Effects of recombination on the evolvability, genetic diversity and mutational robustness of neutrally evolving populations

Alexander Klug and Joachim Krug[1]
Institute for Biological Physics, University of Cologne, Cologne, Germany

ABSTRACT

Many effects attributed to recombination have been invoked to explain the advantage of sex. The most prominent arguments focus on either evolvability, genetic diversity, or mutational robustness to justify why the benefit of recombination overcomes its costs, with partially contradicting results. As a consequence, understanding which aspects of recombination are most important in a given situation remains an open problem for theoretical and experimental research. In this study, we focus on finite populations evolving on neutral networks, which already display remarkably complex behavior. We aim to provide a comprehensive overview of the effects of recombination by jointly considering different measures of evolvability, genetic diversity, and mutational robustness over a broad parameter range, such that many evolutionary regimes are covered. We find that several of these measures vary non-monotonically with the rates of mutation and recombination. Moreover, the presence of lethal genotypes that introduce inhomogeneities in the network of viable states qualitatively alters the effects of recombination. We conclude that conflicting trends induced by recombination can be explained by an emerging trade-off between evolvability and genetic diversity on the one hand, and mutational robustness and fitness on the other. Finally, we discuss how different implementations of the recombination scheme in theoretical models can affect the observed dependence on recombination rate through a coupling between recombination and genetic drift.

The neutral theory of evolution assumes that mutations have either no selective effect or are highly deleterious. This approximation of the distribution of fitness effects was suggested by Kimura based on observations of surprisingly high substitution rates in the amino acid sequence of certain proteins, although their function remained essentially unchanged (Kimura 1968, 1983). Today it is understood that the abundant neutrality in molecular evolution can arise through a wide range of mechanisms. Large portions of the genome are non-coding, allowing mutations to accumulate freely (Nobrega *et al.* 2004). However, also in the coding regions, neutrality is prevalent due to degeneracies in the genotype–phenotype mapping at multiple levels between the blueprint, the DNA, and the final functional structure, which can be a protein, a cell, or an entire organism (Manrubia *et al.* 2021). For example, on the scale of proteins, the degeneracies arise through synonymous mutations and through many different amino acid chains that fold to the same structure (Guo *et al.* 2004; Bloom *et al.* 2005). On the scale of cells, neutrality is observed in regulatory gene networks (Azevedo *et al.* 2006; Ciliberti *et al.* 2007) and metabolic reaction networks (Rodrigues and Wagner 2009). Moreover, recent microbial evolution experiments (Johnson *et al.* 2019) and theory (Reddy and Desai 2021) indicate that populations consistently adapt to regions of the genotype space where diminishing-returns and increasing-costs epistasis are common, which implies that the beneficial effects of mutations are almost neutral, whereas deleterious mutations typically have large negative selection coefficients. Apart from truly neutral mutations, small effect mutations can be effectively neutral if the absolute magnitude of the selection coefficient is smaller than the reciprocal of the population size (Ohta 2002). Therefore neutrality is particularly important for small populations.

[1]Corresponding author: Institute for Biological Physics, University of Cologne, Zülpicher Str. 77, D-50937 Köln, Germany. E-mail: jkrug@uni-koeln.de

The assumption of a binary distribution of fitness effects, where mutations are either selectively neutral or highly deleterious, can be conceptualized as a flat fitness landscape with holes (Gavrilets 1997, 2004) or as a neutral network with varying node degrees (van Nimwegen *et al.* 1999; Wilke 2001). With potentially many loci at which mutations can occur, of which at least a few are selectively neutral, large clusters of viable genotypes connected by point mutations form. These clusters may span the entire sequence space and thus populations can evolve continuously without being trapped at a fitness peak (Gavrilets 2004). In this way large neutral networks are argued to increase evolvability, since populations are able to explore large parts of genotype space leading to ever fitter genotypes (Wagner 2005).

However, for a complete description of evolution on neutral fitness landscapes, also the population dynamics has to be specified. A commonly used simplification is to consider the population as a point on the fitness landscape, implying that only a single genotype is present at a time. Neutral evolution then proceeds as a simple random walk on the neutral network through a sequence of fixation events (Maynard Smith 1970; Gavrilets 1997). This scenario applies in the weak mutation regime where the mutation supply is low (de Visser and Krug 2014). At the opposite end of the spectrum of evolutionary dynamics, quasispecies theory considers populations as continuously distributed clouds of genotypes in sequence space (Jain and Krug 2007; Domingo and Schuster 2016). An important difference compared to the weak mutation regime is that while in a simple random walk, all viable connected genotypes have the same probability of being currently occupied by the population (Hughes 1996), in the quasispecies regime mutationally robust genotypes, i.e. genotypes with an above-average number of viable point mutations, are preferentially occupied (van Nimwegen *et al.* 1999; Bornberg-Bauer and Chan 1999). This effect is strongly enhanced in recombining populations (Szöllősi and Derényi 2008; Klug *et al.* 2019; Singhal *et al.* 2019).

Quasispecies theory is deterministic and, strictly speaking, only applies to infinitely large populations (Wilke 2005). Therefore genetic drift is absent, and all genotypes have a frequency greater than zero by definition. Moreover, in this limit, the population reaches a stationary state determined by a selection-mutation(-recombination) balance, where the frequencies of all genotypes become constant in time. Since the number of genotypes grows exponentially with the number of loci, a shortcoming of this approximation is that it quickly becomes unrealistic for large but finite populations and is only applicable for short sequences, where the population can cover all genotypes. In this case quasispecies theory can approximate finite populations quite well (van Nimwegen *et al.* 1999; Szöllősi and Derényi 2008).

The purpose of this article is to describe and understand neutrally evolving finite populations in large sequence spaces for which the deterministic quasispecies limit does not apply. Within this setting, we explore a broad parameter range, such that all possible evolutionary regimes are covered. In particular, we include recombination and study its effect across a wide range of recombination rates. We believe that such a comprehensive study, which to the best of our knowledge has not been performed previously, is essential for elucidating the conditions under which recombination carries a selective advantage (Weismann 1891; Muller 1932, 1964; Felsenstein 1974; Kondrashov 1988; Feldman *et al.* 1996; Burt 2000; de Visser and Elena 2007; Otto 2009). In previous work we argued that the universal and somewhat underappreciated effect of recombination on mutational robustness may play an important role in this context (Klug *et al.* 2019). Working in the deterministic limit of an infinitely large populations, we showed that mutational robustness increases monotonically with the recombination rate r, independent of model details, and that for low mutation rate μ and small r, mutational robustness grows linearly with r/μ. The deterministic limit allowed us to find precise analytical results, but many relevant questions cannot be addressed in this framework. Here we consider finite populations in large sequence spaces. We are interested in the evolvability of the population and ask how quickly new genotypes are discovered, and how many generations it takes to discover all viable genotypes. The discovery rate of new genotypes can be crucial, e.g., if through environmental perturbations like an immune response certain genotypes become fitter over time or if the majority of the neutral network loses fitness and an escape mutation needs to be found. While the discovery of new genotypes is therefore essential for long-term survival, the accumulation of lethal mutations entails the risk of extinction. This induces an evolutionary trend towards increasing mutational robustness, another measure we investigate.

In order explain how the discovery rate and the mutational robustness of the population change with the parameters, we consider different measures of genotype diversity, such as the mean Hamming distance, the number of segregating mutations and the number of distinct genotypes. Certain properties like the number of segregating mutations and mean Hamming distance are independent of recombination if all genotypes are viable, but become recombination dependent when some genotypes are lethal. Other properties like the number of distinct genotypes and the discovery rate grow monotonically with r if all genotypes are viable, while in the presence lethal genotypes, the dependence becomes non-monotonic. We also find that with recombination, the discovery rate can become non-monotonic in the mutation rate, such that higher mutation rates may lead to reduced evolvability. Furthermore, we discuss different implementations of recombination in the Wright-Fisher model. Depending on the model details, recombination can act as an additional source of genetic drift which matters in small populations or large sequence spaces.

Outline. In the first section Models and methods, we define the structure of the genotype space and the fitness landscape. We consider both finite and infinite-sites settings. We further define the population dynamics and the implementation of recombination. Next we introduce the relevant measures of diversity, robustness and evolvability and describe the visualization of our results. In the second section Results and analysis, we first give an overview of the evolutionary regimes on neutral networks. We then explain our results in the limit of infinite sequence spaces (infinite-sites model) and continue with the results for finite sequence spaces. At the end of this section we discuss aspects of the results that show a non-robust dependence on the implementation of recombination. The results are summarized and conclusions are presented in the last section Discussion. Due to the complexity of the problem, our work relies primarily on extensive numerical simulations, but analytic results are also presented when available.

Models and methods

Genotype space

We consider haploid genomes with L diallelic loci, which can be expressed as sequences

$$\sigma = (\sigma_1, \sigma_2, ..., \sigma_L) \tag{1}$$

of symbols drawn from a binary alphabet $\sigma_i = \{-1, 1\}$. This translates to a genotype space that has the properties of a hypercube $H_2^L = \{-1, 1\}^L$ of dimension L, where each of the 2^L vertices represents a genotype. Genotypes of vertices connected by an edge differ at a single locus and are therefore mutually reachable by a point mutation. The natural metric in this genotype space is the Hamming distance

$$d(\sigma^i, \sigma^j) = \sum_{k=1}^{L} (1 - \delta_{\sigma_k^i, \sigma_k^j}), \tag{2}$$

which quantifies the number of point mutations that separate two genotypes σ^i and σ^j. For our analyses we consider both the so-called infinite-sites model (*ism*) corresponding to the limit $L \to \infty$, and the finite-sites model (*fsm*) with finite L. The *ism* originally introduced by Kimura (1969) is easier to handle analytically, since back mutations do not occur and all mutations are novel. However, certain quantities of interest such as the mutational robustness, for which the number of viable point mutations needs to be computed, and the time until full discovery of all viable genotypes cannot be defined within the *ism*. We therefore consider both models and compare results.

Fitness landscape

In our simulations, we either assume that all genotypes are viable, or that a fraction $1 - p$ of genotypes is lethal. We show that the addition of lethal genotypes strongly alters the structure of the genotype cloud and in particular the effect of recombination. To be maximally agnostic about the distribution of lethal genotypes in sequence space, we assume that each genotype is viable with probability p and otherwise lethal. This kind of fitness landscape is known as a percolation landscape (Gavrilets and Gravner 1997).

In the case of the *fsm* we add the constraint that the resulting network of viable genotypes on the hypercube is connected, i.e. that between any two viable genotypes there is a path of viable point mutations. In our simulations this is achieved by discarding all percolation landscape realizations that do not satisfy this condition; Fig. S1 shows how the fraction of connected landscapes varies with p. The constraint is added in order to avoid situations in which the initial population is trapped in a disconnected cluster of viable genotypes. In the case of the *ism* there is no additional constraint and the fitness of a novel genotype is generated once it has been discovered by the population.

Besides containing only minimal assumptions, we also chose this landscape model because its random nature makes it rich in possible structures, in the sense that it can contain regions with many viable point mutations and regions where genotypes are more often lethal and populations must evolve along a narrow fitness ridge. Furthermore, the choice of this landscape has the benefit of only adding one more parameter p to our analysis. From the point of view of the neutral network, the parameter p determines the degree distribution.

Dynamics

To model the evolutionary forces of selection, mutation and recombination, we use individual-based Wright-Fisher models with discrete, non-overlapping generations and a constant population size N. For the implementation of selection and recombination, we found different computational schemes in the literature. Initial simulations showed that, whereas for large populations in the *fsm* the models become indistinguishable, the model details become apparent for small populations in the *fsm* and at arbitrary population sizes in the *ism*. In the following, the different schemes are explained. For the main part of the article, we show results for only one of the models, but mention important differences when they exist, and discuss the differences in detail in the subsection Recombination-induced genetic drift.

Figure 1 illustrates the course of one generation for three different selection-recombination schemes. In the main text, we use the model that we refer to as *concurrent recombination*. In this model, selection and recombination occur in a single step, whereas in the other two models referred to as *successive recombination* schemes these processes require two separate steps. Despite these differences, the models also share similarities, which we explain first. All models have in common that an individual j has two ancestors in the previous generation with a probability equal to the recombination rate r. If recombination occurs we employ a uniform crossover scheme, which means that at each locus, the allele from one of the two ancestors k,l is chosen with equal probability,

$$R: \ \sigma_i^j \mapsto \begin{cases} \sigma_i^k & \text{with prob. } 1/2, \\ \sigma_i^l & \text{with prob. } 1/2, \end{cases} \quad \forall\ i. \tag{3}$$

Furthermore, the mutation step always occurs last, separate from the two other processes. During the mutation step, each locus of each individual mutates with probability μ to the opposite allele,

$$M: \ \sigma_i^j \mapsto \begin{cases} \sigma_i^j & \text{with prob. } 1 - \mu \\ -\sigma_i^j & \text{with prob. } \mu, \end{cases} \quad \forall\ i, j. \tag{4}$$

In the *ism* we take the joint limits $L \to \infty$ and $\mu \to 0$ at finite genome-wide mutation rate $U = L\mu$. In this limit the number of mutations per individual and generation is Poisson distributed with mean U. Importantly, each mutation is then novel and back mutations cannot occur. This can be expressed by characterising each individual's genotype by the set of acquired novel mutations, e.g.

$$M: \ \sigma^j = \{\tau, \zeta\} \to \begin{cases} \{\tau, \zeta\} \\ \{\tau, \zeta, \lambda\} \\ \{\tau, \zeta, \lambda, \gamma\} \\ ... \end{cases} \quad \forall\ j, \tag{5}$$

denoted by Greek letters. It is necessary to track the mutations carried by each invididual also in the *ism* in order to be able to implement recombination, which combines mutations and breaks them apart. However, once a mutation is fixed in the population, i.e., it is present in all individuals, it can be omitted due to the lack of back mutations, thereby keeping the list of stored mutation finite. Also, the stored list of fitness values of discovered genotypes can be purged of those genotypes whose mutation set does not contain newly fixed mutations, as they cannot be reached anymore.

While the features discussed so far are the same in all models, the differences are the following.

Concurrent recombination: In this case all individuals that are not the product of a recombination event select one ancestor with a probability proportional to the ancestor's fitness. For our neutral landscapes, this implies that viable ancestors are drawn uniformly with replacement. Simultaneously those individuals that are the product of a recombination event select two different ancestors with a probability proportional to their fitness. For a neutral landscape, this implies that each descendant takes a random sample of size two without replacement from the pool of viable ancestors.

Successive recombination: In both successive recombination models, selection occurs first, where all individuals choose one ancestor, again in our case uniformly with replacement among the viable ones. Next, recombination takes place independent of fitness. In the case of the simple successive recombination model, individuals that are a product of recombination choose two different ancestors that survived selection. In the case of successive recombination with mating pairs, all individuals that survived selection and happen to recombine are pooled in groups of two, which then create two offspring individuals. These two offspring individuals are complementary in their recombined material, that is, if one offspring has the allele of the first parent, the other offspring will have the allele of the second parent at the corresponding locus.

Importantly, the three recombination schemes differ in the way in which recombination couples to genetic drift. In the concurrent recombination and the successive recombination with mating pairs model, genetic drift is independent of the recombination rate r, but this is not the case for the simple successive recombination model (see Appendix). Recombination-dependent drift is a confounding factor that needs to be taken into account in the interpretation of the results of the latter model. Since the concurrent and the successive recombination with mating pairs models are implemented in two commonly used open software packages (Haller and Messer 2016; Zanini and Neher 2012), we stick to a non-recombination dependent genetic drift model in the main text and only occasionally refer to differences that would otherwise appear. The recombination-dependent genetic drift model has been used by Nowak *et al.* (2014). Being aware about these differences might be important for the design of experiments, e.g., in the context of *in vitro* recombination (Pesce *et al.* 2016). Of the two non-recombination dependent genetic drift models, we choose the concurrent recombination model because it is somewhat simpler. In particular, in this model the number of recombining individuals does not have to be an even number.

Measures of evolvability, diversity and robustness

To quantify evolvability in the *ism* we consider the discovery rate r_{dis} of novel genotypes and the fixation rate r_{fix} of mutations. The discovery rate r_{dis} is the average number of novel viable genotypes that are discovered in each generation, either through mutation or recombination. The fixation rate r_{fix} of mutations measures the average number of segregating mutations that become fixed in each generation.

In the *fsm* we monitor evolvability through the time t_{fdis} and the number of mutation events N_{mut} until full discovery. Starting from a monomorphic population carrying a single randomly selected viable genotype, we say that full discovery is reached when all viable genotypes have been present in at least one individual in at least one generation. The time is measured

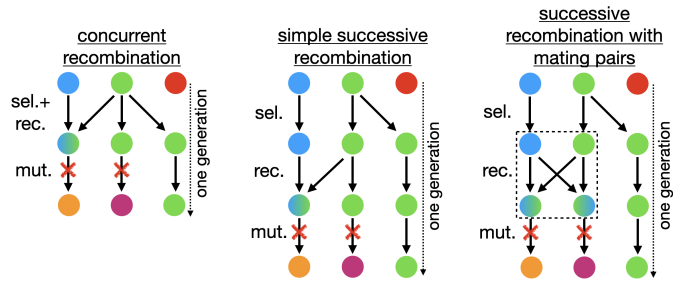

Figure 1 One generation of evolution in the three selection-recombination-mutation schemes discussed in the text. Nodes represent individuals and colors genotypes, with green and blue nodes being fit and red nodes unfit. The arrows show the lineages from the ancestors to the descendants. Individuals with two incoming arrows are the product of uniform recombination. In successive recombination with mating pairs the recombining individuals are grouped in pairs and each pair creates two descendants, which is indicated by the dashed box. Mutations are indicated by red crosses.

in generations and a mutation event occurs if an individual acquires one or multiple mutations during reproduction.

Genetic diversity is characterized through several well-known measures of population genetics, for which the time average is descriptive for the randomly drifting genotype cloud (note that in the regimes of interest here, the population cannot attain an equilibrium state, because the genotype space is larger than the population). Such measures are the pairwise mean Hamming distance between two individuals in the population

$$d_{pw} = \frac{1}{N(N-1)} \sum_{\substack{i,j \\ i \neq j}} d(\sigma^i, \sigma^j), \tag{6}$$

the number of viable distinct genotypes

$$Y = |\{\sigma^i | i \in 1, 2, ..., N \wedge \sigma^i \text{is viable}\}|, \tag{7}$$

and the number of segregating mutations

$$S = \sum_i^L \left(|\{\sigma_i^j\}_{j \in 1,2,..,N}| - 1 \right), \tag{8}$$

i.e., the number of loci at which both alleles are present in the population. These measures are used for the *ism* as well as for the *fsm*.

Additionally, for the *fsm* we consider the mutational robustness m of the population. The robustness m_{σ^i} of an individual i is equal to the fraction of viable point mutations of its genotype σ^i if it is itself viable, and equal to 0 otherwise. The mutational robustness of the population is the average robustness of all individuals,

$$m = \frac{1}{N} \sum_i^N m_{\sigma^i}. \tag{9}$$

This quantity depends on the population distribution in genotype space, and increases if the population moves to genotypes for which most point mutations are viable. Apart from these measures for evolvability, genetic diversity and mutational robustness we also measure the mean fitness defined by

$$\overline{w} = \frac{1}{N} \sum_i^N w_i, \tag{10}$$

as well as the average viable recombination fraction, which represents the fraction of recombination events per generation that generate a viable genotype. The latter can be considered a measure of recombination robustness.

Except for the measures for evolvability in the *fsm*, our numerical results always show the averages in steady state. For the *ism*, we evolve and evaluate the population for $t = \sqrt{10^{10}/U}$ generations for each data point. For the *fsm*, we evolve the population until all genotypes have been discovered once and, except for the measures of evolvability, more than 10^5 mutation events have occurred. This is done for 10^4 landscape realizations and for each data point. We measure all quantities at the end of a generation, i.e., after the mutation step, and denote the averages of measured quantities by overbars.

Illustration of results

3D wireframe plots. In order to represent the numerical results comprehensively, we mostly use 3D wireframe plots. In these plots, the wireframe lines run along either constant recombination rate or constant mutation rate to guide the eye. The color of the wireframe lines depends on their height. Additionally a contour plot is shown below the wireframe. The viewing angles vary and are selected such that the results can be seen in the best possible way.

Graph representation. To visualize the population distribution in sequence space, we use a graph representation. In this graph each genotype in the population is represented as a node, and nodes whose genotypes differ by a single point mutation are connected by an edge. Therefore, the resulting graph only contains information about nearest neighbor relationships. However, as long as the genotype cloud is not distributed too broadly in sequence space, we expect most nodes to have at least one edge, thereby forming clusters of connected components in the high dimensional sequence space. To arrange the nodes in two dimensions we use a forced-based algorithm called *ForceAtlas2* (Jacomy *et al.* 2014). This leads to a configuration in which nodes that share many edges form visual clusters. The frequencies of the genotypes are represented by the node sizes.

Results and analysis

Evolutionary regimes

To organize the discussion of the results, we recall here the distinct evolutionary regimes that can be realized on neutral networks. In the *monomorphic* or *weak mutation* regime $NU \ll 1$, mutations are rare and either fix with probability $1/N$ or go extinct through genetic drift before another mutation originates. In this regime, the population consists of a single genotype most of the time and can be described by a "blind ant" random walker (Hughes 1996; van Nimwegen *et al.* 1999). Without lethal genotypes, a step to one of the current genotype's mutational neighbors is taken with equal probability at rate U independent of N. With lethal genotypes, a step is discarded when the randomly chosen point mutation is lethal. On a connected network this implies that the population occupies on average each genotype, irrespective of its degree, with equal probability (Hughes 1996). In this regime, the effect of recombination is minimal since recombination is fueled by combining segregating mutations, which do not exist most of the time. Furthermore, as long as there are not more than two genotypes in the population within one generation, the population is by definition in linkage equilibrium. Nonetheless, in subsection Recombination-induced genetic drift we show that recombination can have an effect in this regime if it couples to the genetic drift.

In the *polymorphic* regime ($NU \geq 1$), where the population is a cloud of competing genotypes, two subregimes can be distinguished. If the population size is large compared to the number of genotypes ($N \gg 2^L$), all viable genotypes can become occupied and an equilibrium state is reached. In this case genetic drift is irrelevant and the equilibrium distribution can be described by assuming deterministic dynamics of quasispecies type. In the absence of lethal genotypes, all genotypes then have the same frequency in the population, similar to the monomorphic regime. Importantly, though, with lethal genotypes, the population distribution becomes non-uniform, in that robust viable genotypes that have less lethal point mutations exhibit higher frequency (van Nimwegen *et al.* 1999). This imbalance increases dramatically with increasing recombination rate (Klug *et al.* 2019).

Of particular interest in the context of the present work is the second polymorphic subregime, where the population size is smaller than the number of genotypes ($N \ll 2^L$) or even smaller than the number of loci ($N \ll L$), and the population clearly cannot attain an equilibrium state. Instead, it will diffuse as a cloud of genotypes on the neutral network. This subregime has, to the best of our knowledge, not been fully covered in the literature, in particular in the presence of recombination. In the following sections, we first study the population structure in the *ism*, where $N \ll L$ is guaranteed by definition. Subsequently we consider a finite number of sites and study the *fsm* for $N \ll 2^L$.

Infinite-sites model

In this section, we present numerical results that show the impact of recombination and lethal genotypes on the population structure in the *ism*. We keep the population size fixed at $N = 100$ and discuss the dependence on the recombination rate r, the mutation rate U and the fraction of viable genotypes p.

Discovery rate. Figure 2 displays the discovery rate of formerly unexplored viable genotypes for $p = 1$ and $p = 0.5$. Without recombination, the discovery rate is given by

$$r_{dis} = pN(1 - e^{-U}) \overset{U \ll 1}{\approx} pNU \tag{11}$$

since each mutation generates a new genotype which is viable with probability p. The numerical results show that the effect of recombination on the discovery rate depends on the mutation rate U, the product NU and the fraction of viable genotypes p. If $NU \ll 1$, the effect of recombination is minimal, since there are very few segregating mutations that can be recombined. However, for $NU \geq 1$ we notice a rather complex dependence. In the absence of lethal genotypes ($p = 1$), recombination increases the discovery rate monotonically. The relative increase is maximal if $U \ll 1$ but $NU \gg 1$. If U is of order 1, the effect of recombination becomes smaller since the maximum discovery rate is capped by the population size N and almost exhausted through mutations.

In the presence of lethals ($p = 0.5$) the behavior is more surprising. For small U, recombination has almost no effect, even if $NU \geq 1$. As U increases, an intermediate recombination rate becomes optimal for the discovery rate. This intermediate peak shifts to larger r with increasing U and the drop-off becomes sharper until U is of order 1, where the intermediate peak vanishes and the behavior is again similar to $p = 1$. These results indicate three regimes for the effect of recombination in the *ism*: (i) $NU \ll 1$, (ii) $NU \geq 1$ and $U \ll 1$, and (iii) $U \approx 1$.

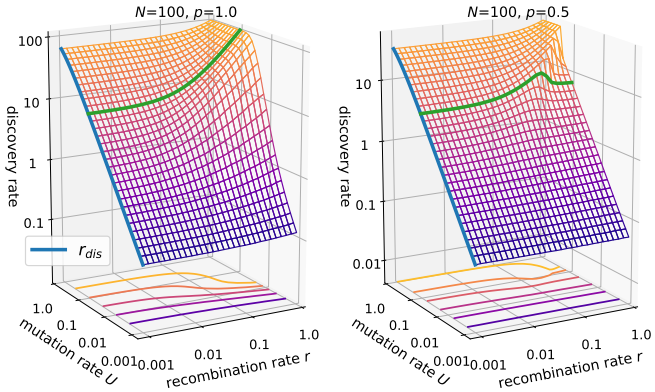

Figure 2 Discovery rate in the *ism* for population size $N = 100$ and $p = 1.0$ (left panel) vs. $p = 0.5$ (right panel). The green line at $U = 0.1$ highlights the different effects of recombination in the absence ($p = 1$) and presence ($p = 0.5$) of lethal genotypes. The blue lines in both panels show Eq. 11.

Fixation rate. Next, we consider the fixation rate of segregating mutations. In the absence of recombination mutations are expected to fix at rate U, since in each generation NU mutations occur, a fraction $1/N$ of which goes to fixation. The results in Fig. 3 for $p = 1$ confirm this expectation and show that concurrent recombination has no effect if all genotypes are viable. However, for $p = 0.5$ the fixation rate is seen to dramatically decline at large recombination rates when $NU \geq 1$ and $U \ll 1$. This disruption of fixation is released only when the mutation rate becomes of order 1. We further notice, that at very large mutation rates, the fixation rate exceeds U.

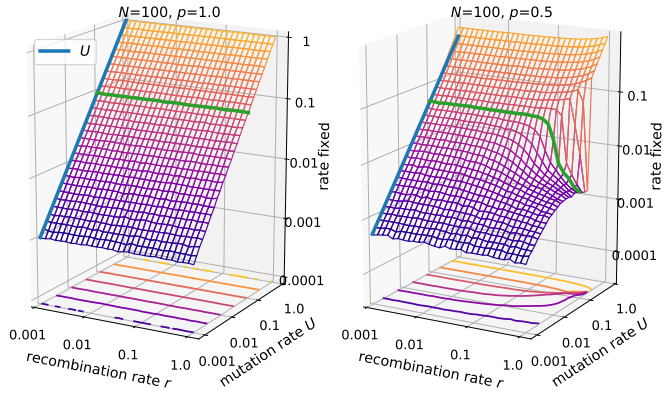

Figure 3 Fixation rate of segregating mutations in the *ism* with $N = 100$ and $p = 1.0$ (left panel) vs. $p = 0.5$ (right panel). The green line is drawn at $U = 0.1$.

Number of distinct genotypes. To understand the results for evolvability, we next consider measures of genetic diversity. The number of distinct genotypes Y is naturally closely related to the discovery rate, since with more novelty discovered in each generation, more distinct genotypes should accumulate, cf. Fig. 4. An analytical expression for the case of no recombination and no lethal genotypes was derived in Ewens (2012):

$$\overline{Y}(r = 0) = \sum_{i=0}^{N-1} \frac{\theta}{\theta + i} \quad \text{with} \quad \theta = 2NU. \tag{12}$$

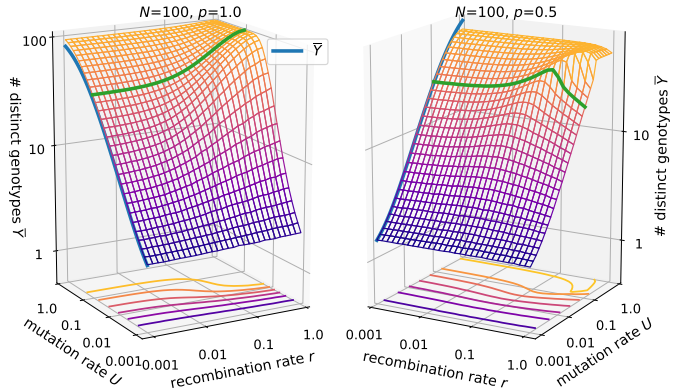

Figure 4 Number of distinct genotypes in the *ism* with $N = 100$ and $p = 1.0$ (left panel) vs. $p = 0.5$ (right panel). The blue lines shows Eq. 12, where θ is replaced by θ^* for $p < 1$. The green line is drawn at $U = 0.1$.

Our numerical results fit this expression and furthermore show that in the absence of recombination, the formula can be extended to include lethal genotypes by replacing θ with

$$\theta^* = 2pNU. \tag{13}$$

The reasoning behind this replacement is that the parent population size is effectively reduced by a factor p in each generation. In general, the effect of recombination on the number of distinct genotypes is similar to that on the discovery rate. A noticeable difference in the case of $p = 0.5$ is that $\overline{Y}$ is not maximal at $U \approx 1$ but at a slightly smaller value. The reason is that at such high mutation rates a significant fraction $(1 - p)U$ of individuals does not survive each generation, such that the population repeatedly goes through a bottleneck, in which distinct genotypes are lost. This leads to the observed increased fixation rate at $U \approx 1$ in Fig. 3.

Number of segregating mutations. Next we consider the number of segregating mutations S (Fig. 5). These span an effective

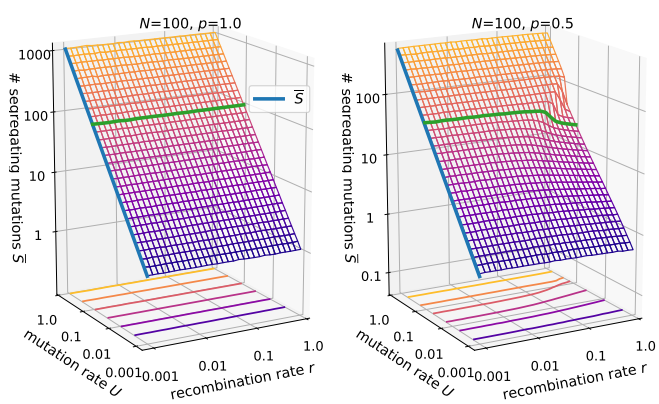

Figure 5 Number of segregating mutations in the *ism* with $N = 100$ and $p = 1.0$ (left panel) vs. $p = 0.5$ (right panel). The blue lines shows Eq. 15. The green line is drawn at $U = 0.1$.

sequence space of size 2^S in which recombination can create novel genotypes. An analytical expression for the number of segregating mutations in the *ism* was developed by Watterson (1975), again assuming no recombination and no lethal genotypes. It can be derived from the expectation of the length of the

genealogical tree to the most recent common ancestor, which is given by (Wakeley 2009)

$$\overline{T}_{MRCA} = 2N \sum_{i=1}^{N-1} \frac{1}{i} \tag{14}$$

for the Wright-Fisher model. The total tree length is equal to the total time, and multiplying by the mutation rate U and the fraction of viable genotypes yields the average number of segregating mutations

$$\overline{S} = \theta^* \sum_{i=1}^{N-1} \frac{1}{i}. \tag{15}$$

The results also show that for $p = 1$ and the concurrent recombination model used here, the number of segregating mutations is independent of r. This is not generally true but depends on the implementation of recombination (see Recombination-induced genetic drift). For $p = 0.5$ and $NU \geq 1$, $U \ll 1$, recombination generally decreases the number of segregating mutations. For example, at $U = 0.1$ the number decreases from around 50 to about 20, which still yields an enormous effective sequence space compared to the population size. This alone cannot explain the decrease in evolvability seen in Figs. 2, 4.

Mean Hamming distance. Since recombination occurs between pairs of individuals we now consider the pairwise mean Hamming distance $\overline{d}_{pw}$ (Fig. 6). From Eq. 14 we conclude that the

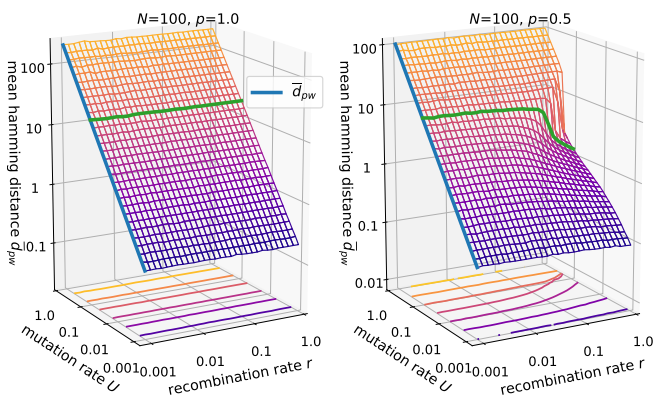

Figure 6 Mean Hamming distance in the *ism* with $N = 100$ and $p = 1.0$ (left panel) vs. $p = 0.5$ (right panel). The blue lines shows Eq. 16. The green line is drawn at $U = 0.1$.

mean length of the genealogical tree to the most recent common ancestor for two random individuals is given by $2N$, which directly leads to the expression

$$\overline{d}_{pw} = \theta^* \tag{16}$$

for the mean Hamming distance in non-recombining populations. The results for $p = 0.5$ show that recombination contracts the population cloud in sequence space in the presence of lethal genotypes. The functional relationship is similar to the number of segregating mutations but the relative change is more dramatic, e.g., for $U = 0.1$ the distance drops from 10 to about 1.

Cross section of the population cloud. To further understand the contraction in sequence space, we took a cross section of the population cloud by measuring the Hamming distance of each individual to the ancestral genotype that contains only fixed mutations, averaged over many generations (Fig. 7). This quantity is equal to the number of segregating mutations in an individual. For $r = 0$ the Hamming distance distribution is seen to follow a hypoexponential distribution, which converts to a Poisson distribution for large r. The hypoexponential distribution follows from the correspondence between the Hamming distance and the time to the most-recent-common-ancestor T_{MRCA}, the distribution of which is well known (Wakeley 2009). With a mutation rate U and a fraction p of viable genotypes, this yields the distribution

$$f_{h_{MRCA}}(d) = \frac{2}{\theta^*} \sum_{i=2}^{N} \binom{i}{2} e^{-\binom{i}{2}\frac{2d}{\theta^*}} \prod_{j=2, j\neq i}^{N} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}}, \tag{17}$$

with mean θ^*. At high recombination rates, segregating mutations become well mixed among all individuals, and the number of segregating mutations acquired by an individual follows a Poisson distribution. For $p = 1$ the mean is independent of r and equal to θ^*, but for $p = 0.5$ we observe a strong contraction of the distance distribution at $r = 1$. This shows that the focal genotype, around which the contraction occurs, is the ancestral genotype that contains only fixed mutations.

Mean fitness and recombination load. The contraction of the genotype distribution described in the preceding paragraphs must be beneficial for the population in the sense that the mean fitness increases. For our binary fitness distribution, the mean fitness is determined by the mutation and recombination load, i.e. the fraction of mutation and recombination events per generation that are lethal. The rate of lethal mutations is always $U(1-p)$ and cannot be optimized in the *ism*. Contrary to that, the outcome of recombination events depends on the genotype composition of the population. If the population is contracted around a focal genotype and most individuals are closely related to each other, the effective sequence space for recombination is much smaller than 2^S. Therefore it becomes likely that a recombination event will not create a novel genotype but a genotype that already exists in the current genotype cloud and that, more importantly, is viable, which benefits the mean fitness.

This connection is shown in Fig. 8. If the population is sparsely distributed which happens at high mutation rates, most recombination events create novel genotypes, which leads to a viable recombination fraction equal to p. Contrary to that in a monomorphic population ($NU \ll 1$) no novelty results from recombination. In the regime $NU \geq 1, U \ll 1$, e.g. at $U = 0.1$, we see that with increasing r the viable fraction initially decreases, as the population becomes more diverse, leading to a decrease in fitness. At large recombination rates this trend reverses as the population becomes concentrated around a focal genotype, which then also leads to a fitness increase. Thus the contraction of the population is an adaptive response to the fitness decline caused by an increased recombination load.

Dependence on the fraction of viable genotypes. The results presented so far were obtained for the two values $p = 1$ and $p = 0.5$ of the fraction of viable genotypes. Figure 9 shows cross sections of the previously shown 3D plots at either fixed recombination rate $r = 1$ (left column) or fixed mutation rate $U = 0.1$ (right column) and four different values of p. For fixed $r = 1$, the lethal genotypes strongly reduce evolvability by contracting the genotype cloud when the mutation rate is low, but the contraction is released once the mutation rate is strong

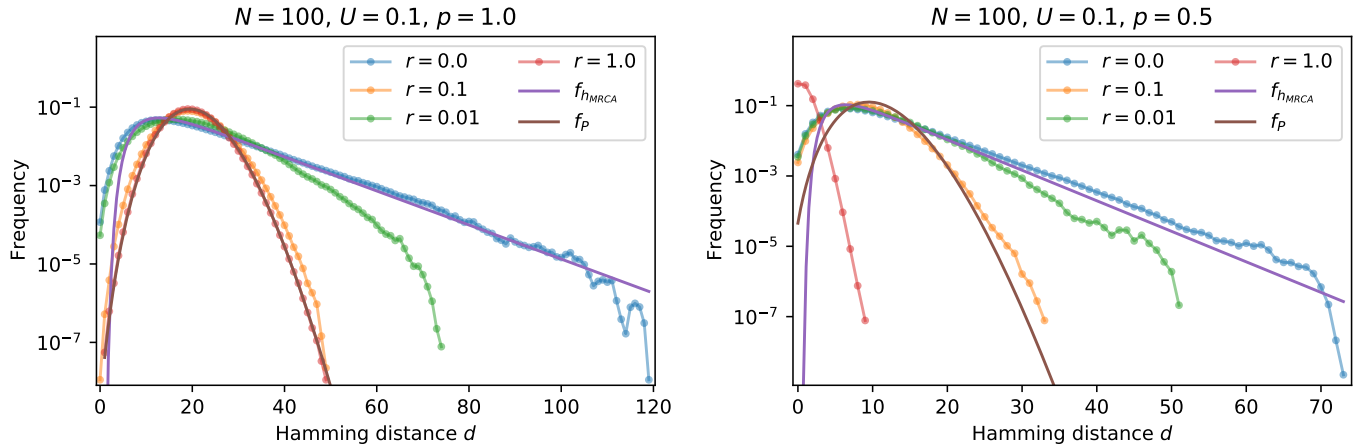

Figure 7 Cross section of the population cloud in the *ism* with $N = 100$, $U = 0.1$ and $p = 1.0$ (left panel) vs. $p = 0.5$ (right panel). The figures show the distribution of the Hamming distance to the genotype containing only fixed mutations. The hypoexponential distribution $f_{h_{MRCA}}$ is given by Eq. 17 and f_P is a Poisson distributions with mean θ^*. The data were accumulated over 10^6 generations.

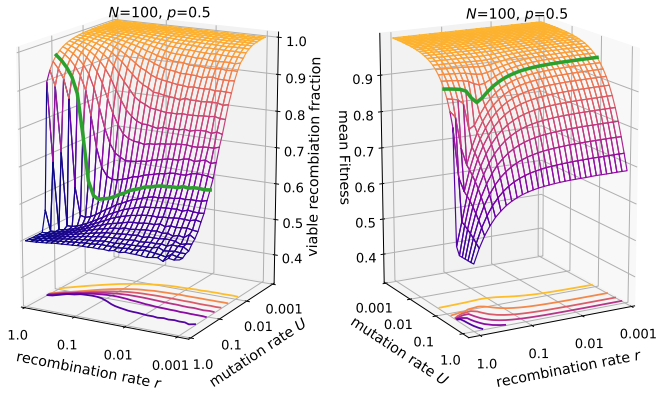

Figure 8 Viable recombination fraction and mean fitness in the *ism* for $N = 100$, $p = 0.5$. The green line is drawn at $U = 0.1$.

enough. At this point the population evolves independent of fitness and therefore the measures coincide with the results for $p = 1$. However, the mean fitness is strongly reduced and equal to p. With smaller p this transition happens at larger U and strikingly at small enough p the numerical results display a discontinuity as a function of U.

The dependence on mutation rate resembles the error threshold phenomenon of quasispecies theory, in which the population delocalizes from a fitness peak in a finite-dimensional sequence space when the mutation rate is increased above a critical value (Jain and Krug 2007; Domingo and Schuster 2016). In the quasispecies context it has been shown that error thresholds do not occur in neutral landscapes with lethal genotypes, at least not in the absence of recombination (Wilke 2005; Wagner and Krall 1993). Moreover, for non-recombining populations the mean population fitness is generally continuous at the error threshold, whereas recombination can induce discontinuous fitness changes and bistability (Boerlijst *et al.* 1996). Although the transfer of results from the infinite population quasispecies model with finite sequence length L to the finite population *ism* is not straightforward, our results are generally consistent with previous work in that there is no discontinuity in the absence of recombination, while at sufficiently large recombination rates we find evidence for a discontinuous error threshold in the percolation landscape with lethal genotypes. From the perspective of quasispecies theory, the shift of the transition to larger U for decreasing p may tentatively be explained as an effect of increased selection pressure in landscapes with a larger fraction of lethal genotypes.

The results for fixed mutation rate $U = 0.1$ displayed in the right column of Fig. 9 show that the contraction of the population only occurs if the recombination rate is sufficiently large, whereas otherwise recombination increases genetic diversity and evolvability. Importantly, with increasing fraction of lethal genotypes the contraction occurs at smaller recombination rates.

Large populations. For large population sizes the number of segregating mutations grows rapidly, such that storing the part of the fitness landscape that could be revisited through recombination becomes computationally challenging. This limits the range of population sizes that can be explored. Nevertheless, the results for the discovery rate for $N = 1000$ displayed in Fig. S2 suggest that, for large populations, the interesting regime with a non-monotonous behavior in r appears in an even larger range of mutation rates than expected from the conditions $U \ll 1$ and $NU \geq 1$.

Summary of ism results. As expected, for $NU \ll 1$ or $U \approx 1$, recombination has almost no effect since the population is either monomorphic or dominated by mutations. In contrast, for $NU \geq 1$ and $U \ll 1$ the behavior is rather complex. While low recombination rates generally diversify the population and increase evolvability, the population can dramatically change its genotype composition at high recombination rates, such that most genotypes are tightly clustered around a focal genotype. In the percolation landscape, the onset of this structural change depends on the fraction of lethal genotypes, whereas for general fitness landscapes we expect it to be determined by the degree distribution of the neutral network. The focal genotype of a tightly clustered population contains no segregating muta-

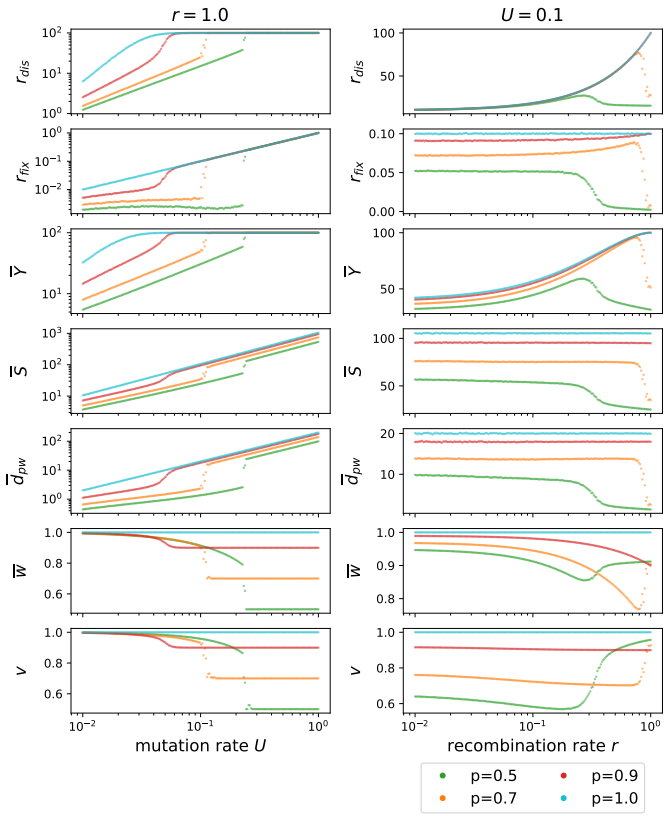

Figure 9 Cross sections of 3D plots of various measures characterizing the *ism* for four different values of the fraction p of viable genotypes. The left column shows the dependence on the mutation rate U for fixed $r = 1.0$, and the right column shows the dependence on the recombination rate r at fixed $U = 0.1$. The last row shows the fraction v of viable genotypes created by recombination events. The population size is $N = 100$.

tions. As a consequence the clustering decreases the discovery rates as well as the fixation rate, but the mean fitness increases. We conclude that recombination does not generally lead to increased evolvability, but may instead reduce diversity in order to increase fitness.

Finite-sites model

In this section, we study the effect of recombination in the finite-sites model. In contrast to the *ism* back mutations are possible, and the number of viable point mutations varies between genotypes. Therefore the population can optimize its mutational robustness to increase fitness. We keep the population size and sequence length fixed at $N = 100$ and $L = 10$, respectively, and investigate the interplay between the mutation rate per site μ and the recombination rate r.

Mutational robustness. The results for the mutational robustness displayed in Fig. 10 show a complex dependence on r and μ, which reflects the different evolutionary regimes described above. Similar to the *ism*, for $NL\mu \ll 1$ the population is essentially monomorphic. In this regime it behaves like a random walker and all genotypes have equal occupation probability, such that the mutational robustness is equal to the average network degree $p = 0.5$. For a monomorphic population, recombination has no effect. With increasing mutation rate ($NL\mu \geq 1, \mu \ll 0.5$), the population becomes polymorphic but also more mutationally robust. Strikingly, with recombination, this effect is strongly amplified, as was observed previously in the quasispecies regime (Klug *et al.* 2019; Szöllősi and Derényi 2008). Similar to the results presented for the *ism* (Fig. 8), the increase in mutational robustness is accompanied by an increase in mean fitness (Fig. S3). In Klug *et al.* (2019) we showed that mutationally robust genotypes are more likely to be the outcome of recombination events, because they have a larger share of potentially viable parents. Therefore increased mutational robustness is a universal feature of recombining populations.

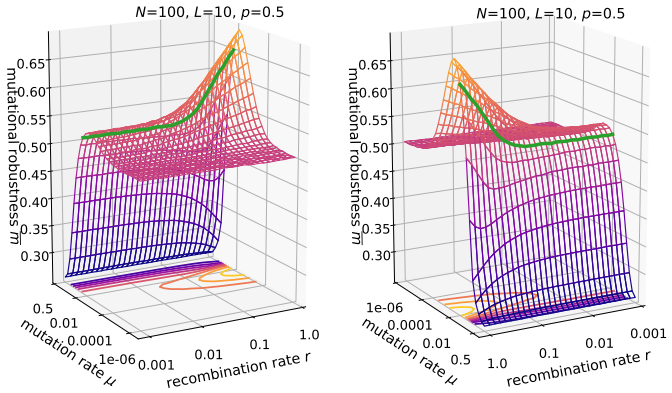

Figure 10 Mutational robustness in the *fsm* for $N = 100$, $L = 10$, $p = 0.5$. The two panels show the same data from two different viewing angles. The green line is drawn at $L\mu = 0.1$.

However, even higher mutation rates ($\mu \approx 0.5$) are detrimental to robustness, and recombination then also has a slightly negative effect. In this regime, the population is not concentrated anymore on a focal genotype but becomes highly delocalized and almost independent of the previous generation. Because of this, recombination events will produce random genotypes; note that for $\mu = 0.5$, all genotypes have the same probability after mutation, independent of viability. Thus, similar to the *ism* we can define three regimes with qualitatively different effects of recombination: (i) $NL\mu \ll 1$, (ii) $NL\mu \geq 1$ and $\mu \ll 0.5$, and (iii) $\mu \approx 0.5$.

Time to full discovery. We quantify evolvability in the *fsm* in terms of the time until all genotypes have been discovered, which can be interpreted as the inverse of the average discovery rate. For ease of comparison with the results for the *ism* shown in Fig. 2, in Fig. 11 we display the inverse of the time to full discovery in the *fsm* for $p = 1$ and $p = 0.5$. For $p = 1$, the overall behavior is similar to the *ism*, but the dependence on the recombination rate is significantly weaker. For $p = 0.5$, we see a decrease in the discovery rate for large r that is much more pronounced than in the *ism*. Whereas in the *ism* the discovery rate never drops below its value in the absence of recombination ($r = 0$), here the time to full discovery *diverges* at large recombination rates when $NL\mu \geq 0$ and $\mu \ll 1$. Furthermore, the dependence on mutation rate becomes non-monotonic at large r, which does not happen in the *ism*. The increase in the time to full discovery coincides with increased mutational robustness (Fig. 10) and a large viable recombination fraction (Fig. S3). Therefore this divergence occurs because the population is focused and entrenched in the highly robust regions of the fitness landscape.

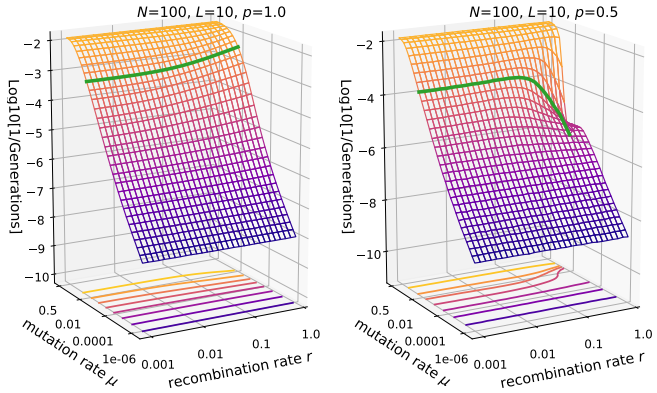

Figure 11 Reciprocal of the time to full discovery in the *fsm* with $N = 100$, $L = 10$ and $p = 1$ (left panel) vs. $p = 0.5$ (right panel). The green line is drawn at $L\mu = 0.1$.

In this way a fitness ridge surrounded by many lethal genotypes can become almost impassable for strongly recombining populations.

In Figure 12 this phenomenon is explored over a wider range of parameters. When the fraction of viable genotypes is closer to $p = 1$, the rate of discovery displays an intermediate maximum as a function of r (panel A), and this behavior becomes more pronounced for longer sequences (panel B). Figure 12 C shows that strongly recombining populations are consistently slower in discovering the remaining undiscovered genotypes, which are presumably those that exhibit low mutational robustness.

Number of mutation events until full discovery. As a second measure of evolvability we consider the total number of mutation events N_{mut} until all viable genotypes have been discovered (Fig. S4). In the random walk regime $NL\mu \ll 1$, N_{mut} is independent of μ and r, since each mutation has the probability $1/N$ to go to fixation. As the mutation rate increases and the population spreads over the genotype space, fewer mutation events are necessary for full discovery. Similar to the time to full discovery t_{fdis}, depending on the fraction of viable genotypes, recombination can be either beneficial or detrimental for evolvability. In fact the two measures are related by

$$N_{mut} = t_{fdis} N L \mu \tag{18}$$

as long double mutations, which we count as a single mutation event, are sufficiently rare ($L\mu \ll 1$).

Genetic diversity. Figure S5 summarizes results for the genetic diversity in the *fsm*. Overall the impact of recombination on the genetic diversity is similar to, but less pronounced than the results for the *ism*. In particular, no discontinuities are observed in the variation of diversity measures with mutation rate (compare to Fig. 9). Because the number of segregating mutations and the mean Hamming distance are bounded in the *fsm*, the capacity of recombination for creating diversity is limited. For example, while in the *ism* there are around 100 segregating mutations for $U = 0.1$, in the *fsm* with $L\mu = 0.1$ this number is limited by the sequence length $L = 10$ (middle row of S5). This suggests that the non-monotonic effect of recombination on evolvability in the *fsm* for $p \leq 1$ is mostly caused by an increase in mutational robustness. For $p = 1$ an analytical expression for the mean Hamming distance is derived in the Appendix. For the

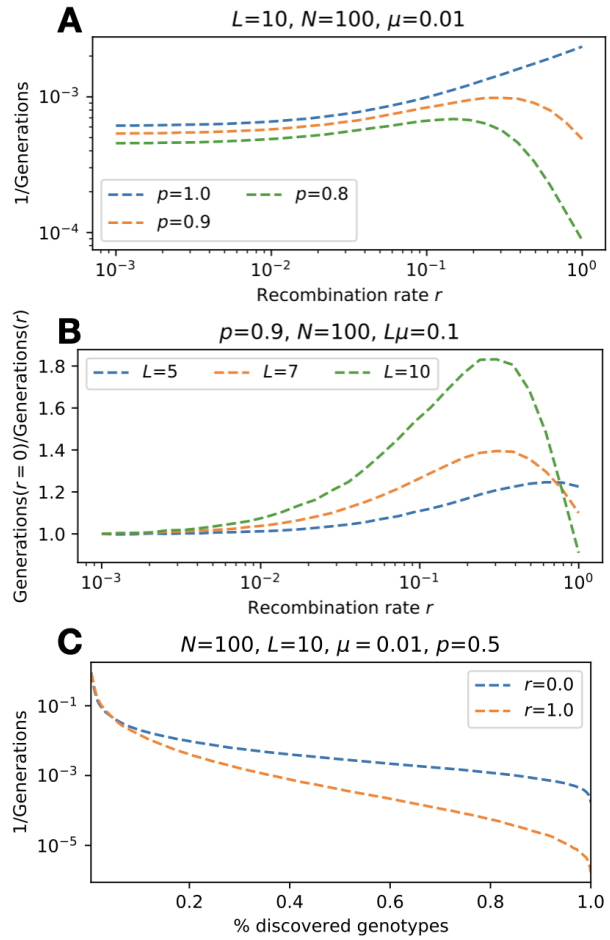

Figure 12 **A**: Reciprocal of the time to full discovery is shown as a function of recombination rate for different values of the fraction p of viable genotypes. **B**: Relative change in the reciprocal of the time to full discovery in obligately recombining ($r = 1$) vs. non-recombining ($r = 0$) populations for different values of the sequence length L and fixed genome-wide mutation rate $U = L\mu = 0.1$. **C**: Reciprocal of the time to discover a given fraction of all viable genotypes for a non-recombining and recombining population. In panel **C** the parameters are the same as in Fig. 11.

concurrent recombination scheme the result given in Eq. 34 is independent of the recombination rate, but this property is model dependent (see Recombination-induced genetic drift for further discussion).

Graph representation. To further illustrate the genotype composition of the population, a graph representation is employed in Fig. 13. These graphs show snapshots of genotype clouds that have evolved for sufficiently many generations to be independent of the initial condition. The population parameters are chosen to be in the regime $NL\mu \geq 1$, $\mu \ll 0.5$. The graphs of the recombining populations consist of significantly more edges representing mutational neighbors, which implies that they form a densely connected component in sequence space. By comparison, the graphs of the non-recombining populations have fewer edges, which implies that the populations are more dispersed. This is consistent with the results for the cross section of the genotype cloud in the *ism* in Fig. 7, which show a narrower distribution for recombining populations.

Figure S6 shows genotype clouds for larger values of L and N, but within the same evolutionary regime. The increased population size allows us to study the frequency distribution of genotypes in the population sorted by their rank. Remarkably, in non-recombining populations ($r = 0$) the distribution is exponential whereas for $r = 1$ we observe a heavy-tailed power law distribution, a feature that appears to be independent of p. The histograms in Fig. S6 also highlight the fact that, depending on the fraction p of viable genotypes, recombination can either increase or decrease the genetic diversity. In terms of mutational robustness, the graph representation shows that genotypes with high frequency exhibit an above-average robustness, thereby increasing the mutational robustness of the population.

Time evolution. So far we have studied the impact of recombination on stationary populations, where the effects of mutation, selection and drift balance on average. However, such a stationary state is generally not reached within a few generations, and in particular in the context of evolution experiments it is also important to understand the transient behavior. Since experiments usually consider a predefined small set of loci and track their evolution, we consider the temporal evolution in the *fsm*.

As an example, Fig. 14 shows the time evolution of mutational robustness m for obligately recombining and non-recombining populations at different mutation rates. The population is initially monomorphic and starts on a random viable genotype. To account for the variability between the trajectories observed in different realizations of the evolutionary process, the shading around the lines showing the average robustness represents the standard deviation of m. Analogous results for measures of evolvability and diversity are shown in Figs. S7, S8, S9 and S10.

As expected, the time scale for the establishment of the stationary regime is determined primarily by the mutation rate, since mutations create the diversity on which selection and recombination can act. At the lowest mutation rate recombining and non-recombining populations behave in the same way, and with increasing mutation rate the evolutionary regimes described above are traversed. This implies in particular that the ordering between the lines representing $r = 1$ and $r = 0$ may change as a function of μ (Figs. 14, S7 and S10). The distinction between recombining and non-recombining populations is often most pronounced at intermediate values of the mutation rate (Figs. S9 and S10).

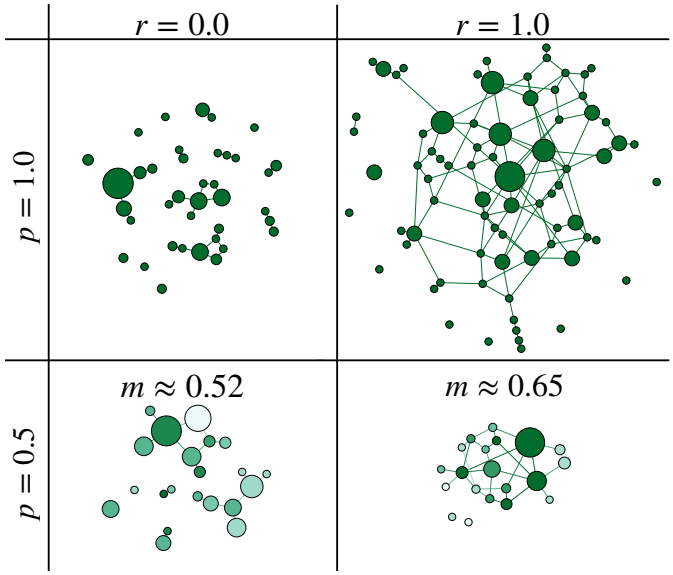

Figure 13 Graph representation of genotype clouds in the *fsm* with $N = 100$, $L = 10$, $\mu = 0.01$ after 10^6 generations of evolution starting from a random viable genotype. Links connect genotypes that differ by a single mutation, and node sizes represent the frequency of the corresponding genotype in the population; see Illustration of results for details. The networks on the left show non-recombining populations ($r = 0$) and on the right obligately recombining populations ($r = 1$). In the top panels all genotypes are viable ($p = 1$), whereas in the bottom panels half of the genotypes are lethal ($p = 0.5$). In the bottom row the mutational robustness of genotypes is shown by color coding with dark green representing high robustness and pale green low robustness, and the average robustness m is also indicated.

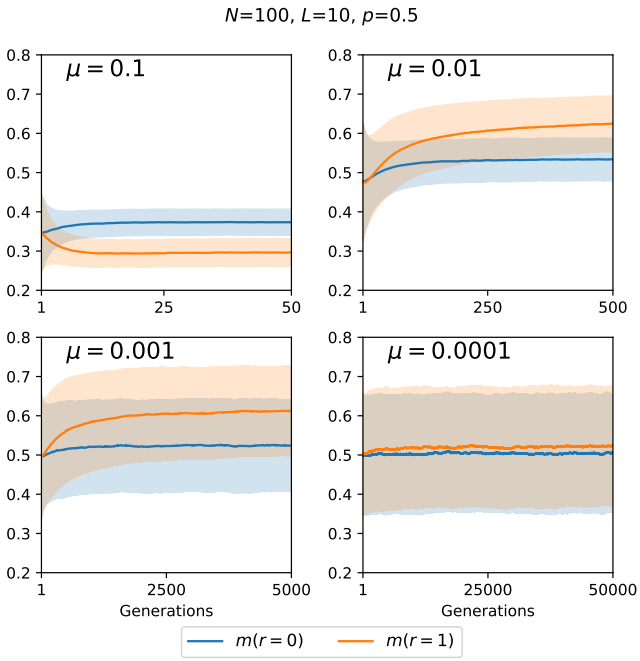

Figure 14 Time evolution of mutational robustness in the *fsm* with $N = 100$, $L = 10$, $p = 0.5$ for different values of the mutation rate μ. Each panel compares obligately recombining ($r = 1$) and non-recombining ($r = 0$) populations. Thick lines represent the mean over 5000 landscape realizations and the shaded areas the corresponding standard deviation.

Recombination-induced genetic drift

If recombination does not occur concurrently with selection but successively, it can act as an additional source of genetic drift. This recombination-induced drift is a confounding factor that needs to be accounted for when interpreting the results obtained with successive recombination models. As an example, Fig. 15 shows results for the numbers of distinct genotypes and segregating mutations for the *ism* with $p = 1$ using the simple successive recombination scheme. While in the concurrent recombination model the number of distinct genotypes is strictly increasing with r at $p = 1$ (Fig. 4), the effect of recombination in the simple successive model is mutation rate dependent and can be non-monotonic even at $p = 1$. This is due to a decrease in the number of segregating mutations and the mean Hamming distance with increasing r which occurs through the additional genetic drift (see Appendix). In the *fsm* the effect is similar and can result in an intermediate peak in the mutational robustness as a function of r when $NL\mu \approx 1$ (Fig. S11). While the effect in the *ism* occurs at all population sizes, in the *fsm* it is only significant for relatively small populations, because the number of segregating mutations is capped at L.

If the design of a simulation model or an *in vitro* experiment requires that recombination and selection occur independently, but recombination should not be an additional source of genetic drift, then successive recombination with mating pairs, as illustrated in Fig. 1, might be an option. Alternatively, one can mitigate the additional genetic drift in the simple successive recombination model by performing the selection and recombination step in each generation within a large population from which a small sample of size N is subsequently drawn.

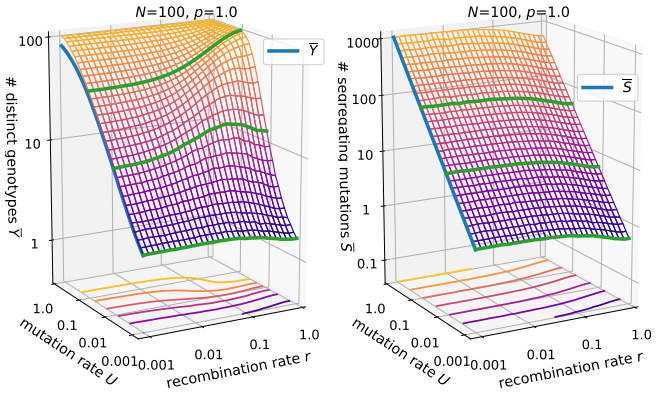

Figure 15 Number of distinct genotypes and segregating mutations in the simple successive recombination model for $p = 1$ in the *ism*. Compare to Figs. 4, 5. Green lines show constant $U = 0.1$, $U = 0.01$ and $U = 0.001$.

Discussion

The most apparent effect of recombination is to shuffle alleles among individuals. Therefore, one might think that recombination always increases genetic diversity and, through that, evolvability, as formulated in Weismann's hypothesis (Weismann 1891). In this study, we have shown that while this is true on a neutral network in the absence of lethal genotypes, the effect of recombination in the presence of lethal or highly deleterious mutations is much more complex.

More precisely, if the fraction of lethal genotypes in the fitness landscape is large enough, we find the emergence of two different regimes for the effect of recombination. While small recombination rates increase diversity in accordance with Weismann's hypothesis, at sufficiently high recombination rates we observe a strong contraction of the genotype cloud and reduced evolvability. This contraction regime is characterized by a clustering of the population in sequence space around a focal genotype, which we have shown to be the most recent common ancestor. Therefore most genotypes have only a few segregating mutations as well as a small pairwise mean Hamming distance, which leads to a reduced number of distinct genotypes. This benefits mean fitness, as recombination events more often lead to viable genotypes. Results for the finite-sites model further reveal, that polymorphic genotype clouds are most dense around mutationally robust genotypes and that a contraction through recombination therefore greatly increases mutational robustness. The trade-off is a reduced evolvability. In the infinite-sites model this is manifested through a reduced discovery rate. Furthermore, the lower frequency of segregating mutations leads to a reduced fixation rate of segregating mutations.

Therefore, even in an infinite-sites setting, recombination can, in some sense, entrench the population. However, as the number of potential mutation sites is unbounded, the discovery rate never falls below the discovery rate in the absence of recombination. In contrast, in the finite-sites model strongly recombining populations can become trapped in mutationally robust regions, such that the time to full discovery diverges. The recombination-induced trapping of the population at fitness peaks is well known from studies on non-neutral fitness landscapes (Eshel and Feldman 1970; Weinreich and Chao 2005; de Visser *et al.* 2009; Jain 2010; Park and Krug 2011; Altland *et al.* 2011), but our results show that a similar phenomenon occurs in

neutral landscapes with fitness plateaus and ridges. In terms of genetic diversity the results in the finite and infinite-sites settings are similar but more gradual in the latter case, as the number of segregating mutations and the mean Hamming distance are capped.

Overall, our numerical simulations show a very consistent increase in mutational robustness with recombination rate in polymorphic populations. Related to this is the observation that recombination leads to a heavy-tailed frequency distribution of genotype abundance, and that the most frequent genotypes have an above-average robustness, thereby increasing the robustness of the whole population. As discussed in previous work, the most frequent genotypes are more likely those that have an above-average fraction of possible viable parent combinations (Klug *et al.* 2019).

The broad scope of our investigation demonstrates that the effects of recombination vary widely across parameter combinations and evolutionary regimes, and this has to be accounted for when interpreting apparent contradictions between different experiments. Furthermore, it is important to distinguish long-term effects of recombination from short-term effects. In this study, we mainly considered long-term effects in stationary populations. Short-term effects can be different, in particular when evolution proceeds in a changing environment (Becks and Agrawal 2012; Nowak *et al.* 2014).

Data availability

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article. Numerical calculations including simulations described in this work were implemented in Python. All relevant source codes will be uploaded on github.

Acknowledgements

This work has been supported by Deutsche Forschungsgemeinschaft within SPP 1590 *Probabilistic Structures in Evolution* and CRC 1310 *Predictability in Evolution*.

Literature Cited

Altland, A., A. Fischer, J. Krug, and I. G. Szendro, 2011 Rare events in population genetics: stochastic tunneling in a two-locus model with recombination. Physical review letters **106**: 088101.

Azevedo, R. B., R. Lohaus, S. Srinivasan, K. K. Dang, and C. L. Burch, 2006 Sexual reproduction selects for robustness and negative epistasis in artificial gene networks. Nature **440**: 87–90.

Becks, L. and A. F. Agrawal, 2012 The evolution of sex is favoured during adaptation to new environments. PLoS Biology **10**: e1001317.

Bloom, J. D., J. J. Silberg, C. O. Wilke, D. A. Drummond, C. Adami, *et al.*, 2005 Thermodynamic prediction of protein neutrality. Proceedings of the National Academy of Sciences **102**: 606–611.

Boerlijst, M. C., S. Bonhoeffer, and M. A. Nowak, 1996 Viral quasi-species and recombination. Proc. R. Soc. Lond. B **263**: 1577–1584.

Bornberg-Bauer, E. and H. S. Chan, 1999 Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space. Proceedings of the National Academy of Sciences **96**: 10689–10694.

Burt, A., 2000 Perspective: sex, recombination, and the efficacy of selection—was weismann right? Evolution **54**: 337–351.

Ciliberti, S., O. C. Martin, and A. Wagner, 2007 Innovation and robustness in complex regulatory gene networks. Proceedings of the National Academy of Sciences **104**: 13591–13596.

de Visser, J. A. G. and S. F. Elena, 2007 The evolution of sex: empirical insights into the roles of epistasis and drift. Nature Reviews Genetics **8**: 139–149.

de Visser, J. A. G. and J. Krug, 2014 Empirical fitness landscapes and the predictability of evolution. Nature Reviews Genetics **15**: 480–490.

de Visser, J. A. G., S.-C. Park, and J. Krug, 2009 Exploring the effect of sex on empirical fitness landscapes. The American Naturalist **174**: S15–S30.

Domingo, E. and P. Schuster, editors, 2016 *Quasispecies: From Theory to Experimental Systems*. Springer Science & Business Media.

Eshel, I. and M. W. Feldman, 1970 On the evolutionary effect of recombination. Theoretical Population Biology **1**: 88–100.

Ewens, W. J., 2012 *Mathematical population genetics 1: theoretical introduction*. Springer Science & Business Media.

Feldman, M. W., S. P. Otto, and F. B. Christiansen, 1996 Population genetic perspectives on the evolution of recombination. Annual review of genetics **30**: 261–295.

Felsenstein, J., 1974 The evolutionary advantage of recombination. Genetics **78**: 737–756.

Gavrilets, S., 1997 Evolution and speciation on holey adaptive landscapes. Trends in ecology & evolution **12**: 307–312.

Gavrilets, S., 2004 *Fitness landscapes and the origin of species*. Princeton University Press.

Gavrilets, S. and J. Gravner, 1997 Percolation on the fitness hypercube and the evolution of reproductive isolation. Journal of theoretical biology **184**: 51–64.

Guo, H. H., J. Choe, and L. A. Loeb, 2004 Protein tolerance to random amino acid change. Proceedings of the National Academy of Sciences **101**: 9205–9210.

Haller, B. C. and P. W. Messer, 2016 SLiM 2: Flexible, Interactive Forward Genetic Simulations. Molecular Biology and Evolution **34**: 230–240.

Hughes, B., 1996 *Random Walks and Random Environments*. Clarendon Press, Oxford.

Jacomy, M., T. Venturini, S. Heymann, and M. Bastian, 2014 Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. PloS one **9**: e98679.

Jain, K., 2010 Time to fixation in the presence of recombination. Theoretical Population Biology **77**: 23–31.

Jain, K. and J. Krug, 2007 Adaptation in simple and complex fitness landscapes. In *Structural Approaches to Sequence Evolution*, edited by H. E. R. Ugo Bastolla, Markus Porto and M. Vendruscolo, Springer.

Johnson, M. S., A. Martsul, S. Kryazhimskiy, and M. M. Desai, 2019 Higher-fitness yeast genotypes are less robust to deleterious mutations. Science **366**: 490–493.

Kimura, M., 1968 Evolutionary rate at the molecular level. Nature **217**: 624–626.

Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61**: 893.

Kimura, M., 1983 *The neutral theory of molecular evolution*. Cambridge University Press.

Klug, A., S.-C. Park, and J. Krug, 2019 Recombination and muta-

tional robustness in neutral fitness landscapes. PLoS computational biology **15**: e1006884.
Kondrashov, A. S., 1988 Deleterious mutations and the evolution of sexual reproduction. Nature **336**: 435–440.
Manrubia, S., J. A. Cuesta, J. Aguirre, S. E. Ahnert, L. Altenberg, *et al.*, 2021 From genotypes to organisms: State-of-the-art and perspectives of a cornerstone in evolutionary dynamics. Physics of Life Reviews **38**: 55–106.
Maynard Smith, J., 1970 Natural selection and the concept of a protein space. Nature **225**: 563–564.
Muller, H. J., 1932 Some genetic aspects of sex. The American Naturalist **66**: 118–138.
Muller, H. J., 1964 The relation of recombination to mutational advance. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis **1**: 2–9.
Nobrega, M. A., Y. Zhu, I. Plajzer-Frick, V. Afzal, and E. M. Rubin, 2004 Megabase deletions of gene deserts result in viable mice. Nature **431**: 988–993.
Nowak, S., J. Neidhart, I. G. Szendro, and J. Krug, 2014 Multidimensional epistasis and the transitory advantage of sex. PLoS Computational Biology **10**: e1003836.
Ohta, T., 2002 Near-neutrality in evolution of genes and gene regulation. Proceedings of the National Academy of Sciences **99**: 16134–16137.
Otto, S. P., 2009 The evolutionary enigma of sex. The American Naturalist **174**: S1–S14.
Park, S.-C. and J. Krug, 2011 Bistability in two-locus models with selection, mutation, and recombination. Journal of Mathematical Biology **62**: 763–788.
Pesce, D., N. Lehman, and J. A. G. M. de Visser, 2016 Sex in a test tube: testing the benefits of in vitro recombination. Phil. Trans. R. Soc. B **371**: 20150529.
Reddy, G. and M. M. Desai, 2021 Global epistasis emerges from a generic model of a complex trait. eLife **10**: e64740.
Rodrigues, J. F. M. and A. Wagner, 2009 Evolutionary plasticity and innovations in complex metabolic reaction networks. PLoS Comput Biol **5**: e1000613.
Serva, M. and L. Peliti, 1991 A statistical model of an evolving population with sexual reproduction. Journal of Physics A: Mathematical and General **24**: L705.
Singhal, S., S. M. Gomez, and C. L. Burch, 2019 Recombination drives the evolution of mutational robustness. Current Opinion in Systems Biology **13**: 142–149.
Szöllősi, G. J. and I. Derényi, 2008 The effect of recombination on the neutral evolution of genetic robustness. Mathematical biosciences **214**: 58–62.
van Nimwegen, E., J. P. Crutchfield, and M. Huynen, 1999 Neutral evolution of mutational robustness. Proceedings of the National Academy of Sciences **96**: 9716–9720.
Wagner, A., 2005 Robustness, evolvability, and neutrality. FEBS letters **579**: 1772–1778.
Wagner, G. and P. Krall, 1993 What is the difference between models of error thresholds and Muller's ratchet? Journal of Mathematical Biology **32**: 33–44.
Wakeley, J., 2009 *Coalescent theory: An introduction*. Roberts & Co. Publishers.
Watterson, G., 1975 On the number of segregating sites in genetical models without recombination. Theoretical population biology **7**: 256–276.
Weinreich, D. M. and L. Chao, 2005 Rapid evolutionary escape by large populations from local fitness peaks is likely in nature. Evolution **59**: 1175–1182.
Weismann, A., 1891 *Essays upon heredity and kindred biological problems*, volume 1. Clarendon press.
Wilke, C. O., 2001 Adaptive evolution on neutral networks. Bulletin of Mathematical Biology **63**: 715–730.
Wilke, C. O., 2005 Quasispecies theory in the context of population genetics. BMC evolutionary biology **5**: 1–8.
Zanini, F. and R. A. Neher, 2012 FFPopSim: an efficient forward simulation package for the evolution of large populations. Bioinformatics **28**: 3332–3333.

Appendix

In the following, analytical expressions for the mean Hamming distance $\overline{d}_{pw}$ are derived for all three recombination models illustrated in Fig. 1. Throughout this appendix these models are abbreviated as *cr* (concurrent recombination), *ssr* (simple successive recombination) and *srmp* (successive recombination with mating pairs), respectively. The derivation is based on the approach of Serva and Peliti (1991), which is generalized to take into account recombination. It is assumed that there are no lethal genotypes ($p = 1$). The results apply to the *fsm* and the *ism* and show that the recombination rate influences $\overline{d}_{pw}$ only in the *ssr* model. The r-dependence decreases with N in the *fsm* but remains independent of N in the *ism*.

To start, in the selection step each individual α picks a parent α' at random from the previous generation. During the mutation step, a mutation occurs at each locus i with probability μ, changing its state from -1 to 1 or vice versa,

$$\sigma_i^\alpha(t+1) = \begin{cases} \sigma_i^{\alpha'}(t), & \text{with prob. } 1-\mu \\ -\sigma_i^{\alpha'}(t), & \text{with prob. } \mu. \end{cases} \tag{19}$$

This can also be written as

$$\sigma_i^\alpha(t+1) = \epsilon_i^\alpha(t)\sigma_i^{\alpha'}(t) \tag{20}$$

with

$$\epsilon_i^\alpha(t) = \begin{cases} +1, & \text{with prob. } 1-\mu \\ -1, & \text{with prob. } \mu. \end{cases} \tag{21}$$

During the recombination step, individuals recombine at rate r. Altogether this leads to

$$\begin{aligned} \sigma_i^\alpha(t+1) =& \epsilon_i^\alpha(t)\Big[\kappa^\alpha(t)\left(\xi_i^\alpha(t)\sigma_i^{\alpha'}(t) + (1-\xi_i^\alpha(t))\sigma_i^{\alpha''}(t)\right) \\ &+(1-\kappa^\alpha(t))\sigma_i^{\alpha'''}(t+1)\Big] \end{aligned} \tag{22}$$

where α', α'' are the parental genotypes in case of recombination and α''' is the parent in case of no recombination. The random variable κ^α determines whether a recombination event occurs and is given by

$$\kappa^\alpha(t) = \begin{cases} 1 & \text{with prob. } r, \\ 0 & \text{with prob. } 1-r. \end{cases} \tag{23}$$

The random variable ξ_i^α determines from which parent the allele is taken in case of a recombination event:

$$\xi_i^\alpha(t) = \begin{cases} 1 & \text{with prob. } 1/2, \\ 0 & \text{with prob. } 1/2. \end{cases} \tag{24}$$

Next the relatedness Q of the population is computed, which is defined by

$$Q = \binom{N}{2}^{-1} \sum_{(\alpha,\beta)} \frac{1}{L} \sum_{i=1}^{L} \sigma_i^\alpha \sigma_i^\beta \tag{25}$$

where the sum runs over all different pairs of individuals (α, β). We are interested in the average relatedness $\overline{Q}$ in the stationary state, and therefore make use of Eq. 22 by evaluating

$$\begin{aligned}
&\overline{\sigma_i^\alpha(t+1)\sigma_i^\beta(t+1)} \\
&= \overline{\underbrace{\epsilon_i^\alpha \epsilon_i^\beta}_{(1-2\mu)^2} \Bigg[\underbrace{\kappa^\alpha \kappa^\beta}_{r^2} \underbrace{\left(\xi_i^\alpha \sigma_i^{\alpha'} + (1-\xi_i^\alpha)\sigma_i^{\alpha''}\right)\left(\xi_i^\beta \sigma_i^{\beta'} + (1-\xi_i^\beta)\sigma_i^{\beta''}\right)}_{(*)}} \\
&\quad + \overline{\underbrace{(1-\kappa^\alpha)(1-\kappa^\beta)}_{(1-r)^2} \sigma_i^{\alpha'''}\sigma_i^{\beta'''}} \\
&\quad + \overline{\underbrace{\kappa^\alpha(1-\kappa^\beta)}_{r(1-r)} \underbrace{\left(\xi_i^\alpha \sigma_i^{\alpha'} + (1-\xi_i^\alpha)\sigma_i^{\alpha''}\right)\sigma_i^{\beta'''}}_{\overline{\sigma_i^{\alpha'}\sigma_i^{\beta'''}}}} \\
&\quad + \overline{\underbrace{(1-\kappa^\alpha)\kappa^\beta}_{r(1-r)} \underbrace{\sigma_i^{\alpha'''}\left(\xi_i^\beta \sigma_i^{\beta'} + (1-\xi_i^\beta)\sigma_i^{\beta''}\right)}_{\overline{\sigma_i^{\beta'}\sigma_i^{\alpha'''}}}} \Bigg].
\end{aligned} \tag{26}$$

To simplify notation, the t-dependence is suppressed on the right hand side. Equation 26 holds for all three models, but differences emerge in the probability that the individuals α, β share a parent as well as in the term marked by $(*)$.

The probability that two non-recombining individuals have the same parent $\alpha''' = \beta'''$ through selection is given by $1/N$ in all models. This leads to

$$\overline{\sigma_i^{\beta'''}(t)\sigma_i^{\alpha'''}(t)} = \frac{1}{N} + \left(1 - \frac{1}{N}\right)\overline{Q}(t). \tag{27}$$

With *cr* and *srwm*, a recombining and non-recombining individual share parents $\alpha' = \beta'''$ with probability $1/N$, leading again to the right hand side of Eq. 27. However, for *ssr*, the probability that they share a common parent $\alpha' = \beta'''$ is $1/N + (1-1/N)1/N \approx 2/N$, as this can occur either during the recombination or selection step. This yields

$$\overline{\sigma_i^{\beta'''}(t)\sigma_i^{\alpha'}(t)} = \overline{\sigma_i^{\beta'}(t)\sigma_i^{\alpha'''}(t)} \approx \frac{2}{N} + \left(1 - \frac{2}{N}\right)\overline{Q}(t). \tag{28}$$

We next turn to the evaluation of the term in Eq. 26 marked by $(*)$. In the *cr* and *ssr* model the random variables ξ_i are not correlated between individuals. Therefore in both models $(*)$ simplifies to

$$(*) = \overline{\sigma_i^{\alpha'}(t)\sigma_i^{\beta'}(t)}. \tag{29}$$

For *cr* this leads to the right hand side of Eq. 27, while for *ssr* we get Eq. 28 using the same argument as before. Similar to *ssr*, for *srwm*, recombining individuals have an increased chance of sharing a parent $\alpha' = \beta'$ since they either can belong to the same mating pair with probability $2/(rN)$ or share a parent during selection with probability $1/N$. However, this is exactly balanced by the constraint that mating pairs are complementary in their recombined material, which is reflected in a correlation of the random variables ξ_i. Simply put, the increased chance that two individuals share the same parent is offset by the constraint that they always inherit the allele of the respective other parental genotype. Therefore $(*)$ again leads to Eq. 27 for *srwm*.

Summarizing, we have

$$\begin{aligned}
\text{suc. rec. pairs:} \quad & \overline{Q(t+1)} = (1-2\mu)^2\left[\frac{1}{N} + (1-\frac{1}{N})\overline{Q}(t)\right]. \\
\text{conc. rec.:} \quad & \overline{Q(t+1)} = (1-2\mu)^2\left[\frac{1}{N} + (1-\frac{1}{N})\overline{Q}(t)\right]. \\
\text{simp. suc. rec.:} \quad & \overline{Q(t+1)} \approx (1-2\mu)^2\Big[r^2(\frac{2}{N} + (1-\frac{2}{N})\overline{Q}(t)) \\
& \quad + (1-r)^2(\frac{1}{N} + (1-\frac{1}{N})\overline{Q}(t)) \\
& \quad + 2r(1-r)(\frac{2}{N} + (1-\frac{2}{N})\overline{Q}(t))\Big].
\end{aligned} \tag{30}$$

Note that the r dependence has vanished with *cr* & *srmp*. Next we compute the stationary relatedness by setting $\overline{Q(t+1)} = \overline{Q(t)}$. For *cr* & *srmp* this yields

$$\overline{Q} = \frac{(1-2\mu)^2}{4(1-\mu)\mu(N-1)+1}, \tag{31}$$

while for *ssr* we get

$$\overline{Q} = \frac{(1-2\mu)^2[1+r(2-r)]}{4(1-\mu)\mu(N-1) - (1-2\mu)^2 r^2 + 2(1-2\mu)^2 r + 1}. \tag{32}$$

The relatedness is connected to the mean Hamming distance through

$$\overline{d}_{pw} = \frac{L(1-\overline{Q})}{2}, \tag{33}$$

which leads in the cases of *cr* & *srmp* to

$$\overline{d}_{pw} = \frac{2(1-\mu)\mu LN}{4(1-\mu)\mu(N-1)+1} \tag{34}$$

and in the case of *ssr* to

$$\overline{d}_{pw} = \frac{2(1-\mu)\mu LN}{4(1-\mu)\mu(N-1) - (1-2\mu)^2 r^2 + 2(1-2\mu)^2 r + 1}. \tag{35}$$

In Figures S5 and S12 we compare the expressions in Eqs. 26 and 35 to numerical simulations and find excellent agreement.

In the deterministic limit $N \to \infty$ with finite L the mean Hamming distance is the same for all recombination models:

$$\overline{d}_{pw} = \frac{L}{2}. \tag{36}$$

However, in the *ism* limit ($L \to \infty, \mu \to 0, L\mu = U$) the result for *cr* & *srmp* leads to

$$\overline{d}_{pw} = 2NU = \theta \tag{37}$$

while the result for *ssr* reads

$$\overline{d}_{pw} = \frac{\theta}{1+r(2-r)}. \tag{38}$$

Hence in this limit the dependence on the recombination rate persists independent of the population size.

To obtain an analytical expression for $\overline{S}$ in the *ism* under *ssr* we adopt a relation between the mean Hamming distance and the number of segregating sites $\overline{S}$ that has been derived for the non-recombining case (Wakeley 2009). This leads us to

$$\overline{S} \approx \overline{d}_{pw} \sum_{i=1}^{N-1} \frac{1}{i} = \frac{\theta}{1+r(2-r)} \sum_{i=1}^{N-1} \frac{1}{i}. \tag{39}$$

Figure S12 shows that, at least for the parameter regime of interest here, Eq. 39 provides an accurate approximation of the simulation results.

Supplementary figures

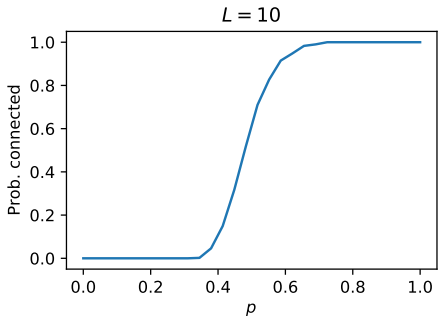

Figure S1 Probability that the network of viable genotypes is connected in a percolation landscape, where genotypes are viable or lethal independently with probability p. The sequence length is $L = 10$.

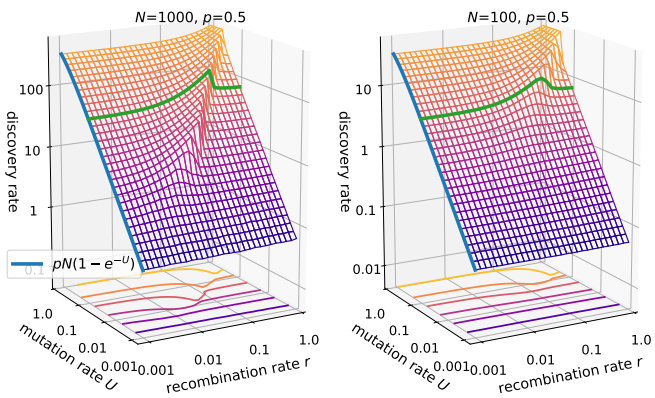

Figure S2 Discovery rate in the *ism* for viable fraction $p = 0.5$ and population size $N = 1000$ (left panel) vs. $N = 100$ (right panel). The green line is drawn at $U = 0.1$ and the blue lines in both panels show Eq. 11. The right panel is identical to that in Fig. 2.

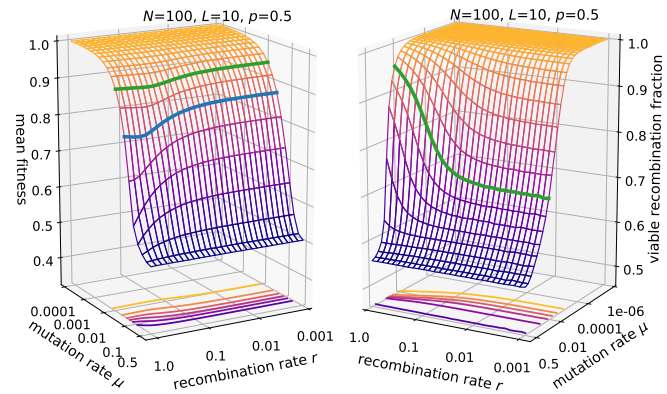

Figure S3 Mean fitness and viable recombination fraction in the *fsm* with concurrent recombination. The green line is drawn at $\mu = 0.01$ in both panels. Similar to the results for the *ism* (Fig. 8), the fitness displays an intermediate minimum at the point where the population structure changes. This is best visible at $\mu = 0.03$ (blue line). Compared to the *ism* the variation in mean fitness and viable recombination fraction is less pronounced.

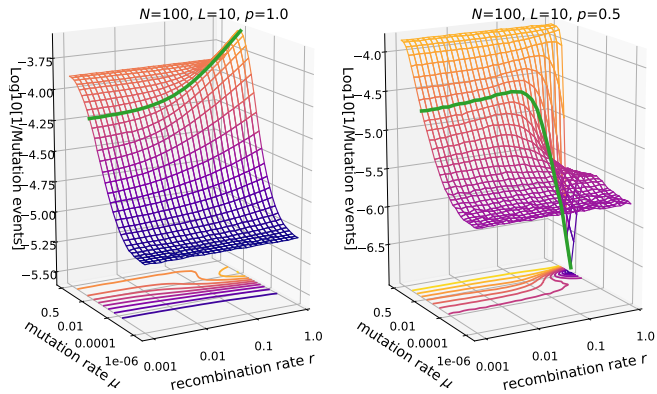

Figure S4 Reciprocal of the total number of mutation events until full discovery in the *fsm* with $N = 100$, $L = 10$ and $p = 1.0$ (left panel) vs. $p = 0.5$ (right panel). The green line is drawn at $L\mu = 0.1$.

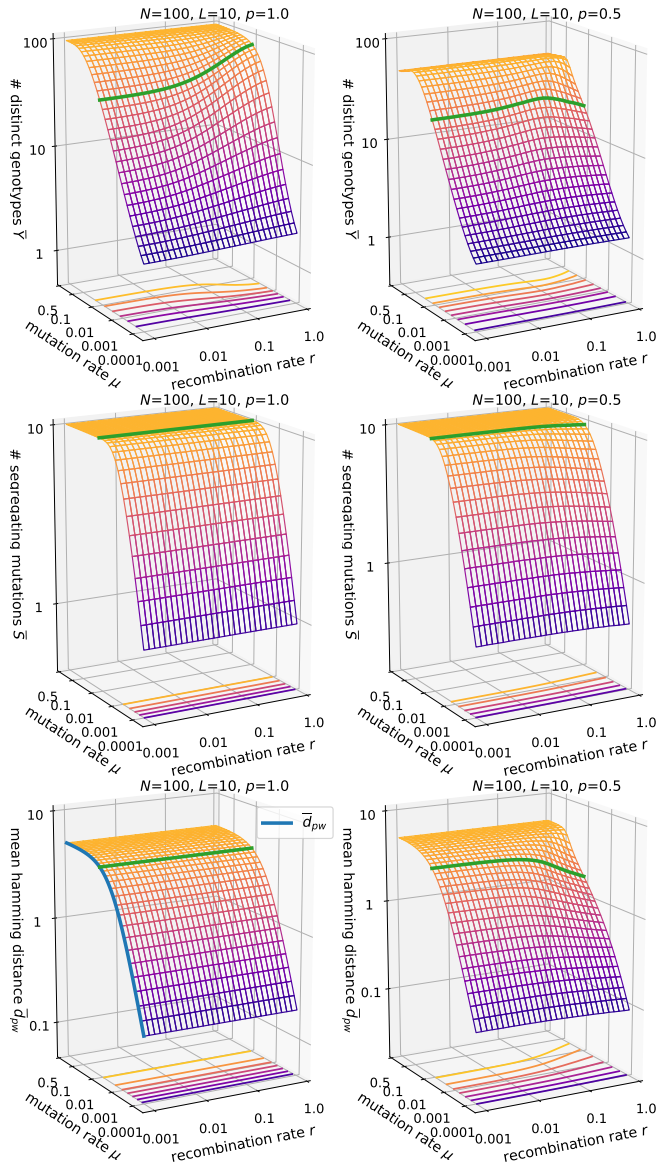

Figure S5 Genetic diversity in the *fsm* with $N = 100$, $L = 10$ and $p = 1.0$ (left column) vs. $p = 0.5$ (right column). The green line is drawn at $L\mu = 0.1$. The blue line in the bottom left panel shows the expression in Eq. 34 for the mean Hamming distance at $p = 1$.

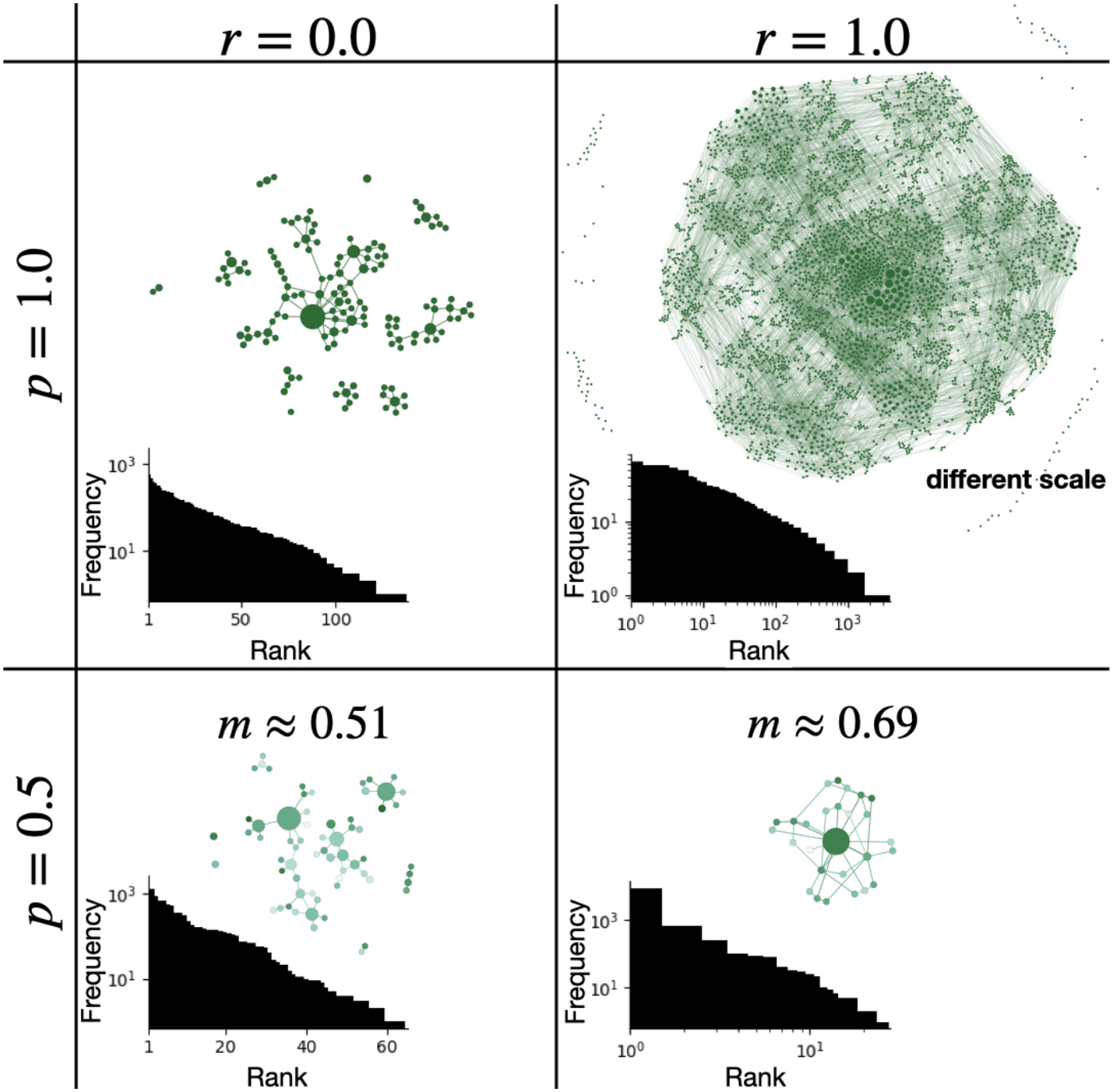

Figure S6 Same as Fig. 13 but with $N = 10000$, $L = 14$, $\mu = 0.0001$ after 10^7 generations. Inset histograms show the frequency distribution sorted by rank. The histograms are in semi-log scale for $r = 0$ and in log-log scale for $r - 1$. Through the number of ranks, the histograms also display the number of existing distinct genotypes. Note that for both values of p recombination makes the frequency distribution heavy-tailed, but it may either increase or decrease the number of distinct genotypes.

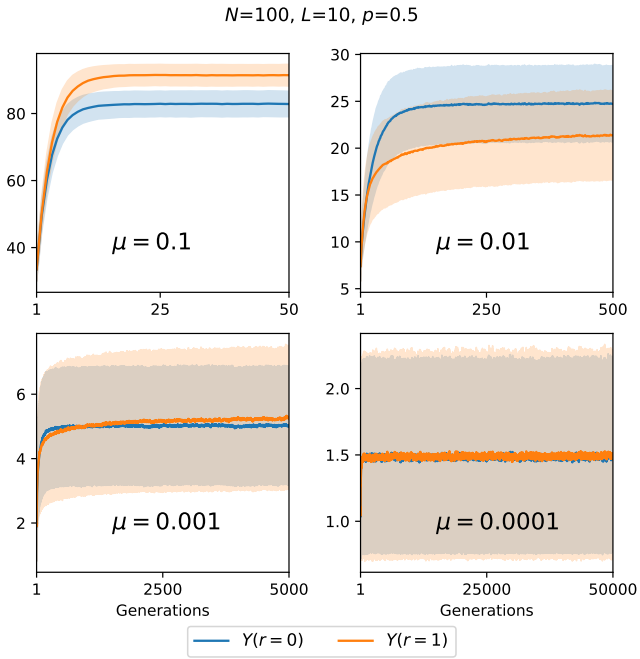

Figure S7 Time evolution of of the number of distinct genotypes in the *fsm* with $N = 100$, $L = 10$, $p = 0.5$ for different values of the mutation rate μ. Each panel compares obligately recombining ($r = 1$) and non-recombining ($r = 0$) populations. Thick lines represent the mean over 5000 landscape realizations and the shaded areas the corresponding standard deviation.

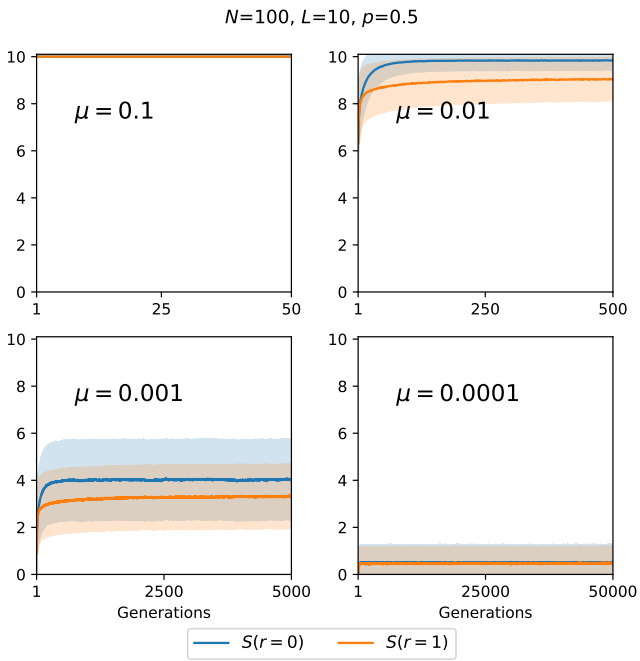

Figure S8 Time evolution of the number of segregating mutations in the *fsm* with $N = 100$, $L = 10$, $p = 0.5$ for different values of the mutation rate μ. Each panel compares obligately recombining ($r = 1$) and non-recombining ($r = 0$) populations. Thick lines represent the mean over 5000 landscape realizations and the shaded areas the corresponding standard deviation.

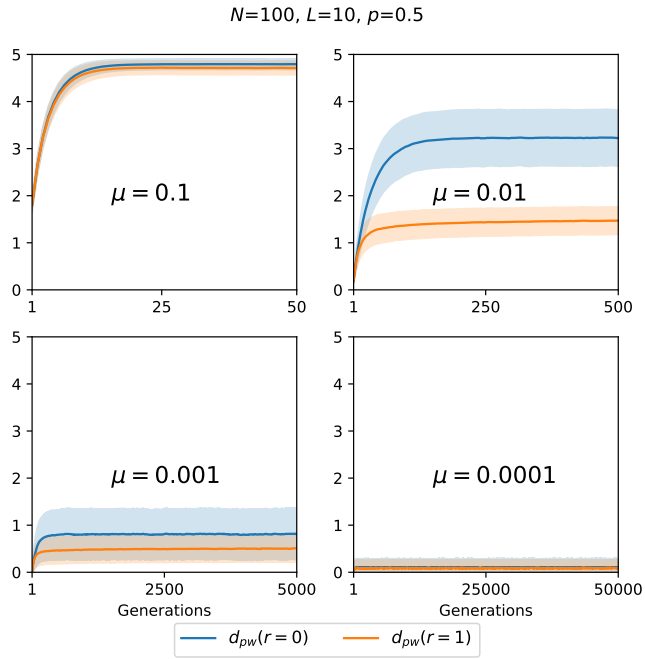

Figure S9 Time evolution of the pairwise mean Hamming distance in the *fsm* with $N = 100$, $L = 10$, $p = 0.5$ for different values of the mutation rate μ. Each panel compares obligately recombining ($r = 1$) and non-recombining ($r = 0$) populations. Thick lines represent the mean over 5000 landscape realizations and the shaded areas the corresponding standard deviation.

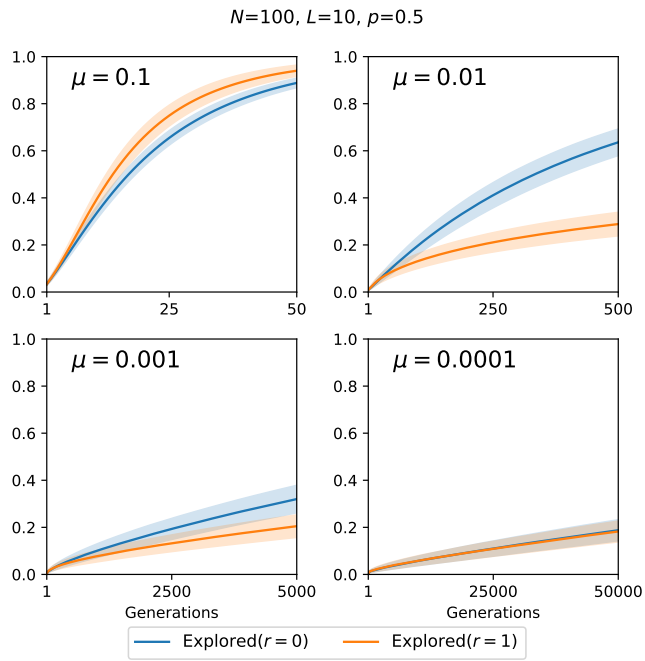

Figure S10 Time evolution of the fraction of explored viable genotypes in the *fsm* with $N = 100$, $L = 10$, $p = 0.5$ for different values of the mutation rate μ. Each panel compares obligately recombining ($r = 1$) and non-recombining ($r = 0$) populations. Thick lines represent the mean over 5000 landscape realizations and the shaded areas the corresponding standard deviation.

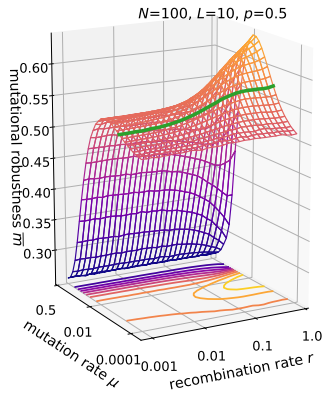

Figure S11 Mutational robustness in the *fsm* with $N = 100$, $L = 10$, $p = 0.5$ and simple successive recombination dynamics. The green line at $\mu = 0.001$ ($NL\mu = 1$) shows a non-monotonic variation with recombination rate, which is caused by recombination-dependent genetic drift.

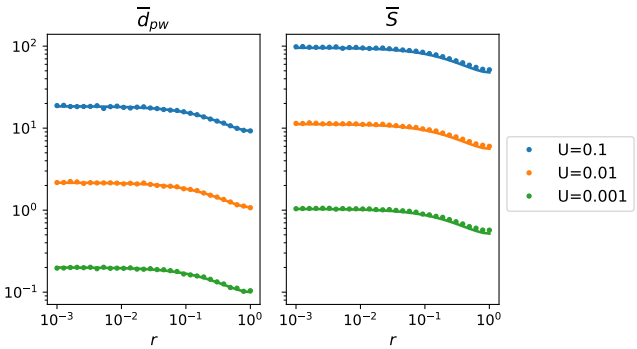

Figure S12 Comparison of numerical results (represented by dots) for the mean Hamming distance $\overline{d}_{pw}$ and the mean number of segregating sites $\overline{S}$ to the analytical expressions in Eqs. 35 and 39 (represented by lines), for the simple successive recombination model. Simulations were carried out using the *ism* with population size $N = 100$ and three different mutation rates. While for $\overline{d}_{pw}$ the fit is perfect, for $\overline{S}$ some deviations are discernible at large r.