Emergence of clones in sexual populations

Emergence of clones in sexual populations

Richard A Neher[1], Marija Vucelja[2], Mark Mezard[3] and Boris I Shraiman[4]

[1] Max Planck Institute for Developmental Biology, D-72076 Tübingen, Germany
[2] Courant Institute for Mathematical Sciences, New York University, New York, NY10012, USA
[3] Laboratoire de Physique Théorique et Modèles Statistiques, CNRS and Universitè Paris-Sud, F-91405 Orsay, France
[4] Kavli Institute for Theoretical Physics and Department of Physics, University of California, Santa Barbara, CA 91306, USA
E-mail: richard.neher@tuebingen.mpg.de, vucelja@cims.nyu.edu, mezard@lptms.u-psud.fr and shraiman@kitp.ucsb.edu

Abstract. In sexual population, recombination reshuffles genetic variation and produces novel combinations of existing alleles, while selection amplifies the fittest genotypes in the population. If recombination is more rapid than selection, populations consist of a diverse mixture of many genotypes, as is observed in many populations. In the opposite regime, which is realized for example in the facultatively sexual populations that outcross in only a fraction of reproductive cycles, selection can amplify individual genotypes into large clones. Such clones emerge when the fitness advantage of some of the genotypes is large enough that they grow to a significant fraction of the population despite being broken down by recombination. The occurrence of this 'clonal condensation' depends, in addition to the outcrossing rate, on the heritability of fitness. Clonal condensation leads to a strong genetic heterogeneity of the population which is not adequately described by traditional population genetics measures, such as linkage disequilibrium. Here we point out the similarity between clonal condensation and the freezing transition in the random energy model of spin glasses. Guided by this analogy we explicitly calculate the probability, Y, that two individuals are genetically identical as a function of the key parameters of the model. While Y is the analog of the spin-glass order parameter, it is also closely related to rate of coalescence

in population genetics: two individuals that are part of the same clone have a recent common ancestor.

Contents

Genetic diversity is the fodder for natural selection and the fuel of evolution. It is generated by mutations and by recombination, which reshuffles genomes and thereby accelerates the exploration of the space of genotypes. The latter consists of all of the 2^L possible combinations of the genetic variants, a.k.a. alleles, present at L (biallelic) polymorphic loci. Of course the number of polymorphic loci, L, itself changes as new mutations arise

forming new polymorphisms while 'older' polymorphisms disappear from the population. The population itself consists of N individuals, which sample only a small fraction of the possible genotypes, i.e. $N \ll 2^L$. The dynamics in genotype space is therefore highly stochastic.

Of particular importance are those genetic polymorphisms that affect the fitness of individuals, the fitness being defined as the expected number of offspring in the next generation. Selection on its own would amplify the number of high fitness individuals and condense the population into a few 'clones' comprising a large fraction of the population. In populations of sexually reproducing organisms, the growth of such clones and the subsequent decline of genetic diversity are checked by recombination. Two parents, if chosen from different clones, produce offspring that are distinct from either parent. In obligate sexually reproducing species, the formation of clones is prevented since reproduction is coupled to recombination and no parent can produce genetically identical offspring. Many species, in particular microbial species and plants, can reproduce both by clonal reproduction (e.g., budding in yeast, selfing or vegetative reproduction in plants) or by sexual propagation. Such facultatively sexual species display a great variety in their mode of propagation, the frequency r of outcrossing, and the heritability of fitness. The latter is very important in sexual populations, as it determines to what extent recombinant offspring benefit from the same fitness advantages that made their parents successful.

The aim of this paper is to describe quantitatively the competing tendencies of natural selection and recombination with regard to genetic diversity, focusing on facultatively sexual organisms. The competition of natural selection with recombination is the dominant mechanism of evolution on relatively short time scales, on which mutational input is negligible compared to diversification by recombination. This situation is particularly relevant to adaptation following a major outcrossing event, or within a so called hybrid zone, where diverged genotypes have come together to generate a hybrid population. As this hybrid population continues to breed within itself, it can give rise to a bout of rapid adaptation, as beneficial alleles from both original populations are combined to form novel fit genotypes that spread within the hybrid population [34].

We will focus on the probability, Y, that two random individuals sampled from the population have the same genotype, i.e., are clones of each other. This quantity is important for population genetics, since it characterizes genetic diversity (its inverse is a measure of the number of dominant clones) as well as the dynamics of coalescence. Whenever two individuals are part of the same clone, they share a recent common ancestor, such that Y is proportional to the rate of pair coalescence. In the canonical theory of neutral coalescence, this rate is equal to the inverse population size. We will find here that Y, and with it the rate of coalescence, is determined by the clonal structure rather than the population size at low outcrossing rates.

Much of our analysis is presented in the context of a facultatively sexual population, where the evolving entities are individuals and their genotypes. Note, however, that some of our considerations also hold for contiguous segments of chromosomal DNA that are short enough to undergo only infrequent recombination even in obligatory sexual reproduction. In that case, we would be interested in the probability of a given chromosomal segment to be identical for a random pair of individuals drawn from the population. In this context, Y is the homozygosity of the population at this extended locus at which many different alleles segregate (this is of course a much weaker condition than clonal relation of whole genomes).

A complementary view of this probability relates it to the 'haplotype diversity', i.e., the number and population distribution of distinct genomic sequences for the chromosomal segment in question. We shall return to this important question in the discussion.

In an earlier publication [29], we have shown that clones are absent in the so called quasi-linkage equilibrium (QLE) 'phase' corresponding to frequent outcrossing limit but appear in a regime of small outcrossing frequency $r < r_c$, with r_c depending on the complexity of the fitness landscape (i.e. the extent of fitness additivity) and weakly dependent on N. We will review this finding and present a more detailed analysis of the time dependence of this condensation phenomenon, as well as its quantitative dependence on fitness additivity and hence heritability. Furthermore, we also study the extent of clonal condensation in a steady state where the fitness distribution is moving towards higher fitness at a constant velocity. We will put these results into the context of the random energy model (REM) of Statistical Physics, introduced and solved by Bernard Derrida [8, 13, 23]. In fact, Y is closely related to the Parisi order parameter and the onset of clonality is closely related to the spin-glass transition observed in simple models of disordered media such as the REM.

Below, we will first draw the analogy between the dynamics of selection in finite populations and the REM. This analogy is particularly simple for $r = 0$. Whereas the condensation transition in the REM occurs below a certain critical temperature, the transition to clonal population structure occurs beyond a certain critical time, $t > t_c(N)$. Hence the population genetic analog of temperature will be the inverse time. We shall then generalize the model in order to include (facultative) recombination and fitness landscapes with varying degrees of epistasis, i.e., genetic interactions. The results for mixed epistatic and additive models enables us to set the analysis into the context of the 'traveling wave' approximation, which has recently emerged as a powerful representation of adaptive population dynamics in genetically diverse populations [6, 9, 14, 32, 38, 42]. Our results therefore provide insight into how recombination and epistasis affect the dynamics and structure of adapting population waves and define conditions under which genetic diversity is maintained or lost.

1. Natural selection and the random energy model

In the absence of recombination or mutation, the frequency of any individual genotype increases if its fitness lies above the population mean and decreases otherwise. Identifying fitness with relative growth rate and ignoring stochastic effects, the expected number of individuals $n_i(t)$ with genotype g_i and fitness F_i obeys:

$$\dot{n}_i(t) = (F_i - \langle F \rangle_t) n_i(t) \Rightarrow n_i(t) = e^{F_i t - \int_0^t dt' \langle F \rangle_{t'}}, \tag{1}$$

where $\dot{n}_i(t)$ is the time derivative of $n_i(t)$ (we assume $n_i(0) = 1$) and the mean fitness $\langle F \rangle_t$ is defined by averaging over the whole population. Defining the rate of growth of clones by differential fitness (relative to the population mean) ensures constant population size: $\sum_i n_i(t) = N$. Since the mean fitness term is shared by every genotype in the population, the frequency of a particular genotype is given by

$$\nu_i(t) = \frac{e^{F_i t}}{\sum_j e^{F_j t}} = Z(t)^{-1} e^{F_i t}, \tag{2}$$

where $Z(t)$ is a time dependent normalization constant known as partition function. Hence the distribution of clones in the population has the Boltzmann form, where inverse time plays the role of temperature. The dynamics of the population will now depend on the initial distribution of fitness values F_i. In the generic case where fitness depends on many loci, each one giving a small contributions to fitness, the density of fitnesses $\rho(F) = \mathcal{N}(0, \sigma^2)$ is approximately Gaussian (with zero average and variance σ^2) and the statistics of clones in the populations is identical to that of the REM [8]. For small t population averages are dominated by the vicinity of the peak of $\rho(F)$, with many individual genotypes contributing. However, for $t > t_c$, the dominant contribution shifts all the way to the leading edge $F_{\max} \approx \sigma\sqrt{2\log N}$, which corresponds to the maximum fitness sampled from the Gaussian $\rho(F)$ in a population of size N. This means that with increasing time, i.e., decreasing 'effective temperature', the population shifts to fitter and fitter genotypes and eventually, for $t > t_c$, 'condenses' into the fittest. This condensation phenomenon manifests itself in a non-negligible probability Y that two randomly chosen individuals from the population have the same genotype. The latter is equal to the average squared genotype frequency in the population at time t. The average participation ratio is obtained by averaging over the fitness values $\{F_i\}$ of the N initial genotypes

$$
\begin{aligned}
\langle Y_t \rangle &= \left\langle \sum_i \left(\frac{n_i(t)}{\sum_j n_j} \right)^2 \right\rangle_{\{F_i\}} \\
&= N \int_0^\infty \mathrm{d}z\, z \int \mathrm{d}F_i\, \rho(F_i) n_i^2(t)\, \mathrm{e}^{-zn_i(t)} \left[\int \mathrm{d}F_j\, \rho(F_j)\, \mathrm{e}^{-zn_j(t)} \right]^{N-1},
\end{aligned}
\tag{3}
$$

where we have used the integral representation of $\Omega^{-2} = \int_0^\infty \mathrm{d}z\, z\, \mathrm{e}^{-z\Omega}$ and the fact that the F_i are the i.i.d. random variables sampled from $\rho(F)$. This calculation is carried out explicitly in [23, 24] and leads to the following quantitative result: in the limit of large populations, $\langle Y_t \rangle$ is given by

$$
\langle Y_t \rangle = \begin{cases} \mathcal{O}(N^{-1}) & t < t_c \\ 1 - \dfrac{t_c}{t} & t > t_c, \end{cases}
\tag{4}
$$

with $t_c \approx \sigma^{-1}\sqrt{2\log N}$. Figure 1 shows how $\langle Y_t \rangle$, measured in a computer simulation (see below), increases in time and compares to the REM prediction. Note that the sharp transition exhibited in (4) is realized only in the limit of large $\log N$, which is hard to achieve both in reality and in a numerical simulation. In both cases one expects to find a crossover rather than a sharp transition. Nevertheless, considering the idealized $N \to \infty$ limit provides a very useful scaffold for the analysis, and the REM also allows us to compute the $1/\log N$ corrections and therefore determine the detailed nature of the crossover.

Recombination and heritability

Sexual reproduction mixes the genetic material from two parents and thereby produces new genotypes from existing genetic variation. To account for recombination, we modify
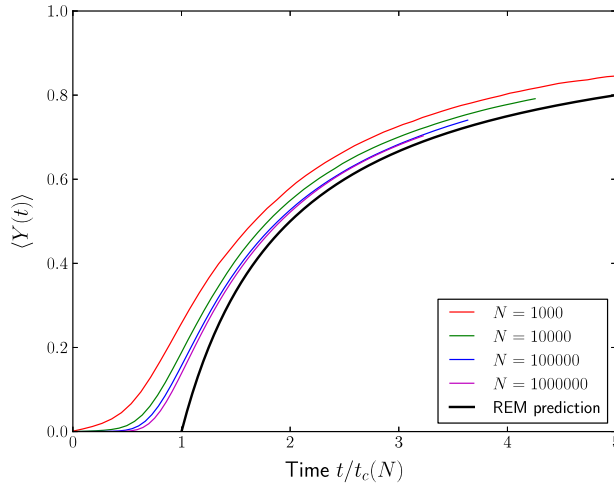
Figure 1. The participation fraction $\langle Y_t \rangle$ in the absence of recombination as a function of time. For each N, the time axis is rescaled by $t_c = \sigma^{-1}\sqrt{2 \log N \sigma}$. The fitness variance $\sigma^2 = 0.0025$ and is included in the expression for t_c to account for stochastic effects. The solid black line shows the $N \to \infty$ asymptotic result predicted by the REM (equation (4)). The convergence of the numerical results to this asymptotic result with increasing N is slow because it is governed by $\log N$. Each curve is averaged over 1000 runs.

the equation describing the evolution of genotypes as follows:

$$\dot{n}(g) = (F(g) - \langle F \rangle)n(g) + r \left[N^{-1} \sum_{g',g''} K(g|g',g'')n(g')n(g'') - n(g) \right] \quad (5)$$

where $n(g)$ is the number of individuals with genome g, and $K(g|g',g'')$ accounts for the probability that genotype g is assembled by recombination from the parental genotypes g' and g''. In the absence of recombination, the only relevant characteristic of a genotype is its fitness, and we could study the evolution of the population along the fitness coordinate instead of on the hypercube of possible genotypes. To achieve a similar simplification with recombination we need to know how the fitness of recombinant offspring relates to that of the parents to map equation (5). The correlation between parental and offspring traits is known as heritability and depends on the underlying genetic architecture. If a trait depends on many loci in the genome in an additive manner, i.e., different loci make independent contributions to the trait, the trait value of offspring will be approximately Gaussian distributed around the parental mean. Such traits are called highly heritable since the correlation between trait values of parents and offspring is high. Conversely, if a trait depends on specific combinations of alleles at many loci (epistasis), these combinations will be disrupted with high probability in sexual reproduction. Such traits have a low heritability in sexual reproduction since the correlation between parents and offspring is low.

The trait we are mainly interested in here is fitness, which in general involves many phenotypes and depends on the environment. We shall set aside the issues associated with fluctuating environment and possible time dependence of selection, and focus on how fitness depends on the genotype. As already noted, the genotype space is a
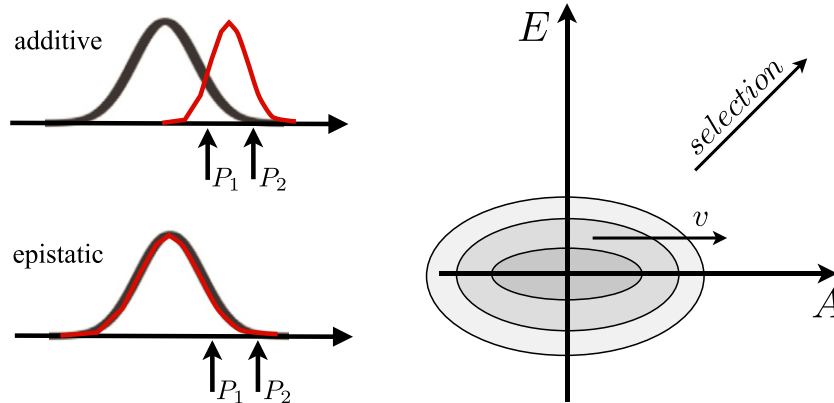
Figure 2. Heritable and non-heritable contributions to fitness. Left: the two panels illustrate heritable and non-heritable fitness functions. The black Gaussian represents the density of states over all 2^L possible genotypes, while the arrows indicate the fitness of parents $P_{1,2}$ sampled from this density of states. For additive fitness functions (top), the fitness distribution of recombinant offspring of parents $P_{1,2}$ (red curve) is centered around the mid-parent value, i.e., fitness is heritable. For completely epistatic fitness functions (bottom), the offspring fitness is a random sample from the density of states and therefore not heritable. Right: fitness is a sum of additive (heritable) and epistatic (non-heritable) components and selection amplifies individuals with large $F = A + E$. In sexual reproduction, only the additive component of fitness is heritable, so only the additive fitness increases in time with rate v.

high dimensional hypercube, $g = (s_1, \ldots, s_L)$, and for present purposes, the fitness is a complicated (but fixed) function of the genotype parameterized as

$$F(g) = f_0 + \sum_i f_i \mathbf{s}_i + \sum_{i<j} f_{ij} \mathbf{s}_i \mathbf{s}_j + \cdots, \tag{6}$$

where the terms $f_i \mathbf{s}_i$ define the additive contribution of locus i to fitness, while higher order terms correspond to epistatic interactions. The relation of the fitness of an offspring relative to that of its parents and the density of states, i.e., the distribution of fitness over all possible genomes, is illustrated in figure 2.

The higher the order of the interactions, the less likely it is that a particular set of loci that contributes to one parent is inherited uninterrupted. As a consequence, the heritability of interaction terms goes down with increasing order. Interaction terms of high order are essentially independent of the parents. This is reminiscent of high order spin-glass models, where the energies of any two configurations that differ at a macroscopic number of spins are uncorrelated. The large p limit of such p-spin glasses is the random energy model, where the energy of each configuration is an independent draw from a Gaussian distribution [8, 13].

2. The model

In general the genetic architecture of fitness is expected to be complex, with additive contributions as well as epistatic contributions of various orders. It is very instructive,

though, to consider a simplified model that includes a heritable component and a random epistatic component. Specifically, let us assume that fitness can be decomposed into an additive component A and an epistatic component E. If two individuals with fitness $F_1 = A_1 + E_1$ and $F_2 = A_2 + E_2$ produce an offspring, its additive component, A, is drawn from a Gaussian with variance $\sigma_A^2/2$ around the parental mean $(A_1 + A_2)/2$, while its epistatic component, E, is independent from that of the parents and is drawn from a Gaussian centered around 0 with variance σ_E^2. Hence the recombination kernel of this model depends only on the additive fitness of the parents

$$K(A, E|A_1, A_2) = \frac{1}{\sqrt{2\pi}\sigma_E\sigma_A} \exp\left[-\frac{(A - (A_1 + A_2)/2)^2}{\sigma_A^2} - \frac{E^2}{2\sigma_E^2}\right]. \tag{7}$$

For much of the analysis below, we will simplify this model even further and assume that the additive fitness of recombinant genotypes is independent of that of the parents and simply drawn from a Gaussian with variance σ_A^2 centered around the population mean $\langle A \rangle$. This distribution is expected if the distribution of A in the population is approximately Gaussian. This simplified model is easier to analyze, while displaying the same qualitative behavior, as has been checked by numerical simulations. The joint distribution of A and E in the population then evolves according to

$$\dot{P}(A, E) = (F - \langle F \rangle - r)P(A, E) + \frac{r}{2\pi\sigma_E\sigma_A} \exp\left[-\frac{(A - \langle A \rangle)^2}{2\sigma_A^2} - \frac{E^2}{2\sigma_E^2}\right]. \tag{8}$$

The ratio of σ_A and σ_E define the extent of fitness heritability in the process of recombination. We define 'heritability', h, which will be one of the key independent parameters of the model, as:

$$h^2 = \sigma_A^2/(\sigma_A^2 + \sigma_E^2). \tag{9}$$

To illustrate the behavior of the model at different recombination rates and different heritabilities, we implemented it as a computer simulation. The computer program keeps track of clones with fitness A_i and E_i and population size n_i. At each generation, the size of a clone is updated by a Poisson distributed number with mean $n_i\, \mathrm{e}^{F_i - \langle F \rangle - r + C}$, where $C = (1 - N/N_0)$ is a density regulating term. A clone is deleted if its size is 0. In addition, at each generation, Nr new clones of size 1 are seeded, with additive fitness drawn from a Gaussian $\mathcal{N}(\langle A \rangle, \sigma_A)$ and epistatic fitness drawn from $\mathcal{N}(0, \sigma_E)$. Alternatively, we can impose a 'velocity" of the additive fitness by drawing its value from $\mathcal{N}(vt, \sigma_A)$ instead of $\mathcal{N}(\langle A \rangle, \sigma_A)$.

Due to the stochastic nature of reproduction, the majority of all initial genotypes will rapidly die out, even if very fit. Of the N initial genotypes, only a fraction $\sim \sigma$ remains, where $\sigma^2 = \sigma_A^2 + \sigma_E^2$ is a measure of the typical strength of selection. Similarly, of the clones produced by recombination, only a fraction σ 'establishes'. Those clones that do establish are on average larger than the deterministic expectation by a factor of σ^{-1}. We will neglect these factors in most of the formulas below. The dominant effect of this stochasticity can be accounted for by rescaling N to σN inside logarithms. We will reinstantiate this correction in comparisons to simulations when necessary. These stochastic effects are of minor importance for the phenomena we study here since they are overshadowed by the randomness inherent in the fitness and seeding time of new clones.
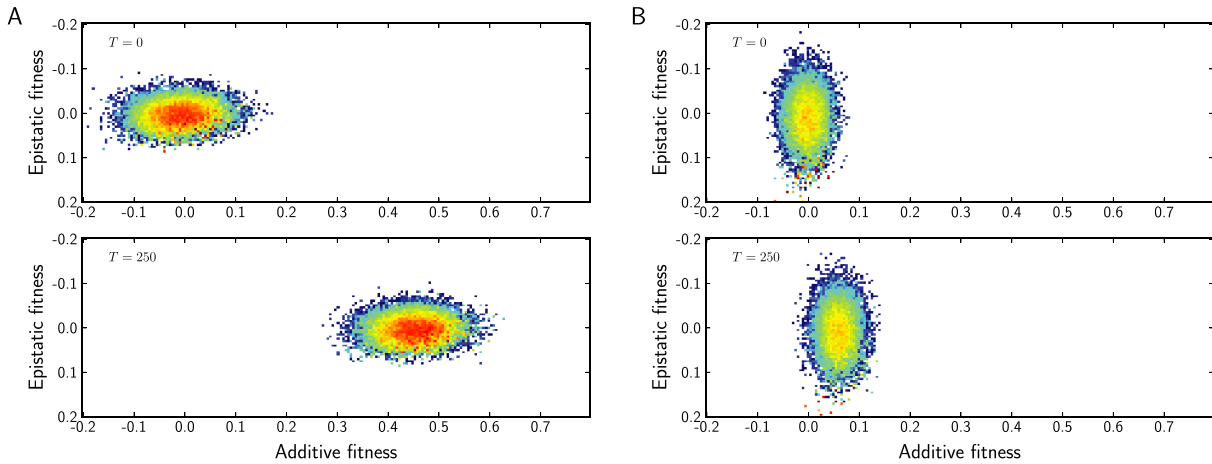
Figure 3. The distribution and dynamics of additive and epistatic fitness in the population at moderate recombination rates $r = \sigma$. Panel A: additive fitness dominates ($\sigma_A^2 = 0.8\sigma^2$, $\sigma_E^2 = 0.2\sigma^2$). Panel B: epistatic fitness dominates ($\sigma_A^2 = 0.2\sigma^2$, $\sigma_E^2 = 0.8\sigma^2$). In both panels, $\sigma^2 = 0.0025$ and $N = 10^5$.

3. Results: structure of adapting populations

Depending on the rate of recombination r and the degree of fitness heritability h, the model formulated above exhibits very different behaviors. Before we present a formal characterization of the population structure and the dynamics of evolution, it is instructive to discuss the dynamics of the model as observed in simulations.

Figure 3 shows the distribution of additive and epistatic fitness at two different times for scenarios where additive (left) or epistatic (right) fitness dominates. If fitness is predominantly additive, the population adapts and moves towards high additive fitness as long a new genotypes are generated by recombination. The velocity is given by the variance of the population along the additive direction. If most of the variance is along the epistatic direction, as in the right panel, the adaptation of the population is much slower. In addition, large fractions of the population tend to be condensed into a small number of clones, as we will discuss at greater length below. As the variance along the additive direction decreases to zero, so does the velocity.

The population structure and dynamics at different recombination rates are illustrated in figure 4. The figure shows the composition of the population as a function of time for different recombination rates and different heritabilities. In each panel, the population was initialized as a diverse sample from the density of states and was subsequently allowed to evolve via selection and recombination. Hence each panel shows a transient, which gives way to a steady state behavior. Each genotype in the population is assigned a specific color, and individuals are ordered according to fitness in each time slice, with the fittest individuals at the bottom.

The left column of figure 4 shows evolution governed by an all additive fitness function ($\sigma_A = \sigma$, $\sigma_E = 0$) for different recombination rates. In this case, the population moves steadily to higher fitness since new genotypes, fitter than their parents, are constantly produced. At low recombination rates, the population consists of a small number of clones that arise at the high fitness edge (bottom part of the panel), grow as time progresses, and

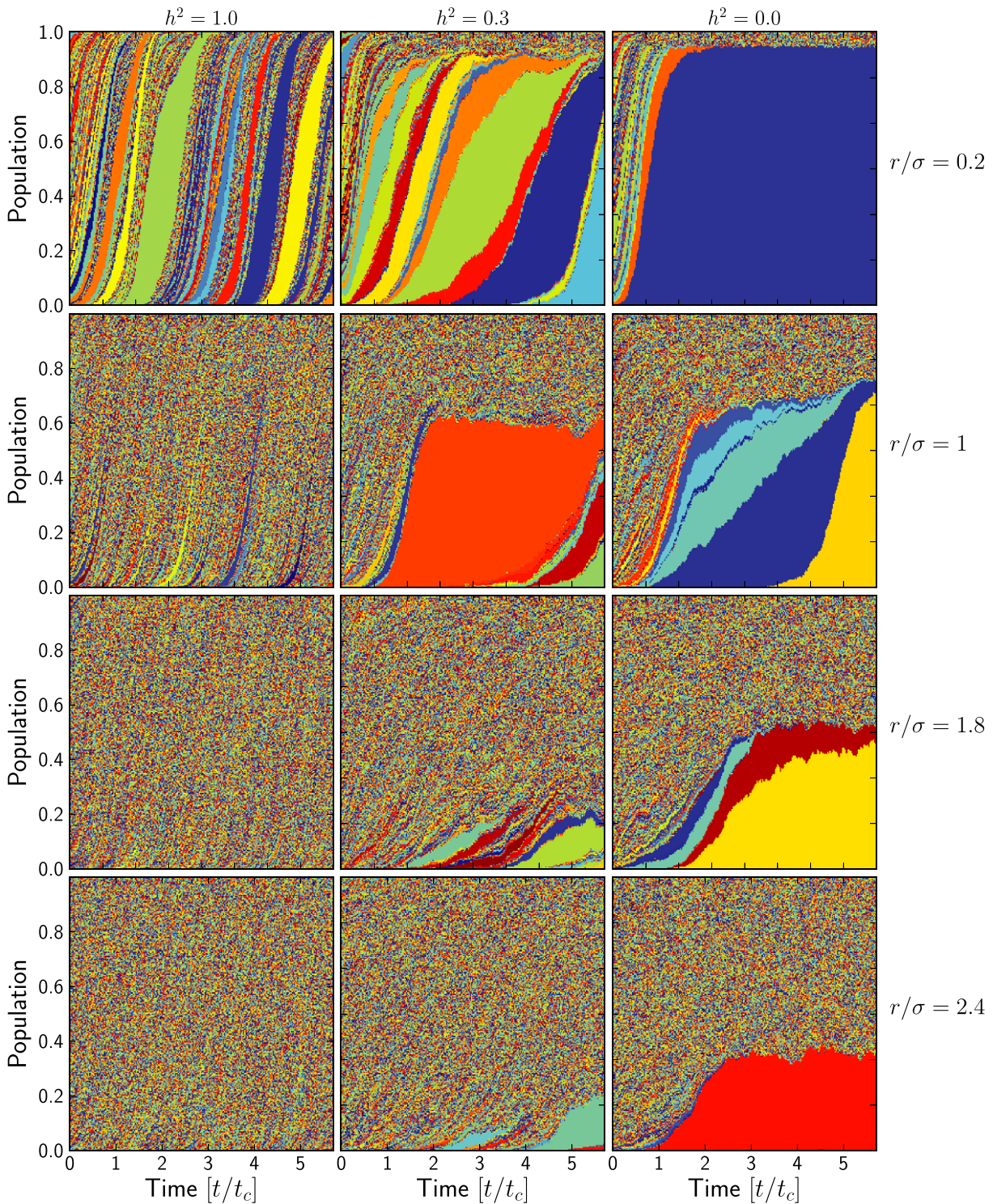**Figure 4.** Structure of population for different recombination rates and heritabilities; see text for discussion. In all panels, $\sigma^2 = 0.0025$ and $N = 10^4$.

shrink again once they fall behind in fitness (i.e., when they move towards the center of the panel). As the recombination rate is increased to values comparable to σ, large clones cease to exist. Most of the population is made from nearly unique genotypes that are

short lived. This pattern becomes even more pronounced at larger recombination rates. The high recombination limit of this dynamics can be understood in mean field theory (MFT), described in section 3.1. The low recombination limit is described in greater detail in section 3.4.

The right column of figure 4 shows the opposite limit, when fitness is not heritable but completely epistatic, corresponding to a fitness function with random high order interactions. In this case, we do not expect the population to move towards higher fitness indefinitely since the epistatic fitness is non-heritable. At low recombination rates, we expect that the fittest genotype in the population will grow while generating new recombinants distributed around zero fitness. With the growing genotypes, the mean epistatic fitness will increase until the selection on the fittest genotype, $E_{\max} - \langle E \rangle$, equals the rate at which it produces recombinants. This behavior is clearly seen in all four panels with $h^2 = 0$ on the right. Since the recombination rate increases from top to bottom, the size at which the fittest genotype stabilizes decreases, while the 'dust'-like fraction of the population that consists of short-lived unfit recombinants increases. Occasionally, recombination generates exceptionally fit genotypes, which have a chance of replacing the previous record holder. As time progresses, these records become rarer and rarer according to the well known results on record dynamics[5]. The clone structure and its dynamics for the fully epistatic case will be discussed in section 3.2. The high recombination limit can again be understood within mean field theory; see section 3.1.

The center column of figure 4 shows the clone structure and dynamics of a case of intermediate heritability. At low recombination rate, the population is dominated by clones, and clones exist even when the additive population is only dust. However, unlike the $h^2 = 0$ case, no clone dominates forever, but new clones are continuously established. The velocity in this regime will depend critically on how often such new clones are produced and on how much they advance the additive fitness. We will investigate this case below in section 3.4.

3.1. Quasi-linkage equilibrium (QLE) and its breakdown

At large r, this model admits a factorized solution $P(A, E) = \vartheta(A, t)\omega(E)$ with

$$\vartheta(A, t) = \frac{e^{-(A - \sigma_A t)^2 / 2\sigma_A^2}}{\sqrt{2\pi\sigma_A^2}} \qquad \text{and} \qquad \omega(E) = \frac{r}{r - \langle E \rangle - E} \frac{e^{-E^2/2\sigma_E^2}}{\sqrt{2\pi\sigma_E^2}} \qquad (10)$$

derived in [29]. This solution travels towards higher additive fitness with a velocity equal to the variance of the additive fitness of recombinant offspring, while the epistatic fitness has a steady distribution, where $\langle E \rangle$ adjusts itself to normalize the distribution. This factorization is a hallmark of quasi-linkage equilibrium [17, 27, 43], where additive components evolve independently, while epistatic components are in a quasi-steady balance between selection and recombination. However, this factorized solution breaks down as soon as there are individuals with epistatic fitness larger than $r + \langle E \rangle$. In that case this solution is no longer normalizable, and additive and epistatic components cease to be independent.

---

[5] Note that a slightly different dynamics is observed if recombination of individuals with the same genotype do not produce a novel genotype. In this case the effective outcrossing rate of large clones goes down and the population quickly condenses into a unique clone.
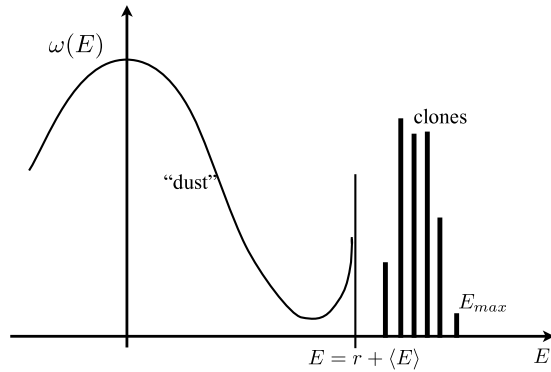
Figure 5. The population consists of mediocre recombinant genotypes (dust) and leading clones. A sketch of the population structure, indicating the smooth quasi-static distribution $\omega(E)$ at $E < \langle E \rangle + r$ and the exponentially growing clones with $E > \langle E \rangle + r$.

Figure 5 illustrates this condensation behavior. The smooth distribution $\omega(E)$ is a deformed Gaussian which diverges at $E = r + \langle E \rangle$. Beyond $r + \langle E \rangle$, the population consists of growing clones whose size depends on when and where they were seeded.

In the following we will characterize the population structure and study how the velocity in the direction of increasing additive fitness depends on parameters. We will begin by studying the purely epistatic case with recombination, which extends the REM by continuous seeding of novel recombinant genotypes. We will then study the condensation phenomenon in the presence of additive fitness, where the population forms a traveling pulse in fitness.

3.2. Clonal condensation for zero heritability

Suppose we start with a diverse sample of size N from the distribution of epistatic fitness and inject new recombinant genotypes with rate Nr. After a time t, the population consists of N initial clones and a Poisson distributed number of new clones. A clone of age τ_i and fitness $F_i = E_i$ will have approximately size

$$n_i(t) = \mathrm{e}^{(E_i - r)\tau_i + \int_{t-\tau_i}^{t} \mathrm{d}t' \, \langle E \rangle_{t'}}, \tag{11}$$

where the mean fitness $\langle E \rangle_{t'}$ can be thought of as a Lagrange parameter that keeps the overall population size constant. To calculate $\langle Y_t(r, h) \rangle$ we have to average over the fitness of the N initial clones and over the fitnesses and seeding times of all subsequently produced recombinant genotypes:

$$\langle Y_t(r, h) \rangle = \int \mathrm{d}z \, z \left[N B_0(z) + \sum_{k=1}^{\infty} \frac{\mathrm{e}^{-Nrt}}{k!} k B_r(z) \right] [1 - C_0(z)]^N C_r(z)^{k-1}, \tag{12}$$

where $B_0(z)$ is the average of $n_i^2(t) \, \mathrm{e}^{-z n_i(t)}$ over the E_i of initial genotypes (note that 'initial' means that $\tau_i = t$). $B_r(z)$ is the corresponding average over recombinant clones, which in addition are averaged over their age τ_i; see appendix A for derivation. Similarly, $C_0(z)$ and $C_r(z)$ are averages of $\mathrm{e}^{-z n_i(t)}$ over the E_i and τ_i. The sum over k—the number of recombinants generated in time t—can be performed easily. These integrals are evaluated

in the appendix B in the limit of large N. One obtains an approximation

$$\langle Y_t(r,0)\rangle \approx \exp[-N\,\mathrm{e}^{-\sqrt{2\log N}[1-(r/r_c)]t\sigma}] \tag{13}$$

valid for $r \sim r_c$, with

$$r_c \approx \sigma\sqrt{2\log N}. \tag{14}$$

We observe that the participation ratio substantially deviates from zero only for $r < r_c$ and only upon reaching $t > t_c(r)$, where for $r \approx r_c$

$$t_c(r) = \sigma^{-1}\frac{\sqrt{\frac{1}{2}\log N}}{1-\frac{r}{r_c}} = \frac{t_c(0)}{2}\frac{r_c}{r_c - r}. \tag{15}$$

This provides us with estimates of the critical time of condensation $t_c(r)$ and the critical r_c below which condensation occurs. Condensation is delayed compared to the case of $r = 0$ and the critical time (analogous to inverse critical temperature) diverges as the recombination rate approaches its critical value r_c. Of course, this divergence only occurs in the limit of $\sqrt{\log N}$ going to infinity. In practice, with realistic population sizes N, one observes not a transition, but merely a crossover between different regimes of behavior. More elaborate expressions for t_c and r_c at finite N can be found in appendix B. Note that the values of t_c and r_c themselves scale with the square root of the logarithm of N.

$\langle Y_t(r,0)\rangle$ is shown in figure 6(A) for different recombination rates. With increasing recombination rate, the early plateau of $\langle Y_t(r,0)\rangle$ is reduced and condensation is delayed. Each individual run condenses rapidly once the dominant clone is large enough that it often recombines with itself, which leaves the genotype intact. Also, the lack of condensation for $r > r_c$ is clearly seen. Figure 6(B) shows the inverse time to condensation for different recombination rates and population sizes, confirming equation (15) for $r \approx r_c$. This transition with increasing r was identified in [29]. Intuitively, the divergence of the time to condensation is due to the fact that the growth rate of best genotypes decreases with increasing recombination load r. As soon as $E - r$ is smaller than zero for all existing genotypes, all clones are short lived and no condensation can occur. Similar behavior has been observed in populations with heterozygote advantage or disadvantage [1, 11].

Population dynamics for $r < r_c$, including $t > t_c$, has the nature of a 'records process' [20, 40]. As time progresses, more and more genotypes are sampled from the density of states and tested by selection. As a consequence, the population will come across fitter and fitter genotypes, resulting in a record process with Nrt trials. Even if initially no genotype with $E - r > 0$ is present, such a genotype will eventually be found and result in condensation. On the other hand, any finite population will eventually reach a final 'record' (with fitness E_f), giving rise to the clone that will eventually take over the whole population. This is because a clone rapidly fixates once it is large enough to frequently recombine with itself. To prevent its fixation a new record would have to be created with the fitness advantage $\Delta E > \tau(E_f - r)^2/\log N$ within the time delay of τ, which is very unlikely beyond a certain E_f.

3.3. Traveling solutions for additive fitness

In the opposite limit when $h^2 = 1$, the population moves towards higher fitness with a velocity that is given by the additive variance $v = \sigma_A^2$ for sufficiently large r. It will be useful to parameterize the ratio between the velocity v and the scale of selection σ^2 as
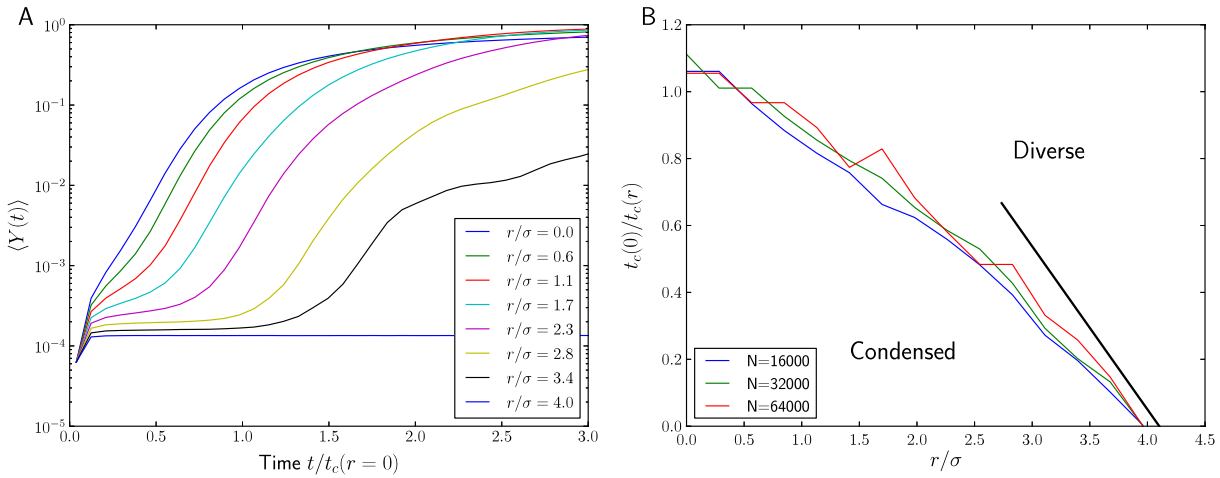
Figure 6. Panel A shows $\langle Y_t(r,0)\rangle$ for different ratios of r/σ for fully epistatic fitness functions as a function of time relative to the critical time with $r = 0$. The curves shown are averages over many runs which exhibit substantial stochasticity. Recombination delays condensation and reduces $\langle Y_t(r,0)\rangle$ early on. The model used to produce the data in this plot accounts for the fact that recombination of identical parents does not produce a novel genotype, which results in rapid condensation once one clone becomes macroscopic. Parameters $N = 16\,000$, $\sigma^2 = 0.005$, and $h^2 = 0$. Panel B shows how the condensation time $t_c(r)$ diverges as r approaches $r_c \approx E_{\max} \approx \sigma\sqrt{2\log N\sigma}$. For times and recombination rates below the lines, the population is condensed into clones; above the lines, no genotype is populated by a macroscopic fraction of the population. Here, $t_c(r)$ is defined empirically by $\langle Y_{t_c}(r)\rangle = 0.1$. The solid black line indicates the prediction of equation (15) for $r \approx r_c$.

$\gamma = v/\sigma^2$. At high recombination rates, we have $\gamma = h^2$, while we generally expect $\gamma < h^2$ at low recombination rates. In contrast to the case $h^2 = 0$, no aging dynamics is observed for $h^2 > 0$. Instead, old genotypes are constantly replaced, and dominant genotypes in the population have a finite characteristic age. Hence the initial genotypes are rapidly forgotten and we can restrict the analysis to recombinant genotypes.

Genotypes seeded at the high fitness nose of the population distribution can grow large and dominate the participation ratio $\langle Y \rangle$. Since $\langle Y \rangle$ is closely related to the rate of coalescence, it is instructive to calculate $\langle Y \rangle$ explicitly assuming $v = \sigma^2$ before considering the case $v < \sigma^2$ at low r or $h^2 < 0$.

In the steady state, a clone is specified by its age τ_j and its initial fitness A_j. Assuming $v = \sigma^2$, we then find

$$n_j = \exp\left[(A_j - r)\tau_j - \frac{v\tau_j^2}{2}\right]. \tag{16}$$

To evaluate $\langle Y(r,1)\rangle$, we have to evaluate the integrals $B_r(z)$ and $C_r(z)$ that appear in equation (12) and are defined in the appendix A, see equation (A.9). The integrals involve averaging over the initial fitness A_i and all possible ages τ_i. For $h^2 = \gamma = 1$, one finds $C_r(z) \approx z$ because relatively young and small clones dominate. In contrast, $B_r(z)$ has a significant contribution from rare old clones, which dominate because of their large size
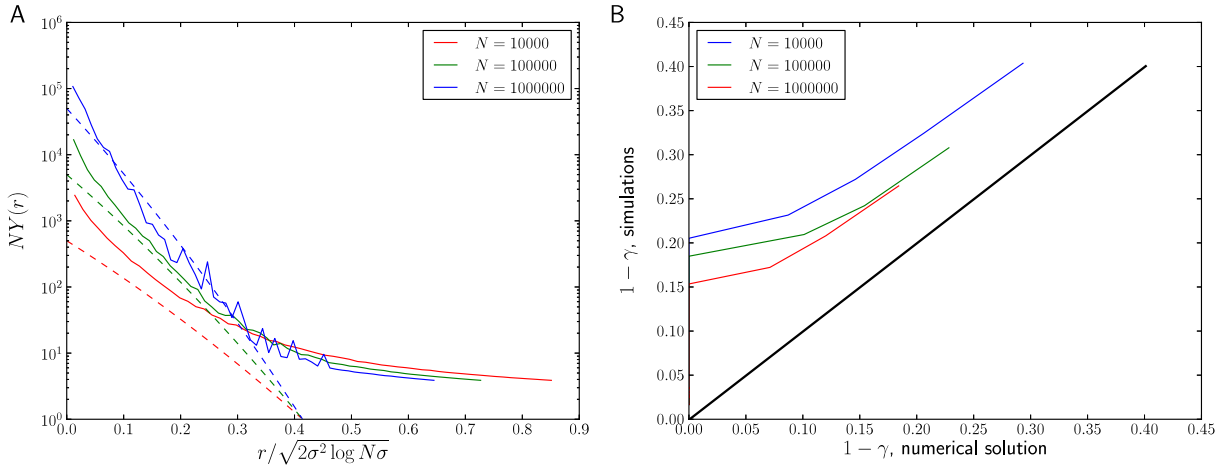
Figure 7. Panel (A): the rescaled average participation fraction $\langle NY(r)\rangle$ as a function of $r/\sqrt{2\sigma^2 \log N\sigma}$. For large r, we find $N\langle Y(r)\rangle \sim 1$, while $N\langle Y(r)\rangle \gg 1$ for small r. The dashed lines indicate the prediction by equation (18), which is expected to be valid for $r < (\sqrt{2}-1)\sqrt{2\sigma^2 \log N}$. The theory curves are scaled such that $Y = N^{-1}$ at the crossover. Panel (B) plots the reduction of velocity observed in simulations against the numerical solution of equation (19) where $\gamma = v/\sigma^2$.

(through the n_i^2 factor). The evaluation of the integrals is detailed in appendix C with the result equation (C.5). We obtain

$$B_r(z) \approx \alpha + z^{-1} \exp\left[-\frac{r}{\sigma}\sqrt{2 \log z^{-1}} - \frac{r^2}{2\sigma^2}\right]\Gamma\left(-\frac{r}{\sigma\sqrt{2 \log z^{-1}}}\right). \qquad (17)$$

The first term is of order 1 and corresponds to the contribution of young clones. Its exact value depends on the details of the stochastic dynamics, such as the offspring distribution. The second term is the contribution from old large clones. The participation ratio therefore becomes

$$\langle Y(r,1)\rangle \approx N \int_0^\infty \mathrm{d}z\, B_r(z)\, \mathrm{e}^{-Nz}$$

$$\sim \begin{cases} \alpha N^{-1} & r/\sigma > (\sqrt{2}-1)\sqrt{2 \log N}\sigma \\ \exp\left[-\dfrac{r}{\sigma}\sqrt{2 \log N\sigma} - \dfrac{r^2}{2\sigma^2}\right] & r/\sigma < (\sqrt{2}-1)\sqrt{2 \log N}\sigma\,. \end{cases} \qquad (18)$$

For small r, $\langle Y(r,1)\rangle$ does not scale as N^{-1}. In other words, the larger the population is, the larger are the clones it is composed of, and those largest clones dominate $\langle Y\rangle$. This result is in agreement with arguments made for rapid coalescence in facultatively sexual populations in [30, 36, 37]. Figure 7(A) shows $\langle NY(r)\rangle$ as a function of $r/\sqrt{2\sigma^2 \log N\sigma}$. As soon as $r/\sqrt{2\sigma^2 \log N\sigma} < 0.4$, $Y(r)$ increases and no longer scales with N, as predicted by the mean field calculation above. The figure also shows the explicit expression in equation (18) as dashed lines. Note, however, that the calculation of $\langle Y(r,1)\rangle$ involved several approximations where $\sqrt{2 \log N}$ was assumed large. As a result, the accuracy of the prefactor is low, and we have dropped all non-exponential parts, while enforcing $\langle Y(r,1)\rangle = 1/N$ at the crossover.

The derivation above is valid for $r\sigma^{-1}\sqrt{2\log N\sigma} > 1$. For smaller r, the velocity drops below the high recombination limit, $\gamma = 1$, and the fitness distribution in the population stops being Gaussian. In this regime, the velocity has to be calculated self-consistently by matching the rate of production of new extremely fit clones to the overall velocity of the population. This problem has been studied by Rouzine and Coffin [35, 37] in the context of HIV evolution.

A simplified version of the arguments of Rouzine and Coffin [35] is given in appendix D. With $\gamma = v/\sigma^2$, we find

$$1 - \gamma = \frac{\log[\frac{\sigma}{r}\sqrt{2(1-\gamma)}]}{\log[\frac{\sigma}{r}\sqrt{2(1-\gamma)}] + \log[\gamma^{-1}rN]}. \tag{19}$$

This implicit equation for γ can either be solved numerically or, in the case of very small r, iteratively. The result is similar to that obtained by Rouzine and Coffin [35]; differences are due to the difference in the model used. The numerical solution is compared to simulation data in figure 7(B). Convergence with increasing N is slow and equation (19) has a solution only for small r, for which we have little data.

The velocity of the traveling wave falls below σ^2 as soon as a few large clones dominate the population. In this case, not only $B_r(z, h)$ but also $C_r(z, h)$ are dominated by those large clones, and $\langle Y(r, h)\rangle$ behaves differently. We show in the appendix that $\langle Y(r, h)\rangle \sim 1 - \gamma$ is of order 1 (see equation (C.7)). This is in agreement with [35], who found that the typical number of large clones is $\sim \log rN\sigma \sim Y(r, h)^{-1}$.

3.4. Traveling wave solutions at intermediate heritabilities

In the high recombination limit, the population moves towards higher fitness with velocity $v = \sigma_A^2$ regardless of any epistatic fitness contributions. However, we have seen above that a clonal structure can emerge even in the absence of epistasis if recombination rates are low enough. Intuition and the simulation in figure 4 show us that this clonal structure is more pronounced in the presence of epistasis and persists at high recombination rates.

The reason for this behavior is the fact that the fitness variance of the recombinant offspring is larger than the velocity $v = \sigma_A^2 < \sigma^2$. In this case, fit genotypes grow above the traveling Gaussian envelope and generate macroscopic clones.

Figure 4 shows that, at low recombination rate and heritability, the population is always dominated by a few large clones with high non-heritable fitness components, which produce a large number of (on average) non-fit offspring. Rarely, such offspring are very fit, and replace their predecessors. The probability that offspring are fitter than their parents, and with it the velocity, increases dramatically with h^2 and r. Conversely, the size of the fit clones and the participation fraction decreases with h^2 and r. Simulation results for Y and v in the steady state are shown in figure 8 for different values of h^2 and r. In the following, we will rationalize and quantify the observed behavior. To begin with, we will assume a constant velocity and calculate Y. Again, these calculations are detailed in appendix C.

After successful establishment, a clone will grow approximately deterministically according to

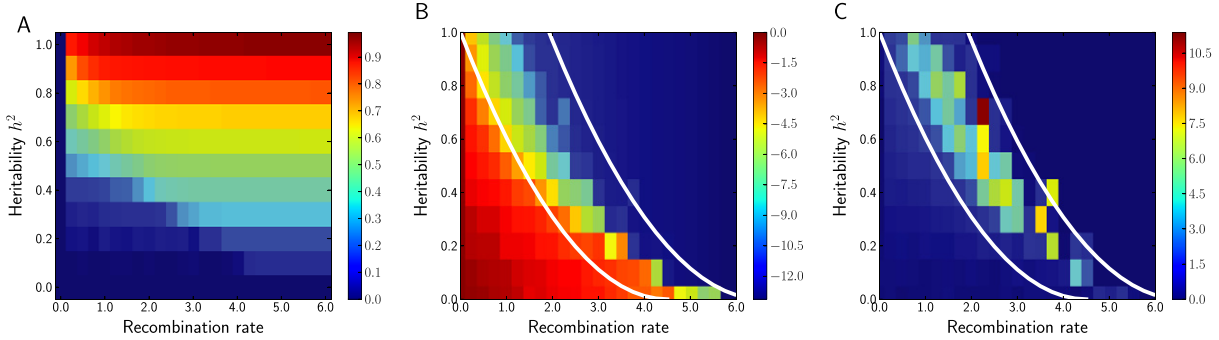$$\frac{\dot{n}_j}{n_j} = F_j - r - v\tau_j - \langle E\rangle \tag{20}$$

Figure 8. Clonal condensation as a function of recombination rate and heritability. Panel (A): ratio of the traveling wave velocity to the total variance. At low heritability and low recombination rates, the average velocity of the traveling wave is much lower than the additive variance. The equality of velocity and additive variance promised by Fisher's fundamental theorem is recovered in the high recombination limit. Panel (B) shows the corresponding participation ratio $\log\langle Y_\infty(r,h)\rangle$. Low velocity goes along with high $\langle Y_\infty(r,h)\rangle$. Panel (C) shows the coefficient of variation of the time trace Y_t, i.e., $\text{std}(Y_t)/\text{mean}(Y_t)$. Y_t fluctuates strongly at intermediate recombination rates between the two critical lines identified in the main text and indicated by the white lines. In all panels, $N = 10^6$ and $\sigma^2 = 0.0025$. The additive fitness is drawn from a Gaussian centered around the current mean additive fitness.

where $F_j = A_j + E_j$ is the sum of the additive fitness A_j, measured relative to vt when the clone was born, and E_j is the epistatic fitness. The advancing mean additive fitness is accounted for by $v\tau_j$, while $\langle E \rangle$ is the mean epistatic fitness. The size of the clone peaks at age $\tau_j^* = (F_j - \tilde{r})/v$ (where we have defined $\tilde{r} = r + \langle E \rangle$) and then decreases until the clone disappears. The maximal size of the clone is $n_j^{\max} \sim e^{-(F_j - \tilde{r})^2/2v}$. Since the fittest genotypes in the population have a fitness $\sqrt{2 \log N}$ above the mean, they grow larger than N if

$$\tilde{r} < r^* = \sigma(1 - \sqrt{\gamma})\sqrt{2 \log N\sigma} \,, \tag{21}$$

which gives us a first indication of the breakdown of the mean field solution, where $v = \sigma_A^2$. We confirm this again by calculating the integrals $C_r(z,v)$ and $B_r(z,v)$ in appendix C; see equation (C.4).

Furthermore, we calculate $\langle Y(r,h^2) \rangle$ in appendix C and find that $\langle Y(r,h^2) \rangle$ starts to be larger than N^{-1} as soon as

$$\tilde{r} < r_{\text{c}} = \sigma(\sqrt{2} - \sqrt{\gamma})\sqrt{2 \log N\sigma}. \tag{22}$$

These two conditions on $\tilde{r}$ define two critical recombination rates r_{c} and r^* at which different features of the mean field solution break down. Note, however, that this expression is not valid close to $v = 0$, since we have assumed a traveling wave with clones of finite lifetime.

Both of these two lines seem to play an important role in the behavior of the population. Between the two lines, we observe a coexistence between condensed populations and non-condensed populations, which results in large fluctuations of $Y(t)$. An example trajectory of $Y(t)$ is shown in figure 9. For a limited amount of time, the
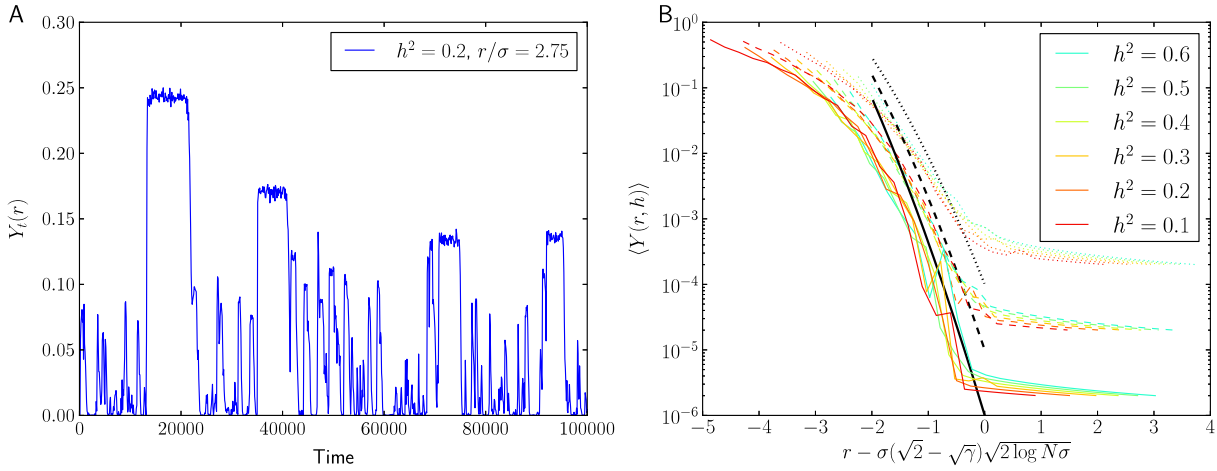
Figure 9. Panel (A) shows a time trace of $Y_t(r,h)$ which exhibits strong fluctuations. For part of the time, the population is dominated by a single clone and does not adapt. After a while, this clone is superseded and the population starts moving again. Panel (B): participation ratio for different heritabilities and different population sizes at an imposed velocity of $v = \sigma_A^2$ (the transition is sharper if v is adjusted self-consistently). Note that above the critical recombination rate, $\langle Y \rangle$ is of order N^{-1}. The analytical predictions (equation (C.7)), up to the prefactor which was fixed to ensure $Y(r_c, h^2) = N^{-1}$, are shown as black lines. Dotted lines: $N = 10^4$, dashed: $N = 10^5$, solid: $N = 10^6$, $\sigma^2 = 0.0025$.

population is condensed with large Y and does not move in the additive direction. It then becomes unstuck and adapts for a while before getting stuck again. The time average of Y is dominated by these intermittent condensates in this meta-stable region and is larger than N^{-1} as soon as $r < r_c = \sigma(\sqrt{2} - \sqrt{\gamma})\sqrt{2 \log N}$. Good simulation evidence for the different transition lines, however, is difficult to obtain because of substantial sub-leading corrections.

To investigate the model in the absence of this stick–slip behavior, we simulated a variant of the model where additive fitness is drawn from a distribution that steadily moves towards higher fitness with velocity v, as is assumed in the calculations. The time averaged $\langle Y(r,h) \rangle$ is shown in figure 9(B) and compared to the predicted onset of condensation by equation (C.7) (black lines).

4. Discussion

4.1. Summary

Correlations between different parts of the genome are usually referred to as linkage disequilibrium, suggesting that due to genetic linkage, i.e., a high probability of coinheritance, the allele frequencies at different loci are not independent. Here, we have used a different measure to characterize correlations in the population. Instead of looking at correlations between individual loci, we have characterized the distribution of clone sizes, or equivalently haplotype frequencies, in adapting populations. Our analysis is not

restricted to additive fitness functions, but we have also analyzed a simple model where fitness is partly heritable and partly random.

While macroscopic condensation, $\langle Y_\infty \rangle \sim \mathcal{O}(1)$, sets in only for $r < \sigma/\sqrt{2 \log N \sigma}$ in the absence of epistasis, we observe condensation of the population at recombination rates of order σ in presence of epistasis. Here, σ is the standard deviation in fitness and sets the strength of selection. The reason for this behavior is the fact that the velocity at which the population adapts is smaller than the fitness variance of the recombinant offspring. The additional epistatic variance allows the seeding of new clones far ahead of the population mean, which is only slowly catching up. Hence fit clones can out-grow any traveling Gaussian. At the same time, condensed clones cause the average epistatic fitness to be significantly greater than 0. Since this epistatic fitness is lost upon outcrossing, the population has a tendency to partition into a few fit clones with high epistatic fitness and a large number of recombinant genotypes with random epistatic fitness. This coexistence between 'condensed' and 'dust' phases is seen in figure 4 in the panels on the right. As long as the heritability, i.e., the fraction of additive variance, is larger than zero, the population seeds new clones even at low recombination rates and the rate of coalescence will be given by $\langle Y \rangle$ times the characteristic turn-over rate of clones.

In the absence of any additive variance, the observed behavior is quite different. In this case, the fitness function is completely random (a.k.a. House-of-cards model [18], or random epistasis/energy model [8, 29]). The only way the population adapts is by sampling fitter and fitter individuals from the same distribution. In other words, the population dynamics amounts to a record process where the total number of samples taken increases as $N(1 + rt)$. Records establish and grow with the rate $E - \langle E \rangle - r$. One additional complication that is of particular importance in the case $h^2 = 0$ is the fact that whenever a clone recombines with itself, it does not generate a novel genotype. This has the tendency to shut off recombination and stabilize clones as soon as they grow large, resulting in a rapid loss of genetic diversity. In a previous publication [29], some of us have studied the onset of condensation in a more descriptive manner. Here, we have extended this work by explicitly calculating Y, both during an initial transient as well as in a steady state where variance is maintained.

The model we have used is extremely simplistic, and one might wonder about its relation to real world populations. Nevertheless, it accounts for a number of features of real populations such as heritabilities between 0 and 1, outcrossing, and mimics a large number of loci in the sense of quantitative genetics. These features give rise to qualitatively different dynamical regimes, which will also be observable in more realistic models. Some facultatively sexual populations are in fact remarkably close to this simple 'E and A' model. Many plants and microbial populations are facultative outcrossers. In the event of outcrossing, a large number of crossovers on many chromosomes produces many independently inherited genomic segments. HIV, for example, recombines via template switching following coinfection at an outcrossing rate of a few per cent [3, 28]. In each of these outcrossing events, roughly 10 crossovers are observed [21]. If populations are polymorphic at many loci, the resulting offspring distribution is approximately described by equation (7). We have made a further simplification by assuming that the fitness of a recombinant offspring is independent of its parents and simply drawn from a comoving distribution. This assumption is expected to approximate recombination processes where the offspring and parent fitness decorrelate rapidly over a few outcrossing events, as for

19

example in equation (7) [32]. Note, however, that loci close together on the chromosome decorrelate only slowly.

The other dramatic simplification made above was the partitioning of fitness into additive and random components corresponding to high order epistatic interactions. More generally we expect to find epistatic interactions of various orders, which are heritable to different extents. However, the 'E and A' model should not be thought of as a parameterization of the genotype fitness map, but rather as a partitioning into variance that can be explained by the best fitting additive model, and the remaining variance [10, 27]. The best fitting additive model is in general time dependent, and the heritability can change as the allele frequencies change [44]. We know rather little about the genetic architecture of fitness, which justifies the use of such simple models. In specific scenarios where the genotype–phenotype map is known, more detailed modeling should be guided by the general conclusions drawn from the 'E and A' model.

The variation that we assume is always present among the recombinant offspring, it is ultimately fueled by mutations. The balance between the influx of beneficial mutations and selection in facultatively sexual populations has been investigated in [7, 32]. Even in sexual populations, the dynamics of mutation can be strongly affected by the selection through the clonal structure of the population.

4.2. Why is Y important?

The participation ratio, Y, is exactly the probability for two individuals to be genetically identical. Therefore, it immediately gives a measure of the clonal structure of the population. If the two genotypes are identical, they have had a common ancestor in the recent past. Hence, $\langle Y \rangle$ is proportional to the rate of coalescence, and as soon as $\langle Y \rangle$ is no longer proportional to N^{-1}, coalescence is greatly accelerated relative to a neutral model. It is well known that selection accelerates coalescence since fit individuals have more offspring and dominate future generations. Recombination tends to reduce the effects of linked selection since it decouples different regions of the genome. We have calculated the rate of coalescence in our model, which is set by a balance between selection and recombination. We have shown that there is a critical recombination rate, where recombination is overwhelmed by selection and the population structure changes qualitatively.

In the case of additive fitness functions, we have found that $\langle Y \rangle \sim \exp\left[-(r/\sigma)\sqrt{2\log N\sigma} - r^2/2\sigma^2\right]$, which is in agreement with earlier work [30, 37]. In this case, the population consists of clones apparent as little 'bubbles" in the representation of figure 4. Any such bubble originates from a common ancestor $\sim \sigma^{-1}\sqrt{2\log N\sigma}$ generations in the past, and $\langle Y \rangle$ is the probability that two individuals belong to the same 'bubble'. This bubble coalescent is similar to ideas developed for structured populations [33] or the fitness class coalescent [45]. Note, however, that the clone size distribution is very long tailed and the bubble coalescent is not in the universality class of the Kingman coalescent [19], but possibly of Bolthausen–Sznitman type [4, 25, 30].

Genetic identity between some of the individuals reduces the effect of outcrossing, since identical parents produce identical offspring. Since the probability of such an occurrence is equal to Y, the effective rate of recombination in the partly clonal population is $r_{\text{eff}} = r(1 - Y(r_{\text{eff}}, h))$. Hence, strictly speaking our calculations of $Y(r, h)$ deep in the clonal condensation regime must be taken through a self-consistency step, which replaces

r that was hereto an independent variable by a dependent variable r_{eff}. This however would not change our estimates for $r_c(h)$ and $t_c(r,h)$, which are defined by the point of first emergence of clones (when Y rises above $O(1/N)$). Going beyond the MFT description of recombination, one may define a participation ratio *density* $\Upsilon(F)$ in terms of which $Y = \int dF\,\Upsilon(F)$, which picks up the fitness dependence of Y: individuals with relatively high fitness are much more likely to be clonally related.

The significance of Y is not limited to the case of exact genetic identity within clones. In particular, mutations would introduce additional polymorphic loci into the 'clones' that were the focus of our study. Yet the genetic structure of the population introduced by clonal condensation still survives: one only needs to distinguish between high-frequency polymorphisms, which are being reshuffled by recombination as approximated by our model, and the low-frequency new polymorphisms due to recent (on the time scale of clonal growth) mutations. The latter would appear on the background that is clonal with respect to the high-frequency alleles. Thus the 'clones' emerging at low r should be thought of as haplotypes [22] and the 'clonal condensation' is the process that suppresses the number of haplotypes on small length scales.

The participation ratio can be readily generalized to allow for a degree of genetic distance within a pair of individuals. As currently defined $Y = \langle \delta(\|g-g'\|) \rangle$ (where $\|g-g'\|$ stands for the genetic distance between g and g'). This is immediately recognizable as a special case of the Parisi order parameter $q(x) = \langle \delta(\|g-g'\|-x) \rangle$ [23]. The latter therefore provides an interesting representation of the haplotypic structure of populations. It would be interesting to understand whether more realistic models of fitness landscapes (with low order, rather than random epistasis) would generate a more complex hierarchical structure of $q(x)$ than the simple 'dust/clone' dichotomy found in our simply model. The relation between the REM and Sherrington–Kirkpatrick models of spin glasses [39] could provide useful guidance and ultimately yield better understanding of haplotype distributions and recombinant coalescents [16, 26, 41].

4.3. Future directions

The analysis of 'clonal condensation' presented above can and should be extended in several ways. Within the confines of the model considered, one may want to obtain a better understanding of the 'mixed phase' lying between the two transition lines identified in figure 8. This phase is characterized by large fluctuations in clone size distribution, and hence in Y, even in the approximation imposing a fixed traveling wave velocity v. In reality, the population sets its own instantaneous rate of change of average fitness, which depends sensitively on the fitness of the leading clones and changes with time as new fitness 'records' are established by fresh recombinants. We have already described in figure 9 the 'stick–slip' dynamics, which is characteristic of the mixed phase regime. (Needless to say, the existence of the mixed phase region corresponds to the 1st order nature of the clonal condensation transition for $h \neq 0, 1$.) A fully quantitative description of this behavior would require going beyond MFT. So far, attempts to describe fluctuations in the dynamics of adaptation have been few and far apart [4, 14, 31]: a quantitative description of the 'stick–slip' progress of adaptation would represent a major step forward.

Another necessary extension of the model involves mutational influx. A non-zero mutation rate would provide an influx of genetic variation and define true statistically stationary states corresponding to adaptive traveling waves [6, 9, 32, 38, 42] or, in the

presence of both deleterious and beneficial mutations, to a dynamic mutation selection balance [12]. In that case, emergent clones become 'fuzzy' as they accumulate mutations, and the participation ratio should be replaced by a more general Parisi order parameter as explained above. The result should provide interesting quantitative insight into the expected genetic structure of facultatively sexual populations.

Perhaps the most interesting and important extension of the present work would involve a generalization of the model and its analysis to linear genomes and obligatory sexual populations. In contrast to the current model where recombination freely re-assorts the loci, a more realistic linear chromosome model would describe recombination in terms of relatively infrequent crossover. This would naturally tie the frequency of recombination to the length of the segment considered. We expect that on sufficiently short scales, where recombination is infrequent, a tendency for haplotype condensation would be manifest if the population is diverse enough. Whether epistasis plays a significant role in the condensation process will depend on the distribution of epistatic interactions along the genome [27]. If there is a strong tendency of mutations to interact with mutations nearby [5], the heritability decreases as smaller and smaller segments are considered. This could result in condensation of mutations into 'super-alleles'. However, the embedding of the haplotype in question into a larger genome gives rise to complications related to Hill–Robertson interference [2, 15, 16]. Transient associations with other genomic regions will either boost (the hitch-hiking process) or suppress (background selection) the spread of a haplotype in the population. This reduces the efficacy of selection and gives rise to a stochastic dynamics with rather different properties than genetic drift [30]. Bridging the different scales and resulting dynamical regimes represent an important challenge for the future.

4.4. Conclusion

In conclusion, we stress the distinction between the QLE and 'clonal condensation' regimes. In the QLE regime, recombination is sufficiently rapid to overwhelm any clonal amplification due to selection, and correlations between alleles at different loci are relatively weak. In this regime, the correlations between loci are well described by linkage disequilibrium, which measures population averaged pair correlation of loci. By contrast, clonal condensation is a non-perturbative, large deviation from linkage equilibrium (under which loci are completely uncorrelated), which in particular results in a stratification of the population depending on its fitness: clones appear predominantly in the upper reaches of the fitness distribution. Strong heterogeneity among individuals along the fitness axis is not captured by measuring linkage disequilibrium and other traditional approaches. Understanding strong interactions in multi-locus systems requires new ideas and tools. We have found simple models such as the REM to be a very useful source of insight into these non-trivial aspects of population genetics.

Acknowledgments

Appendix A. Participation ratio for clonal condensation

Following its definition for the REM we define the participation ratio for a set of clones as the sum of squared frequencies

$$\langle Y_t \rangle = \left\langle \sum_i \left(\frac{n_i(t)}{\sum_j n_j(t)} \right)^2 \right\rangle_{\{F_k\}} = \sum_i \int_0^\infty \mathrm{d}z \, z \, \left\langle n_i^2(t) \, \mathrm{e}^{-zn_i(t)} \prod_{j \neq i} \mathrm{e}^{-zn_j(t)} \right\rangle, \qquad (\text{A.1})$$

where

$$n_{F_i,\tau_i}(t) = \mathrm{e}^{(F_i - r)\tau_i - \int_{t-\tau_i}^t \mathrm{d}\tau' \langle F \rangle_{\tau'}} \qquad (\text{A.2})$$

describes the size of clone i created a time τ_i in the past with fitness F_i. Due to stochastic effects, most clones go extinct early on, and the fraction σ that remains is on average larger by a factor of σ^{-1}. We will ignore this correction and reinstantiate it later when comparing to simulations. The rate of growth of the clone is controlled by the 'replication rate': the fitness relative to the time dependent average fitness of the population $\langle F \rangle(t)$ minus the recombination rate. The average over clones $\langle \cdots \rangle$ implies averaging over F_i and τ_i. It can be decomposed into the average over the initial population (of individuals present at $t = 0$ with $\tau_i = t$) and the average over subsequent recombinants that are generated via Poisson process with rate Nr. Hence

$$\left\langle \prod_j \mathrm{e}^{-zn(t,F_j,\tau_j)} \right\rangle = \left\{ \int \mathrm{d}F \rho(F) \, \mathrm{e}^{-zn_{F,t}(t)} \right\}^N$$

$$\times \sum_{k=0}^\infty \frac{\mathrm{e}^{-Nrt}}{k!} \left\{ Nr \int_0^t \mathrm{d}\tau \int \mathrm{d}F \, \rho(F) \, \mathrm{e}^{-zn_{F,\tau}(t)} \right\}^k \qquad (\text{A.3})$$

$$= \left\{ 1 - \int \mathrm{d}F \, \rho(F)(1 - \mathrm{e}^{-zn_{F,t}(t)}) \right\}^N$$

$$\times \exp\left[-Nr \int_0^t \mathrm{d}\tau \int \mathrm{d}F \, \rho(F)(1 - \mathrm{e}^{-zn_{F,\tau}(t)}) \right] \qquad (\text{A.4})$$

$$\approx \exp\left[-NC_0(z,t) - NC_r(z,t) \right], \qquad (\text{A.5})$$

where we have defined

$$C_0(z,t) = \int \mathrm{d}F \, \rho(F)[1 - \mathrm{e}^{-zn_{F,t}(t)}], \qquad (\text{A.6})$$

which expresses the average over the initial population and

$$C_r(z,t) = r \int_0^t \mathrm{d}\tau \int \mathrm{d}F \, \rho(F)[1 - \mathrm{e}^{-zn_{F,\tau}(t)}] \qquad (\text{A.7})$$

which gives the contribution of all recombinants. If we further define

$$B_0(z,t) = \int \mathrm{d}F \, \rho(F) n_{F,t}^2(t) \, \mathrm{e}^{-zn_{F,t}(t)} = -\frac{\partial^2}{\partial z^2} C_0(z,t), \qquad (\text{A.8})$$

$$B_r(z,t) = r \int_0^t \mathrm{d}\tau \int \mathrm{d}F \, \rho(F) n_{F,\tau}^2(t) \, \mathrm{e}^{-zn_{F,\tau}(t)} = -\frac{\partial^2}{\partial z^2} C_r(z,t), \qquad (\text{A.9})$$

we arrive at the following expression for the participation fraction:

$$\langle Y_t \rangle = N \int_0^\infty \mathrm{d}z \, z [B_0(z,t) + B_r(z,t)] \mathrm{e}^{-N(C_0(z,t)+C_r(z,t))}. \tag{A.10}$$

At sufficiently long times, $rt \gg 1$, the population is dominated by new recombinants so that B_0 and C_0 may be neglected in comparison with B_r and C_r.

The growth and decay of clones, equation (A.2), depends critically on the dynamics of the mean fitness, which in turn depends on the heritability of fitness. We have discussed the different cases ($h^2 = 0, 1$ and intermediate heritabilities) in the main text and will present the calculations pertaining to them in separate appendices below. To streamline the calculation, we will rescale all rates with σ ($r \to r/\sigma$, $A \to A/\sigma$, $E \to E/\sigma$, $F \to F/\sigma$) and multiply times with σ ($t \to t\sigma$, $\tau \to \tau\sigma$). Consequently, v now equals γ. We will reinstantiate σ in the final results to facilitate comparison with formulas in the main text. In addition, the following integrals will prove useful:

$$\int \mathrm{d}x \, \mathrm{e}^{ax+2bx-\mathrm{e}^{bx}} \quad = \int_0^\infty \frac{\mathrm{d}\xi}{b} \xi^{1-(a/b)} \, \mathrm{e}^{-\xi} = \frac{1}{b}\Gamma\left(2 - \frac{a}{b}\right) \tag{A.11}$$

$$\int \mathrm{d}x \, \mathrm{e}^{ax}(1 - \mathrm{e}^{-\mathrm{e}^{bx}}) = \int_0^\infty \frac{\mathrm{d}\xi}{b} \xi^{-1-(a/b)}(1 - \mathrm{e}^{-\xi}) = \left[-\frac{1}{a}\xi^{-a/b}(1 - \mathrm{e}^{-\xi})\right] + \int_0^\infty \frac{\mathrm{d}\xi}{a} \xi^{-a/b} \, \mathrm{e}^{-\xi}$$

$$= \frac{1}{a}\Gamma\left(1 - \frac{a}{b}\right) = -\frac{1}{b}\Gamma\left(-\frac{a}{b}\right) \tag{A.12}$$

where the latter holds only for $a < b$.

Appendix B. Participation ratio for the fully epistatic $h = 0$ case

The fully epistatic case with recombination or mutation is the closest relative of the REM in population genetics, known in the population genetics literature as the 'house-of-cards' model [18]. In this model, every new genotype is drawn from the same distribution, which we will take to be the standard normal distribution, that is, fitness is non-heritable.

Let us consider $rt \gg 1$ so that enough recombinants are produced to dominate over the initial set of genotypes and evaluate $C_r(z,t)$ and $B_r(z,t)$. Furthermore, let us assume that the mean fitness is constant (we will revisit this assumption later). The size of a clone with age τ is then simply $n = \exp[(E - \tilde{r})\tau]$, and $C_r(z,t)$ is given by

$$C_{r,t}(z) \approx r \int_0^t \mathrm{d}\tau \int_{-\infty}^\infty \frac{\mathrm{d}E}{\sqrt{2\pi}} \mathrm{e}^{-E^2/2}[1 - \mathrm{e}^{-z \, \mathrm{e}^{(E-\tilde{r})\tau}}]. \tag{B.1}$$

Our strategy in approximating this integral will be to identify the region where $\Phi = z \, \mathrm{e}^{(E-\tilde{r})\tau} < 1$ and expand the $\mathrm{e}^{-\Phi}$ exponential in that region; outside that region we shall neglect the exponential compared to 1 (in the square bracket). Defining $\theta = E - \tilde{r}$, we have

$$C_{r,t}(z) \approx r \int_{-\infty}^\infty \frac{\mathrm{d}\theta}{\theta\sqrt{2\pi}} \mathrm{e}^{-(\theta+\tilde{r})^2/2} \int_0^{t\theta} \mathrm{d}\tau[1 - \mathrm{e}^{-z\mathrm{e}^\tau}] \tag{B.2}$$

$$\approx \; zr \int_{-\infty}^0 \frac{\mathrm{d}\theta}{\theta\sqrt{2\pi}} \mathrm{e}^{-(\theta+\tilde{r})^2/2}[\mathrm{e}^{\theta t} - 1]$$

$$+ \, zr \int_0^{t^{-1}\log z^{-1}} \frac{\mathrm{d}\theta}{\theta\sqrt{2\pi}} \mathrm{e}^{-(\theta+\tilde{r})^2/2} \int_0^{t\theta} \mathrm{d}\tau \, \mathrm{e}^\tau \tag{B.3}$$

$$+ \; r \int_{t^{-1}\log z^{-1}}^{\infty} \frac{\mathrm{d}\theta}{\theta\sqrt{2\pi}} \, \mathrm{e}^{-(\theta+\tilde{r})^2/2} \int_0^{t\theta} \mathrm{d}\tau[1-\mathrm{e}^{-ze^\tau}] \tag{B.4}$$

$$\approx \; z[1-\mathrm{e}^{-\tilde{r}t}] + zr \int_0^{t^{-1}\log z^{-1}} \frac{\mathrm{d}\theta}{\theta\sqrt{2\pi}} \, \mathrm{e}^{-(\theta+r)^2/2}[\mathrm{e}^{t\theta}-1]$$

$$+ \; \frac{r\,\mathrm{e}^{-E_*^2/2}}{(E_*-r)E_*\sqrt{2\pi}} \left[(1-z)+\frac{t}{E_*}\right] \tag{B.5}$$

$$\approx \; z + \frac{rt\,\mathrm{e}^{-E_*^2/2}}{(E_*-\tilde{r})E_*^2\sqrt{2\pi}}, \tag{B.6}$$

where $E_* = t^{-1}\log z^{-1} + \tilde{r}$ and we have along the way assumed that $t > E_*(t)$, which is realized for $t > t_* = \frac{\tilde{r}}{2}[1+\sqrt{1+(4/\tilde{r}^2)\log z^{-1}}]$. As we shall see below, the dominant contribution to Y comes from the region $z^{-1} \sim N$, so that in the low r regime ($\tilde{r} < \sqrt{2\log N}$) we have $t_* \approx \sqrt{\log N}$. At the end of the calculation we shall check that $t > t_*$ condition is satisfied.

At short times the second term in our expression for $C_{r,t}(z)$ is clearly smaller than the first, so that the dz integration in the expression for Y is controlled by the e^{-Nz} factor, which set the scale for z at $z \sim N^{-1}$. Let us next estimate the time for which the second term in the expression for $C_{r,t}(z)$ becomes larger than the first provided that $z \approx N^{-1}$. We obtain

$$\exp\left[\frac{(t^{-1}\log N + \tilde{r})^2}{2}\right] < \frac{rt^2 N}{(t^{-1}\log N + \tilde{r})^2 \log N \sqrt{2\pi}}. \tag{B.7}$$

Let us define $\hat{t} = t/\sqrt{\frac{1}{2}\log N}$ and $\hat{r} = r/\sqrt{2\log N}$; then

$$N^{-1+(\hat{t}^{-1}+\hat{r})^2} < \frac{\hat{r}\hat{t}^2}{\sqrt{2\log N}(\hat{t}^{-1}+\hat{r})^2\sqrt{2\pi}}. \tag{B.8}$$

For $\hat{r} > \hat{r}_{\mathrm{c}} = 1$ this can only be satisfied if $\hat{t}^2 > \hat{t}_*^2 \sim N^{\hat{r}^2-1}\sqrt{\log N}$, which diverges in the large N limit. Below r_{c}, but still close to it, we can approximate the crossover $\hat{t}_* \gg \hat{r}^{-1}$, and the crossover condition becomes

$$\hat{t}_*^{-1} \approx \frac{1-\hat{r}^2}{2\hat{r}} + \frac{\log\left[\hat{t}_{\mathrm{c}}^2/(4\hat{r}\sqrt{\pi\log N})\right]}{2\hat{r}\log N}, \tag{B.9}$$

which is an approximation valid for $r \sim r_{\mathrm{c}} = \sqrt{2\log N}$.

It remains to calculate the participation ratio as a function of t asymptotically for $t \gg t_{\mathrm{c}}$ and for $t \approx t_{\mathrm{c}}$. To do so, we need to evaluate the B-integral and perform dz integration in the integral representation for $Y(t)$. The B-integral can be obtained by differentiating $C_r(z,v)$ twice with respect to z, which evaluates approximately to

$$B(z,t) = -z^{-2}\left[t^{-2}\frac{\mathrm{d}^2}{\mathrm{d}E_*^2} + t^{-1}\frac{\mathrm{d}}{\mathrm{d}E_*}\right]\left\{\frac{rt\,\mathrm{e}^{-E_*^2/2}}{(E_*-r)E_*^2\sqrt{2\pi}}\right\} \approx z^{-2}\frac{r\,\mathrm{e}^{-E_*^2/2}}{(E_*-r)E_*\sqrt{2\pi}}. \tag{B.10}$$

Hence

$$\langle Y \rangle = N \int_0^\infty \frac{\mathrm{d}z}{z} \frac{r \exp(-E_*{}^2/2)}{(E_* - r)E_*\sqrt{2\pi}} \exp\left[-Nz - N\frac{rt \exp(-E_*^2/2)}{(E_* - r)E_*^2/\sqrt{2\pi}}\right]$$

$$\approx \int_0^\infty \mathrm{d}u \, \mathrm{e}^{-N \, \mathrm{e}^{-t(u-r)}} \frac{\mathrm{d}}{\mathrm{d}u} \exp\left[-N\frac{rt \, \mathrm{e}^{-u^2/2}}{(u-r)u^2/\sqrt{2\pi}}\right] \qquad (B.11)$$

$$\approx \int_0^\infty \mathrm{d}\phi \, \mathrm{e}^{-N \, \mathrm{e}^{-t(u(\phi)-r)} - \phi} \approx \mathrm{e}^{-N \, \mathrm{e}^{-t(u(1)-r)}},$$

where

$$\phi = \frac{Nrt \, \mathrm{e}^{-u^2/2}}{(u-r)u^2\sqrt{2\pi}} \qquad (B.12)$$

and

$$u(\phi) = \sqrt{2\log\left[\frac{Nrt}{\phi(u-r)u^2\sqrt{2\pi}}\right]} \qquad (B.13)$$

so that

$$u(1) \approx \sqrt{2\log N}\sqrt{\left[1 + \frac{\log crt}{\log N}\right]}, \qquad (B.14)$$

which upon substitution into equation (B.11) yields the final result

$$\langle Y \rangle \approx \exp\left[-N\sigma \, \mathrm{e}^{-\sqrt{2\log N\sigma}(1-(r/r_\mathrm{c}))t\sigma}\right]. \qquad (B.15)$$

This result deviates from zero for $t > t_\mathrm{c}(r) = r_\mathrm{c}\sqrt{2\log N\sigma}/2\sigma(r_\mathrm{c} - r)$, consistent with our expected t_c for $r \approx r_\mathrm{c}$. For $r > r_\mathrm{c}$ the participation ratio stays small for all t.

Appendix C. Participation ratio for the general traveling wave

As before, the participation fraction is given by an integral over clones seeded in the past. As opposed to the fully epistatic case, however, we now consider a finite velocity, which limits the lifetime of clones. A clone seeded at time τ in the past at fitness A above the mean has a size

$$n_j = \exp\left[(F_j - \tilde{r})\tau_j - \frac{v\tau_j^2}{2}\right], \qquad (C.1)$$

where $\tilde{r} = r + \langle E \rangle$. New clones get seeded with rate Nr, and to calculate $\langle Y \rangle$, we need to evaluate the integrals $C_r(z,v)$ and $B_r(z,v)$ as defined in equations (A.7) and (A.9). In contrast to the case discussed above with $v = 0$, a finite v causes clones to have a limited lifetime, even if they are initially very fit.

The integral $C_r(z,v)$ over F and τ has one contribution from young (small) clones with average fitness, in which case zn_j is small. This young contribution evaluates to

$$C_r^y(z,v) = zr \int_0^\infty \mathrm{d}\tau \int \frac{\mathrm{d}F}{\sqrt{2\pi}} \, \mathrm{e}^{-F^2/2 + (F-\tilde{r})\tau - v\tau^2/2} \approx \frac{rz}{\tilde{r}}. \qquad (C.2)$$

Here we have evaluated only the contribution from young clones and neglected the fact that there is a growing quadratic term in the exponent. The latter is due to old clones; we will evaluate this term next. If $v < 1$, the integrand starts growing again with τ and θ until it is cutoff at $\log z^{-1} = \theta\tau - v\tau^2/2$. For a given τ, this cutoff is at $\theta^* = \tau^{-1} \log z^{-1} + v\tau/2$, and the integrand has a peak at θ^* if $(1 - v)\tau > \tilde{r}$. In this case, we expand the integrand around θ^*, i.e., $F^* = \theta^* + \tilde{r}$, and find for the old clones

$$C_r^o(z, v) \approx r \int_0^\infty \frac{\mathrm{d}\tau}{\sqrt{2\pi}}\, \mathrm{e}^{-F^{*2}/2} \int \frac{\mathrm{d}\delta\,\theta}{\sqrt{2\pi}}\, \mathrm{e}^{F^*\delta\theta}(1 - \mathrm{e}^{\mathrm{e}^{\tau\delta\theta}}) = -r \int_0^\infty \frac{\mathrm{d}\tau}{\sqrt{2\pi\tau}}\, \mathrm{e}^{-F^{*2}/2}\Gamma\left(-\frac{F^*}{\tau}\right).$$
(C.3)

The remainder can be evaluated by assuming that the non-exponential parts are slowly varying. The integrand peaks roughly at $\log z^{-1}/\tau^{*2} = v/2$ or $\tau^* = \sqrt{2v^{-1}\log z^{-1}}$. This translates into $F^* = \sqrt{2v\log z^{-1}} + \tilde{r}$, and the curvature is $2F^*\log z^{-1}/\tau^{*3} = vF^*/\tau^* \approx v^2$. Hence

$$C_r^o(z, v) \approx -\frac{rz^v\, \mathrm{e}^{-\tilde{r}\sqrt{2v\log z^{-1}} - (\tilde{r}^2/2)}}{\sqrt{2v\log z^{-1}}}\Gamma\left(-v\left(1 + \frac{\tilde{r}}{\sqrt{2v\log z^{-1}}}\right)\right).$$
(C.4)

$B_r(z, v)$ can be evaluated by differentiating C twice, which yields

$$B_r^o(z, v) \approx \frac{rz^{v-2}\mathrm{e}^{-\tilde{r}\sqrt{2v\log z^{-1}} - (\tilde{r}^2/2)}}{\sqrt{2v\log z^{-1}}}\Gamma\left(2 - v\left(1 + \frac{\tilde{r}}{\sqrt{2v\log z^{-1}}}\right)\right).$$
(C.5)

Additive fitness functions

In the additive case with $v = 1$, the cutoff $C_r(z, v)$ is dominated by the young clones for any recombination rate. Its second derivative, however, is dominated by the contribution from old clones if $r < (\sqrt{2} - 1)\sqrt{2\log N}$. In this regime, we find

$$\langle Y(r, 1)\rangle \approx N \int_0^\infty \mathrm{d}z\, \frac{r\exp\left[-r\sqrt{2\log z^{-1}} - r^2/2\right]}{\sqrt{2\log z^{-1}}}\Gamma\left(1 - \frac{r}{\sqrt{2\log z^{-1}}}\right)\mathrm{e}^{-Nz}$$
$$\sim \exp\left[-\frac{r}{\sigma}\sqrt{2\log N\sigma} - \frac{r^2}{2\sigma^2}\right].$$
(C.6)

In the last step, we have dropped all preexponential factors and reinstantiated σ. This correction approximately accounts for the fact that only a fraction σ of the clones that are seeded are successful.

For larger r, even the $B_r(z, h)$ term is dominated by young clones, and $Y \sim N^{-1}$. This behavior holds while the recombination rate is larger than $1/\sqrt{2\log N}$. For smaller r, the velocity starts to deviate from 1. In this case, $C_r(z, h)$ is dominated by old clones. For large enough populations, the powers of z vary much more quickly than all other terms, which we denote collectively by $A(z)$, and we can approximate the integral as

$$\langle Y(r, h)\rangle = -\int_0^\infty \mathrm{d}z\, z^{v-1}A(z)\Gamma\left(2 - \frac{v(\theta^* + r)}{\theta^*}\right)\mathrm{e}^{z^v A(z)\Gamma(-v(\theta^* + r)/\theta^*)}$$
$$= -v^{-1}\frac{\Gamma(2 - v(\theta^* + r)/\theta^*)}{\Gamma(-v(\theta^* + r)/\theta^*)} = \frac{\theta^* + r}{\theta^*}\left(1 - \frac{v(\theta^* + r)}{\theta^*}\right) \approx 1 - v = 1 - \gamma,$$
(C.7)

where in the last step we reintroduced σ by replacing v with $\gamma = v/\sigma^2$. Along the way, we have assumed $r \ll \sigma\sqrt{2\log N\sigma}$, which is consistent with the assumption of small r made above.

Epistatic fitness function

In cases with heritabilities $0 < h^2 < 1$, we generally have $v < 1$ and can be in a regime where $C_r(z, v) \approx z$, while $B_r(z, v)$ is dominated by old clones. In this case, we have

$$\langle Y(r, h) \rangle \approx N \int_0^\infty \mathrm{d}z \, z^{v-1} \frac{r \, \mathrm{e}^{-r\sqrt{2v\log z^{-1}} - r^2/2}}{\sqrt{2v\log z^{-1}}} \, \mathrm{e}^{-Nz} \sim N^{1-\gamma} \, \mathrm{e}^{-r/\sigma\sqrt{2v\log N\sigma} - r^2/2\sigma^2}. \tag{C.8}$$

This starts to be of order N^{-1} as soon as $\tilde{r} < \sigma(\sqrt{2} - \gamma)\sqrt{2\log N\sigma}$. This is similar to the condition identified in [29] based on the factorization of the mean field solution. Note, however, that this expression does not hold near the $v = 0$ line, since it is derived assuming a steady traveling wave.

Appendix D. Velocity in the low recombination limit

With sufficiently frequent recombination, the mean fitness in the population increases with a rate v given by the variance in additive fitness among recombinant offspring. At lower recombination rates, however, the fitness variance in the population gets reduced below that of the distribution of recombinants. With reduced additive variance, v decreases. This effect has been studied in [35] for a model with only additive fitness, and we reproduce this argument for our simplified recombination model.

We will again work in units where the variance of the fitness distribution of recombinant genotypes is 1 and measure fitness a relative to the instantaneous mean fitness in the population. The invariant fitness distribution $P(a)$ in the comoving frame $a = A - vt$ is governed by

$$-v\partial_a P(a) = (a - r)P(a) + \frac{re^{-a^2/2}}{\sqrt{2\pi}}, \tag{D.1}$$

where the last term accounts for the injection of recombinant offspring. This equation has the solution

$$P(a) = rv^{-1} \, \mathrm{e}^{-(a-r)^2/2v} \int_a^{a_c} \frac{\mathrm{d}a'}{\sqrt{2\pi}} \, \mathrm{e}^{-a'^2/2 + (a'-r)^2/2v}. \tag{D.2}$$

Note that this solution becomes negative for $a > a_c$, and only the part $a < a_c$ is of interest. The zero crossing a_c marks the position of the most fit individuals in the population, and its value will depend on the population size.

To determine the velocity of a finite population, Rouzine and Coffin [35] compared the rate at which new genotypes ahead of a_c (records) are created to the speed at which the bulk of the population moves towards higher fitness.

Within the model, recombinant genotypes are generated with rate rN and drawn from a Gaussian distribution with unit variance. Hence genotypes with $a > a_c$ are produced at

rate

$$rN \int_{a_c}^{\infty} \frac{\mathrm{d}a}{\sqrt{2\pi}} \, \mathrm{e}^{-a^2/2} \approx \frac{rN}{a_c\sqrt{2\pi}} \, \mathrm{e}^{-a_c^2/2}. \tag{D.3}$$

Such a genotype seeded a_c ahead of the mean has a chance $a_c\sigma$ to establish and, if established, advances the nose by $1/a_c$. We therefore find for the speed of the nose

$$v_{\mathrm{nose}} \approx \frac{\sigma rN}{a_c} \, \mathrm{e}^{-a_c^2/2}. \tag{D.4}$$

To solve for $v_{\mathrm{nose}} = v$ and a_c, we need an additional relation which can be obtained from the condition that $P(a)$ is normalized. In the limit of small r, we can approximate equation (D.2)

$$P(a) = \frac{r}{v} \exp\left[-\frac{(a-r)^2}{2v}\right] \int_a^{a_c} \frac{\mathrm{d}a'}{\sqrt{2\pi}} \exp\left[\frac{(1-v)a'^2}{2v} - \frac{ra'}{v} + \frac{r^2}{2v}\right]$$

$$\approx \frac{\mathrm{e}^{-(a-r)^2/2v}}{\sqrt{2\pi}} \frac{r}{(1-v)a_c - r} \exp\left[\frac{(1-v)a_c^2}{2v} - \frac{ra_c}{v} + \frac{r^2}{2v}\right]. \tag{D.5}$$

The approximation is valid for $a_c(1-v) \gg r$; otherwise the population distribution is a simple Gaussian, and we find $v \approx 1$. Note that this condition is the same as the one we encountered above in the context of whether $C_r(z,v)$ is dominated by young or old clones. For $a_c(1-v) \gg r$, $C_r(z,v)$ and equivalently the normalization of $P(a)$ are dominated by the old clones. The normalization condition on $P(a)$ now requires

$$\frac{r\sqrt{v}}{(1-v)a_c - r} \exp\left[\frac{(1-v)a_c^2}{2v} - \frac{ra_c}{v} + \frac{r^2}{2v}\right] = 1, \tag{D.6}$$

which yields

$$a_c^2 \approx \frac{2v}{(1-v)} \log \frac{(1-v)a_c}{r\sqrt{v}} \approx \frac{2v}{(1-v)} \log \frac{\sqrt{2(1-v)}}{r}. \tag{D.7}$$

Solving equation (D.4) for a_c^2 and equating the result with equation (D.7) results in

$$\frac{2v}{1-v} \log \frac{\sqrt{2(1-v)}}{r} = 2\log \frac{\sigma rN}{v}. \tag{D.8}$$

After rearranging and replacing v by $\gamma = v/\sigma^2$, we obtain the expression given in equation (19).

Appendix E. Mean field solution

Above, we discussed the large r solution to the equation

$$\dot{P}(A,E) = (F - \langle F \rangle - r)P(A,E) + \frac{r}{\sqrt{2\pi}\sigma_A} \exp\left[-\frac{(A - \langle A \rangle)^2}{2\sigma_A^2}\right] \rho(E), \tag{E.1}$$

where we have introduced the symbol $\rho(E)$ for the distribution of epistatic fitness. At large r, this model admits a factorized solution $P(A, E) = \vartheta(A, t)\omega(E)$ with

$$\vartheta(A, t) = \frac{\mathrm{e}^{-((A-\sigma_A t)^2/2\sigma_A)}}{\sqrt{2\pi\sigma_A^2}} \qquad \omega(E) = \frac{r\rho(E)}{r + C - E}. \tag{E.2}$$

The epistatic part needs both to be normalized, and C needs to equal the mean of the distribution $\omega(E)$. Those two are linked: when one is true, so is the other. Hence

$$\begin{aligned}
\int \mathrm{d}E \, \frac{r\rho(E)}{r + C - E} &= \int \mathrm{d}E \, \rho(E) \sum_{n=0}^{\infty} \left(\frac{E - C}{r}\right)^n \\
&= 1 + \int \mathrm{d}E \, \rho(E) \frac{E - C}{r} \sum_{n=0}^{\infty} \left(\frac{E - C}{r}\right)^n \\
&= 1 + \frac{1}{r} \int \mathrm{d}E \, E \frac{r\rho(E)}{r + C - E} - \frac{C}{r} = 1 + \frac{\langle E \rangle - C}{r} = 1. \tag{E.3}
\end{aligned}$$

Hence the constant C, adjusted to achieve normalization, equals the population mean of E.

References

[1] Barton N, 1983 *Evolution* **37** 454
[2] Barton N, 1994 *Genet Res* **64** 199
[3] Batorsky R, Kearney M F, Palmer S E, Maldarelli F, Rouzine I M and Coffin J M, 2011 *Proc. Nat. Acad. Sci. USA* **108** 5661
[4] Brunet E, Derrida B, Mueller A H and Munier S, 2007 *Phys. Rev.* E **76** 041104
[5] Callahan B, Neher R, Bachtrog D, Andolfatto P and Shraiman B, 2011 *PLoS Gen.* **7** e1001315
[6] Cohen E, Kessler D A and Levine H, 2005 *Phys. Rev.* E **72** 066126
[7] Cohen E, Kessler D A and Levine H, 2005 *Phys. Rev. Lett.* **94** 098102
[8] Derrida B, 1981 *Phys. Rev.* B **24** 2613
[9] Desai M M and Fisher D S, 2007 *Genetics* **176** 1759
[10] Fisher R A, 1930 *The Genetical Theory of Natural Selection* (Oxford: Clarendon)
[11] Franklin I and Lewontin R C, 1970 *Genetics* **65** 707
[12] Goyal S, Balick D J, Jerison E R, Neher R A, Shraiman B I and Desai M M, 2012 *Genetics* **191** 1309
[13] Gross D and Mezard M, 1984 *Nucl. Phys.* B **240** 431
[14] Hallatschek O, 2011 *Proc. Nat. Acad. Sci. USA* **108** 1783
[15] Hill W G and Robertson A, 1966 *Genet. Res.* **8** 269
[16] Hudson R R and Kaplan N L, 1995 *Genetics* **141** 1605
[17] Kimura M, 1965 *Genetics* **52** 875
[18] Kingman J, 1978 *J. Appl. Probab.* **15** 1
[19] Kingman J, 1982 *J. Appl. Probab.* **19A** 27
[20] Krug J and Jain K, 2005 *Physica* A **358** 1
[21] Levy D N, Aldrovandi G M, Kutsch O and Shaw G M, 2004 *Proc. Nat. Acad. Sci. USA* **101** 4204
[22] McVean G A T and Cardin N J, 2005 *Phil. Trans. R. Soc.* B **360** 1387
[23] Mézard M and Montanari A, 2009 *Information, Physics, and Computation* (Oxford: Oxford University Press)
[24] Mézard M, Parisi G and Virasoro M, 1985 *J. Physique Lett.* **46** L217
[25] Mohle M and Sagitov S, 2001 *Ann. Probab.* **29** 1547
[26] Myers S R and Griffiths R C, 2003 *Genetics* **163** 375
[27] Neher R and Shraiman B, 2011 *Rev. Mod. Phys.* **83** 1283
[28] Neher R A and Leitner T, 2010 *PLoS Comput. Biol.* **6** e1000660
[29] Neher R A and Shraiman B I, 2009 *Proc. Nat. Acad. Sci. USA* **106** 6866
[30] Neher R A and Shraiman B I, 2011 *Genetics* **188** 975
[31] Neher R A and Shraiman B I, 2012 *Genetics* **191** 1283

[32] Neher R A, Shraiman B I and Fisher D S, 2010 *Genetics* **184** 467
[33] Nordborg M, 1997 *Genetics* **146** 1501
[34] Ortíz-Barrientos D, Reiland J, Hey J and Noor M A F, 2002 *Genetica* **116** 167
[35] Rouzine I M and Coffin J M, 2005 *Genetics* **170** 7
[36] Rouzine I M and Coffin J M, 2007 *Theor. Popul. Biol.* **71** 239
[37] Rouzine I M and Coffin J M, 2010 *Theor. Popul. Biol.* **77** 189
[38] Rouzine I M, Wakeley J and Coffin J M, 2003 *Proc. Nat. Acad. Sci. USA* **100** 587
[39] Sherrington D and Kirkpatrick S, 1975 *Phys. Rev. Lett.* **35** 1792
[40] Sire C, Majumdar S N and Dean D S, 2006 *J. Stat. Mech.* L07001
[41] Stephens M, Smith N J and Donnelly P, 2001 *Am. J. Human Genet.* **68** 978
[42] Tsimring L, Levine H and Kessler D, 1996 *Phys. Rev. Lett.* **76** 4440
[43] Turelli M and Barton N H, 1994 *Genetics* **138** 913
[44] Turelli M and Barton N H, 2006 *Evolution* **60** 1763
[45] Walczak A M, Nicolaisen L E, Plotkin J B and Desai M M, 2012 *Genetics* **190** 753

J. Stat. Mech. (2013) P01008