A Genetic Landscape Reshaped by Recent Events: Y-Chromosomal Insights into Central Asia

Tatiana Zerjal,[1] R. Spencer Wells,[2] Nadira Yuldasheva,[2,3] Ruslan Ruzibakiev,[3] and Chris Tyler-Smith[1]

[1]Department of Biochemistry, University of Oxford, Oxford; [2]Imperial Cancer Research Fund Cancer and Immunogenetics Laboratory and Wellcome Trust Centre for Human Genetics, University of Oxford, Headington, United Kingdom; and [3]Institute of Immunology, Academy of Sciences, Tashkent, Uzbekistan

Sixteen Y-chromosomal microsatellites and 16 binary markers have been used to analyze DNA variation in 408 male subjects from 15 populations in Central Asia. Large genetic differences were found between populations, but these did not display an obvious geographical or linguistic pattern like that usually seen for Y-chromosomal variation. Nevertheless, an underlying east-west clinal pattern could be detected by the Autocorrelation Index for DNA Analysis and admixture analysis, and this pattern was interpreted as being derived from the ancient peopling of the area, reinforced by subsequent migrations. Two particularly striking features were seen: an extremely high level of Y-chromosomal differentiation between geographically close populations, accompanied by low diversity within some populations. These were due to the presence of high-frequency population-specific lineages and suggested the occurrence of several recent bottlenecks or founder events. Such events could account for the lack of a clear overall pattern and emphasize the importance of multiple recent events in reshaping this genetic landscape.

Introduction

Most human genetic variation, ~85% for autosomal sequences, is found within populations (Lewontin 1972; Barbujani et al. 1997). Nevertheless, the small amount of variation between populations is of great interest for understanding human history and shows geographical structure at a continental level (Cavalli-Sforza et al. 1994). The distribution of classical markers, such as blood groups, has been studied at high resolution in several parts of the world and typically shows smooth geographical distribution patterns, in which nearby populations resemble one another in frequency. These patterns are generally interpreted as resulting from neutral processes, such as migration and genetic drift, with selection acting on only a minority of loci. Genetic differences could have been established during early expansions out of Africa and in some subsequent colonization events, where the number of people taking part is thought to have been small, and frequencies would have been readily influenced by drift. To some extent, subsequent migrations and admixture have acted to reduce these differences.

Much recent work has concentrated on mtDNA and the Y chromosome, haploid non-recombining loci where detailed phylogenies can be constructed. The small effective population size makes both loci more susceptible to drift than are autosomal sequences, and as much as ~30% (for mtDNA [Jorde et al. 2000]) or ~40% (for the Y chromosome [Santos et al. 1999; Hammer et al. 2001]) of the variation has been found between populations. Patrilocality, in which children are born closer to their father's birthplace than their mother's, is more common than matrilocality. This has often led to greater geographical clustering of Y variants at a local scale and may contribute to larger-scale patterns as well.

One of the most studied regions has been Europe, where strong geographical differentiation and clear clinal patterns of Y-chromosomal variation have been detected (Rosser et al. 2000), with less mtDNA differentiation (Richards et al. 1998). There has been debate about the extent to which these patterns were established by the original Paleolithic colonizers or by subsequent Neolithic migration, but they are always considered to be of prehistoric origin. Some events within historical times have had large demographic effects—like the Black Death, which killed 25 million people in 5 years in the 14th century and reduced the European population by one-third—but these do not seem to have left a genetic "scar," perhaps because of the relatively large population size (Cavalli-Sforza et al. 1994). Although other parts of the world have not been studied in such detail, an ancient origin for patterns of diversity is commonly assumed.

Some examples are known of genetic patterns that have been perturbed by more-recent events, mostly associated with the colonization of new areas by Europeans in the past 500 years. For example, Hurles et al.

Received March 25, 2002; accepted for publication May 23, 2002; electronically published July 17, 2002.

Address for correspondence and reprints: Dr. Chris Tyler-Smith, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, United Kingdom. E-mail: chris@bioch.ox.ac.uk

0002-9297/2002/7103-0003$15.00

Table 1

Populations Sampled

Population	Language	Current Population Size[a]	Sampling Location	Historical Subsistence Method
Mongolians	Altaic	2,600,000	Mongolia: Ulaanbaatar	Pastoral nomadism
Kyrgyz	Altaic	2,500,000	Kyrgyzstan: central Kyrgyzstan (mixed)	Pastoral nomadism
Dungans	Sino-Tibetan	38,000	Kyrgyzstan: Alexandrovka, Osh	Agriculturalism
Uyghurs	Altaic	300,000	Kazakstan: Almaty, Lavar	Agriculturalism
Kazaks	Altaic	5,300,000	Kazakstan: Almaty, Katon-Karagay, Karatutuk, Rachmanovsky Kluchi	Pastoral nomadism
Uzbeks	Altaic	16,500,000	Uzbekistan: Kashkadarya region	Agriculturalism
Tajiks	Indo-European	3,300,000	Tajikistan: Penjikent	Agriculturalism
Turkmen	Altaic	3,500,000	Turkmenistan: Ashgabat	Pastoral nomadism
Kurds	Indo-European	50,000	Turkmenistan: Bagyr	Agriculturalism
Georgians	South-Caucasian	4,000,000	Georgia: Kazbegi	Agriculturalism
Ossetians	Indo-European	160,000	Georgia: southern Ossetia	Agriculturalism
Lezgi	North-Caucasian	171,000	Azerbaijan: Azerbaijan/Dagestan border	Agriculturalism
Svans	South-Caucasian	35,000	Georgia: Svanetia	Agriculturalism
Azeri	Altaic	6,000,000	Azerbaijan: Baku	Agriculturalism
Armenians	Indo-European	3,200,000	Armenia: Yerevan	Agriculturalism

[a] See Ethnologue database.

(1998) showed that 33% of Polynesian Y-chromosomal lineages have a European ancestry, and Carvalho-Silva et al. (2001) demonstrated that the vast majority of Y-chromosomal lineages in white male Brazilians were European, with very few of sub-Saharan African origin and none of American Indian origin. This demonstrates that patterns of variation can be dominated by recent events, but the extent to which this has happened outside the context of European colonization has received rather little attention.

We now present a study of Y-chromosomal variation in Central Asia. This is a vast territory, geographically defined on its southeastern edge by the presence of two high mountain ranges, the Tian Shan and the Pamir, and to the north by the Siberian taiga, but not so clearly delimited at its western extreme, where steppes and deserts stretch smoothly into the Middle East and Europe. In the present study, we consider Central Asia to stretch from Mongolia, in the east, to the Caucasus, in the west. The area has supported human life for >1 million years. There are unmistakable remains of Lower, Middle, and Upper Paleolithic cultures, and Mesolithic remains of hunter-gatherer groups have been found even in the Pamir Uplands (Davis and Ranov 1979; Ranov et al. 1995). Major human-population developments with important genetic implications occurred in the Neolithic, as early as the 6th millennium B.C. (Forde 1948; Anthony 1986; Cavalli-Sforza et al. 1994). The development of civilizations in this period has been greatly affected by the harsh climatic conditions of Central Asia. The steppes were a difficult environment for agriculture but were ideally suited to animal husbandry and pastoral nomadism. This kind of economy generally sustains populations of low density and is therefore more sensitive to dramatic demographic fluctuations that lead to genetic drift. At the same time, however, the lack of geographical barriers with the West has allowed a steady movement of individuals and populations (and therefore of genes) from the West into Central Asia and vice versa, as exemplified by the Silk Road (Kato 1992; Cavalli-Sforza et al. 1994).

Previous genetic analyses that used mtDNA and Y-chromosome microsatellites (Perez-Lezaun et al. 1999) have revealed a contrast between the even distribution of mtDNA variation in four populations and the low Y-chromosomal diversity in two high-altitude villages, interpreted as a male founder effect in the settlement of high-altitude lands. A large-scale survey of Eurasian Y-chromosomal diversity, performed with 23 binary markers (Wells et al. 2001), found high diversity in Central Asia and suggested that the region could be the source of at least three major waves of migration that led to Europe, India, and the Americas. Here, we demonstrate not only that an underlying and perhaps ancient east-west gradient of Y-chromosomal variation can be detected in Central Asia but also that this gradient has been disrupted by numerous recent population-specific events. Y genetic structure in this region therefore differs substantially from that in areas supporting large settled populations and may provide a paradigm for the sparsely populated parts of the world.

Material and Methods

DNA Samples

We analyzed a total of 408 male subjects from 15 populations. Between 1993 and 1995, 65 samples from

Mongolia were collected, and the remaining 343, from the other 14 populations, were collected during the EurAsia '98 expedition (see the EurAsia '98 Web site) and additional expeditions to the region. Details are given in table 1. Blood was taken from healthy male donors, mostly living in villages, with appropriate informed consent. Care was taken to avoid related individuals. The first steps of DNA extraction, consisting of selective lysis of the red cells, pelleting of the white cells, and lysis of the white cells in an SDS-containing buffer, were performed during the expedition, immediately after the blood was collected. White-cell lysates were stored at ambient temperature for weeks, until DNA purification could be performed in the laboratory by the salting-out procedure (Wells et al. 2001).

Binary Polymorphisms

Sixteen binary markers known to detect variation in Europe or Asia were used in this study. Typing was as described elsewhere: 12f2, YAP, SRY-8299 (also known as "SRY_{4064}"), sY81 (also known as "M2"), M9, LLY22g/*Hin*dIII, Tat, 92R7, and SRY-1532 (also known as "SRY_{10831}") according to Zerjal et al. (2001); LINE-1 according to Santos et al. (2000); MSY2 according to Bao et al. (2000); and Apt according to Pandya et al. (1998). Primer sequences and conditions for RPS4Y, M48, M20, and M17 are described by Qamar et al. (2002). All samples were typed with all binary markers. This strategy is redundant, but it provides an internal check on the reliability of the typing and will reveal recurrent mutations. With this set of markers, 18 lineages (haplogroups) are defined in global populations, and their relationship, under the assumption of the minimum number of mutational events, generates a unique tree (fig. 1). The nomenclatures reported previously and the new Y-Chromosome Consortium (YCC) nomenclature (Y-Chromosome Consortium 2002) are given.

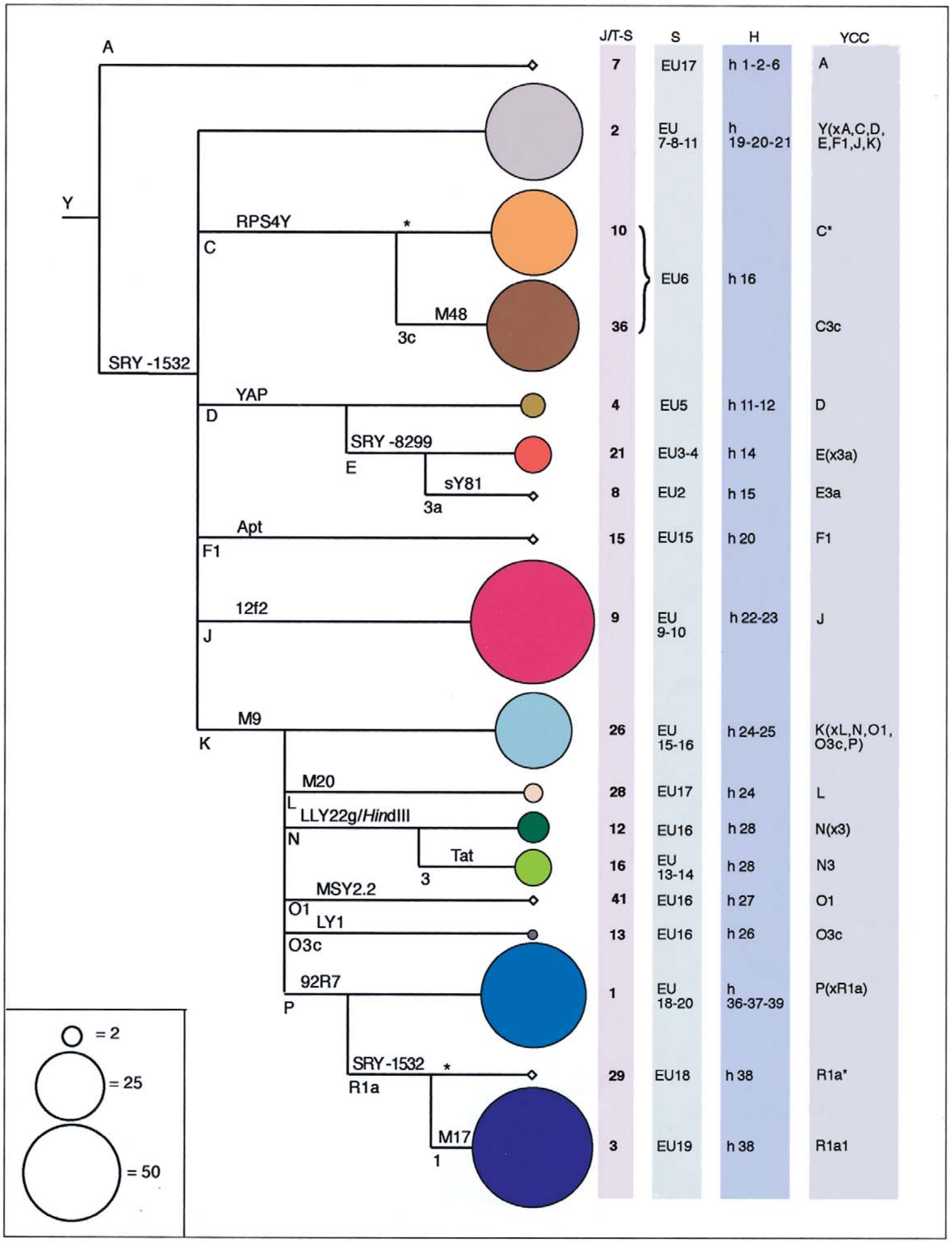
Y-STR Typing

Samples were typed with 16 microsatellites, of which DYS388, DYS392, DYS425, DYS426, and DYS436 are trinucleotides; DYS19, DYS389I, DYS389b, DYS390, DYS391, DYS393, DYS434, DYS435, DYS437, and DYS439 are tetranucleotides; and DYS438 is a pentanucleotide. These 16 microsatellites were organized into three multiplex PCRs and were analyzed on an ABI 377 sequencer, as described elsewhere (Thomas et al. 1999; Ayub et al. 2000; Zerjal et al. 2001). Y-STR alleles are named on the basis of the number of repeat units they contain, as established through the use of sequenced reference DNA samples. Allele lengths for DYS389b were obtained by subtraction of the DYS389I allele length from that of DYS389II. A duplication of DYS19 was encountered in some individuals, leading to the amplification of two alleles that differed in number of repeats. In the samples with duplications, it was impossible to identify the alleles separately. Because of this, DYS19 was omitted from some analyses. The duplication was confined to haplogroup 36 chromosomes, and it was seen in all haplogroup 36 Kazaks, 4 of 5 haplogroup 36 Kyrgyz, and 7 of 13 haplogroup 36 Mongolians. However, we cannot be sure that, in the haplogroup 36 samples in which only one DYS19 allele was detected, the duplication was not present, since the two alleles could have the same length and be indistinguishable. DYS19 duplications were not found outside this haplogroup, and no other Y-STR duplications were observed.

Statistical Analyses

Microsatellite haplotypes were constructed for each sample, and median-joining networks were calculated by use of the program Network 3.0 (Bandelt et al. 1999) and were grouped either according to population or according to haplogroup. A weighting scheme, going from 2 to 8, was adopted on the basis of the molecular variance of each microsatellite, with the weight inversely proportional to the variance. Binary markers, when included, were given a weight of 30. For binary marker and microsatellite haplotypes, genetic distances (as pairwise values of Φ_{ST}), genetic diversities (and their SEs), and the analysis of molecular variance (AMOVA) were calculated by use of the ARLEQUIN 1.1 software (Excoffier et al. 1992). In both cases, a distance matrix was created from the number of differences separating each pair of haplogroups or haplotypes. A multidimensional scaling (MDS) analysis was performed by use of the SPSS 7.0 software package, with pairwise Φ_{ST} distances based on microsatellite haplotypes as variables. Weighted means of within-haplogroup average squared distances (ASDs) were calculated as described elsewhere (Qamar et al. 2002).

We calculated admixture proportions (mY) and their SEs on the basis of 1,000 bootstraps, using the program Admix 2.0 (Dupanloup and Bertorelle 2001), and then we used these values to search for geographical patterns in the genetic data. The Admix 2.0 program allows the use of any number of parental populations and calculates their relative contribution to a hybrid population. We used four parental populations, with haplogroup frequencies taken from the literature (table 2). Each Central Asian population, except the Mongolians, was then considered as a hybrid population, and the admixture proportions were calculated. Admixture values from each parental population were then used to create interpolated maps, by means of a geographical information system (ArcView 3.2a). The area was divided into a grid of cells with size 0.1 degrees, and an interpolated value for each cell, except those that contained the samples,



J/T-S	S	H	YCC
7	EU17	h 1-2-6	A
2	EU 7-8-11	h 19-20-21	Y(xA,C,D, E,F1,J,K)
10	EU6	h 16	C*
36			C3c
4	EU5	h 11-12	D
21	EU3-4	h 14	E(x3a)
8	EU2	h 15	E3a
15	EU15	h 20	F1
9	EU 9-10	h 22-23	J
26	EU 15-16	h 24-25	K(xL,N,O1, O3c,P)
28	EU17	h 24	L
12	EU16	h 28	N(x3)
16	EU 13-14	h 28	N3
41	EU16	h 27	O1
13	EU16	h 26	O3c
1	EU 18-20	h 36-37-39	P(xR1a)
29	EU18	h 38	R1a*
3	EU19	h 38	R1a1

Figure 1 Rooted maximum-parsimony tree of haplogroups defined by binary markers. Marker names are indicated above the lines, and lineage names recommended by the YCC are shown below the lines. Branch lengths are arbitrary. Haplogroups are represented by circles, with an area proportional to frequency. Haplogroup names according to the YCC and former nomenclatures are compared in the right-hand columns. J/T-S = Jobling/Tyler-Smith; S = Semino; H = Hammer.

Table 2

Source Data for the Four Parental Population Groups Used in the Admixture Analysis

Region and Populations	No. of Individuals
Central/Eastern Europe:[a]	
Hungarians	50
Ukrainians	45
Middle East:[b]	
Syrians	88
Turks	72
Saudi Arabians	20
Northeastern Asia:[b]	
Mongolians	148
Buryats	81
Selkups	122
Forest Nentsi	27
Evenks	95
Siberian Eskimos	22
China:[c]	
Northern Han	44
Yizu	43
Southern Han	40
Miao	57

[a] See Semino et al. 2000.

[b] See Hammer et al. 2001.

[c] See Karafet et al. 2001.

was calculated. The interpolation algorithm considered a maximum of the 12 nearest population data points, weighted by the cube of the distance to the cell.

We performed spatial autocorrelation analysis by means of the Autocorrelation Index for DNA Analysis (AIDA) (Bertorelle and Barbujani 1995), using haplogroup frequencies as genetic data. Geographical coordinates for each population were obtained from the Geonames online database.

In populations with a reduced genetic diversity, we wished to determine the time to the most recent ancestor (TMRCA) of particular lineages that seemed to have experienced a bottleneck and could therefore provide information about the history of the population. There are several methods available to date chromosomal lineages, some dependent on a population model and others "model free," based entirely on genetic parameters such as mutation rates and allele length. We applied three of them: BATWING (Wilson and Balding 1998), NETWORK 3.0 (Bandelt et al. 1999; Forster et al. 2000), and Ymrca (Stumpf and Goldstein 2001). The first method is a Bayesian approach that takes into consideration population demographic parameters and allows locus-specific microsatellite mutation rates to be used. The genetic and population parameters adopted here were as described by Qamar et al. (2002). In one case, the lower limit to the 95% CI was 2 years, which is a physically implausible time and indicates a limitation of the program. The other two methods do not explicitly consider any population model. The ρ method is a phylogeny-based statistic that gives relative time estimates by means of ρ, the average distance to the node of interest (Forster et al. 2000). We used, for all loci, the average mutation rate estimate proposed by Kayser et al. (2000) (2.8×10^{-3} per locus per generation) and its 95% CIs (1.72×10^{-3}–4.27×10^{-3}) and a generation time of 30 years. The same mutation rates and generation time were adopted in the last method, which calculates genealogical depth by use of the ASD to the ancestral allele and length-adjusted mutation rates, as described elsewhere (Stumpf and Goldstein 2001).

Results

We first typed the 408 males from 15 Central Asian populations with 16 binary markers, to identify the Y haplogroups present, and with 16 microsatellites, to define more detailed haplotypes.

Frequencies and Distributions of Y-Chromosome Haplogroups in Central Asia

Of the 18 potential haplogroups defined by the set of binary markers, we found 13 (fig. 1), and 7 of these were common. Haplogroup distributions are represented in figure 2. Haplogroups 1 and 3 were common in almost all populations. The highest frequencies of haplogroup 1 were present in the Turkmen and Armenians (52% and 43%, respectively; see table 3). The Kyrgyz and Tajiks had the highest frequency of haplogroup 3 (63% and 64%, respectively), and this haplogroup generally was more frequent in the eastern Central Asian populations than in the western ones, although it was only present at a level of 3% in the Kazaks. Haplogroup 9 is very frequent in the Middle East (Hammer et al. 2000; Semino et al. 2000; Quintana-Murci et al. 2001) and, in Central Asia, was generally more common in the west than in the east. Its presence was high in populations from the Caucasus (among Georgians, 73%; among both Lezgi and Azeri, 53%) and decreased eastward, reaching its lowest frequency among the Mongolians (3%). There were, however, some eastern populations that had a high frequency (i.e., Dungans, Uyghurs, and Uzbeks), and this will be discussed below. Haplogroup 10 and its derivative, haplogroup 36, showed the opposite pattern. Together, they accounted for 54% of the Mongolians (34% and 20%, respectively) and 73% of the Kazaks (11% and 63%, respectively); however, their frequencies decreased sharply as sampling moved westward. Haplogroup 26, present at its highest frequency in Southeast Asia (unpublished observations), was quite frequent in Central Asia, being found in 10 of the 15

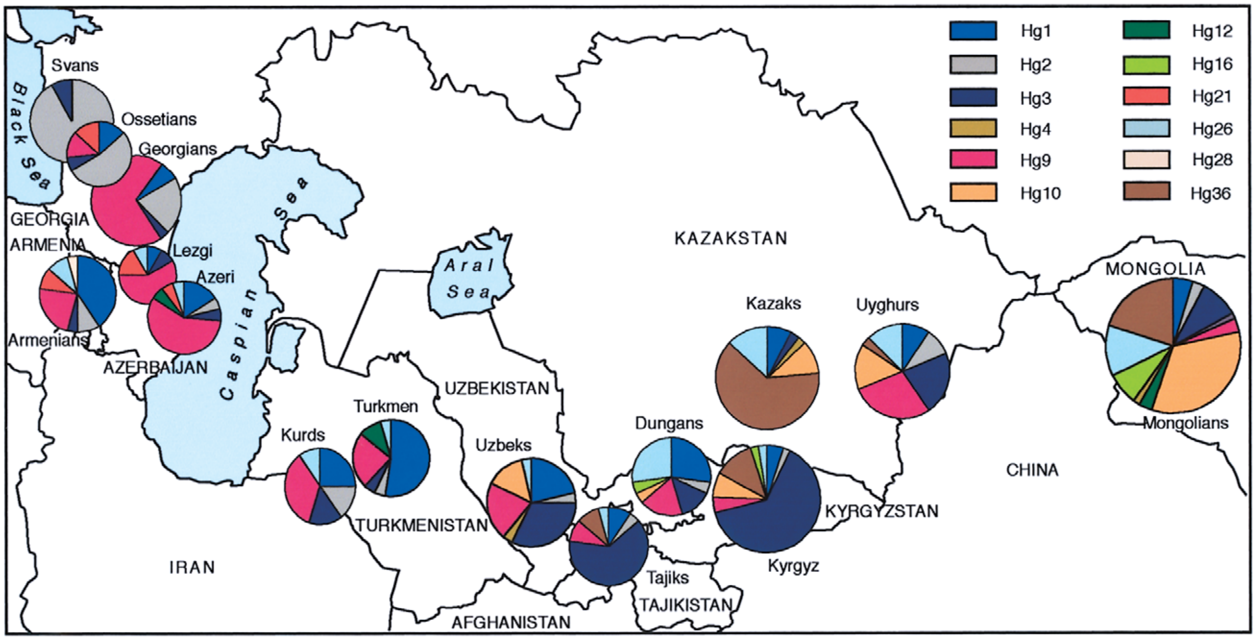


Figure 2 Haplogroup frequencies in the population samples. Circle area is proportional to sample size, and haplogroup color codes are as in figure 1.

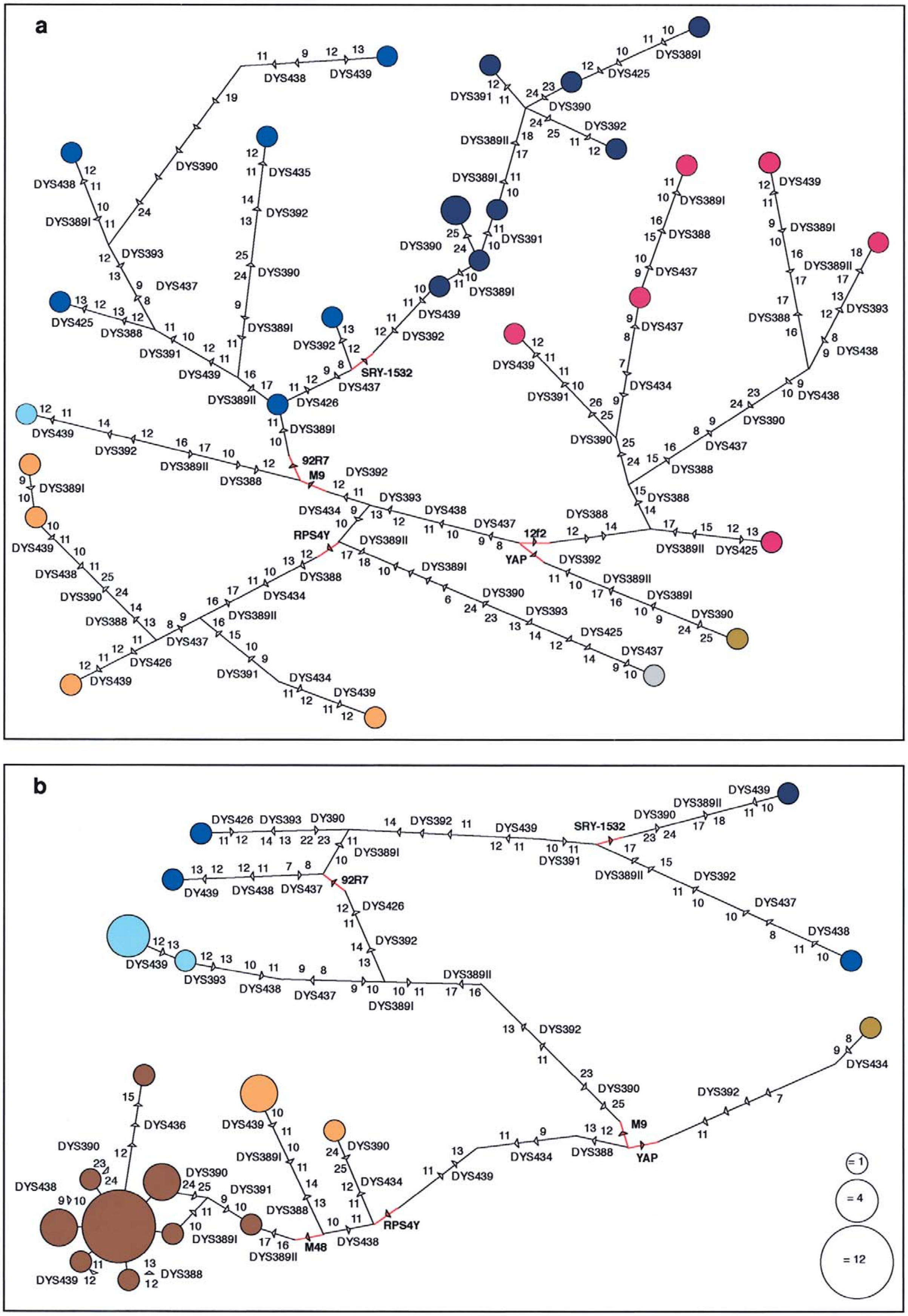
populations analyzed. The highest frequencies were in eastern Central Asia, especially among the Dungans (27%), the Uyghurs (15%), the Mongolians (13%), and the Kazaks (13%), and it decreased, as sampling moved westward, to 5% among both the Armenians and the Azeri and to 0% in the northern Caucasus populations. Haplogroup 2 contains a heterogeneous set of chromosomes that are not necessarily closely related. It is worth noting, however, that 93% of the Svans shared this haplogroup and that its frequencies were often somewhat higher in the west than in the east. Haplogroups 12 and 16, widespread in Siberia and northern Eurasia (Zerjal et al. 1997; Karafet et al. 1999), were rare in Central Asia, with the exception of the Turkmen, among whom haplogroup 12 was present with a frequency of 10%, and the Mongolians, among whom haplogroup 16 accounted for 8% of the chromosomes. Haplogroup 21 was restricted to the Caucasus region, with a frequency of 17% among the Lezgi and 10% among the Armenians. Haplogroups 13, 28, and 41, quite common in different southern parts of the Asian continent, were, in Central Asia, just sporadic or absent.

Thus, many of these haplogroup distributions (including haplogroups 1, 2, 3, 9, and 26, together representing >70% of the sample) span the entire 5,000 km and show large differences in frequency between populations but a striking lack of an overall pattern. A few haplogroups—21 in the west and 10 and 36 in the east—do show geographical clustering, but these are exceptions, representing only 22% of the chromosomes.

Microsatellite Haplotypes and Network Analysis

Y chromosomes were also typed with 16 Y-specific microsatellites. Full data were obtained from all samples, and 304 different haplotypes were identified, among which 264 (87%) were individual specific. No haplotypes were shared between haplogroups.

The haplotype diversity averaged over all populations was 0.96, a value lower than those described in studies in which fewer microsatellites were used (de Knijff et al. 1997; Kayser et al. 2001), whereas a higher haplotype diversity would be expected because of the larger number of microsatellites. Several of the populations we studied did indeed have a haplotype diversity ⩾0.98, but there were others with very low values, such as the Georgians, who had 0.93 ± 0.04, and the Kazaks, Kyrgyz, and Turkmen, who had 0.94 ± 0.02. However, the lowest value of all was 0.84 ± 0.05) in the Svans, who also had an extremely low haplogroup diversity 0.15 ± 0.09. The overall correlation between haplogroup and haplotype diversity was high (r = 0.93). However, no correlation was apparent between haplotype diversity and geography or language, and, often, neighboring populations showed very different levels of haplotype variability. In the Kazaks, only 39% (15/38) of the chro-


a
DYS438
DYS439
DYS390
DYS389I
DYS393
DYS437
DYS435
DYS392
DYS425
DYS388
DYS391
DYS389II
DYS426
SRY-1532
92R7
M9
DYS434
RPS4Y
12f2
YAP
DYS436
b
DYS426
DYS393
DY390
DY439
DYS389I
DYS437
92R7
DYS392
DYS391
DYS389II
DYS439
DYS388
M9
YAP
DYS434
DYS390
M48
RPS4Y
= 1
= 4
= 12

mosomes had a unique haplotype, but, among the neighboring Uyghurs, such chromosomes were present in 73% (24/33) of individuals, in the Mongolians in 72% (47/65), and in the Uzbeks 100%.

To investigate the haplotype variation within each population and within haplogroups, we constructed median-joining networks. Two examples are shown in figure 3. Among the Uzbeks, in whom each chromosome has a different haplotype (fig. 3*a*), the chromosomes are scattered in the network and separated by long branches. In contrast, in the Kazaks, among whom diversity is low (fig. 3*b*), the majority of the haplotypes were clustered together, with many chromosomes sharing the same or closely related haplotypes. The latter pattern is that expected if the population has suffered a bottleneck or founder event (de Knijff 2000). It thus appears that more than one-third of the populations investigated in Central Asia have experienced such events.

We have therefore evaluated a number of statistics that reveal features of the structure within each population (tables 3 and 4 and fig. 4). Low-diversity populations (table 3) have low Θ_k values (table 4), indicating a small effective size of the male populations. In general, they also have low levels of microsatellite haplotype variability, as measured in three additional ways: by ASD, by a weighted mean of within-haplogroup ASD (table 4), and by number of pairwise differences (fig. 4). There are some differences in the ranking of the populations according to the various criteria, and thus there is some ambiguity about the status of the Turkmen, Georgians, and Lezgi, but, overall, a distinct group of six populations stands out as having low diversity: the Kyrgyz, Kazaks, Tajiks, Turkmen, Georgians, and Svans (fig. 4). The low-diversity populations (but not the high-diversity populations) contain population-specific clusters of haplotypes that are likely to have arisen within these populations. The TMRCAs of these clusters therefore are relevant to the history of the populations, and estimates obtained by different methods are summarized in table 5. The three methods gave broadly similar results and show that the times are likely to fall within the historical period. In most such populations, the bottleneck pattern seems quite simple, with one ancestral haplotype at high frequency and several variant haplotypes branching off from it. In the Svans, however, the pattern was complicated by the presence, within haplogroup 2, of two distinct founder lineages with different degrees of diversity. In conclusion, the populations examined can be divided into two groups, one with low diversity and the other with high diversity (table 4).

Haplogroup-specific networks were also calculated for the most-frequent haplogroups. The majority had a similar structure: long branches between haplotypes, most of which were population specific (results not shown). To measure, in a more quantitative way, the network shape, ASD values were calculated for each haplogroup. Most ASDs were between 60 and 90. However, haplogroups 3 and 10 were exceptions to this pattern, with low ASD values: 15 in haplogroup 3 and 31 in haplogroup 10. On a network level, in haplogroup 3, this was reflected by a compact shape, with most haplotypes separated by one mutational step only.

Genetic Distances among Populations and MDS Analysis

Binary marker ascertainment bias can lead to quite different conclusions about the same populations (Su et al. 1999; Karafet et al. 2001), but this should not occur when unbiased markers are used that are variable in all populations. We therefore used microsatellite haplotype frequencies and the molecular differences between haplotypes to compute population genetic distances in the form of values of Φ_{ST}. Pairwise values of Φ_{ST} showed that, in some cases, neighboring populations were significantly different, whereas, in other cases, geographically distant populations had nonsignificant pairwise Φ_{ST} values. To extend this observation, we tested for correlations between genetic and geographical distances. A matrix of pairwise Φ_{ST} values was compared with the same matrix made of pairwise geographical distances, and the correlation was not significant ($r = 0.15$; $P = .13$), indicating that, in this region, genetic distances are not linearly related to geographical distances, as is observed in other cases (Karafet et al. 2001).

Pairwise values of Φ_{ST} were also used to perform an MDS analysis. A good fit between the two-dimension plot and the source data (pairwise values of Φ_{ST}) was obtained, demonstrated by the low stress value obtained (0.16). When all populations were included, the Svans dominated the plot because of their extreme pairwise values of Φ_{ST} (between 0.5 and 0.3), causing a compression of all the remaining populations. They were therefore excluded from subsequent analysis, and figure 5 presents the result. In the lower left corner of the plot, the Kyrgyz and the Tajiks cluster together, well separated

Figure 3 Median-joining networks of microsatellite haplotypes in (*a*) Uzbeks and (*b*) Kazaks. Circles represent haplotypes, with area proportional to frequency and color code as in figure 1. Binary marker mutations are represented by red lines and microsatellite mutations by black lines. Note that DYS19 has not been used in the networks because it is duplicated in some haplogroup 36 individuals; the two Uzbeks with the same 15-locus haplotype are distinguished if DYS19 is used.

Table 3

Haplogroup Frequencies and Y-Chromosomal Diversity

		No. of Individuals from Haplogroup															
Population	n	1	2	3	4	9	10	12	13	16	21	26	28	36	Haplogroup Diversity ± SE	No. Of Microsatellite Haplotypes	Microsatellite Haplotype Diversity ± SE
Mongolians	65	3	2	6	1	2	22	2	1	5		8		13	.82 ± 0.03	54	.99 ± .002
Kyrgyz	41	2	1	26		2	3			1		1		5	.58 ± .09	26	.94 ± .02
Dungans	22	6	1	3		4	1			1		6			.83 ± .04	20	.99 ± .05
Uyghurs	33	3	3	7		9	5					5		1	.84 ± .03	28	.99 ± .01
Kazaks	38	3		1	1		4					5		24	.58 ± .09	22	.94 ± .02
Uzbeks	28	6	1	9	1	6	4					1			.81 ± .04	28	1.00 ± .02
Tajiks	22	2	1	14		2							1	2	.60 ± .11	16	.95 ± .03
Turkmen	21	11	1	1		5		2				1			.68 ± .09	15	.94 ± .02
Kurds	20	5	3	3		7						2			.80 ± .05	19	.99 ± .02
Georgians	26	2	4	1		19									.45 ± .11	19	.93 ± .04
Ossetians	15	2	7	1		3					2				.75 ± .09	15	1.00 ± .02
Lezgi	12	2		1		7					2				.65 ± .13	11	.98 ± .04
Svans	25		23	2											.15 ± .09	14	.84 ± .05
Azeri	19	3	1	1		11		1			1	1			.66 ± .11	19	1.00 ± .02
Armenians	21	9	2	1		5					2	1	1		.79 ± .07	21	1.00 ± .02
Overall	408	59	50	77	3	82	39	5	1	7	7	31	2	45	.66[a]	304[b]	.96[a]

[a] Average value.

[b] Total number of different haplotypes.

from the neighboring Kazaks. On the right-hand side, the Georgians are quite separate from the central group of populations, and the Turkmen are somewhat distinct at the bottom of the plot. The Uzbeks, Uyghurs, and Dungans have close genetic affinities with the populations from the Caucasus. Neither geographical nor linguistic population clusters are apparent in the plot. Instead, the high-diversity populations cluster in the center, and the low-diversity populations lie around the outside. Population size, reflecting historical and social factors, seems to have had a major influence.

Analysis of Molecular Variance

Population genetic structure was analyzed from both data sources (haplogroup and haplotype frequencies) in a hierarchical mode: within populations, among populations, and among groups of populations (table 6). As expected, in all cases, the highest fraction of variability was within populations, but a substantial fraction was due to differences among populations (24% and 19%, by use of haplogroup and haplotype frequencies, respectively [$P < .0001$]), indicating a high level of population differentiation. When the hierarchical approach was taken, populations were pooled together according to geography or linguistic family. In the first case, populations were divided in two geographical groups, using the Caspian Sea as divisor and placing the populations in a western group or an eastern one. In the second grouping strategy, populations were grouped accordingly to their language family (see MDS language division). The amount of variation explained by differences among groups was low ($\Phi_{CT} = 0.09$; table 6), although all the values except one were statistically significant. The amount of variation among populations from the same group was, in all cases, very high (Φ_{SC} ranging from 0.17 to 0.24). This indicates the presence of a great deal of heterogeneity among populations that share either a geographical proximity or a linguistic similarity and, at the same time, the existence of genetic similarities among populations speaking different languages or occupying separate geographical regions.

Two conclusions can be drawn from these analyses. First, despite the ascertainment bias in the binary markers, the AMOVA results, like the diversity values, are broadly similar to those obtained with microsatellites, suggesting that reliable conclusions can be obtained by use of binary markers in other analyses for which they are the most suitable. Second, neither geography nor linguistic similarities are good predictors of the genetic

Table 4

Microsatellite-Based Population Statistics

Population	Θ_k	ASD	Weighted ASD	Population Diversity
Mongolians	107	114	23	High
Kyrgyz	23	85	4	Low
Dungans	101	170	58	High
Uyghurs	85	129	36	High
Kazaks	13	75	4	Low
Uzbeks	360	143	42	High
Tajiks	25	72	11	Low
Turkmen	22	114	35	Low
Kurds	177	137	42	High
Georgians	30	115	40	Low
Ossetians	∞	132	35	High
Lezgi	58	78	17	High
Svans	10	40	19	Low
Azeris	∞	116	54	High
Armenians	∞	147	36	High

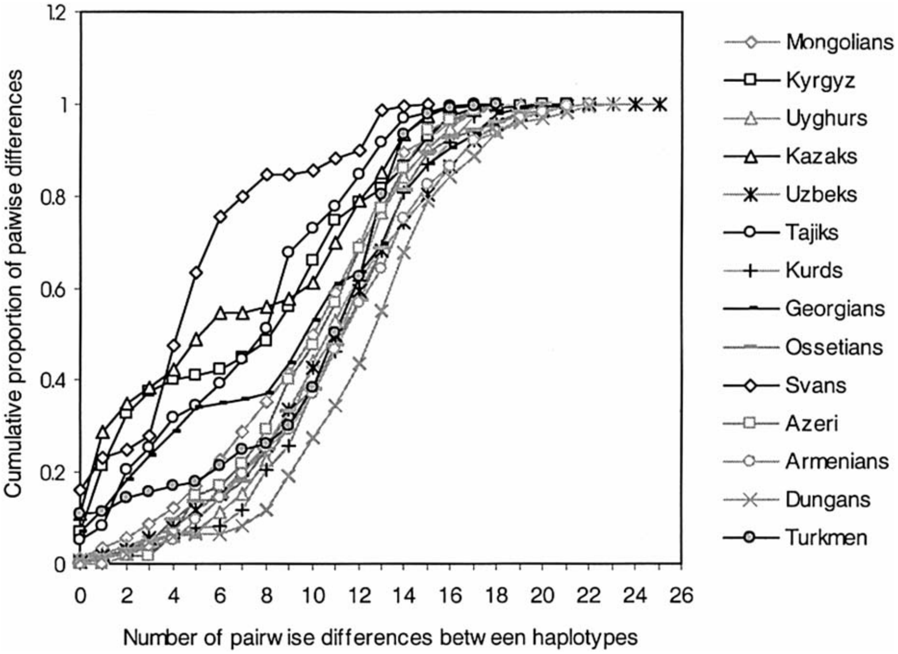


Figure 4 Cumulative pairwise differences between microsatellite haplotypes in populations with sample size ≥15

structure in Central Asian populations; other elements and events must be involved.

Analysis of Haplogroup Diversity by Spatial Autocorrelation

Φ_{CT} values were low, but three of four were significantly different from zero and were higher for geographical grouping than for linguistic grouping. We therefore investigated other methods that might reveal geographical structure. An autocorrelation index for DNA analysis (Moran's *II*) measures the level of similarity between populations as a function of their relative geographical distances. The program AIDA (Bertorelle and Barbujani 1995) was used to examine, in a quantitative way, the geographical pattern of genetic variation (in this case, by use of haplogroup frequencies). The program calculates Moran's *II* index, a normalized similarity index, at different distance classes and assesses the significance of each value by a randomization test. The shape of the correlogram (the plot of *II* indices against geographical-distance classes) is predicted to differ between evolutionary scenarios, as described by Sokal et al. (1989). Figure 6 shows the correlogram obtained. Highly significant positive Moran's *II* values are seen at short distances, which decrease to significantly negative values at long distances, and thus indicate an underlying clinal pattern. However, the ~10-fold reduction of Moran's *II* (from 0.24 to 0.025; $P < .005$) at 500 km and the increase at 1,000 km (0.078; $P < .005$) emphasizes that, in Central Asia, geographically close populations are often more dissimilar than relatively distant ones. This observation was robust to different choices of distance class, and, indeed, negative values of *II* were seen when a very short distance class was used (results not shown). A pattern of this kind can be interpreted as an ancient cline, on top of which more recent events—such as drift and founder effects—are superimposed; in classical spatial autocorrelation terms, it is referred to as "long-distance differentiation" (Sokal et al. 1989; Barbujani et al. 1994). The AIDA analysis thus identified an underlying geographical structure that was hinted at by some of the haplogroup distributions and the AMOVA results but was not apparent from the simple comparison of genetic distance with geographical distance.

Inference of Geographical Patterns from Haplogroup Frequencies

We wished to determine whether the structure identified by AIDA could be confirmed and further elucidated by other methods. Admixture analysis considers multiple lineages to be grouped according to their frequency in selected "source" populations, and we therefore investigated whether it could also reveal an underlying geographical structure. Central Asia cannot be considered solely as a receiver of populations; in this analysis, we were interested more in the method's ability to detect spatial features by consideration of natural groupings of lineages than in identification of migration events into

Table 5

Estimates of TMRCAs for Selected Lineages

Population and Haplogroup	Entire Haplogroup or Sublineage	TMRCA [95% CI] (years)		
		ρ	Ymrca	BATWING
Kyrgyz:				
3	Entire haplogroup	620 [400–1,000]	670 [440–1,090]	820 [360–2,300]
Kazaks:				
36	Entire haplogroup	480 [300–780]	685 [429–1,065]	750 [300–2,000]
Tajiks:				
3	Entire haplogroup	1,340 [882–2,180]	1,700 [1,130–2,800]	1,950 [900–5,000]
Georgians:				
9	Sublineage	980 [650–1,600]	1,250 [819–2,010]	1,600 [660–3,900]
Svans:				
2	Entire haplogroup	2,010 [1,320–3,270]	2,310 [1,500–3,780]	2,090 [1,078–4,140]
2	Sublineage	570 [380–930]	765 [500–1,230]	670 [240–1,670]
2	Sublineage	60 [40–100]	60 [40–100]	100 [2–1,550]

and out of the Central Asian populations. We calculated the admixture components from four potential sources chosen for their geographical locations—the Middle East, Central Europe, northeast Asia, and China—using published data (summarized in table 2). For each Central Asian population, four admixture coefficients were obtained, each corresponding to the relative contribution from one potential source. These coefficients were then used to create interpolated maps, each representing the admixture contribution from one source group. In all Central Asian populations, the contribution from China was not significant, and we therefore show only three interpolated maps (figs. 7*b*–7*d*). There is a major contribution from the Middle East (fig. 7*b*), with high levels of admixture in the Caucasus and in the westernmost Central Asian populations and with decreasing levels to the east. However, in the Uyghurs, the Middle Eastern contribution is moderately high, making them appear as an isolated island in a territory of low Middle Eastern contribution. Figure 7*c* portrays the estimated contribution from Central Europe. The pattern is dominated by the high frequency of haplogroup 3 in the Kyrgyz and Tajiks and probably reflects the strong founder effect in these two populations that is seen in the network analysis (data not shown). Finally, figure 7*d* represents the northeastern contribution. The highest amount of admixture is concentrated in the Kazaks, although some is also detectable in the Turkmen, in other eastern populations, and even in the Armenians in the Caucasus.

We conclude that admixture analysis, in addition to the AIDA results, reveals an underlying geographical structure to the data. Furthermore, it shows that this has a predominantly east-west, rather than north-south, form, although several populations deviate from the underlying trend.

Discussion

The present research was aimed at understanding the Y-chromosomal composition of populations from the Caucasus and Central Asia and identifying the type of evolutionary events that might have produced the current distributions. The region shows a high degree of Y genetic structure, with almost 24% of variation occurring between populations, but several conventional analyses did not reveal either a strong geographical or linguistic pattern to this structure. Nevertheless, an underlying east-west clinal pattern of variation could be detected. What factors could have contributed to the formation of this underlying smooth pattern, and what factors could have disrupted it? Here we discuss what is known of the history of the region and the extent to which known events can explain the distributions observed.

Traces of Ancient Genetic Patterns: Demic Diffusion in Central Asia

Clines are the expected genetic consequence of major demographic processes in which population expansion into new territories is associated with population growth (Cavalli-Sforza et al. 1993; Barbujani et al. 1994). Ways in which a clinal pattern could have been established in Central Asia include the initial peopling of this area from one or more sources during either the Paleolithic, subsequent Neolithic expansions, or a combination of the two. There are thought to have been genetically distinct southern and northern expansions out of Africa, with the former reaching China by 40,000 years ago and the latter extending through western Asia at about the same time. Thus, distinct Paleolithic populations could have moved into Central Asia from both east and west. However, several of the Y haplogroups have coalescence times

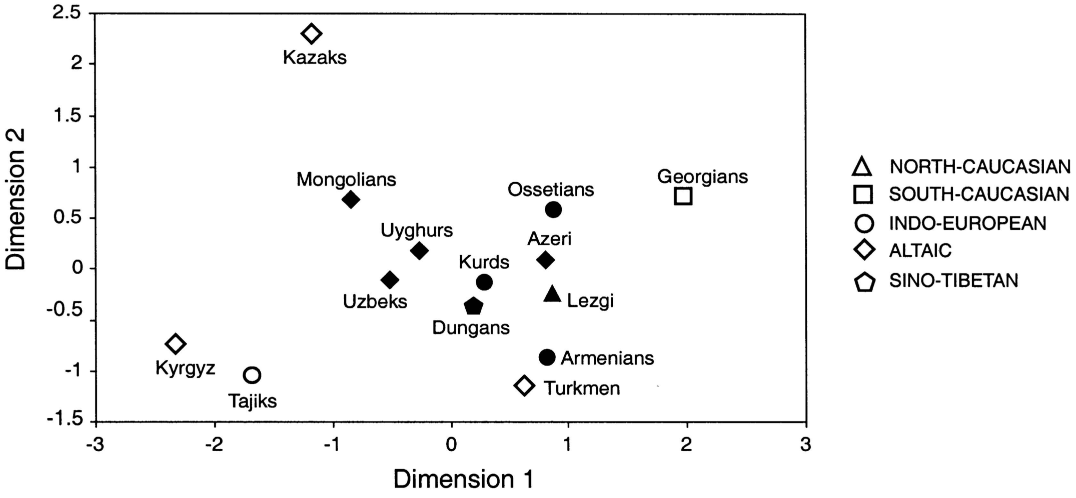


Figure 5 MDS analysis of population pairwise values of Φ_{ST}, based on microsatellite haplotypes. Symbol shapes indicate language affiliation; blackened symbols represent high-diversity populations, and unblackened symbols represent low-diversity populations.

that, although uncertain, are likely to date after the initial colonization, and many (with possible exceptions, such as haplogroup 1) probably arose outside this area. If so, their presence in Central Asia must reflect post-Paleolithic migration events. Quintana-Murci et al. (2001) have shown that genetic traces thought to represent a Neolithic demic expansion from the Middle East are recognizable in southwest Asia by the high frequency of haplogroup 9. How far into Asia has this expansion gone? It could readily have extended into the Caucasus, explaining the haplogroup 9 and Middle Eastern admixture there. The harsh climatic conditions of Central Asia would not have encouraged the spread of agriculture, but it is likely that fertile spots like the Fergana Valley (between Kyrgyzstan and Uzbekistan) and other areas where water was available could have sustained early farming communities. Therefore, Neolithic expansion from the Near East provides a plausible explanation for the overall difference between the western section and the remainder of the region and thus for a major contribution to the observed clinal pattern. Furthermore, the limited number of areas suitable for farming could have contributed to the uneven spread of Neolithic haplotypes.

Influence of Pastoralism

The steppes of Central Asia are the perfect environment for animal husbandry, and archaeological remains of animals such as cattle and goats date back to 6,000 years ago. The domestication of the horse, dating back to the 3rd millennium B.C. in the area between the Dnieper and Volga Rivers (Anthony 1986; Cavalli-Sforza et al. 1994), may have been a particularly important event in the history of the people of the steppes, bringing changes, at all social levels, in subsistence, transportation, and warfare. On a general level, it allowed the development of a more pronounced pastoral nomadism, characterized by seasonal migrations over longer distances, much higher population mobility (Anthony [1986] proposes a factor of five), and, therefore, a higher likelihood of population growth and expansion. Archaeological records (Anthony 1986; Lin 1992; Dexter and Jones-Bley 1997) suggest that several expansion waves of nomadic groups from the Eurasian steppes reached Central Asia. They are described in Chinese historiography as horse-riding, Caucasian-looking, Indo-European–speaking people and are sometimes referred as the "Kurgan Culture," with a homeland said to be in the steppes north of the Black and Caspian Seas (Dexter and Jones-Bley 1997).

Zerjal et al. (1999) postulated that haplogroup 3 could be the most evident male genetic legacy of this population expansion. It is interesting to note that haplogroup 3 is present at relatively low frequency in the Caucasus, as well as in the Middle East (Hammer et al. 2000; Semino et al. 2000). It is possible that, in the expansion of these early nomadic groups, the Caucasus mountain range formed a significant barrier and that the region was already well populated. Instead, they spread easily in eastern Central Asia, and haplogroup 3 is common in those populations, even though its distribution shows differing extremes of frequency—the Kyrgyz and the Tajiks show 63% and 64%, respectively, whereas the Kazaks show only 3%—but these variations are probably the result of drift during population bottlenecks or founder events, as discussed below.

In Siberia (Karafet et al. 1999) and China (Karafet et al. 2001; authors' unpublished observations), the frequency of haplogroup 3 drops quite sharply, suggesting that Central Asia represents the most eastern fringe of this population expansion. A bigger genetic contribution and the introduction of Indo-Iranian languages by these nomadic groups may be detectable in more southerly regions, such as Iran (Quintana-Murci et al. 2001), Pakistan (Qamar et al. 2002), and India (Zerjal et al. 1999). A more extended microsatellite analysis could be informative in disentangling this complex pattern of population expansions.

Later Events: Historical Migrations from the West

Two later historical elements might have enhanced the Middle Eastern contribution to Central Asia: long-distance trade via the Silk Road from the 2nd century B.C. onward (De Rosa 1992; Kato 1992) and the expansion of the Muslim world starting from the 7th century A.D. Traders would have had a variety of origins in the West and could have introduced many lineages to Central Asia at low frequencies; we and others (Hammer et al. 2001; Wells et al. 2001) do, indeed, see a high number of low-frequency Y-chromosomal haplogroups in Central Asian populations (table 3; fig. 1).

The Muslim religion arose in the Middle East and subsequently spread into Asia, but was this a purely cultural phenomenon, with little population movement and therefore little genetic impact—or was it accompanied by significant gene flow? The eastern population that shows relatively high Middle Eastern admixture (fig. 7), the Uyghurs, are Muslims who moved to Central Asia from China within the last couple of centuries (Du and Yip 1993). They—and, to a lesser extent, the Dungans (also Muslims)—have high frequencies of haplogroup 9. This haplogroup is extremely rare in China, but notably it is present in the Hui, the biggest Muslim community in China (unpublished observations), indicating that Chinese Muslim communities formed by the movement of people (or at least of men) from the Middle East. Thus, the Uyghur and Dungan Y-chromosomal haplogroup compositions can be explained by migration from the Middle East during historical times—but indirectly, via China.

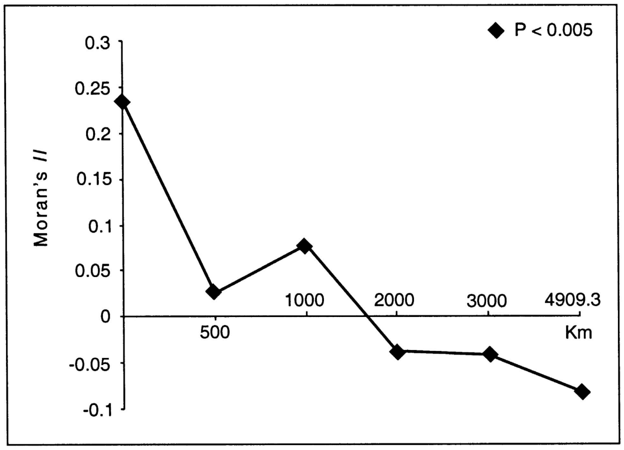


Figure 6 AIDA analysis showing the correlation of genetic distance (Moran's *II*) with geographical distance.

Expansion of Eastern Nomadic Groups

From the 3d century B.C., Central Asia experienced nomadic expansions of Altaic-speaking oriental-looking people, and their incursions continued for hundreds of years, beginning with the Hsiung-Nu (who may be ancestors of the Huns), in ~300 B.C., and followed by the

Table 6

AMOVA Results

	Value for Binary Markers				Value for Microsatellites			
Grouping and Covariance Components	Variance (%)	Φ_{ST}	Φ_{SC}	Φ_{CT}	Variance (%)	Φ_{ST}	Φ_{SC}	Φ_{CT}
No grouping:								
Among populations	23.6				19			
Within populations	76.4	.24			81	.19		
Geography:								
Among groups	9.6			.09	6.4			.06
Among populations within groups	18.1		.2		15.2		.17	
Within populations	72.3	.28			78.5	.22		
Language:								
Among groups	5.8			.06*	4.4			.04
Among populations within groups	20.5		.24		16.7		.17	
Within populations	73.8	.26			79	.21		

NOTE.—All *P* values are <.05, except where noted.
* *P* = .1.

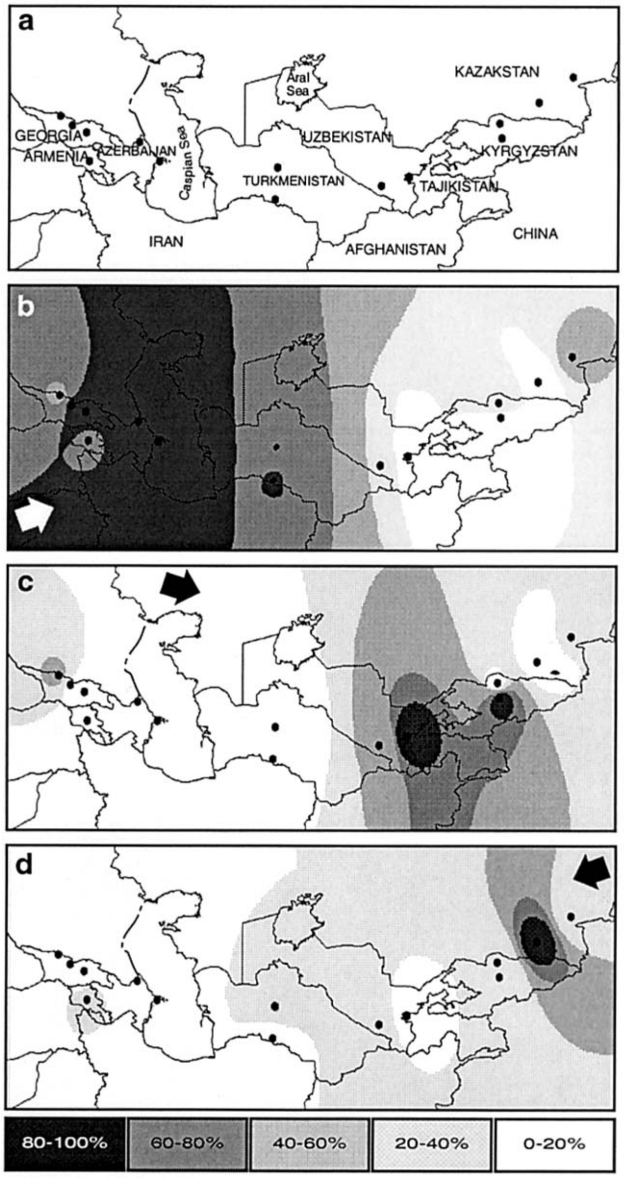


Figure 7 Interpolated maps of admixture proportions. The direction from the source population is shown by the arrow. The estimated proportion of admixture was divided into the five categories summarized by the gray scale in the lowest section.

Turks, in the 1st millennium A.D., and the Mongol expansions of the 13th century. High levels of haplogroup 10 and its derivative, haplogroup 36, are found in most of the Altaic-speaking populations and are a good indicator of the genetic impact of these nomadic groups. The expanding waves of Altaic-speaking nomads involved not only eastern Central Asia—where their genetic contribution is strong, as is shown in figure 7*d*—but also regions farther west, like Iran, Iraq, Anatolia, and the Caucasus, as well as Europe, which was reached by both the Huns and the Mongols. In these western regions, however, the genetic contribution is low or undetectable (Wells et al. 2001), even though the power of these invaders was sometimes strong enough to impose a language replacement, as in Turkey and Azerbaijan (Cavalli-Sforza et al. 1994). The difference could be due to the population density of the different geographical areas. Eastern regions of Central Asia must have had a low population density at the time, so an external contribution could have had a great genetic impact. In contrast, the western regions were more densely inhabited, and it is likely that the existing populations were more numerous than the conquering nomads, therefore leading to only a small genetic impact. Thus, the admixture estimate from northeast Asia is high in the east, but is barely detectable west of Uzbekistan.

Expansions Originating in Central Asia

Previous analyses have identified Asia in general (Hammer et al. 2001), by use of nested cladistic analysis, or Central Asia in particular (Wells et al. 2001), as a source of signals of (1) population-history events, such as range expansions and long-distance colonizations, and (2) population structure processes, such as gene flow restricted by isolation by distance and long-distance dispersals. Several of these signals originate from areas to the east or south of the region examined here, but our results are consistent with the suggested origin of haplogroup 1 in Central Asia (equivalent to M45 and derivatives [Wells et al. 2001]; h36 and derivatives [Hammer et al. 2001]) and thus with Central Asia as a source of population signals as well as a receiver.

Central Asia: The Land of Bottlenecks

Many of the events discussed above would tend to produce smooth distributions. They could account for differences between east and west. Yet, low Y diversity and the presence of population-specific clusters of lineages at high frequency are very striking features of several Central Asian populations. It appears that recent bottlenecks have occurred independently in multiple populations.

The time estimates, particularly those calculated using BATWING, have broad confidence limits, and these overlap between the different lineages, so it is difficult to be certain that they correspond to different times in history. However, it is notable that one of the older estimates is for haplogroup 3 in the Tajiks, who historically were agriculturalists but claim to be the direct descendents of the original European nomadic groups. The TMRCAs for the Kazaks and the Kyrgyz might be explained by the massacres that their ancestors suffered during the expansion of the Mongols under the leadership of Genghis Khan, in the 13th century A.D. In the Caucasus, both the Svans and the Georgians exhibit genetic patterns consistent with multiple bottlenecks or founder events at different times, and this might be a consequence of their small population sizes, with linguistic and geographic isolation making them particularly susceptible to drift and demographic change.

Central Asia: the Social Structure

Even if populations did experience size reductions, it is possible that the effects have been made even more dramatic because of their social structure. Central Asian pastoral nomadic societies are characterized by relatively small groups or clans, often representing patrilineal family units related via the male line. Traditionally, all men are members of their fathers' clan and must marry wives from outside (Forde 1948). In this social context, founder effect and drift must have had a powerful effect because of the small population sizes, further enhanced by patrilocality. Perez-Lezaun and colleagues (1999) found evidence of male founder effects in two high-altitude Central Asian populations (Kazaks and Kyrgyz), which they could explain most satisfactorily by the social structure. The predominant haplotype in their Kyrgyz population sample was also the most frequent in the sample examined here, but, remarkably, the predominant haplotypes found in the two Kazak samples are separated by two mutational steps among six loci that can be compared. The distinction between the two population subsamples is confirmed by the high Φ_{ST} value between them (0.3), which is comparable to intercontinental Φ_{ST} values obtained by Kayser et al. (2001). Our Kazak sample was collected from four villages located in different parts of eastern Kazakstan, and shows no significant difference between the villages. Although this sample represents a large region, it is clearly not a complete sample of the entire Kazak population, which manifests considerable substructure in our analysis.

Conclusions

Central Asia is a land of genetic extremes with very high and very low genetic diversities. The severe bottlenecks found in several of the populations (table 5) show that this phenomenon is not limited to the high-altitude populations in which it was first reported by Perez-Lezaun et al. (1999) but is more widespread in the area and may have an explanation in the small population size of several of the groups, enhanced by the patrilocal social structure of many of the nomadic populations. With such extensive variation between populations, a wider survey is needed to provide a comprehensive view of Y-chromosomal variation in the whole area. Nevertheless, since large regions of the world are sparsely populated, Central Asia may provide a better paradigm for understanding their genetic structure than the dense populations of Europe.

Acknowledgments

We thank all of the donors for making this work possible, Nathaniel Pearson and Matthew Webster for contributions to sample collection, and Marcos Llobera for producing the contour maps. T.Z. was supported by The Wellcome Trust, R.S.W. was supported in part by the Alfred P. Sloan Foundation, and C.T.-S. was supported by the Cancer Research Campaign.

Electronic-Database Information

Admix 2.0, http://www.unife.it/genetica/Isabelle/admix2_0.html

Ethnologue, http://www.ethnologue.com/ (for Ethnologue languages database)

EurAsia '98, http://popgen.well.ox.ac.uk/eurasia/htdocs/index.html

Fluxus Engineering, http://www.fluxus-engineering.com/sharenet.htm (for NETWORK 3.0 phylogenetic network analysis software)

Geonames, http://gnpswww.nima.mil/geonames/GNS/index.jsp

Goldstein Laboratory of Population and Human Genetics, http://www.ucl.ac.uk/biology/goldstein/Gold.htm (for Ymrca program)

Ian Wilson's home page, http://www.maths.abdn.ac.uk/˜ijw (for BATWING program)

Population Genetics and Genetic Epidemiology Group Software Page, http://www.unife.it/genetica/Giorgio/giorgio_soft.html (for AIDA program)

References

Anthony DW (1986) The "Kurgan culture," Indo-European origins, and the domestication of the horse: a reconsideration. Curr Anthropol 27:291–313

Ayub Q, Mohyuddin A, Qamar R, Mazhar K, Zerjal T, Mehdi SQ, Tyler-Smith C (2000) Identification and characterization of novel human Y-chromosomal microsatellites from sequence database information. Nucleic Acids Res 28:e8

Bandelt HJ, Forster P, Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37–48

Bao W, Zhu S, Pandya A, Zerjal T, Xu J, Shu Q, Du R, Yang H, Tyler-Smith C (2000) MSY2: a slowly evolving minisatellite on the human Y chromosome which provides a useful polymorphic marker in Chinese populations. Gene 244:29–33

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

Barbujani G, Pilastro A, De Domenico S, Renfrew C (1994) Genetic variation in North Africa and Eurasia: neolithic demic diffusion vs. Paleolithic colonization. Am J Phys Anthropol 95:137–154

Bertorelle G, Barbujani G (1995) Analysis of DNA diversity by spatial autocorrelation. Genetics 140:811–819

Carvalho-Silva DR, Santos FR, Rocha J, Pena SD (2001) The phylogeography of Brazilian Y-chromosome lineages. Am J Hum Genet 68:281–286

Cavalli-Sforza LL, Menozzi P, Piazza A (1993) Demic expansions and human evolution. Science 259:639–646

——— (1994) The history and geography of human genes. Princeton University Press, Princeton

Davis RS, Ranov VA (1979) Toward a new outline of the Soviet Central Asian Paleolithic. Curr Anthropol 20:249–270

de Knijff P (2000) Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. Am J Hum Genet 67:1055–1061

de Knijff P, Kayser M, Caglia A, Corach D, Fretwell N, Gehrig C, Graziosi G, et al (1997) Chromosome Y microsatellites: population genetic and evolutionary aspects. Int J Legal Med 110:134–149

De Rosa L (1992) Silk and the European economy. In: Umesao T, Sugimura T (eds) Significance of silk roads in the history of human civilizations. National Museum of Ethnology, Osaka, pp 193–205

Dexter MR, Jones-Bley K (1997) The Kurgan culture and the Indo-Europeanization of Europe: selected articles from 1952 to 1993. Institute for the Study of Man, Washington DC

Du R, Yip VF (1993) Ethnic groups in China. Science Press, Beijing

Dupanloup I, Bertorelle G (2001) Inferring admixture proportions from molecular data: extension to any number of parental populations. Mol Biol Evol 18:672–675

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491

Forde CD (1948) Habitat, economy and society: a geographical introduction to ethnology. Methuen & Co, London

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

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

Hammer MF, Redd AJ, Wood ET, Bonner MR, Jarjanazi H, Karafet T, Santachiara-Benerecetti S, Oppenheim A, Jobling MA, Jenkins T, Ostrer H, Bonne-Tamir B (2000) Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. Proc Natl Acad Sci USA 97:6769–6774

Hurles ME, Irven C, Nicholson J, Taylor PG, Santos FR, Loughlin J, Jobling MA, Sykes BC (1998) European Y-chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA. Am J Hum Genet 63: 1793–1806

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

Karafet T, Xu L, Du R, Wang W, Feng S, Wells RS, Redd AJ, Zegura SL, Hammer MF (2001) Paternal population history of East Asia: sources, patterns, and microevolutionary processes. Am J Hum Genet 69:615–628

Karafet TM, Zegura SL, Posukh O, Osipova L, Bergen A, Long J, Goldman D, Klitz W, Harihara S, de Knijff P, Wiebe V, Griffiths RC, Templeton AR, Hammer MF (1999) Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. Am J Hum Genet 64:817–831

Kato K (1992) Cultural exchange on the ancient Steppe route: some observations on Pazyryk heritage. In: Umesao T, Sugimura T (eds) Significance of silk roads in the history of human civilizations. National Museum of Ethnology, Osaka, pp 5–20

Kayser M, Krawczak M, Excoffier L, Dieltjes P, Corach D, Pascali V, Gehrig C, Bernini LF, Jespersen J, Bakker E, Roewer L, de Knijff P (2001) An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. Am J Hum Genet 68:990–1018

Kayser M, Roewer L, Hedman M, Henke L, Henke J, Brauer S, Kruger C, Krawczak M, Nagy M, Dobosz T, Szibor R, de Knijff P, Stoneking M, Sajantila A (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Am J Hum Genet 66:1580–1588

Lewontin R (1972) The apportionment of human diversity. Evol Biol 6:381–398

Lin M-C (1992) Tocharian people: silk road pioneers. In: Umesao T, Sugimura T (eds) Significance of silk roads in the history of human civilizations. National Museum of Ethnology, Osaka, pp 91–96

Pandya A, King TE, Santos FR, Taylor PG, Thangaraj K, Singh L, Jobling MA, Tyler-Smith C (1998) A polymorphic human Y-chromosomal G to A transition found in India. Ind J Hum Genet 4:52–61

Perez-Lezaun A, Calafell F, Comas D, Mateu E, Bosch E, Martinez-Arias R, Clarimon J, Fiori G, Luiselli D, Facchini F, Pettener D, Bertranpetit J (1999) Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. Am J Hum Genet 65:208–219

Qamar R, Ayub Q, Mohyuddin A, Helgason A, Mazhar K, Mansoor A, Zerjal T, Tyler-Smith C, Mehdi SQ (2002) Y-chromosomal DNA variation in Pakistan. Am J Hum Genet 70:1107–1124

Quintana-Murci L, Krausz C, Zerjal T, Sayar SH, Hammer MF, Mehdi SQ, Ayub Q, Qamar R, Mohyuddin A, Radhakrishna U, Jobling MA, Tyler-Smith C, McElreavey K (2001) Y-chromosome lineages trace diffusion of people and languages in southwestern Asia. Am J Hum Genet 68:537–542

Ranov VA, Carbonell E, Rodriguez XP (1995) Kuldara: earliest human occupation in central Asia in its Afro-Asian context. Curr Anthropol 36:337–346

Richards MB, Macaulay VA, Bandelt HJ, Sykes BC (1998) Phylogeography of mitochondrial DNA in western Europe. Ann Hum Genet 62:241–260

Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim A, Amos W, et al (2000) Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am J Hum Genet 67:1526–1543

Santos FR, Pandya A, Kayser M, Mitchell RJ, Liu A, Singh L, Destro-Bisol G, Novelletto A, Qamar R, Mehdi SQ, Adhikari R, de Knijff P, Tyler-Smith C (2000) A polymorphic L1 retroposon insertion in the centromere of the human Y chromosome. Hum Mol Genet 9:421–430

Santos FR, Pandya A, Tyler-Smith C, Pena SD, Schanfield M, Leonard WR, Osipova L, Crawford MH, Mitchell RJ (1999)

The central Siberian origin for Native American Y chromosomes. Am J Hum Genet 64:619–628

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Underhill PA (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. Science 290:1155–1159

Sokal RR, Harding RM, Oden NL (1989) Spatial patterns of human gene frequencies in Europe. Am J Phys Anthropol 80:267–294

Stumpf MP, Goldstein DB (2001) Genealogical and evolutionary inference with the human Y chromosome. Science 291: 1738–1742

Su B, Xiao J, Underhill P, Deka R, Zhang W, Akey J, Huang W, Shen D, Lu D, Luo J, Chu J, Tan J, Shen P, Davis R, Cavalli-Sforza L, Chakraborty R, Xiong M, Du R, Oefner P, Chen Z, Jin L (1999) Y-chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. Am J Hum Genet 65:1718–1724

Thomas MG, Bradman N, Flinn HM (1999) High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. Hum Genet 105:577–581

Wells RS, Yuldasheva N, Ruzibakiev R, Underhill PA, Evseeva I, Blue-Smith J, Jin L, et al (2001) The Eurasian heartland: a continental perspective on Y-chromosome diversity. Proc Natl Acad Sci USA 98:10244–102449

Wilson IJ, Balding DJ (1998) Genealogical inference from microsatellite data. Genetics 150:499–510

Y-Chromosome Consortium (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res 12:339–348

Zerjal T, Beckman L, Beckman G, Mikelsaar AV, Krumina A, Kucinskas V, Hurles ME, Tyler-Smith C (2001) Geographical, linguistic, and cultural influences on genetic diversity: Y-chromosomal distribution in Northern European populations. Mol Biol Evol 18:1077–1087

Zerjal T, Dashnyam B, Pandya A, Kayser M, Roewer L, Santos FR, Schiefenhovel W, Fretwell N, Jobling MA, Harihara S, Shimizu K, Semjidmaa D, Sajantila A, Salo P, Crawford MH, Ginter EK, Evgrafov OV, Tyler-Smith C (1997) Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis. Am J Hum Genet 60: 1174–1183

Zerjal T, Pandya A, Santos FR, Adhikari R, Tarazona-Santos E, Kayser M, Evgrafov OV, Singh L, Thangaraj K, Destro-Bisol G, Thomas M, Qamar R, Mehdi SQ, Rosser ZH, Hurles ME, Jobling MA, Harihara S, Tyler-Smith C (1999) The use of Y-chromosomal DNA variation to investigate population history: recent male spread in Asia and Europe. In: Papiha SS, Deka R, Chakraborty R (eds) Genomic diversity: applications in human population genetics. Plenum, New York, pp 91–101