

High resolution of human evolutionary trees with polymorphic microsatellites

A. M. Bowcock*, A. Ruiz-Linares†, J. Tomfohrde*, E. Minch‡, J. R. Kidd‡ & L. L. Cavalli-Sforza†

* Department of Pediatrics, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, Texas 75235-8591, USA

† Department of Genetics, Stanford University, Stanford, California 94305, USA

‡ Department of Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06510, USA

GENETIC variation at hypervariable loci is being used extensively for linkage analysis¹ and individual identification², and may be useful for inter-population studies²⁻⁵. Here we show that polymorphic microsatellites (primarily CA repeats) allow trees of human individuals to be constructed that reflect their geographic origin with remarkable accuracy. This is achieved by the analysis of a large number of loci for each individual, in spite of the small variations in allele frequencies existing between populations^{6,7}. Reliable evolutionary relationships could also be established in comparisons among human populations but not among great ape species, probably because of constraints on allele length variation. Among human populations, diversity of microsatellites is highest in Africa, which is in contrast to other nuclear markers and supports the hypothesis of an African origin for humans.

We examined 30 microsatellite loci in about ten individuals, from each of 14 indigenous populations chosen from the five continents. As a measure of similarity between the multiple locus genotype of two individuals, we calculated the proportion, P_s , of alleles that they share averaged over loci, where P_s is the number of shared alleles summed over loci/ ($2 \times$ number of loci compared). A distance measure between pairs of individuals was calculated as $(1 - P_s)$. A tree constructed from the pairwise inter-individual distances shows that individuals cluster according to their geographic origin (Fig. 1). Of the 148 individuals examined, 130 (87.8%) form discrete clusters that coincide with the continent of origin of the sample. The position of seven individuals (4.7%) is not clearly defined in the tree, and eleven individuals (7.4%) fall in a continental cluster that does not correspond to their geographic origin. Within continents these samples tend to form sub-clusters that correspond to their population of origin, with the exception of East Asian and some African populations. Nine out of 14 populations form sub-clusters that usually include more than 50% of the individuals from that population.

The tree shown in Fig. 1 contrasts sharply with trees derived from mitochondrial (mt)DNA in which it is rare for individuals of related geographic origin to form discrete clusters^{8,9}. The clustering pattern with microsatellites is not due to continent- or

population-specific alleles. Among the 342 alleles detected, there are 78 that are present in one continent only (43 in Africa, 14 in Oceania, 9 in East Asia, 6 in Europe, 6 in America), but they always have very low frequencies (4.1% on average). Furthermore, there is no indication of high continental differentiation: the average distance ($1 - P_s$) among individuals within the 14 populations is 0.64, versus an average of 0.73 among individuals from different continents, showing that the bulk of the genetic variation is within populations, as seen with other markers^{10,11}.

Construction of trees of individuals with samples of the loci indicates that there is no subset that is particularly informative. However, loci with higher F_{st} values (where for each locus, $F_{st} = \Sigma V_i / [\Sigma P_i(1 - P_i)]$; P_i , V_i are the mean and variance of gene frequencies of allele i summed over all L alleles at the locus) tend to produce trees where clusters are more defined. F_{st} is negatively correlated with locus heterozygosity ($r = -0.49$, $P < 0.01$), suggesting that loci with more extreme diversity and presumably higher mutation rates might on the whole be less informative.

Although most upper segments in the tree of Fig. 1 are very short, the various populations branch in an order similar to that seen in trees derived from classical markers or restriction-fragment length polymorphisms (RFLPs)^{12,13}. The first split separates Africans; this is followed by the close separation of Europeans, East Asians and Pacific populations. Finally, Americans branch off from Asian populations.

Population relationships were also examined using genetic distance measures based on allele frequencies. In the tree of populations (Fig. 2a) there is strong statistical support for the separation of African populations from other world populations (100% of bootstrap resamplings), in agreement with the hypothesis of an African origin for humans. The second split in this tree is supported by 84% of the bootstraps and separates Europeans from other populations, an observation that could reflect a real phylogenetic event, as suggested on the basis of non-DNA markers¹⁴, or could be due to Europeans having received genetic flow from both Africa and Asia¹³. Finally, with lesser statistical support (56% of bootstraps), Amerindians appear most closely related to East Asians, a classical observation but one that has been challenged recently¹⁴.

A subset of ten loci was also examined in the three primates closest to humans (chimpanzees, gorilla and orang). A genus tree constructed from the allele frequencies has very little structure, indicating that microsatellites do not provide adequate information for this type of comparison (Fig. 2b). Table 1 shows that the average length of alleles at the loci examined is quite similar for the four primates. Furthermore, although the non-human primate/human variance ratios are significantly higher than one for five loci, they are low considering that the divergence between primates took place at least 10 times (probably 100) earlier than the divergence among humans¹⁵. This suggests that there is a constraint on allele-length variation, as under an unconstrained

TABLE 1 Variation in allele length at ten microsatellite loci within humans and great apes

Locus	Mean allele length (base pairs)				Variance for non-humans	Variance for humans	F ratio* (d.f.) (non-human/human)
	Human	Chimp	Gorilla	Orang			
ACTC	83.1	78.0	75.9	75.7	21.8	470.88	0.31 (45,263)
D13S133	154.9	144.6	122.4	123.2	132.50	422.76	0.31 (45,265)
D13S137	111.5	105.6	113.5	99.4	93.61	48.53	1.93 (47,267)
D13S227	152.7	152.7	137.6	144.0	54.37	21.41	2.54 (47,259)
D13S193	136.6	134.3	140.6	116.0	123.40	47.37	2.61 (47,273)
O84XC5	83.4	81.5	81.6	87.4	10.19	23.34	0.44 (45,265)
D13S119	132.5	129.0	135.5	124.7	42.46	34.69	1.22 (43,283)
D13S118	193.1	191.5	190.6	190.8	15.17	12.05	1.26 (39,281)
D13S125	147.4	141.4	148.3	141.0	120.57	49.08	2.46 (45,269)
Utsw1523	179.1	175.8	174.9	173.6	18.62	5.23	3.56 (47,259)

* Critical values for F at $P = 5\%$ are between 1.44 and 1.49.

TABLE 2 Genetic variation of human populations detected with different genetic markers (mean heterozygosity $\times 100$)

	110 Classical polymorphisms*	Nuclear RFLPs†	Microsatellites	Microsatellites‡	mtDNA§
Africa	16.3 \pm 1.4	29.7 \pm 0.7	80.7 \pm 1.4	8.9 \pm 0.4	2.32
Europe	20.2 \pm 1.7	37.9 \pm 1.5	73.0 \pm 1.6	7.1 \pm 0.5	0.95
Asia	18.9 \pm 1.9	32.7 \pm 1.2	68.5 \pm 2.1	7.3 \pm 0.4	1.65
Oceania	13.7 \pm 1.8	27.5 \pm 1.2	63.6 \pm 1.8	7.4 \pm 0.5	
America	15.5 \pm 1.8	29.3 \pm 3.8	58.8 \pm 2.3	6.1 \pm 0.4	1.29

An analysis of variance shows significant differences between the heterozygosities of the five continents detected with microsatellites ($F_{4,142} = 20, P < 0.01$). The difference between Africa and Europe is also significant ($t_{57} = 8, P < 0.01$).

* Data compiled elsewhere¹⁵.

† The values for Africa (two populations), Europe (one population), Asia (two populations) and Oceania (three populations) are calculated from data on 79 RFLPs (ref. 23 and Lin *et al.*, manuscript in preparation). Values for America (two populations from North America and three populations from South America) are based on 35 other RFLPs (unpublished data).

‡ Number of alleles detected per locus. Mean number of individuals typed per locus per continent: Africa (27.6), Europe (26), Asia (28.2), Oceania (27.4), America (27.4).

§ Nucleotide diversity (heterozygosity at the nucleotide level) $\times 100$ of the D-loop region²⁴.

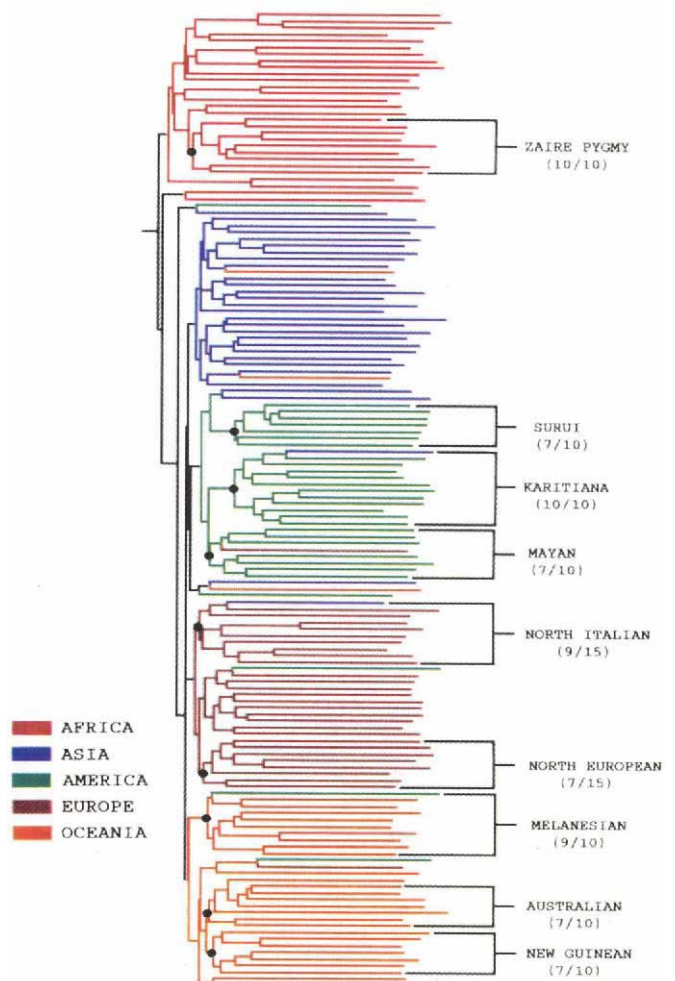
random walk the variance in allele size would increase linearly with time¹⁶. Although the details of the limitations to an unconstrained random walk for microsatellites, in which there is an absorbing or a reflecting barrier at the lower end^{17,18}, remain to be tested, it seems clear that there are limits to free variation.

In Table 2 the variation among human populations detected with microsatellites is compared with that detected with classical polymorphisms (blood groups and proteins), nuclear RFLPs and mtDNA. In the case of classical polymorphisms and nuclear RFLPs, Europeans show the highest level of variation, followed by East Asians, with Africans barely more heterogeneous than Australian aborigines or American Indians. By contrast, micro-

satellites show a significantly higher heterozygosity and number of alleles in Africa compared with other continents. Africa also has a higher diversity with mtDNA, but this finding is at present contested¹⁹. This observation supports the notion of an African origin for humans. A reasonable explanation for the discrepancy among different nuclear markers is the bias introduced by their initial selection in Europeans. This bias is likely to be less serious for markers with large numbers of alleles such as microsatellites.

The probability of distinguishing whether individuals belong to specific populations is probably enhanced in this material because the samples analysed here come from geographically discrete populations. This may generate apparent

FIG. 1 Neighbour-joining tree constructed from the pairwise distances between human individuals²⁵. A total of 148 samples from 14 human populations around the world were examined. The populations were as follows: three are from sub-Saharan Africa (Pygmies from Zaire and the Central African Republic, and the Lisongo population); two from Europe (North European and North Italian); three from East Asia (Chinese, Japanese and Cambodian); three from Oceania (Melanesian, New Guinean and Australian) and three from America (Maya from the Yucatan peninsula, Surui and Karitiana from the Amazon basin). A mean of 27.7 loci were typed on each individual (range 15–30), producing a unique multiple locus genotype for each individual. Distances between pairs of individuals were calculated from the proportion of alleles shared, as indicated in the text. There were no instances of pairs of individuals not sharing alleles. The loci typed by polymerase chain reaction (PCR) consisted of 29 CA repeats and one tetranucleotide repeat, all located on chromosomes 13 or 15. The exact location of two loci is not yet known, 8 other loci are located less than 1 cM away from their nearest neighbour and thus are likely to be in linkage disequilibrium, which with our small sample sizes is difficult to detect. The other 20 loci are located at an average distance of 8.8 cM from their nearest neighbour and presumably would have no detectable linkage disequilibrium except perhaps in very large samples. Methods for genotyping of microsatellites have been described²⁶. Coloured lines represent samples from different continents. Black dots indicate nodes that define population clusters. Numbers in parentheses are the fraction of individuals from the respective population found in a cluster. This tree and those shown in Fig. 2 were generated with programs in the PHYLIP package²⁷. The tree was rooted using midpoint rooting.



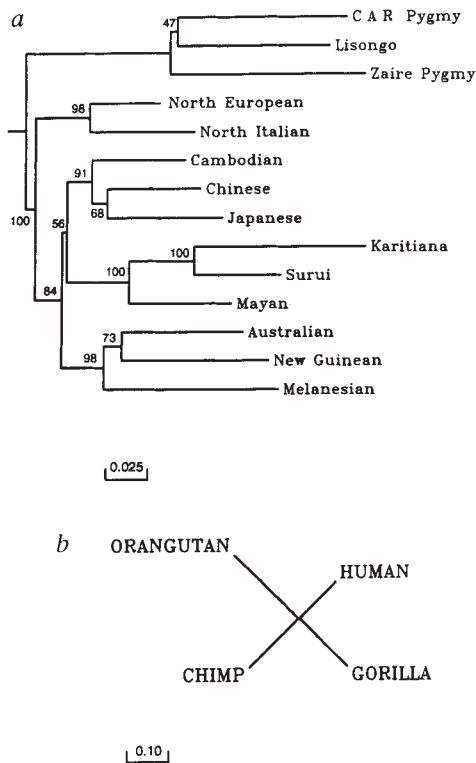


FIG. 2 *a*, Neighbour-joining tree relating the 14 populations examined. This tree was constructed using *f* (chord distance) as an estimate of genetic distance between populations²⁵. The numbers at the nodes are bootstrap values for 100 bootstrap resamplings²⁷ of the 30 loci typed. Other measures of genetic distance produced trees with identical topology but with slightly lower bootstrap values. The root of the tree was placed at the midpoint between the two most distantly related populations (Central African Republic (CAR) Pygmy and Karitiana). *b*, Neighbour-joining tree of the four primate species examined, based on chord distances between taxa (other distances produced the same results). This tree is based on a subset of ten loci typed in eight individuals (on average) for each of three other primate species (chimpanzees, gorillas and orangutans). All loci were polymorphic in all species, with the exception of locus *D13S193*, which was monomorphic in the orangutans tested. The scale is in distance units in both *a* and *b*.

discontinuity^{20,21} but, in general, human genetic geography shows high continuity^{15,22}. A geographically random sample would probably give a more complex and less discontinuous picture. Nevertheless, the results illustrate the resolving power of modern genetic analysis. □

Received 23 November 1993; accepted 24 January 1994.

- Weissenbach, J. et al. *Nature* **359**, 794–801 (1992).
- Pena, S. D. J., Chakraborty, R., Epplen, J. T. & Jeffreys, A. J. (eds) *DNA Fingerprinting: State of the Science* (Birkhauser, Basel, 1993).
- Gilbert, D. A., Lehman, N., O'Brien, S. J. & Wayne, R. K. *Nature* **344**, 764–767 (1990).
- Edwards, A., Hammond, H. A., Jin, L., Caskey, C. T. & Chakraborty, R. *Genomics* **12**, 241–253 (1992).
- Chakraborty, R., Deka, R., Jin, L. & Ferrell, R. E. *Am. J. Hum. Biol.* **4**, 387–397 (1992).
- Smouse, P. E., Spielman, R. S. & Park, M. H. *Am. Nat.* **119**, 445–463 (1982).
- Mitton, J. B. *Am. Nat.* **111**, 203–212 (1977).
- Cann, R. L., Stoneking, M. & Wilson, A. C. *Nature* **325**, 31–36 (1987).
- Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A. C. *Science* **253**, 1503–1507 (1991).
- Lewontin, R. *Evol. Biol.* **6**, 381–398 (1972).
- Nei, M. & Roychoudhury, A. *Science* **177**, 434–436 (1972).
- Cavalli-Sforza, L. L., Piazza, A., Menozzi, P. & Mountain, J. *Proc. natn. Acad. Sci. U.S.A.* **85**, 6002–6006 (1988).
- Bowcock, A. M. et al. *Proc. natn. Acad. Sci. U.S.A.* **88**, 839–843 (1991).
- Nei, M. & Roychoudhury, A. *Molec. Biol. Evol.* **10**, 927–943 (1993).
- Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *History and Geography of Human Genes* (Princeton Univ. Press, Princeton, in the press).
- Karlin, S. *A First Course in Stochastic Processes* (Academic, New York, 1969).
- Walsh, J. B. *Genetics* **115**, 553–567 (1987).
- Tachida, H. & Iizuka, M. *Genetics* **131**, 471–478 (1992).

- Templeton, A. R. *Am. Anthr.* **95**, 51–72 (1993).
- Barbujani, G. & Sokal, R. R. *Proc. natn. Acad. Sci. U.S.A.* **87**, 1816–1819 (1990).
- Zei, G. et al. *Ann. hum. Genet.* **57**, 123–140 (1993).
- Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *Science* **259**, 639–646 (1993).
- Bowcock, A. M. et al. *Gene Geog.* **5**, 151–173 (1991).
- Horai, S. et al. *Molec. Biol. Evol.* **10**, 23–47 (1993).
- Saitou, N. & Nei, M. *Molec. Biol. Evol.* **4**, 406–425 (1987).
- Bowcock, A. et al. *Genomics* **15**, 376–386 (1993).
- Felsenstein, J. *PHYLIP* (Phylogeny Inference Package) Version 3.5c (Department of Genetics, Univ. Washington, Seattle, 1993).
- Cavalli-Sforza, L. L. & Edwards, A. W. F. *Evolution* **32**, 550–570 (1967).

ACKNOWLEDGEMENTS. This work was supported by grants from the NIH (to L.C.S. and K. K. Kidd), the Wenner Gren Foundation (L.C.S.), and the NSF (K. K. Kidd). B. Hewlett helped in obtaining the African cell lines. The North Italian cell lines were made by G. B. Ferrara in Genova with blood samples from Associazione Donatori di Sangue, Bergamo. DNA samples from New Guinea and Australia were provided by the late Allan C. Wilson. We thank the Lucille P. Markey Charitable Trust for funding (to L.C.S.) travel to Africa to collect samples, C. Dunn for technical help, and C. Campbell with help in preparation of the manuscript.

Bcl-2 expression promotes B- but not T-lymphoid development in *scid* mice

Andreas Strasser, Alan W. Harris, Lynn M. Corcoran & Suzanne Cory*

The Walter and Eliza Hall Institute of Medical Research, Post Office, Royal Melbourne Hospital, Victoria 3050, Australia

EXPRESSION of antigen receptors is vital for the development of B and T lymphocytes. In mice with the *scid* mutation^{1,2}, which are unable to make productive rearrangements of their immunoglobulin and T-cell receptor (TCR) genes, lymphopoiesis aborts at an early stage. The death of the immature lymphocytes by apoptosis³ is postulated to result from a failure to receive a survival signal induced by receptor engagement⁴. Consistent with this hypothesis, introduction of immunoglobulin or TCR transgenes into *scid* mice promoted an increase in B- or T-lymphoid cells, respectively^{5–7}. As the protein encoded by the *bcl-2* gene can inhibit cell death^{8,9}, we tested whether lymphopoiesis could be rescued in *scid* mice by crossing in a *bcl-2* transgene. Strikingly, the *bcl-2/scid* mice accumulated almost normal numbers of B-lymphoid cells which lacked surface immunoglobulin but expressed markers of maturity. T-cell development remained blocked. Introducing a TCR transgene enabled *bcl-2/scid* mice to develop normal numbers of CD4⁺8⁺ thymocytes even in the absence of immunological selection, suggesting that T cells become competent to respond to *bcl-2* protein only after the TCR complex is displayed at the cell surface.

The *bcl-2* protein (Bcl-2) is known to enhance the survival capacity of B and T cells at several stages of differentiation^{10–13}, but its effect in pro-B and pro-T cells has not previously been addressed. We introduced the Eμ-*bcl-2*-36 transgene, which is expressed in both lymphoid lineages¹², into *scid* mice by appropriate genetic crosses. The *scid* mice had the normal low number of pro-B and pro-T cells and essentially no mature lymphoid cells^{5,14}, whereas the *bcl-2/scid* mice had large numbers of cells expressing the B-lineage marker CD45R(B220) in the bone marrow, spleen and blood (Table 1), the total number approaching that found in normal mice. The increase was not due to augmented proliferation, because most of the cells were small and quiescent. Neither was it due to increased 'leakiness' of the *scid* mutation², because no immunoglobulin-bearing cells were detectable in the bone marrow (<10⁵ per femur), spleen (<10⁶ per organ), or peripheral blood (<3 × 10⁴ ml⁻¹) (panels I in Fig. 1a and b).

Unexpectedly, most B220⁺ cells in the *bcl-2/scid* mice had developed beyond the pro-B stage. Pro-B and early pre-B cells display various combinations¹⁵ of the markers Thy-1, S7(CD43),

* To whom correspondence should be addressed.