# Clonal Interference Models in Population Genetics

Carl Boettiger

January 10, 2006

**Abstract**

Studies of evolutionary population genetics have long overlooked asexual organisms for historical reasons. Experiments in bacterial populations begun in recent decades now offer extensive data, but theoretical models have only recently begun to address such populations. Here we develop a theoretical background for population genetics of asexual populations and then apply it to exploring a phenomenon unique to such organisms known as clonal interference. By examining two separate papers on the topic we discover that the phenomenon can occur through two separate mechanisms of interference, and what parameter values (regimes) correspond to what mechanisms. Based on our theoretical development we also identify difficulties in the models described by each paper. Additionally, we develop a simulation to further explore the dynamics of the second mechanism.

## 1 A History of Sex Bias

Our current understanding of evolution is encapsulated by the "Modern Synthesis" in which the seemingly contradictory theories of Darwin's evolution and Mendel's genetics were unified into a powerful theoretical model, primarily through the work of R.A. Fisher, J.B.S. Haldane and Sewall Wright. Following this synthesis studies in evolution would fall mostly under the domain of a new field forged by these pioneers: population genetics [3, 4]. This excellent theoretical work was complemented by observational studies (led by T. Dobzhansky and E.B. Ford), and increasingly pursued in experimental laboratory settings following the pioneering work of T.H. Morgan in the fruit-fly *Drosophila* [4]. One historical consequence of these beginnings has been that population genetics has been primarily focused on sexual organisms. For instance, the Hardy-Weinberg law, one of the fruits of early theoretical work which would become the traditional starting point of population genetics[1], applies purely to sexual populations. Likewise, the simple model first introduced by Fisher of a diploid, monoecious population of fixed size $N$ with binomially distributed overspring (Commonly called the "Wright-Fisher" model) became the gold standard of the field and the cornerstone of almost all subsequent work [3].

Just as theoretical work continued to build upon the models of sexual organisms, experimental work quickly became dominated by Morgan's fruitfly despite some early attempts in research in bacterial populations. Only recently have bacteria, now so central in molecular techniques for their ease

[1] Largely thanks to its instrumental role in unifying Mendelian genetics and evolution. Prior to the acceptance of Mendelian genetics, inheritance was believed to operate under the blending hypothesis. Clearly this would reduce the total variation in a population by one half each generation. Since natural populations show great variation, this description must be incomplete, and complicating factors must be hypothesized causing individuals not to resemble their parents. However, if this were the case, natural selection could not operate as the offspring of fit individuals would not likely inherit that fitness. The discrete nature of Mendelian genetics allows inheritance and variation, at ratios predicted by Hardy-Weinberg equilibrium. [3]

of growth and manipulation, realized their potential as ideal organisms in which to study evolution. In addition there easily controllable environments and ease of growing an maintaining in laboratory, bacteria offer rapid generations, large populations essential for much research in the field. Further bacteria can be maintained indefinitely frozen, allowing a more evolved strain to directly compete against its ancestors brought back to life just as they were many of generations ago. This offers a direct measure of fitness change, another vital and often challenging parameter to obtain in population genetics studies. (See the recent *Nature* review article by Lenski *et al.* for a more thorough description of the merits of studying evolution in bacteria [7]). Despite the rising number of long-term bacterial evolution data now available, the theoretical side of the field remains heavily entrenched in its old models of sexual organisms. The existence of certain phenomena unique to asexual populations has prompted several forays into developing theoretical models particularly for such populations. Chief among these is a phenomenon recognized among population genetics' pioneers but undeveloped until recently: the idea that beneficial mutations in an asexual population could be lost due to competition within the population, commonly termed "clonal interference." Such is the topic of this paper. We will focus primarily on two previous studies of clonal interference that present different models, functioning in different regimes and offering different mechanism of interference, and different theoretical conclusions. Both present challenging calculations and suffer oversimplifications and computational troubles. We shall not manage to reach a thorough quantitative understanding of the phenomenon, but instead outline the tools and approaches to address it.

Despite such difficulties, the essence of clonal interference is very simple: mutations can occur faster than they can be fixed in the population. When this is not so, a mutation spreads to the entire population (fix) before the next mutation occurs, and hence the new mutation is guaranteed to occur in an individual having the old mutation. Yet when the mutations occur more rapidly than they fix, an individual can receive the second mutation without having the first. In an asexual organism the decedents of these two are forever distinct. In a sexual population both could increase in frequency until it became probable that a descendant of one would share offspring with a descendant of another, producing an individual with both beneficial mutations that could ultimately fix. In an asexual population however, an individual will only ever benefit from one of the mutations, and the other is guaranteed to eventually be lost.

## 2 Models in Population Genetics

We begin then by a foray into the population genetics of asexual individuals. Models of population genetics are concerned primarily with three values, denoted the population size $N$, mutation rate (per genome per generation) $u$ and the fitness advantage, $s$. Throughout this paper we will consider $u$ to be only the rate at which beneficial mutations occur, and ignore deleterious (which would be selected against), and neutral mutations (which would have no effect on our model as it focuses only on changes in fitness, not merely in gene frequencies). A model containing all three parameters becomes very complicated, as indeed even models with only two of the parameters can become analytically intractable. Hence we follow the tradition of the field by working with the simplest models we can construct, and by working in explicitly in limits where the effect of some of the parameters may be ignored. As we come to understand the behavior of these simple models, we then may include the complicating effects of another parameter. This approach has met with remarkable success in this field, and will serve us well as we extend much of the analysis found in classical theory (see Gillespie [6] for instance), to the case of asexual organism. Many of these results are readily found for sexual populations, but the analogous result beyond the most trivial cases is rarely worked out explicitly for an asexual population, though sometimes the final result is unchanged. Throughout, we shall attempt to apply the notation that is most standard in the field while maintaining consistency and clarity, though this sometimes breaks with modern conventions of statistics and probability.

In our first model we will ignore both mutation and selection (imagine the mutation rate to be too small to introduce any new selective variation over the timescale in which we are currently interested.) Imagine a bacterial population of equal fitness but varying (possibly unique) genotypes, remaining at population size of close to $N$. The key feature of the model is a stochastic reproduction process, where the number of offspring an individual has in the next generation is given by some probability distribution. The probability that a certain genetic allele will have $x$ copies in the next generation depends only on its frequency $p$ in the current generation,

$$P_B(x; N, p) \equiv \binom{N}{x} p^x (1-p)^{N-x} = \frac{N!}{x!(N-x)!} p^x (1-p)^{N-x} \tag{1}$$

The variation in the population is then categorized by the variance $x$, which for the binomial distribution is $Np(1-p)$. Hence the variance of our gene in the next generation, $p' = x/N$ is simply

$$\mathrm{Var}\{p'\} = \frac{\mathrm{Var}\{x\}}{N} = p(1-p)/N \tag{2}$$

The precise expression for the variance can be model dependent. For instance, in a model where individuals either divide or die, $x$ will always be even and only half of the population will have offspring, hence one must use $N/2$ and the variance in equation 2 would be doubled. The variance will figure into many later calculations, and one would have to be careful to work out the necessary consequences of choosing another model. As mentioned before, equation 1 is often referred to as the (simplest form of the) Fisher-Wright Model and forms the cornerstone of most of population genetics. (At this level the model is identical for the canonical diploid, sexual hermaphroditic population as for our asexual one.) Rather than re-derive the results for a each new model one might consider, population geneticists have developed the concept of an effective population size, $N_e$, which is often defined as the value of $N$ that would give the model the same variance as found in the Fisher-Wright model. For the model just suggested (which is also used later in simulation), the effective population size is simply $N/2$. By using the effective population size the hope is to avoid altering the many well-established results build upon this model. Acknowledging this, we will continue to use the expression given by equation 2 for the variance. [2]

In each generation individuals face some probability of leaving no descents in the next generation. Should they be the only individual left in the population carrying a particular trait, that trait would then be lost forever. As no mechanism (such as mutation) exists that would restore variation in the population, we observe that population variance decreases with some probability each generation. To make this precise, allow $\mathcal{G}$ to denote the probability that two individuals in the current generation are genetically identical.[3] This can come about in two ways: either the second has the same parentage as the first, (of probability $1/N$), or each are the offspring of individuals whom themselves were already identical, (which occurs with probability $\left(1 - \frac{1}{N}\right)\mathcal{G}$.) Hence the probability two individuals are identical in the next generation ($\mathcal{G}'$ is given by,

$$\mathcal{G}' = \frac{1}{N} + \left(1 - \frac{1}{N}\right)\mathcal{G} \tag{3}$$

[2] The matter is not actually quite so simple. Following the spirit of what we have just described, population geneticists have come up with at least three separate notions of effective population size, that need not be equivalent, particularly in populations with nonconstant size. See Ewens 2004 for discussion[3].

[3] for convenience we will consider an individual to be defined by its genotype alone and not distinguish between genotypic and phenotypic variation.

Defining the complement probability $\mathcal{H}$ that two individuals are genetically distinct as $\mathcal{H} = 1 - \mathcal{G}$ and substituting we have

$$\mathcal{H}' = \left(1 - \frac{1}{N}\right)\mathcal{H} \tag{4}$$

By iterating this equation, we discover how $\mathcal{H}$ has changed $t$ generations later:

$$\mathcal{H}_t = \mathcal{H}_0 \left(1 - \frac{1}{N}\right)^t \tag{5}$$

From this we can find how long it reach one $e^{\text{th}}$ of the initial variation:

$$\frac{\mathcal{H}_0}{e} = \mathcal{H}_0 \left(1 - \frac{1}{N}\right)^\tau \tag{6}$$

Then solving for time and approximating the logarithm for large $N$, we have

$$\boxed{\tau = \frac{1}{\ln(1 - 1/N)} \approx N} \tag{7}$$

Here we have recovered a fundamental result of population genetics: that stochastically variable numbers of offspring reduce variation in a population on a timescale that goes as the size of the population. This phenomenon is know as genetic drift. For finite populations it is obvious that after a few times $N$ generations all individuals will be identical – every member will share a single ancestor. Whatever traits that individual carried are now found in all individuals of the current population, while the traits of all others in the initial population have been lost forever. In the language of population genetics, the traits of the last common ancestor are said to have fixed in the population.

One commonly discusses the frequency $p$ of a certain trait or gene of interest, which is simply given by $p = n/N$ for $n$ individuals carrying the gene in a total population of size $N$. Clearly throughout the evolution of our population, $p$ will fluctuate between zero and one. These boundaries are absorption points, where the gene has either become extinct or has fixed – once $p$ reaches a boundary value it remains there for all time; its finally outcome can only be one of two states. Consequently we would like to express the probability of the gene fixing rather than going extinct. In our simple model this expression is trivial,

$$\pi(p) = p \tag{8}$$

The probability $\pi(p)$ is just the initial frequency, introducing a standard notation we shall employ later. We see this immediately if we look back from when a gene has fixed and note that without selection each member of the initial population was equally likely to survive to fixation.

We are now ready to introduce the mutation rate $u$ into our model. Knowing that a new mutation will ultimately be lost or become fixed, we would like to know the rate at which new mutations fix in the population, given the rate at which the mutations occur in an individual member. Since each individual in the population receives mutations at a rate of $u$, novel mutations enter the population at a rate $Nu$. However, as we discovered in our earlier model, most variation is lost by genetic drift, and hence the probability that a new mutation survives drift is $\pi(1/N) = 1/N$, and the rate at which novel mutations fix in the population ($\rho$) is given by:

$$\rho = Nu\pi\left(\frac{1}{N}\right) = u \tag{9}$$

Surprisingly this result is independent of $N$, the population size. As we shall see, this is rather different than what we will find even in models with only very weak selection. The smalllest of selective advantages destroys this independence of $N$ (as we will see, this is strictly true only of large $N$, though such an assumption is valid for most populations). As an interesting aside, because the rate of fixation and the rate of mutation may be experimentally estimated for two species [4], equation 9 suggests a method to determine whether or not most evolution is selective or neutral. Observations for a variety of species of varying population size suggesting that $\rho = u$ sparked the exciting and controversial "neutral theory" of evolution that remains much debated to this day.

We now consider the effect of selection in our models. Selection is a subtle and powerful complication, hence we shall begin by considering the effects of selection alone, independent of mutation and drift. Our first calculation will determine how selection increases the frequency of a beneficial mutation, but first we pause for a brief discussion on what is meant by the fitness of an individual. Fitness essentially describes how a mutation's frequency is expected to increase. A gene with no selective advantage is not expected to increase in frequency, $p' = p$, (where prime denotes the next generation) and consequently its fitness relative to the population is defined as $w = 1$. Meanwhile, a gene with a selective advantage relative to the rest of the population is expected to increase by some small amount $s$ and hence has fitness $w = e^s \approx 1 + s$ for small $s$. Hence we can express the mean fitness ($\bar{w}$) of the population as

$$\bar{w} = p(1 + s) + 1 \cdot (1 - p) \tag{10}$$

The expected frequency of a beneficial mutation in the next generation ($p'$) is given by

$$p' = \frac{p(1 + s)}{\bar{w}} \tag{11}$$

This equation captures the essential meaning of fitness and selective advantage. From this we find the expected change in fitness to be

$$\Delta p = p' - p = \frac{p(1 + s) + \bar{w}p}{\bar{w}} \tag{12}$$

Substituting for $\bar{w}$ (equation 10) and simplifying:

$$\Delta p = \frac{ps - p^2 s}{ps + 1} \tag{13}$$

Assuming $s \ll 1$, this can be written as

$$\boxed{\Delta p \approx sp(1 - p)} \tag{14}$$

giving the expected (mean) increase in frequency of a beneficial mutation. Much will follow from this simple result.

Under neutral evolution, we found a very simple expression for the probability that a mutation fixes in a population, and observed that the result was independent of $N$. As soon as we take into account the effect of selection, even at vanishingly small values, we find this conclusion no longer holds. Our calculation will rely on the interaction between drift, governed by $N$, and selection, governed by $s$. Notably, our calculation of the fixation probability will not take into account $u$, that is, we assume that any mutation occurring in the population will be either lost or fixed well before we would expect a new mutation to enter the population, occurring at rate $Nu$. We will later dispense with this

[4] The rate mutations fix may be determined by counting the number of mutations accumulated between two species since their last common ancestor, assuming the time of separation is known from another source, such as the fossil record. The mutation rate might be estimated from the rate of genome proof reading errors, etc measurable in a biological laboratory.

assumption, entering into the regime of clonal interference where all three factors are operating over the relevant timescale. Given that a beneficial mutation has achieved frequency $p$, the probability that it fixes is simply its expected frequency in the next generation:

$$\pi(p) = \langle \pi(\Delta p + p) \rangle \tag{15}$$

For some small increase $\Delta p$ we Taylor expand around $p$:

$$\pi(\Delta p + p) \approx \pi(p) + \pi'(p)\Delta p + \frac{1}{2}\pi''(p)(\Delta p)^2 \tag{16}$$

Since the mean of the sum is the sum of the means, substituting equation 16 into 15 gives

$$\pi(p) \approx \pi(p) + \pi'(p)\langle \Delta p \rangle + \frac{1}{2}\pi''(p)\langle (\Delta p)^2 \rangle \tag{17}$$

The variance is given by $\mathrm{Var}\{\Delta p\} = \langle (\Delta p)^2 \rangle - \langle \Delta p \rangle^2$. If both the mean and variance of $\Delta p$ are small and similar in magnitude, then the mean squared is much smaller than the variance, and may be ignored. To make this approximation explicit, assign for the variance $v(\Delta p) \equiv \langle (\Delta p)^2 \rangle$ and similarly for the mean $m(\Delta p) \equiv \langle \Delta p \rangle$. After subtracting out $\pi(p)$ equation 17 becomes

$$\frac{1}{2}v(\Delta p)\pi''(p) + m(p)\pi'(p) = 0 \tag{18}$$

Which is a linear differential equation for the fixation probability $\pi(p)$, which can be solved given the boundary conditions

$$\pi(0) = 0 \qquad \pi(1) = 1 \tag{19}$$

Transforming to first order by $f(p) \equiv \pi'(p)$ and rearranging:

$$f'(p) + \frac{2m(p)}{v(p)}f(p) = 0 \tag{20}$$

Multiplying by an integrating factor $\alpha = e^B(p)$ where $B(p) \equiv 2\int_0^p \frac{m(x)}{v(x)}\mathrm{d}x$ lets the lefthand side be written as a perfect derivative, which is then integrated out:

$$f(p) = C_1 e^{-B(p)} = \pi'(p) \tag{21}$$

Hence

$$\pi(p) = C_1 \int_0^p e^{-B(y)}\mathrm{d}y + C_2 \tag{22}$$

The boundary conditions $\pi(0) = 0$ and $\pi(1) = 1$ determine $C_1$ and $C_2$,

$$\pi(p) = \frac{\int_0^p e^{-B(y)}\mathrm{d}y}{\int_0^1 e^{-B(y)}\mathrm{d}y} \tag{23}$$

Using the values for the mean $sp(1-p)$( equation 14) and variance $p(1-p)/N$ (equation 2), $B$ becomes

$$B(y) = 2\int_0^y \frac{sp(1-p)}{p(1-p)/N}\mathrm{d}p = 2Nsy \tag{24}$$

Substituting and integrating,

$$\pi(p) = \frac{1 - e^{-2Nsp}}{1 - e^{-2Ns}} \tag{25}$$

Hence for a new mutation of frequency $1/N$

$$\boxed{\pi\left(\frac{1}{N}\right) = \frac{1-e^{-2s}}{1-e^{-2Ns}} \approx 2s} \tag{26}$$

Where the approximation holds for large $N$.[5] This assumption of large $N$ holds for most natural populations and will be especially apt for the bacterial populations we shall consider, making the result notably independent of the population size. This probability will play an important role in the dynamics of models with clonal interference that we shall examine.

The fixation probability can be expressed more generally. For a mutation of frequency $p$, assuming $1 \ll Ns$ equation 25 becomes

$$\pi(p) = 1 - e^{-2Nsp} = 1 - \epsilon \tag{27}$$

where $\epsilon$ is the probability that the mutation does not fix. We can find what frequency our gene must obtain in order to have some arbitrarily small chance $\epsilon$ of not fixing:

$$p \approx \frac{-\ln(\epsilon)}{2Ns} \tag{28}$$

Since $p$ is the number of individuals with the beneficial mutation divided by population size, the population size cancels from both sides and the number of individuals $x$ carrying the beneficial mutation required to have a certain probability of fixing is independent of population size!

$$x = \frac{-\ln(\epsilon)}{2s} \tag{29}$$

For instance, a beneficial mutation with a two percent selective advantage that has reached a population of $x = \frac{-\ln(.01)}{2*.02} = 115$ individuals to have a 99% chance of fixing, regardless of whether it be found in a population of one million or one trillion individuals (recall the approximation holds only for large $N$, while these populations are quite accessible in laboratory experiments with bacteria). It is convienent to refer to a population that has reached this size as being "established," though it is far from fixation (perhaps less than a thousandth of a percent), its eventual fixation has been effectively guarenteed by its fitness advantage. Counterintuitive results such as this highlight the need for careful analysis of such questions. Equation 29 can be used to justifies our earlier regime where we ignored the effect of drift and looked only at the effect of selection. We see that we can effectively ignore the influence of genetic drift once a population has reached a population size of order $1/s$ hence a frequency of order $1/(Ns)$. Starting from this frequency we can use equation 14 to compute the time until the mutation fixes.

Treating our expression 14 as differential equation for $p$ with respect to time (in generations), we can find the expected number of generations required to fix. First, separating and taking the integral of both sides we have

$$t = \int \frac{ps+1}{ps-p^2s}\mathrm{d}p \tag{30}$$

After integrating by partial fractions and some manipulation of logarithms we're left with

$$= \frac{1}{s}\ln\left(\frac{p}{(1-p)^{s+1}}\right) + C \tag{31}$$

[5] Desai, Fisher and Murray simply approximate this as $s$. Gerrish and Lenski present an argument in their appendices that they believe the value should in fact be $4s$, based on a slightly simpler approach that appears to have fallen out of favor in the recent literature. The value of $4s$ would imply a model with half the variance of the model given by equation 1, which need not be the case for bacterial populations generally.

Our constant is determined by our initial condition. Our model requires that the initial frequency be sufficiently large that we can ignore the effect of genetic drift. As we found in equation 29, our beneficial mutation has effectively escaped drift upon reaching a frequency of order $1/(Ns)$. This is the frequency at which we start our clock, $t = 0$ generations, allowing us to determine $C$.

$$C = -\frac{1}{s}\ln\left(\frac{1/(Ns)}{(1-1/(Ns))^{s+1}}\right) \tag{32}$$

Hence the overall equation for the number of generations until fixation under selection alone becomes:

$$t = \frac{1}{s}\ln\left[Nsp\left(\frac{(1-1/(Ns))}{(1-p)}\right)^{s+1}\right] \approx \boxed{\frac{1}{s}\ln\left(\frac{pNs}{1-p}\right)} \tag{33}$$

One immediate observation from this equation is that as $p$ gets closer and closer to one, time goes to infinity. This is because our gene frequency is increasing by selection alone, and once it makes up most of the population it has raised the mean fitness significantly such that it has effectively no fitness advantage. At this stage genetic drift comes into play again, this time in the favor of the beneficial mutation, sweeping it to fixation. We have observed that without any selection, genetic drift will fix a particular genotype with probability equal to its frequency in the population with a timescale of order its number of individuals in the population ($pN$). That is, drift will come into play for when

$$spN \lesssim \ln\left(\frac{pNs}{(1-p)}\right) \tag{34}$$

Clearly drift will not play a strong role until the gene is very near fixation. Hence a reasonable approximation to fixation time will be when $p = 1 - 1/N$, in which case we find,

$$t_{\text{fix}} \approx \frac{1}{s}\ln(sN^2) \tag{35}$$

We will also be interested in computing this time for $p$ of to reach order $N$, that is for $p = 1/2$ we find

$$t_{\text{half}} \approx \frac{1}{s}\ln(Ns) \tag{36}$$

How these timescales compare to the mutation rate will determine the role of clonal interference in the population, and we shall refer back to them shortly.

With the approximations $s + 1 \approx 1$ and $sN \gg 1$. We can also solve for $p$

$$p = \frac{1}{sNe^{-st}+1} \tag{37}$$

Figure 1 shows the plot of a beneficial mutation's frequency as a function of time, as given by equation 37, showing the S shaped fixation curve (in red). Note that this shape is given by selection alone, and does not include the effects of drift. The curve is drawn beginning with a frequency that has already escaped drift ($p = 1/(sN)$) and goes until it is one individual away from fixation, $p = 1 - 1/N$. The figure also includes the results of a simulation of an evolving bacterial population described in a later section (blue curve). The simulated fixation takes longer than expected from our simple prediction. This effect is also readily observed in many experimental bacterial populations [1, 5, 7], and is due to the phenomenon of clonal interference.

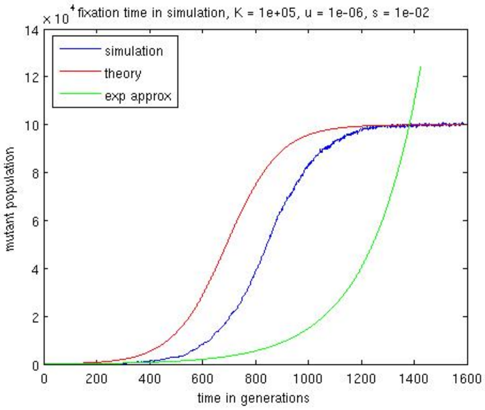


Figure 1: Compare to the red curve predicted by equation 37, the blue curve from simulation and exponential growth at the average rate $s/2$ (see discussion of the model proposed by Desai *et al.*, equation 47 page 14). The delay between the predicted curve (red) and the simulation (blue) is attributable to clonal interference. Before the mutation considered here fixed, a new mutation developed in the mutant population, which established and began to grow. This raised the mean fitness perceived by our mutant, slowing its fixation.

We will see that clonal interference breaks into several regimes, based on the timescales we have derived. In the simplest case, mutations fix much faster than new mutations establish into the population (at rate $2sNu$), no interference occurs, hence we write the noninterfering regime as

$$\ln(sN^2) \ll \frac{1}{2Nu} \tag{38}$$

In such a case, our derivation of fixation probabilities ignoring the possibility of new mutations occurring is accurate, and we find that new mutations will fix in the population at rate:

$$\rho = Nu\pi = \boxed{2Nus} \tag{39}$$

In contrast to our neutral evolution model which depended on $u$ alone, we find the fixation rate of beneficial mutations in the population depends on all three parameters. Each time a beneficial mutation fixes, the mean fitness is raised by $s$, hence the rate at which the mean fitness increases is simply,

$$v = \frac{\mathrm{d}\bar{w}}{\mathrm{d}t} = 2Nu\langle s^2\rangle \tag{40}$$

We shall follow Desai *et al.* by referring to this non-interfering limit as the linear regime, since the rate of evolution increases linearly with population size (similarly, linear in the product $Nu$). As the rate at which mutations enter the population, $Nu$ continues to increase to very large values, the condition given by equation 38 breaks down, and this result (equation 40) no longer holds. The rate of fitness increase rises only weakly with increasing $Nu$ because not all mutations reaching established sizes will fix. There are two ways in which established mutations can be wasted: (a) before a mutation fixes, another mutation occurs in the ancestral population which can compete with it where only one will eventually fix, or (b) a mutation occur in a subpopulation whose fitness lags significantly behind the best mutants (sometimes referred to as the nose or leading edge in deference to its position on the fitness distribution) as to remain doomed to extinction. We consider two papers, each essentially modeling one of these mechanisms of clonal interference. Though each offers some illumination of the mechanism being discussed, it is unclear that either approach offers a precise description of the phenomenon.

# 3 Models of Clonal Interference

## 3.1 Gerrish and Lenski Model

We shall first consider the approach of Gerrish and Lenski [5], considering interference purely as a result of mutations occurring in the ancestral population. Unfortunately Gerrish and Lenski fail to acknowledge the existence of the second mechanism, which comes into effect at about an order of magnitude greater in $Nu$ then required to first enter the interference regime (see discussion beginning section 3.2, page 12. They begin by calculating the number of new mutants expected in the ancestral population from the time the first mutant reaches establishment until it fixes. This is done by integrating the ancestral population, having frequency $1-p$ where $p$ is given by equation 37. This result must then be multiplied by $N$ to convert frequency to a number of individuals and then by $u$ to find the number of mutations introduced in this approximation. The fixation time (the upper bound of the integral) is given by equation 33 for $p = 1 - 1/N$, and the integral (under the approximations specified for equation 37) gives for the expected number of mutations occurring

$$uN \int_0^{\ln(sN^2)/s} \frac{sN}{sN + e^{st}} \mathrm{d}t = \frac{uN \ln(Ns)}{s} \tag{41}$$

Gerrish and Lenski instead find this to be$\frac{Nu}{s}\ln(N)$, which can be shown to in fact be the result of integrating $p$ instead of $1-p$, giving instead the expected number of mutations occurring on top of the initial mutation before it reaches fixation. We note this simple correction and continue along the same vein of their analysis.

Their next step is to assume that such a mutation will only interfere if survives drift (i.e. reaches an established population size) and has a fitness greater than the current beneficial mutation. In order to compute the probability that a mutation will occur with a higher fitness one needs to assume a form for the distribution of beneficial mutations, which they take to be an exponential distribution $\alpha e^{-\alpha s}$ referencing the arguments given previously by Kimura[10]. (Desai *et al.* reject this assumption in favor of a simple average increase value [1].) They determine the probability of a given mutation surviving drift and having selective advantage greater than $s_0$ as the integral of the distribution of $s$ times the probability $\pi(s)$ that the mutation survives drift, integrated from the fitness of the currently fixing mutation, $s_0$, to infinity. This gives the probability that a mutation will interfere as

$$\int_{s_0}^{\infty} \pi(s)\alpha e^{-\alpha s}\mathrm{d}s = \pi\left(s + \frac{1}{\alpha}\right) e^{-\alpha s} \tag{42}$$

There are several difficulties both with this approach and the with assumption that interference occurs (that is, a mutation that establishes fails to fix) only when a more beneficial mutation occurs in the ancestral population. We begin with a pedagogical difficulty in the above equation. Over most of the time interval until fixation ($t_{\text{fix}}$ equation 35) the mutant population is larger than the ancestral population, and most of the further mutations occur in the mutant population (though these double-mutants are not involved in interference, as they resemble the linear regime of a mutation occurring in a population that has essentially already fixed. We note this only to distinguish from other double mutants later that will be involved in interference.) During most of that time mutants that do occur in the ancestral population, even should they have higher fitness, $s > s_0$ will not survive drift with probability $\pi(s)$, but rather with a probability that is function of its advantage relative to the new mean fitness, which has risen to more than $1 + s/2$. This effect may be considered small, since only $uN \ln(2)/s$ of the interfering mutations will occur after time $t_{\mathrm{half}}$ (specified by equation 36) because after that time the ancestral population is no longer in the majority. Such a justification of why the implicit assumption of constant fitness holds ought to offered, and the calculation of Gerrish and Lenski ought to consequently drop mutations for which it the assumption no longer holds by using $t_{\mathrm{half}}$ instead of $t_{\mathrm{fix}}$) in the calculation. Again this implies a simple correction to the above equations 41 and 42. We now question the assumption that an established mutation only fails to fix when a second mutation of better fitness establishes in the ancestral population, and fixes otherwise.

First, consider the case of the interfering mutation, B, with $s_B > s_A$ establishing in the ancestral population sometime after the original mutation A has established but well before it has fixed. If no more mutations enter, mutation B is guaranteed to fix over mutation A: both will grow in frequency until the ancestral population is eliminated, and then A will have no selective advantage while B will continue to grow until it fixes. However, the time for B to reach fixation is not given by the equations we have been considering. Those would hold only if the A population immediately went extinct as soon as B established. Instead, B will not fix at a rate determined by its selective advantage $s$, but rather as a function of $s$ and $s_0$. If A has a reasonable head-start and B has an only slightly greater advantage, its fixation time could be non-negligibly larger. The amount of time until the ancestral population vanishes will not be significantly different than the expected time for either to fix from establishment. However this alone does not comprise the fixation of B, and the time required for B to fix (or at least gain majority) over A could be considerably longer, during which interval a new mutation may arise in A, and B becomes the mutation that fails to fix. These double mutants are distinct from the example earlier in that their occurrence results in an established mutation (B) failing (or at least possibly failing) to fix. (It may be worth pointing out that such a double mutants are also distinct from those we will find in the Desai regime: these are assumed to occur only when the mutation is in the majority of the population, whereas the essential concept of the Desai regime requires these to form while not in the majority).

A second but closely analogous case is that of a mutation B occurring before the fixation (or assumption of the majority) of mutant A, but of slightly smaller fitness advantage $s_B < s_A$. Such a mutation will similarly grow with mutant A to eliminate the ancestral population, but may remain at a reasonable frequency long enough after that to receive a second mutation, hence still successfully interfering. While these effects may be negligible when considering competing mutations with very different finesses, for mutations with similar finesses the small advantage one has over the other may not be able to fix it fast enough to ignore the possibility of a new mutation occurring in one of populations.

Interestingly, Gerrish and Lenski do indeed entertain the probability that a new mutation establishes in the mutant A population before the fixation of mutant B, and claim that the effect is minimal. Unfortunately, they consider the time to fixation of mutant B as the time it would take to fix given its selective advantage over the ancestor. The selective advantage of B over mutant A could be an order of magnitude less, which would consequently require roughly an order of magnitude longer to fix (recall fixation time goes roughly proportional to $1/s$). The essential point here is that in the interference regime it matters not just which mutation has the greater fitness but by how much. Perhaps this oversight is understandable in that this is distinctly not true in the linear regime. As we found in our derivation of the fixation rate $\rho$ in equation 39, even the smallest fitness difference would suffice to guarantee the fixation of an established mutant over other competitors, it just takes longer. However, in the interference regime one can only wait so long before a new mutation enters and changes the game.

Despite corrections for the effects suggested here, it is still instructive to follow through the argument of Gerrish and Lenski in determining the final expression for the rate of fitness increase. The product of the number of mutations occurring (equation  41) and the probability a given mutation will interfere (equation 42) gives the expected interference rate, denoted $\lambda$. The probability a mutation fixes is then the probability that it survives drift ($\pi(s)$) and experiences no interfering mutations, which are Poisson distributed with mean $\lambda$. Integrating this over all $s$ against the distribution of $s$ gives the probability of fixing $P_{\mathrm{fix}}$ (that is, the probability of establishing and experiencing no interference). The rate at which mutations fix in the population ($\rho$), is as always the rate at which mutations enter times the probability with which they fix, just $NuP_{\mathrm{fix}}$ . Meanwhile the expected value $\langle s\rangle$ to fix is then given by integrating $s$ against a normalized version of the the probability that mutant with fitness $s$ fixes. The product of the expected fitness increase $\langle s\rangle$ and the expected rate of fixation $\rho$ gives the rate of fitness increase. The resulting final integral is:

$$\alpha uN \int_0^\infty s\pi(s)e^{-\lambda-\alpha s}\mathrm{d}s \tag{43}$$

which they claim can be shown to have a mathematical maximum that they tentatively refer to as a "speed limit" on the rate of evolution. I have not been able to demonstrate this, and observe that Gerrish and Lenski reword this claim in both the conclusion and discussion as "the rate of fitness increase is an increasing function of both population size and mutation rate, but is only weakly dependent on these parameters when their product is not small" [5] and continue to use only this weaker claim in a later review on the subject [7]. We conclude that this method at least demonstrates that interference results in "diminishing returns" in raising the rate of fitness increase, as more and more mutations that establish are nonetheless wasted. Another argument against the existence of such a speed limit can be made, because as $Nu$ becomes high enough, additional mutations will occur in mutant populations while they are still small, and a mutant will fix by virtue of carrying two or more beneficial mutations. This brings us to the mechanism considered by Desai *et al.*

## 3.2 The Model of Desai, Fisher and Murray

Interference of the form described by Gerrish and Lenski comes into play as soon as the expected number of mutations occurring in the ancestral population before the current mutation fixes becomes non-negligible, i.e. of order unity:

$$Nu\pi \int_{t_{\mathrm{est}}}^{t_{t_{\mathrm{half}}}} (1-p(t))\mathrm{d}t = 2Nu\ln(Ns/2) \simeq 1 \tag{44}$$

Where $\pi$ is approximated from equation 26. Under this regime, mutations enter into populations in majority status before fixation of an allele not in majority status occurs. The consideration of such mutants has been discussed in the previous section, and we shall continue to refer to interference arising out of such mutations as the Gerrish and Lenski mechanism, and thus refer to equation 44 as the Gerrish and Lenski regime.

As $Nu$ increases further, we encounter a new regime, discussed by Desai *et al.* [1]. With high enough mutation rates, it becomes possible for a mutation to occur within the mutant strain well before it reaches fixation, when:

$$Nu\pi \int_{t_{\mathrm{est}}}^{t_{\mathrm{half}}} p(t)\mathrm{d}t = 2Nu\ln(2) \simeq 1 \tag{45}$$

Once such mutations become likely, mutations within the Gerrish and Lenski regime will still occur a factor of $\ln(Ns)$ more frequently than these. However, under the simplifying assumption Desai *et* *al.* that all mutations confer the same fitness advantage $s$, the Gerrish and Lenski mechanism does not operate. In fact, as long as the mutational steps are closer than the stepsize $s$ apart, we will see that once within this regime the fitness advance is determined by establishing higher and higher order mutants which occurs faster than the time required for two similar-fitness mutants to compete for fixation. Because establishment is now occurring much faster than fixation, we are truly in a distinct regime, where dynamics are governed by receiving multiple mutations.

Compare these to our original estimate in equation 38. Here we have used $t_{\text{half}}$ in these integrals instead of $t_{\text{fix}}$. We take a moment to justify this choice. Gerrish and Lenski explicitly consider $t_{\text{fix}}$ just as we have defined it (when the mutant population reaches N-1), and Desai *et al.* use an approximation to the full fixation time of $2\ln(Ns)/s$ which obviously twice our $t_{\text{half}}$ and is approximately our value of $\ln(N^2s)/s$ when $\ln(s) \ll \ln(N)$, (which is generally true in bacterial experimental regimes.) It would also matter little in equation 44 which timescale we used, since most mutations in the ancestral population occur before the mutant reaches large frequencies. However, the value in equation 45 would differ radically (becoming $Nu\ln(N)/s$), which is not what we want since we wish to consider mutations occurring well before the population is fixed. Most of such mutations occur after $t_{\text{half}}$ and have rather different effect on dynamics (in fact, most of these are simply wasted once we enter deeply in the the Desai regime), and hence we must use $t_{\text{half}}$ here. Had we used $t_{\text{half}}$ in equation 38 our condition for no interference would become

$$\ln(sN) \ll \frac{1}{2Nu} \tag{46}$$

Which agrees with Desai *et al.* expression for the linear regime. However, they do not specify that there mechanism gains influence only one passes beyond the Gerrish and Lenski limit, (equation 44) into what we have named the Desai limit, (equation 45). Having (hopefully) clarified the separation between regimes, we now follow the reasoning of Desai *et al.* to understand how this mechanism operates.

The regime begins with the establishment of a double-mutant subpopulation before the mutant population nears fixation, a time $\tau_2$ after the single mutant subpopulation first established, occurring when the single-mutant population is large enough that it reaches frequency $p \approx 1/(Nu)$. Now, mutations in the bulk of the population (still of the ancestral type) are wasted, as they are part of a population doomed to extinction once the established double-mutants fix (though they could interfere with the fixation of single-mutants through the Gerrish-Lenski mechanism). As the double mutant population grows to the order the single mutant population had reached when the double-mutant first appeared, triple mutants emerge, occurring $\tau_3$ generations after the double-mutant first established. Desai *et al.* note that $\tau_3$ will be less than $\tau_2$ for two reasons. First, the double mutant grows faster than the single mutant, reaching the necessary frequency $p \approx 1/(Nu)$ faster than the single mutants; and second, because the triple mutant stands a better chance of surviving drift and will on average take fewer attempts and less time to do so, reaching an established size. The process continues, accelerating with each higher-level mutant, until the single mutant population nears fixation. Now the mean fitness is approximately that of the single mutant, and hence the relative finesses of each mutant are reduced by one. The triple-mutant has only a two-fold advantage now over the mean, etc.

The advancing fitness of the mean checks the advancing fitness of the nose of the fitness distribution (where the newest mutants are attempting to establish), establishing a steadystate distribution illustrated in figure 2. Meanwhile, beneficial mutations occurring in subpopulations other than at the most fit edge are lost and wasted, as these individuals have no chance of surviving against the deterministic growth of their more fit cousins.[6]

[6] It may be worth pointing out that even mutations occurring in the subpopulation directly behind leading edge that would provide their carriers with fitness identical (though perfectly identical finesses composed of different mutations

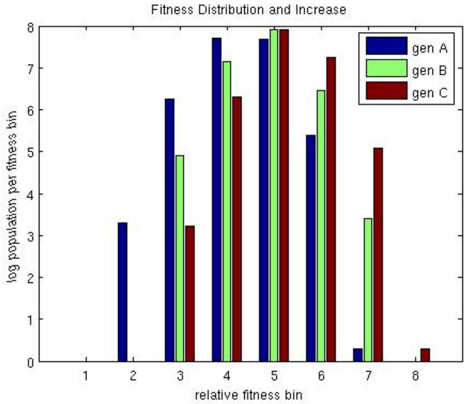


Figure 2: A distribution of fitness in the clonal interference regime. The blue distribution shows a newly-established level seven mutant subpopulation (far right, bin seven). A while later (green bars) this population has grown significantly, while the subpopulation in bin two (double mutants relative to the ancestral strain) have died off entirely. Left of the mean the subpopulations have decreased (green bars shorter than blue), while right of the mean they have increased. After another interval (red bars) the distribution has advanced as a new level-eight mutant population begins to establish. (Data taken from the simulation model described in following section)

Desai *et al.* build their model around the dynamics of this distribution we have just discussed. They then denote the most fit mutant as having accumulated $q$ mutations more than those having the mean fitness, and hence have fitness $qs$. The mean fitness increases as the population with fitness $qs$ approaches order unity (fixation), bringing the mean up to a fitness of $qs$ as it grows. By taking the fixation time to be an exponential of average rate of increase $qs/2$ (starting at rate of $qs$ and going to zero as the mutant fixes), Desai *et al.* determine time required to grow from the established population size of $1/qs$ to a large fraction of $N$ as

$$t_{\text{fix}\approx\frac{2\ln(Nqs)}{qs}} \tag{47}$$

Which we have noted to match reasonably with equation 35 where $qs$ everywhere replaces $s$. Since the population increases by fitness $qs$ in this time, the rate of fitness increase is then consequently,

$$v_m \approx \frac{(qs)^2}{2\ln(Nqs)} \tag{48}$$

This derivation is independent of the mutation rate $u$, and hence does not account for further beneficial mutations entering the population within this time interval. Should all mutations cease once the subpopulation of fitness $qs$ establishes, this would be exactly the result. Our calculation of the time requires the mean fitness increase to depend only on the increase in fitness of the subpopulation being considered, $qs$. Additional mutations offer the following corrections. Under our model, the total flux of beneficial mutations into the population raises the mean fitness by some amount (roughly $Nu$

is only possible in the simplified, single fitness step model) to the leading edge will be wasted, since the population's fitness will be no higher than if they had never occurred. Without any selective advantage over the leading edge mutant they could coexist for a time but being of lower frequency (having established later and then grown at identical rates), these mutants will likely be lost to drift.

multiplied by the time required for a mutant to pass half-way through the distribution, since this is the average number of extra beneficial mutations accumulated in the population. That time interval is roughly the fixation time we seek. Of course realistically this contribution would be a genetic load of deleterious mutations instead). Beneficial mutations in the mutant under consideration establish new populations of even higher fitness, which begin to grow exponentially and raise the mean fitness. Mutants of higher fitness than the mean but lower fitness than our mutant under consideration are also still growing in frequency and raising the mean fitness. These additional contributions to the mean fitness would slow the rate at which the mutation could fix. Further the population is depleted by mutations occurring within it, and bolstered by mutations occurring in individuals with fitness one unit of $s$ below it. (Recall that the latter would normally be considered to interfere through the Gerrish-Lenski mechanism. However we see that for identical fitness steps, and indeed as long as there finesses are closer together than the stepsize, fixation rates will be dominated instead by the number of mutations, and such Gerrish-Lenski interference would be negligible.)

A precise calculation of this speed within the interference regime will require a more careful theoretical treatment of involving the mutation rate. The expected $u$ dependence for equation 48 is left as an area for further investigation.

The correction effect can be explored by simulation as seen in figure 1. The simple theoretical prediction we have discussed predicts the time to fixation by the red curve. This simulation is run in the interference regime, and before the fixation of this mutant, a new mutation established a double-mutant subpopulation that also continued to grow, delaying the near fixation of the original mutant. (Of course the original mutant never entirely fixes in the Desai regime, because before it can do so a new mutant establishes with higher fitness that will eventually out-compete it, and the distribution continues to move steadily forward).

Meanwhile, Desai *et al.* turn their attention to the leading edge of the distribution, whose stochastic dynamics determine the fate of the rest of the distribution following behind. Here, the leading fitness bin (the "nose") of the distribution advances at rate $s/\bar{\tau}_q$, where $\bar{\tau}_q$ is the average time required for the leading edge mutant to be *established.* (Once they are established they enter the deterministic regime with the characteristic time we have already discussed). Through a complicated series of analysis not fully presented and which we will make no attempt to follow, they determine the mean time for establishment of the leading edge, $\bar{\tau}_q$ to be

$$\bar{\tau}_q = \frac{1}{(q-1)s} \ln\left[\frac{(q-1)s\sin(\pi/q)}{u\pi e^{\gamma/q}}\right] \tag{49}$$

Where $s \ll 1$, $Ns \gg 1$, and $\frac{u}{s} \ll 1$ and $\gamma = 0.5772$ is Euler's constant and $\pi$ the familiar constant, not the unfortunately represented fixation probability. Dividing $s$ by this time gives the speed at the nose.

In order to establish a steadystate distribution, these two rates must be equal. Equating $v_n$, the rate fitness increase by establishment at the nose and $v_m$ the rate of increase by fixation at the mean they determine $q$, and can (after more algebra and approximations) then express the fixation speed as a function of our three familiar parameters $u$, $s$ and $N$:

$$v \approx \frac{s^2 \ln\left[N^2 s u^{\frac{2\ln(Ns)}{\ln\left(\frac{s}{u}\right)}}\right]}{\ln^2\left[\frac{s}{u}^{\frac{2\ln(Ns)}{\ln\left(\frac{s}{u}\right)}}\right]} \tag{50}$$

Perhaps this justifies our adding these three variables one by one: together they can make quite a mess. Desai *et al.* describe the overall dependence of the the rate of evolutionary advance under the clonal interference regime more simply as:

$$\boxed{v \propto \ln N} \tag{51}$$

and

$$\boxed{v \propto \frac{1}{\ln(1/u)}} \tag{52}$$

Rather than attempt to follow such analysis in full, we turn to simulation to better understand the dynamics of this advancing fitness distribution.

# 4 Model Exploration by Simulation

Our simulation model is built around the three familiar variables $N$, $u$ and $s$. Here we will introduce $K$, the carrying capacity of the environment, which will determine the average population size, equal to $\langle N \rangle$, allowing $N$ to represent the population size of the current generation. As mentioned in our initial discussion of genetic drift, this serves as a more realistic generalization of a fixed population size, though the dynamics are changed little between the two. The simulation begins with a clonal population of equal fitness. Each generation, individuals first mutate with probability $u$ and then die with probability determined by their fitness and the carrying capacity. Those that survive divide, providing the population for the next generation.

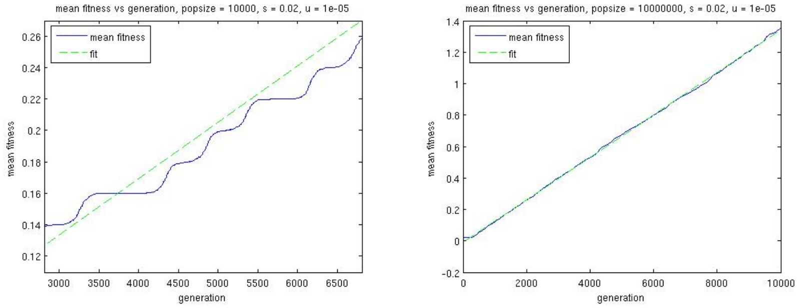
The model is constructed such that populations can be accurately treated probabilistically rather than having to simulate each individual member of the population, making very large values of $N \sim 10^{18}$ accessible. The population is divided only by fitness, and all individuals of the same fitness are lumped together in one bin. Hence the model does not discriminate between individuals that have entered the bin ahead of them on attaining a new mutation and individuals born into that bin, since selection is blind to the distinction. The number of individuals receiving beneficial mutations in a given generation is simply a binomial random variable of size $N_q$ (the population of the fitness bin currently being considered) and probability $u$, and is computed using a condensed table look-up method. The probability of surviving to division is given by

$$\frac{1}{2}\frac{K}{N}\frac{e^{w_i}}{\langle e^{w_i} \rangle} \tag{53}$$

Where $N$ is the current population size and $w_i$ is the fitness of the individual in question ($w_i$ is simply $s$ times the number of beneficial mutations the individual has received). As in the mutational step, this step can be applied statistically as a binomial random number of individuals in each fitness bin that will survive to the next generation.

Many alternatives to this basic model are possible, by permuting the three steps of mutation, death, and replication and choosing whether to include the effect of fitness in either the death step (increasing survival probability) or the replication step (increasing offspring numbers). The concept is the same as either placement allows fitness to increase the mutant's frequency in the next generation, which is the essential feature. In the model described above the effect of fitness is applied in the death process, which occurs in the simulation before division. This results in members of the current generation having either zero or two offspring in the next generation. This both simplifies the model and accelerates the rate at which effects can propagate through the population.

The simulation is run at fixed values of $u$ and $s$ for a fixed number of generations over a range of population sizes from $10^3$ to $10^{18}$. For each population size, the mean fitness is plotted against time in generations and fit to a linear regression whose slope determines the rate of fitness increase



(a)

(b)

Figure 3: Linear fits to the mean fitness increase. 3(a) shows a section of the fit illustrating the mean fitness increasing in a distinctive stepwise fashion found in the linear regime (blue solid line). As with all other runs, a linear regression (green dashed line) is fit over the data. For larger values of $N$ 3(b) we are easily within the Desai interference regime, $1 \ll 2Nu\ln(2)$. After an initial transient lasting less than 500 generations and hence only just visible, the fitness increases in a steady, linear fashion, closely matching the linear fit.

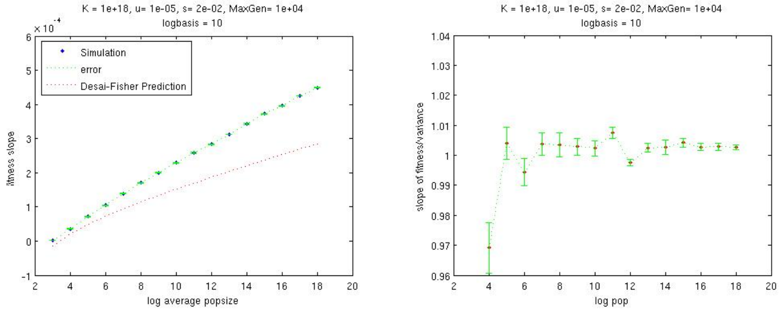
$v$ under the conditions of the run, as seen in figures 3(a) and 3(b). Since the initial population is clonal (all individuals begin with equal fitness) the simulation experiences a short relaxation period to establish the steadystate distribution. This relaxation period is seen as a transient at the beginning of each graph, and is not included in the linear regression. The variance is averaged over the run, again excluding the initial transient. The simulation is run for $10^4$ generations, though shorter periods could probably be used for larger populations since mutations enter faster than in small populations. This time also allows the removal of the initial transient period conservatively without discarding a significant portion of the data. The simulation is run nine times for each set of conditions and the results averaged, providing some measure of the associated uncertainty. Including these repeated runs, the simulation can require around a day to run.[7]

For the relatively small population sizes the regime is approximately linear, and fitness increases in the expected stepwise fashion, as seen in figure 3(a). As a quick check of the regime, we note that in this case $N = 10^4$, $u = 10^{-5}$, and $s = 0.02$, and hence we have $2Nu\ln(Ns) \approx 1$ and we find ourselves at the edge of the interference regime. Figure 3(b) shows the mean fitness increase occurring at an effectively constant rate (excluding the initial transient), as predicted in the Desai regime.

We ultimately desire to confirm that within the interference regime $1 \lesssim 2Nu\ln(Ns)$ the rate of fitness increase only rises with the logarithm of the population size. Plotting the rate of fitness increase against $\ln(N)$ in figure 4(a) we observe that our simulation appears to confirm this result. Also plotted is the final equation derived by Desai *et al.* given by equation 50. This result is also shown in the plot in figure 4(a). This predicted value is notably lower than our calculated value, though supports the same general trend. It differs most at extremely large $N$, which may be anticipated by Desai *et al.* who note the equation fails for $v$ less than $s^2$, while $v$ is at the order of $s^2$ near the upper end of our population values, though the inequality will not actually fail for another seven orders of magnitude given these parameters. Another reason for this difference may be explained by the calculation itself,

[7] All simulations were run on Princeton University's sixtyfour bit server (named *sixtyfour*) in MATLAB® version 7.14 and statistics of the processor in place are available on request.

for instance ignoring the mutation rate in the calculation of the deterministic mean fitness increase. Possible dependence on the explicit model used for replication, mutation and selection can also not be ruled out, though it is not obvious in which direction such corrections would move the predicted value, and a precise explanation of the deviation would require further study.



(a)

(b)

Figure 4: The rate of fitness increase rises only logarithmically with increasing population size (blue points, figure 4(a)). The rate predicted by Desai *et al.* (dotted red line) follows this trend but falls short of the simulated rates. The simulated rates are very reproducible, as suggested by the very narrow error bars (green bars). The plot to the right 4(b) shows the ratio of rate of fitness increase $v$ over the variance, which agrees very well with the predictions of R.A. Fisher's Fundamental Theorem for all points except that at low population sizes which appear the most stochastic.

As a check on the model, the simulation agrees well with R. A. Fisher's Fundamental Theorem of Natural Selection: that the genic variance equals the rate of change of the mean fitness, as seen for simulation runs over various population sizes in figure 4(b)

Our simulation agrees very well with this result, except for a slight deviation for small populations. This deviation can be explained at least in part by the more stochastic nature of such populations. As the rate of fitness for such populations involves fitting a line to a trend that is essentially a step-function as seen in figure 3(a), the determination of the rate fitness increase will necessarily be less accurate for such populations. For the nine runs in the lowest population seen in figure 4(b), the ratio of the speed to the variance was sometimes very close to, sometimes above, and sometimes well below unity. More runs would be necessary to confirm the deviation from the theorem under these parameters. Indeed, that the theorem holds as accurately as it does under these conditions is somewhat surprising, considering that it was not derived for an actively mutating population or even an asexual one. Indeed, it often claimed that this result holds only for single-locus evolution in randomly mating populations, though this is not accurate (see Ewens 2004 for discussion, [3]). While the result has been generalized for non-randomly mating populations and multilocus theory, it is not clear that it should hold for asexual populations in the clonal interference regime, or what order corrections might be necessary within this regime. For now we can do no more than call attention to explaining the accuracy of this prediction within this regime as another area that would merit further investigation.

## 5 Conclusions

We have derived the rate of fitness increase explicitly in the linear regime (equation 40), and specified the conditions under which the linear regime holds (equation 38). When this linear regime no longer holds, we separate the interference into two regimes, one where the Gerrish-Lenski mechanism first becomes applicable (44), and the other where the Desai mechanism comes into play (equation 45). We have derived the probability of a mutation surviving drift (equation 26), which in turn let us calculate the population size a mutation must obtain to have effectively escaped elimination through genetic drift (equation 29). We have derived the timescales for genetic drift (equation 7) and for selective fixation (equations 35 and 36), and argued why the latter is not immediately applicable within interference regimes. This error appears to be made by both models of clonal interference examined here. We have examined two papers on clonal interference, identifying two distinct mechanisms for the phenomenon, and offering modifications where our derivations disagree with the results presented in the papers. Finally, we have outlined and run a simple simulation of the phenomenon over values clearly within the Desai interfence regime, and compared results. Our simulation confirms the trend predicted by Desai *et al.* but not the precise equation. We have also observed R.A. Fisher's Fundamental Theorem to hold very closely for nearly all simulation runs. With this, we have just scratched the surface of investigation into this phenomenon of clonal interference, and hence we conclude with an outline of further investigation.

## 6 Topics for Further Investigation

This subject is rich with areas for further investigation, both in regards to the purely theoretical treatment and in exploring simulations of evolution within the interference regimes. Many such possibilities have been noted in the text as they arise, and a list is given here as a primitive roadmap to continued investigation. Of course experimental laboratory work offers a third avenue of investigation, but remains beyond the scope of this paper and thus is not considered here.

1. Theory-based topics

   (a) Corrections to the Gerrish and Lenski regime. Though we made some corrections all ready to some of the initial calculations, the model could be further modified to account for the correct expected fixation time due to the competition between beneficial mutations, which in turn would allow a more accurate treatment of the effect of double mutants.

   (b) Corrections to the Desai *et al.* regime. For instance, determining the effect of mutation rate on the rate of fitness increase for the mean of the distribution. Reproducing my similar or novel means result for the rate of fitness increase at the nose may also be valuable. Further, integration of this mechanism with the effects of the Gerrish and Lenski mechanism would prove interesting, particularly involving the more complicated treatment of nonidentical fitness mutations. While convenient, the assumption of a single fitness step allows for two different mutants to have identical fitness is nonphysical for beneficial mutations, and has an inherently different dynamic (that of neutral mutations) relative to eachother than do mutations of very similar but non-identical fitness values. Where this approach becomes intractable, simulations may prove the most effective way forward.

   (c) Understanding of the Fundamental theorem result. This result ought to be derived explicitly for the situation considered here, determining the effect of a steady supply of mutations. The theorem must also be considered with the linear regime, where the rate of fitness increase in the long-term is fixed to a constant rate but the fitness of the population does not assume a steadystate distribution, but is rather perturbed by each mutation and then returns to the zero-variance state.

   (d) The effects of deleterious mutations. Deleterious mutations could be included in the theoretical treatment as well, to see if the general dependence on $Nu$ predicted by clonal interference models still holds.

   (e) Clonal interference and the evolutionary advantage of sex. Since recombination offers a way to avoid the waste of beneficial mutations through clonal interference, some have argued that this suggests an evolutionary advantage to sex, while others such as Maynard Smith has refuted such arguments, and an analytic exploration of the question might be addressed in light of these models. See Gerrish and Lenksi for further discussion and references, [5].

2. Simulation-based topics

   (a) Using the simple model presented here several further aspects of the interference phenomenon could be profitably explored:

      i. The time required for an established mutation to reach near fixation raised several difficult questions in the theoretical analysis. This time could be computed directly from the simulation data for each mutation as it establishes, which would offer extensive data on the time required and its dependence on various population sizes.

      ii. Dependence of fixation time on other parameters, such as different mutation rates in a fixed population could also easily be explored in the simulation already developed here, given enough time to run such as simulation.

      iii. The variance plots show small-scale and apparently random oscillations that we have simply averaged out. It would be nice to trace the origins of these oscillations to oscillations in the leading edge. Perhaps there frequency can be mapped to the frequency at which mutations enter the leading edge, which ought to be the what drives the dynamics of the distribution.

   (b) Beyond the simple model: certain modifications to the simulation presented here would also be useful for further investigation. Such additions might include:

      i. incoperating the effects of variable selective advantages, such as the exponential distribution considered under the Gerrish and Lenski regime. This could allow explicit modeling of the Gerrish-Lenski regime and would serve as a check against their calculations.

      ii. Deleterious mutations could easily be added to the simple model simulated here, and its effect on the results presented could then be examined. Such an approach could at least provide a hint as to the relevance of such mutations in the overall behavior before working out a more explicit theoretical model.

      iii. The dynamics of the model could be better studied if it were possible to distinguish the lineages of individual mutations. This would, for instance, allow discrimination between two mutants of equal fitness, one of which has grown up from an established population and another which is the result of a chance mutation in the fitness bin immediately below it. In this manner the effect of mutations entering the bin through mutation could be separated from the growth by replication within that bin.

      iv. For results under small populations and short total generations, it may be possible to simulate all individuals and compare the results to the statistical treatment of populations given by the model presented here.

# References

[1] Michael M. Desai, Daniel S. Fisher, and Andrew W. Murray. The speed of evolution and maintenance of variation in asexual populations. Unpublished work, 2004.

> This article is considered at length in the body of the paper, see the entitled section. More recent versions of this paper have come to my attention to late to receive careful attention: a short summary of the experimental results submitted to Science, "The Speed of Evolution and Maintenance of Variation in Asexual Populations", and a much longer and more theoretical treatment submitted to Nature entitled "Beneficial Mutation-Selection Balance and the Effect of Linkage on Positive Selection." As these papers are not considered in the treatment presented here, I only acknowledge their existence as a reference for further investigation.

[2] W. J. Ewens. *Population Genetics.* Methuen and Co LTD, London, 1969.

> Both earlier and less expansive, this seems to have been largely replaced by Ewens' more expansive text, "Mathematical Population Genetics." However, this is the reference cited by Phillip Gerrish and Richard Lenski in their 1998 article on clonal interference for the probability of a beneficial mutation surviving drift, though their result does not directly follow from this text, which only gives the classical "2s" result. Gale may be consulted for an elaboration of the Branching Process Theory presented in this section and used by Gerrish and Lenksi.

[3] W. J. Ewens. *Mathematical Population Genetics.* Springer, New York, second edition, 2004.

> This text is another standard in the discipline. Written for "graduate students and researchers in Applied Mathematics and Theoretical Biology," it is easily one of the most thorough and rigorous texts of the readily available texts on the topic. Though dense and sometimes challenging to follow, derivations are both careful and rigorous. While the second edition is essentially a reprinting of the 1979 edition as Volume 27 of the Interdisciplinary Applied Mathematics series and the mathematics is essentially unchanged, the discussion has been altered, in some places significantly. For instance, the 1979 interpretation of the Fundamental Theorem of Natural Selection is considered to be incorrect, (though it remains the text-book standard), and a completely revised interpretation is given. As in the 1979 edition, the notation use is not consistant between sections. This text has been recommended to me by several individuals in the field including Leonid Kruglyak, David Stern and Simon Levin, at Princeton University.

[4] J. S. Gale. *Theoretical Population Genetics.* Unwin Hyman, London, 2004.

> A much less well known text, Gale provides a very careful discussion on many of the classical results, introducing many of the common mathematical tools used along the way. The writing style is very clear, though some of the notation appears to be nonstandard. The notation used appears consistent throughout, but it can be difficult to follow a result given from somewhere in the middle without having followed the previous results to at least to understand the symbols used. While fewer topics are addressed than in most references on the subject, the treatment here is thorough both in the mathematical background and the discussion. Exercises are also included at the end of each chapter.

[5] Philip J. Gerrish and Richard E. Lenski. The fate of competing beneficial mutations in an asexual population. *Genetica*, 102/103:127–144, 1998.

> This article is the current standard theoretical treatment of the clonal interference problem. Detailed comments are included in the body of this paper.

[6] John H. Gillespie. *Population Genetics.* John Hopkin University Press, Baltimore, second edition, 2004.

> Gillespie offers an excellent and easily accessible introduction to the basics of population genetics. Along with the more rigorous Ewens, Gillespie appears to be standard reference in the field. The second edition is certainly worthwhile, as not only are some new sections and a new chapter added, but several results have been elucidated and stochastic treatment expanded. Several interesting derivations, such as the diffusion approximation derivation for the survival of beneficial mutant, are also omitted in the first edition. The text remains concise, easy to follow, and introduces most of the basic topics and classic results found in higher level treatments. The book also includes exercises interspersed within each chapter, along with answers at the end, and includes several computer simulation exercises as well. This text has been recommended to me by several individuals in the field including Leonid Kruglyak and David Stern

[7] Richard E. Lenski and Santiago F. Elena. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nature Reviews Genetics*, 4:457–469, 2003.

> This article reviews experimental, theoretical, and modeling work regarding the evolution of asexual organisms. Its theoretical claims regarding clonal interference are largely derived from an article by Gerrish and Lenski, 1998.

[8] Wen-Hsiung Li, editor. *Stochastic Models in Population Genetics.* Dowden, Hutchingon and Ross, Stoudsburg, PA, 1977.

> An excellent collection of many of the seminal papers of Fisher, Wright, Haldane, Kimura, Ewens, and others, this volume includes some of Kimura's papers not found in the other volume. While less recent than the other volumes, (and consequently lacking in more recent topics such as the Neutral Theory), this collection nonetheless provides an excellent feel for how the field has developed and the flavor of its key advances. For anyone desiring to consult the original literature in classical population genetics this volume is an excellent reference.

[9] John Maynard Smith. *Evolutionary Genetics.* Oxford University Press, Oxford, second edition, 1999.

> A simple introduction to the field in its broad sense, containing much greater breadth of topics but without a thorough treatment of many classical results of population genetics. The text is well written, easy to follow, and includes exercises at the end of each chapter.

[10] Naoyuki Takahata, editor. *Population Genetics, Molecular Evolution And the Neutral Theory: Selected Papers of Motoo Kimura.* The University of Chicago Press, Chicago, 1994.

> A nice but not complete selection of the papers of Motoo Kimura. Forward by James Crow. The 1964 paper on"Diffusion Models in Population Gentics" and the 1955

> “Stochastic Processes and Distribution of Gene Frequencies under Natural Selection” are perhaps the most relevant here. Kimura’s papers are often cited but can be very challenging, while most of his well-known results and methods of derivation are found in a more accessible presentation in textbooks on the subject.

# A Code

The Matlab® 7.14 code used for the simulation is provided here for reference and as a starting point to explore the simulation results further.

```
clear
t = clock;
maxReps = 9;
for rep = 1:maxReps

    maxPower = 16;
    runs = zeros(maxPower,2);
    for power = 1:maxPower
       clear history B gen

       logbasis = 10;
       mylog = @(b) log10(b)
       K = 10^(2+power);

       Npop(power) = K;
       u = 10^-5;
       s = 2e-2;
       MaxGen = 1e4;
       relax = 500;

       N = K;
       B(1) = K;
       for gen = 1:MaxGen
          if rem(gen, MaxGen/50) == 0
             fprintf('=')
          end
          nonzeropops = nonzeros(B);
          NewMutants = zeros(1,length(nonzeropops))';
          for i = 1:length(nonzeropops)
             NewMutants(i) = randraw('binom', [nonzeropops(i), u], 1);
             if NewMutants(i) < 0
                NewMutants(i) = 0;
             end
          end
          Mutants = zeros(1, length(B) );  %initialize to full length
          Mutants(find(B)) = NewMutants; %number leaving at each index
          B = B - Mutants; %Remove leaving
          B = [B,0]; %make room for best mutants
```

```
   Mutants = [0, Mutants]; %Mutants entering
   B = B + Mutants; %new mutants entering bins from below
   meanfitness = (exp(((1:length(B))-1)*s)*B')/N;

   dummy = find(B);
   B = B(1:dummy(end));
   nonzerobins = find(B);
   for i = 1:length(nonzerobins)
      j = nonzerobins(i);
         P2(j) = (1/2)*(exp((j-1)*s)/meanfitness)*(K/N);
         B(j) = 2*randraw('binom', [B(j), P2(j)], 1);
   end
   flags(gen,:) = [max(P2),min(P2)];
   N = sum(B);
   popsize(gen) = N;
   history(gen, 1:length(B))  = B;
   best(gen) = s*max(find(B));
end
 width = size(history,2);
 clear novel;
 clear first;
 oldleader = zeros(width, 1);
 for i = 3:width
         novel = find(history(:,i));
         if numel(novel) ~= 0
                 first = novel(1);
                 oldleader(i) = history(first,i-1);
         end
 end
 max_bin(power) = max(oldleader*u);
 ave_bin(power) = mean(oldleader*u);
 error_bin(power) = std(oldleader*u);

fprintf('\n');

record.(genvarname(['history',num2str(power)])) = history;

for gen = 1:MaxGen
   a = nonzeros(history(gen, :)); b = s*find(history(gen,:));
   meanFitness(gen) = b*(a/sum(a));
   v = (b-meanFitness(gen)).^2 .* (a/sum(a))';
   varFitness(gen) = sum(v);
end
bestslope = polyfit(relax:MaxGen, best(relax:MaxGen), 1);
dwdt = polyfit(relax:MaxGen, meanFitness(relax:MaxGen), 1);
dwdtline = @(x) dwdt(1)*(x) + dwdt(2);
runs(power, 1) = dwdt(1);
runs(power, 2) = mean(varFitness(relax:MaxGen));
```

```
    end
    rep_runs(:,:,rep) = runs;
    repeated_ave_bin(:, rep) = ave_bin;
end
mean_runs = mean(rep_runs, 3);
ste_runs = ((mean(rep_runs.^2, 3) - mean_runs.^2).^.5)/sqrt(maxReps-1);
rep_ratio = rep_runs(:,1,:)./rep_runs(:,2,:);

mean_ratio = mean(rep_ratio, 3);
error = ( ( mean(rep_ratio.^2, 3) - mean_ratio.^2)/(maxReps-1) ).^.5;


fprintf(’Running Time = %.0f sec\n’, etime(clock, t))
save run
```