

The Extent of Linkage Disequilibrium and Haplotype Sharing Around a Polymorphic Site

Hideki Innan^{*,†,1} and Magnus Nordborg^{*}

^{*}Molecular and Computational Biology, University of Southern California, Los Angeles, California 90089-1340 and [†]Human Genetics Center, School of Public Health, University of Texas Health Science Center, Houston, Texas 77030

Manuscript received December 11, 2002

Accepted for publication May 22, 2003

ABSTRACT

Various expressions related to the length of a conserved haplotype around a polymorphism of known frequency are derived. We obtain exact expressions for the probability that no recombination has occurred in a sample or subsample. We obtain an approximation for the probability that no recombination that could give rise to a detectable recombination event (through the four-gamete test) has occurred. The probabilities can be used to obtain approximate distributions for the length of variously defined haplotypes around a polymorphic site. The implications of our results for data analysis, and in particular for detecting selection, are discussed.

CHRomosomes that have inherited a particular mutation from a common ancestral chromosome must also have inherited a short chromosomal region surrounding the mutation. The length of the piece that is identical by descent to the ancestral chromosome will vary between the chromosomes in a complex pattern. The phenomenon is not directly observable, but gives rise to linkage disequilibrium and sharing of marker haplotypes, which can be observed in polymorphism data (reviewed, for example, in NORDBORG and TAVARÉ 2002).

There is currently tremendous interest in linkage disequilibrium and haplotype sharing, the primary reason being its importance in association mapping of human disease loci (DALY *et al.* 2001; PATIL *et al.* 2001; DAWSON *et al.* 2002; GABRIEL *et al.* 2002; PHILLIPS *et al.* 2003). The extent of linkage disequilibrium (LD) and haplotype sharing plays an important role in determining the density of single-nucleotide polymorphisms (SNPs) that is required for mapping. In addition, the distribution of haplotype sharing can be important in actual mapping methods (MCPEEK and STRAHS 1999; MORRIS *et al.* 2000, 2002; LIU *et al.* 2001).

Outside of mapping, the distribution of haplotype lengths is of importance to evolutionary biologists, because it is needed to evaluate claims about particular haplotypes being “too long” to fit neutrality. A mutation that has been driven by directional selection (either positive or negative) to its current population frequency will typically have reached that frequency much faster than if it had simply reached that frequency through genetic drift (MARUYAMA 1977). This means that recombination will have had less time to break up the ancestral

haplotype (KAPLAN *et al.* 1989). Mutations that are surrounded by regions of haplotype sharing that are unusually extensive given the frequency of the mutation are therefore candidates for having been influenced by selection. The problem is deciding what “unusual” means.

In this article we derive some basic results concerning the extent of LD and haplotype sharing surrounding a focal mutation (*e.g.*, a SNP) that has been determined to have a certain frequency in a sample. We assume a standard neutral model: our results can therefore serve as a “null model” with which data can be compared.

RESULTS

The model: Consider a sample of n sequences from a population of N diploid individuals that evolves according to a standard neutral model. The coalescent approximation is employed throughout (see, *e.g.*, NORDBORG 2001). Imagine further that the sample contains a diallelic polymorphism at a particular site 0 (*e.g.*, a SNP locus). We assume that the polymorphism was created by a unique mutational event so that the mutation rate at the focal site is zero. It is also assumed that the ancestral allelic state is known. Let i and $j = n - i$ represent the number of the sampled haplotypes with the ancestral allele and those with the mutant allele, respectively. Denote this state $A(i, j)$. We describe the genealogical history of a sample conditional on this configuration. An example genealogy for $A(7, 3)$ is shown in Figure 1. A mutation from “A” to “T” creates the mutant allelic class, which consists of sequences, 8, 9, and 10. During the time between the mutation and present, no coalescence can occur between the two allelic classes. The coalescent conditional on $A(i, j)$ differs from the standard theory (KINGMAN 1982; HUDSON 1983; TAJIMA

¹Corresponding author: Human Genetics Center, School of Public Health, University of Texas Health Science Center, 1200 Hermann Pressler, Houston, TX 77030. E-mail: hinnan@sph.uth.tmc.edu

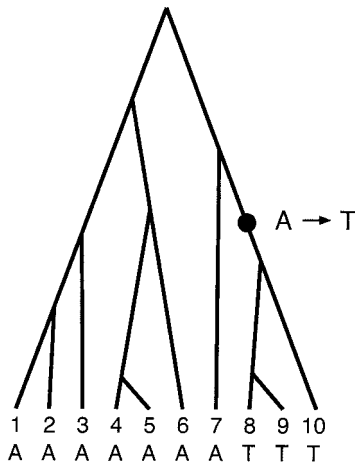


FIGURE 1.—An example of a genealogy in $A(7, 3)$. A mutation changes the ancestral allelic state “A” to the mutant allelic state “T.” Haplotypes 8, 9, and 10 belong to the mutant allelic class. The other seven sequences carry A and belong to the ancestral allelic class.

1983) in many ways. Relevant theory has been developed by INNAN and TAJIMA (1997), GRIFFITHS and TAVARÉ (1998, 1999), and WIUF and DONNELLY (1999).

In this article, we consider how long the genealogical relationship among the sampled sequences at site 0 is conserved along the chromosome. In the absence of recombination, exactly the same genealogy is obtained at any site along the chromosome; recombination causes genealogies to change gradually as we move farther away from site 0.

The probability of no recombination: Consider the genealogy at a site L , L sites away from site 0. Recombination occurs with probability r per site per generation: in the coalescent setting, we use the scaled rate $\rho = \lim_{N \rightarrow \infty} 4Nr$ (e.g., NORDBERG 2001). To simplify expressions, we also define $R = L\rho$.

We first seek the probability that no recombination has occurred between 0 and L in the history of the sample, i.e., before the sample reaches its most recent common ancestor (MRCA). In the absence of recombination the genealogies at these two sites must be identical. We derive this probability using the methods of INNAN and TAJIMA (1997).

Consider the coalescent process at site 0 starting at $A(i, j)$, $j > 1$. When the first coalescence event occurs, $A(i, j)$ changes to either $A(i-1, j)$ or $A(i, j-1)$ with probabilities $(i-1)/(n-1)$ and $j/(n-1)$, respectively (INNAN and TAJIMA 1997, Figure 3A). The waiting time until the first coalescence is exponentially distributed with rate $\binom{n}{2}$. Since recombination occurs at rate $R/2$ per lineage, it follows from standard theory of Poisson processes that the probability that there is no recombination on the genealogy of the whole sample before the first coalescence is

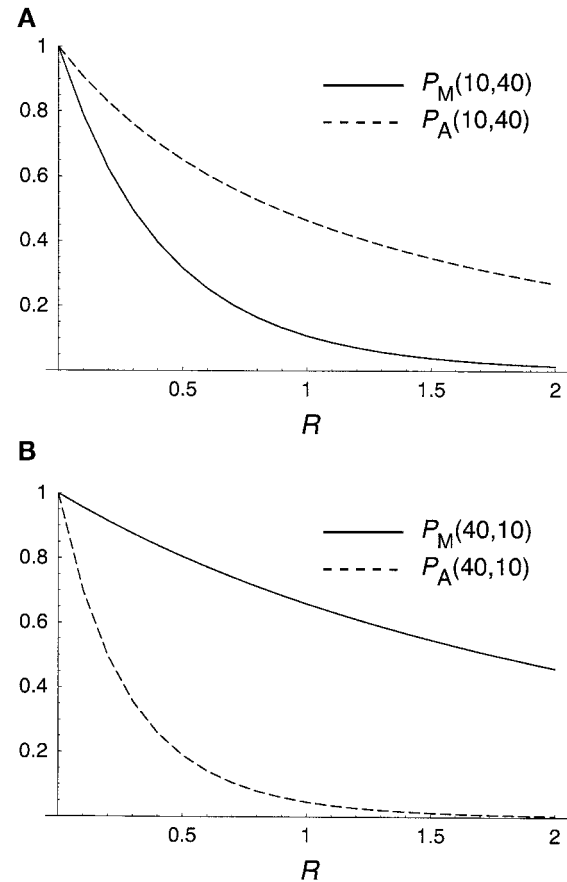


FIGURE 2.—Examples of $P_M(R|i, j)$ and $P_A(R|i, j)$ for $n = 50$. (A) Results for $i = 10$ and $j = 40$. (B) Results for $i = 40$ and $j = 10$.

$$Q_W(R|n) = \frac{\binom{n}{2}}{\binom{n}{2} + nR/2} = \frac{n-1}{n-1+R}. \quad (1)$$

Thus, the probability that there is no recombination on the genealogy of the whole sample given the initial state $A(i, j)$, $j > 1$ obeys

$$P_W(R|i, j) = \left[\frac{i-1}{n-1} P_W(R|i-1, j) + \frac{j}{n-1} P_W(R|i, j-1) \right] Q_W(R|n). \quad (2)$$

The case $j = 1$ is more complicated because it involves the mutational event that created the mutant allelic class. Going backward in time, the next event is either a coalescence within the ancestral allelic class, in which case the process moves from $A(i, 1)$ to $A(i-1, 1)$, or the mutational event that created the mutant allelic class, in which case the process moves from $A(i, 1)$ to $A(i+1)$, where $A(i+1)$ represents the state in which there are i sequences in the ancestral allelic class and one sequence that is the ancestor of the mutant allelic class (INNAN and TAJIMA 1997, Figure 3B). INNAN and TAJIMA (1997) show that the waiting time until the first event is exponentially distributed with rate $(i+1)i/2$

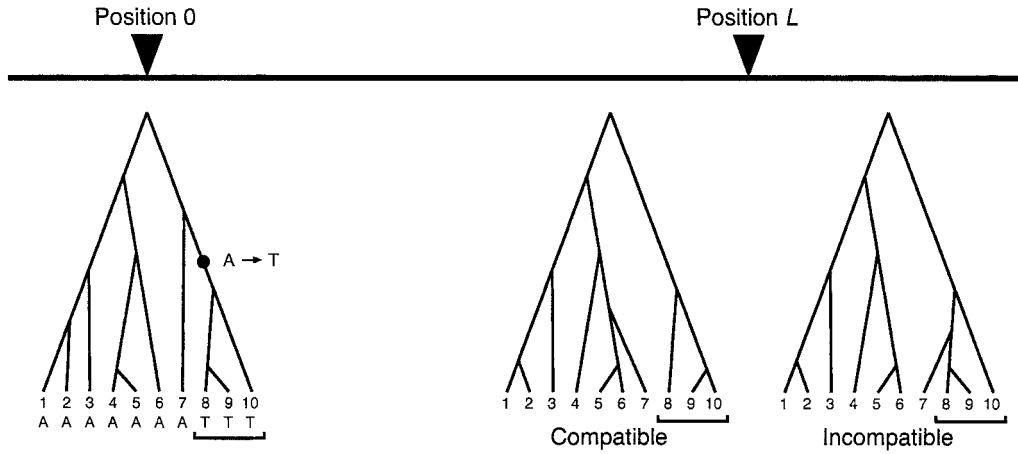


FIGURE 3.—Illustration of the concept of tree compatibility. The sample configuration at 0 is $A(7, 3)$ with 8, 9, and 10 sharing a mutation. The left-hand genealogy at position L is an example of a compatible genealogy, because 8, 9, and 10 are monophyletic. The right-hand genealogy at L is incompatible because of the invasion of 7.

and that probabilities of the two different outcomes are $(i-1)/i$ and $1/i$, respectively. Thus we have

$$P_W(R|i, 1) = \left[\frac{i-1}{i} P_W(R|i-1, 1) + \frac{1}{i} P_W(R|i+1) \right] Q_W(R|i+1), \quad (3)$$

where $P_W(R|i+1)$ is given by

$$P_W(R|i+1) = \prod_{m=2}^{i+1} \frac{m-1}{m+R-1} \quad (4)$$

from Equation 1. By jointly solving Equations 2 and 3, $P_W(R|i, j)$ can now be obtained.

The above results concern the genealogy of the whole sample. If we are interested in questions like the extent of haplotype sharing between members of a particular allelic class, we need results for the genealogy of the appropriate subsample. First we consider the mutant allelic class. Let $P_M(R|i, j)$ be the probability of no recombination between 0 and L in the ancestry of the mutant allelic class, so that the genealogies of the mutant allelic class at the two sites are exactly the same. The same reasoning that led to Equation 2 leads to the following recursion for $P_M(R|i, j)$, $j > 1$,

$$P_M(R|i, j) = \left[\frac{i-1}{n-1} P_M(R|i-1, j) + \frac{j}{n-1} P_M(R|i, j-1) \right] Q_M(R|i, j), \quad (5)$$

where

$$Q_M(R|i, j) = \frac{n(n-1)}{n(n-1) + jR}. \quad (6)$$

$P_M(R|i, j)$ can be calculated from Equation 5, using the initial condition $P_M(R|i, 1) = 1$.

Similarly, we consider $P_A(R|i, j)$ to be the probability of no recombination between 0 and L in the ancestry of the ancestral allelic class. For $j > 1$ we have

$$P_A(R|i, j) = \left[\frac{i-1}{n-1} P_A(R|i-1, j) + \frac{j}{n-1} P_A(R|i, j-1) \right] Q_A(R|i, j), \quad (7)$$

where

$$Q_A(R|i, j) = \frac{n(n-1)}{n(n-1) + iR}. \quad (8)$$

For $j = 1$, the equation analogous to (3) is

$$P_A(R|i, 1) = \left[\frac{i-1}{i} P_A(R|i-1, 1) + \frac{1}{i} P_A(R|i+1) \right] Q_A(R|i+1), \quad (9)$$

where

$$P_A(R|i+1) = \prod_{m=2}^i \frac{m-1}{m+R-1}. \quad (10)$$

Figure 2 shows $P_M(R|i, j)$ and $P_A(R|i, j)$ for $A(40, 10)$ and $A(10, 40)$. When the mutant allelic class is common (Figure 2A), the genealogy of the mutant allelic class decays more quickly than that of the ancestral allelic class. When the mutant allelic class is rare (Figure 2B), the genealogy of the mutant allelic class is conserved over quite a long distance, while that of the ancestral allelic class decays very quickly. Note that $P_M(40, 10) > P_A(10, 40)$: this reflects the fact that the mutant allelic class is younger.

The probability of tree compatibility: In the previous section, we considered the extent of haplotype sharing in the sense of identity by descent with respect to recombination. This cannot of course be observed directly, but must be inferred from data. Under the infinite-site model, unless recombination has occurred between two sites in the history of the sample, there can be at most three out of four possible haplotypes (the “four-gamete” test of HUDSON and KAPLAN 1985) and $|D'|$ (LEWONTIN 1964) must always be 1. These two statistics can be viewed as tests for recombination (in which case they amount to the same thing), but it must be remembered that they have very low power: most recombination events will not be detected. One reason for this is that mutations may not have occurred on the appropriate branches on the genealogy: even if recombination has

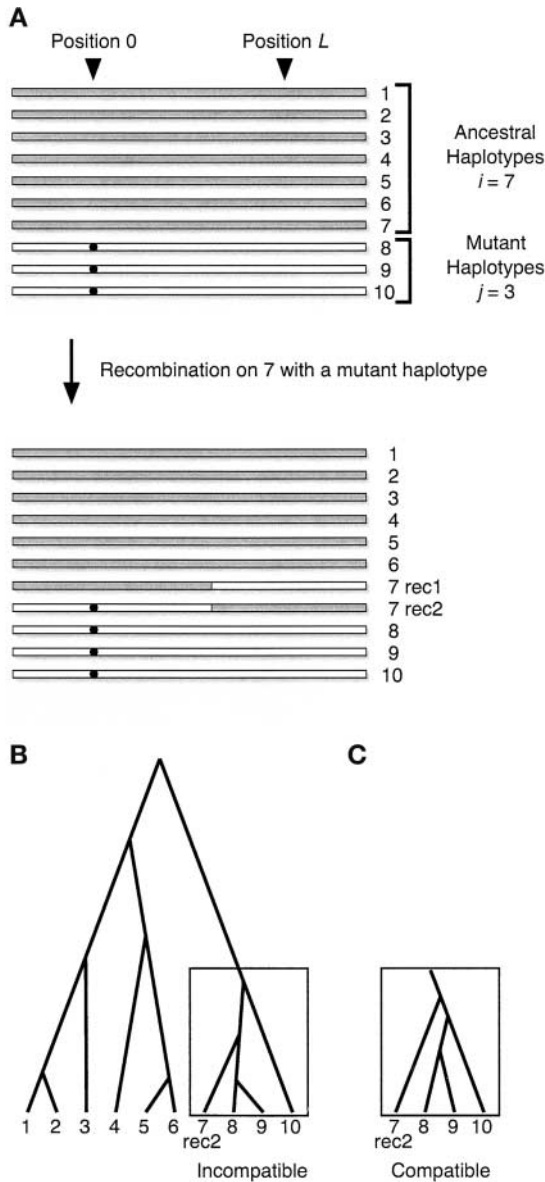


FIGURE 4.—An example of a recombination event that could give rise to incompatible genealogies at 0 and L . See text for discussion.

occurred, any particular pair of SNPs may not reveal it. Another reason is that the underlying genealogy may be such that recombination cannot be detected, even with infinitely many polymorphic sites (HUDSON and KAPLAN 1985). This is particularly important in the present context because closely linked sites will tend to have similar genealogies.

A general analytical treatment of the extent of haplotype sharing visible in data would be nice, but seems extremely difficult. We derive a heuristic approximation for a particular problem, namely the probability that the genealogy at site L is “compatible” with the genealogy at site 0 in the sense that recombination between them

cannot possibly be detected using either of the tests described above. Our definition of tree compatibility is illustrated in Figure 3. The allelic state at site 0 is $A(7, 3)$ because haplotypes 8, 9, and 10 share a mutation. Now consider the genealogy at site L . We define this genealogy as compatible if 8, 9, and 10 are monophyletic at L , *i.e.*, if they are related more closely to each other than to any other sampled haplotype at L . Genealogies that do not have this property are incompatible: with infinitely many polymorphic sites, the recombination between 0 and L would be detected.

Let $P_C(R|i, j)$ be the probability that the genealogy at site L is compatible with that at site 0 where the allelic state is $A(i, j)$. An approximate expression for this probability can be obtained by modifying the equations developed in the previous section. Our approximation is best explained through an example. Consider a sample where the allelic state at 0 is $A(7, 3)$ (Figure 4A). Seven haplotypes belong to the ancestral class and three belong to the mutant class. Going backward in time, a recombination might occur before the first coalescence. Assume that an ancestral haplotype (number 7, say) undergoes recombination and breaks into two haplotypic lineages (for more on the ancestral recombination graph, see, *e.g.*, NORDBERG 2001). Recombination occurs with either an ancestral haplotype or a mutant haplotype. The probabilities of these two events are $1 - X$ and X , respectively, where X is the frequency of the mutant haplotype (allelic class) in the population. Given that the recombination occurred in an ancestral haplotype, only recombination with a mutant haplotype could make the genealogy at site L incompatible with that of site 0. Of course we do not know X , and will use \bar{X} , its expected value given the sample configuration $A(i, j)$, as a proxy. As is shown in the APPENDIX, $\bar{X} = j/(n + 1)$.

Suppose this latter type of recombination occurred, so that there are six unrecombined haplotypes of the ancestral class, three unrecombined haplotypes of the mutant class, and two recombinants derived from sequence 7. As shown in Figure 4A, haplotypes 8–10 and 7rec2 now belong to the mutant allelic class, whereas 1–6 and 7rec1 belong to the ancestral class. Assuming that no further recombination occurs, haplotypes 8–10 and 7rec2 must now be monophyletic. Depending on the order in which they coalesce, the resulting genealogy is either compatible or incompatible in the sense of Figure 3. An example of an incompatible genealogy is shown in Figure 4B; an example of a compatible one is shown in Figure 4C. The latter kind of genealogy occurs if and only if haplotypes 8–10 are monophyletic with respect to 7rec2. Let α be the probability of the outcome exemplified in Figure 4B, given that recombination occurs as just described when there are i ancestral and j nonancestral haplotypes. It is easy to show that $\alpha = 1 - 2/[j(j + 1)]$ (SAUNDERS *et al.* 1984). Therefore, when the process is in $A(i, j)$, recombination that leads

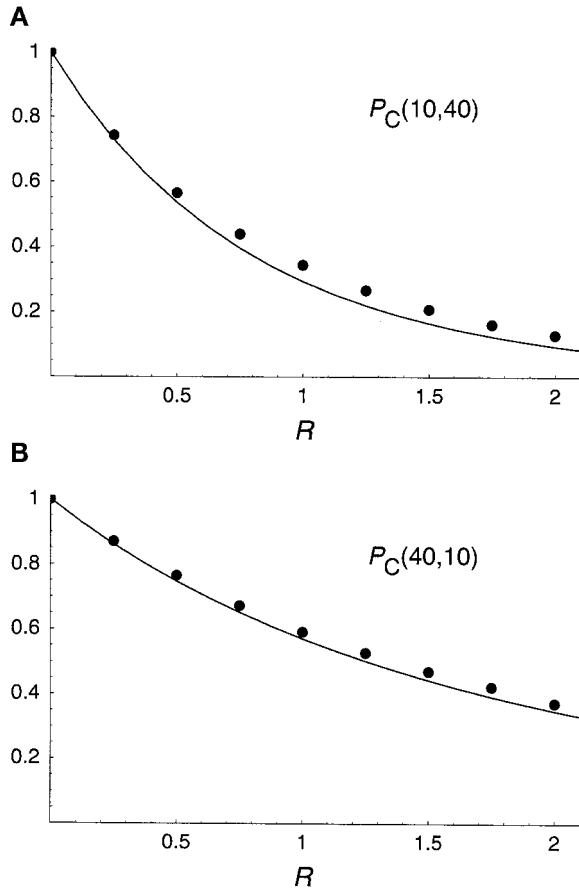


FIGURE 5.—Simulations illustrating the behavior of $P_C(R|i, j)$ for (A) $A(10, 40)$ and (B) $A(40, 10)$. The solid circles represent the expected $P_C(R|i, j)$ obtained from coalescent simulations. The solid curves represent the approximation given by Equation 12.

to an incompatible genealogy at site L occurs approximately at rate $i\bar{X}\alpha R/2$ among the i haplotypes in the ancestral class. The approximation assumes that recombination occurs only once before the lineages coalesce and is therefore likely to work best for small R . It also relies on using \bar{X} instead of integrating over distribution of the random variable X .

Next, we consider recombination in a mutant haplotype. Each such haplotype undergoes recombination with an ancestral haplotype at rate $(1 - X)R/2$. When this happens, incompatibility is highly likely since the mutant haplotypes must coalesce first. Therefore, when the process is in $A(i, j)$, recombination that leads to an incompatible genealogy at L occurs approximately at rate $j(1 - \bar{X})R/2$ among the j haplotypes in the mutant class. Again, this approximation will work best for small R .

Putting all this together, the probability that there is no recombination that gives rise to an incompatible genealogy at site L before the first coalescence in $A(i, j)$ is given by

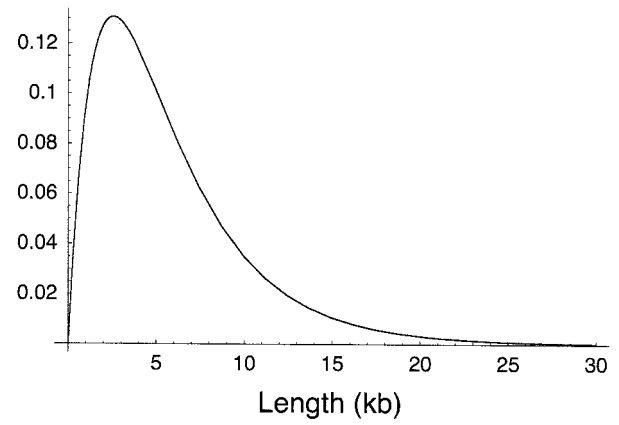


FIGURE 6.—Probability density of the length of the region surrounding a focal mutation present in 15/20 sampled haplotypes in which recombination cannot be detected. The density was obtained from Equation 12 assuming a recombination rate of 0.42/kb (see INNAN *et al.* 2003).

$$Q_C(R|i, j) \approx \frac{\binom{n}{2}}{\binom{n}{2} + i\bar{X}\alpha R/2 + j(1 - \bar{X})R/2} = \frac{(n-1)n(n+1)}{(n-1)n(n+1) + (i+1)jR + i\alpha R}, \quad (11)$$

and we have the recursion

$$P_C(R|i, j) = \left[\frac{i-1}{n-1} P_C(R|i-1, j) + \frac{j}{n-1} P_C(R|i, j-1) \right] Q_C(R|i, j), \quad (12)$$

which can be solved using $P_C(R|i, 1) = 1$.

Simulations indicate that our heuristic approximation works rather well, at least for reasonably large sample size and moderate R . Figure 5 shows some results for $n = 50$. As expected, the lower the frequency of the mutation, the larger the region in which recombination cannot be detected is likely to be. Our approximation tends to be smaller than the real value because we ignore multiple recombination events, which may return an incompatible tree to the compatible state.

The theoretical results shown in this section are based on recursion equations. We have also obtained closed forms for the four probabilities, $P_W(R|i, j)$, $P_M(R|i, j)$, $P_A(R|i, j)$, and $P_C(R|i, j)$ (available upon request), although they are not shown in this article.

DISCUSSION

We have derived several probabilities related to the preservation of an ancestral haplotype along the chromosome. We consider a sample of n haplotypes and focus on a polymorphic site at which i haplotypes carry the ancestral allele, and $j = n - i$ carry a mutant allele. First, we derive the probability of there having been no recombination in the genealogy of the sample (or of a

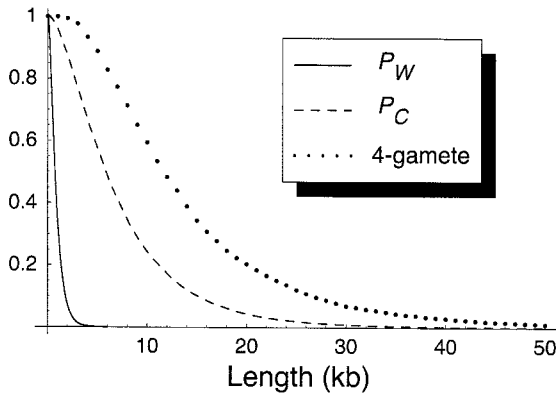


FIGURE 7.—The probability that the extent of haplotype sharing surrounding a focal mutation present in 20 out of 100 sampled haplotypes exceeds a certain length. Three different definitions of haplotype sharing are used; see text for details. In all cases, a recombination rate of 0.42/kb was assumed. The distribution of the extent of haplotype sharing under the four-gamete test was obtained from 10,000 runs of the ancestral recombination graph with a mutation rate of 0.8/kb (INNAN *et al.* 2003).

subsample, such as all the members of a particular allelic class), between this site and an arbitrary linked site. The distribution of the length of the segment (on one side of the focal site) in which no recombination occurred is readily obtained from this probability. By assuming independence of recombination on either side of the focal site, we can also obtain the distribution of the total length. For example, the probability density of the length of this region for the mutant allelic class would be given by the convolution

$$f_M(x|i, j) = \int_0^x P'_M(y|i, j)P'_M(x-y|i, j)dy, \quad (13)$$

where $P'_M(x|i, j)$ denotes the derivative of $P_M(x|i, j)$ with respect to x , which can be calculated using Equation 5. The integral is best evaluated numerically.

Second, we derive an approximation for the probability that a tree at a given distance from the focal site is such that recombination between it and the focal site cannot be detected, in the sense that there will always be less than four gametes, and that $|D'| = 1$. This probability can of course also be used to find an approximation for the distribution of the length of the region in which recombination cannot be detected.

Our results are relevant for understanding the extent of linkage disequilibrium and haplotype sharing around a particular polymorphism. This is important in association mapping of human diseases, where we might be interested in how the pattern of haplotype sharing around a disease allele is expected to depend on the frequency of that allele. It should be noted in this context that whereas other treatments of this problem (KAPLAN *et al.* 1995; THOMPSON and NEEL 1997; SLATKIN and BERTORELLE 2001) are based on assumptions that

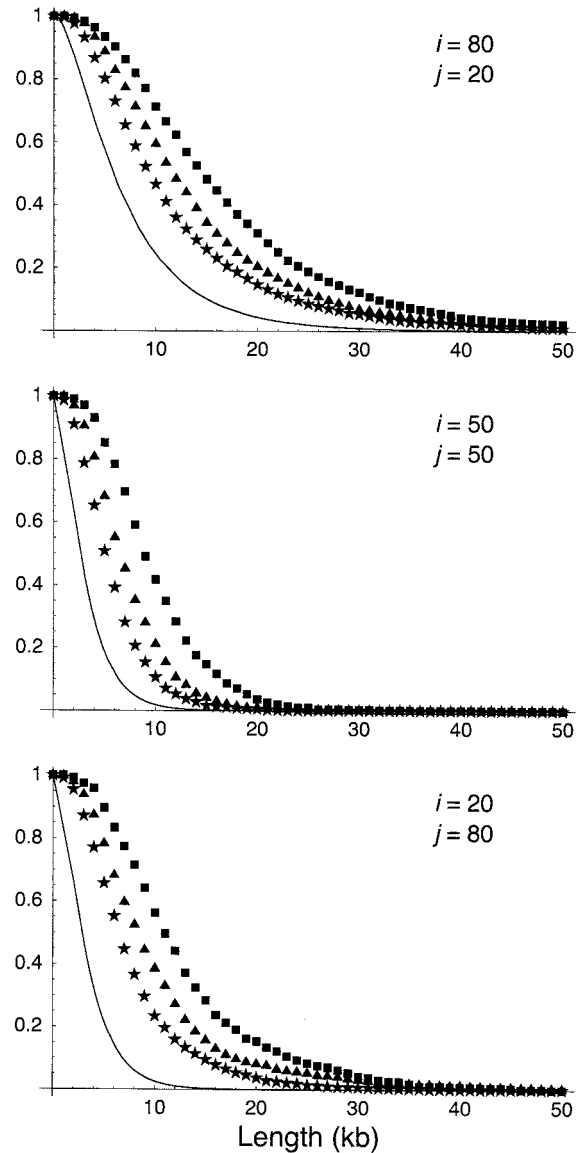


FIGURE 8.—The effect of the density of markers on the power to identify the region of haplotype sharing. All cases assume a total sample size of 100 and a recombination rate of 0.42/kb. Simulation results for mutation rates 0.4, 0.8, and 1.6/kb are presented by boxes, triangles, and stars, respectively. The curve represents P_C . Three mutant allele frequencies, 20/100, 50/100, and 80/100, are investigated.

are valid only for rare alleles, our results are valid for all frequencies. We do, on the other hand, assume a constant population size.

Another application of our results is in evaluating claims of past selection. There is currently great interest in detecting past selection at polymorphic sites by looking at the extent of haplotype sharing surrounding each site and determining whether it is too extensive to be compatible with neutrality (*e.g.*, ANDOLFATTO *et al.* 1999; SABETI *et al.* 2002). Our results are relevant to such questions. For example, INNAN *et al.* (2003) found a local region with a very high level of linkage disequilibrium

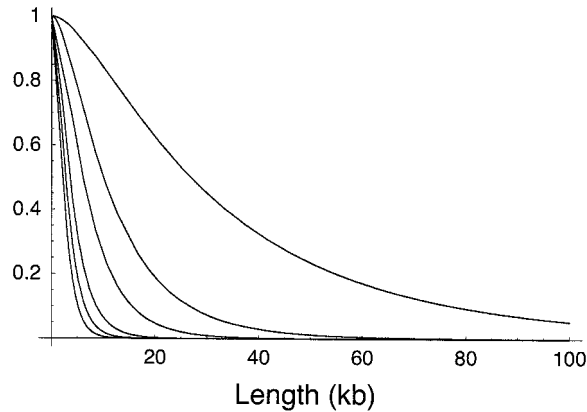


FIGURE 9.—The effect of the frequency of the mutation on the distribution of the length of the compatible region (from P_C). All cases assume a total sample size of 200 and a recombination rate of 0.42/kb. From top to bottom, the curves are for a mutant allele frequency of 5, 10, 15, 25, 50, and 85%, respectively.

rium on human chromosome 21, in the data of PATIL *et al.* (2001). In this ~ 50 -kb region, there is a high-frequency (15 out of 20) haplotype. Since there is almost no variation within this haplotype, it is likely to be derived, as opposed to ancestral. Missing data complicates the interpretation (INNAN *et al.* 2003), but it suffices as an example. Figure 6 shows the probability density of the length of the region in which recombination cannot be detected in this case. An unbroken 50-kb haplotype seems unlikely under the standard neutral model. Alternative explanations include selection, demography, or a local decrease in recombination (INNAN *et al.* 2003).

However, our results should not be used to test for selection by rejecting neutrality. The density in Figure 6 is for the length of the region in which recombination could not possibly be detected, irrespective of the number of markers. With finitely many marker loci, it is necessary to take into account the fact that many loci will not reveal recombination even if recombination has occurred in such a manner that marker loci could potentially reveal it. Figure 7 illustrates the distinction by comparing the distribution of the length of haplotypes in which no recombination (1) occurred (from Equation 2); (2) could have been detected, irrespective of the mutation rate (from Equation 12); or (3) was in fact detected, given a particular mutation rate. The events in 3 are a subset of those in 2, which are a subset of those in 1 (HUDSON and KAPLAN 1985).

The utility of our results lies in the fact that they provide a lower bound for the length of haplotype conservation that might be observed. The difference between this bound and case 3 above is determined by the density of markers, which in turn depends on the mutation rate (Figure 8). Our results allow us to determine whether there is any reason to consider selection as an explanation for a particular data set. They also

provide a very simple method for exploring the effect of the allele frequency on the distribution of haplotype lengths. Figure 9 illustrates the crucial role played by the frequency of the mutant allele in determining the length of the surrounding haplotype. Haplotype sharing surrounding a low-frequency mutant allele can clearly be very extensive even in the absence of selection.

We thank N. Rosenberg and two anonymous reviewers for comments on the manuscript. We also thank P. Donnelly and C. Wiuf for many discussions and for sharing an unpublished manuscript in which they, *inter al.*, study $P_M(R|i, j)$ by simulation.

LITERATURE CITED

- ANDOLFATTO, P., J. D. WALL and M. KREITMAN, 1999 Unusual haplotype structure at the proximal breakpoint of $\ln(2L)t$ in a natural population of *Drosophila melanogaster*. *Genetics* **153**: 1297–1311.
- DALY, M. J., J. D. RIOUX, S. F. SCHAFFNER, T. J. HUDSON and E. S. LANDER, 2001 High-resolution haplotype structure in the human genome. *Nat. Genet.* **29**: 229–232.
- DAWSON, E., G. R. ABECASIS, S. BUMPSTEAD, Y. CHEN, S. HUNT *et al.*, 2002 A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**: 544–548.
- GABRIEL, S. B., S. F. SCHAFFNER, H. NGUYEN, J. M. MOORE, J. ROY *et al.*, 2002 The structure of haplotype blocks in the human genome. *Science* **296**: 2225–2229.
- GRIFFITHS, R. C., and S. TAVARÉ, 1998 The age of a mutant in a general coalescent tree. *Stoch. Models* **14**: 273–295.
- GRIFFITHS, R. C., and S. TAVARÉ, 1999 The ages of mutations in gene trees. *Ann. Appl. Prob.* **9**: 567–590.
- HUDSON, R. R., 1983 Testing the constant-rate neutral allele model with protein sequence data. *Evolution* **37**: 203–217.
- HUDSON, R. R., and N. L. KAPLAN, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**: 147–164.
- INNAN, H., and F. TAJIMA, 1997 The amounts of nucleotide variation within and between allelic classes and the reconstruction of the common ancestral sequence in a population. *Genetics* **147**: 1431–1444.
- INNAN, H., and F. TAJIMA, 1999 The effect of selection on the amounts of nucleotide variation within and between allelic classes. *Genet. Res.* **73**: 15–28.
- INNAN, H., B. PADHUKASAHASRAM and M. NORDBORG, 2003 The pattern of polymorphism on human chromosome 21. *Genome Res.* **13**: 1158–1168.
- KAPLAN, N. L., R. R. HUDSON and C. H. LANGLEY, 1989 The “hitchhiking” effect revisited. *Genetics* **123**: 887–899.
- KAPLAN, N. L., W. G. HILL and B. S. WEIR, 1995 Likelihood methods for locating disease genes in nonequilibrium populations. *Am. J. Hum. Genet.* **56**: 18–32.
- KINGMAN, J. F. C., 1982 The coalescent. *Stoch. Proc. Appl.* **13**: 235–248.
- LEWONTIN, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49**: 49–67.
- LIU, J. S., C. SABATTI, J. TENG, B. J. B. KEATS and N. RISCH, 2001 Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res.* **11**: 1716–1724.
- MARUYAMA, T., 1977 *Stochastic Problems in Population Genetics*. Springer-Verlag, Berlin.
- MCPHEE, M. S., and A. STRAHS, 1999 Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am. J. Hum. Genet.* **65**: 858–875.
- MORRIS, A. P., J. C. WHITTAKER and D. J. BALDING, 2000 Bayesian fine-scale mapping of disease loci, by hidden Markov models. *Am. J. Hum. Genet.* **67**: 155–169.
- MORRIS, A. P., J. C. WHITTAKER and D. J. BALDING, 2002 Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Am. J. Hum. Genet.* **70**: 686–707.
- NORDBORG, M., 2001 Coalescent theory, pp. 179–212 in *Handbook of Statistical Genetics*, edited by D. J. BALDING, M. J. BISHOP and C. CANNINGS. John Wiley & Sons, Chichester, UK.

- NORDBORG, M., and S. TAVARÉ, 2002 Linkage disequilibrium: what history has to tell us. *Trends Genet.* **18**: 83–90.
- PATIL, N., A. J. BERNI, D. A. HINDS, W. A. BARRETT, J. M. DOSHI *et al.*, 2001 Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**: 1719–1723.
- PHILLIPS, M. S., R. LAWRENCE, R. SACHIDANANDAM, A. P. MORRIS, D. J. BALDING *et al.*, 2003 Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat. Genet.* **33**: 382–387.
- SABETI, P. C., D. E. REICH, J. M. HIGGINS, H. Z. P. LEVINE, D. J. RICHTER *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**: 832–837.
- SAUNDERS, I. W., S. TAVARÉ and G. A. WATTERSON, 1984 On the genealogy of nested subsamples from a haploid population. *Adv. Appl. Prob.* **16**: 471–491.
- SLATKIN, M., and G. BERTORELLE, 2001 The use of intraallelic variability for testing neutrality and estimating population growth rate. *Genetics* **158**: 865–874.
- TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**: 437–460.
- THOMPSON, E. A., and J. V. NEEL, 1997 Allelic disequilibrium and allele frequency distribution as a function of social and demographic history. *Am. J. Hum. Genet.* **60**: 197–204.
- WIUF, C., and P. DONNELLY, 1999 Conditional genealogies and the age of a neutral mutant. *Theor. Popul. Biol.* **56**: 183–201.
- WRIGHT, S., 1931 Evolution in Mendelian populations. *Genetics* **16**: 97–159.

Communicating editor: J. HEIN

APPENDIX

In a constant-size population at equilibrium, the probability density function (pdf) of the frequency of the mutant allelic class is given by

$$\Phi(x) = \frac{C}{x},$$

where C is a constant (WRIGHT 1931). In this study, we consider the case where the sample configuration is known; that is, we know i and j , the number of sequences in the mutant and ancestral allelic classes. The conditional pdf of the frequency of the mutant given the sample configuration is

$$\Phi(x|i, j) = \frac{\binom{n}{2} x^j (1-x)^i \Phi(x)}{\int_0^1 \binom{n}{2} y^j (1-y)^i \Phi(y) dy} = \frac{x^{j-1} (1-x)^i}{\int_0^1 y^{j-1} (1-y)^i dy}$$

(*e.g.*, INNAN and TAJIMA 1999). Thus, the expectation of the frequency of the mutant allelic class in $A(i, j)$ is

$$\bar{X} = \int_0^1 x \Phi(x|i, j) dx = \frac{j}{n+1}.$$