

Reductions in genetic variation in *Drosophila* and *E. coli* caused by selection at linked sites

BRIAN CHARLESWORTH* and DAVID S. GUTTMAN

Department of Ecology and Evolution, University of Chicago, 1101 E. 57th Street, Chicago, IL 60637-1573, USA

Abstract. Selection at linked sites has important consequences for the properties of neutral variation and for tests of the predictions of the neutral theory of molecular evolution. We review the theory of the effect of adaptive gene substitutions on neutral variability at linked sites (hitchhiking or selective sweeps) and discuss theoretical results on the effect of selection against deleterious alleles on variation at linked sites (background selection). In *Drosophila melanogaster* there is a clear relation between the frequency of recombination in a given region of the chromosome and the amount of natural variability in that region. Attempts to predict this relation have given rise to models of selective sweeps and background selection. We describe possible methods of discriminating between these models, and also discuss the probable strong influence of selective sweeps on variation in largely nonrecombining genomes, with particular reference to *Escherichia coli*. Finally we present some unresolved questions and possible directions for future research.

Keywords. Selective sweeps; hitchhiking; background selection; *D. melanogaster*; *E. coli*; neutral variation.

1. Introduction

The neutral theory of molecular evolution, as developed by Motoo Kimura and his colleagues, has provided a powerful tool for interpreting data on molecular variation and evolution. In its classical form, the neutral theory assumes that the evolutionary dynamics of allelic variation at a locus can be considered in isolation from events taking place at other loci in the genome. By this means, the rich theory of single-locus population genetics generates predictions about the statistical properties of molecular variation and evolution, which can be tested against DNA and protein sequence data (Kimura 1983; Kreitman 1991). But it was recognized quite early in the development of the theory of molecular population genetics that there may be important consequences of selection at linked sites for the properties of neutral variation (Ohta 1971, 1973; Sved 1972; Maynard Smith and Haigh 1974). This may considerably complicate tests of the predictions of the neutral theory. Our purpose in this paper is to review theories of how selection at linked sites can reduce neutral or nearly neutral variation at a locus under study. We then compare the predictions of these theories with data on molecular variation in *Drosophila* and bacteria.

2. The theory of genetic hitchhiking

Maynard Smith and Haigh (1974) pointed out that the level of polymorphism at a neutral locus can be greatly reduced by the spread of a selectively favourable mutation

*For correspondence

at a linked locus, owing to an increase in the frequency of the neutral allele associated initially with the favourable mutation. They showed that the magnitude of this 'hitchhiking' effect decreases with the ratio of the frequency of recombination (r) between the selected and neutral locus to the selection coefficient (s) at the selected locus, and is negligible when this ratio is of the order of one or more. The most extreme form of hitchhiking thus occurs in asexual or nonrecombining genomes. It was recognized empirically in studies of bacterial chemostat populations as the phenomenon of periodic selection, whereby alleles at a neutral marker locus experience violent fluctuations as successive selective substitutions perturb their frequencies (Atwood *et al.* 1951a, b).

Maynard Smith and Haigh (1974) proposed that hitchhiking might explain the rather loose relation observed between genetic diversity at allozyme loci and species population size, which appears to contradict the predictions of the neutral theory. Hitchhiking events in the neighbourhood of a neutral locus can be viewed as causing a reduction in the effective population size (N_e) perceived by the locus, thereby reducing N_e for an abundant species much below what might be expected from the number of individuals in the species. This effect may obscure the relation between population size and genetic diversity expected under the neutral theory (Kimura and Crow 1964; Kimura 1983).

3. The relation between recombination and variation in *Drosophila*

While Maynard Smith and Haigh's work stimulated a brief flurry of theoretical papers on various aspects of hitchhiking (Ohta and Kimura 1975; Thomson 1977), interest in this problem then lapsed. But at the end of the 1980s, data on DNA variation in *Drosophila* became available which suggested that levels of variability are lower in regions of the genome where rates of recombination per unit physical distance are reduced relative to the rates in other regions (Stephan and Langley 1989; Aguadé *et al.* 1989). Examples of such regions of reduced recombination are the telomeric and centromere-proximal euchromatin, and the small fourth chromosome (Ashburner 1989). In response to these observations, Kaplan *et al.* (1989) published a reanalysis of the hitchhiking model, postulating a steady state in which a constant input of adaptively favourable mutations enters the population at sites scattered randomly over the genome. Under such a model the equilibrium level of nucleotide site diversity under mutation and drift at a neutral locus is reduced below the classical value by an amount that depends on the rate of favourable gene substitutions per nucleotide site, and the rate of recombination per nucleotide site. A neutral locus in a region of reduced recombination is thus expected to have a larger reduction in diversity, since it experiences a higher density of substitutions at closely linked sites than a gene in a more freely recombining region.

In the past few years the intensive effort put into surveying molecular variation in natural populations of *Drosophila*, by the use of both restriction fragment length polymorphism analysis and direct DNA sequencing, has yielded a body of evidence that considerably strengthens the evidence for a relation between the local rate of recombination for a gene and the level of DNA variability that it exhibits (Begun and Aquadro 1992; Aquadro *et al.* 1994; Moriyama and Powell 1996). Figure 1 shows a plot of the relation between variability and the coefficient of exchange (a measure of the rate of recombination per unit physical distance) for genes on the X chromosome of *D. melanogaster*. The possibility that this effect could be generated by higher mutation

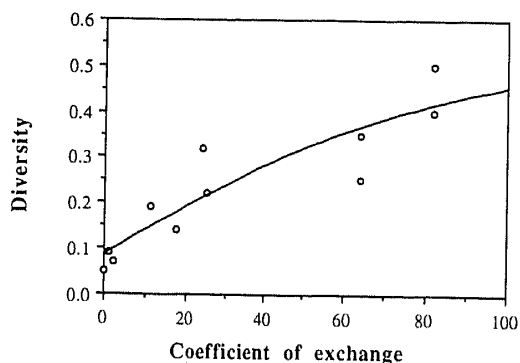


Figure 1. The nucleotide site diversity ($\times 10^2$) for genes on the X chromosome of *D. melanogaster*, plotted against a measure of the coefficient of exchange. The diversity values were obtained from published estimates of the numbers of segregating sites in population samples, using RFLP or SSCP data (Charlesworth 1996). The coefficient of exchange of a gene is estimated as the slope ($\times 10^2$) of the function relating map distance to the proportion of the X euchromatin distal to the gene, at the position of the gene in question. See Charlesworth (1996) for details of this function and of the genes concerned. The curve is the least-squares quadratic fit to the points ($r^2 = 0.81$).

rates associated with higher rates of recombination seems to be ruled out by failure to detect a similar relation between the coefficient of exchange and the extent of interspecies divergence (Begun and Aquadro 1992; Moriyama and Powell 1996).

At first sight, therefore, the evidence seems strong that adaptive gene substitutions must be occurring sufficiently frequently in the *D. melanogaster* genome that hitchhiking events reduce the level of molecular variation below classical neutral expectation, and that this reduction is strongest in regions of the genome where recombination is least frequent. Refinements of the hitchhiking model by Wolfgang Stephan and his collaborators (Stephan *et al.* 1992; Wiehe and Stephan 1993) have generated the following relation between the expected diversity per nucleotide site under Kimura's (1971) infinite-site model, the coefficient of exchange, and the rate of adaptive evolution:

$$\frac{\pi}{\pi_0} = \frac{\rho}{\rho + 2N_e s v k}. \quad (1)$$

Here π is the probability that two randomly sampled chromosomes differ at a given nucleotide site, π_0 the value of π in the absence of hitchhiking, ρ the frequency of recombination per nucleotide site for the region under consideration, s the selection coefficient for a selectively favourable mutation, v the rate of occurrence of selectively favourable substitutions per nucleotide site per generation, and k a constant that is approximately equal to 0.075.

By comparing the fit of equation (1) to data of Kindahl and Aquadro (Aquadro *et al.* 1994) on the relation between the coefficient of exchange and DNA-level variability for loci on the third chromosome of *D. melanogaster*, the composite parameter $2N_e s v$ can be estimated. Stephan (1995) showed that this parameter must be approximately equal to 5×10^{-8} , if recurrent hitchhiking events are the cause of reduced variability in regions of low recombination. If we use the value of $N_e = 10^6$, suggested by data on silent site variation at the *Adh* locus (Kreitman 1983), we have $2s v = 5 \times 10^{-14}$. This corresponds to a substitution rate per nucleotide site of 2.5×10^{-12} , if $s = 0.01$,

a reasonable value. With approximately 10^8 bases in the euchromatic portion of the *Drosophila* genome (Lindsley and Zimm 1992), this would correspond to a substitution rate per genome of 2.5×10^{-4} , or one substitution every 4000 generations. In addition the apparent complete absence of variation at the *ci^P* locus on the small, nonrecombining chromosome 4 suggested that the most recent hitchhiking event on this chromosome occurred about 20,000 generations ago, which is consistent with one adaptive substitution every 2000 generations in the *D. melanogaster* genome as a whole (Berry *et al.* 1991; Charlesworth 1992). Given the uncertainties in the values of the relevant parameters, these two estimates of the substitution rate agree quite well.

4. Background selection

The conclusion that adaptive evolution has left a footprint of reduced variability across the *Drosophila* genome was an exciting one. But, as so often in evolutionary theory, it was not long before alternative explanations of these data were proposed. The main alternative to hitchhiking effects of favourable mutations, now often referred to as 'selective sweeps' (Berry *et al.* 1991), is the effect of hitchhiking by deleterious mutations, termed 'background selection' by Charlesworth *et al.* (1993). A neutral variant that is tightly linked to a deleterious mutation in the process of elimination from the population will also have a high chance of being eliminated before it can unhitch itself by recombination. At first sight this process might seem unimportant. A single locus, subject by mutation to an input of deleterious alleles that are promptly eliminated by selection, will have only a very weak effect on variability at a linked neutral site, since strongly deleterious alleles never reach high frequency in a population of even moderate size, so that there is only a low probability that a neutral variant will encounter a deleterious allele at the selected locus.

But there are numerous loci in the genomes of higher organisms that are subject to mutation to deleterious alleles (Crow 1993), so that the relatively weak effects of individual loci may have a large cumulative effect on neutral variability. This is most easily seen by considering a genomic region in which recombination is absent, such as chromosome 4 of *D. melanogaster*. If the population size is large enough, the selected loci can be assumed to be close to their equilibrium under mutation and selection in an infinite population. Let U be the mean number of new deleterious mutations per zygote for the genomic region in question, and let the selection coefficient against heterozygous carriers of a mutation at a given locus be t (for simplicity t is assumed to be the same for all deleterious mutations). With multiplicative fitness interactions among loci, the equilibrium mean number of mutations per individual is $\bar{n} = U/t$, and the equilibrium frequency of haploid genomes that are free of deleterious mutations is $f_0 = \exp(-\bar{n}/2)$ (Kimura and Maruyama 1966). The irreversible nature of the mutation process means that mutation-free chromosomes are continually being transformed into mutation-carrying chromosomes at rate U per generation, and mutation-carrying chromosomes are being eliminated from the population at the same rate.

Because the mean persistence time of deleterious mutations subject to strong selection is only a few generations (Kimura and Ohta 1969), a pair of chromosomes sampled from the population must have been derived from the mutation-free class relatively recently, compared with the mean time of $2N_e$ generations that they would take to coalesce to a common ancestral chromosome in the absence of background

a reasonable value. With approximately 10^8 bases in the euchromatic portion of the *Drosophila* genome (Lindsley and Zimm 1992), this would correspond to a substitution rate per genome of 2.5×10^{-4} , or one substitution every 4000 generations. In addition the apparent complete absence of variation at the *ci^P* locus on the small, nonrecombining chromosome 4 suggested that the most recent hitchhiking event on this chromosome occurred about 20,000 generations ago, which is consistent with one adaptive substitution every 2000 generations in the *D. melanogaster* genome as a whole (Berry *et al.* 1991; Charlesworth 1992). Given the uncertainties in the values of the relevant parameters, these two estimates of the substitution rate agree quite well.

4. Background selection

The conclusion that adaptive evolution has left a footprint of reduced variability across the *Drosophila* genome was an exciting one. But, as so often in evolutionary theory, it was not long before alternative explanations of these data were proposed. The main alternative to hitchhiking effects of favourable mutations, now often referred to as 'selective sweeps' (Berry *et al.* 1991), is the effect of hitchhiking by deleterious mutations, termed 'background selection' by Charlesworth *et al.* (1993). A neutral variant that is tightly linked to a deleterious mutation in the process of elimination from the population will also have a high chance of being eliminated before it can unhitch itself by recombination. At first sight this process might seem unimportant. A single locus, subject by mutation to an input of deleterious alleles that are promptly eliminated by selection, will have only a very weak effect on variability at a linked neutral site, since strongly deleterious alleles never reach high frequency in a population of even moderate size, so that there is only a low probability that a neutral variant will encounter a deleterious allele at the selected locus.

But there are numerous loci in the genomes of higher organisms that are subject to mutation to deleterious alleles (Crow 1993), so that the relatively weak effects of individual loci may have a large cumulative effect on neutral variability. This is most easily seen by considering a genomic region in which recombination is absent, such as chromosome 4 of *D. melanogaster*. If the population size is large enough, the selected loci can be assumed to be close to their equilibrium under mutation and selection in an infinite population. Let U be the mean number of new deleterious mutations per zygote for the genomic region in question, and let the selection coefficient against heterozygous carriers of a mutation at a given locus be t (for simplicity t is assumed to be the same for all deleterious mutations). With multiplicative fitness interactions among loci, the equilibrium mean number of mutations per individual is $\bar{n} = U/t$, and the equilibrium frequency of haploid genomes that are free of deleterious mutations is $f_0 = \exp(-\bar{n}/2)$ (Kimura and Maruyama 1966). The irreversible nature of the mutation process means that mutation-free chromosomes are continually being transformed into mutation-carrying chromosomes at rate U per generation, and mutation-carrying chromosomes are being eliminated from the population at the same rate.

Because the mean persistence time of deleterious mutations subject to strong selection is only a few generations (Kimura and Ohta 1969), a pair of chromosomes sampled from the population must have been derived from the mutation-free class relatively recently, compared with the mean time of $2N_e$ generations that they would take to coalesce to a common ancestral chromosome in the absence of background

selection (Hudson 1994). Since the effective size of the mutation-free class is $f_0 N_e$, the expected time to coalescence of two chromosomes is close to $2f_0 N_e$. Under the infinite-site model for an autosomal locus (Kimura 1971), the expected nucleotide site diversity (given by twice the product of the mutation rate and the mean coalescence time) is thus

$$\pi = 4f_0 N_e v, \quad (2a)$$

where v is the neutral mutation rate per site (Charlesworth *et al.* 1993; Hudson 1994).

More economically we can write

$$\frac{\pi}{\pi_0} \approx f_0, \quad (2b)$$

where π_0 is the value of π under the classical neutral model. Equation (2b) will apply to any measure of diversity that depends on the expected time to coalescence of a random pair of genomes, such as the variance of repeat lengths in microsatellite sequences (Slatkin 1995), and is thus not restricted to the infinite-site model.

If U is sufficiently large in relation to t , these results imply that genetic diversity could be greatly reduced by background selection in large genomic regions where recombination is absent or severely reduced, such as the centromeric regions of the autosomes of *D. melanogaster*, or in asexual or highly selfing populations (Charlesworth *et al.* 1993). In principle, therefore, we may not need to appeal to relatively frequent selective sweeps to explain the observed relation between recombination rate and genetic diversity.

5. Discriminating between explanations of the *Drosophila* data

To determine the extent to which background selection can account for the patterns observed in *Drosophila*, it is necessary to have a theory that takes into account the effects of recombination between a focal locus exhibiting neutral or nearly neutral variability and all loci that are subject to mutation and selection. Theoretical results developed by Hudson (1994), Hudson and Kaplan (1994, 1995) and Nordborg *et al.* (1996) have facilitated the construction of such a theory. For an autosomal locus i with mutation rate u_i from wild-type to mutant alleles with heterozygous selection coefficient t_i , the equilibrium frequency of mutant alleles in a random-mating population is $q_i = u_i/t_i$, provided that $t_i \gg u_i$ (Haldane 1927). Let the recombination frequency between the neutral locus and the i th selected site be r_i . With multiplicative fitness interactions, Nordborg *et al.* (1996) have shown that the steady-state nucleotide site diversity at the neutral locus in a large but finite population is given by

$$\frac{\pi}{\pi_0} \approx \exp \left(- \sum_i \frac{q_i}{[1 + r_i(1 - t_i)/t_i]^2} \right). \quad (3)$$

The summation is taken over all sites subject to mutation and selection where selection is sufficiently strong in relation to drift that allele frequencies are close to equilibrium. Since the mean number of deleterious mutations per individual is equal to $2\sum q_i$, this result reduces to equation (2b) when there is no recombination.

If we were in the fortunate position of knowing the mutation rates and selection coefficients at all sites in the genome, and the values of the r_i for all loci for which data on putatively neutral variation is available, we could use equation (3) to predict the

relative levels of diversity expected in different parts of the genome, and compare the predictions with the data. In reality only approximations to this can be accomplished, as attempted by Hudson and Kaplan (1995) and Charlesworth (1996) for *D. melanogaster*. These attempts involve use of simplifying assumptions about the relations between physical position of loci on a chromosome and rate of recombination in the surrounding regions, from published data on recombination frequencies between genes with known physical locations. In addition it is assumed that loci subject to strong selection are distributed with uniform density over the euchromatic portions of the chromosomes of *D. melanogaster*, which is where most of the genes of functional importance are found. The summation in equation (3) is then replaced by a double integral over the length of the chromosome and over the distribution of selection coefficients against strongly deleterious alleles, in which the mutation rate per locus is replaced by the density of mutations per unit physical length. This density can be estimated from the mutation rate per genome for deleterious alleles, determined from experiments on accumulation of egg-to-adult viability mutations in *Drosophila* (Mukai *et al.* 1972; Ohnishi 1977; Crow and Simmons 1983; Keightley 1994).

Details of the sources of the estimates of the selection parameters used by Hudson and Kaplan (1995) and by Charlesworth (1996) are given in the original papers. It suffices here to say that selection and mutation parameters that are consistent with available data can generate predicted relative values of π for different positions on a chromosome that are quite similar to the patterns observed in *D. melanogaster* (see figure 2 for an example). But the virtual absence of variability at loci near the centromeres, on chromosome 4, and at the tip of the X chromosome (Kreitman and Wayne 1994; Moriyama and Powell 1996) can only be explained by appealing to the extremely small (of the order of 10^{-4}) average selection coefficients estimated for the deleterious effects of naturally occurring transposable-element insertions (Charlesworth 1996). In addition, there are some notable discrepancies between the predictions of the models and the data: for example, the loci *Tl* and *Mlc2* in the distal part of the right arm of chromosome 3 have diversity values that fall well below the theoretical curve (Charlesworth 1996). This may either reflect the effects of selective sweeps in their neighbourhood, or inadequate estimates of recombination rates in this region.

Given the fact that several uncertainties exist concerning the estimates of mutation and selection parameters used in these models, some degree of scepticism over the meaning of this apparently good fit to the data of the predictions of background selection is in order. An alternative method of discriminating between the selective sweep and background selection models is to examine departures of allele frequency distributions from neutral expectation in regions where variability is strongly reduced. Recovery from a selective sweep that eliminates all or most variation at linked sites is slow, and is expected to produce an excess of segregating sites with low frequencies of new variants for a given level of nucleotide site variation (Charlesworth *et al.* 1993). If enough sequence information is gathered, there is statistical power to detect such departures in a sample of realistic size (Braverman *et al.* 1995; Simonsen *et al.* 1995). In contrast, while background selection causes allele-frequency distributions to be skewed in the direction of low-frequency variants, the effect is weak, and is very difficult to detect in a sample from a population (Charlesworth *et al.* 1995). This implies that significant results of tests for departure from neutral expectation of site frequency spectra at loci with reduced variation, such as Tajima's test (Tajima 1989) and Fu and Li's test (Fu and Li 1993), are potential indicators of a causal role of selective sweeps,

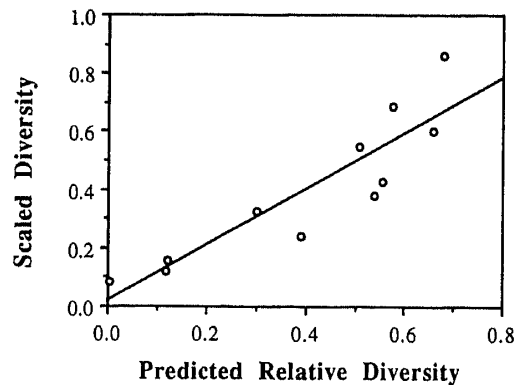


Figure 2. The estimated nucleotide site diversities for the X chromosome genes of figure 1, plotted against the values predicted by the background selection model (Charlesworth 1996). Since the model can only predict relative diversity values, the observed diversities were scaled by dividing by the mean value for the X chromosome, and multiplying by the mean predicted value of π_i/π_0 for the X. This ensures that the mean scaled diversity is equal to the mean predicted relative diversity (Charlesworth 1996). The line is the least-squares linear fit to the points ($r^2 = 0.81$).

whereas a lack of such significance (when the data are adequate to detect deviations) is consistent with background selection.

There are few such indications of departures from neutral expectation for genes in regions of reduced recombination in *D. melanogaster* (Braverman *et al.* 1995; Charlesworth *et al.* 1995), suggesting that selective sweeps have probably not played a major role in causing reduced variation in this species. Of course, this does not necessarily mean that background selection is the causative agent. For example, Gillespie (1994) has shown that a model of temporally fluctuating selection can cause reduced variation at linked sites, without greatly distorting the allele-frequency distribution. It is not known whether this model is capable of explaining the observed relations between recombination rate and diversity in *D. melanogaster*.

6. Selection and variation in bacterial populations

Background selection is likely to be ineffective in bacterial populations owing to the much smaller genome sizes of most microorganisms compared to those of higher eukaryotes. For instance, the commonly studied enteric bacterium *Escherichia coli* has a genome size approximately two orders of magnitude smaller than *Drosophila*'s, and three orders of magnitude smaller than the human genome (Drake 1992; Bird 1995). The smaller genome size of *E. coli* must result in a lower rate of deleterious mutations per genome compared to *Drosophila*, thereby dramatically increasing the equilibrium number of genomes free of deleterious mutations (Drake 1992).

In contrast, bacteria growing in continuous culture commonly show the population dynamics known as periodic selection. Periodic selection refers to the quasiregular and continuous occurrence of selective sweeps (Dykhuizen 1992). It can be observed by following the change in frequency of a neutral marker, which will go through cycles of linear increase and precipitous drops. These cycles are believed to be the result of the regular generation of spontaneous advantageous mutations in large bacterial populations of fixed size. The frequency of the neutral marker increases linearly, as a result of

recurrent mutation, until the population is purged of genetic variation by the rapid increase in frequency of a strain carrying the latest advantageous mutation in a selective sweep. After the fixation, or near-fixation, of this selected strain and the resulting general reduction in variation in the population, recurrent mutation once again increases the frequency of the neutral marker. This model assumes that the population is large enough to avoid genetic drift and that back mutation is negligible.

Periodic selection, originally described in 1951 (Atwood *et al.* 1951a, b), is a common feature of continuous-culture laboratory populations (Dykhuizen and Hartl 1983). It has also been suggested that it plays a primary role in structuring natural bacterial populations (Levin 1981; Milkman and Bridges 1990; Reeves 1992), where it is sometimes referred to as 'epidemic population dynamics' (Maynard Smith *et al.* 1993). Early electrophoretic surveys of natural *E. coli* populations (Milkman 1973, 1975; Selander and Levin 1980) indicated that the level of genetic diversity was substantially lower than would be expected in a species with as large a population size as *E. coli*. More recent data on nucleotide site diversity suggest that the average π for silent sites is approximately 5×10^{-2} in *E. coli* (Whittam and Ake 1993; Hartl *et al.* 1994), only about ten times the value for *D. melanogaster*, which is rather low compared to the values for other *Drosophila* species (Moriyama and Powell 1996). Given a mutation rate of about 4×10^{-10} per nucleotide site in *E. coli* (Drake 1992), equating π to $2N_e v$ gives an N_e of about 6×10^7 . This is consistent with an estimate of $N_e \leq 4 \times 10^8$ from electrophoretic data (Maynard Smith 1991).

Levin (1981) proposed that the discrepancy between the vast census population size of *E. coli* and the significantly smaller effective population size, as inferred from the genetic data, could be due to periodic selection. By modelling the process he concluded that the species *E. coli* is composed of a number of distinct, geographically widespread clones, each with a relatively small effective population size (see also Maynard Smith 1991). This model has recently been extended by Berg (1995, 1996).

Owing to the difficulty of studying evolutionary processes in natural populations of bacteria, the hitchhiking theory went largely untested. Until very recently there has been little direct evidence for the action of periodic selection outside the laboratory. The primary line of support for periodic selection in natural populations, or epidemic population dynamics, came indirectly from multilocus enzyme electrophoresis (MLEE) studies. As allozyme electrophoretic studies accumulated, many bacterial species were found to have relatively few multilocus genotypes occurring at much higher frequencies than expected on the basis of independent assortment. Extremely high levels of linkage disequilibrium between allozymes were observed. Maynard Smith *et al.* (1993) recently compiled and reviewed MLEE data for a wide variety of microorganisms. They measured the extent of multilocus linkage disequilibrium for the purpose of inferring the degree of clonality in bacterial populations. In some instances it was observed that the level of multilocus linkage disequilibrium was significantly higher in the total sample of isolates relative to the level when each electrophoretic type (a clone carrying a specific MLEE profile) was treated as a single individual. In these cases it is believed that a small number of electrophoretic types have recently become extremely abundant and widespread, an indication of epidemic population dynamics or recent selective sweeps.

The only direct genetic evidence of selective sweeps in natural bacterial populations has come from a study of the chromosomal region in and around the *gapA* locus in *E. coli* and *Salmonella*. Guttman and Dykhuizen (1994a) observed that, while both

E. coli and *Salmonella gapA* loci were under roughly the same degree of selective constraint (as indicated by the degree of codon bias), the *E. coli* locus had approximately ten times less genetic variability than the *Salmonella* locus. One possible explanation for this discrepancy was the relatively recent occurrence of a selective sweep in or near the *gapA* locus of *E. coli*. This possibility was investigated by sequencing the *gapA* and the closely linked *pabB* loci in 12 natural *E. coli* isolates and analysing the data for differences in genealogical and selective histories.

These differences can be statistically identified by contrasting the level of intraspecific nucleotide polymorphism with the level of interspecific nucleotide divergence. The neutral theory of molecular evolution (Kimura 1983) predicts that the rate of evolution between two closely related species (divergence), and the rate of evolution within either individual species (polymorphism) should be proportionally the same for all loci with similar genealogical and selective histories. Differing genealogical histories can be inferred when loci are found to have significantly different ratios of divergence to polymorphism. The HKA test (Hudson *et al.* 1987) is a statistical method for testing this prediction. When applied to the *E. coli gapA* and *pabB* data, using the *Salmonella* sequences for the divergence estimates, the HKA test detected no difference between the *gapA* and *pabB* loci, indicating that these two loci had experienced similar genealogical and selective histories. This finding was supported by the extremely high degree of concordance between the *gapA* and *pabB* gene genealogies. HKA comparisons were then made between the *gapA* and *pabB* loci and four other loci dispersed around the *E. coli* genome. These comparisons indicated that the *E. coli gapA* and *pabB* loci had significantly reduced levels of intraspecific polymorphism relative to their divergence from *Salmonella*.

Two important points came out of this work. First, the observations were consistent with a recent selective sweep in or near the *E. coli gapA* and *pabB* loci. Second, the selective sweep did not affect the entire *E. coli* genome as would be expected if the genome were maintained in complete linkage by a strictly clonal population structure. This latter observation raises the question of how much recombination there must be in relation to the selection intensity at the locus undergoing a selective sweep to account for the lack of effect of the selective sweep on the loci that were relatively far from *gapA* and *pabB*. As a rule of thumb we can assume that the recombination frequency must be greater than the selection coefficient if hitchhiking is to have little effect (see section 2). Estimates of the ratio of the recombination frequency per nucleotide site to the mutation rate v give a mean of the order of 5 for *E. coli* (Whittam and Ake 1993). With $v = 4 \times 10^{-10}$, we obtain a recombination rate of about 2×10^{-9} . Given the fact that the surveyed genes that show no effect of the selective sweep are about 200 kb from *gapA* and *pabB* (Guttman and Dykhuizen 1994a), the selection coefficient involved must be less than $200 \times 10^3 \times 2 \times 10^{-9} = 4 \times 10^{-4}$. Using the slightly different mutation and recombination parameters suggested by Guttman and Dykhuizen (1994b), we obtain an estimate of 5×10^{-9} for the recombination rate per nucleotide, which yields an upper bound for the selection coefficient of 10^{-3} .

7. Discussion

The theory and data described above suggest strongly that selection at linked loci may have considerable effects on the amount and pattern of variation at neutral sites.

Especially in species with low-recombination genomes, such as asexual or highly inbreeding species, we may expect to see much less variation than would be expected from census population sizes or that found in related species with breeding systems that permit relatively free recombination. We have already reviewed evidence that variability in *E. coli* is much less than is indicated by its abundance. Evidence for such reduced variation at the DNA level in higher organisms is relatively sparse at present, but allozyme diversity is known to be severely reduced in some species of selfing plants and animals compared with similar outbreeding species (Charlesworth *et al.* 1993; Jarne 1995). In the relatively-low-recombination genome of *D. melanogaster*, background selection is expected to cause up to 30% reduction in variability below classical neutral expectation even in the most freely recombining middle sections of the major chromosomes (Hudson and Kaplan 1995; Charlesworth 1996). Close relatives of *D. melanogaster*, such as *D. simulans* and *D. mauritiana* (Sturtevant 1929; True *et al.* 1996), have substantially higher levels of genetic recombination, particularly in the pericentric regions. It is intriguing to speculate whether this may be a cause of the higher levels of nucleotide site diversity in these species compared to that in *D. melanogaster* (Moriyama and Powell 1996). Organisms such as mammals and flowering plants, with very much higher rates of genetic recombination than *Drosophila*, would not be expected to show such a reduction in variation below the classical neutral value, except in genomic regions with greatly reduced rates of recombination such as the pericentric euchromatin (Nordborg *et al.* 1996).

A number of issues are raised by the results we have discussed, and require further experimental and theoretical work for their resolution.

- (i) Further tests of the respective roles of background selection, selective sweeps and fluctuating selection in generating the patterns seen in *Drosophila* are badly needed. Since there is abundant evidence for deleterious effects of spontaneous mutations (Crow and Simmons 1983), whereas the frequency of adaptively useful variation is an unknown quantity, the background selection model can be viewed as a null hypothesis against which alternatives such as selective sweeps or fluctuating selection can be tested. To have confidence in the predictions of this null hypothesis it is necessary to have better data on the mutation rate per genome to deleterious alleles, and on the distribution of selection coefficients involved. In addition we need predictions of patterns of diversity that take into account the fact that silent sites in *Drosophila* are under some degree of selection to maintain biased codon usage (Akashi 1995).
- (ii) The interpretation of data on bacterial populations is difficult because of the complications due to selective sweeps, extinction and recolonization of local colonies, and adaptations of clones to local environments (Levin 1981; Maynard Smith 1991; Berg 1995, 1996). In addition recombination in bacteria is likely to involve processes such as transformation, conjugation and transduction which involve nonreciprocal transfer of relatively small pieces of DNA (Levin 1988). This means that extrapolations from standard population-genetics models that ignore these complexities may be misleading (Berg 1996). Models that are specifically designed to incorporate them are thus badly needed, as well as further data on the size of the genomic regions over which selective sweeps exert their effects in bacteria.
- (iii) Selective sweeps, background selection and fluctuating selection may affect the rate of evolution of variants that are themselves under selection (Birky and Walsh 1988; Charlesworth 1994a; Peck 1994; Barton 1995). Thus we may expect to see evidence that

natural selection is less efficient at preserving optimal codon usage and amino-acid sequences in regions of reduced recombination. There is evidence that the level of codon bias is reduced in regions of reduced recombination in *D. melanogaster* (Kliman and Hey 1993), and some suggestion that patterns of nucleotide composition in yeast and humans may be related to regional patterns of recombination frequency (Charlesworth 1994b). A complete theory of the effects of selection at linked sites on weakly selected variants remains to be developed, and further empirical evidence on this issue, to parallel the results on patterns of variability, are both needed.

Acknowledgements

The work described here was supported by US PHS Grants GM30201, R01-A132454 and P01-GM-50355-01, NSF Grant DEB-9317683, and by the Darwin Trust of Edinburgh.

References

- Aguadé M., Miyashita N. and Langley C. H. 1989 Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster*. *Genetics* 122: 607–615
- Akashi H. 1995 Inferring weak selection from patterns of polymorphism and divergence at “silent” sites in *Drosophila* DNA. *Genetics* 139: 1067–1076
- Aquadro C. F., Begun D. J. and Kindahl E. C. 1994 Selection, recombination, and DNA polymorphism in *Drosophila*. In *Non-neutral evolution: theories and molecular data* (ed.) B. Golding (London: Chapman and Hall) pp. 46–56
- Ashburner M. 1989 *Drosophila: a laboratory handbook* (Cold Spring Harbor: Cold Spring Harbor Laboratory Press)
- Atwood K. C., Schneider L. K. and Ryan F. J. 1951a Selective mechanisms in bacteria. *Cold Spring Harbor Symp. Quant. Biol.* 16: 345–355
- Atwood K. C., Schneider L. K. and Ryan F. J. 1951b Periodic selection in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 37: 146–155
- Barton N. H. 1995 Linkage and the limits to natural selection. *Genetics* 140: 82–84
- Begun D. J. and Aquadro C. F. 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rate in *Drosophila melanogaster*. *Nature* 356: 519–520
- Berg O. G. 1995 Periodic selection and hitchhiking in a bacterial population. *J. Theor. Biol.* 173: 307–320
- Berg O. G. 1996 Selection intensity for codon bias and the effective population size of *Escherichia coli*. *Genetics* 142: 1379–1382
- Berry A. J., Ajioka J. W. and Kreitman M. 1991 Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. *Genetics* 129: 1111–1117
- Bird A. P. 1995 Gene number, noise reduction and biological complexity. *Trends Genet.* 11: 94–100
- Birky C. W. and Walsh J. B. 1988 Effects of linkage on rates of molecular evolution. *Proc. Natl. Acad. Sci. USA* 85: 6414–6418
- Braverman J. M., Hudson R. R., Kaplan N. L., Langley C. H. and Stephan W. 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphism. *Genetics* 140: 783–796
- Charlesworth B. 1992 New genes sweep clean. *Nature* 356: 475–476
- Charlesworth B. 1994a The effect of background selection against deleterious alleles on weakly selected, linked variants. *Genet. Res.* 63: 213–228
- Charlesworth B. 1994b Patterns in the genome. *Curr. Biol.* 4: 182–184
- Charlesworth B. 1996 Background selection and patterns of genetic diversity in *Drosophila melanogaster*. *Genet. Res.* (in press)
- Charlesworth B., Morgan M. T. and Charlesworth D. 1993 The effect of deleterious mutations on neutral molecular variation. *Genetics* 34: 1289–1303
- Charlesworth D., Charlesworth B. and Morgan M. T. 1995 The pattern of neutral molecular variation under the background selection model. *Genetics* 141: 1619–1632

- Crow J. F. 1993 How much do we know about spontaneous human mutation rates? *Environ. Mol. Mut.* 21: 122–129
- Crow J. F. and Simmons M. J. 1983 The mutation load in *Drosophila*. In *The genetics and biology of Drosophila* (eds.) M. Ashburner, H. L. Carson and J. N. Thompson (London: Academic Press) vol. 3c, pp. 1–35
- Drake J. W. 1992 Mutation rates. *Bioessays* 14: 137–140
- Dykhuizen D. E. 1992 Periodic selection. In *Encyclopedia of microbiology* (San Diego: Academic Press) pp. 351–355
- Dykhuizen D. E. and Hartl D. L. 1983 Selection in chemostats. *Microbiol. Rev.* 47: 150–168
- Fu Y.-X. and Li W.-H. 1993 Statistical tests of neutrality of mutations. *Genetics* 133: 693–709
- Gillespie J. H. 1994 Alternatives to the neutral theory. In *Non-neutral evolution: theories and molecular data* (ed.) B. Golding (London: Chapman and Hall) pp. 1–17
- Guttman D. S. and Dykhuizen D. E. 1994a Detecting selective sweeps in naturally occurring *Escherichia coli*. *Genetics* 138: 993–1003
- Guttman D. S. and Dykhuizen D. E. 1994b Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* 266: 1380–1383
- Haldane J. B. S. 1927 A mathematical theory of natural and artificial selection. Part V. Selection and mutation. *Proc. Cambridge Philos. Soc.* 23: 838–844
- Hartl D. L., Moriyama E. N. and Sawyer S. A. 1994 Selection intensity for codon bias. *Genetics* 138: 227–234
- Hudson R. R. 1994 How can the low levels of DNA sequence variation in regions of the *Drosophila* genome with low recombination rates be explained? *Proc. Natl. Acad. Sci. USA* 91: 6815–6818
- Hudson R. R. and Kaplan N. L. 1994 Gene trees with background selection. In *Non-neutral evolution: theories and molecular data* (ed.) B. Golding (London: Chapman and Hall) pp. 140–153
- Hudson R. R. and Kaplan N. L. 1995 Deleterious background selection with recombination. *Genetics* 141: 1605–1617
- Hudson R. R., Kreitman M. and Aguadé M. 1987 A test of neutral molecular evolution based on nucleotide data. *Genetics* 116: 153–159
- Jarne P. 1995 Mating system, bottlenecks and polymorphism in hermaphroditic animals. *Genet. Res.* 65: 193–207
- Kaplan N. L., Hudson R. R. and Langley C. H. 1989 The “hitch-hiking” effect revisited. *Genetics* 123: 887–899
- Keightley P. D. 1994 The distribution of mutation effects on viability in *Drosophila melanogaster*. *Genetics* 138: 1–8
- Kimura M. 1971 Theoretical foundations of population genetics at the molecular level. *Theor. Popul. Biol.* 2: 174–208
- Kimura M. 1983 *The neutral theory of molecular evolution* (Cambridge: Cambridge University Press)
- Kimura M. and Crow J. F. 1964 The number of alleles that can be maintained in a finite population. *Genetics* 49: 725–738
- Kimura M. and Maruyama T. 1966 The mutational load with epistatic gene interactions in fitness. *Genetics* 54: 1303–1312
- Kimura M. and Ohta T. 1969 The average number of generations until extinction of an individual mutant gene in a population. *Genetics* 63: 701–709
- Kliman R. M. and Hey J. 1993 Reduced natural selection associated with low recombination in *Drosophila melanogaster*. *Mol. Biol. Evol.* 10: 1239–1258
- Kreitman M. 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* 304: 412–417
- Kreitman M. 1991 Detecting selection at the level of DNA. In *Evolution at the molecular level* (eds.) R. K. Selander, A. G. Clark and T. S. Whittam (Sunderland, Mass., USA: Sinauer) pp. 202–221
- Kreitman M. and Wayne M. L. 1994 Organization of genetic variation at the molecular level: lessons from *Drosophila*. In *Molecular ecology and evolution: approaches and applications* (eds.) B. Schierwater, B. Streit, G. P. Wagner and R. DeSalle (Basel: Birkhäuser) pp. 157–184
- Levin B. R. 1981 Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. *Genetics* 99: 1–23
- Levin B. R. 1988 The evolution of sex in bacteria. In *The evolution of sex* (eds.) R. E. Michod and B. R. Levin (Sunderland, Mass., USA: Sinauer) pp. 194–211
- Lindsley D. L. and Zimm G. G. 1992 *The genome of Drosophila melanogaster* (San Diego: Academic Press)
- Maynard Smith J. 1991 The population genetics of bacteria. *Proc. R. Soc. London B245*: 37–41
- Maynard Smith J. and Haigh J. 1974 The hitch-hiking effect of a favourable gene. *Genet. Res.* 23: 23–35

- Maynard Smith J., Smith N. H., O'Rourke M. and Spratt B. G. 1993 How clonal are bacteria? *Proc. Natl. Acad. Sci. USA* 90: 4384–4388
- Milkman R. 1973 Electrophoretic variation in *Escherichia coli* from natural sources. *Science* 182: 1024–1026
- Milkman R. 1975 Allozyme variation of *E. coli* of diverse natural origins. In *Isozymes* (ed.) C. L. Markert (New York: Academic Press) vol. 4, pp. 273–285
- Milkman R. and Bridges M. M. 1990 Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. *Genetics* 126: 505–517
- Moriyama E. N. and Powell J. R. 1996 Intraspecific nuclear DNA variation in *Drosophila*. *Mol. Biol. Evol.* 13: 261–277
- Mukai T., Chigusa S. I., Mettler L. E. and Crow J. F. 1972 Mutation rate and dominance of genes affecting viability in *Drosophila melanogaster*. *Genetics* 72: 335–355
- Nordborg M., Charlesworth B. and Charlesworth D. 1996 The effect of recombination on background selection. *Genet. Res.* 67: 159–174
- Ohnishi O. 1977 Spontaneous and ethyl methanesulfonate-induced mutations controlling viability in *Drosophila melanogaster*. II. Homozygous effects of polygenic mutations. *Genetics* 87: 529–545
- Ohta T. 1971 Associative overdominance caused by linked detrimental mutations. *Genet. Res.* 18: 277–286
- Ohta T. 1973 Effect of linkage on behaviour of mutant genes in finite populations. *Theor. Popul. Biol.* 4: 145–172
- Ohta T. and Kimura M. 1975 The effect of a selected locus on heterozygosity of neutral alleles (the hitch-hiking effect). *Genet. Res.* 25: 313–326
- Peck J. 1994 A ruby in the rubbish: beneficial mutations, deleterious mutations, and the evolution of sex. *Genetics* 137: 597–606
- Reeves P. R. 1992 Variation in O-antigens, niche-specific selection, and bacterial populations. *FEMS Microbiol. Lett.* 100: 509–516
- Selander R. K. and Levin B. R. 1980 Genetic diversity and structure in *Escherichia coli* populations. *Science* 210: 545–547
- Simonsen K. L., Churchill G. A. and Aquadro C. F. 1995 Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* 141: 413–429
- Slatkin M. 1995 Hitchhiking and associative overdominance at a microsatellite locus. *Mol. Biol. Evol.* 12: 473–480
- Stephan W. 1995 An improved method for estimating the rate of fixation of favorable mutations based on DNA polymorphism data. *Mol. Biol. Evol.* 12: 959–962
- Stephan W. and Langley C. H. 1989 Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila ananassae* populations. I. Contrasts between the *vermillion* and *forked* loci. *Genetics* 121: 89–99
- Stephan W., Wiehe T. H. E. and Lenz M. W. 1992 The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *Theor. Popul. Biol.* 41: 237–254
- Sturtevant A. H. 1929 The genetics of *Drosophila simulans*. *Carnegie Inst. Wash. Publ.* 399: 1–62
- Sved J. A. 1972 Heterosis at the level of the chromosome and at the level of the gene. *Theor. Popul. Biol.* 3: 491–506
- Tajima F. 1989 Statistical method for testing the neutral mutation hypothesis. *Genetics* 123: 585–595
- Thomson G. 1977 The effect of a selected locus on linked neutral loci. *Genetics* 85: 753–788
- True J. R., Mercer J. M. and Laurie C. C. 1996 Differences in crossover frequency and distribution among three sibling species of *Drosophila*. *Genetics* 142: 507–523
- Whittam T. S. and Ake S. E. 1993 Genetic polymorphisms and recombination in natural populations of *E. coli*. In *Mechanisms of molecular evolution* (eds.) N. Takahata and A. G. Clark (Sunderland, Mass., USA: Sinauer) pp. 223–245
- Wiehe T. H. E. and Stephan W. 1993 Analysis of a genetic hitchhiking model and its application to DNA polymorphism data. *Mol. Biol. Evol.* 10: 842–854