



ARTICLE

Alu insertion polymorphisms and the genetic structure of human populations from the Caucasus

Ivane Nasidze^{*1}, Gregory M Risch^{2,3,4}, Myles Robichaux^{2,3,4}, Stephen T Sherry^{2,3,4,5}, Mark A Batzer^{2,3,4} and Mark Stoneking¹

¹Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany; ²Department of Pathology, Stanley S. Scott Cancer Center, Neuroscience Center of Excellence, Louisiana State University Health Sciences Center, New Orleans, Louisiana, LA 70112 USA; ³Department of Biometry and Genetics, Stanley S. Scott Cancer Center, Neuroscience Center of Excellence, Louisiana State University Health Sciences Center, New Orleans, Louisiana, LA 70112 USA; ⁴Department of Biochemistry and Molecular Biology, Stanley S. Scott Cancer Center, Neuroscience Center of Excellence, Louisiana State University Health Sciences Center, New Orleans, Louisiana, LA 70112 USA; ⁵National Center for Biotechnology Information, National Institutes of Health, 8600 Rockville Pike, Bethesda, Maryland, MD 20894, USA

An analysis of 8 *Alu* insertion loci (*ACE*, *TPA25*, *PV92*, *APO*, *FXIIB*, *D1*, *A25*, *B65*) has been carried out in six populations from the Caucasus, including Indo-European-speaking Armenians; Altaic-speaking Azerbaijani; North Caucasian-speaking Cherkessians, Darginians, and Ingushians; and South Caucasian (Kartvelian)-speaking Georgians. The Caucasus populations exhibit low levels of within-population variation and high levels of between-population differentiation, with the average F_{st} value for the Caucasus of 0.113, which is almost as large as the F_{st} value of 0.157 for worldwide populations. Maximum likelihood tree and principal coordinate analyses both group the Caucasus populations with European populations. Neither geographic nor linguistic relationships appear to explain the genetic relationships of Caucasus populations. Instead, it appears as if they have been small and relatively isolated, and hence genetic drift has been the dominant influence on the genetic structure of Caucasus populations. *European Journal of Human Genetics* (2001) 9, 267–272.

Keywords: *Alu* insertion; polymorphisms; Caucasus populations

Introduction

The Caucasus (the region situated on the border between Europe and west Asia, between the Black and Caspian Seas) exhibits a high degree of linguistic diversity, with four linguistic families (North Caucasian, South Caucasian, Indo-European, and Altaic) spoken by approximately 50 autochthonous groups. A major geographic barrier, the Caucasus Mountains, separates the northern and southern sub-regions of the Caucasus. The Caucasus is therefore an

interesting region in which to examine the relative influence of linguistic diversity and geographic barriers on the genetic structure of populations.

A number of 'classical' genetic markers (blood groups, serum proteins, and red cell enzymes) have been extensively studied in the Caucasus.^{1,2} Based on these studies, the level of genetic differentiation in the region exceeds that for the rest of Europe, and is only slightly lower than that for all of Asia.^{1–3} Overall, the results of these studies suggest a single ancient origin of Caucasus populations, with subsequent significant subdivision along geographic and linguistic boundaries.

An analysis of mitochondrial DNA (mtDNA) hypervariable region 1 (HVR 1) sequence variability among different groups from the Caucasus showed a similar picture of substantial diversity within and among Caucasus populations, higher

*Correspondence: Dr Ivane Nasidze, Max Planck Institute for Evolutionary Anthropology, Inselstrasse 22, D-04103 Leipzig, Germany. Tel: +49 341 9952505; Fax: +49 341 9952555; E-mail: nasidze@eva.mpg.de
Received 30 August 2000; revised 13 November 2000; accepted 17 November 2000

than for European groups, and slightly lower than for Middle Eastern populations (Nasidze and Stoneking, submitted). The genetic relationships among Caucasus populations, revealed by mtDNA analysis, primarily reflect geographic rather than linguistic relationships. In particular, both Indo-European speaking Armenians and Altaic-speaking Azerbaijanians group with their geographic neighbours in the Caucasus, rather than with their linguistic neighbours (other Indo-European/Altaic-speaking populations), suggesting language replacements involving the Armenian and Azerbaijanian languages. Remarkably, the genetic distance between pairs of populations was significantly correlated with geographic distance, despite the presence of the Caucasus Mountains as a potentially significant barrier to gene flow.

mtDNA is maternally inherited and evolves rapidly. mtDNA variation therefore reflects the maternal history of a population, and the rapid evolutionary rate means that factors causing population differentiation can be identified even on a recent time scale. However, mtDNA is a single genetic locus; multiple loci must be studied to accurately determine the genetic structure and relationships of populations. To that end, we have analysed the genetic diversity associated with several *Alu* insertion polymorphisms in populations from the Caucasus.

Alu insertion polymorphisms are autosomal markers that thus reflect both the maternal and paternal history of a population. Moreover, they are stable markers that reflect unique evolutionary events, namely the insertion of an *Alu* element into a new chromosomal location.^{4–11} Another advantage of these markers is that there is no parallel gain or loss of *Alu* elements at a particular chromosomal location, and thus all chromosomes that carry a polymorphic *Alu* element are identical by descent. Furthermore, the ancestral state of these polymorphisms is known to be the absence of the *Alu* insert, and this information can be used in analyses of population relationships.^{6,7,10} These features make *Alu* insertion polymorphisms more attractive than other autosomal markers (such as classical markers) for analyses of population history and structure.

We present here an analysis of eight *Alu* insertion polymorphisms in six populations from the Caucasus. Among the questions we wish to address are the following: (i) are Caucasus populations genetically heterogeneous; (ii) does genetic differentiation correlate with linguistic differentiation; and (iii) do major geographic boundaries, such as the Caucasus Mountains, demonstrably influence patterns of genetic variation and differentiation in the Caucasus?

Materials and methods

A total of 221 blood samples from healthy, unrelated individuals were collected from the following six autochthonous groups of the Caucasus (Figure 1): Georgians (South Caucasian linguistic family), Armenians (Indo-European linguistic family), Azerbaijanians (Turkic branch of the Altaic

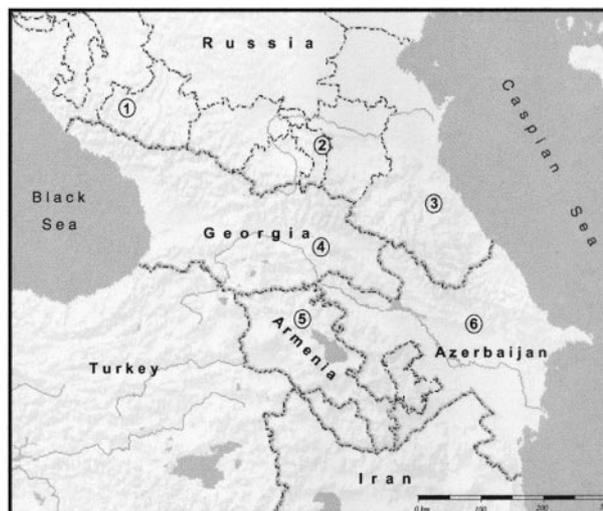


Figure 1 Map of the Caucasus with the locations of the studied populations: 1 – Cherkessians, 2 – Ingushians, 3 – Darginians, 4 – Georgians, 5 – Armenians, and 6 – Azerbaijanians.

linguistic family) and Cherkessians, Ingushians, and Darginians (North Caucasian linguistic family). Total genomic DNA was extracted from whole blood using the IsoQuick DNA extraction kit (Orca Research Inc., Bothell, WA, USA).

In order to minimise the amount of extracted DNA used in subsequent experiments, genomic DNA was amplified using a degenerate oligonucleotide primer (DOP) PCR as described previously.¹² Five microlitres of the DOP amplification product (out of a total volume of 50 μ l) were used for further genotyping of eight *Alu* insertion loci (*ACE*, *TPA25*, *PV92*, *APO*, *FXIIIIB*, *D1*, *A25*, and *B65*), as described previously.^{7,13,14} The consistency of results from genomic DNA and DOP products was verified by typing both genomic DNA and DOP products from several individuals.

Allele frequencies were calculated by direct counting and Hardy-Weinberg equilibrium was evaluated by an exact test¹⁵. Pairwise F_{st} values and corresponding P values were computed between each pair of populations.¹⁶ To create an overall test for the significance of pairwise F_{st} values, we combined the separate probability values for each locus using the formula $P = -2 \sum \ln p_i$, where p_i is the separate probability value associated with the F_{st} value for each locus.¹⁷

The maximum likelihood method was used to construct a tree depicting population relationships, using the PHYLIP 3.5c package.¹⁸ Population relationships were also analysed via principal coordinate analysis (PCA), as implemented in STATISTICA (StatSoft Inc. 1995).

The relative amount of gene flow for each population was assessed by plotting the distance from the centroid *vs* the heterozygosity, as described previously.¹⁹ According to this model, a simple linear relationship is expected between the heterozygosity of a population and the distance of that population from the centroid, which is defined as the

arithmetic mean of the allele frequencies. This analysis was carried out as described previously.^{7,10}

The significance of the correlation coefficient between genetic distance and geographic distance, and between genetic distances based on *Alu* insertion loci and those based on mtDNA, was estimated by 1000 matrix permutations, using the program *Permute!3.4*.²⁰

Hierarchical analysis via the analysis of molecular variance (AMOVA),²¹ grouping populations according to their geographic location or to linguistic affiliation (using different linguistic classifications), was carried out by means of programs in *Arlequin 2.0*.²²

Results

The allele frequencies for eight *Alu* insertion loci in the six Caucasus populations are shown in Table 1. All loci were polymorphic in all groups, with the exception of *FXIII B* and *D1* in Ingushians and *A25* in Azerbaijanians, which were fixed for the absence of the *Alu* element.

Four out of 48 tests for Hardy-Weinberg equilibrium showed a significant departure from equilibrium (*TPA* in Georgians, *FXIII B* in Azerbaijanians, and *B65* in Cherkessians and Ingushians). Since none of the deviations are assigned to a particular locus or population, they probably represent random statistical fluctuations. The average heterozygosity for each locus ranged from 0.09 to 0.47, which is close to the maximum possible heterozygosity of 0.5 for a bi-allelic locus. The average heterozygosity for each population (Table 1) was lower in Ingushians (0.209) than in the other populations (range 0.318–0.394).

A test for homogeneity of allele frequencies across populations indicates substantial heterogeneity among Caucasus populations ($\chi^2=331.93$, $df=35$, $P<0.001$).

The F_{st} values are also substantial (Tables 1 and 2), ranging from 0.018 (*APO* and *A25*) to 0.222 (*FXIII B*). The average F_{st} value for the Caucasus is almost seven times higher than in Europe and twice-that in west Asia (Table 2). This high F_{st} value indicates large differences in the *Alu* insertion frequencies among Caucasus populations.

In order to examine population relationships, the Caucasus groups were compared with a worldwide data set typed for the same eight *Alu* insertion loci.¹⁰ We first constructed a maximum-likelihood tree, rooted with a hypothetical ancestral population that lacked all eight *Alu* elements (Figure 2A). The root resides between African and non-African populations, with Australian and New Guinea populations branching off next. The Caucasus groups form a basal cluster of the Eurasian branch of the tree, with Georgians grouping among the European populations.

In order to further assess population relationships, a principal coordinate (PC) analysis was carried out (Figure 2B). The first two principal components account for 82.07% of the variance observed (61.07 and 21% respectively). The first PC separates the African populations from non-Africans, while the second PC shows differences among non-African populations. In particular, the Caucasus groups are scattered amongst a more compact grouping of European populations, indicating the higher genetic diversity in the Caucasus.

An analysis of molecular variance (AMOVA) shows that when all populations were treated as single groups, 92% of the total genetic variance was within populations and 8% was between populations. When individuals were classified into

Table 2 Average heterozygosity and F_{st} values for Caucasus populations, compared to other geographic regions of the World. Data for the other regions are from Stoneking *et al*.¹⁰

Region	No. of populations	No. of individuals	Heterozygosity	F_{st}
Africa	6	176	0.402	0.086
Americas	4	184	0.381	0.038
Europe	7	334	0.396	0.017
Sahul	3	185	0.308	0.105
Southeast Asia	7	359	0.377	0.067
Western Asia	7	262	0.414	0.053
Caucasus	6	221	0.311	0.113
World	40	1721	0.426	0.157

Table 1 *Alu* insertion frequencies, heterozygosity, and F_{st} values for the Caucasus populations. For each locus, the frequency of the allele with the *Alu* insert is given

Population	Average n*	ACE	TPA25	PV92	APO	FXIII B	D1	A25	B65	Average h
Armeians	40	0.477	0.43	0.013	0.871	0.343	0.151	0.058	0.453	0.327
Azerbaijanians	34	0.216	0.513	0.382	0.943	0.1	0.333	0	0.697	0.318
Cherkessians	40	0.39	0.386	0.167	0.932	0.439	0.167	0.045	0.651	0.394
Darginians	16	0.167	0.361	0.167	0.864	0.143	0.346	0.028	0.321	0.327
Georgians	67	0.354	0.493	0.25	0.934	0.61	0.418	0.088	0.727	0.378
Ingushians	24	0.34	0.224	0.129	0.941	0	0	0.067	0.21	0.209
Average h per locus		0.422	0.467	0.28	0.158	0.321	0.326	0.091	0.429	
F_{st} per locus		0.058	0.045	0.057	0.018	0.222	0.126	0.018	0.207	

*Average number of individuals typed for each locus.

South and North Caucasus groups, 91.4% of the total variance was within populations and 8.6% was between groups. The fraction of the genetic variance determined by differences between populations grouped according to two different linguistic classifications (Indo-European/Altaic/Caucasian; Indo-European/Altaic/North Caucasian/Kartvelian) was similar. When the Caucasus populations were

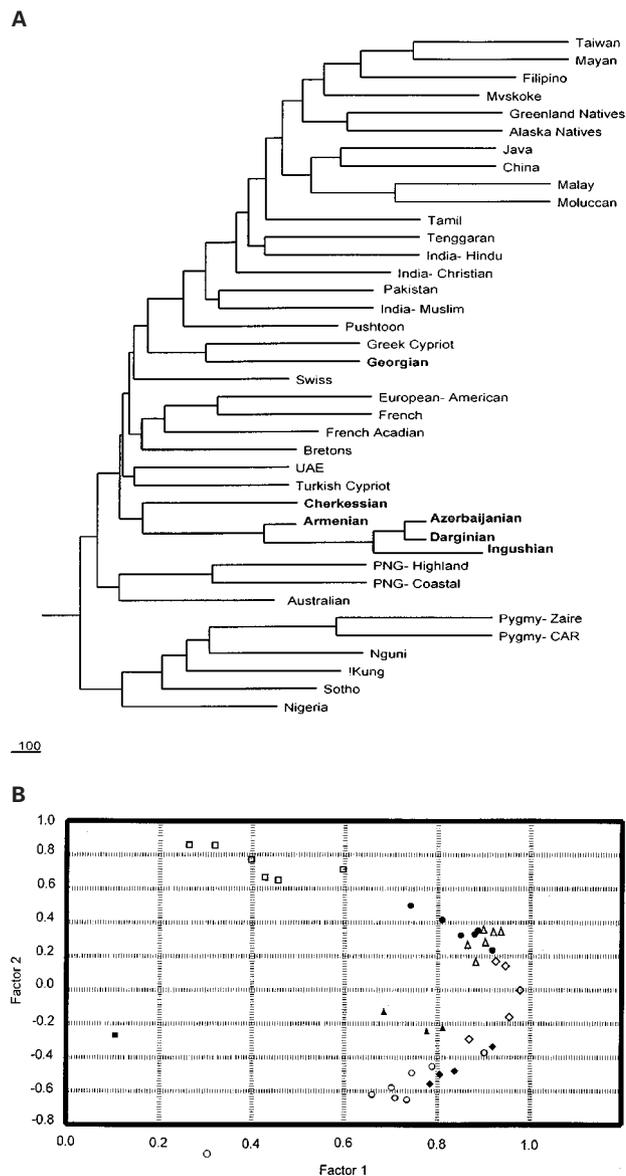


Figure 2 Maximum likelihood tree (A) and PC coordinate plot (B) showing relationships among the Caucasus and Worldwide populations. Symbols on the PC plot correspond to (■) hypothetical ancestral population with absence of *Alu* element insertions; (□) African populations; (●) Caucasus populations; (△) European groups; (◇) West Asian populations; (○) East Asian groups; (▲) Australians and PNG; (◆) New World populations.

compared with European and west Asian populations, the within-population proportion of the total variance was even higher (96.4%). Only 0.8% of the total genetic variance was observed between geographic regions, while 2.8% of the genetic variance was described by differences among populations within geographic groups.

To examine the relative amount of gene flow experienced by Caucasus, European, and west Asian populations, we plotted the distance from the centroid vs heterozygosity (Figure 3), according to the model of Harpending and Ward.¹⁹ Strikingly, all of the Caucasus groups showed less heterozygosity than predicted by the Harpending-Ward model, which suggests that the Caucasus groups have either had a smaller effective population size, or they have been more isolated.

Previously, we showed that F_{st} genetic distances between ten Caucasus groups, based on mtDNA, were significantly correlated with geographic distance (Nasidze and Stoneking, submitted). For the six Caucasus groups studied here, the *Alu* based F_{st} genetic distances exhibit a high (but non-significant) correlation with the mtDNA F_{st} distances ($r=0.65$, $P=0.075$) and a lower, non-significant correlation with geographic distance ($r=-0.36$, $P=0.24$). This may reflect sampling of fewer groups for the *Alu* insertion loci, as the correlation between mtDNA F_{st} distance and geographic distance for the six Caucasus groups analysed for the *Alu* insertion loci was also non-significant ($r=-0.07$, $P=0.58$).

Discussion

The Caucasus populations exhibit, on average, less variability than other populations for the eight *Alu* insertion polymorphisms analysed here. The average heterozygosity is less

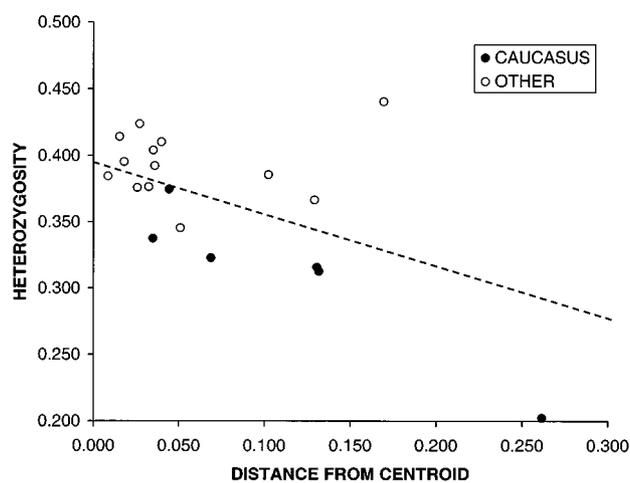


Figure 3 Plot of heterozygosity vs distance from the centroid for Caucasus, European, and western Asian populations. The data for European and western Asian populations are from Stoneking *et al.*¹⁰

than that for any other region of the world, with the exception of Sahul (Table 2). Within the Caucasus, Ingushians have much lower levels of variability than any of the other populations. The Ingushians also showed unusual patterns of mtDNA variation when compared with other Caucasus populations (Nasidze and Stoneking, submitted), which indicates that some feature of the Ingushian population history, or of this particular sample of Ingushians, must be responsible for their different patterns of genetic variation at both mtDNA and the *Alu* insertion loci.

In contrast to the low levels of variation within populations, the Caucasus exhibits very high levels of between-population differentiation. The average F_{st} value among Caucasus populations (0.113) exceeds that of any other geographic region of the world (Table 2) and is comparable to the average value among worldwide populations. The large genetic distances among Caucasus populations are also evident in the genetic tree and PC plots (Figure 2). The picture that thus emerges of Caucasus population structure is one of low variability within populations but high differentiation between populations, indicative of small, isolated populations. Consistent with this view is the analysis of heterozygosity vs distance from the centroid (Figure 3), in which all of the Caucasus populations deviate from the expectations of the Harpending-Ward model by having less heterozygosity than predicted by the model.

Given the geographic location of the Caucasus, intermediate between Europe and west Asia, it is of interest to determine whether European or western Asian populations are more similar to Caucasus populations. Based on the *Alu* insertion loci, the Caucasus populations are most similar to European populations. This can be seen in the tree and PC plots (Figure 2), as well as in a comparison of average pairwise F_{st} values: the average F_{st} between Caucasus-Europe was 0.061, which is lower than that between Caucasus-West Asia ($F_{st}=0.092$). A closer similarity between Caucasus and European populations, as opposed to Caucasus and Middle Eastern populations, was also found in an analysis of mtDNA variation (Nasidze and Stoneking, submitted). However, it is important to note that with the *Alu* insertion polymorphisms there are limited genetic data from Middle Eastern populations, thereby prohibiting a strict comparison between the mtDNA and *Alu* insertion loci results.

A primary goal of this study was to examine the relative influence of geographic vs linguistic diversity on the genetic structure of Caucasus populations. We do not find any significant correspondence between either geography or linguistic diversity and the genetic relationships of Caucasus populations. The within- and between-group components of the total variance do not change when various linguistic groupings are analysed via AMOVA; about 92% of the genetic variance is within populations and about 8% is between populations. With respect to geography, there is no correlation between the genetic distance and the geographic distance separating pairs of Caucasus populations. Nor is

there any evidence that the Caucasus mountains have been a significant factor influencing Caucasus population structure; the average F_{st} value between populations from the northern and southern Caucasus sub-regions is actually slightly lower ($F_{st}=0.078$) than the average values within the northern sub-region ($F_{st}=0.097$) and within the southern sub-region ($F_{st}=0.101$). The lack of correlation of genetic relationships with either geographic or linguistic factors further supports the influence of small population size and genetic drift on the genetic structure of Caucasus populations.

Our previous analysis of mtDNA variation in the Caucasus (Nasidze and Stoneking, submitted) demonstrated that Armenians (who speak an Indo-European language) grouped genetically with their geographic neighbours in the Caucasus, and not with other Indo-European speaking populations. These results led us to suggest that the Armenian language was introduced via a language replacement event, as some linguists have hypothesised.²³ Based on the *Alu* insertion loci, the Armenians do exhibit a slightly lower F_{st} value with respect to Indo-European speaking groups ($F_{st}=0.065$) than with respect to their geographic neighbours in the Caucasus ($F_{st}=0.073$). However, this difference in F_{st} values is not statistically significant, and moreover the tree and PC plots (Figure 2) group the Armenians with other Caucasus populations. Thus, the *Alu* insertion loci also support a language replacement involving the Armenian language, as genetically Armenians are more similar to their geographic neighbours than to their linguistic neighbours.

In conclusion, an analysis of eight *Alu* insertion loci indicates low levels of variability within Caucasus populations, but large genetic differences between populations. The genetic structure of Caucasus populations reflects neither geographic nor linguistic relationships; overall, these results indicate that small population sizes and isolation, leading to genetic drift, have been the primary factors influencing Caucasus population structure. Moreover, there is broad agreement between the *Alu* insertion loci and mtDNA results (Nasidze and Stoneking, submitted).

Future work on the genetic structure of Caucasus populations will focus on additional *Alu* insertion loci and L1 insertion loci,²⁴ to verify the conclusions obtained in the present study, as well as Y-chromosome polymorphisms. The Y-chromosome data will yield insights into the male demographic history of the Caucasus, in comparison with the female demographic history revealed by mtDNA analysis (Nasidze and Stoneking, submitted). An analysis of the Y chromosome based diversity should be of particular interest for the question of language replacements in the Caucasus, since language replacements are frequently male-driven.²³

Acknowledgments

This research was supported by grants from the National Science Foundation (M Stoneking), Louisiana Board of Regents Health Excellence Fund HEF (2000-05)-05 (MA Batzer) and award number

1999-IJ-CX-K009 from the Office of Justice Programs, National Institute of Justice, Department of Justice (MA Batzer). Points of view in this document are those of the authors and do not necessarily represent the official position of the US Department of Justice.

References

- 1 Nasidze I: Genetic polymorphisms of the Caucasus ethnic groups: distribution of some serum protein and red cell enzyme genetic markers. *Gene Geography* 1995; **9**: 91–116.
- 2 Nasidze I: Genetic polymorphisms of the Caucasus ethnic groups: distribution of some blood groups. *Gene Geography* 1995; **9**: 117–167.
- 3 Barbujani G, Nasidze I, Whitehead GN: Genetic diversity in the Caucasus. *Hum Biol* 1994; **66**: 639–668.
- 4 Batzer MA, Deininger PL: A human-specific subfamily of Alu sequences. *Genomics* 1991; **9**: 481–487.
- 5 Batzer MA, Gudi V, Mena JC, Foltz DW, Herrera RJ, Deininger PL: Amplification dynamics of human-specific (HS) Alu family members. *Nucleic Acids Res* 1991; **19**: 3619–3623.
- 6 Batzer MA, Stoneking M, Alegria-Hartman M *et al*: African origin of human-specific polymorphic Alu insertions. *Proc Natl Acad Sci USA* 1994; **91**: 12288–12292.
- 7 Batzer MA, Arcot SS, Phinney JW *et al*: Genetic variation of recent Alu insertions in human populations. *J Mol Evol* 1996; **42**: 22–29.
- 8 Deininger PL, Batzer MA: Evolution of Retroposons; in Heckht MK (ed): *Evolutionary Biology*. New York, 1993, pp 157–196.
- 9 Deininger PL, Batzer MA: SINE master genes and population biology; in Marais RJ (ed): *The Impact of Short Interspersed Elements (SINEs) on the Host Genome*. Georgetown, Texas, R.G. Landes Publishers, 1995, pp 43–60.
- 10 Stoneking M, Fontius JJ, Clifford SL *et al*: Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. *Genome Res* 1997; **7**: 1061–1071.
- 11 Perna NT, Batzer MA, Deininger PL, Stoneking M: Alu insertion polymorphism: A new type of marker for human population studies. *Hum Biol* 1992; **64**: 641–648.
- 12 Cheung V, Nelson S: Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be amplified on less than one nanogram of genomic DNA. *Proc Natl Acad Sci USA* 1996; **93**: 14676–14679.
- 13 Arcot SS, Fontius JJ, Deininger PL, Batzer MA: Identification and analysis of ‘young’ polymorphic Alu element. *Biochim Biophys Acta* 1995; **1263**: 99–102.
- 14 Arcot SS, Wang Z, Weber JL, Deininger PL, Batzer MA: Alu repeats: a source for the genesis of primate microsatellites. *Genomics* 1995; **29**: 136–144.
- 15 Guo S, Thompson E: Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* 1992; **48**: 361–372.
- 16 Weir BS, Cockerham CC: Estimating F-statistics for the analysis of population structure. *Evolution* 1984; **38**, 1358–1370.
- 17 Sokal RR, Rohlf FJ: *Biometry*. New York, Freeman and Company, 1997.
- 18 Felsenstein J: PHYLIP—Phylogeny Inference Package. *Cladistics* 1989; **5**: 164–166.
- 19 Harpending HC, Ward RH: Chemical systematics and human population; in Nitecki M (ed): *Biochemical Aspects of Evolutionary Biology*. Chicago, IL. University of Chicago Press, 1982, pp 213–252.
- 20 Legendre P, Lapointe F-J, Casgrain P: Modeling brain evolution from behavior: a permutational regression approach. *Evolution* 1994; **48**: 1487–1499.
- 21 Excoffier L, Smouse PE, Quattro JM: Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 1992; **131**: 479–491.
- 22 Schneider S, Roessli D, Excoffier L: Arlequin ver. 2.000: A software for population genetics data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland, 2000.
- 23 Renfrew C: *Archaeology and language: The puzzle of Indo-European origins*. London, Cape, 1987.
- 24 Sheen F, Sherry ST, Risch GM *et al*: Reading between the LINES: Human genomic variation induced by LINE-1 retrotransposition. *Genome Res* 2000; **10**: 1496–1508.