DOI: 10.1534/genetics.106.060301

Nucleotide Variation and Haplotype Diversity in a 10-kb Noncoding Region in Three Continental Human Populations

Zhongming Zhao,*,† Ning Yu,‡ Yun-Xin Fu§ and Wen-Hsiung Li‡,1

**Department of Psychiatry, Virginia Commonwealth University, Richmond, Virginia 23298, †Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, Virginia 23284, ‡Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637 and §Human Genetics Center, University of Texas Health Science Center, Houston, Texas 77030*

Manuscript received May 3, 2006
Accepted for publication June 15, 2006

ABSTRACT

Noncoding regions are usually less subject to natural selection than coding regions and so may be more useful for studying human evolution. The recent surveys of worldwide DNA variation in four 10-kb noncoding regions revealed many interesting but also some incongruent patterns. Here we studied another 10-kb noncoding region, which is in 6p22. Sixty-six single-nucleotide polymorphisms were found among the 122 worldwide human sequences, resulting in 46 genotypes, from which 48 haplotypes were inferred. The distribution patterns of DNA variation, genotypes, and haplotypes suggest rapid population expansion in relatively recent times. The levels of polymorphism within human populations and divergence between humans and chimpanzees at this locus were generally similar to those for the other four noncoding regions. Fu and Li's tests rejected the neutrality assumption in the total sample and in the African sample but Tajima's test did not reject neutrality. A detailed examination of the contributions of various types of mutations to the parameters used in the neutrality tests clarified the discrepancy between these test results. The age estimates suggest a relatively young history in this region. Combining three autosomal noncoding regions, we estimated the long-term effective population size of humans to be 11,000 ± 2800 using Tajima's estimator and 17,600 ± 4700 using Watterson's estimator and the age of the most recent common ancestor to be 860,000 ± 258,000 years ago.

GENETIC variation data can be used to study genetic diversity within and between populations, trace migration and population history, and infer population genetics parameters. They are also useful for studying the mechanisms of nucleotide changes and for estimating recombination rate (Hammer *et al.* 2004). However, a population study usually requires the variation data to be obtained from a large sample and from several regions. The Human Genome Project has provided us with millions of single-nucleotide polymorphisms (SNPs), the most abundant form of genetic variation in humans (Venter *et al.* 2001); the largest SNP database contains >10 million SNPs (dbSNP, http://www.ncbi.nlm.nih.gov/SNP/). However, these data, although useful for genotype–phenotype analysis, may not be suitable for population genetics study because of the lack of sampling information and biased selection of regions and SNPs (Zhao *et al.* 2003). Similarly, the HapMap project was designed mainly for genetic association studies. Although it collected >1 million SNPs (phase I) from 269 DNA samples across the human genome, it included only common variants (*i.e.*, alleles with a frequency ≥0.05) and the variants were identified in four specific populations (Nigerians, CEPH, Chinese, and Japanese) (International HapMap Consortium 2005). Much genetic information at each specific locus has been ignored, especially those low-frequency SNPs, which represent the major part of genetic variation (*e.g.*, Yu *et al.* 2001).

Over the past decade, many large-scale investigations of DNA variation in worldwide human populations have been completed. These include surveys of a single locus (or a few loci) in the mitochondrial genome, on the Y and X chromosomes, and on autosomes. A list of these studies is given in Nachman *et al.* (2004). In addition, a few multilocus surveys have been conducted (Yu *et al.* 2002a; Kitano *et al.* 2003). These surveys have greatly enhanced our understanding of the genetic variation in human populations. However, most of these investigations have focused on the nucleotide variation in genic regions because of their functional importance and because of the availability of the DNA sequences at that time. The conclusions based on these data may not hold in general because selection and other factors may perturb the genetic variation patterns from what are expected under the Wright–Fisher neutral model. In comparison, DNA variation data from noncoding regions may more accurately trace the genetic history of humans and better reflect the effects of mutation and random drift, because noncoding regions are usually

[1] *Corresponding author:* Department of Ecology and Evolution, University of Chicago, 1101 E. 57th St., Chicago, IL 60637.
E-mail: whli@uchicago.edu

not directly subject to natural selection. So far, there have been only four global surveys of long noncoding regions: at Xq13.3 (Kaessmann *et al.* 1999), 22q11.2 (Zhao *et al.* 2000), 1q24 (Yu *et al.* 2001), and Xq21.31 (Yu *et al.* 2002b). All these four loci cover an ~10-kb region. While many common features such as excess of low-frequency variants have been observed in these loci, some differences in the level of polymorphism and in the pattern of genetic variation have also been revealed. For example, the proportions of singletons and doubletons varied among loci, which may have some impact on neutrality tests (Yu *et al.* 2001, 2002b). In particular, we found an excess of doubletons at 22q11.2, an observation that has not been found in any other regions. Moreover, the variation patterns in subpopulations varied greatly. For example, at 1q24, there was a deficiency of low-frequency variants in the African sample but an excess in the non-African sample (Yu *et al.* 2001). This pattern was not observed in the other three regions. These inconsistent observations call for the survey of additional noncoding regions, so that a prevailing pattern can be identified (Makova *et al.* 2001). Further, data from more noncoding regions may also help us establish a genomewide and worldwide neutrality standard of nucleotide diversity and investigate the origin and evolution of modern humans.

In this study, we selected an ~10-kb region located at 6p22 and obtained sequence variation data in a worldwide sample as in our previous surveys (Zhao *et al.* 2000; Yu *et al.* 2001, 2002b). This noncoding segment locates in a 1-Mb region where the recombination rate is the same as the genome average. We compared the distribution patterns of allelic variants, genotypes, and haplotypes within and between three continents and compared the new data with the data from other four noncoding regions (22q11.2, 1q24, Xq13.3, and Xq21.31). Overall, the genetic variation patterns suggest that this region is a typical noncoding region. We also used the new data to test the neutrality assumption, to infer the age of the most recent common ancestor of the sequences under study, and to discuss their implications for human evolution.

MATERIALS AND METHODS

The DNA region and human samples: We selected a 12-kb region corresponding to positions 73,015–85,128 at locus HS596H12 (GenBank accession no. AL031347). This locus was mapped on chromosome 6p22. There is no known gene found at this locus from GenBank or the UCSC Genome Browser, although there was a predicted gene/exon by the GENSCAN program. We applied the sequence directly to the GENSCAN server (http://genes.mit.edu/GENSCAN.html) and GrailEXP server (http://compbio.ornl.gov/grailexp/) and found no exon or gene predicted in the 12-kb region that we selected. Further, the nearest known genes were found to be located 927 kb upstream and 615 kb downstream of this 12-kb region, while the nearest predicted genes were located 855 kb upstream and 462 kb downstream of it. In a 1-Mb region where this 12-kb sequence is placed approximately in the middle, the average recombination rate was estimated to be 1.1 cM/Mb by the deCODE genetic map (http://genome.ucsc.edu/). This regional recombination rate is the same as the genome average (Kong *et al.* 2002). Furthermore, a search for the fine-scale recombination rate data from the HapMap project revealed no recombination rate information in this 12-kb region but a possible hotspot was found ~20 kb away (Myers *et al.* 2005). After excluding a fragment containing a poly(A) segment and a fragment containing a LINE/L2, a total of ~10,000 nucleotide sites were selected for sequencing.

A total of 61 unrelated individuals in a worldwide sample were collected, including 14 human subpopulations in three major geographic areas: 20 Africans (5 South African Bantu speakers, 1 !Kung, 2 Mbuti Pygmies, 2 Biaka Pygmies, 5 Nigerians, and 5 Kenyans), 21 Asians (6 Chinese, 3 Japanese, 5 Indians, 3 Yakuts, 2 Cambodian, and 2 Vietnamese), and 20 Europeans (5 Swedes, 2 Finns, 5 French, 5 Hungarians, and 3 Italians). One chimpanzee, one gorilla, and one orangutan were used as outgroups.

DNA sequencing and data collection: Three DNA fragments were separately amplified by PCR. Two pairs of primers were designed to amplify the fragments at the beginning and ending part covering positions 6–2211 and 9085–11,155 in the 12-kb region, respectively. The fragment in the middle part (positions 2262–8508) was amplified by an additional three pairs of primers. Touchdown PCR (Don *et al.* 1991) was performed by the conditions in Zhao *et al.* (2000). The PCR products were purified by a Wizard PCR Preps DNA purification resin kit (Promega, Madison, WI). Sequencing was run on an ABI 377XL DNA sequencer using nested primers according to the protocol of ABI Prism BigDye Terminator sequencing kits (Perkin Elmer, Norwalk, CT).

Raw trace files extracted from the sequencer were evaluated and proofread. The segmented data were automatically assembled using SeqMan in the software package DNASTAR. The assembled files were manually checked using the same program, the sequences were then aligned by MegaAlign in DNASTAR, and variant sites were identified among the aligned sequences. For data quality control, all nucleotides were sequenced with good quality at least once in both directions, and all singleton variants, whose derived alleles appeared only once in the sample, were verified by reamplifying and resequencing the region containing the variant site in both directions. No error was found. We did not verify other types of variants because we found errors occurred rarely in nonsingletons in our previous studies.

Statistical analysis: Human ancestor sequence was inferred by comparing the human sequences with their outgroup sequences according to the parsimony principle. Haplotypes were inferred by the computer program PHASE (Stephens *et al.* 2001). The genetic relationship of these inferred haplotypes was graphically displayed by the network program NETWORK version 3.1.1.1 (Bandelt *et al.* 1999).

The mutation rate per nucleotide per year (v) was estimated by $d/(2t_{div})$, where d is the averaged genetic distance between humans and another species and t_{div} is the divergence time between two species. The mutation rate per sequence per generation (μ) was calculated by vgL, where g is the generation time for humans (20 years) and L is the sequence length (base pairs). Nucleotide diversity (π) within and between populations was calculated as the average of the pairwise nucleotide difference per site between two sequences. The population parameter θ was estimated by the methods of Watterson (1975) and Tajima (1983).

The Hudson–Kreitman–Aguadé (HKA) test (Hudson *et al.* 1987) was performed using the direct mode in the DnaSP 4.10

TABLE 1

Genetic variants at 6p22 and their statistics

Population	Sample size	No. of variants[a]				π	Parameter θ		Neutrality tests[c]		
		Singleton	Doubleton	Others	Total	(%)[b]	θ_w	θ_Π	Tajima's *D*	Fu and Li's *D*	Fu and Li's *F*
All samples	128	40 (12.3)	7 (6.1)	19 (47.6)	66	0.070	12.27	7.29	−1.25 (−1.37)	−5.59** (−1.81)	−4.23** (−1.66)
Africans	40	31 (12.0)	7 (6.0)	13 (33.0)	51	0.069	11.99	7.22	−1.35 (−1.38)	−3.33** (−1.88)	−2.88** (−1.81)
Non-Africans	82	11 (6.0)	1 (3.0)	18 (21.0)	30	0.064	6.03	6.66	0.32 (−1.38)	−1.65 (−1.87)	−0.99 (−1.74)
Asians	42	7 (5.6)	1 (2.8)	16 (15.6)	24	0.061	5.58	6.31	0.42 (−1.42)	−0.47 (−1.88)	−0.16 (−1.77)
Europeans	40	8 (5.9)	1 (2.9)	16 (16.2)	25	0.067	5.88	6.95	0.59 (−1.39)	−0.67 (−1.99)	−0.23 (−1.87)

**Significant at the 1% level.

[a] The numbers of variants expected from $K = \theta a_n$ are given in parentheses.

[b] Nucleotide diversity after excluding indels.

[c] The critical values were obtained from 5000 simulated samples and are given in parentheses.

package (Rozas *et al.* 2003). The interspecific divergence was calculated by comparing the human and chimpanzee sequences.

Selective neutrality for the mutations in the region was tested by Tajima's method (Tajima 1989) and Fu and Li's methods (Fu and Li 1993). A statistical method (Yu *et al.* 2002b) based on analyzing constrained genealogies was used to estimate the age of the most recent common ancestor (MRCA) of the DNA sequences in a sample.

RESULTS

Sequence variation in the total sample: We collected 10,426 nucleotide sites in the selected region for 61 humans, one chimpanzee, one gorilla, and one orangutan. The GC content was 34.9%, considerably lower than the genome average of 40.9% (Zhao *et al.* 2003). Thus, the region we studied is GC poor.

A total of 66 single-nucleotide polymorphic sites were identified among the 61 worldwide human individuals (Table 1). Allele frequencies of the variants were skewed: 40 were observed only once (*i.e.*, singletons), 7 were observed twice (*i.e.*, doubletons), and 19 were observed more than twice (*i.e.*, "others"). Most variants are in low frequency. A comparison of the observed and expected number of variants in each allele frequency class revealed two striking features: (1) a strong excess of singletons as compared to the expected value and (2) a great deficiency of the variants observed in the frequency interval 0.02–0.10 (3 observed *vs.* 19.7 expected, supplemental Figure S1 at http://www.genetics.org/supplemental/).

Sequence variation in subpopulations: The numbers of variant sites after excluding insertions/deletions (indels) in the African, non-African, Asian, and European sequences were 51, 30, 24, and 25, respectively (Table 1). The African sequences carried more variants than the non-African sequences, even though the number of the African sequences is less than half that of the non-African sequences. The extent of singleton excess in Africans (31/12.0 = 2.6) was stronger than those in non-Africans (1.8), Asians (1.3), and Europeans (1.4). Correspondingly, the opposite pattern was observed in the category "others" (Table 1). Unlike the data at 22q11.2 (Zhao *et al.* 2000), the difference between the observed and expected doubletons seems not obvious, probably due to the small values. Furthermore, a total of 36 unique variant sites, including 29 singletons, 6 doubletons, and 1 other, were observed in the African sequences, while only 15, including 11 singletons, 2 doubletons, and 2 others, were found in the non-African sequences.

We observed 10 indels among the 122 human sequences, including 7 singletons, 1 doubleton, and 2 others. Two (singletons) of them were in a 13-bp fragment between one Yakut and the remaining individuals. A careful examination of the fragment indicated that at least one insertion and one deletion event are required to explain the present sequences. This fragment is not inside a repetitive element. Interestingly, this Yakut sample also had a similar unique variation pattern in one of the other two 10-kb noncoding regions in our previous surveys (Yu *et al.* 2001). Note that we used a total of three Yakut individuals in each of these three studies. It remains unknown whether such a pattern is common in Yakuts or just in this individual.

Comparison with variation in the 10 ENCODE regions or the Phase I HapMap SNPs: We compared the global genetic variation in the 3 10-kb noncoding regions (6p22, 22q11.2, and 1q24) with that observed in the 10 500-kb ENCODE regions or the Phase I HapMap SNPs (International HapMap Consortium 2005). We did not include 2 X-linked noncoding regions (Xq13.3 and Xq21.31) because these 10 ENCODE regions were in autosomes and because the number of variant sites in each subpopulation at the two X-linked loci is small. The comparison is summarized as follows. First, we observed a much higher proportion of rare SNPs than that in the ENCODE data. The average proportion of SNPs whose minor allele frequency (MAF) is <5% was 69% in the 3 10-kb noncoding regions. This

was compared with the 46% from the ENCODE data (supplemental Table S1 at http://www.genetics.org/supplemental/). Strikingly, singletons accounted for an average of 42% in the 3 10-kb regions, more than four times that (9%) in the ENCODE regions. Note that the sample size is different between the 3 noncoding regions (~122 chromosomes) and the ENCODE regions (96 chromosomes). Second, in the ENCODE data, 90% of the heterozygous sites in each individual were due to common SNPs. The opposite was observed in the 3 10-kb regions. Third, similar to the ENCODE data, we observed a trend of more rare SNPs in Africans than in Asians and Europeans. However, this pattern varied among the loci (supplemental Table S2 at http://www.genetics.org/supplemental/). For example, at 1q24 the number of SNPs whose derived allele frequency is <10% was very similar for the three subpopulations. Fourth, among those SNPs identified at the 3 noncoding loci, the proportions of the SNPs that are polymorphic in the African, Asian, and European samples were averaged to be 72, 42, and 40%, respectively (supplemental Table S3 at http://www.genetics.org/supplemental/). These are lower than the corresponding values, *i.e.*, 85% in YRI, 75% in CHB + JPT, and 79% in CEU, from the Phase I HapMap data. Finally, among the 180 SNPs identified in the 3 regions, we found 2 sites that were fixed between the Asian and European samples. Among the 1 million Phase I HapMap SNPs, there were 37 SNPs fixed in any two of the three sample panels, including 21 between CEU and CHB + JPT. It appears that the fixed SNPs were more frequently observed in the 3 noncoding regions, in which a more diverse sample was used. Moreover, we observed 2 sites fixed in the African sample, 2 sites fixed in the Asian sample, and 2 sites fixed in the European sample.

Nucleotide diversity: The average pairwise nucleotide difference (π) was 0.070% among all sequences, 0.069% for the African sequences, 0.064% for the non-African sequences, and 0.076% between the African and non-African sequences, respectively, after excluding the indels (Table 1). The nucleotide diversity was 0.061% among the Asian sequences and 0.067% among the European sequences.

Genotype and haplotype distribution: After excluding indels, we observed 46 genotypes among 61 human individuals. The numbers of genotypes were 19, 16, and 14 in the Africans, the Asians, and the Europeans, respectively. Only 6 genotypes were observed more than once in the entire sample. The most frequent genotype was observed eight times, all in non-Africans. No genotype was shared between Africans and non-Africans, although 3 genotypes were shared between Asians and Europeans.

The nonsharing feature was similarly observed in the other two autosomal noncoding regions (supplemental Table S4 at http://www.genetics.org/supplemental/). There were 53 genotypes observed at 22q11.2; however, no genotype was shared between two subpopulations. Among the 40 genotypes identified at 1q24, only 4 were shared between Asians and Europeans.

Forty-eight haplotypes could be inferred using PHASE. Most of these haplotypes were observed only once (35) or twice (4) in the sample (Table 2). The most frequent haplotype was present 32 times, all in non-Africans (15 in Asians and 17 in Europeans). There were 27 different haplotypes in Africans, more than that in Asians (18) or Europeans (13). Since the sample size is nearly the same in the three subpopulations, the large number of haplotypes in Africans reflects a higher genetic diversity. The proportion of haplotypes unique in Africans (85%) or in non-Africans (84%) was remarkably higher than that shared between them (16%). In contrast to genotypes, there were 2 haplotypes shared by Africans and Asians and 4 haplotypes shared by Africans and Europeans.

The haplotype distribution pattern observed above was consistently found at the other four noncoding loci (Table 3). When the five loci were combined, the proportions of haplotypes unique in Africans, non-Africans, Asians, and Europeans were 82, 84, 60, and 53%, respectively. Among these five loci, more haplotypes were shared at Xq13.3.

Figure 1 shows the genetic network of these inferred haplotypes at 6p22. One-third of the African-unique haplotypes (nos. 2, 4, 6, 7, 8, 9, 18, 24, and 25) could be linked to a node that connects to the human ancestral haplotype but only two non-African-unique haplotypes (nos. 36 and 42) could be directly linked to that node. The most frequent haplotype (no. 28) was genetically distant from the ancestral haplotype. Interestingly, all six haplotypes that directly link to haplotype 28 were observed only in non-Africans, a signature of the strong recent population expansion. Some haplotypes (*e.g.*, nos. 2, 44, and 46), as indicated by reticulation, were ambiguously placed in the network. This may reflect possible recombinations at this locus (see MATERIALS AND METHODS) or inaccuracy of haplotype inference.

Mutation rate, parameter θ, and effective population size *N*: The average numbers of nucleotide substitutions were 119.9 between human and chimpanzee sequences, 159.4 between human and gorilla sequences, and 332.8 between human and orangutan sequences. The mutation rate was estimated to be 0.99×10^{-9}/nucleotide/year by using a divergence time of 6 million years (MY) between human and chimpanzee; this corresponds to 2.01×10^{-4}/sequence/generation. The mutation rates were estimated to be 2.01×10^{-4} and 2.43×10^{-4}/sequence/generation using the divergence times of 8 MY between human and gorilla and 14 MY between human and orangutan.

Several methods were used to estimate θ. For all sequences, the θ-values were 7.29 by TAJIMA's (1983) method and 12.27 by WATTERSON's (1975) method

TABLE 2

Haplotype distribution at 6p22

	Polymorphic site position	Frequency				
	111111111 111111122222222233334444455555555566677777888888889999990000000000 011246902345668023402566012355669678167788012235679046788001123357 137858592930486321219729170656899069926609171469675542249140743553 345087065329582549755915485091085720230198145656744283884081114579	Total	Af.	Non-Af.	Eu.	As.
Ancestor:	CATAAAATTACGCTCACCAGGAAATTTCGTCTACACCCCAATGCGCTGTATCTAATGTGAGAAAGC					
Orangutan:	.....Y.....T.....					
Gorilla:	.....T.....G...					
Chimpanzee:	.....					
Haplotype						
1	...G.....G......A.....CGTG..T.....A....C.....G..T	1	1			
2	....G.....G......G.....G.....G...	1	1			
3	.G..GG.....TG.....G.G...TG.....A...G...	1	1			
4	.....G......G.....G...	4	4			
5	...G.....G.....GTG..T.....A.....G..T	7	2	5	1	4
6	.....G..G...C.....G......G.....G..G...	1	1			
7	.....C.....G....T....T...C..G.....G...	1	1			
8	.....C...G.....G......G.....G...	1	1			
9	..C.....G....T.....A.G...	2	2			
10	...G......G.....T.....G......G.....G...	4	4			
11	...G.....A.....C..GTG..T.....A..C.....G..T	2	2			
12	...G.....GTG..T.....A.....G..T	1	1			
13	...GG.....T.....G......G.....G...	1	1			
14	...G.....GTG..T.....A..C.....G..T	4	3	1	1	
15	...G....C.....GTG..T.....A.....G..T	1	1			
16	....G.....T.....G..A...G.G.....G....G...	1	1			
17	T..GG....G...C.....G.G.T.....G.A.....G..T	1	1			
18	.....G.....T.....G.....G......G.....G.....G...	1	1			
19	.....T.....G......G.....G...	4	2	2	2	
20	...G.....G.....G.G..T.....GG...	1	1			
21	....G.....T...A.....G......G.....G...	2	2			
22	.....G.....T.....G......G.....G...	1	1			
23	..CG.....G.....G.G..T......T.....G..T	1	1			
24	.....A.....G....T.....AC.....G.T.	1	1			
25	.....A.....G......G......G.....AC.....G...	1	1			
26	...G.....G.....G.G..T.....A.....G..G..T	1	1			
27	.....T.....G......G.....GT..	4	1	3	1	2
28	...G.....G......A.....TG..T.....A.....G..T	32		32	17	15
29	..C.....A...G....T.....A.G...	12		12	9	3
30	...G.....A.....GTG..T.....A..C.....G..T	1		1	1	
31	...G.....G.....G....GT.....A.....G...	1		1	1	
32	..C.....G.....A...G....T.....A.G...	1		1	1	
33	....G.....T.....G......G.....G...	8		8	3	5
34	..C.....A.G.G....T.....A.G...	1		1	1	
35	...G.....G......A.....TG..T.....T...A.....G..T	2		2	1	1
36	.....T.....T.....G......G.....	1		1	1	
37	...G.....G......A.....TG..T.....A......G.....G..T	1		1		1
38	...G.....G....T.....A..C.....G..T	1		1		1
39	....G.....T.....A...G......G.....G...	1		1		1
40	...G.....G.....GTG....G.....G...	1		1		1
41	...G.....G......A.....TG..T.....A.....G.....G..T	1		1		1
42	.....T.....G......G.....	1		1		1
43	...G.....G......A.....TG..T.....A...T.....G..T	1		1		1
44	....G....G.....T.....G......G.....G...	1		1		1
45	...G.....G.....TG..T.....A.....G..T	1		1		1
46	...G.....G.....G.G..T.....A.....G...	1		1		1
47	....G.C.....T.....G......G.....G...	1		1		1
48	...G.....G......A.....T.....TG..T.....A.....G..T	1		1		1
No. chr		122	40	82	40	42

The frequencies of each haplotype in the total, African (Af.), Non-African (Non-Af.), European (Eu.), and Asian (As.) samples are shown in the right columns.

TABLE 3

Distribution of haplotypes at five noncoding loci

		Unique in subpopulation				Shared between/among subpopulations				
Locus	Total	Af.	Non-Af.	As.	Eu.	Af./non-Af.	Af./As.	Af./Eu.	As./Eu.	Af./As./Eu.
6p22	48	23 (27)	21 (25)	12 (18)	5 (13)	4	2	4	6	2
22q11.2	52	19 (24)	28 (33)	9 (19)	12 (22)	5	3	3	8	1
1q24	36	17 (19)	17 (19)	9 (12)	7 (10)	2	2	2	3	2
Xq13.3	17	7 (11)	6 (10)	3 (6)	2 (7)	4	2	4	3	2
Xq21.31	23	9 (10)	13 (14)	4 (7)	7 (10)	1	1	1	3	1
Combined	176	75 (91)	85 (101)	37 (62)	33 (62)	16	10	14	23	8

Af., Africans; Non-Af., Non-Africans; As., Asians; and Eu., Europeans. The Oceanian sample was excluded in the 22q11.2 and Xq13.3 data sets. The total number of haplotypes in each subpopulation is in parentheses.

(Table 1). Watterson's estimator yielded a larger θ-value mainly because of an excess of singletons. To avoid such a large effect, the singletons were excluded and Watterson's estimator became 4.84.

For an autosomal region, the effective population size may be estimated by $N = \theta/4\mu$. For the following estimates, we used the mutation rate estimated by the divergence time of 6 MY between human and chimpanzee. For the total sample, the *N* was estimated to be 9100 and 15,300 using Tajima's and Watterson's estimators, respectively. The effective size for Africans was estimated to be 9000 and 14,900 by Tajima's and Watterson's estimators, close to that for the entire sample. The *N*-value was estimated lower in the non-African population: 8300 by Tajima's estimator and 7500 by Watterson's estimator.

Next, when the present locus was combined with two other autosomal noncoding loci 22q11.2 and 1q24 (Zhao *et al.* 2000; Yu *et al.* 2001), the *N*-value was averaged to be 11,000 ± 2800 and 17,600 ± 4700 using Tajima's and Watterson's estimators, respectively. Moreover, the *N*-value was slightly higher when these three autosomal loci were combined with an additional two X-linked noncoding loci Xq13.3 and Xq21.31 (Kaessmann *et al.* 1999; Yu *et al.* 2002b) (data not shown). These estimates suggest that the long-term

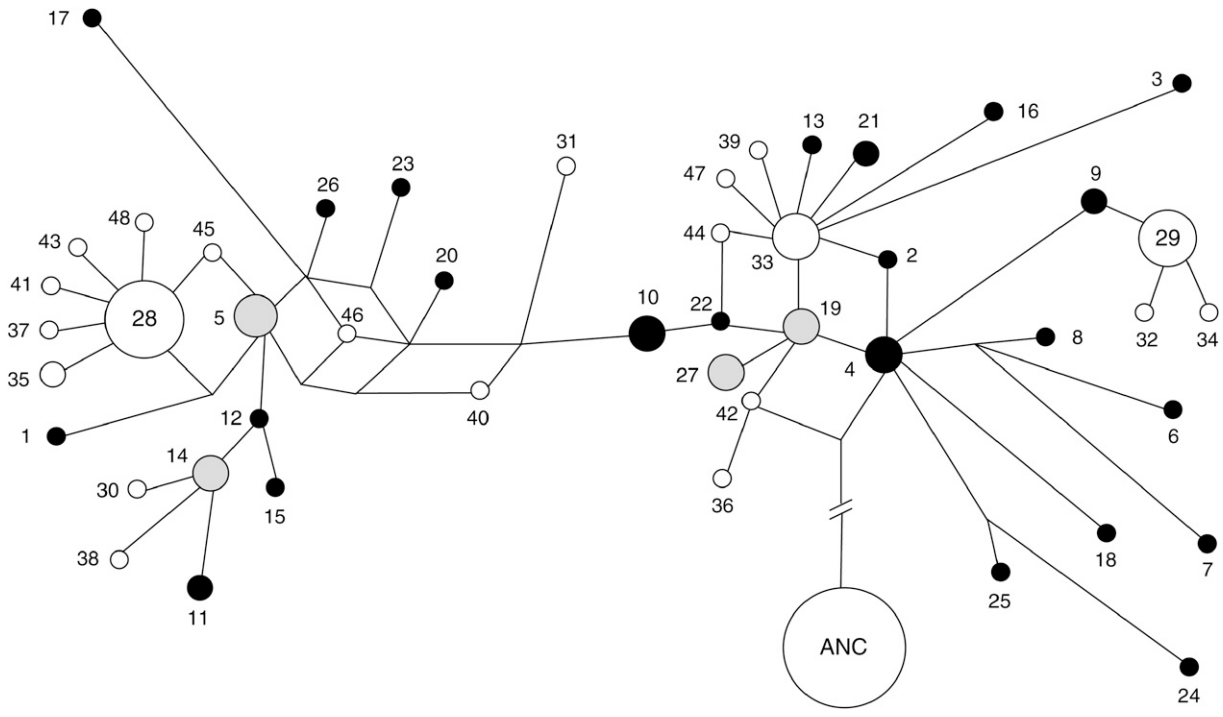

Figure 1.—Network of the inferred haplotypes. The network was obtained by the median-joining algorithm in program NETWORK 3.1.1.1 (Bandelt *et al.* 1999). The number for each circle denotes the haplotype type inferred in each sample set. ANC denotes the human ancestral haplotype. Node areas are proportional to haplotype frequencies and branch length is proportional to the number of mutations. The solid, open, and shaded circles represent the African haplotypes, the non-African haplotypes, and the shared haplotypes, respectively.

TABLE 4

Contribution of each size class of variants to the θ-values

			θ_w			θ_Π		
	Size class	Frequency	Obs.	Exp.	Difference	Obs.	Exp.	Difference
All	Size 1	40	7.44	2.28	5.16	0.66	0.12	0.54
	Size 1–10%	10	1.86	4.80	−2.94	0.46	1.25	−0.79
	Size 10–25%	4	0.74	2.04	−1.30	0.94	1.79	−0.85
	Size 25–75%	11	2.05	2.51	−0.46	5.19	3.67	1.52
	Size 75–99%	1	0.19	0.65	−0.46	0.03	0.46	−0.43
	Total	66	12.27	12.27		7.29	7.29	
Africans	Size 1	31	7.29	2.82	4.47	1.55	0.36	1.19
	Size 1–10%	9	2.12	2.35	−0.23	0.97	0.69	0.28
	Size 10–25%	2	0.47	2.81	−2.34	0.52	1.86	−1.34
	Size 25–75%	9	2.12	3.19	−1.07	4.18	3.80	0.38
	Size 75–99%	0	0	0.82	−0.82	0	0.51	−0.51
	Total	51	11.99	11.99		7.22	7.22	
Asians	Size 1	7	1.63	1.30	0.33	0.33	0.30	0.03
	Size 1–10%	4	0.93	1.40	−0.47	0.54	0.86	−0.32
	Size 10–25%	1	0.23	1.10	−0.87	0.32	1.52	−1.20
	Size 25–75%	11	2.56	1.42	1.14	5.07	3.23	1.84
	Size 75–99%	1	0.23	0.36	−0.13	0.05	0.40	−0.35
	Total	24	5.58	5.58		6.31	6.31	
Europeans	Size 1	8	1.88	1.38	0.50	0.40	0.35	0.05
	Size 1–10%	2	0.47	1.15	−0.68	0.24	0.67	−0.43
	Size 10–25%	2	0.47	1.38	−0.91	0.62	1.79	−1.17
	Size 25–75%	11	2.59	1.57	1.02	5.31	3.65	1.66
	Size 75–99%	2	0.47	0.40	0.07	0.38	0.49	−0.11
	Total	25	5.88	5.88		6.95	6.95	

The observed θ_w and θ_Π were obtained by Watterson's $\theta_w = K/a_n$ and Tajima's $\theta_\Pi = \sum\Pi_{ij}/(n(n-1)/2)$, respectively. The expected θ_w and θ_Π were obtained from $\theta_w = K/a_n$, where K is the expected variant sites in each class, and $\theta_\Pi = \theta\sum(n-i)/(n(n-1)/2)$, respectively.

effective population size of humans may be slightly higher than the commonly accepted size of 10,000.

Tests of selective neutrality: Tajima's *D* (Tajima 1989) and Fu and Li's *D* and *F* (Fu and Li 1993) methods were used to test the null hypothesis that the variations identified in the region are selectively neutral under the Wright–Fisher model with a constant population size. Tajima's test could not reject the neutrality assumption in the samples (Table 1). Fu and Li's tests were not significant in Asians, Europeans, or the combined sample of non-Africans, but were highly significant in the total sample and in Africans.

To understand why Tajima's test and Fu and Li's tests gave different results for the total and African samples, we dissected the variants into classes according to their frequencies. This is to measure the specific contributions of the various sizes of the variants to the θ-values used in the tests. For the purpose of comparison, we grouped the variants into six classes as done in Yu *et al.* (2002b). The singletons were grouped as one class (size 1) due to the strong impact on the test. Table 4 shows the contribution of each class of variants. For both the total sample and the African sample, singletons contributed ~61% (*e.g.*, 7.44/12.27) to θ_w, although they are expected to contribute only ~19% for the total sample and ~24% for the African sample. This significantly inflated the values of external mutations in Fu and Li's tests. As a result, Fu and Li's *D*- and *F*-values were strongly negative and significant. On the other hand, there was an excess of the intermediate-frequency variants and a deficiency of low- and high-frequency variants (not including the singleton class) for the total sample and each subsample. Since the singleton class inflates θ_w and the intermediate class inflates θ_Π, it results in a small difference between θ_Π and θ_w, which were used in Tajima's test. Subsequently, this small difference resulted in the failure of Tajima's test. Finally, for the Asian, the European, or the combined samples, the extent of the difference was moderate, which resulted in the failure of all tests (Table 4, data not shown for non-Africans).

We performed a similar analysis for the other four ~10-kb noncoding regions: 22q11.2 (Zhao *et al.* 2000), 1q24 (Yu *et al.* 2001), Xq13.3 (Kaessmann *et al.* 1999),

TABLE 5

Results of the HKA test

Population	Locus comparison	χ^2	P
All	6p22–22q11.2	0.01	0.91
	6p22–1q24	1.16	0.28
	6p22–Xq13.3	0.03	0.86
	6p22–Xq21.31	1.03	0.31
Africans	6p22–22q11.2	0.04	0.84
	6p22–1q24	0.19	0.66
	6p22–Xq13.3	0.04	0.84
	6p22–Xq21.31	0.50	0.48
Non-Africans	6p22–22q11.2	0.24	0.62
	6p22–1q24	2.49	0.12
	6p22–Xq13.3	0.02	0.88
	6p22–Xq21.31	4.77	0.03

TABLE 6

The estimated age ($\times 10^3$ years) of the MRCA for the human sequences sampled

Population	N	T_{mode}	T_{mean}	95% interval
All	10,000	581	601	~365–1043
	12,000	506	601	~378–893
	15,000	576	615	~400–904
Africans	8,000	499	636	~378–1116
	10,000	544	635	~379–1056
Non-Africans	6,000	397	401	~243–618
	8,000	388	441	~273–670

The averaged mutation rate (2.01×10^{-4}/sequence/generation) estimated by human–chimpanzee (6 MY) and human–gorilla (8 MY) was used.

and Xq21.31 (Yu *et al.* 2002b). The scenario of the contribution from each allele frequency class to its θ-values was different; however, it was generally compatible with the results of neutrality tests in each region (data not shown). For example, similar contribution patterns were observed at loci 6p22, Xq13.3, and Xq21.31, consistent with the similar neutrality test results at these three loci.

Comparison of polymorphism and divergence among noncoding regions: We compared the levels of polymorphism within the human population and divergence between human and chimpanzee at the present locus with those in the other four regions, 22q11.2, 1q24, Xq13.3, and Xq21.31. The 22q11.2 locus showed no rejection of neutral mutation and the Xq13.3 locus showed the same neutrality test results as the 6p22 locus. The comparison was performed using the HKA test. All indels were excluded in the tests. All tests were not significant for the total sample or each subsample except for one in which the loci 6p22 and Xq21.31 were compared for the non-African sample (Table 5). In this case, 30 variant sites were observed from 82 non-African sequences at 6p22 while 34 variant sites were observed from only 42 non-African sequences at Xq21.31. Note that the expectation of polymorphism is $4N\mu$ for an autosomal locus and $3N\mu$ for an X-linked locus. In contrast, the level of divergence between human and chimpanzee at locus Xq21.31 was ~66% of that at 6p22. This difference resulted in a large χ^2-value in the HKA test.

Age of the MRCA: Table 6 shows the estimated age (T) of the MRCA for the entire sample, the African sample, and the non-African sample. To be consistent with our previous studies, we used the averaged mutation rate (2.01×10^{-4}/sequence/generation) estimated between human and chimpanzee (6 MY) and between human and gorilla (8 MY). Given the effective population size of 10,000 for the entire sample, the mode and mean estimates were 581,000 and 601,000 years ago, respectively, and the 95% confidence interval was between 365,000 and 1,043,000 years ago. The estimated age for the African sample was close to that for the entire sample. However, the age was much younger for the non-African sample; for example, the mode estimate (T_{mode}) was 388,000 given the effective population size of 8000. This is also much younger than that in the other two autosomal noncoding regions, 22q11.2 ($T_{mode} = 634{,}000$) (Zhao *et al.* 2000) and 1q24 ($T_{mode} = 672{,}000$) (Yu *et al.* 2001).

DISCUSSION

A typical noncoding region: Although many genomic regions have been studied in human populations, few were from noncoding regions. A genomewide study of genetic variation in a worldwide population is now feasible; however, existing studies (*e.g.*, the HapMap project) are biased toward the genetic causes of diseases, especially common human diseases (International HapMap Consortium 2005). In this study, we selected a noncoding region to reduce the influence of evolutionary forces such as selective constraints. In this 10-kb autosomal region, the nucleotide diversity in humans was 0.070%, close to the previous estimates (Sachidanandam *et al.* 2001). The mutation rate was estimated to be 0.99×10^{-9}/nucleotide site/year, which is typical in the genome. The local recombination rate (in a 1-Mb region) is the same as the genome average. The large proportions of unique variant sites, unique genotypes, and unique haplotypes suggest a large degree of subdivision among continental populations, strong recent population expansion, or both. The results of neutrality tests showed a similar incongruence as in other regions (*e.g.*, Xq13.3 and 1q24), which was likely caused by the strong excess of singletons. The HKA test indicated that the level of polymorphism in this region was not different from that

in the other four noncoding regions. In summary, the present region represents a typical noncoding region in the human genome, although its GC content is low. Therefore, the data may be used for further comparative analysis.

Recent human population expansion: The observed variation pattern provides evidence that the human population has been undergoing rapid expansion in recent history. First, an excess of singletons was observed in the total sample and in each subpopulation in this region. Such strong excess has been observed in many other genomic regions, including autosomal (*e.g.*, Zhao *et al.* 2000; Thorstenson *et al.* 2001; Yu *et al.* 2001, 2002a; Wooding *et al.* 2002; Nakajima *et al.* 2004), X-linked (*e.g.*, Kaessmann *et al.* 1999; Yu *et al.* 2002b; Hammer *et al.* 2004; Nachman *et al.* 2004), and Y-linked (*e.g.*, Shen *et al.* 2000) regions. Note that population subdivision cannot explain the strong excess of singletons, because the opposite was actually predicted by a coalescent simulation under Wright's island model (Yu *et al.* 2001). Moreover, we found a large proportion of unique variant sites in each continent at all five noncoding loci. Second, the contribution of the intermediate-frequency mutations to θ_Π was higher than or close to that expected (Table 4), suggesting that the human population expansion is not very ancient and a bottleneck event is unlikely in recent history. Third, no genotype was shared between Africans and non-Africans and only three were shared between Asians and Europeans. The proportions of unique haplotypes in Africans and non-Africans were remarkably higher than that of the shared ones. The large proportion of genotypes and haplotypes unique in each continental population was consistently observed at the other four 10-kb noncoding loci. Strikingly, the two most frequent haplotypes (nos. 28 and 29, Figure 1) were found only in non-Africans. Haplotype 28 was present on 32 non-African chromosomes, and all the other 7 haplotypes directly linked to it in the genetic network were in non-Africans. The same pattern was observed for haplotype 29. Interestingly, haplotype 28 is genetically distant from the ancestral haplotype, a strong signal of the recent expansion event(s). At 22q11.2, 1q24, and Xq21.31, we found that the highly frequent haplotypes were also more likely observed in the non-Africans, although not so remarkably as that at 6p22 (Kaessmann *et al.* 1999; Yu *et al.* 2002b). This feature is consistent with the observation that the numbers of haplotypes and unique haplotypes in Africans were much larger than those in Asians or Europeans at these loci (Table 3).

Because most genetic variants and haplotypes arose recently, the genetic signatures above should reflect relatively recent population expansions, especially in the non-African population. A recent analysis of SNP density distribution in the human genome under a model including parameters of recombinaiton and population size suggested a collapse followed by a mild population expansion during the Upper Paleolithic period (Marth *et al.* 2003). The scenario of bottleneck and expansion was further examined in Eswaran *et al.* (2005). The observed patterns at the five noncoding loci are compatible with this bottleneck-and-expansion scenario. However, our data suggest the bottleneck to be mild and the population expansion to be rapid and strong, because the contribution of intermediate-frequency SNPs to θ_Π was large at most of these loci and the rare SNPs occurred much more frequently than expected. Eswaran *et al.* (2005) pointed out that a negative Tajima's *D*-value (*e.g.*, ≤ -1.5) at many loci would suggest a population expansion, but such a value is often not found in real data. This criterion may not be appropriate to signal expansion because the two θ-values in Tajima's test likely counteract according to our examination of contribution of each size class of SNPs to the θ-values (Table 4). Finally, a recent nested-clade analysis of 25 loci suggested three major expansions of the human population out of Africa, which occurred ~1,500,000, ~700,000, and ~100,000 years ago, respectively (Templeton 2005). The population expansion we discussed above should reflect the most recent event, likely around or younger than 100,000 years ago.

Age of the MRCA: The age estimates for the entire sample, the African sample, and the non-African sample were 581,000, 544,000, and 388,000 years ago, respectively, given the corresponding effective population sizes of 10,000, 10,000, and 8000. These estimates are significantly younger than those estimated from the data at 22q11.2 (1.29 MY), 1q24 (1.47 MY), PDHA1 (1.86 MY), and MC1R (1.52 MY) (Harris and Hey 1999; Zhao *et al.* 2000; Makova *et al.* 2001; Yu *et al.* 2001). However, they are comparable to other estimates based on the data from the β-globin gene (750,000), Xq13.3 (535,000), and Xq21.31 (710,000) (Harding *et al.* 1997; Kaessmann *et al.* 1999; Yu *et al.* 2002b). Since the mutation rate at this region is close to that at 22q11.2 (2.28×10^{-4}/sequence/generation) and higher than that at 1q24 (1.33×10^{-4}/sequence/generation), the age estimates may indicate different genetic histories at these three loci.

Considering three autosomal noncoding loci (6p22, 22q11.2, and 1q24) altogether, the ages of the MRCA of the entire sample, the African sample, and the non-African sample were averaged to be 860,000 ± 258,000, 803,000 ± 245,000, and 565,000 ± 154,000 years ago, respectively, given the corresponding effective population sizes of 15,000, 10,000, and 8000. Note that the effective population size for the entire sample used here was based on the average of Tajima's and Watterson's estimates in these three regions (see RESULTS). For the commonly accepted size of 10,000, the age of the MRCA would be 1,114,000 ± 470,000.

Origin of modern humans: How modern humans originated is still controversial, although this issue has

been examined with extensive genetic data and simulations in the past two decades. While genetic data from mtDNA, Y chromosomes, microsatellites, minisatellites, *Alu* repeats, and nuclear sequences in general support the "out-of-Africa" hypothesis (*e.g.*, Cann *et al.* 1987; Tishkoff *et al.* 1996), other data and analyses did not support it or even favored the "multiregional" hypothesis (*e.g.*, Jorde *et al.* 1995; Templeton 1997, 2005). The global genetic information in the five 10-kb noncoding regions tends to support the out-of-Africa model but suggests that both models are too simple for several reasons. First, we consistently observed greater sequence and haplotype diversity in the African sample than in the non-African sample, even though the sample size in Africans was smaller than that in non-Africans. This is in agreement with the out-of-Africa hypothesis, but not the multiregional hypothesis, because such a difference cannot be explained by continuous gene flow and migration among the three continental populations (Yu *et al.* 2001). Note that the greater diversity in the African sample could be caused by the larger effective population size in Africa (Nachman *et al.* 1996). However, the smaller effective non-African population size itself might be due to the mild bottleneck event(s) in non-Africans (Marth *et al.* 2003; Eswaran *et al.* 2005).

Second, at 6p22, a large portion of African-unique haplotypes were closely related to the ancestral haplotype, while most non-African-unique haplotypes had the paths to the ancestral haplotype through African-unique or shared haplotypes. Moreover, the most frequent haplotypes were shared only by non-Africans. This pattern cannot be explained by the multiregional hypothesis, unless severe bottleneck event(s) had occurred exclusively in the non-African population and then strong gene flow from Africans to non-Africans was followed by rapid non-African population expansion. Such a scenario is actually against the multiregional hypothesis and favors the out-of-Africa hypothesis.

Third, although the estimated ages of MRCAs for non-African sequences varied greatly among the five loci, they are much older than the emergence date (100,000–130,000 years ago) of modern humans. This ancient genetic history outside of Africa is against the complete replacement of indigenous archaic European and Asian populations by an African founder group, as proposed by the out-of-Africa hypothesis.

Fourth, under the assumptions that archaic non-African populations were completely replaced by an African founder group and, after the complete replacement, gene flow was not strong between African and non-African populations, one would likely observe some subnetworks exclusive for African-unique haplotypes in the haplotype network. While this feature can be generally identified in the haplotype network, it is much less conspicuous than expected, especially at locus Xq13.3.

Finally, genetic patterns varied among the loci. For example, we found that the contribution of the intermediate-frequency mutations in the non-African sample to its θ_Π was higher than or close to that expected at loci 6p22, 22q.11, and Xq21.31 but not at 1q24 and Xq13.3. Therefore, inference of the human population history requires more representative population genetics data sets, especially from noncoding regions.

We thank J. B. Clegg, L. B. Jorde, M. Lin, M. Ramsay, M. Ruvolo, N. Sambuughin, and T. Jenkins for kindly giving us DNA samples. This work was supported by Thomas F. and Kate Miller Jeffress Memorial Trust Fund (to Z.Z.), National Institutes of Health grants (to W.-H. Li) and GM50428 (to Y.-X. Fu), and National Science Foundation grant DEB9707567 (to Y.-X. Fu).

LITERATURE CITED

Bandelt, H. J., P. Forster and A. Rohl, 1999 Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. **16:** 37–48.

Cann, R. L., M. Stoneking and A. C. Wilson, 1987 Mitochondrial DNA and human evolution. Nature **325:** 31–36.

Don, R. H., P. T. Cox, B. J. Wainwright, K. Baker and J. S. Mattick, 1991 'Touchdown' PCR to circumvent spurious priming during gene amplification. Nucleic Acids Res. **19:** 4008.

Eswaran, V., H. Harpending and A. R. Rogers, 2005 Genomics refutes an exclusively African origin of humans. J. Hum. Evol. **49:** 1–18.

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

Hammer, M. F., D. Garrigan, E. Wood, J. A. Wilder, Z. Mobasher *et al.*, 2004 Heterogeneous patterns of variation among multiple human X-linked loci: the possible role of diversity-reducing selection in non-Africans. Genetics **167:** 1841–1853.

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.

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

Hudson, R. R., M. Kreitman and M. Aguade, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

International HapMap Consortium, 2005 A haplotype map of the human genome. Nature **437:** 1299–1320.

Jorde, L. B., M. J. Bamshad, W. S. Watkins, R. Zenger, A. E. Fraley *et al.*, 1995 Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. Am. J. Hum. Genet. **57:** 523–538.

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.

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.

Makova, K. D., M. Ramsay, T. Jenkins and W. H. Li, 2001 Human DNA sequence variation in a 6.6-kb region containing the melanocortin 1 receptor promoter. Genetics **158:** 1253–1268.

Marth, G., G. Schuler, R. Yeh, R. Davenport, R. Agarwala *et al.*, 2003 Sequence variations in the public human genome data reflect a bottlenecked population history. Proc. Natl. Acad. Sci. USA **100:** 376–381.

Myers, S., L. Bottolo, C. Freeman, G. McVean and P. Donnelly, 2005 A fine-scale map of recombination rates and hotspots across the human genome. Science **310:** 321–324.

Nachman, M. W., W. M. Brown, M. Stoneking and C. F. Aquadro, 1996 Nonneutral mitochondrial DNA variation in humans and chimpanzees. Genetics **142:** 953–963.

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.

Nakajima, T., S. Wooding, T. Sakagami, M. Emi, K. Tokunaga *et al.*, 2004 Natural selection and population history in the human angiotensinogen gene (AGT): 736 complete AGT sequences in chromosomes from around the world. Am. J. Hum. Genet. **74:** 898–916.

Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer and R. Rozas, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:** 2496–2497.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein *et al.*, 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409:** 928–933.

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, M., N. J. Smith and P. Donnelly, 2001 A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. **68:** 978–989.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

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

Templeton, A. R., 1997 Out of Africa? What do genes tell us? Curr. Opin. Genet. Dev. **7:** 841–847.

Templeton, A. R., 2005 Haplotype trees and modern human origins. Am. J. Phys. Anthropol. (Suppl. 41): 33–59.

Thorstenson, Y. R., P. Shen, V. G. Tusher, T. L. Wayne, R. W. Davis *et al.*, 2001 Global analysis of ATM polymorphism reveals significant functional constraint. Am. J. Hum. Genet. **69:** 396–412.

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.

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.

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

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.

Yu, N., Z. Zhao, Y. X. Fu, N. Sambuughin, M. Ramsay *et al.*, 2001 Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. Mol. Biol. Evol. **18:** 214–222.

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.

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.

Zhao, Z., Y.-X. Fu, D. Hewett-Emmett and E. Boerwinkle, 2003 Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. Gene **312:** 207–213.

Communicating editor: N. Takahata