6-1-2015

Mongolians in the Genetic Landscape of Central Asia: Exploring the Genetic Relations among Mongolians and Other World Populations

Jane E. Brissenden
University of Toronto, jbrissenden@on.aibn.com

Judith R. Kidd
Yale University School of Medicine, Judith.Kidd@yale.edu

Baigalmaa Evsanaa
Department of Medicine, University of Toronto, abe_pb@yahoo.com

Ariunaa Togtokh
Department of Nephrology, Health Sciences University of Mongolia, ariun_10@yahoo.com

Andrew J. Pakstis
Yale University School of Medicine, Andrew.Pakstis@yale.edu

See next page for additional authors

Authors
Jane E. Brissenden, Judith R. Kidd, Baigalmaa Evsanaa, Ariunaa Togtokh, Andrew J. Pakstis, Françoise Friedlaender, Kenneth K. Kidd, and Janet M. Roscoe

Mongolians in the Genetic Landscape of Central Asia: Exploring the Genetic Relations among Mongolians and Other World Populations

Jane E. Brissenden,[1]* Judith R. Kidd,[2] Baigalmaa Evsanaa,[3,4] Ariunaa Togtokh,[4] Andrew J. Pakstis,[5] Françoise Friedlaender,[6] Kenneth K. Kidd,[5] and Janet M. Roscoe[1,7]

[1]University of Toronto.

[2]Yale University School of Medicine, New Haven, CT.

[3]Department of Medicine, University of Toronto.

[4]Department of Nephrology, Heath Sciences University of Mongolia, Ulaanbataar, Mongolia.

[5]Department of Genetics, Yale University School of Medicine, New Haven, CT.

[6]Independent scientist, Sharon, CT.

[7]The Scarborough Hospital, Toronto, Canada.

*Correspondence to: Jane E. Brissenden, University of Toronto, 3000 Lawrence Ave East, Suite C-11 Scarborough, Ontario M1P2V1 Canada. E-mail: jbrissenden@on.aibn.com.

Conflict of Interest: The authors declare no conflict of interest.

Ethical Approval: All samples were obtained with informed consent for studies of gene frequency variation under protocols approved by the appropriate IRBs. This article does not contain any studies with animals performed by any of the authors.

KEY WORDS: SNP, HAPLOTYPE, MONGOLIAN, TSAATAN, NATIVE AMERICAN.

Short Title: Mongolians in the Genetic Landscape of Central Asia

Abstract Genetic data on North Central Asian populations are under-represented in the literature, especially autosomal markers. In the present study we use 812 single nucleotide polymorphisms that are distributed across all the human autosomes and that have been extensively studied at Yale to examine the affinities of two recently collected, samples of populations: rural and cosmopolitan Mongolians from Ulaanbaatar and nomadic, Turkic-speaking Tsaatan from Mongolia near the Siberian border. We compare these two populations to one another and to a global set of populations and discuss their relationships to New World populations. Specifically, we analyze data on 521 autosomal loci (single SNPs and multi-SNP haplotypes) studied on 57

populations representing all the major geographical regions of the world. We conclude that the North Central Asian populations we study are genetically distinct from all other populations in our study and may be close to the ancestral lineage leading to the New World populations.

North Central Asian (NCA) populations have been studied by many investigators but are underrepresented in more comprehensive population genetic surveys, such as the HGDP (Cann et al. 2002; Rosenberg et al. 2002; Li et al. 2008) and the 1000 Genomes project (www.ncbi.nlm.nih.gov/bioproject). Because the Mongolian population occupies a geographically central area in Eurasia they may provide genetic information about a number of questions: e.g., (1) What are the genetic relationships of the modern Mongolian subset of the North Central Asian population to a global sample of populations? (2) What are the relationships of ancient and modern Mongolian populations to those of the Americas? (3) Has there been temporal continuity of populations of the region? (4) What have been the migration routes both into and out of Mongolia (5) What can we learn about the demography of these contemporary Mongolian populations? (6) Does the intermediate position in Eurasia of Mongolian populations correspond to an intermediate position on a genetic cline across Eurasia? As genetic data are accumulating, answers (and more questions) are emerging. Some of these studies involve uniparental DNA [mitochondrial (mt)

and non-recombining Y chromosome DNA (NRY)] (Chatters et al. 2014; Duggan et al. 2013; Fedorova et al. 2013; Hertzberg et al. 1989; Kemp et al. 2015; Kitchen et al. 2008; Kolman et al. 1996; Lell et al. 2002; Malyarchuk et al. 2011; Malyarchuk et al. 2013; Mulligan et al. 2008; Nasidze et al. 2005; Dulik et al. 2012; Shi et al. 2013; Starikovskaya et al. 2004; Sukernik et al. 2012; Raghavan et al. 2014a; Volodko et al. 2008; Zhong et al. 2011; Zhong et al. 2010), ancient DNA (aDNA, uniparental and autosomal) (Crubezy et al. 2010; Keyser-Tracqui et al. 2003, Keyser-Tracqui et al. 2006; Malyarchuk et al. 2011; Raghavan et al. 2014b). Mitochondrial DNA haplogroups offer a fairly crude inference of continental ancestry, conveying only information regarding possibly one or two top ancestry components while losing other ancestry information (Emery et al. 2015). Other studies used contemporary autosomal DNA (Fedorova et al. 2013; Keyser-Tracqui et al. 2003; Keyser-Tracqui et al. 2006; Kidd et al. 2011a,b; Nasidze et al. 2006; Reich et al. 2012; Tian et al. 2008). These studies were excellent but many more autosomal markers must be investigated, particularly in sparsely studied populations such as Mongolians, especially Outer Mongolians who have been less studied than Mongolians living in Chinese Inner Mongolia whose demographic history is also somewhat different, to clarify population relationships.

Progress in answering the questions posed above is underway though by no means are the answers definitive yet. In this report we shall review some of

the relevant studies that include population data from North Central Asia and contribute our findings to help answer some of those questions.

Relationship of Mongolians Populations to a Global Sample of Populations.

First from studies of ancient DNA; Raghavan et al. (2014a) obtained mitochondrial DNA, Y chromosome sequence, and autosomal DNA from a 24,000 year old person (MA-1) from south-central Siberia and suggested that 14 to 38% of Native American ancestry may have originated through gene flow from populations related to contemporary western Eurasians who had a more north-easterly distribution 24,000 years ago than commonly thought. Zhong et al. (2010) used an ancient Y-chromosome haplogroup C-M130 (Hg C) to propose that ancient migration routes were derived from the African exodus and gradually colonized South Asia, Southeast Asia, Oceania and East Asia by a single Paleolithic migration from Africa to Asia and Oceania which occurred more than 40 KYA, and then moved north in mainland East Asia; likely following the coastline. Hertzberg et al. (1989) used an Asian-specific 9-bp deletion of Mitochondrial DNA to support the thesis that an independent group of pre-Polynesian ancestors who colonized into the Pacific were ultimately derived from East Asia. Duggan et al. (2013) used mtDNA and NRY DNA to investigate the prehistory of Tungusic-speaking reindeer herders and hunter-gatherers (Evenks and Evens) from Northern Asia and other groups in the southern part of Siberia. Their data argue that the Evens and Evenks do stem

from a common ancestral population but genetic drift and differential admixture with neighboring populations have added to the divergence of the two populations.

Second for more modern DNA the Reich et al. (2012) study used autosomal DNA and typed SNPs, both single and when combined into haplotypes to also analyze Native American ancestry. Reich et al. masked genomic segments containing ancestry thought to be non-Native American and analyzed the population structure with Principal Components Analysis. They also computed pair-wise population F_{st}. Reich et al. used multiple applications of their 4 Population Test to demonstrate the existence of a minimum of three gene flow events from Asia to explain the data from Native American populations jointly which they called: First American ancestry and two additional streams, the Eskimo-Aleut and finally Na-Dene speaking Chipewyan. These latter two lineages derive large portions of their genomes from First American ancestors.

The uniparental markers show individuals with different haplogroups, some of which are related to haplogroups common in more eastern populations and some of which, especially in ancient DNA, show relationships to more Western Eurasian populations (Zhong et al. 2011; Duggan et al. 2013; Keyser et al. 2009, Keyser-Tracqui et al. 2003). However, different studies have used different sets of populations for comparisons making a general conclusion difficult. Further, because these uniparental markers represent only a fraction of

genetic ancestry and are subject to strong genetic drift, overall genetic relationships are difficult to quantify. Fedorova et al. (2013) have utilized mtDNA, NRY DNA, and autosomal DNA to add additional Siberian populations to the dataset. These data included autosomal SNPs from 40 newly genotyped individuals, published data on Siberia; and relevant global reference populations now studied with F_{st}, principal component analysis (PCA) and ADMIXTURE. As well, 829 mtDNAs and 375 Y chromosomes from five populations were studied by phylogenetic analysis. Federova et al. (2013) concluded that the indigenous people of the Sakha Republic of Russia descended mostly from the common East Asian gene pool with little West Eurasian influence. Turkic-speaking Yakuts have some relationship to Altai-Sayan populations and distinctive maternal and paternal lineages originating in the Mongolic peoples in the Lake Baikal area or influenced from long timescale residence in Sakha close to Tungusic-speaking Evenks. They conclude that Sakha was colonized by multiple migrations from South Siberia with only minor gene flows from the lower Amur/Southern Okhotsk region and/or Kamchatka (in the South-East).

Relationship of Mongolian Populations to Populations of the Americas.

There is general agreement that the population(s) that founded the Americas originated from ancient Siberian populations (Kitchen et al. 2008; Kolman et al. 1996; Lell et al., 2002; Schurr 2004; Schurr et al. 2010; Malyarchuk et al. 2011; Mulligan et al. 2008; Raghavan et al. 2014a; Reich et al.

2012; Starikovskaya et al. 2003; Sukernik et al. 2012; Volodko et al. 2008), and there is general agreement that the Native American populations were separated from Old World populations by at least 12,000 to 16,000 ybp as they colonized the New World. The group(s) who expanded into the New World may have initially experienced an additional period of about 15,000 years of isolation within Siberia/Beringia (Kitchen et al. 2008; Kolman et al. 1996; Mulligan et al. 2008; Volodko et al. 2008) before entering the subglacial Americas. This timing (Kitchen et al. 2008; Mulligan et al. 2008) makes the evidence for Bronze and Iron Age migration eastward into Siberia much too recent to bear on migration into the New World from Siberia. The particular modern Siberian populations most closely sharing ancestry with Native Americans are not certain.

Migration.

Analyses of 32 ancient DNAs (Keyser et al. 2009) ranging from Middle Bronze Age (1800 BC) to the Iron Age (about 1600 years ago) of Kurgan burials in Siberia between Krasnoyarsk (~91° E, 53° N) and Abakan (~93° E, 53° N) immediately adjacent to Mongolia led to the conclusion that the people represented by these Kurgan burials were light skinned with blond or light brown hair and that 60% had blue or green eyes (European phenotypes). This conclusion was based on a few autosomal phenotypic markers but is also supported by the Y-chromosome haplotypes [op cit] of the Kurgans present today predominantly in Eastern Europe. On the other hand, five of these

individuals (all from Iron Age samples) bore an East Asian mtDNA. These DNA analyses thus support Kurgan migration into Mongolia both from the West and at a somewhat later time period from the East.

Ancient DNA and Temporal Continuity.

Comparisons of ancient DNA with DNA from contemporary populations provides conflicting evidence. Genetic distance analysis of autosomal STR data (Keyser-Tracqui et al. 2006) found no significant differences among three populations, present-day Mongolians, a present-day Egyin Gol population, and an ancient (about 2400 years old) Egyin Gol cemetery population of Xiongnu (i.e., those three populations are genetically indistinguishable). Roughly the same trend was reported for mtDNA and Y chromosome DNA. In a separate paper that ignores their own conflicting STR results, Keyser et al. (2009) conclude that south Siberian Kurgan populations were predominantly of European phenotype during the Bronze and Iron ages. Thus, since modern populations are not European, there would appear to be no continuity over these several thousand years, answering our question 3. Pakendorf et al. (2006) reported uniparental DNA (mt and Y chromosome) data showing Yakut origins to have been in the cis-Baykal region. Crubezy et al. (2010) supported the Pakendorf et al. (2006) data with their analysis of autosomal STRs, Y-chromosomal STRs, SNPs and mtDNA to analyse 58 mummified frozen bodies dating from the fifteenth to the nineteenth century, excavated from Yakutia in Eastern Siberia. The Yakuts had

been assumed to originate from South Siberian populations but the precise origin of their paternal lineages and their admixture rate with indigenous populations is not fully understood. With this study Crubezy et al. (2010) traced the male lineages back to a small group of horse-riders from the Cis-Baikal area. Their mtDNA data showed that intermarriages occurred between the first settlers and Evenks women and established genetic characters during the fifteenth century that exist today.

Kidd et al. (2011a,b) studied phylogenetic origins and ancestry differentiation with SNPs and haplotypes concluding that a single dataset of a small number of markers cannot answer both questions of phylogeny and differentiation. They found that random SNPs and haplotyped loci are best to decide evolutionary relationships on a global level and concluded that more populations must be studied for a single large set of markers to clarify these population relationships. In the present study we are expanding the number of populations and the number of autosomal loci available to help improve our understanding of how Mongolian populations are related to the development of modern human populations. The results we report will provide additional SNPs for ancestry inference, a long time interest of ours (Kidd et al. 2014a,b).

Material and Methods

Population Samples.

Samples of 104 unrelated individuals from Mongolia were the focus of our study using DNA purified from saliva samples. These included 58 Mongolian students attending Health Sciences University of Mongolia (HSUM) in the capital city of Ulaanbaatar and 46 individuals from the isolated, migratory population of Tsaatan reindeer herders. The HSUM students represent individuals from all geographic regions of Mongolia. These students were invited to donate saliva samples for the study and those who agreed to participate were sampled at the University. The Tsaatan participants were drawn from a small population (of about 400 total) reindeer herders who speak Dukha, a Turkic language. All samples were collected with the informed consent of all subjects and under protocols approved by all relevant and appropriate Institutional Review Boards. (Research Ethics Board, The Scarborough Hospital, 3050 Lawrence Ave. East, Toronto, Ontario, M1P 2V5; and Health Sciences University of Mongolia, S. Zorig Street, 3 Health Sciences University of Mongolia, Ulaanbaatar, Mongolia.) The informed consent was also consistent with that required by the IRB at Yale. The consent forms contained questions about the maternal and paternal names and languages and also about those of the maternal and paternal grandparents. These data documented the population attribution of the sampled individuals.

In addition to these two North Central Asian population samples, we have previously genotyped and extensively analyzed 2478 samples from 55

populations described elsewhere (Kidd JR et al. 2011; Kidd KK et al. 2014a,b) which we will refer to as our reference populations here. The total present data set consists of 2582 DNA samples from 57 populations from around the world. Table 1 lists the names of the populations studied, the number of individuals in each population sample, and the general geographical region of origin.

Loci Studied.

Genotype data were collected for 812 single nucleotide polymorphisms (SNPs). The final data set consisted of 157 multiallelic haplotype loci (defined by 448 SNPs) and 364 single nucleotide polymorphisms for a total of 521 loci. The final data set included a panel of individual SNPs that had already been genotyped on a large number of individuals from diverse regions of the world. In almost all cases, supplies to test these markers were already available. These markers had been studied over the years for several different reasons, including studies of normal variation at disease-related genes, as well as evaluations of SNPs for being ancestry informative but that subsequently showed less than optimal variation for that purpose. The ancestry informative markers genotyped included, among others, the 128 SNPs from Kosoy et al. (2009), 55 SNPs from Kidd et al. (2014a and b). Overall other than a bias for higher levels of global gene frequency variation, there should not be a bias for the particular pattern of difference/similarity of these two populations among the existing populations. SNPs chosen for this study were clearly not random but the selection should not

have biased the qualitative relationship between the populations. The remainder were chosen as part of a high global F_{st} project underway in the lab including past typing of the 55 reference populations on the Illumina 650Y chip (Donnelly et al. 2010; Han et al. 2005; Speed et al. 2009; Pakstis et al. 2012). SNPs close enough to show significant linkage disequilibrium with at least moderate to strong values were combined into haplotypes. All individual SNPs are listed in Supplemental Material by their dbSNP 'rs' number. Allele frequencies for the haplotypes and for all individual SNPs are either available in ALFRED (http://alfred.med.yale.edu/) or from the authors upon request.

Genotyping.

The samples from Mongolia were collected in Oragene saliva collection kits available from DNA Genotek and DNA was isolated using the manufacturer's protocol. DNA for all other populations was purified (Sambrook et al. 1989) from lymphoblastoid cell lines (Anderson and Gusella 1984).

All genotyping utilized TaqMan SNP Genotyping Assays. The data on the 55 reference populations were mostly pre-existing or, when one of our reference populations had not been genotyped for a specific SNP in this study, were typed on genomic DNA directly following the manufacturer's protocol in 3 μl reactions. To maximize the number of SNPs that could be typed on the limited amounts of DNA available on the samples of the two Mongolian populations, a

brief pre-amplification was performed. To a pool of 1 µl each of 96 20x TaqMan assays we added 250 µl of Master Mix and 200 µl of water. In a 96 well plate 5 µl of that cocktail were added to approximately 30 ng of DNA for each individual. These samples were then pre-amplified as follows: 10 minute presoak at 95 degrees, followed by 12 cycles at 95 degrees for 15 seconds and 60 degrees for 4 minutes. Each of these pre-amplification reactions was diluted with water to a volume of 250 µl. Then 2 µl of each preamplified sample were aliquoted to a designated well in 384-well plates—one well for each individual, and one plate for each SNP in the assay pool. The material in the plates became the basic template for the individual TaqMan assays following the manufacturer's protocol.

Statistical Methods.

Data on all SNPs and haplotypes were tested for deviations from Hardy-Weinberg ratios in all population samples using simple chi-square tests and/or simulation. The haplotypes were inferred by PHASE (Stephens et al. 2001; Stephens and Scheet 2005). Population clusterings were based on STRUCTURE v 2.3.3 (Pritchard et al. 2000). Each run consisted of 20,000 burn-ins plus 20,000 MCMC with 20 replicates at each K value.

Principal Component Analysis (PCA) of the populations based on the population-specific SNP and haplotype allele frequencies used the software in

XLstat 2010 (version 2009.4.07; Addinsoft SARL, www.xlstat.com/en/company/). Multiple Dimensional Scaling (MDS) also used the matrix of allele frequencies and the functions in XLstat.

A matrix of pairwise Tau genetic distances (Kidd and Cavalli-Sforza 1974) was calculated from the allele frequencies of 521 loci, single SNPs and haplotypes. Under simplifying assumptions these distances are theoretically additive and equal to time in generations divided by twice the effective population size separating two populations; the correct tree should have the pairwise distances equal to the sum of the branches connecting the two populations. The Neighbor Joining (NJ) algorithm in the Phylogeny Inference Package (PHYLIP) software (Felsenstein 1989; 2009) uses such a pairwise genetic distance matrix to generate a bifurcating tree structure and the segment lengths that are an approximate additive least squares (LS) solution to the pairwise distances. Bootstrap values based on 1000 iterations were calculated using the Reynolds distance measure (Reynolds et al. 1983) and programs in PHYLIP.

Results

The SNPs were uniformly typed in all individuals with missing data for no more than 9% of the individuals for any SNP or 10% of the SNPs for any individual

(total 1.7% missing data overall). There were no significant Hardy-Weinberg deviations for the SNP frequencies beyond what would be expected by chance for the number of tests.

MDS (multidimensional scaling) of the populations generated a graphic of the relationships (Figure 1) based on the allele frequency matrix of all 521 loci, individual SNPs and haplotyped loci. The first dimension contrasts the Native South American populations (most specifically the Karitiana) opposite African populations (most specifically Mbuti and Biaka). The second dimension is defined by Europeans and the South West Asians (most specifically Samaritans) opposite Pacific populations (most specifically Ami, Atayal, and Papua New Guinea). The three East Indian populations (Keralite, Thoti, and Kachari) and the Khanty occupy a very central position; the next most central populations are the Tsaatan, Mongolians, and Yakut. We then reran the MDS without the African, SW Asian, and European populations because they are so distinctly different from the Mongolian populations. The resulting plot (Figure 2) shows the two Mongolian populations and the Yakut from Eastern Siberia to be quite central with the East Asians, Native Americans, South Asians, and Pacific populations off to the right, upper left, lower left, and lower right, respectively. (Comparable PCA analyses are in Supplemental Figures 1 and 2.)

Figure 3 illustrates the STRUCTURE analysis solution with the highest likelihood among the largest set of highly correlated solutions (pairwise

similarity statistic--G > 0.9) out of 20 replicates. The column on the right of the figure indicates the frequency of the illustrated solution for each value of K (number of assumed different underlying populations). As shown in Supplemental Figure S3, the likelihoods of the "best" solutions continue to rise from values of K = 2-11. At all K values of seven or greater the two Siberian populations (Khanty and Yakut) cluster with the two Mongolian populations (Mongolian and Tsaatan). There are indications of some small similarities of some individuals to European and/or East Asian clusters. It is clear that African, European, and south Asian populations are largely different with the exception of some European similarity for the Khanty.

Figure 4 displays the outcome of a similar STRUCTURE analysis of a subset of the populations—those from the Ural Mountains eastward into the Americas. As shown in Supplemental Figure 4, the likelihoods for the 20 replicate runs continues to improve to K = 8. In this analysis we can see (1) more clearly the distinction between India (Keralites, Thoti, and Kachari) and the rest of Asia, and (2) the distinction between South East Asia (Lao and Cambodians) from the East Asian groups and from the Yakut, Mongolians, and Tsaatan. As in Figure 3 the column on the right of Figure 4 shows the frequency of the illustrated general solution for each value of K. We note the Khanty (from Western Siberia) now show relationship to South Asia at K = 6 or greater. This is explained as Khanty sharing some Western Asia ancestry that would have shown

up if South-West Asian populations had been included (data not shown) as in the earlier analysis based on all the populations studied.

In summary with reference to point #1 in the introduction the Mongolian and Tsaatan populations share similarity with the Yakut but are clearly distinct from all other populations studied.

Figure 5 is the Neighbor Joining tree. The branches with high bootstrap values are indicated. Overall, there is very strong support for this grouping of populations. The African cluster is separated from the rest of the populations with a 100% bootstrap value. Within the African cluster three segments have 100% bootstrap values and three have values between 98% and 100%. The Kuwaiti are significantly closer (100% bootstrap value) to African populations than to the Middle East and Europe but there are no high bootstrap values within the Middle East - Europe cluster. Moving to the East there are high bootstrap values flanking the western Siberian Khanty and two (Kachari and Thoti) of the three Indian populations as distinct from both the western and eastern populations. The branch into the Americas is separated from all other populations by 100% bootstrap values. The other highly significant branches are the one into Melanesia (100%), the one into the central Pacific (99.67%), and several branches among East Asian populations (98 to 100%). While the major structure of this tree is unambiguous, the branch defining the Mongolian populations and the Yakut has a 94.4% bootstrap value even though it is a

miniscule branch connecting to the Native American branch. The branch connecting Native Americans and the three populations just mentioned to the rest of the tree has an 86.1% bootstrap value. Regarding Introduction point #2 this Native American/Mongolian/Yakut/Tsaatan branch suggests that the Mongolian/Tsaatan/Yakut populations may be close to the ancestral lineage leading to the New World populations. The Pacific populations as a group, and the Malaysians are not well supported.

Discussion

Our study provides information on 517 autosomal loci (composed of single SNPs and multi-SNP haplotypes) from 57 populations. New data on two Mongolian populations, cosmopolitan Mongolians and Tsaatan, with reasonably large sample sizes (58 and 46 samples, respectively) greatly increase the genetic information from North Central Asia. With reference to point #1 in our introduction Structure analysis demonstrates that the Mongolians, Tsaatan, and Yakut are clearly distinguishable from all other world populations in our analysis. These data extend the autosomal SNP studies of Tian et al. (2008) and Reich et al. (2012) and the uniparental marker studies of Schurr et al. 2004, 2010. With respect to global diversity, we found clear similarity between the nomadic, isolated Tsaatan group and the sample of metropolitan Mongolians.

The moderately high bootstrap values of 94.4% grouping the two Mongolian with Yakut and 86.1%, separating these and Native Americans from the rest of the populations we studied are consistent with the idea that modern Mongolian populations are genetically close to the original lineage leading to Native Americans (Point #2). In conjunction with the study of Fedorova et al. (2013) and Schurr et al, 2010, the relationships between Mongolians and Siberian populations are clarifying.

With these populations, STRUCTURE analysis (Figures 3 and 4) shows the Mongolian, Tsaatan, Yakut cluster is distinct from the current East, South, South East, and Pacific clusters and that Native Americans are not similar to any Old World populations. The Neighbor Joining Tree (Figure 5) is compatible with a long period of separation and random genetic drift since the separation of Native Americans from Old World populations. The Native Americans are at a shorter distance from the present day Mongolians etc, than others; this can be due to admixture and not necessarily due to a common founding lineage. What can be stated at the best is that the conclusion that—there was a long period of separation and random genetic drift since a founding lineage—is potentially compatible with the Neighbor Joining Tree found here. These data relate to Point #4 in the introduction and suggest that the migration routes from ancient Siberia to North American may have occurred from around Mongolia and Lake Baikal. With regard to Point #5 our data suggests that the contemporary Mongolian

populations are clearly distinguishable from all other world populations in our analysis. With regard to our last point (Point #6) Mongolia's (including Tsaatan, and Yakut) position in North Central Asia intermediate in Eurasia and at the root of a branch on the Neighbour Joining tree leading to Native Americans likely represents a genetic intermediate position in a genetic cline across Asia. Future genetic studies in the North Central Asian geographic area will clarify and extend these findings.

Conclusion

This study adds to and strengthens the suggestion that North Central Asian populations are distinct from those of Far East Asia and have the shortest genetic distance from New World populations of all our reference populations.

Acknowledgments We thank Ike Zhang, Amanda Brissenden, and Rosemary Palmieri of Toronto and New Haven, respectively, for their excellent technical assistance. We acknowledge the constant financial and logistic support of the Scarborough Nephrology Associates (including Drs. Paul Tam, Robert Ting, Gordon Nagai, Tabo Sikaneta, Paul Ng) without whom this work would not have been possible. We also acknowledge and thank the International Society of Nephrology for educational support of Dr. Baigalmaa Evsanna as an ISN

Nephrology Fellow. Special thanks are due to the many hundreds of individuals who volunteered to give blood or saliva samples for studies of gene frequency variation and to the many colleagues who helped collect the samples. We also want to thank all the collaborators who helped to collect the population samples. Cell lines for some of the populations were obtained from the National Laboratory for the Genetics of Israeli populations at Tel-Aviv University and the Coriell Cell Repositories, Camden, NJ. Some of the cell lines were obtained from the National Laboratory for the Genetics of Israeli Populations at Tel Aviv University and the Coriell Institute for Medical Research, Camden, New Jersey. The work at Yale was funded primarily by Grants 2010-DN-BX-K225 and 2013-DN-BX-K023 to KKK awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice.</ACK>

Received 13 March 2015; revision accepted for publication 20 September 2015.

Literature Cited

Anderson, M., and J. F. Gusella. 1984. The use of cyclosporine A in establishing human EBV-transformed lymphoblastoid cell lines. *Vitro.* 11:856–858.

Cann, H. M., C. de Toma, L. Cazes et al. 2002. A human genome diversity cell line panel. *Science* 296:261–262.

Chatters J.C., D. J. Kennett, Y. Asmerom et al. 2014. Late Pleistocene human skeleton and mtDNA link Paleoamericans and modern Native Americans. *Science* 344:750–754. doi:10.1126/science.1252619.

Crubezy, E., S. Amory, C. Keyser et al. 2010. Human evolution in Siberia: from frozen bodies to ancient DNA. *BMC Evol. Biol*. 10:25. doi:10.1186/1471-2148-10-25.

Donnelly, M. P., P. Paschou, E. Grigorenko et al. 2010. The distribution and most recent common ancestor of the 17q21 inversion in humans. *Am. J. Hum. Genet.* 86:161–171.

Duggan, A. T., M. Whitten, V. Wiebe et al. 2013. Investigating the prehistory of Tungisic peoples of Siberia and the Amur-Ussuri region with complete mtDNA genome sequences and Y-chromosome markers. *PLoS One* 8:E83570. doi:10.137//journal.pone.0083570.

Dulik, M. C., S. I. Zhadanov, L. P. Osipova et al. 2012. Mitochondrial DNA and Y chromosome variation provides evidence for a recent common ancestry between Native Americans and Indigenous Altaians. *Am. J. Hum. Genet*.

90:229–246. doi:10.1016/j.ajhg.2011.12.014. Erratum in: *Am. J. Hum. Genet*. 2012 90:573.

Emery, L. S., K. M. Magnaye, A. W. Bigham et al. 2015. Estimates of continental ancestry vary widely among individuals with the same mtDNA haplogroup. *Am. J. Hum. Genet.* 96:183–193.

Fedorova, S. A., M. Reidla, E. Metspalu et al. 2013. Autosomal and uniparental portraits of the native populations of Sakha (Yakutia): implications for the peopling of Northeast Eurasia. *Evol. Biol.* 13:127–137.

Felsenstein, J. 1989. PHYLIP-Phylogeny Inference Package (Version 3.2). *Cladistics* 5:164–166.

Felsenstein, J. 2009. PHYLIP (Phylogeny Inference Package) version 3.7a. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.*

Han, Y., H. Oota, M. V. Osier et al. 2005. Considerable haplotype diversity within the 23kb encompassing the *ADH7* gene. *Alcohol Clin. Exp.* 29:2091–2100.

Hertzberg, M., K. N. P. Mickleson, S. W. Serjeantson et al. 1989. An Asian-specific 9-bp deletion of mitochondrial DNA is frequently found in Polynesians. *Amer. J. Hum. Genet.* 44:504–510.

Kemp, B. M., J. Lindo, D. A. Bolnick et al. 2015. Anthropology. Response to Comment on "Late Pleistocene human skeleton and mtDNA link Paleoamericans and modern Native Americans." *Science* 347:835. doi:10.1126/science.1261188.

Keyser, C., C. Bouakaze, E. Crubez. 2009. Ancient DNA provides new insights into the history of south Siberian Kurgan people. *Hum. Genet.* 126:395–410.

Keyser-Tracqui, C., E. Crubezy, B. Ludes. 2003. Nuclear and mitochondrial DNA analysis of a 2,000-year-old necropolis in the Egyin Gol Valley of Mongolia. *Amer. J. Hum. Genet.* 73:247–260.

Keyser-Tracqui, C., E. Crubezy, H. Pamzsav. 2006. Population origins in Mongolia: Genetic structure analysis of ancient and modern DNA. *Am. J. Phys. Anthropol.*131:272–281.

Kidd, J. R., F. Friedlaender, A. J. Pakstis et al. 2011a. SNPs and haplotypes in Native American populations. *Am. J. Phys. Anthropol.* 146:495–502.

Kidd, J. R., F. R. Friedlaender, W. C. Speed et al. 2011b. Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 populations. *Investig. Genet.* 2:1–13.

Kidd, K. K. and L. L. Cavalli-Sforza. 1974. The role of genetic drift in the differentiation of Icelandic and Norwegian cattle. *Evolution* 28:381–395.

Kidd, K. K., W. C. Speed, A. J. Pakstis et al. 2014a. Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Sci. Int. Genet*. 10:23–32.

Kidd, K. K., A. J. Pakstis, W. C. Speed et al. 2014b. Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics. *Forensic Sci. Int. Genet.* 12:215–224.

Kitchen, A., M. M. Miyamoto, C. J. Mulligan. 2008. A three-stage colonization model for the peopling of the Americas. *PLoS ONE* 3L:e1596.

Kolman, C. J., N. Sambuughin, and E. Bermingham. 1996. Mitochondrial DNA analysis of Mongolian populations and implications for the origin of new world founders. *Genet*. 142:1321–1334.

Kosoy, R., R. Nassir, C. Tian et al. 2009. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Human. Mutat.* 10: 69–78. DOI:10.1002/hummu, 20822.

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.

Li J. Z., D. M. Absher, H. Tang et al. 2008b. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319: 1100–1104.

Malyarchuk, B., M. Derenko, G. Denisova et al. 2013. Y-chromosome diversity in the Kalmyks at the ethnical and tribal levels. *J. Hum. Genet*. 58: 804–811.

Malyarchuk, B., M. Derenko, G. Denisova et al. 2011. Ancient links between Siberians and Native Americans revealed by subtyping the Y chromosome haplogroup Q1a. *J. Hum. Gen.* 56: 583–588.

Mulligan, C. J., A. Kitchen, M. M. Miyamoto. 2008. Updated three-stage model for the peopling of the Americas. *PLoS One* 3:e3199.

Nasidze, I., D. Quinque, I. Dupanloup et al. 2005. Genetic evidence for the Mongolian ancestry of Kalmyks. *Amer. J. Phys. Anthropol.* 128: 846–854.

Pakendorf, B., I. N. Novgorodov, V. L. Osakovskij et al. 2006. Investigating the effects of prehistoric migrations in Siberia: genetic variation and the origins of Yakuts. *Hum. Genet.* 120: 334–353.

Pakstis, A. J., R. Fang, M. R. Furtado et al. 2012. Mini-haplotypes as lineage informative SNPs and ancestry inference SNPs. *Eur. J. Hum. Genet.* 20:1148–1154.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genet*. 155:945–959.

Raghavan, M., M. DeGiorgio, A. Albrechtsen et al. 2014a. The genetic prehistory of the New World Arctic. *Science* 345:1255832. doi:10.1126/science. 1255832.

Raghavan, M., P. Skoglund, K. Graf et al. 2014b. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* 505:87–91.

Reich, D., N. Patterson, D. Campbell et al. 2012. Reconstructing Native American population history. *Nature* 488:370–374. doi:101038/Nature 11258. *Erratum in* 2012 *Nature* 491:288.

Reynolds, J., B. S. Weir, C. C. Cockerham. 1983. Estimation of the co-ancestry coefficient: Basis for a short-term genetic distance. *Genet.* 105:767–779.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber et al. 2002. Genetic structure of human populations. *Science* 298:2381–2385.

Sambrook, J., E. F. Fritsch, T. M. Maniatis. 1989. *Molecular Cloning: A Laboratory Manual*. Second Edition. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.

Schurr, T. G. 2004. The peopling of the New World: Perspectives from molecular anthropology. *Ann. Rev. Anthropol.* 33:551–583.

Schurr, T. G., L. P. Osipova, S. I. Zhadanov et al. 2010. Genetic diversity in Native Siberians: Implications for the prehistoric settlement of the Cis-Baikal region. In *Prehistoric Hunter-Gatherers of the Baikal Region,*

Siberia: Bio-archaeological Studies of Past Life Ways, edited by A. W. Weber, M. A. Katzenberg, and T. G. Schurr. Philadelphia: University of Pennsylvania Museum of Archaeology and Anthropology.

Shi, H., X. Qi, H. Zhong et al. 2013. Genetic evidence of an East Asian origin and Paleolithic northward migration of Y-chromosome haplogroup N. *PLoS ONE* 8:e66102.

Speed, W. C., S. P. Kang, D. P. Tuck et al. 2009. Global variation in *CYP2C8-CYP2C9* functional haplotypes. *Pharmacogenomics J.* 9:283–290.

Starikovskaya, E. B., R. I. Sukernik, O. A. Derbeneva et al. 2003. Mitochondrial DNA diversity in indigenous populations of the southern extent of Siberia, and the origins of Native American haplogroups. *Ann. Hum. Genet.* 69:67–89.

Stephens, M., N. J. Smith, P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68:978–989.

Stephens, M., P. Scheet. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* 76:449–462.

Sukernik, R. I., N. V. Volodko, I. O. Mazunin et al. 2012. Mitochondrial genome diversity in the Tubalar, Even, and Ulchi: Contribution to prehistory of

native Siberians and their affinities to Native Americans. *Am. J. Phys. Anthropol.* 148:123–138.

Tarazona-Santos, E., and F. R. Santos. 2002. The peopling of the Americas: A second major migration? *Am. J. Hum. Genet.* 70:1377–1380.

Tian, C., R. Kosoy, A. Lee et al. 2008. Analysis of East Asia genetic substructure using genome-wide SNP arrays. *PLos One* 3:e3862.

Volodko, N. V., E. B. Starikovskaya, I. O. Mazunin et al. 2008. Mitochondrial genome diversity in Arctic Siberians, with particular reference to the evolutionary history of Beringia and Pleistocenic peopling of the Americas. *Am. J. Hum. Genet.* 82:1084–1100.

Zhong, H, H. Shi, X.-B. Qi et al. 2010. Global distribution of Y-chromosome haplogroup C reveals the prehistoric migration routes of African exodus and early settlement in East Asia. *J. Hum. Genet.* 55:428–435.

Zhong, H., H. Shi, X.-B. Qi et al. 2011. Extended Y chromosome investigation suggests postglacial migrations of modern humans into East Asia via the Northern route. *Mol. Biol. Evol.* 28:717–727.

Captions for Table and Figures

Table 1. The populations studied (57), the number of individuals in each population sample (total 2582), and the general geographical region of the origin of population groups

Fig. 1. Graphic of the MDS analysis of all 57 populations. In this analysis the first dimension is bounded by South American populations opposite African populations and the second dimension by Europeans and South West Asians opposite Pacific and South East Asian populations. The remaining populations are more centrally located

Fig. 2. Graphic of the MDS analysis of the 28 populations east of the Ural Mountains through the Americas. In this analysis the both dimensions are bounded by the South American populations opposite the Pacific populations. The Mongolians, Tsaatan, Yakut, and Khanty are very central

Fig. 3. Graphic of the highest likelihood of the most common mode STRUCTURE solutions for K values from 2-11 for 57 populations. The ratio on the right indicates the fraction of the 20 independent runs that had the pattern illustrated

Fig. 4. Graphic of the highest likelihood of the most common mode STRUCTURE solutions for K values from 2-8 for 28 populations. As in Figure 3 the ratio on the right indicates the fraction of the 20 independent runs that had the pattern illustrated

Fig. 5. The Neighbor Joining tree. Bootstrap values are indicated on the segments across which the opposite sides of the branch are separated by >98% of the 1000 bootstraps. Note, significance by bootstrap values is not represented by the length of the segments. Bootstrap values represent consistency among loci while segment lengths represent average difference among populations based on fitting an additive model to pairwise genetic distances

Captions of Figures in Supplementary Material

Fig. S1. PCA based on Tau genetic distances of 57 populations. (a) Principal components 1 and 2; (b) principal components 1 and 3.

Fig. S2. PCA of East Asians & the Americas. (a) Principal components 1 and 2; (b) principal components 1 and 3.

Fig. S3. The likelihood plot for STRUCTURE on all 57 populations.

Fig. S4. The likelihood plot for STRUCTURE on 28 populations.

TABLE 1. Populations studied						
Region	**Population**	**N**		**Region**	**Population**	**N**
Africa	Biaka	66		W Siberia	Komi Zyrian	46
	Mbuti	38		E Siberia	Khanty	49
	Lisongo	7			Yakut	49
	Yoruba	77			Tsaatan	46
	Ibo	47			Outer Mongolians	58
	Hausa	38		S Asia	Keralites	30
	Chagga	45			Thoti	13
	Masai	20			Kachari	17
	Sandawe	40		E Asia	Chinese, San Francisco	56
	Zaramo	37			Chinese, Taiwan	48
	Afr Americans	89			Hakka	41
	Ethiopian Jews	31			Koreans	54
					Japanese	44
SWAsia	Yemenite Jews	41		SE Asia	Laotians	118
	Kuwaiti	12			Cambodians	23
	Druze	101			Ami	40
	Samaritans	38			Atayal	41
	Ashkenazi	79		Pacific	Papuans, New Guinea	22
					Nasioi	23
Europe	Adygei	54			Malaysians	10
	Chuvash	41			Samoans	9
	Hungarians	89			Micronesians	34
	Russians, Archangelsk	33		Americas	Pima, Mexico	53
	Russians,Vologda	47			Maya, Yucatan	48
	Finns	34			Guihiba, Colombia	11
	Danes	51			Quechua, Peru	22
	Irish	111			Ticuna	65
	Euro Americans	88			Rondonian Surui	43
	Sardinians	34			Karitiana	50
	Roman Jews	27			**Total Individuals**	2582

Figure 1.

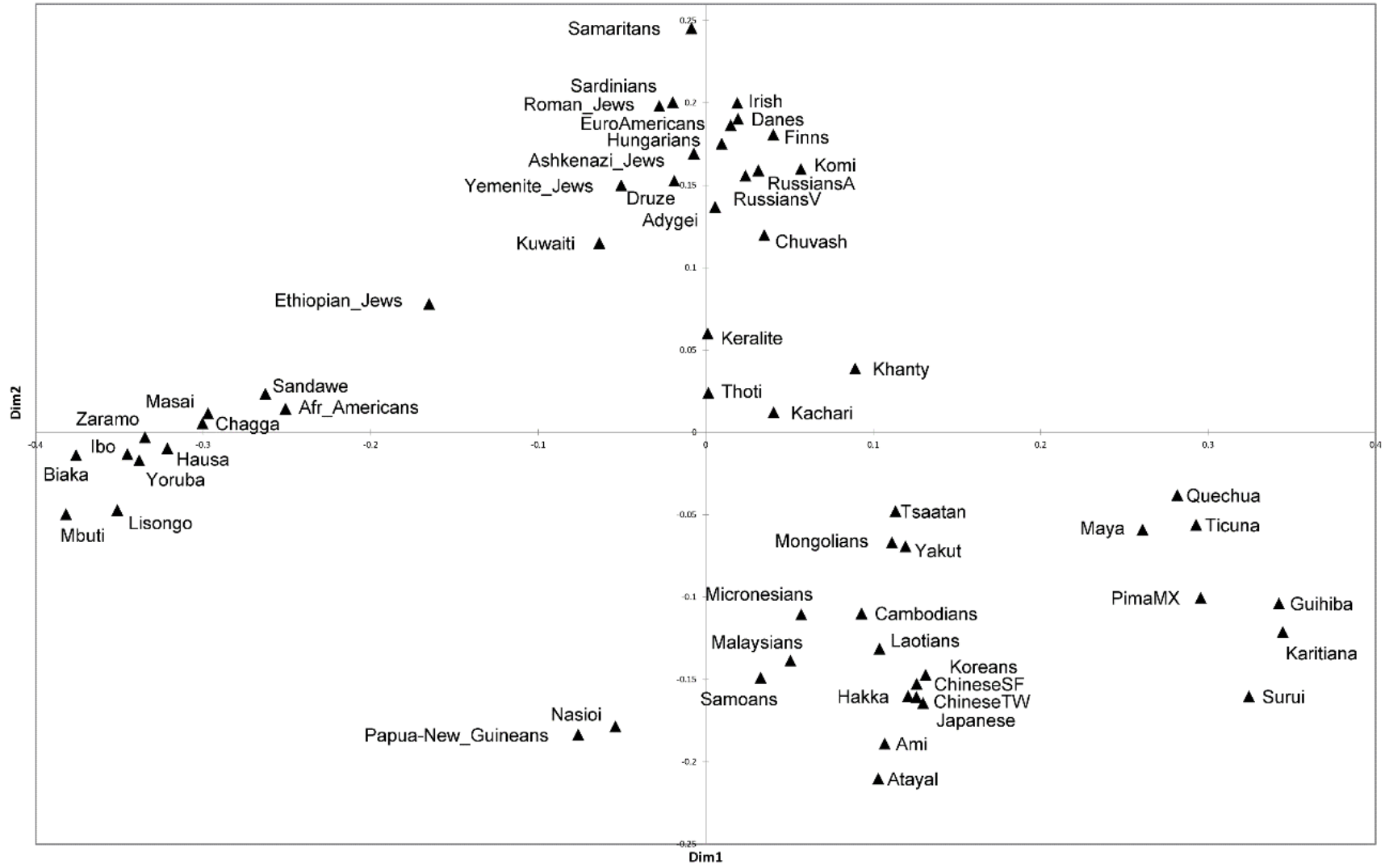


Figure 2

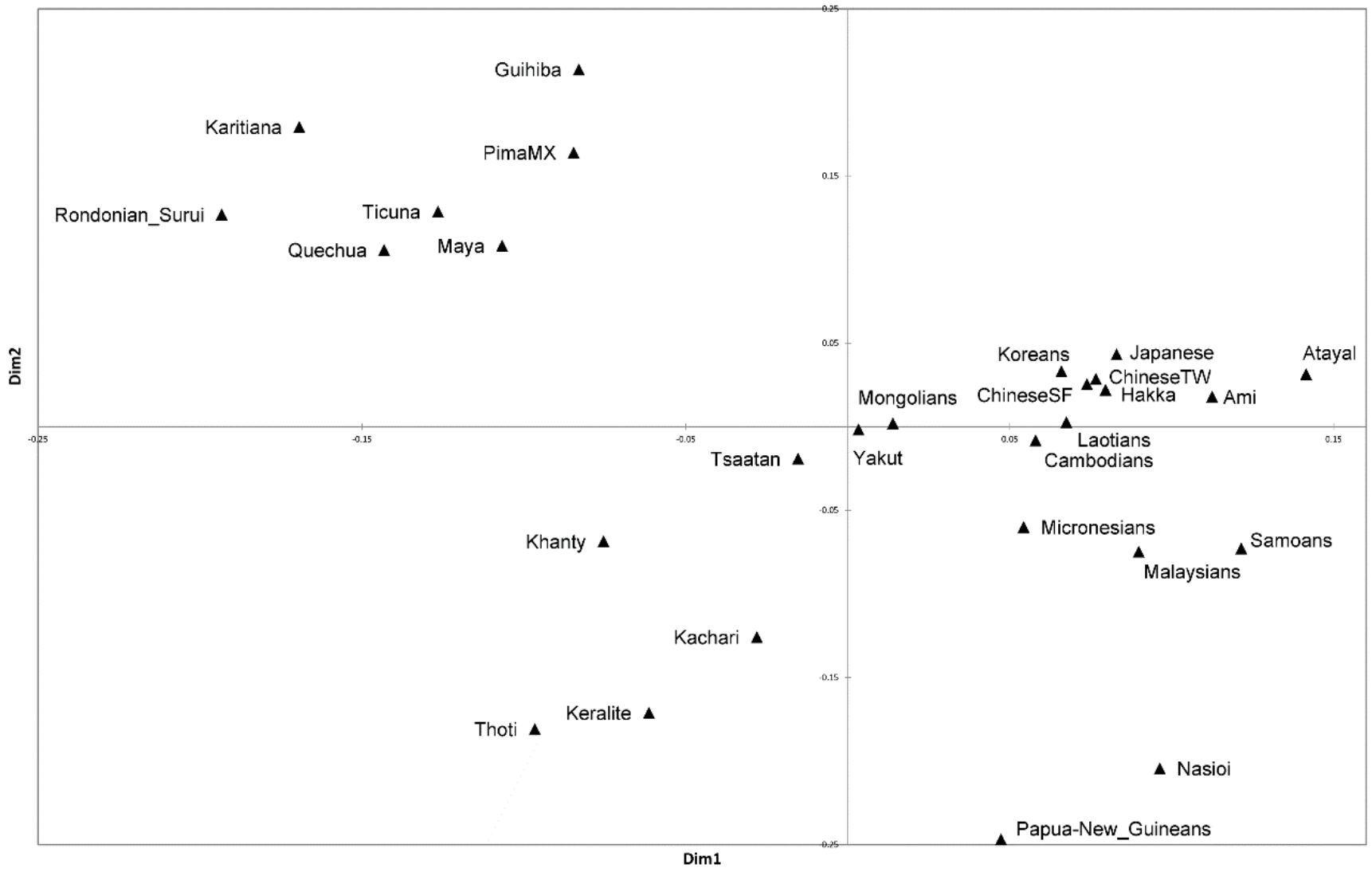


Figure 3

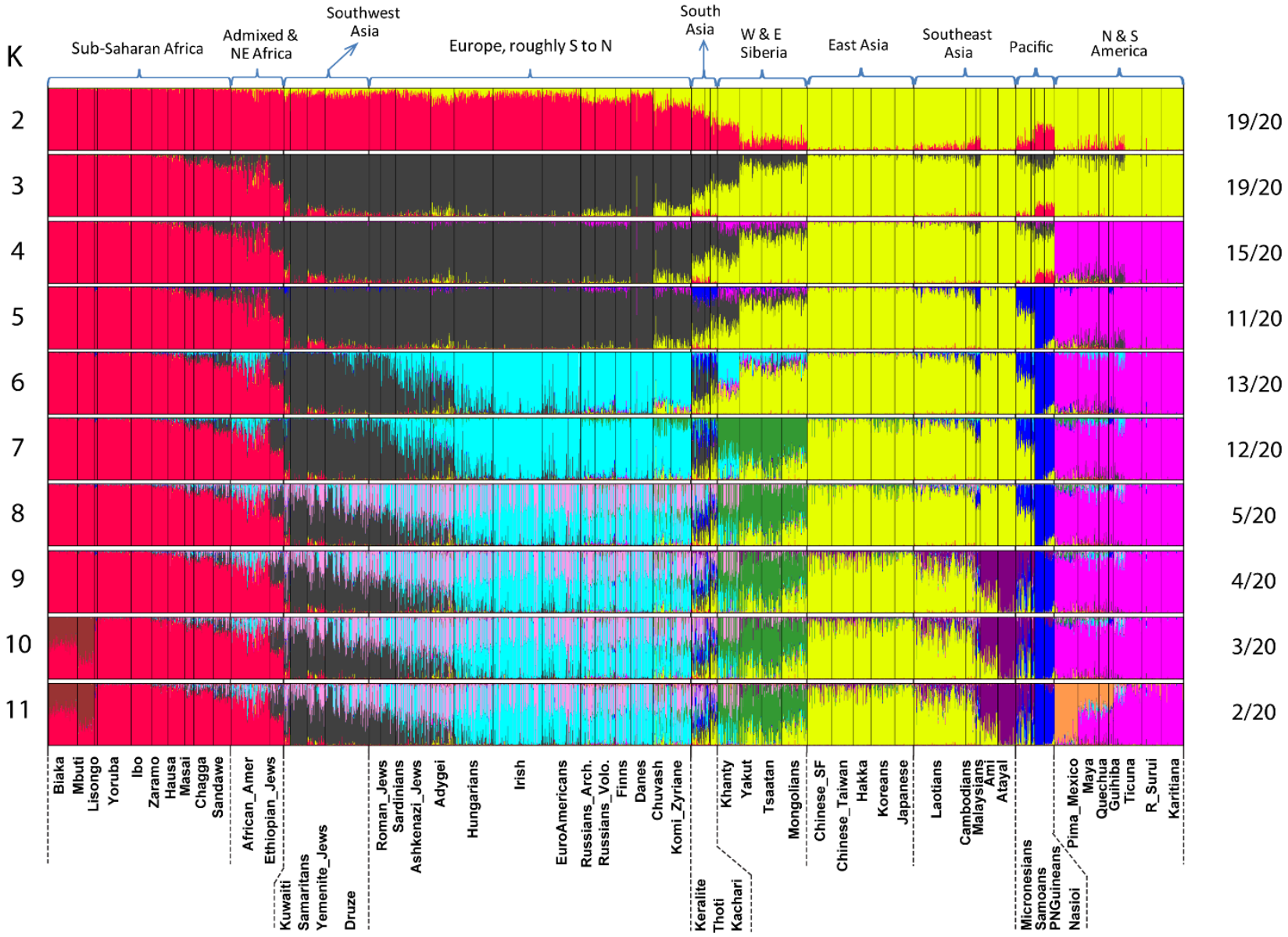


Figure 4.

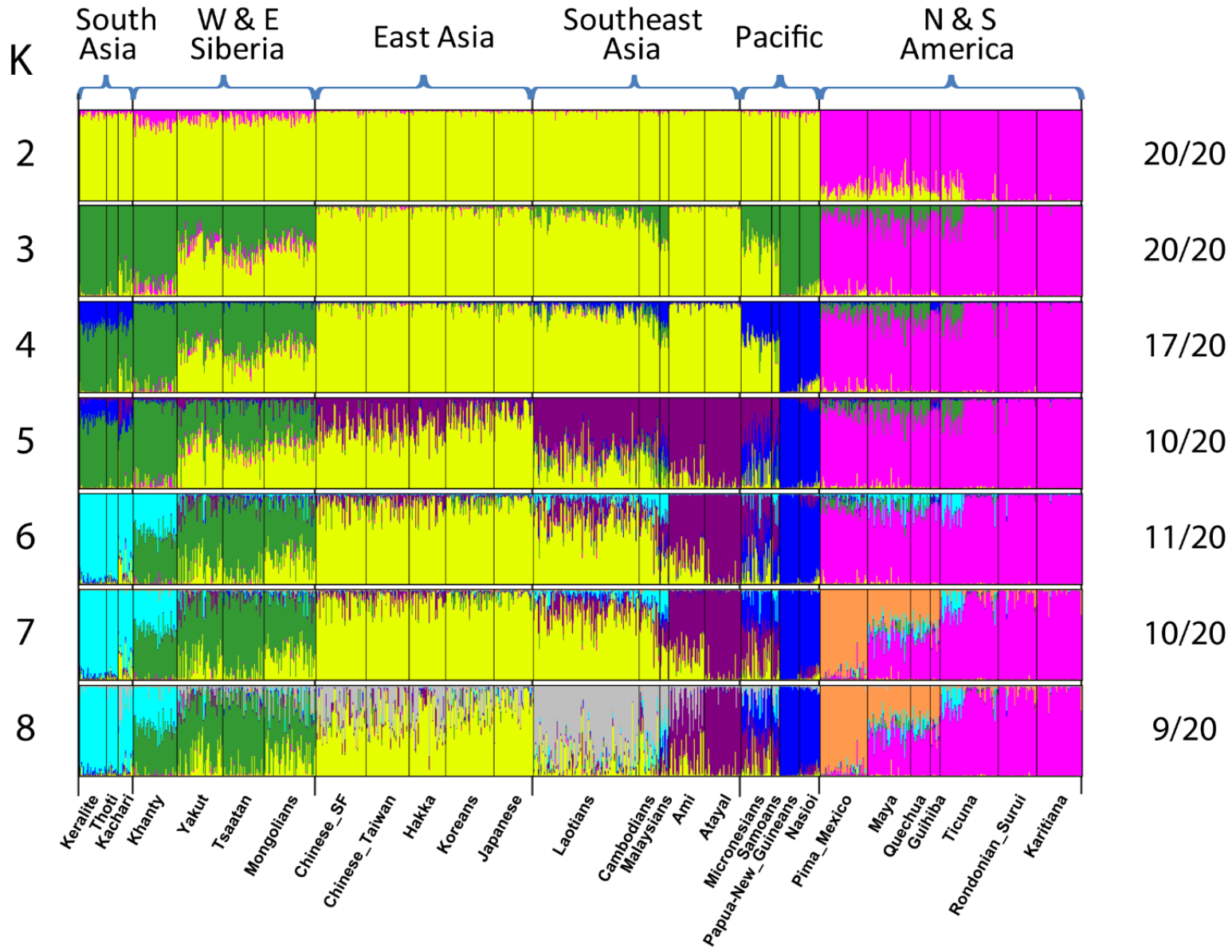
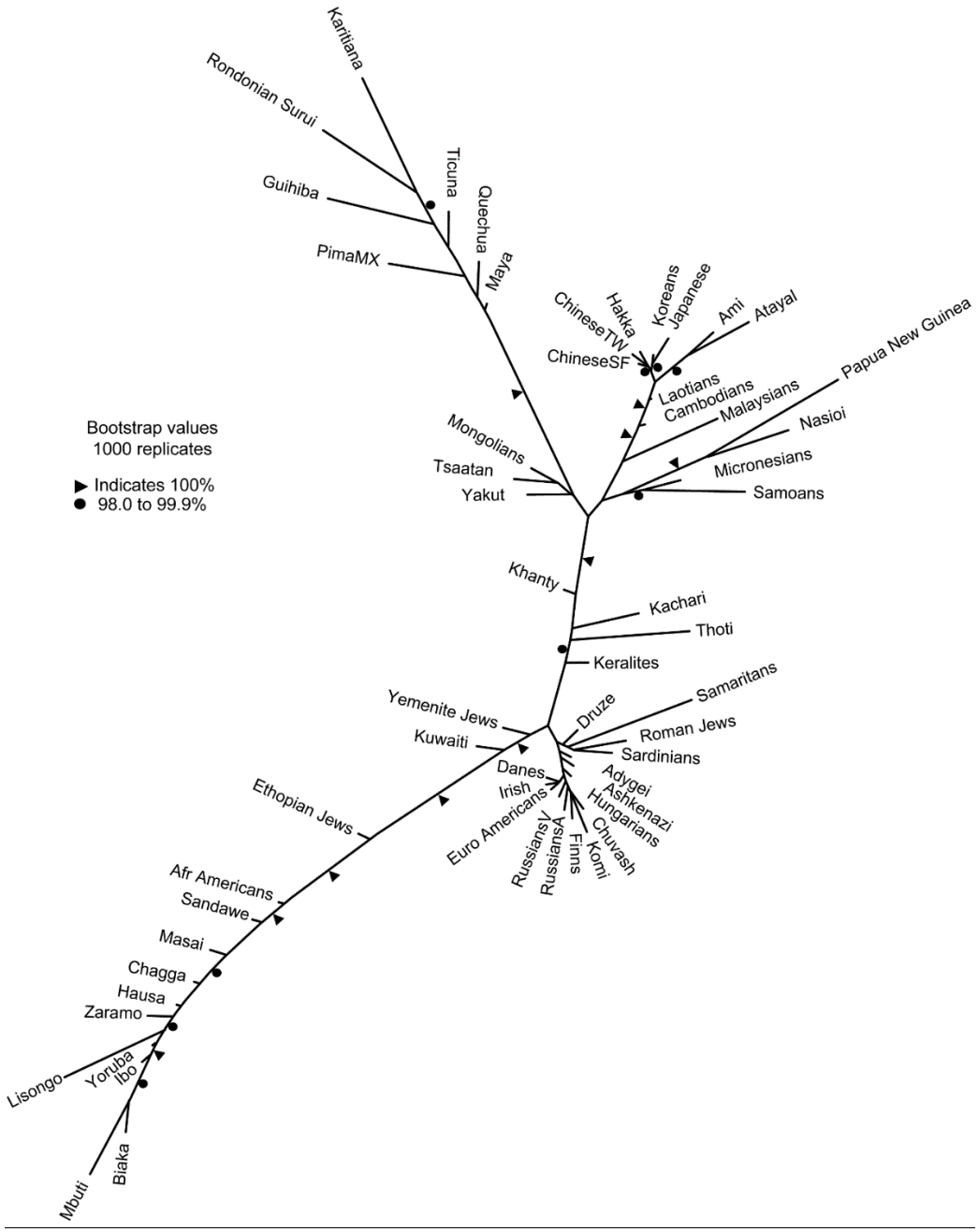


Figure 5.



Supplementary Figures **Figure S1a**

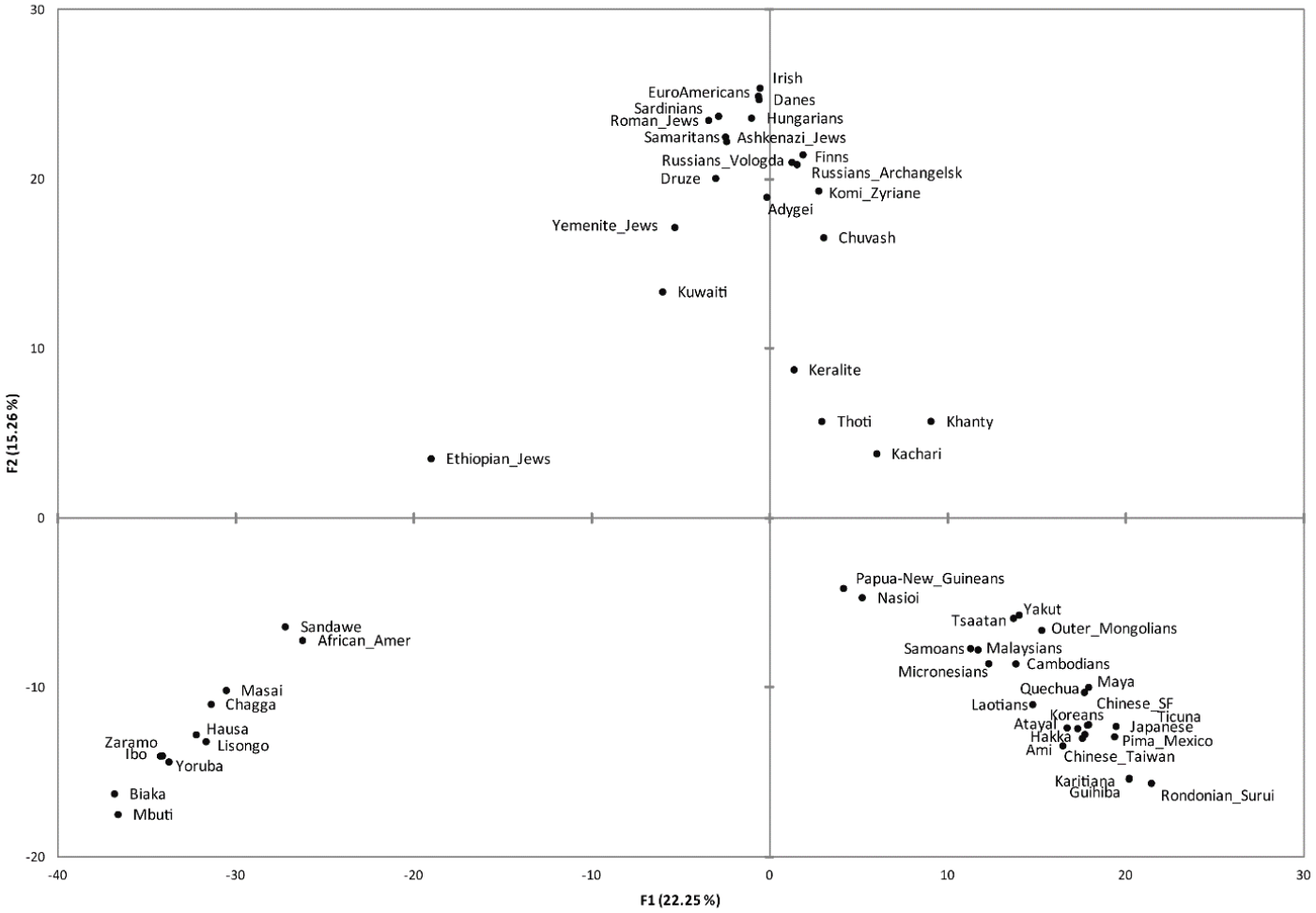


Figure S1b

Figure S2a

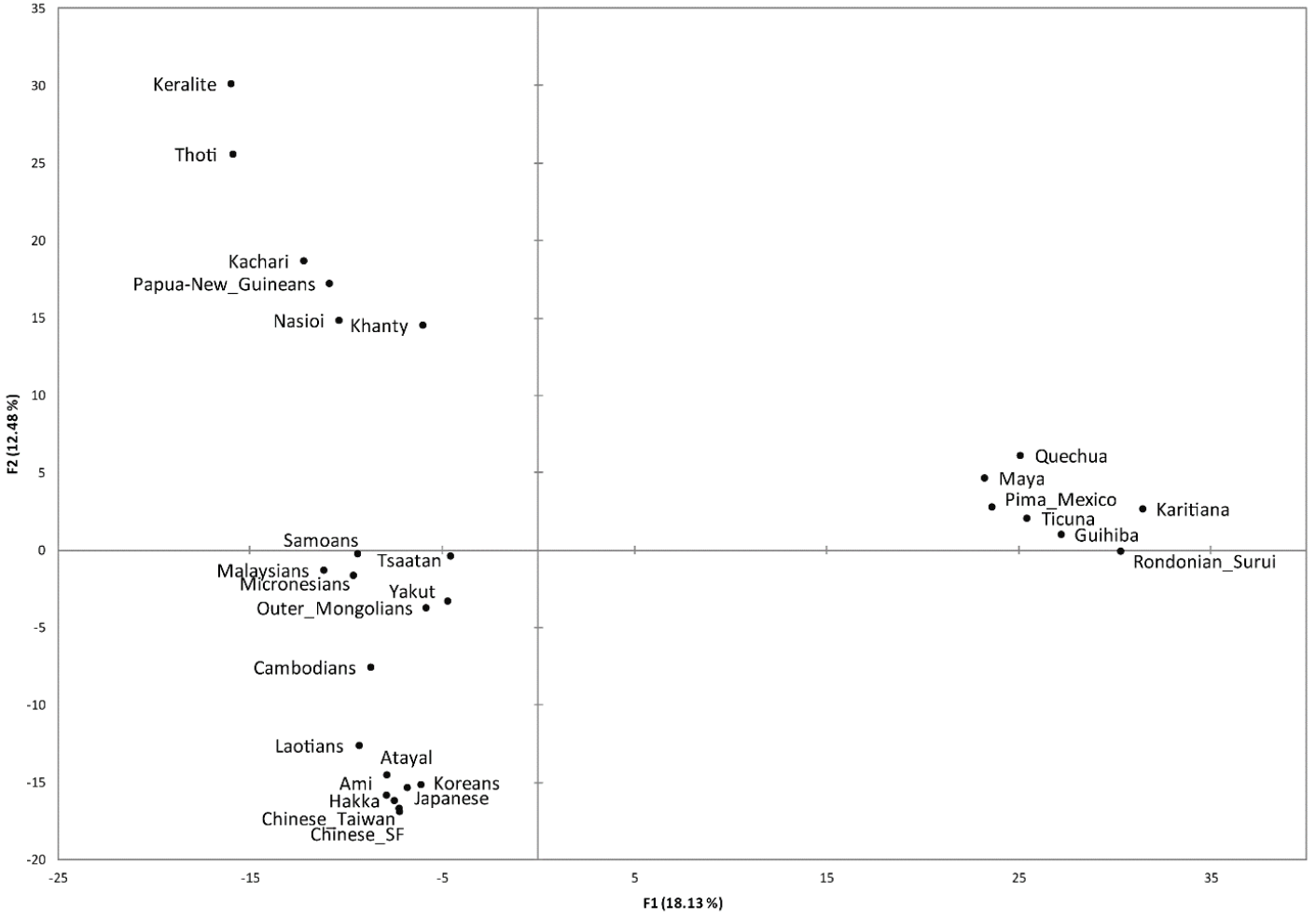


Figure S2b

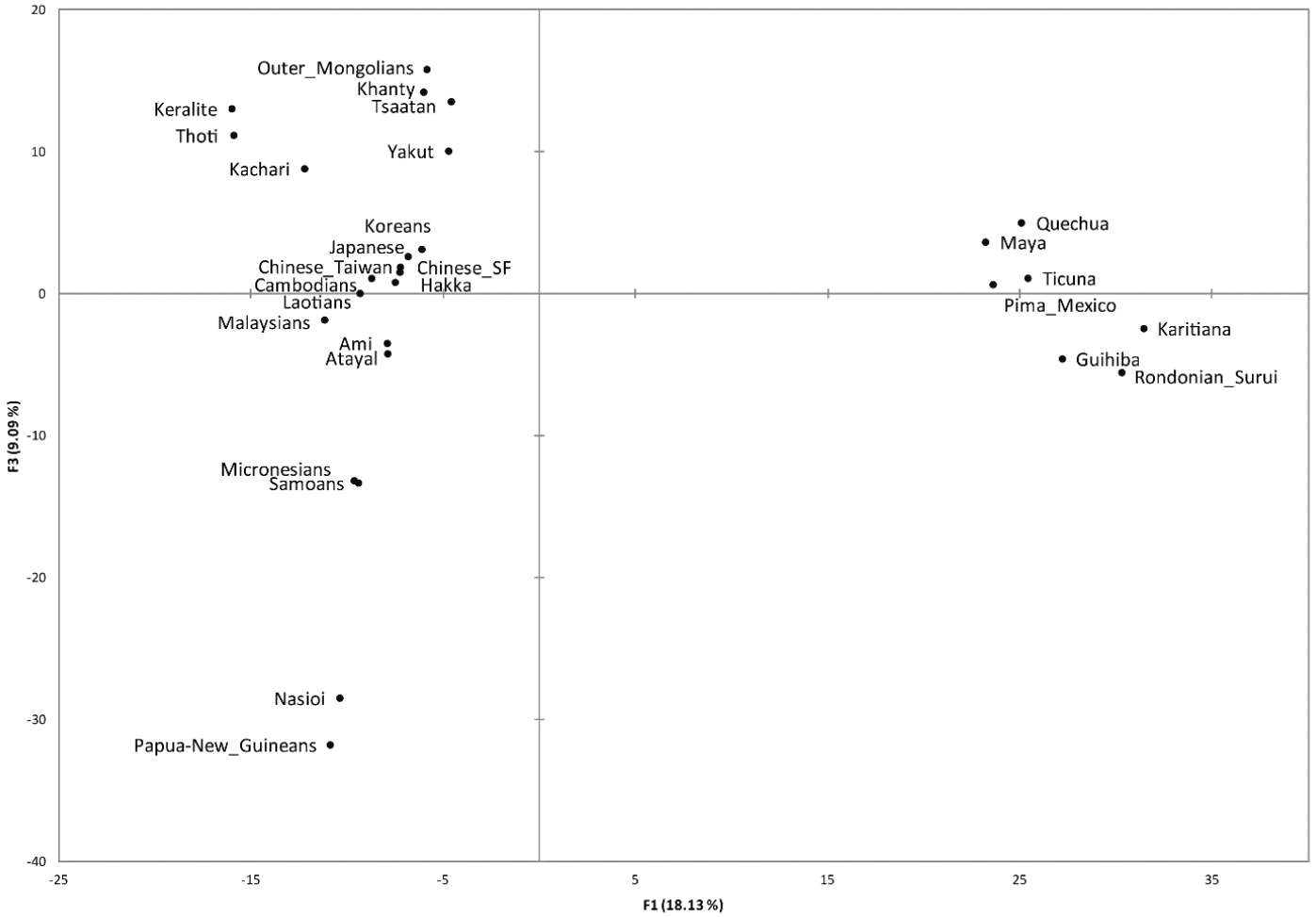


Figure S3

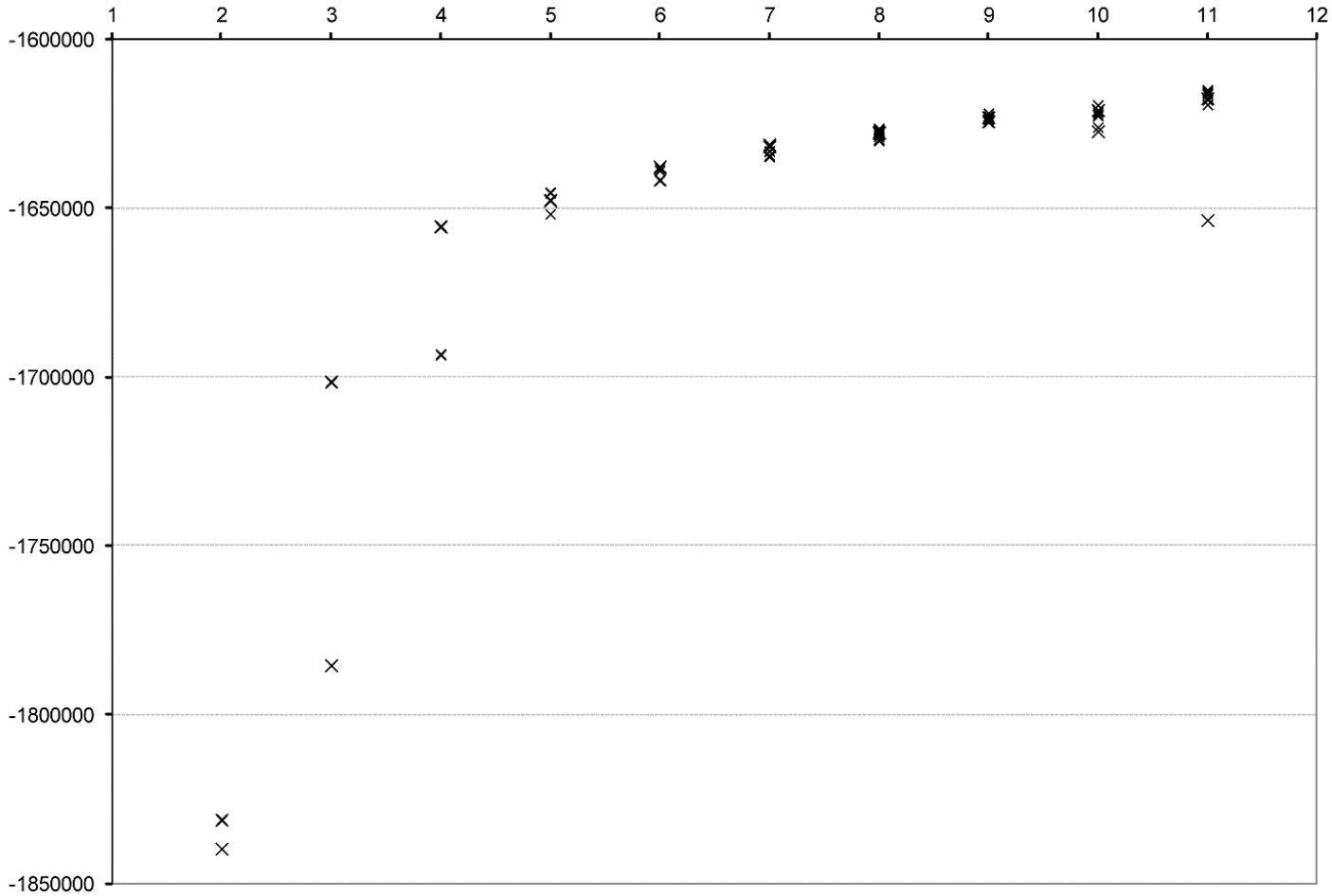


Figure S4