OPEN ACCESS Freely available online

PLoS GENETICS

The Population Genetics of dN/dS

Sergey Kryazhimskiy[1], **Joshua B. Plotkin**[1,2]*

1 Biology Department, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **2** Program in Applied Mathematics and Computational Science, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

Abstract

Evolutionary pressures on proteins are often quantified by the ratio of substitution rates at non-synonymous and synonymous sites. The dN/dS ratio was originally developed for application to distantly diverged sequences, the differences among which represent substitutions that have fixed along independent lineages. Nevertheless, the dN/dS measure is often applied to sequences sampled from a single population, the differences among which represent segregating polymorphisms. Here, we study the expected dN/dS ratio for samples drawn from a single population under selection, and we find that in this context, dN/dS is relatively insensitive to the selection coefficient. Moreover, the hallmark signature of positive selection over divergent lineages, dN/dS>1, is violated within a population. For population samples, the relationship between selection and dN/dS does not follow a monotonic function, and so it may be impossible to infer selection pressures from dN/dS. These results have significant implications for the interpretation of dN/dS measurements among population-genetic samples.

Citation: Kryazhimskiy S, Plotkin JB (2008) The Population Genetics of dN/dS. PLoS Genet 4(12): e1000304. doi:10.1371/journal.pgen.1000304

Editor: Takashi Gojobori, National Institute of Genetics, Japan

Received September 11, 2008; **Accepted** November 10, 2008; **Published** December 12, 2008

Funding: SK and JBP were funded by a grant from the James S. McDonnell Foundation. JBP gratefully acknowledges support by the Burroughs Wellcome Fund and the Defense Advanced Research Projects Agency "Fun Bio" Program (HR0011-05-1-0057).

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jplotkin@sas.upenn.edu

Introduction

The identification of genetic loci undergoing adaptation is a central project of evolutionary biology. With the advent of sequencing technologies, a variety of statistical tests have been developed to quantify selection pressures acting on protein-coding regions. Among these, the dN/dS ratio is one of the most widely used, owing in part to its simplicity and robustness. This measure quantifies selection pressures by comparing the rate of substitutions at silent sites (dS), which are presumed neutral, to the rate of substitutions at non-silent sites (dN), which possibly experience selection. The ratio dN/dS is expected to exceed unity only if natural selection promotes changes in the protein sequence; whereas a ratio less than unity is expected only if natural selection suppresses protein changes [1,2]. This intuitive interpretation of dN/dS is supported by theoretical work on the relationship between the dN/dS statistic and the underlying selection pressure in a Wright-Fisher model [3].

The dN/dS ratio was originally developed for the analysis of genetic sequences from divergent species [1,4,5], the differences amoung which represent fixation events along independent lineages. Theoretical work on the relationship between dN/dS and selection likewise assumes that sequences are sampled from independent, divergent species [3], as do computer packages used to estimate dN/dS from data [6,7]. Nonetheless, the dN/dS ratio test is frequently applied to data that may represent samples from a single population, particularly in the case of microbes (e.g. [3,8–18]). In such cases, the differences between sequences do not represent fixation events along independent lineages, but rather polymorphisms segregating in a single population. It is important, therefore, to understand the relationship between selection pressures and the dN/dS statistic for samples from a single population.

Here we analyze the population genetics of dN/dS. We find that the relationship between the selection pressure and dN/dS is qualitatively different for samples drawn from a single population compared to sampled from divergent lineages. As a result, standard tests for selection based on dN/dS are extremely sensitive to violation of the assumption of divergent lineages. We show that the expected dN/dS ratio within a population is relatively insensitive to selection pressure—a result which helps to explain a body of empirical observations about microbial populations. Moreover, we show that the hallmark signature of positive selection across divergent lineages, dN/dS>1, does not hold within population: strong positive selection is expected to produce dN/dS<1 among population samples. As a result, when applied to intra-specific samples, the standard interpretation of dN/dS is unjustified and may lead to surprising conclusions. This point is illustrated by two recent studies that report dN/dS ratios near 1 among strains of *Salmonella enetrica* serovar Typhi [17,18], and conclude that genetic drift dominates the bacterium's evolution. This conclusion is surprising in light of the large population size of the bacterium (N_e estimated to be on the order of 10^5) and strong selective advantages of antibiotic-resistance mutations [17]. However, our analysis shows that dN/dS values obtained from closely related isolates may be near 1 under both strong positive selection or moderate negative selection, and so parts of the *Salmonella* Typhi genome may well be evolving under considerable selection pressure.

Our presentation begins with a review of the theory underlying the interpretation of dN/dS across divergent lineages. We then develop the appropriate theory for studying selection and dN/dS within a single population. We compare our theoretical expectations to Monte Carlo simulations based on the Wright-Fisher model. We conclude with a discussion of practical implications.

Author Summary

Since the time of Darwin, biologists have worked to identify instances of evolutionary adaptation. At the molecular scale, it is understood that adaptation should induce more genetic changes at amino acid altering sites in the genome, compared to amino acid–preserving sites. The ratio of substitution rates at such sites, denoted dN/dS, is therefore commonly used to detect proteins undergoing adaptation. This test was originally developed for application to distantly diverged genetic sequences, the differences among which represent substitutions along independent evolutionary lineages. Nonetheless, the dN/dS statistics are also frequently applied to genetic sequences sampled from a single population, the differences among which represent transient polymorphisms, not substitutions. Here, we show that the behavior of the dN/dS statistic is very different in these two cases. In particular, when applied to sequences from a single population, the dN/dS ratio is relatively insensitive to the strength of natural selection, and the anticipated signature of adaptive evolution, dN/dS>1, is violated. These results have implications for the interpretation of genetic variation sampled from a population. In particular, these results suggest that microbes may experience substantially stronger selective forces than previously thought.

Results

Time-Scales of Adaptation

There are at least two time-scales on which to investigate adaptive evolution: short time-scales, which apply to genetic variation segregating within a population of conspecifics; and long, or evolutionary, time-scales, which apply when comparing the genomes of divergent species.

Over short time-scales, natural selection at a genetic locus may be inferred by inspecting sequences sampled from a population. Polymorphism data are typically compared to expectations under a neutral null model, such as the Wright-Fisher model that forms the basis of Kingman's coalescent [19] and all coalescent-based tests of neutrality [20–22]. Alternatively, polymorphism data can be compared to expectations under a Wright-Fisher model that incorporates selection—an approach adopted by the Poisson Random Field method of inferring selection coefficients [23,24]. Under both of these approaches, the sequences under analysis share a common ancestor within the past $O(N)$ generations, where N is the population size. Such investigations inform our understanding of the forces that shape genetic variation within a population.

Over long time-scales, by contrast, natural selection is often quantified by comparing orthologous gene sequences from divergent species. In this context, each species is associated with a single representative genetic sequence, and intraspecific polymorphisms are ignored [4]. Instead, the focus is on the rate of substitutions along divergent lineages—i.e. the rate at which mutations arise and subsequently fix. Such investigations inform our understanding of the processes that shape the similarities and differences between the (stereotypical) genomes of divergent species.

Over long time-scales, the dN/dS ratio is an extremely popular measure of adaptive evolution in protein-coding sequences. This measure quantifies selection pressures by comparing the rate of substitutions at silent sites (dS), which are presumed neutral, to the rate of substitutions at non-silent sites (dN), which possibly experience selection. In practice, the dN/dS ratio is commonly estimated from data using, for example, the PAML computer package [7]. Under this approach, the substitution process at a site is described by a continuous-time Markov chain with 61 possible states, corresponding to the 61 sense codons. The instantaneous rate of change from codon i to codon j depends principally on the parameter ω, defined as the relative rate of non-silent versus silent substitutions [2].

The Markov-chain model underlying PAML's calculation of dN/dS explicitly ignores polymorphisms segregating within a population; instead, it represents each divergent species as a single sequence. Furthermore, the Markov-chain model does not describe any details of the process by which a mutation enters a population, changes in frequency, and eventually fixes. Instead, fixation events occur instantaneously in the model, and transient polymorphisms within each divergent population are ignored. These simplifying assumptions are perfectly reasonable when studying substitution rates between long divergent species (e.g. [4]). Over the time-scales of such divergence substitution events are effectively instantaneous.

Given a data set of diverged sequences, and assuming (or simultaneously inferring) their phylogenetic relationship, PAML estimates the parameter ω by maximum likelihood. The likelihood function is derived from the Markov chain, assuming that the substitution process at one site is independent of processes at all other sites. It is critical to emphasize that, by definition, ω describes the relative rate of selected versus neutral *fixation* events. Therefore, it makes sense to estimate ω from a data set of diverged sequences, the differences between which represent fixed substitutions that have accrued along independent branches. But it is not appropriate to estimate ω from a set of conspecific sequences sampled from a single population, because the differences between such sequences represent segregating polymorphisms as opposed to fixed substitutions.

Theory

The Relationship between Selection and dN/dS over Long Time-Scales. Although originally formulated without reference to population genetics *per se*, Yang's Markov-chain model of the substitution process at a site can be derived as an appropriate long-time limit of an underlying Wright-Fisher population process [3]. Such a derivation makes two essential assumptions: (1) sites are independent and thus non-interfering; and (2) there are never more than two alleles segregating in a population at a single nucleotide site. The former assumption, of site independence, is shared by most population-genetic models that incorporate selection, such as the Poisson Random Field model. The latter assumption is justified provided that the population-scaled mutation rate is small enough, so that one allelic variant at a site will always fix or go extinct before another allelic variant is introduced. Under these assumptions, the rate of fixation of new mutations with selection coefficient s is given simply by the product of the population-scaled mutation rate and the probability of fixation [3]:

$$\mu N \frac{2s}{1-e^{-2Ns}}. \quad (1)$$

Rates of this form are used as the instantaneous transition rates in the Markov-chain model of substitutions. As a result, if silent substitutions are assumed neutral and all non-silent mutations experience selection coefficient s, then the expected ratio of their rates, ω, is given by [3]

$$\omega(\gamma) = \frac{2\gamma}{1-e^{-2\gamma}}. \quad (2)$$

where γ is defined as the scaled selection coefficient Ns.

Equation (2) provides an important link between ω, the ratio of substitution rates along independent lineages, and γ, the

underlying selection coefficient in a Wright-Fisher model. This equation was derived using Kimura's expression for the probability that a new mutation will fix in a population, under a Wright-Fisher model. This derivation is appropriate, because dN/dS is defined as the ratio of fixation rates along independent lineages. We can therefore use Equation (2) in the context of divergent sequences, the differences between which represent fixation events. In particular, Equation (2) provides rigorous meaning to the statement that dN/dS is expected to exceed unity only when there is positive selection to promote non-silent changes: according to Equation (2), ω exceeds unity only if γ is positive, and ω is less than unity only if γ is negative.

The Population Genetics of dN/dS. Researchers often compute a dN/dS value when comparing conspecific sequences, whose differences reflect polymorphisms segregating within a population (e.g. [3,8–18]). Equation (2) does not apply to such sequences, because differences among such sequences do not represent fixation events along independent lineages. How, then, are we to interpret dN/dS values measured from intraspecific data? What is the relationship between selection and dN/dS values computed for sequences sampled from a population?

To address this question, we must understand the behavior of the dN/dS statistic within a single population over a relatively short time-scale—i.e. the population genetics of dN/dS. In this context, dN and dS represent, respectively, the number of *non-silent mutations* (as opposed to fixations) per non-silent site and the number of *silent mutations* (as opposed to fixations) per silent site, along the coalescent between individuals sampled from the population.

In principle, calculating these quantities requires knowing the expected coalescent time between sampled individuals. Since the general expression for the coalescent time in the presence of selection is not known, we approximate dN and dS by the number of *differences* between two sampled individuals, at non-silent and silent sites respectively. (While the number of mutations along the coalescent between two individuals can be any integer, the number of differences can be only 0 or 1, depending upon whether the two individuals share the same nucleotide at the focal site.) We operate under the same two simplifying assumptions that Nielsen & Yang used in their analysis of dN/dS and selection [3]: (1) sites are assumed independent and non-interacting; and (2) no more than two mutations are assumed to segregate in the population at a single site. The latter approximation will be accurate provided two individuals are typically separated by at most one mutation along their coalescent—i.e. provided that $\theta = 2N\mu \ll 1$. This approximation is justified for most known biological populations, because θ per site is typically less than unity.

In order to calculate the expected number of differences between two sampled individuals we utilize the stationary allele frequency distribution at a site. If Φ denotes the stationary frequency distribution for polymorphisms that arise at rate μ and experience selection pressure s, then we may calculate the expected number of differences per site, denoted D:

$$D(\gamma,\theta) = \int_0^1 2x(1-x)\Phi(x|\gamma,\theta)dx \qquad (3)$$

Here γ denotes the product Ns, and θ denotes $2N\mu$.

We use diffusion theory to derive an expression for the stationary frequency distribution of polymorphisms at a site, Φ. In the case of recurrent mutation between two alleles with fixed fitnesses 1 and $1+s$, the stationary distribution has been solved classically using a zero-flux condition [25,26]. However, the model of selection analyzed by Yang and other authors (e.g. [3,4,27–35]) in the context of dN/dS is qualitatively different from the classic model of two alleles under recurrent mutation [25].

Strictly speaking, Yang's model of selection is a special case of an infinite-sites model under which subsequent mutations each provide an additional selective advantage (or disadvantage) s. In general, such models are extremely complicated because multiple mutant linages compete with each other [36–41]. However, when the mutation rate is small enough, at most two genotypes segregate in the population at any given time, and so the allele frequency dynamics can be described by a simple two-allele Wright-Fisher model. In this limit, the population is monomorphic for the resident allele until a mutant appears. Each mutant has the same selective advantage (or disadvantage) s over the resident type. The mutant is either lost or fixed before the next mutant type arises. If the mutant fixes, it becomes the new resident type, and a subsequent mutation will experience the same selective advantage (disadvantage) s over the new resident type. This is the model of positive (negative) selection *sensu* Yang [4]. Such a model provides a convenient description of continual positive (or negative) selection at a site, and so we call it the continual selection model.

In the Methods section we derive an expression for the stationary allele frequency distribution under the model of continual selection. The solution is derived by diffusion theory using a constant but non-zero flux condition [42,43], and it deviates from the classical stationary distribution of Wright [26]. The solution for Φ is given by

$$\Phi(x|\gamma,\theta) = Cx^{\theta-1}(1-x)^{\theta-1}e^{2\gamma x}\int_x^1 \xi^{-\theta}(1-\xi)^{-\theta}e^{-2\gamma\xi}d\xi \qquad (4)$$

where C is chosen so that $\int_0^1 \Phi(x|\gamma,\theta)dx = 1$ and $0<\theta<1$.

Equations (3) and (4) provide an analytic approximation for the expected dN/dS ratio between sequences sampled from a single population, which we denote ω_{pop}:

$$\omega_{\text{pop}}(\gamma,\theta) \approx \frac{D(\gamma,\theta)}{D(0,\theta)} \qquad (5)$$

This equation is the single-population analogue of the relationship between selection and dN/dS across long divergent lineages (Equation 2). Note that over long time-scales ω depends only on γ, whereas within a population ω_{pop} depends on both γ and θ.

Comparison of dN/dS over Long and Short Time-Scales

Across divergent lineages there is a simple monotonic relationship between the selection coefficient, γ, and the expected dN/dS ratio, ω (Figure 1). A dN/dS ratio less than unity occurs only under negative selection; and a dN/dS ratio greater than unity occurs only under positive selection. Moreover, the dN/dS ratio is very sensitive to the selection coefficient: for γ less than -4, the expected dN/dS ratio is near zero (less than 0.01); and the dN/dS ratio climbs very rapidly for γ positive.

Within a single population, however, the relationship between selection and dN/dS is markedly different (Figure 1). In the case of negative selection, for example, the expected dN/dS ratio is relatively insensitive to changes in γ. Selective constraints that induce a very low dN/dS value when comparing divergent lineages will produce a less extreme dN/dS value when comparing conspecific samples. For example, very strong negative selection (e.g. $\gamma = -10$) produces an expected dN/dS ratio near zero when comparing divergent lineages, but it produces dN/dS near 0.1 when comparing individuals from a single population. Therefore, the interpretation of an observed dN/dS ratio near 0.1, which is

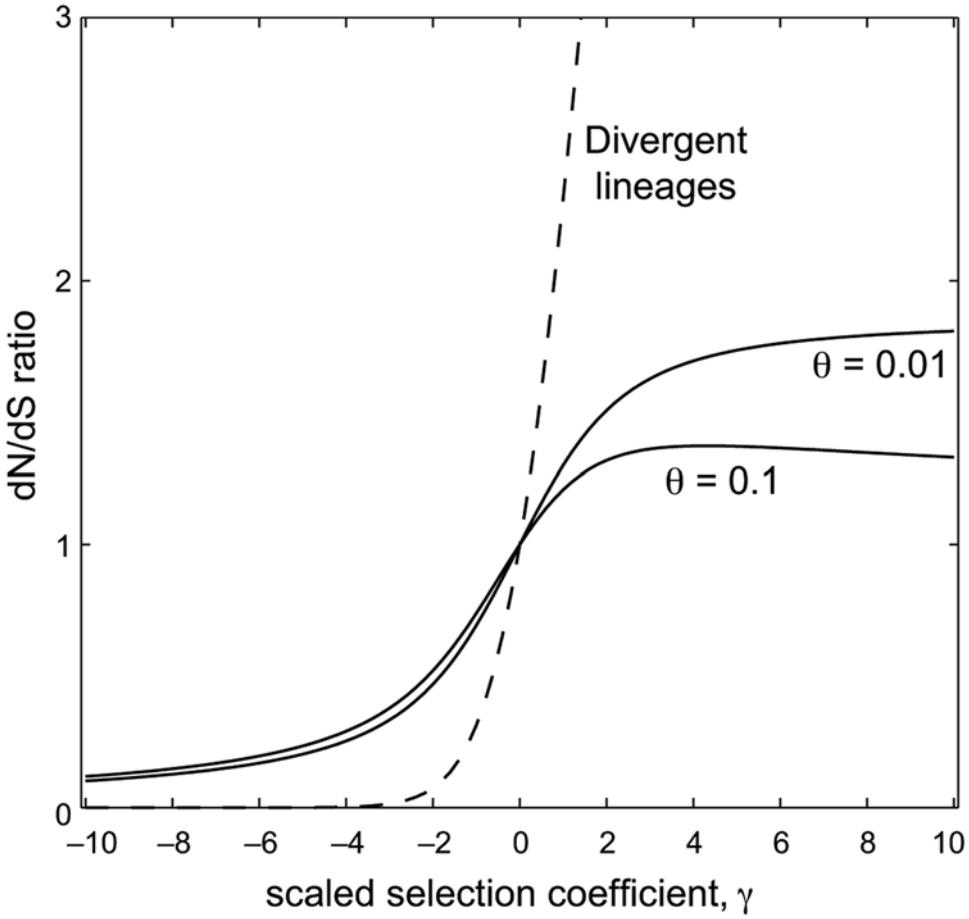

Figure 1. The relationship between the scaled selection coefficient, γ, and the expected dN/dS ratio. The dashed line shows the expected dN/dS ratio for samples from divergent lineages, given by Equation (2). The solid lines show the expected dN/dS ratio for within-population samples, given by Equation (5), under two mutation rates.
doi:10.1371/journal.pgen.1000304.g001

commonly found in practice, depends critically on the time-scale of sequences being compared: within a population such an observation is consistent with strong negative selection, whereas between divergent species such an observation implies weak negative selection.

The difference between short and long time-scales is even more striking in the case of positive selection. Within a population, the dN/dS ratio equals 1 under neutrality ($\gamma = 0$), as usual. But the dN/dS ratio is not a monotonic function of the selection coefficient: for positive selection of moderate strength the expected dN/dS ratio exceeds one, but as γ increases further the dN/dS ratio reaches a maximum value and then starts to descend (Figure 1). In fact, as a standard asymptotic analysis of Equation (5) shows, the expected dN/dS ratio approach zero as γ gets very large. This behavior is verified by Figure 2, which shows that dN/dS falls below unity under very strong positive selection. The exact behavior of dN/dS depends upon the mutation rate (Figures 1 and 2), but in all cases the relationship is non-monotonic.

Compared to the case of divergent lineages, the behavior of dN/dS within a population is so radically different that inferences of positive and negative selection based on dN/dS are problematic or, in many cases, impossible. Whereas $dN/dS<1$ is a faithful indication of negative selection across divergent lineages, the observation of $dN/dS<1$ within a population is consistent with either weak negative or strong positive selection. The intuition behind this result is straightforward: strong positive selection within a population will produce rapid sweeps at selected sites (but not at neutral sites, which are assumed independent). As a result, two individuals sampled from such a population are likely to contain the same allele at each selected site, producing a dN/dS value less than unity. By contrast, selective sweeps along divergent lineages will tend to produce fixed differences between representative individuals sampled from the two independent populations. Thus, the simple interpretation of dN/dS that applies to divergent lineages does not apply within a population.

Numerical Simulations

We performed two sets of Monte Carlo simulations, each based on the Wright-Fisher model with continual selection (i.e. selection *sensu* Yang), for comparison with our analytical results on dN/dS. In the first set of simulations we considered sites that could each assume one of two allelic types, similar to the setup used in our analytical treatment above. We performed a simulation of a single population over a short time-scale, as well as a simulation of two independent populations over a long time-scale (see Methods for details). At the end of each such simulation we sampled a pair of individuals, either from a single population or from each of two independent populations and computed the number of mutations (in the case of single population simulation) or substitutions (in the case of two population simulations) on the lineage separating the

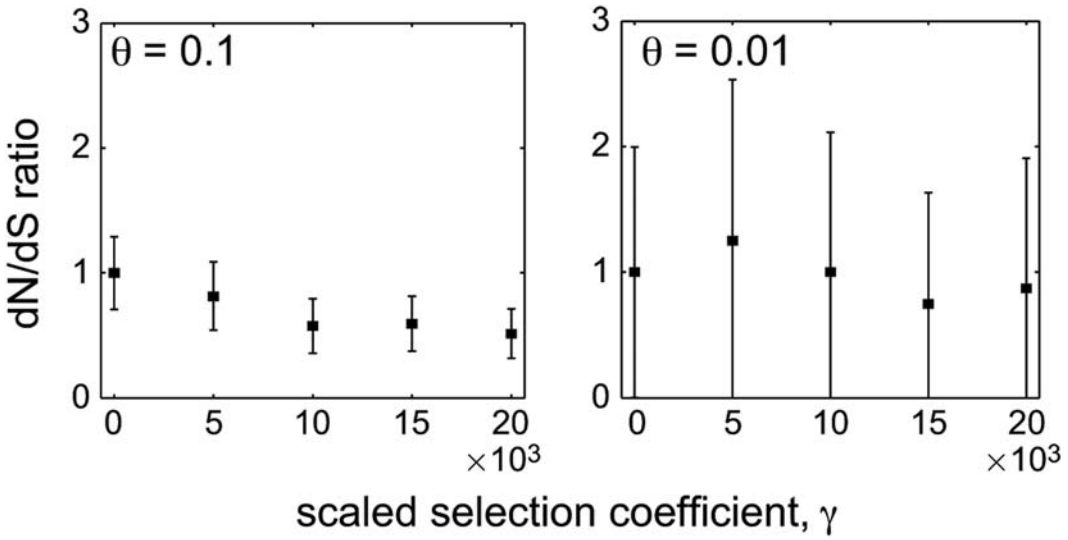

Figure 2. The behavior of the within-population dN/dS ratio for large γ in simulated Wright-Fisher populations. Black squares show the mean±two standard errors of the observed dN/dS ratio. Left panel shows results for $\theta = 0.1$; right panel shows results for $\theta = 0.01$. Simulations were performed at $L = 10^3$ independent sites.
doi:10.1371/journal.pgen.1000304.g002

two sampled individuals. We compared the observed dN/dS values to their theoretical expectations derived above. Figure 3 summarizes the results of these simulations for two values of the mutation rate and across a range of selection coefficients. In the case of a single population, the observed dN/dS value between sampled individuals agreed very well with our theoretical expectation (Equation 5). In the case of two independent populations, the observed dN/dS value agreed with the expectation derived by Neilsen & Yang (Equation 2). The slight departures between the simulations and Equation (2), visible only at $\theta = 0.1$, arise because the theoretical expectations were derived under the assumption that one mutant lineage would fix or go extinct before another mutant lineages is introduced. If we artificially depress the mutation rate to zero whenever two allelic types are segregating in a population we find perfect agreement between theory and simulation, even for $\theta = 0.1$ (Figure S1).

The simulation results confirm our theoretical analysis of dN/dS. The relationship between selection and dN/dS is accurately described by Equation (2) when comparing individuals sampled from two divergent lineages. By contrast, when individuals are sampled from a single population, the relationship between selection and dN/dS is radically different and accurately described by Equation (5) —even though the simulation procedure used for a single population is identical to the procedure used in each of the two independent populations.

In the second set of simulations we considered a slightly more realistic situation based on the true genetic code. These simulations employed the same Wright-Fisher model with continual selection, but in this case 64 allelic types are available instead of two. We compared two sampled individuals, each consisting of 10^4 (single population) or 10^3 (two populations) independent codon sites, and we estimated dN/dS from the sampled sequences using the PAML computer package, as opposed to using the exact ancestry. Thus, these simulations and dN/dS values provide a close representation of data that are likely to be encountered in practice.

Table 1 summarizes the results of the codon-based simulations. As expected, when comparing sequences from two independent populations the estimated dN/dS value increased monotonically with s. Moreover, based on the 95% confidence intervals, dN/dS>1 was rejected in the cases of simulated negative selection ($\gamma = -2$ or $\gamma = -5$); and dN/dS<1 was rejected in the cases of simulated positive selection ($\gamma = +2$ or $\gamma = +5$). In other words, when comparing divergent lineages the magnitude of dN/dS compared to unity is a faithful indicator of the sign of selection. By contrast, when comparing sequences sampled from a single population, dN/dS did not provide a reliable indicator of the strength or sign of selection, even though the length of the sampled sequences was 10 times larger in the single population simulations than in the two population simulations: for both $\gamma = -2$ and $\gamma = -5$ PAML did not reject the possibility that dN/dS>1; and for both $\gamma = +2$ and $\gamma = +5$, PAML did not reject the possibility that dN/dS<1. In fact, in one case of simulated positive selection the most likely estimate of dN/dS was less than unity.

The framework used in our second set of simulations is more realistic than the simple two-allele framework used in our theoretical analyses or those of Nielsen & Yang [3]. These simulations demonstrate the generality of our results: when applied to a single population, dN/dS is not particularly sensitive to the strength of selection and it is not a reliable indicator of the sign of selection.

Discussion

The dN/dS ratio remains one of the most popular and reliable measures of evolutionary pressures on protein-coding regions. Much of its popularity stems from the simple, intuitive interpretation of dN/dS<1 as negative selection, dN/dS = 1 as neutrality, and dN/dS>1 as positive selection. However, this simple interpretation requires that the sequences being compared represent stereotypical samples from divergent populations—an assumption that is also implicit in the methods that estimate dN/dS by maximum likelihood [7]. As we have demonstrated here, the relationship between selection pressure and dN/dS for samples within a population is radically different than the relationship for samples from divergent populations. In particular, within a population dN/dS does not increase monotonically with γ, dN/dS is less sensitive to changes in γ, and dN/dS<1 can occur under both negative and positive selection.

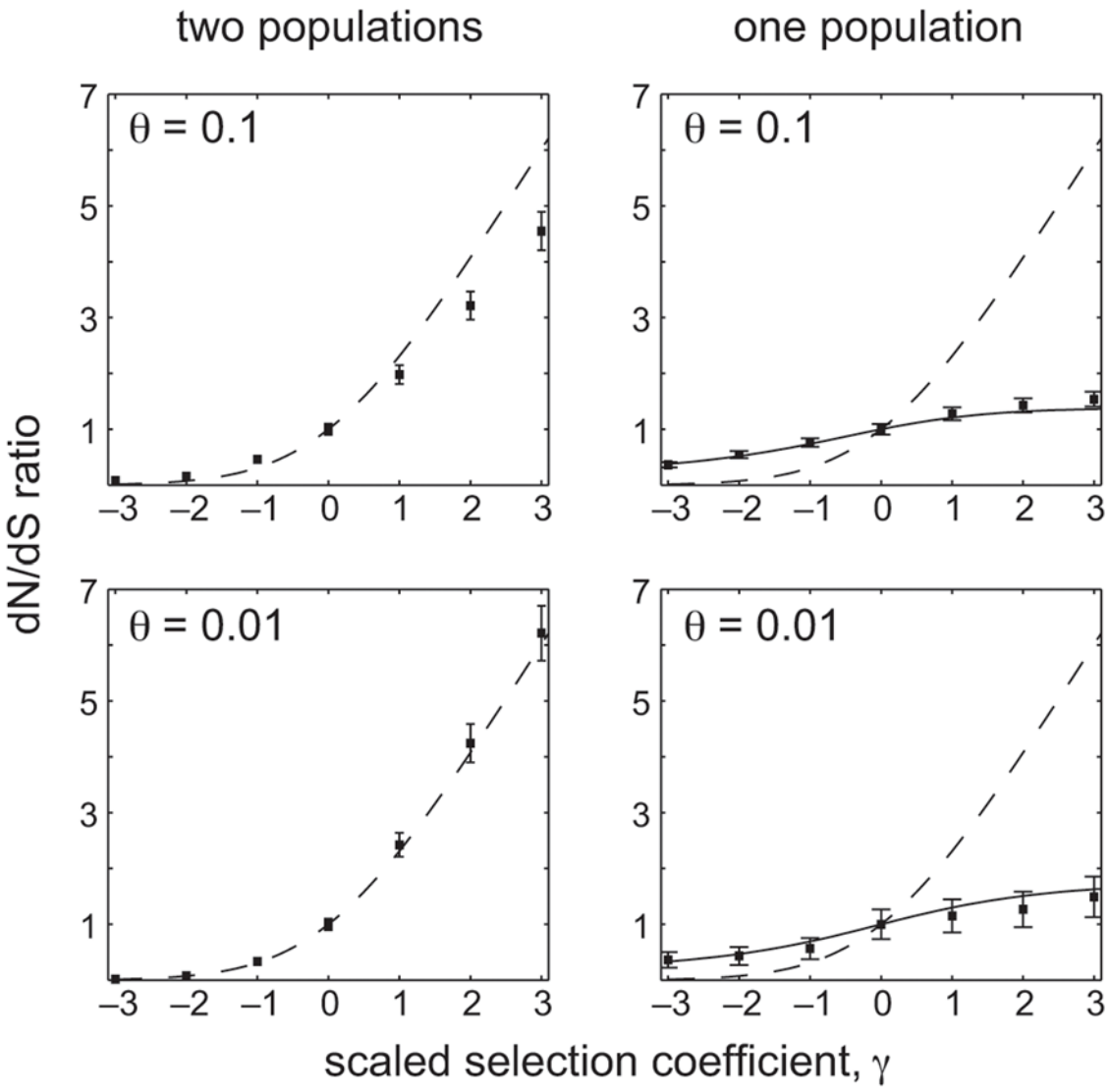

Figure 3. The relationship between the scaled selection coefficient, γ, and the dN/dS ratio in simulated Wright-Fisher populations. Black squares show the mean±two standard errors of the observed dN/dS ratio. The predicted dN/dS ratios for divergent lineages are shown in dashed lines (Equation 2); the predicted dN/dS ratios for a single population are shown in solid lines (Equation 5). Left column corresponds to results for two independent populations; right column corresponds to results for a single population. Top panels show results for $\theta = 0.1$; bottom panels show results for $\theta = 0.01$. The simulations for two populations were performed at $L = 10^3$ independent sites, and the simulations for a single population were performed at $L = 10^4$ independent sites.
doi:10.1371/journal.pgen.1000304.g003

Recently, Rocha et al. have investigated the relationship between divergence time and dN/dS [44]. Those authors considered an infinite-sites model under negative selection, and they presented an expression for the expected dN/dS ratio in an infinite population. By contrast, we have derived an analytic relationship between the selection pressure and dN/dS at a site under the Wright-Fisher model of a finite population, for both negative and positive selection.

The fact that polymorphisms within a population differ from divergences between species is well understood by population geneticists [23,45]. However, this important fact is often neglected in many applications of dN/dS to population data. In fact, one recent study explicitly suggests that dN/dS within a population should be used as a surrogate for dN/dS across divergent species [46]. Moreover, the standard infinite-site analysis of neutral and selected segregating polymorphisms (e.g. [23,47]) would suggest that the ratio pN/pS approaches 2 as γ gets large, whereas in fact the dN/dS ratio within a population approaches zero for strong positive selection (Equation 5). This discrepancy arises because the infinite-site analysis considers only the mean time that an allele spends in each frequency class while segregrating. By contrast, the single-site analysis (Equation 4) accounts for for the increased amount of time that a site spends in the monomorphic state as γ gets large.

Our analysis of selection and dN/dS has assumed independence of sites or, equivalently, free recombination between sites. This assumption is unrealistic in many practical settings. However, the same assumption has been made in prior analytic work on dN/dS [3], and the assumption is expected to be more accurate for small mutation rates, or for weak selection pressures. Outside of this parameter regime, the effects of linkage on dN/dS are difficult to analyze, and they form an important topic for further study.

Table 1. The relationship between the scaled selection coefficient, γ, and the dN/dS ratio as estimated by the PAML package from simulated data.

	Two populations	One population
γ	ω	ω_{pop}
−5	0.002 (0.000, 0.014)	0.001 (0.000, 2.755)
	0.002 (0.000, 0.014)	0.289 (0.068, 0.813)
−2	0.068 (0.040, 0.106)	1.000 (0.000, 19.300)
	0.105 (0.065, 0.159)	0.608 (0.226, 1.399)
0	0.934 (0.712, 1.237)	0.750 (0.000, 11.020)
	1.066 (0.810, 1.412)	0.967 (0.456, 1.934)
2	4.114 (2.821, 5.451)	0.500 (0.025, 5.621)
	3.245 (1.840, 4.868)	1.472 (0.749, 2.796)
5	4.409 (2.942, 6.172)	2.501 (0.396, 14.330)
	2.823 (1.763, 4.023)	1.680 (0.927, 3.024)

Wright-Fisher simulations based on the full genetic code were performed as described for two independent populations (middle column) and a single population (right column). The table shows the most-likely dN/dS value as estimated by PAML for two sampled sequences, as well as a 95% confidence interval obtained from the χ^2 distribution. For each value of γ, the first line corresponds to simulations with $\mu = 10^{-7}$, and the second line corresponds to simulations with $\mu = 10^{-6}$.

doi:10.1371/journal.pgen.1000304.t001

We have focused our analysis on Yang's particular formulation of selection, which stipulates that all mutations experience the same selection coefficient compared to the resident type [3,4,36,40]. Alternative formulations of selection (e.g. those that assume a constant fitness for each allele) can produce different relationships between γ and dN/dS over long time-scales [3]. Our results here, however, do not arise because we have considered a different selective model than Nielsen and Yang [3]; we are studying the same model, but considering samples from a single population instead of divergent populations.

Complications associated with interpreting dN/dS for population samples do not arise in many practical applications of dN/dS—i.e. those involving comparisons among divergent species. However, as sequence data are increasingly available, there is a temptation to apply computer packages such as PAML to intraspecific data—as has been done in many cases already (e.g. [3,8–18]). Published estimates of dN/dS based on samples from a single population are common for microbes and viruses. Inferences about natural selection drawn from such analyses should be interpreted with caution.

Many empirical studies of genes evolving under negative selection have found quizzical results, which our analysis helps to clarify: dN/dS values for such genes are typically closer to 1 when comparing intra-specific samples as opposed to inter-specific samples. This observation holds for bacterial data [11,12,14,16,18], for viral samples isolated from a single host versus viral samples isolated from different hosts [13], for closely related viral samples versus distantly diverged samples [48], and for conspecific versus interspecific mammalian sequences [49,50]. A variety of factors have been suggested to explain the elevation of dN/dS within a population under negative selection [49]: balancing selection, variable population sizes, variable mutation rates, relaxed selective constraint within certain lineages [51,52], statistical artifacts [53], or the prevalence of slightly deleterious mutations [13,48,49,54,55]. Our analysis clarifies these systematic empirical observations: elevated dN/dS values among conspecifics is expected under a model of continual negative selection, in which all protein-coding mutations experience the same selective constraint at all times (Figure 1). It is important to note that this explanation does not require us to assume a separate class of weakly deleterious mutations [13,48,49] or time-varying selective regimes [56].

Our results also have implications for inferences of positive selection based on dN/dS among conspecific samples. Even when samples come from independently evolving populations, the power of the dN/dS statistic to detect positive selection is low when the majority of sites in the protein evolve under purifying selection [28,57,58]. Our results indicate that the power of the dN/dS statistic to detect positive selection is further reduced when samples come from a single population (see Table 1). This lack of power has indeed been observed—and, in some cases, interpreted as a lack of selection—in studies of intrapatient HIV evolution [8,9,59] and genetic variation in *Salmonella* Typhi [17,18].

For higher eukaryotes, the distinction between multiple independent populations versus a single population is usually clearcut: samples from different species represent independent populations, whereas conspecific samples should be treated as arising from a single population (unless they are sampled from regions that have been reproductively isolated for more than N generations). For microbes and viruses, however, the distinction may be more opaque. The central issue is whether or not the sequences being compared represent competing genotypes in the sense of a Wright-Fisher population model. In the case of the human influenza A virus, for example, contemporaneous samples should probably be considered as arising from a single population, because the global population of influenza A strains is known to be well-mixed and genotypes are known to compete for available hosts [60]. When comparing non-contemporaneous samples, however, it is less clear whether the samples should be treated as arising from a single population or independent populations. In some sense, an influenza virus sample from the year 1968 is independent of a sample from year 2000. We might therefore expect that positive selection on influenza's HA locus would produce $\omega > 1$ when comparing non-contemporaneous samples (independent populations), but $\omega \approx 1$ when comparing nearly contemporaneous samples (single population). This type of pattern has indeed been reported [56], but it was interpreted as a signature of time-varying selection pressures on the HA protein. In fact, this kind of pattern would be expected under continual positive selection, given our analysis of dN/dS over short versus long time-scales.

As the discussion above suggests, it may be difficult to determine the appropriate time-scale associated with a dataset of sampled microbial sequences, particularly for a virus sampled at different timepoints. In fact, there may not be a single time-scale that applies to the entire dataset. In such cases, the relationship between the observed dN/dS ratios and the underlying selection coefficients will be described by some (unknown) mixture of Equation (2) and Equation (5). In such cases our central conclusion still holds: the relationship between selection and dN/dS is not necessarily a simple monotonic function, and it may be impossible to infer the selection pressure from the dN/dS measurement.

Methods

Stationary Distribution for a Site under Continual Selection

Here we derive the stationary distribution (4) under Yang's model of continual positive or negative selection. Consider a haploid population of constant size N, where each individual carries one of the two alleles at the focal site. One allele is the resident and confers fitness 1, the other allele is the mutant and

confers fitness $1+s$. Mutations between the resident and the mutant happen at rate μ per generation, and it is assumed that $\theta=2N\mu\ll 1$. The dynamics of the mutant frequency in the population is described by the classical Wright-Fisher model. Continual selection *sensu* Yang is incorporated in this model by setting the number of mutants to zero as soon as the mutant allele goes to fixation (see main text for details).

Within the standard diffusion approximation, the system is described by the frequency x of the mutant allele, which takes values in the interval [0,1]. The probability density $f(x, t; p)$ of the mutant frequency to be x at time t given that it was p at the time zero satisfies the forward Kolmogorov equation

$$\frac{\partial}{\partial t}f(x,t;p)=\frac{1}{2}\frac{\partial^2}{\partial x^2}(b(x)f(x,t;p))-\frac{\partial}{\partial x}(a(x)f(x,t;p)), \quad (6)$$

where $a(x)=\gamma x(1-x)-\theta x/2+\theta(1-x)/2$, $b(x)=x(1-x)$, and t is measured in N generations. Function $f(x, t; p)$ is subject to the following auxilary conditions.

$$f(x,0;p)=\delta(x-p) \quad (7)$$

$$\int_0^1 f(x,t;p)dx=1 \quad (8)$$

$$\lim_{N\to\infty} N\int_{1-1/N}^1 f(x,t;p)dx=0. \quad (9)$$

Equations (7) and (8) are the initial condition and the normalization condition, respectively. The non-standard condition (9) arises in the model of selection *sensu* Yang from the following consideration. In this model, the mutant allele becomes the new resident allele when it fixes in the population. In other words, the population becomes monomorphic for the resident type (the number of mutants is $Nx=0$) immediately upon the fixation of a mutant (the number of mutants is $Nx=N$). Thus, the probability of finding the population in the state where the mutant is fixed, $\int_{1-1/N}^1 f(x,t;p)dx$, must tend to zero with increasing N. Even though this integral does decays to zero, it must do so at least as fast as N^{-1} in order for the diffusion approximation to hold. This leads to Equation (9), which is essentially an absorbing boundary condition at $x=1$. Similar flux conditions have been studied in models of variable selection pressures [43].

It is worth noting that our boundary condition is not the same as a periodic boundary condition. A periodic condition would allow probability flux from state $x=1$ into state $x=0$ as well as in the reverse direction–whereas Yang's model of selection should allow only the former direction of flux.

We are interested in the stationary solution $\Phi(x|\gamma,\theta)$ of Equation (6) subject to conditions (8), (9). It is easy to show that the general stationary solution of (6) is given by

$$\Phi(x|\gamma,\theta)=x^{\theta-1}(1-x)^{\theta-1}e^{2\gamma x}(C_1\Psi(x)+C_2), \quad (10)$$

where

$$\Psi(x)=\int_x^1 \xi^{-\theta}(1-\xi)^{-\theta}e^{-2\gamma\xi}d\xi.$$

Note that, if we put $C_1=0$, we arrive at the classical zero-flux stationary solution by Wright [26]. However, in Yang's model, the probability flows out of $x=1$ into $x=0$, and so we need to satisfy conditions (8) and (9) to determine constants C_1 and C_2. To take the limit in (9), we notice that the following equality is true for any $\alpha\in[0,1)$ and any sufficiently smooth function $f(x)$.

$$\int_{1-1/N}^1 f(x)(1-x)^{-\alpha}dx=\frac{f(1)}{1-\alpha}N^{-1+\alpha}+O\left(N^{-2+\alpha}\right).$$

Therefore, putting $f(x)=x^{\theta-1}e^{2\gamma x}(C_1\Psi(x)+C_2)$ and $\alpha=1-\theta$, we obtain

$$N\int_{1-1/N}^1 \Phi(x|\gamma,\theta)dx=$$
$$\frac{e^{2\gamma}}{\theta}(C_1\Psi(1)+C_2)N^{1-\theta}+O\left(N^{-\theta}\right)=$$
$$\frac{e^{2\gamma}}{\theta}C_2N^{1-\theta}+O\left(N^{-\theta}\right).$$

Thus, in order satisfy condition (9), we must require $C_2=0$. This leads to (4) for $0<\theta<1$. A comparison between this stationary distribution and numerical simulations is shown in Figure S2.

Numerical Simulations

Two-Allele Simulations. We performed Wright-Fisher simulations of a population of constant size N evolving under positive or negative selection *sensu* Yang, in discrete time. In the simulation, each individual carries one of two possible alleles, labeled "0" or "1". At each generation, one of the alleles, called "the resident", confers fitness 1, the other allele, called "the mutant", confers fitness $1+s$ ($s=\gamma/N$ can be negative). However, the labels of the resident and the mutant alleles change over time (see below). During the reproduction round, N individuals are drawn randomly from the population with replacement, with probabilities proportional to their fitnesses. After choosing which individuals will reproduce, we draw the number of mutations to occur in the replication round from the Poisson distribution with mean $\mu N=\theta/2$, and we randomly assign these mutations to individuals. Since we consider only small mutation rates, typically zero or sometimes one mutation occurs in each generation. A mutation does not create a new allele but rather exchanges the allele label (from "0" to "1" or "1" to "0") of the individual in which it arises. Once the next generation is formed, we check whether the number of mutant-type alleles has reached N, in which case the fitness landscape is reversed: the currently fixed mutant allele becomes the new resident type and it is assigned fitness 1, while the currently absent allele (corresponding to the previous resident) becomes the new mutant type and it is assigned fitness $1+s$. Thus, the mutant allele always has fitness $1+s$ relative to the resident allele.

This simulation takes the following parameters as input: N, the population size; s, selection coefficient; μ, mutation rate; T, total number of generations; L, number of replicate populations or, equivalently, the number of independent sites. We initialized all simulations with a population monomorphic for allele "0", defined as the initial resident allele. The following parameter values were used for simulations: $N=1000$, $s\in\{-0.003,-0.002,-0.001,0,0.001, 0.002,0.003\}$, $\mu\in\{5\times10^{-6}, 5\times10^{-5}\}$, These values correspond to $\gamma\in\{-3,-2,-1,0,1,2,3\}$ and $\theta\in\{0.01,0.1\}$. We performed simulations of a single population and also simulations of two independent populations, as described below.

Single population. We used this type of simulation to test our theoretical predictions for the dN/dS ratio for individuals sampled from a single population. We let a population evolve for 2 μ^{-1} generations in order for it to reach the mutation-selection-drift equilibrium. In the last generation, we sampled two individuals and counted the number of mutations that occurred on the lineage connecting them, $d(\gamma,\theta)$. We compute the mean observed dN/dS value as $\hat{\omega}_{\mathrm{pop}}(\gamma,\theta)=\frac{\langle d(\gamma,\theta)\rangle}{\langle d(0,\theta)\rangle}$, where $\langle d(\gamma,\theta)\rangle$ is the average value of $d(\gamma,\theta)$ over L replicate simulations. We compared the observed value $\hat{\omega}_{\mathrm{pop}}(\gamma,\theta)$ with the theoretically expected value $\omega_{\mathrm{pop}}(\gamma,\theta)$.

Two divergent populations. We used this type of simulation to test the predictions for the dN/dS ratio made by Nielsen and Yang [3] (Equation 2) for individuals sampled from two divergent populations. We initialized two populations and let each of them evolve independently for 0.4 μ^{-1} generations, after which we counted the number of substitutions (fixation events) that occurred in each population. The number of substitutions, $s(\gamma,\theta)$, equals the number of mutations that occurred on the lineage connecting the most recent common ancestors of the two final populations. The mean observed dN/dS value is $\hat{\omega}(\gamma,\theta)=\frac{\langle s(\gamma,\theta)\rangle}{\langle s(0,\theta)\rangle}$. We compare the observed $\hat{\omega}(\gamma,\theta)$ with the theoretical prediction $\omega(\gamma,\theta)$.

Codon-Based Simulations and Estimation of dN/dS. We also simulated the evolution of a protein coding sequence consisting of L independent codon sites, in order to produce data that could be analyzed by the the PAML package [7]. We simulated populations for each site independently. In the final generation of each simulation, two individuals were sampled (either from a single population or from two divergent populations), and the corresponding codons were concatenated. A set of such simulations produces a pair of nucleotide sequences of length $3L$.

In each simulation at a site, an individual could carry one of the 64 codons. The mutation probability was μ per nucleotide per generation. The fitness of an individual was determined by the encoded amino acid: we assumed that only two amino acids, alanine and valine, were allowed at the site; one of them was the resident amino acid and conferred fitness 1, the other was the mutant amino acid and conferred fitness $1+s$; codons encoding other amino acids or stop codons were assumed lethal (non-reproductive). In all other respects the codon-based simulation was identical to the two-allele simulation. We initiated all simulations with a population monomorphic for codon GTT, which determined the initial resident allele (valine). The following parameter values were used: $N=1000$, $s \in \{-0.005,-0.002,0,0.002,0.005\}$, $\mu \in \{10^{-7},10^{-6}\}$. We ran the single population simulations for $L=10^4$ sites for $T=5\times10^5$ generations. We ran the two population simulations for $L=10^3$ sites for $T=0.25\ \mu^{-1}$ generations.

We used the CODEML program from the PAML package to infer the most likely dN/dS ratio for each pair of sequences. We used the likelihood ratio test, based on the χ^2 distribution, to determine the 95% confidence interval on the estimated dN/dS ratio.

Supporting Information

Figure S1 The relationship between the selection coefficient, γ, and the dN/dS ratio in simulated Wright-Fisher populations for $\theta=0.1$. Mutations are artificially switched off whenever two alleles segregate in the population. Notations as in Figure 2.
Found at: doi:10.1371/journal.pgen.1000304.s001 (0.3 MB EPS)

Figure S2 Stationary frequency distribution of the mutant allele for the Wright-Fisher model with continual selection. Gray bars show the histrogram obtained from the two-allele simulations with $\theta=0.1$, squares represent the corresponding values expected from Equation (4). Top panel, $\gamma=-3$, bottom panel, $\gamma=3$. Insets show the same data on a different scale.
Found at: doi:10.1371/journal.pgen.1000304.s002 (0.4 MB EPS)

Acknowledgments

SK and JBP were funded by a grant from the James S. McDonnell Foundation. JBP also acknowledges support from the Burroughs Wellcome Fund, the Penn Genome Frontiers Institute, and the Defense Advanced Research Projects Agency\Fun Bio" Program (HR0011-05-1-0057). The authors are grateful to Todd Parsons, Warren Ewens, Michael Desai, and Michael Lässig for discussions on the diffusion approximation, and to Ricky Der for the asymptotic analysis of the expected dN/dS ratio.

Author Contributions

Conceived and designed the experiments: JBP. Performed the experiments: SK. Analyzed the data: SK. Wrote the paper: SK JBP. Designed the research: SK JBP.

References

1. Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267: 275–276.
2. Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15: 496–503.
3. Nielsen R, Yang Z (2003) Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Mol Biol Evol 20: 1231–1239.
4. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725–736.
5. Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11: 715–724.
6. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596–1599.
7. Yang Z (2007) PAML4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586–1591.
8. Plikat U, Nieselt-Struwe K, Meyerhans A (1997) Genetic drift can dominate short-term human immunideficiency virus type 1 *nef* quasispecies evolution in vivo. J Virol 71: 4233–4240.
9. Crandall KA, Kelsey CR, Imamichi H, Lane HC, Salzman NP (1999) Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous rate ratio to detect selection. Mol Biol Evol 16: 372–382.
10. Frost SDW, Günthard HF, Wong JK, Havlir D, Richman DD, et al. (2001) Evidence for positive selection driving the evolution of HIV-1 *env* under potent antiviral therapy. Virology 284: 250–258.
11. Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, et al. (2002) Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J Bacteriol 184: 5479–5490.
12. Feil EJ, Cooper JE, Grundmann H, Robinson DA, Enright MC, et al. (2003) How clonal is *Staphylococcus aureus*? J Bacteriol 185: 3307–3316.
13. Holmes EC (2003) Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. J Virol 77: 11296–11298.
14. Jones N, Bohnsack JF, Takahashi S, Oliver KA, Chan MS, et al. (2003) Multilocus sequence typing system for group B streptococcus. J Clin Microbiol 41: 2530–2536.
15. Meats E, Feil EJ, Stringer S, Cody AJ, Goldstein R, et al. (2003) Characterization of encapsulated and noncapsulated *Haemophilus inuenzae* and determination of phylogenetic relationships by multilocus typing. J Clin Microbiol 41: 1623–1636.
16. Holden MTG, Feil EJ, Lindsay JA, Peacock SJ, Day NPJ, et al. (2004) Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. Proc Natl Acad Sci USA 101: 9786–9791.
17. Roumagnac P, Weill FX, Dolecek C, Baker S, Brisse S, et al. (2006) Evolutionary history of *Salmonella* Typhi. Science 314: 1301–1304.
18. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, et al. (2008) High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. Nat Genet 40: 987–993.
19. Kingman JFC (1982) On the genealogy of large populations. J Appl Prob 19: 27–43.

20. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.
21. Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133: 693–709.
22. Fay JC, Wu CI (2000) Hitchhiking under posiitve Darwinian selection. Genetics 155: 1405–1413.
23. Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. Genetics 132: 1161–1176.
24. Hartl DL, Moriyama EN, Sawyer SA (1994) Selection intensity for codon bias. Genetics 138: 227–234.
25. Ewens WJ (2004) Mathematical population genetics. New York: Springer Science+Business Media, Inc.
26. Wright S (1931) Evolution in Mendelian populations. Genetics 16: 97–159.
27. Yang Z (1996) Maximum-likelihood models for combined analyses of multiple sequence data. J Mol Evol 42: 587–596.
28. Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929–936.
29. Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17: 32–43.
30. Yang Z, Nielsen R, Goldman N, Pedersen AMK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155: 431–449.
31. Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19: 908–917.
32. Mayrose I, Doron-Faigenboim A, Bacharach E, Pupko T (2007) Towards realistic codon models: among site variability and dependency of synonymous and non-synononymous rates. Bioinformatics 23: i319–i327.
33. Forsberg R, Christiansen FB (2003) A codon-based model of host-specific selection in parasites, with an application to the influenza A virus. Mol Biol Evol 20: 1252–1259.
34. Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP (2004) Modeling the site-specific variation of selection patterns along lineages. Proc Natl Acad Sci USA 101: 12957–12962.
35. Yang Z, Nielsen R (2008) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol 25: 568–579.
36. Haigh J (1978) The accumulation of deleterious genes in a population—Muller's ratchet. Theor Pop Biol 14: 251–267.
37. Gordo I, Charlesworth B (2000) The degradation of asexual haploid populations and the speed of Muller's ratchet. Genetics 154: 1379–1387.
38. Gerrish PJ, Lenski RE (1998) The fate of competing beneficial mutations in an asexual population. Genetica 102/103: 127–144.
39. Park SC, Krug J (2007) Clonal interference in large populations. Proc Natl Acad Sci USA 104: 18135–18140.
40. Desai MM, Fisher DS (2007) Beneficial mutation-selection balance and the effect of linkage on positive selection. Genetics 176: 1759–1798.
41. Rouzine IM, Brunet E, Wilke CO (2008) The traveling-wave approach to asexual evolution: Muller's ratchet and speed of adaptation. Theor Pop Biol 73: 24–46.
42. Wright S (1945) The differential equation of the distribution of gene frequencies. Proc Natl Acad Sci USA 31: 382–389.
43. Mustonen V, Lässig M (2007) Adaptations to fluctuating selection in *Drosophila*. PNAS 104: 2277–2282.
44. Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, et al. (2006) Comparison of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol 239: 226–235.
45. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351: 652–654.
46. Liu J, Zhang Y, Lei X, Zhang Z (2008) Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. Genome Biology 9: R69.
47. Charlesworth J, Eyre-Walker A (2008) The McDonald-Kreitman test and slightly deleterious mutations. Mol Biol Evol 25: 1007–1015.
48. Sharp PM, Bailes E, Chaudhuri RR, Rodenburg CM, Santiago MO, et al. (2001) The origins of acquired immune deficiency syndrom viruses: where and when? Phil Trans R Soc Lond B 356: 867–876.
49. Hasegawa M, Cao Y, Yang Z (1998) Preponderance of slightly deleterious polymorphisms in mitochondrial DNA: nonsynonymous/synonymous rate ratio is much higher within species than between species. Mol Biol Evol 15: 1499–1505.
50. Zhang L, Li WH (2005) Human SNPs reveal no evidence of frequent positive selection. Mol Biol Evol 22: 2504–2507.
51. Takahata N, Satta Y, Klein J (1995) Divergence time and population size in lineage leading to modern humans. Theor Pop Biol 48: 198–221.
52. Adachi J, Hasegawa M (1996) Tempo and mode of synonymous substitutions in mitochondrial DNA in primates. Mol Biol Evol 13: 200–208.
53. Jordan IK, Rogozin IB, Wolf YI, Koonin EV (2002) Microevolutionary genomics in bacteria. Theor Pop Biol 61: 435–447.
54. Ohta T (1992) The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst 23: 263–286.
55. Ballard JWO, Kreitman M (1995) Is mitochondrial DNA a strictly neutral marker? Trends Ecol Evol 10: 485–488.
56. Wolf YI, Viboud C, Holmes EC, Koonin EV, Lipman DJ (2006) Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. Biol Direct 1: 34–53.
57. Sharp PM (1997) In search for molecular darwinism. Nature 385: 111–112.
58. Holmes EC, de A Zanotto PM (1997) Genetic drift of human immunodeficiency virus type 1. J Virol 72: 886–887.
59. de A Zanotto PM, Kallas EG, de Souza RF, Holmes EC (1999) Genealogical evidence for positive selection in the *nef* gene of HIV-1. Genetics 153: 1077–1089.
60. Viboud C, Bjørnstad ON, Smith DL, Simonsen L, Miller MA, et al. (2006) Synchrony, waves, and spatial hierarchies in the spread of influenza. Science. pp 447–451.