J Mol Evol (1998) 47:146–155

JOURNAL OF MOLECULAR EVOLUTION

Genetic Structure of the Ancestral Population of Modern Humans

Ewa Ziętkiewicz,[1] Vania Yotova,[1] Michal Jarnik,[1,*] Maria Korab-Laskowska,[1,†] Kenneth K. Kidd,[2] David Modiano,[3] Rosaria Scozzari,[4] Mark Stoneking,[5] Sarah Tishkoff,[2,‡] Mark Batzer,[6] Damian Labuda[1]

[1] Centre de Recherche de l'Hôpital Sainte-Justine, Centre de Cancérologie Charles Bruneau, Département de Pédiatrie, Université de Montréal, Montréal, Québec, H3T-1C5 Canada
[2] Department of Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06510, USA
[3] Fondazione Pasteur Cenci-Bolognetti, Istituto di Parassitologia, Universita ''La Sapienza,'' P. le A. Moro 5, 00185 Rome, Italy
[4] Dipartimento Genetica e Biologia Molecolare, Universita ''La Sapienza,'' P. le A. Moro 5, 00185 Rome, Italy
[5] Department of Anthropology, Penn State University, University Park, PA 16802, USA
[6] Department of Pathology, Stanley S. Scott Cancer Center, Louisiana State University Medical Center, 1901 Perdido Street, New Orleans, LA 70112, USA

Received: 3 February 1998 / Accepted: 9 February 1998

Abstract. Neutral DNA polymorphisms from an 8-kb segment of the dystrophin gene, previously ascertained in a worldwide sample ($n = 250$ chromosomes), were used to characterize the population ancestral to the present-day human groups. The ancestral state of each polymorphic site was determined by comparing human variants with their orthologous sites in the great apes. The ''age before fixation'' of the underlying mutations was estimated from the frequencies of the new alleles and analyzed in the context of these polymorphisms' distribution among 13 populations from Africa, Europe, Asia, New Guinea, and the Americas ($n = 860$ chromosomes in total). Seventeen polymorphisms older tan 100,000–200,000 years, which contributed ~90% to the overall nucleotide diversity, were common to all human groups. Polymorphisms endemic to human groups or continentally restricted were younger than 100,000–200,000 years. Africans (six populations) with 13 such sites stood out from the rest of the world (seven populations), where only 2 population-specific variants were observed. The similarity of the frequencies of the old polymorphisms in Africans and non-Africans suggested a similar profile of genetic variability in the population before the modern human's divergence. This ancestral population was characterized by an effective size of about 10,000 as estimated from the nucleotide diversity; this size may describe the number of breeding individuals over a long time during the Middle Pleistocene or reflect a speciation bottleneck from an initially larger population at the end of this period.

Key words: Human evolution — DNA polymorphisms — Dystrophin locus — Population size — Neutral evolution — Mutation rate

Introduction

Early immunological studies (Goodman 1962; Sarich and Wilson 1967) showed that the African apes were much closer relatives of *Homo sapiens* and that their lineages diverged more recently than classical primate systematists and paleontologists maintained. Protein polymorphisms pointed to the recent origin of modern human groups by indicating that the divergence between Africans and non-Africans had taken place about 110 ky (110,000 years) ago (Nei and Roychoudhury 1974).

* *Present address:* NIAMS, LSBR, Bethesda, MD, USA
† *Present address:* Département de Biochimie, Université de Montréal, Montréal, Quebec, H3T 1C5 Canada
‡ *Present address:* Department of Biology, Pennsylvania State University, University Park, PA 16802, USA
Correspondence to: D. Labuda; *e-mail:* Labuda@ere.Umontreal.ca

During the last decade molecular studies in human population genetics were focused on DNA variability, particularly that of the mitochondrial genome. The resulting information played an important role in shaping our views on the origins of modern humans but it did not settle the debate, which still oscillates between two competing models: ''regional continuity'' and ''recent African origin.'' According to the first model (Weidenreich 1943) different human groups evolved regionally from archaic subpopulations of *H. erectus,* who colonized Eurasia almost 2 million years ago (Swisher et al. 1994; Gabunia and Vekua 1995); in the second model, much of the present-day diversity of our species has been acquired only recently, following dispersal of the anatomically modern humans from Africa. Although new molecular and fossil evidence supports a ''recent African origin'' (Cann et al. 1987; Scozzari et al. 1988; Stringer and Andrews 1988; Bowcock et al. 1991, 1994; Ruvolo et al. 1993; Cavalli-Sforza et al. 1994; Lahr and Foley 1994; Leakey 1994; Mountain and Cavalli-Sforza 1994; Horai et al. 1995; Jorde et al. 1995; Batzer et al. 1996; Knight et al. 1996; Tishkoff et al. 1996; Perez-Lezaun et al. 1997; Shriver et al. 1997; Tattersall 1997), the ''multiregional'' hypothesis still has its proponents (Wolpoff and Caspari 1997). Neither of the two models has been proven and it is well recognized that the extreme forms of the two scenarios are seriously oversimplified.

To characterize the ancestral human population the present-day polymorphisms must be ascertained and analyzed in an unbiased manner in a global array of extant populations (Bowcock et al. 1991; Mountain and Cavalli-Sforza 1994; Cavalli-Sforza et al. 1994; Rogers and Jorde 1996). Since an appropriately deep insight into the past is required, the fact that the average lifetime of nuclear autosomal variability is four times longer (three times in the case of the X chromosome) than that of mitochondrial DNA favors studies of the former (Kimura 1983; Hartl and Clark 1989). Although recombinations in the nuclear genome are a potential source of additional information (Hudson 1983; Hudson and Kaplan 1988; Hey and Wakeley 1997; Labuda et al. 1997), they may constitute a certain drawback in the analysis of non-Y-chromosome nuclear DNA. In most instances recombinations make it difficult to infer detailed genealogies with any reliability (Hudson 1990). This in turn renders difficult the application of coalescence analysis to estimate the time to the most recent common ancestor as well as the age of mutations underling the polymorphisms. This problem can be overcome by restricting the analysis to the regions without recombination, or practically without (Harding et al. 1997), or to a manageable number of recombinants (Hey and Wakeley 1997), but the investigator does not always have such a choice. However, even without recurring to the coalescent, an analysis does not have to be limited to a mere description of the contributing polymorphisms. Given that the ancestral state of a polymorphic site is known, a different approach based on calculating the ''age before fixation'' of new alleles (Kimura and Ohta 1973; Maruyama 1974) can be used to incorporate time estimates into the analysis.

Recently, we have characterized DNA polymorphisms in an intronic segment of the dystrophin gene in a sample of worldwide populations (Zietkiewicz et al. 1997). This DNA segment has characteristics suggesting that it might be a good model of nuclear sequence variation to study human evolution. The dystrophin gene spans several centimorgans of the genetic distance (Nagaraja et al. 1997), which substantially reduces the chances of a hitchhiking effect from any closely linked strongly selected locus (Maynard Smith and Haigh, 1974). Seventy-nine dystrophin exons [14 kilobases (kb) in total] are spread over 2.4 Mb of DNA on chromosome Xp21 (DenDunnen et al. 1992; Roberts et al. 1993). This low gene density is consistent with the dystrophin location in a DNA region of poor G+C content, corresponding to the L1 isochore family (Bettecken et al. 1992). Structural redundancy at the middle region of the protein may also relax the selective pressure on the coding sequence itself; dystrophy patients with internal in-frame deletions of large portions of the gene are affected with a mild (Becker, in contrast to severe Duchenne) form of the disease (Koenig et al. 1989; Simard et al. 1992). None of the polymorphic sites (Zietkiewicz et al. 1997) used in the present study carry disease alleles. This is consistent with the assumption of neutrality of the observed variation.

Here we report on using Kimura and Ohta's (1973) approach to estimate the age of the *dys44* polymorphisms from the frequencies of their new alleles, distinguished from their ancestral state by comparing human variants with the orthologous sites in great apes. The age estimation, combined with the information on the geographical stratification of the polymorphisms in 13 globally distributed populations, allowed us to characterize the ancestral genetic stock from which the present-day human populations originated, thus shedding new light on the ancient history of our species.

Materials and Methods

Human DNA samples [nonnominative, characterized only by their origin (see Zietkiewicz et al. 1997)] included mixed Europeans (175 unrelated chromosomes); Asians—Siberian Nentsi (24), Japanese (67), and Chinese (85); Amerindians—Maya (80) and Karitiana (83); Papuans from New Guinea (69); Africans–Biaka Pygmies (89), M'Buti Pygmies (58), Rimaibe (23), Mossi (25), Nigerians (15), and African Americans (67). Chimpanzee, gorilla, and orangutan DNAs were obtained from ATCC and BIOS or isolated from the peripheral blood specimens obtained from Granby and Québec Zoological gardens.

DNA sequence *dys44* (U94396) spanned dystrophin exon 44 (cDNA positions 6499 through 6646) and its flanking introns between positions −2853 to −1 upstream and 1 to 5034 downstream (total length, 8035 bp). Excluding the terminal primers and short internal

segments where the PCR fragments did not overlap, the total length of the analyzed sequence was 7622 bp. Polymorphisms were detected by single-strand conformational polymorphism (SSCP) combined with heteroduplex analysis (Orita et al. 1989; Zietkiewicz et al. 1992) of *dys44* fragments and characterized by sequencing (Zietkiewicz et al. 1997). In brief, typically 20 chromosomes (15 in Nigerians) from each population were analyzed in this way, amounting to the analysis of almost 2 Mb of DNA to ascertain polymorphisms. The presence of a multiallelic $(T)_n$ microsatellite polymorphism in a fragment encompassing positions 2453 to 2839 rendered SSCP analysis of the adjacent polymorphisms difficult; for this reason, African American and Papuan samples were omitted in the systematic SSCP screening of this fragment. A total of 860 chromosomes was subsequently typed by ASO (allele-specific oligonucleotide) hybridization for all polymorphic sites excluding the $(T)_n$ microsatellite, analyzed by denaturing gel electrophoresis.

Positions orthologous to all human polymorphic sites were examined in DNA from great apes by sequencing PCR-amplified segments including these sites or by ASO hybridization. At this stage the analysis was focused on sequence positions corresponding to the human polymorphic sites. The human allele identical by state to at least two of the great ape orthologues was considered ancestral. In addition, five fragments orthologous to human positions −1590 to −999, −537 to −67, 661 to 1173, 1663 to 2242, and 4347 to 5034 (a total of 2637 bp, excluding primers) were carefully resequenced at their whole length in chimpanzee, gorilla, and orangutan; interspecific genetic distances determined from the comparison of the human and great ape sequences [according to the Kimura two-parameter model, using the DNADIST program from Phylip 3.5 (Felsenstein 1993)] were used to compute the mutation rate in the *dys44* region.

The population parameters, h_i, H_n, F_{ST}, and N_e, were calculated from allele frequencies averaged across 13 (world), 6 (African), or 7 (non-African) corresponding populations. Heterozygosity h_i at each sequence site i is

$$h_i = \sum_{j=1}^{a} p_j(1 - p_j)$$

where p_j is the frequency of the jth allele and a is the number of alleles at the ith site.

Average heterozygosity per nucleotide H_n [since $n \gg 1$, the term $n/(n - 1)$ can be neglected such that H_n corresponds to nucleotide diversity] was calculated as a sum of heterozygosities h_i [excluding $(T)_n$ microsatellite site heterozygosity, reported separately as H_m], divided by the total length L ($L = 7622$) of the DNA segment analyzed:

$$H_n = \sum_{i=1}^{L} h_i/L$$

The standard error of the nucleotide diversity was calculated as the square root of its variance divided by the number of analyzed positions:

$$\mathrm{SE} = \left[\sum_{i=1}^{L} (h_i - H_n)^2/(L-1)L \right]^{1/2}$$

The corresponding H_n values for the groups of populations, such as Africans, non-Africans, and world, were obtained by using allele frequencies averaged across the subpopulations constituting the group.

The F_{ST} statistic was calculated to Hartl and Clark (1989) as

$$F_{ST} = \frac{H_n - \overline{H_n}}{H_n}$$

where H_n is the nucleotide diversity in the group of subpopulations (as above), whereas $\overline{H_n}$ is the average diversity computed by dividing the sum of the nucleotide diversities of the constituting subpopulations by their number.

Assuming random mating, neutrality, constant population size, and the infinite-site model of mutation (Kimura 1983), for $H_n \ll 1$ the long-term effective population size N_e can be obtained from the equation $H_n = 4N_e\nu$, where ν is the mutation rate per nucleotide site per generation. Here, however, since the number of X chromosomes is $1.5N_e$ rather than $2N_e$ (assuming equal contribution of both sexes to N_e), the population size $N_e = H_n/3\nu$. It is important to emphasize that the effective population size is that of the entire species and not of local populations (Kimura 1983). Thus, when N_e was calculated using either African or non-African chromosomes, both of these samples were considered as being representative of the world.

The average age of an allele before fixation, $\bar{a}(p)$, expressed in N_e generations units as a function of the nonancestral allele frequency, was calculated according to (Kimura and Ohta 1973; see also Maruyama 1974; Kimura 1983)

$$\bar{a}(p)/N_e = -3p(1 - p)^{-1}1np$$

where p denotes the frequency of a new, nonancestral allele (a factor of 3 rather than 4 is used because the polymorphisms considered are on the X chromosome). $\bar{a}(p)$ refers to the time span since the appearance of a new mutation until its present frequency p. This estimation assumes the population to be of constant size throughout the relevant period of its history. It has to be considered an approximation which is as good as the fulfillment of the underlying model by the analyzed population. The standard deviation of the age estimate was calculated in units of N_e generations as the square root of its variance:

$$V(\bar{a}) = \overline{a^2(p)} - [\bar{a}(p)]^2$$

where

$$\overline{a^2(p)}/N_e{}^2 = 18\left[p/(1-p) - \int_o^p (1nz/(1-z)dz \right]$$

The test for neutrality was performed according to Tajima (1989) by calculating

$$D = [H_nL - (S/a_1)]/[e_1S + e_2S(S-1)]^{1/2}$$

where H_n,S, and L are as defined earlier, n denotes the number of chromosomes orignally examined by SSCP,

$$a_1 = \sum_{i=1}^{n-1} 1/i,$$

$$a_2 = \sum_{i=1}^{n-1} 1/(i)^2,$$

$$e_1 = [(n+1)/3(n-1) - (1/a_1)]/a_1, \text{ and}$$

$$e_2 = [2(n^2+n+3)/9n(n-1) - (n+2)/a_1n + a_2/(a_1)^2]/(a_1^2 + a_2).$$

Results

Thirty-six segregating sites were found within the 8-kb *dys44* DNA segment encompassing exon 44 of the hu-

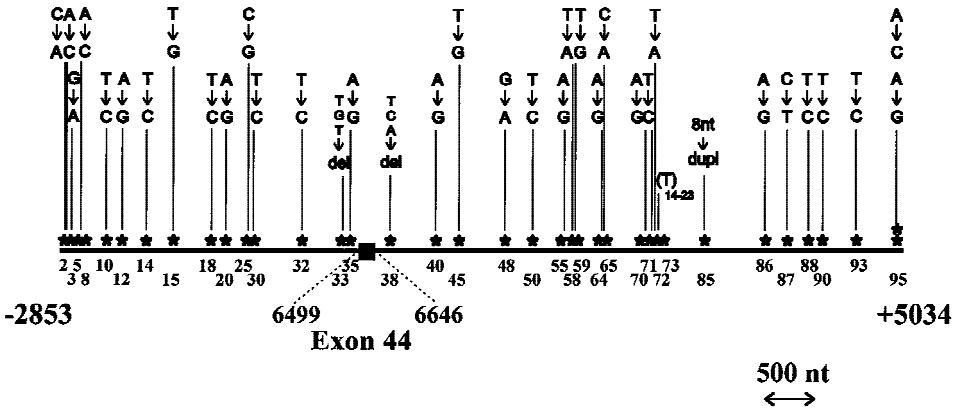

Fig. 1. Polymorphisms in the *dys44* DNA sequence. The sequence analyzed spanned exon 44 of the dystrophin gene (cDNA positions 6499 through 6646; illustrated by the *black box*), and its flanking introns between positions −2853 to −1 upstream and positions 1 to 5034 downstream. The *numbers beneath the horizontal line* denote arbitrary names given to polymorphic sites found in this study. Transversions are shown in the *upper row* and transitions in the *middle row* above the horizontal line; the change from an ancestral to a new allele is indicated by an *arrow*. Note that site 95 has three alleles, while nine alleles were found in the $(T)_n$ microsatellite (site ''73'').

Table 1. Population parameters for the world, Africans, and non-Africans[a]

	World	Africans	Non-Africans
n_{SSCP}	250	115	135
S	35	33	22
$S/\sum_{i=1}^{n-1}(1/i)$	5.74	6.21	4.02
n_{ASO}	860	277	583
H_n (±SE)	0.00101 (±0.00023)	0.00105 (±0.00023)	0.00090 (±0.00022)
H_nL	7.698	7.993	6.860
D	0.962	0.876	2.001
N_e (95% CI)	11,200 (6100–16,200)	11,650 (9100–14,200)	10,000 (5100–15,000)
F_{ST} (%)	0.147	0.072	0.158
H_m	0.587	0.792	0.280

[a] n_{SSCP}—number of chromosomes analyzed by SSCP, i.e., used to ascertain the polymorphisms; S—number of segregating sites (substitutions and small insertion/deletion polymorphisms); n_{ASO}—number of chromosomes analyzed by ASO hybridization, i.e., used to assess geographic distribution and to determine allele frequencies; $L = 7622$ bp.

man dystrophin gene on Xp21 (Zietkiewicz et al. 1997). As illustrated in Fig. 1, one represented a nine-allelic $(T)_{14–23}$ microsatellite. Others, due to substitutions and small insertions/deletions (referred to as point mutation polymorphisms), were biallelic except for site ''95,'' which had three alleles. These polymorphisms were ascertained in a representative sample of 250 worldwide chromosomes (on average, 20 chromosomes from each of the 13 studied populations) to avoid European or any regional bias (Zietkiewicz et al. 1997). Population frequencies of the *dys44* polymorphisms were investigated in an extended sample of 860 chromosomes by ASO hybridization or by denaturing gel electrophoresis.

Because recombination played a significant role in the evolution of the *dys44* segment (work in progress), the available tools of coalescence analysis could not be applied to analyze the age of the constituting polymorphisms (Hudson 1990). We therefore used the approach of Kimura and Ohta (1973) and estimated the so-called ''age before fixation'' (Maruyama 1974) for the underlying mutations. For this we had to distinguish the new allele from the ancestral one and to determine the frequency of the former. To provide the absolute time estimate for the polymorphisms considered [$\bar{a}(p)$ estimates are in N_e generation units], we had to estimate the effective population size N_e from the nucleotide diversity H_n and the mutation rate ν.

Since nuclear genomes of *H. sapiens* and great apes diverged relatively little by point mutations, the ancestral state was inferred by comparison with the orthologous great ape DNA positions (Mountain and Cavalli-Sforza 1994). In the case of point mutation polymorphisms, the allele shared with the chimpanzee, gorilla, and/or orangutan was considered ancestral (see results reported in Fig. 1). Detailed comparative analysis of a 2637-bp fragment of human *dys44* with the orthologous sequences from great apes revealed 46, 49, 55, and 75 differences from two chimpanzee subspecies, gorilla, and orangutan, respectively. The mutation rate of 1.5×10^{-9}/nucleotide/year, similar to values reported elsewhere (Goodman

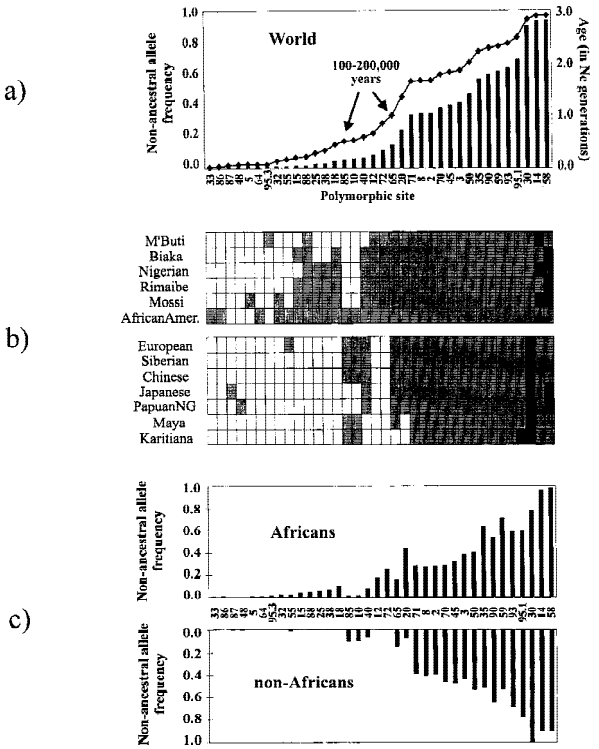

Fig. 2. World distribution of *dys44* point mutation polymorphisms. **a** Histogram of new allele frequencies (*bars*), ordered from the least to the most frequent, and their estimated age ''before fixation'' (*filled diamonds* connected by a line); the scale on the *right* corresponds to a time interval of ~600 ky, equivalent to $3N_e$ generations. **b** Occurrence of a new allele among the studied populations: presence (*gray boxes*), absence (*white boxes*), and fixation (*black boxes*). **c** Frequencies of new alleles in Africans and in non-Africans.

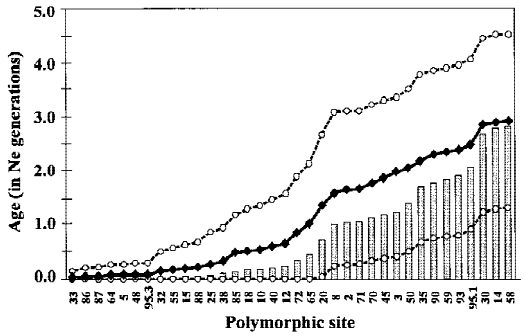

Fig. 3. The values of the age before fixation (*filled diamonds*) plus and minus one standard deviation (*open circles*). Polymorphic sites are ordered as in Fig. 2, with relative frequencies of the new allele shown in the background (*bars*).

1985; Britten 1986; Bailey et al. 1991), was obtained from the human–chimpanzee (an average), human–gorilla, and human–orangutan genetic distances, assuming a divergence of 5, 7, and 12 million years, respectively. The nucleotide diversity H_n of 0.001 (Table 1) was consistent with other studies (Li and Sadler 1991; Zietkiewicz et al. 1992; Fullerton et al. 1994; Harding et al. 1997; Hey 1997) and indicated that, on average, 1 site per 1000 would differ between two randomly chosen X chromosomes. The corresponding long-term effective size N_e of the human population was estimated as 11,200 (95% CI: 6100–16,200), a value consistent with those obtained from other molecular data (Nei and Graur 1984; Li and Sadler 1991; Rogers and Harpending 1992; Takahata 1993; Horai et al. 1995; Rogers 1995). Tajima's (1989) *D* parameter, which compares two measures of the genetic variation at the DNA level (one being a function of the number of segregating sites *S* and the other of the nucleotide diversity H_n), calculated for *dys44* was positive but did not differ significantly from 0 (Table 1). Inclusion of the triallelic site ''95,'' which did not obey the infinite-site model and formally should be disregarded, did not affect this analysis. Our result of Tajima's test, similar to the earlier observations for other nuclear loci (Harding et al. 1997; Hey 1997), is consistent with the neutral evolution of the *dys44* segment.

Figure 2a presents a histogram of the world frequencies of new (nonancestral) alleles ordered from the least to the most frequent (left scale) and the corresponding estimates of the age before fixation (right scale). The age values fell within a time frame of $3 \times N_e$ generations. For N_e of ~10,000, as estimated from our data, and a generation time of ~20 years, this corresponds to about 600 ky. Because of their great variance (Kimura and Ohta 1973; Kimura 1983), these estimates should be taken with caution (note also that they are for the population and not for the sample frequencies). Their uncertainty is illustrated in Fig. 3, showing the corresponding values plus and minus one standard deviation. On the other hand, this approach appears to be more robust than expected when applied to a number of polymorphisms originating from a short DNA segment. Intuitively, when the polymorphisms considered are in allelic association and their frequency changes are correlated, the magnitude of the variance within the analyzed segment should be reduced because the underlying gene trees are either shared or, to a large extent, overlapping (Pluzhnikov and Donnelly 1996). Figure 4 presents the analysis of recently published polymorphisms from a nonrecombinogenic segment of the β-globin locus (Harding et al. 1997). The age before fixation is compared here with the age of mutations estimated by the coalescence approach (Griffiths and Tavaré 1994) as reported by the authors. For most positions the relative ages of the contributing polymorphisms are very similar and the absolute differences between the two age estimates are smaller than could be expected given the great variance of the age before fixation. We have to remember, however, that both models assume constancy of the population size throughout the relevant evolutionary period and thus should be considered with caution.

The distribution of the *dys44* polymorphisms among the populations sampled was not uniform (Fig. 2b). Since the number of segregating sites was significantly higher in Africans than in non-Africans, we considered these two groups separately (see Fig. 2c and Table 1). Nine-

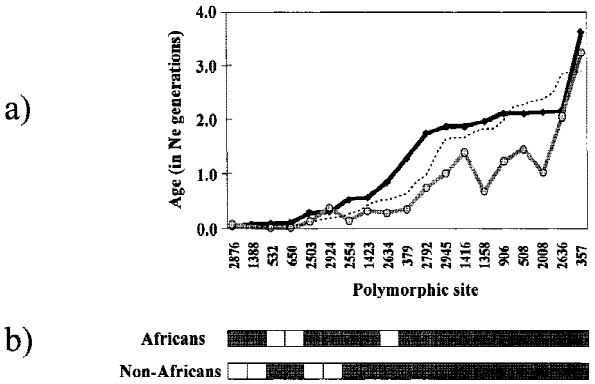

Fig. 4. Timing of the point mutation polymorphisms from the 2.67-kb segment of the β-globin gene (Harding et al. 1997). **a** The estimated age ''before fixation'' of new alleles (*filled diamonds* connected by a solid line) and the mutation age obtained by Harding et al. (1997) from the coalescence analysis (*filled* circles connected by a solid line), assuming that $4N_e$ generations ~800 ky; polymorphisms are ordered as a function of new allele frequency, from the least to the most frequent. The *dotted line* presents ''the age profile'' of *dys44* polymorphisms from Fig. 2a. **b** Occurrence of a new allele among Africans and non-Africans: presence (*gray boxes*) and absence (*white boxes*).

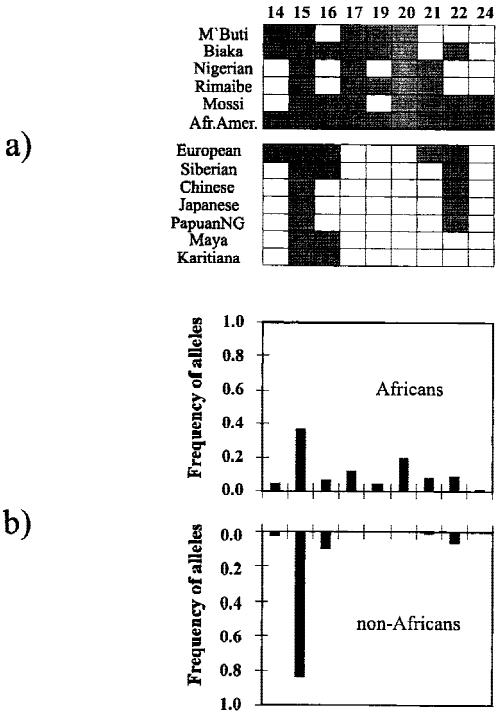

Fig. 5. World distribution of the $(T)_n$ alleles. **a** The presence (*gray boxes*) and absence (*white boxes*) of the alleles among the studied populations; alleles are ordered from the shortest ($n = 14$) to the longest ($n = 23$). **b** Allele frequencies in Africans and in non-Africans.

teen of the 36 new alleles were found both in Africans and in nonAfricans, and 17 of these had a frequency ⩾0.15, corresponding to an estimated age of 200 ky or more (Fig. 2a). At 13 segregating sites found only among Africans, the world frequency of a new allele was ⩽0.1, corresponding to an age of less than 150 ky. In contrast, only two such polymorphisms were observed in non-Africans. The new alleles at sites ''10'' and ''85,'' considered earlier (Zietkiewicz et al. 1997) as the European admixture among African Americans (Chakraborty et al. 1992), were recently found on two different chromosomes from sub-Saharan Africa (not shown), thus confirming their dispersion among continents (Fig. 2b). Interestingly, of the six polymorphisms in the intermediate range (new allele frequency 0.05–0.15; i.e., estimated age of 100–200 ky), two (''12'' and ''72'') were restricted to Africans, two (''10'' and ''85'') were found mostly in non-Africans, and two (''65'' and ''40'') were shared among the continents. This patchy distribution suggests that these sites could have originated in the time period when the ancestral population of the anatomically modern humans started to diverge. With the exception of site ''30,'' where a new allele appeared fixed among seven non-African populations, the relative shortage of rare alleles outside Africa was related to the deficiency of new polymorphisms, rather than to the loss of preexisting ones. The data at the β-globin locus (Harding et al. 1997) show a similar pattern (Fig. 4). New alleles representing 12 variants common to Africans and non-Africans belong to the old category (the age before fixation is greater than N_e generations for 10 variants). At four sites the new allele is restricted to Africa and at three it is found only outside; here, again, the age of new alleles is less than N_e generations (Fig. 4).

The shortage of new alleles outside Africa was more striking at the *dys44* $(T)_n$ site (Fig. 5), although here the ancestral state could be only tentatively associated with the most frequent $(T)_{15}$ variant. The phylogenetic comparison was not informative here: the orthologous sites in chimpanzee [12 chromosomes, $(T)_{10}$] and gorilla [3 chromosomes, $(T)_{11}$] were monomorphic, while orangutan (6 chromosomes) had two alleles, $(T)_{10}$ and $(T)_{12}$.

Discussion

A significant portion of the polymorphic sites ascertained in this study has an estimated age of 200 ky or more (Fig. 2a). Their ubiquity among globally distributed human groups (Fig. 2b) indicates that these sites represent a record of the nucleotide diversity of the ancestral population. Frequency profiles of these polymorphisms do not differ between Africans and non-Africans (Fig. 2c), which suggests that a profile characterizing the population before the divergence was similar to that observed today and that the ancient human population was relatively homogeneos until 100–200 ky ago. Polymorphisms estimated to be younger are restricted to Africans or non-Africans, indicating that after this date these groups must have diverged and subsequently stayed separated. This date is in good agreement with the age of paleontological findings of modern human fossils in Af-

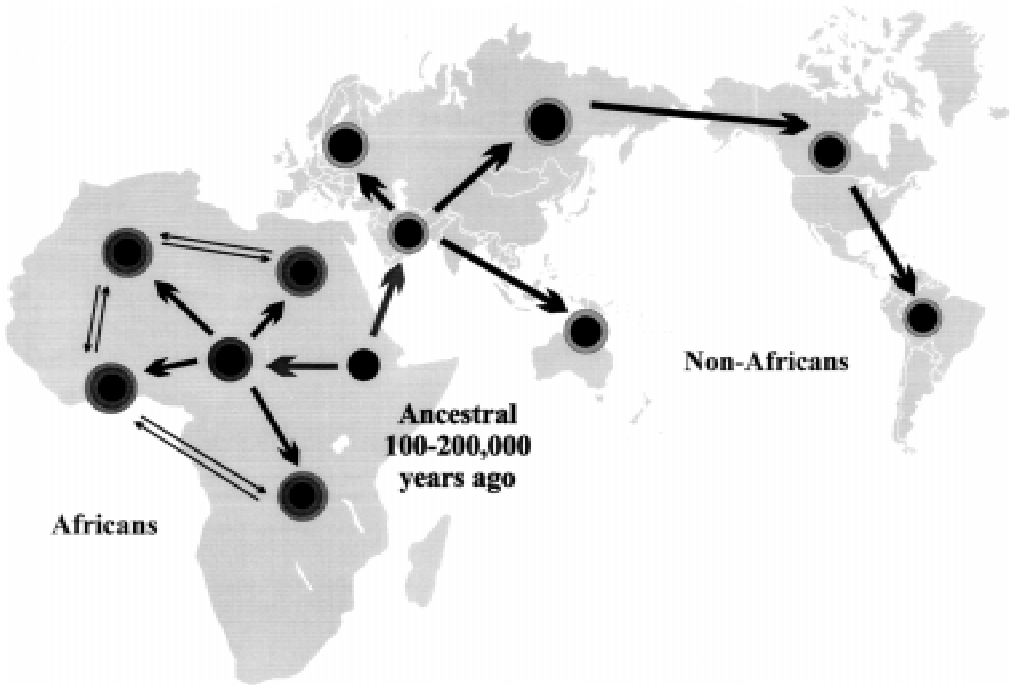

Fig. 6. Model explaining the patterns of *dys44* polymorphisms among human populations. *Black circles* represent globally distributed polymorphisms of the ancestral population (N_e ~10,000 individuals); outer *gray rings* indicate younger polymorphisms, particular for either Africans or non-Africans, that were acquired after divergence of these groups. *Thick arrows* represent geographic expansions of modern humans; *thin arrows* symbolize gene flow between African populations.

rica and the Middle East (Stringer and Andrews 1988; Lahr and Foley 1994; Leakey 1994; Tattersall 1997).

Thus, the N_e of about 10,000 estimated from the nucleotide diversity appears to reflect the size of the ancestral population before its diversification and geographical expansion. Since the ubiquitous polymorphisms contributed 90% to the nucleotide diversity (H_n of 0.0009 was obtained when the sites restricted to Africans and non-Africans were disregarded), the effective population size of ~10,000 was also obtained when either Africans or non-Africans were considered as being representative of the world population (Table 1). Similar N_e values could also be computed from individual nucleotide diversities of the 13 analyzed populations [median $H_n = 0.00084$ (Zietkiewicz et al. 1997)]. Geographic isolation favors speciation, and it could have played such a role in the emergence of *H. sapiens.* The number of 10,000 breeding individuals might thus reflect a speciation bottleneck rather than a long-term size of the ancestral population throughout the Middle Pleistocene [for discussion on the meaning of the effective population size, see Harding (1996)].

Consistent with a number of molecular studies in which the highest genomic variability and/or the deepest phylogenetic branching were seen among Africans (Cann et al. 1987; Scozzari et al. 1988; Ruvolo et al. 1993; Bowcock et al. 1994; Cavalli-Sforza et al. 1994; Mountain and Cavalli-Sforza 1994; Horai et al. 1995; Jorde et al. 1995; Batzer et al. 1996; Knight et al. 1996; Tishkoff et al. 1996; Nei and Takezaki 1996; Perez-Lazaun et al. 1997; Shriver et al. 1997), the highest level of within-population variation at the *dys44* locus was seen in sub-Saharan Africans. This effect was much more pronounced in the number of segregating sites (S) than in H_n (Table 1) and resulted from the accumulation of new polymorphisms in Africa following the divergence of modern humans. However, provided that this observation is not a particularity of *dys44* or merely the result of sampling, it is the paucity of young polymorphisms outside Africa, rather than the relative excess of African polymorphisms, which requires explanation. The greater African diversity is consistent with a scenario in which the ancestral population remained in Africa and non-Africans migrated outside (Tishkoff et al. 1996). Nevertheless, the older age of Africans does not have to be implied to explain the results, since these might reflect the size differences of the diverging *Homo sapiens* populations (Relethford and Harpending 1995; Relethford 1995; Rogers and Jorde 1995), whereby the greater number of young polymorphisms among Africans could be attributed to their greater population size following the divergence from the ancestral stock. At the same time, a higher population turnover outside Africa could account for the differences in the African and non-African diversities observed (Takahata 1994).

The model proposed in Fig. 6 summarizes our conclusions. An ancient population of about 10,000 breeding individuals diverged 100–200 ky ago. An African or Middle Eastern origin of the ancestral human population could not be inferred from our data; placing here the

ancestral population in Africa follows the paleontological evidence (Stringer and Andrews 1988; Lahr and Foley 1994; Leakey 1994; Tattersall 1997). The African population subsequently divided into subpopulations that remained in contact, allowing a certain amount of gene flow, as indicated by the lowest F_{ST} of 0.072 (Table 1). In contrast, non-African populations became isolated following their geographical expansion. This rendered the exchange of genetic material less probable, as reflected both by the higher F_{ST} (0.158) and by the smaller number of rare new polymorphisms shared within this group. While the F_{ST} of 0.147 characterizing the world composed of 13 populations (Zietkiewicz et al. 1997) was similar to estimates reported earlier (Relethford 1995), the F_{ST} obtained assuming the world to be composed of Africans and non-Africans was only 0.036. This indicates that, as expected given the major contribution of the shared polymorphisms and their similar frequency profiles (Fig. 2), most of the global divergence was caused by the differences among the particular populations and not by those between Africans and non-Africans.

Our data from the *dys44* locus contradict the regional continuity hypothesis (Weidenreich 1943; Wolpoff and Caspari 1997). Considering the population size of several thousand individuals, it is difficult to imagine that a population so small could spread over the vast areas of different continents and still maintain unity as a species through gene flow. Furthermore, if the geographic barrier was sufficient to prevent genetic contact between the diversified populations during the Upper Pleistocene, there is no reason to believe that it was ''leaky'' at an earlier period. Finally if, as suggested by the multiregional hypothesis, the ancestral population was composed of distinct subpopulations remaining in contact by migration, the real number of all breeding individuals would be even less than the estimated 10,000 (Takahata 1993); this is too low a number to maintain the allelic diversity observed at the major histocompatibility (*MHC*) locus (Ayala et al. 1994; Takahata et al. 1995).

These arguments are further strengthened by the observation of the closest relatives of *H. sapiens*, great apes. Subspecies of orangutans and those of common chimpanzees and gorillas diverged more than a million years ago and maintained separate identities despite inhabiting relatively restricted geographic areas either in Asis or in subequatorial African (Morin et al. 1994; Garner and Ryder 1996; Zhi et al. 1996). The concept of a primate species, not very numerous, inhabiting diverse geographic areas and evolving for an equally long time without speciation is thus not likely. Applying it to the evolving *H. sapiens* would additionally require our genetic diversity to be in the same range as that of great apes, which is in disagreement with the data (Zhi et al. 1996; Burrows and Ryder 1997). Our results, conforming to earlier molecular studies and to the most recent ones (Krings et al. 1997), are also compatible with the recent paleontological evidence, which suggests that the genus *Homo*, during the Paleolithic until as recently as 30–50 ky ago (Swisher et al. 1996), coexisted as a number of species rather than as regional lineages representing the same species (Tattersall 1997).

A variety of molecular data is needed to characterize the ancient population of modern humans. While studies of nuclear loci will provide a body of new evidence, progress in theoretical work is also required. Models and methods need to be developed that would allow analysis of DNA segments taking into account not only mutations characterized by different rates but also recombinations (see Wiuf and Hein 1997).

Acknowledgments. Luc Desrosiers and Jean-François Bibeau participated in the development and the management of our database. We are grateful to Evelyne Heyer for calculating variances, to Daniel Sinnett for support and discussions, to Jody Hey for his comments, to Tomasz Haertlé for his interest and help, to Lyuda Ossipova for sharing Siberian DNA samples and to Robert Patenaude and Clément Lanthier for the ape samples. This study was supported by the Canadian Genome Analysis and Technology Program and the Medical Research Council of Canada.

References

Ayala F, Escalante A, O'hUigin C, Klein J (1994) Molecular genetics of speciation and human origins. Proc Natl Acad Sci USA 91:6787–6794

Bailey WJ, Fitch DH, Tagle DA, Czelusniak J, Slightom JL, Goodman M (1991) Molecular evolution of the psi eta-globin gene locus: gibbon phylogeny and the hominoid slowdown. Mol Biol Evol 8:155–184

Batzer MA, Arcot SS, Phinney JW, Alegria-Hartman M, Kass DH, Milligan SM, Kimpton C, et al (1996) Genetic variation of recent Alu insertions in human populations. J Mol Evol 42:22–29

Bettecken T, Aissani A, Muller CR, Bernardi G (1992) Compositional mapping of the human dystrophin-encoding gene. Gene 122:329–335

Bowcock AM, Kidd J, Mountain JL, Hebert JM, Carotenuto L, Kidd KK, Cavalli-Sforza LL (1991) Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. Proc Natl Acad Sci USA 88:839–843

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457

Britten RJ (1986) Rates of DNA sequence evolution differ between taxonomic groups. Science 231:1393–1398

Burrows W, Ryder OA (1997) Y-chromosome variation in great apes (letter). Nature 385:125–126

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31–36

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

Chakraborty R, Kamboh MI, Nwankwo M, Ferrell RE (1992) Caucasian genes in American Blacks: new data. Am J Hum Genet 50: 145–155

Den Dunnen JT, Grootscholten PM, Dauwerse JG, Walker AP, Monaco AP, Butler R, Anand R, et al. (1992) Reconstruction of the 2.4 Mb human DMD-gene by homologous YAC recombination. Hum Mol Genet 1:19–28

Felsenstein J (1993) PHYLIP (phylogeny inference package), Version

3.5p [Felsenstein J (1989) PHYLIP—phylogeny inference package (version 3.2)]. Cladistics 5:164–166 (Distributed by the author, Department of Genetics, University of Washington, Seattle)

Fullerton SM, Harding RM, Boyce AJ, Clegg JB (1994) Molecular and population genetic analysis of allelic sequence diversity at the human β-globin locus. Proc Natl Acad Sci USA 91:1805–1809

Gabunia L, Vekua A (1995) A Plio-Pleistocene hominid from Dmanisi, East Georgia, Caucasus. Nature 373:509–512

Garner KJ, Ryder OA (1996) Mitochondrial DNA diversity in gorillas. Mol Phylogenet Evol 6:39–48

Goodman M (1962) Immunochemistry of the primates and primate evolution. Ann NY Acad Sci 102:219–234

Goodman M (1985) Rates of molecular evolution: the hominoid slowdown. Bioassays 3:9–14

Griffiths RC, Tavaré S (1994) Ancestral inference in population genetics. Stat Sci 9:307–319

Happe RP, Rosenboom W, Pierik AJ, Albracht SPY, Bagley KA (1997) Y chromosome variation in great apes. Nature 385:125–126

Harding RM (1996) Using the coalescent to interpret gene trees. In; Boyce AJ, Mascie-Taylor CGN (eds) Molecular biology and human diversity. Cambridge University Press, Cambridge

Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, et al. (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet 60:772–289

Hartl DL, Clark AG (1989) Principles of population genetics. Sinauer Associates, Sunderland, MA

Hey J (1997) Mitochondrial and nuclear genes present conflicting portraits of human origins. Mol Biol Evol 14:166–172

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

Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N (1995) Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc Natl Acad Sci USA 92: 532–536

Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. Theor Popul Biol 23:183–201

Hudson RR (1990) Gene genealogies and the coalescent process. Oxf Surv Evol Biol 7:1–44

Hudson RR, Kaplan NL (1988) The coalescent process in models with selection and recombination. Genetics 120:831–840

Jorde LB, Bamshad MJ, Watkins WS, Zenger R, Fraley AE, Krakowiak PA, Carpenter KD, et al. (1995) Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. Am J Hum Genet 57:523–538

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kimura M, Ohta T (1973) The age of a neutral mutant persisting in a finite population. Genetics 75:199–212

Knight A, Batzer MA, Stoneking M, Tiwari HK, Scheer WD, Herrera RJ, Deininger PL (1996) DNA sequences of Alu elements indicate a recent replacement of the human autosomal genetic complement. Proc Natl Acad Sci USA 93:4360–4364

Koenig M, Beggs AH, Moyer M, Scherpf S, Heindrich K, Bettecken T, Meng G, Muller CR, Lindolf M, Kaariainen H, et al. (1989) Molecular basis for Duchenne versus Becker muscular dystrophy: correlation of severity with type of deletion. Am J Hum Genet 45: 498–506

Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, Paabo S (1997) Neandertal DNA sequences and the origin of modern humans. Cell 90:19–30

Labuda D, Zietkiewicz E, Labuda M (1997) The genetic clock and the age of the founder effect in growing populations: a lesson from French-Canadians and Ashkenazim. Am J Hum Genet 61:768–771

Lahr MM, Foley R (1994) Multiple dispersals and modern human origins. Evol Anthropol 3:48–60

Leakey R (1994) The origin of humankind. BasicBooks, A Division of Harper Collins, New York

Li WH, Sadler LA (1991) Low nucleotide diversity in man. Genetics 129:513–523

Maruyama T (1974) The age of an allele in a finite population. Genet Res Cambr 23:137–143

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res Commun 23:23–35

Morin PA, Moore JJ, Chakraborty R, Jin L, Goodall J, Woodruff DS (1994) Kin selection, social structure, gene flow, and the evolution of chimpanzees. Science 265:1193–1201

Mountain JL, Cavalli-Sforza LL (1994) Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. Proc Natl Acad Sci USA 91:6515–6519

Nagaraja R, MacMillan S, Kere J, Jones C, Griffin S, Schmatz M, Terrell J, et al. (1997) X chromosome map at 75-kb STS resolution, revealing extremes of recombination and GC content. Genome Res 7:210–222

Nei M, Graur D (1984) Extent of protein polymorphism and the neutral mutation theory. Evol Biol 17:73–118

Nei M, Roychoudhury AK (1974) Genetic relationship and evolution of human races. Evol Biol 14:1–59

Nei M, Takezaki N (1996) The root of the phylogenetic tree of human populations. Mol Biol Evol 13:170–177

Orita M, Iwahana H, Kanazawa H, Hayashi K, Sekiya T (1989) Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformational polymorphisms. Proc Natl Acad Sci USA 86:2766–2770

Perez-Lezaun A, Calafell F, Mateu E, Comas D, Ruiz-Pacheco R, Bertranpetit J (1997) Microsatellite variation and the differentiation of modern humans. Hum Genet 99:1–7

Pluzhnikov A, Donnelly P (1996) Optimal sequencing strategies for surveying molecular genetic diversity. Genomics 144:1247–1262

Relethford JH (1995) Genetics and modern human origins. Evol Anthropol 4:53–63

Relethford JH, Harpending HC (1995) Ancient differences in population size can mimic a recent African origin of modern humans. Curr Anthropol 36:667–673

Roberts RG, Coffey AJ, Bobrow M, Bentley DR (1993) Exon structure of the human dystrophin gene. Genomics 16:536–538

Rogers AR (1995) Genetic evidence for a Pleistocene population explosion. Evolution 49:608–615

Rogers AR, Harpending HC (1992) Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol 9: 552–569

Rogers AR, Jorde LB (1995) Genetic evidence on modern human origins. Hum Biol 67:1–36

Rogers AR, Jorde LB (1996) Ascertainment bias in estimates of average heterozygosity. Am J Hum Genet 58:1033–1041

Ruvolo ME, Zehr S, von Dornum M, Pan D, Chang B, Lin J (1993) Mitochondrial COII sequences and modern human origins. Mol Biol Evol 10:1115–1135

Sarich VM, Wilson AC (1967) Immunological time scale for hominid evolution. Science 158:1200–1203

Scozzari R, Torroni A, Semino O, Sirugo G, Brega A, Santachiara-Benerecetti AS (1988) Genetic studies on the Senegal population and mitochondrial DNA polymorphisms. Am J Hum Genet 43: 534–544

Shriver MD, Jin L, Ferrell RE, Deka R (1997) Microsatellite data support an early population expansion in Africa. Genome Res 7: 586–591

Simard L, Gingras F, Delvoye N, Vanasse M, Melancon SB, Labuda D (1992) Deletions in the dystrophin locus: analysis of Duchenne and Becker muscular dystrophy patients in Quebec. Hum Genet 89: 419–424

Stringer CB, Andrews P (1988) Genetic and fossil evidence for the origin of modern humans. Science 239:1263–1268

Swisher CC III, Curtin GH, Jacoby T, Getty AG, Suprijo A, Widiasmoro (1994) Age of the earliest known hominids in Java, Indonesia. Science 263:1118–1121

Swisher CC III, Rink WJ, Anton SC, Schwarcz HP, Curtis GH, Suprijo A, Widiasmoro (1996) Latest *Homo erectus* of Java: potential contemporary with *Homo sapiens* in Southeast Asia. Science 274: 1870–1874

Tajima F (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

Takahata N (1994) Repeated failures that led to the eventual success in human evolution. Mol Biol Evol 11:803–805

Takahata N, Satta Y, Klein J (1995) Divergence time and population size in the lineage leading to modern humans. Theor Popul Biol 48:198–221

Tattersall I (1997) Out of Africa again . . . and again? Sci Am 276:60–67

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonné-Tamir B, et al. (1996) Global patterns of linage disequilibrium at the CD4 locus and modern human origins. Science 271: 1380–1387

Weidenreich F (1943) The skull of *Sinanthropus pekinensis:* a comparative study of a primitive hominid skull. Palaeontol Sinica n.s. D10 (whole series): 127

Wiuf C, Hein J (1997) On a number of ancestors to a DNA sequence. Genetics 147:1459–1468

Wolpoff M, Caspari R (1997) Race and human evolution. Simon and Schuster, New York

Zhi L, Karesh WB, Janczewski DN, Frazier-Taylor H, Sajuthi D, Gombek F, Andau M, et al. (1996) Genomic differentiation among natural populations of orang-utan (*Pongo pygmaeus*). Curr Biol 6:1325–1336

Zietkiewicz E, Sinnett D, Richer C, Mitchell G, Vanasse M, Labuda D (1992) Single strand conformational polymorphisms (SSCP): detection of useful polymorphisms at the dystrophin locus. Hum Genet 89:453–456

Zietkiewicz E, Yotova V, Jarnik M, Korab-Laskowska M, Kidd KK, Modiano D, Scozzari R, Stoneking M, Tishkoff S, Batzer M, Labuda D (1997) Nuclear DNA diversity in worldwide distributed human populations. Gene 205:161–171