DNA Polymorphism in a Worldwide Sample of Human X Chromosomes

Ning Yu,* Yun-Xin Fu,† and Wen-Hsiung Li*

*Department of Ecology and Evolution, University of Chicago, Chicago; and †Human Genetics Center, University of Texas-Houston, Houston, Texas

DNA sequence data from humans can provide insight into the history of modern humans and the genetic variability in human populations. We report here a study of human DNA sequence variation at an X-linked noncoding region of 10,346 bp. The sample consists of 62 X chromosomes from Africa, Europe, and Asia. Forty-four polymorphic sites were found among the 62 sequences, resulting in 23 different haplotypes. Statistical analyses of the data led to the following inferences. (1) There is strong evidence of human population expansion in the relatively recent past, and this population expansion has had a significant effect on the pattern of polymorphism at this locus. (2) Non-African populations were unlikely to have been derived from a very small number of African lineages. (3) There was considerable geographic subdivision in the ancient human population, which could be an important reason why many studies failed to detect population expansion. (4) The long-term effective population size of humans is between 12,000 and 15,000. And (5) a non-African specific variant was found at a frequency of 35% in non-Africans, an estimate supported by the genotyping of additional 80 non-African and 106 African X chromosomes. This variant could have arisen in Eurasia more than 140,000 years ago, predating the emergence of modern humans. Moreover, this haplotype and all other haplotypes coalesced to the most recent common ancestor of the sample, which was estimated to be older than 490,000 years. Therefore, this region may have a long history in Eurasia.

Introduction

The geographic origin and population history of modern humans has been the subject of numerous studies. The molecular approach that examines molecular differences within and between populations has brought a better understanding of this subject but has also created some controversies. One common problem facing both the traditional anthropological approach and the molecular approach is the scarcity of high-quality data. For the former, it is a hurdle hard to get over because high-quality fossils are rarely found. For the latter, it is largely a matter of technology and cost. The Human Genome Project in the last decade has helped to advance much of the DNA technology needed for large-scale population studies that are critical for resolving issues on human origin and history. Indeed, several studies of polymorphism in the human nuclear genome involving regions of ~10 kb or longer and samples of more than 50 sequences have been reported recently (e.g., Nickerson et al. 1998; Kaessmann et al. 1999; Zhao et al. 2000; Yu et al. 2001). But the study of Kaessmann et al. (1999) is to date the only large-scale study on the X chromosome. Notable inferences from that study were that the age of the most recent common ancestor (MRCA) of the sample was ~535,000 years and that the effective population size was ~16,000. Another study on the X chromosome was conducted by Harris and Hey (1999), who investigated a 4,200-bp region in the PDHA1 gene locus.

Many issues about the origin and history of modern humans remain to be resolved. One is whether the recent human population expansion has had a significant effect on the genetic diversity, or whether there is genetic evidence of a relatively recent population expansion at all. To date, the many studies on this issue have yielded conflicting conclusions. Another issue is, “how deep is the genetic history of all modern humans or, more restrictively, the non-Africans?” This issue is controversial partly because a molecular dating of an ancestral event is usually associated with a large standard error and partly because different loci (regions) have different histories. A consensus will unlikely be reached before many more large-scale studies are conducted and the data are carefully analyzed. As a continuous effort to understand global human DNA variation, the human population history, and the mechanisms of maintenance of DNA polymorphism, we report here a study of DNA polymorphism on a 10-kb region of the X chromosome in a worldwide sample of 62 sequences. In comparison with autosomes, the X chromosome offers one distinct advantage, namely that the haplotype sequences can be readily determined by using male individuals. Complete haplotype sequences yield the maximal information that can be achieved from DNA sequencing and permit a finer statistical inference. Although haplotype sequences were also obtained from studies on the Y chromosome and mitochondrial DNA, both Y and mitochondrial DNA behave like a single locus, and both of them might have been influenced by many evolutionary forces, which complicate the inference. Because the X chromosome spends two thirds of its time in females and one third in males, the study of X chromosome polymorphism will provide insight into the modern human history that is slightly more influenced by females than males.

Materials and Methods

Region Selection and Populations Sampled

An ~10-kb region corresponding to nucleotide positions 115273–125804 in locus HS571B2 on human chromosome Xq21.1–21.33 was selected for sequencing (GenBank accession no. AL034406). More specifically, clone dJ571B2 lies in the Xq21.31 region, which is flanked proximally by DXS1168 and distally by [additional text not fully visible].
DXS364. The region is continuous except for an exclusion of a fragment containing a poly-T segment in the very first 3,000 bp. At first, no gene was registered in this region in GenBank, and no potential coding exon in the 15-kb region covering the region we studied was detected by either GenScan or GRAIL-EXP. But upon reexamination we found that the sequences we obtained are parts of the third intron of a newly registered gene (DACH2), which is 683,962-bp long and contains 12 exons. The region we sequenced is about 100 kp downstream of the third exon (112 bp) and about 28 kp upstream of the forth exon (131 bp) of the DACH2 gene.

Sixty unrelated individuals were collected worldwide from 29 human populations in three continents: 20 Africans (five South African Bantu speakers, one !Kung, two Mbuti Pygmies, two Biaka Pygmies, five Nigerians, two Kenyans, one Sun, one Ethiopian, and one Sudanese), 20 Asians (five Chinese, three Japanese, four Indians, one Korean, two Mongolians, two Cambodians, two Vietnamese, one Yakut), and 20 Europeans (two Swedes, two Finns, three French, one German, three Hungarians, one Italian, two Sardinians, one Norwegian, one Portuguese, one Spanish, two Russians, one Ukrainian). All individuals were male, except for one Indian and one Cambodian. One male chimpanzee, one male gorilla, and one female orangutan were used as outgroups.

PCR Amplification and DNA Sequencing

Five primer pairs were designed to amplify one overlapping fragment in the 10-kb region. Touch-down PCR (Don et al. 1991) was used, and the reactions were carried out according to the condition described in Zhao et al. (2000). The PCR products were purified by Wizard PCR Preps DNA Purification Resin Kit (Promega). Sequencing reactions were 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. 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 the nucleotides in the segment sequences were sequenced at least once in both directions. Furthermore, all singletons, doubletons, and tripletons, which are defined as variants that appear, respectively, only once, twice, and thrice in the total sample, were verified by reamplifying the region containing the variant site and resequencing the region in both directions.

Cloning and Genotyping

Three to four primer pairs were designed to amplify two to three overlapping fragments covering the heterozygous sites in the two female samples. Expand High Fidelity PCR System (Roche Molecular Biochemicals, Germany) was used, and the reactions were carried out according to the condition described in the protocol. The PCR products were isolated from the agarose gels and purified with a Gel Purification Kit (Qiagen Inc., Valencia, Calif.). Purified PCR products were cloned using PGEMR-T and PGEMR-T Easy Vector Systems (Promega). At least eight colonies were sequenced.

Analysis Methods

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 parsimony principle. Because a variety of statistical methods were used in analyzing this data, they will be discussed in Results whenever it is appropriate.

Results

The sequenced region was 10,346-bp long. After exclusion of insertions and deletions, 10,158 bp remained and were used in all our analyses. The total sample consists of 62 sequences and 44 mutations were observed, resulting in 23 distinct haplotypes (alleles). We will summarize our analyses of the data with respect to several issues.

Haplotype Distribution

Twenty-three haplotypes were observed among the 62 chromosomes from 29 human populations in three continents. Haplotype designation and their frequencies are shown in table 1. Haplotypes A, B, and C are most common, followed by haplotypes E, F, and G; the remaining 17 haplotypes are singleton alleles. Among the major haplotypes, A and B are unique to the non-Africans sampled, whereas E, F, and G are found only in Africans. Notably, haplotype C is the only haplotype shared by both Africans and non-Africans. The proportion of unique haplotypes in Africans (80%) and non-Africans (79%) are much higher than the shared one (21%). The high proportion (52.4%) of shared haplotypes by Asians and Europeans suggested that non-Africans were derived from a common ancestral population or there has been substantial migration between Europe and Asia.

With the outgroup sequences, which allow us to infer the ancestral nucleotides of each segregating site, a haplotype evolutionary network was constructed using the parsimony principle (fig. 1). We note that only one segregating site (position 1932) has an inferred ancestral nucleotide of the human sequences that is incongruent with those of other sites. A most parsimonious scenario is that since the divergence of human and chimpanzee, the human ancestral lineage mutated at this site from G to A and later since the divergence of human and chimpanzee, the human ancestral lineage mutated back to G. With this scenario, the haplotype network is the most parsimonious, and it does not require recombination to explain the data. Although this does not constitute a proof of no recombination because many recombination events can go without leaving any trace, it does suggest that the recombination
rate in this region is low, and a statistical analysis assuming no recombination will unlikely yield serious bias. Our phylogenetic analysis also suggests that within the human species, each of the 44 segregating sites with the exception of site 1932 (fig. 1) corresponds exactly to one mutation. Therefore, the infinite site model is a reasonable assumption for data analysis.

Figure 1 shows that most non-African and African sequences are intermingled in the network, with one notable exception: haplotype A and its derived haplotypes A1, A2, and A3 are exclusively non-African. Note that haplotype B, the second major non-African haplotype, was derived from the African specific haplotype G. Haplotypes C1–C5 are non-African specific and were derived from haplotype C, the only shared haplotype by Africans and non-Africans. Another noteworthy observation is that the lineage leading to haplotypes D, D1, and D2 has undergone eight mutations and one insertion or deletion. Interestingly, haplotypes D, D1, and D2 include individuals from three different continents. Overall, these data show no clear trend with respect to whether non-Africans were derived from Africans, or vice versa.

### Mutation Rate and Pattern

Together with the outgroup sequences, a total of 417 variable sites were found, seven of which have three segregating nucleotides, but all the segregating sites within the human sequences each have only two segregating nucleotides. The number of mutations is inferred to be 424 by the parsimony principle. This information allows us to estimate the mutation rate as well as the pattern of mutation.

For a locus subject to no natural selection, the mutation rate \( \mu \) per sequence per generation is estimated by \( \mu = n_d \times g \times L/(2T) \), where \( n_d \) is the number of nucleotide substitutions per site between two sequences, \( T \) the divergence time between the two sequences, \( L \) the sequence length (bp), and \( g \) the length of a generation, which is commonly assumed to be 20 years for humans. It is obvious that knowledge on the divergence time \( T \) is crucial in this estimator. Because the divergent dates between human and apes are uncertain, we use multiple species for comparison and derive our estimate as the average over separate estimates.

In our previous studies (Zhao et al. 2000; Yu et al. 2001) we took \( T \) between human and chimpanzee as 6 Myr (see Haile-Selassie 2001) and that between human and gorilla as 8 Myr. We will use the same assumptions here, so that the results are directly comparable with our previous studies. The average rate from the human-chimpanzee comparison and the human-gorilla comparison is \( 1.55 \times 10^{-4} \) per sequence per generation, which corresponds to \( 7.63 \times 10^{-10} \) per site per year.
We examine the pattern of mutations to see if there is any unusual feature that could affect our subsequent analyses. Among the 424 mutations in all the sequences, 234 can be inferred for the direction of mutation, i.e., which nucleotide is the ancestral and which is the mutant. Table 3 shows the pattern of mutations. For the human sequences the number of mutations from nucleotide x to y (x, y = A, G, C, or T) is quite similar to that from y to x. It is also true for the entire data set except for the case x = G and y = A. The transition-transversion ratio is 1.93 for within human sample mutations and 1.95 among all changes. These values are close to the estimated 2:1 ratio for mammalian genomes (Li 1997, p. 31).

We also examine the spatial distribution of mutations. The entire region of 10,346 nucleotides was di-
on 29 July 2018

Wright-Fisher model. Let $\xi$ show the observed frequencies of mutations of various sizes in a sample at a neutral locus. We shall return to the estimation of $\theta$, the expectation of $\xi$, since the MRCA of the sample. We perform several widely used statistical tests: Tajima’s $D$ (Tajima 1989), Fu and Li’s $D$ and $D^*$ (Fu and Li 1993), and Fu’s $F$s (Fu 1997). The first three methods require only the frequencies of mutations of various sizes, whereas the last method requires knowing the number of alleles in the sample. One of the advantages of working on the X chromosome is that the haplotype sequences can be divided into 10 regions of equal length (1,034 pb each) and the occurrences of mutations in the 10 regions are 3, 10, 2, 5, 5, 4, 2, and 4 among the human sequences and 32, 47, 40, 44, 52, 34, 45, 51, 45, and 35 among all sequences. A chi-square test cannot reject the hypothesis that the mutation rates in all the regions are the same, whether we consider all mutations or only mutations in the human sequences.

Frequencies of Mutant Nucleotides in the Sample

We now turn our attention to the frequencies of mutant nucleotides among the human sequences. These frequencies are useful for inferring the evolutionary forces that have operated on the locus and for estimating population parameters. Mutations (mutant nucleotides) in a sample can be classified into different size groups. A mutation in a sample is said to be of size $i$ if there are exactly $i$ mutations in the sample carrying the mutant nucleotide (Fu 1995). For a sample of $n$ sequences, a mutation has a size between 1 to $n - 1$. Tables 4 and 5 show the observed frequencies of mutations of various sizes and the expected frequencies under the neutral Wright-Fisher model. Let $\xi$ be the number of mutations of size $i$ in a sample of $n$ sequences. Fu (1995) showed that the expectation of $\xi$ is $E(\xi) = \theta i$, where $\theta = 3N\mu$, for an X-linked locus, in which $N$ is the effective population size and $\mu$ is the mutation rate per sequence per generation. $\theta$ is an important population parameter because it determines the amount of variation that is expected in a sample at a neutral locus. We shall return to the estimation of $\theta$ in a later section. The expectations in tables 4 and 5 were computed by substituting $\theta$ with Watterson’s (1975) estimate $\theta = K/a_n$, where $K$ is the number of segregating sites and $a_n = 1 + 1/2 + \ldots + 1/(n - 1)$. Tables 4 and 5 reveal a conspicuous excess of singletons (i.e., mutations of size 1) in the total sample and in the subsamples. We shall demonstrate in the next section that the excess is statistically significant.

Neutrality Tests

We are interested in knowing whether this region has evolved under strict neutrality. We shall test the hypothesis that all mutations are selectively neutral and that the population has evolved according to the Wright-Fisher model with a constant effective population size since the MRCA of the sample. We perform several widely used statistical tests: Tajima’s $D$ (Tajima 1989), Fu and Li’s $D$ and $D^*$ (Fu and Li 1993), and Fu’s $F$s (Fu 1997). The first three methods require only the frequencies of mutations of various sizes, whereas the last method requires knowing the number of alleles in the sample. One of the advantages of working on the X chromosome is that the haplotype sequences can be de-

Table 3

| Pair       | (a)   | (b)   | (c) Different pairs |
|------------|-------|-------|---------------------|
|            | A     | G     | C                   | T     | A     | G     | C     |
| A          | 0     | 7     | 2                  | 3     | 0     | 46    | 10    | 8     |
| G          | 6     | 0     | 3                  | 1     | 33    | 0     | 12    | 11    | 67    |
| C          | 0     | 3     | 0                  | 8     | 8     | 12    | 0     | 32    | 19    | 11    |
| T          | 1     | 2     | 8                  | 0     | 9     | 16    | 37    | 0     | 8     | 20    | 65    |

Note.—(a) Within human population changes; (b) all changes whose direction can be inferred; and (c) all changes whose direction cannot be inferred (xy and yx are grouped together where x and y are different nucleotides).
termed relatively easily, allowing the use of more powerful statistical methods. Table 6 shows the results of these statistical tests.

Two of the four tests (Fu and Li’s D and D*) are significant for the total sample, and Fu and Li’s D is significant for the African sample. Because Fu and Li’s D test compares the number of mutations of size 1 with that of other sizes, a significant test result indicates an excess of singletons. This test confirms the impression from a visual inspection of tables 4 and 5. Although the test is not significant for the non-African sample, the result is skewed toward an excess of singletons. We therefore conclude that the locus under study has not evolved according to the neutral Wright-Fisher model. We now face the challenge to dissect the cause(s) for the deviation from the model. Because this should be done by taking into consideration various lines of evidence, we shall defer the task to the Discussion. Instead, we try here to understand why Tajima’s D test and Fu’s Fs test failed to detect access of rare mutants.

To see how various types of mutations contribute to the values of various test statistics, we group mutations into several categories according to the frequency of a mutant in a sample, but mutations of size 1 are considered separately because of their importance in determining the test results. Table 7 shows the actual and expected contributions of mutations of different frequency classes in the total sample and subsamples. In addition to the apparent excess of mutations of size 1 in all the samples, two other notable patterns in table 7 are (1) there is excess of mutations of intermediate frequencies (25%–75%) and (2) there is deficiency of low- and high-frequencies (<25% and >75%, except for mutations of size 1). Note that the largest contribution to $\theta_w$ is from mutations of size 1, whereas the largest contribution to $\Pi$ is from mutations of intermediate frequencies. Because mutations in both groups are in excess, one inflates the value of $\theta_w$ and the other inflates the value of $\Pi$, resulting in a small difference between $\theta_w$ and $\Pi$. As a consequence, Tajima’s test, which compares $\theta_w$ and $\Pi$, failed to show signi-

| Table 6 | Neutrality Tests |
|---------|------------------|
|         | TOTAL SAMPLE     | AFRICANS | NON-AFRICANS |
| Test    | Result | Probability | Result | Probability | Result | Probability |
| Fu and Li’s D    | -2.66  | <0.025      | -1.85  | <0.05       | -1.27  | >0.10       |
| Fu and Li’s D*   | -2.34  | <0.025      | -1.42  | >0.05       | -1.06  | >0.10       |
| Tajima D        | -0.70  | >0.10       | -0.55  | >0.10       | -0.23  | >0.10       |
| Fu’s Fs         | -3.17  | >0.10       | 0.33   | >0.10       | 0.51   | >0.10       |

| Table 7 | Contribution of Different Classes of Mutations to $K_w = K_{\theta_w}$ and $\Pi$ |
|---------|----------------------------------|
| Sample  | Size Classes | Frequency | $K_w$ | Exp. $K_w$ | $\Pi$ | Exp. $\Pi$ |
|---------|--------------|-----------|------|-----------|------|-----------|
| Total sample        | Size 1      | 21        | 4.47 | 2.00      | 0.68 | 0.24      |
|                   | Size 1%–10% | 10        | 2.13 | 2.89      | 0.96 | 1.12      |
|                   | 11%–24%     | 1         | 0.21 | 1.60      | 0.32 | 1.59      |
|                   | 25%–75%     | 12        | 2.56 | 2.32      | 5.35 | 3.90      |
|                   | 76%–99%     | 0         | 0.56 | 0.56      | 0    | 0.46      |
|                   | Total       | 44        | 9.37 | 9.37      | 7.31 | 7.31      |
| African sample      | Size 1      | 17        | 4.79 | 2.38      | 1.70 | 0.72      |
|                   | Size 1%–10% | 1         | 0.28 | 1.19      | 0.19 | 0.68      |
|                   | 11%–24%     | 1         | 0.28 | 1.39      | 0.34 | 1.25      |
|                   | 25%–75%     | 11        | 3.10 | 2.94      | 4.96 | 4.16      |
|                   | 76%–99%     | 0         | 0.55 | 0.55      | 0    | 0.38      |
|                   | Total       | 30        | 8.46 | 8.46      | 7.18 | 7.18      |
| Asian sample       | Size 1      | 14        | 3.84 | 2.03      | 1.27 | 0.68      |
|                   | Size 1%–10% | 0         | 0.00 | 1.02      | 0.00 | 0.65      |
|                   | 11%–24%     | 2         | 0.55 | 1.59      | 0.62 | 1.74      |
|                   | 25%–75%     | 11        | 3.02 | 2.23      | 5.35 | 3.90      |
|                   | 76%–99%     | 0         | 0.54 | 0.54      | 0    | 0.48      |
|                   | Total       | 27        | 7.41 | 7.41      | 7.45 | 7.45      |
| European sample    | Size 1      | 15        | 4.23 | 2.22      | 1.50 | 0.71      |
|                   | Size 1%–10% | 0         | 0.00 | 1.11      | 0.00 | 0.68      |
|                   | 11%–24%     | 2         | 0.56 | 1.30      | 0.54 | 1.24      |
|                   | 25%–75%     | 11        | 3.10 | 2.75      | 5.11 | 4.13      |
|                   | 76%–99%     | 0         | 0.51 | 0.51      | 0    | 0.38      |
|                   | Total       | 28        | 7.89 | 7.89      | 7.14 | 7.14      |

Note.—Mutants are grouped into classes by the percentage of sequences carrying them. $K_w$: the contribution to the value of $K\theta_w$, which is Watterson’s estimate of $\theta$. Exp. $K_w$: the expected contribution based on $\theta = K\theta_w$. $\Pi$: the contribution to the value of $\Pi$, the average number of nucleotide differences between two sequences. Exp. $\Pi$: the expected contribution to $\Pi$ based on $\theta = \Pi$. 

Downloaded from https://academic.oup.com/mbe/article-abstract/19/12/2131/997515 by guest on 29 July 2018
cance. Similarly, the increase in haplotype number due to rare mutants is offset by excess of mutants of intermediate frequency, resulting in failure of detecting departure from neutrality by the Fs test. What is most striking from table 7 is the fact that the patterns discussed above, i.e., excess of mutations of size 1 and mutations of intermediate frequency, and deficit of mutations of other types, are persistent over subsamples.

One possible explanation is that ancient human populations were widely subdivided with limited migration among them and that modern humans were derived from many ancient lineages and have undergone a significant expansion in population size in recent times. This scenario would have created a contrast. On the one hand, there would be excess of singleton mutants reflecting the recent population expansion, whereas on the other hand, there would be excess of mutations of intermediate frequency reflecting ancient population subdivision with less migration than that in modern populations.

Effective Population Size

An essential parameter of a population is \( \theta \). For an X-linked locus, \( \theta = 3N_u \), where \( N \) is the effective population size. \( \theta \) is relevant to almost all the statistics one can compute from the polymorphism of a sample, and thus its value is critical in understanding how the population has evolved. There are many estimators available. Among them, Watterson’s (1975) estimator \( \theta_u \) and Tajima’s (1983) estimator II are widely used because of their simplicity and because they were two of the few estimators available for a long time. Several more sophisticated estimators (e.g., Fu 1994a, 1994b; Griffiths and Tavaré 1994; Kuhner, Yamato, and Felsenstein 1995) have been proposed in the last decade, resulting in better estimates when the assumptions made are met; the assumptions are typically a single random mating population and a constant effective population size.

Our analysis in the previous section provided strong evidence that these assumptions do not hold. Although there are estimators that can incorporate population growth, lacking a good knowledge of the population history, particularly the level of population substructure, for the region makes these methods less useful. Our choice of methods is therefore no longer to achieve the best statistical property under neutrality, but rather we choose to use a number of methods, so that adequate comparisons can be made. As a result, the estimates should be regarded as tentative. Among the sophisticated methods, we choose two estimators UPBLUE (Fu 1994a) and BLUE (Fu 1994b) because of our familiarity with these methods and because they are applicable to subsets of mutations.

UPBLUE and BLUE are based on generalized linear models. UPBLUE obtains its estimate from a sample genealogy estimated by the unweighted pair-group method with arithmetic mean method, whereas BLUE obtains its estimate from the frequencies of mutations of various sizes (such as those in tables 4 and 5). BLUE can also be applied to a subset of mutation classes. This is useful for obtaining an estimate of \( \theta \) when one wishes to exclude certain classes of mutations that are known to be strongly affected by an evolutionary force. In our situation, there is an excess of mutation of size 1, and all the scenarios of the human population history have a large effect on the frequency of mutation of size 1. Therefore, it makes sense to obtain a long-term human effective population size from BLUE by excluding mutations of size 1. Another simple estimator \( (\theta_5) \) can be obtained by excluding mutations of size 1 in Watterson’s estimator, resulting in \( \theta_5 = (K - \xi_1)(a_n - 1) \) (Fu and Li 1993). The results of various estimates are given in table 8.

Consider the total sample first. We can see from table 8 that there is a considerable range in the estimates. For \( \mu = 1.55 \times 10^{-4} \), Tajima and Watterson’s estimates of \( \theta \) suggest that \( N = 15,700 \) and \( 20,200 \), respectively. UPBLUE and BLUE estimates suggest even higher values (22,400 and 25,200, respectively). But when singleton mutations are excluded, BLUE yields an estimate of \( \theta \) equal to 5.88, which corresponds to an effective population size of 12,600. Estimator \( \theta_5 \) yields \( N = 13,400 \). As stated earlier, when there is evidence of a significant excess of singleton mutations, it is logical to obtain an estimate of \( \theta \) that relies little or does not rely on singleton mutations. For this reason, we believe that Tajima’s estimate, BLUE(b) and \( \theta_5 \), particularly the last two, are methods of choice; they give the effective population size 15,700, 12,600, and 13,400, respectively.

Next consider each subsample separately. Neither the Asian nor the European subsample has a smaller estimate of \( \theta \) than that of the African subsample. This is surprising, given that so many studies have shown otherwise. In particular, the estimates of \( \theta \) based on Tajima’s estimator are virtually the same for all the three subsamples as well as for the non-African and the total sample. Estimates from UPBLUE and BLUE(a) for separate subsamples are now closer to those by Tajima’s and Watterson’s estimates. Interestingly, BLUE(b) continues to yield considerably smaller estimates for separate subsamples. This indicates again the deficiency of mutations of smaller sizes except for that of size 1. This further confirms the analysis and conclusion made in the previous section.

### Table 8

| Methods          | Total Sample | Africans | Asians | Europeans | Non-Africans |
|------------------|--------------|----------|--------|-----------|--------------|
| Tajima...........| 7.31         | 7.18     | 7.45   | 7.14      | 7.34         |
| Watterson.........| 9.37         | 8.45     | 7.41   | 7.89      | 7.90         |
| Fu’s UPBLUE ......| 10.41        | 8.98     | 6.23   | 7.27      | 7.09         |
| Fu’s BLUE\(^a\)  | 11.71        | 9.85     | 8.10   | 8.36      | 8.89         |
| Fu’s BLUE\(^b\) | 5.88         | 3.81     | 3.64   | 3.39      | 6.01         |
| Watterson.........| 6.22         | 5.10     | 4.91   | 5.10      | 6.35         |
| UPBLUE...........| 4.66         |          |        |           |              |
| BLUE\(^c\)........| 6.98         |          |        |           |              |

\(^a\) BLUE from Fu (1994a).
\(^b\) BLUE excluding singleton mutations.
\(^c\) Watterson’s estimate excluding singleton mutations (see text).
\(^d\) Computed as \(3N_u / \mu = 10,000\) and \(\mu = 1.55 \times 10^{-4}\).
\(^e\) Computed as \(3N_u / \mu = 15,000\) and \(\mu = 1.55 \times 10^{-4}\).
population size for non-Africans combined is slightly larger than that for the total sample. Also the effective population size (12,600–15,700) from the total sample is much smaller than the sum of those for the three populations. This is natural because the populations are not isolated from each other. One message is clear: if one takes the popular view that non-Africans were derived from Africans, then the analysis above shows that non-African populations were not evolved through a bottleneck from a few ancient lineages in Africa.

Traditionally, nucleotide diversity, the mean differences per site between two sequences, is computed as \( \Pi/L \), where \( L \) is the sequence length, which is 10,158 bp here. A comparison with the other three studies of similar scale shows that the nucleotide diversity in the present region is the highest: it is more than twice as large as that in the \( Xq13.3 \) region (Kaessmann et al. 1999) and larger than those in the two autosomal regions studied by Zhao et al. (2000) and Yu et al. (2001). (For a comparison with an autosomal region, the values need to be multiplied by \( 4/3 \) to take care of the smaller effective population size for an X-linked region.) In conclusion, the nucleotide diversity at the present region may be higher than expected for an X-linked region.

Ages of the MRCA and Specific Mutations

The age of the MRCA and ages of certain mutations in a sample are of interest because they allow dating important past events. There are several recently developed methods for estimating the age of the MRCA based on coalescent theory (reviewed in Li and Fu 1999); they differ in both approaches as well as the amount of information taken into consideration. One sophisticated method is attributed to Griffiths and Tavare (1994). This method requires haplotype sequences, which are available here, and has the advantage of being able to estimate the age of each individual mutation in addition to the age of the MRCA. Similar to the situation of estimating \( \theta \) and effective population size, it is desirable to perform age estimation using more than one method, so that a potential bias can be spotted and corrected. But no alternative method as sophisticated as that of Griffiths and Tavare has been published. We therefore also chose an approach developed by one of us (Fu, unpublished data).

Both methods for estimating ages are based on the assumption that the region under study was not subjected to natural selection or linked to a locus under selection. From the analysis in the previous sections as well as the discussion later, we think that this is a reasonable assumption. A straightforward application of either of these methods also assumes that the population size is constant since the MRCA. This is unlikely because our analysis suggests a significant population expansion. Therefore, the estimates of ages need to be taken as suggestive rather than definite. Fortunately, a very recent population expansion will not bias much the estimates of mutations that are relatively large in size; for a mutation of a small size the assumption of a constant population size will in general slightly overestimate the age because the probability of observing more recent mutations is higher when the recent effective population size is larger. Also, we need to recognize that in addition to the population model used, the mutation rate also plays an important role in age estimation.

Griffiths’ program (GENETREE) for age estimation starts by finding a parsimonious representation of the haplotype sequences and then proceeds to estimate the ages of mutations in the tree. The tree generated by GENETREE is given in figure 2, which is essentially the same as that in figure 1. In this method, when \( \theta \) is defined as \( 4N_\mu \), one unit in the age estimation corresponds to \( 2N \) generations. For an X region, \( \theta \) is defined as \( 3N_\mu = 4(3/4)N_\mu \), so one unit corresponds to \( 3N/2 \) generations. The age of the MRCA is 2.47 units, which corresponds to 741,000 years when \( N = 10,000 \); the standard error is 168,000 years. Similarly, the ages of mutation C5790T (mutation 27 in fig. 2), mutation C8728G (mutation 40), and mutation A10151G (mutation 44) are 93,000 ± 45,000, 51,000 ± 33,000, and 69,000 ± 45,000 years, respectively. Although GENETREE gives the standard error of each age estimate, it is not accurate to construct the confidence interval by assuming a normal distribution for the estimate because the age estimate usually has a distribution with a long tail.

Mutation C5790T is of size 15, which led to all Asian sequences, and mutation A10151G is of size 5, which led to the five African sequences in the sample. Notice that the trees in both figures 1 and 2 do not reveal the branching order among the three lineages represented, respectively, by mutations 27, 44, 40 and the lineage represented by mutations 18, 26, and 33. A coalescent simulation conducted by us showed that the first three lineages usually coalesce first and then some African ancestral lineage then coalesces with lineage (18, 26, 33), resulting in the MRCA of the whole sample. The age of the common ancestor of lineages 27, 40, and 44 is of interest because it sets an upper limit on the time an exclusively non-African lineage was separated from African lineages. Although this age can be explored by Griffiths and Tavare’s approach in principle, it is not readily available from GENETREE.

Table 9 shows age estimates using an alternative approach (Fu, unpublished data), which is based on analyzing genealogies that are constrained, so that mutant variants in the sample can be generated under the infinite site model. Furthermore, the estimation was carried out with the constraint that the lineages 27, 40, and 44 did coalesce together and then joint the other lineage to form the MRCA. The estimates for \( N = 10,000 \) agree well with those from GENETREE. We can see that the common ancestor, C, of lineage 27, 40 and 44 has an average age of 144,000 years with 95% confidence interval (66, 264). When \( N = 15,000 \) the mean age of C is 195,000 years old with the confidence interval (87, 348). Figure 3 shows the distributions of the four age estimates in table 9, from which one can see that the assumption of normality for age estimates is generally invalid.

Discussion and Conclusions

Although statistical tests suggest that the neutral Wright-Fisher model should be rejected, we do not be-
believe that the locus under study has been influenced significantly by natural selection. If it had, the type of natural selection must lead to excess in mutations of size 1. Two types of natural selection are known to have such an effect: genetic hitchhiking and background selection. The former assumes that the locus is genetically linked to another locus at which an advantageous allele has been fixed recently, whereas the later assumes that at the linked locus many mutations were deleterious and eliminated from the population. Neither is likely the cause of significant excess of singletons because there are several lines of evidence against the presence of these two forms of natural selection. First, the locus appears to behave normally in terms of the direction of mutations as well as the spatial distribution of mutations. Second, the sequences we obtained are parts of the third exon of the DACH2 gene, which means it is unlikely to be a direct target of natural selection. Third, if genetic hitchhiking or background selection has been operating, the level of genetic diversity in this locus would have been reduced considerably (see Begun and Aquadro 1992; Charlesworth, D., Charlesworth, B., and Morgan 1995; Fu 1997), but it is clear from our estimations of \( \theta \) that the level of within population polymorphism is higher than predicted using the estimated mutation rate and the commonly assumed effective population size. Although we cannot completely rule out the

### Table 9

Age Estimates (in thousand years) by a New Algorithm with \( \mu = 1.55 \times 10^{-4} \)

| Mutation | \( N \)  | Age | 95% Confidence Interval |
|----------|-------|-----|------------------------|
| MRCA     | 10,000| 710 | (492, 994)             |
| C5790T   | 10,000| 103 | (38, 190)              |
| A10151G  | 10,000| 47  | (6, 118)               |
| C        | 10,000| 144 | (66, 264)              |

**NOTE.**—C: The common ancestor of C5790T, A10151G, and C8728G.

Fig. 2.—A sequence genealogy inferred using GENTREE. The haplotypes are explained in table 1. The ordinate denotes the ages of mutations; each unit corresponds to \((3N)/2\) generations.
presence of natural selection, it is unlikely that natural selection is the main force affecting the pattern of polymorphism at this locus. Therefore, we seek alternative explanations for the cause of the excess of singletons.

The most plausible explanation is a relatively recent population expansion. The rapid increase of the human population size in the recent past is an indisputable fact, but whether it should have left detectable trace in the human genome has been debated. The controversy arose partly because different studies have yielded conflicting signals. Population expansion will affect all loci, but whether the effect can be detected differs widely from locus to locus due to variation in sample genealogies and in mutation rate. Loci with a higher mutation rate usually have a better chance for detecting departure from neutrality. The locus we studied here has a mutation rate that is higher than many of the loci reported. Therefore, it should not be a surprise that we are able to detect an excess of singletons. Interestingly, the data reported by Kaessmann et al. (1999) also shows a significant excess of singletons (Zhao et al. 2000; Yu et al. 2001). Because both data sets are from noncoding regions of the X chromosome, the combined evidence for population expansion is quite strong.

Human populations have been obviously subdivided, but how much effect the subdivision has on the genetic diversity has been debated for years. One important conclusion is that population subdivision cannot be the cause of significant excess of singletons, which is a point we demonstrated earlier (Yu et al. 2001). But it is likely one of the main causes that many studies failed to detect significant excess of rare mutations and thus failed to show evidence of significant population expansion. Population subdivision generally increases the number of mutations of intermediate frequencies, which will inflate the value of $D$, while population expansion will inflate the value of $\theta_w = K\lambda u_w$. When both forces have been operating, Tajima’s $D$ can fail to detect population subdivision and population expansion. A casual application of Tajima’s test will often lead to the conclusion of no evidence of departure from the neutral Wright-Fisher model. This study points to the need for a more careful examination of test results, particularly when several statistical tests show different results. Because the results of Tajima’s test as well as the more powerful Fs test for detecting population expansion are far from significant, a careful examination of the mutation pattern led us to conclude that human population subdivision is likely much more severe in ancient times than in the recent past. Our data do not show evidence that non-Africans were derived from African populations or vice versa. If we assume that the popular “Out of Africa” view is correct, then the non-African populations were very unlikely to have been derived through a bottleneck from few lineages in Africa in the last 100,000 years.

The suggestion that there was substantial ancient human population structure has been put forward before. For example, Harris and Hey (1999) made such a conclusion based largely on two observations. One is that the inferred age of the MRCA in their data is very old and the other is that there is fixed segregating site between African and non-African sequences. Although the fixed difference was no longer true when a larger sample size was used (Yu and Li 2000), Harris and Hey’s data does appear to suggest ancient population structure.

The ages of the mutations that are population specific should be informative in dissecting how and when populations are separated. One mutation in our data, C5790T, leads to exclusively non-African sequences in the sample. Approximately, 35% of the 42 non-African sequences was used (Yu and Li 2000). Harris and Hey’s data does appear to suggest ancient population structure.

The ages of the mutations that are population specific should be informative in dissecting how and when populations are separated. One mutation in our data, C5790T, leads to exclusively non-African sequences in the sample. Approximately, 35% of the 42 non-African sequences carry this mutant. To see if this mutant is indeed specific to non-Africans, an additional sample of 106 Africans were typed but none was found to carry this mutant. We also typed 80 additional non-Africans and found that 31% of them carry the mutant. If this
mutation is truly specific to non-Africans, the following two scenarios are possible. One is that it occurred before the separation of the non-African lineages that carried this mutant and their closest African lineages, but the latter became very low in frequency or extinct. The second possibility is that it occurred in a lineage that was outside of Africa. Given the fact that this mutant has a fairly high frequency among non-Africans, the first scenario is less likely. Therefore, the age of this mutation (confidence interval from 66 to 264 thousand years, assuming an effective population size equal to 10,000) suggests that some of the non-African lineages were separated from African lineages quite long ago, possibly even before the emergence of modern humans (100,000–130,000 years BP). Furthermore, the MRCA (the mean age equal to 710,000 years with \( N = 10,000 \)) of the whole sample is also the MRCA of non-Africans (fig. 1), and so the genetic history at this region in Eurasia may be as deep as that in Africa.

The long-term effective size of the human population \( (N) \) is of great importance not only for inferring human history, but also many other analyses; e.g., the age estimation based on coalescent theory. A classical estimate of \( N \) is 10,000 (e.g., Takahata 1993). Many recent studies have suggested much higher values. Although some estimators also yielded large \( N \) values for our data, we feel that given the evidence of excess of singletons, it is more appropriate to use estimators that rely little on singletons. We therefore suggest that the human long-term effective population size is around 12,500–15,000.

Finally, it is important to recognize that each locus in the human genome can capture only a fraction of the human history, and different loci can have rather different genealogies. Thus, some conclusions from different loci are necessarily conflicting. Only after a sufficient number of studies have been conducted, can we gradually reach a consensus about the history of modern humans. The quality of a study is probably more important than the quantity of studies. One important index of the quality is the sample size. Without a sufficiently large sample size, many analyses will be inconclusive or have a large standard error associated with the estimate. For example, had we sampled 50 or more sequences from Africa, Asia, and Europe, it is likely that we would have been able to detect significant excess of rare mutations in all subpopulations.

Acknowledgments

This work was partly supported by NIH grants GM55757 (W.-H.L.) and GM50428 (Y.X.F.) and by Academia Sinica, Taiwan (W.-H.L.). We thank Drs. J. B. Clegg, Lynn Jordae, T. Jenkins, M. Ramsay, Marie Lin, Pekka Pamilo, Laszlo Patthy, and N. Sambuughin for DNA samples.

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.

Charlesworth, D., B. Charlesworth, and M. T. Morgan. 1995. The pattern of neutral molecular variation under the background selection model. Genetics 141:1619–1632.

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. 1994a. A phylogenetic estimator of effective population size or mutation rate. Genetics 136:685–692.

———. 1994b. 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.

———. 1995. Statistical properties of segregating sites. Theor. Popul. Biol. 48:172–197.

———. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147:915–925.

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

Griffiths, R. C., and S. Tavare. 1994. Ancestral inference in population genetics. Stat. Sci. 9:307–319.

Halle-Selassie, Y. 2001 Late Miocene hominids from the Middle Awash, Ethiopia. Nature 412:178–81.

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

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

Kuhner, M. K., Y. Yamato, and J. Felsenstein. 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics 140:1421–1430.

Li, W. H. 1997. Molecular evolution. Sinauer, Sunderland, Mass.

Li, W. H., and Y. X. Fu. 1999. Coalescent theory and its applications in population genetics. Pp. 45–79 in E. Halloran and S. Geisser, eds. Statistics in genetics. Springer Verlag, New York.

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 implications. J. Mol. Evol. 21:58–71.

Nickerson, D. A., S. L. Taylor, K. M. Weiss, A. G. Clark, R. G. Hutchinson, J. Stengrd, V. Salomaa, E. Vartaiinen, E. Boerwinkle, and C. F. Sing. 1998. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat. Genet. 19:233–240.

Tajima, F. 1983. Evolutionary 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.

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

Yu, N., and W. H. Li. 2000. No fixed nucleotide difference between Africans and NonAfricans at the pyruvate dehydrogenase el alpha-subunit locus. Genetics 155:1481–1483.

Yu, N., Z. Zhao, Y. X. Fu et al. (11 co-authors). 2001. Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. Mol. Biol. Evol. 18:214–222.

Zhao, Z., L. Jin, Y. X. Fu et al. (13 co-authors). 2000. Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc. Natl. Acad. Sci. USA 97:11354–11358.

Keith Crandall, reviewing editor

Accepted July 20, 2002