Published in final edited form as:

Theor Popul Biol. 2007 March ; 71(2): 239–250.

Highly Fit Ancestors of a Partly Sexual Haploid Population

I. M. Rouzine^{†,*} and J.M. Coffin^{†,‡}

[†] *Department of Microbiology and Molecular Biology Tufts University, Boston, MA 02111*

[‡] *HIV Drug Resistance Program, National Cancer Institute, Frederick, MD 21702*

Abstract

Earlier, using the semi-deterministic solitary wave approach, we have investigated accumulation of pre-existing beneficial alleles in genomes consisting of a large number of simultaneously evolving sites in the presence of selection and infrequent recombination with small rate r per genome. Our previous results for the dynamics of the fitness distribution of genomes are now interpreted in terms of the life cycle of recombinant clones. We show that, at sufficiently small r , the clones dominating fitness classes, at the moment of their birth, are nearly the best fit in a population. New progeny clones are mostly generated by parental genomes whose fitness falls within a narrow interval in the middle of the high-fitness tail of fitness distribution. We also derive the fitness distribution for the distant ancestors of sites of a randomly sampled genome and show that its form is controlled by a single composite model parameter proportional to r . The ancestral fitness distribution differs dramatically from the fitness distribution of the entire ancient population: it is much broader and localized in the high-fitness tail of the ancient population. We generalize these results to the case of moderately small r to conclude that, regardless of fitness of an individual, all its distant ancestors are exceptionally well fit.

Keywords

linkage; recombination; selection; multi-locus; fitness; ancestor; traveling wave; semi-deterministic

Evolution of a single site (locus) in a genome, ignoring all others, is readily described mathematically by a model including mutation rate, selective advantage, population size, and other parameters of a population. However, linkage between many evolving sites in a genome (i.e., their inheritance together, as a package) has drastic effects on evolution. The effect of linkage on evolutionary rates has been addressed analytically for few-site models. In particular, Fisher (Fisher 1930) and Muller (Muller 1932) suggested that fixation of advantageous genotypes is slowed by linkage, because emerging clones compete for population space. Later work on two- and three-site models confirmed this effect and offered a variety of interpretations (Hill and Robertson 1966; Felsenstein 1974; Otto and Barton 1997; Hey 1998). The “clonal interference” approach (Gerrish and Lenski 1998; Orr 2000; Wilke 2004) considers competition between fixation of two consecutive beneficial mutations with different fitness advantages and neglects mutations at other sites in mutant clones. When the population size N increases, this approximation predicts saturation of the substitution rate at a level much lower than predicted by a single-site model.

*Corresponding author. Department of Microbiology and Molecular Biology Tufts University, Boston, MA 02111. Phone: 617-636-6759, fax: 617-636-4086, e-mail: irouzine@tufts.edu

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Another major effect of linkage is Muller's ratchet (Felsenstein 1974), which represents enhanced accumulation of deleterious mutations in asexual populations. When genetic drift and mutation operate, the best-fit genotype in the population will eventually be lost from population, despite the action of selection; then the next best-fit genotype will be lost, and so on. Prior analytic studies of Muller's ratchet have considered the limit when a population is very large, focusing on the distribution of genomes over the number of deleterious alleles (or fitness distribution). In that limit, the best-fit class is very large, and the distribution is nearly always close to the equilibrium distribution with a shifted center (Kimura and Maruyama 1966). After the best-fit genotype is lost, the distribution center moves one notch, and equilibrium is regained rapidly. The average time to loss of the best-fit genotype has been calculated using diffusion theory (Haigh 1978;Stephan, Chao, and Smale 1993;Charlesworth and Charlesworth 1997;Gordo and Charlesworth 2000). In this limit, the ratchet rate is very small.

Relatively recently, an approach to multi-site asexual evolution applicable in a broad range of population sizes was proposed (Tsimring, Levine, and Kessler 1996;Rouzine, Wakeley, and Coffin 2003). As in the quoted work on Muller's ratchet, genomes are grouped by their fitness, and all fitness classes are treated deterministically, except for the small best-fit group described either by a simple cutoff or, more accurately, by a diffusion equation (Rouzine, Wakeley, and Coffin 2003). In contrast to the limit of very large N (Haigh 1978;Stephan, Chao, and Smale 1993;Charlesworth and Charlesworth 1997;Gordo and Charlesworth 2000), the best-fit class is not small. The fitness distribution is far from equilibrium and represents a continuously moving traveling ("solitary") wave whose profile is determined from a non-linear balance equation and whose speed is limited by stochastic dynamics of the best-fit class.

The solitary wave method can take into account both deleterious and beneficial mutation and predict overall accumulation of either type, depending on the parameter range (Tsimring, Levine, and Kessler 1996;Rouzine, Wakeley, and Coffin 2003). In particular, in the absence of beneficial mutations or at high average fitness values, the method accurately predicts the rate of Muller's ratchet up to rather small population sizes, $N \sim 10$. In the absence of deleterious mutations or for larger N or smaller average fitness, the approach describes adaptation of asexual populations under the Fisher-Muller-Hill-Robertson effect. Results show that clonal interference between emerging clones (Gerrish and Lenski 1998;Orr 2000;Wilke 2004) is partly resolved by adding more beneficial mutations to the already existing clones. As a result, the substitution rate increases slowly (logarithmically) with N until, at extremely large N , it reaches the value predicted by the single-site model (Rouzine, Wakeley, and Coffin 2003).

Recombination has been proposed as a mechanism that evolved to counteract the adverse effects of linkage on progressive evolution of organisms and to accelerate fixation of beneficial mutations (Fisher 1930;Muller 1932;Maynard Smith 1971;Pamilo, Nei, and Li 1987;Charlesworth 1990;Kondrashov 1993;Barton 1995). To account for a large number of evolving sites, the "solitary wave" method was generalized for sexual models relevant for bacterial (Cohen, Kessler, and Levine 2005) and viral (Rouzine and Coffin 2005) populations. Recently, to study the role of recombination in its pure form, we considered the case when beneficial alleles at all sites pre-exist neglecting new mutation events, and a small fraction of genomes r recombine each generation (Rouzine and Coffin 2005). At small selection coefficients, the transition to the independent-site limit was predicted to occur at much smaller N under recombination than in the asexual case.

In the present work, we build on our previous theory (Rouzine and Coffin 2005), focusing on life cycles of separate clones, including the birth of new recombinant clones, their growth, generation of progeny, and extinction. We study the history of fitness in distant ancestors of a site in a randomly sampled genome and derive the ancestral fitness distribution. We also

generalize our previous results to the case of larger (but still small) recombination parameter r .

The article is organized, as follows. In the *Methods* section, we introduce the model and the solitary wave approach. In *Results and Discussion*, our main results are described in the qualitative form. Formal derivation of these results is presented in *Derivation*.

METHODS

Model

The basic model (Rouzine and Coffin 2005) considers a population of genomes, each with $L \gg 1$ linked sites that carry either a better-fit or a less-fit allele. The relative fitness difference between the two allele types, $s \ll 1$, is assumed to be the same for all sites. Each generation, a genome is replaced by its progeny, whose number obeys Poisson statistics with the average number proportional to the total fitness of the genome, subject to the restriction that the total population size N is constant. A small fraction of the genomes, $r \ll 1$, are not copied directly to progeny but first undergo recombination with other, randomly sampled, genomes. Epistasis between sites is neglected. The total fitness of a genome is given by $\exp(-sk)$, where k is the total number of less-fit alleles in the genome (mutational load).

Beneficial alleles exist at the beginning at low frequency f_{bf} for each site and are distributed randomly among sites and genomes. The number of allele copies per site is assumed to be sufficiently large, $Nf_{bf} \gg 1/s$, that the loss of the alleles at the onset of evolution due to random drift can be neglected. New mutation events are neglected.

Basic approach

The main idea of the approach (Tsimring, Levine, and Kessler 1996; Shpak and Kondrashov 1999; Barton and Shpak 2000; Rouzine, Wakeley, and Coffin 2003; Rouzine and Coffin 2005) is to group genomes according to the less-fit allele number k . Thus, each group consists of genomes with the same fitness $\exp(-sk)$. We define the distribution function $f(k, t)$ as the probability that a genome has less-fit allele number k and treat it in the deterministic approximation, with the exception of the best-fit group, which is small and subject to strong stochastic effects. The rationale behind this semi-deterministic approximation is that, as we show below, the fitness distribution decays very rapidly towards its high-fitness edge, and the next-to-best-fit group is already large enough to be considered as deterministic. (Although the least-fit group is also small, it has little effect on evolution of a population.)

Dynamic equation

The dynamics of $f(k, t)$ is described by a kinetic equation

$$\begin{aligned} \frac{\partial f(k, t)}{\partial t} &= (1 - r) \left\{ e^{-s[k - k_{av}(t)]} - 1 \right\} f(k, t) + r[R(k, t) - f(k, t)] \\ &\approx -s[k - k_{av}(t)] f(k, t) + r[R(k, t) - f(k, t)], \\ \int dk f(k, t) &= \int dk R(k, t) = 1 \end{aligned} \quad (1)$$

The first term on the right-hand side of either line of Eq. 1 describes asexual reproduction and death. The average less-fit allele number k_{av} is defined, as given by

$$\exp[-sk_{av}(t)] \equiv \int dk \exp(-sk) f(k, t),$$

which ensures conservation of the total population size N , i.e., normalization of $f(k,t)$. The selection coefficient s is assumed to be small, so that $s|k - k_{av}| \ll 1$ for all fitness groups, and $k_{av} \approx \int dk k f(k,t)$. The terms $rR(k,t)$ and $-rf(k,t)$ describe, respectively, the generation and loss of genomes with k mutations due to recombination events.

The form of the recombination gain function, $R(k,t)$, depends on the details of the recombination model. We assume that the number of crossovers per recombination event is large, so that a random half of each parental genome is transferred to the progeny recombinant genome. We also need to make some assumptions regarding correlations of the probabilities of different allele types between different sites in a genome (linkage disequilibrium) and between different genomes (inter-genomic correlations).

Recently, we carried out direct Monte-Carlo simulation of populations of long individual genomes in the framework of the present model (Gheorghiu, Coffin, and Rouzine 2006). Although the model naturally includes linkage in the sense defined in our introduction (passing of all sites not separated by recombination crossovers to the progeny as a package), very small linkage disequilibrium was found even for the closest pairs of sites. Under the same conditions (small N), simulation demonstrated strong correlation *between* genomes at homologous sites concurrent with a strong decrease of the substitution rate below the single-site model result (Gheorghiu, Coffin, and Rouzine 2006).

Based on these numeric results, in the present work, we assume that linkage disequilibrium is absent, i.e., that alleles of each type are distributed almost randomly between sites of a genome, restricted only by correlations *between* genomes. The approximation in which linkage disequilibrium is neglected is known as hypergeometric or symmetric model (Shpak and Kondrashov 1999; Barton and Shpak 2000). (Note that the conclusion about the general instability of the state without linkage disequilibrium made in these two works refers to a rather different case, infinite population in equilibrium under stabilizing selection (Gaussian fitness landscape centered at non-zero k). In contrast, in the present work, we consider non-equilibrium dynamics of finite populations under purifying selection (exponential fitness landscape). For purifying selection in the absence of mutation events, the equilibrium state of population is trivial: all genomes have $k=0$.)

We will take into account correlations between genomes arising from identity by descent for separate sites. Let us choose two genomes from a population, and let C be the fraction of homologous sites (sites with the same number in both genomes) that descend from the same ancestor that existed after the onset of evolution. In the absence of new mutations, homologous sites that are identical by descent must be monomorphic, i.e., have alleles of the same type. We can express $R(k,t)$ in terms of the distribution $f(k,t)$, assuming that, (i) for sites not identical by descent, probabilities of specific allele types do not correlate between the two genomes of given fitness, (ii) C is the same for all genome pairs that do not belong to the same clone, and (iii) a population is not too close to full fixation of all beneficial alleles ($k_{av}=0$), so that $|k - k_{av}| \ll k_{av}$ for all relevant k . The last assumption is not an approximation, but only limits the time interval under consideration. Under these assumptions, we have a Gaussian distribution for the relative less-fit allele number of the offspring around the average number of the parents, as given by

$$R(k, t) = \frac{1}{\sqrt{\pi w}} \int dk_1 \int dk_2 f(k_1, t) f(k_2, t) e^{-\frac{[(k_1 + k_2)/2 - k]^2}{w^2}} \quad (2)$$

$$w^2 \equiv (1 - C)k_{av}(1 - k_{av}/L) \quad (3)$$

Eq. 2 can be derived either from Gaussian distribution of k_1, k_2 in parental genomes or from a more general hypergeometric distribution of k in progeny in the limit $k \gg 1$ (Shpak and Kondrashov 1999; Barton and Shpak 2000). The parameter w^2 is the variance of the less-fit allele number of progeny k , given numbers k_1 and k_2 of its parents. Note that w^2 is equal to a half of the average genetic distance between two randomly sampled genomes. In qualitative terms, the less-fit allele number on the inherited half of a parent genome can be larger or smaller than the average, and the fluctuation is transferred to the progeny. Factor $1-C$ in Eq. 3 reflects the fact that sites identical by descent are monomorphic and, hence, not affected by recombination. We study analytically dynamics of identity by decent elsewhere (Rouzine and Coffin 2006); in the present work, C is treated as an external parameter.

Note, that the effects of selection and recombination are described in kinetic equation, Eq. 1, by two independent, additive terms. This approximation is valid, if both the recombination rate and the selection coefficient are small, as given by $r \ll 1$ and $s|k - k_{av}| \ll 1$ for all relevant k . The first condition implies that reproduction proceeds, mostly, in the asexual fashion, through amplification or decay of clones of genomes. The second condition implies that the effect of genome fitness on its chance to recombine is proportional to rs and can be neglected.

As we showed previously (Rouzine and Coffin 2005), Eqs. 1 and 2 have a solution in the form of a localized traveling wave (“solitary wave”), as given by $f(k, t) = \varphi[k - k_{av}(t)]$, $R(k, t) = \rho[k - k_{av}(t)]$. In this case, Eqs. 1 and 2 are reduced to a form

$$V \frac{d\phi}{dx} = -sx\phi(x) + r[\rho(x) - \phi(x)] \quad (4)$$

$$\rho(x) = \frac{1}{\sqrt{\pi}w} \int dx_1 \int dx_2 \phi(x_1) \phi(x_2) e^{-\frac{[(x_1+x_2)/2-x]^2}{w^2}} \quad (5)$$

where the “relative less-fit allele number” $x = k - k_{av}(t)$ is the only independent variable, and $V \equiv -dk_{av}/dt$ is the average substitution rate (the solitary wave velocity).

Note that the “wave profile” $\varphi(x)$ varies slowly in time due to the time dependence of w^2 . The partial time derivative of φ can be neglected under approximation (iii) (Rouzine and Coffin 2005). In the present work, we treat w^2 as a fixed external parameter. The explicit dependence of C and k_{av} on time and the four model parameters (N, s, r , and L), as well as the validity of approximations (i) and (ii) will be addressed elsewhere.

RESULTS AND DISCUSSION

As we showed earlier by means of Monte Carlo simulation (Gheorghiu, Coffin, and Rouzine 2006), the gradual accumulation and fixation of beneficial alleles present at the beginning of the evolutionary process on separate genomes and different sites consists of three phases, as follows.

Phase 1: Initial recombination events form a broad distribution of less-fit allele number.

Phase 2 (solitary wave): The less-fit allele number distribution maintains a stable, slowly changing profile moving at a slowly changing speed towards lower less-fit allele numbers.

Phase 3: The solitary wave arrives at its destination point (complete or partial fixation of beneficial alleles) and collapses into a single clone.

In the present work, we focus on the second phase, which is the longest and determines the overall speed of evolution. We address the case when the total number of recombination events per population per generation is very large, $Nr \gg 1$. The basic qualitative results are as follows.

Fitness distribution for infrequent recombination

While the distribution of the less-fit allele number shifts in time towards smaller numbers (higher fitness), the distribution profile centered at the average less-fit allele number stays nearly constant on a short time scale. We show that the number of less-fit alleles (mutations) in a genome varies across the population according to a Gaussian distribution, just as it would in the limit of infinite population size N or sufficiently large recombination rate r . There are two important effects of finite population size. Firstly, the distribution profile has a cutoff point at a low relative less-fit allele number $x_0 < 0$ that corresponds to the best-fit genomes in the population (Fig. 1a). Secondly, the less-fit allele number distribution is squeezed as compared to the single-site model result due to correlation in fitness among genomes, and due to identity by descent for some sites. The average substitution rate is also decreased to the same degree as the width of the distribution: according to Fisher Fundamental Theorem, the two quantities are uniquely related.

Fitness of most likely parents

In general, the number of mutations in a recombinant genome fluctuates around the average of the less-fit allele numbers of its parents, because positions of mutations within the two parental genomes are random (except for sites identical by descent). New recombinant genomes with different relative less-fit allele numbers are generated at different rates, as described by the “recombinant generation profile” $\rho(x)$. For a solitary wave with a Gaussian shape $\phi(x)$, the recombinant generation profile can be shown to be Gaussian as well, except the second Gaussian, $\rho(x)$, is broader (Fig. 1a, thin line). We show below that, for a given fitness of a recombinant, fitness of its parents is near a most likely value, smaller than progeny fitness (Fig. 1b). Parents with higher fitness are too few, and parents with lower fitness are unlikely to generate such highly fit progeny.

High-fitness edge

Deterministic consideration of fitness groups allows for solitary waves with all possible widths and speeds. The actual speed of the distribution is determined by a finite rate of generation of new best-fit recombinants at the edge (cutoff) of the distribution, that is proportional to the total population size N , which is large, and to the recombinant generation profile at the edge $\rho(x)$, which is very small (Fig. 1a). The overall generation rate at the edge is small. In addition, most new clones are lost due to random drift. Considering a new recombinant a minority variant, and the rest of the population the majority variant, we use the standard two-allele model to calculate the probability that a new recombinant survives to become established in the population. The condition that the solitary wave speed is equal to the speed of its edge extension yields a relationship between the length of the high-fitness tail of the distribution, $|x_0|$ (Fig. 1) and the substitution rate. Below, we show that when the total number of recombination events in a population Nr is large, the tail is always much longer than the standard deviation of the less-fit allele number.

Clone structure of fitness classes

A group of genomes with given fitness is not a mixture of random sequences. In fact, it consists of clones, subgroups of identical sequences that have different size and were born at different times in the past as a result of recombination. After a recombinant clone is born somewhere in the high-fitness tail of distribution, it grows due to selection and, at the same time, decays due

to recombination with other clones. If the recombination rate r is sufficiently small, the case we consider in the first subsections below, decay due to recombination can be neglected.

Clones within the same fitness class have different size and number, as determined by the time elapsed since their birth. Clones born earlier are much larger (because of the exponential growth in time) but also much less numerous than clones born later. Indeed, early clones have unusually high relative fitness (relative to the average genome) at the moment of birth and, hence, their generation rate is small (left tail of $\rho(x)$ in Fig. 1a). The relative contribution of clones born at earlier and later times into a fitness class depends on parameter r . In the subsequent sections, we show that, at sufficiently small r , each fitness group is dominated by clones that were nearly the best-fit at the moment of their birth, and that this fact is responsible for the Gaussian form of the less-fit allele number distribution.

Life cycle of a clone

The life cycle of a representative clone is shown in Fig. 1b in terms of the dynamics of its relative less-fit allele number. Due to the motion of solitary wave (Fig. 1a), a clone born near the edge decreases its relative fitness with respect to the population average, i.e., it moves towards the distribution center. Near a most likely age or parenthood, recombination with a genome from another clone gives birth to a new edge clone that starts a new cycle. Later, the old clone passes through the fitness distribution center, starts to decline, and eventually becomes extinct (Fig. 1b). The condition that each clone, on average, leaves one progeny clone represents the stationary condition for the solitary wave, and yields a second equation relating the substitution rate to the fitness tail length $|x_0|$. From the two relationships, we obtain the average substitution rate in terms of the four model parameters. The result demonstrates a logarithmic increase with N or r and a gradual transition to the limit of unlinked sites.

Ancestor fitness distribution

Different sites of a genome descend from different ancestral genomes in each past generation, but a particular site of a haploid genome has a single ancestor in each generation in the past. Let us choose a site on a genome and consider its distant ancestors.

The life cycle of a representative clone (Fig. 1b) implies a certain type of time dependence of the less-fit allele number of an ancestor (Fig. 2a). The trajectory consists of approximately equal vertical segments corresponding to asexual reproduction (growth of the current ancestor clone), and horizontal segments corresponding to rare recombination events starting new clones. The actual birth point and the actual parenthood point fluctuate around their most likely values, so that strict periodicity is absent. Sufficiently far back in time, one can find the ancestor, with equal likelihood, anywhere in the interval between the edge and the most likely parenthood point (Fig. 2b). Thus, the ancestor distribution in fitness is broad and occupies a good portion of the high-fitness tail of the ancient population distribution. We conclude that only a tiny fraction of individuals well above average fitness perpetuate themselves in posterity.

Fitness distribution and ancestor fitness distribution at intermediate recombination rates

The above argument is essentially restricted to the case of sufficiently small r , so that the growth of clones is not impeded by recombination with other clones. It is also interesting to investigate the case of higher r (but still much smaller than 1, so that reproduction remains mostly asexual). Now, clones decay strongly due to recombination, and, hence, the clones born later, i.e., far from the high-fitness edge provide the main contribution to each fitness group.

Because the most likely ratio between the relative mutation of a child and its parents is preserved in this case, the entire fitness trajectory of ancestors shifts toward the average fitness and becomes less regular (dotted line in Fig. 2a). Although the mathematics of this case is more

complex, fortunately, the correction $h(u)$ to the logarithm of solitary wave profile, as compared to the limit of infinite population size, is small, and depends on a single external parameter, β , proportional to r (Fig. 3a and the legend). Using this fact, we obtained the fitness distribution for remote ancestors at different values of β (Fig. 4). As should be expected from the shifted trajectory of ancestor fitness (Fig. 2a), an increase in the recombination rate leads to larger relative less-fit allele numbers x in the likely ancestors. Nevertheless, over a broad range of r , the ancestor fitness distribution remains localized in the high-fitness tail of an ancient population. Therefore, our conclusion about unusual fitness of ancestors holds in this case as well.

Linkage, linkage disequilibrium and inter-genome correlations: effect on adaptation rate

At finite population size, the substitution rate and, accordingly, the width of fitness distribution, in the present model, are decreased below the single-site model prediction. The decrease represents an effect of linkage, i.e., the fact that many genomic sites are inherited, apart from recombination, all together. When either the recombination rate of the population size increase, linkage is gradually compensated, and all observables approach their single-locus predictions (Rouzine and Coffin 2005; Gheorghiu, Coffin, and Rouzine 2006).

Based on results obtained for two-site and three-site models, linkage disequilibrium, defined as correlation between probabilities of having specific allele types at different sites of a genome, is often considered a measure of the strength of linkage. In the literature, the terms “linkage” and “linkage disequilibrium” are sometimes used interchangeably. In contrast, for the present multi-site model, direct Monte-Carlo simulation shows very small linkage disequilibrium (smaller than the statistical error) even when the linkage effect on the substitution rate is strong (Gheorghiu, Coffin, and Rouzine 2006). This is the reason why we neglect linkage disequilibrium in the present (and previous) work, assuming that alleles at different sites do not correlate (section *Basic Approach*).

At the same time, as the population size becomes smaller, simulation demonstrates increasingly strong correlation *between* different genomes at homologous sites concurrent with a decrease of the substitution rate below the single-site model result (Gheorghiu, Coffin, and Rouzine 2006). The decrease is shown to occur at much larger population sizes than predicted by the approximation neglecting inter-genomic correlations (Rouzine and Coffin 2005). Furthermore, simulation results (Gheorghiu, Coffin, and Rouzine 2006) are consistent with the approximation, in which a pair of homologous sites either do not correlate, or correlate 100% (carry identical alleles with probability 1).

In the present work, we take into account inter-genomic correlations by introducing the fraction of fully correlated homologous sites in an average pair of genomes, C (section *Basic Approach*). Parameter C enters the variance of progeny fitness that determines the width of fitness distribution and, consequently, the composite parameter β controlling the substitution rate, the ancestor fitness distribution and other observables. Intuitively, inter-genomic correlations decrease fitness of better-fit progeny, causing the fitness distribution to shrink and, according to Fisher Fundamental Theorem, adaptation to slow down.

Thus, we determine the observables (substitution rate, the clone structure of fitness groups, and the ancestor fitness distribution) for a fixed value of parameter C . In fact, inter-genomic correlations increase (slowly) in time (Gheorghiu, Coffin, and Rouzine 2006). To obtain closed expressions for the observables, we must be able to derive C as a function of time and model parameters (N , s , r , k_{av} and L). We complete the task in our most recent work (Rouzine and Coffin 2006), where we demonstrate that identity by descent, i.e., the fact that some homologous sites have common ancestors, causes strong correlations of the described kind. Based on that idea, we derive the average coalescent time, parameter C , the substitution rate,

and the final number of less-fit alleles at which evolution stops due to the partial loss of beneficial alleles. The results agree with the simulation (Gheorghiu, Coffin, and Rouzine 2006), supporting the central role for identity by descent.

DERIVATION

Fitness distribution for infrequent recombination

When the recombination rate r per genome is sufficiently small (the exact condition will be given below), but the total number of recombination events per population Nr is still large, Eqs. 4 and 5 have a Gaussian solution with a cutoff at negative $x=x_0$ (Rouzine and Coffin 2005) (Fig. 1a)

$$\phi(x) = \frac{1}{\sqrt{2\pi}pw} e^{-\frac{x^2}{2pw^2}}, \quad x - x_0 \gg w^2/[x_0(1-p)], \quad (6)$$

where the parameter $p < 1$ is defined by

$$V = ps w^2 = s \text{std}_k^2. \quad (7)$$

where std_k is the standard deviation of k in a population. In Eq. 6, we assume strong inequalities $1-p \gg w^2/x_0^2$, $x_0 \gg w$. The values of p and $|x_0|$ and the validity of these inequalities are addressed below. The parameter p reflects the correlation in the total number of deleterious alleles k between genomes, just as the parameter C in Eq. 3 reflects the correlation on the individual site level. Both types of correlation exist due to the combined effect of linkage and finite population size. At $p=1$ and $C=0$, from Eqs. 3 and 6 we obtain the distribution of less-fit allele number characteristic for the deterministic limit, where sites evolve independently of each other, and $V = sk_{av}(1-k_{av}/L)$.

Note that the relation between V and std_k in Eq. 7 (Fisher's Fundamental Theorem) follows directly from the initial kinetic equation and holds even if both standard deviation and the evolution rate are less than their respective deterministic values. Direct Monte-Carlo studies of the present model (Gheorghiu, Coffin, and Rouzine 2006) confirm the relationship with high accuracy, thus demonstrating the validity of the semi-deterministic approximation. Parameter w^2 , defined in Eq. 3, has the meaning of the average pairwise genetic distance. The proportionality between V and w^2 , Eq. 7, as both vary in time, was also confirmed by simulation (Gheorghiu, Coffin, and Rouzine 2006).

Fitness of most likely parents

Substituting Eq. 6 into Eq. 5, we observe, that the integrand has a narrow maximum at $x_1=x_2=xp/(p+1)$. Integrating the Gaussian in x_1 and x_2 , we get the recombinant generation profile (Fig. 1a)

$$\rho(x) = \frac{1}{\sqrt{\pi(1+p)}w} e^{-\frac{x^2}{(1+p)w^2}} \quad (8)$$

Thus, for a recombinant genome with the relative less-fit allele number x , the most likely relative less-fit allele number of both parents is $-xp/(p+1)$ (Fig. 1b).

High-fitness edge

The high-fitness edge of the fitness distribution gradually extends towards smaller k (higher fitness) due to generation of new best-fit recombinants (Fig. 1b). To keep the wave profile constant, the speed of the edge extension must match the speed of the entire wave V , as given by

$$[Nr\rho(x_0)\Delta x](s|x_0|\Delta x) \sim V, \quad \Delta x \equiv (d \ln \rho / dx)_{x_0}^{-1} \quad (9)$$

The three terms in the left-hand side of Eq. 9 are, respectively, the generation rate within a typical interval Δx , the survival probability of a new recombinant in the presence of random drift, and the typical distance between previous and new best-fit variant. From Eqs. 8 and 9, we obtain a first relation between p and x_0 (Rouzine and Coffin 2005)

$$x_0^2 = (1+p)w^2\Lambda_1, \quad \Lambda_1 \equiv \ln \frac{Nr(1+p)}{4p\sqrt{\pi\Lambda_1}} \gg 1 \quad (10)$$

Thus, when the total number of recombination events per population is large, $Nr \gg 1$, the distribution of the less-fit allele number has a relatively long high-fitness tail, $|x_0| \gg w$, just as we assumed above.

Clone structure of fitness classes

As we discussed, each class of genomes with given fitness consists of clones born at different times in the past. Each clone starts as a single recombinant lucky to survive loss by random drift. It is convenient to classify clones by their relative less-fit allele number at the moment of birth, x' (Fig. 1b). We denote the number of survivor clones born with relative less-fit allele numbers within the interval $[x', x'+dx]$, as $m(x')dx'$, where

$$m(x') \equiv [Nr\rho(x')](s|x'|)(1/V), \quad x' < 0 \quad (11)$$

Here the first term is the birth rate of clones per generation, the second term is the survival probability for a clone, and factor $1/V$ relates units of time and x' .

A new clone born at $x' < 0$ has a fitness advantage $\exp(s|x'|)$ with respect to an average genome. The clone is established in the population when and if its size exceeds $\sim 1/(s|x'|)$ and selection starts to dominate over random drift. Then, as the distribution moves towards it, the clone is deterministically amplified by selection (Fig. 1b) and, at the same time, decays due to recombination with genomes from other clones. The size $n(x', x)$ of a clone born at x' and sampled later, when it is located at $x > x'$, is given by

$$n(x', x) \equiv \frac{1}{s|x'|} e^{\frac{s(x'^2 - x^2)}{2V} + \frac{r(x' - x)}{V}} = \frac{1}{s|x'|} e^{\frac{1}{pw^2} \left[\frac{x'^2 - x^2}{2} + \frac{r(x' - x)}{s} \right]} \quad (12)$$

which follows from Eq. 4 in which we dropped the term $r\rho(x)$ responsible for generation of new recombinants, and the initial condition $n(x', x') = 1/(s|x'|)$. The total genome frequency $\phi(x)$ can be written as an integral over clones born at different x' , as given by

$$\phi(x) = \frac{1}{N} \int_{x_0}^x dx' m(x') n(x', x) \quad (13)$$

As one can easily check, $\varphi(x)$ given by Eq. 13 satisfies Eq. 4. We observe that terms $m(x')$ and $n(x',x)$ in the integrand of Eq. 13 both depend exponentially on x' , but in the opposite manner: at negative x' , $m(x')$ is increasing, and $n(x',x)$ is decreasing with x' . At $1-p \gg w^2/x_0^2$, the second factor wins, and the integrand in Eq. 13 has a sharp maximum near the lower limit x_0 . (As we shall see, the inequality is valid when r is not too large; in the same case, the term with r in Eq. 12 describing decay of clones due to recombination can be neglected.) We conclude that all fitness groups comprise, mostly, clones born near the high-fitness edge of the distribution (Fig. 1b). Using this fact, for x not too close to the edge x_0 , we arrive at a non-normalized version of Eq. 6. Using the normalization condition $\int dx \varphi(x) = 1$, we obtain a second relationship between the cutoff point, x_0 , and parameter p

$$x_0^2 = w^2 \frac{2p(1+p)}{1-p} \Lambda_2, \quad \Lambda_2 \equiv \ln \left(\frac{sw^2}{r|x_0|} \Lambda_2 \sqrt{2p(1+p)} \right), \quad (14)$$

The validity condition of Eq. 14 has the form $r|x_0| \ll sw^2$, or $1-p \gg w^2/x_0^2$, which implies that most genomes of a clone never experience recombination and justifies neglecting the recombination term in Eq. 12 above.

From Eqs. 14 and 10, we obtain the desired value of parameter p that determines the substitution rate V , as given by

$$p = \Lambda_1 / (\Lambda_1 + 2\Lambda_2), \quad 1/N \ll r \ll sw^2 / |x_0| \quad (15)$$

where Λ_1 is defined in Eq. 10. As the population size N increases, parameter p approaches the deterministic limit, $p=1$.

Life cycle of a clone

The history of a representative clone dominating a fitness class is shown in Fig. 1b. A clone is born and established near the edge, $x \approx x_0$. As the overall less-fit allele number distribution moves towards smaller k , the value of x decreases, and the clone grows in size. After the clone reaches the point $x=0$ (average fitness), it starts to contract due to negative selection until, eventually, it becomes extinct. When the clone is at most likely parenthood point $x \approx x_0 p / (1+p)$ (see above), one of its genomes recombines with a genome of similar fitness to generate an offspring near the current high-fitness edge of the distribution. The progeny genome starts a new edge clone, and the cycle repeats.

Therefore, if we choose a site in a genome and trace the less-fit allele number k of its ancestors back in time, we obtain a broken line (Fig. 2a) comprising vertical lines (asexual reproduction) and jumps by $\Delta k \approx |x_0|/(1+p)$ due to recombination events separated by time intervals $\Delta t \approx |x_0|/[V(1+p)]$. Further, because a birth point of an edge clone and the point of parenthood fluctuate, long-range periodicity is absent. Therefore, sufficiently far back in time, an ancestor can be found, with equal probability, anywhere in the interval $x_0 < x < x_0 p / (1+p)$ (Fig. 2b). Thus, only genomes whose fitness has not yet reached the most likely parenthood point leave descendants for the future.

Alternatively, speaking in terms of generations of clones rather than generations of individual genomes, $(x-x_0)/V$ represents the age of a clone since its birth by recombination, and $x_0/[V(1+p)]$ is the most likely age of reproduction. Only sufficiently young clones leave descendants.

Ancestor fitness distribution

Let us choose a site on a genome and address the fitness of its distant ancestors (we remind that each site has only one ancestor in each generation). Based on the results of the previous

subsection, the less-fit allele number distribution of the ancestors of the present population is very different from the distribution of an entire ancient population at the same time. While the population distribution is localized symmetrically around average k_{av} in the region $\sim w$, the ancestor distribution is much more broad [has width $\sim w \ln^{1/2}(Nr)$] and occupies an entire upper half of the high-fitness tail of the population distribution. We conclude that, however mediocre an individual genome may be, its ancestors were all exceptional within their populations.

Fitness distribution at intermediate recombination rates

The previous consideration was essentially restricted to the case of infrequent recombination, when the parameter

$$\beta \equiv r |x_0| / V = r |x_0| / (psw^2) \quad (16)$$

where x_0 and p are given by Eqs. 10 and 15, respectively, is much smaller than 1. At larger recombination rates, such that $\beta \sim 1$ or larger, we have $1-p$ on the order of or smaller than w^2/x_0^2 , so that the integrand in Eq. 13 no longer peaks at x_0 . Thus, clones born far from the edge now give important contributions to fitness groups, possibly affecting our conclusions regarding the ancestor history. For this case, we have to use an alternative formalism.

We observe that the difference between $\ln \phi(x)$ and its deterministic limit is on the order of $1-p \sim 1/\ln(Nr)$, Eq. 8. Therefore, we can calculate the difference as a first-order correction in $1/\ln(Nr) \ll 1$. We seek $\phi(x)$ in the form

$$\phi(x) = \frac{1}{\sqrt{2\pi w}} e^{-\frac{x^2}{2w^2} - \epsilon h(u)} \quad (17)$$

$$\epsilon \equiv (x_0^2 / w^2)(1-p) \approx 2\Lambda_1(1-p), \quad u \equiv x / |x_0| \quad (18)$$

Here we rescaled $1-p$ and x to make the rescaled values on the order of 1 for $\beta \sim 1$ and $x \sim |x_0|$. The recombinant generation rate, Eq. 5, takes a form

$$\rho(x) = \frac{1}{\sqrt{2\pi w}} e^{-\frac{x^2}{2w^2} - 2\epsilon h(u/2)} \quad (19)$$

An equation for $h(u)$ that follows from Eqs. 4, 17, and 19

$$\epsilon h(u) = \frac{\epsilon u^2}{2} + \beta \left[u - \int_0^u du' e^{\epsilon h(u') - 2\epsilon h(u'/2)} \right] \quad (20)$$

has β as the only external parameter. The value of ϵ is determined by the condition $\phi(x_0)=0$, i.e., $h(u)$ diverges at $u=-1$. The numeric solution of Eq. 20 with respect to $h(u)$ and ϵ is shown in Fig. 3a for different values of β .

At small β , in agreement with the previous result (Eqs. 6 and 15), we obtain that the function $h(u)$ is quadratic in u except in the vicinity of the edge $u \approx -1$, and

$$\epsilon \approx 4 \ln[(2/\beta) \ln(2/\beta)] \approx 4\Lambda_2, \quad \beta \ll 1$$

Hence, $1-p \approx 2\Lambda_2 / \Lambda_1$ at small β , in agreement with what Eq. 15 predicts at small $1-p$. Two asymptotic expressions for p in two overlapping intervals of r , Eqs. 15 and 18, can be combined in one approximate interpolation formula

$$p = \frac{\Lambda_1}{\Lambda_1 + \epsilon(\beta)/2} \quad (21)$$

with $\epsilon(\beta)$ shown in Fig. 3a. Eq. 21 corrects our previous result for p obtained by extrapolation of Eq. 15 to small $1-p$ (Rouzine and Coffin 2005). Several examples of dependence of p on N for the virus recombination model ($r = N/N_0$) are shown in Fig. 3b. If inter-genomic correlations were absent ($C=0$), $p(N)$ would represent the ratio of the substitution rate to its single-site limit. As we show elsewhere (Rouzine and Coffin 2005; Gheorghiu, Coffin, and Rouzine 2006), C cannot be neglected and increases faster than $1-p$ at small N , thus, dominating the dependence of the substitution rate on the population size. Factor p represents a component of the dependence arising from correlations between fitness of different genomes.

Ancestor fitness at intermediate recombination rates

As we show in this subsection, the ancestor fitness distribution $\varphi(x) \equiv (1/x_0) y(u)$ can be found from the equation

$$y(u) = \beta \int_{\max(2u, -1)}^u du' e^{\epsilon[h(u') - 2h(u'/2)]} y(u'), \quad u < 0 \quad (22)$$

where β is the only external parameter, and $h(u)$ and ϵ are found from Eq. 20. The numeric solution of Eq. 22 (Fig. 4) suggests, in the limit $\beta \rightarrow 0$, a uniform ancestor fitness distribution within the interval $-1 < u < -1/2$, just as we obtained above in that limit (Fig. 2b). At $\beta \sim 1$, the distribution shifts toward low fitness and, at $\beta \gg 1$, assumes a universal asymmetric shape with a scale $|u| \sim 1/\beta$. Thus, unless β is too large, our conclusion that all ancestors are exceptionally well fit remains valid.

Each genomic site has a single ancestor genome in each generation in the past. Consider a site in a genome at time t_{now} and its ancestral genome at time $t < t_{\text{now}}$. The probability density $\varphi(x, t)$ of the relative less-fit allele number of an ancestor $x = k - k_{\text{av}}(t)$ obeys the Markov equation

$$\varphi(x, t) = \int_{x_0}^{\infty} dx' P(x | x') \varphi(x', t+1), \quad x > x_0 \quad (23)$$

where $P(x|x')$ is the probability density of the relative less-fit allele number of a parent x , given the number x' of a progeny genome. Eq. 23 is solved backward in time with the “initial” (actually, final) condition $\varphi(x, t_{\text{now}}) = \delta[x - x(t_{\text{now}})]$. The kernel $P(x|x')$ can be written in the form

$$P(x | x') = A(x') [\phi(x' - V) \delta(x' - x - V) + r \rho(x' - V) P_{\text{sex}}(x | x')] \quad (24)$$

where the first and the second terms in brackets correspond to asexual and sexual reproduction, respectively. $\phi(x)$ and $\rho(x)$ are given by Eqs. 17 and 19; $P_{\text{sex}}(x|x')$ is the normalized probability density of the parental relative less-fit allele number x , given that reproduction is due to recombination, and the progeny genome has the relative less-fit allele number x' ; term $-V$ reflects the shift of the population fitness distribution between adjacent generations. Using the condition that $\varphi(x)$ is zero at $x < x_0$, the normalization factor $A(x)$ can be written as

$$A(x') = \begin{cases} 1/[\phi(x' - V) + r\rho(x' - V)], & x' > x_0 + V \\ 1/[r\rho(x' - V)], & x_0 < x' < x_0 + V \end{cases} \quad (25)$$

Markovian form of Eq. 23 deserves a comment. Hermisson et al (Hermisson et al. 2002) studied an ancestor distribution in asexual populations using a rather general formalism in which a population state was described using the vector of all possible genotypes. These authors obtained a general expression for the transition matrix of the reverse evolution of the ancestor distribution. Noting that the matrix depends on a population state, so that the reverse process is, generally, not Markovian, Hermisson et al focused on the equilibrium state of population.

Here, to address a non-equilibrium case (accumulation of beneficial mutations in partly sexual populations), we employ a reduced description of the population state in terms of the distribution of the less-fit allele number. As the present manuscript and the work quoted in Introduction demonstrate, in a broad time interval, the distribution has a quasi-stable profile moving at a slowly changing speed. By re-centering the distribution at the time-dependent average number of less-fit alleles, we obtain a reverse process with an (almost) *constant* transition matrix $P(x, x')$, Eq. 23. The matrix depends on the constant fitness distribution profile and the constant recombinant generation profile (Eqs. 24 and 25), both derived in the previous subsections.

As we discussed above (Fig. 1b), at $p \approx 1$, $P_{\text{sex}}(x_1|x)$ has a maximum at the most likely parenthood point, $x \approx x'/2$, with the characteristic width on the order of w . The characteristic scale of the ancestor distribution $\varphi(x, t)$ is $|x_0| \gg w$ (Fig. 2b). Therefore, in Eq. 23 we can approximate $P_{\text{sex}}(x_1|x)$ with delta-function, as given by

$$P_{\text{sex}}(x | x') \approx \delta(x - x' / 2) \quad (26)$$

Substituting Eqs. 25 and 26 into Eq. 24, we obtain

$$P(x | x') = [1 - \tilde{r}(x')] \delta(x' - x - V) + \left\{ \tilde{r}(x') + [1 - \tilde{r}(x')] \theta(x_0 + V - x') \right\} \delta(x - x' / 2) \quad (27)$$

$$\tilde{r}(x') \equiv \frac{r\rho(x' - V)}{\phi(x' - V) + r\rho(x' - V)} \quad (28)$$

Due to the inequality $V/|x_0| \sim s|x_0|/\Lambda_1 \ll 1$, we can formally expand the first delta-function and the theta-function in Eq. 27 in small V , which yields

$$\begin{aligned} P(x | x') &= \delta(x' - x) + V[-\delta(x' - x) + \delta(x_0 - x') \delta(x - x' / 2)] \\ &+ \tilde{r}(x') [\delta(x - x' / 2) - \delta(x' - x)] \end{aligned} \quad (29)$$

Substituting Eq. 29 into Eq. 23, approximating $\varphi(x, t+1) - \varphi(x, t) \approx \partial \varphi / \partial t$, and integrating in x , we obtain

$$\begin{aligned} -\frac{\partial \varphi(x, t)}{\partial t} &= V \frac{\partial \varphi(x, t)}{\partial x} + V \varphi(x_0 + 0, t) [\delta(x - x_0 / 2) - \delta(x - x_0)] \\ &+ 2 \tilde{r}(2x) \varphi(2x, t) \theta(x - x_0 / 2) - \tilde{r}(x) \varphi(x, t) \end{aligned} \quad (30)$$

To decrease the number of parameters, we rescale the variables t , x , and φ , as given by

$$\tau \equiv tr / \beta, \quad u \equiv x / |x_0|, \quad y(u, \tau) \equiv |x_0| \varphi(x, t) \quad (31)$$

In the new notation, Eq. 30 takes a form

$$-\partial y(u, t) / \partial \tau = \partial y(u, t) / \partial u + y(-1 + 0, \tau) [\delta(u + 1/2) - \delta(u + 1)] \\ + \beta [2\eta(2u)y(2u, t)\theta(u + 1/2) - \eta(u)y(u, t)], \quad (32)$$

$$\eta(u) \equiv \exp \{ \epsilon [h(u) - 2h(u/2)] \} \quad (33)$$

where we used Eqs. 17, 19 and 28 and inequality $r \ll 1$.

The characteristic time scale in Eq. 32 for is $\tau \sim 1$, or $t \sim \beta/r$. If we go farther back in time, the ancestor fitness distribution takes a stationary form that satisfies equation

$$\frac{dy(u)}{du} = y(-1 + 0) [\delta(u + 1) - \delta(u + 1/2)] \\ + \beta [\eta(u)y(u) - 2\eta(2u)y(2u)\theta(u + 1/2)] \quad (34)$$

Note that $\eta(u)$ diverges at $u \rightarrow -1$ due to divergence of $h(u)$, Eq. 33. To keep derivative dy/du finite, $y(u)$ has to vanish at $u \rightarrow -1$. Therefore, in Eq. 34, the term with delta-functions is zero. Integrating Eq. 34 in u , we arrive at Eq. 22 of the previous subsection.

We conclude that the history of clones dominating fitness classes and the rescaled relative fitness distribution of distant ancestors of sites within ancestral populations is controlled by a single parameter β proportional to the recombination rate and inversely proportional to the selection coefficient. The average substitution rate, the high-fitness tail length, and the width of the fitness distribution of the entire population are expressed in terms of β and of the within-population genetic distance. The genetic distance is decreased by the fraction of sites identical by descent that we derive elsewhere (Rouzine and Coffin 2006).

Acknowledgements

We are thankful to John Wakeley and Claus Wilke for useful comments. We also thank Allen Rodrigo for stimulating comments on the first part of this project. The work was supported by National Institutes of Health grants K25AI01811 to I.M.R. and R01 CA 089441 to J.M.C. J.M.C. was an American Cancer Society Research Professor with support from the George Kirby Foundation.

References

- Barton NH. A general model for the evolution of recombination. *Genet Res, Camb* 1995;65:123–144.
- Barton NH, Shpak M. The stability of symmetric solutions to polygenic models. *Theor Popul Biol* 2000;57:249–263. [PubMed: 10828217]
- Charlesworth B. Mutation-selection balance and the evolutionary advantage of sex and recombination. *Genet Res, Camb* 1990;55:199–221.
- Charlesworth B, Charlesworth D. Rapid fixation of deleterious alleles can be caused by Muller's ratchet. *Genet Res* 1997;70:63–73. [PubMed: 9369098]
- Cohen E, Kessler D, Levine H. Recombination dramatically speeds up evolution of finite populations. *Phys Rev Lett* 2005;94(Art No 098102)
- Felsenstein J. The evolutionary advantage of recombination [review]. *Genetics* 1974;78:737–756. [PubMed: 4448362]
- Fisher, RA. The genetical theory of natural selection. Clarendon Press; Oxford, United Kingdom: 1930. 1958.
- Gerrish PJ, Lenski RE. The fate of competing beneficial mutations in an asexual population. *Genetica* 1998;102/103:127–144. [PubMed: 9720276]

- Gheorghiu, S.; Coffin, JM.; Rouzine, IM. Increasing sequence correlation limits the efficiency of recombination in a multi-site evolution model. 2006. submitted for publication
- Gordo I, Charlesworth B. The degeneration of asexual haploid populations and the speed of Muller's ratchet. *Genetics* 2000;154:1379–1387. [PubMed: 10757777]
- Haigh J. The accumulation of deleterious genes in a population - Muller's ratchet. *Theor Popul Biol* 1978;14:251–267. [PubMed: 746491]
- Hermisson J, Redner O, Wagner H, Baake E. Mutation-selection balance: ancestry, load, and maximum principle. *Theor Popul Biol* 2002;62:9–46. [PubMed: 12056863]
- Hey J. Selfish genes, pleiotropy and the origin of recombination. *Genetics* 1998;149:2089–2097. [PubMed: 9691060]
- Hill WG, Robertson A. The effect of linkage on limits to artificial selection. *Genet Res* 1966;8:269–294. [PubMed: 5980116]
- Kimura M, Maruyama T. Mutational load with epistatic gene interactions in fitness. *Genetics* 1966;54:1337–1351. [PubMed: 17248359]
- Kondrashov AS. Classification of hypotheses on the advantage of amphimixis. *J Hered* 1993;84:372–387. [PubMed: 8409359]
- Maynard Smith JM. What use is sex? *J Theor Biol* 1971;30:319–335. [PubMed: 5548029]
- Muller HJ. Some genetic aspects of sex. *Am Nat* 1932;66:118–128.
- Orr HA. The rate of adaptation in asexuals. *Genetics* 2000;155:961–968. [PubMed: 10835413]
- Otto S, Barton N. The evolution of recombination: removing the limits to natural selection. *Genetics* 1997;147:879–906. [PubMed: 9335621]
- Pamilo P, Nei M, Li WH. Accumulation of mutations in sexual and asexual populations. *Genet Res, Camb* 1987;49:135–146.
- Rouzine I, Wakeley J, Coffin J. The solitary wave of asexual evolution. *Proc Natl Acad Sci U S A* 2003;100:587–592. [PubMed: 12525686]
- Rouzine IM, Coffin JM. Evolution of human immunodeficiency virus under selection and weak recombination. *Genetics* 2005;170:7–18. [PubMed: 15744057]
- Rouzine, IM.; Coffin, JM. Coalescent time and the substitution rate in a multi-site model of haploid populations in the presence of selection and recombination. 2006. submitted for publication
- Shpak M, Kondrashov AS. Applicability of the hypergeometric phenotypic model to haploid and diploid populations. *Evolution* 1999;53:600–604.
- Stephan W, Chao L, Smale JG. The advance of Muller's ratchet in a haploid asexual population: approximate solutions based on diffusion theory. *Genet Res* 1993;61:225–231. [PubMed: 8365659]
- Tsimring LS, Levine H, Kessler D. RNA virus evolution via a fitness-space model. *Phys Rev Lett* 1996;76:4440–4443. [PubMed: 10061290]
- Wilke CO. The speed of adaptation in large asexual populations. *Genetics* 2004;167:2045–2053. [PubMed: 15342539]

Fig 1. Fitness distribution (“solitary wave”) and history of clones

a) Thick line: the average frequency of genomes with k less-fit alleles. Thin line: normalized generation rate of recombinants. Parameters V , $|x_0|$, and $wp^{1/2}$ are the evolution rate, the high-fitness tail length, and the standard deviation of k , respectively. Gray arrow: motion of the distribution in time. b) Solid line: solitary wave profile $\phi(x)$ in logarithmic scale versus the relative less-fit allele number x . Vertical bars: life cycle of a representative clone at different times in the case of small recombination frequencies, $r \ll sw^2/|x_0|$.

Fig 2. Fitness history and fitness distribution of ancestors

a) Thick line: representative trajectory of the less-fit allele number of an ancestor of a site in a genome at low recombination rates, $r \ll sw^2/|x_0|$. Dotted line: the trajectory at higher $r \gg sw^2/|x_0|$. Thin lines: trajectories of the high-fitness edge, the most likely parenthood point, and the average less-fit allele number, respectively. b) Probability distributions of the relative less-fit allele number for an ancestor and for its contemporary population.

Fig 3. Correction to fitness distribution profile and factor p

a) Correction to the fitness distribution profile due to finite population size at intermediate recombination rates. Solid lines: normalized negative correction $h(u)$ to $\ln\phi(x)$ (Eq. 17) at different values of the parameter $\beta = r|x_0|/sw^2$ (shown on the curves), as a function of the normalized relative less-fit allele number u . Results are obtained numerically from Eq. 20. Inset: normalized negative correction to the evolution rate ε (Eq. 18) as a function of β .

b) Examples of the dependence of p on N for $C = 0$, for the virus model ($r = N/N_0$), based on Eq. 21. Parameters are shown.

Fig 4. Probability density of the normalized relative less-fit allele number $u = [k - k_{av}(t)]/|x_0|$ of an ancestor of a site in a genome, $y(u)$, in the limit $\ln(Nr) \rightarrow \infty$

Results are obtained numerically from Eqs. 20 and 22. The values of parameter β (Eq. 16) are shown on the curves. The green line on the right shows the fitness distribution of the entire population from which ancestors are derived, in the limit $\ln(Nr) \rightarrow \infty$.

Table 1

Parameters and variables.

Model parameters:	
N	Population size
s	Selection coefficient (fitness advantage per site)
r	Recombination rate per genome (probability of recombination)
L	Number of sites per genome
Variables:	
k	Number of less-fit alleles in a genome
k_{av}	k averaged over population
std_k	standard deviation of k in a population
$x = k - k_{av}$	Relative less-fit allele number
x_0	x for the best-fit genome (high-fitness edge)
t	Time (generation number)
$V = -dk_{av}/dt$	Substitution rate per genome
w^2	A half of pairwise genetic distance
$p = V/(sw^2)$	Parameter reflecting inter-genome correlation in fitness
C	Pairwise identity by decent for homologous sites
$f(k, t) = \varphi(x)$	Frequency of genomes with given k (or x)
$u = x/ x_0 $	Normalized relative less-fit allele number
$\varepsilon = (x_0^2/w^2)(1-p)$	
$-\varepsilon h(u)$	Correction to the logarithm of $\varphi(x)$ due to uncompensated linkage
$R(k, t) = \rho(x)$	Normalized generation rate of genomes with given k
$\varphi(x) = \gamma(u)/ x_0 $	Probability density of x for an ancestor of a site
$\beta = r x_0 /V$	Clone decay parameter
$\Lambda 1 \approx \ln(Nr)$	The logarithm of the number of recombination events per population, Eq. 10.