Inferring Selection and Mutation from DNA Sequences: The McDonald-Kreitman Test Revisited

Stanley A. Sawyer

Abstract

Aligned DNA sequences from the same species, and from two species that share a sufficiently recent common ancestor, are analyzed for evidence of unidirectional selection. We develop a technique for estimating the direction and magnitude of unidirectional selection that can be used at synonymous sites as well as at replacement codon positions. The model is applied to sample configurations at the ADH locus in two *Drosophila* species, and to the *gnd* locus in *Escherichia coli*. The analysis suggests positive selection for observed amino acid replacements at ADH, and is highly significant for selection against the observed amino acid replacements at *gnd*. Small amount of negative selection are detected in synonymous sites, even after allowance is made for base-dependent mutation.

Introduction

Experiments beginning in the middle 1960s found that many enzymes in natural populations are polymorphic.[1,2] A protracted and often heated debate began about the selective significance of these polymorphisms.[3] Some scientists felt that most of this variation is essentially selectively neutral, with at most negligible effects on fitness.[4,5] Others held that different enzymes were unlikely to be selectively equivalent, and so various forms of selection must be involved.[6,7,8,9,10] Later experiments showed there is even greater *synonymous* or *silent* variation at the DNA level. That is, there is more variation in the DNA itself than in the protein products.

Also puzzling is the phenomenon of *codon bias*, which is the tendency for the different three-base codons that code for the same amino acid to occur in different frequencies.[11,12] One possible reason for codon bias may be biased mutation rates at the DNA level. That is, some of the individual bases that compose DNA may be more mutable than others, or else there may be bias in their rates of mutation to other bases. Selective forces may also favor some codons over others. Particular codons may be necessary for the proper mRNA configuration, and variable tRNA abundances may slow the translation of some synonymous codons.[11,12] However, rates of nucleotide substitution at silent codon sites are similar to those in pseudogenes,[13,14] which suggests that most silent substitutions are nearly neutral. While in general silent DNA changes appear to be under weaker selective constraints than changes that cause amino acid replacement, some silent changes may have a selective effect.

A Quantitative Model for Selection

Our main purpose here is to discuss a method for detecting and estimating selection based on an aligned set of DNA sequences. The method will be applied both to silent site variation and to amino-acid variation. The model of selection will be that all changes to a consensus or an ancestral base or amino acid are either strongly deleterious (and so will never be seen in a natural population) or else change the fitness of the host by an equal amount, equal both for changes to different bases at the same site and for changes at different sites. Mutations at different sites have multiplicative effects on fitness. In particular, this model may not appropriate for detecting balancing or disruptive selection. Of course, no statistical test can detect *arbitrary* forms of selection, since *anything* that you observe could be the result of particular selective forces.

A fringe benefit of this analysis is that you can estimate separate mutation rates for synonymous and for amino-acid replacements. Many changes to a functioning protein or enzyme are presumably lethal or nearly lethal. Amino-acid variation that is common enough to be detected in a sample of DNA sequences is presumably subject to relatively weak selection. An estimate of the ratio of the amino-acid replacement mutation rate to the silent mutation rate should give an estimate of the proportion of amino acids in a protein that can be replaced without lethal or strongly deleterious effects on the host.

A Contingency-Table Test for Unidirectional Selection

Before describing the quantitative model, we first discuss a 2×2 contingency table test that can detect differences in selection between silent and replacement sites.[15] Suppose that we have n aligned DNA sequences from a coding region. Most aligned sites will have a single most common base with at most a few sequences with different bases at that site. If most changes from this consensus base are selectively deleterious, then we would expect relatively few sequences at any particular site to be different from the consensus base. Similarly, if changes from the consensus base were advantageous, we would also expect relatively few differences from the consensus, since otherwise the consensus would be quickly driven from the population. If polymorphic replacement sites are observed to be more bunched (i.e., have fewer deviations from the consensus) than polymorphic silent sites, then one possible explanation is unidirectional selection for or against the replacements.

Specifically, given n aligned DNA sequences, we say that a site is *simply polymorphic* if $n-1$ of the sequences have one base at this site and one sequence has a second base. All other polymorphic sites are called *multiply polymorphic*.

A three-site codon position is called *regular* if the consensus bases code for an amino acid other than leucine or arginine, which is equivalent to saying that the first two positions are nondegenerate (e.g., any replacement changes the amino acid). About half of regular amino acids are fourfold degenerate at the third position, which means that any base can be substituted at the third codon site without changing the amino acid. Most of the other amino acids are twofold degenerate at the third position. Two-fold degenerate amino acids are of two types. The first type can have either of the two pyrimidines T or C at the third position, but any other change at the third site changes the amino acid. For the second type, the third base is either of the two purines A or G. There is one threefold degenerate amino acid (isoleucine), which corresponds to the three codons ATT, ATC, and ATA. Since the codon ATA is extremely rare in most natural populations, we treat isoleucine as twofold degenerate and the rare codon positions with an ATA as irregular.

Consider a 2×2 contingency table with the numbers of silent simply and multiply polymorphic sites at amino-acid monomorphic regular codon positions in the first row, and the numbers of simply and multiply polymorphic sites at the first and second positions of regular codons in the second row (Table 1). Regular codon positions have the potential of supplying two replacement polymorphisms, but this rare, and historically may have been the result of two amino-acid replacements.

When all silent sites are used, the contingency table in Table 1 is highly significant. However, most replacement variation may be lethal or at least subject to highly deleterious selection, so that there may be at most two weakly-selected bases at any amino-acid varying site. Thus it may be fairer to compare replacement polymorphisms with twofold degenerate silent polymorphisms, which are more likely to be simply polymorphic than fourfold degenerate sites. When twofold degenerate sites are used, the contingency table in Table 1 is significant but is not highly significant.

Estimating Mutation Rates from Silent Sites

There are many different ways to estimate the amount of mutation at silent sites in an aligned set of DNA sequences (see e.g. Fu and Li, 1993[18]). The following approach[10,15] has the advantage that it automatically allows for saturation and homoplasy (i.e., parallel or repeated mutations at the same site), and can be adapted to estimate the divergence time between two species as well.

Table 1: 2×2 tables for selection at the *gnd*[a] locus in *Escherichia coli*[b]

	All silent[c]		Two-fold silent[c]	
	simple poly	multiple poly	simple poly	multiple poly
Silent (regular)[c]	60	83	27	31
Replacements (1,2 pos'n regular)	20	7	20	7
	$P = 0.003$[d]		$P = 0.021$[d]	

a – The *gnd* locus transcribes the enzyme 6-phosphogluconate dehydrogenase.

b – 14 strains of *E. coli* (1407bp; GenBank[16,17])

c – See text for definitions.

d – Two-sided Fisher exact test.

We assume that fourfold degenerate sites (for example) have mutation rates μ_T, μ_C, μ_A, and μ_G *to that base* per chromosome per generation. Thus the mutation rate depends on the base, but depends on the *target* base. Note that "mutations" of e.g. T to T that do not change the base are permitted, but will be taken into account before estimating the locus-wide silent mutation rate. While this model is not the most common way to model base-dependent mutation, it is consistent with some models of isochores in mammals,[14] and does lead to computable base-dependent estimates of mutation rates that do not need homoplasy corrections.

Under these conditions, the *population* frequencies of the four bases at that site will have the joint probability density

$$C_\alpha \, p_T^{\alpha_T - 1} \, p_C^{\alpha_C - 1} \, p_A^{\alpha_A - 1} \, p_G^{\alpha_G - 1} \, dp_T dp_C dp_A dp_G \tag{1}$$

where $\alpha_T = 2N_e\mu_T$, $\alpha_C = 2N_e\mu_C$, $\ldots$, where N_e is the haploid effective population size and $C_\alpha = C(\alpha_T, \alpha_C, \ldots)$, under the usual conditions for diffusion approximations.[10,19,20] The density in equation (1) is called a Dirichlet density. The corresponding density for pyrimidine twofold degenerate sites is the beta density $C'_\alpha \, p_T^{\alpha_T - 1} \, p_C^{\alpha_C - 1} \, dp_T dp_C$ for $p_T + p_C = 1$, with a similar expression for the purine twofold degenerate sites.

Now assume that the site is part of an aligned sample of n DNA sequences, and consider the probability that the sample has n_T sequences with the base T at that site, n_C sequences with C, n_A with A, ..., where $n = n_T + n_C + n_A + n_G$. This probability can be obtained by integrating the density in equation (1), and is

$$C_n \frac{\alpha_T^{(n_T)} \alpha_C^{(n_C)} \alpha_A^{(n_A)} \alpha_G^{(n_G)}}{\alpha^{(n)}}, \qquad \alpha = \alpha_T + \alpha_C + \alpha_A + \alpha_G \qquad (2)$$

where $x^{(k)} = x(x+1)\dots(x+k-1)$ and $C_n = n!/(n_T!\,n_C!\,n_A!\,n_G!)$ [10,21]. The corresponding probability for pyrimidine twofold degenerate sites is $C_n' \alpha_T^{(n_T)} \alpha_C^{(n_C)}/(\alpha_T + \alpha_C)^{(n)}$. The probabilities in equation (2) for fourfold degenerate sites, and the corresponding probabilities at twofold degenerate sites, can be combined to obtain maximum likelihood estimators for $\alpha_T, \alpha_C, \alpha_A$, and α_G (Table 2).

Table 2: Maximum likelihood estimates of $\alpha_T, \alpha_C, \alpha_A, \alpha_G$ from the probabilities (2) at silent sites

ADH[a]		*gnd*[b]	
alpha's	4-fold[c]	alpha's	4-fold[c]
$\alpha_T = 0.0080$	0.155	$\alpha_T = 0.128$	0.407
$\alpha_C = 0.0300$	0.610	$\alpha_C = 0.109$	0.288
$\alpha_A = 0.0023$	0.066	$\alpha_A = 0.063$	0.106
$\alpha_G = 0.0097$	0.169	$\alpha_G = 0.057$	0.199
$\mu_{\text{sil}} = 2.05$[d]		$\mu_{\text{sil}} = 30.82$[d]	

a – Pooled likelihoods for 6 *Drosophila simulans* and 12 *D. yakuba* strains[22] (771bp; pooling means that within-species log likelihoods are summed).

b – Likelihoods for 14 *E. coli* strains (1407bp; GenBank[16,17]).

c – Base frequencies at 4-fold degenerate regular silent sites.

d – Locus-wide silent mutation rate scaled by N_e (see text).

Given equation (1), the mean frequency of the base T at fourfold degenerate sites is $E(p_T) = \alpha_T/\alpha$ for α in equation (2). Thus the expected rate of transitions T → C at fourfold degenerate sites in a genetic locus is $N_4 \alpha_T \alpha_C/(2\alpha)$, where N_4 is the number of fourfold degenerate regular codon positions in the locus. (The factor of two is because μ_C is the mutation rate to the base C per N_e generations, while $\alpha_C = 2N_e\mu_C$ in equation (1).) Similarly, the mutation rate at pyrimidine twofold degenerate sites in the locus is $N_{2,TC}\,\alpha_T\alpha_C/(\alpha_T+\alpha_C)$, where $N_{2,TC}$ is the number of pyrimidine twofold degenerate regular codon positions. These considerations lead to a formula for the locus-wide silent mutation rate μ_{sil}[10] (Table 2).

The maximum likelihood method assumes that the distributions at different silent sites can be treated as independent. Recombination and gene conversion both help to insure the independence of site distributions. Independence can be tested by computing the significance of autocorrelations of the events monomorphic/polymorphic for adjacent silent sites. The first three autocorrelations are not significant for either the *E. coli* data in Table 1 nor the two ADH data sets in Table 2. Maximum likelihood theory uses independence only for the central limit theorem for the log likelihoods for the various terms,[23] and so can tolerate some deviation from joint statistical independence.

The methods that are used in this paper assume that each sample from a species is a random sample from a panmictic population. If a sample contains some strains that are significantly different from the others, then model parameters will be estimated incorrectly. For example, this study began with 16 strains of *E. coli* for *gnd*, of which two (labeled r4 and r16) are about as distant from the other *E. coli* strains as they are from *Salmonella*. The remaining *E. coli* strains had an estimated phylogeny that had a more regular appearance. The two aberrant *E. coli* strains were excluded from the analysis.

Estimating Mutation *and* Selection Rates

We now consider a model that will allow us to estimate both the mutation rate μ and the relative selection rate γ for mutants, both scaled by the haploid effective population size N_e. This model is sensitive to saturation and homoplasy, but should give reliable results if the estimate for μ_{sil} (the parameter μ for regular silent sites) is comparable to or greater than the more accurate estimate of μ_{sil} based on the Dirichlet density (1) of the previous section (which, however, assume selective neutrality at silent sites). This model will be applied both for bases at regular silent sites and for amino acids at codon positions.

Consider a flux of mutations (at the rate μ per generation) in the population. Each mutation changes a base in one individual. Most of the resulting new mutant alleles quickly go extinct by drift, but some survive to have appreciable base frequencies in the population. We ignore subsequent mutations at that site. Since we are assuming that all bases (or amino acids) that are not the ancestral base are selectively equivalent, we can ignore mutation between mutant bases at the same site.

Now view the *population frequencies at those sites* for the surviving mutant bases as a *point process* of frequencies on $[0,1]$. Under diffusion approximation conditions, this will be a Poisson point process with expected density[10]

$$2\mu\,\frac{1-e^{-2\gamma(1-p)}}{1-e^{-2\gamma}}\,\frac{dp}{p(1-p)} \qquad \text{(Selection)}$$

$$2\mu\,\frac{dp}{p} \qquad \text{(No selection; i.e. } \gamma = 0)$$

for $0 < p < 1$. In the first case, mutant bases have a relative selective advantage of γ scaled by N_e. Note that these densities are not integrable at $p = 0$. This corresponds to the fact that the population contains a large number of rare mutants at any one time.

Under these assumptions, the counts of the numbers of silent sites that have r sequences with bases different from the ancestral base in a sample of n aligned DNA sequences are independent Poisson with means

$$2\mu \int_0^1 \frac{1-e^{-2\gamma(1-p)}}{1-e^{-2\gamma}} \binom{n}{r} p^r (1-p)^{n-r} \frac{dp}{p(1-p)}, \qquad 1 \le r \le n-1$$

$$\frac{2\mu}{r} \qquad \text{(No selection; i.e. } \gamma = 0)$$

We use the counts for polymorphic silent sites within a species to estimate parameters μ_{sil} and γ_{sil} for silent sites, and the counts for amino-acid polymorphic codon positions to estimate parameters μ_{rep} and γ_{rep} for replacement amino acids. We again use maximum likelihood estimates, but lump together both the counts and the probabilities for r and $n-r$ since we may not know the ancestral base or amino acid (Table 3).

The scaled selection rate $\gamma_{\text{rep}} = -3.66$ for *E. coli* in Table 3 corresponds to a selection rate of $s = -\gamma_{\text{rep}}/N_e$ per generation against replacements, where N_e is the effective population size of *E. coli*. We estimate N_e as follows. The value $\mu_{\text{sil}} = 30.82$ in Table 2 corresponds to $N_{\text{sil}} \times \mu N_e$, where μ is the mutation rate per site per generation and N_{sil} is the number of amino-acid monomorphic codon positions with twofold or fourfold degenerate regular silent sites. The 14 strains of *E. coli* in Table 2 have $N_{\text{sil}} = 367$, and the estimate $\mu = 5 \times 10^{-10}$ per generation[24] implies $N_e = 1.7 \times 10^8$.

This estimate of N_e leads to $s = -\gamma_{\text{rep}}/N_e = 2.2 \times 10^{-8}$ per generation against replacements. Thus the average magnitude of selection per generation that acts against observed amino acid substitutions is quite small. One way in which such a small selection coefficient could be realized is if a substitution is selectively neutral in most environments, but disadvantageous in some rarely-encountered environments.[8]

The estimates $\mu_{\text{rep}} = 12.51$ (for 469 codons, or $2 * 469 = 938$ first and second codon position sites) but $\mu_{\text{sil}} = 30.82$ (for 367 codons) in Table 3 suggests that only about one sixth of amino acid positions in *E. coli* are susceptible to

Table 3: Estimates of the mutation rate μ and the selection rate γ within one species

14 *E. coli* strains, *gnd* locus (1407p):

$\mu_{\text{sil}} = 30.82$	(From Table 2)
$\mu_{\text{sil}} = 33.57 \pm 5.50^{\text{a}}$ $\gamma_{\text{sil}} = -1.34 \pm 0.83^{**\ \text{ac}}$	
$\mu_{\text{rep}} = 12.51 \pm 4.47^{\text{b}}$ $\gamma_{\text{rep}} = -3.66 \pm 2.24^{***\ \text{bc}}$	
$\gamma_{\text{sil}} \neq \gamma_{\text{rep}}$	$(P = 0.029^{*})$

$* \ P < 0.05$ $\quad ** \ P < 0.01$ $\quad *** \ P < 0.001$

a – Estimated from base distributions at polymorphic regular silent sites.

b – Estimated from amino acid distributions at amino-acid polymorphic codon positions.

c – The ranges ± are 95% normal-theory confidence intervals, while P-values are for likelihood ratio tests against $\gamma = 0$.

a weakly-selected replacement. This is roughly consistent with the estimate $s = 1.6 \times 10^{-7}$ against replacement amino acids, which is about seven times as large as the value $s = 2.2 \times 10^{-8}$ obtained above, for a similar *E. coli* model in which *all* codon positions in *gnd* were assumed susceptible to a weakly-selected amino-acid replacement.[15]

The closeness of the two estimates of μ_{sil} in Tables 2 and 3 suggests that saturation or homoplasy do not have a significant effect in the estimates of Table 3. The fitted values for counts at regular silent sites for the data in Table 3 are not significantly different from the observed counts ($P = 0.71$, 5 degrees of freedom, 143 polymorphisms). The fitted values for counts for replacement amino acids resembled the observed counts, but had too many empty cells to carry out a chi-square goodness-of-fit test. A more detailed description of the method of estimation in Table 3 will appear elsewhere.

Fixed Differences Between Two Species

Fixed differences between samples are sites (or amino acids) that are fixed within each sample, but fixed at different bases. Given aligned sequence data from two related species, the fixed differences provide additional data that can be used to estimate mutation and selection rates.[10]

The first step in using this additional data is to estimate the time since the two species diverged. Griffiths (1979)[25,26] derived a time-dependent version of the steady-state Dirichlet density in equation (1) for population frequencies. Griffiths' formula gives the joint probability density of two sets of population base frequencies, one for each species, in terms of two sets of $\alpha_T = 2N_e\mu_T, \ldots$ parameters (for the two species), and a parameter t_{div}, which is the scaled number of generations since the two species diverged.

We fix the $\alpha_T, \ldots$ parameters in both species to the pooled estimates of $\alpha_T, \ldots$ from regular silent sites (as in Table 2), and find the maximum likelihood estimate of t_{div} using the joint configurations at regular silent sites for the two samples of DNA sequences aligned together.[10] The estimate of t_{div} for the two *Drosophila* species in Table 2 is

$$t_{\text{div}} = 5.81 \pm 2.52 \qquad \text{(ADH)}$$

In the mutational flux model of the last section, the scaled rate of fixation of mutant bases at the population level is

$$\begin{aligned} &\mu\,\frac{2\gamma}{1-e^{-2\gamma}} && \text{(Selection)} \qquad (3)\\ &\mu && \text{(No selection; i.e. } \gamma = 0) \end{aligned}$$

If we ignore fixations at the same site in both species, the number of fixed differences will be a Poisson random variable whose mean is $2t_{\text{div}}$ times the relevant expression in equation (3). Technically, the rates in equation (3) are the rates of new mutations that will eventually become fixed. Thus these estimates of fixed differences will overestimate the observed numbers by the number of ancestral polymorphisms that become fixed at different bases, and will underestimate the observed numbers of fixed differences by the number of mutations that will eventually become fixed but remain polymorphic in the present.[10] We ignore both types of errors.

We now obtain maximum likelihood estimates for μ and γ jointly by using polymorphic sites within each species, the estimates for t_{div} above, and the probabilities of fixation by time t_{div} from equation (3) (see Table 4).

The fitted values matched the observed polymorphism counts fairly well, but had a tendency to overestimate counts in one species and underestimate them in the other species.

The mutational flux model assumes either than you combine the observed and theoretical counts for r and $n-r$ differences from the consensus (assuming n DNA sequences), or else that you be able to identify the ancestral base. We followed the first strategy in Table 3 for estimates within each of the species, but attempted to estimate the ancestral bases using the joint configuration

Table 4: Estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ using two species

6 *D. simulans* and 12 *D. yakuba* strains (771bp; ADH[22]):

$\mu_{\rm sil} = 2.05$	(From Table 2)
$\mu_{\rm sil} = 2.48 \pm 0.78^{\rm a}$	
$\gamma_{\rm sil} = -0.53 \pm 0.70^{\rm ac}$	$(P = 0.16)$
$\mu_{\rm rep} = 0.068 \pm 0.050^{\rm b}$	
$\gamma_{\rm rep} = 3.58 \pm 9.51^{\rm bc}$	$(P = 0.12)$
$\gamma_{\rm sil} \neq \gamma_{\rm rep}$	$(P = 0.051)$

$** \ P < 0.01$ $*** \ P < 0.001$

a,b,c – See footnotes to Table 3.

data for the two species in Table 4. If both species had the same consensus base at a site, then that base was assumed to be ancestral. The likelihood was summed for the two bases at fixed differences (which eliminates the need to estimate the ancestral base between those two bases), but this was not practical for all polymorphic sites with different consensus bases in the two species. The ancestral bases at those sites was assigned randomly in Table 4 for those sites, and appeared to give similar estimates for different random assignments.

Other strategies would be to combine the counts over r and $n-r$ within each species, as in Table 3, or perhaps to change the assignment with the sign of γ so that the most common base at that site is the one that is selected. The random-assignment strategy may not be the best one. The optimum strategy on this point will be the object of further research.

A McDonald-Kreitman Table

McDonald and Kreitman (1991)[22] introduced a 2×2 contingency table for selective differences between silent sites and replacement sites in aligned samples between two species. They include the number of sites that are fixed differences and the number of sites that are polymorphic within either species (or both) as the two columns. In the first row we take the numbers of regular silent sites (that are fixed differences or polymorphic) and the similar number of sites within amino acid polymorphic codon positions in the second row. If there is no selective difference between silent and replacement sites, then the 2×2 table should be nonsignificant (Table 5).

Table 5: A "McDonald-Kreitman[22]" table for ADH (*D. simulans* and *D. yakuba*[a])

	fixed at diff. bases	poly. in either spp.
Silent (regular only)	17	21
Replacement (sites)	6	0

$P = 0.022$[b]

a – 6 strains of *D. simulans* and 12 strains of *D. yakuba* (771bp).[22]

b – Two-sided Fisher exact test.

The contingency table in Table 5 is significant using regular silent sites ($P = 0.022$), while the two-species analysis in Table 4 just misses significance ($P = 0.051$). However, the methods of analysis are quite different, and it would be quite possible for all of the estimates in Table 4 to be significant while the McDonald-Kreitman table above is not significant.

Acknowledgements

This work was partially supported by National Science Foundation Grant DMS-9108262 and National Institutes of Health research grant GM-44889.

References

1 Lewontin, R. C. and Hubby, J. L. (1966) *Genetics* 54, 595–609

2 Harris, H. (1966) *Proc. Royal Soc. London Ser. B* 164, 298–310

3 Lewontin, R. C. (1991) *Genetics* 128, 657–662

4 Kimura, M. (1968) *Nature* 217, 624–626

5 Kimura, M. (1983) *The Neutral Theory of Molecular Evolution.* Cambridge University Press

6 Wills, C. (1973) *Amer. Naturalist* 107, 23–34

7 Lewontin, R. C. (1974) *The Genetic Basis of Evolutionary Change.* Columbia University Press

8 Hartl, D. L. (1989) *Genetics* 122, 1–6

9 Hartl, D. L. and Sawyer, S. A. (1991) *J. Evol. Biol.* 4, 519–532

10 Sawyer, S. A. and Hartl, D. L. (1992) *Genetics* 132, 1161–1176

11 Li, W.-H. and Graur, D. (1991) *Fundamentals of molecular evolution.* Sinauer Associates

12 Hartl, D. L. and Clark, A. (1989) *Principles of population genetics* (2nd Ed) Sinauer Associates

13 Li, W.-H., Wu, C.-I. and Luo, C.-C. (1985) *Mol. Biol. Evol.* 2, 150–174

14 Wolfe, K., Sharp, P., and Li, W.-H. (1989) *Nature* 337, 283–285

15 Sawyer, S. A., Dykhuizen, D., and Hartl, D. L. (1987) *Proc. Nat. Acad. Sci. USA* 84, 6225–6228

16 Dykhuizen, D. E., and Green, L. (1991) *J. Bacteriol.* 173, 7257–7268

17 Bisercic, M., Feutrier, J. Y., and Reeves, P. R. (1991) *J. Bacteriol.* 173, 3894–3900

18 Fu, Y.-X. and Li, W.-H. (1993) *Genetics* 134, 1261–1270

19 Wright, S. (1949) pp365–389 in *Genetics, Paleontology, and Evolution*, edited by G. Jepson, G. Simpson, and E. Mayr. Princeton Univ. Press

20 Kingman, J. F. C. (1980) *Mathematics of Genetic Diversity* CBMS-NSF Regional Conf. Ser. Appl. Math **34**

21 Watterson, G. (1977) *Genetics* 85, 789–814

22 McDonald, J. H. and Kreitman, M. (1991) *Nature* 351, 652–654

23 If you maximize the product of the individual likelihoods (instead of the unknown true joint likelihood), but assume the central limit theorem for the individual log likelihoods, then the maximum likelihood estimator has the correct asymptotic behavior.

24 Ochman, H. and Wilson, A. C. (1987) in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology* eds Ingraham, J. L., Low, K. B., Magasanik, B., Neidhardt, F. C., Schaechter, M. and Umbarger, H. E. (American Society of Microbiology Pubs.)

25 Griffiths, R. C. (1979) *Adv. Appl. Probab.* 11, 310–325

26 Tavaré, S. (1984) *Theor. Popul. Biol.* 26, 119–164