DOI: 10.1534/genetics.103.025361

Heterogeneous Patterns of Variation Among Multiple Human X-Linked Loci: The Possible Role of Diversity-Reducing Selection in Non-Africans

Michael F. Hammer,*,†,[1] Daniel Garrigan,*,† Elizabeth Wood,*,† Jason A. Wilder,*,† Zahra Mobasher,* Abigail Bigham,* James G. Krenz† and Michael W. Nachman†

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

Manuscript received December 5, 2003
Accepted for publication April 13, 2004

ABSTRACT

Studies of human DNA sequence polymorphism reveal a range of diversity patterns throughout the genome. This variation among loci may be due to natural selection, demographic influences, and/or different sampling strategies. Here we build on a continuing study of noncoding regions on the X chromosome in a panel of 41 globally sampled humans representing African and non-African populations by examining patterns of DNA sequence variation at four loci (*APXL*, *AMELX*, *TNFSF5*, and *RRM2P4*) and comparing these patterns with those previously reported at six loci in the same panel of 41 individuals. We also include comparisons with patterns of noncoding variation seen at five additional X-linked loci that were sequenced in similar global panels. We find that, while almost all loci show a reduction in non-African diversity, the magnitude of the reduction varies substantially across loci. The large observed variance in non-African levels of diversity results in the rejection of a neutral model of molecular evolution with a multi-locus HKA test under both a constant size and a bottleneck model. In non-Africans, some loci harbor an excess of rare mutations over neutral equilibrium predictions, while other loci show no such deviation in the distribution of mutation frequencies. We also observe a positive relationship between recombination rate and frequency spectra in our non-African, but not in our African, sample. These results indicate that a simple out-of-Africa bottleneck model is not sufficient to explain the observed patterns of sequence variation and that diversity-reducing selection acting at a subset of loci and/or a more complex neutral model must be invoked.

PATTERNS of variation at multiple loci can be used to infer the history of human migration patterns, subdivision, and changes in population size. These patterns can also shed light on the relative importance of different population genetic processes (*e.g.*, mutation, genetic drift, selection, and recombination) and thus provide clues to the mechanisms of evolutionary change at the molecular level. A current challenge is to distinguish the signature of natural selection from those of neutral demographic processes associated with changes in population size, distribution, and structure. Often selective and demographic processes produce identical patterns of sequence variation at a given locus. For example, an excess of rare mutations over neutral, equilibrium expectations could be a signature of either recent directional selection at the locus under investigation or recent population growth. One approach to distinguishing between selective and neutral demographic effects on genome variability is to sample multiple independent loci: natural selection is expected to affect variation in small regions (*i.e.*, at sites linked to those under selection), while demographic processes tend to affect all loci in the genome similarly.

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), while more recent single-locus studies have focused on the X chromosome (Nachman *et al.* 1998; Harris and Hey 1999; Kaessmann *et al.* 1999; Nachman and Crowell 2000; Gilad *et al.* 2002; Saunders *et al.* 2002; Yu *et al.* 2002b) and on the autosomes (reviewed in Przeworski *et al.* 2000; Excoffier 2002). Two major features to emerge from this body of work are (1) substantial heterogeneity among genes in overall patterns of variation, including differences in the level of nucleotide diversity, the amount of linkage disequilibrium, and the distribution of allele frequencies, and (2) clear differences in levels and patterns of variation among populations. For example, there is mounting evidence that African populations have more genetic variation (Vigilant *et al.* 1991; Tishkoff *et al.* 1996; Przeworski *et al.* 2000; Hammer *et al.* 2001), harbor more rare alleles (Wall and Przeworski 2000), and have lower levels of linkage disequilibrium (Reich *et al.* 2001) than non-African populations.

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

These genetic patterns have led to contrasting inferences of human demographic history. Results from loci with an excess of rare polymorphisms (*i.e.*, with large negative Tajima's *D* values; Tajima 1989) have been used to support models in which humans expanded dramatically from small initial size (Harpending and Rogers 2000; Shen *et al.* 2000; Wooding and Rogers 2000; Alonso and Armour 2001; Rogers 2001). On the other hand, many nuclear loci have positive Tajima's *D* values so they do not provide evidence of population growth (Harding *et al.* 1997; Hey 1997; Zietkiewicz *et al.* 1997; Przeworski *et al.* 2000). Using data from the 12 nuclear loci then available, Wall and Przeworski (2000) tested several simple models of population growth and found that the different patterns among loci were not compatible with any of their models. This led to the suggestion that various forms of selection have influenced a subset of loci (Wall and Przeworski 2000; Excoffier 2002). It is also becoming clear that more complex models of human demography must be considered, such as those incorporating geographic structure and changes in population size (Pluzhnikov *et al.* 2002; Ptak and Przeworski 2002).

One of the major challenges for interpreting the contrasting patterns observed among human loci comes from the sampling strategies used by different investigators. Studies of nuclear sequence variation vary greatly with regard to the scheme for sampling populations (from population-based to global "grid" sampling), the type of genomic regions studied (from coding to noncoding), and the molecular methods of variation detection employed (Przeworski *et al.* 2000; Ptak and Przeworski 2002). This diverse array of strategies has made it difficult to compare results across studies. Recently, several surveys have sampled DNA sequences from multiple loci in a common set of individuals (Frisse *et al.* 2001; Harris and Hey 2001; Stephens *et al.* 2001; Yu *et al.* 2002a,b; Carlson *et al.* 2003). These studies have typically focused on multiple autosomal loci from either noncoding regions exclusively (Frisse *et al.* 2001) or regions encompassing exons (Stephens *et al.* 2001). Studies of the X chromosome have typically focused on coding regions and/or only a few loci (Harris and Hey 1999, 2001; Stephens *et al.* 2001; Kitano *et al.* 2003). These studies also vary in the way humans are sampled, ranging from panels of individuals from the United States (Stephens *et al.* 2001), to panels containing many globally dispersed samples (Harris and Hey 1999, 2001), to panels with multiple individuals from a limited number of human populations (Frisse *et al.* 2001).

Here we build on a continuing study of noncoding regions on the X chromosome (Figure 1) in a panel of 41 globally sampled humans representing African and non-African populations by examining patterns of variation at four loci (*APXL*, *AMELX*, *TNFSF5*, and *RRM2P4*) and comparing these patterns with those previously reported at *DMD* (*i.e.*, introns 7 and 44; Nachman and Crowell 2000), *G6PD* and *L1CAM* (Saunders *et al.* 2002), and *MSN* and *ALAS2* (Nachman *et al.* 2004). We also include comparisons with patterns of noncoding variation seen at five additional X-linked loci that were sequenced in similar global panels: *PDHA1* (Harris and Hey 1999), *Xq13.3* (Kaessmann *et al.* 1999), *FIX* (Harris and Hey 2001), *MAO-A* (Gilad *et al.* 2002), and *Xq21.3* (Yu *et al.* 2002b). Because we studied X-linked loci only, we were able to avoid some of the complications that arise when comparing loci with different modes of inheritance and effective population sizes, such as those associated with the Y chromosome, autosomes, or the mitochondrial genome (Fay and Wu 1999; Przeworski *et al.* 2000; Hellmann *et al.* 2003). Our results indicate that, despite a common sampling strategy, there is still substantial heterogeneity in patterns of variation among loci on the human X chromosome. This degree of heterogeneity does not appear to be compatible with a simple demographic model and may reflect the effects of recent diversity-reducing selection acting on a subset of loci.

SUBJECTS AND METHODS

Subjects: Human genomic DNAs were isolated from lymphoblastoid cell lines that were established by the Y Chromosome Consortium (2002) at the New York Blood Center from blood donated by volunteers who gave informed consent. All sampling protocols were according to procedures approved by the New York Blood Center and University of Arizona Human Subjects Committees. A total of 41 men were sampled, including 10 Africans (2 Tsumkwe San from Namibia, 1 West Bantu Herero, and 1 East Bantu Pedi, 1 East Bantu Sotho, 2 Biaka Pygmies from CAR, and 3 Mbuti Pygmies from Zaire), 11 Asians (3 Han Chinese, 2 Siberian Yakuts, 1 Cambodian, 3 Japanese, and 1 Pakistani, and 1 Nasioi from Melanesia), 10 Europeans/Middle Easterners (2 Ashkenazi Jews, 1 British, 1 Adygean from Krasnodar, 3 Germans, 2 Western Russians, and 1 Turk), and 10 Native Americans (1 Navajo, 1 Tohono O'Odham, 1 Poarch Creek, 2 Karitianans, and 2 Surui from Brazil, 1 Mayan, and 2 Amerindians of unknown tribal affiliation). This sample was chosen as part of a long-term project in our labs to survey nucleotide variability at a number of loci throughout the genome using a common set of individuals (Nachman *et al.* 1998, 2004; Nachman and Crowell 2000; Saunders *et al.* 2002). A single male common chimpanzee (*Pan troglodytes*) was surveyed from DNAs provided by O. Ryder. By sequencing X chromosomes in males we were able to avoid problems associated with sequencing and scoring heterozygous sites and we were also able to recover haplotypes directly among all sites in the sample.

Choice of loci: We chose to sequence *APXL* (apical protein-like *Xenopus laevis*), *AMELX* (amelogenin, X-linked), *RRM2P4* (ribonucleotide reductase M2 polypeptide pseudogene 4), and *TNFSF5* (tumor necrosis factor ligand superfamily, member 5) because they map to telomeric regions with moderate to high rates of recombination. These loci complement our existing database of six other genes (sequenced in the same global panel) mapping to X chromosome regions with a range of recombination rates (Figure 1): *DMD* intron 7, *DMD* intron 44 (Nachman and Crowell 2000), *G6PD* and *L1CAM* (Saunders *et al.* 2002), and *MSN* and *ALAS2* (Nachman *et al.* 2004). Approximately 5 kb of intron from each gene was sequenced.

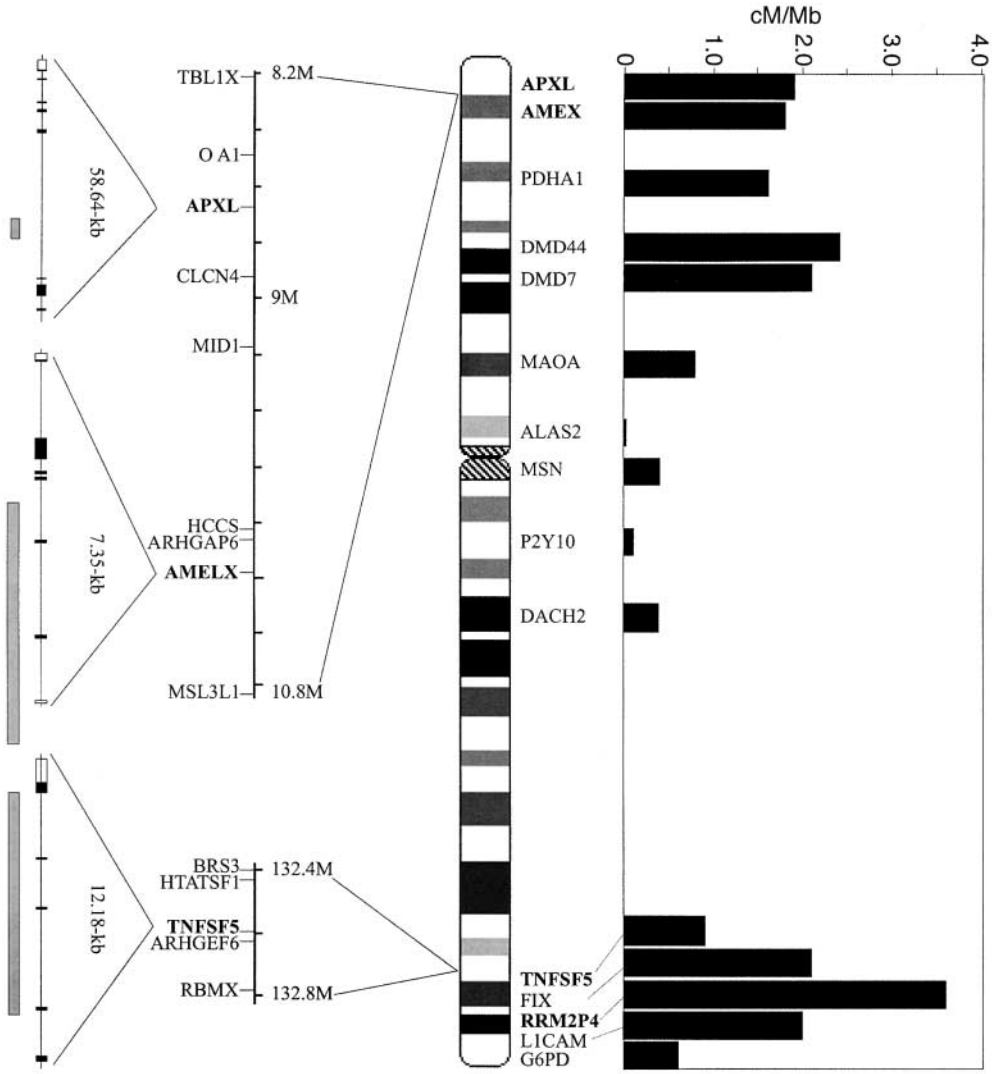

FIGURE 1.—Recombination rates and the genomic regions compared in this study. (Top) Recombination rates (centimorgans per megabase) as estimated from KONG *et al.* (2002; see SUBJECTS AND METHODS) for 15 loci in Table 2. (Middle) Schematic of the human X chromosome and approximate genomic positions of loci (four regions sequenced in this study are in boldface type). (Bottom) Schematic of three genes. Exons are marked by solid boxes and regions sequenced are indicated by shaded boxes.

With the exception of *G6PD* in Africa (SAUNDERS *et al.* 2002), none of these loci was *a priori* believed to be influenced by selective forces. In addition, five published X chromosome DNA sequence data sets were used for comparisons with the 10 loci examined in our global panel. These loci included *PDHA1* (HARRIS and HEY 1999), *FIX* (HARRIS and HEY 2001), Xq13.3 (KAESSMANN *et al.* 1999; referred to here as *P2Y10*), Xq21.1 (YU *et al.* 2002b; referred to here as *DACH2*), and *MAO-A* (GILAD *et al.* 2002). We did not include *ZFX* (JARUZELSKA *et al.* 1999) or *DYS44* (ZIETKIEWICZ *et al.* 1997) because the polymorphism data were ascertained mainly by single-strand conformation polymorphism rather than by DNA sequencing.

PCR amplification and sequencing: DNA was PCR amplified in 25-μl volumes with 40 cycles. Conditions for each of the fragments described below varied slightly and are available from the authors upon request. Amplification primers were designed from published sequences for *APXL* (AC002365), *AMELX* (AY040206), *RRM2P4* (NG_000871; HSJ169P22), and *TNFSF5* (NT_011786) and are available upon request. Internal primers (also available upon request) were used to generate overlapping sequence runs on an ABI3730 automated sequencer. Contiguous sequence that included coding and noncoding regions (4885, 5331, 5240, and 2385 bp for *APXL*, *AMELX*, *TNFSF5*, and *RRM2P4*, respectively) was assembled for each individual and aligned using the computer program Sequencher (GeneCodes). Sequences have been submitted to GenBank under accession nos. AY694820–AY694987.

Data analysis: Nucleotide diversity, π (NEI and LI 1979), WATTERSON's (1975) estimator of θ, and F_{ST} (HUDSON *et al.* 1992) were calculated using the program DNAsp 3.99 (ROZAS and ROZAS 1999), excluding insertion-deletion polymorphisms. Under neutral equilibrium conditions both π and θ estimate the neutral parameter $3N_e\mu$ for X-linked loci, where N_e is the effective population size and μ is the neutral mutation rate. To test for deviations from a neutral equilibrium frequency distribution, Tajima's *D* (TAJIMA 1989), Fu and Li's *D* with an outgroup (FU and LI 1993), and Fay and Wu's *H* (FAY and WU 2000) were also calculated using DNAsp 3.99 (*P* values were determined by 1000 replicates of Monte Carlo simulation of the coalescent process under a neutral panmictic model with no recombination). Ratios of polymorphism to divergence were compared with the expectations under a neutral model using a multilocus Hudson-Kreitman-Aguadé (HKA) test (HUDSON *et al.* 1987) with the software "HKA" (J. Hey; http://lifesci.rutgers.edu/heylab/). This program does not take account of intragenic recombination and therefore the resulting *P* values are slightly inflated (FRISSE *et al.* 2001). Divergence data were derived for each of these loci by estimating the net divergence (D_A; NEI 1987) between homologous sequences from a chimpanzee and all 41 human sequences. The times

to most recent common ancestors (TMRCAs) among sampled sequences were estimated by dividing Watterson's estimator of $3N_e\mu$ (Watterson 1975) by the locus-specific rates of neutral mutation, estimated from the interspecific divergence. We assumed a human-chimpanzee divergence of 6 million years and a 20-year human generation time. Female estimated recombination rates were taken from the University of California, Santa Cruz, web site (http://www.genome.ucsc.edu) using the July 2003 freeze of the Human Genome Project Working Draft. The recombination rates represent average rates for a window of 1 Mb around each locus estimated through a comparison of the sequence of the human X chromosome with the deCODE Genetics map (Kong *et al.* 2002), which is based on 5136 microsatellite markers in 146 families with a total of 1257 meioses.

RESULTS

Patterns of variation at four telomeric loci: Polymorphic sites within the *APXL*, *AMELX*, and *TNFSF5* introns, as well as in the *RRM2P4* pseudogene, are shown in Table 1. Numbers of segregating sites, nucleotide diversity, measures of the frequency distribution, levels of divergence, TMRCA, and F_{ST} values are summarized in Table 2. The number of segregating sites ranges from 13 to 19 for the four loci. The average nucleotide diversity for the three gene regions (consisting mainly of introns) is slightly lower than that in the pseudogene (Table 2). The average level of Homo-Pan divergence is also lower at two of the three genes compared with the pseudogene (Table 2). A four-locus HKA test does not reject the null model ($P = 0.64$). Tajima's D (TD) and Fu and Li's D (FLD) values are negative for all three gene regions; however, none is statistically significant. *RRM2P4* is the only locus with a positive FLD value (+0.55) indicating a low proportion of singletons. Interestingly, *TNFSF5* has an excess of high-frequency-derived polymorphisms as reflected in the statistically significant negative H value (Table 2).

Global patterns of variation at 10 X-linked loci sampled in the same individuals: Global nucleotide diversity levels exhibit a large range of values—from a high value of 0.00143 at *DMD44* to a low value of 0.00035 at *ALAS2* (Table 2). However, a single 10-locus HKA test does not reject the null model of equal rates of molecular evolution ($P = 0.53$; data not shown). Interestingly, all 10 loci surveyed in our global panel have negative TD values. The mean global TD and FLD values for these 10 loci are −0.759 and −1.135, respectively. The average negative value of TD and FLD does not reveal the great extent to which frequency spectra vary among loci. Two loci have TD and FLD values close to zero (*DMD44* and *RRM2P4*), while three loci (*ALAS2*, *DMD7*, and *L1CAM*) have statistically significant negative TD values. There is an excess over neutral expectations of singleton polymorphisms at 2 of these 10 loci: *DMD7* and *L1CAM* as indicated by the significantly negative FLD values (Table 2).

Comparisons with patterns of variation at other X-linked loci: Summary statistics for the 10 loci sampled in the same panel of 41 individuals are compared with those from five additional X-linked loci surveyed in global samples in Table 2. Average global levels of variation at these five loci (average $\theta = 0.00076$ and $\pi = 0.00068$) are very similar to the 10 others in Table 2 (Mann-Whitney test, $P = 0.668$ and 0.951, respectively), as are summaries of the frequency spectra (average TD = −0.552, FLD = −1.401; Mann-Whitney test, $P = 0.582$ and 0.951, respectively). Heterogeneity among loci is also apparent: levels of variation at *FIX*, *MAO-A*, and *P2Y10* are low, while those at *DACH2* and *PDHA1* are average and high, respectively. *P2Y10*, *FIX*, and *DACH2* show an excess of rare and/or singleton polymorphisms (Table 2). In contrast, *PDHA1* and *MAO-A* have positive TD values. As expected given the variation in levels of polymorphism, estimates of the TMRCA also vary considerably among all 15 loci in Table 2. For example, the TMRCA for *FIX* and *MAO-A* is <500 KY, while 6 loci have TMRCAs >1 MY.

Kitano *et al.* (2003) recently surveyed sequence variation at 10 X-linked genes that contain mutations known to cause mental retardation. Global patterns of intron variability within these genes are similar to those reported in Table 2, although the average global level of diversity at their 10 loci ($\theta = 0.00051$ and $\pi = 0.00035$) is ~40% lower than that for the 15 loci in Table 2 ($\theta = 0.00081$ and $\pi = 0.00061$). Their sequences also had an ~30% reduction in human-chimpanzee divergence (0.699%) relative to the 15 X-linked loci in Table 2 (1.01%) and a lower mean TMRCA (474 *vs.* 1004 KY, respectively), possibly reflecting higher levels of selective constraint on loci involved in human cognitive function (Kitano *et al.* 2003).

Nucleotide diversity and recombination rate: Global nucleotide diversity (θ) and recombination rate for the 10 loci sampled in the $N = 41$ panel are positively correlated (Pearson linear correlation, two-tailed t-test, $R^2 = 0.648$, $P = 0.005$). When all 15 loci in Table 2 are considered, there is still a positive but weaker correlation ($R^2 = 0.315$, $P = 0.029$; Figure 2A). This increased scatter in the larger sample may reflect variation in levels of diversity that is attributable to different sampling strategies. Similarly, in *Drosophila melanogaster*, the correlation between diversity and recombination rate observed in heterogeneous samples (Begun and Aquadro 1992) is stronger when studied from multiple loci in a single sample (Aquadro *et al.* 1994). When we consider the relationship of human-chimpanzee divergence levels to human recombination rates, there is no statistically significant relationship for either the set of 10 or the set of 15 loci in Table 2 ($R^2 = 0.073$, $P = 0.449$; $R^2 = 0.082$, $P = 0.303$, respectively), which may reflect the small number of loci investigated (Figure 2B).

African and non-African levels of diversity: Table 3 summarizes the numbers of segregating sites, nucleotide diversity, measures of the frequency distribution, and TMRCAs within African and non-African samples. Consistent with many other studies (*e.g.*, Yu *et al.* 2002a), most of the 15 X-linked loci exhibit a pattern of reduced

TABLE 1

Polymorphic sites at four X-linked genes

				APXL	*AMELX*	*TNFSF5*	*RRM2P4*
				0 0 0 0 0 0 0 1 1 2 2 2 2 2 2 2 3 4 4	0 0 0 1 1 1 1 1 1 2 2 2 2 3 3 3 3	0 0 1 1 1 2 3 3 3 4 4 4 5 5 5 5	0 0 0 1 1 1 1 1 1 1 1 2 2 2
				0 2 3 4 5 7 8 6 7 0 1 3 5 5 6 8 9 0 4	2 2 8 0 3 5 7 8 9 0 3 4 9 0 0 9 9	6 9 3 9 9 0 4 6 9 1 4 4 0 0 1 2	0 3 8 0 2 3 3 6 6 7 9 0 2 3
				2 2 9 8 0 6 7 7 1 7 1 4 0 3 2 0 2 9 1	5 7 7 7 5 5 0 1 4 1 7 9 3 0 8 8 8	7 3 0 1 4 7 8 5 9 1 0 3 1 8 8 2	1 9 4 8 6 4 4 3 3 3 5 2 6 0
				2 1 2 5 0 9 3 5 9 3 9 7 8 8 0 2 4 9 2	0 5 6 5 0 2 6 8 7 1 7 0 6 7 3 4 7	1 7 2 9 9 9 2 7 6 2 6 5 6 7 0 1	2 7 7 67 1 3 4 4 9 3 1 0 3 4
Continent	Country	Ethnic/language group	YCC no.	Consensus A C G G C C C C T T G G C A G G A T C	C C G T A T A C A G C C A T T G A	C T C T G T A A T G G C G C C G	C G C G A A T G T G G T G C C
Africa	Namibia	San	38	. . . . . G . A . . . . . G	. . . . . C C C	. . . . . T . . . . C . A	
	Namibia	San	22	. . . . .	. . . . . T . . . . C C . .	T . . . . C . A A . A . T	. . . . .
	South Africa	E. Bantu	32	. . . . . T A . . . . . G	. . . . . T . . . . C C . .	. . . . . T . . . . . T . . A	
	South Africa	E. Bantu	33	. . . . . T . . . C	T . . . . A . T . . . . . C C . .	T . . . . . A . . T . . . . . C . A	
	South Africa	W. Bantu	40	. . . . . C .	T . . . . A . T . . . . . C C . .	. . . . . A . . T . . . . . C . A	
	CAR	Biaka	7	. . . . . G . . . C .	T . . . . A . T . . . . . C C . .	. . . . . C . A . . . . . A . . T -- . . . C . A	
	CAR	Biaka	6	. . . . . T . . . . . G . A	. . . . . C C A G	. . . . . C C A	C . . . . .
	Zaire	Mbuti	8	. . . . . T . . . G . .	T . . . . A . T . . . . . C C . .	. . . . . C	. . . . . C . A . . . .
	Zaire	Mbuti	65	. . . . . T . . . G	. . . . . T . . . . C C A . .	. . . . . G	T -- . . . C . A . . . .
	Zaire	Mbuti	9	. . . . . C . . . G T . G	. . . . . T . . . . C C A . .	. . . . .	. . . . .
Europe	England	English	26	G T . A . . . . . A A . . . . G	. . . . .	. . . . .	. . . . .
	Germany	German	61	G T . A . . . . . A . . . . G	. . . . . T . . . . C C A . .	. . . . .	. . . . .
	Germany	German	64	. . . . . T . . . . . G	. . . . . T . . . . C C A . .	. . . . .	. . . . .
	Germany	German	62	G T . A . . . . . A . . . . G	. . . . . T	. . . . .	. . . . .
	E. Europe	Ashkenazi	24	. . . . . T	. . . . . C	C	T -- . . . C . A . . . .
	Poland	Ashkenazi	59	G T . A . . . . . A . . . . G	. . . . . T . . . . C C A . .	. . . . .	. . . . .
	Russia	Adygean	56	. . . . . T	. . . . .	. . . . .	. . . . .
	Russia	Russian	72	G T . A . . . . . A . . . . G	G . . . . . T . . . . A . .	. . . . .	. . . . .
	Russia	Russian	71	G T . A . . . . . A . . . . G	. . . . . T . . T . C C . .	. . . . .	. . . . .
	Turkey	Turkish	79	. . . . . T	. T . . . . . T . . . . C C A . .	. . . . .	. . . . .
Asia	Japan	Japanese	78	. . A . . . . T	. . . . .	. . . . . T . . . . T . . . . . A . . A A G A A T	
	Japan	Japanese	76	. . . . .	. . . . . T . . . . C C A . .	. . . . .	. . . . .
	Japan	Japanese	77	G T . A . . . . . A . . . . G	. . . . . T	. . . . .	. . . . .
	China	S. Han	66	. . . . . T	. . . . .	. . . . . T . . . . . A . . A A G A . T	
	China	S. Han	67	. . . . . G	. . . . . T	. . . . .	. . . . .
	China	S. Han	68	. . . . . T	. . . . . T	. . . . .	. . . . .
	Cambodia	Cambodian	69	G T . A . . . . . A . . . . G	. . . . . T	. . . . .	. . . . .
	Pakistan	Pakistani	57	G T . A . . . . . A . . . . G	. . . . . T	A . . . . . T . . G . . C . A	
	Siberia	Yakut	49	. . . . . T	A	. . . . .	. . . . .
	Siberia	Yakut	51	. . . . .	A . . . . . T	A . . C . A	. . . . .
	Melanesia	Nasioi	10	. . . . T . . . . . G	. . . . . T	A . GC . A . . . . . T . . . . . A . . A A G A A T	
Americas	U.S.A.	Poarch Creek	27	. . . . . T	. . . . .	C	. . . . .
	U.S.A.	Tohono O'Odham	25	. . . . . T	. . . . .	A . GC . A . . . . . T . . G . . C . A	
	U.S.A.	Navajo	23	G T . A . . . . . A . . . . G . . . . A . . . . . T . . . . C C A . .	. . . . .	. . . . . T . . G . . C . A	
	U.S.A.	Amerindian	2	. . . . . T	. . . . . T . . . . C C A . .	. . . . . A . GC . A . . . . . T . . G . . C . A	
	U.S.A.	Amerindian	4	. . . . . T	. . . . .	C	T . . G . . C . A . . . .
	Mexico	Mayan	17	G T . A . . . . . A . . . . G	. . . . . T	C	T . . G . . C . A . . . .
	Brazil	Karatiana	12	G T . A . . . . . A . . . . G	. . . . .	C	. . . . .
	Brazil	Karatiana	13	G T . A . . . . . A . . . . G	. . . . .	C	. . . . .
	Brazil	Surui	16	. . . . . T	. . . . .	A	T . . G . . C . A . . . .
	Brazil	Surui	14	. . . . . T	. . . . .	. . . . .	T . . G . . C . A . . . .
				Pan	. . . . . T . . . . C C A . .	. . . . . C C A A	T . . . . . A A . . . T

Polymorphisms at *APXL*, *AMELX*, *TNFSF5*, and *RRM2P4* among 41 humans. The consensus sequence is shown at the top, and the chimpanzee sequence is shown at the bottom.

nucleotide diversity in non-Africans relative to Africans. Two loci, *RRM2P4* and *MSN*, show the opposite pattern (Table 3). Mean levels of nucleotide diversity are almost twice as high in Africans ($\theta = 0.00082 \pm 0.00040$, $\pi = 0.00076 \pm 0.00049$) as in non-Africans ($\theta = 0.00045 \pm 0.00036$, $\pi = 0.00040 \pm 0.00042$). These differences are statistically significant (Mann-Whitney test, $P = 0.002$ and 0.004, respectively).

To test for relationships between polymorphism and divergence, we performed a single 15-locus HKA test on the entire global sample, as well as on Africans and non-Africans separately (Table 4). We found that the null model was rejected in non-Africans only ($P = 0.007$). When we repeated the HKA tests using only the 10 loci sequenced in the panel of 41 individuals, similar results obtained ($P = 0.030$; data not shown). We also note that the correlation between recombination rate and nucleotide diversity was stronger in non-Africans ($R^2 = 0.521$, $P = 0.019$) compared with Africans ($R^2 = 0.308$, $P = 0.096$; data not shown).

TABLE 2

Summary statistics for 15 X-linked loci in worldwide samples

Locus	Base pairs	*N*	*S*	θ (%)	π (%)	TD	FLD	FWH	*D* (%)[a]	TMRCA[b]	F_{ST}	Reference
APXL	4,638	41	19	0.096	0.080	−0.541	−1.170	0.124	1.39	840	0.226	This study
AMELX	5,331	41	17	0.075	0.061	−0.601	−1.759	−0.282	0.80	1,178	0.343	This study
TNFSF5	5,239	41	16	0.071	0.035	−1.634	−1.488	−4.987*	0.64	1,340	0.062	This study
RRM2P4	2,385	41	13	0.127	0.119	−0.206	0.548	−3.577	0.89	1,714	0.079	This study
ALAS2	4,742	41	7	0.035	0.015	−1.527*	−1.873	0.659	0.55	656	0.305	Nachman *et al.* (2004)
MSN	4,576	41	9	0.046	0.035	−0.662	−1.956	−2.046	0.91	605	0.636	Nachman *et al.* (2004)
Dmd44	2,959	41	19	0.150	0.143	−0.160	−0.332	−2.734	0.93	1,938	0.043	Nachman and Crowell (2000)
Dmd7	2,383	41	9	0.088	0.034	−1.778*	−3.306*	−0.905	1.83	576	0.275	Nachman and Crowell (2000)
L1CAM	2,087	41	6	0.067	0.018	−1.925*	−2.264*	0.371	1.07	762	0.061	Saunders *et al.* (2002)
G6PD	2,918	41	10	0.080	0.039	−1.512	−1.048	0.293	1.32	692	0.296	Saunders *et al.* (2002)
PDHA1	4,069	35	24	0.143	0.180	0.886	0.916	−2.763	0.98	1,744	0.609	Harris and Hey (1999)
FIX	3,728	36	6	0.039	0.014	−1.713*	−1.327	0.508	1.00	470	0.013	Harris and Hey (2001)
MAO-A	18,820	56	39	0.038	0.043	0.429	−0.522	−8.345	1.01	449	0.125	Gilad *et al.* (2002)
P2Y10	10,163	69	33	0.068	0.033	−1.633	−3.073*	−3.219	0.91	898	0.061	Kaessmann *et al.* (1999)
DACH2	10,215	62	44	0.092	0.072	−0.731	−2.738*	−2.024	0.92	1,201	0.006	Yu *et al.* (2002b)
Average	5,617	45	18	0.081	0.061	−0.887	−1.472	−1.926	1.01	1,004	0.209	

Only SNPs were included in all analyses except the *H* test where indels were also used. Underlined loci were sequenced in same panel of 41 men. *Two-tailed $P \leq 0.025$ in Monte Carlo simluation.

[a] Human-chimpanzee sequence divergence.

[b] TMRCA values are based on Watterson's (1975) estimator of $3N_e\mu$ and are in units of 1000 years (KYA).

African and non-African frequency spectra: When we estimate TD and FLD values separately in African and non-African samples, both estimators are less negative than those in the global panel (see above): mean African TD and FLD = −0.401 and −0.552, respectively, and mean non-African TD and FLD = −0.512 and −0.846, respectively. TD and FLD values for all loci are consistent with neutral equilibrium expectations in Africans, while four loci showed an excess over neutral expectations of rare polymorphisms and/or singletons in non-Africans: *ALAS2*, *MSN*, *DMD7*, and *L1CAM* (Table 3). There is also an excess of high-frequency-derived polymorphisms at *MSN* and *DMD7* in non-Africans as reflected in statistically significant negative Fay and Wu's *H* (FWH) values (Table 3). African TD values ranged from −1.72 to 0.80, while those of non-Africans ranged from −2.06 to 0.72. The mean TD for non-Africans (−0.512 ± 0.906) was slightly more negative than that of Africans (−0.401 ± 0.736).

Under a model of population growth TD and FD are negatively correlated with sample size (Ptak and Przeworski 2002; Hammer *et al.* 2003). To determine whether the more negative mean TD value in our non-African sample compared with our African sample was the result of a larger mean sample size (*i.e.*, 32.3 *vs.* 12.3), we reanalyzed our non-African data by resampling each locus 100 times and making the sample size equal to the number of Africans sequenced at the locus. In other words, for the $n = 41$ data set, 10 non-Africans were resampled 100 times. Resampled data sets with no variation (*i.e.*, $S = 0$) were thrown out and an additional resampling was performed. The *FIX* locus was not resampled because equal numbers of Africans and non-Africans were surveyed initially (Harris and Hey 2001). The mean TD value for the resampled non-African data set was still more negative (TD = −0.639) despite having an identical size as the African sample. This suggests that the more negative TD value in non-Africans compared with Africans is unrelated to sample size differences.

To explore the relationship between frequency spectra and recombination rate we plotted African or non-African FLD *vs.* recombination rate for each locus in Table 3. For Africans there is no relationship either for the set of 10 loci sampled in the same set of 41 individuals or for all 15 loci ($R^2 = 0.014$, $P = 0.747$ and $R^2 = 0.014$, $P = 0.675$, respectively; Figure 3A). In contrast, non-African FLD values exhibit a statistically significant positive correlation for both the 10 loci and 15 loci

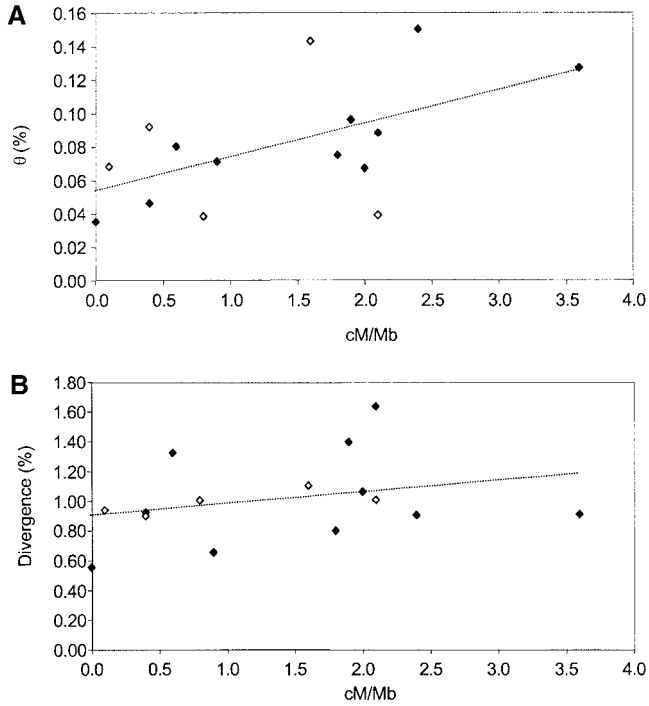

FIGURE 2.—Relationship between human nucleotide diversity or human-chimpanzee divergence and recombination rate. (A) Scatterplot of human nucleotide diversity levels (θ_W) *vs.* recombination rate for loci in Table 2. Solid diamonds represent 10 loci sequenced in same panel of $N = 41$ samples (linear regression: $R^2 = 0.648$, $P = 0.005$). Open diamonds represent 5 additional X-linked loci sequenced in similar global panel (see text; linear correlation for 15 loci: $R^2 = 0.315$, $P = 0.029$). (B) Scatterplot of human-chimpanzee divergence *vs.* recombination rate (10-locus linear regression: $R^2 = 0.073$, $P = 0.449$; 15-locus linear regression: $R^2 = 0.082$, $P = 0.303$). Trend lines are based on the sample of 15 loci.

data sets ($R^2 = 0.520$, $P = 0.019$ and $R^2 = 0.454$, 0.006, respectively; Figure 3B). Similar trends were observed with TD; however, the non-African correlation was not statistically significant ($P = 0.138$). Interestingly, there is also a positive relationship between non-African FLD and nucleotide variability (θ) for the set of 10 loci sampled in the same individuals ($R^2 = 0.678$, $P = 0.003$), while no such relationship was observed for the African FLD values ($R^2 = 0.001$, $P = 0.936$; data not shown). This relationship is only marginally statistically significant for the full set of 15 loci in non-Africans ($R^2 = 0.235$, $P = 0.067$).

DISCUSSION

We obtained DNA sequence data from four loci in regions of high recombination on the X chromosome and compared patterns of variation at these loci with those from six additional loci sequenced in the same panel of 41 global samples. Despite the fact that all loci were X-linked and sampled in the same set of individuals, we found substantial heterogeneity in levels and patterns of variation among these 10 loci. We also compared patterns at these 10 loci with those from five additional X-linked genes that were sequenced in similar global panels (HARRIS and HEY 1999, 2001; KAESSMANN *et al.* 1999; GILAD *et al.* 2002; YU *et al.* 2002b). Nucleotide diversity varies by more than an order of magnitude among loci (Table 2): *ALAS2*, *L1CAM*, and *FIX* exhibit some of the lowest levels of nucleotide diversity seen in the human genome, while *PDHA1*, *DMD44*, *RRM2P4* are higher than the autosomal average (LI and SADLER 1991; PRZEWORSKI *et al.* 2000). Likewise, the distribution of mutation frequencies differs considerably among loci, with several harboring an excess (over neutral predictions) of low-frequency polymorphisms (*e.g.*, *ALAS2*, *DMD7*, *L1CAM*, *FIX*), and others with an abundance of high-frequency (*e.g.*, *TNFSF5*) or intermediate-frequency (*e.g.*, *PDHA1*, *DMD44*, *RRM2P4*)-derived polymorphisms.

Patterns of heterogeneity seen among loci are different between African and non-African samples. There was a wide range of F_{ST} values, with two loci (*PDHA1* and *MSN*) exhibiting some of the highest known levels of differentiation among populations and others with extremely low levels of differentiation (*e.g.*, *DACH2*, *FIX*, and *DMD44*; ROMUALDI *et al.* 2002). As previously documented for autosomal, Y-linked, and mitochondrial loci (VIGILANT *et al.* 1991; PRZEWORSKI *et al.* 2000; SHEN *et al.* 2000; HAMMER *et al.* 2003), X-linked loci are more variable in sub-Saharan African populations than in non-African populations. However, the extent of the reduction in non-African diversity at X-linked loci appears to be greater than that observed on the autosomes. For example, the average non-African reduction in θ for the 15 X chromosome loci in Table 3 is 45%, while the average non-African reduction in θ on the autosomes is 30% (HALUSHKA *et al.* 1999; FRISSE *et al.* 2001; STEPHENS *et al.* 2001). We also note that there is substantial variability among X-linked loci in the degree of reduction in non-African variation, with some loci having <10% of African diversity (*e.g.*, *PDHA1*, *ALAS2*, and *L1CAM*). Similar disparities in African and non-African autosomal levels of diversity have not been reported (PRZEWORSKI *et al.* 2000; ALONSO and ARMOUR 2001; FRISSE *et al.* 2001). Mounting evidence suggests that, for many loci, African populations contain more rare alleles than non-African populations (WALL and PRZEWORSKI 2000). We found that our African sample has TD values similar (*i.e.*, slightly negative on average) to those reported in the literature for African populations (PRZEWORSKI *et al.* 2000). However, our non-African sample exhibits an unusual pattern, whereby the mean TD value is slightly more negative than the mean TD value in our African sample (−0.512 and −0.401, respectively). The more negative average TD value for non-Africans held even after subsampling 10 non-Africans to control for differences in sample sizes between Africans and non-Africans. This is driven, in part, by sharply negative TD values at *MSN* and *DMD7* and fewer loci with positive TD outside Africa (Table 3).

TABLE 3

Patterns of variation in African and non-African individuals

Locus	Geographic region	*N*	*S*	θ (%)	π (%)	TD	FLD	FWH	TMRCA[a]
APXL	Afr	10	13	0.091	0.078	−0.660	−0.476	0.533	781
	NAf	31	11	0.049	0.070	1.344	−0.479	0.869	421
AMELX	Afr	10	9	0.060	0.064	0.301	−0.410	0.356	942
	NAf	31	10	0.047	0.045	−0.103	−0.894	−2.224	738
TNFSF5	Afr	10	10	0.067	0.056	−0.726	−0.215	−3.200	1,264
	NAf	31	8	0.038	0.027	−0.910	−0.692	−1.871	717
RRM2P4	Afr	10	6	0.089	0.090	0.068	0.082	−1.244	1,201
	NAf	31	11	0.115	0.124	0.224	1.540*	−3.441	1,552
ALAS2	Afr	10	6	0.045	0.044	−0.106	−0.596	0.800	860
	NAf	31	1	0.005	0.001	−1.145*	−1.731*	0.062	102
MSN	Afr	10	4	0.031	0.035	0.566	0.372	0.356	407
	NAf	31	8	0.044	0.014	−2.057*	−2.079	−4.927*	578
Dmd44	Afr	10	14	0.167	0.175	0.212	−0.018	−0.978	2,157
	NAf	31	15	0.127	0.129	0.058	0.238	−1.024	1,640
Dmd7	Afr	10	6	0.089	0.080	−0.409	−1.723	0.711	583
	NAf	31	4	0.042	0.011	−1.889*	−2.138*	−1.684*	275
L1CAM	Afr	10	5	0.085	0.063	−1.035	−0.951	1.067	967
	NAf	31	1	0.012	0.003	−1.145	−1.731*	0.062	136
G6PD	Afr	10	8	0.097	0.082	−0.687	−0.094	1.867	792
	NAf	31	3	0.026	0.016	−0.929	−1.580	−0.879	310
PDHA1	Afr	16	22	0.163	0.195	0.802	0.727	2.417	1,988
	NAf	19	2	0.014	0.011	−0.485	−0.645	0.357	171
FIX	Afr	18	6	0.047	0.023	−1.664	−1.713	0.781	566
	NAf	18	1	0.008	0.006	−0.529	0.640	0.183	96
MAO-A	Afr	7	20	0.043	0.041	−0.374	0.434	−6.333	508
	NAf	49	29	0.035	0.042	0.717	−0.306	−8.992	414
P2Y10	Afr	23	24	0.065	0.035	−1.718	−2.102*	−0.537	845
	NAf	46	17	0.038	0.031	−0.586	−1.420	0.094	502
DACH2	Afr	20	30	0.083	0.071	−0.590	−1.603	−2.674	1,097
	NAf	42	34	0.078	0.072	−0.243	−1.413	−2.179	1,022
Average	Afr	12.3	12.1	0.081	0.075	−0.401	−0.552	−0.405	997
	NAf	32.3	10.2	0.045	0.040	−0.512	−0.846	−1.707	578

Only SNPs were included in all analyses except in the *H* test where indels were also used. *Two-tailed $P \leq 0.025$ in Monte Carlo simulation. Afr, African; NAf, non-African. Underlined loci were sequenced in the same panel of 41 men.

[a] TMRCA values are based on Watterson's (1975) estimator of $3N_e\mu$ and are in units of 1000 years (KYA).

In summary, there is substantial heterogeneity in patterns of variation among loci on the X chromosome, even when sampled in the same set of individuals. Previous observations of heterogeneity among loci have been interpreted as evidence for selection. For example, Wall and Przeworski (2000) tested whether patterns of variation observed at a number of nuclear loci (including some of those examined here) were compatible with a variety of demographic models. They found that the low TD values at some loci (including *DMD7* and *P2Y10*) and the high TD values at other loci (including *PDHA1* and *DMD44*) together were not consistent with a model of constant size or with a model of constant size followed by exponential growth. Even after incorporating more complex demographic components (such as a bottleneck or geographic structure), none of their models could account for the patterns of variation seen at all loci. To explain these contrasting patterns, they suggested that selection influenced variation at several of the loci studied.

Here we consider our observations in light of two alternative demographic models incorporating selection put forward by Wall and Przeworski (2000). The first is a model with long-term population growth (Harpending *et al.* 1998) that is expected to lead to an excess of rare variants (*i.e.*, negative TD) at all loci that are not subject to selection. Under this model, the negative TD seen at some loci reflects population growth while the positive TD values, or those close to zero, observed at other loci reflect the action of diversity-enhancing selection (Harpending and Rogers 2000; Wall and Przeworski 2000; Rogers 2001; Excoffier 2002). The second model has constant population size (*i.e.*, the onset of human population growth is too recent to leave a signature in the nuclear genome). Under this model, TD values near zero reflect constant population size

TABLE 4

Observed and expected number of polymorphic sites within humans in HKA test

	Global				Africans				Non-Africans			
Locus	*N*	Observed	Expected	Deviation[a]	*N*	Observed	Expected	Deviation[a]	*N*	Observed	Expected	Deviation[a]
APXL	41	19	19.85	0.013	10	12	13.14	0.028	31	9	11.24	0.209
AMELX	41	17	13.79	0.336	10	9	8.69	0.004	31	10	7.78	0.357
TNF	41	16	11.67	0.791	10	10	7.48	0.347	31	8	6.27	0.292
RRM2P	41	13	8.05	1.775	10	6	4.68	0.197	31	11	4.85	5.233
ALAS	41	7	8.51	0.153	10	6	6.03	0.000	31	1	4.55	1.895
MSN	41	9	11.98	0.360	10	4	7.88	0.759	31	8	7.51	0.018
DMD44	41	18	11.06	2.205	10	13	7.20	1.955	31	14	6.47	5.294
DMD7	41	9	12.33	0.430	10	6	8.47	0.274	31	4	7.14	0.802
L1CAM	41	6	6.55	0.029	10	5	4.60	0.019	31	1	3.44	1.286
G6PD	41	10	12.92	0.308	10	8	9.07	0.046	31	3	7.22	1.426
PDHA1	35	24	15.25	2.044	16	22	12.65	2.456	19	2	6.08	1.528
FIX	35	6	9.78	0.756	18	6	8.60	0.366	18	1	5.01	1.918
MAO-A	56	39	54.91	1.671	7	20	31.62	0.482	49	29	35.83	0.331
p2y10	69	33	32.22	0.006	24	24	25.01	0.11	46	17	17.88	0.017
DACH2	61	44	35.12	0.621	20	30	25.88	0.153	42	34	20.75	3.004
Degrees of freedom:				14				14				14
Chi-square value:				18.22				10.54				30.32
Probability from chi-square distribution:				0.197				0.722				0.007

[a] Deviation is (observed − expected)^2/variance. Variance is not shown. Underlined loci were sequenced in the same panel of 41 men.

while significantly negative TD values at other loci reflect the recent effects of directional selection.

Results from previous analyses of X-linked loci have been interpreted to support both models. Nachman and Crowell (2000) sampled variation at two *DMD* introns and showed that *DMD7* has much lower levels of nucleotide diversity, many more rare polymorphisms, higher levels of linkage disequilibrium, and different African *vs.* non-African patterns, compared with *DMD44*. They suggested that patterns of variation at *DMD44* are consistent with a neutral equilibrium model of molecular evolution and that those at *DMD7* were shaped by recent directional selection (especially out of Africa). Harris and Hey (2001) compared patterns of variation at *PDHA1* and *FIX* and posited that the much lower global nucleotide diversity and skew in the frequency distribution at *FIX* was the result of a history of positive directional selection, or background selection, acting at or near *FIX*. In an earlier report, Harris and Hey (1999) demonstrated that *PDHA1* had an unusual pattern of sequence variation and suggested that this locus experienced some form of diversity-reducing selection outside of Africa. Similarly, Nachman *et al.* (2004) present evidence that *MSN* and *ALAS2* have patterns of variation that reflect a history of diversity-reducing selection, with stronger effects outside of Africa.

In contrast, other authors reached very different conclusions on the basis of analyses of some of these very same loci, as well as others on the X chromosome (Harpending and Rogers 2000; Excoffier 2002). Rogers (2001) took the opposite view of Nachman and Crowell (2000) by suggesting that patterns of variation at *DMD7* reflect demography (*i.e.*, expansion of population size) while those at *DMD44* reflect a long history of balancing selection. However, it is difficult to see how a long history of balancing selection could create the patterns of variation seen at *DMD44* because there is little linkage disequilibrium among sites at *DMD44*. Wooding and Rogers (2000) argued that even though Tajima's *D* test did not reject the null hypothesis of constant population size at the *P2Y10* locus (Kaessmann *et al.* 1999), the significantly negative Fu's F_s value at this locus does support a model of a Pleistocene population expansion. Yu *et al.* (2002b) interpreted a significant Fu and Li's *D* test to indicate a population expansion signature at *DACH2*, despite a failure of Tajima's *D* and Fu's F_s tests to reject neutrality. They suggested that ancient population subdivision must be taken into account to interpret these tests properly. We note that when the African and non-African *P2Y10* and *DACH2* data sets are considered separately, an excess of rare and/or singleton alleles is found only in the African samples (Table 3). Therefore, these loci do not support the simplest model of population expansion.

Non-African levels of polymorphism reject a neutral, constant-size model by the conservative HKA test (Table 4), indicating that the variance in polymorphism among the 15 X-linked loci is too large outside of Africa. Is this

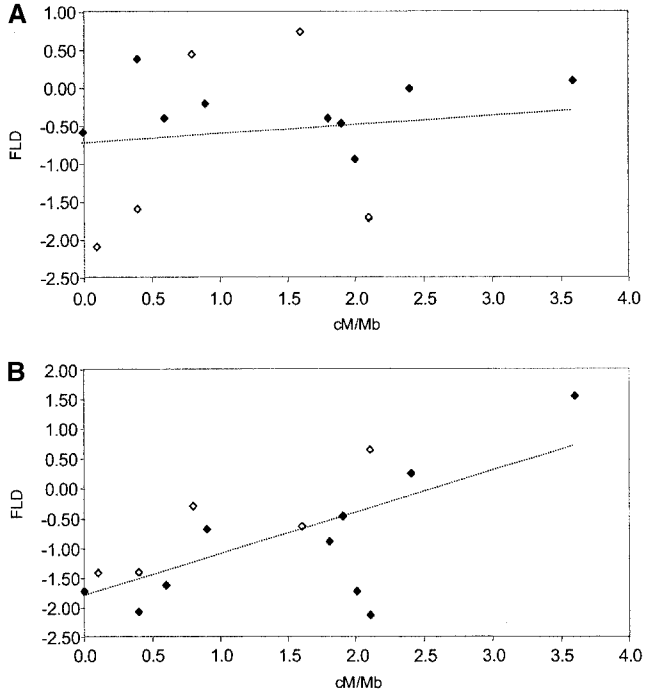

FIGURE 3.—Relationship between frequency spectra and recombination rate in Africans and non-Africans. (A) Scatterplot of FLD values in Africans *vs.* recombination rate for loci in Table 3. Solid diamonds represent 10 loci sequenced in the same panel of $N = 41$ samples (linear regression: $R^2 = 0.014$, $P = 0.747$). Open diamonds represent 5 additional X-linked loci sequenced in similar global panel (linear correlation for 15 loci: $R^2 = 0.014$, $P = 0.675$). (B) Scatterplot of FLD values in non-Africans *vs.* recombination rate (10-locus linear regression: $R^2 = 0.520$, $P = 0.019$; 15-locus linear regression: $R^2 = 0.454$, $P = 0.006$). Trend lines are based on the sample of 15 loci.

large variance in polymorphism across loci primarily the product of either natural selection acting upon a subset of loci or a history of nonequilibrium demography out of Africa (*e.g.*, a population bottleneck)? Although HUDSON *et al.* (1987) initially dismissed the possibility that a bottleneck systematically influences the HKA test, we have conducted coalescent simulations of intermediate strength bottlenecks that result in an increase in the variance of polymorphism across unlinked loci. These simulations were not intended to estimate bottleneck parameters from the data, but instead were used to examine the effects of a simple population bottleneck on the outcome of the HKA test. Similar to the bottleneck model of FAY and WU (1999), the model we examined assumes constant population size ($N = 10^4$) until 3000 generations ago, when a 40-fold bottleneck is imposed upon the population for 500 generations, after which the population reverts to its original size of 10^4 individuals (*i.e.*, the bottleneck reduces the effective population size to 250 for 10,000 years). These bottleneck parameters were chosen to (1) produce the observed reduction in non-African diversity and (2) maximize the variance in diversity among loci (data not shown). Maximizing variance among loci produces a simulated null distribution of the HKA statistic that is likely to be conservative when assessing the impact of a bottleneck on the test. We also incorporated estimates of the population recombination rate at each locus (data not shown) in the 15 locus simulations, which were replicated 1000 times. We find that our observed non-African HKA test statistic is still significantly too high ($P = 0.038$) when compared with the null distribution generated by the conservative bottleneck model (Figure 4). This result is compelling evidence that a simple population bottleneck out of Africa is insufficient to account for the increased variance in polymorphism across loci, although more complex demographic models might account for these observations.

We also observed a positive relationship between recombination rate and nucleotide diversity (Figure 2). This relationship may be caused by either positive or negative selection at linked sites (MAYNARD SMITH and HAIGH 1974; CHARLESWORTH *et al.* 1993), by variation in underlying mutation rate, or by some combination of these factors. A simple test of the idea that variation in underlying mutation rate is responsible for the correlation between nucleotide diversity and recombination rate is to compare recombination rate with interspecific divergence. Several different studies have documented a significant positive correlation between recombination rate in humans and interspecific divergence (LERCHER and HURST 2002; WATERSTON *et al.* 2002; HARDISON *et al.* 2003; HELLMANN *et al.* 2003) and, thus, it seems likely that variation in mutation rate accounts for some of the variation in nucleotide diversity. In this study we observed a significant positive correlation between nucleotide diversity and recombination rate but not between interspecific divergence and recombination rate (Figure 2), although both showed positive trends. The stronger association between nucleotide diversity and recombination rate here, compared with other studies, is noteworthy in two respects. First, many studies sample single nucleotide polymorphisms (SNPs) in a heterogeneous pool of individuals with small sample sizes instead of a common sample for all loci. For example, the average sample size for most genomic regions in the SNP consortium data that are analyzed in LERCHER and HURST (2002) and WATERSTON *et al.* (2002) is two (ALTSHULER *et al.* 2000). Here the effects of sampling can also be seen: the correlation between nucleotide diversity and recombination rate is stronger among the 10 loci sampled in the same set of individuals than among all 15 loci (Figure 2). A similar effect of sampling has been observed in *D. melanogaster* (AQUADRO *et al.* 1994). Second, we have studied X-linked loci, while all previous studies have focused mainly or exclusively on autosomal loci. One interesting possibility is that selection at linked sites may be more important on the X chromosome than on the autosomes. While the effects of background selection are expected to be weaker on the X chromosome

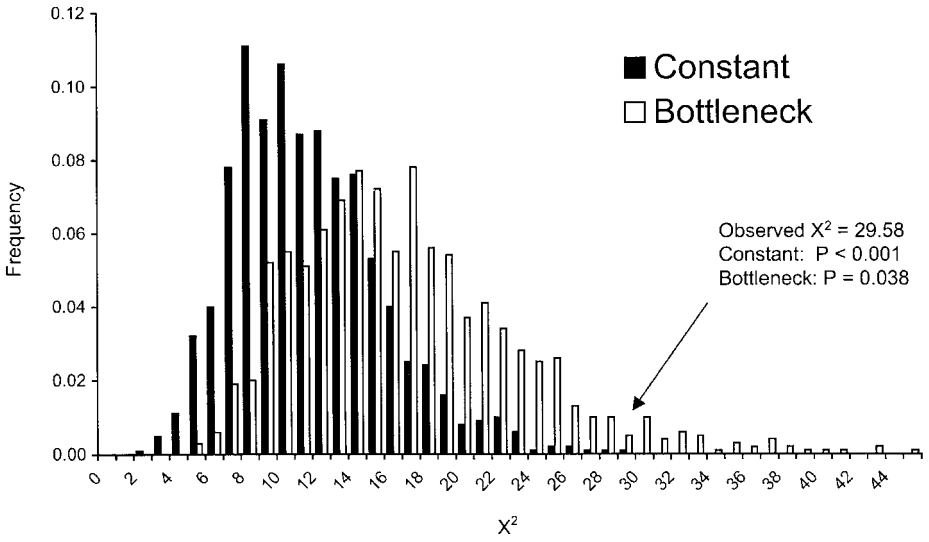

FIGURE 4.—Distribution of χ^2 values obtained from 1000 simulated 15-locus HKA tests performed on a population of constant size (solid bars) and a population experiencing a bottleneck of intermediate strength (open bars). The population recombination (ρ) and mutation parameters (θ) per locus were estimated simultaneously from the African samples using the method of FEARNHEAD and DONNELLY (2001). Using the coalescent program of Hudson (http://home.uchicago.edu/rhudson1/source/mksamples.html), the non-African bottleneck was simulated with the following parameters: (a) between species divergence: $30N$ generations (corresponding to 6 MYA, assuming $N = 10^4$ and generation time = 20 years); (b) ρ and θ per locus are estimated from the observed data; (c) the bottleneck ends $0.25N$ generations ago; and (d) the bottleneck imposes a 40-fold reduction of the ancestral African population size and lasts for $0.03N$ generations. Changes in population size are assumed to be instantaneous. The observed χ^2 value for the non-African data is indicated by an arrow.

(CHARLESWORTH *et al.* 1993), hitchhiking effects are expected to be stronger (discussed in BEGUN and WHITLEY 2000) as a consequence of either higher fixation rates (CHARLESWORTH *et al.* 1987) or shorter sojourn times (AVERY 1984).

We also found a positive relationship between recombination rate and frequency spectra in our non-African sample, but not in our African sample (Figure 3). Such a relationship is not expected under a neutral equilibrium model (PRZEWORSKI *et al.* 2001). One explanation for this observation is that diversity-reducing selective forces (*i.e.*, hitchhiking or background selection) have led to an excess over neutral expectations of singletons at loci in regions of lower recombination (CHARLESWORTH *et al.* 1993; BRAVERMAN *et al.* 1995). However, ANDOLFATTO and PRZEWORSKI (2001) demonstrated that a similar positive correlation between the summary of the frequency spectrum of polymorphic mutations (both TD and FLD) and the recombination rate in *D. melanogaster*, while expected under simple hitchhiking models, was unlikely under a model of background selection. Neither is there an expectation that diversity-enhancing selection would lead to a positive correlation between FLD and recombination rate. Moreover, the expected signature of long-term balancing selection—a peak of polymorphism surrounding a selected site—has not been observed at loci with high levels of variation and positive TD or FLD values (WALL and PRZEWORSKI 2000). As mentioned above, the correlation between recombination rate and nucleotide diversity was also stronger in non-Africans compared with Africans. The combined data suggest that positive directional selection (*i.e.*, hitchhiking) may be a more important factor influencing X chromosome variation outside Africa.

Finally, our data set included nine intronic regions within functional genes and a pseudogene, which may be less perturbed by selection than introns of genes. We chose the *RRM2P4* pseudogene in particular because it maps to a region of high recombination and low gene density on Xq27.3 and thus should provide good estimates of neutral parameters. We found that levels and patterns of variation at this pseudogene are similar to those at other loci exhibiting high levels of variation (*e.g.*, *DMD44*, *DACH2*, *AMELX*, *APXL*). This region has the third highest level of diversity, exhibits no skew in the frequency spectrum, and harbors similar levels of variation in African and non-African samples. This supports the hypothesis that similar patterns of variation at other high variation loci reflect neutral demographic processes.

If we accept that these five X-linked regions, as well as *P2Y10* and *TNFSF5*, are relatively free from the influences of natural selection, then what can we discern about human demography from patterns of variation at these loci? There is only a minor reduction in non-African diversity (*i.e.*, ~20%), a slightly negative TD in Africa, and a TD close to zero out of Africa. These data do not provide evidence for long-term population growth outside Africa. Rather, they are consistent with a larger effective population size for Africans and the possibility that non-Africans experienced a phase of population size reduction during which rare variants were lost more quickly than common variants (ZIETKIEWICZ *et al.* 1997; PRZEWORSKI *et al.* 2000). While it is possible that both diversity-reducing selection and population expansion have left signatures on X-linked loci, it is difficult to explain the heterogeneous patterns observed here by a simple model of population expansion without selection. We note that after removing the five loci showing low variation from our analyses (*ALAS2*, *MSN*,

DMD7, *L1CAM*, and *FIX*) there is still considerable heterogeneity among loci, especially in the non-African samples. More realistic models of human demography might include more complex patterns of subdivision and population size changes (Pluzhnikov *et al.* 2002), changing migration rates over time (Wakeley 1999) and/or low levels of admixture with archaic Homo. Finally, the unexpected finding of several X-linked loci with a putative signature of selection (Przeworski 2002) is consistent with the possibility that the colonization of novel environments by modern humans as they migrated out of Africa in the last ~50,000 years may have coincided with a burst of adaptive evolution (Payseur *et al.* 2002; Kayser *et al.* 2003; Mishmar *et al.* 2003).

Publication of this article was made possible by grants BCS-9906362 (to M.W.N. and M.F.H.) from the National Science Foundation (NSF) and GM-53566 from the National Institute of General Medical Sciences (to M.F.H.). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NSF or the National Institutes of Health.

LITERATURE CITED

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.

Altshuler, D., V. J. Pollara, C. R. Cowles, W. J. Van Etten, J. Baldwin *et al.*, 2000 An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature **407:** 513–516.

Andolfatto, P., and M. Przeworski, 2001 Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. Genetics **158:** 657–665.

Aquadro, C. F., D. J. Begun and E. C. Kindahl, 1994 Selection, recombination, and DNA polymorphism in *Drosophila*, pp. 46–55 in *Non-neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, New York.

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

Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature **356:** 519–520.

Begun, D. J., and P. Whitley, 2000 Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. Proc. Natl. Acad. Sci. USA **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 polymorphisms. Genetics **140:** 783–796.

Carlson, C. S., M. A. Eberle, M. J. Rieder, J. D. Smith, L. Kruglyak *et al.*, 2003 Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. Nat. Genet. **33:** 518–521.

Charlesworth, B., J. Coyne and N. Barton, 1987 The relative rates of evolution of sex chromosomes and autosomes. Am. Nat. **130:** 113–146.

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

Excoffier, L., 2002 Human demographic history: refining the recent African origin model. Curr. Opin. Genet. Dev. **12:** 675–682.

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.

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

Fearnhead, P., and P. Donnelly, 2001 Estimating recombination rates from population genetic data. Genetics **159:** 1299–1318.

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.

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.

Halushka, M. K., J. B. Fan, K. Bentley, L. Hsie, N. Shen *et al.*, 1999 Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat. Genet. **22:** 239–247.

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

Hammer, M. F., T. M. Karafet, A. J. Redd, H. Jarjanazi, S. Santachiara-Benerecetti *et al.*, 2001 Hierarchical patterns of global human Y-chromosome diversity. Mol. Biol. Evol. **18:** 1189–1203.

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.

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.

Harpending, H., and A. Rogers, 2000 Genetic perspectives on human origins and differentiation. Annu. Rev. Genomics Hum. Genet. **1:** 361–385.

Harpending, H. C., M. A. Batzer, M. Gurven, L. B. Jorde, A. R. Rogers *et al.*, 1998 Genetic traces of ancient demography. Proc. Natl. Acad. Sci. USA **95:** 1961–1967.

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. Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

Hudson, R. R., M. Slatkin and W. P. Maddison, 1992 Estimation of levels of gene flow from DNA sequence data. Genetics **132:** 583–589.

Jaruzelska, J., E. Zietkiewicz, M. Batzer, D. E. Cole, J. P. Moisan *et al.*, 1999 Spatial and temporal distribution of the neutral polymorphisms in the last ZFX intron: analysis of the haplotype structure and genealogy. Genetics **152:** 1091–1101.

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.

Kayser, M., S. Brauer and M. Stoneking, 2003 A genome scan to detect candidate regions influenced by local natural selection in human populations. Mol. Biol. Evol. **20:** 893–900.

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.

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.

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.

Mishmar, D., E. Ruiz-Pesini, P. Golik, V. Macaulay, A. G. Clark

et al., 2003 Natural selection shaped regional mtDNA variation in humans. Proc. Natl. Acad. Sci. USA **100:** 171–176.

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

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.

Nachman, M. W., S. L. D'Agostino, C. R. Tillquist, Z. Mobasher and M. F. Hammer, 2004 Nucleotide variation at *Msn* and *Alas2*, two genes flanking the centromere of the X chromosome in humans. Genetics **167:** 423–437.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

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.

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.

Pluzhnikov, A., A. Di Rienzo and R. R. Hudson, 2002 Inferences about human demography based on multilocus analyses of noncoding sequences. Genetics **161:** 1209–1218.

Przeworski, M., 2002 The signature of positive selection at randomly chosen loci. Genetics **160:** 1179–1189.

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

Przeworski, M., J. D. Wall and P. Andolfatto, 2001 Recombination and the frequency spectrum in Drosophila melanogaster and Drosophila simulans. Mol. Biol. Evol. **18:** 291–298.

Ptak, S. E., and M. Przeworski, 2002 Evidence for population growth 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.

Rogers, A. R., 2001 Order emerging from chaos in human evolutionary genetics. Proc. Natl. Acad. Sci. USA **98:** 779–780.

Romualdi, C., D. Balding, I. S. Nasidze, G. Risch, M. Robichaux *et al.*, 2002 Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. Genome Res. **12:** 602–612.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

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.

Shen, P., F. Wang, P. A. Underhill, C. Franco, W. H. Yang *et al.*, 2000 Population genetic implications from sequence variation in four Y chromosome genes. Proc. Natl. Acad. Sci. USA **97:** 7354–7359.

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.

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., 1999 Nonequilibrium migration in human history. Genetics **153:** 1863–1871.

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.

Wooding, S., and A. Rogers, 2000 A Pleistocene population X-plosion? Hum. Biol. **72:** 693–695.

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

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

Yu, N., Y. X. Fu and W. H. Li, 2002b DNA polymorphism in a worldwide sample of human X chromosomes. Mol. Biol. Evol. **19:** 2131–2141.

Zietkiewicz, E., V. Yotova, M. Jarnik, M. Korab-Laskowska, K. K. Kidd *et al.*, 1997 Nuclear DNA diversity in worldwide distributed human populations. Gene **205:** 161–171.

Communicating editor: L. Excoffier