Journal of Human Genetics (2009) 54, 603–611

www.nature.com/jhg

ORIGINAL ARTICLE

Insights on human evolution: an analysis of *Alu* insertion polymorphisms

Maria C Terreros[1,5], Miguel A Alfonso-Sánchez[2,5], Gabriel E Novick[1,5], Javier R Luis[3], Harlette Lacau[1], Robert K Lowery[1,4], Maria Regueiro[1] and Rene J Herrera[1]

We analyzed the genetic profile of 563 individuals from 12 geographically targeted human populations from Europe, Asia and Africa using 27 human-specific polymorphic *Alu* insertions. Phylogenetic analyses indicated a clear correspondence between genetic profiles and historical patterns of gene flow and genetic drift. Sub-Saharan African populations (Benin, Cameroon, Kenya and Rwanda) formed a visibly differentiated cluster, indicating the role of the Sahara desert as a strong natural barrier to gene flow. Moreover, a higher than expected genetic affinity between populations from Europe, North Africa and Asia was detected, probably reflecting the homogenizing effects of bidirectional migratory processes between Eurasia and North Africa during the Plio-Pleistocene and Neolithic periods or the insensitivity of these markers in discriminating between these groups. The Ami aborigines of Formosa present a distinctive degree of genetic uniqueness from all the other groups, consistent with a pattern of isolation by distance, small population size and, accordingly, substantial genetic drift. We further tested all 27 *Alu* loci for their potential usefulness as ancestry informative markers (AIMs). On the basis of differences between weighted allelic frequencies (δ-values) and F_{ST} values, we propose that 11 of the 27 *Alu* elements could be useful as part of the current AIM panels to assess phylogenetic relationships.
Journal of Human Genetics (2009) **54,** 603–611; doi:10.1038/jhg.2009.86; published online 11 September 2009

Keywords: ancestry informative markers; genetic structuring; gene flow; human evolution; isolation by distance model; PAIs

INTRODUCTION

The *Alu* family of repetitive elements was originally defined as a fraction of renatured repetitive DNA that was distinctively cleaved with the restriction enzyme *Alu*I.[1] *Alu* elements are derived from the 7SL RNA gene whose transcript is an essential constituent of the endoplasmic reticulum signal recognition particle. *Alus* share about 90% sequence homology with the 7SL RNA.[2] They represent close to 11% of the human genome, and they are present in excess of 1 000 000 copies per haploid genome with an average distribution of one copy every 4 kb.[3]

Alu elements are dimeric sequences[4] in which the left half contains the typical internal RNA polymerase III split promoter.[5] They possess a short A-rich linker between the two dimers and a 3′-oligo-dA-rich tail up to 100 bp in length (dependent on the locus) characteristic of all SINEs.[6] *Alu* repeats are characteristically flanked by direct repeats derived by duplication of target sequences at the site of integration. These elements mobilize by retroposition through an RNA polymerase III transcript intermediate[7] and their chromosomal distribution shows a certain preference for R bands or AT-rich areas.[8] *Alus* are thought to retrotranspose using an L1-encoded reverse transcriptase.[9] The origin of *Alu* elements can be traced to the radiation of primates some 65 million years ago,[10] and only a few source genes, termed 'master genes,' are still undergoing amplification at a rate of approximately 8×10^{-3} *de novo Alu* insertions per year.[11,12]

Owing to changes in the master *Alu* genes during evolution, several families and subfamilies have been generated[13–16] that are classified as Old (Jo and Jb subfamilies), Intermediate and Young (Y). The Y family contains approximately 100 000 members. Of them, between 500 to 2000 copies belong to several closely related Y subfamilies (Yc1, Yc2, Ya5, Ya8, Yb8 and Yb9) containing almost all recently inserted, unfixed, human-specific *Alu* members, not found at orthologous positions in the genomes of the great apes.[15,17–21] Insertion polymorphisms of these subfamilies exhibit a biallelic, codominant pattern of insertion–lack of insertion inheritance reflecting common ancestry, the absence of the insertion being the ancestral state.[15,22,23] Therefore, *Alu* insertions shared by different individuals are identical by descent not just by state. This means that if two individuals share an insertion, it is most likely that they share a common ancestor in whom the insertion took place. Furthermore, there is no known mechanism for the complete and specific removal of an element;[24] therefore, the lack of insertion in all likelihood represents the ancestral state. Also, as the rate of insertion and fixation of new *Alu* elements is

[1]College of Medicine, Florida International University, Miami, FL, USA; [2]Servicio de Investigación Genómica: Banco de ADN, Facultad de Farmacia, Universidad del País Vasco (UPV/EHU), Alava, Spain; [3]Departamento de Bioquimica, Xenetica e Inmunoloxia. Universidade de Vigo, Vigo, Spain and [4]Department of Biological Sciences, Florida International University, FL, USA
[5]These authors contributed equally to this work.
Correspondence: Dr RJ Herrera, College of Medicine, Florida International University, University Park, OE 304, Miami, FL 33199, USA.
E-mail: herrerar@fiu.edu
Received 24 April 2009; revised 4 August 2009; accepted 5 August 2009; published online 11 September 2009

about 100–200 per Myr,[13,25] the probability of two independent *Alu* elements being inserted by chance in the same genomic location is virtually nil.[22,26]

These three properties of recently inserted polymorphic *Alu* insertions (PAIs), identity by descent, lack of insertion as the ancestral state and no known mechanism for their precise and complete removal, make them unique markers to investigate human evolution and conduct population genetic studies.

The insertion of new *Alu* elements has been a constant process during the evolution of the human lineage, and depending on the evolutionary age of the insertion, different *Alu* polymorphisms offer different perspectives on the history of human evolution and a different window in the time continuum. Several studies have been performed addressing a wide range of phylogenetic questions including global and regional relationships among populations, confirming the reliability of PAIs for the accurate reconstruction of the evolutionary history of human population groups. For instance, the use of these polymorphisms in a worldwide survey of human populations has given a strong support to the African origin of modern humans.[22,23,27,28] Earlier investigations from our laboratory using limited numbers of loci have provided further support for an Out of Africa migration[27,29,30] and a distinctive genetic makeup of populations located north and south of the Sahara desert within the African continent.[30]

This study examines the genetic profiles of six populations from Africa, four from Asia and two from Europe, using 27 PAIs, 16 loci from the *Alu* Ya5, five loci from the *Alu* Ya8, four loci from the *Alu* Yb8 and two from the *Alu* Y subfamilies, to gain insights on human evolutionary history. Bearing in mind that the 12 geographically targeted populations are representatives from the three major human ethnic groups, we also explored the potential usefulness of specific *Alu* loci as ancestry informative markers (AIMs). AIMs, formerly called population-specific alleles, are genetic markers that are capable of detecting differences between populations, so that they can be used to estimate biogeographical ancestry at the level of groups, subgroups and among individuals.[31–34] In a broad sense, biogeographical ancestry is the quantitative representation of the effects of all the factors that have influenced human migration and mating patterns in the past, thereby contributing to the modeling of the present-day worldwide distribution of genetic variation. Knowledge of the proportion of recent genetic ancestry that a given individual shares with members of one or more groups can be very important in forensic, clinical and other scientific applications.[34–36] Our results further support the Out of Africa hypothesis for the origin of modern humans, the genetic segregation of the sub-Saharan groups from the rest of the African populations and the role of geographic proximity in shaping the genetic blueprint of the groups under study. We further tested, for the first time, 27 polymorphic human-specific *Alu* insertions for their potential usefulness as AIMs.

MATERIALS AND METHODS

Subjects

A total of 563 blood samples from healthy, unrelated individuals were collected from the following populations: North Africa (Morocco, N=40; Egypt, N=40), Eastern Africa (Bantu from Kenya, N=40), Central Africa (Hutus from Rwanda, N=46), Western Africa (Benin, N=40; Cameroon, N=16), Asia (Madras, N=42; Ami from Formosa, N=41; Oman, N=66; United Arab Emirates, N=61) and Europe (Galicia, N=73; Georgia, N=58). Table 1 provides the population designations in the form of abbreviations as well as the site of collection and linguistic and ethnic affiliations.

Collection of samples and DNA isolation

All samples were collected as whole blood in EDTA Vacutainer tubes. The ancestry of individuals was assessed by biographical information traced back at least two generations. Each collection was arranged through and supervised by the leaders of the region. Samples were collected according to the ethics guidelines as indicated by Florida International University's Institutional Review Board. The blood cells were lysed and leukocyte nuclei were separated from the rest of the blood components as previously reported.[37] DNA was extracted from leukocyte nuclei as described earlier using proteinase-K digestion and standard organic phenol-chloroform extraction.[28] All samples were stored at −80 °C when not in use.

DNA amplification

In this study, a total average of 184 individuals from three of the 12 populations (Georgia, Oman and UAE) were genotyped for 27 *Alu* insertion polymorphisms: ACE, APO, A25, B65, COL3A1, D1, F13B, HS2.43, HS4.14, HS4.32, HS4.65, HS4.75, HS4.69, HS3.23, HS4.59, HS2.25, NBC1, NBC4, NBC6, NBC60, Sb19.3, Sb19.12, Sb19.10, PR1, PV92, TCR and TPA25. A total average of 379 individuals from the remaining nine populations (Benin, Kenia, Morocco, Rwanda, Egypt, Cameroon, Ami from Formosa, Madras and Galicia) were typed for 15 of the 27 loci examined herein (ACE, A25, D1, F13B, HS4.69, HS3.23, HS4.59, HS2.25, NBC1, NBC4, NBC6, NBC60, Sb19.10, PR1 and TCR). Data for the remaining 12 loci (APO, B65, COL3A1, HS2.43, HS4.14, HS4.32, HS4.65, HS4.75, Sb19.3, Sb19.12, PV92 and TPA25) for these populations were previously reported.[22,28,30,38] Amplification reactions were carried out in 15 µl volumes with 1× buffer (Applied Biosystems, Foster City, CA, USA), 1.5 mM $MgCl_2$, 0.1 mM dNTPs (Applied Biosystems), 250 nM of each primer and 1.0 U Ampli-*Taq* DNA polymerase (Perkin-Elmer, Waltham, MA, USA). Samples were cycled as described in earlier works.[15,30,38]

Polymerase chain reaction products were analyzed by electrophoresis on 3% agarose 1× TAE gels. DNA bands were visualized by staining with ethidium bromide and photographed under ultraviolet light as described earlier.[28]

Statistical analysis

Allelic frequencies for the 27 *Alu* loci in the 12 populations examined were assessed by the direct counting method.[39] To test for the Hardy–Weinberg equilibrium (HWE) expectations, Fisher's exact probability test was conducted to estimate P-values[40] using the Arlequin software, version 3.0.[41]

To analyze genetic affinities among the collections in our study, allelic frequencies of the *Alu* insertions were used to compute F_{ST} unbiased genetic distances[42] between all pairs of populations. From the resultant F_{ST} genetic distance matrix, a dendrogram based on the Neighbor-Joining (NJ) method[43] was constructed using the Phylip v3.2 program.[44] The reliability of the consensus NJ tree was ascertained by means of the bootstrap resampling method.[45] In addition, nonmetric multidimensional scaling (MDS) analysis was performed to represent the F_{ST} genetic distance matrix in two-dimensional space using the SPSS v13.5 statistical package (SPSS Inc., Chicago, IL, USA).

To determine the fraction of the genetic variability due to differences within and among populations as well as among groups of populations, genetic variance was hierarchically apportioned through the analysis of molecular variance (AMOVA)[46] using the Arlequin program. In this statistical analysis, a permutation procedure allows assessment of the significance of the fixation indices F_{CT}, F_{SC} and F_{ST} that measure the relative contribution of the genetic variation between groups, between populations within groups and within populations, respectively. AMOVA tests were performed first for the whole set of populations considered and later for two different population clusters classified according to geographic criteria. In the latter case, we also established an overall test (including all the *Alu* loci) to check the statistical significance of F_{CT} values by combining the separate probability values for each locus through the equation,

$$\chi^2_{[2k]} = -2\sum_{i=1}^{k} \ln p_i$$

where k indicates the number of loci and p_i the separate probability value associated with the F_{CT} values for each i locus.[47–49]

Table 1 Populations analyzed

Population	*Code*	*Ethnic groups*	*Linguistic affiliation*	*Geographical coordinates*
North Africa				
Egypt	EGY	Arabs/Berbers	Afro-Asiatic/Semitic	31°00′ N; 30°00′ E
Morocco	MOR	Arabs/Berbers	Afro-Asiatic/Semitic and Berber	32°00′ N; 5°00′ W
Eastern Africa				
Kenya	KEN	Bantu	Niger-Congo/Benué-Congo/Bantu	2°00′ S; 37°30′ E
Central Africa				
Rwanda	RWA	Hutu	Niger-Congo/Benué-Congo/Bantu	2°00′ S; 30°00′ E
Western Africa				
Benin	BEN	Fon	Niger-Congo/Volta-Congo/Fon	9°30′ N; 2°15′ E
Cameroon	CAM	Bantu	Niger-Congo/Benué-Congo/Bantoid and Bantu	5°00′ N; 12°00′ E
Asia				
Formosa	AMI	Ami	Austric/Austronesian/Formosan	23°30′ N; 121°00′ E
Madras	MAD	Saurashtra	Indo-European/Indo-Aryan/Gujarati	13°4′ N; 80°15′ E
Oman	OMN	Arabs	Afro-Asiatic/Semitic	21°00′ N; 57°00′ E
United Arab Emirates	UAE	Arabs	Afro-Asiatic/Semitic	24°00′ N; 54°00′ E
Europe				
Galicia	GAL	General population	Indo-European/Italic/Romance	41°00′ N; 8°00′ W
Georgia	GEO	General population	Caucasian/South Caucasian	42°00′ N; 43°30′ E

With the aim of interpreting the genetic diversity observed among population clusters, we used two methods to assess its congruency in relation to the geographic and genetic coordinates. First, to obtain a consensus topogenetic map, the first two eigenvectors of the nonmetric MDS analysis were extracted and rotated to maximum congruence with geographic coordinates using methods described by Lalouel.[50] The second method used was the matrix comparison test devised by Mantel[51] and modified by Smouse *et al.*,[52] which can be used to compare the distance and similarity or dissimilarity matrices provided that they are calculated from independent data sets.[53]

The population structure was inferred from the *Structure* v. 2.3.1 program using genotype data for the whole set of populations.[54] We first used the admixture model, performing clustering without population of origin information and with the number of ancestral populations fixed at K=3 and 4 to assess whether the clustering correlates with the NJ and MDS analyses. The same model was then implemented using population information (geographic sampling location) and with the number of populations fixed at K=12 and 13 to examine the possibility of population substructure in the studied populations. In both analyses, we ran *Structure* under the assumption that the allele frequencies in the populations are independent. Analyses were performed with a length of burn-in period of 20 000 and 20 000 Markov chain Monte Carlo repetitions after burn-in. The software applied for this analysis is available at http://pritch.bsd.uchicago.edu/structure.html.

To evaluate the effectiveness of the 27 *Alu* markers examined as AIMs, we first calculated the weighted *Alu* insertion frequencies for each one of the geographical population groups (see Table 1). Then, the weighted frequencies were used to compute the δ-value between population clusters. For dimorphic markers, $\delta=|p_1-p_2|$, where p_1 and p_2 are the weighted frequencies of the insertion in population groups 1 and 2, respectively. Earlier studies have established the threshold δ as a frequency differential of 30%.[55,56]

RESULTS

Intrapopulation diversity

Averages of 563 individuals from 12 populations were examined using 27 PAIs. Populations, as well as the numbers of individuals, loci analyzed and observed allelic frequencies, are presented in Supplementary Table 1. Likewise, heterozygosity values and results of Fisher's exact probability test for the HWE expectations are presented in Supplementary Table 2.

All loci are polymorphic in all populations, with the exception of APO in Morocco and the Ami, and HS4.14, HS4.75 and NBC6 in the Ami, which are all fixed for the presence of the *Alu* insertion as well as HS2.43 and PR1 in Benin, Kenya, Cameroon and the Ami, Sb19.10 in Benin and Cameroon, and A25 and TCR in the Ami, which are all fixed for the absence of the *Alu* element.

Of all the populations analyzed, the one with the highest observed heterozygosity is Cameroon (0.403), and that with the lowest value is the population of the Ami (0.205). When grouped by continent, a slightly higher level of observed heterozygosity is present in the African populations (0.309) than in the Asian (ASI) (0.288) and European (0.268) groups.

Using Fisher's exact probability test, significant departures from the HWE expectations were observed in 42 of the 324 tests analyzed (12.9%). Only 26 tests (8.0%), however, remained in significant disagreement with the expected values when the Bonferroni correction was applied. The loci with more deviations from the expected values were D1 in five populations, and HS4.14 and Sb19.12 in three populations each (Supplementary Table 2). In two instances, involving HS4.69 in Madras and Galicia, an excess of heterozygosity was observed. In the remaining 24 of the 26 significant tests (92.3%), deviations from the HWE expectations were due to heterozygous deficit. The populations possessing a larger number of *Alu* markers in disequilibrium were Egypt (eight loci) and Galicia (seven loci). On the other hand, Cameroon, Morocco and Rwanda showed no loci with significant deviations from the expected heterozygosity values.

Interpopulation diversity

Phylogenetic relationships among populations were examined using NJ analysis (Figure 1). The overall topology of the resultant

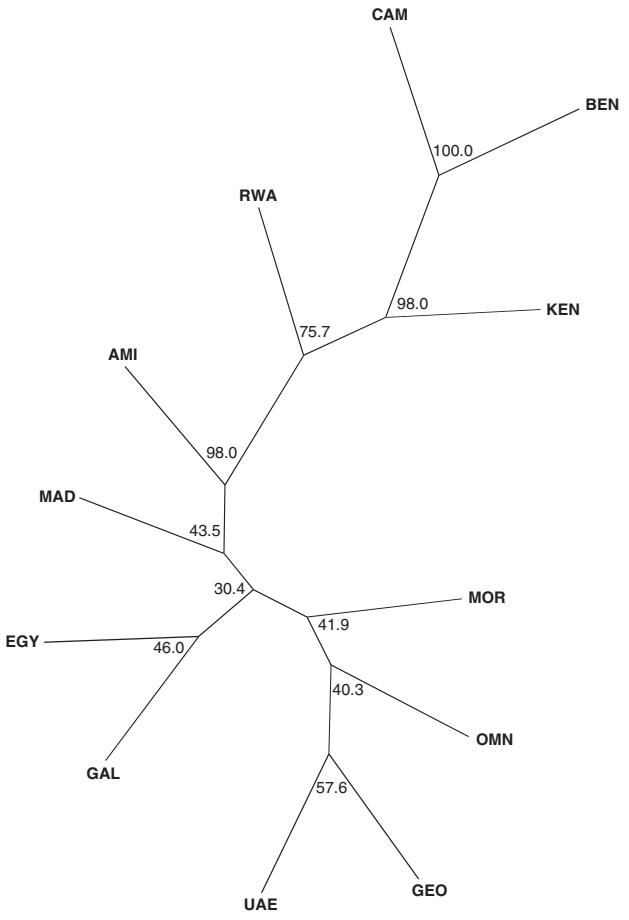

Figure 1 Neighbor-joining tree based on the F_{ST} genetic distance matrix. Genetic distances were computed from allelic frequencies of 27 *Alu* insertions of 12 worldwide populations. Figures in tree nodes are percentage bootstrap values, estimated from 1000 replicates. BEN, Benin; CAM, Cameroon; KEN, Kenya; RWA, Rwanda; EGY, Egypt; GEO, Georgia; MAD, Madras; MOR, Morocco; OMN, Oman; UAE, United Arab Emirates; AMI, Ami from Formosa; GAL, Galicia.

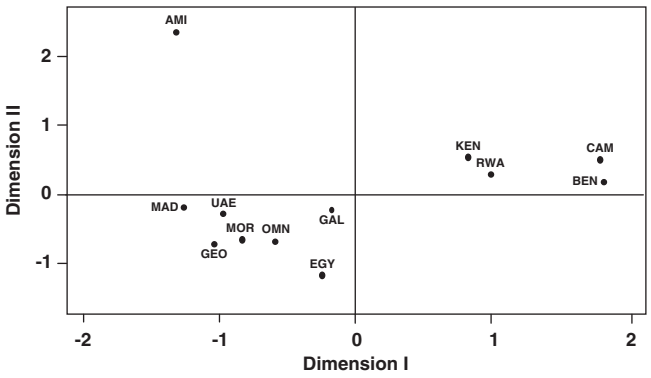

Figure 2 Nonmetric multidimensional scaling (MDS) based on the Reynolds (F_{ST}) genetic distance matrix. Genetic distances were computed from allelic frequencies of 27 *Alu* insertions of 12 worldwide populations. BEN, Benin; CAM, Cameroon; KEN, Kenya; RWA, Rwanda; EGY, Egypt; GEO, Georgia; MAD, Madras; MOR, Morocco; OMN, Oman; UAE, United Arab Emirates; AMI, Ami from Formosa; GAL, Galicia. The total variance accounted for is 97.4% and the coefficient of stress 0.075.

phylogenetic tree is congruent with the geographical distribution of the collections examined. One markedly distinctive cluster grouped all SSA populations (Benin, Cameroon, Kenya and Rwanda). The Ami partitions intermediate between the SSA cluster and the rest of the populations. Madras also segregates on the same branch, distant from the SSA groups and proximal to the rest of the populations. Both the position of the sub-Saharan cluster and the Ami on the dendrogram are statistically supported by the high bootstrap values estimated for their corresponding tree nodes, based on 1000 replications. The remaining populations (Egypt, Morocco, Galicia, Oman, UAE and Georgia) segregate as a second pole on the dendrogram. The geographical heterogeneity of this cluster (North Africa, Europe, Arabian Peninsula and Asia) seems to be reflected in the moderate bootstrap values obtained for the consensus NJ tree.

Figure 2 illustrates the results of nonmetrical MDS applied to the F_{ST} matrix. Compatible with NJ data, populations cluster according to geography. Statistically, the results of the MDS analysis are robust accounting for 97.4% of the total variance, with a coefficient of stress of 0.075.

As can be noted, populations segregated along both dimensions, with all the SSA populations (Benin, Cameroon, Kenya and Rwanda) concentrated in the positive segment of both axes. Noteworthy is the partition of the West African populations of Cameroon and Benin from the East African collection of Kenya and the Central African group of Rwanda. All the remaining populations, all Caucasians, plotted in the negative segment of Dimensions I and II, with Galicia located very close to the center of the plot. The exception to this general topology was the Ami, which showed the greatest genetic divergence, plotting on the negative quadrant of Dimension I and the positive quadrant of Dimension II as an outlier.

To confirm that the results obtained with all 27 markers were not affected by the loci showing a large number of deviations from the expected values (D1 in five populations, and HS4.14 and Sb19.12 in three populations each, see Supplementary Table 2), NJ and MDS analyses were performed again, this time excluding the D1, HS4.14 and Sb19.12 loci. The overall topologies of both the NJ tree (Figure 3) and the MDS plot (Figure 4) proved to be very similar to the ones obtained by considering these three loci, indicating no substantial contribution of these loci to the phylogenetic relationships of the populations included in the analysis.

On the basis of the NJ and MDS results, which indicate a certain congruency of *Alu* diversity pattern with geographical location, we analyze how the observed genetic heterogeneity is spatially structured by hierarchical AMOVA. In a first step, AMOVA tests were performed for each locus considering the whole set of populations (Supplementary Table 3). Most *Alu* insertions show significant frequency differences among populations, as can be inferred from the F_{ST} fixation indexes. The exceptions were ACE ($P>0.226$), B65 ($P>0.191$), HS2.43 ($P>0.211$), NBC1 ($P>0.752$) and Sb19.12 ($P>0.052$). The highest F_{ST} values were observed in APO, PV92 (18.4 each), Sb19.10 (16.8) and NBC6 (16.7). The overall AMOVA test performed by considering the frequencies of all *Alu* elements (the 27 loci) also indicates highly significant differences among populations ($P<0.001$).

Further AMOVA analyses were performed to ascertain maximum genetic variance between population groups. On the basis of the results of the NJ tree, populations were divided into two geographic clusters: (1) sub-Saharan Africa (Benin, Cameroon, Kenya and Rwanda), and (2) North Africa, Europe and Asia (Morocco, Egypt, Georgia, Galicia, Oman, UAE, Madras and Ami from Formosa). On assignment of the populations within these two broad geographic

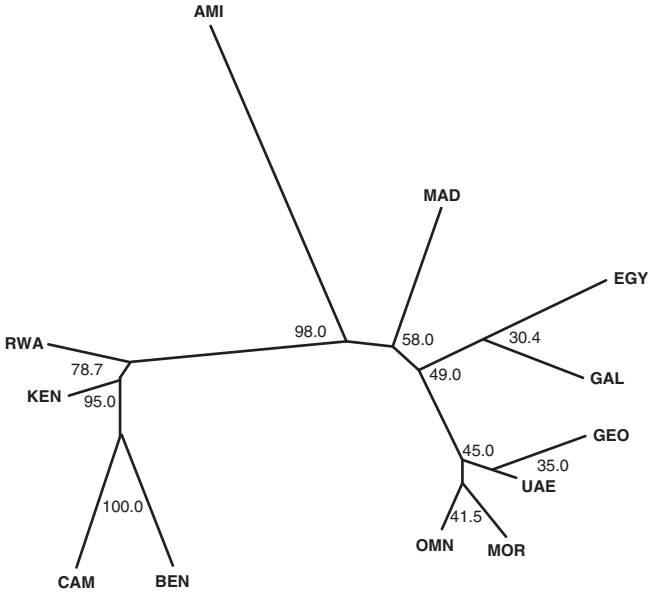

Figure 3 Neighbor-joining tree based on the F_{ST} genetic distance matrix. Genetic distances were computed from the allelic frequencies of 24 *Alu* insertions (excluding D1, HS4.14 and Sb1912 loci) of 12 worldwide populations. Figures in tree nodes are percentage bootstrap values, estimated from 1000 replicates. BEN, Benin; CAM, Cameroon; KEN, Kenya; RWA, Rwanda; EGY, Egypt; GEO, Georgia; MAD, Madras; MOR, Morocco; OMN, Oman; UAE, United Arab Emirates; AMI, Ami from Formosa; GAL, Galicia.

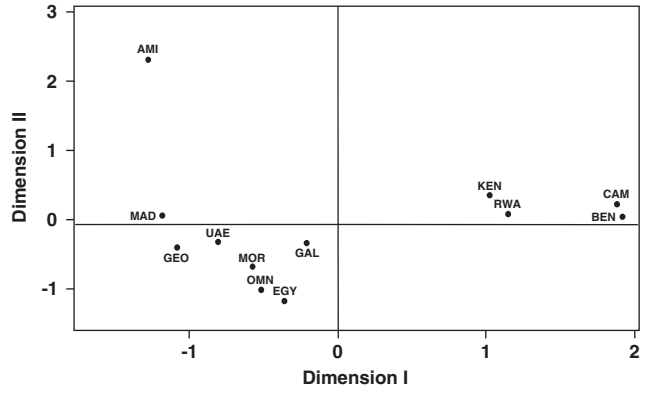

Figure 4 Nonmetric multidimensional scaling (MDS) based on the Reynolds (F_{ST}) genetic distance matrix. Genetic distances were computed from the allelic frequencies of 24 *Alu* insertions (excluding D1, HS4.14 and Sb1912 loci) of 12 worldwide populations. BEN, Benin; CAM, Cameroon; KEN, Kenya; RWA, Rwanda; EGY, Egypt; GEO, Georgia; MAD, Madras; MOR, Morocco; OMN, Oman; UAE, United Arab Emirates; AMI, Ami from Formosa; GAL, Galicia. The total variance accounted for is 97.8%, and the coefficient of stress 0.082.

regions, AMOVA analyses were performed for each of the 27 PAIs (Table 2). We detected statistically significant differences between population groups (F_{CT}) for 17 of the PAIs. Of them, F_{CT} values for APO (19.81, $P<0.01$), B65 (4.41, $P<0.01$), COL3A1 (23.12, $P<0.01$), F13B (12.50, $P<0.05$), HS2.25 (3.17, $P<0.05$), HS2.43 (1.96, $P<0.05$), HS3.23 (4.39, $P<0.05$), HS4.59 (4.56, $P<0.05$), HS4.69 (5.10, $P<0.05$), HS4.75 (15.22, $P<0.01$), NBC4 (14.90, $P<0.01$), NBC6 (29.22, $P<0.01$), NBC60 (7.00, $P<0.05$), PR1 (6.13, $P<0.05$),

Table 2 Fixation indices (F_{CT}, F_{SC}, F_{ST}) generated from a hierarchical analysis of molecular variance for 27 polymorphic *Alu* insertions, considering two population groups

			Fixation indices		
Alu *marker*	*Frequency range*	*Mean frequency*	F_{CT} *(%)*	F_{SC} *(%)*	F_{ST} *(%)*
A25	0.000–0.409	0.194	−1.20 NS	6.05***	4.92***
ACE	0.263–0.530	0.373	−0.62 NS	0.91 NS	0.30 NS
APO	0.488–1.000	0.840	19.81**	10.45***	28.19***
B65	0.474–0.763	0.579	4.41**	−1.12 NS	3.34 NS
COL3A1	0.022–0.500	0.149	23.12**	3.04*	25.45***
D1	0.053–0.591	0.315	−1.20 NS	6.05***	4.92***
F13B	0.135–0.881	0.398	12.50*	8.68***	20.09***
HS2.25	0.075–0.417	0.214	3.17*	3.14**	6.20**
HS2.43	0.000–0.087	0.029	1.96*	−0.26 NS	1.71 NS
HS3.23	0.758–0.975	0.835	4.39*	0.63 NS	4.99*
HS4.14	0.469–1.000	0.659	−0.04 NS	7.35***	7.32***
HS4.32	0.184–0.850	0.553	5.40 NS	8.29***	13.24***
HS4.59	0.375–0.709	0.594	4.56*	0.73 NS	5.25*
HS4.65	0.017–0.286	0.130	3.13 NS	2.50*	5.55***
HS4.69	0.075–0.518	0.362	5.10*	1.27 NS	6.30*
HS4.75	0.650–1.000	0.854	15.22**	3.23*	17.95***
NBC1	0.531–0.800	0.701	−0.11 NS	−0.75 NS	−0.86 NS
NBC4	0.500–0.946	0.747	14.90**	4.21**	18.48***
NBC6	0.531–1.000	0.823	29.22**	0.79 NS	29.78***
NBC60	0.375–0.819	0.599	7.00*	1.39 NS	8.30***
PR1	0.000–0.152	0.072	6.13*	0.82 NS	6.90**
PV92	0.125–0.913	0.337	−3.48 NS	19.53***	16.73***
Sb19.3	0.312–0.868	0.678	12.88**	3.74**	16.14***
Sb19.10	0.000–0.594	0.245	23.13**	6.17***	27.88***
Sb19.12	0.088–0.406	0.240	−0.12 NS	1.55 NS	1.44 NS
TCR	0.000–0.438	0.181	1.32 NS	5.75***	6.99***
TPA25	0.275–0.700	0.469	3.43**	0.74 NS	4.14*

Abbreviation: NS, nonsignificant. Groups are: sub-Saharan Africa (Benin, Cameroon, Kenya and Rwanda) and North Africa, Europe and Asia (Morocco, Egypt, Georgia, Galicia, Oman, United Arab Emirates, Madras and Formosa). F_{CT}, genetic variation among groups. F_{SC}, genetic variation among populations within groups. F_{ST}, genetic variation among individuals within populations. Statistical significance for $^{*}P<0.05$, $^{**}P<0.01$, $^{***}P<0.001$.

Sb19.3 (12.88, $P<0.01$), Sb19.10 (23.13, $P<0.01$) and TPA25 (3.43, $P<0.01$) are higher than the genetic variance among populations within groups (F_{SC}); thus, these PAIs can be considered as those contributing the most to the differentiation of geographic clusters. By contrast, 10 out of the 27 *Alu* loci do not show significant partitioning along geographic lines. The overall test for the significance of F_{CT}, which combines separate probability values for each locus, generated statistically significant differences ($P<0.01$), indicating substantial genetic structuring between the two geographical clusters. The high number of F_{ST} tests resulting in significant departure may reflect the high genetic diversity shown by the populations examined in this study (Table 2). These findings are not surprising given that several of the populations analyzed originate from regions that lie within major migratory routes.

The topology of genetic structuring, that is, the concordance between the F_{ST} genetic distance matrix and the geographic distance matrix, was assessed using the Mantel test of matrix correspondence. To that end, the first two eigenvectors generated by the MDS plot were fitted up to maximum congruity with the geographical coordinates of the samples' origin (Figure 5). The results of the Mantel test revealed that both matrices correlated significantly ($r=0.528$, $p<0.001$). These

Figure 5 Matrix fitting of the geographic and genetic coordinates for 12 populations from Africa, Asia and Europe. Genetic coordinates were estimated from the allelic frequencies of 27 *Alu* insertions. Full circles represent the geographic locations of the targeted populations. Arrows indicate the location predicted by genetic kinship. BEN, Benin; CAM, Cameroon; KEN, Kenya; RWA, Rwanda; EGY, Egypt; GEO, Georgia; MAD, Madras; MOR, Morocco; OMN, Oman; UAE, United Arab Emirates; AMI, Ami from Formosa; GAL, Galicia.

data are congruent with genetic heterogeneity among the populations analyzed, as they are patterned mainly by isolation by distance. This is particularly perceptible for the group of SSA collections, clearly separated from the rest of the populations and possessing the lowest differences between the genetic and geographic coordinates. Remote from the main population cluster is the Ami collection, whose genetic coordinates drift apart from the rest of the samples examined, coinciding with the NJ and MDS results. Interestingly, when genetic and geographic topologies are adjusted (Figure 5), the rest of the populations under study (Galicia, Morocco, UAE, Oman, Egypt, Georgia and Madras) show a degree of genetic kinship higher than expected according to the model of isolation by distance. In other words, the genetic coordinates of the cited populations tend to be closer than expected according to their geographic distribution.

Weighted allele frequencies for the 27 PAIs in the three population clusters established according to geography are listed in Table 3. Differences in weighted frequencies (δ values) are also provided. According to δ values, the population cluster showing the greater intergroup differences is sub-Saharan Africa (SSA). The lower δ values appeared in the comparison between the ASI (Asia) and Europe-North African (EUR) population groups. The mean δ values (as percentage) for the comparisons between SSA/ASI, SSA/EUR and ASI/EUR are 15.9, 16.3 and 6.1%, respectively.

With respect to individual PAIs, δ values of 20% or higher were obtained for a total of 13 *Alu* loci. Of these, three presented $\delta \geqslant 30\%$, the threshold value established to differentiate population groups.[55] These three loci are: (i) F13B (35.1%) in the pair wise comparison of SSA/ASI, (ii) Sb19.10, with the greatest δ value registered in the analysis (SSA/ASI: 40.2%) and (iii) NBC6, the only *Alu* marker showing $\delta \geqslant 30\%$ in more than one comparison between population clusters (SSA/ASI: 32.1%, SSA/EUR: 30.1%). Of the remaining 10 loci, five exhibit $\delta > 20\%$, five show differences in weighted frequency values for both the SSA/ASI and the SSA/EUR comparisons. These *Alu* loci are A25 (26.0 and 29.1%, respectively), APO (21.0 and 26.9%), COL3A1 (23.1 and 20.8%), NBC4 (24.8 and 25.5%) and Sb19.3 (22.0 and 26.8%). The remaining five loci within this category ($\delta > 20\%$) show this frequency difference threshold in one of the comparisons: HS4.32 (SSA/ASI: 23.3%), HS4.69 (SSA/EUR: 20.4%), HS4.75 (SSA/ASI: 22.4%), NBC60 (SSA/EUR: 21.8%) and PV92 (ASI/EUR: 23.3%). PV92 is the only locus with a particularly high δ value in the comparison between the geographic groups ASI/EUR.

Phylogenetic relationships evident in both the NJ and MDS graphs (Figures 1 and 2, respectively) are mirrored by the *Structure* bar plots assuming three and four ancestral populations (K=3 and 4). Individual ancestry proportions inferred by the *Structure* algorithms (Supplementary Figure 1a) are consistent with the clustering patterns observed in the NJ dendrogram and MDS plot discriminating among the SSAs, the North African, the Arabian Peninsula and the European populations. Similarly, the Ami, which occupies an intermediate position within the tree but is relatively isolated from the other two groups in the MDS, forms a distinct clade in the *Structure* plot indicating high proportions of a single ancestral population (that is,

Table 3 Weighted allelic frequencies of 27 Alu insertions in three population groups classified according to geography and differences in weighted frequencies (δ) between them

	Weighted allele frequencies			*Difference in frequencies (δ)*		
Alu marker	*SSA*	*ASI*	*EUR*	*SSA/ASI*	*SSA/EUR*	*ASI/EUR*
A25	0.379	0.119	0.088	0.260	0.291	0.031
ACE	0.368	0.427	0.319	0.059	0.049	0.108
APO	0.680	0.890	0.950	0.210	0.269	0.059
B65	0.680	0.510	0.542	0.170	0.138	0.031
COL3A1	0.276	0.045	0.068	0.231	0.208	0.023
D1	0.294	0.367	0.317	0.073	0.024	0.049
F13B	0.202	0.552	0.393	**0.351**	0.192	0.159
HS2.25	0.121	0.203	0.283	0.081	0.161	0.080
HS2.43	0.007	0.026	0.068	0.019	0.060	0.042
HS3.23	0.916	0.806	0.775	0.110	0.141	0.031
HS4.14	0.599	0.704	0.685	0.104	0.086	0.018
HS4.32	0.419	0.652	0.588	0.233	0.168	0.065
HS4.59	0.477	0.630	0.644	0.153	0.167	0.014
HS4.65	0.188	0.134	0.058	0.054	0.129	0.076
HS4.69	0.259	0.397	0.462	0.139	0.204	0.065
HS4.75	0.728	0.952	0.900	0.224	0.172	0.052
NBC1	0.687	0.732	0.711	0.045	0.024	0.021
NBC4	0.588	0.836	0.843	0.248	0.255	0.007
NCB6	0.621	0.942	0.922	**0.321**	**0.301**	0.021
NBC60	0.491	0.651	0.709	0.160	0.218	0.058
PR1	0.007	0.085	0.134	0.078	0.127	0.049
PV92	0.360	0.415	0.182	0.055	0.178	0.233
Sb19.3	0.533	0.753	0.801	0.220	0.268	0.049
Sb19.10	0.022	0.424	0.303	**0.402**	0.280	0.121
Sb19.12	0.265	0.284	0.163	0.019	0.102	0.121
TCR	0.228	0.108	0.176	0.120	0.051	0.069
TPA25	0.364	0.515	0.510	0.151	0.146	0.006

Population groups are as follows: SSA, sub-Saharan Africa (Benin, Cameroon, Kenya and Rwanda), ASI, Asian populations (Oman, United Arab Emirates, Madras, Ami), EUR, Europe and Northern Africa (Morocco, Egypt, Georgia and Galicia).
Bold type indicates δ-values >0.30.

represented by the color red in Supplementary Figure 1a) across all individuals examined. The membership of individuals within the graphical representation inferred by $K=12$ and 13 revealed no population substructure (Supplementary Figure 1b).

DISCUSSION

This study provides for the first time a comprehensive survey on the value of a set of 27 polymorphic human-specific *Alu* insertions as AIMs.

A common origin of all human populations and the recent divergence of the human species into continental groups result in the fact that the vast majority of the genetic variation (80–90%) among humans is interindividual.[39,57] Results of the various genetic analyses based on the 27 PAIs in 12 worldwide populations show that human genetic variation tends to be spatially (geographically) structured, in concurrence with historical patterns of gene flow and genetic drift. Of these two evolutionary forces, genetic drift is presumed to have had a major role in the genetic makeup of human groups as a result of the partial isolation of populations during much of their evolutionary history[58] and the sequential nature of the original human migration that populated the planet. Along these lines, our analyses reveal the sensitivity of the *Alu* markers under study to detect the effects of isolation, genetic drift and admixture on the genetic background of the populations examined. Moreover, our findings show the reliability of some of the PAIs screened as AIMS.

Both the genetic affinities observed in the results of the NJ, MDS and *Structure* analyses, as well as the spatial structuring of the *Alu* frequencies inferred from AMOVA, are essentially summarized in the topology of the *Alu* diversity displayed in Figure 3. All of them are revealing of three major trends, namely: (1) the polarization between the SSA populations (Benin, Cameroon, Kenya and Rwanda) and the rest of the samples considered in the study, including those of Northern Africa, (2) the genetic uniqueness shown by the Ami of Formosa, which is perceptibly separated from the other collections and (3) bearing the geographical position in mind, a higher than expected genetic connection between populations from Europe, North Africa and Asia (Galicia, Georgia, Egypt, Morocco, UAE, Oman and Madras).

The sharp genetic discontinuity between the North African population cluster (Egypt, Morocco) and the sub-Saharan groups is most likely determined by the presence of the Sahara Desert, which is thought to have had a key role in shaping the genetic landscape of the African continent and beyond. This desert might have constituted a strong physical barrier to the gene flow, thereby promoting a prominent genetic differentiation among human groups settled at opposite sides of this obstacle.[30,59,60]

A thorough examination of the results of the several genetic analyses performed reveals that although all sub-Saharan collections segregated in a well-differentiated cluster from the remaining of the geographically targeted worldwide populations, a clear partition is perceptible between West African (Benin and Cameroon) and East/Central Africa (Kenya and Rwanda) populations. This observation is congruent with earlier studies on genetic variation in Africa, based on autosomal protein markers[61] and on Y-chromosome haplotypes.[59] These latter authors have suggested a large component of the Khoisan gene pool in East Africa to explain the observed geographic structuring of the haplotypic diversity. In the same study, northern Cameroonians were found to be clearly distinct from a cluster formed by a group of poorly differentiated Niger-Congo-speaking populations from western, central western and southern Africa, indicating genetic isolation for a considerable period of time. Yet some investigators have affirmed that, despite some correspondence between language affiliation and genetic similarity, geographic proximity seems to be a better predictor of genetic affinity among African populations.[62] The findings from this study give credence to this notion.

The noticeable genetic singularity of the Ami of Formosa is reflected in the lowest heterozygosity level of all populations under study (0.212). The prolonged isolation and the small effective population size that characterize the Ami aborigines would have promoted local genetic microdifferentiation by genetic drift, accumulating distinctive allele frequencies and tending to fixation.[57,63]

The higher than expected genetic affinity between populations from Europe, North Africa and Asia (Galicia, Georgia, Egypt, Morocco, UAE, Oman and Madras) could be attributed to the homogenizing effect of gene flow between regions connected by two major migratory passageways in recent human dispersals: the Levantine Corridor and the Horn of Africa.[60,64,65] Northern Africa (Morocco and Egypt), the Arabian Peninsula (UAE and Oman) and the Caucasus region (Georgia) were all key geographic regions of major importance to human evolution during the Plio-Pleistocene, for the emergence of anatomically modern humans, for bidirectional migrations between Africa and Eurasia, and for understanding transcontinental gene flow and dispersal patterns in 'Out-of-Africa' models.[66,67] However, another plausible explanation for these findings may be limited

resolving power of the 27 *Alu* loci discriminating among North Africans, Europeans and Asians.

A deficiency of heterozygotes and an excess of homozygotes may suggest the presence of a population substructure (that is, Wahlund effect).[68] To explore this possibility, the *Structure* analysis was performed. When we set the number of assumed populations at $K=3$ and 4, the inferred clusters were congruent with the patterns observed in the NJ and MDS plots. Overall, no population substructure was observed at $K=12$ and 13, revealing that all groups belong to predefined populations, thereby providing no explanation concerning the deviation from the HWE expectations.

Bearing in mind that we were analyzing populations from different continents, the level of genetic differentiation among groups was expected to be high. Thus, we took advantage of the high number of *Alu* insertions examined and the wide geographical distribution of the target populations to explore the potential usefulness of these *Alu* elements as AIMs. Three *Alu* markers were found to show notable differences ($\delta \geqslant 30\%$) in weighted allelic frequencies between population groups: F13B, Sb19.10 and NBC6. All of them can be considered AIMs to differentiate between the SSA and ASI groups, whereas NBC6 is also efficient in distinguishing between the SSA and Caucasian (EUR) populations. AMOVA results strongly corroborate the discrimination power of these PAIs, as they render statistically significant differences in allele frequencies between the two population groups, and possess high values for Wright's fixation index, F_{ST} (F13B: 14.7%, Sb19.10: 16.8% and NBC6: 16.7%). Of these three *Alu* AIMs, F13B has already been included in some AIM panels.[55] We report here, for the first time, two additional PAIs with AIM characteristics that may be incorporated into panels for identification purposes.

There are 10 *Alu* markers in total that present moderate δ levels ($20\% < \delta < 30\%$). Among them, APO, PV92 and Sb19.3 have also been used earlier in AIM panels.[55,56] Interestingly, PV92 is the only *Alu* marker with a δ value of over 20% in the ASI/EUR comparison. On the other hand, APO, Sb19.3, A25, COL3A1 and NBC4 show moderate δ levels ($20\% < \delta < 30\%$) in the SSA/ASI and SSA/EUR comparisons. The potential efficacy of these PAIs to detect genetic differences between distinct ethnic groups is also supported by the F_{ST} values obtained with the AMOVA values for the set of all populations, all of them close to or over 10%: 10.3, 18.4, 12.1, 10.9, 18.4 and 9.2% for A25, APO, COL3A1, NBC4, PV92 and Sb19.3, respectively. The remaining group of four Alu markers includes in HS4.32 and HS4.75 with moderate differences in allele frequencies between SSA and ASI, and HS4.69 and NBC60, which could be considered as AIMs for SSA/EUR discrimination. However, the fixation indices for HS4.69 and NBC60 are below 5% (3.4 and 4.3%, respectively), indicating that more than 95% of the variance in allele frequency corresponds to within-group variance. The utilization of these loci as AIMs should be discouraged.

CONCLUSIONS

Alu elements have been shown to be robust markers for evolutionary and phylogenetic studies due to of their unique mechanism of insertion, which confers on each locus a genetic polarity, allowing the inference of ancestral states and sound assumptions of migration patterns. Furthermore, some PAIs are diagnostic markers for the detection of interpopulation differences. In the quest for understanding the relationship between populations, ancestry, population genetic profile, population structure, demographic history and geographic origin, we believe the use of PAIs represents a reliable, informative and cost-effective experimental method to establish those relationships, further opening new grounds for epidemiological, medical and forensic studies. Our results support the use of 11 *Alu* markers in AIM panels: (i) A25, APO, COL3A1, NBC4, NBC6, Sb19.3 and Sb19.10 to distinguish SSA from both ASI and European samples, (ii) F13B, HS4.32 and HS4.75 for comparisons between SSA and ASI collections and (iii) PV92 to differentiate between ASI and European samples. Conversely, our results support the idea of not including the remaining markers in human population study panels, the low δ values obtained and the *F*st results derived from AMOVA.

The findings of our study also indicate that a limited number of *Alu* markers could be helpful to infer the population structure in other human groups. Overall, these analyses corroborate that *Alu* loci with higher F_{ST} values possess greater resolving power and produce more consistent genetic distance estimates, in accord with the findings of an earlier study.[69]

ACKNOWLEDGEMENTS

MA Alfonso-Sánchez was funded by Research Projects GIU 05/51 from the University of the Basque Country (UPV/EHU) and by IT-424-07 from the Basque Government.

1 Houck, C. M., Rinehart, F. P. & Schmid, C. W. A ubiquitous family of repeated DNA sequences in the human genome. *J. Mol. Biol.* **132,** 289–306 (1979).
2 Ullu, E. & Tschudi, C. *Alu* sequences are processed 7SL RNA genes. *Nature* **312,** 171–172 (1984).
3 Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. & the human genome consortium Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
4 Deininger, P. L., Jolly, D. J., Rubin, C. M., Friedmann, T. & Schmid, C.W. Base sequence studies of 300 nucleotide renatured repeated human DNA clones. *J. Mol. Biol.* **151,** 17–33 (1981).
5 Paolella, G., Lucero, M. A., Murphy, M. H. & Baralle, J. E. The *Alu* family repeat promoter has a tRNA-like bipartate structure. *EMBO J.* **2,** 691–696 (1983).
6 Economou, E. P., Bergen, A. W., Warren, A. L. & Antonarakis, S. E. The polydeoxyadenylate tract of *Alu* repetitive elements is polymorphic in the human genome. *Proc. Natl Acad. Sci. USA* **87,** 2951–2954 (1990).
7 Rogers, J. The origin and evolution of retroposons. *Int. Rev. Cytol.* **93,** 187–279 (1985).
8 Korenberg, J. R. & Rykowski, M. C. Human genome organization: *Alu*, LINE and the molecular structure of metaphase chromosome bands. *Cell* **53,** 391–400 (1988).
9 Mathias, S. L., Scott, A. F., Kazazian, H. H. Jr., Boeke, J. D. & Gabriel, A. Reverse Transcriptase encoded by a human transposable element. *Science* **254,** 1808–1810 (1991).
10 Deininger, P. L. & Daniels, G. R. The recent evolution of mammalian repetitive elements. *Trends Genet.* **2,** 76–80 (1986).
11 Zuckerkandl, E., Latter, G. & Jurka, J. Maintenance of function without selection: *Alu* sequences as 'cheap genes'. *J. Mol. Evol.* **29,** 504–512 (1989).
12 Deininger, P. L., Batzer, M. A., Hutchison, C. A. & Edgell, M. H. Master genes in mammalian repetitive DNA amplification. *Trends Genet.* **8,** 307–311 (1992).
13 Willard, C., Nguyen, H. T. & Schmid, C. W. Existence of at least three distinct *Alu* subfamilies. *J. Mol. Evol.* **26,** 180–186 (1987).
14 Matera, A. G., Hellmann, U. & Schmid, C. W. A transpositionally and transcriptionally competent *Alu* subfamily. *Mol. Cell. Biol.* **10,** 5424–5432 (1990).
15 Batzer, M. A. & Deininger, P. L. A human-specific subfamily of *Alu* sequences. *Genomics* **9,** 481–487 (1991).
16 Batzer, M. A. & Deininger, P. L. *Alu* repeats and human genomic diversity. *Nat. Rev. Genet.* **3,** 370–379 (2002).
17 Arcot, S. S., DeAngelis, M. M., Sherry, S. T., Adamson, A. W., Lamerdin, J. E., Deininger, P. L. *et al.* Identification and characterization of two polymorphic Ya5 *Alu* repeats. *Mutat. Res.* **382,** 1084–1092 (1997).
18 Xing, J., Salem, A. H., Hedges, D. J., Kilroy, G. E., Watkins, W. S., Schienman, J. E. *et al.* Comprehensive analysis of two *Alu* Yd subfamilies. *J. Mol. Evol.* **57**(Suppl 1), S76–S89 (2003).
19 Carter, A. B., Salem, A. H., Hedges, D. J., Keegan, C. N., Kimball, B., Walter, J. A. *et al.* Genome-wide analysis of the human *Alu* Yb-lineage. *Hum. Genomics* **1,** 167–178 (2004).
20 Otieno, A. C., Carter, A. B., Hedges, D. J., Walter, J. A., Ray, D.A., Garber, R. K. *et al.* Analysis of the human *Alu* Ya-Lineage. *J. Mol. Biol.* **342,** 109–118 (2004).
21 Garber, R. K., Hedges, D. J., Herke, S. W., Hazard, N. W. & Batzer, M. A. The *Alu* Yc1 subfamily: sorting the wheat from the chaff. *Cytogenet. Genome. Res.* **110,** 537–542 (2005).
22 Batzer, M. A., Stoneking, M., Alegria-Hartman, M., Bazan, H., Kass, D. H., Shaikh, T. H. *et al.* African origin of human specific polymorphic *Alu* insertions. *Proc. Natl Acad. Sci. USA* **91,** 12288–12292 (1994).

23 Stoneking, M., Fontius, J. J., Clifford, S. L., Soodyall, H., Arcot, S. S., Saha, N. *et al.* *Alu* insertions polymorphisms and human evolution: evidence for a larger population size in Africa. *Genome Res.* **7,** 1061–1071 (1997).
24 Edwards, M. C. & Gibbs, R. A. A human dimorphism resulting from loss of an *Alu*. *Genomics* **14,** 590–597 (1992).
25 Britten, R. J. Mobile elements inserted in the distant past have taken on important functions. *Gene* **205,** 177–182 (1997).
26 Roy-Engel, A. M., Carroll, M. L., El-Sawy, M., Salem, A. H., Garber, R. K., Nguyen, S. V. *et al.* Non-traditional *Alu* evolution and primate genomic diversity. *J. Mol. Biol.* **316,** 1033–1040 (2002).
27 Batzer, M. A., Arcot, S. S., Phinney, J. W., Alegria-Hartman, M., Kass, D. H., Milligan, S. M. *et al.* Genetic variation of recent *Alu* insertions in human populations. *J. Mol. Evol.* **42,** 22–29 (1996).
28 Antunez de Mayolo, G., Antunez de Mayolo, A., Antunez de Mayolo, P., Papita, S. S., Hammer, M., Yunis, J. J. *et al.* Phylogenetics of worldwide human populations as determined by polymorphic *Alu* insertions. *Electrophoresis* **23,** 3346–3356 (2002).
29 Novick, G. E., Batzer, M. A., Deininger, P. L. & Herrera, R. J. The mobile genetic element *Alu* in the human genome. *BioScience* **46,** 32–41 (1996).
30 Terreros, M. C., Martinez, L. & Herrera, R. J. Polymorphic Alu insertions and genetic diversity among African populations. *Hum. Biol.* **77,** 675–704 (2005).
31 Shriver, M. D., Smith, M. W., Jin, L., Marcini, A., Akey, J. M., Deka, R. *et al.* Ethnic-affiliation estimation by use of population-specific DNA markers. *Am. J. Hum. Genet.* **60,** 957–964 (1997).
32 Shriver, M. D., Parra, E. J., Dios, S., Bonilla, C., Norton, H., Jovel, C. *et al.* Skin pigmentation, biogeographical ancestry and admixture mapping. *Hum. Genet.* **112,** 387–399 (2003).
33 Bamshad, M., Wooding, S., Salisbury, B. A. & Stephens, J. C. Deconstructing the relationship between genetics and race. *Nat. Rev. Genet.* **5,** 598–609 (2004).
34 Shriver, M. D. & Kittles, R. A. Genetic ancestry and the search for personalized genetic histories. *Nat. Rev. Genet.* **5,** 611–618 (2004).
35 Parra, E. J., Marcini, A., Akey, J., Martinson, J., Batzer, M. A., Cooper, R. *et al.* Estimating African American admixture proportions by use of population- specific alleles. *Am. J. Hum. Genet.* **63,** 1839–1851 (1998).
36 Frudakis, T., Venkateswarlu, K., Thomas, M. J., Gaskin, Z., Ginjupalli, S., Gunturi, S. *et al.* A classifier for the SNP-based inference of ancestry. *J. Forensic. Sci.* **48,** 771–782 (2003).
37 Ausubel, F. M. *Current Protocols in Molecular Biology* (John Wiley & Sons, Inc, 1987).
38 Novick, G. E., Novick, C. C., Yunis, J., Yunis, E., Antunez de Mayolo, P., Scheer, W. D. *et al.* Polymorphic *Alu* insertions and the Asian origin of Native American populations. *Hum. Biol.* **70,** 23–39 (1998).
39 Nei, M. *Molecular Population Genetics* (Columbia University Press, New York, 1987).
40 Guo, S. W. & Thompson, E. A. Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics* **48,** 361–372 (1992).
41 Excoffier, L., Laval, G. & Schneider, S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* **1,** 47–50 (2005).
42 Reynolds, J., Weir, B. S. & Cockerham, C. C. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* **105,** 767–779 (1983).
43 Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425 (1987).
44 Felsenstein, J. PHYLIP: Phylogeny inference package (version 3.2). *Cladistics* **5,** 164–166 (1989).
45 Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39,** 783–791 (1985).
46 Excoffier, L., Smouse, P. E. & Quattro, JM. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131,** 479–491 (1992).
47 Sokal, R. R. & Rohlf, F. J. *Biometry. The Principles and Practice of Statistics in Biological Research* (WH Freeman and Company, New York, 1997).
48 García-Obregón, S., Alfonso-Sánchez, M. A., Pérez-Miranda, A. M., Vidales, C., Arroyo, D. & Peña, J. A. Genetic position of Valencia (Spain) in the Mediterranean basin according to Alu insertions. *Am. J. Hum. Biol.* **18,** 187–195 (2006).
49 García-Obregón, S., Alfonso-Sánchez, M. A., Pérez-Miranda, A. M., de Pancorbo, M. M. & Peña, J. A. Polymorphic Alu insertions and the genetic structure of Iberian Basques. *J. Hum. Genet.* **52,** 317–327 (2007).
50 Lalouel, J. M. Topology of population structure. in *Genetic Structure of Populations* (ed. Morton, N.E.) 139–149 (University of Hawaii Press, Honolulu, 1973).
51 Mantel, N. A. The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27,** 209–220 (1967).
52 Smouse, P. E., Long, J. C. & Sokal, R. R. Multiple regression and correlatione rxtensions of the Mantel test of matrix correspondence. *Syst. Zool.* **35,** 627–632 (1986).
53 Dietz, E. J. Permutation tests for association between two distance matrices. *Syst. Zool.* **32,** 21–26 (1983).
54 Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using Multilocus Genotype Data. *Genetics* **155,** 945–959 (2000).
55 Bonilla, C., Parra, E. J., Pfaff, C. L., Dios, S., Marshall, J. A., Hamman, R. F. *et al.* Admixture in the Hispanics of the San Luis Valley, Colorado, and its implications for complex trait gene mapping. *Ann. Hum. Genet.* **68,** 139–153 (2004).
56 Luizon, M. R., Mendes-Junior, C. T., De Oliveira, S. F. & Simões, A. L. Ancestry informative markers in Amerindians from Brazilian Amazon. *Am. J. Hum. Biol.* **20,** 86–90 (2008).
57 Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A. *et al.* Genetic structure of human populations. *Science* **298,** 2381–2385 (2002).
58 Jorde, L. B. & Wooding, S. P. Genetic variation, classification and 'race'. *Nat. Genet.* **36,** S28–S33 (2004).
59 Cruciani, F., Santolamazza, P., Shen, P., Macaulay, V., Moral, P., Olckers, A. *et al.* A back migration from Asia to sub-Saharan Africa is supported by high- resolution analysis of human Y-chromosome haplotypes. *Am.J. Hum. Genet.* **70,** 1197–1214 (2002).
60 Luis, J. R., Rowold, D. J., Regueiro, M., Caeiro, B., Cinnioğlu, C., Roseman, C., *et al* The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migrations. *Am. J. Hum. Genet.* **74,** 532–544 (2004).
61 Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton University Press, Princeton, 1994).
62 Scozzari, R., Cruciani, F., Santolamazza, P., Malaspina, P., Torroni, A., Sellitto, D. *et al.* Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. *Am. J. Hum. Genet.* **65,** 829–846 (1999).
63 Sewerin, B., Cuza, F. J., Szmulewicz, M. N., Rowold, D. J., Bertrand-Garcia, R. L. & Herrera, R. J. On the genetic uniqueness of the Ami aborigines of Formosa. *Am. J. Phys. Anthropol.* **119,** 240–248 (2002).
64 Olivieri, A., Achilli, A., Pala, M., Battaglia, V., Fornarino, S., Al-Zahery, N. *et al.* The mtDNA legacy of the Levantine early upper Palaeolithic in Africa. *Science* **314,** 1767–1770 (2006).
65 Pérez-Miranda, A. M., Alfonso-Sánchez, M. A., Peña, J. A. & Herrera, R. J. Qatari DNA variation at a crossroad of human migrations. *Hum. Hered.* **61,** 67–79 (2006).
66 Lahr, M. M. & Foley, RA. Multiple dispersals and modern human origins. *Evol. Anthropol.* **3,** 48–60 (1994).
67 Stringer, C. Palaeoanthropology: coasting out of Africa. *Nature* **405,** 24–27 (2000).
68 Wahlund, S. Zusammensetzung von Population und Korrelationserscheinung vom Standpunkt der Vererbungslehre aus betrachtet. *Hereditas* **11,** 65–106 (1928).
69 Watkins, W. S., Rogers, A. R., Ostler, C. T., Wooding, S., Bamshad, M. J., Brassington, A. M. *et al.* Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. *Genome Res.* **13,** 1607–1618 (2003).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)