Divergent Haplotypes and Human History as Revealed in a Worldwide Survey of X-Linked DNA Sequence Variation

Makoto K. Shimada,† Karuna Panchapakesan,‡ Sarah A. Tishkoff,‡ Alejandro Q. Nato Jr,* and Jody Hey**

*Department of Genetics, Rutgers University; †Japan Biological Information Research Center, Japan Biological Informatics Consortium, Koto-ku, Tokyo, Japan; and ‡Department of Biology, University of Maryland

The population genetic history of a 10.1-kbp noncoding region of the human X chromosome was studied using the males of the HGDP-CEPH Human Genome Diversity Panel (672 individuals from 52 populations). The geographic distribution of patterns of variation was roughly consistent with previous studies, with the major exception that 1 highly divergent haplotype (haplotype *X*, *hX*) was observed at low frequency in widely scattered non-African populations and not at all observed in sub-Saharan African populations. Microsatellite (short tandem repeat) variation within the sequenced region was low among copies of *hX*, even though the estimated time of ancestry of *hX* and other sequences was 1.44 Myr. The estimated age of the common ancestor of all *hX* copies was 5,230 years (95% consistency index: 2,000–75,480 years). To further address the presence of *hX* in Africa, additional samples from Chad and Tanzania were screened. Five additional copies of *hX* were observed, consistent with a history in which *hX* was present in Africa prior to the migration of modern humans out of Africa and with eastern Africa being the source of non-African modern human populations. Taken together, these features of *hX*—that it is much older than other haplotypes and uncommon and patchily distributed throughout Africa, Europe, and Asia—present a cautionary tale for interpretations of human history.

Introduction

Many questions on human evolution depend on the development of a full historical demographic portrait, including answer to questions on where our ancestors lived at different times and how many there were and of the timing and magnitude of migrations of populations to new regions and environments. All of these questions can, in principle, be addressed with data on genetic variation and with population genetic analyses. This is true even though humans are not highly variable at the DNA level. The origin of modern humans and their spread throughout the world were comparatively recent events, in evolutionary terms, and the gene trees for many loci extend to times long before these events (Tishkoff and Verrelli 2003) and thus afford us the potential for developing a population genetic portrait of our ancestors.

Genetic studies of the history of modern human populations are largely consistent with a model of recent African origin (Cavalli-Sforza et al. 1994; Rosenberg et al. 2002; Tishkoff and Verrelli 2003; Ramachandran et al. 2005). However, human genes vary considerably in their individual histories (Przeworski et al. 2000), suggesting a history of our species that includes not only the formation of new populations but also migration among populations (Richards et al. 2003; Kivisild et al. 2004) and admixture among divergent populations (Hammer and Horai 1995; Labuda et al. 2000; Harding and McVean 2004; Garrigan, Mobasher, Kingan, et al. 2005; Garrigan, Mobasher, Severson, et al. 2005; Evans et al. 2006).

If we are to discern this complex history on a global scale, then we require data from a very large number of individuals and populations. It is also necessary that the loci under study not be strongly affected by ascertainment issues. Here we meet both of these critieria in a DNA sequence–based population genetic study of a 10-kbp (kilo basepair) region of the human X chromosome. The sample set is the Human Genome Diversity Panel of the Centre d'Etude du Polymorphisme Humain (HGDP-CEPH), which has been made available to investigators in hopes that a common sample set might be studied by different investigators using different loci (Cann et al. 2002). The sequenced region is noncoding and shows little evidence of recombination. The region also contains 2 short tandem repeat (STR) loci and so offers the opportunity to directly contrast STR variation with flanking sequence haplotypes that lie upon the same historical genealogy.

Key words: human populations, population genetics, microsatellite (STR), HGDP-CEPH Diversity Panel, sequence variation, population structure, coalescent.

E-mail: hey@biology.rutgers.edu.

Mol. Biol. Evol. 24(3):687–698. 2007
doi:10.1093/molbev/msl196
Advance Access publication December 14, 2006

Materials and Methods

DNA Samples

Six hundred and seventy-two male individuals were selected from the HGDP-CEPH DNA Panel (Cann et al. 2002), including 6 populations from sub-Saharan Africa, 4 populations from Middle East and North Africa, 8 populations from Europe, 8 populations from Central and South Asian, 18 populations from East Asian populations, 2 populations from Oceania, and 5 populations from New World, as detailed in table 1. The full HGDP-CEPH set includes 685 male samples; however, only 678 are independent because of duplicate samples (Mountain and Ramakrishnan 2005). The total sample size of the study was constrained to what could be accommodated in seven 96-well plates. Therefore, 6 individuals selected at random were not included.

The entire HGDP-CEPH has been the subject of studies involving hundreds of STR loci (Rosenberg et al. 2002; Zhivotovsky et al. 2003; Ramachandran et al. 2004, 2005). Different types of analyses on these data, including clustering on the basis of departures from Hardy–Weinberg (Rosenberg et al. 2002) and using principal components (Zhivotovsky et al. 2003), found that the data were consistent with a model of 5 major geographic regions, including Africa, Eurasia (including the Middle East and North Africa, as well as Central and South Asia), East Asia, Oceania, and the Americas. Several of our analyses make use of these same regional designations.

To identify the distribution of haplotype *X* (*hX*) across Africa, 3 portions of the Xp11.22 region that contained

Table 1
Polymorphism Levels by Population

Population	Region[a]	n[b]	$\hat{\theta}$/bp[c]	π/bp[d]	Tajima's D[e]	STR1[f] $(\delta u)^2$	STR 2[f] $(\delta u)^2$
Bantu, Kenya	Africa	11	0.00031	0.00032	0.549	19.6	2.6
Bantu, South Africa	Africa	8	0.00045	0.00047	0.098	2.3	3.9
Biaka, Pygmy	Africa	27	0.00044	0.00041	−0.351	9.1	7.4
Mandenka, Senegal	Africa	16	0.00058	0.00043	−1.023	9.3	4.5
Mbuti, Pygmy	Africa	13	0.00029	0.00027	−0.053	9.5	1.4
San, Namibia	Africa	7	0.00008	0.00012	1.099	2.3	15.9
Yoruba, Nigeria	Africa	13	0.00036	0.00039	0.706	28.2	1.8
Adygei, Caucasus	Europe	7	0.00026	0.00032	0.919	61.9	1.1
French	Europe	12	0.00130	0.00073	−3.062*	35.8	0.5
French Basque, France	Europe	16	0.00026	0.00027	0.073	51.2	0.7
Northern Italian	Europe	8	0.00004	0.00006	0.792	3.4	1.1
Orcadian	Europe	7	0.00148	0.00108	−2.331*	48.7	1.9
Russian	Europe	16	0.00021	0.00026	0.601	52.0	1.7
Sardinian, Italy	Europe	16	0.00023	0.00030	0.741	48.8	1.3
Tuscan, Italy	Europe	6	0.00023	0.00026	0.845	48.3	3.7
Bedouin, Israel	Middle East	28	0.00112	0.00078	1.274	33.6	1.6
Druze, Israel	Middle East	14	0.00120	0.00070	−3.360*	45.2	0.3
Mozabite, Algeria	Middle East	20	0.00121	0.00060	−3.603*	30.0	1.2
Palestinian, Israel	Middle East	17	0.00027	0.00023	0.375	34.4	0.1
Balochi, Pakistan	Central and South Asia	25	0.00028	0.00030	−1.34	46.6	0.7
Brahui, Pakistan	Central and South Asia	25	0.00025	0.00029	−0.993	45.5	0.4
Burusho, Pakistan	Central and South Asia	20	0.00027	0.00023	−0.858	38.2	0.3
Hazara, Pakistan	Central and South Asia	24	0.00017	0.00022	0.485	34.4	0.9
Kalash, Pakistan	Central and South Asia	20	0.00108	0.00054	−3.871*	37.2	0.3
Makrani, Pakistan	Central and South Asia	19	0.00016	0.00024	1.237	44.5	0.5
Pathan, Pakistan	Central and South Asia	19	0.00107	0.00055	−4.122*	44.2	0.6
Sindhi, Pakistan	Central and South Asia	21	0.00027	0.00025	0.297	35.4	1.7
Uygur, China	Central and South Asia	8	0.00028	0.00020	0.405	21.1	0.6
Melanesian	Oceania	8	0.00128	0.00081	−2.453*	12.3	1.1
Papuan, New Guinea	Oceania	13	0.00028	0.00031	0.916	39.6	0.2
Cambodian	East Asia	6	0.00029	0.00033	0.389	56.6	0.0
Dai, China	East Asia	7	0.00004	0.00005	0.831	1.2	0.0
Daur, China	East Asia	7	0.00025	0.00029	0.919	53.9	0.3
Han, China	East Asia	24	0.00021	0.00017	−0.091	29.3	0.0
Hezhen, China	East Asia	4	0.00034	0.00031	0.782	44.0	0.0
Japanese	East Asia	20	0.00015	0.00019	1.229	34.6	0.0
Lahu, China	East Asia	7	0.00009	0.00005	−1.593	2.0	0.3
Miaozu, China	East Asia	7	0.00008	0.00008	−0.247	3.6	0.0
Mongola, China	East Asia	7	0.00030	0.00029	−0.081	40.6	0.0
Naxi, China	East Asia	8	0.00021	0.00020	0.689	38.3	0.0
Oroqen, China	East Asia	7	0.00023	0.00019	1.482	26.3	0.0
She, China	East Asia	7	0.00022	0.00017	1.482	34.3	0.0
Tu, China	East Asia	7	0.00009	0.00009	−0.247	0.6	0.3
Tujia, China	East Asia	9	0.00019	0.00013	1.393	23.5	0.5
Xibo, China	East Asia	8	0.00020	0.00025	0.689	40.5	1.0
Yakut, Siberia	East Asia	18	0.00018	0.00021	0.566	30.7	0.0
Yizu, China	East Asia	9	0.00024	0.0002	0.138	32.0	0.0
Colombian	America	5	0.00015	0.00017	1.448	81.6	0.0
Karitiana, Brazil	America	10	0.00004	0.00002	−1.352	0.8	0.0
Maya, Mexico	America	3	0	0	n.a.	4.7	0.0
Pima, Mexico	America	12	0.00025	0.00018	0.829	21.6	0.0
Surui, Brazil	America	9	0.00020	0.00028	1.393	84.0	0.0

NOTE.—n.a., not available.

[a] Regions correspond to the 7 population regions identified by Rosenberg et al. (2002).

[b] The number of sequenced samples.

[c] Watterson's estimate (Watterson 1975) of the population mutation rate for an X-linked locus, 3 Nu, per basepair.

[d] Nucleotide diversity (Nei and Tajima 1981) per basepair.

[e] Tajima (1989).

[f] Estimate of the population mutation rate for microsatellites (Goldstein et al. 1995).

* $P < 0.05$.

diagnostic variants of *hX* were sequenced in East African samples. Samples included 95 male Tanzanians (19 individuals from each of the following tribes: Maasai, Sandawe, Turu, Hadza, and Burunge) as well as 5 (4 males and 1 female) from the Laka tribe in Chad.

Selection of a Sequencing Region

A region of the X chromosome was selected for resequencing based on the following criteria: 1) it should be noncoding, to help avoid the effects of natural selection; 2) recombination should be low, so that gene tree and coalescent-based analyses could be applied; 3) it should contain one or more STR sites; and 4) it should not be atypical with regard to repeat insertion or GC content. The criterion of low recombination necessarily represented a tradeoff—permitting a wider array of analyses, but increasing the chance that selection on linked regions could have affected the history. Approximate levels of recombination along the X chromosome were assessed using single nucleotide polymorphism (SNP) data for African populations (Clark et al. 2003).

We selected the region corresponding to bases 50,583,087–50,593,170 on National Center for Biotechnology Information (NCBI) Build 36.1, which lies within Xp11.22. This region does not contain any known gene but has 26 repeats composed of 11 short interspersed nuclear (SINE) elements, 7 portions of long interspersed nuclear (LINE) elements, 3 portions of long terminal repeat (LTR) elements, 3 DNA transposons, and 2 STRs (Kent et al. 2002). The region was divided into 22 segments for direct sequencing. Polymerase chain reaction (PCR) amplification and bidirectional sequencing were performed by Genaissance Pharmaceuticals, Inc. (New Haven, CT).

Assembling and Alignment of Sequence Data

Sequence data were trimmed, assembled, and aligned using phred/phrap/consed/polyphred (Nickerson et al. 1997; Ewing and Green 1998; Ewing et al. 1998; Gordon et al. 1998) coupled with automated shell scripts and Perl programs made for this study. Sequence with data quality value (QV) PHRED scores (Ewing et al. 1998) lower than 35 were not used. Singleton SNPs (i.e., sites where a single individual is different from all others) with QV scores lower than 65 were not used. This corresponds to an expected error rate of 3.16×10^{-7} or about one mistake per 3 million basepairs. Thus, the entire data set of 6.7 mega basepairs (Mbp) would be expected to have about 2 false singleton SNPs. The GenBank accession numbers are AB268589–AB269260.

STR Genotyping

We genotyped 2 dinuclotide repeat STR regions at chrX: 50,585,063–505,85,102 (STR2) and chrX: 50,588,637–50,588,675 (STR1) on NCBI Build 36.

For STR2, the primer sequences were as follows: forward primer, STR2L (5′-TCTCAGGATCTTGGCTTCG T-3′) and reverse primer, STR2R (5′-CCAACCACCACA-GATAGCAG-3′). For STR1, they were as follows: forward primer, STR1–2L (5′-TGGGCAGCAAAAGCACT AT-3′) and reverse primer, STR1–2R (5′-GCCTGGAT-TTGGCTTTCAG-3′).

The STR1–2R and STR2L primer sequences also included an M13 sequence at the 5′ end. Using IRD41 dye–labeled M13 forward and reverse primers (LI-COR, Inc., Lincoln, NE), the double loci amplifications were carried out in a final reaction volume of 20 μl according to the following PCR cycling conditions: 95 °C for 2 min, 33 cycles at 95 °C for 30 s, 59 °C for 30 s, and 72 °C for 30 s, plus a final extension at 72 °C for 5 min. Amplification products were run in an LI-COR automated electrophoresis unit, model 4200 under the Global IR^2 system. The Gene ImagIR and SAGA analysis software were used to collect and analyze fragment size. STR genotypes are available as Supplementary Material online.

Definition of DNA Sequence Haplotypes

In order to define haplotypes, in a context where some data are missing, we follow the concept of ambiguous and unambiguous haplotypes proposed by Patil et al. (2001) and adopted by Zhang et al. (2002, 2003). In this framework 2 haplotypes are incompatible if they differ at one or more base positions at which neither is missing data, and a haplotype is considered ambiguous if it is compatible (because of missing data) with 2 or more other haplotypes that are themselves incompatible.

In DNA sequence–based population genetic studies, the question arises of how best to handle data at base positions for which a subset of sequences are not resolved. The simple solution of not including such sites becomes less practical in proportion to the number of sequences in the study. This is because with more sequences more sites will be invalidated by virtue of ambiguity in at least one sequence. Alternatively, ambiguous bases can just be scored individually as missing data. Under an assumption that missing data (or data with low QV scores) are a random sample of all data with respect to the levels and patterns of DNA sequence variation, and so long as each sample in the data set is represented by a substantial portion of the sequenced region, most analyses will not be affected by missing data. This is the approach used here. In this study, approximately 9% of the samples (58) are missing data for a portion of one of the 22 PCR segments because of a failed sequencing reaction on one or both strands. The locations and frequencies of unresolved bases are provided in Supplementary Material online.

Resequencing studies with large numbers of samples do offer an advantage for resolving previously unknown SNPs. This is because the false positive rate of a SNP is roughly the error rate for individual basepairs raised to the power of the number of times the low-frequency base was observed. SNPs for which the low-frequency base occurs more than once will have a very low false positive rate, and the proportion of such SNPs is higher for studies with large samples. For example, with a data set of n sequences, under an assumption of neutrality and constant population size, the proportion of SNPs that are singletons has an expected value of $1/\sum_{i=1}^{n-1} 1/i$ (Watterson 1975). If $n = 10$, this is 0.35, and if $n = 672$, it is 0.14 (although in this study the fraction of singletons turned out to be considerably higher). In fact, singleton SNPs can play a major role in analyses of changing population size, and for data sets with small numbers of sequences, their frequency strongly shapes statistics that are used as indicators of population size change (or selection). However, in large data sets, SNPs occur across a more finely grained frequency range and frequency-based statistics are not as sensitive to the numbers of singletons.

Identification of *hX* in East African Samples

PCR primers used to amplify a 314-bp product from the 5′ region (5A) containing diagnostic site (A/G at position 50,583,786 in NCBI Build 36.1) were 5′-TGT GGGTATTAGGCCAAAGC-3′ and 5′-CAATCTTTGA GCCAGCACAG-3′; primers used to amplify a 315-bp region containing the middle diagnostic site (MA) (TT/AG at position 50,588,247–50,588,248) were 5′-TGAATCAG-GACTCTTTTGCTTG-3′ and 5′-GGAACTTTCGCTGG-GACTTT-3′; and primers used to amplify a 436-bp region containing the 3′ diagnostic site (3A) (G/C at position 50,592,308) were 5′-TGGTGTGATCATGGCTCATT-3′ and 5′-GCCAGGGGTGTGACATTTAT-3′. A total of 50 ng of genomic DNA was amplified in a final reaction volume of 25 μl according to the following PCR cycling conditions: for the 5′ region—94 °C for 1 min, 14 cycles at 94 °C for 1 min, 62 °C for 38 s, and 72 °C for 1 min 30 s and then 20 cycles at 94 °C for 45 s, 57 °C for 40 s, and 72 °C for 1 min 30 s, plus a final extension at 72 °C for 5 min. For the middle region: 94 °C for 1 min, 30 cycles at 94 °C for 1 min, 60 °C for 1 min, and 72 °C for 1 min, plus a final extension at 72 °C for 5 min. For the 3′ region, 94 °C for 1 min, 30 cycles at 94 °C for 1 min, 62 °C for 1 min, and 72 °C for 1 min plus a final extension at 72 °C for 5 min. Amplification products were run on a 1.8% agarose gel. Primers 5A, MA, 3A were used for sequencing directly from PCR products using the ABI Big Dye v3.1 terminator kit, followed by SAP-EXO purification of products, which were run on a 3730*xl* automated sequencer (Applied Biosystems, Foster City, CA). Sequence files were aligned and SNPs identified using the Sequencher v. 4.0.5 program (Gene Codes, Ann Arbor, MI).

Haplotype Tree Estimation

We constructed a minimum evolution (ME) tree (Rzhetsky and Nei 1992) using the MEGA 3.0 program (Kumar et al. 2004) with a chimpanzee sequence as outgroup. The chimpanzee sequence was obtained through the University of California at Santa Cruz Genome Browser (Kent et al. 2002), from the Arachne draft assembly (Build 1 Version 1, November 2003, http://genome.ucsc.edu/goldenPath/credits.html#chimp_credits). We excluded chimpanzee sequence with sequence quality scores below 40. The total length of aligned chimpanzee sequence in the genome build with high QV scores was 7,075 bp. Distances were estimated using the Jukes–Cantor model (Jukes and Cantor 1969). A relative rate test (Tajima 1993) did not show significant difference between the longest and shortest branch lengths from the chimpanzee sequence ($P > 0.05$).

To estimate the time to the most recent common ancestral sequence, the ME tree was linearized and calibrated assuming a root time of 6 MYA (Chen and Li 2001; Brunet et al. 2002). If the Xp11.22 region actually has a more recent common ancestry time, as has been suggested for X-chromosomal genes (Patterson et al. 2006), then the estimated dates can be rescaled proportionately.

Diversity Measures

Measures of variation, including Watterson's estimator (Watterson 1975) and the average number of pairwise differences, or π (Nei and Tajima 1981), as well as Tajima's *D* (Tajima 1989) were calculated using the SITES program (Hey and Wakeley 1997). We used the Arlequin computer program (http://anthro.unige.ch/arlwquin) (Excoffier et al. 2005) to estimate F_{ST} and conduct a Mantel test of association between F_{ST} values for the Xp11.22 region and for a large STR data set (Rosenberg et al. 2002).

Coalescent Analysis of Common Ancestor Time

A Bayesian estimate of the time of common ancestry for all of the sequences and for particular haplotypes was conducted using the IM computer program (Hey and Nielsen 2004). This method implements a population divergence model, but can be readily adapted to a single population model. A high prior upper bound on the population mutation rate, $\theta = 3\,Nu$ for X-linked loci, was selected to exceed the estimated upper limit of the posterior density for θ (Hey and Nielsen 2004). The program was modified to record the common ancestor time for an entire sample, as well as for a particular subset of sequences of the same haplotype, assuming the infinite-site mutation model (Kimura 1969). By scaling by the divergence observed between humans and chimpanzees, the method generates posterior probability density estimates of the time of the most recent common ancestor (TMRCA). Four different locus models were considered: sequence alone under the infinite-site model, sequence and linked STR 1, sequence and linked STR 2, and sequence and both linked STR regions. The accommodation of models with perfectly linked loci of different mutation models is described in Hey et al. (2004). The method assumes that there has not been recombination during the history since the TMRCA. The 4-gamete test was used to check for the presence of recombination (Hudson and Kaplan 1985).

Results

Among the 672 sequences, we identified 106 SNPs. Indel variation was low, with nearly all length variations associated with single- or 2-base changes, often associated with short runs of repeats or with 1 of the 2 STR regions. The longest unambiguous indel (i.e., away from an STR region) was 9 bp and occurred in just 1 sequence from the Makrani sample from Pakistan.

The sequence data revealed almost no sign of recombination in the history of the sample. The 4-gamete test of Hudson and Kaplan (1985) suggested just a single recombination event across the entire data set. The total number of distinct haplotypes was 67, and there were 100 sequences for which the haplotype was ambiguous because of an ambiguous base in a position that distinguished one or more haplotypes. There are 7 haplotypes (*h18*, *h21*, *h26*, *h33*, *h39*, *h40*, and *hX*) that occurred in more than 1 geographic region. Three of these haplotypes were observed many times, with 95, 119, and 219 copies for *h33*, *h39*, and *h40*, respectively, for a total of 433, or 75.7% of all unambiguous haplotypes. One unusual haplotype differed at 27 positions from all other haplotypes. Designated *hX*, this haplotype was found once in Melanesia (Oceania) and 8 times in widespread locations in Eurasia, including the Orkney Islands, Pakistan, Algeria, Israel, and France. Contrary to the typical pattern found in many genes, in which the variation in non-African populations is a subset of the variation in sub-Saharan Africa, *hX* was not observed among

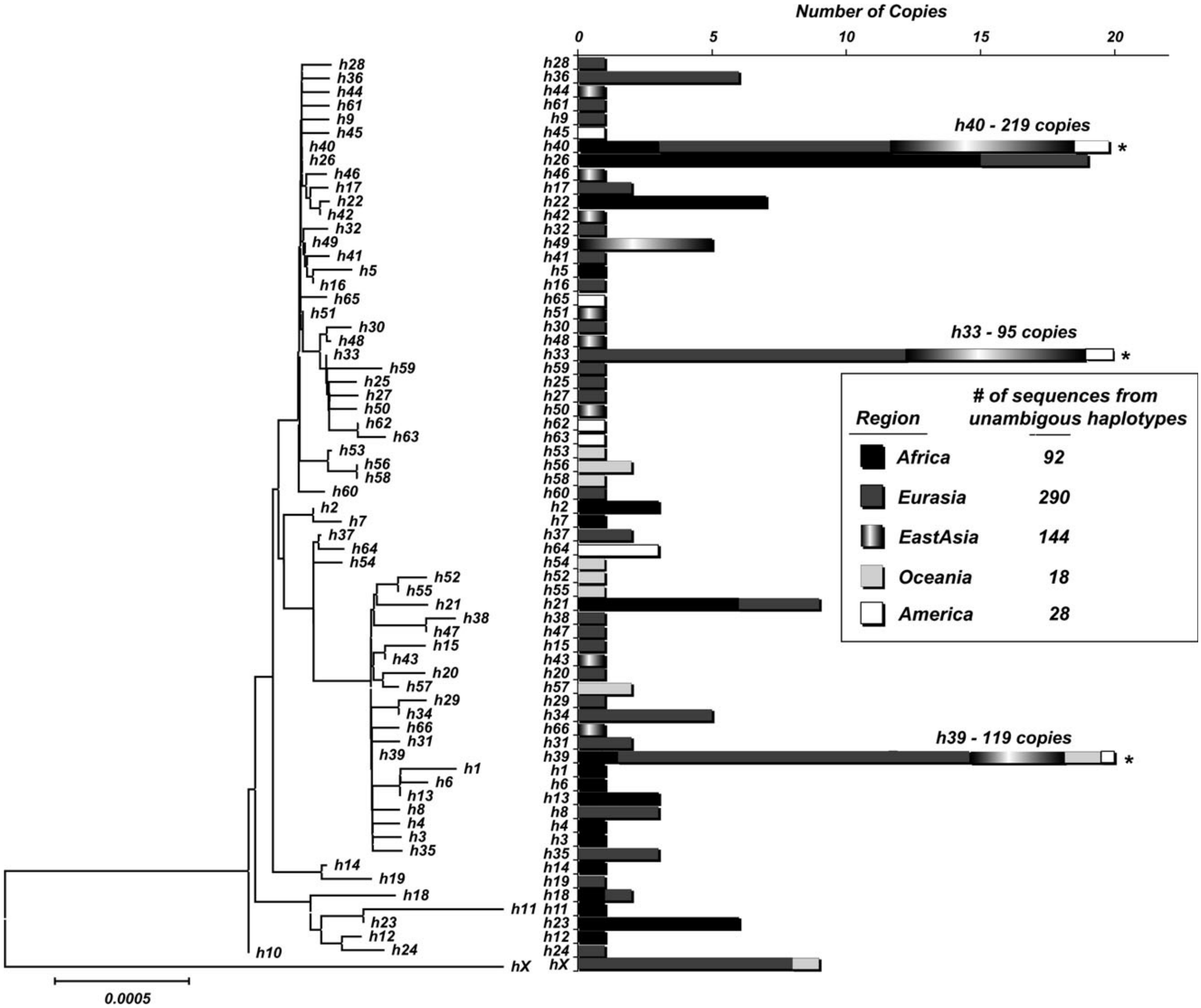

FIG. 1.—A ME tree (Rzhetsky and Nei 1992) of haplotypes showing the numbers for different geographic regions. For clarity, we have grouped Europe together with the Middle East and North Africa as well as Central and South Asia, to make a single Eurasian group, as suggested by Rosenberg et al. (2002).

the HGDP-CEPH males from sub-Saharan Africa (98 individuals).

Figure 1 shows the gene tree estimate for the 67 unambiguous haplotypes, together with their frequencies and geographic regions in which they occurred. The average of Jukes–Cantor distances between the chimpanzee and the human haplotypes is 0.0133 changes per site, which corresponds to a mutation rate of 1.11×10^{-9} per year per site, assuming a 6 MYA divergence since human/chimpanzee common ancestry. Based on divergence from the Chimpanzee sequence, the estimated date of ancestry of the human haplotypes (assuming a molecular clock) is 1.44 MYA. The large majority of this time depth is due to the presence of *hX*. Without the *hX*, the TMRCA is about 0.23 Myr (fig. 1).

Overall, the predominant pattern is one of a few common widespread haplotypes, with many rare haplotypes that are very similar to the few common widespread haplotypes. Haplotype *h33* is widespread outside of Africa, but this and related sequences were not observed in the sub-Saharan African samples. However, *h33* is quite similar to the most common haplotype, *h40*, which is widespread both within and outside of Africa. In contrast, *hX* that is highly divergent from all other haplotypes was not observed in sub-Saharan African samples and was found at a low frequency in populations where it was observed even though those populations with *hX* are widely dispersed. The presence of *hX* creates a highly unbalanced tree with just 9 gene copies on one side of the basal node (fig. 1).

Measures of Variability

Table 1 summarizes the variation in each of 52 populations, and figure 2 shows the values of 2 common estimators of the population mutation rate, both with and without inclusion of *hX* for 7 geographic regions. As is common for many loci in human populations, there tends to be an excess of low-frequency variation causing an elevation of $\hat{\theta}$ relative to π (Tishkoff and Verrelli 2003). This can be caused by population growth as well as by some patterns of population structure. When *hX* is excluded

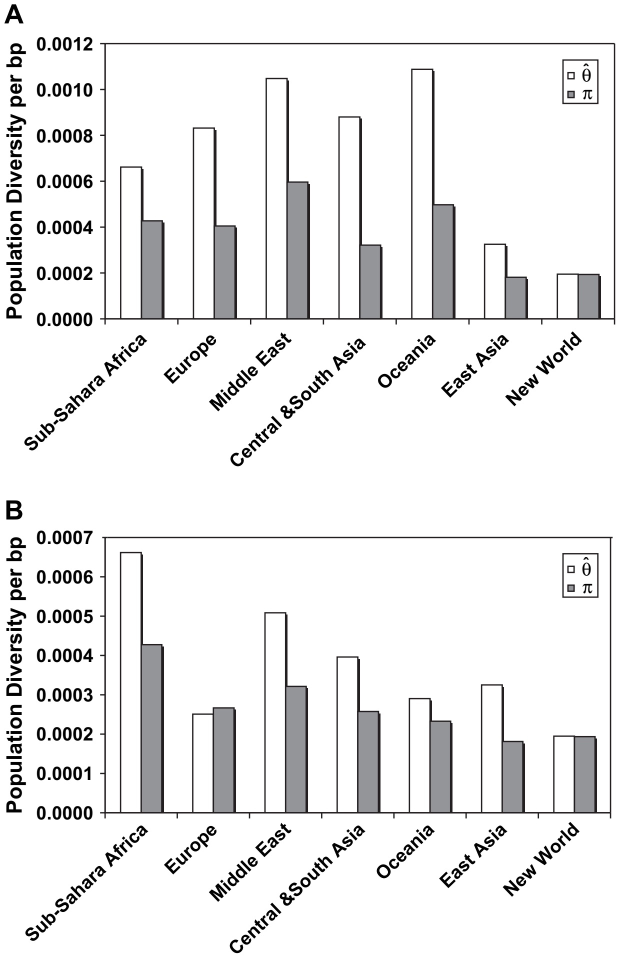

FIG. 2.—Sequence diversity by region. Estimates of the population mutation rate, based on Watterson's estimator $\hat{\theta}$ (Watterson 1975) and nucleotide diversity π (Nei and Tajima 1981). (*A*) All sequences were included for each of the 7 geographic regions of Rosenberg et al. (2002). (*B*)The highly divergent haplotype, *hX*, was excluded from these estimates.

(fig. 2*B*), sub-Saharan Africa is more variable than other regions of the world; however, the pattern changes markedly when *hX* is included (fig. 2*A*), in which case some other regions are more variable than sub-Saharan Africa.

The elevated counts of low-frequency polymorphic sites, which are reflected in the differences between $\hat{\theta}$ and π, are also reflected in negative values of Tajima's *D* statistic (Tajima 1989), as seen in figure 3. The significant values of Tajima's *D* for individual populations (table 1) are due to the presence in those population samples of single copies of *hX*.

STR Typing

The number of repeats in the 2 STR loci in all samples examined is shown in the Supplementary Material online. STR 1 had an average repeat number of 17.3, whereas the average for STR 2 was 19.0. Frequency distributions for the 2 STRs are shown in figure 4, and it can be seen that STR 1 is much more variable (see also table 1). Figure 4 also shows the STR allele frequencies observed in 9 copies of *hX*. Despite the apparent age of *hX*, only 3 different alleles were observed at STR 1 and only 2 different alleles were observed at STR 2. For neither STR does *hX* have the same modal allele as found in the entire population sample.

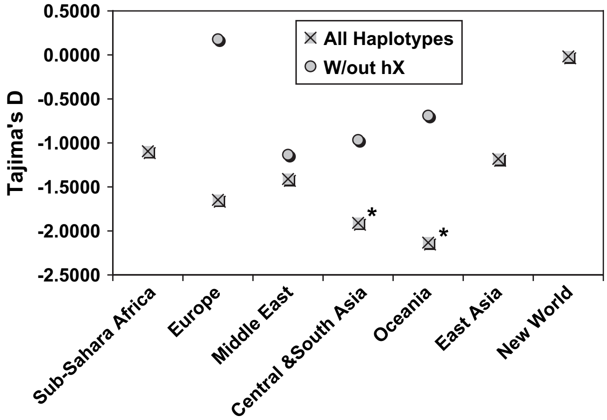

FIG. 3.—Tajima's *D* by region is shown for each of the 7 geographic regions for all sequences and with *hX* excluded from the 4 regions where *hX*s were observed. "*" indicates the values of Tajima's *D* that were statistically significant at the $P < 0.05$ level based on table 2 of Tajima (1989).

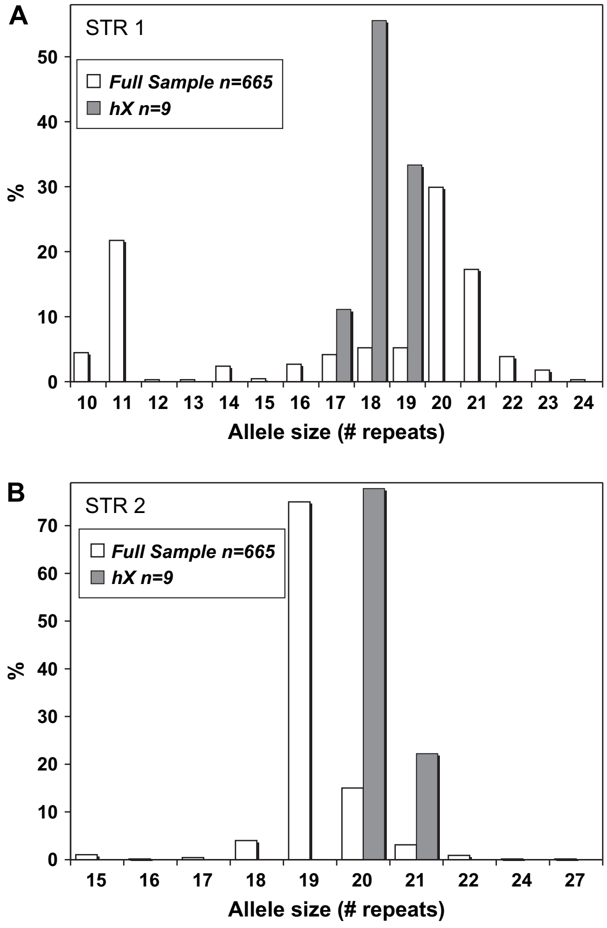

FIG. 4.—Frequency distribution of STR alleles, by number of repeat units. Frequencies are shown for both the entire sample, as well as for just the 9 individuals carrying *hX*.

Table 2
F_{ST} Estimates between Pairs of Geographic Regions

	Africa	Europe	Middle East	Central South Asia	Oceania	East Asia	America
Africa		0.145	0.125	0.164	0.248	0.266	0.290
Europe	0.045		0.016	0.007	0.063	0.054	0.145
Middle East	0.029	0.000		0.022	0.143	0.065	0.151
Central South Asia	0.051	0.004	0.000		0.113	0.038	0.114
Oceania	0.144	0.110	0.091	0.178		0.220	0.432
East Asia	0.089	0.084	0.032	0.038	0.327		0.070
America	0.098	0.079	0.034	0.034	0.325	0.000	

NOTE.—Estimates from STRs shown above the diagonal. Estimates from DNA sequences shown below the diagonal.

SNP Database Search

We examined the correspondence of a public SNP database dbSNP (b126) (http://www.ncbi.nlm.nih.gov/SNP/) to the SNPs discovered in this study using the NBCI Build 36.1 of the human genome and the UCSC genome browser. In our sequenced region, 24 RefSNP Clusters, including 4 rsIDs representing virtually identical indel sites, were recorded, but 2 (rs17249475, rs5961152) were not observed in our data. It is noteworthy that 4 rsIDs (rs17249482, rs17315387, rs17315394, and rs17249496) are unique SNPs caused by *hX*, whose variant allele frequencies in the European panel (PERLEGEN: AFD_EUR_PANEL, $n = 24$) were 0.043, 0.043, 0.045, and 0.062, respectively, but were not found in the African panel (PERLEGEN: AFD_AFR_PANEL, $n = 23$) or the Chinese panel (PERLEGEN: AFD_CHN_PANEL, $n = 24$). This pattern is consistent with our data in terms of allele frequencies in global populations, that is, we observed 2 copies of *hX* in 66 European chromosomes (0.03), but none in the African or East Asian samples of the HGDP-CEPH panel. The HapMap project (Rel #21, on NCBI B35, dbSNP b125) (http://hapmap.org/) genotyped 6 SNPs (rs5915321, rs2226047, rs17315373, rs6614353, rs5961209, and rs5961210) in this region, although these do not contain any *hX* unique SNPs. The HapMap data suggest strong linkage disequibrium among these 6 SNPs in all 4 populations included in the HapMap project. In particular, $|D'| = 1$ and $r^2 = 1$ in all pairs among the 3 SNPs with minor allele frequencies greater than 17% (i.e., rs5915321, rs5961209, and rs5961210), including the pair having the greatest separation distance (i.e., 5,930 bp between rs5915321 and rs5961210).

Population Structure

Table 2 shows the matrix of pairwise F_{ST} estimates calculated from DNA sequences (Hudson et al. 1992) and using both STRs in the Xp11.22 region, on the basis of the geographic regions identified by Rosenberg et al. (2002). This table corresponds to supplemental table 2 of Rosenberg et al. (2002), which was calculated using 377 autosomal STRs for all 1,056 individuals (males and females) in the HGDP-CEPH. The correspondence between F_{ST} estimates in the 2 studies is shown in figure 5. The Mantel test of association between the F_{ST} estimates based on our STR data and for the STR data of Rosenberg et al. was statistically significant. However, no significant association was observed for F_{ST} estimates based on the Xp11.22 sequence data and for the STR data of Rosenberg et al. (2002). In general, the patterns of population structure for the Xp11.22 region are similar to the larger study of autosomal STRs. However, population structure appears to be stronger for the Xp11.22 region than for the autosomal loci (fig. 5) as suggested by estimated linear regression slopes greater than 1.

Time of the Most Recent Common Ancestor

Haplotype *hX* was found in geographically diverse populations, but not in the sub-Saharan African samples. The *hX* haplotype connects to the root of the human gene tree for the Xp11.22 region, and in this sense it is quite old, yet it is also relatively rare and not very variable. No haplotypes similar to *hX* were found (i.e., all 9 copies of *hX* were identical at the sequence level) and *hX* harbors little STR variation (fig. 4). The pattern is suggestive of a history in which *hX* entered the non-African populations at a low frequency from some non-African source. Similar patterns at other loci have been interpreted as suggestive of admixture between modern humans and archaic *Homo sapiens* outside of Africa. (Garrigan, Mobasher, Kingan, et al. 2005; Garrigan, Mobasher, Severson, et al. 2005; Evans et al. 2006).

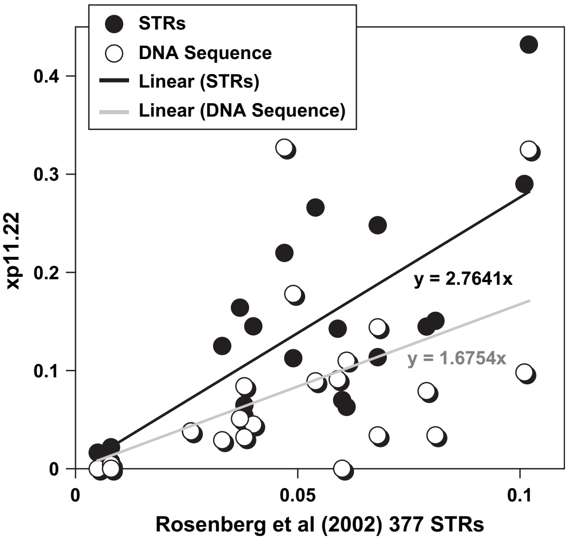

FIG. 5.—F_{ST} comparison between studies A plot of the values in table 2 and supplementary table 2 of Rosenberg et al. (2002).

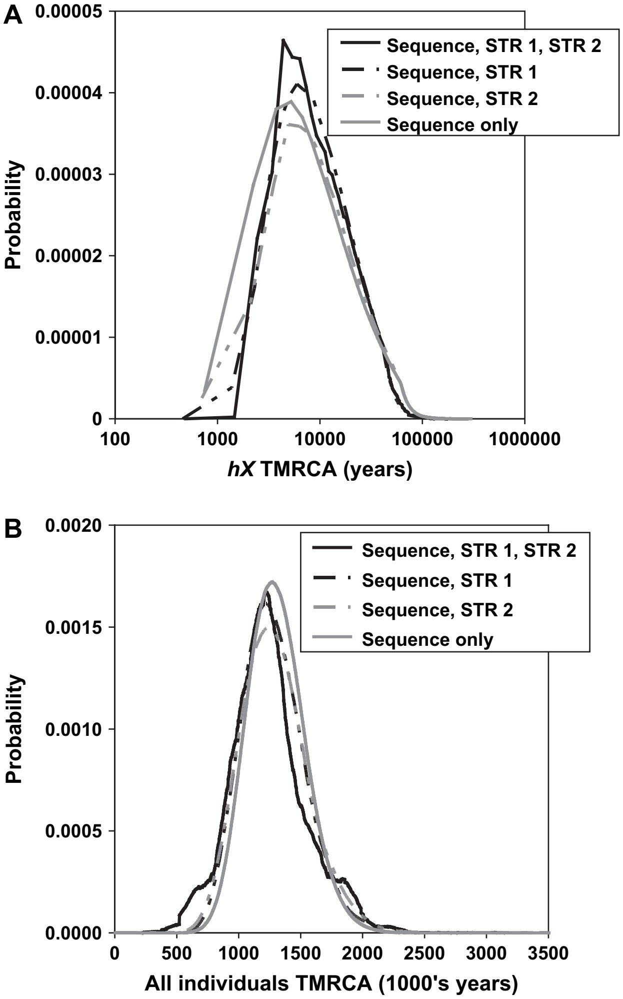

FIG. 6.—Estimated posterior probability of the time of common ancestry. Each figure shows estimates based on gene copies with nonambiguous haplotypes. Results are shown for the sequence alone, sequence with STR 1, sequence with STR 2, and sequence with both of the STR regions. (*A*) TMRCA distribution for *hX*, plotted on a log scale of years. (*B*) TMRCA distribution for the entire sample plotted on a linear scale in thousands of years.

Analyses of the posterior probability of the TMRCA were conducted in order to assess how long ago the common ancestor of *hX* is likely to have existed. Over the course of the Markov chain Monte Carlo simulation, the TMRCA of all gene copies as well as for just those gene copies bearing *hX* were recorded (see Materials and Methods). To provide estimates in terms of years, the coalescent time was scaled by the estimated human/chimpanzee divergence rate. Figure 6 shows the resulting posterior density estimates, which were very similar regardless of whether or not the STR data were included in the analyses. For just the analyses using only DNA sequences (i.e., no STR data), the estimated time of common ancestry for *hX* was 5,230 years (95% confidence interval [CI]: 2,000–75,480 years) and the estimated time for the common ancestry of all the sequences was 1.271 MYA (95% CI: 0.882 to 1.792 MYA). However, the posterior probability curves were very similar regardless of whether one or both STR regions were included with the sequence. It is useful to emphasize that even though the estimated TMRCA for *hX* is quite recent, the posterior density is skewed and extends far to the right (fig. 6*A*, note the log scale of the *x* axis). Given that appreciable probability is also associated with times older than 50,000 years, the TMRCA analyses are not inconsistent with a model in which *hX* moved out of Africa with an out-of-Africa migration of modern humans sometime within the past 100,000 years. If this model is correct, then we would expect to find *hX* in a broader sampling of sub-Saharan African populations.

Extending Sampling from Africa

Given the observed frequency of *hX* outside of sub-Saharan Africa (1.6%), we can estimate the probability of not observing *hX* in our sample of 92 sub-Saharan gene copies. That probability is 0.226 (i.e., $(1 - 0.016)^{92}$), which means that the absence of *hX* in the sub-Saharan HGDP-CEPH male samples is not particularly a strong evidence that *hX* does not occur there at low frequencies similar to what has been observed outside of sub-Saharan Africa.

To better explore the question of whether *hX* occurs in sub-Saharan Africa, we sequenced 3 portions of the Xp11.22 region that are diagnostic for *hX* in an additional set of samples from Tanzania and Chad. Out of 101 chromosomes, 5 copies (0.0495%) of *hX* were found. These included 1 from the Laka tribe in Chad and 4 from Tanzania: 1 from the Burunge tribe, 2 from the Turu tribe, and 1 from the Sandawe tribe. There is an apparent contrast between the absence of *hX* in males from the HGDP-CEPH panel, for which African samples come from southern and western sub-Saharan Africa, and the presence of *hX* in samples from northern and eastern sub-Saharan Africa. Fisher's exact test of a difference in frequencies between the 2 sets of samples (n = 95 0 copies of *hX* for HGDP-CEPH sub-Saharan African Samples; n = 101, 5 copies of *hx* for Chad and Tanzania) is on the border of statistical significance (P = 0.0599).

STR 1 and STR 2 were genotyped in the samples of *hX* from Chad and Tanzania to assess whether they are divergent from those copies of *hX* found in the HGDP-CEPH panel. Only modest evidence of divergence was found: STR 1 had 4 copies of length 18 and 1 of length 17 (compare with fig. 4) and STR 2 had 3 copies of length 20 and 2 copies of length 22. Like the copies of *hX* found in the HGDP-CEPH panel, the samples from Chad and Tanzania had low variation at STR 1 and STR 2, and at both STR regions, the most common allele found in the Chad and Tanzania samples was also the most common allele found in the HGDP-CEPH samples.

Discussion

This study extends the general approach of using DNA resequencing for answering questions about the population genetic history of humans to a worldwide sample of individuals (672) and populations (the 52 populations sampled in the HGDP-CEPH panel). The survey of a 10.1-kbp region from the human X chromosome revealed a widespread rare haplotype, *hX*, that is highly divergent from the other copies of the region and that might not have been observed in a more modest sampling scheme. This study also incorporated linked STR variation for 2 repeat regions within the Xp11.22 sequenced region. The use of compound haplotypes like these has often been used in the past for comparison of mutation processes (Orti et al. 1997; Colson and Goldstein 1999; Schlötterer 2000; Blankenship et al. 2002; Estoup et al. 2002). Less often have STRs and flanking sequences been used conjointly to infer population genetic history (Tishkoff et al. 1996; Rogers et al. 2000; Hey et al. 2004; Won et al. 2005).

Levels and Patterns of Variation

With the exception of 1 haplotype, *hX*, the Xp11.22 region is not highly variable. The estimated depth of the haplotype tree, for that portion not including *hX*, is just 0.23 Myr, similar to what has been reported for a 10-kbp region in xq13.3 (Kaessmann et al. 1999), but less than that for other X-linked loci (Nachman et al. 1998; Harris and Hey 1999, 2001; Jaruzelska et al. 1999; Nachman and Crowell 2000; Gilad et al. 2002; Verrelli et al. 2002; Nachman et al. 2004). The low level of variation among non-*hX* haplotypes is consistent with what has been observed in resequencing studies of other regions that are fairly close to Xp11.22. The *Alas2* gene region on Xp11.21, which is about 4.5 Mbp away, and the *Msn* gene region on Xq12, separated from Xp11.22 by about 10 Mbp, both showed low levels of variation in samples of 41 chromosomes (10 African and 31 non-African) (Nachman et al. 2004). The patterns of population structure were similar to those observed in a large study of autosomal STR loci on the HGDP-CEPH panel (Rosenberg et al. 2002), with levels of structure being somewhat higher, perhaps because the Xp11.22 is sex linked and is expected to have a higher rate of genetic drift compared with autosomal loci.

Inclusion of *hX* essentially adds 1 Myr to the depth of the haplotype tree and greatly increases the overall level of variation. Although *hX* is always found at low frequency, it was found in 9 copies among males of the HGDP-CEPH panel, and these were scattered widely throughout the non–sub-Saharan African samples. The *hX* haplotype differed from all other sequences at 27 base positions, although among the copies of *hX* there was no sequence variation and STR diversity was low. The STR alleles that were found among copies of *hX* were also common among the other haplotypes (fig. 4). From the perspective of single STRs, copies of the Xp11.22 region that carry *hX* are nearly indistinguishable from non-*hX* types. However, at the level of STR haplotypes (joint for STR 1 and STR 2), there is some differentiation. Table 3 shows the frequencies of STR haplotypes that were common in the HGDP-CEPH, as well as the frequencies of the STR haplotypes found among copies of *hX*. The most common *hX* STR haplotype (18, 20) was not found among the non-*hX* types, and only one of the *hX* STR haplotypes (17, 20) was found in appreciable frequency (0.03%) among non-*hX* types.

A coalescent analysis provided estimates of the posterior probability density of the TMRCA of *hX* and of the full set of data (fig. 6). The distribution for *hX* had an extended tail to the right and a peak located at about 5,200 years—a very recent time that is reflective of the overall lack of variability among copies of *hX*.

Table 3
STR Haplotype Frequencies (percentages in parentheses)

	STR	# In Non-*hX*	# In *hX*	# In *hX*	# In *hX*
STR 1	2	Sequences	HGDP	C and T[a]	Total
Common STR haplotypes (frequency > 0.05 in HGDP-CEPH Panel)					
18	19	24 (5.2)	0	0	0
22	19	26 (5.6)	0	0	0
19	19	28 (6.1)	0	0	0
21	20	30 (6.5)	0	0	0
21	19	72 (15.7)	0	0	0
20	19	167 (36.4)	0	0	0
hX STR haplotypes					
17	20	13 (0.03)	0	1 (0.20)	1 (0.07)
17	21	1 (0.002)	1 (0.11)	0	1 (0.07)
18	20	0	4 (0.44)	2 (0.40)	6 (0.43)
18	21	0	1 (0.11)	0	1
18	22	0	3 (0.33)	2 (0.40)	2 (0.14)
19	20	1 (0.002)	0	0	3 (0.21)

[a] Samples from Chad and Tanzania.

[#] Number observed.

What *hX* Tells Us about the History of Modern Human Populations

The patterns presented by the HGDP-CEPH sequences and by the 9 copies of *hX* among those sequences are puzzling: *hX* is highly divergent from other gene copies, extending the gene tree depth to several times what it would be otherwise; *hX* is rare but widespread out of sub-Saharan Africa; *hX* was not found in the sub-Saharan African HGDP-CEPH samples; and a coalescent analysis suggests a recent common ancestor time of the 9 copies of *hX*. On the basis of these observations alone, the pattern is consistent with what has been seen at some other loci and that has been interpreted as evidence of limited admixture outside of Africa between modern humans and archaic humans (Garrigan, Mobesher, Severson et al. 2005; Evans et al. 2006). However, *hX* does indeed occur in sub-Saharan Africa as shown by the extended sampling of populations in Chad and Tanzania (populations not included in the HGDP-CEPH). If we were to still consider a model of recent admixture between modern and archaic humans outside of Africa, then the finding of *hX* in Chad and Tanzania means that we would also have to include gene flow of *hX* into eastern Africa following that admixture.

The presence of *hX* makes for a very unbalanced gene tree (fig. 1). The low frequency and the fact that *hX* connects to the base of the human Xp11.22 gene tree suggests some kind of departure from selective neutrality or from panmixia. Under a random branching model, such as the standard neutral coalescent, the proportion of sampled sequences that fall to 1 side of the basal node of a tree is a uniform random variable (Harding 1971). Given our sample of 672 gene copies, there are 671 possible different counts of sequences to 1 side of the basal node, each with equal probability (under the standard neutral coalescent). The probability of an imbalance equal to or greater than what was observed (i.e., 9 gene copies on either side of the tree) is 2 × 9 × (1/671) = 0.0268. In other words, the gene tree is more unbalanced than we expect by chance under a simple neutral, panmictic model. One possible selective explanation is that some or all of the non-*hX* gene copies are from a class of haplotypes that are selectively favored and that are slowly increasing in frequency at the expense of *hX*. Alternatively, it is possible that selection (e.g., on an adjacent tightly linked region) has played a role in maintaining *hX* consistently at low frequencies. However, this would not easily explain the estimated recent young age of the common acestor of the different copies of *hX*.

Another explanation of the unbalanced tree is that our sample does not come from a panmictic population. This is certainly true given our general understanding of human population structure and history. However, discerning the kind of population structure that best explains *hX* and the unbalanced tree is not a simple matter. One possible historical model that could generate this pattern supposes that differences between *hX* and other haplotypes arose in the presence of population structure that allowed for the divergence among Xp11.22 haplotypes. Even if we discount the possibility of a non-African archaic human population as the source of *hX*, the age and low frequency of the haplotype does suggest a history in which *hX* persisted and diverged in a separate refugium population (either a separate modern human population or possibly an archaic human population). Models of this type, which suppose the presence of old population strutcture among African populations, have been suggested based on evidence from other regions of the genome (Tishkoff et al. 1996; Harding et al. 1997; Labuda et al. 2000; Tishkoff et al. 2000; Zietkiewicz et al. 2003; Garrigan, Mobesher, Kingan, et al. 2005).

But whether or not the divergence of *hX* and other copies of the Xp11.22 region is specifically due to old population structure, the close relatedness of different copies of *hX* and their widespread occurrence in Eastern Africa, Europe, and Asia are consistent with Eastern Africa being a recent source of non-African modern human populations (Quintana-Murci et al. 1999; Underhill et al. 2000; Semino et al. 2002; Tishkoff and Verrelli 2003).

Supplementary Material

STR genotypes, the locations and frequencies of unresolved bases, and the number of repeats in the 2 STR loci in all samples examined are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are grateful to Yong-Jin Won and Arjun Sivasundar for their assistance and discussions. This study was financially supported by the Department of Genetics, Division of Life Sciences of Rutgers University and by National Science Foundation grant BCS-0196183 and a Packard Career Award to S.T. This study was approved by the Institutional Review Boards of Rutgers University and the University of Maryland.

Literature Cited

Blankenship SM, May B, Hedgecock D. 2002. Evolution of a perfect simple sequence repeat locus in the context of its flanking sequence. Mol Biol Evol. 19:1943–1951.

Brunet M, Guy F, Pilbeam D, et al. (38 co-authors). 2002. A new hominid from the Upper Miocene of Chad, Central Africa. Nature. 418:145–151.

Cann HM, de Toma C, Cazes L, et al. (41 co-authors). 2002. A human genome diversity cell line panel. Science. 296:261–262.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. The history and geography of human genes. Princeton (NJ): Princeton University Press.

Chen F-C, Li W-H. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet. 68:444–456.

Clark AG, Nielsen R, Signorovitch J, Matise TC, Glanowski S, Heil J, Winn-Deen ES, Holden AL, Lai E. 2003. Linkage disequilibrium and inference of ancestral recombination in 538 single-nucleotide polymorphism clusters across the human genome. Am J Hum Genet. 73:285–300.

Colson I, Goldstein DB. 1999. Evidence for complex mutations at microsatellite loci in Drosophila. Genetics. 152:617–627.

Estoup A, Jarne P, Cornuet JM. 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Mol Ecol. 11:1591–1604.

Evans PD, Mekel-Bobrov N, Vallender EJ, Hudson RR, Lahn BT. 2006. Evidence that the adaptive allele of the brain size gene microcephalin introgressed into Homo sapiens from an archaic Homo lineage. Proc Natl Acad Sci USA. 103:18178–18183.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8:186–194.

Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8:175–185.

Excoffier L, Laval G, Schneider S. 2005. Arlequin ver. 3.0: an integrated software package for population genetics data analysis. Evol Bioinform Online. 1:47–50.

Garrigan D, Mobasher Z, Kingan SB, Wilder JA, Hammer MF. 2005. Deep haplotype divergence and long-range linkage disequilibrium at xp21.1 provide evidence that humans descend from a structured ancestral population. Genetics. 170:1849–1856.

Garrigan D, Mobasher Z, Severson T, Wilder JA, Hammer MF. 2005. Evidence for archaic Asian ancestry on the human X chromosome. Mol Biol Evol. 22:189–192.

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

Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW. 1995. An evaluation of genetic distances for use within microsatellite loci. Genetics. 139:463–471.

Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing. Genome Res. 8:195–202.

Hammer MF, Horai S. 1995. Y chromosomal DNA variation and the peopling of Japan. Am J Hum Genet. 56:951–962.

Harding EF. 1971. The probabilities of rooted tree-shapes generated by random bifurcation. Adv Appl Probab. 3:44–77.

Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, Clegg JB. 1997. Archaic African and Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet. 60:772–789.

Harding RM, McVean G. 2004. A structured ancestral population for the evolution of modern humans. Curr Opin Genet Dev. 14:667–674.

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

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

Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. Persimilis.* Genetics. 167:747–760.

Hey J, Wakeley J. 1997. A coalescent estimator of the population recombination rate. Genetics. 145:833–846.

Hey J, Won Y-J, Sivasundar A, Nielsen R, Markert JA. 2004. Using nuclear haplotypes with microsatellites to study gene flow between recently separated cichlid species. Mol Ecol. 13:909–919.

Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics. 111:147–164.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics. 132:583–589.

Jaruzelska J, Zietkiewicz E, Batzer M, Cole DE, Moisan JP, Scozzari R, Tavar S, Labuda D. 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.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press. p. 21–132.

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

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res. 12:996–1006.

Kimura M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics. 61:893–903.

Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, Villems R. 2004. Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet. 75:752–770.

Kumar S, Tamura K, Nei M. 2004. Mega3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform. 5:150–163.

Labuda D, Zietkiewicz E, Yotova V. 2000. Archaic lineages in the history of modern humans. Genetics. 156:799–808.

Mountain JL, Ramakrishnan U. 2005. Impact of human population history on distributions of individual-level genetic distance. Hum Genomics. 2:4–19.

Nachman MW, Bauer VL, Crowell SL, Aquadro CF. 1998. DNA variability and recombination rates at X-linked loci in humans. Genetics. 150:1133–1141.

Nachman MW, Crowell SL. 2000. Contrasting evolutionary histories of two introns of the duchenne muscular dystrophy gene, dmd, in humans. Genetics. 155:1855–1864.

Nachman MW, D'Agostino SL, Tillquist CR, Mobasher Z, Hammer MF. 2004. Nucleotide variation at msn and alas2, two genes flanking the centromere of the X chromosome in humans. Genetics. 167:423–437.

Nei M, Tajima F. 1981. DNA polymorphism detectable by restriction endonucleases. Genetics. 97:145–163.

Nickerson DA, Tobe VO, Taylor SL. 1997. Polyphred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res. 25:2745–2751.

Orti G, Pearse DE, Avise JC. 1997. Phylogenetic assessment of length variation at a microsatellite locus. Proc Natl Acad Sci USA. 94:10745–10749.

Patil N, Berno AJ, Hinds DA, et al. (22 co-authors). 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science. 294:1719–1723.

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature. 441:1103–1108.

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

Quintana-Murci L, Semino O, Bandelt HJ, Passarino G, McElreavey K, Santachiara-Benerecetti AS. 1999. Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nat Genet. 23:437–441.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci USA. 102:15942–15947.

Ramachandran S, Rosenberg NA, Zhivotovsky LA, Feldman MW. 2004. Robustness of the inference of human population structure: a comparison of X-chromosomal and autosomal microsatellites. Hum Genomics. 1:87–97.

Richards M, Rengo C, Cruciani F, Gratrix F, Wilson JF, Scozzari R, Macaulay V, Torroni A. 2003. Extensive female-mediated gene flow from sub-Saharan Africa into near eastern Arab populations. Am J Hum Genet. 72:1058–1064.

Rogers EJ, Shone AC, Alonso S, May CA, Armour JA. 2000. Integrated analysis of sequence evolution and population history using hypervariable compound haplotypes. Hum Mol Genet. 9:2675–2681.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic structure of human populations. Science. 298:2381–2385.

Rzhetsky A, Nei M. 1992. A simple method for estimating and testing minimum-evolution trees. Mol Biol Evol. 9:945–967.

Schlötterer C. 2000. Evolutionary dynamics of microsatellite DNA. Chromosoma. 109:365–371.

Semino O, Santachiara-Benerecetti AS, Falaschi F, Cavalli-Sforza LL, Underhill PA. 2002. Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. Am J Hum Genet. 70:265–268.

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

Tajima F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics. 135:599–607.

Tishkoff SA, Dietzsch E, Speed W, et al. (15 co-authors). 1996. Global patterns of linkage disequilibrium at the cd4 locus and modern human origins. Science. 271:1380–1387.

Tishkoff SA, Pakstis AJ, Stoneking M, et al. (12 co-authors). 2000. Short tandem-repeat polymorphism/alu haplotype variation at the plat locus: implications for modern human origins. Am J Hum Genet. 67:901–925.

Tishkoff SA, Verrelli BC. 2003. Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet. 4:293–340.

Underhill PA, Shen P, Lin AA, et al. (21 co-authors). 2000. Y chromosome sequence variation and the history of human populations. Nat Genet. 26:358–361.

Verrelli BC, McDonald JH, Argyropoulos G, Destro-Bisol G, Froment A, Drousiotou A, Lefranc G, Helal AN, Loiselet J, Tishkoff SA. 2002. Evidence for balancing selection from nucleotide sequence analyses of human g6pd. Am J Hum Genet. 71:1112–1128.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 7:256–275.

Won YJ, Sivasundar A, Wang Y, Hey J. 2005. On the origin of Lake Malawi cichlid species: a population genetic analysis of divergence. Proc Natl Acad Sci USA. 102:6581–6586.

Zhang K, Deng M, Chen T, Waterman MS, Sun F. 2002. A dynamic programming algorithm for haplotype block partitioning. Proc Natl Acad Sci USA. 99:7335–7339.

Zhang K, Sun F, Waterman MS, Chen T. 2003. Haplotype block partition with limited resources and applications to human chromosome 21 haplotype data. Am J Hum Genet. 73:63–73.

Zhivotovsky LA, Rosenberg NA, Feldman MW. 2003. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. Am J Hum Genet. 72:1171–1186.

Zietkiewicz E, Yotova V, Gehl D, et al. (13 co-authors). 2003. Haplotypes in the dystrophin DNA segment point to a mosaic origin of modern human diversity. Am J Hum Genet. 73:994–1015.

Anne Stone, Associate Editor

Accepted December 11, 2006