Original Paper

Human Heredity

Hum Hered 2006;62:30–46
DOI: 10.1159/000095851

Received: April 18, 2006
Accepted after revision: July 25, 2006
Published online: September 21, 2006

Human Population Genetic Structure and Diversity Inferred from Polymorphic *L1 (LINE-1)* and *Alu* Insertions

D.J. Witherspoon[a] E.E. Marchani[a] W.S. Watkins[a] C.T. Ostler[a] S.P. Wooding[a] B.A. Anders[b] J.D. Fowlkes[b] S. Boissinot[c] A.V. Furano[d] D.A. Ray[b] A.R. Rogers[e] M.A. Batzer[b] L.B. Jorde[a]

[a]Department of Human Genetics, University of Utah Health Sciences Center, Salt Lake City, Utah, [b]Department of Biological Sciences, Louisiana State University, Baton Rouge, La., [c]Department of Biology, Queens College, Flushing, N.Y., [d]Laboratory of Molecular and Cellular Biology, NIDDK, National Institutes of Health, Bethesda, Md., [e]Department of Anthropology, University of Utah, Salt Lake City, Utah, USA

Key Words

Genetics/population genetics · Evolution · Bioinformatics/computational biology

Abstract

Background/Aims: The *L1* retrotransposable element family is the most successful self-replicating genomic parasite of the human genome. *L1* elements drive replication of *Alu* elements, and both have had far-reaching impacts on the human genome. We use *L1* and *Alu* insertion polymorphisms to analyze human population structure. ***Methods:*** We genotyped 75 recent, polymorphic *L1* insertions in 317 individuals from 21 populations in sub-Saharan Africa, East Asia, Europe and the Indian subcontinent. This is the first sample of *L1* loci large enough to support detailed population genetic inference. We analyzed these data in parallel with a set of 100 polymorphic *Alu* insertion loci previously genotyped in the same individuals. ***Results and Conclusion:*** The data sets yield congruent results that support the recent African origin model of human ancestry. A genetic clustering algorithm detects clusters of individuals corresponding to continental regions. The number of loci sampled is critical: with fewer than 50 typical loci, structure cannot be reliably discerned in these populations. The inclusion of geographically intermediate populations (from India) reduces the distinctness of clustering. Our results indicate that human genetic variation is neither perfectly correlated with geographic distance (purely clinal) nor independent of distance (purely clustered), but a combination of both: stepped clinal.

Introduction

The *LINE-1* (long interspersed element 1, or *L1*) retrotransposable element family is by far the most successful and enduring self-replicating genomic parasite of the human genome. *L1*s became established in the ancestors of mammals ~120 million years ago (mya), and today remnants of over half a million *L1*s constitute one-fifth of the human genome [1–4]. Intact *L1*s are ~6 kb in length and encode the proteins required for their own replication, which proceeds through a target-primed reverse transcription (TPRT) mechanism [5]. As a result of this mode of retrotransposition, many *L1* elements are severe-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0001–5652/06/0621–0030$23.50/0

Accessible online at:
www.karger.com/hhe

David J. Witherspoon
Department of Human Genetics, Eccles Institute of Human Genetics
University of Utah, 15 North 2030 East, Room 7225
Salt Lake City, UT 84112-5330 (USA)
Tel. +1 801 587 3094, Fax +1 801 585 9148, E-Mail david.witherspoon@utah.edu

ly truncated upon insertion, rendering them incapable of catalyzing their own replication. Fewer than one hundred *L1*s, mostly of the *L1Hs Ta* and *L1 preTa* subfamilies, continue to replicate and thereby create polymorphic insertions in the human population [6–12].

Alu elements are the most common *SINE*s (short interspersed elements) in the human genome. They are dimeric 300-bp sequences that evolved from the 7SL RNA component of the signal-recognition particle ~65 mya and became extremely successful parasites of *L1*s. They rely on retrotransposition proteins encoded by active *L1* elements in order to replicate [13]. The human genome now contains more than one million *Alu* insertions, accounting for about a tenth of the genome [2]. As with *L1*s, some young *Alu* elements continue to replicate and have spawned subfamilies of *Alu* insertions, many of which are polymorphic for their presence or absence [14]. Like the numerous and highly polymorphic canine *SINEC_Cf* repeats [15], the *L1* and *Alu* families of mobile elements have had a significant impact on the composition and structure of their host genome. *L1* and *Alu* elements continue to generate mutations by triggering ectopic recombination events and chromosomal rearrangements and by insertional disruption of genes [16].

LINE and *SINE* insertions have two uniquely valuable properties as markers for phylogenetic and population genetic analyses. First, they are virtually free of homoplasy: every observed insertion at a specific locus is identical by descent to the insertion created by a single transposition event. The probability that two insertions of the same element sequence will occur at the same site and then drift to any appreciable frequency is extremely small due to the low rate of insertion relative to the vast number of potential insertion sites [11, 17–22]. Since insertions are almost never precisely deleted [23], homoplasy due to reversion is also extremely rare [18].

The second advantage of *LINE* and *SINE* insertion markers is that the ancestral state of the locus is known to be the absence of the insertion [24]. Other often-used genetic marker types, such as single nucleotide polymorphisms (SNPs), restriction site polymorphisms (RSPs) and short tandem repeat polymorphisms (STRPs) suffer from higher probabilities of homoplasy and greater difficulty of confidently establishing the ancestral marker state. As a result, mobile element insertion polymorphisms have found increasing application in phylogenetic analyses [20–22, 25–29].

Alu insertion polymorphisms have been used to study patterns of human genetic diversity and to illuminate human demographic history. The unambiguously known ancestral state of *Alu* insertion loci allows the root of a population tree to be determined. This fact aided analyses of small numbers of *Alu* loci to strongly support an African origin for modern humans and allowed estimation of the effective size of the human population [24, 30, 31]. A much larger set of 100 polymorphic *Alu* insertion loci demonstrated the utility of *Alu* markers in inferring the continent of origin of individuals in a worldwide population sample [32] and further supported an African origin of modern humans [33]. We have previously shown strong correspondence between results obtained from RSP, STRP, and mtDNA markers and results obtained from 35 *Alu* insertion polymorphisms, supporting their utility in population genetic analyses [34].

To date, *L1* markers have not been widely used for population genetic studies. Sheen and coworkers [8] identified six polymorphic *L1* insertion loci and determined the insertion frequencies in a set of populations. The expected absence of authentic insertion homoplasy has been demonstrated for *L1*s in primates [18]. Recent studies have focused on identifying recent *L1* insertions and analyzing their genomic environment and chromosomal distribution [11, 35].

L1 insertion polymorphisms differ from other commonly investigated polymorphisms in several unique and important ways. Their mutational mechanism is unique. In particular, the *L1* and *Alu* subfamilies studied here were transpositionally active between 1 and 5 mya and are largely quiescent now [6, 11, 12, 25, 35–39]. Some *L1* elements, particularly those of full length, may not be selectively neutral [40]. And, as we describe here, the ascertainment of *L1* insertion polymorphisms in human populations has not been unbiased.

In this work, we obtain genotypes for a set of *L1* loci in a large set of individuals from many diverse populations worldwide. With the resulting data we examine the distribution of human genetic diversity, the distinctness and relatedness of human populations, and the congruence of conclusions drawn from *L1* polymorphisms with those drawn from *Alu* markers in the same populations. We demonstrate that, in spite of the unique features of *L1* polymorphisms, population structure results based on a large collection of *L1* polymorphisms are remarkably consistent with those based on other types of polymorphic elements.

Central to discussions of human genetic diversity is the question of whether human population structure is best described as 'clinal' or 'clustered' [e.g. 41, 42]. We address this question and explore how the statistical power of the data set influences the answer.

Subjects, Materials and Methods

Populations, Ascertainment and Genotyping

L1 and *Alu* Diversity Panels

A total of 75 unlinked polymorphic autosomal *L1* insertion loci of the *Ta* and *preTa L1* families were genotyped in 317 individuals from diverse populations. Subjects for whom genotypes could not be obtained for at least 90% of these loci were removed, leaving 272 individuals. One hundred polymorphic *Alu* loci had been previously genotyped in 445 individuals from the same populations, including 246 of the individuals in whom the *L1* loci were genotyped (26 individuals were typed at the *L1* loci only). Genotypes were available for at least 90% of the 100 *Alu* loci in each of the 445 individuals in that sample [33]. The locations of the African populations range from equatorial to southern Africa [see map of 33]. These samples (table 1) constitute our 'diversity panels' and are the basis for our population structure analyses.

L1 Ascertainment

Most of the *L1* and *Alu* loci genotyped in our diversity panels were ascertained first in 'ascertainment panels' as follows.

Fifty-four of the 75 *L1* loci genotyped in the Diversity Panel were chosen from a larger set originally identified by searching the draft human genome sequence for *L1* elements of the recently-active *Ta* [12] and *preTa* [11] subfamilies. This set of 54 loci was genotyped in ascertainment panels of 80 individuals, 20 from each of four geographical regions.

The *Ta L1* loci were genotyped in: (1) African Americans from Michigan (ALFRED Allele Frequency Database sample SA000494R [31]); (2) East Asians (Chinese, Taiwanese and Malaysian, SA000536O, SA000534M, SA000530I [31, 43]) or, for some loci, in Native Alaskans (Inuit, Aleut, and Native Amerindian, SA000497U [24, 44]); (3) Germans [22], and (4) Egyptians (Nile River Valley [36]). The *preTa L1* loci were genotyped in the African American, East Asian, and Egyptian panels as above, but in Swiss and French [45, 46] instead of Germans.

The remaining 21 *L1* loci were ascertained directly for polymorphism in panels of diverse individuals, rather than for presence in the human genome sequence and then polymorphism in ascertainment panels. Twenty of 21 were ascertained for presence in a four-member panel (Biaka, Druze, Chinese, and Melanesian) using an anchored PCR and cloning method designed to identify all *Ta L1* insertions [35], then screened for polymorphism in an eight-member panel composed of the above four plus a Maya, Mbuti, Cambodian, and a Karitiana individual. One additional locus was identified by differential display in the above ascertainment panels (unpublished data; method based on [8]).

Alu Ascertainment

The 100 autosomal *Alu* loci were ascertained as follows [33, 47, 48].

Twelve *Ya5*-subfamily *Alu* polymorphisms were identified due to their presence in well-studied genes. The remainder were chosen from larger sets constructed by searching genome sequence databases: 34 *Ya5* and 47 *Yb8* [47–49] and 2 *Yb9* and 5 *Yc1* [48]. The *Yb9* and *Yc1* loci were genotyped in 20 individuals each from the African American, Asian, Swiss/French, and Egyptian samples (above), while the *Yb8* and most *Ya5* loci were genotyped in the same panel, except that a sample of Greenland Natives (SA000496T [45]) was used in lieu of the Asian sample. Some *Ya5* loci were genotyped in a panel of Americans of northern European descent from Michigan instead of the above Swiss and French sample (SA000493Q [45, 48]).

Ascertainment-panel genotypes for 121 polymorphic autosomal *L1* loci (of the *Ta*, *Ta-0*, *Ta-1*, *preTA* and *preTA2* subfamilies) and 515 polymorphic autosomal *Alu* loci (*Ya5*, *Ya5a2*, *Yb8*, *Yb9*, and *Yc1* subfamilies) were obtained from the above references and from dbRIP [50]. *L1* and *Alu* polymorphisms of intermediate frequency in the ascertainment panels were chosen for genotyping in the diversity panels [33]. For the *Alu* data set, polymorphisms with minor allele frequencies (MAFs) as low as 5% across the ascertainment panel were accepted. For the later *L1* data set, MAFs >10% were required.

Genotyping

We used locus-specific primers and internal *L1* primers designed in previous studies [11, 12, 35] to perform genotyping by PCR and gel electrophoresis. Short (severely truncated) *L1* insertion loci were genotyped using a single PCR with two primers flanking the insertion site, whereas longer insertions required two reactions: one using primers flanking the insertion site to amplify empty alleles, and another using one flanking primer and one primer internal to the *L1* insertion to diagnose insertion alleles [8].

PCR conditions were based on [12, 35], with PCR extension times and annealing temperatures modified to optimize the results for the latter set. Generally, PCR was performed in 25 μl reaction volumes using 5 or 25 ng of template DNA, 2.5 pmol of each primer and 1.0 U Taq DNA polymerase in a solution of 200 μM of dNTPs, 50 mM KCl, 10 mM Tris-HCl (pH 8.4) and 1.5 mM $MgCl_2$. Samples were initially denatured at 94°C for 2 min, then cycled 25 to 35 times as follows: denaturing at 94°C for 15 s, annealing between 55 and 61°C for 30 s, extension at 72°C for 30–70 s, followed by a final extension period of 3 min at 72°C. The expected product sizes range from ~100 to 1,400 bp. Denaturing and extension times were doubled for product sizes >500 bp. PCR products were electrophoresed in 2 or 4% agarose gels, stained with 0.05 mg/ml ethidium bromide and visualized by UV fluorescence.

Data Analysis

Gene diversities *(h)* were estimated using Nei's [51] unbiased measure. Tests of Hardy-Weinberg equilibrium for genotypes within continental groups were performed using an unconditional Bayesian exact test [52] with a Bonferroni correction for multiple tests. Bifurcating trees of population relatedness with bootstrap support values were constructed from the pairwise matrix of genetic distances (Nei's genetic distance [53]; using the NEIGHBOR, GENDIST, and SEQBOOT programs of PHYLIP [54].

Structure Analysis

To examine population structure in the *L1* and *Alu* data sets, *structure* v. 2.1 [55, 56] analyses were performed five times each for each data set and parameter combination (below), with 50,000 burn-in iterations followed by 1,000,000 iterations to estimate model likelihoods for different numbers of ancestral populations, *k*. From each set of five runs, only the results of the run with the highest likelihood were used. Admixture was assumed to occur between populations (*noadmix* = 0). We used the '*F* model' of *structure* (*freqscorr* = 1) to model the expected correlation of allele frequencies between populations. The number of ancestral popu-

lations, *k*, was varied between one and seven. Defaults were used for all other parameters.

Structure Parameter Variations

We studied the effects of model choice and data set on the ability of *structure* to consistently infer individual ancestries. For these analyses, the *L1* and *Alu* data sets were combined to yield 175 loci typed in 246 individuals.

In order to examine the effect of the number of loci typed, data sets were constructed by resampling $L = 10, 20, \ldots, 160, 175$ loci at a time, with replacement (17 values). Allele frequencies can be assumed to be independent or correlated across populations. *Structure* can estimate either the probability that an individual is a member of a given ancestral population (if no admixture is assumed, *noadmix* = 1), or the proportion of that individual's genome that is derived from each population (with admixture allowed, *noadmix* = 0). We examined the results of all four combinations of these parameter choices. For each of the $17 \times 4 = 68$ parameter combinations, we applied *structure* to 20 data sets generated by resampling the specified number of loci (thus generating $68 \times 20 = 1{,}360$ data sets). In order to eliminate the effect of some *structure* trials that fail to approach an optimal solution, three *structure* runs were performed for each of the 1,360 combinations, and only the run with the highest likelihood of the three was retained for analysis. Each run began with 10,000 burn-in iterations and collected data through 20,000 more to allow convergence [42]. We repeated the above procedure with a data set lacking the 40 Indian individuals to study the effect of analyzing only more geographically separate populations.

Results

The *L1* data set consists of 75 *L1* loci genotyped in 272 individuals (table 1). A larger set of nearly 900 individuals, including nearly all of those genotyped here, was previously genotyped at 100 *Alu* insertion loci [table 1 in 33]. Table 2 shows the mean frequencies and gene diversities of the *L1* and *Alu* loci in the diversity panels, by continental group. Also shown are the mean frequencies and genetic diversities of *L1* and *Alu* loci previously typed in the 80-member ascertainment panels from which these loci were chosen. Most loci (97%) do not deviate significantly from Hardy-Weinberg equilibrium, although some deviation is expected due to the pooling of populations within continents. Figure 1A shows the distributions of the frequencies of *L1* and *Alu* loci in the diversity panels, and figure 1B shows the frequency spectra of the larger pool of loci previously genotyped in ascertainment panels.

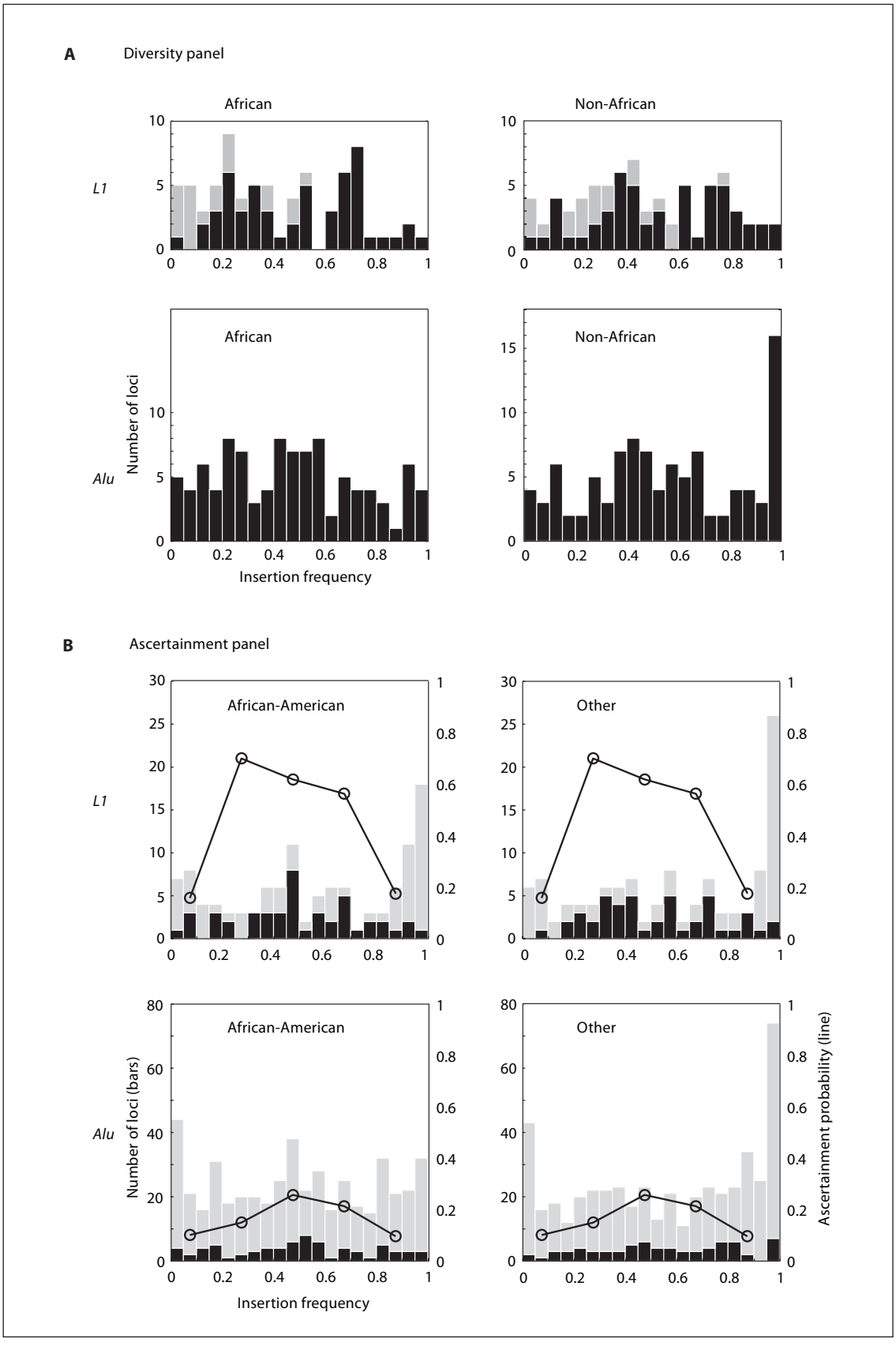
Ascertainment

Nearly all the *L1* and *Alu* loci represented in figure 1B were identified due to their presence in the human genome sequence. This genome is a composite constructed from the DNA of volunteers in California [2], so African alleles are probably underrepresented in that sample relative to our *L1* diversity panel. Therefore, *L1* and *Alu* polymorphisms at high frequency in non-African populations will be overrepresented relative to those that would be identified by complete ascertainment in our diversity panels.

The frequencies of the *L1* and *Alu* loci selected for further work from those genotyped in the ascertainment panel are shown in black in figure 1B. To maximize power for population structure analyses, we chose mostly polymorphisms of intermediate frequency in the ascertainment panel (see Methods). Figure 1B shows estimates of the ascertainment functions, computed as the fraction of initially ascertained loci selected for further genotyp-

Table 1. Sample sizes by population and data set

Continental group	Population	Individuals data	
		L1	*Alu*
Africa	Tswana, Sotho, Pedi or Xhosa	12	27
	Tsonga	7	6
	Nguni	11	9
	!Kung (San)	12	10
	Alur	10	9
	Hema	18	18
	Nande	18	18
	Pygmy (Tchabi, Idoho, Lolwa and Itende)	20	33
		108	130
East Asia	Cambodian	10	12
	Chinese	16	16
	Japanese	13	16
	Malaysian	6	6
	Vietnamese	6	9
		51	59
Europe	Northern European (from Utah)	40	59
	French (CEPH)	15	20
	Polish	5	9
	Finnish	13	19
		73	107
India	Brahmin (Andhra Pradesh)	10	60
	Madiga	10	29
	Irula	10	33
	Khonda Dora	10	27
		40	149
Total		272	445


A
Diversity panel
African
Non-African
L1
Alu
Number of loci
Insertion frequency
B
Ascertainment panel
African-American
Other
Number of loci (bars)
Ascertainment probability (line)
Insertion frequency

ing from each frequency quintile. The ascertainment effect can be seen in the difference between the more U-shaped spectra in the ascertainment panels (fig. 1B) compared to the flatter shape of the spectra in the diversity panels (fig. 1A). The effect is stronger in the set of *L1* loci than in the *Alu* set.

Allele Frequencies and Gene Diversities

L1 insertions have a significantly lower frequency in Africa than they do in Europe, India, or in the pooled non-African populations (table 2; $p < 0.05$, pairwise Wilcoxon-signed rank tests). *L1* frequencies are higher in East Asians (but not significantly so, $p \sim 0.056$, due to smaller sample size) and do not differ significantly among Asian, European, and Indian populations. *Alu* insertion frequencies do not differ significantly among Asian, European, and Indian populations, but they are significantly lower in African populations (table 2; $p < 0.0002$ in each pairwise comparison, as above) [see also 33].

The trend towards higher insertion frequencies in non-African samples is present in both the *L1* and *Alu* data sets, and even to some extent in the ascertainment panels between African-American and other populations. A much larger number of *Alu* loci are fixed or near fixation (frequency > 0.95) in non-African populations in the diversity panel. Partly because of this, the average gene diversity *(h)* of the *Alu* loci is higher in African samples [consistent with 33]. Gene diversity is also higher in African-American samples in the ascertainment panel (table 2).

In contrast, gene diversity at *L1* loci in the diversity panel is nearly constant across the four continental groups (table 2). This is due largely to the 21 *L1* loci ascertained for polymorphism in small panels. Since these loci were ascertained mainly in non-African individuals, it is not surprising that they have a lower frequency in the African populations (mean frequency $f = 0.19$) compared to the non-African samples ($f = 0.30$). The lower frequencies (further from 0.5) in the African populations result in lower gene diversity. The *Alu* data set has no counterpart to this low-frequency polymorphism set, and does not show this effect.

The proportion of genetic variation explained by differences between populations and continental population groups *(F_{ST})* is given in table 3. All F_{ST} values are significantly greater than zero ($p < 0.01$ by resampling). The results for *L1* and *Alu* loci are consistent with each other and with previous findings. As expected, the highest F_{ST} values arise in comparisons between African and non-African populations.

Genetic Distances

Pairwise genetic distances between populations (Nei's distance [53]) are very similar between the *L1* and *Alu* data sets. The Mantel matrix correlation coefficient for the two distance matrices (not including distances to the hypothetical root, which sharply inflate the correlation) is 0.88 ($p < 10^{-6}$ by resampling [57]). The pattern of genetic distances between populations is visualized using neighbor-joining trees in figure 2. The tree topologies based on the *L1* and *Alu* data are remarkably similar. In both trees, sub-Saharan Africa populations are strongly separated from non-African populations, and the 'root' population (constructed by setting the frequency of every insertion to zero, the ancestral state) is located in the African cluster adjacent to the !Kung. Asian and European populations form clusters with strong bootstrap support from the *Alu* data and weaker support from the *L1* data, while Indian populations branch out between the European and Asian groups, consistent with their geographical and historical origins.

Structure Analysis

The *structure* algorithm [55] was used to investigate how reliably individuals can be assigned to their continents of origin and to determine the number of distinct

Fig. 1. A Histograms of the frequencies (frequency spectra) of 75 *L1* and 100 *Alu* insertions in the human genetic diversity panel of African (left) and non-African populations (right; bin size 0.05). Non-African populations from Asia, Europe and India were pooled, since allele frequencies are not significantly different between those continental groups. The contributions of the 21 *L1* loci ascertained directly for polymorphism, rather than first for presence in the human genome sequence and then for polymorphism, are shown in gray. Since previously-known polymorphisms were excluded from this set, they were effectively ascertained for lower frequency. Consistent with this, their mean frequency in the diversity panel is 0.26, compared to 0.52 for the other 54 *L1* loci ($p < 4 \times 10^{-5}$, Wilcoxon-signed rank test). **B** Frequency spectra of 121 autosomal *L1* and 515 autosomal *Alu* insertion alleles in an ascertainment panel of African Americans (left, 20 individuals) and in a pool of East Asians, Native Americans, Europeans and Egyptians (right, 60 individuals), shown in light gray. Where data permits, loci selected for genotyping in the diversity panel (**A**) are shown in black (46 *L1* and 72 *Alu* loci shown). The fraction of loci chosen for further genotyping computed for each frequency quintile (the estimated ascertainment function) is shown according to the scale on the right-hand axes. Not shown are 299 fixed present *L1* insertions and 350 fixed present *Alu* insertions that were identified along with the polymorphic insertions.

Table 2. Mean frequencies and gene diversities of *L1* and *Alu* insertions in ascertainment and diversity panels

	Population				
	African	East Asian	European	Indian	Overall
*Insertion frequency in diversity panel**					
L1 (75)	0.41	0.47	0.46	0.48	0.44
Alu (100)	0.46	0.56	0.56	0.55	0.53
	Population				
	African American	East Asian/ Greenland Native/ Native Alaskan	Swiss/French/ European American/ German	Egyptian	Overall
Insertion frequency in ascertainment panel					
L1 (142)	0.57	0.60	0.59	0.58	0.59
Alu (157)	0.42	0.43	0.43	0.44	0.43
	Population				
	African	East Asian	European	Indian	Overall
Gene diversity in diversity panel					
L1 (75)	0.34	0.34	0.34	0.35	0.38
Alu (100)	0.35	0.31	0.30	0.31	0.35
	Population				
	African American	East Asian/ Greenland Native/ Native Alaskan	Swiss/French/ European American/ German	Egyptian	Overall
Gene diversity in ascertainment panel					
L1 (142)	0.28	0.24	0.25	0.25	0.26
Alu (157)	0.26	0.25	0.24	0.23	0.24

* Sample sizes for diversity panel in table 1. The ascertainment panels all consist of 80 individuals total, 20 in each of four groups given in the table.

Table 3. F_{ST} estimates for *L1* and *Alu* loci for several population groupings

	Africa, Asia, Europe, India	Asia, Europe, India	All 21 populations	8 African populations	13 Non-African populations
L1	0.122	0.0753	0.128	0.0586	0.0799
Alu	0.109	0.0519	0.107	0.0411	0.0527

populations required to explain the genetic structure in the data (fig. 3). Assuming four ancestral populations ($k = 4$), individual ancestry proportions inferred by *structure* coincide with the four continental groups (fig. 3A, B). If we assign each individual to the population that is estimated to contribute most to that individual's ancestry [as in 32], we find that their assignments correspond very well with their known geographic origins. On the basis of the *L1* data, nearly all Africans (99%) are assigned to a single population to which no other individual is as-

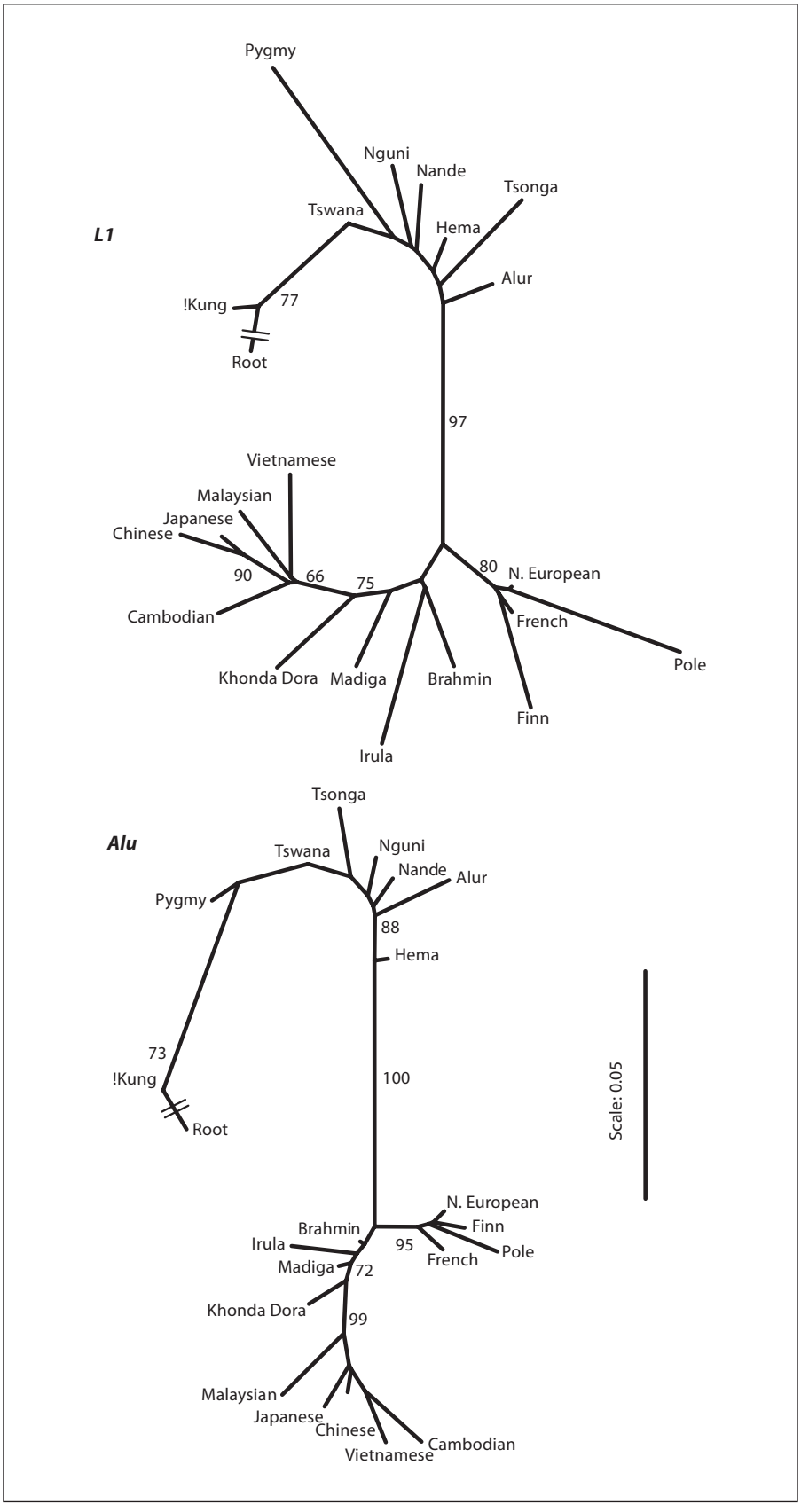
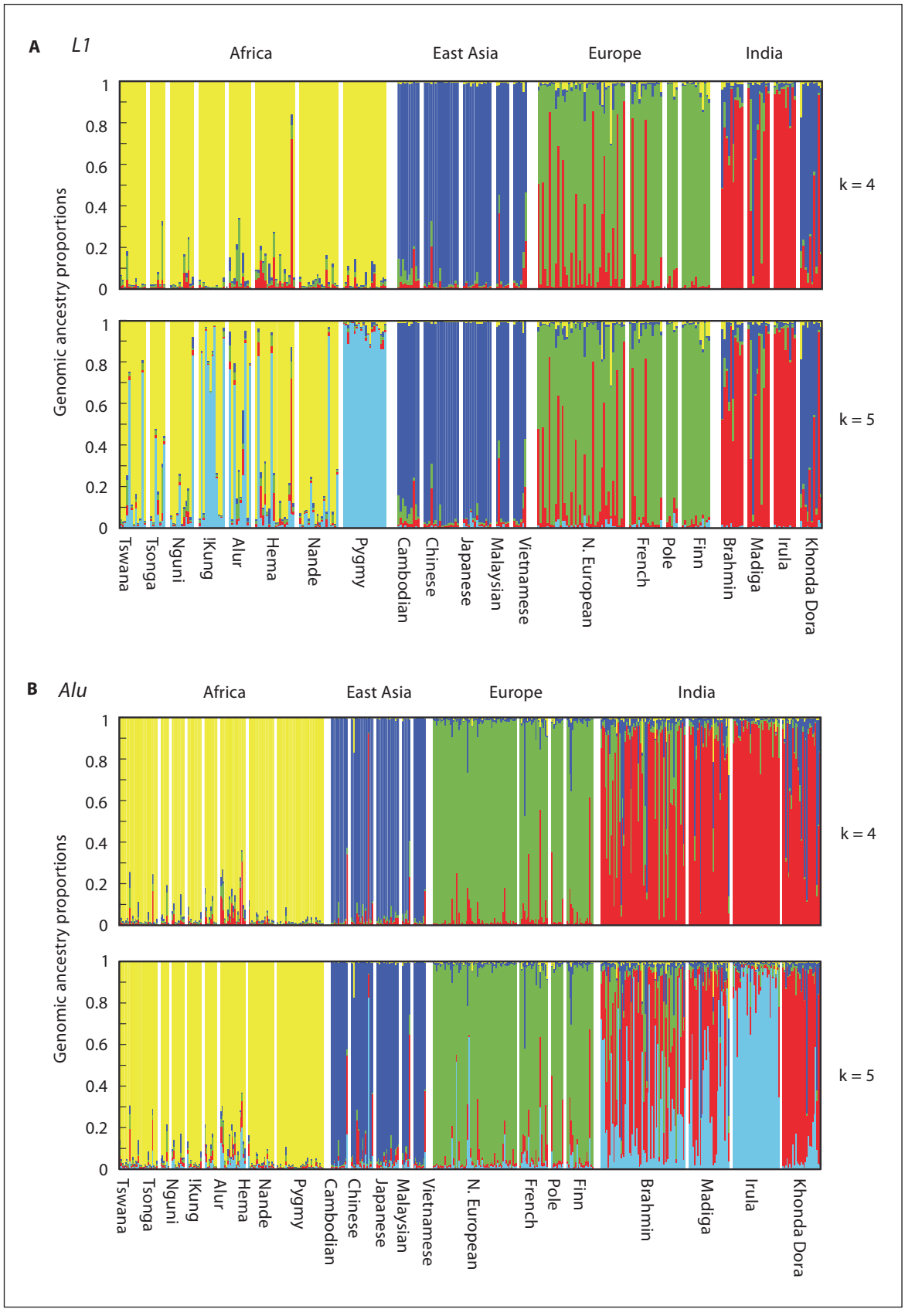


Fig. 2. Neighbor-joining networks based on pairwise Nei's genetic distances between populations calculated from *L1* and *Alu* loci. Bootstrap values (in percentages) are shown for clades present in >70% of 5,000 replicate trees generated from data sets constructed by resampling over loci.


A L1
Africa
East Asia
Europe
India
Genomic ancestry proportions
k = 4
k = 5
Tswana
Tsonga
Nguni
!Kung
Alur
Hema
Nande
Pygmy
Cambodian
Chinese
Japanese
Malaysian
Vietnamese
N. European
French
Pole
Finn
Brahmin
Madiga
Irula
Khonda Dora
B Alu
Africa
East Asia
Europe
India
Genomic ancestry proportions
k = 4
k = 5
Tswana
Tsonga
Nguni
!Kung
Alur
Hema
Nande
Pygmy
Cambodian
Chinese
Japanese
Malaysian
Vietnamese
N. European
French
Pole
Finn
Brahmin
Madiga
Irula
Khonda Dora

signed. All of the East Asian individuals are assigned to another population, 84% of Europeans are assigned to a third, and 73% of Indians are assigned to a fourth. These percentages are higher for the *Alu* data set (100, 98, 98, and 80%), probably due to the larger number of loci and individuals.

The greater uncertainty of assignment for Indians reflects their intermediate geographic position between Europeans and East Asians, which is also apparent in the genetic distance trees (fig. 2). In particular, many Khonda Dora (8 of 10 for the *L1* data, 6 of 27 for the *Alu* data) are grouped with East Asians rather than with other Indian populations.

The optimal number of ancestral populations for both the *L1* and *Alu* data is five, not four: the posterior probabilities of the models are essentially one for $k = 5$ and zero otherwise (assuming a uniform prior on $k = \{1–7\}$). Figure 3 shows the estimated ancestry proportions of individuals when five ancestral populations are assumed. For the *L1* data, all the Pygmy and some !Kung individuals are assigned to the fifth population, whereas the *Alu* data assign nearly all Irula (a non-caste 'tribal' population in India) and some other Indian individuals to a fifth population. In both cases, *structure* uses the fifth group to accommodate a sizable population sample that is relatively distinct from other populations (see fig. 2).

The classifications of individuals and the clustering of the populations inferred from the *L1* and *Alu* data by *structure* are consistent with previous results [32, 33, 42, 58, 59]. Sub-Saharan African individuals are readily distinguished from non-Africans, and East Asian, European, and Indian individuals are generally assigned into groups congruent with their populations of origin with high proportions of inferred ancestry in those groups.

Effect of Parameter Choices on Structure Results

Figure 4 shows the effects of modeling choices on population structure inferences. Three parameters were varied: the number of loci, the choice of populations, and whether or not allele frequencies are assumed to be independent or correlated between populations (the '*F*-model' [56]). Three result variables were examined for each individual: proportions of ancestry derived from each of k ancestral populations, assuming admixture; probabilities of membership in those populations, assuming no admixture; and the accuracy of their assignment to a population based on that probability [32].

The number of loci strongly affects all three measures. Ancestry proportions are relatively low when only 10 or 20 loci are used, indicating that most individuals are inferred to have substantial ancestry from several populations (red and yellow histograms, fig. 4A, B). Probabilities of membership are also low, indicating low certainty of assignment (light and dark blue histograms, fig. 4A, B). Similarly, the proportions of individuals correctly assigned to their continents of origin average only ~0.65 (green histograms, fig. 4C, D). The variances for all three measures are high when only 10 or 20 loci are used. However, as loci are added, the means increase substantially and variances decrease, nearing their asymptotic values with 50 loci.

Population choice also has a powerful impact on the results. Without the geographically intermediate populations of India, maximal proportions of ancestry, maximal probabilities of membership, and proportions of individuals correctly assigned rise more rapidly with increasing loci, and often to a higher asymptotic value (compare panels A and C to B and D in fig. 4). As in the *structure* analyses in figure 3, the Indian individuals are not as reliably classifiable.

Modelling allele frequencies as correlated or independent has no significant effect on any measure of structure in these analyses (no more than the expected number of signed-rank tests on paired result sets yield $p < 0.05$). Small effects might go undetected, given the modest number of replicates used [20], but large effects can be ruled out for these data.

Fig. 3. Posterior mean estimates of individual ancestry proportions in each of k presumed ancestral populations for the *L1* and *Alu* data sets. Each individual's genome is represented by a vertical bar composed of colored sections, where each section represents the proportion of an individual's ancestry derived from one of the k ancestral populations. Individuals are arrayed horizontally and grouped by population and continent of origin as indicated, separated by gaps for clarity (continental groups separated by wider gaps).

Discussion

The current working model of recent human origins – the Recent African Origin (RAO) model – posits that anatomically modern humans evolved from a population (N_e ~10,000) in Africa and then migrated out of Africa ~50,000–100,000 years ago, replacing pre-modern human populations with little or no genetic admixture [60, 61]. The African population remained sizable and thus retained substantial genetic diversity. Subdivi-

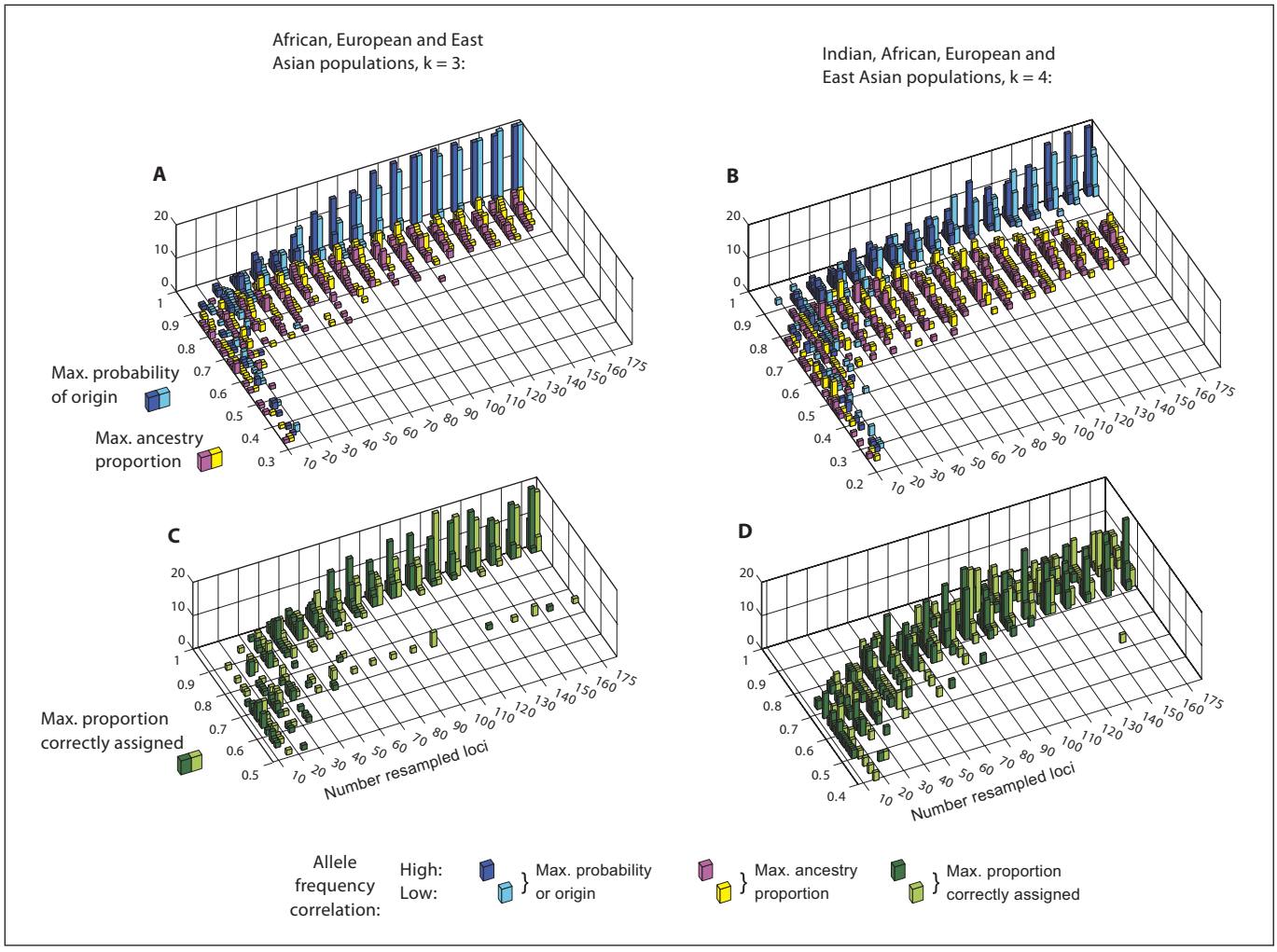


Fig. 4. Histograms of *structure* results (one histogram per set of 20 *structure* runs for a given number of loci and combination of parameters). In each run, *structure* estimates ancestry proportions (if admixture is assumed) or probabilities of membership (if admixture is not allowed) for each individual in each assumed ancestral population. For each individual, we use only the maximum ancestry proportion or probability of membership, and these maximal values are averaged across all individuals for each *structure* run. Average maximum ancestry proportions (red and yellow bars, bin size 0.05) and average maximum membership probabilities (light and dark blue bars) are shown in panels **A** and **B**. Panels **C** and **D** (green bars) show the proportion of individuals assigned to groups that correspond to their known geographical origins. The mappings of *structure*-inferred groups to continental groupings were chosen to minimize the number of mis-assigned individuals. The panels on the left (**A**, **C**) show results obtained using 206 individuals from Africa, East Asia, and Europe, with three ancestral populations presumed; panels on the right (**B**, **D**) show the results obtained using an additional 40 individuals from India and assuming four ancestral populations. The few persistent low values (near 0.75) in panel **C** are due to unusual *structure* runs in which one inferred population corresponds to the African Pygmy populations. This forces another inferred group to represent the remaining African individuals, and the third pools Europeans and East Asians, resulting in misclassification in that group.

sions either arose or were maintained within the African population, further preserving genetic diversity. The populations that migrated out of Africa lost genetic diversity to drift during repeated or severe population size bottlenecks. We assess our results in this context.

The pattern of human genetic diversity has been studied with a wide variety of genetic polymorphisms in diverse populations worldwide [reviewed in 61]. These markers include classical protein polymorphism loci [62, 63]; mtDNA haplotypes [64, 65]; Y chromosome STRPs, SNPs, and haplotypes [66–68]; autosomal RSPs

[69, 70]; autosomal STRPs [71–74]; autosomal SNPs [75, 76]; non-STRP, non-repetitive element indels [77]; *Alu* insertion polymorphisms [33, 34, 38, 78]; and a few *L1* insertion polymorphisms [8]. The degree of agreement on worldwide-scale questions across these marker types, analysis methods and population samples is striking. That congruence is also seen in our sample of *L1* loci, which represent another category of genetic system, distinct from others due to its own unique evolutionary dynamics.

Genetic Diversity

Studies of autosomal polymorphisms show that most genetic variation is found within populations, not between them: F_{ST} at the continental level ranges from 0.09 to 0.16 [71–73, 79]. Our estimates of F_{ST} from the autosomal *L1* loci fall comfortably in the established range. This level of differentiation is incompatible with large long-term population size or high gene flow between continental population groups. It also seems too high to fit a model of recent expansion (in population size and range) of humans from a homogeneous founder population [80, 81]. This suggests that the founding population was subdivided for some time before the archaeologically-recorded expansion (the 'weak Garden of Eden' hypothesis) or that human populations suffered repeated bottlenecks during their worldwide expansion [81].

Genetic diversity is generally highest in Africa. African mtDNA and NRY haplotypes are more divergent than non-African ones [67], and autosomal markers (whether protein polymorphisms, STRPs, RSPs, indels or *Alu* and *L1* insertion loci) have higher average heterozygosity in African than in non-African populations [34, 63, 71, 77]. This is consistent with larger long-term population size in African populations, reduced effective population size in non-African ones, and long-standing subdivisions among African populations [82, 83].

A few exceptions to the pattern of higher gene diversity in Africa have been noted, but these are due to ascertainment bias for high diversity in non-African populations [69, 80, 84]. The *Alu* data show the expected higher diversity in African populations, but our *L1* data do not. This is due to the inclusion of 21 low-frequency (and thus low-diversity) *L1* loci ascertained for polymorphism in small panels of mostly non-African individuals. Ascertainment of lower-diversity loci ($h < 0.35$) in small samples inflates estimates of diversity in the ascertainment population [85], so these 21 loci show higher diversity (as well as higher frequency) in the non-African populations (genetic diversity $h = 0.25$ in Africa, compared to $h = 0.30$ to 0.34 in East Asian, Indian or European groups). The remaining 54 loci show higher diversity in the African populations ($h = 0.38$ compared to $h = 0.34$ to 0.35), so the net result across all 75 loci is even diversity across continental groups.

Although the ascertainment biases affect the polymorphism frequency distributions in complex ways that render the data less useful for quantitative demographic inference [81, 82, 84, 85], the qualitative congruence remains. In particular, classification methods such as that implemented in *structure* are unlikely to be strongly affected by ascertainment biases.

An African Root

The deepest bifurcations in mtDNA and NRY haplotype trees are between lineages that are most common in Africa, and lineages that are more common outside Africa appear to be recently derived from African lineages [67, 76]. Genetic distance trees based on frequencies of RSP, STRP, *Alu* and now *L1* polymorphisms routinely show the longest branch separating African from non-African populations [34, 86]. Consistent with this, we find that F_{ST} estimates are largest in comparisons of African and non-African populations (table 3).

The *Alu* and *L1* data allow a hypothetical root population to be constructed by setting all insertion frequencies to zero, the known ancestral state. In both data sets, the !Kung population is nearest to the root (fig. 2) [see also 33]. The same was observed with a set of RSP markers [34]. Since genetic distance reflects the amount of drift that has occurred during the evolution of two populations, these results suggest that the !Kung population has maintained a larger long-term effective size in comparison with other human populations, and therefore has not drifted as far from the ancestral human population.

Comparable data are scarce for African populations [87]. Because of differences in mutation rates and effective population size, patterns of mtDNA and nonrecombining Y variation are not directly comparable to patterns of SNP, RSP, and insertion polymorphisms. Nonetheless, the !Kung have been placed on the earliest branch in a network of Y STRPs [66], and the oldest mtDNA lineage is most common in the !Kung and Biaka populations [76].

Clines and Clusters

Debate continues as to whether the fraction of human genetic variation attributable to population rather than

individual differences better fits a model of genetic isolation by distance (IBD) or an island model [63, p. 19]. Under IBD, frequent migration between adjacent subpopulations maintains smooth clines in allele frequencies over geographic distance. Under the island model, discontinuous shifts in allele frequencies will arise wherever linguistic, cultural, or discrete environmental barriers separate populations, resulting in genetic clusters. A model of strong IBD (e.g. low migration rates) maintained over a long period effectively becomes an island model with a little migration. Neither extreme is *a priori* likely, and both clinal and clustered patterns are observed at regional and worldwide scales [33, 34, 42, 63, 75, 78, 88–90].

Genetic distance and geographic distance are not independent, but neither are they perfectly correlated, as evidenced by the genetic distance trees. That broad correlation is clear evidence of a cline, the result of the original migration of anatomically modern humans out of Africa. Nonetheless, among pairwise comparisons of the four continental groups analyzed here, the ratio of genetic distance to geographic (great circle) distance – the slope of the putative genetic cline – varies by more than threefold [not shown; see also figure 7 in 33].

At a local scale, for example, the African-derived Siddi of southern India are genetically distant from their nearest geographical neighbors, resulting in a very steep local cline [78]. Recent mass migrations from Europe and Africa to the Americas have created still more disparities in the relationship between genetic and geographic distance. Therefore, to describe human population genetic variation purely in terms of clines would require many clines with different slopes: in some geographic regions, the cline would be flat; but across geographic or cultural barriers, much steeper clines are observed. Overall, the pattern is one of shallow clines within broad geographical regions, circumscribed by steeper steps between [33, 42, 88].

These discontinuities between populations are detected by *structure*, which displays them as large discontinuities in inferred ancestry proportions (fig. 3). *Structure* should infer the existence of distinct ancestral populations even from data that contains only smooth allele frequency clines [55, p. 956]. However, *structure* should not infer a discontinuous pattern of ancestry proportions for the sampled individuals (the pattern observed in fig. 3 and implied by the high average ancestry proportions in fig. 4). Instead, *structure* infers smooth gradients of ancestry proportions, mirroring the underlying allelic frequency clines. The simulation results in figure 5 demonstrate these expected gradients [see also 56]. Sampling populations only from the left, right, and center of the gradient of figure 5 would produce a very discontinuous pattern of ancestry proportions. Sampling from ten evenly spaced populations would replace those large discontinuities with nine smaller steps that approximate the smooth gradients in figure 5.

Our population samples are not so widely spaced or tightly clumped [see map in 33] as to create a large discontinuities where none exist. The CEPH Diversity Panel used by Rosenberg et al. [42] is still more evenly distributed. Including Indian populations in our analysis (compare left and right panels in fig. 4) or Middle Eastern populations in the analyses of Rosenberg et al. [42] reduces measures of population discreteness. This is consistent with the position of these populations on the known worldwide correlation of genetic and geographical distance [e.g. 33, 63, 88]. Nonetheless, discontinuities remain, with smaller distances between them, which require steeper gradients to bridge.

Serre and Paäbo [41] have argued that human genetic diversity is distributed in 'gradients of allele frequencies that extend over the entire world, rather than discrete clusters', and that findings of relatively discrete clusters [e.g. 58] are artifacts of sampling and model assumptions. However, attempting to infer population structure using a small number of loci will yield the appearance of clinal (or at least indistinct) variation from populations that are in fact distinguishable (see results for 10 or 20 resampled loci, fig. 4). This can explain the wide range of ancestry proportions inferred from the 20-locus data set collected by Serre and Paäbo [fig. 1C in 41]. As can be seen in figure 4A, 20 loci lack the statistical power to retrieve the very high ancestry proportions that are inferred with larger data sets for the same individuals. Twenty loci also fail to reliably classify individuals into the African, European and East Asian populations [see also 32, 42, 91].

Correlation of Allele Frequencies Across Populations

In our data, assuming correlated or independent allele frequencies between populations does not affect the results. This contrasts with the results of Rosenberg et al. [58], whose data were reanalyzed by Serre and Paäbo [41] under the assumption of independent allele frequencies, and with the more comprehensive analysis of Rosenberg et al. [42], who found that assuming correlated frequencies resulted in stronger clustering. The critical difference is that the population sample of Rosenberg et al. [42] includes more populations and more pairs of populations that are close geographic neighbors, whereas our diver-

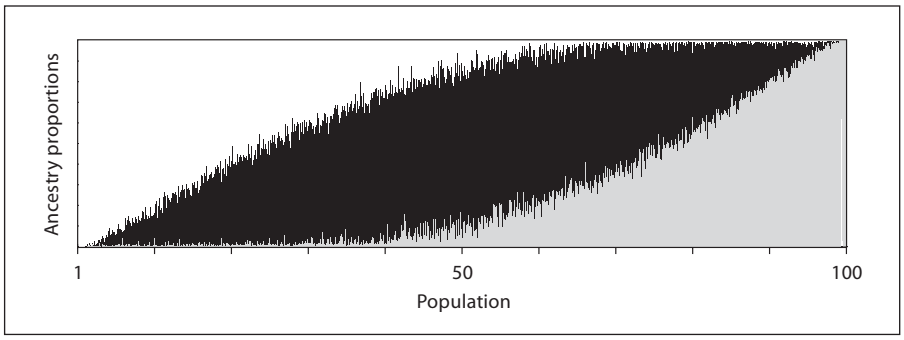


Fig. 5. Individual ancestry proportions inferred by *structure* for a simulated data set sampled evenly from a metapopulation with smooth allele frequency clines at 1,000 loci. The metapopulation consists of a linear array of 100 subpopulations of 5,000 haploid individuals each. Allele frequencies in the first and last subpopulations were resampled from the frequencies of the *L1* insertions in our African and East Asian samples, respectively, and held constant thereafter. Frequencies in the 98 intermediate populations were initialized by linear interpolation between the extremes, then allowed to vary through 3,000 generations of migration and drift. Each intermediate population was replaced every generation, with half of the new alleles sampled from the original population and one-quarter from each of the two adjacent populations. Ten individuals were sampled from each subpopulation to construct a data set for *structure* analysis. Correlated frequencies were assumed, k was varied from one to five, and 10,000 replicates following 5,000 burn-in replicates were performed. The optimum k in this case is three; lower migration rates and smaller population sizes produce greater isolation by distance between subpopulations and therefore higher optimal k.

sity panel is dominated by more distant populations. Allele frequency between adjacent populations are generally higher (e.g. Middle Eastern and European populations, with correlations coefficients as high as 0.96 [42]). The prior probability distribution of F (an F_{ST}-like parameter) used by default in *structure*'s model of correlation is centered on 0.01, implying very high allele correlations among populations. The values of F inferred from our *L1* and *Alu* data are more than an order of magnitude greater, indicating that the default prior had little influence on the results because the data strongly supported higher F values. Allele frequencies are in fact correlated even across major continental groupings in our data (Pearson's $r^2 > 0.55$ in all continental pairs), but the F-model is only expected to make a difference with very similar populations [56], such as some of those used by Rosenberg et al. [42]. The assumption of uncorrelated allele frequencies [as in 41] is highly unrealistic, but has little or no impact for our diversity panel.

Conclusion

The local discontinuities implied by the *structure* results, the broad correlation between genetic and geographic distance, and the many exceptions to that correlation point to a genetic cline created by the original migration of humans out of Africa, with initial genetic discontinuities created by founding events. Migration between populations was not sufficient to erase those discontinuities, and isolation and genetic drift may have enhanced them. Subsequent mass migration events have created still further discontinuities which have not been erased by admixture with preexisting populations. Thus the human population shows a pattern of clines with steps in them, which fits a model in between the pure IBD and island models.

Sampling of more geographically diverse populations, more individuals from each population, and more loci from each individual will allow inference of finer details of human demographic history. Analyses of population structure will benefit enormously from such data, and should eventually resolve the degree to which human populations are clinally and discontinuously related.

The relatively qualitative analyses presented here are not strongly affected by the varying strategies used to ascertain polymorphisms. However, in order to precisely estimate quantitative parameters in models of human demographic change or the evolution of different genetic marker systems, careful attention must be given to ascertainment. In particular, improved ascertainment strategies would allow a better understanding of *L1* and *Alu* transposition dynamics [e.g. 38], which in turn would al-

low better use of the unique information inherent in these genetic systems.

The congruence of answers to broad, worldwide questions across so many types of genetic markers, including markers with unusual mutational and evolutionary properties, lends confidence in their accuracy. The data continue to support the RAO model, although some admixture between migrating modern humans and resident archaic populations cannot be excluded [92].

Acknowledgments

We thank several anonymous reviewers for their accurate and helpful comments on this work. This research was supported by National Science Foundation BCS-0218338 (MAB), BCS-0218370 (LBJ), EPS-0346411 (MAB), National Institutes of Health GM-59290 (LBJ and MAB) and Louisiana Board of Regents Millennium Trust Health Excellence Fund HEF (2000-05)-05 (M.A.B.), (2000-05)-01 (M.A.B.) and (2001-06)-02 (M.A.B.), and by the Intramural Research Program of the NIH, NIDDK.

References

1 Smit AF, Toth G, Riggs AD, Jurka J: Ancestral, mammalian-wide subfamilies of *LINE-1* repetitive sequences. J Mol Biol 1995; 246:401–417.

2 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al: Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.

3 Kazazian HH Jr, Moran JV: The impact of *L1* retrotransposons on the human genome. Nat Genet 1998;19:19–24.

4 Kazazian HH Jr: Mobile elements: drivers of genome evolution. Science 2004;303:1626–1632.

5 Luan DD, Korman MH, Jakubczak JL, Eickbush TH: Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: A mechanism for non-LTR retrotransposition. Cell 1993;72:595–605.

6 Sassaman DM, Dombroski BA, Moran JV, Kimberland ML, Naas TP, DeBerardinis RJ, Gabriel A, Swergold GD, Kazazian HH Jr: Many human *L1* elements are capable of retrotransposition. Nat Genet 1997;16:37–43.

7 Boissinot S, Chevret P, Furano AV: *L1* (*LINE-1*) retrotransposon evolution and amplification in recent human history. Mol Biol Evol 2000;17:915–928.

8 Sheen FM, Sherry ST, Risch GM, Robichaux M, Nasidze I, Stoneking M, Batzer MA, Swergold GD: Reading between the *LINE*s: Human genomic variation induced by *LINE-1* retrotransposition. Genome Res 2000;10:1496–1508.

9 Ovchinnikov I, Rubin A, Swergold GD: Tracing the *LINE*s of human evolution. Proc Natl Acad Sci USA 2002;99:10522–10527.

10 Badge RM, Alisch RS, Moran JV: ATLAS: a system to selectively identify human-specific *L1* insertions. Am J Hum Genet 2003;72: 823–838.

11 Salem AH, Myers JS, Otieno AC, Watkins WS, Jorde LB, Batzer MA: *LINE-1 preTa* elements in the human genome. J Mol Biol 2003;326:1127–1146.

12 Myers JS, Vincent BJ, Udall H, Watkins WS, Morrish TA, Kilroy GE, Swergold GD, Henke J, Henke L, Moran JV, Jorde LB, Batzer MA: A comprehensive analysis of recently integrated human *Ta L1* elements. Am J Hum Genet 2002;71:312–326.

13 Dewannieux M, Esnault C, Heidmann T: *LINE*-mediated retrotransposition of marked *Alu* sequences. Nat Genet 2003;35: 41–48.

14 Batzer MA, Deininger PL: *Alu* repeats and human genomic diversity. Nat Rev Genet 2002;3:370–379.

15 Wang W, Kirkness EF: Short interspersed elements (*SINE*s) are a major source of canine genomic diversity. Genome Res 2005;15: 1798–1808.

16 Deininger PL, Batzer MA: *Alu* repeats and human disease. Mol Genet Metab 1999;67: 183–193.

17 Batzer MA, Deininger PL: A human-specific subfamily of *Alu* sequences. Genomics 1991; 9:481–487.

18 Ho HJ, Ray DA, Salem AH, Myers JS, Batzer MA: Straightening out the *LINE*s: *LINE-1* orthologous loci. Genomics 2005;85:201–207.

19 Roy-Engel AM, Carroll ML, El-Sawy M, Salem AH, Garber RK, Nguyen SV, Deininger PL, Batzer MA: Non-traditional *Alu* evolution and primate genomic diversity. J Mol Biol 2002;316:1033–1040.

20 Ray DA, Xing J, Hedges DJ, Hall MA, Laborde ME, Anders BA, White BR, Stoilova N, Fowlkes JD, Landry KE, Chemnick LG, Ryder OA, Batzer MA: *Alu* insertion loci and platyrrhine primate phylogeny. Mol Phylogenet Evol 2005;35:117–126.

21 Xing J, Wang H, Han K, Ray DA, Huang CH, Chemnick LG, Stewart CB, Disotell TR, Ryder OA, Batzer MA: A mobile element based phylogeny of Old World monkeys. Mol Phylogenet Evol 2005;37:872–880.

22 Salem AH, Kilroy GE, Watkins WS, Jorde LB, Batzer MA: Recently integrated *Alu* elements and human genomic diversity. Mol Biol Evol 2003;20:1349–1361.

23 van de Lagemaat LN, Gagnier L, Medstrand P, Mager DL: Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. Genome Res 2005;15:1243–1249.

24 Batzer MA, Stoneking M, Alegria-Hartman M, Bazan H, Kass DH, Shaikh TH, Novick GE, Ioannou PA, Scheer WD, Herrera RJ, et al: African origin of human-specific polymorphic *Alu* insertions. Proc Natl Acad Sci USA 1994;91:12288–12292.

25 Otieno AC, Carter AB, Hedges DJ, Walker JA, Ray DA, Garber RK, Anders BA, Stoilova N, Laborde ME, Fowlkes JD, Huang CH, Perodeau B, Batzer MA: Analysis of the human *Alu Ya*-lineage. J Mol Biol 2004;342: 109–118.

26 Terai Y, Takezaki N, Mayer WE, Tichy H, Takahata N, Klein J, Okada N: Phylogenetic relationships among East African haplochromine fish as revealed by short interspersed elements (SINEs). J Mol Evol 2004; 58:64–78.

27 Okada N, Shedlock AM, Nikaido M: Retroposon mapping in molecular systematics. Methods Mol Biol 2004;260:189–226.

28 Okada N: SINEs: Short interspersed repeated elements of the eukaryotic genome. Trends Ecol Evol (Amst) 1991;6:358–361.

29 Shedlock AM, Okada N: *SINE* insertions: powerful tools for molecular systematics. Bioessays 2000;22:148–160.

30 Sherry ST, Harpending HC, Batzer MA, Stoneking M: *Alu* evolution in human populations: using the coalescent to estimate effective population size. Genetics 1997;147: 1977–1982.

31 Stoneking M, Fontius JJ, Clifford SL, Soodyall H, Arcot SS, Saha N, Jenkins T, Tahir MA, Deininger PL, Batzer MA: Alu insertion polymorphisms and human evolution: Evidence for a larger population size in Africa. Genome Res 1997;7:1061–1071.

32 Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB: Human population genetic structure and inference of group membership. Am J Hum Genet 2003; 72:578–589.

33 Watkins WS, Rogers AR, Ostler CT, Wooding S, Bamshad MJ, Brassington AM, Carroll ML, Nguyen SV, Walker JA, Prasad BV, Reddy PG, Das PK, Batzer MA, Jorde LB: Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. Genome Res 2003;13:1607–1618.

34 Watkins WS, Ricker CE, Bamshad MJ, Carroll ML, Nguyen SV, Batzer MA, Harpending HC, Rogers AR, Jorde LB: Patterns of ancestral human diversity: an analysis of Alu-insertion and restriction-site polymorphisms. Am J Hum Genet 2001;68:738–752.

35 Boissinot S, Entezam A, Young L, Munson PJ, Furano AV: The insertional history of an active family of *L1* retrotransposons in humans. Genome Res 2004;14:1221–1231.

36 Roy AM, Carroll ML, Kass DH, Nguyen SV, Salem AH, Batzer MA, Deininger PL: Recently integrated human Alu repeats: Finding needles in the haystack. Genetica 1999; 107:149–161.

37 Batzer MA, Kilroy GE, Richard PE, Shaikh TH, Desselle TD, Hoppens CL, Deininger PL: Structure and variability of recently inserted Alu family members. Nucleic Acids Res 1990;18:6793–6798.

38 Hedges DJ, Cordaux R, Xing J, Witherspoon DJ, Rogers AR, Jorde LB, Batzer MA: Modeling the amplification dynamics of human alu retrotransposons. PLoS Comput Biol 2005;1:e44.

39 Batzer MA, Rubin CM, Hellmann-Blumberg U, Alegria-Hartman M, Leeflang EP, Stern JD, Bazan HA, Shaikh TH, Deininger PL, Schmid CW: Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. J Mol Biol 1995;247:418–427.

40 Boissinot S, Davis J, Entezam A, Petrov D, Furano AV: Fitness cost of *LINE-1* (*L1*) activity in humans. Proc Natl Acad Sci USA 2006; 103:9590–9594.

41 Serre D, Paabo S: Evidence for gradients of human genetic diversity within and among continents. Genome Res 2004;14:1679–1685.

42 Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW: Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. PLoS Genet 2005;1:e70.

43 Melton T, Peterson R, Redd AJ, Saha N, Sofro AS, Martinson J, Stoneking M: Polynesian genetic affinities with Southeast Asian populations as identified by mtDNA analysis. Am J Hum Genet 1995;57:403–414.

44 Newman WP, Middaugh JP, Propst MT, Rogers DR: Atherosclerosis in Alaska Natives and non-natives. Lancet 1993;341: 1056–1057.

45 Batzer MA, Arcot SS, Phinney JW, Alegria-Hartman M, Kass DH, Milligan SM, Kimpton C, Gill P, Hochmeister M, Ioannou PA, Herrera RJ, Boudreau DA, Scheer WD, Keats BJ, Deininger PL, Stoneking M: Genetic variation of recent Alu insertions in human populations. J Mol Evol 1996;42:22–29.

46 Monson KL, Moisan JP, Pascal O, McSween M, Aubert D, Giusti A, Budowle B, Lavergne L: Description and analysis of allele distribution for four VNTR markers in French and French Canadian populations. Hum Hered 1995;45:135–143.

47 Carroll ML, Roy-Engel AM, Nguyen SV, Salem AH, Vogel E, Vincent B, Myers J, Ahmad Z, Nguyen L, Sammarco M, Watkins WS, Henke J, Makalowski W, Jorde LB, Deininger PL, Batzer MA: Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. J Mol Biol 2001;311:17–40.

48 Roy AM, Carroll ML, Nguyen SV, Salem AH, Oldridge M, Wilkie AO, Batzer MA, Deininger PL: Potential gene conversion and source genes for recently integrated Alu elements. Genome Res 2000;10:1485–1495.

49 Carter AB, Salem AH, Hedges DJ, Keegan CN, Kimball B, Walker JA, Watkins WS, Jorde LB, Batzer MA: Genome-wide analysis of the human Alu Yb-lineage. Hum Genomics 2004;1:167–178.

50 Wang J, Song L, Grover D, Azrak S, Batzer MA: dbRIP: A Highly Integrated Database of Retrotransposon Insertion Polymorphism in Human. Submitted 2006.

51 Nei M: Molecular Evolutionary Genetics-New York, Columbia University Press, 1987.

52 Montoya-Delgado LE, Irony TZ, de B Pereira CA, Whittle MR: An unconditional exact test for the Hardy-Weinberg equilibrium law: Sample-space ordering using the Bayes factor. Genetics 2001;158:875–883.

53 Nei M: Genetic distance between populations. Am Nat 1972;106:283–292.

54 Felsenstein J: PHYLIP: Phylogenetic Inference Package, Version 3.5. 1993.

55 Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. Genetics 2000;155:945–959.

56 Falush D, Stephens M, Pritchard JK: Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 2003;164: 1567–1587.

57 Smouse PE, Chakraborty R: The use of restriction fragment length polymorphisms in paternity analysis. Am J Hum Genet 1986; 38:918–939.

58 Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW: Genetic structure of human populations. Science 2002;298:2381–2385.

59 Yang N, Li H, Criswell LA, Gregersen PK, Alarcon-Riquelme ME, Kittles R, Shigeta R, Silva G, Patel PI, Belmont JW, Seldin MF: Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: Application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine. Hum Genet 2005;118:382–392.

60 Stringer CB, Andrews P: Genetic and fossil evidence for the origin of modern humans. Science 1988;239:1263–1268.

61 Tishkoff SA, Verrelli BC: Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet 2003;4:293–340.

62 Corbo RM, Scacchi R: Apolipoprotein E (APOE) allele distribution in the world. Is APOE*4 a 'thrifty' allele? Ann Hum Genet 1999;63(Pt 4):301–310.

63 Cavalli-Sforza L, Menozzi P, Piazza A: The History and Geography of Human Genes. Princeton, Princeton University Press, 1994.

64 Torroni A, Rengo C, Guida V, Cruciani F, Sellitto D, Coppa A, Calderon FL, Simionati B, Valle G, Richards M, Macaulay V, Scozzari R: Do the four clades of the mtDNA haplogroup L2 evolve at different rates? Am J Hum Genet 2001;69:1348–1356.

65 Chen YS, Olckers A, Schurr TG, Kogelnik AM, Huoponen K, Wallace DC: mtDNA variation in the South African Kung and Khwe-and their genetic relationships to other African populations. Am J Hum Genet 2000;66:1362–1383.

66 Forster P, Rohl A, Lunnemann P, Brinkmann C, Zerjal T, Tyler-Smith C, Brinkmann B: A short tandem repeat-based phylogeny for the human Y chromosome. Am J Hum Genet 2000;67:182–196.

67 Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonne-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi SQ, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-Sforza LL, Oefner PJ: Y chromosome sequence variation and the history of human populations. Nat Genet 2000;26:358–361.

68 Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, Soodyall H, Zegura SL: Hierarchical patterns of global human Y-chromosome diversity. Mol Biol Evol 2001;18:1189–1203.

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

70 Jorde LB, Bamshad MJ, Watkins WS, Zenger R, Fraley AE, Krakowiak PA, Carpenter KD, Soodyall H, Jenkins T, Rogers AR: Origins and affinities of modern humans: A comparison of mitochondrial and nuclear genetic data. Am J Hum Genet 1995;57:523–538.

71 Deka R, Jin L, Shriver MD, Yu LM, DeCroo S, Hundrieser J, Bunker CH, Ferrell RE, Chakraborty R: Population genetics of dinucleotide (dC-dA)n.(dG-dT)n polymorphisms in world populations. Am J Hum Genet 1995;56:461–474.

72 Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL: High resolution of human evolutionary trees with polymorphic microsatellites. Nature 1994; 368:455–457.

73 Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA: The distribution of human genetic diversity: A comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 2000;66:979–988.

74 Jorde LB, Rogers AR, Bamshad M, Watkins WS, Krakowiak P, Sung S, Kere J, Harpending HC: Microsatellite diversity and the demographic history of modern humans. Proc Natl Acad Sci USA 1997;94:3100–3103.

75 Shriver MD, Mei R, Parra EJ, Sonpar V, Halder I, Tishkoff SA, Schurr TG, Zhadanov SI, Osipova LP, Brutsaert TD, Friedlaender J, Jorde LB, Watkins WS, Bamshad MJ, Gutierrez G, Loi H, Matsuzaki H, Kittles RA, Argyropoulos G, Fernandez JR, Akey JM, Jones KW: Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. Hum Genomics 2005;2:81–89.

76 Chen J, Sokal RR, Ruhlen M: Worldwide analysis of genetic and linguistic relationships of human populations. Hum Biol 1995; 67:595–612.

77 Weber JL, David D, Heil J, Fan Y, Zhao C, Marth G: Human diallelic insertion/deletion polymorphisms. Am J Hum Genet 2002; 71:854–862.

78 Watkins WS, Prasad BV, Naidu JM, Rao BB, Bhanu BA, Ramachandran B, Das PK, Gai PB, Reddy PC, Reddy PG, Sethuraman M, Bamshad MJ, Jorde LB: Diversity and divergence among the tribal populations of India. Ann Hum Genet 2005;69:680–692.

79 Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL: An apportionment of human DNA diversity. Proc Natl Acad Sci USA 1997; 94:4516–4519.

80 Rogers AR, Jorde LB: Genetic evidence on modern human origins. Hum Biol 1995;67: 1–36.

81 Harpending H, Rogers A: Genetic perspectives on human origins and differentiation. Annu Rev Genomics Hum Genet 2000;1: 361–385.

82 Relethford JH: Mutation rate and excess African heterozygosity. Hum Biol 1997;69:785–792.

83 Relethford JH, Jorde LB: Genetic evidence for larger African population size during recent human evolution. Am J Phys Anthropol 1999;108:251–260.

84 Eller E: Population substructure and isolation by distance in three continental regions. Am J Phys Anthropol 1999;108:147–159.

85 Rogers AR, Jorde LB: Ascertainment bias in estimates of average heterozygosity. Am J Hum Genet 1996;58:1033–1041.

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

87 Tishkoff SA, Williams SM: Genetic analysis of African populations: human evolution and complex disease. Nat Rev Genet 2002;3: 611–621.

88 Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL: Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci USA 2005;102:15942–15947.

89 Barbujani G, Sokal RR: Genetic population structure of Italy. I. Geographic patterns of gene frequencies. Hum Biol 1991;63:253–272.

90 Barbujani G, Sokal RR: Zones of sharp genetic change in Europe are also linguistic boundaries. Proc Natl Acad Sci USA 1990; 87:1816–1819.

91 Turakulov R, Easteal S: Number of SNPS loci needed to detect population structure. Hum Hered 2003;55:37–45.

92 Eswaran V, Harpending H, Rogers AR: Genomics refutes an exclusively African origin of humans. J Hum Evol 2005;49:1–18.