A Method for Detecting Recent Selection in the Human Genome From Allele Age Estimates

Christopher Toomajian,*[,1,2] Richard S. Ajioka,[†] Lynn B. Jorde,[‡] James P. Kushner[†] and Martin Kreitman*[,§]

**Committee on Genetics and* [§]*Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637 and* [†]*Division of Hematology/Oncology and* [‡]*Department of Human Genetics, University of Utah, Salt Lake City, Utah 84112*

Manuscript received January 8, 2003
Accepted for publication May 21, 2003

ABSTRACT

Mutations that have recently increased in frequency by positive natural selection are an important component of naturally occurring variation that affects fitness. To identify such variants, we developed a method to test for recent selection by estimating the age of an allele from the extent of haplotype sharing at linked sites. Neutral coalescent simulations are then used to determine the likelihood of this age given the allele's observed frequency. We applied this method to a common disease allele, the hemochromatosis-associated *HFE* C282Y mutation. Our results allow us to reject neutral models incorporating plausible human demographic histories for *HFE* C282Y and one other young but common allele, indicating positive selection at *HFE* or a linked locus. This method will be useful for scanning the human genome for alleles under selection using the haplotype map now being constructed.

THE ability to detect mutations under positive selection while they are still segregating in populations is an exciting prospect because these are variants that determine the fitness difference among individuals. Some of these variants may be found among human disease alleles that segregate at unexpectedly high frequencies. For example, the case has been made for positive selection for alleles at cystic fibrosis transmembrane conductance regulator (Romeo *et al.* 1989), phenylalanine hydroxylase (Kidd 1987) and genes involved in lysosomal lipid storage disease (Rotter and Diamond 1987), sickle cell, and other hereditary anemias (Motulsky 1964).

This article deals with DNA polymorphism on different scales (base pairs, kilobase pairs, megabase pairs) so it helps to clarify our terminology. We study alleles defined at the base pair level and additional variable sites linked to this focal polymorphism at the megabase pair level. We focus on the derived, or novel, allele that is produced when a DNA mutation occurs and for convenience may refer to this as a "mutation" or simply as "the allele." Since allele often implies a single base pair or locus (approximate kilobase pair), we refer to the set of homologous sequences (approximate megabase pair) carrying a particular derived allele (base pair) as an "allele class." Each sequence in this allele class, a copy of the allele, will be identical at the site defining the class, but may or may not be identical to one another at other linked sites. "Haplotype" refers to the particular configuration of the polymorphisms at two or more variable sites, such as occurs on each of two homologs of a diploid chromosome. Thus, an allele class may contain more than one different haplotype. We use the term "haplotype sharing" as a quantitative measure of the extent to which two or more sequences in an allele class resemble each other at linked sites.

Positive selection can be inferred for an allele on the basis of the relationship between its age and frequency in a population. In the absence of selection, higher-frequency alleles are expected to be older than lower-frequency alleles (Kimura and Ohta 1973), whereas under positive selection a young allele can quickly rise to a high frequency. As alleles age, mutation leads to the accumulation of more linked variation, and recombination causes the breakdown of linkage disequilibrium (LD) with existing linked alleles. Both of these attributes, linked variation segregating within an allele class (as defined by the presence of a specific mutation) and LD, can be used to estimate an allele's age. The haplotype test (Hudson *et al.* 1994) and its subsequent modifications (Kirby and Stephan 1995; Andolfatto *et al.* 1999) aim to detect selection by identifying alleles that show less linked single nucleotide polymorphism (SNP) variation than expected given their frequency. While these tests have had success in Drosophila, the approach is less useful in humans because of the low background level of SNP variation against which to contrast a positively selected allele.

As an alternative, one can include LD information from haplotypes to estimate allele age, taking a genealogical view of LD and uncovering historical patterns of

[1] *Corresponding author:* Molecular and Computational Biology, University of Southern California, 835 W. 37th St., SHS 172, Los Angeles, CA 90089-1340. E-mail: cmtoomaj@alumni.uchicago.edu

[2] *Present address:* Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089-1340.

recombination that can reflect an allele's age (Nordborg and Tavaré 2002). Multilocus data generated from gene mapping studies provide estimates of LD that can be used in conjunction with a genetic map to infer an allele's age (Serre *et al.* 1990; Risch *et al.* 1995) and in some cases test for positive selection (Slatkin and Bertorelle 2001; Sabeti *et al.* 2002). Although these data are widely available in humans, methods of estimating allele age from LD suffer from various problems. They often use very limited information, such as an LD estimate between alleles at only two loci. The short physical distance between two loci can make the estimate of genetic distance between them inaccurate. Current multilocus methods based on pairwise LD inappropriately assume independence between linked alleles (Guo and Xiong 1997; Reich and Goldstein 1999). Other methods that consider the length of a region in complete LD with a mutation (Slatkin and Bertorelle 2001) ignore a substantial amount of information apparent in the level of LD at more distal markers. The recent method of Sabeti *et al.* (2002) provides an improvement on earlier methods that allow the addition of loci at increasing distances but, as applied, does not permit the use of highly mutable loci such as simple sequence repeats (SSRs), which may be especially valuable in tracing the history of potentially selected haplotypes.

To avoid these problems, an allele's age can be estimated by the decay of ancestral haplotype sharing (DHS; McPeek and Strahs 1999), an approach developed for fine-scale genetic mapping. Ancestral haplotype is a term for the multilocus configuration of linked segregating sites that were present in the first chromosome to carry a particular allele (*i.e.*, mutation). The length of an ancestral haplotype retained between different copies of an allele shortens over time due to recombination and mutation, so a negative correlation is expected between an allele's frequency and the extent of haplotype sharing at linked loci. The expected genetic distance to which the ancestral haplotype is preserved is also the reciprocal of the time in generations back to the most recent common ancestor (TMRCA) of the allele class. The DHS method estimates this parameter using tightly linked multilocus data that can include both highly variable SSRs and SNPs. It also explicitly models the dependencies across loci within haplotypes due to their expected correlated histories. Thus the DHS method has many desirable features for application to the problem of detecting positive selection by comparing an allele's frequency to its age.

Our method to test whether alleles at a region of interest are compatible with models of neutral evolution can be summarized as follows. Collect haplotype data from a population using markers that flank a region of interest (*e.g.*, a gene or exon) and span up to 1 cM. Sort these haplotypes on the basis of the alleles they carry at the region of interest. Estimate the ages of these alleles with the DHS method from the observed decay of haplotype sharing at the flanking markers within this haplotype set and compared to the background LD and allele frequencies found at the marker loci on the remaining haplotypes. Compare the frequency of the alleles at the region of interest with the estimated ages of these alleles to identify young alleles at unexpectedly high frequencies. Because natural selection specifically discriminates between alternative alleles at a locus, multiple alleles help resolve whether a pattern is due to selection or to processes that act more generally in regions, such as heterogeneity in recombination rate, or, in populations, such as genetic drift and demographic history. Evaluate the compatibility of the observed relationship between allele age and frequency in the region to neutral evolutionary models by simulating haplotypes that include an allele at the region of interest and at flanking markers and by estimating the age of these alleles using the DHS method. These simulations model uncertainty in the genealogy of alleles and provide an appropriate statistical comparison for the observed alleles. Compare the age of observed alleles with the distribution of ages for simulated alleles at the same frequency produced under different demographic models. Alleles that are younger than the vast majority or all of the simulated alleles are unlikely to occur by chance under the neutral models and indicate the possible role of selection.

As a demonstration, we applied this method to alleles at the human *HFE* gene. *HFE* (Feder *et al.* 1996) provides a model with which to investigate the occurrence of a disease allele at a higher-than-expected frequency, possibly due to positive selection. Mutations in the *HFE* gene are responsible for most cases of hereditary hemochromatosis, an autosomal recessive disease characterized by iron overload and common in populations of European descent. When homozygous, one *HFE* allele, C282Y, is responsible for >80% of this disease and occurs with a frequency of 5–10% in northern European populations (Merryweather-Clarke *et al.* 1997; Lucotte 2001). There has been speculation on whether the C282Y mutation has reached its present frequency through genetic drift, expansion by a bottleneck, hitchhiking, or positive selection to prevent iron deficiency in iron poor diets (Rotter and Diamond 1987; Fairbanks 2000). Like many of the other disease genes posited to be positively selected, little progress has been made in resolving this question.

Our method benefits from the implementation of three approaches that overcome some of the difficulties of previous methods in detecting patterns characteristic of recent positive selection: (1) estimating allele age by the DHS method; (2) testing empirical allele age estimates against neutral models that include realistic demographic parameters with the use of coalescent simulations; and (3) comparing results between eight *HFE* alleles that act as a control in interpreting patterns for any individual allele.

MATERIALS AND METHODS

DNA samples: Samples were collected from Utah and surrounding states (Edwards *et al.* 1988). Because the samples come from pedigrees and most markers in the study are highly polymorphic, the full haplotype for these markers could be inferred for each chromosome. The Institutional Review Board of the University of Chicago approved this project.

SNP typing: We typed a subset of highly informative SNPs at the *HFE* locus (Toomajian and Kreitman 2002) and from them inferred the total 11-kb sequence for our 210 samples. The published haplotype structure of the *HFE* locus (Toomajian and Kreitman 2002) guided our choice of SNPs to type and the classification of each chromosome into a haplotype class. The specific SNPs typed for each sample varied, as the typing was done hierarchically and the results of the first SNPs typed were used to decide which additional SNPs to type. Depending on the specific SNP, typing was done by one of the following methods: (1) sequencing diploid PCR products of ~1 kb to determine the SNP genotype (Toomajian and Kreitman 2002) or (2) analyzing diploid PCR products of between 200 and 500 bp by denaturing high-performance liquid chromatography (DHPLC) to determine SNP genotypes (WAVE System, Transgenomic, Omaha). DHPLC temperature and gradient conditions were chosen on the basis of predictions of the WAVEMaker software (Transgenomic) and were adjusted as necessary. To assign homozygous samples to the correct allele and to verify that heterozygous samples were heterozygous at the specific SNP allele of interest, we mixed additional PCR product from each sample with a known sequence control.

Inference of the 11-kb sequence: We used pedigrees to resolve haplotypes for the SNP alleles and place these haplotypes in the context of the broader marker haplotypes. This process was simplified in individuals heterozygous for the C282Y mutation, because they almost always carry the common C282Y haplotype (22). SNP haplotypes for the non-C282Y-bearing chromosome could then be resolved. Eight non-C282Y chromosomes had been fully sequenced (Toomajian and Kreitman 2002) for the *HFE* region. For the other 132 non-C282Y chromosomes, a minimum of seven SNPs were phased. For 123 chromosomes that carried the C282Y allele, a minimum of six SNPs were phased. When the phase of some SNPs could not be determined from segregation in the pedigree, the phase consistent with two previously observed haplotypes was assumed. This amount of SNP typing was sufficient to infer the haplotypes on the basis of previously defined haplotypes (Toomajian and Kreitman 2002) and allowed a high probability of detecting new recombinant haplotypes.

Estimating allele age: We used DHSMAP v. 1.04 software (McPeek and Strahs 1999) to generate a summary of LD around the C282Y allele and other *HFE* SNP alleles by assessing the decay of haplotype sharing at multiple markers flanking a mutation. The program takes as input a set of marker haplotypes from chromosomes that carry the SNP allele of interest, haplotypes from a control population (in our case, the random sample of chromosomes that lack this SNP allele), a genetic map of the markers, and a mutation rate for each marker. It then infers the ancestral haplotype surrounding the specified SNP allele and produces as output a maximum-likelihood estimate of the TMRCA in generations for the allele class. We refer to these TMRCA values as allele ages for the purpose of this study but acknowledge the distinction between the two terms. The method assumes that there is no selection at any of the marker loci and makes no assumption about selection at the allele-class-defining locus. Information on the 23 markers outside of the 11-kb *HFE* region used in creating multilocus haplotypes is listed in Table 1. Possible errors in inferring the 11-kb haplotypes would have minor effects on the DHSMAP output, since it is primarily the flanking markers that determine the level of haplotype sharing for each allele.

Our pedigree samples are biased for the C282Y allele. For the purpose of estimating allele age, we constructed a sample that includes 140 independent non-C282Ychromosomes and 11 randomly chosen independent C282Y chromosomes. This constructed random sample corrects for the C282Y allele bias, so that its frequency is ~7%, the estimate for the population from which these samples were drawn (Edwards *et al.* 1988). In estimating the age of each allele except C282Y, only alleles from this set of 151 chromosomes were used. For the C282Y allele, 123 independent C282Y-bearing chromosomes were used, while the remaining 140 independent chromosomes served as the control class. This should produce a better age estimate for the C282Y allele than for the other *HFE* alleles but does not bias the age estimation process.

Coalescent simulations: Simulations were performed by modifying MS, a C program from R. Hudson, which incorporates recombination, population size changes, and population subdivision (Hudson 1990). We adapted additional code provided by R. Hudson that simplifies the way in which recombination is handled and integrated this into the MS framework. Rather than allowing crossing-over events between adjacent bases in a contiguous sequence with uniform probability, crossing over occurs between loci used in the study (not within) at a rate proportional to the genetic distance between the loci, thus limiting the number of genealogies the program must follow. We produce mutations at discrete loci separated by fixed genetic distances. A diallelic polymorphism is created at the focal site by randomly placing a single mutation on the genealogy of the focal site. Individual coalescent samples are later weighted by the length of the total genealogy at the focal site. At all other loci, mutations occur according to a Poisson process with rate equal to the per-generation mutation probability for that locus. Mutations at these loci change the allele size and occur according to a one-step mutation model, so that the allele size of all descendants carrying this mutation increases or decreases by one with equal probability (Ohta and Kimura 1973). The DHSMAP analysis does not assume a particular mutation model; it records only whether or not alleles are identical, so that our choice of a one-step mutation model is inconsequential here. In particular, it keeps track of whether alleles at linked loci are the inferred ancestral allele for a focal mutation, although it does allow for mutation or recombination to reintroduce the ancestral allele onto a chromosome that had previously carried a nonancestral allele.

Estimating the ages of simulated alleles: When a comparison was made between simulated and *HFE* SNP alleles, the age estimate based only on SSRs was used. When the assumption of constant population size is relaxed, the expected TMRCA for two randomly drawn alleles will change and affect the expected variance in SSR allele size. We empirically estimated this deviation by measuring the average variance in allele size of a SSR in a simulated sample that had undergone population size changes and comparing it to the constant population size expectation. To mirror the observed variance in allele size for each SSR in our simulations of demographic models, we adjusted the mutation rate to compensate for the change in the TMRCA for two randomly drawn samples.

RESULTS

We generated the data needed to estimate the ages of several alleles at the *HFE* locus by two steps: (1) determine individual haplotypes for the SNPs that segregate in an 11-kb region encompassing the *HFE* locus in a Caucasian sample (Toomajian and Kreitman 2002)

TABLE 1

Marker type, spacing, allele number, heterozygosity, and mutation rate

Name	Type[a]	Centimorgans to next locus[b]	No. of alleles	Heterozygosity	VAS-estimated μ ($\times 10^{-4}$)[c]
HLA-B	Serotype	0.04	26	0.911	
HLA-C	Southern probe	0.138	3	0.584	
Y104	Southern probe	0.03	2	0.441	
HLA-E	Southern probe	0.0552	3	0.461	
Y129	Southern probe	0.012	2	0.410	
Y158	Southern probe	0.03	3	0.608	
HLA-A	Serotype	0.01	15	0.829	
D6S265	SSR	0.046	7	0.781	1.019
HLA-G	Southern probe	0.0232	2	0.507	
HLA-F	Southern probe	0.4	2	0.302	
D6S306	SSR	0.09	7	0.676	1.081
D6S464	SSR	0.01	11	0.661	4.019
D6S105	SSR	0.12	9	0.822	1.731
D6S1260	SSR	0.03	9	0.671	1.439
D6S1558	SSR	0.13	8	0.510	1.057
D6S2231	SSR	0.06132	12	0.849	4.799
D6S2238	SSR	0.00741	9	0.812	2.719
HFE	SNP haplotypes	0.0129			
D6S2239	SSR	0.1744	3	0.656	0.379
D6S2241	SSR	0.15	4	0.500	0.144
D6S1621	SSR	0.4	7	0.710	1.339
GATA	SSR	0.7	8	0.761	0.806
D6S1545	SSR	1	6	0.710	0.894
D6S1691	SSR		16	0.889	11.800

[a] Reference for markers are given in Ajioka *et al.* (1997). All SSRs are dinucleotide repeats except for GATA, which is a tetranucleotide repeat.

[b] Genetic distances were calculated primarily from physical distances used in Ajioka *et al.* (1997); genome sequence information was used to recalculate physical distance for markers closest to *HFE*. On the basis of experimentally measured recombination in the region (Ajioka *et al.* 1997; Malfroy *et al.* 1997), the conversion 1 cM = 5 Mb was used centromeric to *HFE* while the conversion 1 cM = 1 Mb was used telomeric. Estimates of recombination throughout the HLA region (*i.e.*, farther centromeric to HLA-A) are not likely as low as 1/5 cM/Mb (Carrington 1999). Underestimates in these rates affect only the age estimates of alleles with the most haplotype sharing, producing slightly overestimated ages (making our comparisons conservative).

[c] Mutation rates were estimated on the basis of the *v*ariance in *a*llele *s*ize (VAS) of SSRs under a one-step mutation model (Valdes *et al.* 1993). A constant effective population size of 10^4 was assumed.

and (2) construct longer haplotypes for these chromosomes by determining variants segregating at many additional sites spread over a multi-megabase region surrounding the locus. We took advantage of a large sample of individuals from Utah with extended pedigrees for which haplotypes had previously been deduced for 24 markers spread over 8 Mb (Ajioka *et al.* 1997) around the *HFE* locus. For these chromosomes we determined *HFE* haplotypes by genotyping SNPs in the 11-kb *HFE* region (Toomajian and Kreitman 2002; Figure 1). Specifically, we genotyped 210 individuals from 65 hereditary hemochromatosis pedigrees (Edwards *et al.* 1988). The samples represent 123 independent chromosomes that carry the C282Y allele and 140 independent chromosomes without this allele. Figure 2 shows a mutational network of *HFE* SNP haplotypes inferred for these samples on a background of worldwide haplotypes (Toomajian and Kreitman 2002).

Estimating allele ages: The above data allowed us to determine the age of alleles in the *HFE* locus, including *HFE* C282Y, by assessing the decay of haplotype sharing to more distant sites. We selected eight SNP alleles for age estimation by DHSMAP, a program that implements the DHS method (McPeek and Strahs 1999). Alleles were chosen to minimize the correlation due to their tight physical linkage. The individual genealogies of each of these alleles are less correlated, as most of these SNP alleles mark a different haplotype uniquely and tend to fall on external branches of the *HFE* network (Figure 2). In most cases, the allele classes they define represent almost completely nonoverlapping portions of the sample, but two exceptions exist. To include the relatively common haplotype 3, we estimate the age of the 1972T allele, although this allele is also found in the C282Y haplotype (22). The age estimate of 1972T may be influenced by the homogeneity of the C282Y haplotype samples with regard to the flanking markers; *i.e.*, the age of 1972T may be underestimated. We esti-

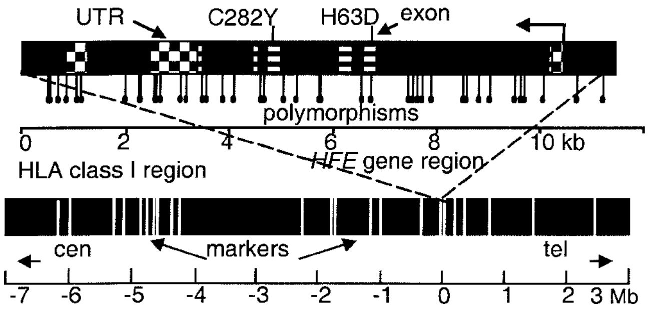

FIGURE 1.—Map of *HFE* gene region showing the location of markers used in this study. For orientation, the chromosome 6p telomere is to the right. A detailed map of the region surveyed for *HFE* polymorphisms, including the intron/exon structure of *HFE* and the location of polymorphisms, is shown at the top.

mate the age of 7522C to include haplotype 9, although haplotype 8, which carries the identifying 709T allele, also carries this allele. In this case, similar results are obtained for 7522C and 709T, likely due to the small number (2) of haplotype 9 samples.

Table 2 shows the frequency of these alleles and their age estimates for a range of assumed SSR mutation rates (10^{-5}–10^{-3}). Our best estimate for the age of C282Y is 138 generations, with a range from 88 to 156. When interpreting the frequency of an allele in light of its estimated age, the comparison with many presumably neutral alleles provides a control for the effects of population demography and heterogeneity in mutation and recombination rates in the genomic region under study. Relative to other alleles, C282Y is indeed young, given its frequency. However, the 7633A allele produced an even more remarkable result, as it is at higher frequency and has a younger age estimate (78 generations). The next youngest allele after 7633A and C282Y is 4600G at 529 generations, with the remaining alleles ~1000 generations or older when estimating SSR mutation rates from their variance in allele size (Table 2, row A). Deviations from the one-step mutation model of SSRs, which is assumed in estimating these mutation rates, would mean that our mutation rate estimates may be high, although the true value would likely still be within the range of values we consider. Age estimates depended on the assumed marker mutation rate because DHSMAP estimates age from the effect on haplotype sharing of both recombination and mutation. Alleles with much haplotype sharing (7633A) were less sensitive to misspecification of mutation rate than those with little haplotype sharing (H63D, Table 2 rows D, F, and H). Also, the difference between ages estimated with mutation rates 10^{-4} and 10^{-5} were small compared to the difference for 10^{-4} and 10^{-3}. If the true mutation rate for the SSRs is closer to 10^{-3}, then age estimates decrease. This makes the young age estimates of 7633A and C282Y more extreme but is probably unlikely since all *HFE* alleles <30% would appear fairly young (TMRCA of <550 generations).

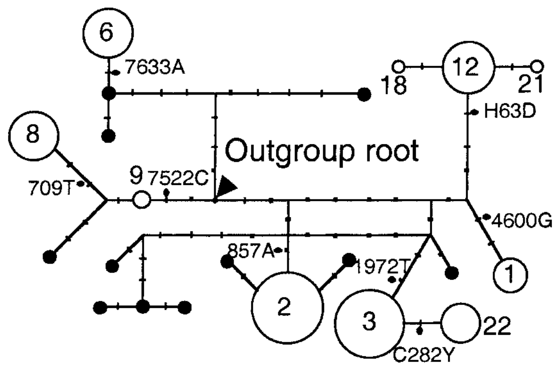

FIGURE 2.—Reduced median network (BANDELT *et al.* 1995) of *HFE* sequence haplotypes from a worldwide population. Haplotypes are numbered according to their frequency in the worldwide population (1 is most frequent). Open circles represent haplotypes sampled from hemochromatosis pedigrees, with size proportional to frequency in pedigree samples. The frequency of haplotype 22, the C282Y-bearing haplotype identified by sequencing (TOOMAJIAN and KREITMAN 2002), is corrected for biased sampling. Mutational differences between haplotypes are indicated on the branches of the network, and the locations of sites used to define allelic classes for age estimation are labeled. The network has been modified to include two haplotypes (21 and 22) found in hemochromatosis pedigrees that are one mutational step away from haplotypes found in a worldwide sample (TOOMAJIAN and KREITMAN 2002). Five haplotypes found in six non-C282Y chromosomes and one C282Y chromosome of the pedigree samples and likely due to recombination are not displayed.

Test of neutrality using coalescent simulations: Having estimated young ages for two common alleles at *HFE*, we tested whether these two alleles were compatible with models that assume no positive selection. To assess the likelihood under neutrality of age estimates for individual *HFE* alleles, we simulated haplotypes under a range of parameter values (SSR mutation rate, recombination rate, and different demographic scenarios) using coalescent theory (program available upon request). The coalescent simulations model the genealogy of the entire population and account for uncertainty in the number of lineages present in past generations for particular alleles (SLATKIN and RANNALA 2000). Haplotypes consisted of one focal SNP and 14 flanking SSR loci spaced on a genetic map as the 14 SSRs from the observed data were. We simulated replicates of 151 chromosomes (the size of our constructed random sample of observed *HFE* haplotypes) for each combination of parameter values, and for each replicate estimated the age of the focal SNP allele with DHSMAP. We used SSR mutation rates estimated from the variance in allele size observed at each locus and assumed these same mutation rates in estimating age. Simulated haplotypes, comparable to the observed haplotypes around *HFE*, served as a null distribution for the relationship between allele frequency and age estimated from haplotype sharing. Ob-

TABLE 2

DHSMAP allele age estimates (in generations) for different mutation rates

	Details of estimate		*HFE* SNP alleles/frequency							
	Marker set[a]	μ[b]	4600G/ 0.06	C282Y/ 0.07	709T/ 0.11	7633A/ 0.11	7522C/ 0.13	H63D/ 0.16	857A/ 0.24	1972T/ 0.30
A	1 + 3	VAS[c]	529	138	939	78	1030	1722	986	855
B	1	VAS	516	138	1095	85	1496	1830	1265	785
C	1 + 3	10^{-4}	564	141	1207	81	1316	2066	1301	1162
D	1	10^{-4}	548	143	1545	90	2029	2272	1441	1070
E	2 + 4	10^{-3}	366	88	558	47	666	1029	477	543
F	1	10^{-3}	249	N/A[d]	455	58	455	538	315	357
G	1 + 3	10^{-5}	652	151	1499	84	1667	2709	1526	1687
H	1	10^{-5}	638	156	2262	97	2931	3866	1779	1578

[a] Marker sets: 1, All SSRs; 2, centromeric SSRs; 3, non-SSR markers (centromeric markers and *HFE* SNPs); 4, *HFE* SNPs.

[b] SSR mutation rate assumed; the mutation rate of all non-SSR loci was set to 5×10^{-7}.

[c] Mutation rate estimated from variance in allele size (Valdes *et al.* 1993); for most markers this was ~10^{-4} (see Table 1).

[d] N/A: DHSMAP failed to estimate an allele age.

served alleles were compared to the null distribution to find the probability of the allele age estimate given allele frequency.

Simulation replicates from a constant population size model included at least 1000 representatives in each of 41 allele-frequency classes (from 6/151 to 46/151). Figure 3 displays the probability distribution for each allele-frequency class as percentile values of allele age. Points corresponding to the ages and frequencies of alleles at *HFE* are overlaid for comparison. Both the C282Y and 7633A alleles are at the tail of the distribution for their frequency classes and are thus significantly younger than expected under this neutral model ($P < 0.01$ and $P < 0.001$, respectively). In random population samples, different numbers of chromosomes carrying C282Y or 7633A will be found due to sampling variance. Given our age estimates for these alleles, at least nine copies of C282Y in a sample of 151 chromosomes must be found for the age of this allele to fall below the 2.5 percentile value of the simulated haplotypes, and at least nine or six copies of 7633A must be found for the age of this allele to fall below the 1 and 2.5 percentile value, respectively, of the simulated haplotypes. Also, the reported P values should not be taken at face value since multiple tests have been performed. These tests are not independent, so the standard Bonferroni correction is too conservative and not appropriate. Nevertheless, we note that while the C282Y result would not remain significant after this correction for multiple tests, the 7633A result would.

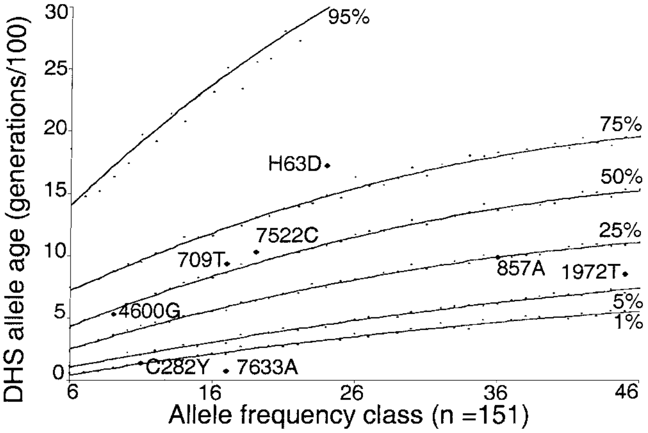

Figure 3.—Comparison of the observed relationship between *HFE* allele age and frequency with that produced by simulation. Plotted for each allele-frequency class are six percentiles of the estimated allele age distribution. Placed in the context of these percentile values are the frequencies and estimated ages for eight alleles found at *HFE*.

Alternative demographic models: We simulated samples under different demographic models (Table 3) to explore their effects on the estimation of allele ages by DHSMAP. The significance of *HFE* allele ages under the various demographic models is indicated by the empirical cumulative probability distributions of ages for alleles in three different frequency classes (Figure 4). Under the simple exponential growth model, the age distributions shifted toward higher values and the significance of the C282Y estimate increased. This is expected, since LD should be less in a fast-growing population (Slatkin 1994). Studies have found that population bottlenecks can increase the level of LD (Ardlie *et al.* 2002). The old bottleneck model produced an average age similar to the constant population size model, but with a reduced range of ages. Therefore, both age extremes were less probable for this bottleneck model and the significance of the C282Y allele's young age increased. For the extreme case of a recent bottle-

TABLE 3

Demographic models

	Initial size	Final size	Start of growth (generation)	Bottleneck size	Length of bottleneck (generation)
Constant	10^4	10^4			
Exponential growth	10^4	10^5	2500		
Following a bottleneck					
Old	10^4	10^6	500	10^3	500
Recent	10^4	10^6	50	10^3	500

neck only 50 generations ago, the age of the C282Y allele lost its significance. In fact, only age estimates <750 generations were very likely for this model. If an event such as a recent bottleneck were responsible for C282Y's young age, several *HFE* alleles might also show levels of haplotype sharing that point to similar ages. This model was inconsistent with age estimates of most other *HFE* alleles, suggesting that the model is inappropriate for this population.

The significance of the young age estimate for C282Y is also insensitive to error in the estimate of recombination rates throughout the region studied. Figure 4 shows the empirical cumulative probability distribution produced when samples were simulated assuming the genetic distances between all markers were only half of the values used to estimate allele ages by DHS (1/2 ρ). In this case, the age of the C282Y allele is borderline significant ($P < 0.06$). However, the low recombination rate used implies a genetic distance between human histocompatibility-A (HLA-A) and *HFE* that lies below the 95% confidence interval estimated from recombination events observed between these loci (Ajioka *et al.* 1997; Malfroy *et al.* 1997).

Comparison with other *HFE* alleles: The estimated age of 7633A is significantly younger than expected on the basis of its frequency for all models considered. Only a very extreme founder effect would produce such young age estimates with any likelihood. Yet no evidence exists for a severe founder effect on the basis of estimated ages of other *HFE* alleles. For example, Figure 4b includes the age of the 709T allele, found at the same frequency as allele 7633A. It is older than 97% of the ages estimated for simulated alleles in this frequency class under the recent bottleneck model. Figure 4c gives the example of the H63D allele at 16% frequency. This allele is not consistent with either bottleneck model tested, with ~95 and 99.5% of simulated alleles having younger ages under an old bottleneck and a recent bottleneck, respectively. If a recent and severe founding effect occurred in this population, then the old age estimated by DHS at H63D would suggest that it was preferentially preserved through the founding bottleneck. Thus, the comparison of a broad range of alleles from the *HFE* locus and their patterns of haplotype-sharing decay can jointly determine likely demographic parameters of this sample and detect particular alleles

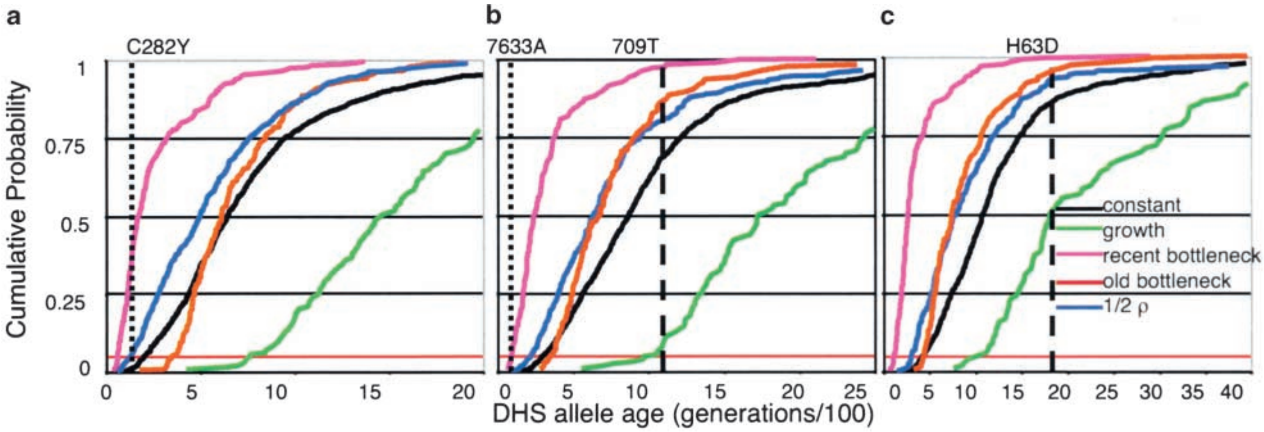

Figure 4.—The effect of population size changes or overestimation of recombination rates on the significance of the *HFE* allele ages. Shown are the empirical cumulative probability distributions of ages for alleles in frequency class 11 (a), 17 (b), and 24 (c) from a sample of 151 haplotypes. These ages were estimated from samples produced by coalescent simulation under various demographic models (Table 3) and by a simulation in which the recombination rate (ρ) between each marker was half of the level used in estimating age. The dashed vertical lines indicate the age of the *HFE* alleles estimated from the haplotype sharing observed at only the flanking SSR markers.

that depart significantly from the expectations of neutral demographic models.

DISCUSSION

Our investigation of the empirical relationship between the decay of haplotype sharing around different *HFE* alleles and their frequency supports the hypothesis that *HFE* C282Y has increased in frequency due to selection and detects an additional selected haplotype in the *HLA-HFE* region. Multilocus LD estimated from the decay of haplotype sharing provides a clear indication of which haplotypes have been affected by natural selection, and our method may be able to indicate recent positive selection too subtle to be detected by current tests that emphasize levels of variation rather than allele association (Hudson *et al.* 1994). Due to the phenomenon of hitchhiking, however, evidence for the selective sweep of a haplotype does not indicate with certainty the specific genetic change under selection (Maynard Smith and Haigh 1974).

Selective events that are detectable by haplotype sharing may also be detected using methods that focus on the reduction in variation in an allele class (*e.g.*, the haplotype test of Hudson *et al.* 1994). But our method has certain advantages. Haplotype sharing and haplotype variability tests use different aspects of empirical data, primarily LD information *vs.* variation levels, respectively. Allele classes subject to genetic drift alone that are at low or moderate frequencies are expected to have reduced variability at linked sites. In addition, if one considers the variability level of an allele class for which there is, say, a 5% probability of observing that number or fewer segregating sites by drift alone, this level can be extremely low for low-frequency allele classes. Lower-frequency allele classes that are subject to positive selection, therefore, may require haplotype information from impractically large regions to exhibit a statistically significant reduction in variation. The haplotype-sharing approach, in contrast, leverages how variability is patterned into haplotypes and the extent to which reduced variability centers around a focal site. This, we believe, allows for a more economical test of very recent selection in terms of linked variants needed to achieve significance.

Highly polymorphic SSRs in a haplotype test may not be as informative as in a haplotype-sharing test because the high mutation rate of SSRs will rapidly increase the number of variants within an allele class. As an example, the SSRs we use are quite variable, and there is not an overwhelming absence of variability in the C282Y allele class. Our study included SSRs not because they increase variability, but because they were the markers that had already been typed in the appropriate samples. Without SSRs, more SNPs would probably be needed to provide the same amount of information on how far ancestral haplotypes extended along chromosomes, but their lower mutation rate would provide a clearer picture of ancestral haplotype extent and result in more accurate age estimates. Our method could also be applied to previously ascertained SNPs, but doing the same with the haplotype test may be problematic because it would require, impractically, many ascertained markers to detect a significant reduction in variation. In some cases, detecting a significant difference in SNP variability between selected *vs.* neutral allele classes may be more difficult than detecting a significant difference in the level of haplotype sharing.

The age of *HFE* C282Y: Our study emphasized the comparison of haplotype sharing for multiple observed SNP alleles and alleles produced in neutral coalescent simulations, so that no conclusions were based on absolute ages alone. Still, comparisons with two previous studies that have used the observed decay of LD to estimate the C282Y allele age show that the DHS method performs well. Our estimate (138 generations) fell within the 95% confidence interval (27–161 generations) for the estimate (59 generations) of Ajioka *et al.* (1997). This interval would be even larger if the uncertainty in the pairwise LD between C282Y and HLA-A were taken into account. Our estimate was closer to the multilocus estimate of Thomas *et al.* (1998), although it lay outside their confidence interval. Their average estimate over a set of markers telomeric to *HFE* was 62 generations, and they reported confidence intervals presumably calculated from the variance among estimates for individual markers. They also produced an estimate of 77 generations from the best fit among the marker loci. Their confidence interval is too narrow, as it is produced with DISMULT and DISLAMB (Terwilliger 1995) by assuming that age estimates using different linked markers are independent.

Aside from the technical limitations of these age estimates, directly interpreting the young age of an allele, given its frequency, as clear evidence of positive selection is inappropriate without formal statistical testing. Therefore, the fact that the DHS method produces estimates of the TMRCA of an allele class rather than its true age is inconsequential. Because both observed and simulated haplotypes were evaluated by the same method, the conclusions from our statistical testing are robust to any biases inherent in the way we estimate allele age. Inaccuracies in allele age estimates do not affect our ability to evaluate the contribution of natural selection and population demography to the current frequencies of young alleles. We used the DHS method to measure the decay of LD because it is designed to use information from several linked markers, can account for the dependence across loci within a haplotype, and is an improvement over other multilocus LD methods that do not (Terwilliger 1995). Still, given the wide interest in improving the performance of LD mapping, several new multipoint LD methods have recently been described and could be adapted for detecting patterns of selection evident in LD data (Liu *et al.* 2001; Rannala and Reeve 2001).

The accuracy of genetic distances is important in all methods of estimating the age of an allele on the basis of patterns of LD. The genetic distances we used in our *HFE* example (Table 1) relied on the measured recombination rates in the *HFE* region, where the rate centromeric to *HFE* is estimated to be one-fifth the rate telomeric (Malfroy *et al.* 1997). Our age estimates produced using markers only centromeric or only telomeric to *HFE* gave similar results when we made this assumption (data not shown), supporting its validity. Comparisons between alleles of the same locus should be relatively insensitive to global errors in estimated genetic distances, but recombinational heterogeneity on the scale of the haplotype sharing for some alleles, such as from hotspots (Jeffreys *et al.* 2001), may affect the comparison of age estimates between alleles. Results for 7633A and C282Y would be insensitive to hotspots, as their ancestral haplotypes span long distances. If recombination were clustered near *HFE* and very rare farther away from it, the short ancestral haplotypes of the other *HFE* alleles would be consistent with younger ages. This scenario seems unlikely, and the 7633A and C282Y alleles are still clearly younger than other alleles. Further study of fine-scale patterns of recombination in the human genome will only improve our ability to make evolutionary inferences about the level of haplotype sharing around alleles.

Alternative demographic models: Recent data on human variation indicate that complex demographic histories should be used as null models in the analysis of human population data (Przeworski *et al.* 2000). One of the strengths of our method is the ability to explore alternative demographic models. The demographic models we test in this example were not meant to be exhaustive but instead identified which types of models make the observation of larger regions of haplotype sharing more likely. The estimated age of the C282Y allele is consistent with neutrality for a bottleneck on the order of 100 generations ago. When little independent data exist about the demographic history of a population, the full set of alleles at a locus can provide the basis to exclude certain demographic scenarios as inconsistent with the general pattern of LD decay. For example, patterns of haplotype sharing for most other alleles at *HFE* do not support a recent bottleneck. Also, a recent and severe bottleneck is expected to affect levels of nucleotide diversity and the frequency spectrum, neither of which is evident in nucleotide polymorphism data for a collection of European samples (Toomajian and Kreitman 2002).

In a study of LD in 19 random genomic regions, Reich *et al.* (2001) reported results consistent with a major bottleneck in the same population. However, assuming nearly the same bottleneck severity as we have, they estimated that it occurred between 800 and 1600 generations ago. This bottleneck model would not change our conclusion that the C282Y-bearing haplotype has been positively selected. LD can also be increased by other factors such as inbreeding, population structure, and admixture (Pritchard and Przeworski 2001). The inclusion of these factors in more complicated demographic models may help to explain more of the results from studies of SNP variation and LD in northern European populations. Coalescent methods like the one we used have been adapted to incorporate these factors, so that as more data emerge, more realistic null models can be tested against with our general approach.

***HFE* and natural selection:** Independent data on HLA haplotypes suggest that the *HFE* 7633A allele has hitchhiked on a particularly large chromosomal region. The 7633A mutation is located in an intron and has no known effect on HFE levels or function. The inferred ancestral allele at HLA-B for the *HFE* 7633A allele class is B8. Most Caucasians that carry HLA-B8 share alleles on a common haplotype that extends throughout the HLA region and includes HLA A1. This extended haplotype has reached 10% frequency in northern European populations while remaining more or less intact (Price *et al.* 1999). Strong LD over multiple loci is a common finding in the HLA region, but the particular homogeneity of the A1-B8 extended haplotype has led to the conclusion that it has a very recent origin, likely due to strong selection (Bugawan *et al.* 2000). Our evidence for this extended haplotype in the *HFE* region is consistent with the finding (Worwood *et al.* 1997) that this haplotype extends to SSR loci distal to *HFE*. Although physical suppression of recombination specific to the A1-B8 haplotype (Pichon *et al.* 1996) could explain the extent of this frequent HLA haplotype, a study of a large number of HLA recombination events does not reject the null hypothesis of equal recombination rates for each haplotype (Termijtelen *et al.* 1995).

Direct positive selection on the C282Y mutation is a viable possibility, given what is known about its biological effect on HFE function (Feder *et al.* 1998; Salter-Cid *et al.* 1999). However, since the locus was chosen with the knowledge that the frequency of one allele (C282Y) was high relative to its apparent age and phenotypic effect, we must be cautious in assessing the significance of our results *vs.* models of neutrality. In addition, alternative hypotheses about the specific target of positive selection within the HLA A3-B7-DR2 extended haplotype (which includes the C282Y allele; Worwood *et al.* 1997) cannot be ruled out. This extended haplotype is not as homogeneous as the A1-B8 haplotype and must therefore have a more ancient common ancestor. Because both the A3-B7 haplotype and the *HFE* C282Y allele appear older than the HLA A1-B8 haplotype, it is difficult to make conclusions about the number and location of targets of selection on this haplotype without further work.

It should be possible to apply our method described here to several regions within a potentially selected long-range haplotype such as the *HFE* 7633A-HLA A1-B8 example above to find evidence of the specific target of selection. However, it is not always clear how results should be compared between loci. For example, we estimated the age of alleles at the HLA-A locus and the adjacent SSR D6S265 (data not shown). In each case, the alleles estimated to be ancestral on the C282Y and 7633A haplotypes either were the youngest of the set of alleles at the locus or appeared young given their frequency, consistent with the hypothesis that these two haplotypes have been affected by positive selection. Although at neither locus was the ancestral allele age for either haplotype younger than the age estimated for the respective *HFE* allele, it would not be appropriate to conclude that in both cases the target of selection is an allele at the *HFE* gene. The locus spacing of this study is more suited to producing accurate age estimates for alleles at *HFE*, with many close and highly polymorphic markers surrounding the gene. SSRs pose additional problems since one cannot assume that all sequences in an allele class are descended from a unique mutation event. The nature of allele designations at the HLA-A locus (our alleles are determined by serotype) and its high polymorphism level also make comparisons with *HFE* alleles difficult. Further work and the inclusion of additional markers in the analysis may provide more resolution to this question.

The possible action of negative selection on an allele must also be considered when explaining an allele age lower than expected given its frequency. Selection on an allele reduces the expected age of the allele given its frequency regardless of the direction of selection (Maruyama 1974). For selection to greatly influence the age of an allele at the frequency of C282Y, this allele should have a fitness effect observable in both homozygotes and heterozygotes. The hypothetical fitness advantage of increased iron stores may occur in both heterozygotes and homozygotes of the C282Y allele. However, there is little evidence that this mutation can have a deleterious effect in heterozygotes (Bulaj *et al.* 1996), and a recent study by Beutler *et al.* (2002) calls into question whether the mutation has an appreciable deleterious effect on fitness in homozygotes. It also seems unlikely that a deleterious allele could reach the frequency of several percent in many European populations, as C282Y has (Lucotte 2001). Although we consider it unlikely that negative selection can explain the specific pattern seen for the C282Y allele, we expect that many alleles at 1% or lower frequency that show haplotype-sharing patterns inconsistent with neutrality will be due to negative selection.

Conclusion: We have described a general method for testing human LD data *vs.* neutral theory predictions about the expected relationship between allele age and frequency under demographic models that can be simulated with the coalescent approach. With an effort now underway to determine the haplotype structure for the entire human genome, methods such as ours are likely to have widespread application. Humans have almost certainly undergone strong recent selection for many different traits, not the least of which is resistance to infectious disease, and methods to detect positively selected haplotypes might provide positional information about the selected allele, not unlike LD mapping itself. Although in most cases our method will not allow identification of the specific locus under positive selection, the number of loci in many cases will be small enough to allow a candidate to be identified. The human data that will be available soon, and the application of methods to detect selection, will dramatically improve our understanding of both human history and molecular evolutionary processes acting in natural populations.

We thank A. Strahs and M. S. McPeek for providing the source code and assistance in implementing DHSMAP and R. Hudson for providing computer code used in the coalescent simulations. Thanks go to R. Hudson, E. Stahl, J. Comeron, M. Antezana, and M. S. McPeek for helpful discussions on the specific form of the simulations and tests of the algorithm's performance. Two reviewers provided helpful comments on the manuscript. This work was supported by National Institutes of Health (NIH) grants GM39355 to M.K., GM-59290 to L.B.J., R01 DK20630-21 and M01 RR 00064 to R.S.A. and J.P.K., and a National Science Foundation Doctoral Dissertation Improvement Grant (DEB-0073297) to C.T. and M.K. C.T. was partially supported by a Howard Hughes Medical Institute predoctoral fellowship and by NIH training grant no. T32 GM07197.

LITERATURE CITED

Ajioka, R. S., L. B. Jorde, J. R. Gruen, P. Yu, D. Dimitrova *et al.*, 1997 Haplotype analysis of hemochromatosis: evaluation of different linkage-disequilibrium approaches and evolution of disease chromosomes. Am. J. Hum. Genet. **60:** 1439–1447.

Andolfatto, P., J. D. Wall and M. Kreitman, 1999 Unusual haplotype structure at the proximal breakpoint of In(2L)t in a natural population of *Drosophila melanogaster*. Genetics **153:** 1297–1311.

Ardlie, K. G., L. Kruglyak and M. Seielstad, 2002 Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. **3:** 299–309.

Bandelt, H. J., P. Forster, B. C. Sykes and M. B. Richards, 1995 Mitochondrial portraits of human populations using median networks. Genetics **141:** 743–753.

Beutler, E., V. J. Felitti, J. A. Koziol, N. J. Ho and T. Gelbart, 2002 Penetrance of 845G→A (C282Y) HFE hereditary haemochromatosis mutation in the USA. Lancet **359:** 211–218.

Bugawan, T. L., W. Klitz, A. Blair and H. A. Erlich, 2000 High-resolution HLA class I typing in the CEPH families: analysis of linkage disequilibrium among HLA loci. Tissue Antigens **56:** 392–404.

Bulaj, Z. J., L. M. Griffen, L. B. Jorde, C. Q. Edwards and J. P. Kushner, 1996 Clinical and biochemical abnormalities in people heterozygous for hemochromatosis. New Engl. J. Med. **335:** 1799–1805.

Carrington, M., 1999 Recombination within the human MHC. Immunol. Rev. **167:** 245–256.

Edwards, C. Q., L. M. Griffen, D. Goldgar, C. Drummond, M. H. Skolnick *et al.*, 1988 Prevalence of hemochromatosis among 11,065 presumably healthy blood donors. N. Engl. J. Med. **318:** 1355–1362.

Fairbanks, V. F., 2000 Hemochromatosis: population genetics, pp. 42–50 in *Hemochromatosis: Genetics, Pathophysiology, Diagnosis and*

Treatment, edited by J. C. Barton and C. Q. Edwards. Cambridge University Press, Cambridge, UK.

Feder, J. N., A. Gnirke, W. Thomas, Z. Tsuchihashi, D. A. Ruddy *et al.*, 1996 A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. Nat. Genet. **13:** 399–408.

Feder, J. N., D. M. Penny, A. Irrinki, V. K. Lee, J. A. Lebrón *et al.*, 1998 The hemochromatosis gene product complexes with the transferrin receptor and lowers its affinity for ligand binding. Proc. Natl. Acad. Sci. USA **95:** 1472–1477.

Guo, S. W., and M. Xiong, 1997 Estimating the age of mutant disease alleles based on linkage disequilibrium. Hum. Hered. **47:** 315–337.

Hudson, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

Hudson, R. R., K. Bailey, D. Skarecky, J. Kwiatowski and F. J. Ayala, 1994 Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster.* Genetics **136:** 1329–1340.

Jeffreys, A. J., L. Kauppi and R. Neumann, 2001 Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat. Genet. **29:** 217–222.

Kidd, K. K., 1987 Phenylketonuria: population genetics of a disease. Nature **327:** 282–283.

Kimura, M., and T. Ohta, 1973 The age of a neutral mutant persisting in a finite population. Genetics **75:** 199–212.

Kirby, D. A., and W. Stephan, 1995 Haplotype test reveals departure from neutrality in a segment of the white gene of *Drosophila melanogaster.* Genetics **141:** 1483–1490.

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.

Lucotte, G., 2001 Frequency analysis and allele map in favor of the Celtic origin of the C282Y mutation of hemochromatosis. Blood Cells Mol. Dis. **27:** 549–556.

Malfroy, L., M. P. Roth, M. Carrington, N. Borot, A. Volz *et al.*, 1997 Heterogeneity in rates of recombination in the 6-Mb region telomeric to the human major histocompatibility complex. Genomics **43:** 226–231.

Maruyama, T., 1974 The age of an allele in a finite population. Genet. Res. **23:** 137–143.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23–35.

McPeek, 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.

Merryweather-Clarke, A. T., J. J. Pointon, J. D. Shearman and K. J. H. Robson, 1997 Global prevalence of putative haemochromatosis mutations. J. Med. Genet. **34:** 275–278.

Motulsky, A. G., 1964 Hereditary red cell traits and malaria. Am. J. Trop. Med. Hyg. **13:** 147–158.

Nordborg, M., and S. Tavaré, 2002 Linkage disequilibrium: what history has to tell us. Trends Genet. **18:** 83–90.

Ohta, T., and M. Kimura, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22:** 201–204.

Pichon, L., G. Carn, P. Bouric, T. Giffon, B. Chauvel *et al.*, 1996 Structural analysis of the HLA-A/HLA-F subregion: precise localization of two new multigene families closely associated with the HLA class I sequences. Genomics **32:** 236–244.

Price, P., C. Witt, R. Allcock, D. Sayer, M. Garlepp *et al.*, 1999 The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunolopathological diseases. Immunol. Rev. **167:** 257–274.

Pritchard, J. K., and M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. **69:** 1–14.

Przeworski, M., R. R. Hudson and A. Di Rienzo, 2000 Adjusting the focus on human variation. Trends Genet. **16:** 296–302.

Rannala, B., and J. P. Reeve, 2001 High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. Am. J. Hum. Genet. **69:** 159–178.

Reich, D. E., and D. B. Goldstein, 1999 Estimating the age of mutations using variation at linked markers, pp. 129–138 in *Microsatellites: Evolution and Applications,* edited by D. B. Goldstein and C. Schlötterer. Oxford University Press, Oxford.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti *et al.*, 2001 Linkage disequilibrium in the human genome. Nature **411:** 199–204.

Risch, N., D. de Leon, L. Ozelius, P. Kramer, L. Almasy *et al.*, 1995 Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. Nat. Genet. **9:** 152–159.

Romeo, G., M. Devoto and L. J. V. Galietta, 1989 Why is the cystic fibrosis gene so frequent? Hum. Genet. **84:** 1–5.

Rotter, J. I., and J. M. Diamond, 1987 What maintains the frequencies of human genetic diseases? Nature **329:** 289–290.

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.

Salter-Cid, L., A. Brunmark, Y. Li, D. Leturcq, P. A. Peterson *et al.*, 1999 Transferrin receptor is negatively modulated by the hemochromatosis protein HFE: implications for cellular iron homeostasis. Proc. Natl. Acad. Sci. USA **96:** 5434–5439.

Serre, J. L., B. Simon-Bouy, E. Mornet, B. Jaume-Roig, A. Balassopoulou *et al.*, 1990 Studies of RFLP closely linked to the cystic fibrosis locus throughout Europe lead to new considerations in population genetics. Hum. Genet. **84:** 449–454.

Slatkin, M., 1994 Linkage disequiibrium in growing and stable populations. Genetics **137:** 331–336.

Slatkin, M., and G. Bertorelle, 2001 The use of intraallelic variability for testing neutrality and estimating population growth rate. Genetics **158:** 865–874.

Slatkin, M., and B. Rannala, 2000 Estimating allele age. Annu. Rev. Genomics Hum. Genet. **1:** 225–249.

Termijtelen, A., J. D'Amaro, J. J. van Rood and G. M. T. Schreuder, 1995 Linkage disequilibrium in HLA cannot be explained by selective recombination. Tissue Antigens **46:** 387–390.

Terwilliger, J. D., 1995 A powerful method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am. J. Hum. Genet. **56:** 777–787.

Thomas, W., A. Fullan, D. B. Loeb, E. E. McClelland, B. R. Bacon *et al.*, 1998 A haplotype and linkage disequilibrium analysis of the hereditary hemochromatosis gene region. Hum. Genet. **102:** 517–525.

Toomajian, C., and M. Kreitman, 2002 Sequence variation and haplotype structure at the human *HFE* locus. Genetics **161:** 1609–1623.

Valdes, A. M., M. Slatkin and N. B. Freimer, 1993 Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics **133:** 737–749.

Worwood, M., R. Raha Chowdhury, K. J. H. Robson, J. Pointon, J. D. Shearman *et al.*, 1997 The HLA A1-B8 haplotype extends 6 Mb beyond HLA-A: associations between HLA-A, B, F and 15 microsatellite markers. Tissue Antigens **50:** 521–526.

Communicating editor: W. Stephan