High Levels of Y-Chromosome Differentiation among Native Siberian Populations and the Genetic Signature of a Boreal Hunter-Gatherer Way of Life

TATIANA M. KARAFET,[1,3] LUDMILA P. OSIPOVA,[3] MARINA A. GUBINA,[3]
OLGA L. POSUKH,[3] STEPHEN L. ZEGURA,[2] AND MICHAEL F. HAMMER[1,2]

Abstract We examined genetic variation on the nonrecombining portion of the Y chromosome (NRY) to investigate the paternal population structure of indigenous Siberian groups and to reconstruct the historical events leading to the peopling of Siberia. A set of 62 biallelic markers on the NRY were genotyped in 1432 males representing 18 Siberian populations, as well as nine populations from Central and East Asia and one from European Russia. A subset of these markers defines the 18 major NRY haplogroups (*A*-*R*) recently described by the Y Chromosome Consortium (YCC 2002). While only four of these 18 major NRY haplogroups accounted for ~95% of Siberian Y-chromosome variation, native Siberian populations differed greatly in their haplogroup composition and exhibited the highest Φ_{ST} value for any region of the world. When we divided our Siberian sample into four geographic regions versus five major linguistic groupings, analyses of molecular variance (AMOVA) indicated higher Φ_{ST} and Φ_{CT} values for linguistic groups than for geographic groups. Mantel tests also supported the existence of NRY genetic patterns that were correlated with language, indicating that language affiliation might be a better predictor of the genetic affinity among Siberians than their present geographic position. The combined results, including those from a nested cladistic analysis, underscored the important role of directed dispersals, range expansions, and long-distance colonizations bound by common ethnic and linguistic affiliation in shaping the genetic landscape of Siberia. The Siberian pattern of reduced haplogroup diversity within populations combined with high levels of differentiation among populations may be a general feature characteristic of indigenous groups that have small effective population sizes and that have been isolated for long periods of time.

There are several important reasons to study genetic variation in native Siberian populations. First, Siberia occupies the greatest part of North Asia and extends

[1]Division of Biotechnology, University of Arizona, Tucson, Arizona 85721.
[2]Department of Anthropology, University of Arizona, Tucson, Arizona 85721.
[3]Laboratory of Human Molecular and Evolutionary Genetics, Institute of Cytology and Genetics, Novosibirsk, Russia.

Human Biology, December 2002, v. 74, no. 6, pp. 761–789.

KEY WORDS: Y-CHROMOSOME HAPLOGROUPS, SIBERIA, LANGUAGE FAMILIES, HUNTER-GATHERERS, GENETIC DRIFT, LONG DISTANCE COLONIZATION

from the Ural Mountains in the west to the Pacific watershed in the east and from the Arctic Ocean in the north to Kazakhstan, Mongolia, and China in the south. As such, it forms an important geographic link between the Asian and North American continents and between North Asia and the Japanese Archipelago. Because Siberia serves as a source and/or transit point for human dispersals to the Americas and to Japan, it is crucial that this region's own settlement history be understood so that testable data-based migration models can be formulated for both gateway and destinations. Second, Siberia is among the few places in the world where, until recently, most people lived a foraging lifestyle. Siberia's economic and cultural patterns have links traceable to Paleolithic and Neolithic subsistence strategies (Rychkov and Sheremet'eva 1980). Population density in Siberia has been quite low partly because of resource limitations, and traditional Siberian life-ways reflect common features of hunter-gatherer existence throughout much of the Arctic and sub-Arctic ecosystems. Thus, it has been postulated that surveys of genetic variation in indigenous groups (such as those in Siberia) will provide the opportunity to investigate aspects of population structure that have characterized humans from the Pleistocene to the present (Birdsell 1973; Cavalli-Sforza 1986). These surveys will also help to test archaeological- and language-based hypotheses about the history of Siberian populations.

The archaeological record suggests that the early settlement of Siberia was a complex, lengthy process with at least four proposed source regions: Central Asia, Mongolia, North China, and southern Russia/eastern Europe (Okladnikov 1981, 1983; Vasil'ev 1993; Derev'anko 1998a). The first Siberian Paleolithic archaeological site was discovered in 1871 (Derev'anko 1998a). Although a considerable amount of material has been unearthed since then, controversy and unsolved problems complicate the interpretation of Siberian ancient remains. For instance, the precise antiquity of a human presence in the Arctic is still an open question. The earliest C^{14}-dated Asian Upper Paleolithic industries occur in the Altai Mountains of Southwest Siberia at 43,300 ± 1,600 years BP (Goebel et al. 1993; Kuzmin and Orlova 1998). There are indications of earlier habitations; however, neither the site chronologies nor the identities of the tools (or tool-makers) have been sufficiently clarified to permit their wide acceptance.

The recorded history of Siberia begins with the Russian invasions in the late 16th century AD. Today, 31 different populations are indigenous to Siberia, of which 26 are considered to be small "ethnic" groups. These small ethnohistoric/linguistic groups constitute only 2% of the population in Siberia. While it is estimated that the number of linguistic communities encountered by early Russian colonists was on the order of 120 (Levin and Potapov 1964), today there are only approximately 35 indigenous languages recognized in Siberia. Although differing in their origin, language, and culture, most Siberian populations are characterized by common types of economic activities involving hunting, fishing, reindeer breeding, and cattle herding. These traditional occupations are closely linked to nomadic and seminomadic ways of life. Most Siberian indigenous groups share a number of common sociocultural features such as clan structure, polygamous marriages, the levirate (the compulsory marriage of a widow to a younger brother of her deceased husband), and a high level of endogamy. Until the 1970s many Siberian peoples had not experienced much gene flow from nonnative populations, although interindigenous population gene flow seems to have been more common.

Earlier genetic studies based on blood groups and classical markers consistently showed a high degree of between-group heterogeneity for Siberian populations often attributed to low population densities (Rychkov and Sheremet'eva 1980; Szathmary 1981; Sukernik et al. 1986; Posukh et al. 1990; Novoradovsky et al. 1993; Cavalli-Sforza et al. 1994; Osipova et al. 1996), as well as the existence of statistically significant relationships among geographic, linguistic, and genetic variation (Crawford and Enciso 1982; Cavalli-Sforza et al. 1994; Karafet et al. 1994), and a clear demarcation line between eastern and western Siberian populations (Sukernik et al. 1978, 1981; Karafet et al. 1981; Szathmary 1981; Osipova et al. 1996).

Recent genetic surveys of Siberian populations have mainly involved the haploid regions of the genome, and have primarily focused on the peopling of the Americas (Shields et al. 1993; Torroni et al. 1993; Karafet et al. 1997, 1999; Lell et al. 1997, 2002; Starikovskaya et al. 1998; Santos et al. 1999; Derenko et al. 2001), on the history of early human migrations in Eurasia (Zerjal et al. 1997, 1999) and East Asia (Su et al. 2000; Ke et al. 2001), or on particular regions of Siberia (Schurr et al. 1999; Derenko et al. 2000a, b; Stepanov and Puzyrev 2000; Pakendorf et al. 2002). Mitochondrial DNA (mtDNA) studies revealed that modern indigenous southern Siberian populations show traces of an eastward expansion from Central Asia (Derenko et al. 2000a, b), while eastern Siberian populations from the Kamchatka Peninsula and the Amur River/Sea of Okhotsk region have stronger genetic affinities with East Asian populations (Schurr et al. 1999). The NRY study of Su et al. (1999) suggested that mainland Southeast Asia was the homeland for all eastern Asian populations including those now living in Siberia, while that of Karafet et al. (2001) underscored an important role for Central Asia in the early peopling of Siberia.

Because the NRY is subject to higher levels of genetic drift than other regions of the genome (Hammer and Zegura 1996), it is an excellent tool for detecting between-group variation and for reconstructing the history of human migrations. This study was designed to characterize patterns of NRY variation in a set of contemporary hunter-gatherers and to provide a baseline for similar comparisons of other hunter-gatherer groups. In addition to technology and economy there are a number of supplementary factors that contribute to population structure (Fix 1999). For instance, geographic barriers and cultural/ethnic/language boundaries are well-recognized phenomena that limit migration. Below we examine large-scale patterns of SNP variation on the Y chromosomes of North Asians, including markers that define all 18 major haplogroups on the recently published Y Chromosome Consortium haplogroup tree (YCC 2002), to investigate (1) the population structure of Siberian foraging groups; (2) the correspondences among genetic, linguistic, and geographic variation; and (3) the history of the early colonization of Siberia.

Subjects and Methods

Populations and DNA Samples. We analyzed 62 binary polymorphisms on the Y chromosomes of 902 males from 18 Siberian populations. Ten additional populations thought to have had contacts with Siberian groups were also included in our analyses. Thus, our total sample comprises 1432 Y chromosomes from 28 populations. These 28 populations were divided into six regional groupings based on arbitrary geographic criteria as follows: (1) Northwest Siberia (10 groups), (2) Central-South Siberia (6 groups), (3) Northeast Siberia (4 groups), (4) Central Asia (5 groups), (5) East Asia (2 groups), and (6) European Russia (1 group). The approximate geographic locations of the sampling sites are shown in Figure 1.

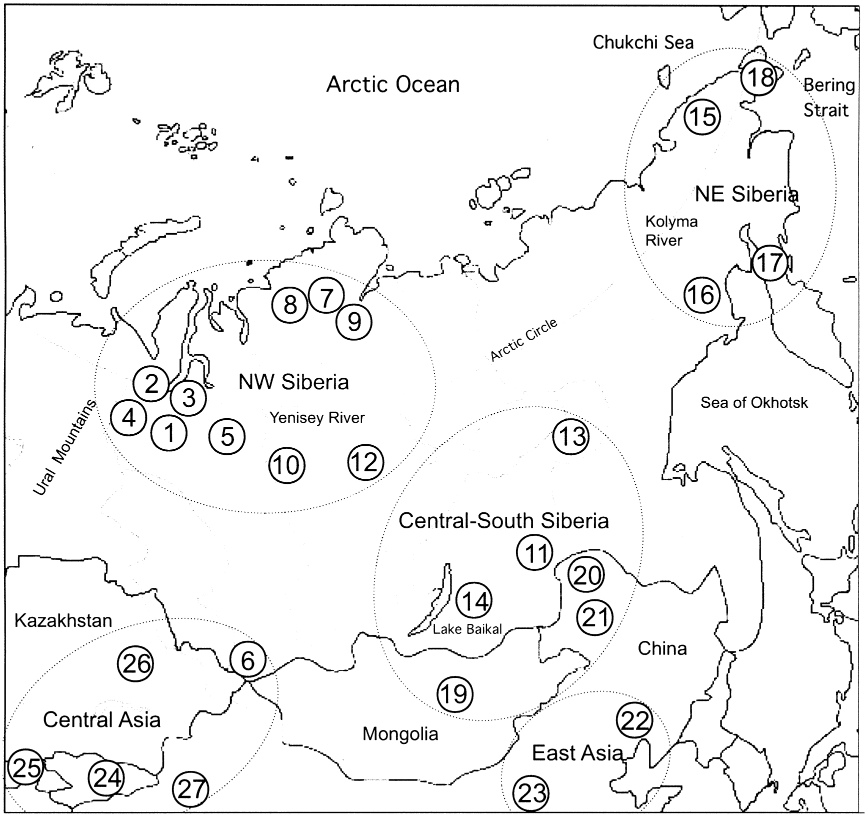

Figure 1. Map of 27 North Asian sampling localities divided into five regional groupings. See Table 1 for names, sizes, and linguistic affiliations of population samples. The European Russian population (28) representing the sixth regional grouping is not shown.

Regional geographic membership, linguistic affiliations, sample sizes, population size estimates, genetic diversity statistics, and three-letter and numerical population codes are given in Table 1. Note that in accord with Russian ethnographic tradition, the Altais in the Central Asia geographic grouping are considered to be "Siberians" for all ensuing statistical analyses, while the Mongolians, Chinese Evenks, and Oroqens in the Central-South Siberia regional grouping are not. Many of the samples analyzed here were included in our previous studies (Karafet et al. 1999, 2001; Hammer et al. 2001). New samples from the Khants, Komi, Nganasans, Dolgans, and Entsi were collected by T. Karafet and L. P. Osipova of the Institute of Cytology and Genetics in Novosibirsk with informed consent during 1999–2000 in the Yamal-Nenets and Taymyr Autonomous Districts. Additional genomic DNAs from the Kets, Forest Nentsi, and Yakut-Sakha populations were collected by the laboratory of L. P. Osipova in Russia. All samples were collected in traditional settlements. Demographic and pedigree data were obtained along with blood samples. From demographic and genealogical information we were able to identify paternally unrelated males (for at least three to six generations). All sampling protocols were approved by the Human Subjects Committee at the University of Arizona.

Genetic Markers. The polymorphic sites in our survey included a set of 52 previously published binary NRY markers (Karafet et al. 2001). Additionally, we genotyped seven polymorphisms (M52, M60, M86, M91, M128, M178, and M201) reported by Underhill et al. (2000), the LLY22g polymorphism (Zerjal et al., in preparation), and two newly discovered polymorphisms (see below). These last three markers represent subclades of three common haplogroups found in North Asia (Karafet et al. 2001) and should provide useful information to address questions about the population structure and history of Siberian peoples. The C→T transition at M86 was genotyped by digestion with the *Dra*I enzyme (New England Biolabs), the T→C transition at M178 was typed using allele-specific polymerase chain reaction (PCR), and the other five M markers were typed as reported in Underhill et al. (2000, 2001). Information for typing LLY22g will be presented by Zerjal et al. elsewhere. We followed the hierarchical typing strategy explained in Underhill et al. (2000) and Hammer et al. (2001), wherein additional genotyping of a sample was restricted to markers on the appropriate branch of the haplogroup tree.

We also typed a G→A transition at position 72,425 of the arylsulfatase D pseudogene (*ARSDP*) and a C→T transition at position 20,784 of the 16E4 clone (AC003094). These two new mutations (subsequently referred to as P36 and P43, respectively) were discovered in a panel of 57 Y chromosomes from sub-Saharan Africa, Asia, Europe, Oceania, and the Americas (Hammer et al. 2001) using denaturing high performance liquid chromatography (DHPLC). The P36 G→A mutation was typed by allele-specific PCR. The following primers were used to amplify the 238–base pair (bp) allele-specific band: P36U (5′-TGAAG-GACAGTAAGTACACA-3′), P36AL (5′-TATCTATCCATTATTCTCTCTA-3′),

Table 1. Sample Composition, Linguistic Affiliations, and Genetic Diversity for 28 Populations

Population	*Linguistic Affiliation*[a]	*Sample Size*	*Population Size*[b]	*h* ± *SE*	*p* ± *SE*
Northwest Siberia ($n = 534$)					
1. Forest Nentsi (FNE)	Uralic (N. Samoyed)	89	1,500	0.55 ± 0.00	1.66 ± 0.10
2. Tundra Nentsi (TNE)	Uralic (N. Samoyed)	59	10,000	0.39 ± 0.01	1.22 ± 0.10
3. Komi (KOM)	Uralic (Finno-Ugric)	28	7,000	0.62 ± 0.01	2.78 ± 0.29
4. Khants (KHA)	Uralic (Finno-Ugric)	47	22,300	0.71 ± 0.00	3.29 ± 0.25
5. Selkups (SEL)	Uralic (S. Samoyed)	131	2,000	0.52 ± 0.00	2.23 ± 0.11
7. Nganasans (NGA)	Uralic (N. Samoyed)	38	750	0.20 ± 0.01	1.02 ± 0.11
8. Entsi (ENE)	Uralic (N. Samoyed)	9	120	0.58 ± 0.06	1.39 ± 0.31
9. Dolgans (DOL)	Altaic (Turkic)	67	6,600	0.78 ± 0.00	5.93 ± 0.35
10. Kets (KET)	Yeniseian (isolate)	48	500	0.12 ± 0.01	0.96 ± 0.10
12. Western Evenks (WEV)	Altaic (Tungus)	18	29,000	0.65 ± 0.01	3.78 ± 0.45
Central-South Siberia ($n = 402$)					
11. Eastern Evenks (EEV)	Altaic (Tungus)	78	29,000	0.65 ± 0.01	4.55 ± 0.27
13. Yakuts-Sakha (YAK)	Altaic (Turkic)	35	296,000	0.11 ± 0.01	1.00 ± 0.12
14. Buryats (BUR)	Altaic (Mongolian)	81	314,000	0.61 ± 0.00	4.25 ± 0.24
19. Mongolian-Khalks (MON)	Altaic (Mongolian)	145	1,500,000	0.82 ± 0.00	4.91 ± 0.20
20. Chinese Evenks (CEV)	Altaic (Tungus)	40	26,000	0.86 ± 0.01	5.54 ± 0.43
21. Oroqens (ORO)	Altaic (Tungus)	23	7,000	0.49 ± 0.02	1.71 ± 0.22
Northeast Siberia ($n = 76$)					
15. Yukaghirs (YUK)	Uralic (Yukaghir)	11	100	0.82 ± 0.02	5.13 ± 0.81
16. Evens (EVN)	Altaic (Tungus)	31	17,000	0.60 ± 0.02	3.88 ± 0.36
17 Koryaks (KOR)	Chukchi-Kamchatkan (Chukchi)	12	8,900	0.80 ± 0.02	5.36 ± 0.80
18. Eskimos (ESK)	Eskimo-Aleut (Yupik)	22	600	0.64 ± 0.02	3.10 ± 0.36
East Asia ($n = 96$)					
22. Manchu (MAN)	Altaic (Manchu)	52	9,820,000	0.88 ± 0.00	4.44 ± 0.31
23. Northern Han (NHA)	Sino-Tibetan	44	1,042,482,000	0.80 ± 0.01	2.98 ± 0.24

Central Asia (*n* = 263)					
6. Altais (ALT)	Altaic (Turkic)	98	57,000	0.72 ± 0.00	4.77 ± 0.24
24. Uzbeks (UZB)	Altaic (Turkic)	54	9,200,000	0.92 ± 0.01	5.26 ± 0.35
25. Kirghiz (KIR)	Altaic (Turkic)	13	2,500,000	0.62 ± 0.04	4.00 ± 0.59
26. Kazakhs (KAZ)	Altaic (Turkic)	30	4,200,000	0.89 ± 0.01	5.19 ± 0.47
17. Uygurs (UYG)	Altaic (Turkic)	68	7,200,000	0.91 ± 0.00	4.86 ± 0.29
European Russia (*n* = 61)					
28. Russians (RUS)	Indo-European	61	107,000,000	0.76 ± 0.01	3.93 ± 0.26

a. See Ruhlen (1991); Ruhlen (1998); Greenberg (2000).

b. See Ruofu and Yip (1993); Karafet et al. (1994); formal census data and unpublished estimates.

and the 553-bp control band: U (5′-ACCCTTCCCTTCATATTTT-3′), L (5′-GGCATAAACTACCTGGAAA-3′). Ten ng of genomic DNA was amplified in 15 µL final volume containing 0.2 m*M* of each dNTP, 1 µ*M* of each primer, 0.046 µ*M* of TaqStart Antibody (Clontech), 0.0016 µ*M* of Taq DNA polymerase (Eppendorf), 3.0 m*M* $MgCl_2$, 50 m*M* KCl, and 10 m*M* Tris-HCl (pH 8.3). The cycling conditions were 94°C for 3 min, followed by 20 touchdown cycles with –0.5°C/cycle increments at 94°C, 63°→53°C, 72°C for 30 s, then 20 cycles of standard amplification at 94°C, 53°C, 72°C for 30 s, with final extension step at 72°C for 2 min.

A 519-bp segment encompassing the polymorphic site at P43 was PCR-amplified using the following primer pairs: P43-R (5′-GAAGCAATACTCTGAAAAGT-3′) and P43-F (5′-TTTGGAGGGACATTATTCTC-3′). The PCR conditions for this amplification were 94°C for 3 min, followed by 35 cycles at 94°C, 53°C, 72°C for 30 s, with final extension step at 72°C for 2 min. Reactions were run in a final volume of 15 µL containing 10 ng of genomic DNA, 0.2 m*M* each dNTP, 1 µ*M* each primer, 0.046 µ*M* of TaqStart Antibody, 0.0016 µ*M* of Taq DNA polymerase, 4.0 m*M* $MgCl_2$, 50 m*M* KCl, and 10 m*M* Tris-HCl (pH 8.3). The P43 C→T polymorphism at position 268 was genotyped by digestion with the *Nla*III enzyme.

Terminology. We follow the terminology recommended by the Y Chromosome Consortium (YCC 2002) for naming NRY lineages. A cladistic tree of lineages defined by mutational events on the NRY can be seen as a series of nested monophyletic clades. A hierarchical system was devised by the YCC (2002) to enable clades at all levels of the nested series to be named unambiguously with respect to the markers typed in any particular study. Capital letters A-R are used to identify 18 major clades or haplogroups. The letter Y is reserved for the most inclusive haplogroup encompassing the entire tree. Lineages that are not defined on the basis of a derived character represent interior nodes of the tree and are potentially paraphyletic (i.e., comprised of basal lineages and monophyletic subclades). Thus, the term "paragroup" rather than haplogroup is used to describe these lineages. Paragroup names are distinguished by using the * star symbol. Lineages excluded from a haplogroup/paragroup are listed after an initial "x" symbol within parentheses after the name of the haplogroup/paragroup (or the last derived marker in the case of the mutation-based nomenclature). We opted to omit the "x" notation and parenthetical system for mutation-based names. See Table 2 for a list of the lineage- and mutation-based names of the Siberian haplogroups/paragroups defined by the markers used in this study.

Statistical Analyses. Population genetic structure indices (molecular variances and Φ statistics) and measures of haplogroup diversity including Nei's heterozygosity (*h*) and the mean number of pairwise differences among haplogroups (*p*) were estimated by ARLEQUIN 2.000 software (Schneider et al. 2000). Both haplogroup frequencies and molecular differences among haplogroups are taken into account with this approach. Variance components due to different sources of

Table 2. Lineage- and Mutation-Based Names of 23 Siberian Haplogroups/Paragroups in Figure 1

Lineage-Based	*Mutation-Based*	*Derived*	
Name	*Name*[a]	*State at*	*Ancestral State at*
C3*(xC3c)	C-M217*	M217	M86
C3c	C-M86	M86	
D*(xD1)	D-M174*	M174	M15
E3*(xE3a)	E-P2*	P2	P1
F*(xG,H,I,J,K)	F-P14*	P14	M201, P19, 12f2a, M9, M52
G*(xG2)	G-M201*	M201	P15
I*(xI1b)	I-P19*	P19	P37b
I1b	I-P37b	P37b	
J2*(xJ2e)	J-M172*	M172	M12
J2e	J-M12	M12	
K*(xL,M,N,O,P)	K-M9*	M9	LLY, M20, M4, M175, P27
N*	N-LLY22g*	LLY22g	P43, Tat, M128
N2	N-P43	P43	
N3a	N-M178	M178	
O1	O-M119	M119	
O3*(xO3c,O3e)	O-M122*	M122	LINE-1, M134
O3e	O-M134	M134	
O3c	O-LINE-1	LINE-1	
Q*(xQ3)	Q-P36*	P36	M3
Q3	Q-M3	M3	
R*	R-M207*	M207	M173
R1a	R-SRY_{10831b}	SRY_{10831b}	
R1b	R-P25	P25	

a. Abbreviated without parenthetical system.

variation were estimated, and their significances were tested using a nonparametric permutation procedure. The relationships among genetic, geographic, and linguistic structure were assessed by the Mantel test also employing ARLEQUIN 2.0 software. Geographic distances were calculated between populations from latitude and longitude data for the sample sites. The matrix of pairwise linguistic distances among populations was constructed according to the method described by Excoffier et al. (1991) and Poloni et al. (1997). Language classifications were adopted primarily from Ruhlen (1991). Populations related within a linguistic family were set to distances from 0 to 5. A distance of 6 was assigned to any pair of populations belonging to different language families. We performed nonmetric multidimensional scaling (MDS) (Kruskal 1964) on Slatkin's linearized Φ_{ST} distances using the software package NTSYS (Rohlf 1998). Multidimensional scaling is an ordination technique that transforms a similarity-dissimilarity matrix into distances in a Euclidean n-dimensional graph. The goodness of fit between the distances in the graphic configuration and the monotonic function of the original distances is measured by a statistic called "stress." Spatial autocorrelation analysis was performed using the autocorrelation index for DNA analysis (AIDA)

(Bertorelle and Barbujani 1995). AIDA is a form of spatial autocorrelation analysis that summarizes genetic variation among individuals as a function of their distance in space. Measures of molecular similarity are estimated within arbitrary intervals, and their departure from null expectations is tested by randomization. Two populations are considered similar if the same haplogroups occur at similar frequencies in the two localities, and when the haplogroups differ by only a small number of substitutions. Much like correlation coefficients, autocorrelation statistics are positive when individuals are genetically similar at a certain distance, are negative when they are dissimilar, and are expected to be zero under the hypothesis of spatial randomness. A haplotype-based analog to Moran's I, symbolized as II, was employed as the measure of spatial effects on haplogroup frequencies (Fix 1999). We used a method developed by Harpending and Ward (1982) to estimate the relative roles of genetic drift and founder effect versus gene flow in causing population differentiation. Under their model, genetic heterozygosity is negatively correlated with genetic distances from the gene frequency centroid (the overall mean gene frequencies of the population system). Those populations that have undergone systematic migrations will show greater heterozygosity than predicted by the regression line, while those groups that are more isolated will exhibit lower-than-predicted heterozygosity. The ANTANA program computed these distances from the centroid (Harpending and Rogers 1984). We used Geodis, version 2.0 (Posada et al. 2000) to conduct Nested Cladistic Analysis (NCA). This method attempts to clarify statistically significant associations between haplogroups and geography in terms of population history and population structure factors. The null hypothesis of NCA is that there is no association between haplogroups and geographic location. When this hypothesis is rejected, a decision key is used to discriminate among the various population structure processes (i.e., recurrent gene flow restricted by isolation by distance versus long-distance dispersal) and/or population history events (i.e., contiguous range expansion, long-distance colonization, or fragmentation). The mean and standard deviation of the time to the most recent common ancestral (TMRCA) Y-chromosome sequence, as well as the ages of each of the mutations in our cladogram, were estimated using the program GENETREE (Karafet et al. 1999; Bahlo and Griffiths 2000).

Results

NRY Haplogroup/Paragroup Distribution in Siberia. Figure 2 presents an evolutionary tree showing the relationships among 56 global haplogroups/paragroups defined by 62 binary markers. This tree reflects the new YCC standardized hierarchical nomenclature system (YCC 2002). The 23 haplogroups/paragroups present in the 18 Siberian populations fall into 12 of the 18 major haplogroup divisions. We refer to each haplogroup/paragroup using the appropriate capital letter followed by a dash and the name of the terminal mutation that defines a given haplogroup (see YCC 2002 for a complete description of this shorthand "muta-

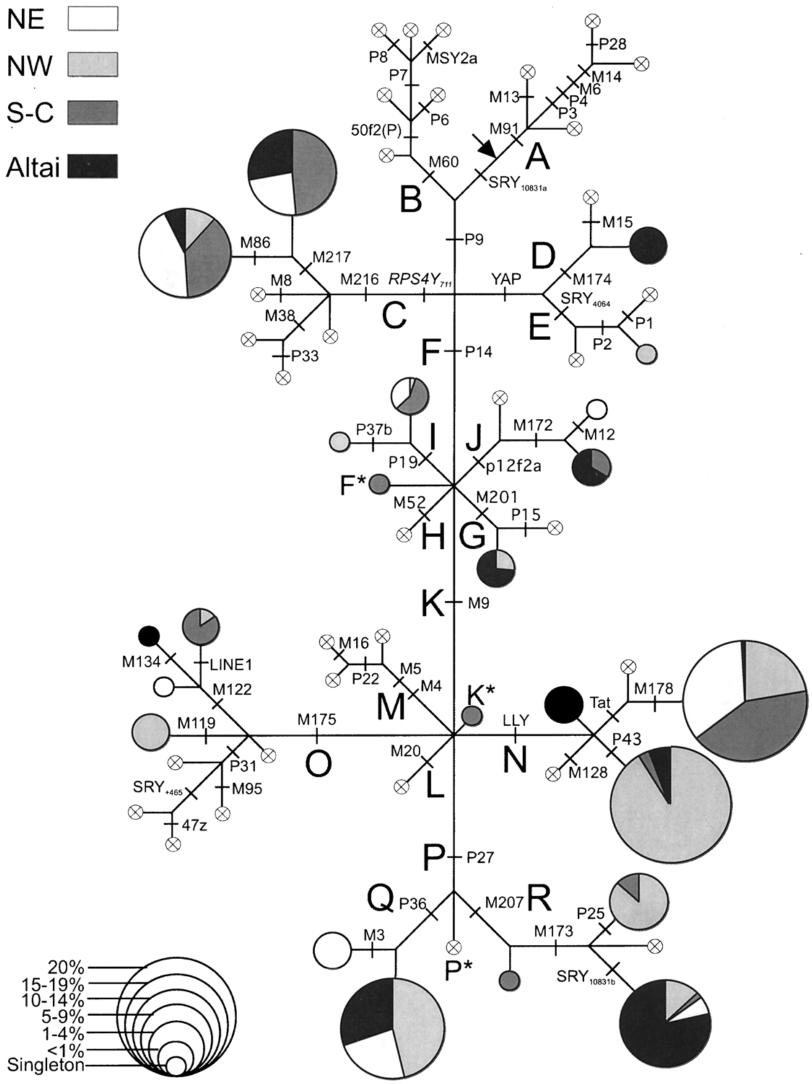

Figure 2. Evolutionary tree for 56 NRY haplogroups/paragroups. The root of the haplogroup tree is denoted by an arrow. The 62 mutational events are shown by cross-hatches. Capital letters *A-R* correspond to 18 major haplogroups (YCC 2002) and are placed on the tree in proximity to the mutation defining the respective major clade. Three paragroups (*F*, K**, and *P**) represent internal nodes on the tree (YCC 2002). Haplogroups/paragroups are coded in black, white, and two shades of gray according to geographic region (see figure for key); open circles with an "X" denote absence in our Siberian sample. The pie charts represent the frequency of occurrence (weighted by regional sample size) of the haplogroups/paragroups within each of the three Siberian geographic regions shown in Figure 1 plus the Altais. The overall size of each pie chart corresponds to one of seven frequency classes (see figure for frequency class key) and represents the frequency of that haplogroup/paragroup in the total sample of 902 Siberian chromosomes.

tion-based" nomenclature). Table 2 lists the formal lineage-based names as well as the mutation-based names for these 23 haplogroups/paragroups.

Only six haplogroups/paragroups were present at frequencies greater than 9% across Siberia, while eight were singletons (haplogroup and paragroup frequencies in all 28 populations surveyed here are available from T. Karafet). The vast majority (96.4%) of Siberian Y chromosomes belong to only four of the 18 major haplogroups (*N* = 42.7%; *C* = 22.5%; *Q* = 18.8%; and *R* = 12.4%). The most frequent single haplogroup/paragroup, *N-M178* (22.7%), was found in 15 of the 18 Siberian populations, reaching its highest frequency (94.3%) in the Yakuts. The widespread distribution of *N-M178* transcends both geographic and linguistic boundaries. Interestingly, this haplogroup is limited almost entirely to northern Eurasia and is absent or only marginally present in other regions of the globe (unpublished data).

The second most frequent haplogroup/paragroup, *N-P43* (19.7%), defined by a newly discovered polymorphism, was prevalent in the Northwest Siberia region (32.6%), infrequent in Central-South Siberia, and absent in Northeast Siberia. Among the seven Uralic-speaking populations in Northwest Siberia, the frequency of *N-P43* was a relatively high 40.6%; moreover, 91.6% of the Siberians with *N-P43* were Uralic-speakers. Only the Selkups, where the *N-P43* frequency was much lower (6.9%), did not fit this pattern.

A second newly discovered polymorphism (P36) defines one of the 18 major haplogroups, *Q*. Haplogroup *Q* chromosomes were present in 18.8% of the Siberian samples and were distributed primarily across Northwest and Northeast Siberia. The vast majority of haplogroup *Q* chromosomes (79.5%) occurred in only two Siberian populations, the Kets and the Selkups, with frequencies of 93.8% and 66.4%, respectively.

Haplogroup *C-M86* (with a frequency of 13.0%) was widely distributed throughout Siberia and Central Asia in the 18 Siberian populations and was concentrated mainly in Altaic-speaking populations. Haplogroup $R\text{-}SRY_{10831b}$ (with a Siberian frequency of 10.3%) was primarily a Central Asian lineage (Karafet et al. 1999; Wells et al. 2001) that was found at relatively high frequencies in the Altai (46.9%), Kirghiz (61.5%), and European Russians (42.6%), at moderate frequencies in some Northwest Siberian populations, and at low frequencies throughout Central-South Siberia, Northeast Siberia, and East Asia. Although absent in Northwest Siberian populations, the Siberian frequency of paragroup *C-M217** was 9.5%. Of the 117 *C-M217** chromosomes in the 28 populations in this survey, all but nine occurred in Altaic-speaking groups.

With the exception of *R-P25* (with a frequency of 2.1%), all of the remaining lineages were present at frequencies of less than 1% in Siberia. It is interesting to note that no single polymorphism was unique to Siberia, and unlike the case for Native American populations (Karafet et al. 1999), Siberian populations were not characterized by a discrete set of founder NRY haplogroups/paragroups.

Y-Chromosome Diversity. Y-chromosome diversity values for each of the 28 populations are given in Table 1. Nei's (1987) diversity statistic (*h*), which is based on the frequency and number of haplogroups/paragroups, ranged from 0.11 in the Yakuts to 0.92 in the Uzbeks. The low diversity in the Yakuts is consistent with the results of Pakendorf et al. (2002), which inferred a low effective size for male Yakuts. The mean number of pairwise differences among haplogroups/paragroups (p) ranged from 0.96 in the Kets to 5.93 in the Dolgans. Both diversity statistics exhibited a similar pattern of high values in East and Central Asia, whereas Northwest Siberia had by far the lowest values for these two statistics. When Siberia as a whole is contrasted with other major regions of the globe, its h value (0.89) was higher than those in South Asia, Europe, or the Americas, and its p value (5.46) was higher than that of any non-African region except Central Asia (Hammer et al. 2001; data not shown). However, when considering only the 18 individual Siberian populations, the mean h value was 0.55 and the mean p value was 3.13. These relatively low average diversity values may be explained by the fact that most Siberian populations possess only one or two predominant haplogroups/paragroups, and these lineages are often one- or two-step mutation neighbors.

Nonmetric Multidimensional Scaling (MDS). Results of MDS based on Φ_{ST} genetic distances are shown in Figure 3 (for which the stress value was 0.17). The

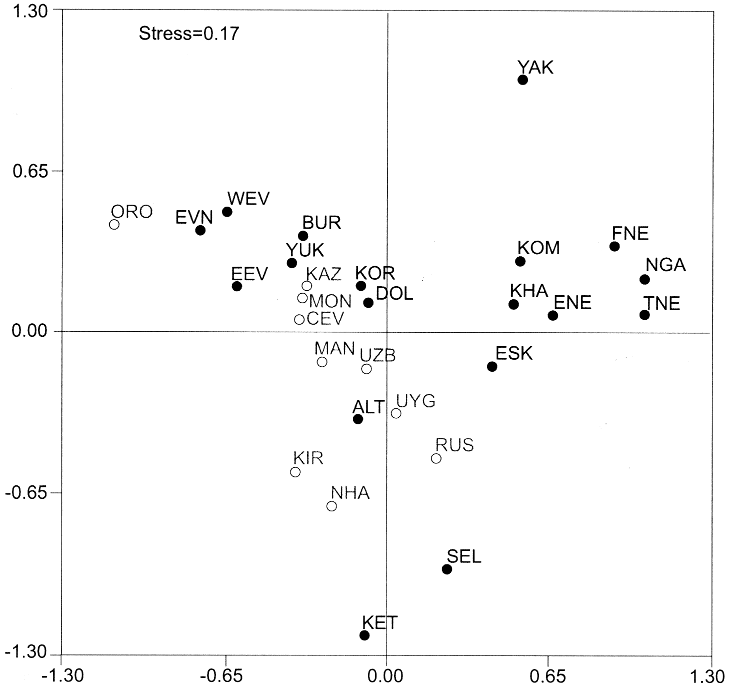

Figure 3. MDS plot of 28 populations based on Φ_{ST} genetic distances. For three-letter population codes, see Table 1. Siberian populations are shown as solid black circles, while other populations are shown as open circles.

18 Siberian populations (black circles) formed three clusters, two of which corresponded with linguistic affiliation. For example, five of the seven Altaic-speaking Siberian populations formed a loose cluster on the upper left side of Figure 3, while six of the eight Uralic-speaking Siberian populations fell within a cluster on the upper right side. Two of the four clear exceptions to this pattern were the position of the Uralic-speaking Yukaghirs within the Altaic cluster, and the presence of the Chukchi-Kamchatkan-speaking Koryaks in this same cluster. The small third cluster was comprised of the Uralic-speaking Selkups and the linguistic isolate, the Yeniseian-speaking Kets. Finally, the Altaic-speaking Yakuts were an extreme outlier, while the Eskimos occupied a position close to the Uralic cluster. When considering all 28 populations it is evident that the Altaic-speaking Central Asian, Chinese, and Mongolian populations fall within, and actually extend, the Altaic cluster.

Analysis of Molecular Variance (AMOVA). The Φ_{ST} value for the 18 Siberian populations was 0.41 (Table 3), indicating a significant degree of population differentiation within Siberia. When all 28 populations were included, the Φ_{ST} value dropped to 0.35. Interestingly, these two values bracketed our global Φ_{ST} value of 0.36 based on 43 binary polymorphisms typed in 2858 individuals from 50 populations (Hammer et al. 2001). Among-population differentiation in Siberia is even higher than that recently reported for a set of 22 sub-Saharan, northern, and eastern African populations ($\Phi_{ST} = 0.34$), a figure that was previously the highest NRY-based Φ_{ST} value for any continent (Cruciani et al. 2002). When the 18 Siberian populations were divided into four geographic groupings (Northwest Siberia, Northeast Siberia, Central-South Siberia, and the Altai from Southwest Siberia), the Φ_{ST} value of 0.41 was the same as it was without geographic subdivision. On the other hand, when these 18 populations were divided into five linguistic families, the Φ_{ST} value rose to 0.45. Similarly, the geographically based Φ_{CT} value was –0.01 ($p = 0.55$), whereas the language-based Φ_{CT} was 0.16 ($p = 0.01$). Indeed, the only Φ-statistic in Table 3 that was not significant was the geographically based $\mathit{\Phi}_{CT}$. Thus, between-group variation was much more striking when Siberian populations were grouped by language than by geography for both Φ_{ST} and Φ_{CT}. The Φ_{SC} values show the reverse pattern where higher values were observed by geography than by language (geographic $\Phi_{SC} = 0.42$; language $\Phi_{SC} = 0.34$).

Mantel Test. Correlation and partial correlation coefficients between genetic, geographic, and linguistic distances are presented in Table 4. Genetic and geographic distances among Siberian populations did not reveal any significant correlation ($r = -0.060$, $p = 0.695$). The partial correlation of genetics and geography with language held constant was also nonsignificant ($r = -0.157$, $p = 0.922$). In contrast, genetics and language were significantly correlated ($r = 0.256$, $p = 0.002$). Moreover, the partial correlation between genetics and language with geography held constant demonstrated an even stronger relationship ($r = 0.292$, $p <$

Table 3. Analysis of Molecular Variance (AMOVA)

				Within Populations		*Among-Populations Within Groups*		*Among Groups*	
Group	*N*	*No. of Populations*	*No. of Groups*	*Variance (%)*	Φ_{ST}	*Variance (%)*	Φ_{CT}	*Variance (%)*	$\Phi_{CT}{}^{a}$
All populations	1432	28	1	65.0	0.35				
Siberia	902	18	1	58.8	0.41				
Geographic groups[b]	902	18	4	59.0	0.41	41.8	0.42	–0.9	–0.01
Linguistic groups[c]	902	18	5	55.3	0.45	28.2	0.34	16.5	0.16

a. All Φ-statistics *p* values are less than 0.01, except for the Siberian geographic groups Φ_{CT}, for which $p = 0.547$.

b. Southwest (ALT), Northwest (KOM, TNE, FNE, KHA, SEL, DOL, NGA, ENE, KET, WEV), Central-South (EEV, YAK, BUR), and Northeast (EVN, ESK, YUK, KOR).

c. Altaic (DOL, ALT, YAK, BUR, EEV, WEV, EVN), Uralic (TNE, FNE, KHA, ENE, NGA, SEL, KOM, YUK), Yeniseian (KET), Eskimo-Aleut (ESK), Chukchi-Kamchatkan (KOR)

Table 4. Correlation and Partial Correlation Coefficients between Siberian Genetic, Geographic, and Linguistic Distance[a]

Distance Comparison	*Correlation Coefficient*	*p*
Genetics and geography	–0.060	0.695
Genetics and language	0.256	0.002
Geography and language	0.327	<0.001
Genetics and geography, language held constant	–0.157	0.922
Genetics and language, geography held constant	0.292	<0.001

a. Distances between distinct language families = 6.

0.001). Thus, 8% of the variance in the genetic data was explained by language, while only 0.1% was determined by geography. Following Poloni et al. (1997), we reexamined the correlations after modifying the distance values between distinct language families. We found that genetics-language correlations changed only slightly when larger (i.e., eight versus six) distance values were used (r = 0.258 versus 0.256, 0.294 versus 0.292, for correlations and partial correlations, respectively). To test whether our results may have been biased by the small number of Siberian populations, we repeated the Mantel test including 10 additional populations. The correlation between genetic and geographic distances was still nonsignificant (r = –0.013 p = 0.544), while the correlation between genetics and language remained significant (r = 0.238, p = 0.000) (data not shown).

Spatial Structure of Y Chromosomal Variation. The Mantel test correlation analysis between genetic and geographic distances did not reveal any significant association for Y-chromosome structure in Siberia; nevertheless, spatial autocorrelation analysis (AIDA) rejected the null hypothesis of random geographic distribution of NRY haplogroup frequencies in space. Two correlograms were derived from different data sets (Figure 4a and Figure 4b). Very high positive autocorrelation at distance 0 was observed for the 18 Siberian populations (Figure 4a), indicating that individuals within the same populations in Siberia resembled each other. The II coefficient at distance 0 was nearly two times higher than that found in East Asia (Karafet et al. 2001). This result is consistent with the small number of haplogroups and low diversity indices in individual Siberian populations. In both data sets (Figures 4a and 4b), the II values were positive for nonzero distances <2500 km and negative for distances >2900 km. Increasing autocorrelation around 6300 km is probably due to substantial paternal genetic similarities among Eastern Evenks, Western Evenks, and Evens, despite the fact that these populations are presently separated by long distances. The pattern for the 28-population analysis (Figure 4b) was clearly clinal and exhibited a decrease of autocorrelation indices from significantly positive to significantly negative as geographic distance increased (Barbujani et al. 1994). Although the correlogram constructed for the 18 Siberian populations (Figure 4a) did not show a monotonic

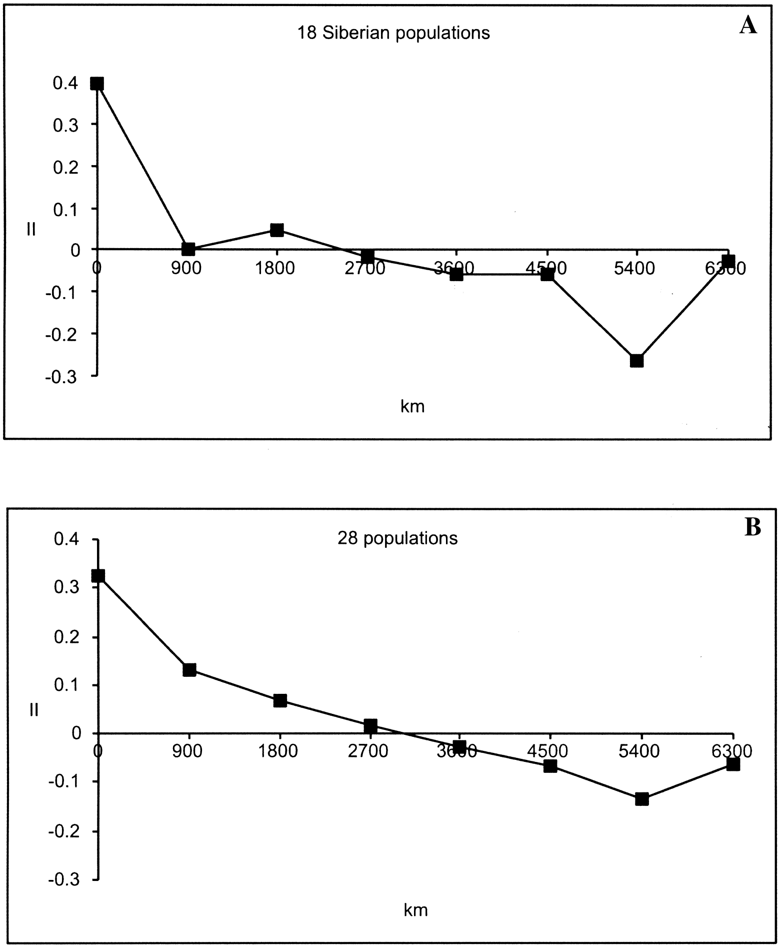

Figure 4. Spatial autocorrelation plots. **A**. 18 Siberian populations. **B**. 28 populations.

decrease, the pattern was still clinal, albeit disturbed by a "depression" at the 5400-km distance class and by an upward fluctuation at the 1800-km distance class. This kind of pattern has been characterized as "long distance differentiation" (Sokal et al. 1989; Barbujani et al. 1994).

Heterozygosity versus Distance from the Centroid. Figure 5 represents a plot of haplogroup diversity regressed against distance from the centroid (r_{ii}). All Siberian populations except the Koryaks exhibited lower than predicted heterozygosity, suggesting genetic differentiation due primarily to geographic isolation and genetic drift. On the other hand, Central Asian populations, the Manchu, and the Evenks from China were characterized by higher than theoretically expected

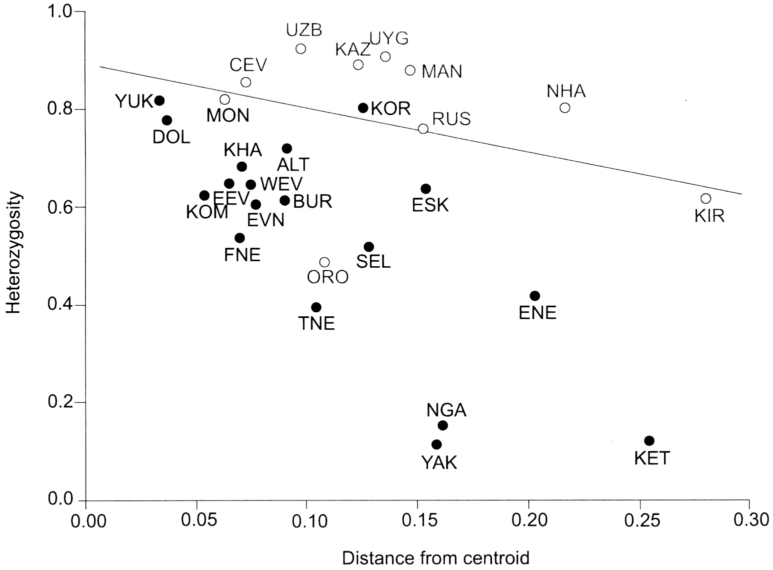

Figure 5. Regression of genetic heterozygosity (*h*) on distance from centroid (r_{ii}). Siberian populations are shown as solid black circles, while other populations are shown as open circles. The solid line represents the theoretical regression line as described in Harpending and Ward (1982).

heterozygosity combined with low or moderate r_{ii} values and, therefore, have probably experienced substantial gene flow.

Nested Cladistic Analysis (NCA). The nesting design produced 18 one-step clades, 6 two-step clades, 2 three-step clades, and a clade representing the total cladogram (data not shown). Highly significant associations between clades and geographic location were found for 14 of the 27 clades. With the help of the key published on the Geodis 2.0 website (http://bioag.BYU.edu/zoology/crandall_lab/geodis.htm) (Posada et al. 2000), we were able to infer the presumable causes of these 14 patterns (Table 5). Similar to our previous publications (Hammer et al. 1998; Karafet et al. 1999; Hammer et al. 2001; Karafet et al. 2001), both population structure and population history factors influenced the NRY haplogroup distribution. Interestingly, unlike any of our previous nested cladistic analyses, most of these signals (10 out of 14) involved historical events operating at the population level (i.e., contiguous range expansions [*n* = 4] and long-distance colonizations [*n* = 6]). Only four phylogeographic associations were the result of population structure processes such as gene flow restricted by isolation by distance.

Discussion

Siberian Hunter-Gatherer Population Structure. Human populations exhibit a wide range of population densities and mobilities. For example, hunter-

Table 5. Inferences from Nested Cladistic Analysis

Clade	*Inference Chain*
1-1	1→2→3→4→No→RGF/IBD
1-7	1→2→11→RE→12→CRE
1-9	1→2→3→4→No→RGF/IBD
1-13	1→2→11→17→4→No→RGF/IBD
1-15	1→2→3→5→6→13→LDC
1-16	1→2→11→RE→12→CRE
1-17	1→2→3→4→No→RGF/IBD
2-3	1→2→11→RE→12→CRE
2-4	1→2→3→5→6→13→LDC
2-5	1→2→11→RE→12→CRE
2-6	1→2→3→5→6→13→LDC
3-1	1→2→11→RE→12→13→LDC
3-2	1→2→11→RE→12→13→LDC
Total	1→2→11→RE→12→13→LDC

Note: RGF/IBD = gene flow restricted by isolation by distance, CRE = contiguous range expansion, LDC = long distance colonization.

gatherers from Australia and Central Africa have densities ranging from <0.1 to <1.0 persons per square kilometer, respectively, while some Asian wet rice farmers have densities of nearly 800 persons per square kilometer (Fix 1999). It has often been assumed that the discovery of common patterns of genetic diversity in groups with low densities will reveal aspects of population structure that have characterized humans throughout prehistory from the Pleistocene to the present (Birdsell 1973; Cavalli-Sforza 1986). On the other hand, Fix (1999) has cautioned against generalizing about all of human evolution from what is known about historic and contemporary hunter-gatherers, especially given their large range of population densities and mobilities. Still, it is useful to classify human populations in terms of these and other characteristics (e.g., degree of endogamy and contact with nonforagers, intensity of land use, size of social group, clan/dialect group structure, marriage practices, etc.), to provide a framework for discovering similarities and differences among different categories of foraging and nonforaging groups (Fix 1999). It might be expected that foraging groups with a particular level of sociocultural integration share certain features of population structure, while those exhibiting differing levels may possess different patterns of technology, subsistence economics, and/or social organization. Indeed, there is a dearth of comparative genetic data in sets of related populations practicing a hunter-gatherer lifestyle. This study represents one of the first investigations of the paternal genetic structure of contemporary hunter-gatherer populations from a single major ecological region of the world.

It has been claimed based on classical marker data that high values of F_{ST} are found primarily among tropical agriculturists of restricted mobility, with substantially lower values characterizing hunter-gatherers of greater mobility (Harpending and Jenkins 1974; Jorde 1980). Our results, however, show that Siberian hunter-gatherer groups (i.e., all Siberians in our sample except the Altai, Buryats, and Yakuts, who are sedentary) exhibit a considerable amount of NRY differentiation ($\Phi_{ST} = 0.44$). When comparing these results with similar analyses of global populations, it can be seen that this degree of differentiation is higher than for any region of the world (Hammer et al. 2001; Cruciani et al. 2002), and even higher than the worldwide Φ_{ST} value of 0.36 (Hammer et al. 2001). Another unusual pattern of genetic structure in Siberia is the high level of among-populations within-groups variance relative to the among-groups variance (see below). These patterns may be explainable, in part, by the low population densities of Siberian native groups. Just after the 17th-century Russian invasion, approximately 227,000 indigenous people of Siberia (Forsyth 1991) were distributed over territory exceeding 13 million square kilometers. This average density of 0.017 individuals per square km is one of the lowest estimates for any hunter-gatherer population (Hassan 1981; Fix 1999). Furthermore, the graph of heterozygosity versus distance from the centroid (Figure 5) revealed a deficiency of heterozygosity among Siberian populations compared with Central Asian populations. This pattern suggests smaller effective population size and/or less gene flow among populations in Siberia. In summation, we infer that intra- and intergenerational genetic drift (resulting from high population mobility and small population size, respectively), were key evolutionary forces leading to the high levels of genetic differentiation observed among Siberian foraging groups.

An assortment of cultural and demographic factors may also have influenced the paternal genetic structure of Siberian populations. For NRY polymorphisms this differentiation is expected to be strongly elevated by cultural factors such as clan systems, polygamy, and the levirate. Siberian local residence groups were organized on the basis of kinship and clan affiliation. Each local camp or band (the level of organization most important for hunter-gatherer communities) is a group of people who might be related matrilineally, patrilineally, or bilaterally. Because most Siberian populations practiced patrilocality, people in a band were usually paternally related. Every band, therefore, could potentially be responsible for a strong founder effect detectable in Y-chromosome data.

If genetic drift and short-range dispersals were the key factors shaping spatial variation of NRY structure in Siberia, isolation by distance might be expected. Spatial autocorrelation analysis of Siberian populations revealed strong patterning of genetic variation compatible with a clinal distribution (Figure 4). One of the explanations for this type of observed pattern is a series of founder effects taking place in a phase of population expansion not accompanied by admixture, but followed by local gene flow (Barbujani et al. 1994). Loss of genetic variation through repeated founder effects has been invoked as the likely cause of clines in several studies of natural populations (Fix 1999). When the clines extend over long distances and originate from fairly large initial gene frequency differences, they will be remarkably stable over time, as has been shown by Wijsman and Cavalli-Sforza (1984). In addition, four of the statistically significant NCA signals were due to a population structure process: specifically, gene flow restricted by isolation by distance (Table 5). The main thrust of the NCA results, however, requires that population history events must be considered. Six long-distance colonizations and four contiguous-range expansions were detected. Therefore, based on these NCA signals, as well as on the results from diversity statistics, AMOVA, and spatial autocorrelation patterns, there is an indication that both founder effect/genetic drift and long-range population movements influenced the genetic structure of Siberian indigenous groups.

Associations among Geography, Language, and NRY Variation. The majority of Siberian peoples speak languages from the Altaic and Uralic linguistic families, although populations in Northeast Siberia speak languages in the Chukchi-Kamchatkan family, Siberian Eskimos speak languages belonging to the Eskimo-Aleut family, and the Kets speak the only surviving member of the Yeniseian family (Ruhlen 1991, 1998; Greenberg 2000). A common language usually signifies a common origin for two populations and a related language indicates a common origin farther back in time (Ruhlen 1991). It is important to note that language differences are themselves barriers to free gene flow (Barbujani 1991), thereby reinforcing genetic differentiation. Because both linguistic and genetic patterns result from the biological and social interactions of individuals, genetic and linguistic differentiation should demonstrate considerable similarity if they occurred synchronously and at comparable rates (Chen et al. 1995). Despite several factors complicating language inheritance (i.e., horizontal transmission, cultural assimilation, and elite dominance), autosomal-based studies pointed to the conclusion that genetics and language are interrelated in global populations (Cavalli-Sforza et al. 1988; Barbujani and Sokal 1990; Excoffier et al. 1991; Sokal et al. 1992). Subsequent mtDNA studies, however, did not demonstrate clear correlations between genetic and language structures (Torroni et al. 1992; Ward et al. 1993; Watson et al. 1996; Bonatto and Salzano 1997).

When AMOVA was performed on Siberian populations combined in accordance with geographic criteria (Northwest Siberia, Northeast Siberia, Central-South Siberia, and the Altai from Southwest Siberia), a hierarchical analysis of variance revealed that Φ_{CT}, the parameter that estimates among-group differentiation, was considerably lower than Φ_{SC} (an estimator of variation among populations within a group) (Table 3). Unlike Siberia, global data showed the opposite pattern: groups were more different among themselves than were populations within these groups (Poloni et al. 1997; Hammer et al. 2001; Cruciani et al. 2002). When Siberian populations were divided by language family rather than by geography, the Φ_{ST} value rose to 0.45, the Φ_{CT} changed from a statistically nonsignificant value of –0.01 to a statistically significant value of 0.16, and the Φ_{SC} value declined from 0.42 to 0.34. Taken together these results indicate that language affiliation might be a better predictor of the genetic affinity among Siberians than their present geographic position. Mantel tests also support the existence of NRY genetic patterns that are correlated with language in Siberia. The language-genetics correlation ($r = 0.256$) was statistically significant. Interestingly, the partial correlation between genetics and language, holding geography constant actually increased in value to $r = 0.292$. Thus, approximately 8% of the paternal Siberian genetic variation is explained by linguistic affiliation. Moreover, when we excluded the Selkup and Yakut populations (clear outliers for their respective linguistic families in the MDS plot), the correlation between genetics and language increased sharply to $r = 0.380$ (data not shown).

Two recent NRY studies (Poloni et al. 1997; Rosser et al. 2000) employed Mantel procedures to test the significance of the correlation between linguistic and paternal genetic systems. Poloni et al. (1997) showed a statistically significant association of Y chromosomal population structure in Europe and Africa with language family. For example, the partial correlation between genetics and language, holding geography constant, was $r = 0.323$ ($p < 0.001$). On the other hand, European populations in the Rosser et al. (2000) survey exhibited a low, nonsignificant partial correlation between genetics and language, holding geography constant ($r = 0.088$). In both studies genetics and geography were strongly and significantly correlated, even when language was held constant. Rosser et al. (2000) explained their contrasting results by suggesting that Poloni et al. (1997) employed geographically and linguistically more globally distributed samples, while the European samples of Rosser et al. (2000) were all located within a single continent, and mostly spoke Indo-European languages.

In Siberia we found a statistically significant correlation of language with both paternal genetics and geography, but a notable absence of correlation between NRY genetic structure and geography (Table 4). Here, the patterns of observed paternal genetic variation cannot be explained by a simple isolation by distance model with short-range gene flow. One reason for a weak association between geography and NRY variation in Siberia may be the occurrence of extensive genetic drift. However, it may also be the case that highly mobile, endogamous populations will not show associations between genetic variation and geography, a fact that was demonstrated for Jewish populations that had recently radiated out of the Middle East (Hammer et al. 2000). Directed dispersals, range expansions, and long-distance colonizations bound by common ethnic and linguistic affiliation have most probably been of utmost importance in fashioning the genetic landscape of Siberia.

Paleolithic Colonization of Siberia: Insights into the Initial Peopling of Siberia. During the Late Pleistocene (and more specifically, the early Upper Paleolithic), most of Siberia was free of continental ice sheets, and mountain glaciation was quite limited. Even during periods of maximum cold, the vegetation in southern Siberia was tundra and forest tundra with light larch forests. There were no natural obstacles such as continental or large mountain glaciers to prevent human migrations toward and within Siberia (Kuzmin and Orlova 1998). The earliest dated North Asian Upper Paleolithic industries occur in the Altai Mountains in southwest Siberia (43,300 ± 1600 years BP). Paleolithic industries originally developed in the Altai region subsequently (i.e., from 34,000 BP to 21,000 BP) colonized southern Siberia including the Sayan Mountains, the Angara River basin, the Trans-Baikal, and Mongolia (Derev'anko 1998b; Goebel 1999). Early Upper Paleolithic stone tool industries were centered on the production of macroblades similar to points found in initial Upper Paleolithic industries in western Asia and eastern Europe (Kuzmin and Orlova 1998; Goebel 1999), suggesting continued ties between Siberia and western Eurasia during that time.

Later Siberian Paleolithic sites (i.e., postdating 20 ky ago) tend to share an abundance of microblades and wedge-shaped cores. The individual sites are distributed throughout Siberia and the Russian Far East with unequivocally dated microblade industries in the Yenisei River basin already present by 23 ky ago (Kuzmin and Orlova 1998). Late Upper Paleolithic people seem to have formed small groups of highly mobile hunter-gatherers. There is clear evidence of transport of material over great distances. Goebel (1999) has suggested rapid recolonization and possible replacement of early Siberian Upper Paleolithic people by microblade-making human populations from the Lake Baikal, Yenisei River, and Lena River basin regions. The origin of the Siberian microblade industry is unclear. This assemblage differs in detail from its west Asian and European counterparts (Klein 1999). A great number of sites in Mongolia, North China, Japan, and Korea contain evidence for this core type. Many scholars find filial connections among these industries (Derev'anko 1998b). Whether they represent evolution of microblade technologies out of local ancestors or trace migrations from farther south and east cannot be determined conclusively with the available archeological evidence (Goebel 1999).

We estimated the ages of the major Siberian NRY haplogroups to investigate the genetic history of Siberian populations. The most frequent lineages in Siberia belong to four major haplogroups: *C* (*C-M217** and *C-M86*), *N* (*N-M178* and *N-P43*), *Q*, and *R* (*R-SRY$_{10831b}$*) (Figure 2). Although globally distributed, the *N-P43* and *N-M178* haplogroups were found at their highest frequencies in Siberia. The ages of these mutations were estimated as 3500±300 and 2180±105 years old, respectively. The LLY22g mutation which defines haplogroup *N* may be as old as 6910±1480 years, suggesting that the expansion of the *N-P43* and *N-M178* haplogroups probably occurred much later than the first migrations of anatomically modern human into Siberia. Haplogroups *N-P43* and *N-M178* may have entered Siberia from Mongolia and North China (Zerjal et al. 1997) and later spread west, and then northeastward within Siberia.

The ages of haplogroups *C* and *P* (the haplogroup that contains *Q* and *R*) were estimated to be 27,500 ± 10,100 and 29,900 ± 4200 years old, respectively. This estimate of the age of haplogroup *C* agrees with that of Bergen et al. (1999) (27,000–33,000 years) which was based on the variance in repeat numbers at nine Y-chromosome STRs (Y-STRs). The age of the M45 marker that also defines haplogroup *P* was estimated by Wells et al. (2001) as 40,000 years old based on only six Y-STRs. When we use the same approach as Wells et al. (2001) with our data from 11 Y-STRs, we estimate an age for haplogroup *P* of 30,000–37,000 years, depending on whether we assume 20 or 25 years per generation (data not shown). Thus, the age estimates of haplogroups *C* and *P* are consistent with the age of the Siberian Upper Paleolithic, albeit somewhat younger than the oldest Paleolithic sites. We must caution, however, that these age estimates depend on sampling and on other parameters in the model that are difficult to measure such as effective population size and mutation rate.

Archeological evidence suggests that the Altai Mountains were the first habitat of anatomically modern humans in Siberia. Both haplogroups *C* and *P* are found in the Altai. Y-STR analyses indicate that haplogroup *P* is about three times more diverse (considering the variance in STR repeat numbers) than haplogroup *C* in the Altai (0.757 versus 0.280, respectively) (data not shown). Therefore, haplogroup *P* might represent the oldest lineage in this area. The candidate source populations for haplogroup *P* most likely include Central Asian populations—the most diverse in Eurasia (Hammer et al. 2001; Wells 2001). Our conclusion is consistent with the inference of Wells et al. (2001) that early settlement of Central Asia 40,000–50,000 years ago was followed by subsequent migrations into Europe, India, and Siberia. This finding also supports archeological evidence for a Central Asian source of the first colonization of anatomically modern humans in Siberia. We hypothesize that the first Siberians, with a macroblade industry and carrying NRY haplogroup *P*, settled in the Altai region and subsequently moved to the east.

Two descendant lineages of haplogroup *P*, *R-SRY*$_{10831b}$ and *Q-P36*, were also detected in the Altai. The estimated age of *R-SRY*$_{10831b}$ (roughly 4000 years) is well after early human dispersals into Siberia. It has been suggested that *R-SRY*$_{10831b}$ likely traces a population migration originating somewhere in southern Russia and the Ukraine, perhaps stemming from the Kurgan culture (Zerjal et al. 1999; Rosser et al. 2000; Semino et al. 2000; Wells et al. 2001). The presence of *R-SRY*$_{10831b}$ in western Siberia probably chronicles known migrations originating in the Altai and Sayan Mountains. The low frequency of this haplogroup in several Central and East Siberian populations is most likely due to admixture with recent migrants of European descent.

Haplogroup *Q*, with an estimated age of 17,700 ± 4800 years, was found at moderate frequencies in our Altai sample, as well as in remote regions of Northeast Siberia (i.e., among Eskimos, Yukagirs, and Koryaks). The extremely high frequency of haplogroup *Q* in the Selkups and Kets may be due to intergenerational genetic drift coupled with founder effects. This is supported by very low levels of Y-STR diversity associated with haplogroup *Q* in both populations (0.149 and 0.159, respectively). This haplogroup is present at low frequencies in other Northwest Siberian populations and is absent in Central Siberia.

The highest Y-STR diversity associated with haplogroup *C* chromosomes was found in East Asia (including Mongolia), followed by Siberia and Central Asia (0.954, 0.940, and 0.461, respectively). Two haplogroup *C* members were found in Siberia at moderate frequencies: *C-M217** dated at 11,900 ± 4800 years and its relatively recent descendant *C-M86*. Our time estimate for the M86 mutation is 2,750 ± 1370 years. Mongolia and/or the Lake Baikal region might represent the source of this rather recently derived haplogroup in Siberia.

The combined archaeological and NRY data lead to the following scenario for the early peopling of Siberia. The first migration(s) of anatomically modern humans to the Altai Mountains from Central Asia brought haplogroup *P* Y chromosomes, and these people later dispersed throughout the southern part of Siberia including the Sayan Mountains, the Angara River basin, the Trans-Baikal, and Mongolia. They also produced the early Siberian Upper Paleolithic stone tool industries that were centered on macroblade technology. Eventually they became the first colonists of the Americas. Another migration from Mongolia and/or North China to the Baikal region may have been associated with carriers of haplogroup *C*. These mobile hunter-gatherers with a microblade industry initially colonized southern Siberia, and later the subarctic and arctic zones of North America, perhaps arriving there after the last Glacial Maximum and thus representing a second dispersal to the New World.

In sum, the level of among-population variation in NRY diversity for contemporary Siberian populations outpaces that for any other region of the world. This underscores the fact that foraging populations adapted to boreal climates in the northernmost regions inhabited by humans are genetically subdivided, and that genetic drift has played a key role in shaping patterns of variation in Siberia. These results also emphasize the large-scale coherence of family-level language affiliation and the role of long-distance range expansions in Siberia.

Acknowledgments We thank Svetlana Reznikova and Ji Park for excellent technical assistance, Alan Redd for helpful comments, and Tatiana Zerjal and Chris Tyler-Smith for providing LLY22g typing information. We also thank Richard Diebold for his continued interest in this research. This work was supported by the National Science Foundation through grant OPP-9806759 and by the National Institute of General Medical Sciences through grant GM53566-06 (to MFH), and grants RFBR98-04-49626a and INTAS93-0035 (to LPO).

Received 1 July 2002; revision received 5 September 2002.

Literature Cited

Bahlo, M., and R.C. Griffiths. 2000. Inference from gene trees in a subdivided population. *Theor. Popul. Biol.* 57:79–95.

Barbujani, G. 1991. What do languages tell us about human microevolution? *Trends. Ecol. Evol.* 6:151–155.

Barbujani, G., A. Pilastro, S. De Domenico et al. 1994. Genetic variation in North Africa and Eurasia: Neolithic demic diffusion vs. Paleolithic colonization. *Am. J. Phys. Anthropol.* 95:137–154.

Barbujani, G., and R.R. Sokal. 1990. Zones of sharp genetic change in Europe are also linguistic boundaries. *Proc. Natl. Acad. Sci. USA* 94:4516–4519.

Bergen, A.W., C.Y. Wang, J. Tsai et al. 1999. An Asian-Native American paternal lineage identified by RPS4Y resequencing and by microsatellite haplotyping. *Ann. Hum. Genet.* 63:63–80.

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

Birdsell, J.B. 1973. A basic demographic unit. *Curr. Anthropol.* 14:337–356.

Bonatto, S.L., and F.M. Salzano. 1997. A single and early migration for the peopling of the Americas supported by mitochondrial sequence data. *Proc. Natl. Acad. Sci. USA* 94:1866–1871.

Cavalli-Sforza, L.L. 1986. *African Pygmies*. Orlando, FL: Academic Press.

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

Cavalli-Sforza, L.L., A. Piazza, P. Menozzi et al. 1988. Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. *Proc. Natl. Acad. Sci. USA* 85:6002–6006.

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

Crawford, M.H., and V.B. Enciso. 1981. Population structure of circumpolar groups of Siberia, Alaska, Canada, and Greenland. In *Current Developments in Anthropological Genetics,* Vol. 2*,* M.H. Crawford and J.H. Mielke, eds. New York, NY: Plenum Press, 51–91.

Cruciani, F., P. Santolamazza, P. Shen et al. 2002. A back migration from Asia to Sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am. J. Hum. Genet.* 70:1197–1214.

Derenko, M.V., G.A. Denisova, B.A. Malyarchuk et al. 2000a. Mitochondrial DNA variablility in Turkic-speaking populations of the Altai and Sayan region from South Siberia. *Am. J. Hum. Genet. Suppl.* 67:A1161.

Derenko, M.V., T. Grzybowski, B.A. Malyarchuk et al. 2001. The presence of mitochondrial haplogroup X in Altaians from South Siberia. *Am. J. Hum. Genet.* 69:237–241.

Derenko, M.V., B.A. Malyarchuk, I.K. Dambueva et al. 2000b. Mitochondrial DNA variation in two South Siberian Aboriginal populations: Implications for the genetic history of North Asia. *Hum. Biol*. 72:945–973.

Derev'anko, A.P. 1998a. A short history of discoveries and the development of ideas in the Paleolithic of Siberia. In *The Paleolithic of Siberia: New Discoveries and Interpretations,* A.P. Derev'anko, ed. Urbana, IL: University of Illinois Press, 5–12.

Derev'anko, A.P. 1998b. Human occupation of nearby regions and the role of population movements in the Paleolithic of Siberia. In *The Paleolithic of Siberia: New Discoveries and Interpretations,* A.P. Derev'anko, ed. Urbana. IL: University of Illinois Press, 336–351.

Excoffier, L., R.M. Harding, R.R. Sokal et al. 1991. Spatial differentiation of RH and GM haplotype frequencies in Sub-Saharan Africa and its relation to linguistic affinities. *Hum. Biol.* 63:273–307.

Fix, A.G. 1999. *Migration and Colonization in Human Microevolution*. Cambridge, UK: Cambridge University Press.

Forsyth, J. 1991. The Siberian native peoples before and after the Russian conquest. In *The History of Siberia: From Russian Conquest to Revolution,* A. Wood, ed. London, UK: Routledge, 69–91.

Goebel, T. 1999. Pleistocene human colonization of Siberia and peopling of the Americas: An ecological approach. *Evol. Anthropol*. 8:208–227.

Goebel, T., A.P. Derevianko, and V.T. Petrin. 1993. Dating the Middle-to-Upper-Paleolithic transition at Kara-Bom. *Curr. Anthropol.* 34:452–458.

Greenberg, J. 2000. *Indo-European and Its Closest Relatives: The Eurasiatic Family: Grammar*. Stanford, CA: Stanford University Press.

Hammer, M.F., T. Karafet, A. Rasanayagam et al. 1998. Out of Africa and back again: Nested cladistic analysis of human Y chromosome variation. *Mol. Biol. Evol.* 15:427–441.

Hammer, M.F., T.M. Karafet, A.J. Redd et al. 2001. Hierarchical patterns of global human Y-chromosome diversity. *Mol. Biol. Evol*. 18:1189–1203.

Hammer, M.F., A.J. Redd, E.T. Wood et al. 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.

Hammer, M.F., and S.L. Zegura. 1996. The role of the Y chromosome in human evolutionary studies. *Evol. Anthropol*. 5:116–134.

Harpending, H., and T. Jenkins. 1974. !Kung population structure. In *Genetic Distance,* J. Crow and C. Denniston, eds. New York, NY: Plenum, 137–161.

Harpending, H., and A. Rogers. 1984. *ANTANA: A package for multivariate data analysis*. Bosque Farms: H. Harpending and A. Rogers.

Harpending, H.C., and R.H. Ward. 1982. Chemical systematics and human populations. In *Biochemical Aspects of Evolutionary Biology,* M. Nitecki, ed. Chicago, IL: University of Chicago Press, 213–256.

Hassan, F.A. 1981. *Demographic Archaeology*. New York, NY: Academic.

Jorde, L.B. 1980. The genetic structure of subdivided human populations: A review. In *Current Developments in Anthropological Genetics,* J.H. Mielke and M.H. Crawford, eds. New York, NY: Plenum, 135–208.

Karafet, T.M., O.L. Posukh, and L.P. Osipova. 1994. Results and perspectives of human population studies in Siberia. *Siberian Journal of Ecology* 2:105–118.

Karafet, T.M., R.I. Sukernik, L.P. Osipova et al. 1981. Blood groups, serum proteins, and red cell enzymes in the Nganasans Tavghi—Reindeer hunters from the Taimir Peninsula. *Am. J. Phys. Anthropol*. 56:139–145.

Karafet, T.M., L. Xu, R. F. Du et al. 2001. Paternal population history of East Asia: Sources, patterns, and microevolutionary processes. *Am. J. Hum. Genet.* 69:615–628.

Karafet, T.M., S.L. Zegura, O. Posukh et al. 1999. Ancestral Asian sources of new world Y-chromosome founder haplotypes. *Am. J. Hum. Genet*. 64:817–831.

Karafet, T., S.L. Zegura, J. Vuturo-Brady et al. 1997. Y chromosome markers and trans-Bering Strait dispersals. *Am. J. Phys. Anthropol*. 102:301–314.

Ke, Y., B. Su, X. Song et al. 2001. African origin of modern humans in East Asia: A tale of 12,000 Y chromosomes. *Science* 292:1151–1153.

Klein, R.G. 1999. *The Human Career: Human Biological and Cultural Origins.* Chicago, IL: The University of Chicago Press.

Kruskal, J.B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29:1–27.

Kuzmin, Y., and L.A. Orlova. 1998. Radiocarbon chronology of the Siberian Paleolithic. *J. World. Prehist*. 12:1–53.

Lell, J.T., M.D. Brown, T.G. Schurr et al. 1997. Y chromosome polymorphisms in Native American and Siberian populations: Identification of native American Y chromosome haplotypes. *Hum. Genet*. 100:536–543.

Lell, J.T., R.I. Sukernik, Y.B. Starikovskaya et al. 2002. The dual origin and Siberian affinities of Native American Y chromosomes. *Am. J. Hum. Genet*. 70:192–206.

Levin, M.G., and L.P. Potapov. 1964. *The Peoples of Siberia.* Chicago, IL: The University of Chicago Press.

Nei, M. 1987. *Molecular Evolutionary Genetics*. New York, NY: Columbia University Press.

Novaradovsky, A.G., V.A. Spitsyn, R. Duggirala et al. 1993. Population genetics and structure of Buryats from the Lake Baikal region of Siberia. *Hum. Biol.* 65:689–709.

Okladnikov, A.P. 1981. *The Paleolithic of Central Asia*. Novosibirsk, Russia: Nauka.

Okladnikov, A.P. 1983. The Paleolithic of Mongolia in the light of the new studies. In *Late Pleistocene and Early Holocene Connections between Asia and Americas,* R.S. Vasil'evsky, ed. Moscow, Russia: Nauka, 8–21.

Osipova, L.P., O.L. Posukh, E.A. Ivakin et al. 1996. Gene pool of indigenous people of Samburg tundra. *Genetika* 32:830–836.

Pakendorf, B., B. Morar, L.A. Tarskaia et al. 2002. Y-chromosomal evidence for a strong reduction in male population size of Yakuts. *Hum. Genet.* 110:198–200.

Poloni, E.S., O. Semino, G. Passarino et al. 1997. Human genetic affinities for Y-chromosome P49a,f/*Taq*I haplotypes show strong correspondence with linguistics. *Am. J. Hum. Genet*. 61:1015–1035.

Posada, D., K.A. Crandall, and A.R. Templeton. 1999. Geodis. URL= http://bioag.byu.edu/zoology/crandall_lab/programs.htm.

Posukh, O.L., V.P. Wiebe, R.I. Sukernik et al. 1990. Genetic study of the Evens, an ancient human population of Eastern Siberia. *Hum. Biol*. 62:457–465.

Rohlf, F.J. 1998. *NTSYS-pc: Numerical Taxonomy and Multivariate Analysis System*. Setauket, NY: Exeter Software,.

Rosser, Z.H., T. Zerjal, M.E. Hurles 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.

Ruofu, D., and V.F. Yip. 1993. *Ethnic Groups in China.* Beijing, China: Science Press.

Ruhlen, M. 1991. *A Guide to the World's Languages.* Vol. 1: *Classification*. London, UK: Edward Arnold.

Ruhlen, M. 1998. The origin of the Na-Dene. *Proc. Natl. Acad. Sci. USA* 95:13994–13996.

Rychkov, Y.G., and V.A. Sheremet'eva. 1980. The genetics of circumpolar populations of Eurasia related to the problem of human adaptation. In *The Human Biology of Circumpolar Populations,* F. Milan. ed. London, UK: Cambridge University, 37–80.

Santos, F.R., A. Pandya, C. Tyler-Smith et al. 1999. The Central Siberian origin for Native American Y chromosomes. *Am. J. Hum. Genet*. 64:619–628.

Schneider, S., D. Roessli, and L. Excoffier. 2000. *ARLEQUIN ver.2.000: A Software for Population Genetic Analysis*. Geneva, Switzerland: Genetics and Biometry Laboratory, University of Geneva.

Schurr, T.G., R.I. Sukernik, Y.B. Starikovskaya et al. 1999. Mitochondrial DNA variation in Koryaks and Itel'men: Population replacement in the Okhotsk Sea–Bering Sea region during the Neolithic. *Am. J. Phys. Anthropol*. 108:1–39.

Semino, O., G. Passarino, P.J. Oefner et al. 2000. The genetic legacy of Paleolithic *Homo sapiens sapiens* in extant Europeans: A Y chromosome perspective. *Science* 290:1155–1159.

Shields, G.F., A.M. Schmiechen, B.L. Frazier et al. 1993. MtDNA sequences suggest a recent evolutionary divergence for Beringian and northern North American populations. *Am. J. Hum. Genet.* 53:549–562.

Sokal, R.R., R. M. Harding, and N. L. Oden. 1989. Spatial patterns of human gene frequencies in Europe. *Am. J. Phys. Anthropol.* 80:267–294.

Sokal, R.R., N. L. Oden, and B.A. Thomson. 1992. Origins of the Indo-Europeans: Genetic evidence. *Proc. Natl. Acad. Sci. USA*. 89:7669–7673.

Starikovskaya, Y.B., R.I. Sukernik, T. G. Schurr et al. 1998. MtDNA diversity in Chukchi and Siberian Eskimos: Implications for the genetic history of Ancient Beringia and the peopling of the New World. *Am. J. Hum. Genet.* 63:1473–1491.

Stepanov, V.A., and V.P. Puzyrev. 2000. Y-chromosome microsatellite haplotypes demonstrate absence of subdivision and presence of several components in the Tuvinian male gene pool. *Genetika* 36:377–384.

Su, B., C. Xiao, R. Deka et al. 2000. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. *Hum. Genet.* 107:582–590.

Su, B., J. Xiao, P. Underhill et al. 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.

Sukernik, R.I., T.M. Karafet, and L. P. Osipova. 1978. Distribution of blood groups, serum markers and red cell enzymes in two human populations from Northern Siberia. *Hum. Hered*. 28:321–327.

Sukernik, R.I., S.V. Lemza, T.M. Karafet et al. 1981. Reindeer Chukchi and Siberian Eskimos: Studies on blood groups, serum proteins, and red cell enzymes with regard to genetic heterogeneity. *Am. J. Phys. Anthropol.* 55:121–128.

Sukernik, R.I., L.P. Osipova, T.M. Karafet et al. 1986. Genetic and ecological studies of aboriginal inhabitants of North-Eastern Siberia. 2. Polymorphic blood systems, immunoglobulin allotypes, and other genetic markers in Asian Eskimos. Genetic structure of the Bering Sea Eskimos. *Genetika* 22:2369–2380.

Szathmary, E.J.E. 1981. Genetic markers in Siberian and northern North American populations. *Yrbk. Phys. Anthropol.* 24:37–74.

Torroni, A., T.G. Schurr, C-C. Yang et al. 1992. Native American mitochondrial DNA analysis indicates that the Amerind and the Nadene populations were founded by two independent migrations. *Genetics* 130:153–162.

Torroni, A., R.I. Sukernik, T.G. Schurr et al. 1993. MtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans. *Am. J. Hum. Genet.* 53:591–608.

Underhill, P.A., G. Passarino, A.A. Lin et al. 2001. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.* 65:43–62.

Underhill, P.A., P. Shen, A.A. Lin et al. 2000. Y chromosome sequence variation and the history of human populations. *Nat. Genet.* 26:358–361.

Vasil'ev, S.A. 1993. The Upper Paleolithic of northern Asia. *Curr. Anthropol.* 34:82–92.

Ward, R.H., A. Redd, D. Valencia et al. 1993. Genetic and linguistic differentiation in the Americas. *Proc. Natl. Acad. Sci. USA* 90:10663–10667.

Watson, E., K. Bauer, R. Aman et al. 1996. MtDNA sequence diversity in Africa. *Am. J. Hum. Genet.* 59:437–444.

Wells, R.S., N. Yuldasheva, R. Ruzibakiev et al. 2001. The Eurasian heartland: A continental perspective on Y-chromosome diversity. *Proc. Natl. Acad. Sci. USA* 98:10244–10249.

Wijsman, E.M., and L.L. Cavalli-Sforza. 1984. Migration and genetic population structure with special reference to humans. *Ann. Rev. Ecol. Syst.* 15:279–301.

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

Zerjal, T., B. Dashnyam, A. Pandya et al. 1997. Genetic relationships of Asians and northern Europeans, revealed by Y-chromosome DNA analysis. *Am. J. Hum. Genet*. 60:1174–1183.

Zerjal, T., A. Pandya, F.R. Santos et al. 1999. The use of Y-chromosomal DNA variation to investigate population history: Recent male spread in Asia and Europe. In *Genomic Diversity: Applications in Human Population Genetics,* S.S. Papiha and R. Deka, eds. New York, NY: Plenum, 91–102.