Archaic African *and* Asian Lineages in the Genetic Ancestry of Modern Humans

Rosalind M. Harding,[1] S. M. Fullerton,[1,*] R. C. Griffiths,[2] Jacquelyn Bond,[1,**] Martin J. Cox,[1,***] Julie A. Schneider,[1] Danielle S. Moulin,[1] and J. B. Clegg[1]

[1]MRC Molecular Haematology Unit, Institute of Molecular Medicine, John Radcliffe Hospital, Oxford; and [2]Department of Mathematics, Monash University, Clayton, Australia

Summary

A 3-kb region encompassing the β-globin gene has been analyzed for allelic sequence polymorphism in nine populations from Africa, Asia, and Europe. A unique gene tree was constructed from 326 sequences of 349 in the total sample. New maximum-likelihood methods for analyzing gene trees on the basis of coalescence theory have been used. The most recent common ancestor of the β-globin gene tree is a sequence found only in Africa and estimated to have arisen ~800,000 years ago. There is no evidence for an exponential expansion out of a bottlenecked founding population, and an effective population size of ~10,000 has been maintained. Modest differences in levels of β-globin diversity between Africa and Asia are better explained by greater African effective population size than by greater time depth. There may have been a reduction of Asian effective population size in recent evolutionary history. Characteristically Asian ancestry is estimated to be older than 200,000 years, suggesting that the ancestral hominid population at this time was widely dispersed across Africa and Asia. Patterns of β-globin diversity suggest extensive worldwide late Pleistocene gene flow and are not easily reconciled with a unidirectional migration out of Africa 100,000 years ago and total replacement of archaic populations in Asia.

Introduction

Many types of DNA polymorphism have been used to examine human evolutionary history, and phylogenetic methods provide one way of analyzing these data (Brookfield 1994). Alternatively, population genetic analyses for inferences on coalescent time depth (Ruvolo et al. 1993) and on demographic history (Harpending et al. 1993; Sherry et al. 1994) have become popular, with the accumulation of sequence data, initially from the highly variable mtDNA D-loop (Vigilant et al. 1991). Coalescence times have also been estimated for Y chromosome sequences, which, like mtDNA, are haploid (Dorit et al. 1995; Hammer 1995). The assembly of sequence data from the diploid nuclear genome for population genetic studies is a more formidable task. Here, we present and analyze the first extensive survey of allelic sequence variation for a single copy gene at a nuclear autosomal locus, a 3-kb region on chromosome 11 encompassing the β-globin gene, introns, and flanking sequences. More than 1 Mb of DNA have been sequenced to determine 349 alleles, sampled from nine populations in Africa, Asia, and Europe.

The β-globin locus was chosen for these molecular and population genetic analyses to take advantage of the great deal that is already known about its structure and variability. These studies have been motivated by functional variation at the β-globin locus, causing a variety of hemoglobinopathies, including HbS, HbC, HbE, and the thalassemias (Weatherall and Clegg 1981; Orkin and Kazazian 1984). Frequencies of these alleles are elevated where falciparum malaria is endemic, because of the selective advantage conferred on heterozygous genotypes (Haldane 1948; Allison 1954). Geographic distributions of these polymorphisms and of their background haplotypes for the 30-kb β-globin complex suggests that malarial selection is evolutionarily recent, estimated at <10,000 years by Flint et al. (1993).

The primary difficulty in determining sequence alleles at autosomal loci is diploidy. To overcome this problem, a technique of allele-specific amplification was developed for a study of β-globin polymorphism in the Melanesian population of Vanuatu by Fullerton et al. (1994). They reported an average age of sequence divergence of 450,000 years, consistent with a total coalescence time of 900,000 years. This estimate is approximately fourfold greater than the time depth for mtDNA,

Received July 15, 1996; accepted for publication January 10, 1997.

Address for correspondence and reprints: Dr. Rosalind M. Harding, MRC Molecular Haematology Unit, Institute of Molecular Medicine, John Radcliffe Hospital, Headington, Oxford, OX3 9DU, United Kingdom. E-mail: rharding@vax.ox.ac.uk

*Present affiliation: Department of Anthropology, University of Durham, Durham, United Kingdom.

**Present affiliation: Department of Biology and Biochemistry, Brunel University, Uxbridge.

***Present affiliation: Ecosystems Analysis and Management Group, Department of Biological Sciences, University of Warwick, Coventry.

0002-9297/97/6004-0006$02.00

in line with expectations for an autosomal locus assuming neutrality (Takahata 1995). A recombination hot spot between the δ-globin and β-globin genes (Antonarakis et al. 1982) was found to elevate haplotype diversity substantially in the 5′ flanking region of the β-globin gene but to have no significant effect on numbers of haplotypes in either the gene or 3′ flanking sequence. These findings encouraged us to consider using methods developed in application to the mtDNA genome for estimating the time to the most recent common ancestor (TMRCA) of β-globin diversity, assuming neutrality and an infinitely-many-sites mutation model (Griffiths and Tavaré 1994*a*). Our expectation of a greater coalescence time for diversity from β-globin compared with that from mtDNA suggested that a detailed analysis of β-globin would permit a more informative study of the ancestors of modern humans in the middle to late Pleistocene period.

Material and Methods

We examined nucleotide and haplotype polymorphism in the same 3-kb region encompassing the β-globin gene as studied by Fullerton et al. (1994), for an additional eight populations. Sequence haplotypes from 349 chromosomes are presented in the appendix. There are four groups of sequence haplotypes identified as A, B, C and D, and related to the RFLP frameworks, I, II, III-Asian, and III, respectively, defined by Orkin et al. (1982). After sequencing the diploid genotype, linkage relationships in all compound heterozygotes represented in table A1A were determined by sequencing individual alleles specifically amplified by the amplification refractory mutation system (Newton et al. 1989). Because only small quantities of DNA were available for the samples in table A1B, haplotypes were inferred using linkage information from table A1 after diploid sequencing.

We included 61 chromosomes from the Melanesian inhabitants of the islands of Espiritu Santo and Maewo in Vanuatu (VAN) from the study by Fullerton et al. (1994). A second Melanesian population is represented by 24 chromosomes from the Southern Highlands of Papua New Guinea (PNG). A Southeast Asian sample of 41 chromosomes is from Palembang, Sumatra, Indonesia (SUM). All of these DNA samples were collected during the course of ongoing surveys for α- and β-thalassemia. The 31 Gambian (GAM) chromosomes derive from immunological studies of host resistance to malaria. Forty-six chromosomes from Oxfordshire, England (UK), were obtained from DNAs collected for a study of blood pressure and hypertension. Forty-eight chromosomes from the Nuu-Chah-Nulth Amerindians of the Pacific Northwest of the United States (NCN) are from a set previously analyzed for mtDNA polymorphisms (Ward et al. 1991). Data for 96 chromosomes from another three populations were included to add detail to an emerging pattern of haplotype distribution that contrasts Asia with Africa. There are 24 from Mongolia (M), 24 from the Biaka Pygmies of the Central African Republic (CAR), and 48 from the Luo of Kenya (KEN). The Luo moved into the Nyanza region of Kenya within the last 200 years, having migrated down the Nile River valley out of an area in what is now Sudan ~3000 years ago (Newman 1995). The African samples used in this study represent three major sub-Saharan areas, west (GAM), central (CAR), and northeast (KEN).

New methods using a coalescent model for the ancestral history of a sample of genes (Kingman 1982*a*, 1982*b*, 1982*c*) were used to infer the time scale of the origin and evolution of polymorphic variation in a 2.67-kb region encompassing β-globin. In this model, unique mutations are supposed to occur along ancestral lineages by a Poisson process of rate $\theta/2$, where $\theta = 4N_e\mu$, μ is the mutation rate per sequence per generation, and N_e is the effective population size. The genealogy is taken from a Wright-Fisher population under equilibrium, with a time scaling given by the usual formulation of N_e as a constant value (Kimura 1983). Coalescent models do, however, permit other demographies to be investigated, and time scaling does depend on the assumed demography. Clear and accessible discussions of this theoretical background have been presented by Donnelly and Tavaré (1995) and Donnelly (1996).

It is not difficult to simulate the coalescent process to show the time scaling associated with a given value of the parameter θ for a variety of demographic and mutation assumptions. For a restricted set of assumptions, the TMRCA can be estimated using an estimate of θ together with the maximum pairwise sequence difference in a sample of data (Tajima 1983). This approach was taken for analyses of mtDNA by Ruvolo et al. (1993), Y-chromosomal DNA by Hammer (1995), and autosomal nuclear gene sequences by Takahata et al. (1995). The reason for using the difference between a pair of sequences is that mathematical formulation becomes complicated when the number of sequences sampled is more than two (Takahata 1995). The distinction of the coalescence methods used in this study is that they are formulated for the full sample available, presented as a gene tree, preserving information from all relationships among the sequences and not just the pairwise relationships. Simulating a coalescent process complete with time information, conditional on a specified gene tree with a given θ (Griffiths and Tavaré 1994*a*, 1994*b*), is intuitively comparable to, but more mathematically complex than, simulating the process alone. One aim of this simulation approach is to describe the likelihood surface for θ, from which a maximum-likelihood estimate can be determined.

The full information in a sample of sequences can be equivalently represented by a unique gene tree describing the mutation history of the sequences, provided that all segregating sites in the sample have arisen from single point mutations. If the ancestral type at each site is known, the tree is rooted. Construction of the gene tree from the sequences is a perfect phylogeny problem using the types of bases at segregating sites as characters in a phylogenetic sense. Construction algorithms are given in Gusfield (1991) and Griffiths (1987). Because there is a unique tree, these algorithms follow an exact graph-theoretic procedure, not a statistical one. It is straightforward to tell whether a sample of sequences could have been obtained from single point mutations, by checking whether a consistency condition based on the mutation pattern at pairs of sites is satisfied. To estimate θ and compute the TMRCA, a fully stochastic coalescent model is required. The infinitely-many-sites model of mutation is assumed, requiring that all point mutations have occurred at sites that have never previously segregated in the population.

A maximum-likelihood estimate of θ using the full information in the sequence data set, or, equivalently, the gene tree, was found using a computational method proposed by Griffiths and Tavaré (1994*a,* 1994*b*) and implemented in a computer program, *ptreesim,* written by R. C. Griffiths. We made estimates of θ for both a constant population-size model and a model where there is exponential population growth (Griffiths and Tavaré 1994*b*). A program, *timesim,* also written by R. C. Griffiths, gave estimates of the TMRCA of samples of sequences, scaled in units of N_e generations, by use of the maximum-likelihood estimate of θ as a parameter, conditional on the gene tree and θ (Griffiths and Tavaré 1994*a,* 1994*b*). Estimates for the ages of mutations were also made using *timesim.*

The computational technique used by *timesim* is to represent the likelihood of a gene tree as the expected value of a functional on a stochastic process that has ancestor gene trees as its state space. If there are r runs in a simulation generating (likelihood,time) pairs (l_1,t_1) . . . (l_r,t_r), then an estimate of the expected time to the root, conditional on the data is

$$\sum_{i=1}^{r} l_i t_i \Big/ \sum_{i=1}^{r} l_i \, .$$

The expected ages of the mutations and the expected TMRCA were computed at the same time in comparable procedures. Estimates of expected ages are weighted averages, with respect to simulated likelihoods, of the ages observed in the simulation runs.

All estimates of θ, TMRCA, and ages of mutations were made on individual population samples and the combined data set, on the assumption of random mating and random sampling. Analysis of a world data set is appropriate because we have assumed that each mutation has occurred once only in the history of the modern human population. However, because population structure may lead to underestimation of coalescence times from the world data set, the estimates from individual populations are given primary consideration.

In addition to the coalescent analyses, a number of statistics were computed to provide measures of diversity within and between populations. The number of segregating sites for the sample size (s_n) and the average pairwise sequence difference ($\hat{k}$) provided alternative estimators of θ under the same assumptions as required to estimate θ from the gene tree. To flag departures from these assumptions, the standardized difference between s_n and $\hat{k}$ (D) as proposed by Tajima (1989*b*) was examined. Nucleotide diversity, π, was estimated from average pairwise sequence difference (Kimura 1983) over numbers of effectively silent sites (Kreitman 1983). Expected numbers of haplotypes were estimated on the assumption of an infinitely-many-alleles model of mutation (Ewens 1972) and using an estimate of θ based on s_n. Population relationships were examined in a minimum spanning tree imposed on a multidimensional scaling (Rohlf 1993) of a matrix of net sequence differences, $d_{ij} = k_{ij} - (k_i + k_j)/2$ for pairwise sequence comparisons between populations i and j (Nei and Li 1979). A nonparametric test suggested by Hudson et al. (1992) was used to check for significance of population structure. These tests were based on comparison of a weighted average of the sequence pairwise differences, k, within populations i and j as $K_s = (k_i.k_j)/2$, against a distribution for K_s generated from 1,000 random partitions of the $i + j$ sequences for each pair of populations.

Results

Linkage Patterns

For the full 3-kb region, there are larger numbers of haplotypes in each population than expected under the assumption of an infinitely-many-alleles mutation model (table 1), reflecting recombination (Strobeck and Morgan 1978; Hudson 1983). This effect is confined to the 5′ 330 bp of the 5′ flanking region containing the variable $(AT)_xT_y$ repeat sequence (Trabuchet et al. 1991). Downstream of site 330, there are no insertion and deletion polymorphisms and no excess haplotypes in individual populations. Of 349 sequences for this 2.67-kb region, only 23 recombinant haplotypes were found.

The frequency of recombination events decreases from the recombination hot spot along a gradient through the 5′ flanking region up to site 906 in the first exon. The few recombinant haplotypes in the 2.67-kb region were readily identified after large numbers of al-

Table 1

Diversity Statistics

Sequence and Statistic	Population									
	CAR	GAM	KEN	MON	NCN	SUM	PNG	VAN	UK	World
Full 3-kb sequences:										
No. of sequences	...	31	...	...	48	43	24	61	46	
s_n	...	6.51	...	...	4.06	4.39	4.28	4.27	3.87	
Expected no. of haplotypes | s_n	...	11.8	...	...	10.8	10.9	8.5	12.1	10.4	
SD	...	2.44	...	...	2.56	2.53	2.56	2.76	2.71	
Observed no. of haplotypes	...	23[a]	...	...	16[b]	20[a]	9 ns	17 ns	15 ns	
$\hat{k}$	...	6.26	...	...	5.62	6.68	5.97	6.03	5.52	
$\pi = \hat{k}/2{,}650$ bp	...	.24%	...	...	.21%	.25%	.23%	.23%	.21%	
Tajima's D[c]	...	−.134 ns	...	...	1.222 ns	1.689 ns	1.398 ns	1.269 ns	1.353 ns	
5′ 330-bp sequences:										
No. of sequences	...	31	...	...	48	43	24	61	46	
s_n	...	3	...	...	1.58	1.62	1.87	1.5	1.37	
Expected no. of haplotypes | s_n	...	7.8	...	...	6	5.9	5.4	6.1	5.4	
SD	...	2.12	...	...	1.96	1.93	1.76	2.02	1.86	
Observed no. of haplotypes	...	14[b]	...	...	12[a]	12[a]	8 ns	10 ns	10[b]	
$\hat{k}$	...	3.17	...	...	1.71	2.27	2.63	2.05	2.4	
$\pi = \hat{k}/330$ bp	...	.96%	...	...	.52%	.69%	.80%	.62%	.73%	
Tajima's D[c]	...	.17 ns	...	...	.232 ns	1.092 ns	1.254 ns	.952 ns	1.961 ns	
2.67-kb sequences:										
No. of sequences	24	31	48	24	48	43	24	61	46	349
s_n	1.61	3.5	3.15	3.48	2.48	2.77	2.41	2.78	2.5	2.95
Expected no. of haplotypes | s_n	5	8.5	9.3	7.7	8	8.3	6.3	9.2	7.9	14.7
SD	1.69	2.19	2.4	2.02	2.25	2.26	1.87	2.45	2.23	3.34
Observed no. of haplotypes	7 ns	11 ns	12 ns	8 ns	6 ns	8 ns	4 ns	10 ns	5 ns	30[d]
$\hat{k}$	2.11	3.1	2.58	4.4	3.91	4.41	3.34	3.97	3.09	4.2
$\pi = \hat{k}/2{,}320$ bp	.09%	.13%	.11%	.19%	.17%	.19%	.14%	.17%	.13%	.18%
Tajima's D[c]	.945 ns	−.386 ns	−.555 ns	.874 ns	1.690 ns	1.793 90%	1.265 ns	1.247 ns	.690 ns	1.062 ns
Tree-compatible 2.67-kb sequences:										
No. of sequences	22	28	42	22	46	39	24	57	46	326
s_n	1.65	3.34	3.25	3.57	2.5	2.37	2.41	2.82	2.5	2.83
Expected no. of haplotypes | s_n	4.9	8.0	9.1	7.5	7.9	7.3	6.3	9.1	7.9	14.0
SD	1.67	2.1	2.34	1.97	2.23	2.12	1.87	2.42	2.23	3.26
Observed no. of haplotypes	6 ns	8 ns	7 ns	7 ns	5 ns	5 ns	4 ns	7 ns	5 ns	16 ns
$\hat{k}$	1.95	2.87	2.32	4.45	3.96	4.35	3.34	3.89	3.09	4.14
$\pi = \hat{k}/2{,}320$ bp	.08%	.12%	.10%	.19%	.17%	.19%	.14%	.17%	.13%	.18%
Tajima's D[c]	.577 ns	−.467 ns	−.897 ns	.882 ns	1.724 90%	2.516 95%	1.265 ns	1.118 ns	.69 ns	1.158 ns

NOTE.—ns = not significant.

[a] Confidence limit for a significant departure of observed number of haplotypes from expected is ±3 SD.

[b] Confidence limit for a significant departure of observed number of haplotypes from expected is ±2 SD.

[c] D is the standardized difference between $\hat{k}$ and s_n; confidence limits are from a beta distribution given by Tajima (1989*b*).

[d] Confidence limit for a significant departure of observed number of haplotypes from expected is ±4 SD.

Table 2

Numbers of β-Globin Sequence Haplotypes in Each Population Sample

Haplotype[a]	No. of Haplotypes									
	CAR (n = 24)	GAM (n = 31)	KEN (n = 48)	MON (n = 24)	NCN (n = 48)	SUM (n = 43)	PNG (n = 24)	VAN (n = 61)	UK (n = 46)	World (n = 349)
A1	9	8	12	3	15	8	1	25	23	104
A2	...	...	...	...	...	...	...	1	...	1
A3	1	2	5	...	...	...	...	...	...	8
A4	...	...	...	1	...	...	...	...	...	1
B1	6	5	9	3	2	10	12	16	16	79
B2	4	6	8	...	...	...	...	...	...	18
B3	1	2	6	...	...	...	...	...	...	9
B4	...	3	...	...	...	...	...	...	...	3
B9	1	...	...	...	...	...	...	...	1	2
B11	...	...	1	...	...	...	...	...	...	1
B5 (5A1[5′357]xB4 rcb.+1388[b])	...	1	...	...	...	...	...	...	...	1
B8 (6B2[5′357]xB1 rcb.)	...	1	2	...	...	...	...	...	...	3
B10 (?B1[5′357]xB4 rcb.)	...	...	1	...	...	...	...	...	...	1
B14 (?B2[5′357]xB9 rcb.)	2	...	...	...	...	...	...	...	...	2
B7 (1C1[5′379]xB1 rcb.)	...	...	...	...	...	2	...	2	...	4
B6 (2C3[5′508]xB1 rcb.)	...	...	...	...	...	...	...	1	...	1
B12 (?D1[5′508]xB2 rcb.)	...	...	1	...	...	...	...	...	...	1
B13 (?D1[5′906]xB1 rcb.)	...	...	1	...	...	...	...	...	...	1
C1	...	1	...	4	22	14	...	7	...	48
C2	...	...	...	2	6	1	...	...	...	9
C3	...	...	...	...	...	...	7	3	...	10
C7	...	...	...	6	...	6	4	1	2	19
C6 (2B2[5′508]xC1 rcb.)	...	1	1	...	...	...	...	...	...	2
C8 (?C1[5′379]xC7 rcb.)	...	...	...	2	...	...	...	...	...	2
C5 (?A1/?B1[5′508]xC1 rcb.)	...	...	...	...	2	...	...	...	...	2
C4 (A1/B1 906 conversion)	...	...	...	...	...	1	...	...	...	1
D1	...	...	1	3	1	...	...	4	4	13
D2	...	1	...	...	...	...	...	...	...	1
D3 (A1/B1 906 conversion)	...	...	...	...	...	1	...	1	...	2

[a] rcb. = recombinant.

[b] Additional mutation.

leles were sequenced, because they showed up as rare haplotypes characterized by a new 5′ linkage pattern among common polymorphisms. These haplotypes probably have been generated by interallelic recombination, resolving as either flanking exchange or gene conversion (table 2). After these 23 recombinants were omitted, numbers of haplotypes were found to lie within 2 SD of expected numbers for all samples, including the combined world sample (table 1).

Patterns of Haplotype and Nucleotide Diversity

Although several 2.67-kb β-globin haplotypes have global distributions, haplotype frequencies clearly distinguish African and Asian populations (fig. 1). Significant population structure was found between African and Asian populations but not within sub-Saharan Africa (CAR, GAM, KEN), within east Asia (MON, NCN, SUM), between Vanuatu and the United Kingdom, or between the United Kingdom and CAR Pygmies. Diversity in Asia comprises a variety of C haplotypes, together with A1, B1, and D1 haplotypes. In Africa, A1, B1, and a variety of exclusively African B haplotypes create diversity (table 2). The presence of B9 in the U.K. sample, together with the low frequency of C haplotypes (C7 only), contributes to the U.K. sample's closer affinity with Africa than with Asia.

The patterns of diversity summarized by average pairwise difference, $\hat{k}$; nucleotide diversity, π; numbers of segregating sites, s_n; and Tajima's D, reflect the different haplotype compositions in Africa and Asia described above. The presence in Asia of C diversity, in addition to A1 and B1 haplotypes, apparently enhances $\hat{k}$ relative to s_n, to the extent of a significant discordance in the Sumatran sample, as indicated by Tajima's D (table 1). Consequently, genetic diversity measured by $\hat{k}$ is greater in Asia than in Africa. Note that C1 haplotypes common in Asia are distinguished from A1 and B1 haplotypes by seven and eight mutational steps, respectively, and that the A1 and various B haplotypes in Africa are connected through steps of only one or two mutations. In Africa, there is little discordance between s_n and $\hat{k}$. Diversity measured by numbers of segregating sites in the 2.67-kb region indicates marginally higher diversity in Africa than in Asia for all partitions of the data, including the recombinogenic 5′ 330 bp region (table 1). This same pattern distinguishing Africa and Asia is clear among both the 3-kb and 2.67-kb haplotypes. Comparison of Tajima's D between the full set of 2.67-kb haplotypes and the subset of tree-compatible sequences (table 1) shows that no bias is introduced by omitting the recombinants for the gene tree analyses.

The major patterns of β-globin diversity are lack of significant structure within Africa and east Asia, but significant structure between Africa and Asia, deep pairwise sequence diversity in Asia, including a significant value of Tajima's D for Sumatra, and the opposite pattern in Africa, with increased diversity resulting from a larger number of segregating sites among a closely related set of haplotypes. The contrasting patterns between Africa and Asia cannot be explained by malarial selection, which maintains HbS in the three African populations sampled and HbE and β-thalassemia polymorphisms in Sumatra (appendix). Although Tajima's D is significant for Sumatra, malarial selection cannot explain why the pattern of diversity in Sumatra is more similar to that in Mongolia, where there is no malaria, than to that in Africa. A possible explanation for $\hat{k}$ larger than s_n in Asia and significantly so in Sumatra, is that N_e has not been constant and has been reduced from a larger value (Tajima 1989*a*). It is likely that different patterns of diversity between Africa and Asia have arisen as a result of different demographic histories and a level of gene flow between them that is restricted compared with expectations of random mating.

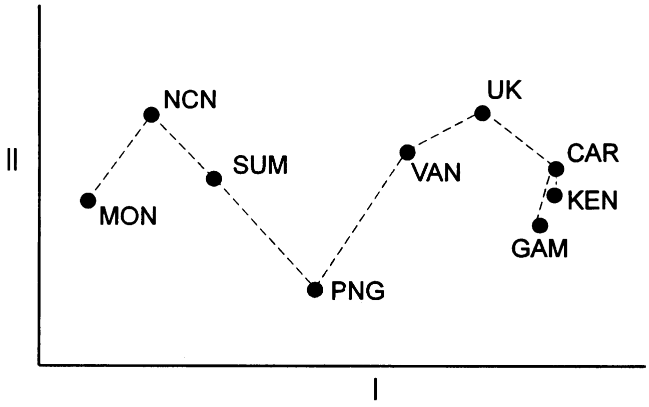

Figure 1 Nonmetric multidimensional scaling of net sequence differences, showing population relationships.

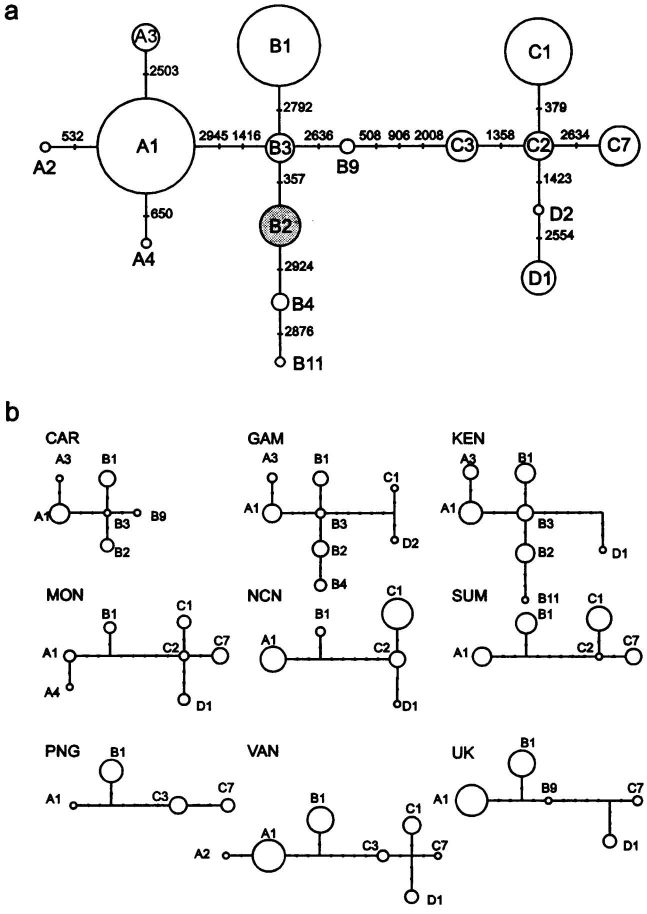

Figure 2 β-Globin gene trees showing relationships between haplotypes. Circle sizes are proportional to haplotype frequencies. *a,* Gene tree for 326 sequences from all populations. B2 (*shaded*) is the root. *b,* Gene trees for individual populations. Abbreviations for population names are given in full in the text.

Gene Tree Analyses

A unique unrooted tree compatible with 326 2.67-kb sequences (table 2) is presented in figure 2*a*. All segregating sites in this gene tree were assumed to have arisen from unique point mutations. The number of mutations was not sufficient to resolve the ordering of sites 2945 and 1416 and, likewise, 508, 906, and 2008. As a result of the worldwide distribution of A1, B1, C1, and D1 haplotypes, the main connectivity pattern shown in figure 2*a* is represented in trees for individual populations (fig. 2*b*).

Examination of β-globin sequences from other primates allows estimation of the root and mutation rate. The root of the tree was determined as either haplotype B2 or B3, by use of a chimpanzee sequence to define the ancestral nucleotides at each polymorphic site. Both B2 and B3 are found in all three African samples and are absent from all other samples. The chimpanzee sequence was not informative at the site homologous to 357, where there is a G and the human polymorphism is A/C. We determined the ancestral variant at site 357 to be C, by comparison with gorilla (and gibbon) sequences, indicating B3 as the root. (B2 is the maximum-likelihood root.) The nucleotide mutation rate was estimated as 1.34×10^{-9} per year from 31 sites that are monomorphic in humans but that vary between humans and chimpanzees, (1.34% divergence) on the assumption of a 5-million-year split. This is a low rate of mutation by comparison with other nuclear genes (Wolfe et al. 1989) and may explain why we found no evidence of recurrent mutation. A neutral mutation rate to new alleles of 6.2×10^{-5} per generation was determined, on the assumption of generations of 20 years and 2,320 effectively silent sites (Kreitman 1983) in the 2.67-kb β-globin region. This estimate of the number of effectively silent sites excludes the majority of nucleotides in the three exons and a small additional number at splicing sites in the introns and in regulatory sequences that are under functional constraint.

The gene tree shows that the D1 haplotype is derived through the C lineage. Although its position at a tip of the tree and far from the root indicates that it must be a relatively recent haplotype, it has a worldwide distribution (table 2). Copies of D1 were found in samples from the United Kingdom, Mongolia, Nuu-Chah-Nulth Amerindians, Vanuatu, and Africa. D1 is represented by a single copy in the Kenyan Luo, where there are also two recombinant B haplotypes with 5′ flanking sequence probably from a D1 (or possibly a C1) haplotype. In addition, an ancestral haplotype for D1 (D2) was found in the Gambia.

Demographic Analysis

Maximum-likelihood estimates of θ were made with a model of exponential growth for comparison with a constant population-size model (Griffiths and Tavaré 1984*a*, 1984*b*). This analysis examines the gene tree, shown by simulation studies to appear more starlike if there has been population expansion (Harding 1996). The best-fitting expansion model was found for the Gambia gene tree with $\theta = 4N_0\mu$ of 4.5 and an expansion parameter of 0.7. In this model, there is an exponential decline backward in time of the population size at rate 0.7 from a current size $N_0 = 18{,}087$.

$$N_t = N_0 e^{-0.7t}$$

gives the size of the population at time t back. An improved fit compared with the constant population-size model has to be expected because of the addition of a parameter to the model, but this improvement was not judged to be significant by a log-likelihood test ratio. Accordingly, there is no evidence for a population-size expansion out of a small number of founders. Any population-size expansion that has occurred probably has been too recent (Takahata 1995) to be detectable in the surveyed patterns of β-globin diversity.

Maximum-likelihood estimates of θ, conditional on β-globin gene trees, assuming constant population size, reflect slightly higher genetic diversity within Africa, as was indicated by comparing numbers of segregating sites. A value for N_e was computed from $\theta/4\mu$. These estimates show that higher genetic diversity in Africa compared with Asia is a consequence of larger N_e (table 3). From the pattern of haplotype composition described above, it seems likely that N_e in Asia is smaller, not because of expansion in Africa, but because of reduction in Asia. This reduction may have been simply numerical or, alternatively, may reflect a change in the level of population structure.

TMRCA and the Ages of Mutations

The gene tree for the world data set, scaled in units of $2N_e$, is shown in figure 3. The expected TMRCA (table 3) and also the ages of the mutations (fig. 4) were estimated for each population as well as for the world data set. Estimating the TMRCA of the world data set gave a value of 750,000 years with a 95% confidence interval of 400,000–1,300,000, encompassing all of the TMRCA values from individual populations. Although coalescence times are asymmetrically distributed with a longer right tail, we found that the 95% confidence interval is well approximated by ±2 SD, and the latter are given for the ages of mutations. The differences between the individual population estimates are due both to patterns of connectivity and to allele-frequency differences. Among the population gene trees, the largest TMRCA estimate is 1.1 million years for the Mongolian sample. This tree has five nearly equally frequent alleles at its tips. The Gambia shares a very similar composition of sequence haplotypes with the Kenyan Luo, and the TMRCA estimates for these populations are both ~800,000 years. These trees have common alleles in lineages A and B, but few C or D alleles. There is no evidence for greater time depth in African populations compared with Asian populations. There is no evidence that malarial selection acting on β-globin variation in Africa and Sumatra increases or decreases the estimated ages compared with the other Asian and U.K. samples.

The confidence region around an expected TMRCA represents the effects of sampling error for estimating the gene tree, of evolutionary variance among gene trees for estimating the population genealogy, and of the random nature of the mutation process. The TMRCA is estimated in units of $2N_e$, and, in a final translation, these units are converted to years, on the assumption of a value for N_e, represented as a constant and estimated from the same genetic diversity used to estimate coalescence times. The effect of uncertainty in the estimation

Table 3

Statistics from Gene Tree Analyses

	Population									
Statistic[a]	CAR	GAM	KEN	MON	NCN	SUM	PNG	VAN	UK	World
θ	1.75	3.7	3.35	3.35	2.05	1.9	2.1	2.55	2.35	2.9
N_e	7,037	14,877	13,470	13,470	8,243	7,640	8,444	10,253	9,449	11,661
TMRCA	4.7×10^5	8.4×10^5	8.3×10^5	1.1×10^6	8.8×10^5	8.4×10^5	7.7×10^5	9.0×10^5	7.2×10^5	7.7×10^5
SD	1.7×10^5	2.4×10^5	2.4×10^5	3.0×10^5	2.6×10^5	2.5×10^5	2.4×10^5	2.6×10^5	2.0×10^5	2.1×10^5
Time to site 1358	...	1.8×10^5	2.2×10^5	4.8×10^5	5.8×10^5	5.7×10^5	2.1×10^5	3.1×10^5	3.1×10^5	1.4×10^5
SD	...	9.6×10^4	1.1×10^5	1.6×10^5	1.3×10^5	1.3×10^5	7.4×10^4	8.4×10^4	8.6×10^4	4.1×10^4
Time to site 2554	...	...	6.9×10^4	1.0×10^5	4.3×10^4	...	...	6.3×10^4	6.7×10^4	2.8×10^4
SD	...	...	6.8×10^4	7.2×10^4	4.7×10^4	...	...	4.0×10^4	4.3×10^4	1.6×10^4

[a] For all statistics but θ and N_e, data are given in years.

of N_e and μ were investigated by replacing the maximum-likelihood estimate for θ by a range of alternative values. These alternative values represent variation in estimates of N_e as 5,000, 10,000 or 50,000 and in the estimates for μ, given by divergence times between humans and chimpanzees of 4, 5, and 7 million years. TMRCA estimated conditional on the world gene tree together with the values for θ are presented in table 4. All but one of these alternative TMRCA estimates are within the 95% confidence region around the estimate when the maximum-likelihood θ value is used.

The dates estimated for mutations show wide variation between different population gene trees and in comparison with the world data set (fig. 4 and table 3). The variation is widest on sites where a sample does not allow the full gene tree to be inferred. In the Pygmy sample, for instance, the C and D lineages are absent. Consequently, the time estimated for site 2636 from the Pygmies is much more recent than from other samples where this site is ancestral to complete C and D lineages. Gene trees for Sumatra and New Guinea do not extend to the D lineage. The absence of the B3 allele from the non-African populations, and subsequent repositioning of the root, also contributes to variable estimates for ages of some sites in the Asian populations compared with those obtained from the African and world trees. We cannot exclude sampling error and evolutionary variance as a sufficient explanation for the broad range of dates. However, it is likely that differences in demographic history contribute to variable estimates between African and Asian population samples and that population structure accounts for lower time estimates from the world gene tree than from gene trees for individual populations.

We find half the sites have expected ages of ⩾200,000 years (fig. 4), and they indicate that the A1, B1, B2, B3, B9, C2, and C3 alleles are probably also of this age or older. Alleles A1 and B1 are globally distributed, B2 and B3 are localized to Africa, B9 is found in both CAR Pygmy and the U.K. samples, and C2 and C3 are Asian localized. The sites with expected ages of <150,000 years indicate that the following alleles are as young or younger: A2 (Vanuatu), A3 (Africa), A4 (Mongolia), B4 (Gambia), B11 (Kenya), C1 (global, but rare outside of Asia), C7 (Asia and United Kingdom), D2 (Gambia), and D1 (global, but low frequency).

The age estimates of the Asian-localized C2 and C3 haplotypes and of globally distributed C1 and D1 are of particular interest. D1 appears to be younger than C1. Analyses of individual populations where the D1 allele was found indicate an expected age of ~60,000 years for site 2554 in an age range of 0–200,000 years for ±2 SD (table 3). With these confidence limits it is possible that D1, as well as C1, alleles were carried worldwide by a dispersal of modern humans within the last 200,000 years. These analyses also give expected ages for site 1358 indicating that the Asian-localized C2 and C3 alleles are >200,000 years old (table 3). An alternative analysis of the world data set under an assumption of global random mating suggests younger ages for the C2, C3, C1, and D1 haplotypes (fig. 3). The estimate for site 1358 of 137,000 ± 81,500 years is not inconsistent with the evolution of the C lineage in Asia subsequent to a dispersal of modern humans within the last 200,000 years. This analysis also indicates an age for site 2554 of 28,000 ± 32,000 years, which implies that the modern human dispersal continued as substantial worldwide gene flow between established populations in Africa and Asia up to a recent date in the late Pleistocene.

Discussion

Most of the findings of this population genetic analysis of nonfunctional β-globin diversity are concordant with those of other studies. In one respect they differ,

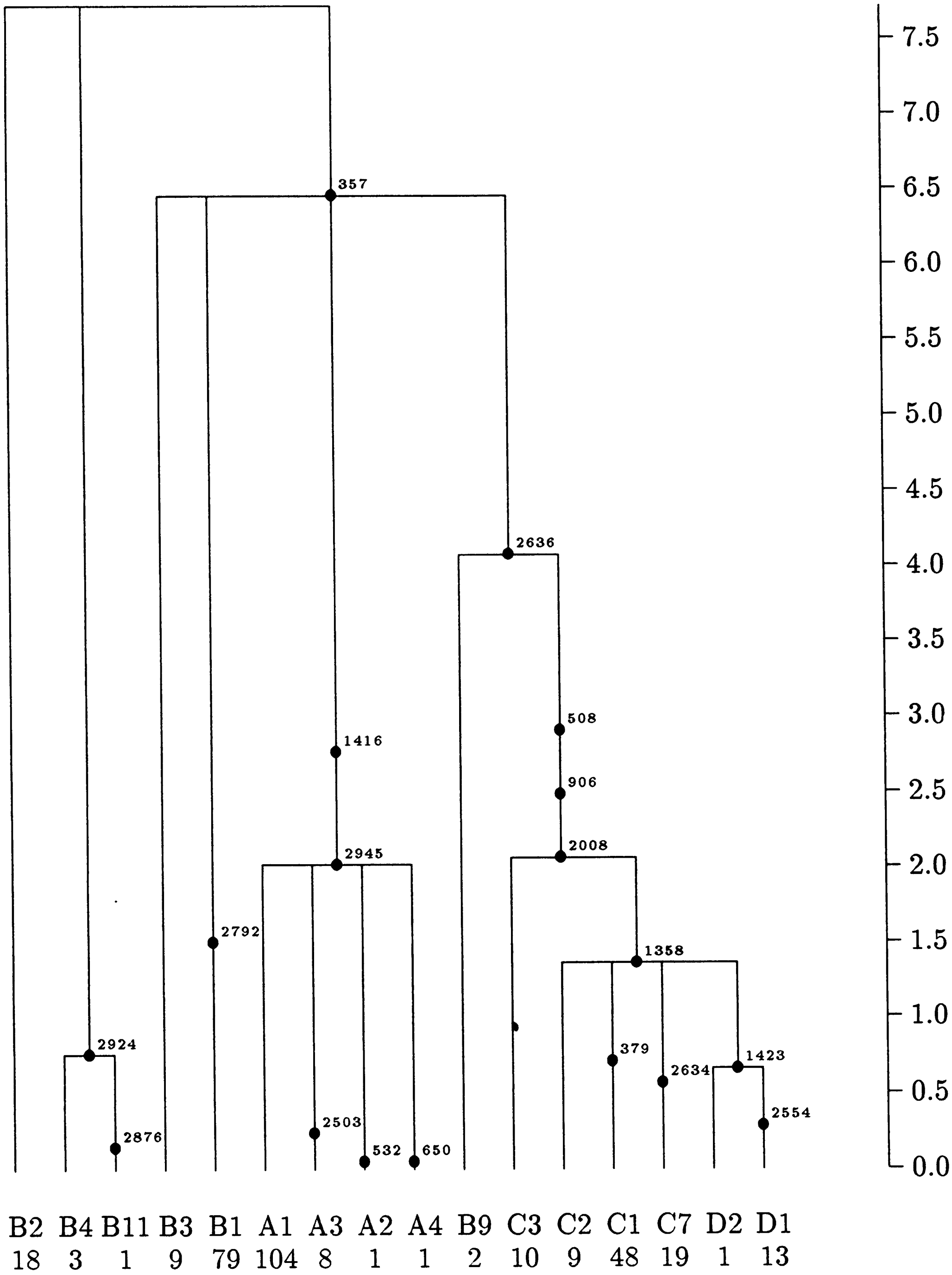

Figure 3 Scaled coalescent tree for β-globin haplotypes, showing ages of mutations estimated from the world set of 326 sequences. Time scale in units of 10^5 years.

by suggesting that modern human populations carry old Asian diversity. Not unexpectedly, the estimated TMRCA of the β-globin gene tree is ~800,000 years, and the level of β-globin diversity maintained over the last 800,000 years indicates N_e of ~10,000. The ancestral sequence for the total sample was found only in Africa. There is no evidence for an exponential expansion out of a bottlenecked founding population 200,000 years ago. More surprisingly, genetic diversity measured by pairwise sequence difference is greater in Asia than in Africa and highest in Mongolia, where extensive diversity recently has been reported for mtDNA (Kolman et al. 1996) and Y haplotypes (C. Tyler-Smith, personal communication). However, genetic diversity measured

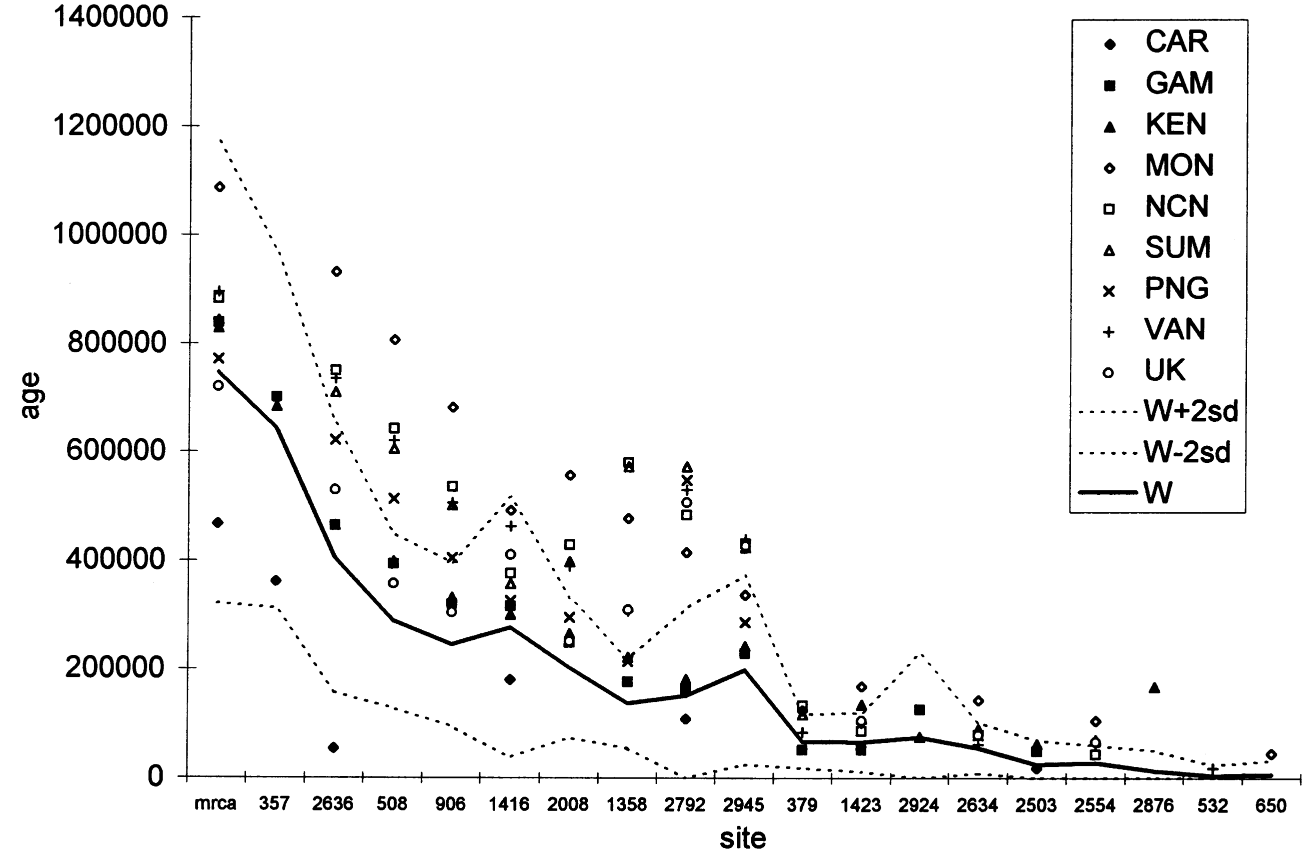

Figure 4 Scatterplot comparing ages for mutations, estimated from world (W) and individual population trees

by numbers of segregating sites is higher in Africa. Gene tree analyses suggest that N_e is greater for Africa than for Asia and that differences in N_e between Africa and Asia, as reflected by β-globin diversity, are not likely to be due to population expansion in Africa but may be due to population-size reduction in Asia.

Table 4

Estimates of TMRCA Conditional on Gene Tree and Alternative Estimates for θ

	Mutation Rate/bp/Year, Given Human-Chimp Divergence Rate		
Statistic	5 Million and 1.34 × 10^{-9}	4 Million and 1.675 × 10^{-9}	7 Million and 9.58 × 10^{-10}
θ:			
N_e = 5,000	1.2	1.6	.9
N_e = 10,000	2.5	3.1	1.8
N_e = 50,000	12.4	15.5	8.9
TMRCA (years):			
N_e = 5,000	4.8 × 10^5	3.4 × 10^5	4.1 × 10^5
N_e = 10,000	6.9 × 10^5	5.8 × 10^5	7.7 × 10^5
N_e = 50,000	1.2 × 10^6	1.0 × 10^6	1.5 × 10^6

These patterns of diversity observed in the β-globin data are consistent with those described for many other polymorphisms. First, estimates from β-globin of N_e and TMRCA are concordant with estimates from other nuclear loci (Li and Sadler 1991; Takahata 1995). Second, as for β-globin, ancestral alleles for nuclear loci judged by comparison with chimpanzee sequences typically root diversity in Africa (Mountain and Cavalli-Sforza 1994; Nei and Takezaki 1996). Third, broadly comparable levels of genetic diversity between populations as observed for β-globin have been reported from studies of classical blood groups, enzyme polymorphisms, and unlinked RFLPs (Bowcock et al. 1994; Mountain and Cavalli-Sforza 1994; Jorde et al. 1995; Takahata 1995; Nei and Takezaki 1996).

In contrast to the finding of comparable genetic diversity between populations in the studies listed above, highly allelic loci, including mtDNA (Chen et al. 1995), HLA (Hill et al. 1992), α- and β-globin RFLP haplotypes (Wainscoat et al. 1986; Martinson et al. 1995), CD4 microsatellite haplotypes (Tishkoff et al. 1996), and minisatellites (Armour et al. 1996) show substantially greater diversity in Africa than in Asia or Europe. Our analyses show that diversity differences cannot be attributed to greater time depth in Africa but may be due to the maintenance of a larger N_e in Africa. Highly allelic loci are also likely to reflect demographic expansions and contractions in Africa and Asia during the late Paleolithic and Neolithic periods not detected by β-globin diversity (Takahata 1995).

The β-globin diversity sampled for this study does not show significant structure within Africa or east Asia; nor does it reveal enhanced diversity within populations consistent with malarial selection or population-size expansion. This is probably because these processes have operated mainly within the last 10,000 years (Flint et al. 1993; Takahata 1995) and this is too recent to be detectable in random samples of the size taken, given the slow mutation rate for β-globin. It is likely that any effect of recent selection, population expansion, or structure on the estimates made from the gene trees for individual populations has been subsumed by the large evolutionary variances associated with 800,000 years of mutation and drift. This study also shows that, despite their proximity to a recombinogenic region, autosomal diploid DNA sequences can contain stretches of tightly linked polymorphism appropriate for constructing gene trees. Indeed, if recombination events are distributed nonrandomly throughout the human genome, hot spots may flag not an overall higher rate of recombination but clusters of recombination events flanked by regions of tight linkage that are ideal for analyses of gene trees.

One assumption of the gene tree analyses that is not supported is worldwide random mating. Rather, the sampled β-globin data provide evidence for a level of population structure between Africa and Asia. If random mating on a global scale is nonetheless assumed, the estimated age range of the Asian C alleles does not exclude their origin in a population derived from Africa within the last 200,000 years. However, this model requires levels of gene flow sufficient to disperse worldwide a mutation that arose ~28,000 years ago. It is difficult to assess whether such gene flow is feasible, but, more certainly, the assumption of random mating is not consistent with either the data presented here or the finding of multilocus phylogenetic studies of an African–non-African split in the modern human population. The evidence for population structure between Africa and Asia suggests that the ages of the C alleles are more reliably estimated from the Asian samples than from the world sample. It is clear that C diversity has flourished in Asia and that its relationship to B diversity in Africa is made distant by several mutations. Finding further alleles ancestral to the C/D lineage in addition to B9 and C3 will give more information about where and when C diversity evolved. On the data available, the estimated ages and distribution pattern of C haplotypes together suggest that the ancestral human population was located in Asia, as well as in Africa, >200,000 years ago.

The worldwide distribution of D alleles is particularly interesting, in the light of their low frequencies and estimated recency, whether or not random mating is assumed. It is likely that there has been an accelerated rate of multidirectional migratory activity and admixture beginning in the terminal middle and/or late Pleistocene period as suggested by Pope (1992). Although gene flow apparently has been globally pervasive, it is reasonable to infer gene flow at a rate that allows a level of population structure. The observed divergence between African and Asian patterns of β-globin polymorphism is consistent with gene flow restricted in an isolation-by-distance model (Templeton 1993, 1996).

The value of this study is that it describes allelic sequence variation for an autosomal nuclear locus, allowing inferences on more than one ancestral lineage in the human population living >200,000 years ago (Harding et al., in press). Nevertheless, a single-locus study such as this one belongs to the pilot-study phase of human evolutionary studies. Many more autosomal loci must be investigated to reduce the large evolutionary variance surrounding estimates of time depth and to sort out the uncertain effects of variable population size, nonrandom mating, selection, and recombination.

The new methods used here to estimate the ages of mutations in gene trees challenge some of the currently favored interpretations of human genetic diversity regarding the ancestral history of contemporary populations. These and other population genetic methods will be applied to nuclear sequence data as they become increasingly available, and it is likely that there is much more to learn about the evolutionary history of modern humans. Our conclusions from this study of allelic β-globin sequences are that there has been substantial multidirectional global gene flow within the last 100,000 years and that modern humans have both African and Asian ancestry dating to >200,000 years ago. We infer an earlier evolution and dispersal out of Africa by the ancestors of modern humans than indicated by some interpretations of the fossil data (Stringer and Andrews 1988) and, therefore, inclusion in the ancestral gene pool of non-African population groups identified morphologically as archaic or pre-sapiens.

Acknowledgments

We thank A. V. S. Hill, P. J. Ratcliffe, D. Tumen, A. S. M. Sofro, and R. J. Ward for DNA samples; A. J. Boyce, P. Donnelly, J. Flint, A. V. S. Hill, P. Marjoram, and J. J. Martinson for helpful discussions; and D. J. Weatherall for longstanding support. This work was funded in part by the Wellcome Trust, The Royal Society, and *The British Journal of Haematology* Research Trust.

Appendix

Here we present the allelic sequence variation found in a 3-kb region around the β-globin locus. Polymorphic sites are indicated by a number 1–3000. Polymorphic sites 130–650 lie in the 5′ flanking sequence of the β-globin gene, 906 is a twofold degenerate site in exon 1 where there is a synonymous mutation, sites 1358–2008 are in intron 2, and 2503–2945 are in the 3′ flanking sequence. GB Ref is the β-globin GenBank reference sequence. Sequence labels include the following information. A number following a letter code classifies alleles by linkage of polymorphisms 357–2945. A number preceding a letter code subclassifies alleles by linkage of polymorphisms 130–328. Functional polymorphisms are excluded from our evolutionary analyses but are indicated by allele labeling as follows: sickle-cell (β^s) at site 917 (exon 1, position 17, A→T) denoted by "s"; hemoglobin E (β^E) at site 976 (exon 1, position 66, G→A) denoted by "e"; and thalassemia (β^+) at site 994 (intron 1, position 5, G→C) denoted by "t." Ellipses (. . .) denote base deleted. In part A are the full 3-kb sequence haplotypes for 253 chromosomes from six populations; abbreviations are given in the text. The numbers of sequences carrying functional polymorphisms are proportional to their population frequencies. In part B are 2.67-kb haplotypes for 96 chromosomes randomly chosen from samples available on three populations. These samples include 2 CAR Pygmy chromosomes and 10 Kenyan Luo chromosomes carrying β^s. We could not infer haplotype linkage for these sickle-cell mutations, but it is clear that β^s is carried on both A1 and B2 haplotypes, as found in the Gambia.

Table A1

Allelic Sequence Variation for the β-Globin Locus

	SITE AND GB REF.																			
	130	138	145	284a	297	305	307	318	319a	319b	320	322	325	325a	327	328	357	379	508	532	650
	G	T	T	...	T	C	T	A	...	...	T	T	T	...	C	T	A	T	T	T	A
HAPLOTYPE	A. Full 3-kb Sequence Haplotypes for 253 Chromosomes																				
1A1																					
1A1t																					
2A1														T							
3A1											A			T							
4A1s		G							AT						...	...					
5A1					C																
6A1					C				AT	AT	A										
7A1			C		C																
8A1			C		C			T													
9A1			C		C	T															
10A1					C	T															
11A1			C		C				AT		A										
1A2			C		C				AT		A									C	
1A3					C			T													
1B5[a]					C																
1B4						T					A		...				C				
2B4			C			T											C				
3B4			C														C				
1B3											A										
1B2s						T					A						C				
2B2						T					A		...				C				
3B2				AT		T							...				C				
4B2			C			T							...				C				
5B2			C			T											C				
6B2			C														C				
1B1																					
2B1			C		C																
3B1	A		C		C																
4B1											A										
5B1			C		C				AT												
6B1					C				AT		A										
7B1			C		C				AT		A										
8B1			C		C				AT	AT	A										
9B1									AT		A				...	...					
1B6[a]									AT		A				...	...			C		
1B7[a]											A							C			
1B8[a]			C														C				
1B9									AT	AT	A	A									
1C3			C		C				AT		A								C		
2C3									AT		A				...	...			C		
3C3									AT		A	A			...	...			C		
1C2											A					...			C		
2C2											A								C		
3C2					C				AT		A								C		
4C2					C	T					A	A							C		
5C2					C				AT		A	A							C		
6C2			C		C				AT		A								C		
1C1											A							C	C		
2C1																		C	C		
3C1									AT									C	C		
4C1											A			T				C	C		
5C1							C				A							C	C		
6C1							C											C	C		
7C1			C								A							C	C		

Site and GB Ref.														Population Distribution						
906	1358	1388	1416	1423	2008	2503	2554	2634	2636	2792	2876	2924	2945							
C	C	T	G	C	T	G	G	G	C	A	G	T	G							
A. Full 3-kb Sequence Haplotypes for 253 Chromosomes														VAN	PNG	SUM	GAM	UK	NCN	Total
														4		2		12	5	23
														1		1				2
																	2			2
																	1			1
																	2			2
																	2	3		5
																		1		1
														12	1	5		5	8	31
																	1			1
																			1	1
																			1	1
														8				2		10
														1						1
						A											2			2
		C	T									C	T				1			1
			T									C	T				1			1
			T									C	T				1			1
			T									C	T				1			1
			T										T				2			2
			T										T				1			1
			T										T				1			1
			T										T				1			1
			T										T				1			1
			T										T				1			1
			T										T				1			1
			T							T			T					3		3
			T							T			T	1		1		1		3
			T							T			T			1				1
			T							T			T			1	3		1	5
			T							T			T	2	1	1		1		5
			T							T			T	2	4	2			1	9
			T							T			T	11	6	4	2	6		29
			T							T			T					5		5
			T							T			T		1					1
			T							T			T	1						1
			T							T			T	2		2				4
			T							T			T				1			1
			T						A				T					1		1
T			T		C				A				T	3						3
T			T		C				A				T		6					6
T			T		C				A				T		1					1
T	G		T		C				A				T						1	1
T	G		T		C				A				T						2	2
T	G		T		C				A				T						1	1
T	G		T		C				A				T						1	1
T	G		T		C				A				T						1	1
T	G		T		C				A				T			1				1
T	G		T		C				A				T	6		4			19	29
T	G		T		C				A				T			3				3
T	G		T		C				A				T			1				1
T	G		T		C				A				T				1			1
T	G		T		C				A				T	1		1				2
T	G		T		C				A				T			5				5
T	G		T		C				A				T						3	3

(*continued*)

Table A1 (continued)

	Site and GB Ref.																				
	130	138	145	284a	297	305	307	318	319a	319b	320	322	325	325a	327	328	357	379	508	532	650
	G	T	T	...	T	C	T	A	...	...	T	T	T	...	C	T	A	T	T	T	A
Haplotype	A. Full 3-kb Sequence Haplotypes for 253 Chromosomes																				
1C4[a]																		C	C		
1C5[a]			C		C				AT		A										
2C5[a]									AT		A										
1C6[a]						T					A		...				C				
1C7																			C		
2C7			C		C				AT		A	A							C		
3C7															...				C		
4C7											A				...				C		
5C7			C								A				...				C		
6C7t											A				...				C		
7C7e											A				...				C		
1D2					C			T											C		
1D1			C		C				AT	AT	A	A							C		
2D1											A								C		
3D1			C																C		
4D1																			C		
5D1																			C		
1D3[a]																			C		
	B. 2.67-kb Sequence Haplotypes for 96 Chromosomes																				
A1																					
A3																					
A4																					G
B10[a]																					
B11																	C				
B3																					
B12[a]																			C		
B2																	C				
B8[a]																	C				
B1																					
B13[a]																			C		
B9																					
B14[a]																	C				
C6[a]																	C				
C2																			C		
C1																		C	C		
C7																			C		
C8[a]																		C	C		
D1																			C		

[a] Recombinant sequence.

References

Allison AC (1954) Protection afforded by sickle cell trait against subtertian malarial infection. Br Med J 1:290–294

Antonarakis SE, Boehm CD, Giardina PJV, Kazazian HH Jr (1982) Nonrandom association of polymorphic restriction sites in the β-globin gene cluster. Proc Natl Acad Sci USA 79:137–141

Armour JAL, Anttinen T, May CA, Vega EE, Sajantila A, Kidd JR, Kidd KK, et al (1996) Minisatellite diversity supports a recent African origin for modern humans. Nat Genet 13: 154–160

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

Brookfield JFY (1994) A new molecular view of human origins. Curr Biol 4:651–652

Site and GB Ref.														Population Distribution						
906	1358	1388	1416	1423	2008	2503	2554	2634	2636	2792	2876	2924	2945							
C	C	T	G	C	T	G	G	G	C	A	G	T	G							
A. Full 3-kb Sequence Haplotypes for 253 Chromosomes														VAN	PNG	SUM	GAM	UK	NCN	Total
	G		T		C				A				T			1				1
T	G		T		C				A				T						1	1
T	G		T		C				A				T						1	1
T	G		T		C				A				T				1			1
T	G		T		C			A	A				T			1				1
T	G		T		C			A	A				T					2		2
T	G		T		C			A	A				T		1					1
T	G		T		C			A	A				T	1	3	2				6
T	G		T		C			A	A				T			1				1
T	G		T		C			A	A				T			1				1
T	G		T		C			A	A				T			1				1
T	G		T	T	C				A				T				1			1
T	G		T	T	C		C		A				T					1		1
T	G		T	T	C		C		A				T						1	1
T	G		T	T	C		C		A				T	1						1
T	G		T	T	C		C		A				T	3						3
T	G		T	T	C		C		A				T					3		3
	G		T	T	C		C		A				T	1		1				2
B. 2.67-kb Sequence Haplotypes for 96 Chromosomes																	CAR	KEN	MON	Total
																	9	12	3	24
						A											1	5		6
																			1	1
			T									C	T					1		1
			T								A	C	T					1		1
			T										T				1	6		7
			T										T					1		1
			T										T				4	8		12
			T							T			T					2		2
			T							T			T				6	9	3	18
T			T							T			T					1		1
			T						A				T				1			1
			T						A				T				2			2
T	G		T		C				A				T					1		1
T	G		T		C				A				T						2	2
T	G		T		C				A				T						4	4
T	G		T		C			A	A				T						6	6
T	G		T		C			A	A				T						2	2
T	G		T	T	C		C		A				T					1	3	4

Chen Y-S, Torroni A, Excoffier L, Santachiara-Benerecetti AS, Wallace DC (1995) Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. Am J Hum Genet 57:133–149

Donnelly P (1996) Interpreting genetic variability: the effects of shared evolutionary history. In: Chadwick D, Cardew G (eds) Variation in the human genome. John Wiley & Sons, Chichester, pp 25–40

Donnelly P, Tavaré S (1995) Coalescents and genealogical structure under neutrality. Annu Rev Genet 29:401–421

Dorit, RL, Akashi H, Gilbert W (1995) Absence of polymorphism at the ZFY locus on the human Y chromosome. Science 268:1183–1185

Ewens, WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol 3:87–112

Flint J, Harding RM, Clegg JB, Boyce AJ (1993) Why are some genetic disease common? Distinguishing selection from other processes by molecular analysis of globin gene variants. Hum Genet 91:91–117

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

Griffiths RC (1987) An algorithm for constructing genealogical trees. Statistics Research Report 163, Department of Mathematics, Monash University

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

——— (1994*b*) Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc Lond B 344:403–410

Gusfield D (1991) Efficient algorithms for inferring evolutionary trees. Networks 21:19–28

Haldane JBS (1948) The rate of mutation of human genes. Proceedings of the 8th International Congress of Genetics. Hereditas Suppl 35:267–273

Hammer MF (1995) A recent common ancestry for human Y chromosomes. Nature 378:376–378

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, pp 63–80

Harding RM, Fullerton SM, Griffiths RC, Clegg JB. A gene tree for β-globin sequences from Melanesia. J Mol Evol (in press)

Harpending HC, Sherry ST, Rogers AR, Stoneking M (1993) The genetic structure of ancient human populations. Curr Anthropol 34:483–496

Hill AVS, Allsopp CEM, Kwiatkowski D, Taylor TE, Yates SNR, Anstey NM, Wirima JJ, et al (1992) Extensive genetic diversity in the HLA class II region of Africans, with a focally predominant allele, *DRB1*1304*. Proc Natl Acad Sci USA 89:2277–2281

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

Hudson RR, Boos DD, Kaplan NL (1992) A statistical test for detecting geographic subdivision. Mol Biol Evol 9:138–151

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 57:523–539

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

Kingman JFC (1982*a*) The coalescent. Stochastic Processes Appl 13:235–248

——— (1982*b*) Exchangeability and the evolution of large populations. In: Koch G, Spizzichino F (eds) Exchangeability in probability and statistics. North-Holland, Amsterdam, pp 97–112

——— (1982*c*) On the genealogy of large populations. J Appl Prob 19A:27–43

Kolman CJ, Sambuughin N, Bermingham E (1996) Mitochondrial DNA analysis of Mongolian populations and implications for the origin of New World founders. Genetics 142:1321–1334

Kreitman M (1983) Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature 304:412–417

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

Martinson JJ, Excoffier L, Swinburn C, Boyce AJ, Harding RM, Langaney A, Clegg JB (1995) High diversity of α-globin haplotypes in a Senegalese population, including many previously unreported variants. Am J Hum Genet 57:1186–1198

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

Nei M, Li W-H (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5269–5273

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

Newman JL (1995) The peopling of Africa: a geographic interpretation. Yale University Press, New Haven

Newton CR, Graham A, Heptinstall LE, Powell SJ, Summers C, Kalsheker N, Smith JC, et al (1989) Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). Nucleic Acids Res 17:2503–2516

Orkin SH, Kazazian HH Jr (1984) The mutation and polymorphism of the human β-globin gene and its surrounding DNA. Annu Rev Genet 18:131–171

Orkin SH, Kazazian HH Jr, Antonarakis SE, Goff SC, Boehm CD, Sexton JP, Waber PG, et al (1982) Linkage of β-thalassaemia mutations and β-globin gene polymorphisms with DNA polymorphisms in the human β-globin gene cluster. Nature 296:627–631

Pope GG (1992) Cranioacial evidence for the origin of modern humans in China. Yearbook Phys Anthropol 35:243–298

Rohlf FJ (1993) NTSYS-pc: numerical taxonomy and multivariate analysis system. Exeter Software, Setauket, NY

Ruvolo M, 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

Sherry ST, Rogers AR, Harpending H, Soodyall H, Jenkins T, Stoneking M (1994) Mismatch distributions of mtDNA reveal recent human population expansions. Hum Biol 66:761–775

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

Strobeck C, Morgan K (1978) The effect of intragenic recombination on the number of alleles in a finite population. Genetics 88:829–844

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460

——— (1989*a*) The effect of change in population size on DNA polymorphism. Genetics 123:597–601

——— (1989*b*) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Takahata N (1995) A genetic perspective on the origin and history of humans. Annu Rev Ecol Syst 26:343–372

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

Templeton AR (1993) The "Eve" hypothesis: a genetic critique and reanalysis. Am Anthropol 95:51–72

——— (1996) Gene lineages and human evolution. Science 272:1363

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

Trabuchet G, Elion J, Baudot G, Pagnier J, Bouhass R, Nigon VM, Labie D, et al (1991) Origin and spread of β-globin gene mutations in India, Africa, and Mediterranea: analysis of the 5′ flanking and intragenic sequences of β^S and β^C genes. Hum Biol 63:241–252

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. Science 253:1503–1507

Wainscoat JS, Hill AVS, Boyce AJ, Flint J, Hernandez M, Thein SL, Old JM, et al (1986) Evolutionary relationships of human populations from an analysis of nuclear DNA polymorphisms. Nature 319:491–493

Ward RH, Frazier BL, Dew-Jager K, Pääbo S (1991) Extensive mitochondrial diversity within a single Amerindian tribe. Proc Natl Acad Sci USA 88:8720–8724

Weatherall DJ, Clegg JB (1981) The thalassaemia syndromes. Blackwell Scientific, Oxford

Wolfe KH, Sharp PM, Li W-H (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283–285