [11]. These scales determine whether a sweep will be "hard" or "soft" within a finite range of given size.

When clones grow as compact, connected domains, growth is interrupted when the advancing sharp boundary of the clone encounters a different allele. However, for clones growing via long-range dispersal events, a sharp boundary no longer exists, and small obstacles can be traversed by jumps. The picture of jump-driven growth that we have developed suggests that haloes belonging to different clones can freely overlap, whereas core regions cannot. Therefore, new mutations arising within the halo region of a growing clone do not significantly impede its growth. Instead, the crucial factor restricting the growth of a clone is when its high-occupancy *core* encounters a different clone, as depicted schematically in Fig. 2. Since $\ell(t)$ defines precisely the time-evolution of the core extent of a single clone, we define t^* as the time interval for which exactly one mutation is expected to occur in the
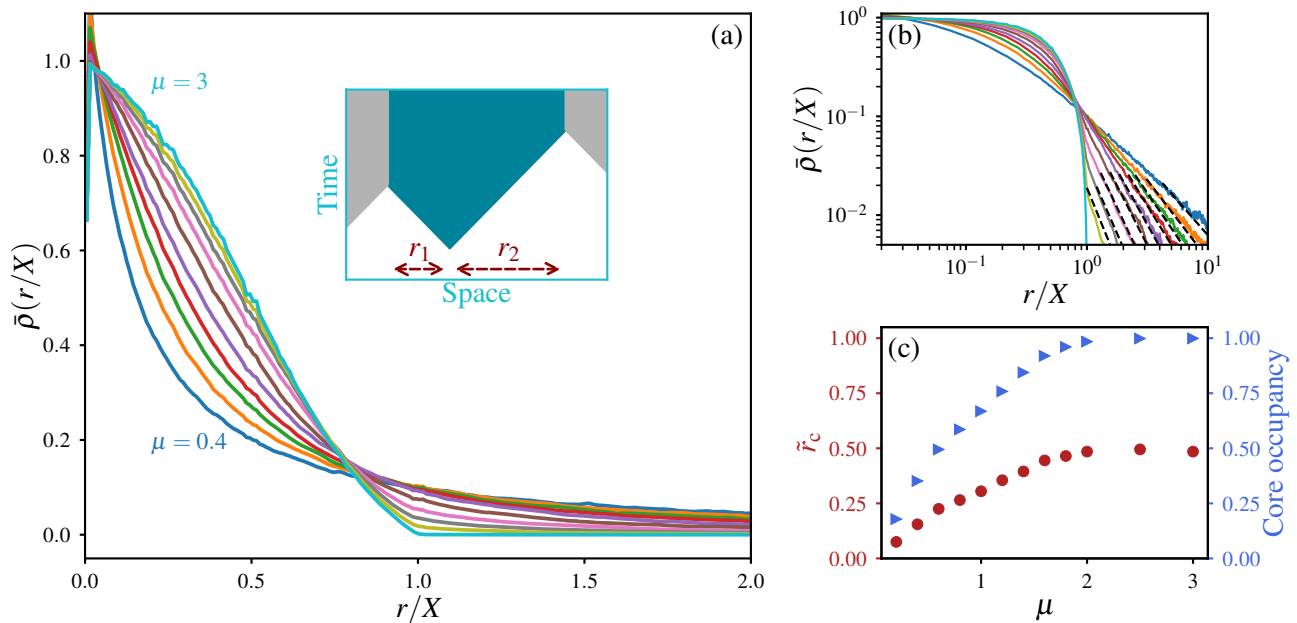
FIG. 3 **Occupancy profiles of clones show the reduced prominence of the core with increased long-range dispersal.** (a) Averaged occupancy profiles of mutant clones in 1D, with $L = 10^6$ and $\tilde{u} = 10^{-5}$. Colours signify different dispersal kernels, with exponents $\mu = \{0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2, 3\}$ in order of increasing occupancy at the origin. Curves are obtained by averaging over all clones with total mass $X > 100$ from 100 independent simulations for each dispersal kernel. Inset illustrates the origin of the variation in occupancy for $r/X < 1$ in the wavelike growth regime: the mutational origin need not be positioned at the centre of mass of the contiguous domain, giving rise to an occupancy profile $\rho(r) = \{1, 0 < r \leq r_1; \quad 1/2, r_1 < r \leq r_2; \quad 0, r > r_2\}$. (b) Same data as in (a) on logarithmic scales. The dashed lines show the power-law dependence $\bar{\rho} \propto r^{-(\mu+d)}$. (c) Rescaled average core radius, $\tilde{r}_c$, defined as the value of r/X at which the profile in (a) drops below $1/2$. The core occupancy, defined as the fraction of the averaged occupancy that lies within $0 < r < 2\tilde{r}_c$, is also shown.

space-time region swept out by the growing core:

$$\tilde{u}t^*\omega_d\ell^d(t^*) = 1, \qquad (1)$$

where ω_d is the volume of the d-sphere of radius 1 ($\omega_1 = 2$, $\omega_2 = \pi$). The corresponding characteristic extent, $\chi \equiv \ell(t^*)$, matches the length scale introduced in Ref. 11 to characterize spatial soft sweeps.

Rough estimates for t^* and χ can be obtained by using the long-time asymptotic forms for $\ell(t)$ in the different growth regimes, see Table I. These estimates highlight the vastly different functional dependences of the characteristic scales on the rescaled mutation rate as the kernel exponent is varied. For quantitative tests, we use scaling forms for $\ell(t)$ derived in Ref. 17, which are much more accurate at short times when μ is close to d. All theoretical forms for $\ell(t)$ include unspecified multiplicative factors A and B for the length and time variables, which are determined by fitting the functional forms to the growth of isolated clones starting from a single seed, see Appendix B for details.

Fig. 4 compares the measured average clone sizes from 1D simulations with $\chi(\tilde{u})$ computed using the theoretical forms for $\ell(t)$ for different system sizes and mutation rates. For clarity, the expected scaling with mutation rate in the wavelike spreading limit, $\chi \propto 1/\sqrt{\tilde{u}}$,

is divided out. At each value of μ, the computed $\chi(\tilde{u})$ quantitatively captures the dependence of X_{ave} on $\tilde{u}$ over many orders of magnitude, up to a $\tilde{u}$-independent constant factor (the factor depends on model details and is treated as a fitting parameter, but it turns out to not vary significantly with μ). These results confirm that the mutation-growth balance captured in Eq. 1, first identified in Ref. 11, remains relevant for characterizing spatial soft sweeps when long-range dispersal is prominent.

Many features of spatial soft sweeps are captured in the characteristic scales defined above. The length scale χ provides the appropriate rescaling of space to compare two sweeps with different mutation rates but the same growth rule $\ell(t)$. The precise definition, Eq. 1, also suggests that χ is the size attained by the core of a typical clone under mutation-expansion balance. This interpretation must be used with care, because we will see that the actual size distribution of clones in a spatial soft sweep is often very broad so that no one size dominates. However, it is very unlikely that the core region will grow to a radius much greater than χ, so that the largest clones in the range—which dominate the sampling statistics for random samples from the range—are of order χ. Ranges that are much larger than χ will display multiple clones with high probability, whereas ranges much smaller than
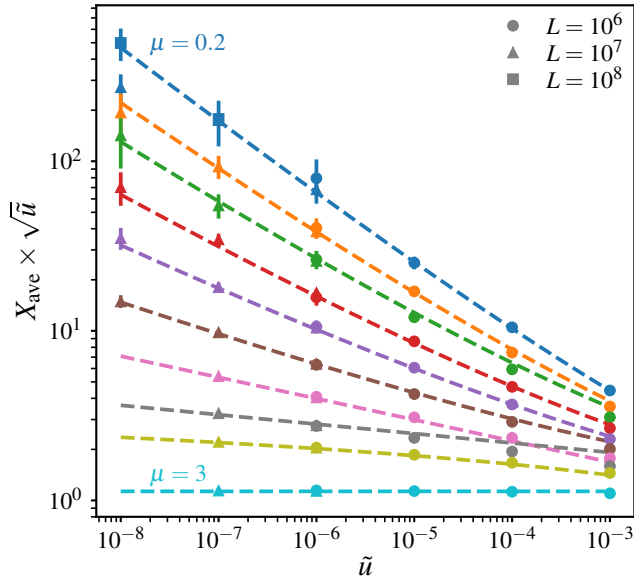
FIG. 4 **Average clone size is set by mutation-expansion balance.** The average size of mutant clones as measured from simulations in 1D as a function of rescaled mutation rate, scaled by the expected dependence ($\propto \sqrt{\tilde{u}}$) for wavelike growth of clones. Results from different system sizes (symbols) are presented for each dispersal kernel quantified by μ (colours); the values are (from top to bottom) $\mu = \{0.2, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2, 3\}$. Each point represents an average over 20 or more independent simulations. Error bars denote measured standard deviation across repetitions. Dashed lines show the theoretical predictions for $\chi(\tilde{u})$ for each dispersal kernel (see Appendix B), multiplied by a μ-dependent magnitude factor whose value is 1.5 for $1 \le \mu < 2$, 1.6 for $\mu > 2$, and 1.65 for all other values.

FIG. 5 **The spatial extent of clones deviates strongly from the average size.** Spatial extent of clones, defined as the distance between leftmost and rightmost demes belonging to a particular clone, averaged over all clones and normalized by the average clone size X_{ave}. Each point is an average over values from 20 independent simulations. Dotted lines connect simulation data points. Dashed lines show the theoretical expectations for r_{max}/χ (with $\tilde{u} = 10^{-3}$) and ψ/χ in the ranges $\mu < 1$ and $1 < \mu < 2$ respectively, multiplied by model-dependent numerical factors. For $\mu < 1$, the fall in the measured ratio at low values of $\tilde{u}$ is due to finite-size effects, as seen by comparing the two system sizes: the measured values of Δr in this range are below the values that would be measured in an infinitely large system.

χ will likely be dominated by one clone.

Similarly, t^* sets the time scale for the sweep to be completed, independently of the range size: if both sides of Eq. 1 are multiplied by the range size N, it equates to a condition for the range to be completely filled by mutations accumulated at a rate $N\tilde{u}$ over a time t^* without interference, each of which grows to the characteristic size $\ell(t^*)$. The expected number of mutational origins scales as $N\tilde{u}t^* = N/\omega_d\chi^d$, which implies that the average clone size must be proportional to χ^d.

2. Characteristic extent of clones

The analysis of the spatial structure of individual clones (Section III.A) showed that the extent of clones (i.e. the portion of the range over which demes belonging to the clone can be found) can be many times larger than the expected extent for a compact clone, especially for broad dispersal kernels. Therefore, the average clone
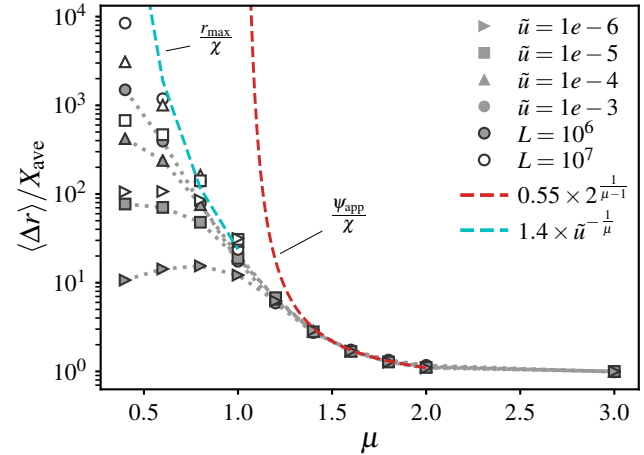
size may not be representative of the spatial extent of typical clones over the range, pointing to the potential relevance of additional length scales when characterizing jump-driven spatial soft sweeps. To quantify this effect, we measure the extent, Δr, of a clone in our 1D simulations as the largest distance between any pair of individuals belonging to that clone. Fig. 5 shows the ratio between $\langle\Delta r\rangle$ (brackets denote averaging over all clones in the range) and the average clone size X_{ave}. We find that the average spatial extent is much larger than the average clone size, except in the regime of compact clones ($\mu > d + 1$). Their ratio appears independent of the rescaled mutation rate for kernels narrower than the marginal value $\mu = d$, whereas it appears $\tilde{u}$-dependent, and grows dramatically in magnitude, for $\mu < d$. At the smallest values of μ, the measured extent becomes limited by the system size (the maximum allowed extent is $L/2$ due to periodic boundary conditions), which artificially suppresses the ratio $\langle\Delta r\rangle/X_{\text{ave}}$ as seen by comparing the measurements at $L = 10^6$ and $L = 10^7$.

To explain these features, we return to the core-halo picture of clone structures. We had argued that, in the vicinity of the critical value $\mu = d$, a clone whose core has grown to some size $\ell(t)$ will have likely established satellites of a significant size out to a distance $\ell(2t)$. Having

derived a characteristic time scale t^* for the growth of a typical clone restricted by mutation-expansion balance, a rough estimate for the extent of its halo is provided by the quantity $\psi \equiv \ell(2t^*)$, which quantifies the extent of established satellites of the typical clone (at the cost of ignoring interference among haloes of different clones). An alternative argument, which balances the rate of key jumps out of the core with the expected rate of mutations arising in the target region of the jumps, produces a similar estimate for the typical halo extent (see Appendix C).

The length scale ψ is a lower bound to the spatial extent of clones, since the growing core at time t also sends out offspring far beyond the range of the key jumps, which would not become part of the core at a future time $2t$. Since our time units are set by the migration rate, the net number of jumps out of a typical clone which grows over a time t^* is roughly $t^* \ell^d(t^*)$, which scales as $1/\tilde{u}$ by the definition of t^*. The fraction of these jumps which end up beyond a distance l is $\int_l^\infty J(r) \, d^d r = l^{-\mu}$. A crude estimate for the outer limit $r_{\max}$ of these rare events is obtained by requiring the net number of jumps from the core to distances $r_{\max}$ or greater to be equal to 1:

$$r_{\max}^{-\mu}/\tilde{u} = 1 \Rightarrow r_{\max} = \tilde{u}^{-1/\mu}. \qquad (2)$$

The extremal measure Δr is sensitive to these isolated offspring, even if they do not form satellite clusters of significant size. However, the scaling for $r_{\max}$ does not account for the fact that many of the long-distance rare jumps would fail to establish because they land on the high-occupancy cores of competing clones, which fill up the range over the same time scale t^*. Therefore, $r_{\max}$ is likely to be relevant only when the sparse halo regions become dominant over the cores, i.e. for $\mu < d$.

We now use the asymptotic growth rules (Table I) to gauge the qualitative behaviour of the candidate length scales ψ and $r_{\max}$ for the typical spatial extent of clones. For power-law growth in 1D $(1 < \mu < 2)$, ψ and $r_{\max}$ show the same scaling behaviour as a function of the rescaled mutation rate, making the distinction between the two length scales irrelevant. The asymptotic growth rule $\ell(t) \propto t^{1/(\mu-1)}$ predicts the mutation-rate-independent ratio $\psi_{\mathrm{app}}/\chi \sim 2^{1/(\mu-1)}$, an approximation that becomes increasingly inaccurate as $\mu \to 1$. Having already established that $X_{\mathrm{ave}} \propto \chi$, we therefore expect the ratio $\langle \Delta r \rangle / X_{\mathrm{ave}}$ to follow this functional form. The simulation results are consistent with this behaviour for $\mu > 1.2$, up to an overall proportionality constant that is not captured by the scaling arguments (red dashed line in Fig. 5). The divergence of the approximation ψ_{app} as $\mu \to 1$ reflects the inaccuracy of the long-time asymptotic form for $\ell(t)$ as the marginal kernel exponent is approached.

For $\mu \leq 1$, $r_{\max}$ quickly dominates ψ in magnitude due to the strong divergence in Eq. 2 as $\mu \to 0$ for $\tilde{u} < 1$. Upon using the more accurate forms for $\ell(t)$ from Appendix B to compute ψ and χ (see Appendix C for details), we find that the ratio ψ/χ underestimates the measured values in Fig. 5 by many orders of magnitude. By contrast, the alternative length scale $r_{\max}$ (Eq. 2) successfully captures the variation in $\langle \Delta r \rangle / X_{\mathrm{ave}}$ for the largest rescaled mutation rate $\tilde{u} = 10^{-3}$ (cyan dashed line in Fig. 5). These results confirm that the spatial structures associated with soft sweeps are qualitatively different in the stretched-exponential and power-law regimes, and provide quantitative predictions for the scaling of the typical extent of clones in both regimes. For the lower mutation rates, the predicted $r_{\max}$ values are larger than the largest system size $L = 10^7$, pointing to the need for significantly larger-scale simulations to avoid finite-size effects.

C. Clone size distributions vary with dispersal kernel and influence global sampling statistics

Unlike our simulations, studies of real populations do not have access to complete allelic information over the entire range. Instead, the allelic identity of a small number of individuals is sampled from the population. The likelihood of detecting a soft sweep in such a random sample is determined not only by the total number of distinct clones in the range, but also by their size distribution: if the range contains many clones, but all but one are at extremely small frequency, the sweep is likely to appear "hard" in a small random sample which would with high probability contain only the majority allele. Long-range dispersal can therefore influence soft sweep detection not only by setting the average clone size, but also by modifying the distribution of clone sizes around the average. Having already established that the dispersal kernel has a significant effect on the average clone size (Fig. 4), we now analyze its effects on the clone size distribution and the consequences for soft sweep detection.

Clone size distributions were quantified by computing the *allele frequency spectrum* $f(x)$, defined such that $f(x) \, \delta x$ is the expected number of alleles which have attained frequencies between x and $x + \delta x$ in the population [18]. The allele frequency spectrum is related to the average probability distribution of clone sizes, but has a different normalization $\int_{1/N}^1 x f(x) \, dx = 1$ which allows sampling statistics to be expressed as integrals involving $f(x)$ (we will exploit this fact in Section III.C.1 below). Analytical results for $f(x)$ can be derived for the deterministic wavelike growth limit $\mu \gg d + 1$ in 1D by mapping the spatial soft sweep on to a grain growth model [19], and for the panmictic limit $\mu \to 0$ in any dimension via a different mapping to an urn model [20]. The resulting functions, termed f_{w} and f_∞ for the two limits respectively, provide bounds on the expected frequency spectra at intermediate μ. Details of the map-

pings and complete forms for the functions $f_{\rm w}$ and f_{∞} are provided in Appendix D.

Fig. 6(a) shows allele frequency spectra computed from the outcomes of 1D soft sweep simulations for system size $L = 10^7$ and mutation rate $\tilde{u} = 10^{-4}$. We find that the frequency spectra vary strongly with the dispersal kernel, and approach the exact forms f_{∞} and $f_{\rm w}$ for small and large μ respectively. Generically, spectra become broader as the kernel exponent is reduced: as $\mu \to 0$, more high-frequency clones are observed. Although this broadening is partly explained by the increase in the average clone size due to accelerated expansion, which would lead to more high-frequency alleles, there are also systematic changes in the overall shapes of the distribution as the dispersal kernel is varied. Upon reducing the rescaled mutation rate to $\tilde{u} = 10^{-6}$ [Fig. 6(b)], all frequency spectra broaden due to the increase in the average clone size, but the variations in shapes of the $f(x)$ curves with μ remain consistent across the two mutation rates. These observations suggest that spatial soft sweep patterns with similar numbers of distinct alleles in a range might nevertheless have vastly different clone size distributions due to different dispersal kernels, with implications for sampling statistics.

To uncover variations due to long-range dispersal beyond changes in the average clone size, we rescaled the frequency spectra by the expected dependence on $X_{\rm ave}$, which we have already established as being set by the mutation-expansion balance. Under the assumption that modifying the mutation rate for a given dispersal kernel only affects the characteristic length scale $\chi \propto X_{\rm ave}$, we expect that

$$f(x) = (N/X_{\rm ave})^2 F_\mu(Nx/X_{\rm ave}), \qquad (3)$$

where F_μ depends only on the dispersal kernel and not the rescaled mutation rate. Although this is only exact in the regime of power-law growth (for which ψ is proportional to χ, so that the entire spatial pattern is determined by one length scale), we find that the rescaling suggested by Eq. 3 leads to a decent collapse of frequency spectra for the two mutation rates across all regimes, see Fig. 6(c). The existence of the additional length scale ψ, which is not merely proportional to χ in the stretched-exponential and marginal cases, introduces weak corrections that are logarithmic in mutation rate and system size and become strongest as $\mu \to 0$, see Appendix E.

The scaled frequency spectra show that broader dispersal kernels favour broader allele frequency spectra even after accounting for changes in the average clone size. At $\mu = 4$, the steep decline in the frequency spectrum occurs near the frequency expected of an average clone, $x \approx X_{\rm ave}/L$. As μ is reduced, the falloff occurs at higher frequencies; at $\mu = 0.4$, for instance, clones with frequencies an order of magnitude higher than the average clone are still likely. Qualitatively, this trend is a result of the increased nonlinearity of the growth functions $\ell(t)$

for broader dispersal. If we assume no interference among distinct clones until the time t^*, the size of an allele which arrives at time t_i is proportional to $\ell^d(t^* - t_i)$. For a given spread of arrival times of mutations, the spread of final clone sizes is significantly enhanced by nonlinearity in $\ell(t)$. Therefore, the increased departure from linear growth in $\ell(t)$ as $\mu \to 0$ gives rise to broader clone size distributions. Deterministic approximations to the clone size distributions expected for a given $\ell(t)$, described in Appendix E, support this heuristic picture.

Although we do not have analytical expressions for the frequency spectra at intermediate μ, the measured curves and deterministic calculations suggest a simple approximate form with two parameters:

$$f(x) = \begin{cases} \dfrac{p+2}{x_{\rm c}^{p+2}} x^p, & x < x_{\rm c} \\ 0, & x > x_{\rm c}, \end{cases} \qquad (4)$$

i.e. a power-law behaviour characterized by exponent p, up to some maximal frequency $x_{\rm c}$, with the constant of proportionality determined by the normalization. The values p and $x_{\rm c}$ are determined from the numerical data, but are also consistent with theoretical arguments (Appendix E). The small-x behaviour of the two limiting spectra, $f_{\infty}(x) \sim x^{-1}$ and $f_{\rm w}(x) \sim x$ as $x \to 0$, imply that p is restricted to vary from -1 to 1 as μ increases from zero. Despite its simplicity, this approximation can be used to quantify the relationships among various features of the clone size distributions as we show in Appendix F. For instance, the power-law ansatz predicts a relation between the average clone size and the cut-off frequency, $Lx_{\rm c}/X_{\rm ave} = (p+1)/(p+2)$, which matches the trends observed in the rescaled frequency spectra, see inset to Fig. 6(c).

1. Sampling statistics

The utility of $f(x)$ in the context of soft sweep detection is made apparent by noting that the probability $P_{\rm hard}(j)$ of finding only one unique allele in a sample of size $j \geq 2$ drawn randomly from the population (i.e. detecting a *hard* sweep) is [18]

$$P_{\rm hard}(j) = \int_0^1 x^j f(x)\, dx. \qquad (5)$$

The probability of observing a *soft* sweep in a sample of size j is simply $P_{\rm soft} = 1 - P_{\rm hard}(j)$ [1]. Since $xf(x)$ does not diverge as $x \to 0$ for all observed frequency spectra, the integral in Eq. 5 is dominated by contributions from

_____

[1] Although $P_{\rm soft}$ might be more relevant to soft sweeps, we deal with $P_{\rm hard}(j)$ in the following sections because it is more straightforward to compute and manipulate mathematically.
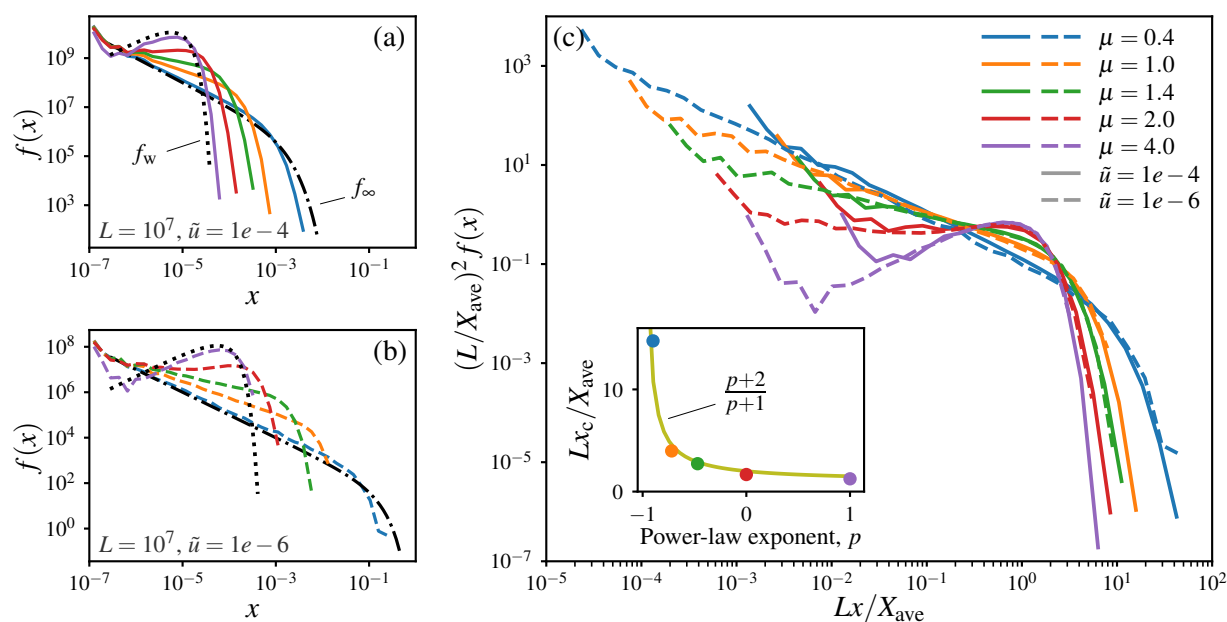
FIG. 6 **Variation of allele frequency spectra with mutation rate and dispersal kernel.** (a) Allele frequency spectra estimated from 1D simulations with $L = 10^7$ and $\tilde{u} = 1e-4$. Each curve is the average of frequencies measured from 20 independent simulations. Curves are coloured by dispersal kernel according to the legend in (c). The uptick for the lowest bin is an artifact of the logarithmic bin sizes together with the hard lower cutoff in allele frequency at $x = 1/L$. Black dotted and dash-dotted lines show the analytical distributions for wavelike spreading and panmictic limits respectively. (b) Same as in (a) except with $\tilde{u} = 1e-6$. (c) Frequency spectra from (a) and (b), rescaled to remove the expected variation due to changes in average clone size. Inset, dependence of the cut-off frequency x_c on the exponent p when frequency spectra are approximated by a power law $f(x) \propto x^p$ for $x \leq x_\text{c}$ and $f(x) = 0$ for $x > x_\text{c}$. Dots show the cutoff estimated numerically as the value for which the second derivative of the rescaled curves first drops below -4, plotted against the observed exponents $p \approx \{-0.9, -0.72, -0.47, 0, 1\}$ for the small-frequency behaviour.
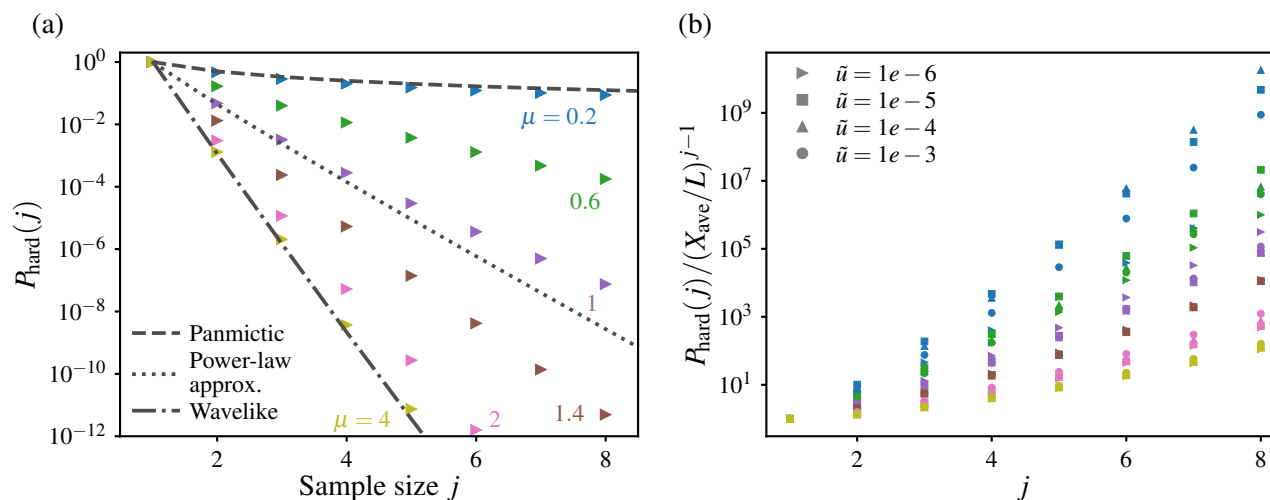

FIG. 7 **Probability of observing only one allele in a finite sample.** (a) The quantity $P_\text{hard}(j)$ computed from simulated clone size distributions for different dispersal kernels (colours, labeled) with $L = 10^6$ and $\tilde{u} = 10^{-6}$ in 1D. Three analytical forms are shown as dotted lines (from top to bottom): the Ewens' sampling result for the panmictic case, the approximate form derived using a hard-cutoff ansatz for the allele frequency spectrum for $\mu = 1$ (Eq. F3), and P_hard calculated from the exact $f(x)$ for the wavelike spreading limit (Appendix G). (b) The same quantity computed across a range of rescaled mutation rates (symbols), and scaled by the expectation for a range with all clones having the same size and hence the same frequency X_ave/L.

the high-frequency region of $f(x)$ and is therefore highly sensitive to the breadth of the frequency spectrum. Using the power-law ansatz for the frequency spectrum, Eq. 4, gives $P_{\text{hard}}(j) \propto x_c^{j-1}/j$ for large j (see Appendix F): the dominant behaviour is an exponential decay with sample size, with the decay scale set by the high-frequency cutoff x_c. At a given rescaled mutation rate, this cutoff falls by many orders of magnitude as μ is increased, as we saw in Fig. 6(a)–(b). As a consequence, the probability of finding a monoallelic sample also falls dramatically with increasing μ, see Fig. 7(a). Analytical calculations of P_{hard} using $f_\infty(x)$ and $f_w(x)$ in the $\mu \to 0$ and $\mu \gg d + 1$ limits (dashed and dash-dotted lines) provide bounds on the variation (see Appendix G for explicit forms of $P_{\text{hard}}(j)$ in these limits).

We have seen that increased long-range dispersal broadens the frequency spectrum both by increasing the average clone size, and by enhancing the spread of clone sizes around the average. To distinguish the effect of the shape of $f(x)$ from that of the overall size of clones, we scale $P_{\text{hard}}(j)$ by the idealized expectation $(\chi/L)^{j-1} \propto (X_{\text{ave}}/L)^{j-1}$ for a range of rescaled mutation rates, see Fig. 7(b). If the sampling statistics were determined primarily by the average clone size (which in turn is set by χ) and the effect of variations in the shape of $f(x)$ were insignificant, we would expect the rescaled $P_{\text{hard}}(j)$ for different kernels to all collapse on the same curve. Instead, we find that the sampling statistics vary significantly with μ even when accounting for differences in average clone size. Whereas the rescaling captures a significant amount of the variation in $P_{\text{hard}}(j)$ *within* each value of μ (with a residual $\tilde{u}$-dependence that differs for the different regimes of $\ell(t)$, and is due to the relevance of the additional length scales outside the power-law growth region), the rescaled curves vary widely among the different dispersal kernels.

Fig. 7(b) quantifies the influence of long-range dispersal on soft sweep detection beyond merely setting the average size of clones: if mutation rates are adjusted so that the characteristic length scales χ and hence the average clone sizes are comparable for different dispersal kernels, soft sweeps continue to be *less* likely to be detected for broader kernels (smaller μ). This happens because the range has a larger contribution from high-frequency clones with $x > \chi/L$ (or $(\chi/L)^d$ in general) for broader dispersal kernels, making monoallelic sampling more likely. Therefore, when ranges with similar overall diversity (as evaluated by the total number of unique alleles in the range) are compared, long-range dispersal skews the distribution of clone sizes in a way that disfavours soft sweep detection in small samples. Since a wide range of possible outcomes separates the two limits of panmictic ($\mu \to 0$) and wavelike spreading ($\mu \gg d + 1$), predictions based on these extremes might perform poorly in making inferences from sampling statistics in populations with intermediate long-range dispersal.

D. Local sampling protocols are highly sensitive to the core-halo structure

Population genomic studies are often limited not only in the number of independent samples available, but in their geographic distribution as well. Samples tend to be clustered in regions chosen for a variety of reasons such as anthropological or ecological significance, or practical limitations. The analysis of the last section would apply to comparing samples across different regions, provided that these relatively well spread out in the range. Here we focus on the variation within local samples from a subrange of the entire population. As illustrated by the wide variation in local diversity within the highlighted subranges (dashed boxes) in Fig. 1(a), inferences based on local sampling can be significantly different from inferences based on global information, and may be very sensitive to modes of long-range dispersal.

Long-range dispersal enhances local diversity. When clones extend over a much wider spatial range than required by their mass (Fig. 5), local subranges contain alleles whose origins lie far away from the subrange, and are consequently more diverse than expected from the diversity of the range as a whole. To quantitatively illustrate this effect, we compute sampling statistics for different dispersal kernels and subrange sizes from 1D simulations with a global range size much larger than the characteristic length scale χ (Fig. 8). (Subrange size and extent are equivalent in our 1D simulations.) We observe that the smaller clones expected at higher values of μ favour the detection of soft sweeps globally (Fig. 8a), but the diversity is less detectable in samples from subranges that are smaller than the characteristic size shared by the compact domains at $\mu = 4$. By contrast, samples from smaller subranges continue to show signatures of soft sweeps for broader dispersal kernels (Fig. 8b–c).

To compare the sensitivity of soft sweep detection to subrange size across different dispersal kernels and mutation rates, we focus on the probability of detecting the same allele in a *pair* of individuals randomly sampled from a subrange, $P_{\text{hard,s}}(2)$ (also called the species homoallelicity of the subrange). This probability is high only when the subrange is mostly occupied by the *core* of a single clone; it is low if the subrange contains cores belonging to different clones, or a combination of cores and haloes. Therefore, we expect χ (or equivalently the average clone size X_{ave}, which we may use as a proxy for χ in 1D) to also be the relevant scale to compare L_s values across different situations. Fig. 9(a) shows the dependence of $P_{\text{hard,s}}(2)$ on L_s/X_{ave} for different dispersal kernels and mutation rates in the $X_{\text{ave}} \ll L$ limit. As with the global sampling probabilities reported in Fig. 7(b), we find that the rescaling of subrange size with X_{ave} captures much of the variation among different mutation rates (symbols) for a given dispersal kernel. In contrast with the global sampling statistics, however, hard sweep de-
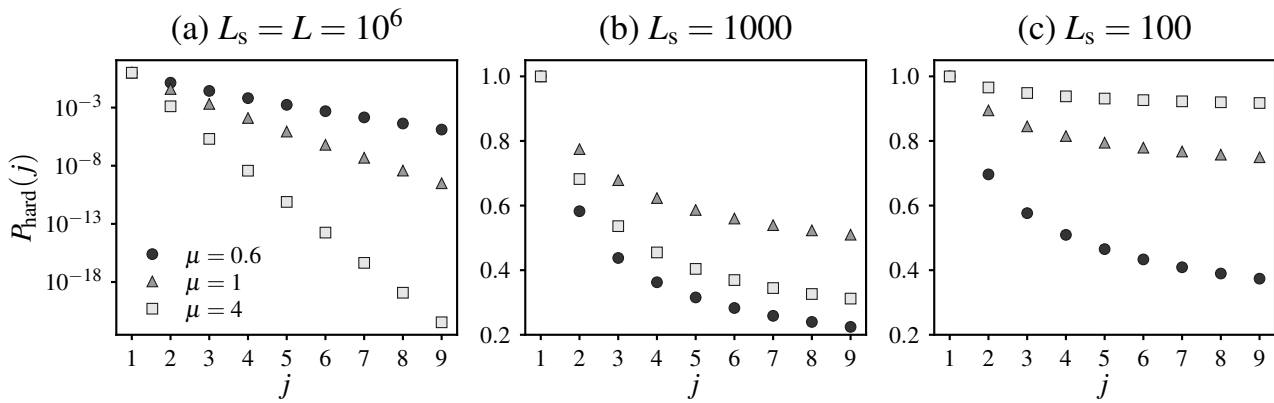
FIG. 8 **Dispersal promotes soft sweep detection in small subranges.** (a)–(c) Probability of observing a hard sweep in j samples randomly chosen from contiguous subranges of different sizes L_s in simulated 1D ranges of size $L = 10^6$, with rescaled mutation rate $\tilde{u} = 10^{-6}$. At the same mutation rate, broader dispersal kernels lead to a larger average clone size ($X_{ave} \approx 980$, 1.6×10^4, 4×10^4 for $\mu = \{4, 1, 0.6\}$ respectively), which reduces the total number of alleles and favours hard sweep signatures when the sampling is done over the entire range $[L = L_s$, (a)]. However, when L_s is reduced [(b)–(c)], the detection of soft sweeps become increasingly likely for the broader dispersal kernels; the broken-up structure of clones compensates for their smaller overall number. For small enough subranges, the order of values of $P_{hard}(j)$ with increasing μ is inverted compared to the values at $L_s = L$.
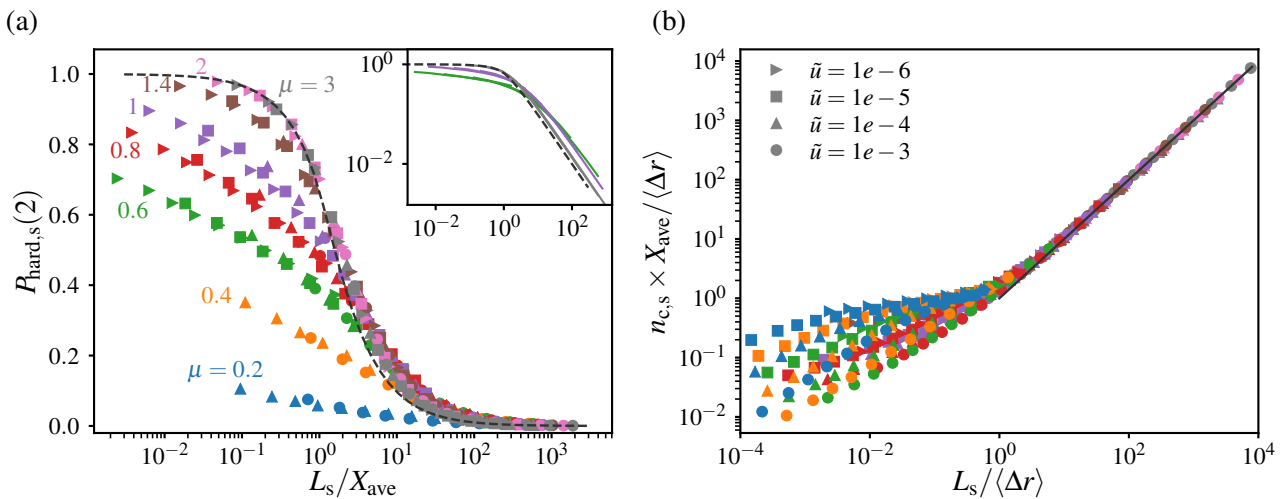

FIG. 9 **Different measures of diversity within a subrange are sensitive to different characteristic scales.** (a) Probability $P_{hard,s}(2)$ of observing a single allele in a pair drawn from a subrange of size L_s for different dispersal kernels (colours, labeled) and mutation rates [symbols, see legend in panel (b)], for 1D simulations with $L = 10^6$, as a function of the ratio L_s/X_{ave}. In all cases, the population range was chosen to be many times larger than the characteristic size χ and harbours many distinct alleles. The dashed line is the prediction if all clones are of the same size X_{ave}, in which case geometry dictates that $P_{hard,s}(2) = \{1 - x/3, x < 1; 1/x - 1/(3x^2), x \geq 1\}$. The inset shows data for $\mu = \{0.6, 1.0, 3.0\}$ on log axes. (b) Number of distinct alleles $n_{c,s}$ observed in a subrange of size L_s, shown as a function of the ratio $L_s/\langle\Delta r\rangle$. Values are scaled by $\langle\Delta r\rangle/X_{ave}$, the expected number of clones in the area occupied by the average halo. The solid line corresponds to $n_{c,s}X_{ave}/\langle\Delta r\rangle = L_s/\langle\Delta r\rangle$, or equivalently $n_{c,s} = L_s/X_{ave}$.

tection probabilities are suppressed (or equivalently, soft sweeps are *easier* to detect for the same rescaled subrange size) as the jump kernel is broadened. At high values of μ in the wavelike expansion limit, the shape of the curves is well-approximated by the null expectation for an idealized clone size distribution where all clones are perfectly contiguous segments of equal size X_{ave}. As μ falls below

$d+1$, the prevalence of overlapping haloes increases local diversity at the scale of satellite clusters, much smaller than the typical clone size would dictate. The effect is especially strong in the marginal and stretched-exponential growth regimes ($\mu \leq d$), which was associated with the halo dominating over the core (Figs. 3 and 5).

A different measure of subrange diversity is the total

number of distinct alleles present in a subrange on average, which we call $n_{c,s}$. Unlike the subrange homoallelicity, which was dominated by the most prevalent clone in the subrange, this measure gives equal weight to all clones, and is sensitive to haloes that overlap with the subrange. The expected number of distinct cores in the subrange is L_s/X_{ave}; in the absence of haloes, we would expect $n_{c,s}$ to be equal to this value. However, haloes of clones whose cores are outside the subrange would cause $n_{c,s}$ to exceed the number of cores in the subrange. This enhancement in diversity due to encroaching haloes would be expected to occur only when L_s is smaller than the average clone extent $\langle \Delta r \rangle$: when $L_s > \langle \Delta r \rangle$, the cores of clones whose haloes contribute to $n_{c,s}$ also lie within the subrange, and are accounted for in L_s/X_{ave}. This expectation is confirmed in Fig. 9(b). When the subrange size is rescaled by the extent of the clone including the halo, the average number of distinct alleles in the subrange follows $n_{c,s} = L_s/X_{ave}$ (solid line) in all cases, provided $L_s/\langle \Delta r \rangle > 1$. For smaller subrange sizes, $n_{c,s}$ lies above this estimate, reflecting the enhancement of local diversity due to encroaching haloes.

IV. DISCUSSION

Adaptation in a spatially extended population often uses different alleles in different geographic regions, even if the selection pressure is homogeneous across the entire range. The probability of such *convergent adaptation* [21] and the patterns of spatial soft sweeps that result depend on two factors: the potential for the population to recruit adaptive variants from either new mutations or from the standing genetic variation, and the mode of dispersal. Previous work has focused on the two extremes of dispersal phenomena: panmictic populations without spatial structure [3, 4, 5] or wavelike spreading due to local diffusion of organisms [11, 21]. However, gene flow in many natural populations does not conform strictly to either limit. Many species experience some long-distance dispersal either through active transport or through passive hitchhiking on wind, water, or migrating animals including humans [12, 13, 14]. For populations with a large range, even rare long-distance dispersal can strongly influence the adaptive dynamics [22].

We have described spatial patterns of convergent adaptation for a general dispersal model, with jump rates taken from a kernel that falls of as a power-law with distance. Although the underlying analysis is applicable to more general dispersal kernels, our specific choice of kernel allows us to span a wide range of outcomes using a single parameter. We have shown that long-range dispersal tends to break up mutant clones into a core region dominated by the clone, surrounded by a disconnected halo of satellite clusters and isolated demes which mingle with other alleles. A key result of our analysis is that al-

though the total mass of a clone is well-captured by the extent of the core region, the sparse halo can extend out to distances that are significantly larger than the core, sometimes by orders of magnitude. Therefore, understanding clone masses alone provides incomplete information about spatial soft sweep patterns, and can vastly underestimate the true extent of mutant clones.

By analyzing the balance between the jump-driven expansion of single clones and the introduction of new mutations, we have identified three characteristic length scales that quantify the spatial relationships between core and halo: the characteristic core extent χ, which sets the average clone mass; the radial extent ψ within which well-developed satellite clusters are expected; and the outer limit r_{max} within which both satellite clusters and isolated demes are typically found. As the kernel exponent μ is varied, these length scales demarcate three regimes with qualitatively different core-halo relationships: compact cores with insignificant haloes, similar to the case of wavelike growth, for $\mu > d + 1$; a dominant high-occupancy core surrounded by a halo of well-developed satellite clusters which extend to a size-independent multiple of the core radius ($r_{max} \sim \psi \propto \chi$) when $d < \mu < d + 1$; and a halo including a significant number of isolated demes in addition to satellite clusters, which may extend over a region orders of magnitude larger than the core ($r_{max} \gg \psi \gg \chi$) when $\mu < d$.

We have also studied the signatures left behind by these patterns on population samples that are taken either from a local region, or globally from the entire range. Under which conditions, and for which types of samples, can we expect to observe a "soft sweep" with more than a single adaptive variant? We have found that when ranges with similar overall diversity (as judged by the number of distinct clones in the entire range) are compared, broadening the dispersal kernel has opposing effects on soft sweep detection at global and local scales: soft sweeps become harder to detect in a global random sample, but easier to detect in samples from smaller subranges.

Besides having consequences for detecting and interpreting evidence for spatial soft sweeps, the breakup of mutant clones by long-range dispersal also impacts future evolution after the soft sweep has completed. Our analysis describes the spatial patterns arising in the regime of strong selection, where the large advantage of beneficial mutants over the wildtype dominates the evolutionary dynamics. Once the entire population has adapted to the driving selection pressure, smaller fitness differences among the distinct alleles will become significant, and modify the spatial patterns on longer time scales. Selection is most sensitive to these fitness differences at the boundaries separating demes belonging to different clones. For the same global diversity, the total length of these boundaries is strongly influenced by the connectivity of clones, and grows significantly as the kernel exponent is reduced, thereby modifying the post-sweep

evolution of the population. The post-sweep evolution could also favour well-developed satellite clusters over isolated demes of one allele within a region dominated by another: isolated demes are likely to be taken over by their surrounding allele through local diffusion of individuals. Therefore, the characteristic length ψ may prove to be a relevant spatial scale for the post-sweep evolution, even in the regime $\mu < d$ where $r_{\max}$ sets the extent of the halo in the sweep patterns.

Although a quantitative evaluation of our model using real-world genomic data is beyond the scope of this work, some qualitative features of long-range dispersal can be identified in previous studies of spatial soft sweeps. The evolution of resistance to widely-adopted drugs in the malarial parasite *Plasmodium falciparium* is a well-studied example of a soft sweep arising in response to a broadly applied selective pressure. While multiple mutant haplotypes conferring resistance to pyrimethamine-based drugs have been observed across Africa and Southeast Asia, the number of distinct haplotypes is smaller than would have been expected if resistance-granting mutations were confined to their area of origin [23]; this feature has been linked to long-distance migration of parasites through their human hosts, which allowed individual haplotypes to quickly spread across disconnected parts of the globe [24]. Within the same soft sweep, high levels of spatial mixing of distinct resistant lineages was also observed in some sub-regions [25]. These observations are consistent with the contrasting effects of long-range dispersal we have quantified in our model: at a given rescaled mutation rate, dispersal reduces diversity globally, but increases the mixing of alleles locally. Advances in sequencing technology have driven rapid improvements in the spatiotemporal resolution of drug-resistance evolution studies [26], making them a promising candidate for quantitative analysis of the spatial soft sweep patterns we have described.

Many interesting questions remain to be explored, even within the confines of our lattice-based model. We report simulation results primarily in $d = 1$, although we expect the differences among the various growth regimes to also hold in higher dimensions. We have also focused on the limit in which the average clone size is many times smaller than the entire range. It would also be interesting to study the statistics of soft sweeps when the extent of the range is comparable to the characteristic length scale χ, making a soft sweep an event of low but significant probability which may vary significantly with the dispersal kernel.

The model can also be extended to include additional mechanisms involved in parallel adaptation. Besides recurring mutations, standing genetic variation (SGV) in the population is a important source of diversity for soft sweeps [3]. Long-range dispersal could impact both the spatial distribution of SGV before selection begins to act, and the spreading of alleles from distinct variational ori-

gins during the sweep [21]; both situations can be explored through extensions of our model. In the latter case, we expect the distinct regimes of core-halo patterns for different jump kernels to persist, but with the characteristic core size set by the initial distribution of variational origins rather than mutation-expansion balance.

The necessity of including heterogeneity motivates a natural set of extensions of the model. When soft sweeps arise due to mutations at different loci producing similar phenotypic effects, some variation in fitness among the distinct variants is inevitable. In panmictic models, fitness variations do not significantly affect the probability of observing a soft sweep, provided that the variations are small relative to the absolute fitness advantage of mutants over the wildtype [5]. Since spatial structure restricts competition to the geographic neighbourhood of a clone, we expect the effect of fitness variation to be even weaker than for panmictic populations, and our results should be robust to a small amount of variation in fitness effects. However, more significant variations could have differing impacts in the distinct regimes of spatial patterns. Similarly, spatial heterogeneities in the environment could lead to so-called "patchy" landscapes which lead to certain mutations being highly beneficial in some patches but neutral or even deleterious in others [27]. Convergent adaptation on patchy landscapes is likely to be significantly impacted by long-range dispersal which would allow mutations to spread efficiently to geographically separated patches.

Finally, the assumptions of strong selection and weak mutation/migration allowed us to ignore the dynamics of introduction of beneficial mutations within a deme. Relaxing these assumptions would lead us to a more general model with an additional time scale characterizing the local well-mixed dynamics at the deme level. The interplay between this time scale and the time scales governing the large-scale dynamics driven by long-range dispersal could lead to new patterns of genetic variation during convergent adaptation.

ACKNOWLEDGMENTS

Appendix A: When are rare LRD events significant?

For most species, even those that primarily spread through short-range spreading events, rare LRD events cannot be excluded. For instance, plants may deposit most of their seeds locally, but some seeds might be trans-

planted over many miles by animals. The effect of such rare events on genetic diversity can only be ignored if the typical waiting time between LRD events is longer than the time taken for the wavelike advance of beneficial mutations due to the short-range spreading dynamics to sweep through the population range. In the simplest case of species in a 1D range of linear size L with two dispersal modes — short-ranged, giving rise to a wavelike spreading at a speed v, and long-ranged jumps at a uniform rate j per unit length of the origin and of the target — the LRD events can be ignored provided $\int_0^{L/v} j(vt)(d - vt)\, dt \ll 1 \Rightarrow jL^3/(6v) \ll 1$. For expansions in d dimensions, the generalized criterion is $jL^{2d} \ll v/L$ up to a dimensionless constant of order 1.

Appendix B: Forms for $\ell(t)$ and χ in 1D

Here we describe the analytical forms for $\ell(t)$ used to compute the predictions for the characteristic length scale χ in main text Fig. 4. Ref. 17 derived asymptotic growth forms for the long-time limit of the domain core $\ell(t)$ (i.e. the region within which the occupancy of the range by an isolated domain is of order 1) for dispersal kernels with tails that fall off as $r^{-(\mu+d)}$. The forms for $d = 1$ are:

$$A \exp(Bt^\eta), \quad \eta = \log[2/(1+\mu)]/\log 2, \quad 0 < \mu < 1, \tag{B1}$$

$$A \exp\left[\frac{\log^2(Bt)}{4\log 2}\right], \qquad \mu = 1, \tag{B2}$$

$$At^{1/(\mu-1)}, \qquad 1 < \mu < 2, \tag{B3}$$

$$At \log(Bt), \qquad \mu = 2, \tag{B4}$$

$$At, \qquad \mu > 2. \tag{B5}$$

Here, A and B are magnitude scales for ℓ and t that depend on μ and on details of the dispersal kernel. (In the wavelike growth regime, $\mu > 2$, A is the front velocity of the growin domain.) To extract A and B for the specific kernels used here, we performed separate simulations in which domains were grown from a single seed at the origin at $t = 0$ up to a size of order 10^8 for $\mu \leq 1$ and 10^5 for $\mu > 1$, with the background mutation rate turned off. For each value of μ, 20 independent simulations were performed and the mass evolution over time, averaged over the independent runs, was equated to $2\ell(t)$. The measured $\ell(t)$ was fit to the growth forms to obtain A and B. Using the total mass as a proxy for $\ell(t)$ leads to an overestimate of the true size of the core, because it also counts individuals in the inevitable "halo" that exists due to jumps from the core to regions outside it during the stochastic growth process. The halo contains a fraction of the individuals in the core, which falls as μ increases. This correction is expected to provide a multiplicative constant of order 1 to $\ell(t)$, which is inconsequential to

the prediction of X_{ave} which itself equals χ only up to an overall constant for each μ.

The asymptotic forms only agree with the measured single-allele growth profiles when $\ell(t)$ has grown beyond a certain size. However, this threshold size becomes extremely large (i.e. order of the simulation range or larger) for values of μ close to 1 [17], making the asymptotic forms of limited utility to predict χ. Ref. 17 also derives an analytical scaling form for the behaviour of $\log_2 \ell(t)$ over a much broader range of times for μ close to d, which reads

$$\log_2 \ell(t) \approx \log A + \frac{2d}{\delta^2}\left[(Bt)^\zeta - \zeta \log(Bt) - 1\right], \quad (B6)$$

where $\delta = \mu - d$ and

$$\zeta = -\frac{\delta}{2d\log 2}, \qquad \delta > 0, \tag{B7}$$

$$\zeta = -\frac{\log(1 + \delta/2d)}{\log 2}, \qquad \delta < 0. \tag{B8}$$

From our fits to the single allele growth forms, we find that the scaling form is significantly more accurate than the asymptotic forms for $\mu \leq 1.4$ (except $\mu = 1$) and use it for our predictions of χ. In all cases, the forms for $\ell(t)$ with fitted values for A and B are accurate to within a few percent for $\ell(t)$ of order 20 and larger.

Once A and B are determined from the fit either to Eq. B1 or Eq. B6, the relation defining the characteristic length, Eq. 1 (main text), is solved to obtain $t^*(u)$ and $\chi_\mu(u) = \ell_\mu(t^*)$. The functional forms for χ derived assuming that $\ell(t)$ follows the asymptotic forms is provided in Table 1. For the more complex scaling form, Eq. B6, a functional form of $\chi(u)$ can be derived in terms of Lambert W-functions. For each dispersal kernel, the solution $\chi_\mu(u)$ is analytically determined taking only μ, A and B as inputs.

The characteristic length scale χ quantifies the balance between domain growth and mutations that sets the average domain size X_{ave} up to a multiplicative constant of order 1; the precise relationship between χ and X_{ave} is determined by the distribution of domain sizes about the characteristic size which is in turn established by the complete growth dynamics. We have an explicit form for the domain size distribution in the constant-velocity wavelike growth regime, $\mu > 2$ (Eqs. D2 and D3), which allows us to derive $X_{\mathrm{ave}} = 2\sqrt{2/\pi}\chi \approx 1.6\chi$ in this regime. In Fig. 4, we find that multiplicative constants close to 1.6 also lead to agreement between $X_{\mathrm{ave}}(u)$ and $\chi(u)$ for other values of μ, over many orders of magnitude of u. The agreement is weakest for high u which corresponds to small domains (average clone sizes of 100 or smaller); here the functional forms of $\ell(t)$ are least accurate and stochastic effects begin to dominate the deterministic growth implied by $\ell(t)$.
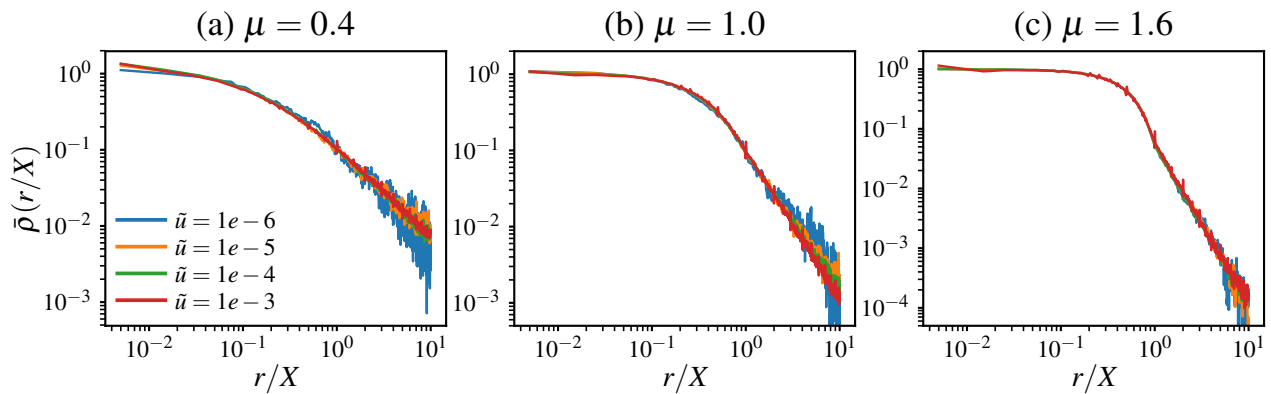
FIG. A1 **Occupancy profiles for different mutation rates collapse when the radial coordinate is rescaled by clone size.** Each panel shows the averaged occupancy profile, $\bar{\rho}(r/X)$ for a given dispersal kernel at different rescaled mutation rates. Each curve is itself an average over clones of different sizes, and the average clone sizes vary by orders of magnitude among the different values of $\tilde{u}$. Despite this variation, the profiles for a given dispersal kernel collapse onto a single curve, confirming the validity of the rescaling of the distance variable r with the total clone mass X. The smallest and largest average clone sizes (at $\tilde{u} = 1e - 3$ and $\tilde{u} = 1e - 6$ respectively) are $(130, 5.8 \times 10^4)$ for $\mu = 0.4$; $(84, 1.6 \times 10^4)$ for $\mu = 1.0$; and $(56, 4100)$ for $\mu = 1.6$.

Appendix C: Secondary length scales ψ and r_{max} in 1D

Here, we provide an alternative estimate for the length scale ψ that sets the extent of the halo of a "typical" clone, which agrees with the estimate proposed in the main text. The iterative scaling picture of Ref. 17 argues that, for growth in the marginal regime near $\mu = d$, key jumps that land at a distance $\ell(t)$ from the mutational origin typically occurred around time $t/2$ and spanned a distance of roughly $\ell(t)$ connecting source and target regions each of size $\sim \ell^d(t/2)$ (Fig. 2b). The core extent at a given time constrains the expected number of these key jumps that have contributed to the core boundary by that time: they can be neither too rare (in which case the core would not have reached the purported boundary) nor too common (which would imply that the region should have been filled much earlier). Since the number of key jumps is itself set by the extent of the core (the source for the jumps) together with the jump kernel, the above constraint equates to a self-consistency requirement on $\ell(t)$ [17]:

$$t \, \ell^{2d}(t/2) \, J[\ell(t)] \sim 1.$$

In the soft-sweep model, key jumps compete with new mutations in the target region, which occur at a rate of order $\tilde{u}\ell^d(t/2)$. The growth of the halo is obstructed by new clones when the rate of mutations arising in the target region becomes comparable to the rate of key jumps into it from the expanding core. This requires

$$\tilde{u}\ell^d(t/2) \sim J[\ell(t)]\ell^{2d}(t/2) \sim 1/t \Rightarrow t\ell^d(t/2) \sim 1/\tilde{u}. \tag{C1}$$

Up to factors of order unity, the above scaling relation is satisfied by $t = 2t^*$, where t^* was the solution to Eq. 1.

Appendix D: Exact allele frequency spectra in the panmictic and wavelike spreading limits

a. Panmictic limit

The panmictic limit in our lattice model would correspond to jumps being attempted from the source deme to a randomly chosen deme in the entire range. The allele frequency spectrum and related sampling probabilities can be computed exactly in this limit by mapping to an urn process. To see this, consider the evolution of allele frequencies in our lattice model when the fraction of wildtype sites is w and mutants occupy the lattice with individual fraction f_i for mutant i. At the next time step, the probability weight associated with picking a wildtype site to introduce a new mutation is $\tilde{u} \times Nw = \theta w$, where $\theta = \tilde{u}N$ is the initial mutation rate for the empty lattice. By contrast, the probability weight associated with picking a site of mutant type i for an attempted dispersal event is Nf_i, but only a fraction w of these attempted dispersal events is successful since the mutant only fixes in the target deme if it contains the wildtype. Therefore the probability weight of a successful reproduction of mutant i is Nwf_i. The final statistics of clone sizes is determined by the *relative* rate of mutation to reproduction at each time step [5] (unlike the times for the appearance of new clones which depends on the absolute rates), which is θ versus $n_i = Nf_i$ at all times since the wildtype fraction drops out.

The genealogy of new mutants in this model is identical to that of a stochastic process called Hoppe's urn [20], which begins with an urn containing a single black ball with an assigned probability weight θ. At any time step, a ball is picked from the urn with probability proportional to its weight. If the black ball is chosen, it is

returned along with a ball with a new colour and probability weight 1 (a new mutant). If a coloured ball is chosen, it is returned along with one copy of itself. The relative rate of mutation to the duplication of a ball with colour i is θ versus n_i at each turn, thus establishing the equivalence to our lattice model. The distributions of mutant frequencies in this urn model are the same as those for the infinite allele model at equilibrium [18]. In particular, the allele frequency spectrum is

$$f_\infty(x) = \frac{\theta}{x}(1-x)^{\theta-1}. \qquad (D1)$$

Fig. 6 shows that panmictic simulations reproduce the theoretical limit, which also persists for $\mu \approx 0.5$ in two dimensions.

The average clone size in the panmictic limit can be obtained from the allele frequency spectrum by computing the expected number of distinct clones n_c. The smallest possible clone frequency is $1/N$. Therefore, the expected number of distinct clones, n_c, is the sum of all allowed allele frequencies, i.e. $n_c = \int_{1/N}^{1} f(x)\,dx$, which can be evaluated exactly using $f(x)$ from Eq. D1. For large N, we have $n_c \approx \tilde{u}[-1+\theta+N\log N - N(\gamma+\psi_0(\theta))]$, where γ is the Euler-Mascheroni constant and ψ_0 is the digamma function. A further simplification, valid for $\theta \gg 1$, is $n_c \approx \theta \log(1/\tilde{u})$ [28]. Once n_c is computed, the average clone size is N/n_c.

b. Wavelike spreading limit

For $\mu > d+1$, domains are predicted to grow in radially expanding waves, whose speed depends on the details of the dispersal kernel. The statistics of soft sweeps in this limit was previously explored by Ralph and Coop [11], who observed the equivalence of the process in the wavelike limit to Kolmogorov-Johnson-Mehl-Avrami (KJMA) models of grain growth. KJMA models track the evolution of isotropic domains which nucleate at random positions in space at a constant rate. Nucleated domains grow isotropically at a constant front velocity until they run into other domains, leaving a boundary separating domains that nucleated at different origins. The final pattern of domains matches the spatial pattern of clones in the mutation-expansion model, where individual domains correspond to distinct mutants.

In one dimension, the final grain size distribution for a KJMA process in which each nucleation gives rise to a unique domain is known exactly [19]. Using this result, we obtain the allele frequency spectrum for wavelike growth in 1D ($\mu > 2$) as

$$f_{\mathrm{w}}(x) = \left(\frac{L}{\sqrt{2}\chi}\right)^2 p\left(\frac{Lx}{\sqrt{2}\chi}\right), \qquad (D2)$$

where $\chi = \sqrt{v/2u}$ is the characteristic length scale for domains growing with front speed v, and

$$p(s) = \frac{\sqrt{\pi}}{4}(1-\mathrm{erf}(s))\left[\sqrt{2\pi}e^{\frac{s^2}{2}}\left(s^2+1\right)\mathrm{erf}\left(\frac{s}{\sqrt{2}}\right)+2s\right], \qquad (D3)$$

where erf is the error function. The result is valid as long as the domain sizes are not limited by the range size, i.e. $L \gg \chi$.

The front velocity for arbitrary $\mu > d+1$ is not known analytically, but its limiting value for very large μ in the lattice model is known. In the limit $\mu \gg d+1$, practically all attempted jumps land exactly one lattice site away from the source (this is the lower cutoff for allowed jump distances). Isolated domains grow *via* jumps from the demes situated at the edges, only half of which are successful in advancing the front (the other half land on the occupied side of the front and have no effect). Therefore, the front velocity is $1/2$ a lattice site per generation in the large-μ limit. The frequency spectra for $\mu > d+1$ approach this limit as μ increases, see Fig. 6. We can also extract the μ-dependent front speed by a one-parameter fit of Eq. D2 to the observed frequency spectra, and obtain consistent results when performing fits at different values of the mutation rate for any given μ, as shown in Fig. A2.

Appendix E: Deterministic approximation to allele frequency spectra

The analysis of the panmictic limit in the main text revealed that the distribution of alleles as $\mu \to 0$ was identical to that of Hoppe's urn process. The continuous-time analogue of Hoppe's urn is the Yule process with immigration, in which new alleles enter the population as a Poisson process with rate θ, and already-present individuals give birth to offspring at rate 1 without death. Yule's process generates the same distribution of allele sizes as Hoppe's urn, but the continuous-time description has the advantage that the dynamics of different alleles are independent: the population of allele i at time t is proportional to e^{t-t_i} where t_i was the time at which it entered the population. Statistical properties of the allele frequencies, such as the frequency spectrum $f_\infty(x)$, can be derived efficiently within this viewpoint.

In our simulations, the growth rate of alleles is *not* constant over time even if we assume panmictic migration; the success of each birth event is proportional to the wild-type fraction w which falls as the simulation progresses. However, as we saw in the main text, the mapping to Hoppe's urn/Yule process remains exact because the rate of generation of new alleles is also proportional to w and the relative rates of birth and migration remain constant throughout the duration of the simulation in the panmictic limit. This is no longer true for $\mu > 0$ when domains grow somewhat contiguously, because the likely targets for migrants become correlated with the occupancy of the
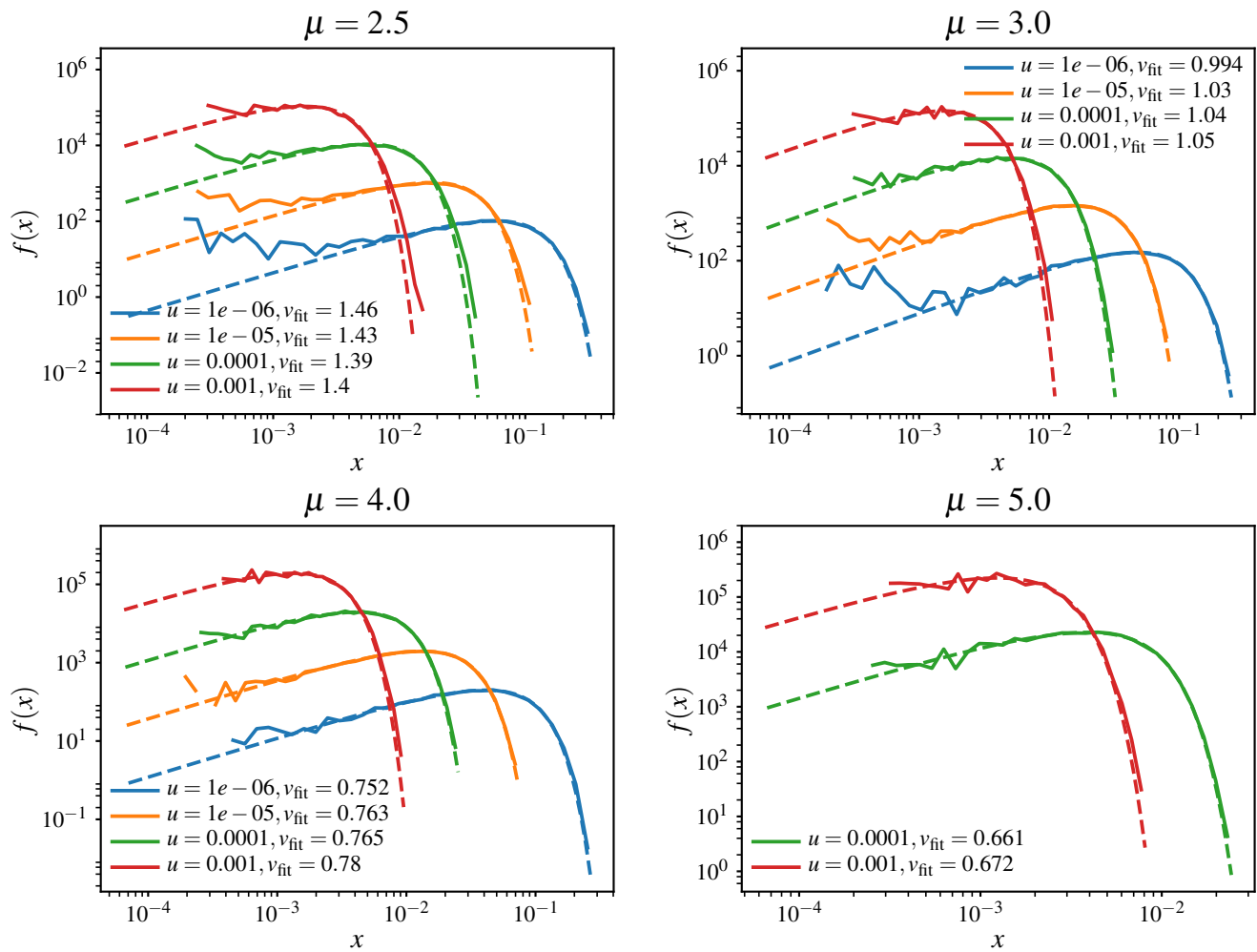
FIG. A2 **Fits to the exact frequency spectrum in the wavelike growth limit.** The measured allele frequency spectra from 1D simulations in the wavelike growth regime ($\mu > d+1$) are shown along with the theoretical form from Eqs. D2–D3. The unknown front speed v is extracted using a one-parameter nonlinear fit, and reported in units of lattice steps per generation. The fit values are consistent with v being determined by μ and independent of $\tilde{u}$. The front speed approaches the limit of $1/2$ lattice steps per generation as μ increases.

lattice and the reduction in growth rate may not simply be given by the fraction w. If we ignore these correlations, we arrive at the following approximate continuous-time model for the allele frequency spectra: process, new alleles enter the population at a constant rate θ, and grow according to the growth rule $\ell(t)$ for the particular dispersal kernel.

We can make analytical headway if we further assume that the arrival of new alleles is deterministic rather than Poisson: the kth allele enters the population at time $t_k = k/\theta$, and hence the size of the kth clone is $n_k = \ell(t - k/\theta)$. The total number of alleles, K, is fixed by the range size: $N = \sum_{k=1}^{K} n_k$. In this deterministic model, the strict time ordering of alleles implies that there are k alleles with size greater than or equal to n_k; i.e. if we can invert the n_k relation to get $k(n_k)$, this is the survival function associated with the probability distribution of

n_k and hence $x = n_k/N$. The probability distribution of x is precisely the allele frequency spectrum up to a normalization.

Below, we summarize the outcome of computing $f(x)$ according to this deterministic approximation upon using the asymptotic functional forms for $\ell(t)$ in the different regimes in 1D, summarized in Table I.

1. Power-law growth

The deterministic approach can be used to compute an approximate frequency spectrum for the growth form $\ell(t) = At^{1/(\mu-1)}$, which is the asymptotic growth rule for $1 < \mu < 2$. In this case, we have a frequency spectrum that decays as a power law: $f(x) \sim x^{\mu-2}$, up to a hard cutoff at a maximal value determined by the value of K that fills the entire range. Furthermore, the
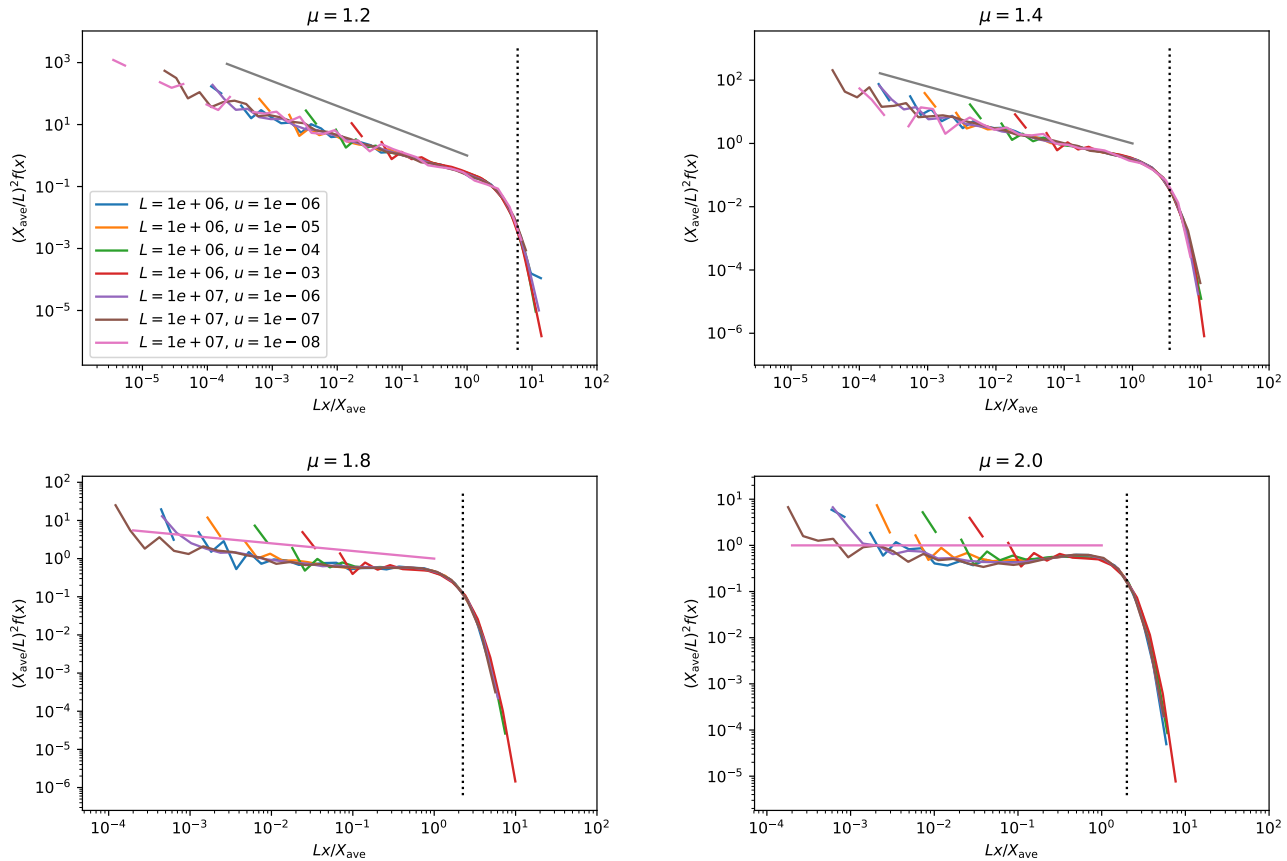
FIG. A3 **Deterministic approximation to allele frequency spectra.** Allele frequency spectra in the power-law growth regime for different mutation rates and system sizes. The rescaling is suggested by the deterministic calculation, it corresponds to a clone size distribution whose only scale is the characteristic length scale χ or equivalently the average clone size X_{ave}. The solid line is the prediction $f(y) = y^{\mu-2}$ and the vertical dashed line indicates the maximal rescaled allele frequency $\mu/(\mu-1)$ from the deterministic approximation.

form admits a rescaling that ought to collapse frequency spectra across different system sizes and mutation rates: $f(x) = (L/X_{\mathrm{ave}})^2 F(Lx/X_{\mathrm{ave}})$, where $F(y) = y^{\mu-2}$ up to the cutoff $y_{\max} = \mu/(\mu-1)$, which is the same as Eq. 3 in the main text. Fig. A3 shows that the collapse works very well across different mutation rates and two system sizes. The predicted power law for $f(x)$ is near-quantitative for all μ except $\mu = 1.2$, which is too close to the marginal case $\mu = 1$ for the asymptotic growth rule to be relevant. The predicted cutoff frequency captures the rough location of the dropoff in $f(x)$, but the deterministic approximation fails to capture the "soft shoulder" or the clones at very large frequency, which may have an outsize influence on sampling statistics.

Note that the deterministic approximation predicts a flat frequency spectrum $f(x) = $ const. for linear growth $\ell(t) = vt$, whereas the exact result for wavelike growth in 1D from the Axe and Yamada results, which we have seen to be quantitatively accurate for $\mu \gg 2$, predict a linear increase in the power spectrum $f(x) \propto x$ for small x. The difference is due to the fact that the de-

terministic approximation assumes that growth happens symmetrically toward both the left and the right at all times, whereas the wavelike growth limit is characterized by the left and right edges of the domain being interrupted independently as they run into other domains, so that one edge always advances for longer than the other. We can also explicitly include the $\log t$ correction to linear growth exactly at $\mu = 2$, and we find that the low-x behaviour is unaffected (i.e. $f(x) \sim$ const. as $x \to 0$) but there are contributions at higher x. These arise in the "shoulder" region of the spectrum, which is not captured by the deterministic analysis.

2. Marginal growth

If we use the growth form for $\mu = 1$ in the deterministic calculation, we no longer get a simple power law for $f(x)$; the functional form is instead $f(x) \sim \exp(\sqrt{a + b\log x}/\sqrt{a + b\log x}/x$ where a and b depend on the prefactors associated with $\ell(t)$ and on θ and K.
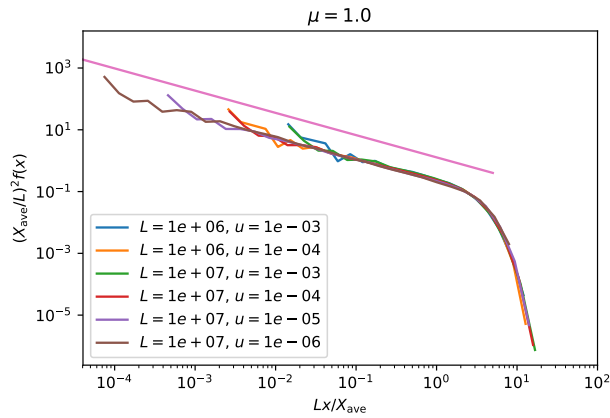
FIG. A4 **Comparison of allele frequency spectra over a range of system sizes for $\mu = 1$.** Allele frequency spectra for $\mu = 1$ over different values of L and u, rescaled according to the assumption that the only length scale for the domains is χ. The low-frequency behaviour is consistent with a power-law decay that goes as $x^{-0.72}$ (straight line).

This form is not a strict power law in x. However, when the various coefficients are computed using the full expression for $\ell(t)$ measured from the growth of single domains (Appendix B), we find that $f(x)$ behaves similar to a power law over a wide range of n_k, with an effective exponent between -0.65 and -0.85. Using the same rescaling as for the power-law growth for the measured $f(x)$ gives reasonable collapse over a range of values of u and L (Fig. A4) with a power law decay $f(x) \sim x^{-0.72}$. We note that $f(x)$ measured from simulations appears closer to a power-law form for $x \to 0$ than the deterministic approximation.

3. Stretched exponential growth

In the stretched-exponential growth regime $\mu < d$, the rescaling of the frequency spectra for a specific kernel proposed in Equation 3 is no longer exact. The rescaling assumed that χ set all length scales in the problem; this was true for power-law growth because the halo-dependent scales ψ and $r_{\max}$ were proportional to χ (with proportionality factors that depended only on μ and not on χ). By contrast, for stretched-exponential growth the additional length scales depend on the average clone sizes and hence on $\tilde{u}$. However, Fig. 6 showed that the rescaling captured much of the variation in $f(x)$ across two well-separated mutation rates, down to $\mu = 0.4$.

Although we could compute approximate frequency spectra using the deterministic calculation outlined above, they are less revealing in this regime. Instead, we gauge the inaccuracy of the proposed scaling in the panmictic limit $\mu \to 0$ where we know the exact frequency spectrum f_∞. When $N\tilde{u} = \theta \gg 1$, we have

$X_{\text{ave}} \approx -1/(\tilde{u} \log \tilde{u})$ in the panmictic limit. Using this result and the form for f_∞ in Eq. 3, we find that

$$F_\infty(y) = \frac{-1}{y \log \tilde{u}} \left(1 + \frac{y}{\theta \log \tilde{u}} \right)^{\theta - 1} \approx \frac{-1}{y \log \tilde{u}} \left(1 + \frac{y}{\log \tilde{u}} \right),$$
(E1)

where $y = Nx/X_{\text{ave}}$ and we have used $\theta \gg 1$ in the second step. We find that the function after rescaling has a residual dependence on $\log \tilde{u}$, both in the overall magnitude and in the value $y_c \sim \log \tilde{u}$ of the dropoff in f. The gentle logarithmic correction implies that the proposed rescaling still captures much of the variation with mutation rates for a given kernel, even if $\tilde{u}$ is varied by orders of magnitude, thus explaining the decent collapse of curves at different mutation rates in Fig. 6 even for $\mu < d$.

Appendix F: Allele frequency spectra with a hard cutoff

The measured allele frequency spectra display a power-law behaviour $f(x) \sim x^p$, $p > -1$ for small values of x. For cores growing as contiguous domains, balancing growth and mutation rates gives rise to a characteristic linear domain size χ (and corresponding clone size χ^d) for domain growth before a cone encounters a new mutation. In a finite range of size L^d, such growth would imply an upper bound on the allowed allele frequency at some value $x_c \sim (\chi/L)^d$. These observations suggest the ansatz for the allele frequency spectra introduced in the main text:

$$f(x) = \begin{cases} \dfrac{p+2}{x_c^{p+2}} x^p, & x < x_c \\ 0, & x > x_c, \end{cases}$$
(F1)

where the prefactor is determined by the normalization condition $\int_0^1 x \, f(x) dx = 1$.

This ansatz ignores contributions from higher-frequency clones, which are clearly significant especially for small values of μ. We can evaluate the significance of these contributions by comparing measured quantities to expectations from the hard-cutoff ansatz below.

The average clone size $X_{\text{ave}} \equiv N/n_c = N/\int f(x) dx$ can be evaluated for all $p > -1$ as

$$X_{\text{ave}} = \frac{p+1}{p+2} N x_c.$$
(F2)

The sampling probability of observing only one allele in a sample of size j evaluates to

$$P(1; j) = \int_0^1 x^j f(x) dx = \frac{p+2}{p+j+1} x_c^{j-1}$$
(F3)

which deviates weakly from the exponential falloff $P(1; j) = x^{*j-1}$ expected if all clones are of the same size and hence the same frequency x^*.

Appendix G: Sampling statistics in panmictic and wavelike growth limits

In the panmictic limit, $\mu \to 0$, sampling probabilities are known analytically for all sample sizes [18]. Using $f_\infty(x)$ in Eq. 5 gives $P_{\text{hard}} = \theta(j-1)!\Gamma(\theta)/\Gamma(j+\theta)$ [5, 18] (where Γ denotes the gamma function). The result has two distinct behaviours depending on the value of $\theta = N\tilde{u}$. When $\theta \gg 1$, an exponential falloff $P_{\text{hard}} \sim (1/\theta)^j \theta \Gamma(\theta)$ is recovered for large j, whereas for $\theta \ll 1$, $P_{\text{hard}}(j)$ falls slower than $1 - \theta \log j$.

For 1D wavelike growth with constant front velocity, Ref. [19] provides the exact form for the allele frequency spectrum, Eqs. D2–D3. The probability of observing only one allele in a random sample of size j is then $P(1;j) = \int_0^1 x^j f(x)\,dx = (\sqrt{2}\chi/L)^{j-1} \int_0^{L/\sqrt{2}\chi} s^j p(s)$. The latter integral cannot be evaluated in a closed form, even when we consider $L/\chi \gg 1$ so that the upper limit can be replaced by $s = \infty$. However, by tracking the position of the maximum value of the integrand which occurs at $x \approx \sqrt{j}$, and using Laplace's method to approximate the integral, we arrive at $\int_0^\infty s^j p(s) \approx 2j^{j/2}p(\sqrt{j})$, which provides a correction to the leading approximation $P(i;j) \approx (\sqrt{2}\chi/L)^{j-1}$.

REFERENCES

[1] H. Innan and Y. Kim, Proceedings of the National Academy of Sciences 101, 10667 (2004).

[2] M. Przeworski, G. Coop, and J. D. Wall, Evolution 59, 2312 (2005).

[3] J. Hermisson and P. S. Pennings, Genetics 169, 2335 (2005).

[4] P. S. Pennings and J. Hermisson, PLoS Genetics 2, 1998 (2006).

[5] P. S. Pennings and J. Hermisson, Molecular Biology and Evolution 23, 1076 (2006).

[6] P. W. Messer and D. A. Petrov, Trends in Ecology and Evolution 28, 659 (2013).

[7] J. Hermisson and P. S. Pennings, Methods in Ecology and Evolution 8, 700 (2017).

[8] D. P. Kwiatkowski, The American Journal of Human Genetics 77, 171 (2005).

[9] S. A. Tishkoff, F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, J. S. Silverman, K. Powell, H. M. Mortensen, J. B. Hirbo, M. Osman, M. Ibrahim, S. A. Omar, G. Lema, T. B. Nyambo, J. Ghori, S. Bumpstead, J. K. Pritchard, G. A. Wray, and P. Deloukas, Nature Genetics 39, 31 (2007), arXiv:NIHMS150003.

[10] B. L. Jones, T. O. Raga, A. Liebert, P. Zmarz, E. Bekele, E. T. Danielsen, A. K. Olsen, N. Bradman, J. T. Troelsen, and D. M. Swallow, American Journal of Human Genetics 93, 538 (2013).

[11] P. Ralph and G. Coop, Genetics 186, 647 (2010), arXiv:1005.0554.

[12] M. Kot, M. A. Lewis, and P. van den Driessche, Ecology 77, 2027 (1996).

[13] J. Clobert, M. Baguette, T. Benton, and J. Bullock, Dispersal Ecology and Evolution (OUP Oxford, 2012).

[14] J. M. Bullock, L. Mallada González, R. Tamme, L. Götzenberger, S. M. White, M. Pärtel, and D. A. P. Hooftman, Journal of Ecology 105, 6 (2017).

[15] D. Mollison, in Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 3: Probability Theory (The Regents of the University of California, 1972).

[16] M. A. Lewis and S. Pacala, Journal of Mathematical Biology 41, 387 (2000).

[17] O. Hallatschek and D. S. Fisher, Proceedings of the National Academy of Sciences 111, E4911 (2014), arXiv:arXiv:1403.4639v1.

[18] W. J. Ewens, Theoretical Population Biology 3, 87 (1972).

[19] J. D. Axe and Y. Yamada, Physical Review B 34, 1599 (1986).

[20] F. M. Hoppe, Journal of Mathematical Biology 20, 91 (1984).

[21] P. L. Ralph and G. Coop, The American Naturalist 186, S5 (2015), arXiv:/dx.doi.org/10.1101/009803 [http:].

[22] R. Nathan, Science 313, 786 (2006).

[23] C. Roper, R. Pearce, B. Bredenkamp, J. Gumede, C. Drakeley, F. Mosha, D. Chandramohan, and B. Sharp, Lancet 361, 1174 (2003).

[24] C. Roper, R. Pearce, S. Nair, B. Sharp, F. Nosten, and T. Anderson, Science 305, 1124 (2004).

[25] R. J. Pearce, H. Pota, M.-S. B. Evehe, E.-H. Bâ, G. Mombo-Ngoma, A. L. Malisa, R. Ord, W. Inojosa, A. Matondo, D. A. Diallo, W. Mbacham, I. V. van den Broek, T. D. Swarthout, A. Getachew, S. Dejene, M. P. Grobusch, F. Njie, S. Dunyo, M. Kweku, S. Owusu-Agyei, D. Chandramohan, M. Bonnet, J.-P. Guthmann, S. Clarke, K. I. Barnes, E. Streat, S. T. Katokele, P. Uusiku, C. O. Agboghoroma, O. Y. Elegba, B. Cissé, I. E. A-Elbasit, H. A. Giha, S. P. Kachur, C. Lynch, J. B. Rwakimari, P. Chanda, M. Hawela, B. Sharp, I. Naidoo, and C. Roper, PLoS Medicine 6, e1000055 (2009).

[26] L. C. Okell, J. T. Griffin, and C. Roper, Scientific Reports 7, 1 (2017).

[27] P. L. Ralph and G. Coop, PLOS Genetics 11, e1005630 (2015).

[28] G. Watterson, Theoretical Population Biology 7, 256 (1975).