Clonal interference in large populations

Su-Chan Park and Joachim Krug*

Institut für Theoretische Physik, Universität zu Köln, Zülpicher Strasse 77, 50937 Köln, Germany

Edited by Richard E. Lenski, Michigan State University, East Lansing, MI, and approved September 25, 2007 (received for review June 20, 2007)

Clonal interference, the competition between lineages arising from different beneficial mutations in an asexually reproducing population, is an important factor determining the tempo and mode of microbial adaptation. The standard theory of this phenomenon neglects the occurrence of multiple mutations as well as the correlation between loss by genetic drift and clonal competition, which is questionable in large populations. Working within the Wright–Fisher model with multiplicative fitness (no epistasis), we determine the rate of adaptation asymptotically for very large population sizes and show that the standard theory fails in this regime. Our study also explains the success of the standard theory in predicting the rate of adaptation for moderately large populations. Furthermore, we show that the nature of the substitution process changes qualitatively when multiple mutations are allowed for, because several mutations can be fixed in a single fixation event. As a consequence, the index of dispersion for counts of the fixation process displays a minimum as a function of population size, whereas the origination process of fixed mutations becomes completely regular for very large populations. We find that the number of mutations fixed in a single event is geometrically distributed as in the neutral case. These conclusions are based on extensive simulations combined with analytic results for the limit of infinite population size.

microbial adaptation | substitution process | rate of adaptation | index of dispersion | Wright–Fisher model

If two different beneficial mutations happen to occur from the wild-type simultaneously in an asexually reproducing organism and both survive against genetic drift, how will the population evolve? At first, both mutations will independently struggle against the wild-type and ultimately eliminate it. Once the wild-type becomes insignificant in the population, the mutants now compete with each other for fixation and eventually the mutation that has the larger fitness will be fixed. This phenomenon has been referred to as clonal interference (CI) (1), and it has a long history in the discussion about the evolution and maintenance of sex (2–6). In a sexual population, different beneficial mutations, rather than outcompeting each other, can recombine into a single genome, which implies an advantage compared with the asexual reproduction mode.

The key parameter governing the occurrence of CI is the number of beneficial mutations per generation $N\mu$, where N is the population size and μ is the mutation probability per individual. When $N\mu$ is small, beneficial mutations arise and fix one by one and the population evolves by "periodic selection" (7, 8). In this regime, the substitution rate and the rate of adaptation (RA; see Eq. **1**) are linear functions of both N and μ. The onset of CI for $N\mu \gg 1$ implies, first, that the RA slows down (compared with the periodic selection limit), and second, that the selective advantage conferred by a mutation that does become fixed (and thus has survived the competition with other clones) is larger than that of typical beneficial mutations. There is considerable evidence for both effects from evolution experiments on viral and bacterial populations (8–12). A more subtle prediction of CI theory is a certain temporal regularity in the process of substitution events, in the sense that the index of dispersion for counts (IDC) of the corresponding time series (the ratio of the variance to the mean of the number of events up to that time) becomes much smaller than unity for large populations (13).

The first systematic statistical description of CI was developed by Gerrish and Lenski (GL) in 1998 (1) and has since been elaborated by several authors (11, 13–16). The GL theory is based on two important approximations. First, it neglects the occurrence of multiple mutations, i.e., all mutations are assumed to arise from the wild-type. This is plausible only when $N\mu$ is small, so that the time between subsequent mutation events is long compared with the time during which the mutant clone destined for fixation makes up a significant fraction of the population, but it must surely break down in the CI regime where $N\mu \gg 1$ (17). Second, the survival of a mutation against genetic drift is assumed to be independent of its success in the clonal competition process. The following argument shows that this assumption implies an overestimation of the survival probability of a superior mutation arising during the fixation process of an earlier established clone. Consider a situation where the frequency of the wild-type and that of a mutant with selection coefficient $s > 0$ are $1 - f$ and f, respectively. Suppose that a new mutation with selection coefficient $s' > s$ occurs from the wild-type background. What is the fixation probability of the new mutant, provided no further mutations are generated? According to GL, because of the independence of genetic drift and clonal interference, the survival probability is simply $\pi(s')$, the fixation probability of the beneficial mutation against the wild-type background. However, the new mutation has to compete with a background population whose average selection coefficient is $fs > 0$, and hence the fixation probability is reduced to $\pi(s' - fs)$. The difference between the two expressions can be appreciable if the selection coefficients involved are large. Hence, the waste of mutations with large selective advantage is more probable than predicted by the GL theory, which clearly reduces the RA.

Whereas the mechanism described above gives rise to a quantitative correction to the GL theory, the effect of including multiple mutations is of a more fundamental nature, because they imply a conceptual ambiguity in the very definition of substitution events. According to a rigorous study of the neutral case (18, 19), once multiple mutations are allowed, a fixation event can involve several mutations which fix simultaneously. In this situation, the processes of origination and fixation need to be distinguished, where the origination process consists of the events when a mutation destined to be fixed first appears in the population (20, 21). Because the population is always polymorphic when $N\mu$ is large, a fixation event should be understood here as a change in the genotype of the most recent common ancestor of the whole population. After such an event, the set of mutations that all individuals share has increased by one or several mutations, which are thus fixed simultaneously. In contrast to the case of periodic selection ($N\mu$ small), these fixation

Author contributions: S.-C.P. and J.K. designed research; S.-C.P. and J.K. performed research; and S.-C.P. and J.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Abbreviations: CI, clonal interference; IDC, index of dispersion for counts; GL, Gerrish–Lenski; RA, rate of adaptation.

*To whom correspondence should be addressed. E-mail: krug@thp.uni-koeln.de.

This article contains supporting information online at www.pnas.org/cgi/content/full/0705778104/DC1.

events are not necessarily accompanied by selective sweeps. Rather, the genetic variability of the population is essentially stationary in time.

In the work of Gerrish (13) on the timing of substitution events multiple mutations were not allowed for. This guarantees that a single mutation is fixed in each substitution, and that the population is monomorphic immediately after the event. As a consequence, the sequence of substitution events can be treated as a renewal process (22), because the population structure is reset to its initial state after each fixation. In the presence of multiple mutations, this simplification is not possible, and the statistical properties of the origination and fixation processes turn out to be markedly different (see below).

The purpose of this article is to examine the consequences for the dynamics of asexual adaptation when the two key approximations of the GL theory, as outlined above, are relaxed. In all other respects, we maintain the basic setting of the GL theory. In particular, we assume an unlimited supply of beneficial mutations, we assign a different randomly drawn fitness to each new mutation (infinite sites model), and we take the fitness effects of the mutations to be multiplicative (no epistasis). Below we present our results for the RA and the statistics of substitution events, which is followed by a discussion of the limitations of this work and a brief summary. A detailed description of the Wright–Fisher model of asexually reproducing populations that underlies our investigations can be found in *Methods*.

Results

Rate of Adaptation. For the Wright–Fisher model with multiplicative fitness effects at finite population size N, it can be rigorously proven that the asymptotic increase rate of the mean logarithmic fitness (referred to herein as the rate of adaptation or RA) is the sum of the change produced by mutations in one generation and an entropy-like functional of the relative fitness distribution in the population (23), that is,

$$\lim_{t\to\infty}\frac{\langle \ln \bar{w}(t)\rangle}{t} = \langle \ln(1+s)\rangle + \left\langle \sum_i \frac{1}{N}(\chi_i - 1)\ln\chi_i \right\rangle, \quad [1]$$

where $\chi_i = w_{i,t}/\bar{w}(t)$ is the relative fitness, $w_{i,t}$ is the fitness of the ith individual at time t and $\bar{w}(t)$ its population average; throughout $\langle \ldots \rangle$ means an average over independent samples. The distribution of selection coefficients s is taken to be of the form

$$g_b(s) = s_b^{-1}\exp(-s/s_b), \quad [2]$$

with mean s_b. The choice of an exponential distribution is standard in the field, and can be motivated by arguments from extreme-value theory (24, 25). In our setting, the increase by mutations is $\langle \ln(1+s)\rangle \approx \mu s_b$, which is usually negligible for our choice of parameters $s_b = 0.02$, $\mu = 10^{-6}$ (see *Methods*). The second term on the right side of Eq. **1** is related to Fisher's fundamental theorem of natural selection (2, 26). Indeed, if the χ_i values are not far away from their mean value of unity, we can approximate $\ln\chi_i \simeq \chi_i - 1$ and see that the second term in Eq. **1** reduces to the variance of the relative fitness.

In supporting information (SI) *Appendix*, we show how the Wright–Fisher model can be solved exactly in the infinite population limit along the lines of (27). In this limit the fitness increases for long times according to

$$\ln\bar{w}(t) \simeq t(\ln(s_b t) - 1) + \frac{1}{2}\ln(2\pi\mu^2 t) + \frac{1}{s_b} = O(t\ln t), \quad [3]$$

which shows that the RA $\ln\bar{w}(t)/t$ diverges logarithmically as $t\to\infty$. This is in contrast to the case of finite populations, where the relative fitness distribution χ_i becomes stationary and the right side of Eq. **1** attains a finite limit; data illustrating the convergence are presented in *SI Appendix*. Fig. 1 shows the time dependence of the mean logarithmic fitness for finite and infinite populations. For our set of parameters, the different curves begin to deviate at $t \approx 50$.

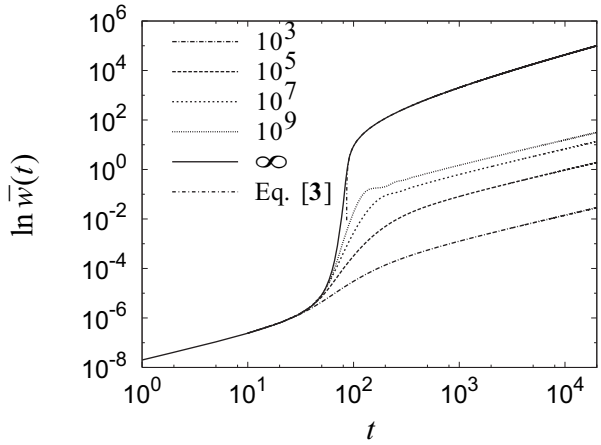

Fig. 1. Mean logarithmic fitness as a function of time for finite and infinite populations. At long times the mean logarithmic fitness increases linearly for finite populations but superlinearly in the infinite population limit. The asymptotic expression (Eq. **3**) is indistinguishable from the full solution of the infinite population equation for $t > 87$.

Within GL theory, one computes separately the expected rate of substitution and the expected selection coefficient of fixed mutations which, following the notation of (15), will be denoted by $E[k]$ and $E[s]$, respectively. Denoting by $\pi(s)$ the survival probability against genetic drift, the substitution rate reads (1)

$$E[k] = N\mu\int_0^\infty \pi(s)e^{-\lambda(s)-s/s_b}ds/s_b \quad [4]$$

with

$$\lambda(s) = \frac{\mu}{s}N\ln N\int_s^\infty \pi(u)e^{-u/s_b}du/s_b, \quad [5]$$

and the average selection coefficient of a fixed mutation is (1)

$$E[s] = \frac{\int_0^\infty s\pi(s)\exp(-\lambda(s) - s/s_b)ds}{\int_0^\infty \pi(s)\exp(-\lambda(s) - s/s_b)ds}. \quad [6]$$

From $E[k]$ and $E[s]$, the RA of the Wright–Fisher model with multiplicative landscape is predicted to be (15)

$$\frac{\langle \ln\bar{w}(t)\rangle}{t} \to E[k]\ln(1+E[s]). \quad [7]$$

Here, we compare these predictions to our simulation results. The integrals in Eqs. **4** and **6** were evaluated numerically using the expression $\pi(s) = 1 - \exp(-2s)$ (28) for the survival probability which (in contrast to the commonly used approximation $\pi(s) \approx 2s$) remains valid also for large s (16). An asymptotic analytic evaluation of the integrals using the full expression for $\pi(s)$ is given in *SI Appendix*. We simulated systems with population sizes from 10^3 to 10^9 and the number of independent samples ranged from 10^8 for $N = 10^3$ to 32,000 for $N = 10^9$.

As discussed above, in large populations the notion of a substitution event becomes ambiguous and one has to distinguish be-

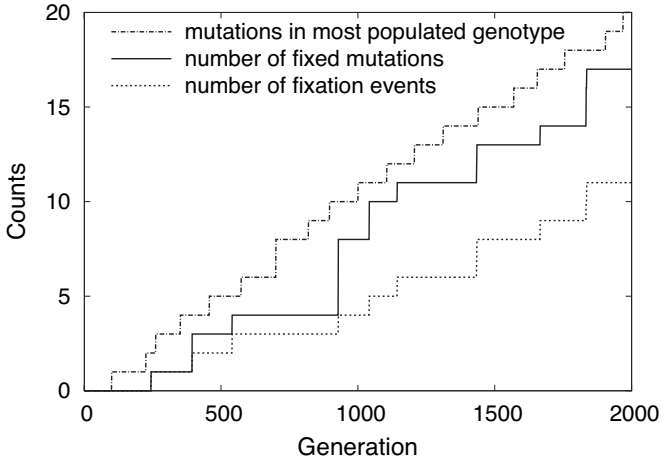

Fig. 2. An example plot showing different quantities characterizing the rate of substitution, as explained in the text. Population size is $N = 10^9$. The apparent two-step jump of the fixation curve is due to the small time gap between two consecutive fixation events.

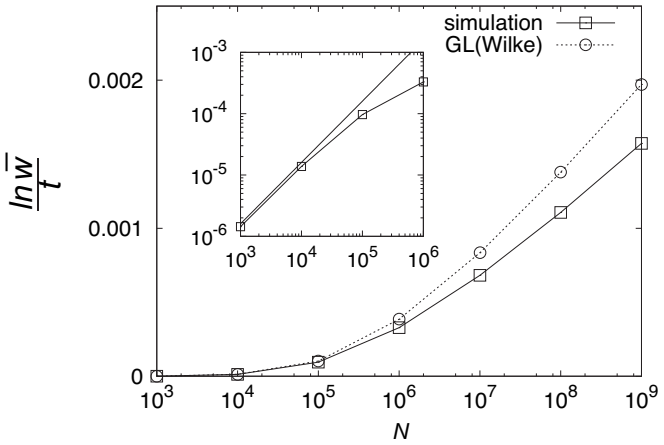

Fig. 3. Semilogarithmic plots of the RA $\ln \bar{w}/t$ vs. N from simulations (square) and GL prediction of Eq. **7** (circle) with Eqs. **4**–**6**. Significant deviations appear for population sizes exceeding $N = 10^5$. (*Inset*) Comparison of the simulation with the periodic selection result $4s_b^2 N\mu$ (straight line) on double-logarithmic scales. The discrepancy due to CI is observable from $N = 10^4$.

tween the fixation and origination processes that have in general different rates of occurrence (21). The question then arises to which of these processes the GL substitution rate $E[k]$ should be compared. Because the focus of the GL theory is on the fixation of single mutations, the origination process, which involves one mutation per event, rather than the fixation process, which can include multiple mutations, will be used here for the comparison.

To count the number of origination events up to a given time is rather cumbersome, although not difficult in principle: We have to keep track of the birth date for every mutation, and at its fixation (if it occurs), we have to increase the number of origination events at that time. Because we are interested here in the long time behavior of the origination process, we take an easier route and simply count the number of mutations fixed up to time t. Clearly, this number is smaller than the number of origination events, but the ratio of these two numbers approaches unity when the observation time is large. We also measured the number of mutations of the most populated genotype, which is the most directly available quantity in experiments. This number is clearly smaller (larger) than the number of origination events (number of fixed mutations), so again its rate of increase in the asymptotic limit is equal to the rate of substitution. Because the variance in the number of mutations in the population is bounded, the number of mutations in the most populated genotype is not much different from the mean number of mutations, which can be calculated analytically in the limit of infinite population size (see below). In Fig. 2, we present an example which shows the difference and similarity between these measures of the rate of substitution (for further details, see *SI Appendix*).

To obtain the mean selection coefficient of fixed mutations in the simulation, we wait until a fixation event occurs and calculate the selection coefficient(s) of the fixed mutation(s), comparing the fitness of each mutant to its direct precedent. The distribution of selection coefficients of fixed mutations can be found in *SI Appendix*. The RA can then be measured directly or indirectly using the relation in Eq. **7**. The two procedures give identical results (see *SI Appendix*).

In Fig. 3, simulation results for the RA are compared with the predictions of GL theory. For population sizes $N < 10^4$, the evolution proceeds by periodic selection and the RA is given by the simple expression $4s_b^2 N\mu$ (see Fig. 3 *Inset*). Clonal interference starts to reduce the RA for $N > 10^4$, and significant deviations between theory and simulation appear for $N \geq 10^6$ (note that these numbers depend on the chosen values of μ and s_b). However, even for populations as large as $N = 10^9$, the discrepancy is not dramatic, amounting to no more than 25%.

To gain further insight into the origins of the discrepancy, we show in Fig. 4 the relative errors of the RA, $E[k]$ and $E[s]$ as functions of N. On first glance these data seem to indicate that the substitution rate $E[k]$ is quite well described by GL theory, whereas the prediction for the mean selection coefficient $E[s]$ becomes increasingly inaccurate as N increases. However, the following considerations show that the agreement in $E[k]$ is fortuitous and that the GL theory eventually strongly underestimates the substitution rate in extremely large populations. As shown in *SI Appendix*, the GL expression (Eq. **4**) for $E[k]$ approaches s_b for $N \to \infty$, a result that also follows from a simple extreme value argument. According to extreme value theory (see *SI Appendix* for a brief summary), the largest selection coefficient among mutations drawn from the distribution (Eq. **2**) during one generation should be of order $s_b \ln(N\mu)$, which yields the fixation time for such a mutation to be $t_{\text{fix}} \approx \ln(N)/(s_b \ln(N\mu)) \to 1/s_b$ for large N. Thus, because multiple mutations are disallowed, the rate of substitution must approach s_b within the GL framework. In view of the exact result

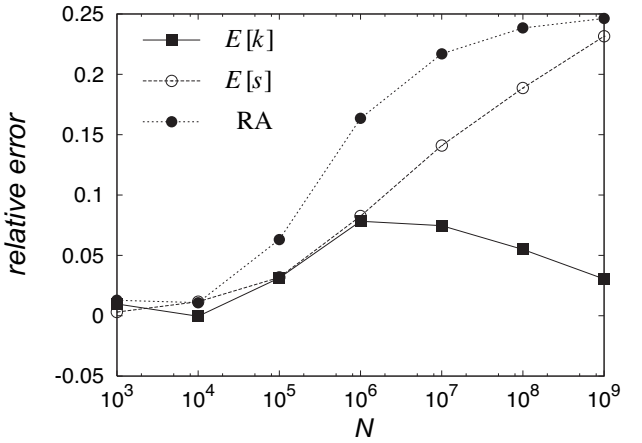

Fig. 4. Semilogarithmic plot of the relative error of the GL prediction (=(GLprediction − simulation)/simulation) vs. N for $E[k]$ (filled square), $E[s]$ (open circle), and the RA $\ln \bar{w}(t)/t$ [$= E[k] \ln(1 + E[s])$; filled circle], respectively. The prediction for the substitution rate is surprisingly good compared with that for the selection coefficient.

$$\lim_{N\to\infty} E[k] = 1 \qquad [8]$$

derived in *SI Appendix*, we thus expect that, as a consequence of multiple mutations, the relative error of $E[k]$ in Fig. 4 becomes negative and approaches the value $s_b - 1 \approx -1$ for very large N. The fact that our simulation results for $E[k]$ lie below the GL prediction shows that the effect of multiple mutations on the rate of substitutions is irrelevant in this range of population sizes.

However, despite the significant and growing discrepancy in $E[s]$ that is evident from Fig. 4, we now argue that for extremely large populations the GL estimate of this quantity will again become accurate. Recall Eq. **8**, which means that, in every generation, one mutation appears that is destined to be fixed. Most probably, the fixed mutation will be the one with the largest selection coefficient, which on average is equal to $s_b \ln(N\mu)$. This is precisely the $N \to \infty$ limit of $E[s]$ obtained by Wilke (15) within the GL theory [see *SI Appendix* for a derivation using the full expression for $\pi(s)$]. It is also consistent with the infinite population calculation presented in *SI Appendix*, which shows that the RA is determined by the upper bound of the support of the distribution of selection coefficients; in a finite population with unbounded selection coefficient, s_b, $\ln(\mu N)$ effectively plays the role of this bound. Thus the relative error of $E[s]$ in Fig. 4 is expected to reach a maximum and then decrease to 0 as N becomes large.

Two general conclusions can be drawn from these considerations. First, despite its inherent approximations, the GL theory provides a reasonable description of the rate of substitution and the RA in the experimentally relevant range of population sizes (at least up to $N = 10^9$). Second, the deviations between simulations and theory that can be observed in this regime (as depicted in Fig. 4) are a poor guide to the true asymptotic behavior as $N \to \infty$. In this sense, population sizes of order 10^9 are actually still rather small. This observation is further underscored by the fitness evolution curves in Fig. 1. It might not be fair to compare the RA of a finite population with that of the infinite population, because the latter unlike the former increases, although slowly, indefinitely. Still, it is evident that the overlap between the fitness evolution of the infinite population and that of the finite population with $N = 10^9$ in Fig. 1 is restricted to very short times.

Substitution Events. We now turn to the temporal statistics of substitution events in the CI regime. This issue was first addressed by Gerrish (13), who argued that CI renders the substitution process more regular than the random Poisson process through which the beneficial mutations arise. Based on the GL theory, he predicted that the IDC of the substitution process approaches a universal value for large populations, which is significantly smaller than the value of unity characteristic of a Poisson process. In view of the distinction between the number of fixed mutations and the number of fixation events that becomes necessary in the presence of multiple mutations, it is obviously important to clarify to which (if any) of the two processes Gerrish's argument applies.

In our simulations, we observe the mean and variance of the number of fixed mutations and of the number of fixation events to increase linearly in time. It is then convenient to introduce the increase rates of the mean and variance of the origination (fixation) process, which will be denoted by E_O (E_F) and V_O (V_F), respectively. Hence, the average number of fixation events until generation t is $\approx E_F t$ and so on. The IDC of each process is $I_X = V_X/E_X$, where X stands for either F or O. In the previous section, E_O was used for comparison with $E[k]$ predicted from the GL theory.

To understand how the fixation and origination processes are related, we introduce the probability distribution $J(k)$ of the number of mutations (k) fixed in a given fixation event. Fig. 5 shows that $J(k)$ follows a geometric distribution with success probability q

$$J(k) = q(1-q)^{k-1}. \qquad [9]$$

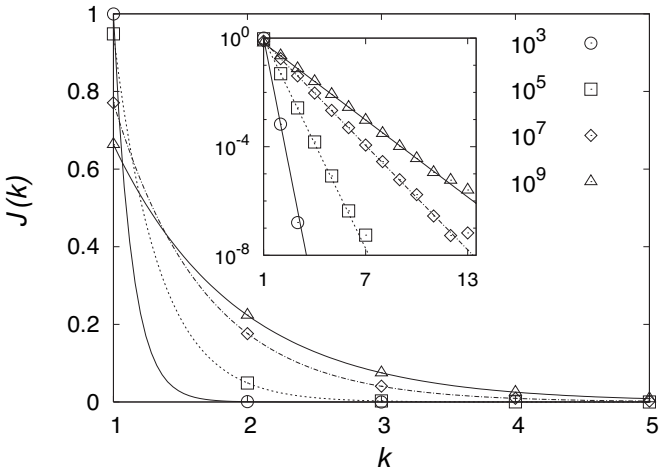

Fig. 5. The distribution $J(k)$ of the number of simultaneously fixed mutations for $N = 10^3$, 10^5, 10^7, and 10^9. The curves are the fits to the geometric distribution Eq. **9**. (*Inset*) Same, but in semilogarithmic scales.

This is similar to the known behavior in the case of neutral mutations, where $q = 2/(2 + N\mu)$ for the Moran model (18, 19). Fig. 6 depicts the N-dependence of q. For $N < 10^4$ we have $q \approx 1$, because the fixation of multiple mutations is very rare in the regime of periodic selection. In the CI regime, q decreases with increasing N, similarly to but much more slowly than in the neutral case. For the largest population size, $q \approx 2/3$, which implies that the average number of mutations fixed in a single event is $1/q \approx 1.5$.

Now let us proceed to the analysis of the IDC. In Fig. 7, the rates of increase of the various quantities related to the substitution process are depicted. Two branching points are conspicuous. At $N = 10^4$, E_X starts to deviate from the V_X (as before X represents either F or O), which means that the sequence of substitution events begins to deviate from the random Poisson process generating the mutations, and, in turn, the low population limit expression $4s_b^2\mu N$ for the RA becomes invalid (see Fig. 3 *Inset*). From around $N = 10^5$, the rates associated with the fixation and origination processes start to deviate. Hence, multiple mutations become important in the evolution, which is also reflected in the deviation of the GL prediction from the simulation data in Fig. 3.

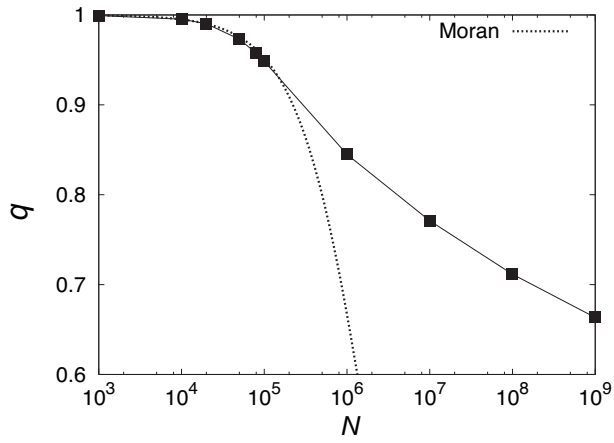

Fig. 6. Semilogarithmic plot of the parameter q of the geometric distribution vs. N. To illustrate the smooth change occurring between 10^4 and 10^5, we have added data for $N = 2 \times 10^4$, 5×10^4, and 8×10^4, which were not used in the other figures. Dotted curve shows the result for the Moran model in the neutral case (18, 19).

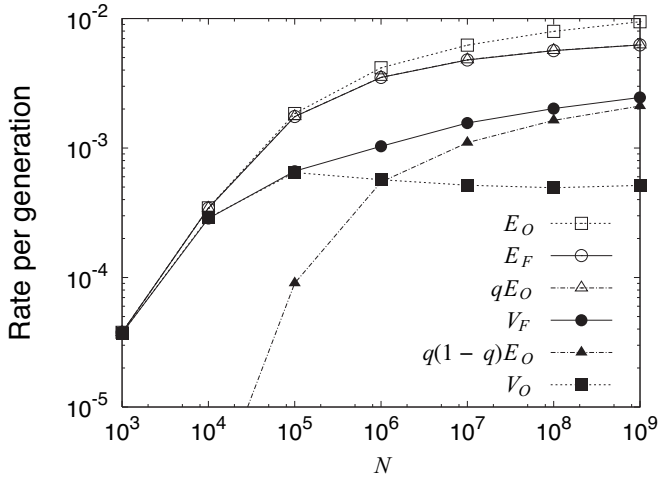

Fig. 7. Plots of E_O, V_O (open and filled squares) and E_F, V_F (open and filled circles) as a function of N in double-logarithmic scales. For comparison, plots for q E_O and $q(1-q)E_O$ (open and filled triangle) with the values for q from Fig. 6 are also drawn (see text). After about $N = 10^5$, the origination and fixation processes are discernible.

It is evident from Fig. 7 that the IDC of the fixation process is different from that of the origination process if $N > 10^5$. The data displayed in Fig. 8 show that Gerrish's prediction (13) of an asymptotically constant, nonzero but sub-Poissonian IDC applies neither to the origination process, nor to the fixation process. Instead, the qualitative nature of the IDC of the two processes becomes distinct for $N > 10^6$, with the origination (fixation) process becoming more regular (more random) as the population size becomes large. The question then naturally arises what to expect as $N \to \infty$.

For the origination process, the simulation suggests $I_O \to 0$. We now argue that this is indeed the case. As discussed before, the number of origination events, the number of mutations of the most populated genotype, and the population average of the number of mutations are interrelated. Hence, the IDC of the origination process can be calculated from that of the average number of mutations in the population. However, as the population size becomes large, the distribution of the number of mutations in the population becomes deterministic and evolves according to the infinite population evolution equations given in *SI Appendix*. Hence, the origination process also becomes completely deterministic and the IDC that reflects the variation between the number of origination events among different realizations of the stochastic evolution must vanish.

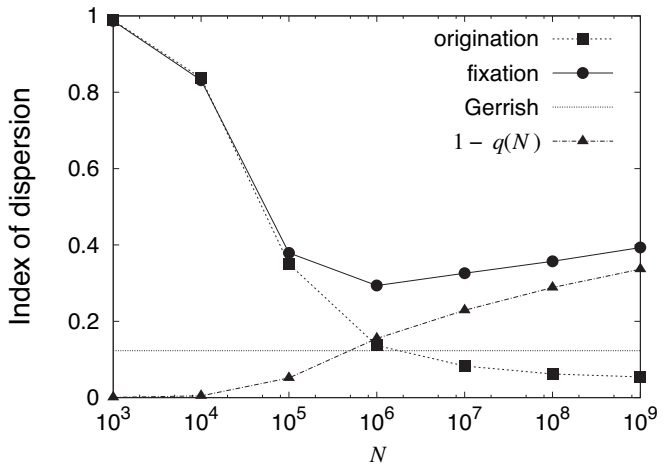

Fig. 8. Plots of the IDC of fixation and origination processes as a function of N in semilogarithmic scales. For comparison, we draw the straight line, which indicates the number (≈0.123) predicted by Gerrish (13). None of the curves approach the Gerrish number. After $N \approx 10^5$, fixation events and the number of mutations show different statistical behavior, which implies that the probability for fixation of multiple mutations becomes substantial. $1-q(N)$ is drawn for comparison with the fixation process (see text).

Based on this observation along with that of the geometric distribution $J(k)$, we can develop a phenomenological prediction as to how I_F behaves as $N \to \infty$. Because the origination process becomes deterministic for large populations, we can rescale the generation time in terms of the rate of origination events (in fact, the two rates become identical for $N \to \infty$, because $E_O = E[k] \to 1$). The fixation process can then be approximated as a Bernoulli process consisting of repeatedly performing independent but identical Bernoulli trials with success probability $q(N)$ at each (scaled) generation. This picture implies a set of simple relations between the quantities characterizing the two processes. First, because the mean number of mutations fixed in an event is $1/q$, the relation $E_F = qE_O$ holds as an identity. Second, because the IDC of the Bernoulli process is $1-q$ (22), we expect that $I_F \approx 1-q$ and $V_F \approx q(1-q)E_O$ for large N. It can be seen from Figs. 7 and 8 that these relations are rather well satisfied already for moderate population sizes $N > 10^7$. This agreement is quite remarkable, because even for the largest accessible population size $N = 10^9$ the rate E_O of the origination process is two orders of magnitude smaller than the asymptotic limit (Eq. **8**).

Having established a link between I_F and q, we now argue that $q(N)$ will decrease to zero as $N \to \infty$. Because two fixation events are on average separated by $1/q$ origination events (or scaled generations), the fixation time for any one of these originating mutations is of order $1/q$. As discussed above, the initial genotype (or wild type) will be removed at time $\approx 1/s_b$, because the largest selection coefficient is $\approx s_b \ln(\mu N)$. After the wild type is removed, the subsequent fixation events are determined by clonal interference. Because the mean difference between the largest and the second largest selection coefficient in the population is s_b (see *SI Appendix* for calculation), the selective advantage of the fittest clone against the second fittest clone is only $\approx 1/\ln(\mu N)$. Hence, fixation will occur after a time of order $\approx O((\ln N)^2)$, which diverges for $N \to \infty$, and it follows that $q \to 0$ in this limit.

According to the above argument, q decreases logarithmically for very large N, which explains the slow decrease seen in Fig. 6. This should be compared with the neutral mutation model, where the fixation time is of order N and accordingly q decreases proportional to the inverse of the population size (18, 19). The slow decrease of q appears to be linked to the presence of mutations with a distribution of selection coefficients. Our preliminary study of a model (17) in which all mutations confer the same selective advantage indicates a faster decay of $q(N)$. This finding is consistent with the fact that in such a setting there is a high degree of degeneracy in the fitness of different genotypes, which brings the behavior closer to the neutral case.

Summary and Discussion

Limitations of the Model. The model studied here may appear to have two significant shortcomings: the neglect of deleterious mutations and the assumption of an unlimited supply of beneficial mutations. Regarding the first issue, in the large populations of interest here, Muller's ratchet (5, 29, 30) can be safely disregarded and a mutation–selection equilibrium can be assumed to exist for the deleterious mutations. This is consistent with the deterministic calculation in the infinite population limit, which shows that the increase of the fitness in the asymptotic regime is only governed by the effect of the beneficial mutations irrespective of the strength and the rate of the deleterious mutations (see *SI Appendix*). According to the studies by Orr (14) and Wilke (15), the population

EVOLUTION

size N in our model can then be interpreted as the population size of the genotype without deleterious mutations, that is, $N \mapsto N \exp(-U/s_d)$, where U is the rate of the deleterious mutation and s_d is the strength of a deleterious mutation (30).

In the absence of Muller's ratchet, deleterious mutations can be fixed only by hitchhiking with beneficial mutations. If $U/s_d < 1$, the subpopulation without deleterious mutations is larger than that with one deleterious mutation (30). Using extreme value statistics, it is then easy to see that most of the fixed beneficial mutations arise from the genotype without deleterious mutations. On the other hand, if $U/s_d \gg 1$, beneficial mutations occurring in genotypes with a few deleterious mutations can have larger fitness than those in the genotype without deleterious mutation. In this regime, fixation of deleterious mutation by hitchhiking can frequently happen and may change the statistics of the fixation and origination processes. Hence, the results of this paper remain applicable when U/s_d is sufficiently small.

The assumption of an unlimited supply of beneficial mutations is common to most theoretical work on clonal interference (1, 13–15, 17), but it is strictly true only for an infinite number of sites. A population evolving on a finite space of genotypes sooner or later approaches a fitness peak and the supply of beneficial mutations accordingly dwindles (31, 32). At the same time, beneficial mutations become recurrent, which alleviates the effect of CI (33, 34). As long as the fitness landscape remains static, the applicability of our results is therefore limited to an early time regime where the finite extent of the genome is not yet felt. On the other hand, the continuous supply of beneficial mutations can also be thought to mimic a situation where the environment changes over time. In this case, the comparison of the fitness between two genotypes adapted at different environment is meaningless, and hence the overall RA cannot be read off from an increase of the mean fitness. Our results for the rate and statistics of substitution events may nevertheless be applicable, provided the environmental change is slower than the time scale of the fixation and this change does not affect the mean selective advantage significantly.

Summary. We have presented a detailed study of asexual adaptation in large populations, where CI is common. Two key simplifying assumptions of the established theory of CI (1) were identified and shown to have opposite effects on the rate of adaptation. For the population sizes that are accessible to our simulations (up to $N = 10^9$), the correlation between clonal interference and survival against drift leads to a moderate decrease of the adaptation rate compared with the Gerrish-Lenski prediction, but for larger populations (or larger values of $N\mu$) we predict a significant speedup of adaptation due to multiple mutations. In contrast, the effect of multiple mutations on the statistics of substitution events is important throughout the range of population sizes where CI occurs. We have shown that Gerrish's prediction of "rhythmic" adaptation, in the sense of a decreased index of dispersion for counts of substitution process (13), is qualitatively correct as far as the origination process is concerned, but the distinction between the processes of origination and fixation is crucial. The two are linked through the geometric distribution of the number of mutations fixed in a single fixation event characterized by a single parameter q, which succinctly encapsulates the statistical structure of the substitution process.

Methods

Our numerical work is based on the Wright–Fisher model of asexually reproducing organisms with fixed population size N. The reproduction scheme is as follows: Each individual i is assigned fitness $w_{i,t}$ ($i = 1, \ldots, N$) at generation t. Initially, all individuals have the same genotype and accordingly same fitness $w_{i,0} = 1$. The probability that a parent of an individual in the next generation is i is $w_{i,t}/(\bar{w}(t)N)$, where $\bar{w}(t) = \sum_{i=1}^{N} w_{i,t}/N$ is the mean fitness at generation t. In the actual simulation, we did not discern different progenitors if they have the same genotype. Instead, the number of progeny of a given genotype is determined from the multinomial distribution with the probability also proportional to the population of that genotype. The multinomial distributed numbers are chosen by sampling correlated binomial random numbers, as described in *SI Appendix*; see also ref. 35.

Once an offspring has chosen its parent, a mutation can change its genotype with probability μ. For simplicity, every mutation is assumed to change only one nucleotide and we neglect the effect of deleterious mutations, that is, every mutation is beneficial (a discussion of the consequences of including deleterious mutations can be found in *Summary and Discussion*). When a mutation occurs in an individual whose parent is i, its fitness becomes $w_{i,t}(1 + s)$, where s is a random number drawn from the exponential distribution (Eq. **2**). If no mutation occurs, the offspring simply inherits the fitness of its parent. The above steps are repeated until the end of the observation time, which is set to 20,000 generations in this paper. In the simulations presented here, the mutation probability and the average selection coefficient are set to the values $\mu = 10^{-6}$ and $s_b = 0.02$, respectively, and the focus is on the variation of the population size N.

This work was supported by Deutsche Forschungsgemeinschaft within SFB 680 Molecular Basis of Evolutionary Innovations. S.-C.P. acknowledges partial support by National Science Foundation Grant PHY99-07949 during a visit at the Kavli Institute of Theoretical Physics (Santa Barbara, CA).

1. Gerrish PJ, Lenski RE (1998) *Genetica* 102/103:127–144.
2. Fisher RA (1930) *The Genetical Theory of Natural Selection* (Clarendon, Oxford).
3. Muller HJ (1932) *Am Nat* 66:118–138.
4. Crow JF, Kimura M (1965) *Am Nat* 99:439–450.
5. Felsenstein J (1974) *Genetics* 78:737–756.
6. Otto SP, Lenormand T (2002) *Nat Rev Genet* 3:252–261.
7. Atwood KC, Schneider LK, Ryan FJ (1951) *Proc Natl Acad Sci USA* 37:146–155.
8. de Visser JAGM, Rozen DE (2006) *Genetics* 172:2093–2100.
9. de Visser JAGM, Zeyl CW, Gerrish PJ, Blanchard JL, Lenski RE (1999) *Science* 83:404–406.
10. Miralles R, Gerrish PJ, Moya A, Elena SF (1999) *Science* 285:1745–1747.
11. Rozen DE, de Visser JAGM, Gerrish PJ (2002) *Curr Biol* 12:1040–1045.
12. Perfeito L, Fernandes L, Mota C, Gordo I (2007) *Science* 317:813–815.
13. Gerrish PJ (2001) *Nature* 413:299–302.
14. Orr HA (2000) *Genetics* 155:961–968.
15. Wilke CO (2004) *Genetics* 167:2045–2053.
16. Barrett RDH, M'Gonigle LK, Otto SP (2006) *Genetics* 174:2071–2079.
17. Desai MM, Fisher DS, Murray AW (2007) *Curr Biol* 17:385–394.
18. Watterson GA (1982) *J Appl Prob* 19A:59–70.
19. Watterson GA (1982) *Adv Appl Prob* 14:206–224.
20. Gillespie JH (1993) *Genetics* 134:971–981.
21. Gillespie JH (1994) *The Causes of Molecular Evolution* (Oxford Univ Press, Oxford).
22. Feller W (1968) *An Introduction to Probability Theory and Its Applications* (Wiley, New York), Vol I, 3rd Ed, pp 303–341.
23. Guess HA (1974) *Ann Prob* 2:14–31.
24. Gillespie JH (1983) *Theor Popul Biol* 23:202–215.
25. Orr HA (2003) *Genetics* 163:1519–1526.
26. Guess HA (1974) *Theor Popul Biol* 5:417–430.
27. Eshel I (1971) *Theor Popul Biol* 2:209–236.
28. Kimura M (1962) *Genetics* 47:713–719.
29. Muller HJ (1964) *Mut Res* 1:1–9.
30. Haigh J (1978) *Theor Popul Biol* 14:251–267.
31. Orr HA (2002) *Evolution (Lawrence, Kans)* 56:1317–1330.
32. Jain K, Krug J (2007) *Genetics* 175:1275–1288.
33. Kim Y, Orr HA (2005) *Genetics* 171:1377–1386.
34. Bollback JP, Huelsenbeck JP (2007) *Mol Biol Evol* 24:1397–1406.
35. Devroye L (1986) *Non-Uniform Random Variate Generation* (Springer, New York), pp 588–589.