Nucleotide Variation at *Msn* and *Alas2*, Two Genes Flanking the Centromere of the X Chromosome in Humans

Michael W. Nachman,*[,1] Susan L. D'Agostino,* Christopher R. Tillquist,[†] Zahra Mobasher[†] and Michael F. Hammer*[,†]

**Department of Ecology and Evolutionary Biology and [†]Genomic Analysis and Technology Core, Division of Biotechnology, University of Arizona, Tucson, Arizona 85721*

Manuscript received December 25, 2003
Accepted for publication January 29, 2004

ABSTRACT

The centromeric region of the X chromosome in humans experiences low rates of recombination over a considerable physical distance. In such a region, the effects of selection may extend to linked sites that are far away. To investigate the effects of this recombinational environment on patterns of nucleotide variability, we sequenced 4581 bp at *Msn* and 4697 bp at *Alas2*, two genes situated on either side of the X chromosome centromere, in a worldwide sample of 41 men, as well as in one common chimpanzee and one orangutan. To investigate patterns of linkage disequilibrium (LD) across the centromere, we also genotyped several informative sites from each gene in 120 men from sub-Saharan Africa. By studying X-linked loci in males, we were able to recover haplotypes and study long-range patterns of LD directly. Overall patterns of variability were remarkably similar at these two loci. Both loci exhibited (i) very low levels of nucleotide diversity (among the lowest seen in the human genome); (ii) a strong skew in the distribution of allele frequencies, with an excess of both very-low and very-high-frequency derived alleles in non-African populations; (iii) much less variation in the non-African than in the African samples; (iv) very high levels of population differentiation; and (v) complete LD among all sites within loci. We also observed significant LD between *Msn* and *Alas2* in Africa, despite the fact that they are separated by ~10 Mb. These observations are difficult to reconcile with a simple demographic model but may be consistent with positive and/or purifying selection acting on loci within this large region of low recombination.

THE amount and distribution of genetic variation in human populations is a central issue in population genetics. With the completion of the human genome sequence (Lander *et al.* 2001; Venter *et al.* 2001), a major goal now is to identify variation among individuals and among populations. A detailed description of this variation will provide the necessary background for the design of efficient association studies to uncover genes involved in complex diseases. Studies of human molecular variation also shed light on the relative importance of different population genetic processes (*e.g.*, mutation, drift, selection, and recombination) and thus provide clues to the mechanism of evolutionary change at the molecular level. Finally, patterns of nucleotide variation across the genome help reveal human evolutionary history, including relationships among major ethnic groups, patterns of migration and range expansions, and changes in population size.

Considerable work over the past decade has documented DNA sequence variation in humans. Early studies focused primarily on mitochondrial DNA (Vigilant *et al.* 1991) and the Y chromosome (Hammer 1995; Whitfield *et al.* 1995; Underhill *et al.* 2000), while more recent single-locus studies have focused on the X chromosome (*e.g.*, Nachman *et al.* 1998; Harris and Hey 1999; Kaessmann *et al.* 1999; Nachman and Crowell 2000a; Gilad *et al.* 2002; Saunders *et al.* 2002; Verrelli *et al.* 2002; Yu *et al.* 2002) and on the autosomes (*e.g.*, Harding *et al.* 1997; Clark *et al.* 1998; Rieder *et al.* 1999; Fullerton *et al.* 2000; Hamblin and Di Rienzo 2000; Harding *et al.* 2000; Zhao *et al.* 2000; Alonso and Armour 2001; Bamshad *et al.* 2002; Enard *et al.* 2002; Toomajian and Kreitman 2002; Wooding *et al.* 2002). One of the clear results to emerge from this body of work is the substantial heterogeneity among genes in overall patterns of variation, including differences in the level of nucleotide diversity, the amount of linkage disequilibrium, and the frequency distribution of alleles. For example, Li and Sadler (1991) suggested 13 years ago that the average level of nucleotide heterozygosity is quite low in humans ($\pi = 0.1\%$), and subsequent work has largely confirmed this result (Przeworski *et al.* 2000). However, it is clear that this average value masks substantial variation in levels of heterozygosity among different genes; some loci exhibit almost no variation (*e.g.*, Xq13.3, $\pi = 0.036\%$; Kaessmann *et al.* 1999) while others exhibit variation more than four times higher than average (*e.g.*, 16p13.3, $\pi = 0.46\%$; Alonso and Armour 2001). A portion of these

[1]*Corresponding author:* Department of Ecology and Evolutionary Biology, Biosciences West Bldg., University of Arizona, Tucson, AZ 85721. E-mail: nachman@u.arizona.edu

differences may be accounted for by the differences in effective population size associated with the autosomes, X chromosome, Y chromosome, or mitochondrial DNA; however, heterozygosity still varies by more than one order of magnitude once effective population size differences are taken into account (Nachman 2001). Similarly, some regions of the genome exhibit nonrandom associations [*i.e.*, linkage disequilibrium (LD)] between single-nucleotide polymorphisms (SNPs) over hundreds of kilobases (Reich *et al.* 2001; Sabeti *et al.* 2002; Saunders *et al.* 2002), while other regions exhibit no associations over distances of <1 kb. Likewise, the distribution of allele frequencies differs significantly among loci (Hey 1997). For example, mitochondrial loci and many nuclear loci harbor an excess (over neutral predictions) of low-frequency alleles (*e.g.*, Hey 1997; Kaessmann *et al.* 1999; Nachman and Crowell 2000a; Stephens *et al.* 2001; Ptak and Przeworski 2002), while some nuclear genes show the opposite pattern, with an excess of intermediate-frequency alleles (Harding *et al.* 1997; Harris and Hey 1999; Bamshad *et al.* 2002).

These and other patterns of DNA sequence variation are context dependent in at least two important ways. First, the distribution of genetic variation is a property of populations and, as such, is expected to vary among populations with different histories. For example, Reich *et al.* (2001) report higher levels of LD in non-African compared with African populations. There is also considerable evidence suggesting that, in general, non-African populations harbor less genetic variation than African populations (*e.g.*, Vigilant *et al.* 1991). For many loci, African populations harbor more rare alleles than non-African populations (Wall and Przeworski 2000), although for some loci, the opposite pattern is seen (Nachman and Crowell 2000a). Attempts to identify population-specific patterns were hampered initially by the lack of a common sampling scheme for the loci under comparison. More recently, however, several impressive studies have sampled multiple loci in a common set of individuals (*e.g.*, Frisse *et al.* 2001; Patil *et al.* 2001; Stephens *et al.* 2001; Yu *et al.* 2002; Kitano *et al.* 2003; Seattle SNPs 2003). More studies of this sort will help disentangle locus-specific effects, such as selection, from population-specific effects or patterns that are a consequence of a particular sampling strategy.

A second way in which context is important is in the genomic position of genes. Different regions of the genome differ in many important attributes, including gene density, local rate of recombination, mutation rate, and base composition. With the human genome sequence in hand, we can now begin to quantify some of these parameters more precisely and ask how they influence patterns of genetic variation. For example, nucleotide heterozygosity is positively correlated with recombination rate and negatively correlated with gene density (Payseur and Nachman 2002), a result that is expected under different models invoking the joint effects of selection and linkage (Maynard Smith and Haigh 1974; Charlesworth *et al.* 1993; Galtier *et al.* 2000). These models in turn make different predictions about the frequency distribution of alleles. Likewise, the mutation rate in mammalian genomes is likely to vary as a function of base composition. For example, mutation rates at CpG sites are ~10 times higher than the average as a consequence of deamination of 5-methylcytosine (Cooper and Krawczak 1993; Sommer and Ketterling 1996; Nachman and Crowell 2000b). Mutation rate appears to vary in a nonlinear way with overall GC content, and mutation rate (as reflected in interspecific divergence) is also positively correlated with both recombination rate and SNP density (Lercher and Hurst 2002; Waterston *et al.* 2002; Hardison *et al.* 2003; Hellmann *et al.* 2003).

To understand the determinants of nucleotide variation in humans, we have initiated a long-term study of DNA sequence polymorphism in different regions of the human genome in a common set of samples (Nachman and Crowell 2000a; Saunders *et al.* 2002; Hammer *et al.* 2003). Our sample includes 41 individuals, with 10 each from Africa, Europe, and the Americas, and 11 from Asia. Much of our effort is focused on the X chromosome, which, because of its hemizygosity in males, allows us to study long-range patterns of linkage disequilibrium. Here, we report on two genes located on either side of the X centromere, *Msn* and *Alas2* (Figure 1). Both genes are situated in regions of very low recombination and low gene density. Patterns of variation are remarkably similar at these two loci: we found overall low levels of nucleotide heterozygosity, an excess of rare alleles, considerably less variation in the non-African samples than in the African sample, and no evidence for intragenic recombination.

SUBJECTS AND METHODS

Samples: Forty-one men were chosen for the initial sequencing of *Msn* and *Alas2* (see below), including 10 from Africa, 10 from Europe, 11 from Asia (including 1 from Melanesia), and 10 from the Americas. Human genomic DNAs were isolated from lymphoblastoid cell lines established by the Y Chromosome Consortium (2002) at the New York Blood Center from blood donated by volunteers who gave informed consent. To measure LD in a larger sample, we also resequenced ~750 bp from each gene to capture several informative sites in 110 men from sub-Saharan Africa (29 South African Bantu speakers, 1 Biaka, 13 Cameroonians, 21 Gambians, 32 Khoisan, 1 Mbuti, and 13 Tanzanians). All sampling protocols were according to procedures approved by the New York Blood Center and University of Arizona Human Subjects Committees. A single male common chimpanzee (*Pan troglodytes*) and a single male orangutan (*Pongo pygmaeus*) were also surveyed from DNAs provided by O. A. Ryder.

PCR amplification and sequencing of *Msn* and *Alas2*: A map of the centromeric region of the human X chromosome is shown in Figure 1. *Msn* (moesin, membrane-organizing extension spike protein) and *Alas2* (aminolevulinate, delta-, synthase 2) are separated by ~10 Mb of DNA in the assembled sequence of the human genome (Lander *et al.* 2001). The exact distance between these loci is uncertain because of incomplete sequence assembly across the centromere of the X chromosome. Both *Msn* and *Alas2* lie in genomic regions experiencing low rates of recombination (<1 cM/Mb; Payseur and Nachman 2000; Yu *et al.* 2001; Kong *et al.* 2002). DNA was PCR amplified in 25-μl volumes with 40 cycles of 94° 1 min, 55° 1 min, and 72° 2 min. Amplification primers were designed from published sequence for *Msn* exon 2 and intron 2 (GenBank accession no. Z98946) and for *Alas2* from introns 8 and 10 (GenBank accession no. AF068624). Products were cycle sequenced on both strands and run on an ABI 377 automated sequencer. For *Msn*, a total of 4578 bp was sequenced in our worldwide sample of 41 individuals, entirely from intron 2 (the first base in our sequence corresponds to the first base of intron 2 in GenBank accession Z98946). For *Alas2*, portions of introns 8 and 10, and all of exon 9, intron 9, and exon 10, were sequenced in our worldwide sample of 41 individuals; of the *Alas2* sequence, 4697 bp represent introns. To investigate LD in our sample of 120 Africans, we resequenced nucleotides 664–1413 of *Msn* (750 bp) and 2895–3682 of *Alas2* (789 bp). Sequences have been submitted to Gen Bank under accession nos. AY530963–AY531005 (*Msn*) and AY532068–AY532109 and AY532641 (*Alas2*).

Data analysis: Sequences were aligned by eye, and the numbers and frequencies of all polymorphisms were counted. Two measures of nucleotide variability were calculated: π (Nei and Li 1979) and θ (Watterson 1975). Nucleotide diversity, π, is based on the average number of nucleotide differences between two sequences randomly drawn from a sample, and θ is based on the proportion of segregating sites in a sample. Under neutral equilibrium conditions, both π and θ estimate the parameter $3N_e\mu$ for *X*-linked loci, where N_e is the effective population size and μ is the neutral mutation rate. Departures from a neutral steady-state frequency distribution of polymorphisms were evaluated using three approaches (Tajima 1989; Fu and Li 1993; Fay and Wu 2000). Tajima's (1989) test compares the average number of nucleotide differences between sequences (π) with the proportion of polymorphic sites (θ) in a sample; Fu and Li's (1993) test is based on the number of singletons in a sample; and Fay and Wu's (2000) test is based on the number of high-frequency-derived polymorphic nucleotides in a sample. Both Tajima's (1989) and Fu and Li's (1993) tests may reject the null model because of selection or because of demographic processes (such as a population bottleneck); however, Fay and Wu's (2000) test is unlikely to reject the null model except in cases where selection is operating (but see also Przeworski 2002). Linkage disequilibrium (D') was calculated for a set of independent pairwise comparisons between nonunique polymorphic sites (Lewontin 1964, 1995), and the significance of D' was assessed using Fisher's exact tests (FET). Ratios of polymorphism within humans to divergence between human and chimpanzee were compared with expectations under a neutral model using the Hudson-Kreitman-Aguadé (HKA) test (Hudson *et al.* 1987). Polymorphism was based on variation segregating among the 41 human alleles and divergence was based on a single randomly chosen human allele and a single chimpanzee allele. The program Genetree v 9.0 (Bahlo and Griffiths 2000) was used to infer the root of the tree and to estimate the time to the most recent common ancestor (TMRCA). Analysis of molecular variance (AMOVA) was used to infer population structure (Excoffier *et al.* 1992).

RESULTS

Levels of polymorphism and divergence: A total of ~9.6 kb was sequenced from *Msn* and *Alas2* in a sample of 41 globally dispersed humans (Figure 1). Because all of the sequence from *Msn* is from introns, we have excluded the short exon sequences from *Alas2* in all the analyses that follow. This increases the likelihood that all comparisons are among genomic regions experiencing similar levels of selective constraint. Thus, most analyses and discussion refer only to the 9281 bp of intron sequences (Table 1). Polymorphic sites for *Msn* and *Alas2* introns are shown in Table 2. Numbers of segregating sites, nucleotide diversity, measures of the distribution of allele frequencies, and levels of divergence are summarized in Table 1 for the complete data set of 41 individuals. Nine segregating sites were observed in *Msn*, while 7 segregating sites and two single-base insertion-deletion polymorphisms were observed in *Alas2*. *Msn* also had a variable poly(A) tract ranging in length from 17 to 25 bp. Nucleotide diversity was low at both *Msn* (π = 0.00035) and *Alas2* (π = 0.00015), as was Watterson's θ (0.00046 and 0.00035, respectively).

Divergence between humans and chimpanzee for both *Msn* (0.0092) and *Alas2* (0.0055) was comparable to results from previous studies of X-linked introns (average divergence for seven loci = 0.0072; Nachman *et al.* 1998), suggesting that the neutral mutation rate at these loci is similar to genomic average values ($\sim 2 \times 10^{-8}$/site/generation; Nachman and Crowell 2000b). Nonetheless, the divergence at *Msn* was >50% higher than the divergence at *Alas2*. Similarly, levels of polymorphism at *Msn* were higher than levels of polymorphism at *Alas2*. The fact that *Msn* is more variable than *Alas2* both within and between species is consistent with a higher mutation rate (or lower level of constraint, or both) for *Msn* compared to *Alas2*.

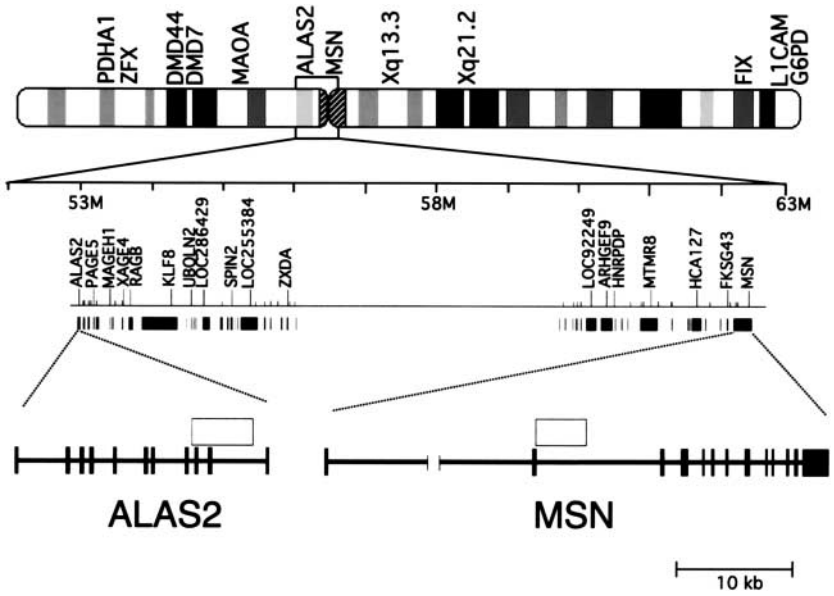

FIGURE 1.—Map of the human X chromosome. Genes for which published polymorphism data are available are shown above the chromosome. The centromeric region flanked by *Msn* and *Alas2* is shown below the chromosome, with all known genes in this 10-Mb region. Open boxes immediately above *Alas2* and *Msn* indicate the regions that have been sequenced in this study.

We compared levels of polymorphism and divergence at *Msn* and *Alas2* to levels of polymorphism and divergence at two other X-linked loci using the HKA test (HUDSON *et al.* 1987) to test the neutral prediction that these ratios should be the same. Polymorphism data include SNPs in humans, and divergence is based on a randomly chosen allele from humans and a randomly chosen allele from chimpanzees. *DmdI44* (NACHMAN and CROWELL 2000a) and *Pdha1* (HARRIS and HEY 1999) were chosen as reference loci in these comparisons because both loci reside in genomic regions with moderate to high rates of recombination and thus should be relatively free of the effects of selection at linked sites. *DmdI44* was surveyed in the same set of individuals used in the present study. In the total sample, *Msn* and *Alas2* showed marginally significantly lower variation than expected ($0.04 < P < 0.10$) relative to both *DmdI44* and *Pdha1* (Table 3). The combined data from *Msn* and *Alas2* showed significantly lower variation than expected relative to both *DmdI44* and *Pdha1* ($P = 0.03$; Table 3). In the African sample alone, *Msn* showed significantly lower variation than expected while *Alas2* did not; in the non-African sample, both *Msn* and *Alas2* showed low variation relative to *DmdI44* but not relative to *Pdha1* (Table 3). This is consistent with the known low level of variation at *Pdha1* in non-African populations (HARRIS and HEY 1999).

Frequency distribution of polymorphisms: The frequency distribution of all polymorphisms is plotted in Figure 2. The ancestral state of each polymorphic site was inferred by comparison with the chimpanzee and orangutan sequences, and the frequency of the derived state is shown for each polymorphism. The distribution is characterized by a large number of both low-frequency- and high-frequency-derived polymorphisms. We compared the observed distribution with the distribution expected under the standard neutral model using Tajima's *D*, Fu and Li's *D*, and Fay and Wu's *H* tests (Tables 1 and 4). In the total sample, all three of these tests take on negative values for *Msn* alone, *Alas2* alone,

TABLE 1

Nucleotide polymorphism and divergence at *Msn* and *Alas2*

Locus	Length (bp)	Sample size	S^a	π (SD) (%)[a]	θ (SD) (%)[a]	Tajima's D^b	Fu and Li's D^b	Fay and Wu's H^b	Divergence of Homo-Pan[c] (%)	Divergence of Homo-Pongo[c] (%)
Msn	4584	41	9	0.035 (0.006)	0.046 (0.020)	−0.85	−1.63*	−2.05*	0.92	2.25
Alas2	4697	41	7	0.015 (0.004)	0.035 (0.017)	−1.42**	−1.90**	−0.65	0.55	2.66[d]
Msn-Alas2	9281	41	16	0.025 (0.004)	0.040 (0.015)	−1.27*	−2.17**	−2.70	0.73	2.45[d]

*$P < 0.10$; **$P < 0.05$.

[a] Polymorphism statistics are based on single-nucleotide polymorphisms only and do not include indels.

[b] Measures of the frequency distribution of segregating variation are based on both SNPs and indels; values based on SNPs alone are nearly identical and do not change significance levels for any of the tests.

[c] Divergence based on a randomly chosen allele from humans (YCC 49).

[d] Homo-Pongo divergence for *Alas2* is based on 4439 bp.

TABLE 2

Individual samples and polymorphic sites at *MSN* and *ALAS2*

Continent	Country	Ethnic/language group	YCC no.	HI	Polymorphic sites: *MSN* 120	245	426	1046	1312	1414	1956	2514	4271	4325	Cen	*ALAS2* 495	567	2144	3126	3203	3416	3740/3946	4588
					T	A	G	G	T	G	[a]	A	T	A		T	G	C	G	—	G	—	A	C
Africa	Namibia	Khoisan	38	E1	.	.	.	.	.	C	.	.	.	T		.	.	T	.	T	T	.	.	.
	Namibia	Khoisan	22	E2	.	C	.	.	.	C	.	.	.	T		.	.	T	.	T	T	.	.	.
	S. Africa	E. Bantu	32	D5	.	.	.	A	C	C	[b]	.	.	T		.	.	.	.	.	.	.	.	T
	S. Africa	E. Bantu	33	E2	.	C	.	.	.	C	.	.	.	T		.	.	T	.	T	T	.	.	.
	S. Africa	W. Bantu	40	D4	.	.	.	A	C	C	[b]	.	.	T		C	.	.	.	.	.	.	.	.
	C. African Repub.	Biaka	7	D3	.	.	.	A	C	C	[c]	.	.	T		.	.	.	.	.	.	.	G	.
	C. African Repub.	Biaka	6	C1	.	.	.	.	.	C	.	.	.	T		.	.	.	.	.	.	.	.	.
	C. African Repub.	Mbuti	8	E3	.	.	.	.	.	C	.	.	.	T		.	A	T	.	T	T	.	.	.
	C. African Repub.	Mbuti	65	E3	.	.	.	.	.	C	.	.	.	T		.	A	T	.	T	T	.	.	.
	C. African Repub.	Mbuti	9	D2	.	.	.	A	C	C	[c]	C	.	T		.	.	.	.	.	.	.	.	.
Europe/ Middle East	United Kingdom	English	26	B1	.	T	.	.	.	C	.	.	.	T		.	.	.	.	.	.	T	.	.
	Germany	German	61	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	Germany	German	64	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	Germany	German	62	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	E. Europe	Ashkenazi	24	A5	.	.	.	.	.	.	.	.	.	.		.	.	.	A	.	.	.	.	.
	Poland	Ashkenazi	59	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	S. W. Russia	Adygean	56	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	Russia	Russian	72	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	Russia	Russian	71	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	Turkey	Turkish	79	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.

(*continued*)

TABLE 2

(Continued)

					Polymorphic sites																				
					MSN											*ALAS2*									
Continent	Country	Ethnic/language group	YCC no.	HI	120	245	426	1046	1312	1414	1956	2514	4271	4325	Cen	495	567	2144	3126	3203	3416	3740[f]	3946	4588
Asia	Japan	Japanese	78	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	Japan	Japanese	76	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	Japan	Japanese	77	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	China	S. Han	66	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	China	S. Han	67	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	China	S. Han	68	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	Cambodia	Cambodian	69	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	Pakistan	Pakistani	57	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	Siberia	Yakut	49	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	Siberia	Yakut	51	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	Melanesia	Nasioi	10	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
Americas	USA	Poarch Creek	27	D1	.	.	.	A	C	C	[d]	.	.	T		.	.	.	.	.	.	.	.	.
	USA	Tohono O'Odham	25	A4	.	.	.	.	.	.	.	.	G	.		.	.	.	.	.	.	.	.	.
	USA	Navajo	23	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	USA	Amerindian	2	A3	C	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	USA	Amerindian	4	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	Mexico	Mayan	17	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	Brazil	Karitiana	12	A2	.	.	A	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	Brazil	Karitiana	13	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	Brazil	Surui	16	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
	Brazil	Surui	14	A1	.	.	.	.	.	.	.	.	.	.		.	.	.	.	.	.	.	.	.
Pan troglodytes					.	.	.	.	C	C	[d]	.	.	T		.	.	.	.	T	.	.	.	.
Pongo pygmaeus					.	.	.	.	C	C	[e]	.	.	T		.	C	.	.	T	.	.	A	.

Polymorphisms at *Msn* and *Alas2* among 41 humans are shown. The consensus sequence is shown at the top, and the chimpanzee and orangutan sequences are shown at the bottom. HI, haplotype identification as in Figure 3. Cen, centromere. Dots indicate identity with the consensus sequence.

[a] Poly(A) tract of 17 bases.
[b] Poly(A) = 23.
[c] Poly(A) = 25.
[d] Poly(A) = 21.
[e] Poly(A) = 16.
[f] Exon.

TABLE 3

HKA test results comparing *Msn* and *Alas2* to *DmdI44* and *Pdha1*

Geographic region	Locus comparison	HKA χ^2	*P*-value
World	*Msn - DmdI44*	3.71	0.05
	Msn - Pdha1	4.06	0.04
	Alas2 - DmdI44	2.67	0.10
	Alas2 - Pdha1	2.99	0.08
	Msn+Alas2 - DmdI44	4.48	0.03
	Msn+Alas2 - Pdha1	4.91	0.03
Africa	*Msn - DmdI44*	4.31	0.04
	Msn - Pdha1	4.39	0.04
	Alas2 - DmdI44	1.15	0.28
	Alas2 - Pdha1	1.26	0.26
	Msn+Alas2 - DmdI44	4.07	0.05
	Msn+Alas2 - Pdha1	4.24	0.04
Non-Africa	*Msn - DmdI44*	3.99	0.05
	Msn - Pdha1	0.59	0.44
	Alas2 - DmdI44	5.51	0.01
	Alas2 - Pdha1	0.00	1.00
	Msn+Alas2 - DmdI44	6.94	0.01
	Msn+Alas2 - Pdha1	0.48	0.49

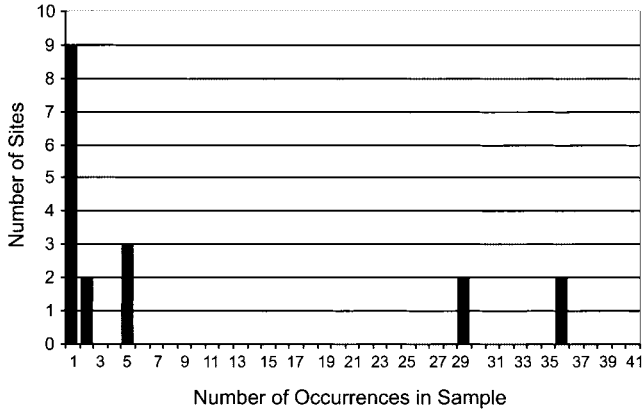

FIGURE 2.—The frequency distribution of polymorphisms at *Msn* and *Alas2* among 41 humans. Both SNPs and indels are included. The frequency of the derived state is shown, with polarity assessed in comparison to the chimpanzee and orangutan.

and for the combined data (*Msn* + *Alas2*), and most of these values are either significant or marginally significant (Table 1). When different geographic regions are considered separately, all three test statistics are ~0 in Africa, consistent with a neutral model, but are strongly negative in non-African populations (Table 4). Thus, much of the deviation observed in the total sample appears to be due to deviations largely in the non-African populations. The direction of the deviation is consistent with positive selection, a population expansion, background selection (depending on the strength of selection), and/or some form of population structure (see DISCUSSION).

Linkage disequilibrium: Complete linkage disequilibrium was observed among all sites within *Msn* and among all sites within *Alas2*. For example, when all pairwise comparisons were made among nonsingleton segregating sites, none of the comparisons between pairs of sites in *Msn* ($N = 10$) or *Alas2* ($N = 5$) contained all four gametic types (*i.e.*, $D' = 1$ in all cases). LD was also observed between *Msn* and *Alas2*; none of the 20 comparisons between pairs of sites across *Msn* and *Alas2* contained all four gametic types. We tested the significance of LD by comparing pairs of sites in order along the chromosome; this provides a set of statistically independent comparisons for tests of significance (LEWONTIN 1995). We excluded singletons and doubletons and compared the following seven sites in order: M1046, M1312, M1414, M4325, A2144, A3203, and A3416. Significant linkage disequilibrium was observed in each of the six sequential comparisons involving these sites (FET, $P < 0.01$ for each, after Bonferroni correction for multiple tests). To see if this degree of LD was driven by the differences between African and non-African samples (see *Geographic variation*), we conducted the same analysis on the African sample alone. Despite the small sample size ($n = 10$), significant linkage disequilibrium was observed in each of the sequential comparisons involving polymorphic sites in Africa alone (FET, $P < 0.05$ for each). This analysis could not be performed on the non-African sample alone because of the absence of intermediate-frequency polymorphisms.

We were surprised to find LD between *Msn* and *Alas2* in Africa because these genes are separated by ~10 Mb. For example, in a study of 19 genomic regions using an African sample from Nigeria, REICH *et al.* (2001) found that LD typically decayed to half of its maximum value within 5 kb. Our African sample included only 10 individuals. To explore the possibility that recombinant haplotypes are present in Africa and to better document LD between *Msn* and *Alas2*, we genotyped an additional 110 African individuals for four informative sites: *Msn* 1046, *Msn* 1312, *Alas2* 3203, and *Alas2* 3416. To genotype these SNPs, we PCR amplified and sequenced ~750 bp from each gene. Table 5 shows all of the polymorphisms among these 120 Africans (the original 10 plus 110 new individuals). Three observations are noteworthy. First, $D' = 1$ between sites within each gene in the total sample of 120 Africans. Second, in comparisons between *Msn* and *Alas2*, we observed all four gametic types at appreciable frequencies, suggesting that the absence of some haplotypes in the smaller set of 10 individuals (Table 2) was simply a consequence of the small sample size. Third, despite the presence of these new haplotypes in the larger sample, we observed significant LD between sites at *Msn* and sites at *Alas2* (*Msn* 1046–*Alas2* 3203, FET $P = 0.01$, $D' = 0.58$; *Msn* 1046–*Alas2* 3416, FET $P = 0.004$, $D' = 0.86$). The difference in D' between our sample of 10 ($D' = 1$) and our sample

TABLE 4

Amount and distribution of polymorphisms at *Msn* and *Alas2* by geographic region

Locus	Geographic region	Sample size	S^a	π (SD) (%)[a]	θ (SD) (%)[a]	Tajima's D^b	Fu and Li's D^b	Fay and Wu's H^b
Msn	Africa	10	4	0.035 (0.006)	0.031 (0.019)	0.50	0.35	0.36
	Non-Africa	31	8	0.014 (0.006)	0.044 (0.020)	−1.93***	−2.68***	−2.99**
	Europe	10	3	0.013 (0.010)	0.023 (0.015)	−1.39**	0.06	−3.02**
	Asia	11	0	0.000 (0.000)	0.000 (0.000)	—	—	—
	Americas	10	7	0.031 (0.013)	0.054 (0.029)	−1.65***	−1.42*	−2.31*
Alas2	Africa	10	6	0.044 (0.006)	0.042 (0.025)	0.23	−0.30	0.80
	Non-Africa	31	1	0.001 (0.001)	0.005 (0.005)	−1.40**	−2.32**	0.12
	Europe	10	1	0.004 (0.003)	0.008 (0.008)	−1.24*	−1.69**	0.36
	Asia	11	0	0.000 (0.000)	0.000 (0.000)	—	—	—
	Americas	10	0	0.000 (0.000)	0.000 (0.000)	—	—	—
Msn-Alas2	Africa	10	10	0.040 (0.005)	0.038 (0.019)	0.37	−0.04	1.16
	Non-Africa	31	9	0.008 (0.003)	0.024 (0.011)	−2.05***	−3.17***	−2.87*
	Europe	10	4	0.009 (0.005)	0.015 (0.009)	−1.56**	−0.88	−2.67**
	Asia	11	0	0.000 (0.000)	0.000 (0.000)	—	—	—
	Americas	10	7	0.015 (0.007)	0.027 (0.014)	−1.65***	−1.42*	−2.31*

*$P < 0.10$; **$P < 0.05$; ***$P < 0.01$.

[a] Polymorphism statistics are based on single-nucleotide polymorphisms only and do not include indels.

[b] Measures of the frequency distribution of segregating variation are based on both SNPs and indels; values based on SNPs alone are very similar and do not change significance levels for any of the tests.

of 120 ($D' = 0.58$) in comparisons between *Msn* 1046 and *Alas2* 3203 highlights the importance of using large samples to make inferences concerning LD.

The LD in this data set can also be seen in the phylogenetic analysis. Using parsimony, the 18 human polymorphisms in Table 2 were mapped onto a single shortest tree of length 19 (*Msn* site 245 includes two mutations resulting in three segregating nucleotides; Figure 3A).

TABLE 5

Haplotypes defined by variant sites in subregions of *Msn* and *Alas2* and their frequencies among 120 African individuals

	Polymorphic sites										
Haplotype I.D.	*Msn* 837	*Msn* 850	*Msn* 1046	*Msn* 1094	*Msn* 1312	*Msn* 1412	*Msn* 1414	*Alas2* 2947	*Alas2* 3203	*Alas2* 3416	No. of individuals
	G	A	G	G	T	C	C	C	—	G	
C1	.	.	.	.	.	.	.	.	.	.	24
C2	.	.	.	.	.	.	G	.	.	.	4
C3	A	.	.	.	.	.	.	.	.	.	1
C4	.	.	.	C	.	.	.	.	.	.	1
D1	.	.	A	.	C	.	.	.	.	.	39
D6	.	.	A	.	C	.	.	T	.	.	1
E1	.	.	.	.	.	.	.	.	T	T	18
E4	.	G	.	.	.	.	.	.	T	T	1
E5	.	.	.	.	.	.	G	.	T	T	1
F1	.	.	A	.	C	.	.	.	T	T	6
G1	.	.	A	.	C	.	.	.	T	.	11
H1	.	.	.	.	.	.	.	.	T	.	11
H2	.	G	.	.	.	.	.	.	T	.	1
H3	.	.	.	.	.	G	.	.	T	.	1
Pan	.	.	.	.	C	.	.	.	T	.	1

Haplotypes are grouped into six types on the basis of intermediate-frequency polymorphisms at sites *Msn* 1046, *Msn* 1312, *Alas2* 3203, and *Alas2* 3416. Consensus sequence is shown at the top. Dots indicate identity with the consensus sequence.

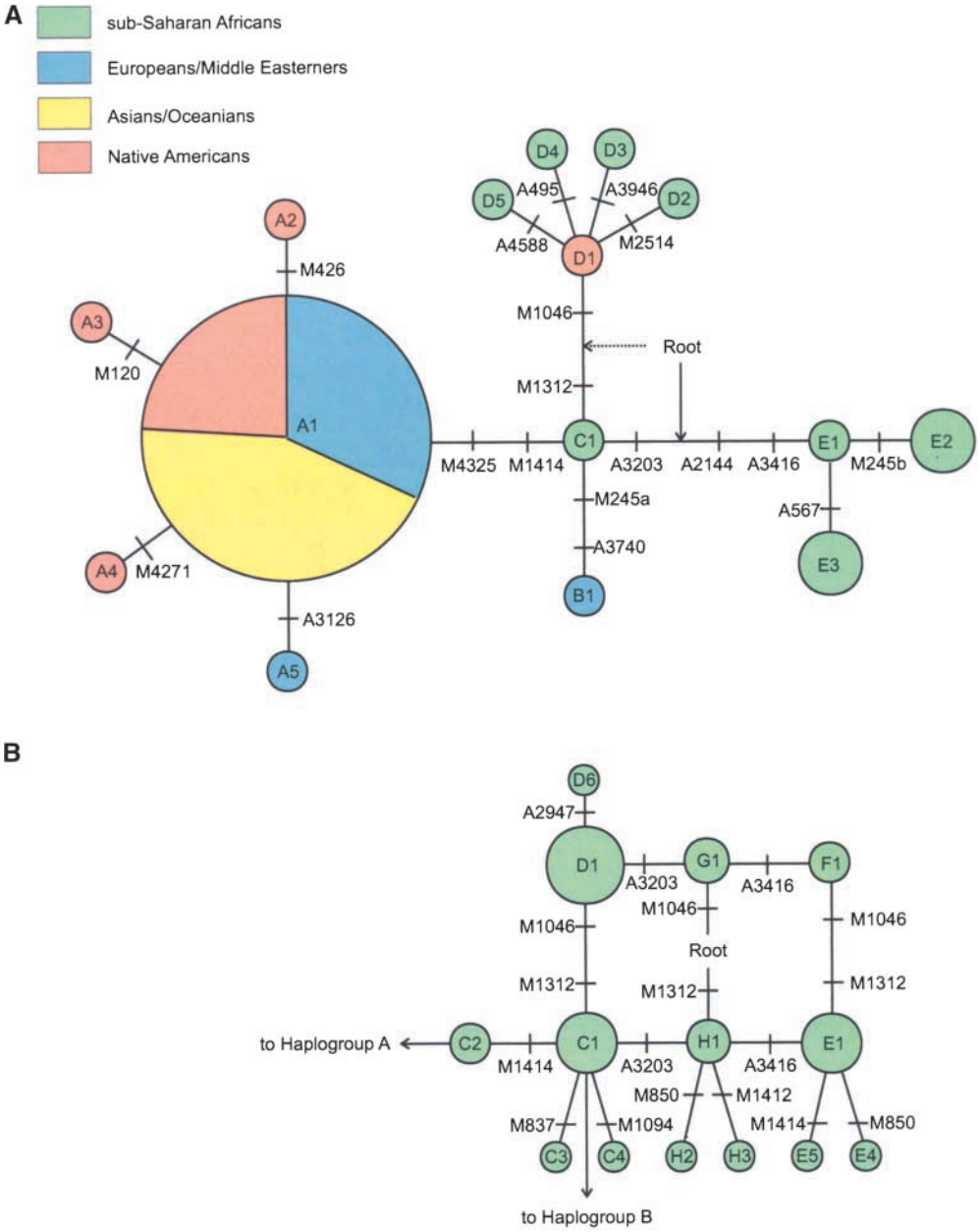

FIGURE 3.—(A) Haplotype network based on 9281 bp combined from *Msn* and *Alas2* in a worldwide sample of 41 individuals. Mutations are indicated on branches and correspond to the numbers in Table 2. The size of each circle is proportional to the frequency of the haplotype in the total sample. (B) Haplotype network for 120 African individuals based on nucleotide sites *Msn* 664–1413 (750 bp) and *Alas2* 2895–3682 (789 bp).

In this tree, there are two equally parsimonious placements of the root. A haplotype network of the 120 Africans based on subregions of *Msn* and *Alas2* is shown in Figure 3B. The reticulation in Figure 3B is consistent with the recombinant haplotypes in Table 5. In this network, there is a single most parsimonious placement of the root between haplotypes G and H. Two alternative evolutionary hypotheses for the evolution of the six African haplogroups (C–H) are shown in Figure 4; both hypotheses involve four mutations and one recombination event.

Geographic variation: The geographic distribution of nucleotide variation at *Msn* and *Alas2* is shown in Table 4. For both genes, nucleotide diversity is substantially

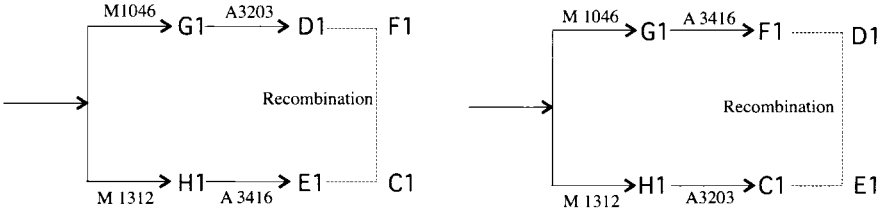

FIGURE 4.—Alternative evolutionary hypotheses for the origin of the six major African haplogroups. Both hypotheses require four mutations and one recombination event.

TABLE 6

AMOVA showing population differentiation for a set of loci sampled in the same individuals

Group	*N*	No. of populations	No. of groups	Within populations		Among populations within groups		Among groups	
				Variance (%)	Φ_{ST}	Variance (%)	Φ_{SC}	Variance (%)	Φ_{CT}
Msn-Alas2									
All populations	41	4	1	55.0	0.45				
Africa/non-Africa	41	4	2	37.0	0.63	−2.1	−0.06	65.1	0.65
DmdI7[a]									
All populations	41	4	1	71.3	0.29				
Africa/non-Africa	41	4	2	53.2	0.47	−3.8	−0.08	50.6	0.51
DmdI44[a]									
All populations	41	4	1	94.0	0.06				
Africa/non-Africa	41	4	2	92.5	0.08	4.3	0.04	3.3	0.03
NRY[b]									
All populations	41	4	1	66.7	0.33				
Africa/non-Africa	41	4	2	54.1	0.46	8.9	0.14	54.1	0.37

All Φ-statistic *P*-values are <0.01, except for *Msn-Alas2* Φ_{SC} for which $P = 0.327$.

[a] Nachman and Crowell (2000a).

[b] Hammer *et al.* (2003).

lower in the non-African than in the African samples. The distribution of haplotypes in the combined *Msn* and *Alas2* data ($N = 41$, Table 2 and Figure 3A) is quite different in the African and non-African samples. In Africa, haplotypes are present in only one or two individuals; in the non-African sample, a single very common haplotype (A1) is shared among 25 of 31 individuals (Figure 3A). This haplotype, represented by the consensus sequence in Table 2, is present in each of the non-African continents surveyed. The tree in Figure 3A is largely split into an African clade and a non-African clade. One exception to this pattern is a single Native American [Y Chromosome Consortium (YCC) 27] with a *Msn* haplotype otherwise found only in Africa. This same individual also contained a polymorphism at *DmdI7* otherwise found only in Africa (Nachman and Crowell 2000a). Both observations are consistent with recent admixture between the Poarch Creek and African Americans. Another individual (YCC 26) from the United Kingdom showed two polymorphisms (*Msn* 1414 C and *Msn* 4325 T) otherwise found only in Africa. In our sample of 120 Africans, 115 individuals had a "C" at site *Msn* 1414 while only 5 individuals had a "T" at this site. This is largely consistent with the division of the tree in Figure 3A into a mostly African clade and a mostly non-African clade. It is interesting to note that, as recently observed for another X-linked locus (Nachman and Crowell 2000a), Native Americans have more diversity than Asians at *Msn*. This is still true even when the Poarch Creek chromosome is removed from the analysis.

The differentiation between Africans and non-Africans was reflected in patterns of within- and between-group variation in an AMOVA. When the four population samples (*i.e.*, from each continent) were clustered in a single group, Φ_{ST} was 0.45 (Table 6). When populations were divided into Africans and non-Africans, the Φ_{ST} value increased to 0.63. Interestingly, 100% of this between-group variation was partitioned between Africans and non-Africans (*e.g.*, $\Phi_{CT} = 0.65$ and $\Phi_{SC} = -0.06$; Table 6). The difference between Africans and non-Africans at *Msn* + *Alas2* is greater than the level of differentiation observed for other loci on the X or Y chromosomes sampled in these same individuals (Table 6).

The network in Figure 3A contains three major haplogroups (C–E) in sub-Saharan Africa. Haplogroup C was found in a single Pygmy, haplogroup D (D1–D5) was found in South African Bantu speakers and Pygmies (as well as in the Poarch Creek; see above), and haplogroup E (E1–E3) was found in Khoisan and in Pygmies. Thus, Pygmies exhibit the highest level of diversity in this small sample of sub-Saharan Africans.

Comparisons with the chimpanzee sequence place the root of the tree in Figure 3A either on the branch between haplogroups C and D or on the branch between haplogroups C and E; these alternative roots are equally parsimonious in our sample of 41 individuals. The larger sample of 120 African individuals places the root between haplogroups G and H (Figure 3B). In both cases, African samples occur on both sides of the deepest node in the tree. Thus, Africa is the most likely location of the ancestral *Msn-Alas2* sequence. We used two approaches to estimate the time to the most recent common ancestral *Msn-Alas2* sequence. First, we calculated the average number of differences between the root of the tree in Figure 3A and each haplotype and multiplied this by two. This reflects the average distance between haplotypes through the root of the tree. We assumed a human-

chimpanzee divergence of 6 million years and calculated the ratio of the average distance between haplotypes through the root of the human tree to the human-chimpanzee *Msn-Alas2* divergence. The average number of mutations between sequences across the base of the human gene tree was 6.2, while divergence between the human consensus and the chimpanzee sequences was 73. This leads to an estimated time of human sequence divergence of 510,000 years. This time estimate is slightly larger than the one generated from the TMRCA obtained from maximum-likelihood simulations (Bahlo and Griffiths 2000). The maximum-likelihood estimate for the TMRCA was $1.2 \times 1.5N$ generations (assuming a panmictic population of constant size). Substituting a value of $N = 10{,}000$ (Hammer 1995) and assuming 20 years/generation, the TMRCA was 360,900 years ± 98,200 years.

DISCUSSION

We investigated the levels and patterns of nucleotide variation in noncoding sequences from two genes mapping on either side of the X chromosome centromere in a worldwide sample of 41 humans. These genes are located ~10 Mb apart and both lie in genomic regions with low rates of recombination and low gene density. We were interested in exploring the effects of this "genomic context" on patterns of variation at these genes. Two major results emerge from this study. First, we found significant linkage disequilibrium across the X chromosome centromere. Second, levels and patterns of variation at both genes show significant departures from a standard neutral model of evolution. We discuss each of these in turn.

Linkage disequilibrium across the centromeric region of the X chromosome: We observed significant LD within *Msn*, within *Alas2*, as well as significant LD between these genes. LD was seen in our worldwide sample of 41 individuals (containing only 10 Africans), and it was also seen in our sample of 120 African individuals.

It is instructive to compare the LD seen in this study to the LD observed at *Dmd* and *G6pd*, two other X-linked loci that have been surveyed in these same 41 individuals. For example, at *DmdI44*, recombinants (*i.e.*, all four gametic types) are seen between nucleotides separated by <200 bp. *Dmd* lies in a genomic region with high rates of recombination (>2 cM/Mb), and this likely contributes to the difference in LD seen between *Dmd* and *Msn* + *Alas2*. *G6pd* lies in Xq28 in a region of moderate recombination (~1–2 cM/Mb). Several mutations at *G6pd* are known to confer resistance to malaria and thus are under selection in regions of the world where malaria is common. In Africa, the *G6pd* A− allele is in linkage disequilibrium with mutations at *L1cam*, a locus situated ~500 kb from *G6pd*, and this long-range LD is likely caused by selection acting on the *G6pd* A− allele (Sabeti *et al.* 2002; Saunders *et al.* 2002). The distance between *Msn* and *Alas2* over which significant LD is seen is ~10 times greater than the LD seen near *G6pd*, a locus known to have unusually high LD as a consequence of selection. Because these estimates come from the same sample of individuals, the differences in LD among loci are unlikely to be due to population-level effects or to sampling strategy.

It is also useful to compare the LD seen in this study to the LD observed in other samples. Reich *et al.* (2001) studied the decay of LD throughout the genome in both African and non-African populations. By measuring D' between SNPs at regular intervals, they were able to measure the "half-length" of LD (the distance at which the average D' drops below 0.5) for 19 genomic regions spanning a range of recombination rates. The average half-length of LD was 60 kb in a population of European ancestry and was <5 kb in a Yoruban population from Nigeria. This supported earlier suggestions that LD was generally lower in African than in non-African populations (*e.g.*, Tishkoff *et al.* 1996). In contrast, we observed $D' > 0.5$ among 120 Africans between SNPs separated by ~10 Mb.

What is the cause of the high LD near the centromere of the X chromosome? Because this amount of LD is not seen at other loci sampled in the same individuals, nor in other samples from the same general geographic regions, it is unlikely to be due to population-level effects such as a bottleneck or admixture. Two other factors may contribute to the observed LD. The first is simply that the centromeric region of the X chromosome experiences low rates of recombination. Estimates of recombination rate in this region are on the order of 0.1–0.6 cM/Mb (Yu *et al.* 2001; Kong *et al.* 2002; Payseur and Nachman 2000), suggesting that *Msn* and *Alas* are separated by 1–6 cM. While the overall low level of recombination likely contributes to the observed patterns, LD over 1–6 cM is still highly unusual in the human genome (Huttley *et al.* 1999) and is otherwise unknown in African populations, except in cases where selection is known to be acting (Sabeti *et al.* 2002; Saunders *et al.* 2002). The second factor that may contribute to LD in our data is natural selection. Either positive or purifying selection at linked sites could increase the level of LD, as discussed below.

Rejection of the standard neutral model: Several observations suggest that patterns of variation at both *Msn* and *Alas2* are not consistent with the standard neutral model. First, there are low levels of variability at both loci despite typical levels of divergence. In the total data set (*Msn* + *Alas2*, $N = 41$), the HKA test rejects the null model in comparison with two other X-linked loci, *DmdI44* and *Pdha1*, chosen because they reside in genomic regions with above-average rates of recombination and thus are likely to be free of the effects of selection at linked sites (Table 3). The inference of significantly lower variation at *Msn* + *Alas2* comes with two caveats. One is that we have performed multiple HKA tests but

TABLE 7

Nucleotide variability at X-linked genes in humans

Locus	Length (bp)	Sample size	*S*	π (%)	θ (%)	Tajima's *D*	Reference
Pdha1	4153	35	25	0.189	0.146	1.03	Harris and Hey (1999)
Zfx	1089	336	10	0.082	0.143	−0.95	Jaruzelska *et al.* (1999)
DmdI44	3000	41	19	0.141	0.148	−0.16	Nachman and Crowell (2000a)
DmdI7	2389	41	9	0.034	0.088	−1.79	Nachman and Crowell (2000a)
Mao-a	18820	56	41	0.050	0.047	0.34	Gilad *et al.* (2002)
Alas2	4697	41	7	0.015	0.035	−1.42	This study
Msn	4584	41	9	0.035	0.046	−0.85	This study
Msn + Alas2	9281	41	16	0.025	0.040	−1.27	This study
Xq13.3	10163	69	33	0.036	0.068	−1.61	Kaessmann *et al.* (1999)
Xq21.3	10346	62	44	0.071	0.091	−0.70	Yu *et al.* (2002)
F9	3731	36	6	0.014	0.039	−1.71	Harris and Hey (2001)
L1cam	2087	41	6	0.020	0.070	−1.93	Saunders *et al.* (2002)
G6pd	2918	41	10	0.040	0.080	−1.51	Saunders *et al.* (2002)

have not corrected the significance level for multiple comparisons; thus the reduction should be interpreted as modest. The other is that the significance of the HKA test, and the subsequent inference of selection on particular loci, depends on the choice of reference loci. For example, compared with other loci showing little variation (such as *Xq13.3*, Kaessmann *et al.* 1999; or *F9*, Harris and Hey 2001), neither *Msn* nor *Alas2* rejects the null model in an HKA test (data not shown). Table 7 lists X-linked loci in humans for which polymorphism data are available from large samples. *Msn* and *Alas2* are among the least variable loci surveyed on the X chromosome. Considering Watterson's θ, *Alas2* is the least variable locus, and *Msn* is the third least variable locus. Most other genes that show low variability, such as *F9* (Harris and Hey 2001), *Mao-a* (Gilad *et al.* 2002), and *G6pd* (Saunders *et al.* 2002), are believed to be influenced by selection. Thus, even though the *P*-values associated with the HKA tests in Table 3 are not strongly significant, it is true that *Msn* and *Alas2* are among the least variable genes in the human genome (Nachman 2001).

In addition to the reduction in variability at *Msn* and *Alas2*, we observed a significant skew in the distribution of allele frequencies, particularly in non-African populations. This is seen in the negative values for Tajima's *D*, Fu and Li's *D*, and Fay and Wu's *H* statistics (Tables 1 and 4). These observations are certainly consistent with selection, but may also be consistent with some demographic explanations. For example, a population expansion is expected to lead to negative values of Tajima's *D* and Fu and Li's *D* and may help account for the values in Table 4. Fay and Wu's *H*, which is based on the frequency of derived polymorphic nucleotides, is not expected to reject the null model under a simple population expansion (Fay and Wu 2000). However, some forms of population structure may lead to significantly negative values of this test statistic even in the absence of selection (Wakeley and Aliacar 2001; Przeworski 2002). For example, Przeworski (2002) showed that Fay and Wu's *H* might lead to a rejection under a symmetric two-island model with moderate migration even if individuals are sampled from only one of the populations. In our combined *Msn + Alas2* data, Fay and Wu's *H* is significantly negative in the non-African but not in the African sample (Table 4). This result is entirely driven by three polymorphisms at *Msn*: 1312, 1414, and 4325. At each of these sites, the ancestral nucleotide is common or fixed in Africa and is represented in the non-African sample in only two individuals (YCC 26 and YCC 27), one of whom may be partially of African origin (see above). Thus it is possible that the significant Fay and Wu's *H* test derives in part from admixture of African haplotypes in non-African populations. A similar effect of admixture may contribute to the negative values of Tajima's *D* in non-African populations.

A third unexpected observation is the long-range LD seen between *Msn* and *Alas2*. As discussed above, it is not easy to account for this by any simple model of population structure, since LD is seen in the total data set and in the African sample alone. It is also not easy to account for this by the reduced recombination rate in this genomic region, since the total genetic distance between *Msn* and *Alas2* is 1–6 cM.

Finally, we observed a very high level of population structure in our data, mostly driven by differences between African and non-African samples. Two sites show a nearly fixed difference between Africa and the rest of the world (*Msn* 1414 and *Msn* 4325), and the resulting Φ_{ST} for the combined data is 0.45, a value greater than that for other loci surveyed in these individuals (Table 6). Akey *et al.* (2002) estimated F_{ST} for 26,530 SNP markers throughout the genome sampled in 42 East Asians, 42 African Americans, and 42 European Americans. The

mean F_{ST} for these data was 0.123, and ~6% of the markers had $F_{ST} \geq 0.40$. Thus *Msn* + *Alas2* show more population structure than most loci in the genome. However, autosomal loci, which constitute most of the data in Akey *et al.* (2002), are typically expected to show less population structure than X-linked loci because of differences in population size.

The standard neutral model is based on a population of constant size at mutation-drift equilibrium and in principle may be rejected because of selection, population processes, or both. At *Msn* and *Alas2* we observe low variability, a skew in the frequency spectrum, high LD, and high Φ_{ST}. Can we distinguish selection from demography as the cause of these patterns? One standard approach for distinguishing population-level processes (or artifacts of sampling) from locus-specific effects is to compare multiple loci, ideally sampled in the same set of individuals. Viewed in the context of other X-linked loci (Tables 6 and 7), *Msn* and *Alas2* are unusual in many but not all respects. *Msn* and *Alas2* show less variability than most loci and greater LD than virtually all loci. They show a strong skew in the frequency distribution with an excess of rare variants, but this is also seen at a handful of other loci. Likewise, they show considerable population structure, but this too is seen at some other loci. It appears difficult to reconcile a single demographic model with this combination of results. For example, while a population expansion out of Africa coupled with subsequent migration might explain the negative values for Tajima's *D*, Fu and Li's *D*, and Fay and Wu's *H*, it does not account for the unusually high levels of LD both in the total sample and in Africa nor does it account for the significantly lower variability at *Msn* and *Alas2* but not in other genes sampled in these same individuals.

On the other hand, many of our observations for *Msn* and *Alas2* are consistent with the action of selection at linked sites near the centromere of the X chromosome. The combination of low variability, a skew in the frequency spectrum, high LD, and high Φ_{ST} could be explained by background selection, positive directional selection, or some combination of these processes.

Background selection (Charlesworth *et al.* 1993) has the effect of reducing the effective population size of the chromosomal region in question. This reduced population size will lead to reduced levels of variation, increased levels of LD, and increased levels of differentiation among populations. In addition, when selection is weak, background selection can lead to a skew in the distribution of allele frequencies with an excess of rare alleles (Charlesworth *et al.* 1993). The overall low levels of variability at *Msn* and *Alas2* and the difference in the number of haplotypes seen in African and non-African populations might be consistent with background selection coupled with a founder effect out of Africa.

Likewise, genetic hitchhiking (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989) can lead to reduced levels of variation and a skew in the distribution of allele frequencies with an excess of rare variants (Braverman *et al.* 1995). Positive selection can lead to increased population differentiation either as a consequence of local adaptation or through the fixation of the same beneficial allele in different subpopulations with low levels of gene flow (Slatkin and Wiehe 1998). Positive selection can increase levels of LD in several ways. For example, if a selective sweep is partial or is geographically restricted, LD may be generated among linked sites on the selected chromosome (*e.g.*, Stephan *et al.* 1998; Sabeti *et al.* 2002; Saunders *et al.* 2002). Presumably, complete selective sweeps can also generate LD near the target of selection simply as a consequence of a localized reduction in population size, although this has not been studied in detail (Pritchard and Przeworski 2001). It is important to bear in mind that effects of selection at linked sites are not expected to extend far unless selection is strong. For example, the probability of a linked neutral site escaping hitchhiking is high when the selected and neutral sites are separated by more than $(0.1)s/c$ bp, where s is the selection coefficient and c is the recombination rate per base (Kaplan *et al.* 1989). Thus, if $c = 10^{-9}$ for the centromeric region of the X chromosome, and $s = 0.1$, sites >10 Mb away are likely to have recombined off of a selected chromosome. If positive selection is responsible for the observed patterns, the exact nature of this selection is not clear from our data. Significant rejections of the neutral model are seen for both African and non-African subsets of the data, although the patterns of variation in these two subsets of the data clearly differ. One straightforward explanation for the observation of multiple haplogroups in Africa and a single lineage out of Africa is a partial selective sweep of the common haplogroup (A) in non-African populations. Such a sweep could have occurred concomitant with or following the movement of anatomically modern humans out of Africa. However, selection could also be responsible for the significant reduction of variation in Africa (Table 3) as well as the significant LD in Africa, since these features are not seen at other loci.

Two recent genomic scans for selection, in each case based on different data and approaches, suggest that positive selection has acted recently near the X chromosome centromere (Akey *et al.* 2002; Payseur *et al.* 2002). Akey *et al.* (2002) estimated F_{ST} for >26,000 SNPs to identify genomic regions with unusually high levels of population differentiation that might be indicative of selection acting differently in Asia, Africa, or Europe. Payseur *et al.* (2002) compared observed and expected distributions of allele frequencies for >5000 microsatellites in a population of European origin to identify genomic regions with an excess of rare alleles that might be indicative of directional selection. Both studies identified the centromeric region of the X chromosome as

containing multiple markers showing the signature of positive selection.

The observations presented here are difficult to reconcile with a simple demographic model. However, numerous aspects of our data seem consistent with both background selection and hitchhiking models, and we emphasize that both processes may be important. In principle, it might be possible to distinguish between them by surveying microsatellite variation in this region of the X chromosome. Background selection predicts a reduction in levels of variability, even for markers with high mutation rates such as microsatellites, while genetic hitchhiking only predicts reduced variation at microsatellites for very recent selective sweeps (Slatkin 1995; Wiehe 1998; Payseur and Nachman 2000).

We thank the members of the Nachman and Hammer labs for useful discussions and comments on the manuscript. We also thank M. Noor and two anonymous reviewers for comments. This work was supported by the National Science Foundation.

LITERATURE CITED

Akey, J. M., G. Zhang, K. Zhang, L. Jin and M. D. Shriver, 2002 Interrogating a high-density SNP map for signatures of natural selection. Genome Res. **12:** 1805–1814.

Alonso, S., and J. A. Armour, 2001 A highly variable segment of human subterminal 16p reveals a history of population growth for modern humans outside Africa. Proc. Natl. Acad. Sci. USA **98:** 864–869.

Bahlo, M., and R. C. Griffiths, 2000 Inference from gene trees in a subdivided population. Theor. Popul. Biol. **57:** 79–95.

Bamshad, M. J., S. Mummidi, E. Gonzalez, S. S. Ahuja, D. M. Dunn *et al.*, 2002 A strong signature of balancing selection in the 5′ cis-regulatory region of CCR5. Proc. Natl. Acad. Sci. USA **99:** 10539–10544.

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., M. T. Morgan and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289–1303.

Clark, A. G., K. M. Weiss, D. A. Nickerson, S. L. Taylor, A. Buchanan *et al.*, 1998 Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am. J. Hum. Genet. **63:** 595–612.

Cooper, D. N., and M. Krawczak, 1993 *Human Gene Mutation.* Bios Scientific, Oxford.

Enard, W., M. Przeworski, S. E. Fisher, C. S. Lai, V. Wiebe *et al.*, 2002 Molecular evolution of FOXP2, a gene involved in speech and language. Nature **418:** 869–872.

Excoffier, L., P. E. Smouse and J. M. Quattro, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics **131:** 479–491.

Fay, J. C., and C.-I Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

Frisse, L., R. R. Hudson, A. Bartoszewicz, J. D. Wall, J. Donfack *et al.*, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. **69:** 831–843.

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

Fullerton, S. M., A. G. Clark, K. M. Weiss, D. A. Nickerson, S. L. Taylor *et al.*, 2000 Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am. J. Hum. Genet. **67:** 881–900.

Galtier, N., F. Depaulis and N. H. Barton, 2000 Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics **155:** 981–987.

Gilad, Y., S. Rosenberg, M. Przeworski, D. Lancet and K. Skorecki, 2002 Evidence for positive selection and population structure at the human MAO-A gene. Proc. Natl. Acad. Sci. USA **99:** 862–867.

Hamblin, M. T., and A. Di Rienzo, 2000 Detection of the signature of natural selection in humans: evidence from the duffy blood group locus. Am. J. Hum. Genet. **66:** 1669–1679.

Hammer, M. F., 1995 A recent common ancestry for human Y chromosomes. Nature **378:** 376–378.

Hammer, M. F., F. Blackmer, D. Garrigan, M. W. Nachman and J. A. Wilder, 2003 Human population structure and its effects on sampling Y chromosome sequence variation. Genetics **164:** 1495–1509.

Harding, R. M., S. M. Fullerton, R. C. Griffiths, J. Bond, M. J. Cox *et al.*, 1997 Archaic African and Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. **60:** 772–789.

Harding, R. M., E. Healy, A. J. Ray, N. S. Ellis, N. Flanagan *et al.*, 2000 Evidence for variable selective pressures at MC1R. Am. J. Hum. Genet. **66:** 1351–1361.

Hardison, R. C., K. M. Roskin, S. Yang, M. Diekhans, W. J. Kent *et al.*, 2003 Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome Res. **13:** 13–26.

Harris, E. E., and J. Hey, 1999 X chromosome evidence for ancient human histories. Proc. Natl. Acad. Sci. USA **96:** 3320–3324.

Harris, E. E., and J. Hey, 2001 Human populations show reduced DNA sequence variation at the Factor IX locus. Curr. Biol. **11:** 774–778.

Hellmann, I., I. Ebersberger, S. E. Ptak, S. Paabo and M. Przeworski, 2003 A neutral explanation for the correlation of diversity with recombination rates in humans. Am. J. Hum. Genet. **72:** 1527–1535.

Hey, J., 1997 Mitochondrial and nuclear genes present conflicting portraits of human origins. Mol. Biol. Evol. **14:** 166–172.

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.

Kaessmann, H., F. Heissig, A. von Haeseler and S. Paabo, 1999 DNA sequence variation in a non-coding region of low recombination on the human X chromosome. Nat. Genet. **22:** 78–81.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

Kitano, T., C. Schwarz, B. Nickel and S. Paabo, 2003 Gene diversity patterns at 10 X-chromosomal loci in humans and chimpanzees. Mol. Biol. Evol. **20:** 1281–1289.

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson *et al.*, 2002 A high-resolution recombination map of the human genome. Nat. Genet. **31:** 241–247.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860–921.

Lercher, M. J., and L. D. Hurst, 2002 Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet. **18:** 337–340.

Lewontin, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **43:** 419–434.

Lewontin, R. C., 1995 The detection of linkage disequilibrium in molecular sequence data. Genetics **140:** 377–388.

Li, W.-H., and L. A. Sadler, 1991 Low nucleotide diversity in man. Genetics **129:** 513–523.

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

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

Nachman, M. W., and S. L. Crowell, 2000a Contrasting evolutionary histories of two introns of the duchenne muscular dystrophy gene, Dmd, in humans. Genetics **155:** 1855–1864.

Nachman, M. W., and S. L. Crowell, 2000b Estimate of the mutation rate per nucleotide in humans. Genetics **156:** 297–304.

Nachman, M. W., V. L. Bauer, S. L. Crowell and C. F. Aquadro, 1998 DNA variability and recombination rates at X-linked loci in humans. Genetics **150:** 1133–1141.

Nei, M., and W.-H. Li, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA **76:** 5269–5273.

Patil, N., A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi *et al.*, 2001 Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science **294:** 1719–1723.

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

Payseur, B. A., and M. W. Nachman, 2002 Gene density and human nucleotide polymorphism. Mol. Biol. Evol. **19:** 336–340.

Payseur, B. A., A. D. Cutter and M. W. Nachman, 2002 Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. Mol. Biol. Evol. **19:** 1143–1153.

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

Przeworski, M., 2002 The signature of positive selection at randomly chosen loci. Genetics **162:** 2053.

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

Ptak, S. E., and M. Przeworski, 2002 Evidence for population gowth in humans is confounded by fine-scale population structure. Trends Genet. **18:** 559–563.

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.

Rieder, M. J., S. L. Taylor, A. G. Clark and D. A. Nickerson, 1999 Sequence variation in the human angiotensin converting enzyme. Nat. Genet. **22:** 59–62.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419:** 832–837.

Saunders, M. A., M. F. Hammer and M. W. Nachman, 2002 Nucleotide variability at g6pd and the signature of malarial selection in humans. Genetics **162:** 1849–1861.

Seattle SNPs, 2003 NHLBI Program for Genomic Applications, UW-FHCRC, Seattle (http://pga.mbt.washington.edu).

Slatkin, M., 1995 Hitchhiking and associative overdominance at a microsatellite locus. Mol. Biol. Evol. **12:** 473–480.

Slatkin, M., and T. Wiehe, 1998 Genetic hitchhiking in a subdivided population. Genet. Res. **71:** 155–160.

Sommer, S. S., and R. P. Ketterling, 1996 The factor IX gene as a model for analysis of human germline mutations: an update. Hum. Mol. Genet. **5** (Spec. No.): 1505–1514.

Stephan, W., L. Xing, D. A. Kirby and J. M. Braverman, 1998 A test of the background selection hypothesis based on nucleotide data from *Drosophila ananassae*. Proc. Natl. Acad. Sci. USA **95:** 5649–5654.

Stephens, J. C., J. A. Schneider, D. A. Tanguay, J. Choi, T. Acharya *et al.*, 2001 Haplotype variation and linkage disequilibrium in 313 human genes. Science **293:** 489–493.

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

Tishkoff, S. A., E. Dietzsch, W. Speed, A. J. Pakstis, J. R. Kidd *et al.*, 1996 Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science **271:** 1380–1387.

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

Underhill, P. A., P. Shen, A. A. Lin, L. Jin, G. Passarino *et al.*, 2000 Y chromosome sequence variation and the history of human populations. Nat. Genet. **26:** 358–361.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural *et al.*, 2001 The sequence of the human genome. Science **291:** 1304–1351.

Verrelli, B. C., J. H. McDonald, G. Argyropoulos, G. Destro-Bisol, A. Froment *et al.*, 2002 Evidence for balancing selection from nucleotide sequence analyses of human G6PD. Am. J. Hum. Genet. **71:** 1112–1128.

Vigilant, L., M. Stoneking, H. Harpending, K. Hawkes and A. C. Wilson, 1991 African populations and the evolution of human mitochondrial DNA. Science **253:** 1503–1507.

Wakeley, J., and N. Aliacar, 2001 Gene genealogies in a metapopulation. Genetics **159:** 893–905.

Wall, J. D., and M. Przeworski, 2000 When did the human population size start increasing? Genetics **155:** 1865–1874.

Waterston, R. H., K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril *et al.*, 2002 Initial sequencing and comparative analysis of the mouse genome. Nature **420:** 520–562.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Whitfield, L. S., J. E. Sulston and P. N. Goodfellow, 1995 Sequence variation of the human Y chromosome. Nature **378:** 379–380.

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

Wooding, S. P., W. S. Watkins, M. J. Bamshad, D. M. Dunn, R. B. Weiss *et al.*, 2002 DNA sequence variation in a 3.7-kb noncoding sequence 5′ of the CYP1A2 gene: implications for human population history and natural selection. Am. J. Hum. Genet. **71:** 528–542.

Y Chromosome Consortium, 2002 A nomenclature system for the tree of Y chromosomal binary haplogroups. Genome Res. **12:** 339–348.

Yu, A., C. Zhao, Y. Fan, W. Jang, A. J. Mungall *et al.*, 2001 Comparison of human genetic and sequence-based physical maps. Nature **409:** 951–953.

Yu, N., F. C. Chen, S. Ota, L. B. Jorde, P. Pamilo *et al.*, 2002 Larger genetic differences within Africans than between Africans and Eurasians. Genetics **161:** 269–274.

Zhao, Z., L. Jin, Y. X. Fu, M. Ramsay, T. Jenkins *et al.*, 2000 Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc. Natl. Acad. Sci. USA **97:** 11354–11358.

Communicating editor: M. A. F. Noor