Spatial and Temporal Distribution of the Neutral Polymorphisms in the Last ZFX Intron: Analysis of the Haplotype Structure and Genealogy

Jadwiga Jaruzelska,* 1 Ewa Zietkiewicz,¹ 1 Mark Batzer,¹ David E. C. Cole,§ Jean-Paul Moisan,** Rosaria Scozzari,†† Simon Tavaré‡‡ and Damian Labuda***

*Institute of Human Genetics, Polish Academy of Sciences, 60-479 Poznan, Poland, ¹Centre de Cancérologie Charles-Bruneau, Centre de Recherche de l’Hôpital Sainte-Justine and ²Département de Pédriatric, Université de Montréal, Montréal, Quebec H3T 1C5, Canada, ³Department of Pathology, Stanley S. Scott Cancer Center, Louisiana State University Medical Center, New Orleans, Louisiana 70112, ³³Department of Laboratory Medicine and Pathology, Banting Institute, Toronto, Ontario M5G 1L5, Canada, **Laboratoire de Génétique Moléculaire, Plateau Technique de l’Hôtel-Dieu, Centre Hospitalier Régional et Universitaire, 44035 Nantes, Cédex, France, ‡‡Dipartimento Genetica e Biologia Molecolare, Universita “La Sapienza,” 00185 Rome, Italy and ††Departments of Biological Sciences and Mathematics, University of Southern California, Los Angeles, California 90089-1113

Manuscript received November 19, 1998
Accepted for publication March 19, 1999

ABSTRACT

With 10 segregating sites (simple nucleotide polymorphisms) in the last intron (1089 bp) of the ZFX gene we have observed 11 haplotypes in 336 chromosomes representing a worldwide array of 15 human populations. Two haplotypes representing 77% of all chromosomes were distributed almost evenly among four continents. Five of the remaining haplotypes were detected in Africa and 4 others were restricted to Eurasia and the Americas. Using the information about the ancestral state of the segregating positions (inferred from human-great ape comparisons), we applied coalescent analysis to estimate the age of the polymorphisms and the resulting haplotypes. The oldest haplotype, with the ancestral alleles at all the sites, was observed at low frequency only in two groups of African origin. Its estimated age of 740 to 1100 kyr corresponded to the time to the most recent common ancestor. The two most frequent worldwide distributed haplotypes were estimated at 550 to 840 and 260 to 400 kyr, respectively, while the age of the continentally restricted polymorphisms was 120 to 180 kyr and smaller. Comparison of spatial and temporal distribution of the ZFX haplotypes suggests that modern humans diverged from the common ancestral stock in the Middle Paleolithic era. Subsequent range expansion prevented substantial gene flow among continents, separating African groups from populations that colonized Eurasia and the New World.

A widespread effort to document the amount and geographic distribution of genetic variation in our species is motivated by our curiosity about the origins and prehistory of human populations and our interest in the genetic basis of different diseases. These issues are closely related because the genetic bases of human conditions are a function of the present-day structure of human populations, which by itself cannot be understood without knowing the mechanisms shaping the present genetic variation. Indeed, a starting point for studies that aim to explain the role of genetic variation in disease risk is a description of the quality, quantity, and organization of genetic variation within and between human populations. With the advent of new techniques of DNA analysis it is possible to investigate this variability directly. The greatest progress has been achieved in the analysis of mitochondrial DNA (see Jorde et al. 1998 for a recent review). Nuclear DNA studies have been focused on microsatellite markers (Bowcock et al. 1994; Deko et al. 1995; Jorde et al. 1995, 1997; Di Rienzo et al. 1998; Kimmel et al. 1998; Reich and Goldstein 1998) and on insertion polymorphisms resulting from Alu retropositions (Batzer et al. 1994, 1996; Stoneking et al. 1997). Studies of nonrecombining genetic portions of the Y chromosome are just now gaining momentum (see Jobling and Tyler-Smith 1995; Hammer and Zegura 1996 for the reviews; Underhill et al. 1997; Poloni et al. 1997; Scozzari et al. 1997; Hammer et al. 1998; Malaspina et al. 1998).

Genetic systems differ in their capacity to reveal the information pertinent to current structure and likely population history. Due to differences in effective population size, the time depth of the autosomal diversity is expected to be four times greater and the X-chromosomal diversity three times greater than that of the Y-chromosome or mitochondrial genome. How far one can look back in time depends upon mutation rates. The fast mutation rate of microsatellites makes them suitable for tracing recent evolutionary events, while the presence of recurrent parallel mutations makes homo-
nuclear DNA segments have been systematically surveyed in human DNA samples (non-nominate, characterized only for one human group). Neglecting polymorphisms that were initially characterized in a small number of samples primarily of European origin, which led to genotyping under study is now widely recognized (Bowcock et al. 1991; Mountain and Cavalli-Sforza 1994; Rogers and Jorde 1996), only a few non-Y-chromosome polymorphisms have been systematically surveyed for the presence of sequence polymorphisms in a number of distantly related populations. Variable sites were assessed by direct sequencing in an ~3-kb segment of the β-globin locus (Fullerton et al. 1994; Harding et al. 1997) and a 10-kb portion of the lipoprotein lipase gene (Clark et al. 1998) in 349 and 142 chromosomes, respectively. In another survey of 13 populations from four continents, we screened ~20 independent chromosomes from each of these populations (total of 250) by single-strand conformational polymorphism (SSCP) analysis for polymorphisms in an 8-kb segment of the dystrophin locus (Zietkiewicz et al. 1997). Most of these studies indicate the effective population size to be in the range of ~10,000 individuals, which is an order of magnitude lower than that concluded from diversity data at the major histocompatibility complex (Takahata 1993). Mitochondrial and microsatellite data suggest recent demographic expansion; this, however, is not the case of simple nuclear polymorphisms. The highest molecular diversity is consistently seen in sub-Saharan Africans, but it is variably interpreted as being due to the older age of this population, the higher population size in Africa, and/or the greater gene flow on this continent. Because the history of a single locus is not sufficient to make conclusions about the history of populations, more variability data from independent nuclear loci are needed.

Studying DNA diversity on the X chromosome offers an advantage of straightforward and unequivocal determination of haplotypes in hemizygous (male) samples. The X-chromosome-specific zinc finger protein (ZFX) locus became of particular interest because of the absence of DNA diversity in its Y chromosome homologue, the ZFY gene (Dorit et al. 1995; Huang et al. 1998). Therefore, we have characterized DNA variation in the last introns of ZFX and ZFY in a worldwide sample of human chromosomes (J. Jaruzelska, E. Zietkiewicz and D. Labuda, unpublished results; see also Huang et al. 1998). In the present study, the ZFX polymorphisms were combined into haplotypes. The frequencies of ZFX haplotypes and their geographical stratification in 15 globally distributed populations were analyzed. The age of the underlying mutations and the time to the most recent common ancestor (TMRCA) were estimated from the coalescent haplotype tree. The common distribution of the most frequent haplotypes indicates a diversity that characterized the ancestral population of modern humans. Differential patterns of haplotype variability in Africans and non-Africans reflect more recent events. The data are consistent with the divergence of modern human groups from a common ancestral population ~100 kyr ago, followed by range expansion leading to isolation by distance.

Materials and Methods

Human DNA samples (non-nominate, characterized only by their origin) represented 15 human populations from four continents. Europeans included Polish (30 unrelated chromosomes), French-Canadians from the Province of Quebec (23), Italians (19), and French (21); Asians were represented by Siberian Nentsi (25), Japanese (21), and Chinese from mainland China (25); Amerindians by Ojibwa (19), Maya (23), and Karitiana from Brazil (16); Africans by Biaka Pygmies from Central Africa Republic (22), M’Butu Pygmies from Congo (22), Rimaibe (19), and Mossi from Burkina Faso (27), and African-Americans from Michigan (24).

DNA variation within 1089 bp of the last intron of the ZFX gene on Xp21.3 (Shimmin et al. 1993) in the above sample of 336 worldwide chromosomes was investigated by SSCP/heteroduplex approach, as described earlier (Zietkiewicz et al. 1992a, 1997). Briefly, the analyzed segment was amplified in three overlapping fragments using the following primer pairs: (i) 5′-CGGCAAGCTGGCTAACAA and 5′-ATGCCATTAAATTTTGGAGG (349 bp, annealing temperature 55°C); (ii) 5′-AGGACATGGCTGAACAT and 5′-GTGACAATTTTTCG (306 bp, 55°C); (iii) 5′-CCGGTAACTTACGTTAAG and 5′-CTGTCCCCATTTTTCC (666 bp, 50°C); amplification of the first two fragments was done in the presence of 2% formamide. Prior to electrophoresis in a 6% polyacrylamide gel (acrylamide to bisacrylamide ratio 30:1 and 30:1), the data were digested with EcoRI into fragments of 319 and 347 bp. Mutations underlying gel mobility variants were identified by dideoxy sequencing.

The mutation rate in the 1089-bp ZFX segment, estimated from the human-chimpanzee and human-orangutan comparisons, was 2.55 (±0.44) × 10⁻⁵ per generation assuming a generation time of 20 yr. The ancestral state of the segregating sites that were due to nucleotide substitutions was inferred by comparison with the orthologous positions in chimpanzee and orangutan DNA. Given the mutation rate in the order of 10⁻⁹ per nucleotide position per year, there is a very small chance that a recurrent event has taken place since the divergence of these species from human lineage 5 and 12 mya. In other words, the probability is negligible that a human allele identical with the corresponding site in chimpanzee and orangutan results from a back mutation or that these identical mutations occurred independently after the separation of these lineages. The identity by state of a human allele and the corresponding positions in great apes is in practice tanta...
mount to their identity by descent. Thus, the human allele identical by state with the chimpanzee and orangutan orthologues was considered ancestral, whereas the other allele of the same polymorphic site was considered new.

The derivation of haplotypes from the genotypes was straightforward in hemizygous males (76 chromosomes), in homozygous female samples (112 chromosomes), and those heterozygous at a single position (92 chromosomes). In multiple heterozygotes (female samples having distinct alleles at more than one polymorphic site), the genotype data were easily resolved into haplotypes when two polymorphisms were present within a single SSCP-analyzed fragment (21 individuals). The remaining haplotypes (7 chromosome pairs) were inferred assuming most likely combinations of the unequivocally resolved haplotypes and taking into account their distribution among populations.

A number of summary statistics (haplotype diversity \( G \), mean number of pairwise differences \( \Pi_2 \), nucleotide diversity \( \pi \), and different estimates of \( \theta \), the scaled mutation parameter) were compared using the Arlequin package v.1.1. (Schneider et al. 1997). Arlequin was also used to perform neutrality tests according to Ewens (1972), Watterson (1978), Tajima (1989), and Chakraborty (1990), as well as to assess population genetic structure by analysis of molecular variance (AMOVA; Excoffier et al. 1992). For detailed description of the parameters and tests please refer to the Arlequin help manual (http://anthropologie.unige.ch/arlequin/).

The maximum-likelihood estimate of \( \theta \) using the full information in the sequence data set (a haplotype tree, including the information about the ancestral state) was found using a computational method proposed by Griffiths and Tavaré (1994), implemented in the program genetree and available from the mathematical genetics group web page at http://stats.ox.ac.uk/. genetree was also used to find the distribution of the TMRCA and the distribution of the ages of the underlying mutations, conditional on the gene tree, using the methods described in Griffiths and Tavaré (1999); alternatively, the average age of the mutations was computed from the frequencies of their new (nonancestral) alleles (Kimura and Ohta 1973; Griffiths and Tavaré 1998; Zietkiewicz et al. 1998).

Maximum-likelihood estimates of \( \theta \) obtained with a constant population size model were compared with those using a model of exponential growth. In the exponential growth model there is a decline of the population size backward in time from a current size \( N_s = \theta / 3u \) (22,000 for the maximum-likelihood \( \theta \) estimate of 1.66) such that the population size at time \( t \) ago is \( N_s = N_0 e^{-\beta t} \). Using the genetree program with the expansion rate parameter \( \beta \) of 0.1, 0.2, and 0.3 and the generating \( \theta \) value of 1.66, we obtained no improvement in the log-likelihood compared with the constant population size model (i.e., no evidence for exponential population size expansion was found). The arbitrary \( \beta \) values selected were lower than those suggested in the genetree manual (0.5 and above) to keep them close to more realistic range.

RESULTS

Ten single-nucleotide polymorphisms within the 1089-bp sequence of the last ZFX intron were previously ascertained in a sample of 336 worldwide-distributed chromosomes (J. Jaruzelska, E. Zietkiewicz and D. Labuda, unpublished results). The ancestral and the new allele at each site were inferred by comparison with the orthologous great ape sequences (see materials and methods). The new allele frequencies in the world population varied from 0.3 to 99.1% (Table 1). The average heterozygosity per polymorphic site was 8.9%, corresponding to the nucleotide diversity \( \pi \) (or the mean number of pairwise differences per nucleotide position) of 0.082%. The latter values ranged from 0.06% in Europeans to 0.10% in Asians, while African and Amerindian groups were characterized by \( \pi \) of 0.08% (Table 2). The number of polymorphisms in particular populations ranged from two to four (Figure 1B). One polymorphic site (700) was shared by all the groups; at another (1093), the new allele was fixed in all non-Africans. Three sites (514, 632, and 757) had the new allele present only in non-Africans, and taking into account their distribution among populations.

Nucleotide diversity parameters reported here could be underestimated because of limitations of the SSCP method. On the other hand, use of various gel conditions, analyzing partially overlapping fragments, as well as screening a large sample of chromosomes largely improve the detection rate of DNA polymorphisms. We estimate the efficiency of the SSCP/heteroduplex approach at well above 80%, especially given our long experience in using this technique (e.g., Zietkiewicz et al. 1992a, 1997; Akalin et al. 1994). This is a conservative estimate in light of the success rate of up to 98% reported by other authors (Hayashi 1991; Ravnik-Glavac et al. 1994; Welsh et al. 1997). Moreover, Huang et al. (1998) sequenced the same ZFX segment in a sample of 29 worldwide chromosomes revealing only the ubiquitous site 700. In this context the probability that a frequent polymorphism was missed in our analysis is very low.

The 10 polymorphic positions described above segregated as 11 haplotype variants (Table 1). The phylogenetic relations between these variants are represented by a tree (Figure 1A) that summarizes the underlying history of mutations. The tree is rooted at the ancestral haplotype (H0), in which the alleles at all the polymorphic positions are identical by state with the orthologous positions in nonhuman primates. Two most common haplotypes, H1 and H2 are separated from the ancestral one by one and two mutational steps, respectively, and from each other by a single mutation. The only exception is the single copy of H5/1, which could be regarded as a recombinant of H5 and H1. Given the mutation rate in the order of \( 10^{-3} \) per nucleotide per year, the alternative routes of generating H5/1 through reverse/recurrent mutation at position 700 or at position 502 are much less likely.

The worldwide distribution of haplotype frequencies (Table 1), with H1 and H2 representing 77% of chromosomes, fits very well the infinite-allele model according
TABLE 1

Haplotypes generated from 10 polymorphic sites in the last intron of the ZFX gene

<table>
<thead>
<tr>
<th>Haplotype variant</th>
<th>Haplotype variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>H0</td>
<td>H0</td>
</tr>
<tr>
<td>H1</td>
<td>H1</td>
</tr>
<tr>
<td>H2</td>
<td>H2</td>
</tr>
<tr>
<td>H3</td>
<td>H3</td>
</tr>
<tr>
<td>H4</td>
<td>H4</td>
</tr>
<tr>
<td>H5</td>
<td>H5</td>
</tr>
<tr>
<td>H6</td>
<td>H6</td>
</tr>
<tr>
<td>H7</td>
<td>H7</td>
</tr>
<tr>
<td>H8</td>
<td>H8</td>
</tr>
<tr>
<td>H9</td>
<td>H9</td>
</tr>
<tr>
<td>H5/1</td>
<td>H5/1</td>
</tr>
</tbody>
</table>

New allele frequency (%) 1.19 0.60 0.60 5.06 0.30 8.63 67.26 1.19 5.65 99.11

<table>
<thead>
<tr>
<th>Sequence position</th>
<th>Counts</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>203 T</td>
<td>3</td>
<td>0.89</td>
</tr>
<tr>
<td>400 C</td>
<td>75</td>
<td>22.32</td>
</tr>
<tr>
<td>491 G</td>
<td>183</td>
<td>54.46</td>
</tr>
<tr>
<td>502 A</td>
<td>20</td>
<td>5.65</td>
</tr>
<tr>
<td>514 G</td>
<td>16</td>
<td>4.76</td>
</tr>
<tr>
<td>632 C</td>
<td>4</td>
<td>1.19</td>
</tr>
<tr>
<td>700 A</td>
<td>4</td>
<td>1.19</td>
</tr>
<tr>
<td>716 C</td>
<td>2</td>
<td>0.60</td>
</tr>
<tr>
<td>757 G</td>
<td>1</td>
<td>0.30</td>
</tr>
<tr>
<td>1093 T</td>
<td>1</td>
<td>0.30</td>
</tr>
</tbody>
</table>

| New allele frequency (%) | 1.19 | 0.60 | 0.60 | 5.06 | 0.30 | 8.63 | 67.26 | 1.19 | 5.65 | 99.11 |

| Numbering is that of GenBank sequence HSZFXIN1-X58925; proposed ancestral and new alleles are denoted by upper- and lowercase, respectively. |
| Excluded from statistical analysis (see the text). |

In the absence of recombinations, the haplotype diversity within the ZFX segment results solely from mutation. Leaving aside the H5/1 recombinant, the remaining haplotypes represent sequences that can be treated formally as mitochondrial DNA and/or haplotypes in the nonrecombining portion on the Y chromosome. The lower overall rate of evolution and lower diversity of the ZFX intron, as compared to other genomic segments, suggests that selection may have acted on this sequence (J. Jaruzelska, E. Zietkiewicz and D. Labuda, unpublished results). However, this effect appears too small to influence the population variability of ZFX, which, according to Ewens-Watterson and Tajima tests (not shown and Table 2, respectively), does not differ from that of a neutral locus.

We used a coalescent model to infer the time scale of the origin and evolution of polymorphic variation within the ZFX segment. The distribution of coalescence times and the ages of mutations depend on the mutation rate θ; θ = 2Nμ, where N is the number of chromosomes in the population (for the X-linked locus, N corresponds to 1.5Ne, assuming equal contribution of both sexes to Ne), and μ is the mutation rate per DNA segment per generation. The Ne estimate (of 20,000 individuals) based upon θ is almost twice as large (Table 2). The maximum-likelihood estimate of θ conditional on the ZFX gene tree (θML) is 1.66 (±0.43), and the corresponding Ne of ~22,000 is close to that based on θ. In what follows, we compare the distribution of mutational ages using different estimates of θ (and hence Ne). In principle it is possible to allow for the effect of variability in θ when inferring the ages of mutations and coalescence times (cf., Tavare et al. 1997), but here we treat the values of θ as known.

The mutation tree shown in Figure 2 summarizes the results obtained by applying the approach of Griffiths and Tavaré (1999) to infer mutational ages. This tree is rooted because the ancestral state at each site is known. The distribution of haplotypes among the continental groups is indicated below the tree. Using the
<table>
<thead>
<tr>
<th>Interpretation of molecular diversity at the ZFX locus</th>
</tr>
</thead>
<tbody>
<tr>
<td>World</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>335</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>

**Data**
- Number of chromosomes ($n$):
  - World: 335
  - African: 113
  - Non-African: 222
  - Asian: 71
  - Amerindian: 58
  - European: 93
- Number of polymorphic sites ($S$):
  - World: 10
  - African: 7
  - Non-African: 7
  - Asian: 5
  - Amerindian: 5
  - European: 4
- Number of haplotypes ($k$):
  - World: 10
  - African: 7
  - Non-African: 7
  - Asian: 6
  - Amerindian: 5
  - European: 5

**Population statistics**
- Haplotype diversity, $G$ (SD):
  - World: 0.64 (0.02)
  - African: 0.63 (0.04)
  - Non-African: 0.63 (0.03)
  - Asian: 0.63 (0.04)
  - Amerindian: 0.64 (0.04)
  - European: 0.57 (0.04)
- Mean no. pairwise differences, $P$ (SD):
  - World: 0.89 (0.62)
  - African: 0.84 (0.60)
  - Non-African: 0.89 (0.62)
  - Asian: 1.12 (0.74)
  - Amerindian: 0.82 (0.60)
  - European: 0.67 (0.52)
- Nucleotide diversity (%), $\pi$ = $P/L$ (SD):
  - World: 0.082 (0.045)
  - African: 0.077 (0.043)
  - Non-African: 0.082 (0.049)
  - Asian: 0.103 (0.061)
  - Amerindian: 0.075 (0.049)
  - European: 0.061 (0.041)
- Expected $\theta$ (based on $P$), $\theta_{II} = II$ (SD):
  - World: 0.89 (0.69)
  - African: 0.84 (0.67)
  - Non-African: 0.89 (0.69)
  - Asian: 1.12 (0.82)
  - Amerindian: 0.82 (0.66)
  - European: 0.67 (0.57)
- Expected $\theta$ (based on $S$), $\theta_S = S\Sigma^{-1} (1/i)$ (SD):
  - World: 1.56 (0.57)
  - African: 1.32 (0.57)
  - Non-African: 1.17 (0.45)
  - Asian: 1.03 (0.52)
  - Amerindian: 1.08 (0.47)
  - European: 0.78 (0.43)
- Tajima’s estimator, $D$:
  - World: -0.95
  - African: -0.83
  - Non-African: -0.51
  - Asian: +0.18
  - Amerindian: -0.57
  - European: -0.30
- Expected $\theta$ (maximum likelihood, $\theta_{ML}$) (SD):
  - World: 1.66 (0.43)
  - African: 1.36 (0.62)
  - Non-African: 1.06 (0.49)
  - Asian: 1.08 (0.49)
  - Amerindian: 0.91 (0.47)
  - European: 0.81 (0.46)
- Expected number haplotypes, $E(k)$:
  - World: 8.0
  - African: 6.4
  - Non-African: 7.3
  - Asian: 5.7
  - Amerindian: 5.7
  - European: 5.1

**Effective population size**
- $N_{ML}$ (SD):
  - World: 11,600 (9,000)
  - African: 11,000 (8,800)
  - Non-African: 11,600 (9,000)
  - Asian: 14,600 (10,700)
  - Amerindian: 10,700 (8,800)
  - European: 8,800 (7,500)
- $N_{ML}$ (SD):
  - World: 20,400 (7,500)
  - African: 17,300 (7,500)
  - Non-African: 13,100 (5,900)
  - Asian: 13,500 (6,800)
  - Amerindian: 11,200 (6,100)
  - European: 10,200 (5,600)
- $N_{ML}$ (SD):
  - World: 21,700 (5,600)
  - African: 17,800 (8,100)
  - Non-African: 13,900 (6,400)
  - Asian: 14,100 (6,400)
  - Amerindian: 11,900 (6,100)
  - European: 10,600 (6,000)
- Coalescent time, $T_{ML}$ (SD):
  - World: 2.52 (0.90)
  - African: 2.70 (1.02)
  - Non-African: 2.57 (1.05)
  - Asian: 2.62 (1.05)
  - Amerindian: 2.61 (1.14)
  - European: 2.79 (1.22)

* Excluding variant H5/1.
* Mutation rate, $2.55 \times 10^{-5}$ per 1089-bp segment per 20-yr generation.
* In $N_e$ generation units, using $\mu_{ML}$ ($2 \times 10^6$ simulation runs).
The expected TMRCA values of 2.5 and 3.1 Ne generation units correspond to 1100 and 740 kyr, respectively. TMRCA also represents the expected age of the ancestral haplotype H0. Two globally distributed haplotypes H1 and H2, which arose as the result of mutation at sites 1093 and 700, are 550 to 840 kyr and 200 to 400 kyr old, respectively. The estimated ages of the remaining haplotypes, i.e., those with the restricted or patchy geographical distribution, vary from 5±9 to 120±180 kyr (Figure 2). The average age of each polymorphism estimated from the full data is very similar to the average age computed from just the frequencies of its new (non-ancestral) alleles (Kimura and Ohta 1973; Griffiths and Tavare 1998; Zietkiewicz et al. 1998). Figure 3 compares the average ages of polymorphisms (in Ne generation units) and associated 5 and 95% percentiles obtained from the new allele frequencies with those obtained from the full data using the estimates \( \theta_M \) and \( \theta_N \) of \( \theta \). The expected age of the recombinant H5/1 is the same as that of haplotype H9.

Table 2 summarizes the population parameters estimated for the world, continents, and all non-Africans together. The estimates of population parameters based on \( \pi \) are generally (except for Asians) lower than those based on \( S \), but this difference is not statistically significant. This holds, whether we consider the whole sample or its subsets as representative of the world population. Thus, these population groups look quantitatively similar in spite of the qualitative differences in distribution of young haplotypes. The respective parameters calculated for each of the 15 populations analyzed (not shown) are within the range indicated by the standard deviation values for the larger, continental groups. It is noteworthy that the coalescent time (Table 2) for the world and for the continental samples is almost identical, suggesting again that these groups spread recently from a common ancestral stock and share most of their earlier genetic history (see also Di Rienzo et al. 1998).

To assess the level of population structure, we estimated the \( F_{ST} \) parameter, describing the contribution of variance among populations to total variance. The \( F_{ST} \) values were computed using an AMOVA test (Excoffier et al. 1992) based on the pairwise distances between haplotypes. The significance of the components of variance was assessed by comparing observed levels with the distribution of 16,000 values obtained by randomization. \( F_{ST} \) calculated for the world population, considered to be composed of 15 subpopulations, was 0.081, indicating that only 8.1% of the total variance resulted from the differences among populations and \( \sim 92\% \) was due to the variance within populations. When the world was divided into four continental populations or only into Africans and non-Africans, \( F_{ST} \) values were 0.067 and 0.048, respectively. The \( F_{ST} \) value for Africans was 0.064 and for the cluster combining the remaining groups,
Figure 2.—Gene tree of the ZFX polymorphisms in the sample of 336 worldwide chromosomes. Mutations are indicated on the branches at the distances proportional to their estimated average age ($2 \times 10^7$ simulation runs), using $\theta_{ml}$ (maximum-likelihood estimate of $\theta$) or $\theta_{up}$. The order of mutations at positions 400 and 491 cannot be resolved. The values at the right correspond to the coalescent times in $N_e$ generations or, between parentheses, in years (with a generation time of 20 yr and $N_e$ conditional on $\theta_{ml}$ or $\theta_{up}$, i.e., $N_e = 22,000$ or 12,000, respectively). The corresponding haplotypes are indicated below the tree; the presence of a specific haplotype in the continental population is indicated by a filled box whereas a hatched box indicates the presumed admixture (see text).

0.062. The results for Europe, Asia, and the Americas were nonsignificant, indicating the absence of population structure in these groups, which is consistent with the conclusions based on other summary statistics above.

DISCUSSION

The geographical distribution of the DNA diversity in the 1.1-kb ZFX segment is not uniform. Two out of 11 haplotypes (H1 and H2) are ubiquitous and represent 77% of the worldwide sample of 336 chromosomes. Similar prevalence of common haplotypes was also found at other genomic loci. At the 2.67-kb $\beta$-globin segment (Harding et al. 1997), 5 common haplotypes contributed $\approx 77\%$ to the worldwide diversity in a sample of 349 chromosomes, and at the 8-kb-long dys44 segment, 6 frequent haplotypes represented $\approx 76\%$ of the worldwide sample of 973 chromosomes (our unpublished data). The remaining 9 haplotypes at the ZFX locus have patchy or continentally restricted distribution. There is good reason to believe that the occurrence of haplotype H6 in African-Americans is due to European admixture in this population (Chakraborty et al. 1992), although this will have to be confirmed by studies of a greater number of local populations and/or chromosomes. Similarly, haplotype H5 in Maya may be the result of an African-American admixture. The
observation of Seielstad et al. (1994), who found evidence of African admixture in the same Mayan samples, reinforces our conclusion; African-American admixture was also observed among other Amerindian populations by A. Ruiz-Linares (personal communication). We can therefore tentatively assign 5 of the haplotypes (H0, H5, H7, H8, and H5/1) to populations from Sub-Saharan Africa. Haplotypes H3, H4, H6, and H9 appear specific to the group of non-African populations. Quantitatively, the haplotype diversity (6 values in Table 2) is similar in Africa and the remaining continents. Summary statistics based on the mean number of pairwise differences (θm) indicate the greatest diversity in Asia. However, similar to what was observed at the β-globin locus (Harding et al. 1997), the estimates based on the number of segregating sites as well as maximum-likelihood estimates from the whole data set show that the diversity is greatest in Africa (θ0 and θMLE).

All continentally restricted haplotypes except for the ancestral H0 are estimated to have arisen between 120 and 180 kya and the present, on the basis of a coalescent analysis of the full data, and between 160 kya and the present using the mutation ages estimated from the new allele frequency (Figures 2 and 3). In contrast, variants H1 and H2 estimated at 550 to 840 and 200 to 400 kyr, respectively, are found at high proportions in all populations except from Chinese (Figure 2). The absence of H1 from Chinese, where it is replaced by the much younger haplotype H3, may reflect founder effect associated with peopling of Asia and its subregions (e.g., Wang et al. 1991). The presence of H1 in two other Asian populations, the paucity of H3 in Europeans, as well as the tree structure in Figure 1A (indicating that H1 “missing” in Chinese is an intermediary form between H0 and both H2 and H3) preclude the interpretation that the two most common haplotypes in the Chinese reflect ancestral haplotype frequencies. Rather, it is the presence of H1 and H2 and their frequency profile that are similar across the continents (Figure 1C), together with the oldest variant H0 that is still detectable in Africa, which can be regarded as the ancestral configuration.

Considering the continents separately, one obtains the same coalescent time (Table 2) as for the pooled sample representing the world. A similar observation was made with microsatellite data (Di Rienzo et al. 1998), which are consistent with the shared roots of the contributing populations throughout the substantial period of their earlier genetic history. Worldwide similarity of the frequency profiles of the most common haplotypes could reflect extensive gene flow between the continents. An AMOVA test performed to provide estimates of variance components and of the F-statistics representing the correlation of haplotypes at different levels of hierarchical groupings revealed low variation among constituting populations or groups of populations as compared to the variation within populations. This is similar to the observation at the dystrophin locus (Zietkiewicz et al. 1997) and could again indicate a high degree of gene flow connecting the populations as well as continents. However, if such intercontinental genetic exchange existed during the last 100 to 200 kyr, why would a number of distinct young haplotypes persist in isolation, either in Africa or outside, in Eurasia, and the Americas? Likewise, why would the oldest H0 occur at detectable frequency only among Sub-Saharan Africans? We argue that if substantial genetic exchange occurred, it was at the time of ancestral population divergence, preceding the expansion of modern groups. Thereafter, gene flow was restricted by geographic isolation and physical distance between populations colonizing new continents.

With reduced genetic drift, population growth could be conducive to maintaining the frequency profile of polymorphisms characteristic of the ancestral stock from which expanding populations diverged. The excess of the number of segregating sites over the nucleotide diversity (i.e. the negative value of Tajima’s D parameter) and the excess of the observed over expected number of haplotypes, E(k), according to Ewens (1972) sampling (see World, Table 2) may suggest that a weak trace of the recent demographic growth is reflected in our sample. However, selection could also be evoked to explain these observations (J. Jaruzelska, E. Zietkiewicz and D. Labuda, unpublished results). Moreover, other non-Y-chromosome nuclear loci studied in a similar way provide no evidence for population expansion. Hey (1997) indicated that most nuclear genes show slightly positive skew in frequency spectrum, and studies of β-globin, dystrophin, and LPL loci (Clark et al. 1998; Harding et al. 1997; Zietkiewicz et al. 1997, 1998) reported positive D associated with the non-Y-chromosome nuclear diversity. On the other hand, these observations do not allow firm conclusions. The positive D values reported for β-globin, dystrophin, and LPL, as well as the negative ones observed here for ZFX were not significantly different from zero. This is further illustrated by the fact that the D values calculated here for individual populations were both positive and negative (not shown). The coalescent estimates from β-globin data (Harding et al. 1997) did not perform better under a model of expanding population than under a constant population size model, which was also the case for ZFX (see materials and methods). Thus, if there was any population size expansion, it was not detected by the exponential model or it was probably too recent to be detectable in the observed patterns of ZFX and/or β-globin diversity. On the other hand, a model that assumes expansion from the Middle/Upper Paleolithic up to the present is suggested by paleontological studies (Lahr and Foley 1998). A recent population growth is also consistent with genetic studies of mitochondrial DNA or nuclear microsatellites (Di Rienzo and Wilson 1991; Slatkin and Hudson 1991;
Rogers and Harpending 1992; Jorde et al. 1997; Di Rienzo et al. 1998; Kimmel et al. 1998). These genetic systems, because of the higher mutation rates, are more suitable for detecting expansion at relatively short evolutionary time scales in the range of 100 kyr. In conclusion, most nuclear loci characterized for simple nucleotide polymorphisms imply a different past in comparison to mitochondrial or microsatellite studies (Jorde et al. 1995; Hey 1997).

Interestingly, the answer to these discrepancies may lie in the analysis of extended haplotypes, where nuclear diversity generated by rare mutational events would be enhanced by recombinational events, giving insight into more recent population histories, and corresponding to the time frame of mitochondrial and microsatellite mutations. It should be emphasized that the virtual absence of recombinants in ZFX does not necessarily indicate lack of recombination in this region. Assuming the genomic average of 44 crossovers per female meiosis per generation (Brøman et al. 1998) and taking into account that only two-thirds of the X chromosome population participates in female meioses, an average genomic crossover rate is $10^{-8}$ per base pair per generation, i.e., $\sim 10^{-2}$ per 1089 bp in the ZFX segment analyzed. Yet, informative chromosomal pairs that would lead to detectable recombinants represent only 3% of all ZFX haplotype combinations (in other words, 97% of recombinations between “noninformative” ZFX segments remain undetected). Thus, an apparent recombination rate in ZFX is $3 \times 10^{-7}$, a hundred times less than the mutation rate of $2.6 \times 10^{-5}$ in the same segment.

The ZFX data truncation (a single chromosome representing the recombined haplotype 5I/1 was excluded) introduced for the sake of the coalescent analysis was minimal. In some loci, however, the contribution of crossovers to haplotype diversity may be non-negligible. Such an effect is very pronounced in the recently studied LPL locus (Clark et al. 1998; Nickerson et al. 1998), as well as in the dys44 locus (our unpublished data). In this context the good fit between the mutational age estimates based on the full ZFX data and those using just the frequencies of the new allele (Figure 3) is noteworthy. Similar concordance was shown for the β-globin data of Harding et al. (1997; see Figure 4 in Zietkiewicz et al. 1998). Although the latter age estimate has a larger variance, it can be calculated for each mutation in the haplotype without knowing the underlying tree.

The genetic diversity measures, pointing to the population size of $\sim 10,000$ to 20,000, presumably reflect the size of the ancestral population or speciation bottleneck rather than the harmonic mean of the effective sizes of the world population from Middle Pleistocene to the present. Speciation bottleneck and/or divergence and range expansion from a relatively small ancestral population would have a greater impact on the number of segregating sites than on the nucleotide diversity. If such an event had taken place relatively recently, it would have left no time (depending also upon the size of the expanding populations) for the system to fully recover and reach the equilibrium. As a result, we may observe either the shortage of segregating sites, asi non-African populations at the dys44 locus (Zietkiewicz et al. 1998) or a number of new sites that contribute to $\theta$, but very little to nucleotide diversity, asi Africans at dys44 (Zietkiewicz et al. 1998) and in this study. The net effect on the world diversity reflected by $n$ and $S$ may be that the data fit well and can be thus analyzed in terms of a constant population size model.

Our interpretation of the ZFX data is supported by similar results obtained for the β-globin (Harding et al. 1997) and dys44 (Zietkiewicz et al. 1998) loci. Taken together with mitochondrial, microsatellite, and paleontological evidence, these data suggest the divergence of modern human groups from a common ancestral population $\sim 100$ kyr ago (taking into account that population trees are usually shorter than these of the underlying genes) followed by range expansion leading to the isolation by distance. What is the geographic origin of the ancestral population? Greater genetic diversity in Africans has frequently been taken as evidence of this population’s antiquity (Nei et al. 1975; Batzer et al. 1994; Horai et al. 1995; Tishkoff et al. 1996). The fact that genetic diversity outside of Africa can be interpreted as being a subset of the African set has been seen as even more compelling evidence for the out-of-Africa origin (Stoneking et al. 1997; Jorde et al. 1998). However, in the dys44 study, the excess of the African diversity was shown to have accumulated only following the divergence of modern humans (Zietkiewicz et al. 1998). Greater diversity in one group may be due to its better representation in the worldwide sample analyzed (Bowcock et al. 1991; Mountain and Cavalli-Sforza 1994; Rogers and Jorde 1996) or to the way we group or define the populations. It may also reflect greater long-term effective population size following expansion (Relethford 1995; Relethford and Harpending 1995; Rogers and Jorde 1995) or particular history of founder effects, extinctions, and expansions of local populations examined (Takahata 1994).

Assuming that the ancestral population diverged and underwent range expansion in the Middle/Upper Paleolithic, old haplotypes could have been carried away from the place of origin and preferentially preserved in populations that colonized new areas conducive to relatively uninterrupted growth and demographic prosperity. Another model, which can be described as a single-species model (see Foley 1998), favors extensive gene flow among distant hominid populations over the whole Paleolithic and thus the contribution of dispersed ancient populations to modern human origins (reviewed in Templeton 1997). The argument becomes circular if assuming gene flow we subsequently argue that the presence of a haplotype in a geographically limited sample points to its ancestral place of origin.
there. Thus, a demonstration of an archaic lineage on one or another continent does not prove anything under any of the above scenarios. With migration and range expansion playing an important role in all the underlying models, the geographic location of the modern human origin cannot be inferred from the genetic evidence alone. Paleontological findings and paleoclimatic data (Tattersall 1995; Lahr and Foley 1998) place the origin of the ancestral population in Africa or the Middle East, between the Klasies River and Qafzeh. Our data neither contradict nor prove this scenario. We believe, however, that integrating data from a variety of genetic systems and interpreting them in the context of the outside evidence will be more illuminating in this respect. The need for more extensive sampling of human populations from different geographic areas and of different loci from the nuclear genome appears more than evident, and this study is one of the contributions to fulfill this task.

Thanks to Dominik Gehl for his assistance in computing and to Raffaella Ballaran for typing the manuscript. This work was supported by National Science Foundation grants DMS 95-04393 to S.T. and SBR-9610147 to M.B., by the Canadian Genome Analysis and Technology Program, and by the Medical Research Council of Canada grant to D.L.

LITERATURE CITED


ZFX Haplotypes in Human Populations


Communicating editor: A. G. Clark