

# Minisatellite diversity supports a recent African origin for modern humans

John A.L. Armour<sup>1</sup>, Tiiu Anttinen<sup>2</sup>, Celia A. May<sup>1</sup>, Emilce E. Vega<sup>1</sup>, Antti Sajantila<sup>2</sup>, Judith R. Kidd<sup>3</sup>, Kenneth K. Kidd<sup>3</sup>, Jaume Bertranpetit<sup>4</sup>, Svante Pääbo<sup>2</sup> & Alec J. Jeffreys<sup>1</sup>

In a study of human diversity at a highly variable locus, we have mapped the internal structures of tandem-repetitive alleles from different populations at the minisatellite MS205 (*D16S309*). The results give an unusually detailed view of the different allelic structures represented on modern human chromosomes, and of the ancestral relationships between them. There was a clear difference in allelic diversity between African and non-African populations. A restricted set of allele families was found in non-African populations, and formed a subset of the much greater diversity seen on African chromosomes. The data strongly support a recent African origin for modern human diversity at this locus.

<sup>1</sup>Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, UK

<sup>2</sup>Zoologisches Institut der Universität München, P.O. Box 2021 36, D-80021 München, Germany

<sup>3</sup>Department of Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06520-8005, USA

<sup>4</sup>Laboratori d'Antropologia, Facultat de Biologia, Universitat de Barcelona, Av Diagonal 645, E-08028 Barcelona, Spain

E.E.V. Current address: Institute of Molecular Medicine, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK

\*Correspondence should be addressed to J.A.L.A.: Department of Genetics, University of Nottingham, Queen's Medical Centre, Nottingham NG7 2UH, UK

A wide variety of polymorphic loci has been used to investigate the extent of human genetic diversity and to delineate the relationships between modern human populations<sup>1–8</sup>. In addition to mitochondrial variation<sup>1,2</sup>, nuclear loci used include: classical blood group and serological markers<sup>3</sup>, RFLPs<sup>4,5</sup>, microsatellites<sup>6,7</sup>, retrotransposon insertions<sup>8</sup> and haplotypes of closely linked polymorphisms<sup>9,10</sup>. When different alleles exist in different populations, additional information on population relationships can be obtained if cladistic information can be used to define the evolutionary relationships among the alleles<sup>9–11</sup>. For example, the analysis of haplotypes of closely linked polymorphisms has allowed the inference of phylogenetic relationships between the different states encountered and thus the reconstruction of a probable history for that chromosomal segment, in which different haplotypic states differ either by new mutation or recombination<sup>9,10</sup>.

Tandemly repeated minisatellite loci include the most variable loci described to date in humans. By virtue of their extensive length variation between different alleles, they have found many applications in genetic analysis, including the establishment of individual identity and family relationships<sup>12–14</sup>. The high levels of population variability at these loci are due to a high rate of germline mutation to new allelic states, at frequencies (up to 15% per gamete) high enough to measure by direct observation in pedigrees and single molecule analysis of germline DNA<sup>15–17</sup>. Analysis of genetic diversity by simple measurement of allele length does not provide information on the cladistic relationships between specific alleles; information on these relationships can, however, be obtained by deter-

mining the interspersion pattern of variant repeats along individual alleles<sup>18–22</sup>.

These interspersion patterns can be conveniently determined using minisatellite variant repeat-PCR (MVR-PCR<sup>19–22</sup>), and have been used to study the mechanisms underlying the generation of new allelic diversity by germline mutation<sup>17,22</sup>. The striking feature of these studies at some<sup>17,19–21</sup> but not all<sup>22,23</sup> hypervariable minisatellite loci is that mutational rearrangements do not occur randomly but are preferentially located at or near one extremity of the repeat array. At such loci, germline mutation, and therefore recently generated variation, is polar — one extremity of the locus is extremely variable, undergoing rearrangement at high frequency, whereas other parts of the locus are more stable in the germline and have a smaller repertoire of population variability.

The minisatellite locus MS205 (*D16S309*) combines a number of properties useful in the study of human diversity. It is a highly variable nuclear locus, with an estimated true heterozygosity in Europeans of approximately 99.7% (ref. 21). The germline mutation rate has been estimated from extensive paternity casework as approximately 0.4% per gamete<sup>24</sup>. Unlike many other hypervariable loci, MS205 is relatively short, with all alleles observed to date less than 5 kb long (up to 87 repeats of a 45–54 bp repeat unit), so that full-length allelic structures can be obtained for all alleles by MVR-PCR<sup>21</sup>. MVR-PCR analysis of new mutant alleles<sup>17</sup> and of alleles selected at random from a European population<sup>21</sup> has shown that the mutation process is highly polar, with all pedigree mutations so far involving mutational change within five repeat

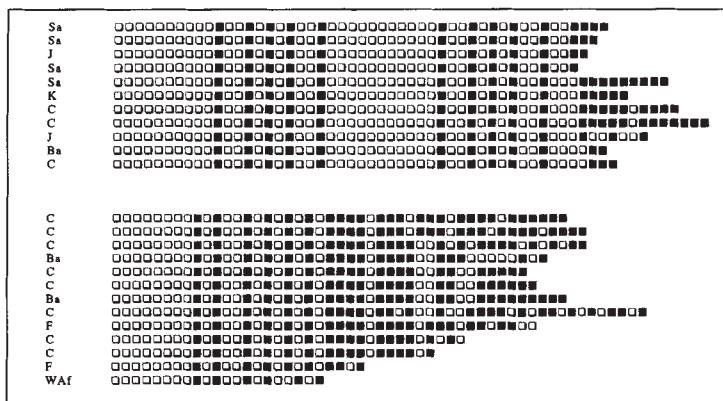
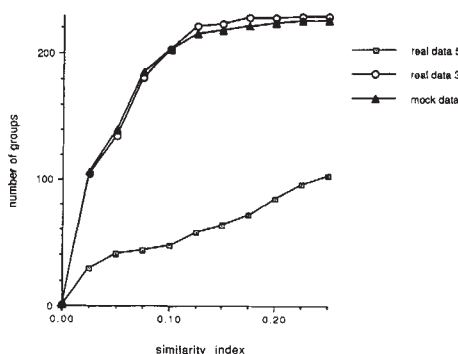


Fig. 1 Selected examples of similar, presumably closely related alleles at minisatellite MS205, demonstrating preferential variation within each group at the 3' end of the locus, here shown on the right. A-type repeat units<sup>21</sup> are shown as black boxes, T-type as white boxes. The population from which each allele is drawn is shown (Sa = Saami; J = Japanese; K = Kenyan; C = North European; Ba = Basque; F = Finnish; Waf = West African).

units of one extremity of the locus. Instability preferentially acting at one end of the locus has the effect that new mutations repeatedly 'over-write' the effects of older changes. By contrast, away from this highly unstable region, the rest of the locus allows some of the deeper evolutionary history of an allele to be retained in its structure without being obscured by subsequent mutation (Fig. 1). This can be used as a simple basis for inferring relationships between alleles: differences between two alleles that are confined to the unstable region of the locus suggest recent common ancestry, whereas differences at the more stable part of the array suggest a more distant relationship<sup>21</sup>. Thus by contrast with many other types of markers used in the study of human population diversity, at MS205 there exist simple rules, derived from direct observation of germline mutations, that allow the delineation of affinities between different allelic states; diversity within and between populations can therefore be analysed at the level both of individual alleles and of larger families of structurally related alleles sharing common ancestry. We have used MVR-PCR to survey allelic structures at MS205 from a geographically diverse selection of human populations.

### Groups of closely related alleles

Interspersion patterns of A-type and T-type minisatellite variant repeats were mapped in a total of 330 alleles from Finns, Saami (=Lapps), North Europeans (from the CEPH families<sup>21</sup>), Moksha (Finno-Ugric



speakers of the Urals), Basques, West Africans, Kenyans, Zimbabweans, Japanese, Melanesians and Rondonian Surui (from Western Amazonia). Overall, 229 different structures were found; observed frequencies of individual alleles within each population sample ranged from 0.009 (alleles seen only once among the 106 CEPH alleles) to 0.75 (allele CE1 [see below] in Surui).

Simple visual inspection of allele structures, or an analysis of the frequency of repeat unit motifs in different alleles that makes no assumptions about the mutational process (not shown), showed clearly that allelic variation at MS205 was distributed non-randomly between the populations studied. Methods were therefore developed to allow objective grouping of related alleles.

Previous studies of European MS205 alleles<sup>21</sup> and direct analysis of germline mutations<sup>17</sup> both suggested that differences between closely related alleles most frequently accumulated at the 3' end of the locus (shown on the right in Fig. 1). Mutations in other locations have presumably arisen on occasion in the history of the locus, and the data set was therefore analysed for internal evidence of other patterns of mutation (see Methods). Examination of all possible pairwise combinations of alleles failed to demonstrate any clearly related pairs of alleles differing only by changes at other locations; alignment of alleles was therefore carried out at the 5' end, since this would most frequently pair up alleles along regions of common ancestry.

To assemble similar alleles into groups, pairwise comparisons of allelic structures were made after alignment at the 5' end, and a pairwise similarity score between each pair derived. Groups of alleles sharing a given minimum similarity score were then assembled (see Methods). Comparisons were made of the number of groups found over a range of different similarity scores with the same analysis performed on 229 randomised 'mock' alleles, and also with the real data set but with pairs of alleles aligned from the opposite (3') end (see Methods). The results (Fig. 2) confirm the impression from visual inspection that there are significant similarities between alleles aligned from their 5' ends. The curve for real alleles aligned from the opposite (3') end follows the trajectory of the curve for randomised data, further supporting the idea that evolution at MS205 is well modelled by 5' alignment of alleles, and that different alleles rarely if ever show identity restricted to the 3' end. A similarity score of 0.125 gave the greatest difference between the real and

Fig. 2 Overall results of comparisons between allelic structures at MS205. The numbers of groups found at different criterion similarity scores are compared between (i) 5' aligned alleles from the data set (ii) 3' aligned alleles and (iii) 5' aligned mock alleles, made by randomly permuting the repeat units of each of the 229 different alleles. A similarity score of 0.125 was subsequently used as a lower threshold of similarity in the analysis of groups of alleles.

'mock' data, and was used in the analysis discussed below. The use of other similarity scores (0.05–0.15) did not affect the overall conclusions (data not shown).

### Population diversity: Africans and non-Africans

Analysis of ancestral relationships between alleles was carried out by assembling groups of similar, presumably closely related alleles after 5' alignment (Figs 3, 4). Common motifs were found among alleles from non-Africans studied, even from geographically very distant populations such as Melanesians and Saami. Among 242 alleles observed in non-Africans, 226 fell into groups 1–6; these groups in turn show clear inter-group relationships, with a shared central motif (shown in red in Fig. 3), suggesting a relatively recent common ancestry for these non-African populations. By contrast, a wider variety of groups was represented in the African alleles, including some groups (7, 8, 9 and 12) of which members were not observed at all outside Africa, and a large number of 'singleton' alleles which did not group with any other non-identical allele. Of the 88 alleles sampled from Africans, 37 belong to those groups (1–6) that predominate in non-Africans, consistent with the origin of the non-African populations as a subgroup of an ancestral African population. Additional detail from mapping autoradiographs was used to examine African alleles with relatively uninformative structures (generally classified as ungrouped singletons), and suggested that while some show structural affinities, most have distinct patterns of variant T-type repeat units (data not shown; see Methods); these alleles do not therefore exaggerate the number of different allele lineages represented in Africans.

The simple demonstration of higher diversity among Africans has been noted in other systems and is compatible with a number of hypotheses that do not assume an African origin; for example, greater diversity could result if effective population sizes have historically been higher in Africa<sup>25</sup>. Because we can apply the observed mechanisms of germline mutation to deduce groups of alleles sharing common ancestry, however, our data go further than this: there are more distinct lines of descent represented in Africa, and the lineages present outside Africa form a subset of the global variation. Furthermore, the same limited set of lineages is shared by the non-African populations studied.

These data are very difficult to reconcile with a simple hypothesis of multi-regional development; such a model does not predict the sharing of a small number of lineages by (separately developing) populations as far afield as Scandinavia and South America, nor does it account for the marked difference observed between African and other populations. We can consider three basic explanations for the observed patterns of allelic diversity at MS205. First, it is possible that mutation rates at MS205 are higher in some lineages, and thus that there may have been preferential allelic diversification among the alleles found in Africa. We have no evidence to exclude this possibility, but the predicted recent clusters of closely related alleles are not seen in the African data, and this hypothesis fails to explain why these highly unstable lineages are not also represented outside Africa. Second, selective constraints at

this locus may have specifically acted to maintain higher diversity in Africa (or to reduce diversity outside Africa); if selection were to account for reduced variation outside Africa, the model would have to explain the selection of the same restricted set of alleles in widely dispersed populations. While the MS205 tandem repeat array may be evolving without selective constraint, it is not possible to exclude local indirect selective forces which may have shaped the evolution of the locus. However, the concordance of the basic difference found here between African and non-African populations with the conclusions of other studies, most notably from RFLPs<sup>4,5</sup>, mitochondrial DNA<sup>1,26</sup>, microsatellite allele frequencies<sup>6</sup> and haplotype analysis<sup>9,10</sup>, would on the contrary suggest that the observed pattern reflects the true pattern of human origins rather than being a peculiarity of this minisatellite locus. Most recently, the analysis of haplotypes at the *CD4* locus<sup>10</sup> has demonstrated a similar pattern to that found here, in which the existence of a small subset of lineages in different non-African populations suggests that these populations may have a recent common origin. Overall, and taken together with the results of other studies, our data are most simply consistent with a third explanation, that there has been a major division in the origin of modern humans between African and non-African populations, with a founder non-African population arising as a subset of a larger population found in Africa. So far, the different systems used to investigate human diversity have given results consistent with the general conclusion of an African origin, rather than demonstrating the exclusion of all other possibilities<sup>25</sup>; a single substantiated counter-example would be sufficient to cast serious doubt on this view of human origins.

### Timing divergence: alleles shared between populations

If all MS205 allele maps could be aligned to create a single phylogeny incorporating all mutation events, and if mutation rates both for polar and non-polar changes were known, then it would be feasible to date allele lineages and to estimate the time of the split between African and non-African populations. Problems of allele alignment associated with superimposed mutations, and lack of information on non-polar mutation rates makes such an approach impossible. There is however an alternative approach to estimating population divergence times, using information on identical alleles shared between different populations together with mutation rate data.

Numerous examples exist of identical MS205 alleles shared between different African populations and between different non-African populations (Table 1). There are by contrast only three instances of identical alleles shared between African and non-African populations (Table 1, Fig. 5), out of 88 and 242 alleles analysed in the two groups respectively. To estimate the maximum time required to reduce allele sharing to this level between two diverging populations, we made the worst-case assumption that the ancestral population contained only three different and equally frequent alleles. Computer simulations with the measured mutation rate of 0.4% (ref. 24) showed that the rate of decay of allele sharing between diverging

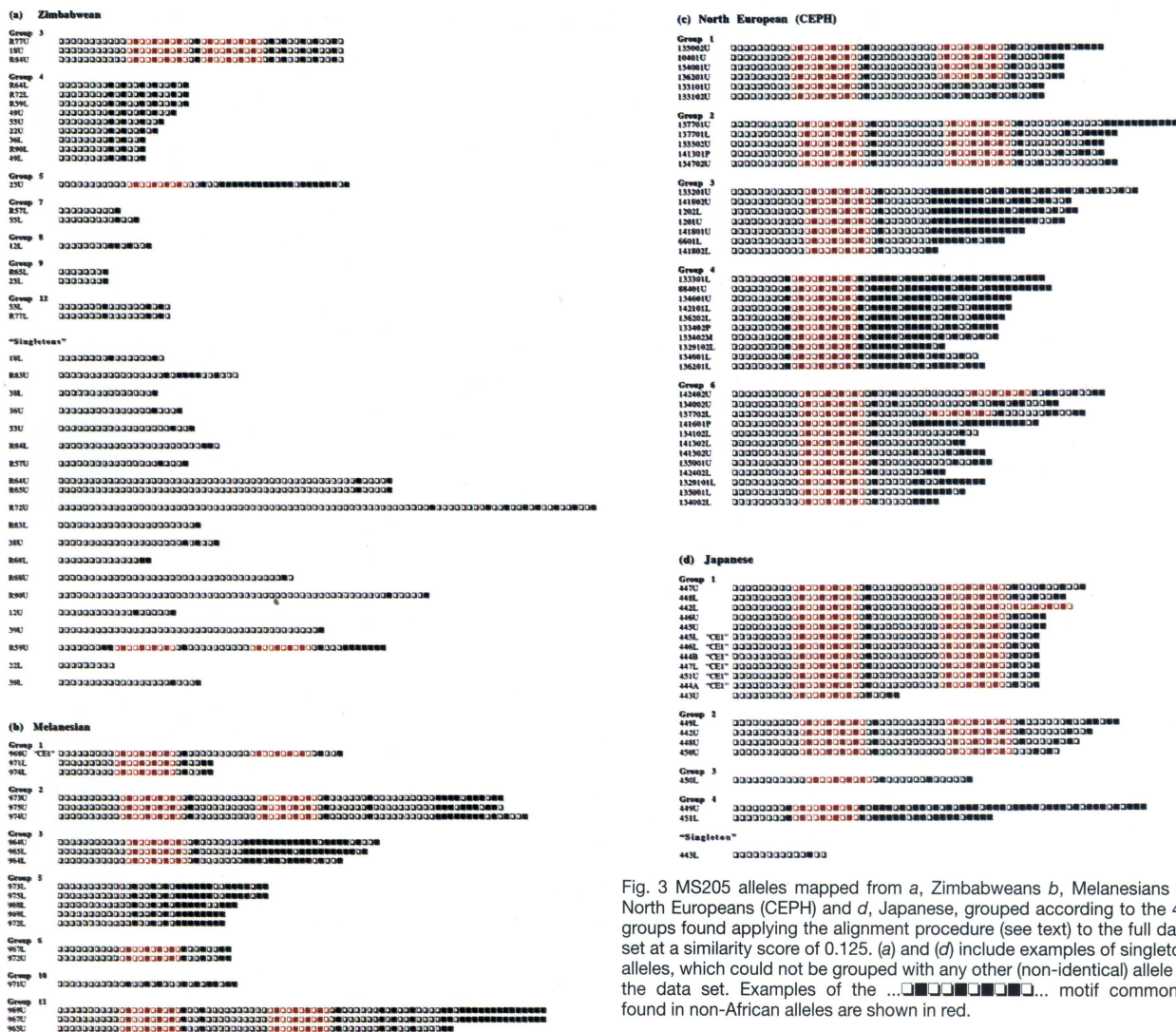


Fig. 3 MS205 alleles mapped from *a*, Zimbabweans *b*, Melanesians *c*, North Europeans (CEPH) and *d*, Japanese, grouped according to the 44 groups found applying the alignment procedure (see text) to the full data set at a similarity score of 0.125. (a) and (d) include examples of singleton alleles, which could not be grouped with any other (non-identical) allele in the data set. Examples of the ...□□□□□□□□... motif commonly found in non-African alleles are shown in red.

populations was largely independent of population size above 1,000, being driven by mutation rate rather than genetic drift (see Methods). Decay to the observed level of three instances of African/non-African allele sharing required an average of 770 generations (15,000 years; 95% upper confidence limit 22,000 years). This estimated divergence time for the split between African and non-African populations is much more recent than other estimates, for example based on mitochondrial DNA<sup>1</sup>. There are several possible explanations why the method adopted here might underestimate the true age. First, the examples of shared alleles may not be identical by descent but have arisen by convergent evolution; this is unlikely, since even variations in band intensity observed on mapping autoradiographs agree between the different copies (Fig. 5). Although convergence from widely differing ancestral alleles is thus unlikely, it is not possible to exclude convergence to the same state from a more recent common ancestor, because there is a preponderance of mutations

involving small changes at the unstable end of the locus<sup>17</sup>. Second, there has been gene flow, perhaps very recently, between the African and non-African popula-

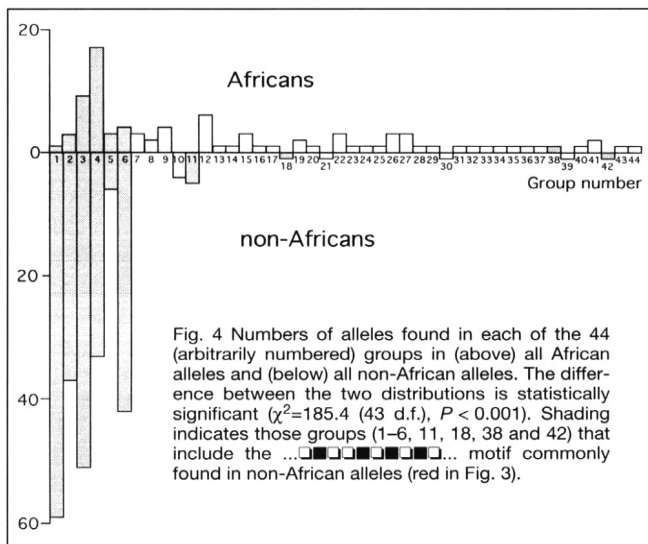


Fig. 4 Numbers of alleles found in each of the 44 (arbitrarily numbered) groups in (above) all African alleles and (below) all non-African alleles. The difference between the two distributions is statistically significant ( $\chi^2=185.4$  (43 d.f.),  $P < 0.001$ ). Shading indicates those groups (1-6, 11, 18, 38 and 42) that include the ...□□□□□□□□... motif commonly found in non-African alleles (red in Fig. 3).

**Table 1 Allele sharing between different populations sampled**

	K (15)	Waf (24)	Mok (15)	Ba (15)	C (87)	F (17)	Sa (18)	J (14)	Mel (15)	Su (4)
Z (31)	4	7	-	-	-	-	-	-	-	1
K (15)		2	-	-	-	1	-	-	-	-
Waf (24)			-	-	1	-	-	-	-	-
Mok (15)				1	-	-	1	1	1	1
Ba (15)					1	-	-	-	-	-
C (87)						1	2	-	-	-
F (17)							1	-	-	-
Sa (18)								2	1	1
J (14)									2	1
Mel (15)										1

The figures show the number of distinct alleles found in each of a pair of populations: Z = Zimbabwean; K = Kenyan; Waf = West African; Mok = Moksha; Ba = Basque; C = North European (CEPH); F = Finnish; Sa = Saami; J = Japanese; Mel = Melanesian; Su = Surui. Figures in parentheses show the number of distinct alleles found in each sample. The rectangle delineated contains the comparisons between each African/non-African pair of populations sampled.

tions sampled — this also seems unlikely, particularly between the geographically remote Zimbabweans and Surui (Fig. 5a), or between Kenyans and Finns (Fig. 5c). Another possibility is allelic heterogeneity in mutation rate, as already described at minisatellite MS32 (*DIS8*)<sup>27</sup>, with alleles shared between populations being biased towards those showing unusual germline stability. As the rate of decay of allele sharing is proportional to mutation rate (see Methods), a split between African and non-African populations about 150,000 years ago would require that shared alleles have mutation rates reduced by about 10-fold over the average of 0.4%/gamete. The existence of alleles shared between Africans and non-Africans therefore suggests that there may be considerable mutation rate heterogeneity at MS205, and that direct measurement of mutation rates for the relevant alleles (see Discussion) may help place more precise upper limits on the age of the observed divergence.

The structure marked 'CE1' (Fig. 3b, 3d) is found at high frequency in Saami ( $f = 0.21$ ) and Japanese ( $f = 0.3$ ) and at very high frequency in Surui ( $f = 0.75$ ), and as single examples each from the Moksha and Melanesian samples among all the other populations studied. The high frequency of this allele may be the result either of recent divergence and expansion in these populations, or else of unusual germline stability of this allele. If we accept that the Saami moved westwards from Siberia to reach their present location<sup>3</sup>, and that the ancestors of the indigenous populations of North and South America arrived from Asia via Alaska<sup>3</sup>, then a geographical unity is suggested for these high frequencies of allele *CE1* as a marker of chromosomes present in ancient populations of North and East Asia; this pattern of distribution also correlates with a mitochondrial variant (16298T) found at high frequencies in Saami, Japanese and Amerindian populations<sup>28</sup>.

**Discussion**

The data presented here demonstrate a marked difference between the diverse lineages found in African populations and the limited subset that predominate throughout non-African populations. For this reason, populations that inhabit intermediate regions between Africans and non-Africans, such as North East Africa and the Middle East, may be of particular

interest in delineating features of the initial subgroup that diverged and expanded 'out of Africa'. Other features characteristic of particular population groups have also been found, such as the allele *CE1* common in Saami, Japanese and Surui, and these may be of

use in tracing nuclear lineages in other geographical locations; for example, the stepwise dispersal of humans across Polynesia may have involved dramatic 'bottlenecks' in population size, as suggested by studies of minisatellite allele length<sup>29</sup>, and the ability to trace the relationships between alleles predominating in different locations may be highly informative.

A number of questions are raised by these data about possible differences between the mutation rates of particular alleles or allelic lineages, especially of those alleles shared between African and non-African populations (Fig. 5) and of the allele (*CE1*) at high frequency in Saami, Japanese (Fig. 3d) and Surui. In principle, these questions can be addressed directly by analysis of male germline DNA using small-pool PCR<sup>17</sup> and work is in progress to develop such methods for MS205 (C.A.M., J.A.L.A. and A.J.J., unpublished data). Since MS205 undergoes mutation predominantly in the male germline (11/12 germline mutations so far verified are paternal in origin (ref. 17 and J.A.L.A., unpublished data)), analysis of sperm mutations should give a largely representative view of overall mutational behaviour for any allele tested.

In this study, no attempt has been made to construct

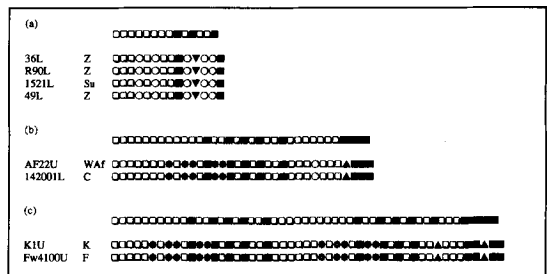


Fig. 5 The three examples of identical alleles found in both African and non-African populations in this study. a, Z=Zimbabwean; Su=Surui; b, Waf=West African; C= North European; c, K = Kenyan; F = Finnish. In each case the shared binary map is shown above maps which incorporate additional information from intensity differences between repeat units seen on autoradiographs. Repeat units that give reduced signals (either with forward or reverse mapping primers<sup>21</sup>) are indicated by additional symbols: ■ = A-type; ▲ = weak A-type; ▼ = very weak A-type; □ = T-type; ○ = weak T-type; ● = very weak T-type. The agreement between the structures of the maps shared between African and non-African populations even with this additional detail suggests that they are not simply examples of fortuitous convergence.

a complete 'tree' of alleles mainly because only the most recent branchings are described by the mutational processes for which we have direct observational evidence, and the deeper branches will involve rarer (perhaps non-polar) events for which much less descriptive data can be obtained. Nevertheless, future work may allow analysis of affinities *between* the small clusters of recently diverged alleles by studying their association with polymorphic sites in the DNA flanking the repeat array; although gene conversion-like processes appear to have disrupted associations between minisatellite alleles and flanking polymorphisms at the (active) 3' end of the locus, associations between the tandem repeat array and 5' flanking polymorphisms appear to be more robust<sup>21</sup>, and may be a useful basis for understanding the detailed longer-term histories and affinities of entire groups of alleles. Similarly, no population tree based on allele frequencies has been derived because the sample sizes from each population are relatively small in this study; larger surveys may allow better estimates of allele or group frequencies.

Can other human loci give a similarly clear and detailed view of divergence between different lineages of a chromosomal segment? The chief reasons for the clarity and detail of these data are that MS205 (i) is a highly *unstable* locus, with many recently generated new alleles, (ii) undergoes *compartmentalized* (polar) mutation, so that more recent mutation does not often act to obscure the deeper ancestry of an allele and (iii) is relatively *short*, so that complete structural analyses can be made of even the longest alleles encountered. Many other minisatellite loci fail to meet one or other of these essential criteria, usually because they are either too long, like *DIS8* (ref. 19) or *D7S21* (ref. 20), or because their mutations are frequently non-polar, like *D2S90* (ref. 22) and *D5S110* (ref. 23). It may nevertheless be possible to find other loci with the right combination of properties to act as similarly efficient reporters of mutation and divergence at additional chromosomal locations.

## Methods

**Mapping MS205 alleles by MVR-PCR.** Minisatellite variant repeat (MVR) interspersions patterns were mapped as described<sup>21</sup>, except that a new primer 205TAG-T (5'-TCAT-GCGTCCATGGTCCGGACTCACCYGCCCGTACAC-3') was used in forward mapping instead of primers 205TAG-N and 205TAG-N2, to give better detection of T-type repeat units. In all, 330 alleles were mapped from 165 individuals of the following origins: North European (Indo-European speakers from the CEPH families, 106 alleles<sup>21</sup>), Finnish (18 alleles), Saami (=Lapps, from both Sweden and Finland, 24 alleles), Moksha (Finno-Ugric speakers of the Urals, 16 alleles), Basque (18 alleles), West African (from Ibadan, Nigeria, 30 alleles), Kenyan (Mijikenda from the Kilifi district, 18 alleles), Zimbabwean (from Harare, 40 alleles), Japanese (from Nagoya, 20 alleles), Melanesian (Nasioi speakers from Bougainville, 20 alleles) and Rondonian Surui (Lupi speakers from Western Amazonia, 20 alleles).

**Alignment of alleles according to 5' similarity.** To arrive at an alignment procedure in which alleles would be grouped as accurately as possible according to their ancestry, the data set was examined in detail for evidence of the structural basis of mutational change. Previous studies<sup>17,21</sup> suggested that mutational differences between closely related alleles most frequently accumulated at the 3' end of the locus (shown on the right in

Fig. 1). Thus the 5' end of the locus will frequently escape mutation, thereby preserving structural evidence of common ancestry between related alleles, and an algorithm that aligns alleles according to similarity at the 5' end should generally assemble true families of alleles sharing common ancestry. To check for evidence of other patterns of mutation the data set was analysed for pairs of similar alleles that differed at locations other than the mutationally most active 3' extremity of the array. Examination of all possible pairwise combinations of alleles (data not shown) confirmed that mutations away from the preferentially unstable 3' end of the locus are sufficiently uncommon in the evolution of MS205 that groups of alleles assembled using alignment at the 5' end will be largely monophyletic.

**Comparison between alleles and assembly into groups.** To assemble alleles into groups sharing 5' similarity (and therefore probable common ancestry), pairwise comparisons of allelic structures were made after alignment at the 5' end, expressing a similarity score between each pair as the number of switches of repeat type (from 'A'-type to 'T'-type, or vice versa) matched at the 5' end, divided by the mean allele length. Switches of repeat type were counted to align alleles using their structurally most informative features. Groups of alleles sharing a given similarity score were then assembled by a stepwise process, building up groups by starting with the most similar alleles and gradually lowering the similarity required for inclusion until the threshold was reached (details and programs available on request). Groups were fully mutual, in that to be included an allele had to fulfill the similarity criterion with every other member of the group. The validity of the groups so assembled was assessed by determining the number of groups found using a range of different similarity scores. At a shared similarity score of 0, all alleles can be placed together in one group; as the required score increases, a greater number of smaller groups will be formed, until finally all non-identical alleles are in different groups. Alignment results were compared with the same analysis performed on 229 mock alleles, in which each allele in the data set was randomly permuted to give a new structure of identical length and composition, and similar internal complexity (frequency of doublets, triplets, quartets of repeat units). A comparison was also made with the real data set but with pairs of alleles aligned from the opposite (3') end.

The comparison between the real and mock data (Fig. 2) showed a clear difference, presumably reflecting authentic structural similarity in the real data, with the greatest difference at a similarity score of 0.125; as it appeared to maximize the signal-to-noise ratio from the data, this score was used in the analyses presented. Tests of the robustness of the grouping procedure were carried out by randomizing the order of the input data prior to grouping at a cut-off of 0.125. Four of these five randomized runs gave identical results (as shown in this manuscript). One run differed only in the distribution of some of the alleles originally placed in groups 2, 6, and 10; the total number of groups, and the grouping of other alleles was unchanged.

**Additional detail from mapping autoradiographs.** The intensity of bands produced on MVR autoradiographs can show reproducible heterogeneity even within one type of repeat unit<sup>21</sup>. As only some of the observed sequence variation between MS205 repeats is assayed by the MVR mapping procedure<sup>21</sup>, these reproducible variations in intensity almost certainly indicate further, uncharacterized repeat sequence variants. This further detail in the MVR maps was used to provide additional information both for African alleles containing long runs of 'T'-type repeat units, and for alleles shared between African and non-African populations.

Some African alleles contained long runs of 'T'-type repeat units; these uninformative alleles may not reveal their true affinities in our analysis, and yield ungrouped singletons following allele grouping. Analysis of additional information from variant 'T'-type repeat units generally shows no simple

affinities between these alleles, and we conclude that these relatively 'bland' alleles do not simply form one large family, but instead belong to diverse lineages. The three examples of identical alleles found in both African and non-African populations were examined for additional information from intensity differences on autoradiographs. No differences were found between the maps even at this more detailed level of analysis (Fig. 5), confirming true common ancestry.

**Simulating mutational decay in allele frequency.** The mutation rate at MS205 has been measured at 0.4% per gamete<sup>24</sup>. Assuming that identical structures will only be very rarely recreated by new mutation, it is possible to predict the rate of decay in the frequency of an allele with time. Decay in allele frequency was simulated using population sizes up to 15,000 (30,000 alleles). Integer arrays with each position representing one chromosome were set to an initial value of 1, and 'mutation' (to 0) and random sampling into the next generation carried out repeatedly until the frequency of the intact allele (value 1) fell below a threshold level (programs available on request). As population size increased, the effects of drift became less significant, so that at large population size (>1000) the time taken to decay to a particular level was mainly dependent upon mutation rate. In the absence of drift, the frequency  $f$  after  $n$  generations will be given by  $f = (1 - \mu)^n$ , from which  $n = [\log_e f] / [\log_e (1 - \mu)]$ . Thus the times ( $n_1$  and  $n_2$ ) taken to decay

to a given level  $f$  at two different mutation rates  $\mu_1$  and  $\mu_2$  will be in the ratio  $n_1/n_2 = [\log_e (1 - \mu_2)] / [\log_e (1 - \mu_1)]$ . At low values of  $\mu$ , this approximates to  $n_1/n_2 = \mu_2/\mu_1$ , indicating that the rate of decay in allele frequency is proportional to mutation rate.

Decay simulations were then carried out to model the experimental observations, using a founder population with only three, initially equally frequent, alleles, from which two sub-populations were derived. Population divergence was continued until samples of 88 alleles from one sub-population and 242 from the other no longer gives as many as three shared alleles.

**Acknowledgements**

We thank K. Tamaki, Y. Katsumata, A.D. Nkomo, S.B. Kanoyangwa, C. Tyler-Smith, J. Wainscoat, R. Neumann, M. Jobling and M. Webb for their help and contribution to this work. We are also grateful to R. Trembath for access to computing facilities. This work was made possible by grants from the Wellcome Trust (038225/Z/93/Z), CEC (EV5V-CT910585), MRC, Royal Society, DFG (Pa452/2-1, to S.P.) and U.S. National Science Foundation (SBR9408934, to J.R.K.). The work of A.J.J. was also supported by an International Research Scholar's Award from the Howard Hughes Medical Institute.

Received 13 February; accepted 29 March 1996.

1. Cann, R.L., Stoneking, M. & Wilson, A.C. Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987).
2. Templeton, A.R. Human origins and the analysis of mitochondrial DNA sequences. *Science* **255**, 737 (1992).
3. Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes*. (Princeton University Press, Princeton, 1994).
4. Wainscoat, J.S. *et al.* Evolutionary relationship of human populations from an analysis of nuclear DNA polymorphisms. *Nature* **319**, 491–493 (1986).
5. Kidd, K.K. & Kidd, J.R. A nuclear perspective of human evolution. in *Cambridge Symposium on Molecular Biology and Human Diversity*. (ed. Boyce, A.J.) (Cambridge University Press, in the press).
6. Bowcock, A.M. *et al.* High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**, 455–457 (1994).
7. Edwards, A., Hammond, H.A., Jin, L., Caskey, C.T. & Chakraborty, R. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics* **12**, 241–253 (1992).
8. Batzer, M.A. *et al.* African origin of human-specific polymorphic Alu insertions. *Proc. Natl. Acad. Sci. USA* **91**, 12288–12292 (1994).
9. Castiglione, C.M. *et al.* Evolution of haplotypes at the DRD2 locus. *Am. J. Hum. Genet.* **57**, 1445–1456 (1995).
10. Tishkoff, S.A. *et al.* Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**, 1380–1387 (1996).
11. Templeton, A.R., Routman, E. & Phillips, C.A. Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics* **140**, 767–782 (1995).
12. Jeffreys, A.J., Wilson, V. & Thein, S.L. Hypervariable 'minisatellite' regions in human DNA. *Nature* **314**, 67–73 (1985).
13. Nakamura, Y. *et al.* Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* **235**, 1616–1622 (1987).
14. Armour, J.A.L. & Jeffreys, A.J. Biology and applications of human minisatellite loci. *Curr. Opin. Genet. Dev.* **2**, 850–856 (1992).
15. Jeffreys, A.J., Royle, N.J., Wilson, V. & Wong, Z. Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* **332**, 278–281 (1988).
16. Vergnaud, G. *et al.* The use of synthetic tandem repeats to isolate new VNTR loci: cloning of a human hypermutable sequence. *Genomics* **11**, 135–144 (1991).
17. Jeffreys, A.J. *et al.* Complex gene conversion events in germline mutation at human minisatellites. *Nature Genet.* **6**, 136–145 (1994).
18. Jeffreys, A.J., Neumann, R. & Wilson, V. Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell* **60**, 473–485 (1990).
19. Jeffreys, A.J., MacLeod, A., Tamaki, K., Neil, D.L. & Monckton, D.G. Minisatellite repeat coding as a digital approach to DNA typing. *Nature* **354**, 204–209 (1991).
20. Neil, D.L. & Jeffreys, A.J. Digital DNA typing at a second hypervariable locus by minisatellite variant repeat mapping. *Hum. Mol. Genet.* **2**, 1129–1135 (1993).
21. Armour, J.A.L., Harris, P.C. & Jeffreys, A.J. Allelic diversity at minisatellite MS205 (D16S309): evidence for polarized variability. *Hum. Mol. Genet.* **2**, 1137–1145 (1993).
22. Buard, J. & Vergnaud, G. Complex recombination events at the hypermutable minisatellite CEB1 (D2S90). *EMBO J.* **13**, 3203–3210 (1994).
23. Armour, J.A.L., Crosier, M. & Jeffreys, A.J. Distribution of tandem repeat polymorphism within minisatellite MS621 (D5S110). *Ann. Hum. Genet.* **60**, 11–20 (1996).
24. Royle, N.J., Armour, J.A.L., Webb, M., Thomas, A. & Jeffreys, A.J. A hypervariable locus D16S309 located at the distal end of 16p. *Nucl. Acids Res.* **20**, 1164 (1992).
25. Templeton, A.R. 'Eve': hypothesis compatibility versus hypothesis testing. *Am. Anthropol.* **96**, 141–147 (1994).
26. Zischler, H., Geisert, H., von Haeseler, A. & Pääbo, S. A nuclear 'fossil' of the mitochondrial D-loop and the origin of modern humans. *Nature* **378**, 489–492 (1995).
27. Monckton, D.G. *et al.* Minisatellite mutation rate variation associated with a flanking DNA sequence polymorphism. *Nature Genet.* **8**, 162–170 (1994).
28. Sajantila, A. *et al.* Genes and languages in Europe: an analysis of mitochondrial lineages. *Genome Res.* **5**, 42–52 (1995).
29. Flint, J., Boyce, A.J., Martinson, J.J. & Clegg, J.B. Population bottlenecks in Polynesia revealed by minisatellites. *Hum. Genet.* **83**, 257–263 (1989).