Searching for Evidence of Positive Selection in the Human Genome Using Patterns of Microsatellite Variability

Bret A. Payseur, Asher D. Cutter, and Michael W. Nachman

Department of Ecology and Evolutionary Biology, Biosciences West Building, University of Arizona, Tucson

Both natural selection and nonequilibrium population-level processes can lead to a skew in the frequency distribution of polymorphisms. Population-level processes are expected to affect all loci in a roughly equal fashion, whereas selection will affect only some regions of the genome. We conducted a sliding-window analysis of the frequency distribution of microsatellite polymorphisms across the human genome to identify regions that may be under positive selection. The analysis was based on a published data set of 5,257 mapped microsatellites in individuals of European ancestry. Observed and expected numbers of alleles were compared under a stepwise mutation model (SMM) using analytical formulae. Observed and expected heterozygosities were compared under a SMM using coalescent simulations. The two sets of analyses gave similar results. Approximately one-fourth of all loci showed a significant deficit of heterozygosity, consistent with a recent population expansion. Forty-three windows were identified with extreme skews in the frequency distribution of polymorphisms (in the direction of a deficit of heterozygosity, given the number of alleles). If these extreme windows are tracking selection at linked sites, theory predicts that they should be more common in regions of the genome with less recombination. We tested this prediction by comparing recombination rates in these extreme windows and in other regions of the genome and found that extreme windows had a significantly lower recombination rate than the genomic average. The proportion of extreme windows was significantly higher on the X chromosome than on the autosomes. Moreover, all the windows with extreme skews on the X chromosome were found in two clusters near the centromere; both these clusters exhibit markedly reduced recombination rates. These analyses point to regions of the genome that may recently have been subject to positive selection. These results also suggest that the effects of positive selection may be more pronounced on the X chromosome than on the autosomes in humans.

Introduction

A fundamental goal of population genetics is to understand the frequency and strength of positive selection. At the molecular level, the detection of positive selection relies on theoretical predictions of patterns of variation under neutrality (Kreitman and Akashi 1995). For example, assuming that mutations are neutral, the ratio of polymorphism to divergence should be the same for different genes (Hudson, Kreitman, and Aguade 1987) or for different categories of sites within the same gene (McDonald and Kreitman 1991). Genes (or categories of sites) that harbor more or fewer polymorphisms than expected on the basis of their divergence levels represent candidates for positive selection. Some tests of neutral, equilibrium models focus on the frequency spectrum of polymorphisms (Watterson 1978; Tajima 1989; Fu and Li 1993; Fu 1997). For instance, Ewens (1972) showed that under the infinite alleles model, the allelic configuration of a population at steady state is specified by the sample size and the observed number of alleles. This result motivates a test of neutrality in which homozygosity is predicted using the observed number of alleles (Watterson 1978). This expected value of homozygosity is then compared with the observed value, and discrepancies indicate the potential action of positive selection. Because this test is based on an equilibrium model, deviations may also be caused by any departure from equilibrium, such as a population expansion or contraction.

In addition to leaving a signature on the target of selection, positive selection should affect patterns of linked neutral variation to an extent determined by the local recombination rate and the strength of selection (Maynard Smith and Haigh 1974). Two modes of positive selection, directional and balancing, are expected to impact molecular diversity at linked regions in different ways. Positive directional selection can lead to an overall reduction of linked variation and an excess of alleles, given the observed heterozygosity (Maynard Smith and Haigh 1974; Watterson 1978; Kaplan, Hudson, and Langley 1989). In contrast, balancing selection can lead to an elevation of linked polymorphism and a deficit of alleles, given the observed heterozygosity (Watterson 1978; Hudson, Kreitman, and Aguade 1987).

The idea that selection can lead to a skew in the frequency distribution at linked loci suggests a way to identify individual genomic regions experiencing positive selection. With a set of loci throughout the genome mapped at sufficiently high density, clusters of loci exhibiting similar skews in the frequency distribution of alleles may pinpoint genomic regions that have recently experienced positive selection. In principle, such an approach could be used to identify regions with either an excess of rare alleles (indicative of directional selection) or an excess of intermediate frequency alleles (indicative of balancing selection). Because most human populations are not at equilibrium and have undergone recent expansions (e.g., Rogers and Harpending 1992), localized skews in the frequency distribution for particular genomic regions must be interpreted against the background of the genomic average skew in the frequency distribution created by population-level processes.

Key words: positive selection, directional selection, balancing selection, genetic hitchhiking, microsatellites, human genome.

Address for correspondence and reprints: Bret A. Payseur, Department of Ecology and Evolutionary Biology, Biosciences West Building, University of Arizona, Tucson, Arizona 85721.
E-mail: payseur@email.arizona.edu.

Mol. Biol. Evol. 19(7):1143–1153. 2002

In addition to localized skews in the frequency distribution, positive selection may cause different patterns of variability at X-linked and autosomal loci (Begun and Whitley 2000). If adaptive mutations are recessive on an average, fixation rates are expected to be higher for the X chromosome relative to the autosomes because recessive X-linked mutations will be visible to selection at every generation in males (Charlesworth, Coyne, and Barton 1987). In addition to having faster fixation rates, X-linked beneficial mutations may experience shorter transit times than their autosomal counterparts. Because X-linked mutations spend one-third of their time in a haploid state, their fitness variance is larger than that of autosomal mutations, making selection more effective and the increase in frequency more rapid for X-linked beneficial mutations (Avery 1984). Thus, there are several reasons why beneficial mutations arising on the X chromosome may be affected more dramatically by selection than mutations arising on the autosomes (discussed in Begun and Whitley 2000). However, if selection acts on standing variation that is in mutation-selection balance, fixation rates for X-linked mutations are not expected to be faster (Orr and Betancourt 2001). Assuming similar recombinational environments for the X chromosome and the autosomes, the frequency distributions at neutral loci can be used to test the hypothesis that X-linked beneficial mutations experience faster fixation rates or shorter transit times (or both) than autosomal mutations. This hypothesis predicts that X-linked loci may show a greater excess of rare alleles on average.

In this article, we use published polymorphism data from 5,257 genetically mapped microsatellites to search the human genome for evidence of positive selection in two ways. First, we use a sliding-window approach to look for genomic regions showing unusually high numbers of loci deviating from equilibrium. Second, we compare the frequency distributions of alleles at X-linked and autosomal microsatellites.

Materials and Methods

The Frequency Distribution of Polymorphisms

The sample consisted of 56 autosomes and 108 X chromosomes from unrelated individuals of European descent, genotyped at 5,257 microsatellite loci (Dib et al. 1996). All loci were dinucleotide $(CA)_n$ repeats. Allele size and frequency, heterozygosity, number of alleles, and genetic position were obtained for each locus from published data (Dib et al. 1996; ftp://ftp.genethon.fr/pub/Gmap/Nature_1995/data/). Variance in allele size was estimated as

$$\left(\frac{n}{n-1}\right)\sum_{i=1}^{n} f_i(x_i - \bar{x})^2$$

(Sokal and Rohlf 1995, p. 52), where f_i is the frequency of the ith allele, x_i the number of repeats at the ith allele, $\bar{x}$ the average number of repeats weighted by frequency, and n the number of alleles.

We calculated the expected allele numbers and heterozygosities by assuming a stepwise mutation model (SMM; Ohta and Kimura 1973). The SMM appears to be appropriate for many human dinucleotide repeat microsatellites (Valdes, Slatkin, and Friemer 1993; Weber and Wong 1993). In the data used in this study (Dib et al. 1996), about 90% of the observed mutations produced repeat changes of size one or two.

We investigated departures from the frequency distribution expected using the SMM in two ways. First, we used the formula of Ohta and Kimura (1973) to estimate the expected number of alleles, given the observed levels of heterozygosity, for each locus separately. Second, we used the coalescent method (and BOTTLENECK v. 1.2.02 software) of Cornuet and Luikart (1996). This approach simulates microsatellite evolution under a specified mutational model (in this case, 1,000 replicates per locus using the SMM), assuming a constant population size at mutation-drift equilibrium, and calculates the expected heterozygosity at a locus, given the observed number of alleles. By completing many replicate simulations, the difference between observed and expected heterozygosity can be assigned a probability. This approach is formally equivalent to Watterson's (1978) homozygosity test (in reverse). The critical value for the probabilities was set at 0.05.

Clustering of Nonequilibrium Loci

Positive selection may induce a nonrandom spatial distribution of loci with unusual frequency distributions. We tested this prediction in two ways. First, we conducted simple spatial autocorrelation analyses, asking whether the statistics of neighboring loci (i.e., lag of one locus) are more similar than those expected by chance. Autocorrelations were estimated separately for each chromosome and in analyses including all chromosomes. Autocorrelation tests were performed using the observed minus expected heterozygosity, as well as the number of alleles, heterozygosity, and variance in allele size.

Second, we identified genomic regions exhibiting clustering of significant loci. We defined a sliding window, 5 cM in size, beginning with each microsatellite. Huttley et al. (1999) demonstrated that linkage disequilibrium rarely extends beyond 5 cM for this data set; hence, this genetic distance provides a reasonable upper bound for selection effects mediated by linkage. Next, we counted the number of loci in each window showing significant differences between observed and expected heterozygosity using coalescent simulations (Cornuet and Luikart 1996). We then used the genome-wide average proportion of significant loci (or the chromosomal average) to predict the proportion of significant markers that would be found in each window if significant loci were randomly distributed throughout the genome (or along an individual chromosome). The probability of the observed (or more extreme) proportion was estimated using the binomial distribution as

Table 1
Descriptive Statistics for Measures of Variation and Skews in the Frequency Distribution for 5,257 Microsatellites

	Mean	Standard Deviation	Standard Error of the Mean	Minimum	Maximum
Variance in allele size	13.85	20.55	0.28	0.19	474.29
Heterozygosity	69.83	12.04	0.17	16.00	94.00
Number of alleles	7.58	2.78	0.04	2.00	27.00
Observed minus expected number of alleles	0.28	1.76	0.02	−4.51	11.39
Observed minus expected heterozygosity	−7.30	8.76	0.12	−54.10	18.80

$$P = \sum_{i=k}^{m} \binom{m}{i} y^i (1 - y)^{m-i}$$

where y is the average proportion of significant markers per 5 cM window, m the total number of markers in the window, and k the number of significant markers in the window. These calculations were performed separately for heterozygosity deficits (genome-wide $y = 0.275$) and heterozygosity excesses (genome-wide $y = 0.009$).

The computation of more than 5,000 binomial tests suggests that the statistical critical value should be reduced. We employed two independent corrections for multiple tests. First, we applied a Bonferroni correction (Sokal and Rohlf 1995, p. 240). This suggested a critical value of $9.5 \times 10^{-6} = (0.05/5{,}257)$. We also used the randomization procedure of Churchill and Doerge (1994), which was designed to deal with a similar multiple test problem: the identification of regions of the genome related to quantitative trait variation (QTL mapping). Holding the number of loci per window and their genetic positions constant, the significance state (significant vs. not significant) was shuffled across loci throughout the genome. Binomial P-values for all resulting windows were calculated, and the minimum P-value in the genome was recorded. We repeated this procedure 10,000 times to generate a distribution of minimum P-values under the null hypothesis that significant loci are randomly distributed throughout the genome. Binomial probabilities observed for actual windows were then compared with this distribution. We conducted a comparable test using the chromosome as the level of the experiment (23 times). Here, the null distribution was formulated on the basis of 1,000 replicates, and a separate mean proportion of significant loci per window was used for each chromosome (rather than the genome-wide mean).

This binomial test assumes that adjacent loci are independent under a neutral model. Because closely linked loci will have correlated histories, this assumption is not strictly correct. Therefore, this test may overestimate the significance of some genomic regions. However, autocorrelation analyses of the skew in the frequency distribution and of the measures of variation suggest little evidence for shared histories among neighboring loci overall (see *Results*).

The Frequency Distribution of Polymorphisms and Recombination Rate

Theory predicts that positive selection will affect nearby neutral polymorphism most strongly in regions of reduced recombination (Maynard Smith and Haigh 1974). Therefore, we also assessed the relationship between measures of skew in the frequency distribution and local recombination rate. Recombination rates were estimated for a subset of the microsatellites considered in this article by comparing integrated genetic and physical maps of the human genome (Payseur and Nachman 2000). We used Kendall's correlation tests to evaluate the association between measures of skew in the frequency distribution and recombination rate.

Comparisons Between the X Chromosome and the Autosomes

We compared the frequency distributions of the X-linked and autosomal microsatellites using Mann-Whitney U-tests. Results from both Ohta and Kimura's (1973) formulae and from Cornuet and Luikart's (1996) method were analyzed in these tests.

Results

The Frequency Distribution of Polymorphisms

Statistics describing overall microsatellite variation and the skew in the frequency distribution are summarized in table 1. As previously reported (Dib et al. 1996), the average number of alleles is 7.58 and the average heterozygosity is 69.83%. The mean variance in allele size (repeat number) is 13.85. Measures of skew in the frequency distribution suggest that, conditioned on the observed heterozygosity, across the genome there are more alleles observed than expected (table 1 and fig. 1). Similarly, coalescent simulations indicate that, conditioned on the observed allele number and sample size, more than one-fourth (1,447 of 5,257) of all loci exhibit a significant deficit of heterozygosity ($P < 0.05$), whereas only 48 loci show a significant excess of heterozygosity. Because this pattern is seen at so many loci, it is probably best explained by a recent population expansion (Kimmel et al. 1998; Torroni et al. 1998).

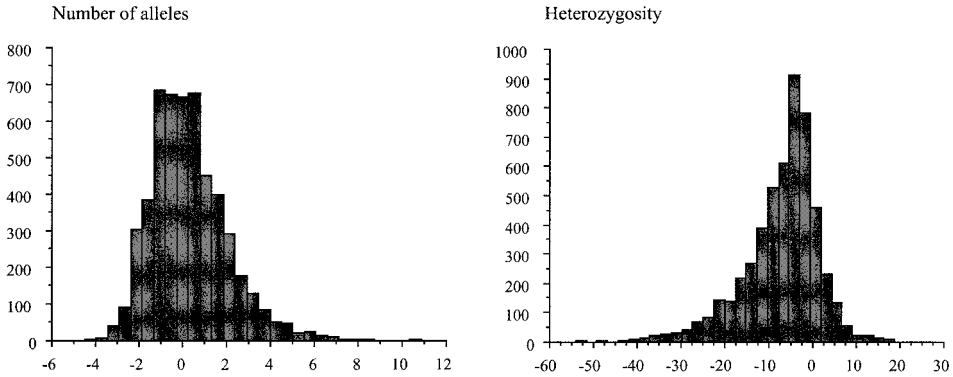

FIG. 1.—Histograms of observed minus expected alleles and observed minus expected heterozygosity. The *y*-axis is counts. A few observations at the far negative end of the range are not shown for the difference between observed and expected heterozygosity for ease of visual presentation.

Table 2
Genomic Regions with Extreme Skews in the Frequency Distribution of Microsatellite Alleles

Chromosome	Genethon Map Position (cM)[b]	Probability[c]	Deficit-Excess of Heterozygosity	Recombination Rate (cM/Mb)
1....	245.7	0.010	Deficit	0.60
1....	247.6	0.008	Deficit	0.92
2....	217.2	0.006	Deficit	1.49
4....	169.1	0.007	Deficit	2.94
5....	132.8	0.002	Deficit	0.93
5....	134.4	0.002	Deficit	0.85
5....	134.9	0.002	Deficit	0.87
7....	68.4	0.010	Deficit	0.85
7....	154.0	0.007	Deficit	1.34
10....	62.0	0.007	Deficit	0.61
10....	157.2	0.006	Deficit	3.28
17....	100.4	0.002	Deficit	1.13
20....	50.7	0.002	Deficit	0.14
20....	51.7	0.002	Deficit	0.62
20....	53.6	0.010	Deficit	0.68
X	82.6	0.001	Deficit	0.86
X	84.9	0.001	Deficit	0.77
X	85.0	0.001	Deficit	0.77
X	85.1	0.001	Deficit	0.77
X	85.2	0.001	Deficit	0.77
X	85.3	0.002	Deficit	0.77
X	86.5	0.005	Deficit	0.64
X	86.6	0.005	Deficit	0.64
X	86.7	0.002	Deficit	0.64
X	86.8	0.001	Deficit	0.64
X	86.9	0.002	Deficit	0.64
X	87.4	0.007	Deficit	0.48
X	95.1	0.008	Deficit	0.67
X	95.9	0.006	Deficit	0.67
X	96.1	0.008	Deficit	0.67
6....	157.5	0.004	Excess	2.35
10....	44.8	0.007	Excess	1.14
10....	46.2	0.008	Excess	0.84
10....	46.7	0.004	Excess	0.89

[a] A total of 43 windows with binomial probabilities less than 0.01 were identified. Because some markers map to the same position, there are several extreme windows that contain slightly different sets of loci and still map to the same position. These largely overlapping windows are not shown; only the 34 windows that map to unique genetic positions are shown.

[b] Map position of the first locus in the 5 cM window.

[c] Binomial probability of the observed ratio of significant loci to total loci, given the genome-wide average ratio.

Autocorrelations

The average distance between loci in this study is approximately 0.5 cM. In genome-wide analyses, microsatellite frequency distributions show very weak evidence of autocorrelation at this scale (observed minus expected heterozygosity: $\tau = 0.02$; $P = 0.02$). The magnitude of this autocorrelation is very small, and its detection is likely a consequence of measuring associations using more than 5,000 loci. Consistent with this interpretation, only loci on chromosome 17 exhibit a significant degree of autocorrelation ($\tau = 0.13$; $P = 0.009$) in analyses of individual chromosomes. The observation that frequency spectra are not correlated with recombination rate overall (see subsequently) suggests that these weak autocorrelations are not driven by recurrent selective sweeps occurring in similar recombinational environments. Measures of microsatellite variation, including allele number, heterozygosity, and variance in allele size, are not autocorrelated ($P > 0.05$ in all tests), a result which also provides little evidence for shared histories, in general, among neighboring loci (although some linkage disequilibrium has been observed between microsatellites at this scale in humans; reviewed in Pritchard and Przeworski 2001).

Sliding-Window Analysis

The sliding-window binomial probabilities for a deficit of heterozygosity are plotted against genetic map position for all the 23 chromosomes in figure 2. Several patterns emerge from this analysis. First, there is considerable variance in binomial *P*-values within individual chromosomes. Chromosomal regions with multiple adjacent windows containing low *P*-values may represent portions of the genome recently subject to positive directional selection. Second, the X chromosome, in general, harbors a greater proportion of windows with low *P*-values than do the autosomes. For example, 7.7% of all windows on the X chromosome are significant at the 0.01 level compared with 0.5% of the windows on the autosomes. Finally, when either a Bonferroni correction or the randomization method of Churchill and Doerge (1994) using genome-wide means is used to account for 5,257 tests, none of the individual sliding-

Table 3
Numbers of Highly-Skewed ($P < 0.01$) and Not–Highly-Skewed ($P > 0.01$) Windows Exhibiting a Deficit of Heterozygosity on the X Chromosome and on the Autosomes

	X Chromosome	Autosomes
Sliding windows with $P < 0.01$	16	27
Sliding windows with $P > 0.01$	193	5,021

NOTE.—Fisher's Exact Test, $P < 10^{-10}$.

Table 4
Numbers of Highly-Skewed ($P < 0.01$) and Not–Highly-Skewed ($P > 0.01$) Windows Exhibiting a Deficit of Heterozygosity on the X Chromosome in Regions of Low Recombination (<1.0 cM/Mb) and in Regions of Average to High Recombination (>1.0 cM/Mb)

	Low Recombination	High Recombination
Sliding windows with $P < 0.01$.	16	0
Sliding windows with $P > 0.01$.	27	166

NOTE.—Fisher's Exact Test, $P < 10^{-10}$.

window probabilities remain significant for either an excess or deficit of heterozygosity. When the randomization method of Churchill and Doerge (1994) is used with chromosome means, one window contains a significant deficit of heterozygosity, located at 100.4 cM on chromosome 17 (although the observation that chromosome 17 exhibits a significant degree of autocorrelation suggests caution in the interpretation that this window contains a target of selection). These corrections for multiple tests may be conservative and may overlook regions of biological importance, as discussed below. They also do not take into account adjacent clusters of windows with low *P*-values.

Individual sliding windows showing an excess or deficit of heterozygosity with probabilities below 0.01 are listed in table 2. These are the regions of the genome showing the most extreme skew in the frequency distribution. Each of these windows contains 13 markers, on an average, and 9–13 of the markers within each window show a significant skew in the frequency distribution. There is a highly significant nonrandom association between the proportions of these extreme windows and their location (X chromosome vs. autosomes; table 3, Fisher's Exact Test, $P < 10^{-10}$). Because some of these windows are overlapping, they do not constitute independent observations. We therefore analyzed the association in table 3 using: (1) only nonoverlapping windows, and (2) individual loci rather than windows of multiple loci. Both analyses give the same result as depicted in table 3 (Fisher's Exact Test, $P < 0.01$ for both).

Interestingly, the strong skew in the frequency distribution on the X chromosome is found in two distinct clusters near the centromere. The first cluster (spanning 82.6–92.4 cM) contains 12 windows, and the second cluster (spanning 95.1–101.1 cM) contains three windows. These clusters correspond to the regions of the X chromosome with the lowest recombination rates (<1.0 cM/Mb; Payseur and Nachman 2000).

The hypothesis that genomic windows with the most severe skews in the frequency distribution represent candidate regions for positive directional selection also predicts that microsatellites in these windows will show reduced variation. The mean heterozygosity is reduced in these extreme windows relative to the genome-wide average, and this effect is significant (Mann-Whitney *U;* $P < 0.0003$). However, extreme windows do not show reduced variance in allele size ($P > 0.05$) and do exhibit a higher mean number of alleles than other windows ($P < 0.02$). This discrepancy between different measures of variation may suggest that microsatellite heterozygosity is a more sensitive indicator of recent selective sweeps than allele number or variance in allele size. Although we cannot offer a simple demographic scenario that could produce this discordance, these results may also indicate that some of the identified windows are not tracking selective events.

Recombination Rates and the Frequency Spectrum

Previous work revealed no correlation between recombination rate and levels of microsatellite polymorphism in these data (Payseur and Nachman 2000). We compared variation in the recombination rate with measures of the skew in the frequency distribution. Overall, recombination rate is not correlated with the difference between observed and expected heterozygosity ($P > 0.05$) or with the difference between observed and expected number of alleles ($P > 0.05$). However, there is evidence that windows with extreme skews toward a deficit of heterozygosity exhibit reduced recombination rates relative to the genome-wide average (table 4; extreme windows: average = 0.8 cM/Mb; genome-wide: average = 1.5 cM/Mb; Mann Whitney *U*, $P < 0.05$). This effect is most pronounced for windows on the X chromosome, where all the windows with extreme skews are found in regions with the lowest rates of recombination. This nonrandom association between skews in the frequency distribution and recombination rate on the X chromosome is highly significant (table 4, Fisher's Exact Test, $P < 10^{-10}$). An analysis of this pattern using only nonoverlapping windows or using individual loci gives the same result (Fisher's Exact Test, $P < 0.05$ for both).

Comparisons Between the X Chromosome and the Autosomes

We compared data for 209 X-linked microsatellites with those for 5,048 loci mapped to autosomes. The average difference between the observed and expected heterozygosity is −10.30 for X-linked loci and −7.17 for autosomal loci. This difference between the X chromosome and the autosomes is statistically significant (Mann-Whitney *U*-test, $P < 10^{-4}$). Using Ohta and Kimura's formula, the average difference between the observed and expected number of alleles is 1.00 for X-linked loci and 0.25 for autosomal loci. This difference is also significant (Mann-Whitney *U*-test, $P < 10^{-4}$).

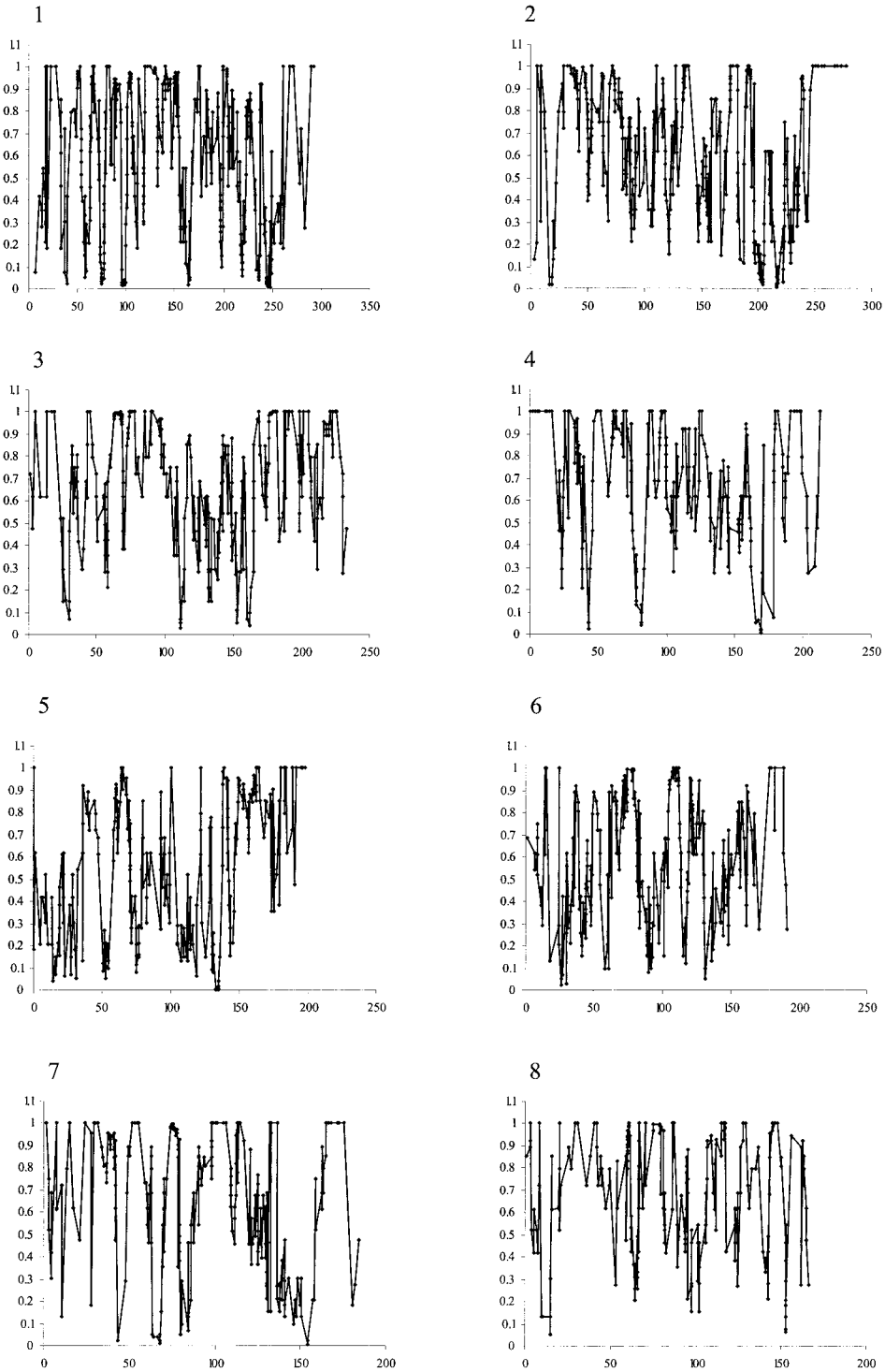

FIG. 2.—Sliding-window plots of binomial *P*-values associated with a deficit of heterozygosity (*y*-axis) versus genetic position (cM; *x*-axis) for each chromosome. The binomial probabilities are calculated from the genome-wide proportion of significant loci (see *Materials and Methods*).

Furthermore, there is weak evidence that X-linked loci are less variable than autosomal loci. Using the SMM, the observed heterozygosity (H) and variance in allele size (V) can be used to estimate the neutral parameter θ_A for the autosomes as

$$\theta_A = \frac{1}{2}\left[\frac{1}{(1 - H)^2} - 1\right] \quad \text{and} \quad \theta_A = 2V + 4\mu$$

where $\theta_A = 4N_e\mu$, N_e is the effective population size, and μ is the neutral mutation rate (Ohta and Kimura 1973;

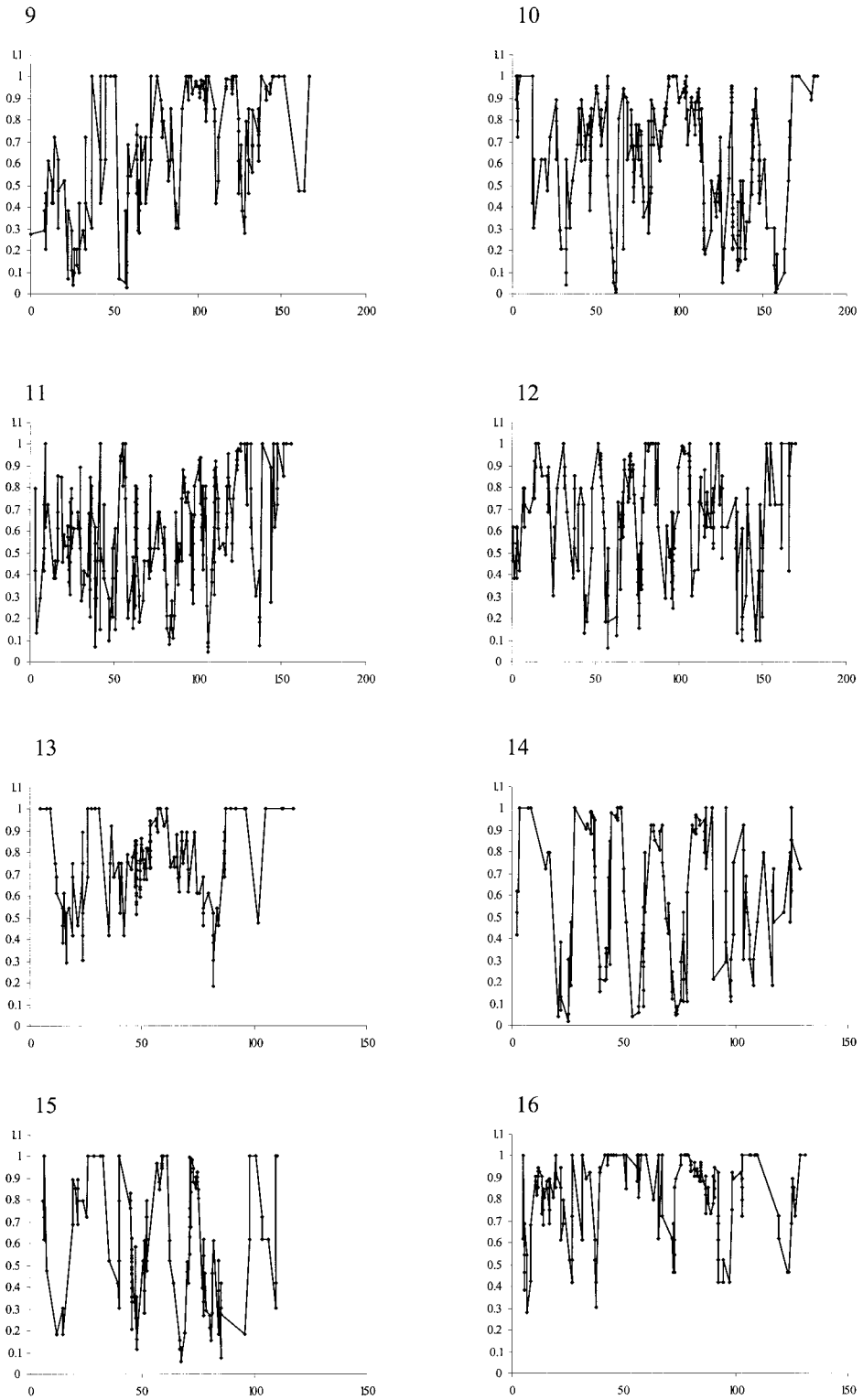

FIG. 2 *(Continued)*

Moran 1975). Under neutral, equilibrium conditions, assuming equal mutation rates on the X chromosome and the autosomes, a sex ratio of one, and equal variances in mating success between males and females, the X chromosome has three-fourths of the effective population size of the autosomes; therefore, the expected value of θ_X/θ_A is 0.75. Using observed heterozygosities to estimate this ratio leads to a value of 0.71. Using observed variances in allele size, the estimated ratio is 0.68. Although we cannot reject the hypothesis that these ratios are equivalent to 0.75, the observed variances in allele size suggest a weak reduction in the level of variation observed on

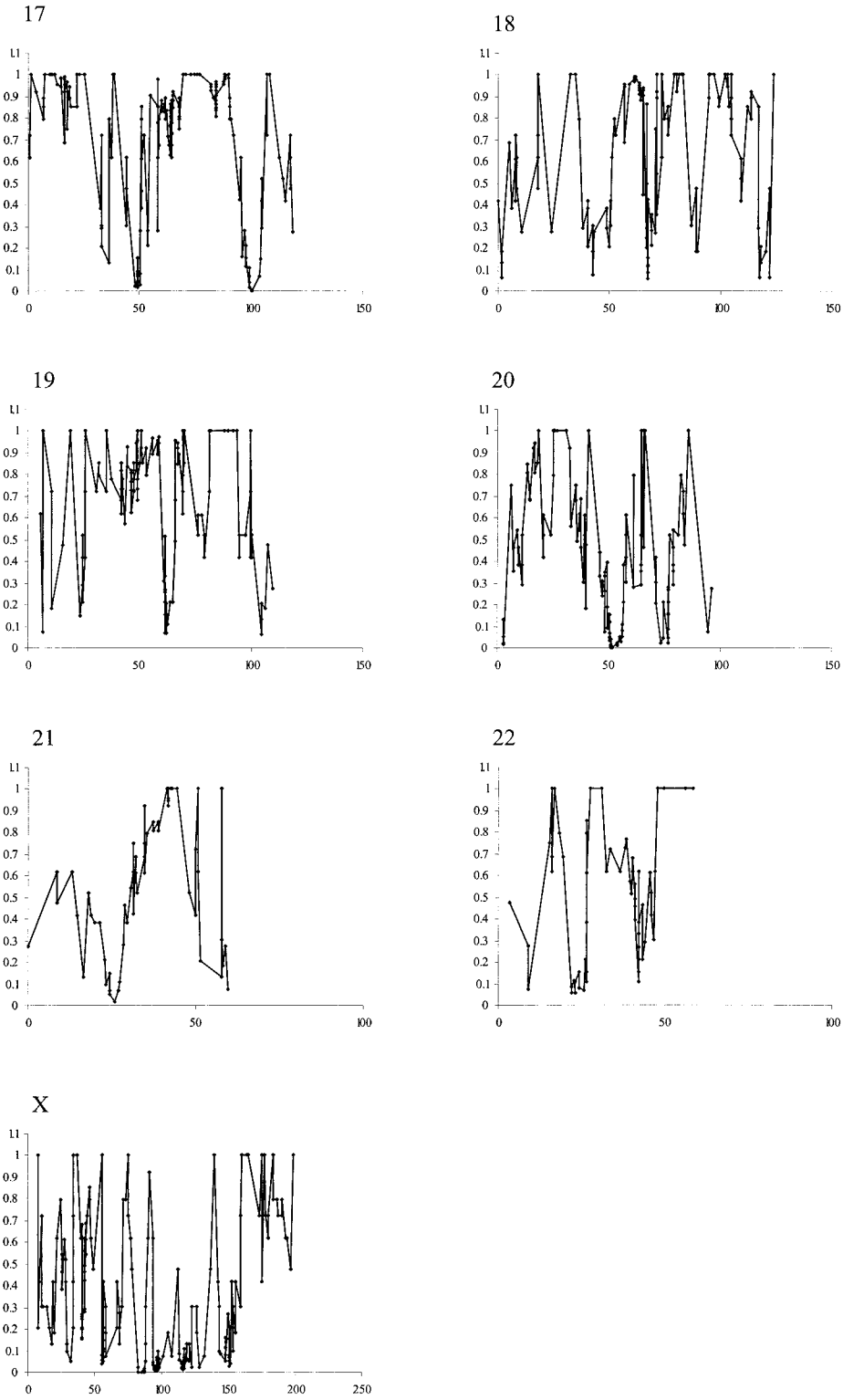

FIG. 2 *(Continued)*

the X chromosome relative to the autosomes, after accounting for differences in the effective population size. If the variance in male mating success exceeds that of females, the expected value of θ_X/θ_A would be greater than 0.75, suggesting that using a correction factor of $\frac{4}{3}$ may be conservative relative to our interpretation.

Discussion

We searched the human genome for evidence of positive selection by looking for departures from the neutral, equilibrium frequency spectrum, and several clear patterns emerged. First, there appears to be an

overall signature of population expansion on patterns of microsatellite polymorphism, as previously noted (e.g., Kimmel et al. 1998). Second, a sliding-window analysis identified 43 windows with extreme skews in the frequency distribution. Although the deviations in these windows are not individually significant after correcting for 5,257 tests, these windows represent candidate regions for positive selection. Third, the distribution of these extreme windows is highly nonrandom with respect to recombination rate. Extreme windows occur predominantly in regions of the genome with less recombination. Fourth, the frequency distribution of microsatellite polymorphisms is significantly different in comparisons between the X chromosome and the autosomes. We discuss each of these issues in turn.

Population Structure and the Frequency Distribution of Polymorphisms

The average microsatellite in this population is about 7% less heterozygous than expected when heterozygosity is predicted with allele-based coalescent simulations assuming the SMM. Over 25% (1,447/5,257) of the tests on individual loci show significant deficits of heterozygosity. As expected, most of these loci also show an excess of rare alleles using Ohta and Kimura's (1973) formula. It seems unlikely that most of these loci are tracking individual events of directional selection or that the excess of rare alleles indicates widespread maintenance of slightly deleterious alleles. Such a genome-wide pattern is more parsimoniously interpreted as evidence of a recent population expansion in Europe. An alternative explanation is that the SMM is inappropriate for many human microsatellites and that it (systematically) overpredicts the expected heterozygosity. However, given the genetic evidence (Kimmel et al. 1998; Torroni et al. 1998) for the recent European population size increase and the support for the SMM in humans (Valdes, Slatkin, and Friemer 1993; Weber and Wong 1993), the demographic explanation seems most reasonable.

Sliding-Window Analysis

The sliding-window analysis is predicated on the notion that the spatial distribution of markers can provide information that is useful in identifying regions of the genome under selection. There are a number of biological and statistical issues that must be confronted in interpreting the results of these analyses, including the background signature of a population expansion, the likelihood that selection will impact patterns of microsatellite variability, and the statistical problems of assuming independence among linked loci and of conducting multiple tests.

First, it is challenging to identify locus-specific skews in the frequency distribution against the background of a genome-wide skew in the same direction. Thus, it is likely to be more difficult to detect positive directional selection in an expanding population than in one of a constant size.

Second, although many models of selection deal with simple, strong effects (e.g., Maynard Smith and Haigh 1974; Simonsen, Churchill, and Aquadro 1995), weak or fluctuating selection may be important in shaping patterns of variation at the molecular level (Gillespie 1991, p. 142; Wayne and Simonsen 1998; Przeworski, Hudson, and di Rienzo 2000). A simple, catastrophic selective sweep in which a newly arising adaptive mutation is quickly driven to fixation is expected to leave a clear signature on the frequency spectrum of segregating polymorphisms, and this has been best studied with nucleotide polymorphisms (Braverman et al. 1995; Simonsen, Churchill, and Aquadro 1995). In these studies, the power to detect a sweep from the frequency spectrum is high only during a fairly brief interval after the sweep. Outside of this interval, or if selection is weak, the power to detect a sweep is quite low. Although similar power analyses have not been carried out with microsatellite data, it is likely that the power to detect sweeps will be similarly low outside of a narrow time interval. An added concern is that the high mutation rate of microsatellites will quickly obscure past selective events. In a simple hitchhiking model with $s = 0.01$ and $N = 10^4$, Wiehe (1998) and Schlötterer and Wiehe (1999) have shown that a reduction in microsatellite variability is expected only if microsatellites are tightly linked (<0.15 cM) and if microsatellite mutation rates are low (<0.01). Thus, we might expect, a priori, that it will be difficult to detect a skew in the frequency distribution, except in regions of low recombination where any given microsatellite will presumably be linked to more genes. For example, the recombinational distance of less than approximately 0.15 cM required to detect hitchhiking may correspond to a physical distance of less than 25 kb in high-recombination regions of the genome and a distance of 1 Mb or more in low-recombination regions. The average gene density across the human genome is approximately 10 genes per Mb (International Human Genome Sequencing Consortium 2001). It is noteworthy that we observe windows with extreme skews in the frequency distribution mostly in regions of the genome with little recombination (table 2). This suggests that these windows are tracking selective events. The power to detect selection may be even lower under more complex models of selection. For example, some models of selection in fluctuating environments can result in little skew in the frequency distribution of polymorphisms (Gillespie 1994).

Statistical issues may also complicate our attempts to identify genomic regions under selection. First, the binomial test we applied assumes statistical independence among loci. Even under a neutral model, however, closely linked loci may share histories, leading to correlations among their frequency spectra. As a result, our analyses may overestimate the statistical significance of some genomic windows. However, the observation that measures of variation are not autocorrelated and that frequency spectra are only weakly autocorrelated suggests that this problem is unlikely to be severe.

Second, given the corrections for multiple tests, a sliding window will be significant only if its binomial

probability is less than about 10^{-5}. Because of the genome-wide skew toward a deficit of heterozygosity, this occurs only if approximately 90% of the loci within the window individually show significant skews in the frequency distribution. One way to improve the power of the test would be to increase the window size. Although this approach would be statistically defensible, it is less biologically reasonable because selection is unlikely to exert effects over long distances. Therefore, the identification of genomic regions affected by selection when large numbers of loci are analyzed remains a difficult problem.

The distribution of extreme windows is highly nonrandom, both with respect to X-autosome differences and with respect to recombination rate. As noted previously, these observations suggest that many or most of these windows are tracking selective events. Further indication that some of the windows in table 2 are tracking selection comes from studies of human nucleotide polymorphism in these same regions. For example, the two regions of low recombination near the X chromosome centromere that are identified in table 2 also contain genes showing reduced nucleotide heterozygosity and a skew in the frequency spectrum with negative values of Tajima's (1989) *D* statistic (Nachman 2001). Furthermore, one of these regions contains a microsatellite (at 99.7 cM) that shows an unusually high level of divergence in allele frequencies between European Americans and African Americans (Smith et al. 2001). These observations are consistent with recent positive directional selection.

Recombination Rates and the Frequency Spectrum

Theory predicts that positive selection will impact linked neutral variation most severely in regions of reduced recombination (Maynard Smith and Haigh 1974). Although we do not observe an overall correlation between recombination rate and measures of the frequency distribution, windows with extreme skews are found predominantly in low recombination regions. The mean recombination rate for windows with extreme skews is significantly below the genomic average. This effect is particularly pronounced on the X chromosome (table 4).

Comparisons Between the X Chromosome and the Autosomes

There is a greater skew in the frequency distribution for X-linked loci than for autosomal loci. The existence of this difference is not sensitive to the mechanism of mutation: a similar result is obtained when an infinite alleles model is assumed (results not shown). Assuming that the X chromosome and the autosomes experience similar average recombination rates, this result has both demographic and selective explanations. First, differences in effective population size may cause X-linked and autosomal loci to be differentially affected by demographic events. For example, Fay and Wu (1999) showed that a similar discordance in the frequency spectrum between mitochondrial and autosomal loci in humans could be caused by a population bottleneck. The effect is expected to be less severe for X-autosome comparisons because the X chromosome has three-fourths of the effective population size of autosomes, whereas the comparable value for the mitochondrial genome is one-fourth. Differences between the sexes in demographic factors, such as migration rate and age structure, may also leave different signatures on X-chromosomal and autosomal frequency spectra.

Alternatively, this discrepancy may indicate a relatively higher fixation rate or shorter transit time for beneficial mutations on the X chromosome (Avery 1984; Charlesworth, Coyne, and Barton 1987; Begun and Whitley 2000). Weak support for the hypothesis that fixation rates are higher or transit times are lower for adaptive mutations on the X chromosome comes from the observation that the overall levels of variability are slightly lower on the X chromosome than on the autosomes. From the observed heterozygosity and the observed variance in allele size, we estimated the ratio θ_X/θ_A as 0.71 and 0.68, respectively, both slightly below the expected value of 0.75. Similarly, the density of single nucleotide polymorphisms is lower on the X chromosome than on the autosomes (International SNP Map Working Group 2001). Reduced levels of variation on the X chromosome relative to the autosomes have also been noted in several other species, including mice (Hedrick and Parker 1997) and *Drosophila simulans* (Begun and Whitley 2000).

At present, we cannot clearly distinguish between the competing demographic and selective explanations for the discrepancy in frequency spectra between the X chromosome and the autosomes. Additional theoretical work could help resolve which model(s) best explain(s) this pattern.

Predictions

The availability of the complete genome sequence of humans may make it possible to test some predictions from this study. For example, we might expect to find some genes in the windows in table 2 that show high ratios of nonsynonymous to synonymous substitutions in interspecific comparisons. There is an inherent difficulty, common to mapping studies, in identifying the underlying genes: it is easiest to detect a signal in regions of reduced recombination specifically because the number of genes contained in these regions is expected to be large, but the large number of genes makes it more difficult to pinpoint individual genes of interest. For this reason, it may be most promising to look first at the few windows in table 2 that show high rates of recombination. The markers in these windows may lie close to the genes under selection.

Acknowledgments

We thank Bruce Walsh for useful discussions and statistical advice. We also appreciate the feedback of members of the Nachman lab during the course of the project. Wolfgang Stephan and two anonymous reviewers provided helpful comments on this article. This work was funded by the National Science Foundation.

LITERATURE CITED

AVERY, P. J. 1984. The population-genetics of haplo-diploids and X-linked genes. Genet. Res. **44**:321–341.

BEGUN, D. J., and P. WHITLEY. 2000. Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. Proc. Natl. Acad. Sci. **97**:5960–5965.

BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY, and W. STEPHAN. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphims. Genetics **140**:783–796.

CHARLESWORTH, B., J. A. COYNE, and N. H. BARTON. 1987. The relative rates of evolution of sex chromosomes and autosomes. Am. Nat. **130**:113–146.

CHURCHILL, G. A., and R. W. DOERGE. 1994. Empirical threshold values for quantitative trait mapping. Genetics **138**:963–971.

CORNUET, J.-M., and G. LUIKART. 1996. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics **144**:2001–2014.

DIB, C., S. FAURE, C. FIZAMES et al. (14 co-authors). 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature **380**:152–154.

EWENS, W. J. 1972. The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3**:87–112.

FAY, J. C., and C.-I. WU. 1999. A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. Mol. Biol. Evol. **16**:1003–1005.

FU, Y.-X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147**:915–925.

FU, Y.-X., and W.-H. LI. 1993. Statistical tests of neutrality of mutations. Genetics **133**:693–709.

GILLESPIE, J. H. 1991. The causes of molecular evolution. Oxford University Press, Oxford.

———. 1994. Alternatives to the neutral theory. Pp. 1–17 *in* B. GOLDING, ed. Non-neutral evolution: theories and molecular data. Chapman and Hall, New York.

HEDRICK, P. W., and J. D. PARKER. 1997. Evolutionary genetics and genetic variation of haplodiploids and X-linked genes. Ann. Rev. Ecol. Syst. **28**:55–83.

HUDSON, R. R., M. KREITMAN, and M. AGUADE. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics **116**:153–159.

HUTTLEY, G. A., M. W. SMITH, M. CARRINGTON, and S. J. O'BRIEN. 1999. A scan for linkage disequilibrium across the human genome. Genetics **152**:1711–1722.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

INTERNATIONAL SNP MAP WORKING GROUP. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409**:928–933.

KAPLAN, N. L., R. R. HUDSON, and C. H. LANGLEY. 1989. The ''hitchhiking effect'' revisited. Genetics **123**:887–899.

KIMMEL, M., R. CHAKRABORTY, J. P. KING, M. BAMSHAD, W. S. WATKINS, and L. B. JORDE. 1998. Signatures of population expansion in microsatellite repeat data. Genetics **148**:1921–1930.

KREITMAN, M., and H. AKASHI. 1995. Molecular evidence for natural selection. Ann. Rev. Ecol. Syst. **26**:403–422.

MAYNARD SMITH, J., and J. HAIGH. 1974. The hitch-hiking effect of a favorable gene. Genet. Res. **23**:23–35.

MCDONALD, J. H., and M. KREITMAN. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature **351**:652–654.

MORAN, P. A. P. 1975. Wandering distributions and the electrophoretic profile. Theor. Popul. Biol. **8**:318–330.

NACHMAN, M. W. 2001. Single nucleotide polymorphisms and recombination rate in humans. Trends Genet. **17**:481–485.

OHTA, T., and M. KIMURA. 1973. The model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic population. Genet. Res. **22**:201–204.

ORR, H. A., and A. BETANCOURT. 2001. Haldane's sieve and adaptation from the standing genetic variation. Genetics **157**:875–884.

PAYSEUR, B. A., and M. W. NACHMAN. 2000. Microsatellite variation and recombination rate in the human genome. Genetics **156**:1285–1298.

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.

ROGERS, A. R., and H. HARPENDING. 1992. Population-growth makes waves in the distribution of pairwise genetic-differences. Mol. Biol. Evol. **9**:552–569.

SCHLÖTTERER, C., and T. WIEHE. 1999. Microsatellites, a neutral marker to infer selective sweeps. Pp. 238–248 *in* D. B. GOLDSTEIN and C. SCHLÖTTERER, eds. Microsatellites: evolution and applications. Oxford University Press, Oxford.

SIMONSEN, K. L., G. A. CHURCHILL, and C. F. AQUADRO. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141**:413–439.

SMITH, M. W., J. A. LAUTENBERGER, H. D. SHIN, J.-P. CHRETIEN, S. SHRESTHA, D. A. GILBERT, and S. J. O'BRIEN. 2001. Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. Am. J. Hum. Genet. **69**:1080–1094.

SOKAL, R. R., and F. J. ROHLF. 1995. Biometry. W. H. Freeman, New York.

TAJIMA, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**:585–595.

TORRONI, A., H.-J. BANDELT, L. D'URBANO et al. (11 co-authors). 1998. MtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. Am. J. Hum. Genet. **62**:1137–1152.

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

WATTERSON, G. A. 1978. The homozygosity test of neutrality. Genetics **88**:405–417.

WAYNE, M. L., and K. L. SIMONSEN. 1998. Statistical tests of neutrality in the age of weak selection. Trends Ecol. Evol. **13**:236–240.

WEBER, J. L., and C. WONG. 1993. Mutation of short human tandem repeats. Hum. Mol. Genet. **2**:1123–1128.

WIEHE, T. 1998. The effect of selective sweeps on the variance of the allele distribution of a linked multiallele locus: hitchhiking of microsatellites. Theor. Popul. Biol. **53**:272–283.

WOLFGANG STEPHAN, reviewing editor

Accepted February 28, 2002