

Global Patterns of Human DNA Sequence Variation in a 10-kb Region on Chromosome 1

Ning Yu,* Z. Zhao,† Y.-X. Fu,† N. Sambuughin,‡ M. Ramsay,§ T. Jenkins,§ E. Leskinen,|| L. Patthy,¶ L. B. Jorde,** T. Kuromori,* and W.-H. Li*

*Department of Ecology and Evolution, University of Chicago; †Human Genetics Center, University of Texas at Houston; ‡Neurology Research, Phoenix, Arizona; §Department of Human Genetics, South African Institute for Medical Research, Johannesburg, South Africa; ||Department of Biology, University of Oulu, Finland; ¶Institute of Enzymology, Hungarian Academy of Sciences, Budapest, Hungary; and **Department of Human Genetics, University of Utah

Human DNA variation is currently a subject of intense research because of its importance for studying human origins, evolution, and demographic history and for association studies of complex diseases. A ~10-kb region on chromosome 1, which contains only four small exons (each <155 bp), was sequenced for 61 humans (20 Africans, 20 Asians, and 21 Europeans) and for 1 chimpanzee, 1 gorilla, and 1 orangutan. We found 52 polymorphic sites among the 122 human sequences and 382 variant sites among the human, chimpanzee, gorilla, and orangutan sequences. For the introns sequenced (8,991 bp), the nucleotide diversity (π) was 0.058% among all sequences, 0.076% among the African sequences, 0.047% among the Asian sequences, and 0.045% among the European sequences. A compilation of data revealed that autosomal regions have, on average, the highest π value (0.091%), X-linked regions have a somewhat lower π value (0.079%), and Y-linked regions have a very low π value (0.008%). The lower polymorphism in the present region may be due to a lower mutation rate and/or selection in the gene containing these introns or in genes linked to this region. The present region and two other 10-kb noncoding regions all show a strong excess of low-frequency variants, indicating a relatively recent population expansion. This region has a low mutation rate, which was estimated to be 0.74×10^{-9} per nucleotide per year. An average estimate of ~12,600 for the long-term effective population size was obtained using various methods; the estimate was not far from the commonly used value of 10,000. Fu and Li's tests rejected the assumption of an equilibrium neutral Wright-Fisher population, largely owing to the high proportion of low-frequency variants. The age of the most recent common ancestor of the sequences in our sample was estimated to be more than 1 Myr. Allowing for some unrealistic assumptions in the model, this estimate would still suggest an age of more than 500,000 years, providing further evidence for a genetic history of humans much more ancient than the emergence of modern humans. The fact that many unique variants exist in Europe and Asia also suggests a fairly long genetic history outside of Africa and argues against a complete replacement of all indigenous populations in Europe and Asia by a small Africa stock. Moreover, the ancient genetic history of humans indicates no severe bottleneck during the evolution of humans in the last half million years; otherwise, much of the ancient genetic history would have been lost during a severe bottleneck. We suggest that both the "Out of Africa" and the multiregional models are too simple to explain the evolution of modern humans.

Introduction

Human DNA variation is currently a subject of intense research for several reasons. First, DNA variation within and between human populations is of great interest to human geneticists and evolutionists. Second, there is much interest in the utility of single nucleotide polymorphisms (SNPs) in molecular medicine, because SNP markers may be useful in association studies of complex diseases, assessment of individuals' predisposition to diseases, and tailoring of therapies. Third, the availability of long genomic sequences generated by the Human Genome Project and the advent of inexpensive DNA sequencing techniques have made large-scale population studies feasible. In fact, there are now many large-scale population studies of human DNA variation. These include regions containing the genes for, respectively, β -globin (Harding et al. 1997), lipoprotein lipase (Nickerson et al. 1998), angiotensin-converting enzyme (Rieder et al. 1999), and pyruvate dehydrogenase E1

(PDHA1) (Harris and Hey 1999) and an 8-kb intron segment of the dystrophin gene (Zietkiewicz et al. 1998), a 10-kb noncoding region in Xq13.3 (Kaessmann et al. 1999), the replication origin IR and the flanking and coding regions of the β -globin gene (Fullerton et al. 2000), and a 10-kb noncoding region in 22q11.2 (Zhao et al. 2000). To date, however, the number of large-scale worldwide surveys of DNA sequence variation is still small, especially for noncoding regions.

We have been pursuing human DNA variation studies in noncoding regions for two purposes. First, we wish to establish a genomewide and worldwide neutrality standard of nucleotide diversity. By "neutrality standard," we mean the level of nucleotide diversity expected in a region in which all mutations are neutral and not directly subject to natural selection. This standard will be a very useful reference, especially for comparison with the levels of nucleotide diversity in coding regions. Obviously, such a standard requires data from many genomic regions, because the level of nucleotide diversity in a region is subject to strong stochastic effects. Second, we wish to study the origin and evolution of modern humans. DNA sequence data from noncoding regions may more accurately reflect human history than data from coding regions, because noncoding regions

Key words: nucleotide diversity, DNA variation, human evolution, unique variants.

Address for correspondence and reprints: Wen-Hsiung Li, Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, Illinois 60637. E-mail: whli@uchicago.edu.

Mol. Biol. Evol. 18(2):214–222. 2001

© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

are not directly subject to natural selection. The majority of past studies on human DNA variation, which are mainly from mitochondrial (mt) DNA, microsatellite DNA, and the Y chromosome, have largely given the impression of a relatively shallow genetic history of humans. These observations have been taken as evidence for the Out of Africa model for the origin of modern humans, which postulates that a founder group of modern humans emigrated from Africa about 100,000 years ago to Europe and Asia and completely replaced all the indigenous populations outside of Africa (Cann, Stoneking, and Wilson 1987; Stringer and Andrew 1988). However, recent studies of the β -globin and the PDHA1 gene regions (Harding et al. 1997; Harris and Hey 1999) and a 10-kb noncoding region on chromosome 22 (Zhao et al. 2000) have revealed an ancient genetic history of humans and suggested that human evolution has been more complex than depicted by the simple Out of Africa model. To attain a better understanding of this issue, it is necessary to obtain sequence data from other non-coding regions.

For the above purposes, we selected a \sim 10-kb region on chromosome 1 and obtained sequence data from worldwide populations. This region contains mostly introns, although it also includes four short exons. The new data were compared with the data from Xq13.3 and 22q11.2 and other regions to study the features of sequence variation within and between populations and were used to infer the genetic history of human evolution.

Materials and Methods

Region Selection and Populations Sampled

A 12-kb region corresponding to nucleotide positions 18333–30332 in locus HS125H23 on human chromosome 1q24 was selected (GenBank accession number Z94054) because no gene was registered in this region at GenBank. After excluding a region containing a polyA segment and a region containing an MER33, an incomplete *AluSx* repeat, and an MIR2 repeat, a total of \sim 10 kb of nucleotide sites were selected for sequencing. Initially, no potential coding region was detected in the 12-kb region by XGRAIL. However, upon reexamination, three strong potential coding regions (exons) and a weak potential coding region were detected by both GenScan and GRAIL-EXP (see *Results*). As these potential exons are short (\leq 155 bp), the region selected covers largely introns.

Sixty-one individuals were collected worldwide from 14 human populations in three major geographic areas: 20 Africans (5 South African Bantu speakers, 1 !Kung, 2 Mbuti Pygmies, 2 Biaka Pygmies, 5 Nigerians, 5 Kenyans), 20 Asians (8 Chinese, 3 Japanese, 6 Indians, 3 Yakuts), and 21 Europeans (6 Swedes, 2 Finns, 5 French, 5 Hungarians, 3 Italians). One chimpanzee, one gorilla, and one orangutan were used as outgroups.

PCR Amplification and DNA Sequencing

Five primer pairs were designed to amplify three overlapping fragments covering positions 9–6141 and

two overlapping fragments covering positions 6880–11022 in the 12-kb region. Touchdown PCR (Don et al. 1991) was used, and the reactions were carried out under the conditions described in Zhao et al. (2000). The PCR products were purified with the Wizard PCR Preps DNA Purification Resin Kit (Promega). Sequencing reaction was performed according to the protocol of ABI Prism BigDye Terminator Sequencing Kits (Perkin Elmer) modified by quarter reaction. The extension products were purified by Sephadex G-50 (DNA grade, Pharmacia) and run on an ABI 377XL DNA sequencer using 4.25% gels (Sooner Scientific).

ABI DNA Sequence Analysis 3.0 was used for lane tracking and base calling. The data were then proofread; the fluorescence traces were reread manually and heterozygous sites were detected as double peaks. The segment sequences were assembled automatically using SeqMan in DNASTAR. The assembled files were carefully checked manually using the same program, and variant sites were identified in the aligned sequences in MegAlign in DNASTAR. All of the nucleotides in each segment were sequenced at least once in both directions. Furthermore, all singletons and doubletons, which are defined as variants that appear, respectively, only once and twice in the total sample, were verified by reamplifying the region containing the variant site and resequencing the region in both directions using new internal primers that were close to the site.

Data Analysis

The sequences were aligned by MegAlign in the DNASTAR software package. The human consensus sequence was obtained from the alignment using DNASTAR. The human ancestral sequence was inferred by comparing the human sequences with the outgroup sequences using the maximum-parsimony principle.

For a DNA sequence subject to no natural selection, the mutation rate per sequence per generation (μ) is estimated by

$$\mu = dgL/(2t), \quad (1)$$

where d is the number of nucleotide substitutions per site between two sequences, t is the divergence time between two sequences, L is the sequence length (bp), and g is the generation time ($g \approx 20$ years for humans). Watterson's (1975), Tajima's (1983), and Fu's (1994) methods were used to estimate $\theta = 4N_e\mu$, where N_e is the effective population size.

Tajima's (1989) test and Fu and Li's (1993) tests were used to test the selective neutrality of the region studied; a program is available at <http://hgc.sph.uth.tmc.edu/fu>. The critical points (values) for the neutrality tests were obtained from 5,000 simulated samples. Fu's (1996) and Fu and Li's (1997) methods were used to estimate the age of the most recent common ancestor (MRCA) of the DNA sequences in a sample. We computed both the mode and mean of the age (T) of the MRCA in years.

Note that all of the above computations require only segregating site data but do not require haplotype data.

Table 1
Numbers of Variant Sizes (excluding indels) in each DNA Region Among Sequences

TYPE OF VARIANT	10 KB IN 1Q24			10 KB IN 22Q11.2 ^a			10 KB IN XQ13.3 ^b		
	All (122) ^c	Africans (40)	Non-Africans (82)	All (128)	Africans (40)	Non-Africans (88)	All (69)	Africans (23)	Non-Africans (46)
Singleton	19 (3) ^d	7 (3)	12 (0)	20 (2)	8 (1)	12 (1)	19 (0)	11 (0)	8 (0)
Doubleton. . . .	7 (0)	2 (0)	5 (0)	23 (1)	19 (0)	4 (1)	5 (0)	5 (0)	0 (0)
Others	22 (1)	20 (1)	11 (0)	32 (1)	27 (1)	28 (0)	9 (0)	8 (0)	9 (0)
Total	48 (4)	29 (4)	28 (0)	75 (4)	54 (2)	44 (2)	33 (0)	24 (0)	17 (0)

^a From Zhao et al. (2000).

^b From Kaessmann et al. (1999).

^c Number of sequences studied.

^d The numbers of indels are given in parentheses.

Results and Discussion

Sequence Data

We sequenced ~9,626 nucleotide sites in the selected region in 61 humans (AF310265–AF310325), 1 chimpanzee (AF310683), 1 gorilla (AF310682), and 1 orangutan (AF310681). The human consensus and ancestral sequences were obtained as explained in *Materials and Methods*. The GC contents of the human consensus, human ancestral, chimpanzee, gorilla, and orangutan sequences were ~31.5%, which is much lower than the genome average of 42%. Thus, the region studied was GC-poor.

Three strong potential coding regions (exons) in the selected 12-kb segment were predicted by GenScan (each with a probability >95.5%) and GRAIL-EXP; these potential exons were at sites 23522–23632, 26131–26285, and 27542–27679 in locus HS125H23 (Z94054), respectively. In fact, a BLAST search showed that the amino acid sequence translated from these three exons was 86% similar to a segment of a human membrane protein CH1 (GenBank accession number AF097535). Later on a BLAST search of the GenBank with the nucleotide sequence of the ~10 kb region indicated that four exons are similar to the human membrane protein CH1 (similarity > 99%) and the amino acid sequence translated from these four exons is identical to that translated from this gene. These exons are from site 21,773 to 21,888, site 26,128 to 26,285, site 27,542 to 27,679, and site 27,902 to 28,056 in locus HS125H23 (Z94054), respectively.

Pattern of Sequence Variation

A total of 48 variant sites were found in the alignment of human sequences; 19 of them were observed only once (i.e., singletons), 7 were observed twice (i.e., doubletons), and 22 were observed more than twice (i.e., others) (table 1). Two variant sites (one synonymous and one nonsynonymous singleton) were observed in the second exon, while all of the remaining 46 variant sites were found in introns. All singletons and doubletons were verified as explained in *Materials and Methods*, and no error was found. In addition to the 48 single-nucleotide variants, we found 4 insertions/deletions (indels) among the 122 human sequences. On average, ~5

variant sites per 1,000 bp were found in the region studied.

The numbers of variant sites (excluding indels) in the African, Asian, and European sequences were 29, 20, and 16, respectively (table 1). Thus, Africans had the largest number of variants. The pattern of sequence variation in Africans differed somewhat from that in non-Africans. For example, less than one third of the variant sites in Africans were singletons, whereas close to one half of the variant sites in non-Africans were singletons. Interestingly, the proportions of unique variant sites in each of the three continents were high: 20 (69%, including 7 singletons), 11 (55%, 8 singletons), and 8 (50%, 4 singletons) among the African, Asian, and European sequences, respectively. This observation suggests a substantial degree of isolation between continents.

Table 1 also includes the patterns of sequence variation in the 10-kb noncoding regions on chromosomes X and 22 (Kaessmann et al. 1999; Zhao et al. 2000). Note that among the African sequences, the number of low-frequency variants (i.e., singletons and doubletons) was smaller than that of high-frequency variants (i.e., others) in the present region, whereas the opposite was true for the two other regions. On the other hand, among non-African sequences, the number of low-frequency variants was larger than that of the high-frequency variants in the present region, whereas the opposite was true for the other two regions. Thus, although a stronger excess of low-frequency variants in Africans than in non-Africans was observed in the previous two regions, the opposite was found in the present region. This difference in pattern notwithstanding, two features common to the three regions were noted. First, there were more variants in the African sample than in the non-African sample, despite the number of Africans studied being less than half that of non-Africans studied. Thus, Africans were considerably more polymorphic than non-Africans, in agreement with previous observations (Cann, Stoneking, and Wilson 1987; Kaessmann et al. 1999; Zhao et al. 2000). Second, the number of low-frequency variants (singletons and doubletons) in the total sample was larger than the number of high-frequency variants, e.g., 26 versus 22 in the present region. This excess of low-frequency variants in all three regions is

Table 2
Distribution of the Numbers of Variant Sites Observed in the Human and Outgroup Sequences

TYPE OF VARIATION	SUBREGIONS (~1 kb each)										Total
	1	2	3	4	5	6	7	8	9	10	
10-kb region in 1q24 (present study)											
Substitutions . . .	41	42	36	30	35	32	38	35	28	21	338
Indels	5	6	5	4	7	8	4	3	1	1	44
Total	46	48	41	34	42	40	42	38	29	22	382
Human	8	6	4	6	6	9	6	3	2	2	52
10-kb region in 22q11.2 (Zhao et al. 2000)											
Substitutions . . .	37	36	33	51	41	45	40	48	49	36	416
Indels	2	14	10	2	4	5	3	2	2	3	47
Total	39	50	43	53	45	50	43	50	51	39	463
10-kb region in Xq13.3 (Kaessmann et al. 1999)											
Substitutions . . .	20	29	20	16	18	20	26	14	29	24	216
Indels	1	5	2	0	1	2	0	3	2	0	16
Total	21	34	22	16	19	22	26	17	31	24	232

in sharp contrast to the situations for the dystrophin and PDHA1 genes (Zietkiewicz et al. 1998; Harris and Hey 1999) and suggests a relatively recent population expansion, because such an excess is not expected from an equilibrium Wright-Fisher population.

The present region was considerably less polymorphic than the region on chromosome 22—it had fewer high-frequency variants, especially among non-Africans, and a much smaller number of doubletons. The relatively low polymorphism may be due to a lower mutation rate (see below) and selection in the gene containing these introns or in genes linked to this region. The X-linked region was even less variable (table 1). This may be because an X-linked region has a smaller effective population size than an autosomal region ($3N_e/4$ vs. N_e) and because the X region has a low recombination rate (Kaessmann et al. 1999), so that compared to the other two regions, it is subject to stronger background selection (i.e., effects of deleterious mutations in genes linked to the region) or selective sweep (effects of positive selection in genes linked to the region) (Begun and Aquadro 1992; Charlesworth 1994).

A comparison of all sequences, including the chimpanzee, gorilla, and orangutan sequences, revealed 382

variant sites, 44 of which were indels. The 382 variant sites were evenly distributed in this region ($\chi^2 = 14.4$, $df = 9$, $P = 0.11$) (table 2). The number of variant sites in human populations was also evenly distributed ($\chi^2 = 9.9$, $df = 9$, $P = 0.36$).

Among the 44 indels in our data, 13, 7, 10, and 10 were 1-, 2-, 3-, and 4-nt indels, respectively. The remaining 4 indels involved 5, 8, or >10 nt. Therefore, the majority of these indels were short.

Mutation Pattern

Comparing the human, chimpanzee, gorilla, and orangutan sequences, we were able to infer the direction of 172 mutations (table 3); the proportion of transitional changes was 66%. For the 169 mutations for which the direction could not be inferred, the proportion of transitions was 65% (table 3). This proportion was between the values (59% and 70%) observed in pseudogenes (Li, Wu, and Luo 1984) and for the 10-kb region in 22q11.2 (Zhao et al. 2000). For those mutations whose direction could be inferred, the number of G/C-to-A/T mutations was 57, while that of A/T-to-G/C mutations was 90. According to the GC and AT contents of 68.5% and 31.5%, the expected numbers of G/C-to-A/T mutations and A/T-to-G/C mutations are 100.7 and 46.3, respectively, and a comparison with the observed numbers gives $\chi^2 = 3.61$ and $P = 0.057$, which is close to significant. This result suggests that G/C-to-A/T mutations might occur more frequently than A/T-to-G/C mutations, similar to the situation for mammalian pseudogenes (G/C-to-A/T, 64.5%; A/T-to-G/C, 35.5%) (Li 1997).

Neutrality Tests

The assumption that the region under study is subject to no natural selection was tested using the sequence data. Using the critical points obtained from the 5,000 samples we simulated, we found that Tajima's and Fu and Li's tests could not reject the neutral Wright-Fisher model when each continent was considered separately (table 4). However, when we used data from more than one continent, the results became different. Although

Table 3
Mutation Pattern

Direction of Mutation Can Be Inferred	Direction of Mutation Cannot Be Inferred	
A→C	A↔G	66
A→G	C↔T	44
A→T	A↔C	14
C→A	A↔T	17
C→G	G↔C	12
C→T	G↔T	16
G→A		
G→C		
G→T		
T→A		
T→C		
T→G		
Subtotal		169
Transitions (%)		113 (66%)
		110 (65%)

Table 4
Neutrality Tests

Test	No. of Sequences	θ	Test Value	Critical Value ($P = 0.05$)	P
Asian sequences					
Tajima's test	40	4.26	-0.293	-1.43	>0.10
Fu and Li's D	40	4.26	-1.230	-1.86	>0.10
Fu and Li's F	40	4.26	-0.997	-1.78	>0.10
European sequences					
Tajima's test	42	4.14	0.145	-1.39	>0.10
Fu and Li's D	42	4.14	-0.021	-1.91	>0.10
Fu and Li's F	42	4.14	0.042	-1.78	>0.10
African sequences					
Tajima's test	40	6.82	0.000	-1.43	>0.10
Fu and Li's D	40	6.82	-0.051	-1.84	>0.10
Fu and Li's F	40	6.82	-0.036	-1.76	>0.10
Non-African sequences					
Tajima's test	82	4.17	-0.768	-1.43	>0.10
Fu and Li's D	82	4.17	-2.226*	-1.90	$0.02 < P < 0.03$
Fu and Li's F	82	4.17	-1.889*	-1.77	0.04
All sequences					
Tajima's test	122	5.29	-1.224	-1.41	$0.05 < P < 0.10$
Fu and Li's D	122	5.29	-2.601*	-1.86	$0.01 < P < 0.02$
Fu and Li's F	122	5.29	-2.325*	-1.70	$0.01 < P < 0.02$
All sequences (including indels)					
Tajima's test	122	5.83	-1.200	-1.38	$0.05 < P < 0.10$
Fu and Li's D	122	5.83	-2.994*	-1.79	$0.005 < P < 0.01$
Fu and Li's F	122	5.83	-2.564*	-1.64	$0.005 < P < 0.01$

NOTE.—The critical points (values) were obtained from 5,000 simulated samples. The indels were excluded in each test except for the last three tests.
* Significant at the 5% level.

Tajima's test remained nonsignificant, Fu and Li's tests for non-Africans and for all samples were significant (table 4); the conclusion became even stronger when indels were included in the tests. The rejection of neutrality was largely due to the high proportion of low-frequency variants. As the region studied was largely noncoding, the rejection of neutrality may be due to two factors: (1) a relatively recent population expansion, which can increase the number of low-frequency variants, and (2) natural selection in the exons or in genes linked to this region. In fact, the region is ~44 kb and ~110 kb away from a functional gene at its 5' and 3' ends, respectively.

As the rejection of the neutrality assumption was largely due to the excess of low-frequency variants, one might wonder whether the excess was due to pooling of data from different populations. However, data pooling actually tends to increase, rather than decrease, the proportion of low-frequency variants, as implied by the re-

Table 5
Expected Proportion of Singleton Mutations in the Total Sample when 10 Sequences Are Sampled from each Subpopulation

NO. OF SUBPOPULATIONS	MIGRATION RATE ($4Nm$)		RANDOM MATING (no subdivision)
	0.1	1.0	
2	0.074	0.210	0.281
4	0.017	0.100	0.235
10	0.004	0.036	0.190

sult of Fu (1996), who showed that Fu and Li's D test tends to be positive under population subdivision. Since the D test compares the number of singletons with the total number of mutations, Fu's (1996) result indicates that the proportion of singletons is reduced under population subdivision. To show this, we conducted a coalescent simulation using Wright's island model. The expected number of singleton mutations was assumed to be equal to the expected sum of lengths of external branches in the sequence genealogy, which is expressed in units of $\theta = 4N_e\mu$, where N_e is the effective size of the entire population and μ is the rate of mutation per sequence per generation. The results are shown in table 5, in which each entry is the ratio of the expected sum of length of external branches and the expected total tree length of the simulated sequence genealogy; the ratio is independent of θ . Under the assumption of a single random mating population with the same θ , the expected sums of lengths of external branches and all branches are 1 and $1 + \frac{1}{2} + \dots + 1/(n - 1)$, respectively, where n is the sample size. As can be seen from table 5, the more subpopulations or the less migration, the lower the proportion of singletons relative to the total number of mutations in the sample. Clearly, population subdivision tends to reduce the proportion of low-frequency variants.

Nucleotide Diversity

Nucleotide diversity (π) is defined as the average number of nucleotide differences per site between two

Table 6
Nucleotide Diversity Within and Between Populations in Noncoding Regions

REGION	TYPE OF SEQUENCE	NO. OF SEQUENCES	LENGTH (bp)	NUCLEOTIDE DIVERSITY (%)							REFERENCES ^a
				Total	Africa	Asia	Europe	Africa/Asia	Africa/Europe	Asia/Europe	
Autosomal ^b											
1q24	Introns	122	8,991	0.058	0.076	0.047	0.045	0.065	0.063	0.046	
22q11	Noncoding	128	9,834	0.088	0.085	0.075	0.077	0.083	0.108	0.091	1
β-globin	5' flanking	349	814	0.318	0.323	0.331	0.232	0.393	0.309	0.331	2
	Introns	349	980	0.140	0.072	0.142	0.116	0.171	0.096	0.182	2
	3' flanking	349	582	0.264	0.217	0.273	0.255	0.304	0.242	0.319	2
Subtotal			21,201	0.091	0.093	0.081	0.076	0.097	0.099	0.092	
β-globin IR	5' flanking	36	2,902	0.075	0.136		0.029		0.104		3
	IR	36	1,300	0.232	0.287		0.159		0.290		3
	3' flanking	36	1,874	0.141	0.119		0.147		0.140		3
X-linked ^b											
Xq13.3	Noncoding	69	9,564	0.034	0.035	0.025	0.034	0.038	0.034	0.035	4
PDHA1	Introns	35	3,530	0.169	0.216	0	0.009	0.248	0.251	0.005	5
Dys44	Introns	847	7,185	0.102	0.109 ^c	0.085	0.071	0.107 ^c	0.110 ^c	0.084	6
ZFX	Introns	336	782	0.021	0.053 ^c	0.004 ^c	0.000	0.030 ^c	0.028 ^c	0.002	7
Subtotal			21,061	0.079	0.091	0.040	0.041	0.096	0.096	0.045	
Y-linked ^b											
ZFY	Introns	205	676	0.001							8
SRY	Introns	5	~17,685	0.008							9
Subtotal			18,361	0.008							
Autosomal	<i>Alu</i>		1,138	0.263							1, 10
X-linked	<i>Alu</i>		1,455	0.125							4-7

^a 1 = Zhao et al. (2000); 2 = Harding et al. (1997); 3 = Fullerton et al. (2000); 4 = Kaessmann et al. (1999); 5 = Harris and Hey (1999); 6 = Zietkiewicz et al. (1998); 7 = Jaruzelska et al. (1999); 8 = Jaruzelska, Zietkiewicz, and Labuda (1999); 9 = Whitfield, Sulston, and Goodfellow (1995); 10 = Nickerson et al. (1998).

^b *Alu* sequences were excluded.

^c African Americans were excluded from the analysis.

randomly chosen sequences from the population. The π value was calculated using the program DNASP, version 3.0 (Rozas and Rozas 1999). For the introns sequenced, the π value was 0.058% among all sequences, 0.076% among the African sequences, 0.047% among the Asian sequences, and 0.045% among the European sequences (table 6). Thus, the π value was largest among Africans. These π values were considerably lower than those for the 10-kb noncoding region in 22q11.2 (Zhao et al. 2000) and the average value (0.11%) for fourfold-degenerate sites in coding genes in white Americans (Li and Sadler 1991). The low nucleotide diversity in this region may be due to a lower mutation rate (see below) and/or background selection and selective sweep.

Table 6 presents a list of the π values for various noncoding regions. The π values are usually highest in Africans but are quite similar in Asians and Europeans. For example, the average π values for the autosomal regions are 0.093%, 0.081%, and 0.076% for the Africans, Asians, and Europeans, respectively. These values are somewhat lower than the average π value (0.11%) at fourfold-degenerate sites in 49 genes. However, the number of noncoding regions studied is small, so no general conclusion can be drawn yet.

A large variation in π is seen among regions (table 6). In particular, the 5' and 3' flanking regions of the β -globin gene and the β -globin replication origin initiation region (IR) have high π values (Harding et al. 1977; Fullerton et al. 2000). The high π value in the IR has

been speculated to be due to a high mutation rate because of the peculiar feature of the DNA unwinding element in the IR (Fullerton et al. 2000). On average, the autosomal regions have the highest π value (0.091%), the X-linked regions have a somewhat lower π value (0.079%), and the Y-linked regions have a very low π value (0.008%). These differences may be partly due to the fact that the relative effective population sizes are N_e , $3N_e/4$, and $N_e/4$ for an autosomal, an X-linked, and a Y-linked sequence, respectively. However, the extremely low π value for Y-linked sequences may be mainly due to background selection and selective sweep, because there is no recombination in the Y chromosome except for the pseudoautosomal region. Background selection and selective sweep should have, on average, stronger effects on an X-linked region than on an autosomal region because of a lower average recombination rate in the X chromosome than in an autosome, partly accounting for the lower π value for X-linked regions. The number of *Alu* sequences studied is small, but the data suggest a higher average π value for *Alu* than for other regions. This is not surprising, because *Alus* have higher mutation rates due to the presence of a high frequency of CpG dinucleotides.

Mutation Rate, θ , N_e

The average numbers of nucleotide substitutions per site were 0.62% between human and chimpanzee

Table 7
Estimates of the Mutation Rate per Site per Year (ν), Parameter θ , and Effective Population Size (N_e)

PARAMETER	ESTIMATED VALUES								
	Chimpanzee vs. Human			Gorilla vs. Human			Orangutan vs. Human		
Divergence time (Myr)	5	6	7	7	8	9	12	14	16
ν ($\times 10^9$)	0.62	0.52	0.44	0.76	0.67	0.59	1.02	0.87	0.76
μ ($\times 10^4$) ^a	1.11	0.93	0.80	1.37	1.20	1.07	1.83	1.57	1.37
θ ($N_e = 10,000$) ^b	4.46	3.72	3.19	5.50	4.81	4.28	7.31	6.27	5.48
N_e (W) ^c	19,200	23,000	26,800	15,500	17,800	20,000	11,700	13,600	15,600
N_e (T) ^d	11,800	14,100	16,500	9,600	10,900	12,300	7,200	8,400	9,600
N_e (BLUE) ^e	14,100	16,900	19,700	11,500	13,100	14,700	8,600	10,000	11,500

^a μ = mutation rate per sequence per generation, assuming 20 years per generation.

^b θ was estimated as $\theta = 4N_e\mu$, with $N_e = 10,000$.

^c The N_e value estimated as $\theta/4\mu$ using the θ value (8.55) estimated by Watterson's (1975) method.

^d The N_e value estimated by $\theta/4\mu$ using the θ value (5.26) estimated by Tajima's (1983) method.

^e The N_e value estimated by $\theta/4N$ using θ value 6.30 estimated by Fu's (1994) BLUE method.

sequences, 1.07% between human and gorilla sequences, and 2.44% between human and orangutan sequences; here, we exclude the four exons. The mutation rates (ν) were estimated to be 0.52×10^{-9} , 0.67×10^{-9} , and 1.02×10^{-9} per nucleotide site per year based on divergence times of 6 Myr between humans and chimpanzees, 8 Myr between humans and gorillas, and 12 Myr between humans and orangutans, respectively; other divergence dates are also considered in table 7. The first two values are considerably smaller than the third, but the differences may be largely due to stochastic fluctuations. The average for the three estimates is 0.74×10^{-9} . This value is considerably lower than the estimate (1.15×10^{-9}) obtained from the 10-kb region on chromosome 22, but it is consistent with the lower nucleotide diversity in this region than in the 10-kb region on chromosome 22. It is possible that this region has a lower (neutral) mutation rate.

Several methods are available for estimating the parameter $\theta = 4N_e\mu$, where $\mu = \nu gL$, g is the generation time (20 years), and $L = 8,991$ bp (see eq. 1). If we use the commonly used value 10,000 for N_e (Takahata 1993), θ varies with the assumption of divergence dates (table 7). For the average mutation rate obtained above, we obtained $\theta = 5.32$. For the two commonly used methods, one known as Watterson's (1975) estimator and the other as Tajima's (1983) estimator, we obtained

$\theta = 8.55$ and 5.26 , respectively. Note that these two methods are not optimal in terms of minimizing variance. The minimum variance estimator (BLUE) by Fu (1994) gave $\theta = 11.50$. BLUE and Watterson's (1975) estimator yielded larger θ values, mainly because there was an excess of singletons (table 1). To avoid the effect of the excess of low-frequency variants, we excluded singletons and doubletons from analysis and obtained the BLUE estimate of $\theta = 6.30$, which is similar to the other two estimates.

If we know the mutation rate, we can estimate the effective population size N_e from θ . As the estimate of the mutation rate varies with the assumption of the divergence dates, the estimate of N_e also varies (table 7). Moreover, it also depends on the estimation methods used (table 7). If we use the average mutation rate obtained above and the average (6.70) of the θ values estimated by Watterson's, Tajima's, and the BLUE methods, we obtain $N_e = 12,600$, which is not far from the commonly used value of 10,000 (Takahata 1993).

Age of the MRCA

To estimate the age (T) of the MRCA of the sequences in a sample, the values of both N_e and mutation rate per sequence per generation (u) are required. As mentioned above, the estimate of mutation rate depends on the species pair and the divergence dates used. For simplicity, we shall use the average mutation rate obtained above, i.e., $\nu = 0.74 \times 10^{-9}$ changes per site per year and $u = 1.33 \times 10^{-4}$ changes per sequence per generation in humans; the estimate of T increases with decreasing u . Table 8 presents the estimates of T for several values of effective population sizes for the entire sample, the subsample of sequences from Africa only, and the subsample of non-African sequences only, respectively. If the commonly used $N_e = 10,000$ was assumed, the mode estimate (T_{mode}) and mean estimate (T_{mean}) were, respectively, 1,376,000 and 1,559,000 years for the entire sample. These estimates were comparable to our previous estimates based on the polymorphism data from a 10-kb region on chromosome 22 (Zhao et al. 2000). The estimates based on the African

Table 8
The Age (T , 10^3 years) of the MRCA of the Human Sequences Sampled

Sequences	N_e	T_{mode}	T_{mean}	95% Interval
All samples	10,000	1,472	1,503	728~2,392
	12,000	1,238	1,383	643~2,314
	15,000	1,080	1,229	564~2,136
Africans	6,000	1,061	1,096	523~1,795
	8,000	922	1,039	480~1,766
	10,000	832	987	448~1,712
Non-Africans	6,000	677	805	341~1,445
	7,000	655	779	330~1,422
	8,000	672	757	320~1,382

NOTE.—An average mutation rate of 1.33×10^{-4} per sequence per generation was used.

sample were only somewhat smaller than those based on the entire sample, while those based on the non-African sample were the smallest. This pattern was also consistent with our previous study (Zhao et al. 2000).

It should be pointed out that the method used to estimate the age of the MRCA makes the assumption that the sample was taken from a random-mating population with a constant effective population size. However, the fact that there are many unique variants in each of the three continents (see above) suggests that this assumption may not be appropriate. The existence of more singletons than expected would tend to inflate our estimates of the age of the MRCA. Methods that make a better use of segregating sites of different types are needed for correcting such biases, and one of us is in the process of developing such methods. Note that the alternative method developed by Griffiths and Tavaré (1994) is not applicable here because we do not have haplotype sequences. Nevertheless, it is unlikely that the revised estimates will be smaller than half of the values given in table 8.

In summary, like the data from the PDHA1 locus (Harris and Hey 1999) and the 10-kb region on chromosome 22 (Zhao et al. 2000), the present data also provide evidence for a genetic history of humans that is much more ancient than the emergence of modern humans. The observation that both the region on chromosome 22 and the present region show an ancient genetic history outside of Africa argues against a complete replacement of all indigenous populations in Europe and Asia by an African stock. Moreover, the ancient genetic history of humans indicates no severe bottleneck during the evolution of humans in the last half million years, because much of the ancient genetic history would have been lost during a severe bottleneck. On the other hand, the fact that most available nuclear DNA variation data, as well as mitochondrial DNA data, show a considerably shallower genetic history in Asia and Europe than in Africa suggests that human evolution has not occurred in parallel in different parts of the Old World, as depicted by the multiregional model. Thus, both the Out of Africa and the multiregional models appear to be too simple to explain the evolution of modern humans.

Acknowledgments

We thank Drs. J. B. Clegg, Marie Lin, and Maryellen Ruvolo for DNA samples. This work was supported by NIH grants GM55759 (W.-H.L.), GM50428 (Y.-X.F.), GM59290 (L.B.J.) and NSF DEB9707567 (Y.-X.F.).

LITERATURE CITED

- 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.
- CANN, R. L., M. STONEKING, and A. C. WILSON. 1987. Mitochondrial DNA and human evolution. *Nature* **325**:31–36.
- CHARLESWORTH, B. 1994. The effect of background selection against deleterious alleles on weakly selected, linked variants. *Genet. Res.* **63**:213–228.
- 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.
- FU, Y. X. 1994. Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. *Genetics* **138**:1375–1386.
- . 1996. Estimating the age of the common ancestor of a DNA sample using the number of segregating sites. *Genetics* **144**:829–838.
- FU, Y. X., and W. H. LI. 1997. Estimating the age of the common ancestor of a sample of DNA sequences. *Mol. Biol. Evol.* **14**:195–199.
- . 1993. Statistical tests of neutrality of mutations. *Genetics* **133**:693–709.
- FULLERTON, S. M., J. BOND, J. A. SCHNEIDER, B. HAMILTON, R. M. HARDING, A. J. BOYCE, and J. B. CLEGG. 2000. Polymorphism and divergence in the β -globin replication origin initiation region. *Mol. Biol. Evol.* **17**:179–188.
- GRIFFITHS, R. C., and S. TAVARÉ. 1994. Ancestral inference in population genetics. *Stat. Sci.* **9**:307–319.
- HARDING, R. M., S. M. FULLERTON, R. C. GRIFFITHS, J. BOND, M. J. COX, J. A. SCHNEIDER, D. S. MOULIN, and J. B. CLEGG. 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.
- JARUZELSKA, J., E. ZIETKIEWICZ, M. BATZER, D. E. C. COLE, J.-P. MOISAN, R. SCOZZARI, S. TAVARE, and D. LABUDA. 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.
- JARUZELSKA, J., E. ZIETKIEWICZ, and D. LABUDA. 1999. Is selection responsible for the low level of variation in the last intron of the ZFY locus? *Mol. Biol. Evol.* **16**(11):1633–1640. *J. Mol. Evol.* **21**:58–71.
- KAESSMANN, H., F. HEIßIG, A. VON HAESLER, and S. PÄÄBO. 1999. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nat. Genet.* **22**:78–81.
- LI, W. H. 1997. *Molecular evolution*. Sinauer, Sunderland, Mass.
- LI, W. H., and L. SADLER. 1991. Low nucleotide diversity in man. *Genetics* **129**:513–523.
- LI, W. H., C.-I. WU, and C.-C. LUO. 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implication. *J. Mol. Evol.* **21**:58–71.
- NICKERSON, D. A., S. L. TAYLOR, K. M. WEISS, A. G. CLARK, R. G. HUTCHINSON, J. STENGARD, V. SALOMAA, E. VARTAINEN, E. BOERWINKLE, and C. F. SING. 1998. DNA sequence diversity in a 9.7-kb region of the human lipoprotein gene. *Nat. Genet.* **19**:233–240.
- RIEDER, M. J., S. L. TAYLOR, A. G. CLARK, and D. A. NICKERSON. 1999. Sequence variation in the human angiotensin converting enzyme. *Nat. Genet.* **22**:59–62.
- ROZAS, J., and R. ROZAS. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**:174–175.

- STRINGER, C. B., and P. ANDREWS. 1988. Genetic and fossil evidence for the origin of modern humans. *Science* **139**: 1263–1268.
- TAJIMA, F. 1983. Evolution relationship of DNA sequences in finite populations. *Genetics* **105**:437–460.
- . 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**:585–595.
- TAKAHATA, N. 1993. Allelic genealogy and human evolution. *Mol. Biol. Evol.* **10**:2–22.
- WATERSON, G. A. 1975. On the number of segregation sites. *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.
- ZHAO, Z., J. LI, Y.-X. FU et al. (13 co-authors). 2000. Worldwide DNA sequence variation in a 10 kb noncoding region on human chromosome 22. *Proc. Natl. Acad. Sci. USA* **97**: 11354–11358.
- ZIETKIEWICZ, E., V. YOTOVA, M. JARNIK, et al. (11 co-authors). 1998. Genetic structure of the ancestral population of modern humans. *J. Mol. Evol.* **47**:146–155.

KEITH CRANDALL, reviewing editor

Accepted October 9, 2000