The spatial distribution of rare deleterious alleles

Daniel P. Rice Christian Porras John Novembre

June 22, 2020

Abstract

[TODO]

Introduction

[TODO]

Methods

Overview

We model the movement, reproduction, and death of the carriers of a rare deleterious allele. These carriers are generated by mutations in a much larger population of wild-type individuals, whose movements and reproduction we otherwise ignore. By explicitly modeling only the rare carriers, we can approximate the evolution of their spatial distribution as a *superprocess*. Recent developments [**?**, **?**] in the theory of superprocesses allow us to find the stationary moment generating functional at mutation/selection/migration/drift balance. Using this result, we can compute the expected (joint) site frequency spectra for arbitrary spatial sampling schemes.

Population genetic model

We consider a population of organisms living in a habitat $\mathcal{D}$. We are primarily interested in populations that live in $d-$dimensional continuous habitats ($\mathcal{D} \subset \mathbb{R}^d$) or in networks of discrete demes ($\mathcal{D} \subset \mathbb{Z}$), but in the following, $\mathcal{D}$ may be an arbitrary Polish space. Let $\mathcal{N}$ be the population density so that the number of individuals living in a region $A \subset \mathcal{D}$ is given by $N_A = \int_A \mathcal{N}(x)dx$. We assume that $\mathcal{N}$ is large and stably maintained by ecological forces so that we can neglect random density fluctuations due to migration, births, and deaths. While our model may be extended to include time-varying population densities, here we will focus on spatial variation by computing stationary distributions of time-homogeneous models.

We are interested in tracking the number and spatial distribution of individuals who carry a rare deleterious variant with fitness cost $s > 0$. For $s \ll 1$, the carriers of the deleterious allele have $1 - s$ offspring on average, compared to 1 for the wild-type. (By focusing on rare alleles, we will neglect dominance effects.) We assume that wild-type individuals undergo mutation to the deleterious variant at rate μ per generation. Thus, we model the influx of de novo mutations as a Poisson point process with intensity $\mu\mathcal{N}(x)$. In general, the density of wild-type individuals is less than the total population density, $\mathcal{N}$, because $\mathcal{N}$ includes carriers. However, if $\mu \ll s$, the effect of selection dominates the resulting reduction in mutation supply and we can neglect it.

For mathematical tractability, we will assume that carriers of the deleterious allele reproduce, die, and move about the habitat independent of one another and of the background of wild-type individuals. This assumption is justified as long as the deleterious allele remains rare. In a continuous habitat—or a lattice with low occupancy—the allele is common in any sufficiently small neighborhood around a carrier. We can account for this technicality with a suitable definition of 'rare'. To wit, let ℓ be the typical distance an individual will move in its lifetime. We assume that the number of carriers in a ball $B_\ell(x)$ with radius ℓ about any point $x \in \mathcal{D}$ is small compared to $N_{B_\ell(x)}$. As ℓ is the shortest length scale in the model, it is reasonable to assume that fluctuations in allele frequency below this scale will be short-lived and not contribute to the long-term evolution of the population.

Given the rare-allele assumption, we model the movement of an individual carrier as a continuous-time Markov process on $\mathcal{D}$ with infinitessimal generator $\mathcal{L}$, independent of the positions of other carriers. If $\mathcal{D}$ is a set of n discrete demes, $\mathcal{L} = M$, an $n \times n$ matrix of migration rates where M_{ij} is the migration rate from deme i to deme j and M_{ii} is the total rate of migration out of deme i. If $\mathcal{D} \subset \mathcal{R}^d$ and movement is diffusive, $\mathcal{L} = \nabla \cdot D(\mathbf{x})\nabla$, where $D(\mathbf{x})$ is the local diffusion tensor at position $\mathbf{x} \in \mathcal{D}$. The dimension of D is length^2/time, and if time is measured in generations, $D^{1/2} = \ell$, the typical distance an individual will move in its lifetime. For translation-invariant, isotropic diffusion, D is a scalar constant and $\mathcal{L} = D\nabla^2$, the Laplacian operator.

Having defined the mutational process by which carriers are generated, and the migration process by which they move around the habitat, it remains to define the mechanism by which they die and reproduce. Consistent with our assumption that carriers behave independently, we will model reproduction as a continuous-state branching process. Combining this with our Markov process for movement yields a superprocess model of the evolution of the spatial distribution of carriers. We now define these processes and then give the main result (Eqs. 5–7) needed to calculate sample properties.

Continuous-state branching processes

A branching process models a set of particles independently undergoing random "branching" events (see [**?**]). When a particle branches, it leaves behind a random number (possibly zero) of identical offspring. A branching

process is thus defined by two parameters: the branching rate, and the distribution of offspring number. Sample paths of branching processes are functions $N(t) : [0, \infty) \to \mathbb{N}$ that count the number of extant particles as a function of time.

A continuous-state branching process (CSBP) is a stochastic process that preserves the independent branching property of the discrete branching process, but whose state space is $\mathbb{R}_{\geq 0}$ rather than $\mathbb{N}$. Formally, the CSBP is defined as a limit of branching processes with suitably rescaled time and particle counts so that sample paths converge to right continuous functions with left limits [**?**]. A CSBP is characterized by a *branching mechanism*: $\psi(u) = au + bu^2$, where a measures the rate of exponential growth or decay, and b measures the randomness introduced by branching. In our model, it will be convenient to measure time in generations, which corresponds to setting $b = 1$. In these time units, $a = -s$, the fitness cost of the deleterious mutation. (More general branching mechanisms are allowed, corresponding to highly skewed offspring number distributions, but we will not consider these here.)

There are two reasons to prefer the CSBP to a discrete branching process. The first is computational convenience: we can use the known properties of the CSBP (and the corresponding superprocess) to map our problem onto a partial differential equation (Eq. 5), which we can solve numerically using standard methods. The second is universality: a CSBP with a particular branching mechanism ψ is the limiting process for a large number of different discrete branching process models. Rather than specifying the entire distribution of offspring numbers, we need only to specify a single parameter: s. We will see that the unrealistic continuous state is washed out by our finite sampling process: sample properties depend only on the moments of the carrier distribution. In this sense, the CSBP plays the same role as the Wright-Fisher diffusion does vis-á-vis standard population genetic models. Our approach is thus related to recent methods that project the Wright-Fisher transition density onto a finite basis [**?**, **?**].

The superprocess

If the particles of a branching process move independently through a habitat $\mathcal{D}$ according to a Markov process with generator $\mathcal{L}$, the stochastic process that tracks their number and positions is known as branching diffusion. Taking the continuous-state limit of a branching diffusion, we arrive at the *superprocess* [**?**, **?**]. For overviews of the superprocess and it properties, see [**?**, **?**, **?**]. If we add an influx of particles according to the mutational point process with intensity $\mu\mathcal{N}$, we have a superprocess with immigration [**?**]. (Confusingly, in the superprocess literature, our mutation process is called "immigration" and the migration process is sometimes called "mutation".)

A superprocess $\{Z_t\}$ is a measure-valued random process. That is, the value of Z_t at a particular time is a measure on the habitat $\mathcal{D}$. Measures are defined by how they integrate functions over their domain. Accordingly, we introduce the inner product $\langle Z, f \rangle \in \mathbb{R}$, defined as:

$$\langle Z, f \rangle = \int_{\mathcal{D}} f(x) dZ(x), \tag{1}$$

for $Z \in \mathcal{M}_f(\mathcal{D})$, the set of finite measures on $\mathcal{D}$, and $f \in \mathcal{B}_+(\mathcal{D})$, the

set of nonnegative Borel measurable functions on $\mathcal{D}$ [**?**]. (If $\mathcal{D}$ is discrete, measures are sums of point masses, and the integral is a sum over $\mathbb{Z}$). In our model, Z_t counts the number of carriers in a region of space. For a region $A \subseteq \mathcal{D}$, we define the characteristic function $\chi_A(x) = 1$ for $x \in A$ and zero otherwise, so that:

$$\langle Z_t, \chi_A \rangle = \text{\# of carriers in } A, \text{ and} \tag{2}$$

$$\langle Z_t, \chi_{\mathcal{D}} \rangle = \text{total \# of carriers.} \tag{3}$$

By analogy, we can think of $\langle Z_t, f \rangle$ for an arbitrary nonnegative f as counts of carriers weighted by their positions according to f.

The probability distribution of a superprocess is characterized by its *moment generating functional* (MGF), $\Phi : \mathcal{B}_+ \to \mathbb{R}$, defined as:

$$\Phi_t[f] = \mathbb{E}\left[\exp\left(\langle Z_t, f \rangle\right)\right]. \tag{4}$$

Just as differentiating the moment generating function of a random variable gives its moments, the functional derivatives of Φ_t with respect to f give moments of the inner product $\langle Z_t, f \rangle$. Alternatively, Φ_t is the Laplace transform of the Markov transition kernel of the process.

We are now ready to state the main result that we need from the superprocess literature. [**?**] and [**?**] have shown that subcritical superprocesses (i.e., ones where $a < 0$ so that the measure decays exponentially) with immigration tend to a stationary distribution, subject to various technical conditions. In particular, for our process with mutational supply intensity $\mu\mathcal{N}$, Markov generator $\mathcal{L}$, and branching mechanism $\psi(u) = -s(x)u + u^2$,

$$\lim_{t\to\infty} \Phi_t[f] = \Phi[f] \equiv \exp\left(\int_0^\infty \langle \mu\mathcal{N}, u_t \rangle \, dt\right), \tag{5}$$

where u_t is the solution to the semilinear PDE

$$\frac{\partial}{\partial t} u_t(x) = \mathcal{L}u - su + u^2, \tag{6}$$

subject to initial condition

$$u_0(x) = f(x). \tag{7}$$

The stationary MGF, Φ, completely characterizes the counts and spatial distribution of carriers of the deleterious alleles of a population at steady-state. These patterns are due to the balance between the forces of mutation, selection, genetic drift, and migration. In the next two sections, we will show how Eq. 5 can be used to calculate the expected site frequency spectrum for a spatially localized sample from the population. The remainder of the Methods are devoted to numerical solution of Eqs. 6–7.

Spatial sampling

For a sample of n haploid genomes from the population, not labeled by their geographic origin, define the site frequency spectrum (SFS), $\xi_k^{(n)}$, to be the probability that there are k copies of the deleterious allele. In a polygenic

context, $\xi_k^{(n)}$ corresponds to the expected fraction of sites with k copies of the deleterious allele. Assuming that the samples are taken independently from the population with replacement, the number of copies of the deleterious allele is the sum of n Bernoulli trials with a random probability of success:

$$\xi_k^{(n)} = \mathbb{E}\left[\binom{n}{k} P^k (1-P)^{n-k}\right], \tag{8}$$

where P is a random variable representing the probability any particular sampled allele is deleterious. Thus, the SFS is a linear combination of the moments of P, and is completely determined by the moment generating function of P.

The distribution of P depends on (1) the locations of carrier individuals characterized by the superprocess Z, and (2) the probability we sample a carrier given its location. For a sample taken uniformly from a region $A \subseteq \mathcal{D}$, P is the fraction of carriers in A:

$$P(A) = \frac{\#\text{ of carriers in } A}{\text{total }\#\text{ of individuals in } A} = \frac{\langle Z, \chi_A \rangle}{N_A}. \tag{9}$$

If, instead, the sample is taken by choosing among a countable set of regions $\{A_i\}$ according to probabilities $\{\rho_i\}$ and then sampling uniformly within the chosen region,

$$P = \sum_i \rho_i \frac{\langle Z, \chi_{A_i} \rangle}{N_{A_i}} = \left\langle Z, \sum_i \frac{\rho_i \chi_{A_i}}{N_{A_i}} \right\rangle \equiv \left\langle Z, \tilde{\rho} \right\rangle, \tag{10}$$

where the second equality uses the linearity of inner products. For $\mathcal{D} \subset \mathbb{R}^d$, we can take the limit that the sampling probabilities vary continuously, i.e., that $\sum_i \frac{\rho_i \chi_{A_i}}{N_{A_i}} \to \tilde{\rho}(x) \in \mathcal{B}_+(\mathcal{D})$. If, additionally, the population density is constant, we can write $\tilde{\rho}(x)$ in terms of a sampling density function as $\rho(x)/\mathcal{N}$.

In any case, the sampling probability is an inner product of the random measure Z with the population-scaled sampling function $\tilde{\rho}$,

$$P = \left\langle Z, \tilde{\rho} \right\rangle. \tag{11}$$

Therefore, using Eqs. 4, 5, and 11, the moment generating function of P at steady-state is given by

$$\begin{aligned} \phi(z) &= \mathbb{E}\left[\exp\left(zP\right)\right] \\ &= \mathbb{E}\left[\exp\left(\left\langle Z, z\tilde{\rho} \right\rangle\right)\right] \\ &= \Phi[z\tilde{\rho}]. \end{aligned} \tag{12}$$

According to Eqs. 5–7, we can compute the right hand side of Eq. 12 by solving the PDE in Eq. 6 with initial condition $u_0 = z\tilde{\rho}$ and integrating over time. For small samples, it is practical to use automatic differentiation to compute a truncated power series for ϕ, and thus the moments of P. The SFS then follows from Eq. 8. For large samples, we develop a saddle point approximation to the probability density function of P, as described in the next section.

Geographically labeled samples

So far, we have considered only samples where no information is retained about the geographic origin of the sampled individuals. In practice, we often have partial geographic information in the form of, e.g., country or population labels. For such samples, it is possible to compute a *joint* site frequency spectrum of allele counts in each labeled subsample. Denoting the vector of sample sizes in each subsample as $\mathbf{n}$ and the deleterious allele counts as $\mathbf{k}$, the joint SFS is given by:

$$\xi_{\mathbf{k}}^{\mathbf{n}} = \mathbb{E}\left[\prod_i \binom{n_i}{k_i} P_i{}^{k_i}(1-P_i)^{n_i-k_i}\right], \tag{13}$$

where $\mathbf{P}$ is a random vector of probabilities of sampling the deleterious allele in each subsample, generalizing P.

Because the $\{P_i\}$ are all dependent on the underlying random measure Z, the expectation does not factorize and we need to find the joint moment generating function,

$$\begin{aligned}\phi(\mathbf{z}) &= \mathbb{E}\left[\exp\left(\mathbf{z}\cdot\mathbf{P}\right)\right] \\ &= \mathbb{E}\left[\exp\left(\left\langle Z, \sum_i z_i\tilde{\rho}_i\right\rangle\right)\right] \\ &= \Phi\left[\sum_i z_i\tilde{\rho}_i\right],\end{aligned} \tag{14}$$

where $\{\tilde{\rho}_i\}$ are the population-scaled sampling densities for each subsample. The mixed moments in Eq. 13 can be found by taking mixed partial derivatives of ϕ with respect to z_i.

Eqs. 13 and 14 cover the case where the geographic origin of every sampled individual is known. In that case $n_i = 2$ for all subsamples ($n_i = 1$ for haploids), and $\tilde{\rho}_i$ represents the uncertainty in the sampling location of individual i. Of particular interest is $\mathbf{n} = (2, 2)$, which gives the sample configurations for a pair of diploids sampled according to $(\tilde{\rho}_1, \tilde{\rho}_2)$. Below, we use this case to estimate the genetic variance explained by polygenic scores derived from one samples in region applied to individuals living in another region.

Saddlepoint approximation to the SFS

For small sample sizes, it suffices to use the Binomial theorem to expand Eq. 8 as a polynomial in P and compute the SFS for n chromosomes as a linear combination of the first $n + 1$ moments of P. We can compute these moments by automatic differentiation of the MGF or, in the spatially and temporally homogeneous case, by n convolutions of Green's functions. A difficulty arises for large n. The terms of the binomial expansion have alternating signs and widely varying magnitudes. Summing the series is thus subject to catastrophic cancellation.

We avoid this difficulty by approximating the probability density of the spatially sampled allele frequency, $\xi(p)$, rather than computing its moments. There are three advantages to this approach. First, it avoids the catastrophic cancellation described above. Second, for $n \to \infty$, $k/n \to p$, $\xi_k^{(n)} \to \xi(p)$. Therefore, $\xi(p)$ captures the shape of the SFS for large sample sizes, independent of n, and is consequently a more fundamental theoretical quantity. Finally, for a given n and k, we can compute the finite-sample SFS $\xi_k^{(n)}$ from $\xi(p)$ by quadrature using Eq. 8.

We approximate the probability density of P, $\xi(p)$, from its moment generating function, $\phi(z)$, by a standard saddlepoint approximation. Like the Gaussian approximation to a probability density, the saddlepoint approximation uses a second-order Taylor series expansion of the log MGF. However, where a Gaussian approximation uses the expansion of $\log \phi(z)$ around $z = 0$, the saddlepoint approximation expands $\log \phi(z)$ about a saddlepoint $z_s(p)$, which varies as a function of p. This affords much greater flexibility in the functional forms for $\xi(p)$ that the saddlepoint approximation can accurately approximate. Let $\phi(z) = \exp(K(z))$. Then, the saddlepoint appoximation to ϕ is given by:

$$\phi(p) \approx \frac{1}{\sqrt{2\pi K''(z_s)}} \exp\left(K(z_s) - z_s p\right), \tag{15}$$

where the saddlepoint z_s is the solution to $K'(z_s) = p$. Eq. 15 is not guaranteed to integrate to one, but may be normalized appropriately if it important to have a proper probability density.

Before moving on, we gather our important results from the superprocess and spatial sampling sections above. Using Eqs. 12 and 5–7, we have

$$K(z) = \int_0^\infty \left\langle \mu \mathcal{N}, u_t \right\rangle dt \tag{16}$$

$$\frac{\partial}{\partial t} u_t(x) = \mathcal{L} u - s u + u^2 \tag{17}$$

$$u_0(x) = z\tilde{\rho}(x). \tag{18}$$

For a given value of p, we could (1) compute K, K', and K'' by automatic differentiation, (2) solve for z_s, and (3) substitute into Eq. 15 to find $\phi(p)$. The difficulty of this procedure is that it must be redone for every p we are interested in. Instead, in the remainder of this section, we develop an approach that yields a parametric approximation to $\phi(p)$. The result is numerically convenient and interpretable in terms of population genetic intuition.

Saddlepoint approximation in the single-deme case

To demonstrate the usefulness of the saddlepoint approximation and guide our approach to the full spatial model, we first solve the single-deme case, i.e., no spatial structure. The advantage of this case is that it is exactly solvable, and, in fact, the saddlepoint approximation gives the exact solution.

In the case of a single deme with population size N, Eqs. **??**–18 reduce to:

$$K(z) = \mu N \int_0^\infty u_t dt \tag{19}$$

$$u_t' = -s u_t + u_t^2 \tag{20}$$

$$u_0 = z/N. \tag{21}$$

Eq. 20 is a Bernoulli equation, which can be solved in closed form to give:

$$u_t = \frac{sz}{e^{st}(Ns - z) + z} \tag{22}$$

$$K(z) = N\mu \log\left(\frac{Ns}{Ns - z}\right) \tag{23}$$

$$K'(z) = \frac{N\mu}{Ns - z} \tag{24}$$

$$K''(z) = \frac{N\mu}{(Ns - z)^2}. \tag{25}$$

We can now find $\phi(p)$ two ways. First, exponentiating Eq. 23 gives $\phi(p) = \left(\frac{Ns}{Ns-z}\right)^{N\mu}$, which we recognize as the moment-generating function of a Gamma distribution with rate parameter Ns and shape parameter $N\mu$. Second, we can apply the saddlepoint approximation. The first step is to solve $K'(z_s) = p$ for the saddlepoint using Eq. 24. Substituting the result, $z_s = Ns - \frac{N\mu}{p}$, into Eq. 15 gives the SFS:

$$\phi(p) \propto p^{N\mu - 1} \exp(-Nsp), \tag{26}$$

which agrees with the exact result up to a normalizing constant.

In the single-deme case, we can see the effect of our rare allele branching process approximation on the SFS. In the standard Wright-Fisher diffusion symmetric back-mutation in one deme, the steady-state population SFS is given by

$$\phi_{WF}(p) \propto \big(p(1-p)\big)^{N\mu - 1} \exp(-Nsp) \tag{27}$$

[**?**]. The branching process result varies from the Wright-Fisher result in two ways. First, it lacks the $(1-p)^{N\mu-1}$ factor, which represents saturation and back-mutation. This factor is negligible for rare mutations, $p \ll 1$. Furthermore, when $Ns \gg 1$, the exponential term dominates as p increases, so that—outside of a singularity at $p \to 1$ that represents fixation of the deleterious allele—the absolute agreement of the equations will be very good. The second disagreement between the models is that the Gamma density has support $(0, \infty)$, whereas the Wright-Fisher SFS is constrained to $(0, 1)$. This disagreement represents the fact that an allele modeled by an independent branching process cannot fix. In practice, we can patch this inconsistency by truncating $\phi(p)$ at $p = 1$. For $Ns \gg 1$, the mass in this tail is rendered negligible by the exponential term. Even for $Ns = 0$, the truncated result is sensible: it is the SFS in a neutral infinite-sites model. [NOTE: This truncation issue is another reason to use the saddlepoint approximation. If we compute the moments directly, sufficiently large moments start 'feeling' the weight above $p = 1$.]

A scaling hypothesis

1. $\mu \to 0$ limit
2. Ansatz

Numerical methods

Solving the cumulant generating functional PDE

We implement a pseudospectral PDE solver in Julia using the ode library. We then solve the nondimensionalized PDE with initial data corresponding to the sampling density, integrate the result, and rescale back to the original parameters.

Identifying the critical behavior

We explore the space of initial data to find the boundary of the region where the PDE blows up in finite time. We then use a logarithmic grid of initial conditions approaching the boundary to extract the critical exponent and parameters.

Moment computations

[TODO] We use julia's autodiff library to compute moments by differentiating the cumulant generating function. The finite-sample SFS is a polynomial in these moments.

Results

Critical behavior of the cumulant generating function

The spatially sampled SFS

The decay of explained variance with distance

Discussion