

Variation After a Selective Sweep in a Subdivided Population

Enrique Santiago^{*,1} and Armando Caballero[†]

^{*}Departamento de Biología Funcional, Facultad de Biología, Universidad de Oviedo, 33071 Oviedo, Spain and [†]Departamento de Bioquímica, Genética e Inmunología, Facultad de Biología, Universidad de Vigo, Campus Universitario, 36310 Vigo, Spain

Manuscript received July 12, 2004

Accepted for publication October 8, 2004

ABSTRACT

The effect of genetic hitchhiking on neutral variation is analyzed in subdivided populations with differentiated demes. After fixation of a favorable mutation, the consequences on particular subpopulations can be radically different. In the subpopulation where the mutation first appeared by mutation, variation at linked neutral loci is expected to be reduced, as predicted by the classical theory. However, the effect in the other subpopulations, where the mutation is introduced by migration, can be the opposite. This effect depends on the level of genetic differentiation of the subpopulations, the selective advantage of the mutation, the recombination frequency, and the population size, as stated by analytical derivations and computer simulations. The characteristic outcomes of the effect are three. First, the genomic region of reduced variation around the selected locus is smaller than that predicted in a panmictic population. Second, for more distant neutral loci, the amount of variation increases over the level they had before the hitchhiking event. Third, for these loci, the spectrum of gene frequencies is dominated by an excess of alleles at intermediate frequencies when compared with the neutral theory. At these loci, hitchhiking works like a system that takes variation from the between-subpopulation component and introduces it into the subpopulations. The mechanism can also operate in other systems in which the genetic variation is distributed in clusters with limited exchange of variation, such as chromosome arrangements or genomic regions closely linked to targets of balancing selection.

IT is generally accepted that the spread of an advantageous mutation reduces the genetic variation at linked neutral loci: the mutation drags linked alleles in its way to fixation, and most of the original variation is eliminated. The magnitude of the effect depends on the recombination rate and the selective value of the mutation (MAYNARD-SMITH and HAIGH 1974; WIEHE and STEPHAN 1993). After this selective sweep, most of the neutral alleles at closely linked loci are lost and, afterward, the neutral variation is recovered very slowly by mutation. During this time, the spectrum of gene frequencies is dominated by rare alleles until, if there is enough time without any other “disturbance,” a new mutation-drift balance is reached. These predictions have been widely used to discriminate between selective sweep and other selective models. Particularly, the background selection model (CHARLESWORTH *et al.* 1993) predicts a reduction of variability with a distribution of gene frequencies following a nearly neutral spectrum, and balancing selection models predict an increase in neutral variation that is represented by genes at intermediate frequencies. However, most of the previous theory of hitchhiking considers only the effect on a single panmictic population. Natural populations, in general, are

more appropriately described as an arrangement of partially differentiated subpopulations. Here we show that the effect of a selective sweep on the neutral variation of a subdivided population can be very different from that predicted by the previous theory. Under particular combinations of recombination frequency, selective values, and population subdivision, the genetic variability increases at loci linked with the selected gene. At these loci, an excess of genes at intermediate frequencies is expected.

THE BASIC THEORY

We assume a model of two subpopulations with $2N$ monoecious haploid individuals each, which is equivalent to N diploid individuals. The number of individuals is constant over discrete generations and there are no extinction-recolonization events. Every generation, a proportion m of individuals migrates from each subpopulation to the other one and there is a random association of individuals within subpopulations to accomplish sexual reproduction; *i.e.*, pairs of random individuals are temporarily combined in different meiosis to generate individuals of the next generation. We also consider a neutral locus for which mutation follows an infinite-allele model, so that every mutation creates a new allele not present before in the population. As a consequence of the isolation, there is a genetic correlation F_{ST} (WRIGHT

¹Corresponding author: Departamento de Biología Funcional, Facultad de Medicina, Universidad de Oviedo, 33071 Oviedo, Spain.
E-mail: esr@uniovi.es

1951) between gene copies at the neutral locus within subpopulations. The value of this correlation can be given as

$$F_{ST} = \frac{\theta_2 - \theta_3}{1 - \theta_3}$$

(COCKERHAM 1969, 1973), where θ_2 is the average probability of identity by descent for pairs of genes sampled within subpopulations and θ_3 is the probability of identity by descent for pairs of genes sampled from different subpopulations, respectively. The term θ_1 , normally used to indicate the genetic identity within individuals, has no meaning here as individuals are haploid.

Because mutation generates new alleles, identity by descent and identity by state are equivalent. The expected heterozygosities within subpopulations (H_s) and for the whole population (H_t) can be expressed as functions of the probabilities of identity,

$$H_s = 1 - \theta_2,$$

$$H_t = 1 - \frac{\theta_2 + \theta_3}{2}.$$

A single copy of a favorable allele a occurs by mutation at a linked locus in one of the two subpopulations. We refer to this subpopulation as the “first subpopulation.” At this moment, neutral variation is not necessarily at mutation-migration-drift balance. The selective coefficient of the new mutation is s and the recombination frequency between the selected and the neutral locus is r . With time, the mutation will be lost or fixed in the first subpopulation. As we are interested in studying the effect of hitchhiking on variation, only the populations in which the mutation is fixed are considered. Eventually, the favorable mutation is passed by migration to the other subpopulation (the “second subpopulation”) and fixed in both subpopulations.

The effect of fixation of a favorable mutation on existing heterozygosity at a neutral linked locus in a single population is given by Equation 19 of STEPHAN *et al.* (1992). If the migration rate is not too large, this equation is also applicable to predict the heterozygosity in the first subpopulation after fixation. This is because most of the small amounts of variability introduced from the second to the first subpopulation when the mutation goes to fixation will be swept by the hitchhiking effect if r is small. Although this article deals with the effect of hitchhiking on the whole population and, particularly, on the second subpopulation, we initially consider the effect of the selective sweep on heterozygosity in the first subpopulation.

Let c be the neutral copy originally associated with a when this first appears by mutation as a single copy. Let q_i be the frequency of allele a in the first subpopulation i generations after mutation (*i.e.*, $q_0 = 1/2N$) and p_i be the proportion of chromosomes carrying copies of c

within the subset of chromosomes carrying copies of a . Obviously, $p_0 = 1$ and its value decreases due to recombination until a is fixed in the subpopulation at generation f . At this generation, a proportion p_f of the copies at the neutral locus will be replicates of the original neutral copy c .

As the mutation rate at the neutral locus is too small to affect variability when the selected gene is segregating, the identity by descent at the neutral locus after the hitchhiking process in the first subpopulation can be approximated in the following way. Two genes randomly taken from the subpopulation have a probability p_f^2 of being identical copies of c . Accordingly, the probability of one copy of c and one copy of another ancestral gene at generation 0 is $2p_f(1 - p_f)$, the expected identity between these copies being that before the selective sweep (θ_2). Finally, the probability of none of the two copies coming from c is $(1 - p_f)^2$, the probability of identity of these copies (x) being larger than the initial identity θ_2 because of the drift process during the selective sweep. Averaging over the three values, the expected identity in the first population after the selective sweep (θ'_2) is

$$E(\theta'_2) = 1E(p_f^2) + \theta_2E(2p_f(1 - p_f)) + E(x(1 - p_f)^2).$$

If the subpopulation is large and the proportion $(1 - p_f)$ is not small, then genetic drift will not increase in a significant way the original θ_2 value within the set of chromosomes that do not carry c . Therefore, x will be close to θ_2 . In contrast, if $(1 - p_f)$ is small the x value will be larger than the original θ_2 , but the third term of the equation will be negligible. Therefore, a simplification can be obtained by substituting the third term for $\theta_2E((1 - p_f)^2)$ and, after rearrangement,

$$E(\theta'_2) = \theta_2 + (1 - \theta_2)E(p_f^2).$$

Now, the prediction is extended to the second subpopulation. Here we assume that one single copy of the favorable mutation is transferred from the first to the second subpopulation. In the following generations, selection increases the frequency of the neutral copy c coming from the first subpopulation. As the mutation goes to fixation, recombination tends to remove the association with this neutral copy, and the expected values of p_f and p_f^2 are the same as in the first subpopulation. Two genes randomly taken from the second subpopulation have a probability p_f^2 of being identical copies of c . The probability of one copy of c and one copy of another ancestral gene at generation 0 is $2p_f(1 - p_f)$, the expected identity between these copies being the identity between subpopulations before the selective sweep (θ_3). Finally, the probability of none of the two copies coming from c is $(1 - p_f)^2$ and the probability of identity of these copies is θ_2 . Averaging these three identities we obtain the expected identity after fixation in the second subpopulation,

TABLE 1

Average simulated (sim) and predicted (E) values of p_f and p_f^2 for different combinations of population size (N), recombination frequency (r), and selective coefficient (s) of the favorable mutation

$2N$	r	s	sim p_f	$E(p_f)$	(A1)	sim p_f^2	$E(p_f^2)$	(A2)
2×10^4	0.0005	0.005	0.5789	0.5567	0.5694	0.4118	0.3785	0.3289
10^5	0.0007	0.02	0.7299	0.7335	0.7353	0.5719	0.5771	0.5417
10^4	0.01	0.1	0.4618	0.4422	0.4523	0.2596	0.2388	0.2075
10^5	0.02	0.1	0.1300	0.1234	0.1354	0.0253	0.0227	0.0193
10^6	0.01	0.1	0.2716	0.2790	0.2854	0.0916	0.0951	0.0826
10^5	0.001	0.2	0.9394	0.9457	0.9457	0.8950	0.9034	0.8944
10^4	0.01	0.2	0.6378	0.6423	0.6457	0.4504	0.4560	0.4185
10^4	0.02	0.2	0.4136	0.4126	0.4220	0.2114	0.2079	0.1807
10^5	0.05	1	0.4397	0.5282	0.5310	0.2153	0.3084	0.2830

(A1) and (A2) are predictions of p_f and p_f^2 using Equation 17 from OTTO and BARTON (1997) and Equation 4 from BARTON (1998), respectively.

$$E(\theta_2'') = 1E(p_f^2) + \theta_3 E(2p_f(1 - p_f)) + \theta_2 E((1 - p_f)^2) \\ = \theta_2 + E(p_f^2)(1 + \theta_2 - 2\theta_3) - 2E(p_f)(\theta_2 - \theta_3).$$

As the third term in the right-hand side of the equation is negative, for some combinations of $E(p_f)$, $E(p_f^2)$, θ_2 , and θ_3 , it is possible for the identity by descent to decrease in the second subpopulation after the selective sweep. In general this is expected when

$$F_{ST} > \frac{E(p_f^2)}{2E(p_f) - E(p_f^2)}.$$

Finally, after the selective sweep, the expected value of the identity θ_3' of two genes, one from the first subpopulation and the other from the second subpopulation, is the average of the probabilities of sampling the second subpopulation from the area p_f or the area $(1 - p_f)$,

$$E(\theta_3') = E(p_f)\theta_2' + (1 - E(p_f))\theta_3.$$

OTTO and BARTON (1997, Equation 17) and BARTON (1998, Equation 4) give approximations for $E(p_f)$ and $E(p_f^2)$, which are, respectively, the expected change in allele frequency of the neutral allele initially linked to the beneficial mutation and the probability of coalescence with the original neutral allele linked to the beneficial mutation in the first generation. The approximation for $E(p_f^2)$ clearly underestimates the true value and turns out to be too close to $E(p_f)^2$ (see Table 1). This is probably because the only source of variance considered was the variation in times to fixation of the favorable allele. In the APPENDIX we develop simple alternative predictions for $E(p_f)$ and $E(p_f^2)$, considering the effect of drift and using a combination of deterministic and diffusion approaches. $E(p_f)$ has a weak dependence on N , and $E(p_f^2)$ moves away from $E(p_f)^2$ as s becomes smaller:

$$E(p_f) = (7Ns)^{-r/s},$$

$$E(p_f^2) = (2.6Ns)^{-2(r/s)}.$$

Substituting these values in the equations above, predictions for $E(\theta_2')$, $E(\theta_2'')$, and $E(\theta_3')$ can be made for any combination on N , r , and s .

ANALYSIS OF EXPRESSIONS

Predictions of variability in the second subpopulation are quite different depending on the structure of the population (Figure 1). In a single panmictic population (we could say that the population is equivalent to a set of subpopulations with a high migration rate and, therefore, $F_{ST} = 0$), hitchhiking will reduce the variability at linked neutral loci. The reduction will be large in a chromosome region closely linked to the selected gene (for $r/s = 0.05$ the variability is halved; see Figure 1). This is also the expectation for the first subpopulation

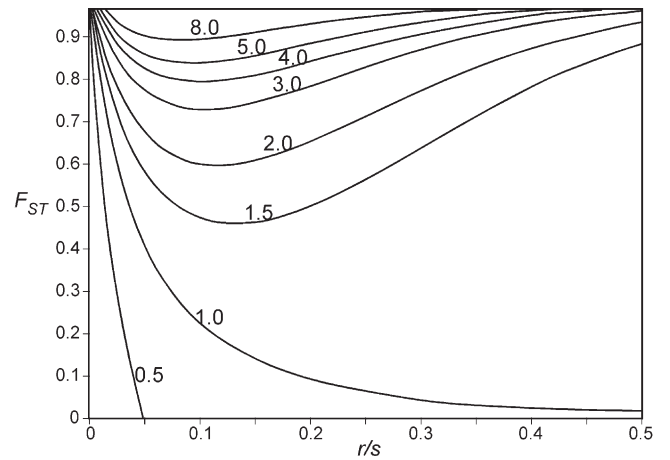


FIGURE 1.—Expected variability in the second subpopulation after hitchhiking ($2Ns = 1000$) as a function of r/s (where r is the recombination frequency and s is the selection coefficient of the mutation) and F_{ST} (the level of differentiation). Labels on level lines represent the proportion of the original variability maintained after hitchhiking.

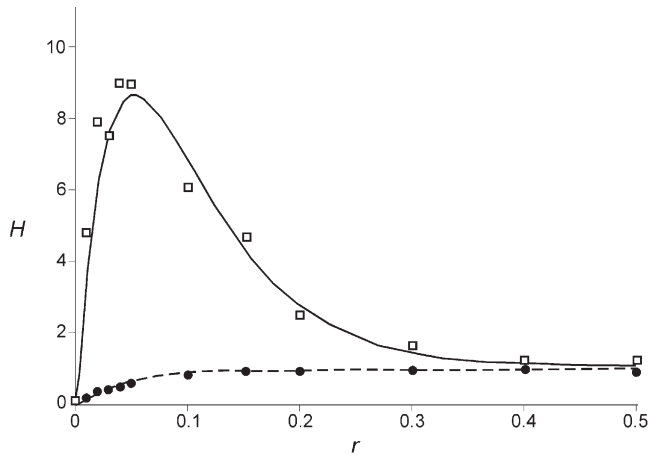


FIGURE 2.—Proportion of the original variation within subpopulations after hitchhiking given as a function of the recombination frequency (r) between the selected and the neutral genes. Lines are predictions of the model for the first (dashed line) and the second (solid line) subpopulations ($2N = 1000$, $s = 0.5$). Circles and squares are simulated values for the first and the second subpopulations, respectively, with parameters $2N = 1000$, $s = 0.5$, $m = 0.00001$, and $\mu = 0.0000112$. This combination of parameters yields an F_{ST} value of 0.908 before the sweep.

in our model for any value of F_{ST} . However, the effect on the second subpopulation can be very different when $F_{ST} > 0$. The genome region affected by the decline of variability is narrowed around the selected gene; *i.e.*, the larger the magnitude of the genetic divergence between subpopulations, the smaller the region of reduced variability. Additionally, if the divergence of the subpopulations is large, the genetic variability in the second subpopulation could even increase over the previous level before the hitchhiking. This increase affects a region in the vicinity of the selected gene but not very close.

Figure 2 shows the effect of a selective sweep on the variability of both subpopulations for neutral genes at different distances from the selected gene. As the effect on the first subpopulation is equivalent to the effect on a single panmictic population, the comparison of the predictions for both subpopulations reveals the great difference between the predictions for a single or a subdivided population. The genetic variability for neutral loci increases as the genetic distance decreases and, only for loci closely linked with the selected gene, the genetic variability drops below the original level.

SIMULATIONS

Predictions were checked by Monte Carlo simulations. A population with two random-mating subpopulations of 1000 haploid individuals each was reproduced under constant rates of migration and neutral mutation. The neutral loci were distributed over a genome represented by one single chromosome. After 10^5 generations it was considered that the neutral loci were practically

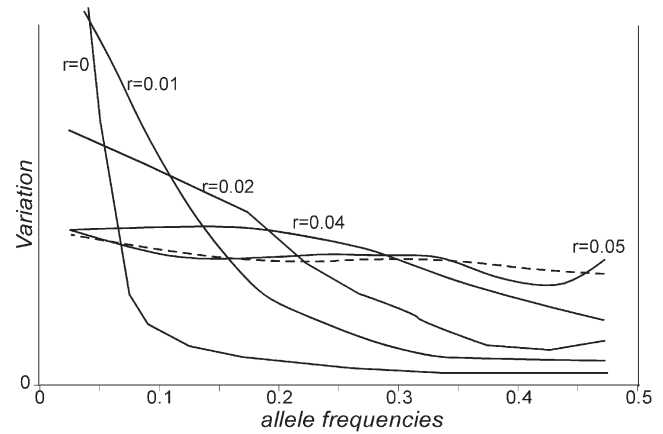


FIGURE 3.—Contribution of neutral loci at different frequencies to the variation in the first subpopulation. The dashed line represents the contribution of loci before hitchhiking. Solid lines are the contributions of loci at different distances (r) from the selected gene. $2N = 1000$, $m = 0.00001$, $s = 0.5$, and $\mu = 0.0000112$.

at mutation-migration-drift equilibrium (note that this equilibrium is not necessarily assumed in the derivations). Thereafter, a single copy of a favorable mutation was assigned to a chromosome of an individual randomly taken from the first subpopulation. The frequency of the new mutation increased until it was fixed in the first subpopulation (simulations in which the favorable mutation was lost were discarded). As individuals migrated between subpopulations, copies of the favorable mutation were eventually transferred and fixed in the second subpopulation too. Then, a number of parameters were computed for neutral loci at different distances from the selected gene: heterozygosity, F_{ST} values, and the distribution of the spectrum of variability in both subpopulations. At the end, 5000 simulations were carried out for each particular combination of values of population size, selection coefficient, recombination, and migration rates. Figure 2 represents an example of the agreement between the observed heterozygosities in both subpopulations after hitchhiking and the predictions of the model.

Figures 3 and 4 show simulation results of the contributions of neutral loci at different frequencies to the variation of the first and second subpopulations, respectively. Before the selective sweep (broken line), the distribution of neutral variability is nearly uniform over the whole span of allele frequencies. There are few loci with alleles at intermediate frequencies, but they contribute as much variation as the large number of loci with alleles at low frequencies. In other words, if the range of allele frequencies from 0 to 1 were split into equal segments of allele frequencies and loci were assigned to segments according to their frequencies, all the segments would contribute nearly the same to the variation. This is the expectation of the neutral model (see GALE 1990, Chap. 9). There is a deviation from

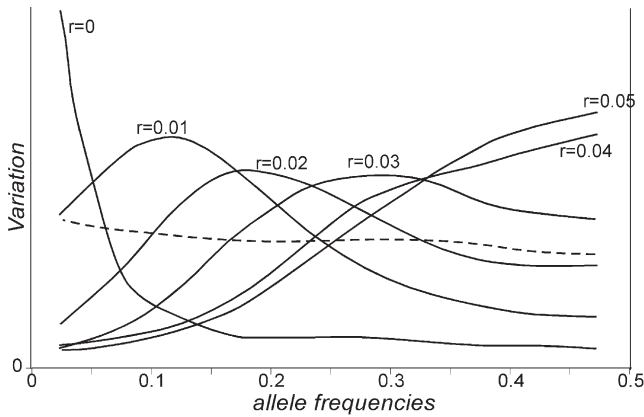


FIGURE 4.—Contribution of neutral loci at different frequencies to the variation in the second subpopulation. The dashed line represents the contribution of loci before hitchhiking. Solid lines are the contributions of loci at different distances (r) from the selected gene. $2N = 1000$, $m = 0.00001$, $s = 0.5$, and $\mu = 0.0000112$.

the uniform distribution due to migration, but this deviation is relatively small when compared with the effect of hitchhiking. After the selective sweep, the distribution of variability of neutral loci is very different in both subpopulations. The effect in the first subpopulation (Figure 3) is equivalent to that expected in a single panmictic subpopulation; *i.e.*, there is a decrease in heterozygosity for neutral linked loci and this reduction follows a characteristic pattern: neutral genes at intermediate frequencies contribute to variation less than expected in a neutral model, and genes at extreme frequencies contribute proportionally more than expected (see Figure 3). This effect is larger as linkage with the selected gene gets closer. TAJIMA'S (1989) D and other statistics were designed to detect this distortion of the spectrum of gene frequencies.

The effect on the second subpopulation is quite different (Figure 4). For the chromosome region with the largest increase in genetic variance after hitchhiking ($r = 0.05$ for the combination of parameters given in Figure 2), there is a maximum contribution of genes at frequencies ~ 0.5 and the set of genes at low frequencies contributes less to variation (see Figure 4). As the neutral gene becomes closer to the selected gene, the maximum contribution moves from 0.5 to lower frequencies. For a wide range of recombination frequencies, the maximum contribution corresponds to neither intermediate nor low frequencies, but to frequencies ~ 0.25 .

To assess the effect of this peculiar distortion of the spectrum of gene frequencies on the capability of Tajima's D to detect hitchhiking, 5000 simulations were carried out in the same way as before. On each simulation, D was computed from samples of 1000 individuals from each subpopulation for chromosome segments at different distances from the selected gene. Before the hitchhiking, 11% of the samples gave significant values

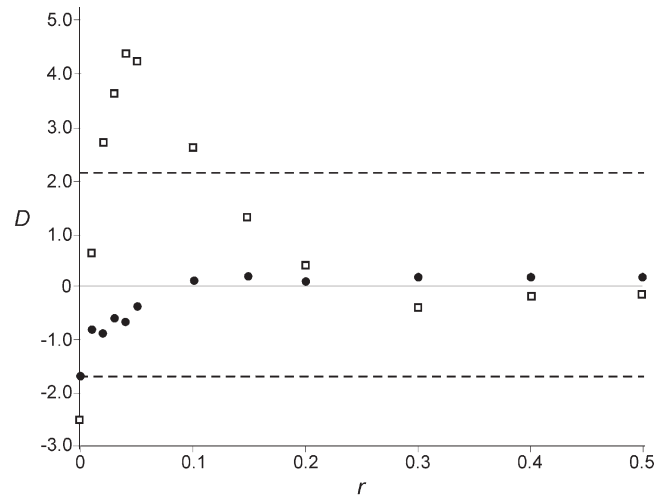


FIGURE 5.—Average Tajima's D values from simulations ($2N = 1000$, $m = 0.00001$, $s = 0.5$, and $\mu = 0.0000112$) after hitchhiking in the first (circles) and in the second (squares) subpopulations. Each of the points represents an average of 5000 simulations. Horizontal dashed lines correspond to the upper and the lower limits of confidence at the 5% level.

of D at the 5% level (critical values from Table 2 of TAJIMA 1989), about half of them over the upper limit of confidence and half below the lower one. This means that, although the populations were run for enough time to reach the neutral equilibrium, genetic drift and population structure cause deviations from the expectation of Tajima's model, increasing the proportion of replicates with D values falling out of the limits of confidence.

As expected after hitchhiking in the first subpopulation, the number of samples with significant D values below the confidence limit increased for close linkage (say $r/s < 0.1$), but the average D value was always over the lower limit of confidence (see Figure 5). The increase of variation in the second subpopulation when linkage was not very tight was associated with an increase in the number of significant D values over the confidence limit. For the combination of parameters given in Figure 5, the increase in the average value of D affected a broad region around the selected gene.

DISCUSSION

The fundamental conclusion of our analysis is that hitchhiking in a subdivided population can lead to an increase in neutral variation at particular subpopulations. Although we have considered a simple model with two subpopulations, this conclusion can be extended to models with any number of differentiated subpopulations. An excess of neutral loci with alleles at intermediate frequencies accompanies the increase in variation. These predictions are made for chromosome segments, which are expected to have reduced variability under the simple hitchhiking model. A reduction in genetic

variance is expected only for a narrow section of chromosome around the selected gene. The larger the F_{ST} value between subpopulations is before hitchhiking, the smaller the section with decreased variability is for subpopulations other than that where the favorable allele first appeared by mutation. The effect can be seen as a consequence of the increase in the rate of effective migration: the selected gene enforces the “migration of linked haplotypes” between differentiated subpopulations, causing increases in diversity under very low migration rates.

To formulate predictions, we have derived approximations for the expected value of p_j , which is the frequency after hitchhiking of copies derived from the original copy associated with the favorable mutation, and for $E(p_j^2)$. Although $E(p_j^2)$ is not conceptually equal to the identity generated by the hitchhiking process (because p_j^2 does not include genetic drift within the set of neutral copies different from c), their values are very close and the equation for $E(p_j^2)$ has a similar form to the equation for identity of BARTON (1998, Equation 13). The reason is that almost all the identity generated by the selective sweep is due to the increase in frequency of the allele originally associated with the selected mutation.

It is difficult to assess the extent to which the effect predicted with our model is responsible for the distribution of neutral variation in natural populations. Most of the published research finds reduced diversity in chromosome regions with low recombination (DEPAULIS *et al.* 1999, 2000; LANGLEY *et al.* 2000; YI and CHARLESWORTH 2000; and review by ANDOLFATTO 2001). Although there is a negative correlation between the recombination rate and the amount of variation contributed by rare alleles in *Drosophila melanogaster* (ANDOLFATTO and PRZEWORSKI 2001), the number of publications reporting significant and negative Tajima’s D values in regions of low recombination is small in other species. HAMBLIN and AQUADRO (1996) reported high levels of variability in regions of low recombination in *D. simulans*. NACHMAN and CROWELL (2000) reported differences in the level of variation and the spectrum of gene frequencies between closely linked regions. These results can be explained by our model of local differentiation and selective sweeps affecting the whole population. In other cases there are evidences of a recent hitchhiking with reduction of variation, but there are not significant deviations from the spectrum of gene frequencies (HAMBLIN and RIENZO 2000; VIEIRA *et al.* 2001, PAYSEUR and NACHMAN 2002). These observations can be explained by a combination of local sweeps, which reduce the neutral variability, move the spectrum of gene frequencies toward an excess of rare alleles, and increase the differentiation of subpopulations; and sporadic events of selective sweeps affecting the whole population, which increase the variation within subpopulations and move the spectrum in the opposite direction.

In our model, hitchhiking behaves like a system that takes variation from the between-subpopulation component and injects it into the within-subpopulation one. Thus, F_{ST} is expected to be reduced after hitchhiking and all our simulations confirm this statement. For example, in a simulation of 5000 replicates, F_{ST} drops from 0.90 to 0.18 when $r = 0.001$, $m = 0.00001$, $s = 0.5$, $\mu = 0.0000112$, and $2N = 1000$ individuals in each subpopulation. These observations seem to be in contradiction with SLATKIN and WIEHE’s (1998) prediction of an increase of F_{ST} after hitchhiking. The disagreement is a consequence of the differences in the initial assumptions of both models. In our model, particular migration and mutation rates are responsible for both the differentiation between subpopulations before hitchhiking and the transfer of the favorable allele from one subpopulation to another. In Slatkin and Wiehe’s model, there is neither mutation nor drift, and all subpopulations have the same allele frequencies before hitchhiking ($F_{ST} = 0$). Their only parameter for migration is that one single copy of the favorable allele is passed from subpopulation to subpopulation. Therefore, any random process (including hitchhiking) will certainly produce an increase in F_{ST} . Another difference is the way of computation of the expected F_{ST} values in both models. Slatkin and Wiehe compute these expectations as $1 - (\overline{H}_s/\overline{H}_t)$, where the bar represents the unweighted average over replicates. This is equivalent to averaging F_{ST} values weighted by their corresponding H_t , the common procedure for averaging estimates of F_{ST} over loci (REYNOLDS *et al.* 1983; WEIR and COCKERHAM 1984). In contrast, we obtain the expected value F_{ST} as the unweighted average over replicates. We think that this averaging method is more appropriate for the design of experiments measuring hitchhiking, as these deal with a particular genome region and a particular set of subpopulations. Thus, the interest is focused on the expected distribution of F_{ST} values for a specific region under the effect of hitchhiking rather than on estimating an averaged differentiation over the whole genome.

Although we have focused on the effect of hitchhiking in subdivided populations, the mechanism can operate on any genetic system where the variation is structured in clusters. For example, the distribution of genetic variation in *Drosophila* is often correlated with chromosome arrangements. Gene flow between karyotypes is strongly restricted and the sporadic events of recombination or gene conversion would be equivalent to migration in our two-subpopulations model. Any favorable mutation would sweep the karyotype where it first appeared and, eventually, will be propagated to the other chromosomal types. Here, karyotypes are equivalent to subpopulations. Hitchhiking could explain the observed differences in the amount of variation of different chromosome arrangements and the nonnegative D values could be evidences of recent selective sweeps (DEPAULIS *et al.* 1999; VIEIRA *et al.* 2001). Gene variation

is also structured around loci with balanced polymorphism: if the polymorphism has been maintained for a long time, a genetic differentiation of haplotypes associated with balanced alleles is expected. Selective sweeps caused by other genes linked to the balanced polymorphism could probably explain the observed differences in the levels and spectra of variation associated with the alleles of the balanced site (BALAKIREV *et al.* 2002).

We are grateful to two referees for useful comments. This work was supported by Universidad de Vigo, Ministerio de Educación y Cultura (PB1998-0814-C03-02) and Ministerio de Ciencia y Tecnología, and Fondos Feder (BMC2003-03022).

LITERATURE CITED

- ANDOLFATTO, P., 2001 Adaptative hitchhiking effects on genome variability. *Curr. Opin. Genet. Dev.* **11**: 635–641.
- ANDOLFATTO, P., and M. PRZEWSKI, 2001 Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. *Genetics* **158**: 657–665.
- BARTON, N. H., 1998 The effect of hitch-hiking on neutral genealogies. *Genet. Res.* **72**: 123–133.
- BALAKIREV, E. S., E. I. BALAKIREV and F. J. AYALA, 2002 Molecular evolution of the *Est-6* gene in *Drosophila melanogaster*: contrasting patterns of DNA variability in adjacent functional regions. *Gene* **282**: 167–177.
- CHARLESWORTH, B., M. T. MORGAN and D. CHARLESWORTH, 1993 The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289–1303.
- COCKERHAM, C. C., 1969 Variance of gene frequencies. *Evolution* **23**: 72–84.
- COCKERHAM, C. C., 1973 Analyses of gene frequencies. *Genetics* **74**: 679–700.
- DEPAULIS, F., L. BRAZIER and M. VEUILLE, 1999 Selective sweep at the *Drosophila melanogaster* *Suppressor of Hairless* locus and its association with the *In(2L)t* inversion polymorphism. *Genetics* **152**: 1017–1024.
- DEPAULIS, F., L. BRAZIER, S. MOUSSET, A. TURBE and M. VEUILLE, 2000 Selective sweep near the *In(2L)t* inversion breakpoint in an African population of *Drosophila melanogaster*. *Genet. Res.* **76**: 149–158.
- GALE, J. S., 1990 *Theoretical Population Genetics*. Unwin Hyman, London.
- HAMBLIN, M. T., and C. F. AQUADRO, 1996 High nucleotide sequence variation in a region of low recombination in *Drosophila simulans* is consistent with the background selection model. *Mol. Biol. Evol.* **13**: 1133–1140.
- HAMBLIN, M. T., and A. RIENZO, 2000 Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am. J. Hum. Genet.* **66**: 1669–1679.
- LANGLEY, C. H., B. P. LAZZARO, W. PHILLIPS, E. HEIKKINEN and J. M. BRAVERMAN, 2000 Linkage disequilibria and the site frequency spectra in the *su(s)* and *su(w^o)* regions of the *Drosophila melanogaster* X chromosome. *Genetics* **156**: 1837–1852.
- MAYNARD-SMITH, J., and J. HAIGH, 1974 The hitchhiking effect of a favorable gene. *Genet. Res.* **23**: 23–35.
- NACHMAN, M. W., and S. L. CROWELL, 2000 Contrasting evolutionary histories of two introns of the Duchenne muscular dystrophy gene, *Dmd*, in humans. *Genetics* **155**: 1855–1864.
- OTTO, S. P., and N. H. BARTON, 1997 The evolution of recombination: removing the limits to natural selection. *Genetics* **147**: 879–906.
- PAYSEUR, B. A., and M. W. NACHMAN, 2002 Natural selection at linked sites in humans. *Gene* **300**: 31–42.
- REYNOLDS, J., B. S. WEIR and C. C. COCKERHAM, 1983 Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* **105**: 767–779.
- SLATKIN, M., and T. WIEHE, 1998 Genetic hitch-hiking in a subdivided population. *Genet. Res.* **71**: 155–160.
- STEPHAN, W., T. H. E. WIEHE and M. W. LENZ, 1992 The effect of

strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *Theor. Popul. Biol.* **41**: 237–254.

- TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.
- WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.
- VIEIRA, J., B. F. MCALLISTER and B. CHARLESWORTH, 2001 Evidence for selection at the *fused1* locus of *Drosophila americana*. *Genetics* **158**: 279–290.
- WIEHE, T. H. E., and W. STEPHAN, 1993 Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Mol. Biol. Evol.* **10**: 842–854.
- WRIGHT, S., 1951 The genetical structure of populations. *Ann. Eugen.* **15**: 323–354.
- YI, S., and B. CHARLESWORTH, 2000 A selective sweep associated with a recent gene transposition in *Drosophila miranda*. *Genetics* **156**: 1753–1763.

Communicating editor: M. VEUILLE

APPENDIX

Prediction of p_j : As the selected allele a goes to fixation, the association with the original neutral copy c is reduced by recombination at a constant rate r . Within the set of chromosomes carrying a , which has a frequency q_i at generation i , p_i represents the proportion of chromosomes carrying c . In the next generation, the value of p_{i+1} can be predicted from the observed value at the previous generation,

$$p_{i+1} = p_i(1 - (1 - q_i)r). \quad (A1)$$

This equation assumes that recombination destroys ac chromosomes but never generates new ac chromosomes back again. Under the conditions for hitchhiking, the consequence of the generation of new ac chromosomes by recombination is negligible and an unnecessary complication as individuals heterozygous for both loci in repulsion are very infrequent in the subpopulation.

The favorable mutation goes to fixation following an S-shaped series of q_i frequencies. Thus, the equation for p_j after fixation can be simplified to

$$E(p_j) = p_0 \prod_{i=0}^f (1 - (1 - q_i)r) \approx e^{-r \sum_{i=0}^f (1 - q_i)} = e^{-r(f/2)}.$$

The fixation time f depends mainly on the effective population size and the selective coefficient of the mutation. A good approximation for the mean value is given by GALE (1990, p. 265),

$$f = \frac{3.9 + 2 \log N + 2 \log s}{s} - 2 \approx \frac{3.9 + 2 \log N + 2 \log s}{s}.$$

Substituting this equation in the previous one, the prediction for p_j becomes

$$E(p_j) = (7Ns)^{-r/s}.$$

It is worth noting that this prediction is not deterministic as the approximation of Gale considers random events during the fixation process. The prediction is more simple than that obtained by OTTO and BARTON (1997; Equations 17 and 18).

Prediction of p_i^2 : At generation $i + 1$, the expected value of p_{i+1}^2 can be given as a function of the value in the previous generation:

$$E(p_{i+1}^2) = p_i^2(1 - 2r(1 - q_i)) + \frac{p_i - p_i^2}{2Nq_i}.$$

The first term is the square of Equation A1 and represents the effect of change in the mean value of p_i . The second term is the effect of drift within the set of chromosomes carrying the favorable allele a .

The solution for the corresponding differential equation is very complex as the expectation $E(p_i^2)$ changes with time as allele a goes to fixation. To simplify the approach, the p_i^2 values of consecutive generations are scaled (divided) by the square of the expected p_i values, which is represented by $(p_i)^2$ in the equation

$$E\left(\frac{p_{i+1}^2}{(p_{i+1})^2}\right) = \left(\frac{p_i^2}{(p_i)^2}\right) + \frac{1}{2Nq_i(1 - 2r(1 - q_i))} \left(\frac{1}{p_i} - \left(\frac{p_i^2}{(p_i)^2}\right)\right). \quad (\text{A2})$$

This equation reaches a nearly constant value when the number of copies of allele a ($2Nq_i$) is large enough. Note that p_i is the expected frequency of c over all the possible populations and its value in consecutive generations can be approximated by Equation A1. Both p_i and $(p_i)^2$ are considered to be deterministic variables. In contrast, the term p_i^2 is considered a random variable. As the number of copies of the favorable mutation ($2Nq_i$) increases, the second term of Equation A2 decreases, and the equation reaches a constant value.

Numerical simulations have shown that the distribution of p_i^2 (after fixation) over replicates is mainly dependent on the random process within the set of chromosomes carrying the favorable mutation: the random fluctuations of the frequency of the favorable mutation (q_i) have little effect, perhaps because of compensation of deviations around its expectation. Therefore, we have simplified the derivations considering that the frequency of the favorable mutation a increases deterministically. To approximate the consecutive frequencies over generations we considered that the evolution of the number of copies k_i ($q_i = k_i/2N$) of allele a follows a branching process when $k_i \ll 2N$; *i.e.*, the copies are propagating independently. Let us assume that we have an infinite number of identical and independent populations with k_i copies. Eventually, allele a will be lost in a proportion $(1 - 2s)^{k_i}$ of all the populations and fixed in a proportion $1 - (1 - 2s)^{k_i}$.

In the next generation, the expected number of copies of allele a in all populations is

$$k_i(1 + s).$$

For the set of populations in which allele a will be eventually lost, the expected number of copies of a is

$$\left[\sum_{x=0}^{\infty} \frac{e^{-k_i(1+s)} (k_i(1+s))^x}{x!} (1-2s)^x \right] \Big/ \left[\sum_{x=0}^{\infty} \frac{e^{-k_i(1+s)} (k_i(1+s))^x}{x!} (1-2s)^x \right] = k_i(1-s),$$

where x is the number of copies of allele a in the next

generation. Therefore, in one generation time, the expected increase of the number of copies of a in the set of populations where the allele will be fixed is

$$k_{i+1} - k_i = 1 + k_i s - \frac{1 - (1 - 2s)^{k_i} (1 + 2k_i s)}{1 - (1 - 2s)^{k_i}}.$$

The first term on the right-hand side (the number 1) is relevant only in early generations when $k_i < 1/s$. It defines a linear period in which the average number of copies of allele a increases at a constant rate of one copy per generation. The second term represents the exponential rate of increase in copy number that is relevant later when $k_i > 1/s$. The third term is a correction that smoothes the transition between the linear and the exponential periods. To simplify the solution of Equation A2, we split the evolution of the number of copies of a into these two consecutive periods: the period from one single copy to $1/s$ copies and the period from $1/s$ copies to fixation. This simplification works better if r is very small when compared with s , which is the combination of parameters needed for the hitchhiking effect. In the first period, we consider that the number of copies of the favorable mutation increases at an average rate of one copy per generation in the set of populations where the mutation is going to be fixed, independently of the selection coefficient of the mutation. The simplification for the first period becomes

$$E\left(\frac{p_{i+1}^2}{(p_{i+1})^2}\right) = \left(\frac{p_i^2}{(p_i)^2}\right) + \frac{1}{i} \left(\frac{1}{(1-r)^i} - \left(\frac{p_i^2}{(p_i)^2}\right)\right).$$

In the second period from $1/s$ to fixation, the favorable mutation increases deterministically at a rate $(1 + s)$ per generation,

$$E\left(\frac{p_{i+1}^2}{(p_{i+1})^2}\right) = \left(\frac{p_i^2}{(p_i)^2}\right) + \frac{1}{(1/s)(1+s)^i} \left(\frac{1}{(1-r)^{(1/s)+i}} - \left(\frac{p_i^2}{(p_i)^2}\right)\right).$$

This is not true when the frequency is high, but this does not affect the prediction because, as we have shown, the expectation becomes constant when the number of copies of the favorable gene is large.

Connecting both equations and solving the corresponding differential equation, the solution after fixation is

$$E\left(\frac{p_f^2}{(p_f)^2}\right) = e^{2(r/s)}.$$

As we have assumed that p_i changes deterministically, $(p_i)^2$ can be considered a constant equal to $(7Ns)^{-2(r/s)}$. Therefore

$$E(p_f^2) = (p_f)^2 e^{2(r/s)} = (7Ns)^{-2(r/s)} e^{2(r/s)} = (2.6Ns)^{-2(r/s)}.$$

Intensive simulations were carried out to check the equations. A single favorable mutation was put in a population associated with a particular neutral copy. After fixation, the frequency of the neutral allele was

computed in the population. After 10,000 simulations for each particular combination of N , r , and s , the average frequency of the neutral allele (observed p_j) and the average value of the squared frequencies (observed

p_j^2) were calculated (Table 1). The accuracy of the predictions is very high. Deviations of the observed values are $\sim 2\%$ for p_j and $\sim 5\%$ for p_j^2 when $s < 1$, above or below the expectation with no general trend.

