Genetics and Population History of Caucasus Populations

KAZIMA BULAYEVA,[1] LYNN B. JORDE,[2] CHRISTOPHER OSTLER,[2] SCOTT WATKINS,[2] OLEG BULAYEV,[1] AND HENRY HARPENDING[3]

Abstract We describe aspects of genetic diversity in several ethnic populations of the Caucasus Mountains of Daghestan using mitochondrial DNA sequences and a sample of 100 polymorphic *Alu* insertion loci. The mitochondrial DNA (mtDNA) sequences are like those of Europe. Principal coordinates and nearest neighbor statistics show that there is little detectable structure in the distances among populations computed from mtDNA. The *Alu* frequencies of the Caucasus populations suggest that they have undergone more genetic drift than most other groups since the dispersal of modern humans. Genetic differences among these populations are not large; instead, they are of the same order as distances among populations of Europe. We compare two methods of inference about the demography of ancient colonizing populations from Africa, one based on conventional F_{ST} statistics and one based on mean *Alu* insertion frequencies. The two approaches agree reasonably well if we assume that there was demographic growth in Africa before the diaspora of ancestors of contemporary regional human groups outside Africa.

In this paper we describe patterns of genetic differentiation among several populations of the Caucasus Mountains of Daghestan and compare them with a larger sample of human groups. The Caucasus Mountains, between the Black and Caspian seas, are astride what must have been a major corridor of movement since the expansion of modern humans. The inaccessible mountains may have functioned as a refuge and cul-de-sac off these migration streams. Today, ethnic groups in the Caucasus are characterized by extreme cultural and linguistic differentiation in a small geographic area. The groups are thought to be of great antiquity.

It is known from previous work (Barbujani et al. 1994) that Caucasus populations are not part of the system of gene frequency clines extending from Anatolia across Europe to the northwest. The inference is that they are not descendants of the Neolithic farmers whose expansion across Europe is responsible for

[1] Daghestan Branch, Russian Academy of Sciences, Makhachkala, Daghestan, Russia.
[2] Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah.
[3] Department of Anthropology, University of Utah, Salt Lake City, Utah.

Human Biology, December 2003, v. 75, no. 6, pp. 837–853.

KEY WORDS: DAGHESTAN, *ALU* INSERTION POLYMORPHISM, MTDNA, POPULATION STRUCTURE

the gene frequency clines. They may be, instead, descendants of the earlier "layer" of the population of Europe. In this they are like Basques, and indeed some linguists see a genetic relationship between Basque and Caucasian languages. The most extreme lumping (Ruhlen 1994) places Caucasian languages with Basque, a Siberian language with few speakers, Chinese and related languages, and the Athapascan languages of North America.

Daghestan is a southern Russian republic between the Black and Caspian seas. The southern two-thirds of Daghestan is in the Caucasus Mountains, reaching 2000–4000 meters above sea level. The northern third is a flat plain that extends along the western shores of the Caspian Sea (Figure 1). The republic of roughly 50,000 square kilometers has a population of about two million people. While many of them are urban, there remain many isolated ethnic groups that rely on subsistence agriculture, herding, and craft production, especially in the difficult Caucasus Mountains.

Many rural people live in remote mountain villages, known as auls, which have been geographically and reputed to be genetically isolated for thousands of years (Bulayeva 1991; Gammer 1994). These auls often exhibit unique customs, languages and dialects, and architectural styles. They are characterized by elevated rates of inbreeding, encouraged by Muslim traditions of marriages within families. Migration from highland to lowland regions has occurred for some of the groups, leading to outbred populations residing either in large lowland agricultural villages or cities. Within the auls valuable properties (e.g., farming terraces and sheep) usually were kept in the same family from one generation to the next by arrangement of marriages within the family.

The region has been predominantly Muslim since the 12th to 14th centuries. Before the introduction of Islam many groups were Christian (Aglarov 1988; Gadjiev 1971), but there was little Byzantine presence in the region. In the latter part of this millennium Daghestan was a locus of conflict among the Ottoman, Persian, and Russian empires.

The mountain auls have undergone remarkable linguistic and ethnic differentiation. There are auls of goldsmiths, woodcarvers, tinsmiths, boot-makers, dancers, singers, and many more. But the main occupations of highlanders are growing crops, primarily on hillside terraces, and stock raising, primarily sheep. Despite the harsh environment highlander groups have persisted for many centuries. In fact, some of them may have contributed to the initial exploitation of some important world crops on the hillside terraces (Vavilov 1936).

Materials and Methods

Populations. We describe HVS-I mtDNA sequences from five Daghestan populations: Kubachi, Novo-Kurush, Novo-Mehelta, Urkarah, and Stalskoe, as well as *Alu* insertion frequencies in a partially overlapping sample of populations:

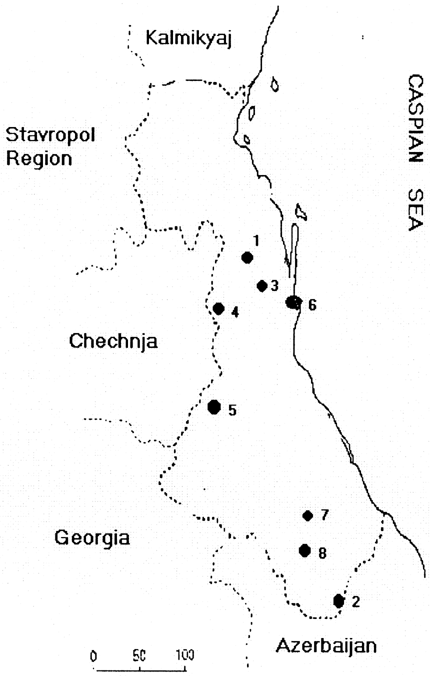

Figure 1. Map of Daghestan showing groups in the sample. 1: Novo-Kurush; 2: Kurush (Ethnic Lezgins); 3: Stalskoe (Ethnic Kumiks); 4: Novo-Mehelta; 5: Mehelta (Ethnic Avars); 6: Makhachkala (Mixed Ethnic, the capital city of the Daghestan); 7: Urkarah (Ethnic Dargins); 8: Kubachi (Ethnic Kubachians).

Kubachi, Urkarah, Stalskoe, Nogais, and Makhachkala. We compare both sets of data with comparable data from other populations.

- Makhachkala is the capital city of Daghestan, populated by people of all the ethnic groups of Daghestan and of many other Caucasian and Russian groups.
- Urkarah is one of the regional centers of Ethnic Dargins. The population of Urkarah is about 3000, half of whom are immigrants from neighboring smaller villages. Most of the population farms and raises sheep.
- Kubachi is a village of goldsmiths and silversmiths, well known in Europe and the East for this craft and, earlier, for arms and armor since the 11th century A.D. With a population of 2500, it is in the highlands at 2000 meters. In the last nine generations there were only ten marriages outside the village: nine men and one woman married out. Most marriages are between first and second cousins.
- Novo-Mehelta. Mehelta is one of the regional centers of largest Daghes-

tan Ethnic Group, Avars. In 1944 about half of the population of Mehelta was moved to a new settlement in the lowlands called Novo-Mehelta.

- Kurush is a highland aul of ethnic Lezgins at 3000 meters above sea level. They speak a unique dialect of the Lezgin language. In 1957 many inhabitants were forced to relocate to a new lowland village called Novo-Kurush.
- Stalskoe is a village of an aboriginal lowland ethnic group called Kumiks. They speak a dialect related to Turkish. Kumiks have a relatively low degree of inbreeding since they traditionally have been more open to intervillage marriages than other groups.
- Nogais are reputed to be descendants of the Nogai horde, a relict of the Mongol invasions of the early part of the last millennium.

Mitochondrial DNA. We have sequenced 410 base pairs of mitochondrial DNA (mtDNA) HVS-I in 114 individuals from five Daghestan groups. For comparison we used several hundred sequences from Europe, East Asia, Africa, and India described in Jorde et al. (1995); some additional Central African sequences from the Jorde laboratory from Hema, Alur, and Pygmies; Mongolian sequences from samples furnished by Ews Zeitkowics; and Georgians, Ingushians, Chechenians, Abazinians, Armenians, Azerbaijanians, and Cherkessians from the Caucasus region published by Nasidze and Stoneking (2001) and available at www.hvrbase.org. Altogether we had 1131 HVS-I mtDNA sequences. We then eliminated missing values and uninformative sites from the data by deleting any nucleotide position at which there were more than five sequences with missing values or at which the sample was monomorphic, then eliminating any sequence with any missing value. There remained 219 nucleotide positions in 1100 individuals for statistical analysis. Table 1 gives the sample sizes for each population.

Our statistical analysis follows Harpending and Jenkins (1973), treating each nucleotide position as a locus. Nucleotide frequencies at each position are normalized by division by $\sqrt{p(1-p)}$, where p is the world mean nucleotide frequency, yielding a $k \times l$ matrix Z. The k rows correspond to populations, while each of the l columns corresponds to an allele or, in the case of DNA sequences, a nucleotide position. The singular vectors of Z, each multiplied by the corresponding singular value, are then principal coordinates that can be plotted to show least squares optimum pictures of genetic distances among populations, while the distances themselves are computed simply as squared Euclidean distances between population centroids along the normalized frequency axes. A convenient way of doing the calculation is to compute $r = ZZ^t/\mathrm{l}$: the diagonal entries of r are genetic distances of each population from the overall centroid, the average of these is the statistic R_{ST} that we treat as an estimator of Wright's F_{ST}, and the genetic distance between populations i and j is just

$$d_{ij} = r_{ii} + r_{jj} - 2r_{ij}. \tag{1}$$

Table 1. Source Populations of MtDNA Sequences, Sample Sizes, Genetic Distances from the World Centroid, Nearest Neighbor, and Distance to Nearest Neighbor

Population	*Sample Size*	*Distance to Centroid*	*Nearest Neighbor*	*Distance to Nearest Neighbor*
Mongolians	19	0.15	French	0.18
Chinese	16	0.06	Middle caste	0.09
Japanese	20	0.08	Middle caste	0.11
Kubachi	27	0.17	French	0.19
Malay	6	0.12	N. European	0.15
Vietnamese	9	0.07	N. European	0.11
Hema	18	0.11	Nande	0.14
Novo-Mehelta	31	0.03	Georgian	0.08
Cambodian	12	0.09	French	0.12
Upper caste	61	0.06	Middle caste	0.04
Nande	18	0.05	Nigerian	0.10
Stalskoe	28	0.04	N. European	0.07
Abazinian	74	0.13	N. European	0.06
Middle caste	112	0.06	Upper caste	0.04
Alur	9	0.03	Pygmy	0.14
Finns	20	0.04	N. European	0.06
Georgian	53	0.06	N. European	0.03
Armenian	76	0.05	N. European	0.05
Lower caste	67	0.06	Middle caste	0.07
Azerbaijanian	32	0.05	Armenian	0.06
Cherkessian	44	0.06	Georgian	0.06
Italians	17	0.04	N. European	0.04
Urkarah	29	0.05	N. European	0.05
Poles	10	0.04	N. European	0.04
N. European	69	0.03	French	0.03
Chechenian	23	0.05	N. European	0.05
Ingushian	35	0.06	N. European	0.05
French	20	0.03	N. European	0.03
Nigerian	24	0.11	Nande	0.10
Novo-Kurush	24	0.06	N. European	0.05
San	14	0.17	Nguni	0.15
Tsonga	14	0.09	Nguni	0.06
Nguni	13	0.10	Tsonga	0.06
Pygmy	37	0.19	Alur	0.14
Sotho/Tawa	19	0.17	Nguni	0.09

This computation procedure is algebraically equivalent to other standard procedures for studying sequence data. For example, the genetic distance between populations is the mean pairwise difference between them less the mean within-population pairwise difference, divided by the overall mean pairwise difference.

Many statistical analyses suppose that populations are drawn from a larger universe of populations, leading to bias corrections of various kinds. We treat the sample as a world and do not do any such bias corrections. Instead, we view the analysis simply as geometry in several dimensions.

***Alu* Insertion Polymorphisms.** We describe frequencies at 100 *Alu* insertion polymorphisms from 184 individuals of five Daghestan populations. Details of the ascertainment and typing procedures along with comparative data from European, African, Indian, and East Asian populations are given in Watkins et al. (2002). These loci, scattered widely over the nuclear genome, were ascertained by finding them in sequence from the Human Genome Project; that is, they were each ascertained in a single human chromosome. An important characteristic of *Alu* markers is that the polarity of the locus is always known: the ancestral state is the absence of the *Alu*. The ascertainment mechanism together with the polarity must be accounted for in the analysis of these loci, so some of our methods may be unfamiliar.

Rogers and Harpending (Rogers and Harpending, in preparation) discuss a model in which an array of populations is descended from an ancestral source population. *Alu* insertions in this source population varied in frequency according to some distribution determined by population size and history. If in the ancestral population a large sample of *Alu* loci were discovered or ascertained by scanning a single chromosome, then the mean frequency of the insertion in the ascertained loci is called the "biased mean" frequency Π of *Alu*s in the ancestral population. In a population that has been of constant size for a long time, the distribution of the biased frequencies is uniform so that the mean insertion frequency is Π = 0.5.

We cannot observe this ancestral frequency. Instead we scanned for *Alu*s in single chromosomes derived from a contemporary ascertainment population, then tabulated insertion frequencies in this population, finding that the mean insertion frequency is P_a. Rogers and Harpending show that

$$P_a = \Pi + (1 - \Pi)r_{aa}, \tag{2}$$

where r_{aa} is the normalized or Wahlund variance of the ascertainment population, proportional to the total amount of genetic drift since the separation of the population from the ancestor (Harpending and Jenkins 1973). Similarly, the mean *Alu* frequency in another population b that is not the source of the ascertainment chromosome panel is

$$P_b = \Pi + (1 - \Pi)r_{ab}, \tag{3}$$

where r_{ab} is the normalized or Wahlund covariance between populations *a* and *b*. Notice that if daughter populations *a* and *b* have been separated with no intermixture since their origin, then the covariance is zero and the mean frequency in population *b* of *Alu*s ascertained in chromosomes from population *a* is an estimate of the ancestral biased frequency Π.

***Alu* Simulations.** We simulated *Alu* insertion frequencies using a standard coalescent algorithm (Hudson 1990) that allows stepwise changes in population size modified to simulate several subpopulations among which gene flow occurs. Our procedure was to repeatedly generate a gene tree and to choose a location uniformly distributed along the total branch length of the tree for an *Alu* insertion to occur. Each simulated tree was accepted with probability equal to the frequency of the *Alu* insertion in the ascertainment population in order to mimic our ascertainment procedure. When computing statistics about the sample of collected trees we weighted each tree according to the total branch length of the tree, since the probability of any *Alu* insertion is proportional to branch length.

Results

Mitochondrial Results. Figure 2 shows the least squares best two-dimensional picture of mtDNA genetic distances among the six major population groups: East Asia, Europe, Africa, India, Caucasus, and Daghestan. It is apparent from the figure that both our mtDNA sequences from Daghestan and those of Nasidze and Stoneking (Nasidze and Stoneking 2001) are essentially European. Finer scale analysis of principal coordinates of our sample is rather uninformative. If we drop Africa from the computation, for example, the portrayal of distances is dominated by the difference between Asia and all the others. Dropping Asia, the dominant feature is the separation of the Daghestan population of Kubachi from all the others. We find that each successive coordinate essentially describes a single population.

Table 1 shows basic characteristics of our sample including, for each population, genetic distance to the world centroid and genetic distance to the nearest neighbor. Half the populations are closer to the world centroid than to the nearest neighbor. Further, nearest neighbors seem not to fall into any coherent pattern except within Africa. For example, the sequences most similar to those of Mongolians are those of the French. The Malays and Vietnamese are closest to northern Europeans, while the Chinese and Japanese are closest to middle-caste Indians. The pattern is that almost all the populations are sitting on something like a high-dimensional sphere and there is little or no coherent grouping. Since mtDNA is a single locus, we should not be surprised to see such poor resolution of population relationships. This view of genetic distances computed from mtDNA shows that while interpretable patterns always emerge from principal coordinates analysis, they should be viewed with caution when they are derived from what is essentially a single locus.

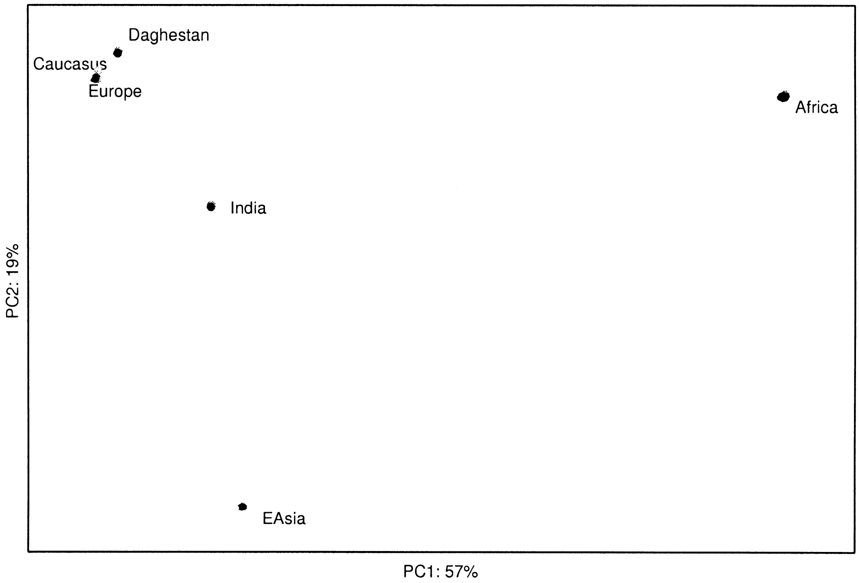

Figure 2. Principal components diagram from mtDNA sequence differences showing genetic distances among five population groups. "Caucasus" is the centroid of the samples from Nasidze and Stoneking (2001), "Daghestan" is the centroid of our Daghestan populations, and the others are the centroids of groups given in Jorde et al. (1995).

Estimates of F_{ST} from mtDNA should not be directly compared to estimates from nuclear markers. The effective size of mtDNA is roughly a quarter of that of nuclear loci, and the mutation rate is much higher. The former should make F_{ST} larger, the latter should make it smaller. Computed from the mtDNA sequence differences, F_{ST} among all populations is 0.081, while among the six group centroids it is 0.026. Among the populations in our Daghestan sample, F_{ST} is 0.073, among the Caucasus populations it is 0.025, and among the European populations in the Jorde sample it is 0.043. For comparison, among African populations it is 0.105, among East Asian populations it is 0.075, and among Indian populations 0.012. Mitochondrial diversity with Daghestan is high, second only to that with the African populations, while diversity within the Caucasus sample of Nasidze and Stoneking is lower than that within Europe. (This contradicts the Nasidze and Stoneking finding that mtDNA diversity within their Caucasus sample was higher than diversity within Europe. Our sample of populations that we take to represent Europe is different from their sample of European populations.)

The relatively high between-population diversity among the Daghestan groups supports the hypothesis that they have been small and genetically isolated from each other for a long time. On the other hand, we show below using 100 *Alu* loci that there is no elevated among-group diversity in Daghestan at all. Mitochondrial and nuclear markers might respond differently to brief bottlenecks because of the one-to-four difference in effective size, and they might differ because of sex differences in migration among groups. However, simulations of F_{ST} in subdivided populations show that the variance from locus to locus is very large and that the apparent discrepancy here between the mtDNA and the *Alu* findings most likely reflects statistical fluctuation.

Figure 3 shows mtDNA mismatch distributions from the Jorde laboratory with the Daghestan populations labeled with capital letters. The pattern is that mismatch means are highest in Africa, lower in Asia, and lowest in Europe. The populations from Daghestan show higher mean pairwise differences than those from Europe, but lower than those from Africa, supporting the idea that Daghestan populations are in a sense "older" than European populations to their west. This idea is in accord with the observation that Daghestan populations are not part of the large-scale cline across Europe thought to represent the expansion of Neolithic farmers.

***Alu* Results.** Equations (2) and (3) show that biased gene frequencies of *Alu*s can be regarded as covariances rather than as simple gene frequencies. On the other hand, our simulations show that it is not unreasonable to treat them as ordinary gene frequencies for the purposes of computing genetic distances, principal coordinates, and such, but there is no good theory yet about what the results mean. For example, we have found from simulations that F_{ST} statistics computed from biased *Alu* frequencies are essentially the same as those computed from nuclear markers other than *Alu*s.

We show in Figure 4 a principal coordinates diagram portraying genetic distances among five Daghestan populations together with group centroids of Europe, Asia, and India. The dominant feature along the first coordinate is the separation of Asia from the other populations, and the dominant feature along the second is the separation of India. The Daghestan populations are very close to Europe except for two, the Nogais and the Kubachians. The Nogais are much closer than the Kubachians to Asia in agreement with their reputed origin as a relict of the Mongol invasions. The Kubachians are simply divergent from the other Daghestan populations as well as the continental centroids, and this divergence was also apparent in their unusual mtDNA sequences.

With all the separate populations in our groups included, the computed F_{ST} for the world is 0.132, while among the five major group centroids it is 0.086: approximately two-thirds of the diversity is among our major groups, while the rest is within them. F_{ST} values within major groups are: Africa, 0.075; Asia, 0.049; India, 0.047; Europe, 0.031; and Daghestan, 0.029. Daghestan shows no more among-population diversity than Europe and conspicuously less than the other continents. The indication from mtDNA of elevated among-population diversity in Daghestan is likely to be a statistical artifact, but it might reflect sex differences

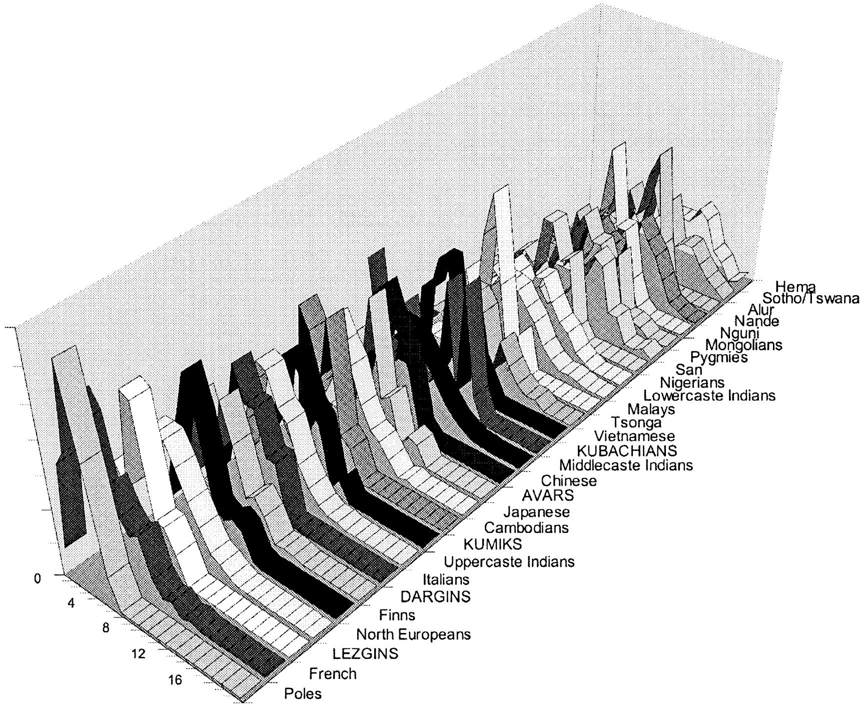

Figure 3. Comparison of some world mtDNA mismatch distributions. The Daghestan populations of our study are labeled in capital letters.

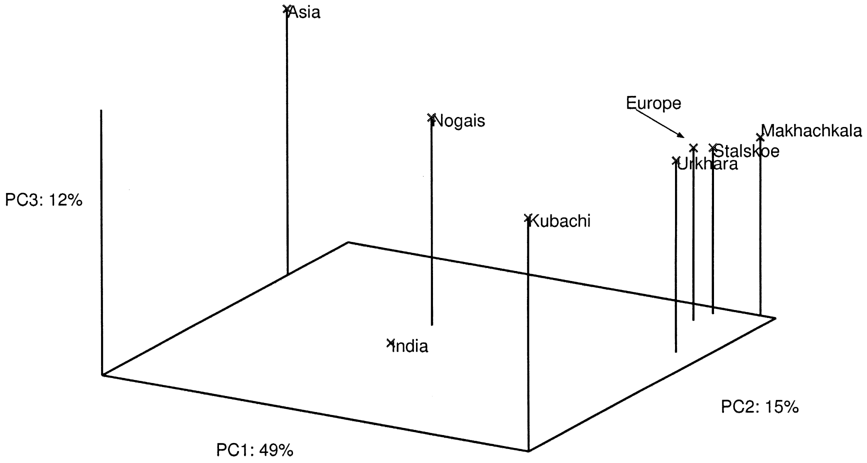

Figure 4. Principal components diagram showing genetic distances among five Daghestan populations and the group centroids of populations of Europe, Asia, and India.

in migration or the fourfold difference in effective size between nuclear and mitochondrial DNA.

Table 2 shows the mean frequency of *Alu* insertions in each of the populations we have studied along with group means in boldface. The high mean *Alu* frequencies of the Daghestan populations are striking. In spite of the apparent similarity in terms of genetic distance to populations of Europe, mean *Alu* frequencies in Daghestan populations are conspicuously higher than those of other European populations and are matched only by the Vietnamese and Cambodians. African populations, on the other extreme, have low mean frequencies. The world seems to be divided into an African clump and a non-African clump, with the only intermediate populations the tribal populations of India, Mariagond, and Santal. African populations in our sample have mean *Alu* frequencies around 0.46, as shown in Table 2, while the mean frequency of *Alu*s in European populations is 0.56 and for Daghestanis slightly higher. This is a suggestion that these Indian tribal groups are representative of some earlier "layer" of colonization of that subcontinent because they have a lower covariance with the other large populations of Europe, Asia, and India (see equation 3). The Bushmen have the world's lowest mean *Alu* frequency, but the difference between them and other African populations is not statistically significant.

Discussion

Mitochondrial Sequences. The results from analysis of mtDNA sequence diversity do not merit much discussion. While principal coordinates do display some structured relationships among populations, examination of nearest neighbors of populations showed that the strongest inference that ought to be made is that African populations are more similar to each other than they are to populations outside Africa. While F_{ST} among Daghestan populations is higher than the same statistic among the populations of Europe, there was no such pattern when we computed F_{ST} from 100 *Alu* loci.

The mean pairwise divergence of Daghestan mtDNA sequences is higher than that of all other European populations, suggesting that the Daghestan populations were established earlier than those of Europe. Europeans, according to mtDNA, are considerably "younger" than the populations of Asia, India, and Daghestan.

Average *Alu* Frequencies and the Multiregional Hypothesis. There are two aspects of the *Alu* insertion frequency data that need to be explained. First, why is the biased mean *Alu* frequency roughly 0.5, and, second, why is the mean *Alu* frequency low in Africa and high in Daghestan and in several populations from southeast Asia?

Standard coalescence theory has an answer for the first question. In the case

Table 2. Population Mean *Alu* Frequencies[a]

Mean	*Population*
0.43	San
0.44	Mbuti, Tsonga
0.45	Biaka, Sotho
0.46	Zaire Pygmy. **Africa,** Nguni
0.47	Nande
0.48	Mariagond
0.49	Alur
0.50	Hema, Santal
0.51	
0.52	
0.53	Khonda Dora
0.54	Vysya, **India,** Japanese
0.55	Brahmin, Mixed Asian, Relli, Madiga, Mala, Poles, Malays, Finns, Kapu
0.56	Stalskoe, **Europe,** Chinese, French, Northern European, Kshatriya, Yadava, Asia
0.57	Kubachi, Irula, Urkarah, Vietnamese, **Daghestan,** Nogais
0.58	Makhachkala, Cambodian

[a] Group means are shown in boldface. The standard deviation of any population's position on this chart is approximately 0.03, that is, three lines up or down. The standard deviation of the difference between any two groups is about 0.02. These are computed from the differences among loci and do not include the contribution of error in gene frequency estimation within groups, which is small in comparison.

of a constant size population it is well known (Watterson 1975) that the frequency spectrum of neutral *Alu* insertions should follow

$$G'(p) \propto 1/p = \frac{1}{pC_N} \tag{4}$$

where

$$C_N = \sum_{i=1/N}^{(N-1)/N} 1/i.$$

This is the unbiased spectrum, which we could not observe unless we examined every single chromosome in a population for *Alu* insertions. Instead, we find loci with probability equal to the *Alu* frequency at that locus so that the biased frequency spectrum $G(\pi)$ would be uniform,

$$G(\pi) = \frac{1}{N-1}, \quad \pi = 1/N \ldots (N-1)/N.$$

with mean 0.5, in reasonable agreement with the world data. This is an important observation since it falsifies the multiregional model of the origin of modern humans (Hawks et al. 2000).

Inferences from many genetic systems agree that the long-term effective size of humanity is ten to twenty thousand (Harpending et al. 1998; Harpending and Rogers 2000). One perhaps unlikely interpretation of this figure is that there were only several tens of thousands human ancestors during the middle and upper Pleistocene.

According to the multiregional hypothesis, our apparent small effective size is a consequence of bottlenecking during the genesis and dispersal of *Homo erectus* approximately 1.8 mya, so that many nuclear gene trees coalesce just prior to this time. The expected time to the most recent common ancestor of a nuclear locus is 4N generations, according to standard coalescence theory. Assuming 25 years per generation, 1.8 mya corresponds to 72,000 generations, yielding an estimate of human effective size of 18,000. According to this model the population ancestral to modern humans may have been large during the middle and upper Pleistocene, but the bottleneck associated with the appearance of *Homo erectus* compresses the tops of nuclear gene trees, thereby mimicking shorter coalescence trees generated by a small population.

Alu frequencies now provide a simple test of this hypothesis because gene trees with shallow tops should lead to an excess of low frequency *Alu*s in contemporary populations. With the top of the gene tree drastically shortened, there would be a relative shortage of old common insertions in contemporary populations.

The effect of multiregional evolution on *Alu* frequencies is simple to evaluate with the simulation. If we postulate that the effective size of our ancestors was 30,000 throughout the Pleistocene, rather than 10,000, then without the dispersal event at 1.8 mya the mean time to MRCA for gene trees would be 120,000 generations or 3.75 million years. Under the multiregional hypothesis these trees are "flattened" at 1.8 million years ago. If we simulate a population that has been 30,000 since 1.8 million years ago and was 3000 before that, the world mean *Alu* frequency would be between 0.3 and 0.4. If the effective size since the event was 100,000, then we would observe a contemporary mean *Alu* frequency of 0.19. Our finding that the contemporary human mean *Alu* frequency is about 0.50 falsifies the multiregional hypothesis, since full multiregional evolution since the dispersal of *Homo erectus* would correspond to an effective size of a quarter million or so (Harpending et al. 1993). The *Alu*s tell us that that did not happen.

Population Differences in Mean *Alu* Frequency. A conventional calculation (Harpending and Jenkins 1973) of normalized variances, covariances, and genetic distances is given in Table 3 with the world centroid taken to be the simple mean of mean *Alu* frequencies for each major group of populations. The similarity of the Daghestan group to Europe is apparent.

Table 3. Genetic Statistics Describing the Similarities of Five Major Population Groups[a]

	Europe	*Africa*	*Daghestan*	*Asia*	*India*
Europe	0.049	0.343	0.031	0.162	0.087
Africa	–0.060	0.196	0.313	0.399	0.301
Daghestan	0.022	–0.045	0.027	0.112	0.055
Asia	–0.018	–0.064	–0.004	0.076	0.086
India	–0.004	–0.037	0.001	0.010	0.030

a. Normalized variances and covariances *r* around the contemporary world mean are on the diagonal and below, while genetic distances are above the diagonal. The distances are computed as $d_{ab} = r_a + r_b - 2r_{ab}$.

If we assume that ancestors of Africans were the source population of modern humans and that African ancestors were demographically successful before the diaspora, then it may be more appropriate to use the African mean *Alu* frequencies as the world centroid. Table 4 shows covariances, variances, and genetic distances computed in this way. With this perspective F_{ST} measures dispersal from Africa, and it is identical to the mean genetic distance from Africa: it is 0.32 computed in this way. Note that genetic distances among the populations are essentially unchanged in spite of the change of assumption about the inferred ancestral population.

Since the normalized covariances in Table 3 are around a contemporary world mean while those in Table 4 are around a hypothesized ancestral mean, we refer to the former as *r*s and the latter as *R*s. We can also use differences in mean *Alu* frequency among populations to estimate the *R*s, as we discuss next.

In order to use the model of biased *Alu* frequencies as products of ascertainment discussed above, we need to identify the population from which ascertainment chromosomes were drawn. The source of the chromosomes is not publicly known, so we are forced to proceed with the assumption that they were European. Another likely possibility is that they were a mixture of European and Asian chromosomes, in which case we would simply merge Asian and Europeans into a category "Eurasian" in the discussion that follows.

Equations (2) and (3) show that the mean of current biased *Alu* frequencies should be equal to or higher than that of the biased ancestral distribution *G,* and that the difference can be used to estimate the amount of drift of the ascertainment population since the diaspora. For example, if ascertainment chromosomes were drawn from Europeans (a possibility) and if the ancestral ascertained mean were 0.46 (assuming African populations have been large and drifted little since the diaspora), then an estimate of the total drift of Europeans from the ancestral mean is

$$R_{\text{European}} = \frac{0.56 - 0.46}{0.55} = 0.18,$$

Table 4. Genetic Statistics Describing the Similarities of Four Major Population Groups[a]

	Europe	*Daghestan*	*Asia*	*India*
Europe	0.338	0.030	0.159	0.085
Daghestan	0.308	0.308	0.110	0.054
Asia	0.286	0.296	0.393	0.085
India	0.274	0.275	0.302	0.296

[a] This table is like Table 3 except that the world centroid is the contemporary African mean. Normalized covariances are on the diagonal and below, while genetic distances are above the diagonal. The distances are computed as $d_{ab} = r_a + r_b - 2r_{ab}$.

which should be compared with the diagonal entry for Europe of 0.338 in Table 4. If the ancestral biased mean *Alu* frequency was as low as 0.43, as it is among Bushmen today, the corresponding estimate of European R is 0.25, closer to the estimate from contemporary frequency differences of 0.34.

Calculations based on contemporary mean *Alu* frequencies provide us with a partial picture of human differentiation since the diaspora. We have the normalized variance of one population, that from which ascertainment chromosomes were drawn, and normalized covariances between the ascertainment population and the others. This means that they give us one row and column of the matrix R. We can fill in the rest of the matrix from genetic distances among the populations. The justification is that the Rs are covariances around an ancestral set of gene frequencies, and this covariance matrix can be broken down into the sum of covariances around current gene frequencies and the normized squared difference between current and ancestral mean frequencies, that is

$$R = r + (1 - r_{ST})\, R_0,$$

where r_{ST} is an average of the diagonal entries of r and R_0 is a number that is the normalized mean squared difference between current and founding frequencies. This is equivalent to Wright's hierarchical decomposition of F statistics.

Following up with the assumption of European ascertainment, equation (3) together with the genetic distances given in either Table 3 or Table 4 allows us to "fill in" the normalized covariance matrix implied by population differences in mean *Alu* frequency. The average diagonal entry of this reconstructed matrix is 0.22, which should be comparable to the mean of the diagonal in Table 4, which is 0.33. If we assume that the ancestral biased mean *Alu* frequency was 0.40 instead of 0.46, we obtain comparable estimates from the two approaches.

At any rate, world *Alu* frequencies are in accord the accepted view of a primarily African origin of our species and with mild bottlenecking associated with the exodus from Africa. There is weaker support for southern African Khoisan speakers as the best genetic representative we have of the source for the populations of Daghestan, as well as for several southeast Asian groups as genetic representatives of derived populations with population histories of extensive genetic drift. We expect that New World populations will have even higher mean frequencies of insertions of this *Alu* panel.

For an isolated population, R follows

$$R(t) = 1 - e^{-t/2N},$$

where t is time in generations and N is the effective size of the population (Crow and Kimura 1970). Substitution of $R = 0.18$ for Europe yields $t/2N = 0.22$. If the diaspora occurred 40,000 years ago, equivalent to 1600 generations, our estimate of the effective size of humanity outside Africa is about 4000, corresponding to a census size of perhaps 10,000. One good possibility is that this number reflects the size of the "wavefront" of a wave of advance of modern humans out of Africa. Eswaran (2002) shows that during a wave of advance of a new advantageous genotype the wavefront is relatively genetically isolated, leading to its acting as a rolling genetic bottleneck that drastically reduces genetic diversity as it passes. A mechanism like this is the current best explanation for the apparent low human effective size shown by our DNA.

Acknowledgments This research was supported by the National Institutes of Health through grant GM-59290 and by the National Science Foundation through grant SBR-9818215. The mtDNA sequences and *Alu* frequency tables used in this paper are available at: http://harpend.dsl.xmission.com/daghestan_data.

Received 5 January 2002; revision received 16 July 2003.

Literature Cited

Aglarov, M. 1988. *A Rural Community in Highland Daghestan in XVII and Beginning XIX Centuries.* Nauka, Moscow.

Barbujani, G., I. Nasidze, and G. Whitehead. 1994. Genetic diversity in the Caucasus. *Hum. Biol.* 66(4):639–668.

Bulayeva, K. 1991. *Genetic Basis of Human Psychophysiology.* Nauka, Moscow.

Crow, J.F., and M. Kimura. 1970. *An Introduction to Population Genetics Theory.* New York, NY: Harper and Row.

Eswaran, V. 2002. A diffusion wave out of Africa—The mechanism of the modern human revolution? *Curr. Anthropol.* 43:749–774.

Gadjiev, A. 1971. *Anthropology of Small Daghestanian Populations.* Daghestan Branch of the USSR Academy of Sciences, Makhachkala.

Gammer, M. 1994. *Muslim Resistance to the Tsar: Shamil and the Conquest of Chechnia and Daghestan.* London, UK: Frank Cass.

Harpending, H., and T. Jenkins. 1973. Genetic distance among southern African populations. In *Methods and Theories of Anthropological Genetics*, M. Crawford and P. Workman, eds. Albuquerque, NM: University of New Mexico Press, 177–199.

Harpending, H.C., M.A. Batzer, M. Gurven et al. 1998. Genetic traces of ancient demography. *Proc. Natl. Acad. Sci. USA* 95:1961–1967.

Harpending, H.C., and A.R. Rogers. 2000. Genetic perspectives on human origins and differentiation. *Annual Review of Genomics and Human Genetics*, 361–365.

Harpending, H.C., S.T. Sherry, A.R. Rogers et al. 1993. The genetic structure of ancient human populations. *Curr. Anthropol.* 34:483–496.

Hawks, J. K. Hunley, S.H. Lee et al. 2000. Population bottlenecks and Pleistocene human evolution. *Mol. Biol. Evol.* 17(1):2–22.

Hudson, R.R. 1990. Gene genealogies and the coalescent process. In *Oxford Series in Evolutionary Biology,* D. Futuyma and J. Antonovics, eds. Oxford, UK: Oxford University Press 7:1-44.

Jorde, L.B., M.J. Bamshad, W.S. Watkins et al. 1995. Origins and affinities of modern humans: A comparison of mitochondrial and nuclear genetic data. *Amer. J. Hum. Genet.* 57:523–538.

Nasidze, I., and M. Stoneking. 2001. Mitochondrial DNA variation and language replacements in the Caucasus. *Proc. Roy. Soc. B.* 268(1472):1197–1206.

Rogers, A.R., and H.C. Harpending. In preparation. How to interpret *Alu* frequencies in populations.

Ruhlen, M. 1994. *The Origin of Language.* New York, NY: John Wiley and Sons.

Vavilov, N.I. 1936. Worldwide experience of the highland regions consumed. *Priroda* (Moscow) 2:76–85.

Watkins, W., A. Rogers, C. Ostler et al. 2002. Genetic variation among world populations using 100 *Alu* insertion polymorphisms. Submitted.

Watterson, G.A. 1975. On the number of segregating sites in genetical models without recombination. *Theoret. Pop. Biol.* 7:256–276.