TEXT S1. SUPPLEMENTARY MATERIALS

Validation of the Admixture History Graph method In order to validate the AHG method, we generated in R populations with all the possible combinations of admixture proportions A, B and C such that in the populations reconstructed as they were at the time of admixture the initial ancestry component A was allowed to vary around 10%, 20%, 30%, 40% and 50% (B=1-A), and the later contribution of C similarly varied from around 10%, 20%, 30% etc to around 90%. For each of the different admixture histories 100,000 outcomes were generated. We then ran the covariance test described in the Materials and Methods section of the main text to determine the sequence of admixture events. For sample sizes of 10 individuals the sequence of admixture events was inferred correctly for all values of A (100% of outcomes with the correct history inferred), unless the initial proportion of A or C was less than 10%, or the initial proportion of C was over 90%, bringing the final proportion of A to below 5%. When the final proportion of ancestry A in a population dropped to below 5%, the performance of the method diminished slightly (Fig.S12A).

To determine if the test is useful for more complicated admixture scenarios we added another ancestry component (D) to the hypothetical populations already containing ancestries A, B, and C at various mixture proportions. We tested various possible mixture combinations that would allow for the final proportions of any ancestry to be above 5% (Fig.S12B). For sample sizes of 10, the admixture history is inferred correctly more than 95% of the time for a wide range of possible permissible values of A and C (0.2-0.8) and for values of D up to 70%. (Fig.S12B).

Additionally, we explored a scenario where one admixture event between two already admixed populations leads to the appearance of a population with four sources of ancestry ((AB)(CD)) (Fig.S12C). Again, for sample sizes of 10, the admixture history is inferred correctly more than 98% of the time for a wide range of possible permissible values of A and C and for different mixing ratios of AB and CD (Fig.S12C).

Validation of AHG results for South Siberian populations

Altaians

For the Altaians, the configuration most supported by the AHG calculation is (EUR, WestSIB) (CentralSIB, ASI), shown in Fig. 5A. This graph is supported by two out of four sets of covariance tests (see Main Text and Table S2), while the remaining sets of tests are inconclusive. Because the admixture history suggested by the AHG results is different for the Altaians and Tuvans, despite their apparent similarity in ancestry composition as inferred by ADMIXTURE, we tested the admixture tree supported by the AHG for the Altaians (EUR,WestSIB)(CentralSIB,ASI) against the tree suggested for the Tuvans (((EUR CentralSIB)WestSIB)ASI). The question here was: if we simulate the ancestral population under the (((AB)C)D) admixture model using as input parameters the means and variances in A, B, C and D components as observed in the Altaians, how likely are we to recover the (AB) (CD) graph using the AHG method? Based on the results of 100,000 simulations we estimate the probability of this error to be around 5%.

Tuvans

Because Tuvans have two out of four ancestries present at low proportion (namely Western Siberian and European), the covariance tests were largely inconclusive (Table S2), and we had to rely only on covariance tests which did not involve both of these ancestries simultaneously as part of the tested trio. Based on these more reliable tests, out of fifteen possible configurations only three were supported. These three configurations are: (((EUR WestSIB)CentralSIB)ASI), (((EUR CentralSIB)WestSIB)ASI), and (((WestSIB CentralSIB)EUR)ASI).

Nevertheless, analysis of ancestry block structure reveals that blocks of European and Central Siberian ancestry are the smallest and therefore are likely to be the oldest in Tuvans. However, there is clear evidence of recent additional Central Siberian gene flow into Tuvans, as in addition to small blocks they exhibit an abundance of wide blocks not seen in the Altaians (Fig.5B). On the other hand, the average width of Western Siberian ancestry blocks

is only slightly smaller than that of Asian ancestry blocks, and although this implies that the Asian ancestry is more recent (in agreement with all the AHG results), the difference is too small to distinguish in terms of inferred admixture dates (the date inferred for these two ancestries is ~1,980 ya). Given this additional evidence, we conclude that the configuration best supported for Tuvans is (((EUR CentralSIB)WestSIB)ASI), and it was this configuration that we tested against the graph inferred for Altaians (EUR, WestSIB)(CentralSIB, ASI). We tried to simulate the admixture history (AB)(CD) using as the input for A, B, C and D the sample means and variances for the EUR, CentralSIB, WestSIB, and ASI ancestry components observed in the Tuvans. However, as part of the simulation we generate the parameter space of all possible combinations of A, B, C and D available for the population at the time of admixture under a given admixture model and observed admixture proportions, and the Tuvan set of parameters was outside such a parameter space, i.e. the (AB)(CD) admixture model (at least in its simplest form without additional gene flow) inferred for Altaians is not possible for Tuvans. For example, if we assume that the Central Siberian and Asian ancestry we observe in Tuvans originated from the same source population (as in Altaians), then we can estimate the initial parameters in the admixing population. For Tuvans, let

mc=0.37 (observed mean frequency of Central Siberian ancestry)

md=0.39 (observed mean frequency of Asian ancestry)

ma=0.11 (observed mean frequency of European ancestry)

mb=0.08 (observed mean frequency of West Siberian ancestry)

(Note that the Tuvans also have 3% Far Eastern and 2% Eskimo ancestry components, which do not enter into the calculations). From these values we can first estimate the initial frequency of the Central Siberian ancestry in the initial admixing population:

$$mz = mc / (1-(ma+mb)).$$

In this example, the initial mean frequency of the component mc was 0.46. As mz is distributed in the interval (0,1) and follows a truncated gamma distribution, the maximal expected variance is ((1-mz)^2)/3 or 0.097. We can compare this value for variance with the one that would yield the observed variances in mc and md :

vc= 0.0005 (observed variance of Central Siberian ancestry)

vd=0.002 (observed variance of Asian ancestry)

vz = (1-2*mz) /((vd+md^2) / (vc+mc^2) -1) - mz^2

In this example, the value of vz we obtain is 0.46 and this value violates the restriction of 0.097 calculated above.

In addition, we can simulate the admixture history inferred for Tuvans ((AB)C)D) only if the A component assumes the mean and variance of the European component as observed in Tuvans (and B is calculated as 1-A), while if we express A through the Central Siberian component, this admixture graph becomes impossible. In fact the Central Siberian component with the mean and variance as observed in Tuvans is not possible under any simple model of 4-way admixture given the other parameters seen in this population. We therefore suggest this as complementary indirect evidence for more recent additional Central Siberian gene flow into this population.

Validation of the modifications to the admixture dating method based on wavelet transform analysis.

We constructed different sets of artificially admixed chromosome 1 from the phased genotypes of unrelated Yoruban (YRI) and Han Chinese (CHB) individuals and individuals of European ancestry (CEU), downloaded from the International HapMap project home page [1]. First, we simulated admixture between YRI and CEU haploid chromosomes. Admixed genomes were constructed as described previously [2]. Briefly, we started by building a recombination map by drawing from an exponential distribution with weight λ, such that an ancestry switch occurred with probability $1 - e^{-\lambda g}$ for each distance of g Morgans. Starting at the beginning of each simulated chromosome and at each of the recombination points from the recombination map, we sampled CEU ancestry with probability α and YRI ancestry with probability $1 - \alpha$, where the value of α was sampled from a beta distribution with mean 0.30 and standard deviation 0.10. The following values of λ were simulated: 80, 100, 120. We constructed 20 admixed chromosomes for each value of λ. These artificial chromosomes were then used as one of the parental populations for another simulated admixture event, involving

CHB. New hybrids were constructed exactly as above, but the values of λ simulated for this more recent admixture event were: 10, 20, 40, and 60. As before, 20 haploid chromosomes were simulated per population. Thus, in total 12 artificial data sets with different admixture histories were simulated (each recent admixture time point is simulated three times, each time using a different admixture background). To recover these 24 admixture dates (two dates of admixture for each of these artificial populations), we first inferred ancestry along the chromosomes using PCAdmix with the same settings as for the empirical data. Wavelet decomposition of the admixture signal as inferred by PCAdmix was performed following the workflow described for the empirical data. The results are reported in Fig.S15. λ is a parameter that reflects the time of admixture through the expected number of ancestry switches (or breakpoints) that occur given the rate of admixture and recombination [2]. In our previous study [2] we showed that for simulated data, the observed number of breakpoints starts to deviate from the expected value beginning around 50-100 generations since the admixture event (depending on the rate of admixture). Since for real populations our WT method takes this deviation into account, in these artificial datasets for higher values of λ (λ =80) the time since admixture tends to be overestimated (the λ is inferred correctly, but at this time depth the estimated number of breakpoints given this λ implies older admixture). On the other hand, the fact that admixture that took place more than 100 generations ago is underestimated underlines the limitations of the method: our resolution is circumscribed by the number of markers we have available for the analysis (for chromosome 1 we have ~ 25,000 SNPs distributed over 1024 windows).

As a further part of the validation scheme we inferred the dates of admixture in Russians as revealed by ADMIXTURE (Fig.1) using two different methods: the one described in this study, and the previously published StepPCO method [2]. Although both methods use wavelet transform analysis for the time of admixture estimation, the inference of ancestry along the chromosome is based on different approaches in these two methods. The method used here requires phased data (which potentially introduces phasing switch error [3,4]), and the PCAdmix algorithm [5], which uses HMM to assign ancestry for small non-overlapping windows across the genome. The second method uses the StepPCO approach (which does not require the data to be phased), and the windows taken across the genome are larger and overlapping; furthermore, the assignment of ancestry within each window is based on the PC1 coordinates of the SNPs within the window. In both instances we used Italians and Khanty as

proxies for the parental populations (chosen based on the results of ADMIXTURE, Fig.1). Based on the results of PCAdmix the admixture event in Russians is estimated to have occurred 870 years ago (CI: 810-1,020), and the StepPCO method yields a similar date of 750 years ago (CI: 600-960). It has been suggested previously that in the course of their ethnogenesis Russians assimilated and admixed with Finno-Ugric tribes of northeastern Europe [6,7], and the time we infer here with both methods (beginning to middle of the 13th century) is consistent with this potential source of the admixture signal [8].

Simulations of additional gene flow

To understand the effect on estimates of time of admixture in situations when the same ancestral population contributes to the admixed population multiple times, we first simulated admixture between YRI and CEU haploid chromosomes at 120, 100 and 80 generations ago (as described in the previous section). Then, on each of these artificial backgrounds one additional admixture event involving haploid chromosomes from CHB was introduced at 60 generations ago (as before), thus producing three sets of artificially admixed populations. For each of these three sets we then introduced additional gene flow from CEU haploid chromosomes. The values of λ simulated for this additional admixture event were: 20, 30 and 40. Furthermore, for each of these time points we tested scenarios involving four different rates of admixture, with α being 0.1, 0.2, 0.3 or 0.4. This produced 36 different admixture histories. As before, to recover the dates of admixture (two dates for each of these admixture histories), we first inferred ancestry along the chromosomes using PCAdmix with the same settings as for the empirical data. Wavelet decomposition of the admixture signal as inferred by PCAdmix was performed following the workflow described for the empirical data. The results are reported in Fig.S8As expected, additional gene flow lowers the estimated dates of admixture, as it introduces new wider ancestry blocks into the population, thereby altering older narrower block structure and erasing information on past admixture events. Higher rates of additional gene flow amplify this effect, and the underestimates are most severe for scenarios involving the most recent gene flow (λ =20), and the highest admixture rates (α = 0.4). Furthermore, the confidence intervals for the dates inferred for the earlier vs. later event of admixture overlap considerably. This means that in some situations an earlier admixture event might be inferred to be younger than the more recent admixture event.

Addressing other concerns with dates of admixture inference

We assessed the possibility of overestimating the dates of admixture due to potential errors in ancestry estimation by PCAdmix. Indeed, if small errors are being introduced during ancestry estimation (with small blocks of "inappropriate" ancestry being inserted within long continuous blocks), such errors would cause the signal of admixture to appear older. We calculated the widths of ancestry blocks, as inferred by PCAdmix, in Anabar Dolgans and scrutinized the genomewide distribution of block widths for each individual and each ancestry. The results are summarized in Fig.S16. We do not observe an excess of short ancestral blocks, and hence find no evidence for erroneous insertions by PCAdmix.

In addition, for every individual and every window, PCAdmix outputs posterior probabilities obtained from the HMM of assignment to a particular ancestry. To further explore if errors in PCAdmix ancestry estimation affect our dates of admixture inference, in another set of experiments we performed admixture dating using only those PCAdmix windows where the ancestry was inferred with high confidence (at least 0.8 probability in at least 80% of samples). Comparison of dates of admixture inferred using a full set of windows vs. only high confidence windows (Fig.S17A) reveals no tendency to overestimate admixture dates when all the data are being used, and again argues against PCAdmix being the source of systematic error in our dates of admixture inference.

Since it is known that haplotype phasing at the level of the entire chromosome is subject to phasing (switch) errors [3,4], we wanted to assess to what extent such potential switch errors might affect our results. For two-way admixture we can directly compare results obtained from phased and unphased data (see the above discussion of similar dates obtained for admixture in Russians using two different methods). For admixture scenarios involving more than two sources of ancestry, the direct comparison is impossible, since the StepPCO approach (which uses unphased data) cannot handle more than two parental populations and PCAdmix requires phased data. Therefore, to test the effect of phasing on admixture dating in a population with more than two sources of ancestry, we re-phased part of the data using a different phasing algorithm, SHAPE-it [9]. SHAPE-it has been shown to have a lower error rate than BEAGLE [10], which was the phasing algorithm employed for this study. The

expectation here was that if phasing switch error has a strong influence on the inferred dates of admixture, then a different switch error rate should result in a different date of admixture estimate. Our results show that dating a three-way admixture event using BEAGLE-phased vs. SHAPE-it-phased data yields almost identical dates (Fig.S17B).

In another set of experiments we took the PCAdmix output and "unphased" it by collapsing each pair of windows corresponding to two phased chromosomes of each individual into one, and then running wavelet transform analysis on such "unphased" chromosomes. Although this is not the same as using unphased data (as "unphasing" is done on windows, not on genotypes), we still expect the small ancestry blocks potentially introduced by switch error to disappear and their effect on dates of admixture estimation to be removed. For example, if the ancestry inferred by PCAdmix in six consecutive windows on two phased parental chromosomes is 1 2 1 2 0 1 and 2 1 2 1 0 1 (ancestry from parental populations 0, 1 or 2), then our "collapsed" chromosome will have the following assignment for these windows: 3 3 3 3 0 1 (3 coding a window with ancestry assigned to both parental populations 1 and 2). It is easy to see that such "unphasing" converts four potentially erroneous small ancestry blocks into one longer ancestry block. If we indeed have such erroneously inferred small blocks of ancestry in our data, "unphasing" should result in a more recent admixture date. This however is not the case, and the dates of admixture inferred using phased and "unphased" data were almost identical (Fig.S17A). In summary, phasing switch error does not appear to have a strong influence on our results.

TEXT S2. SUPPLEMENTARY NOTE ON HISTORY AND ARCHAEOLOGICAL RECORD

South Siberia (Altaians, Tuvans, and Buryats)

The earliest archaeological monuments in the Altai region that postdate the Last Glacial Maximum belong to the Afanasevo culture, which flourished from around 5,000 years ago (ya) [11]; morphologically the skeletal remains associated with this culture show affinities with Europeans [12]. Around 3,500-4,000 ya the Afanasevo culture was succeeded or replaced by the Okunev culture [13,14], which is thought to have been brought to the Altai region by immigrating peoples [13,15]. The date of 3,630 ya inferred here for the admixture between the European and West Siberian substrates in Altaians (Fig.5B) might reflect this prehistoric migration. Around Lake Baikal, to the east of the Afanasevo and Okunev cultures and contemporary with them, thrived an independent archaeological culture called Serovo-Glazkovo (dated to 6,200-3,000 ya; Weber2010). Analysis of ancient mtDNA lineages from Serovo-Glazkovo sites of the Baikal region show affinities with modern-day South and Central Siberian and East Asian populations, but not with western Eurasians [16]. Yet artifacts from Okunev and Serovo-Glazkovo archaeological sites show reciprocal influence of both traditions on each other, and it is not surprising that this cultural contact would have been accompanied by genetic contact as well [17]. The archaeological record of Mongolia for the same time period attests to the emergence of a new grave culture, known as the Slab Grave culture. This cultural tradition persisted in Mongolia from 3,100 to 2,300 ya, and is attributed to the ancestors of the Xiongnu people (also known as Asian Huns), who around 2,000 ya created a powerful empire which for three centuries dominated central and eastern Asia and extended from eastern Kazakhstan to western Manchuria [18]. There are notable similarities between the Bronze Age cultures of the Altai region (Okunev culture and its successors) and the Slab Grave culture of Mongolia, thus suggesting extensive cross-regional contacts between Altai, Mongolia and Lake Baikal ([19] and references therein). Nuclear and mitochondrial DNA extracted from the skeletal remains contained in a 2,000 year old Xiongnu necropolis in northern Mongolia [20] show affinities to modern-day South Siberian populations and Mongolians [21,22]. And although the dates of around 2,000 ya estimated for the appearance of the Asian signal in Buryats and Tuvans perfectly correspond to the time

period of the Xiongnu empire, we suggest, based on the distribution of the ancestral Asian block widths in populations of South Siberia (Fig.5C), that Buryats must have received additional gene flow from some East Eurasian source. This additional gene flow is likely to have occurred 800-600 ya, in the times of the Mongol empire under Genghis Khan and his successors, which stretched far beyond the modern borders of Mongolia. Hence, we suggest that the dates of Asian admixture estimated for Buryats are composite and reflect: an older admixture event of around 2,700 ya, as observed in Altaians and as consistent with the dates for the appearance of the Slab Grave Culture in Mongolia associated with the ancestors of the Xiongnu: and a more recent admixture of around 800-600 ya, consistent with the Mongol expansion under Genghis Khan. Furthermore, mtDNA analysis of skeletal remains from Bronze Age archaeological sites in the eastern and western Altai shows that 5,000-3,000 ya mtDNA lineages in the eastern (Russian) Altai were exclusively of western Eurasian origin, while the mtDNA lineages in the western (Mongolian) Altai were exclusively eastern Eurasian. In contrast, all analysed skeletal remains from the Iron Age (2,800-1,800 ya) sites from the eastern and western Altai as well as from Kazakhstan (Pazyryk culture, traditionally associated with the Eastern Scythians and also with the Tarim mummies of western China [23]), reveal the presence of both western and eastern Eurasian mtDNA haplogroups [22], in agreement with the AHG inferred for Altaians. Since the peoples with western and eastern Eurasian ancestries co-existed in the Altai for a prolonged period of time [22], it is plausible that the admixture date of ~2,300 ya estimated for the Altaians captures admixture between these local groups and is not associated with any additional migration from the outside.

Tungusic-speaking populations

Archaeological sites along the Amur are numerous, and the earliest are dated to at least 8,000 ya. Interestingly, the oldest dates of ca. 3,100 ya for the Far Eastern and Central Siberian admixture inferred by our analysis correspond to the latest Neolithic material culture of the region, which is dated to 4,000-3,000 ya and is associated with the appearance of fundamentally new stone-working techniques, which are ascribed to the arrival of new peoples [24]. The Asian component dated to around ~1,860 ya could be attributed to the expansion of the Xiongnu people (see above), as although the Xiongnu empire was centred

around modern-day Mongolia, it stretched far eastward to the Amur River tributaries and Western Manchuria [25].

Samoyedic populations.

The sequence of admixture events and the dates of admixture estimated for the Selkup and Nenets (Fig.S11A) are relatively close to those obtained for Tuvans (see Fig.5A). Although the dates for the prehistoric mixture between the European and the Central Siberian components are more recent for the Samoyedic populations (~2500-2700 ya) than the dates estimated for Tuvans (~3100 ya), this is most likely due to later additional gene flow into the Selkup and Nenets from some population of predominantly Central Siberian ancestry, as discussed in the main text, and also due to probable additional recent admixture with Russians (Fig.S11B). On the other hand, the Asian component we observe in Tuvans (and Altaians) is not seen in any of the Samoyedic groups, and its arrival in the Sayan plateau (and the Altai region) predates the West Siberian admixture estimated for the Samoyeds. Our analyses therefore suggest the following history for the Selkup and Nenets: it is likely that modern-day Selkup, Nenets, and Tuvans trace their ancestry to the same source population, and either modern-day Tuvans migrated to the Sayan plateau after their split from the ancestors of modern-day Selkup and Nenets, or the Sayan plateau itself was the ancestral homeland of these populations. Before later migrations brought Asian genes into this area of South Siberia, the ancestors of the Samoyeds started to expand north (e.g. it is thought that the northward migration of the Kulai people of the southeastern part of Western Siberia, which some suggest were the ancestors of the modern-day Samoyedic populations, took place approximately 1500 ya [26]). Later admixture with some population of primarily West Siberian ancestry would have added further West Siberian ancestry to the genepool of the Samoyedic groups, resulting in a younger date for the West Siberian gene flow than that estimated for the Tuvans. The date for this mixture in the Selkup and Nenets of 1,590 ya coincides with what is known as the Great Migration Period, which took place in Eurasia 1,700-1,200 years ago. This was characterized by massive movements of people across the continent and could have been precipitated by climate deterioration [27] or the rise of the Hunnic Empire under Attila the Hun and the invasion of Europe by the Huns of Central Asia [18]; its impact on the genetic landscape of Europe and Asia was profound [18]. The Nenets subsequently experienced

additional gene flow from some population of Central Siberian ancestry (such as the Nganasan), which probably happened not before 1000 ya, after the Nenets settled on the Yamal and Taimyr [28]. Blocks of European ancestry that are on average wider in Nenets than anywhere else in Siberia, and a larger variance around the mean, also suggest additional later gene flow, probably from Russians.

Populations of the Russian Far East: Koryaks, Chukchi and Naukan Yupik

The presence of tribes of hunters and fishers on the Kamchatka and Chukotka Peninsulas is documented archaeologically from around 13,000 ya [29,30], yet cultural traits that can be attributed to the putative ancestors of Chukchi and Koryaks do not appear in the archaeological record until around 3,000 ya [31]. The origin of the Naukan Yupik is still widely debated. While Paleo-Eskimo populations migrated into the Bering Sea region from Siberia around 5,500 ya [32,33], the Naukan Yupik are the descendants of the pre-Inuit Neo-Eskimo populations which have been recently shown to have a migration history independent of Paleo-Eskimos [34].While some authors suggest that the Naukan Yupik are a group that never crossed the Bering Strait, but who persisted in Siberia (e.g. [31]), others think that their presence in Chukotka is the result of a back-migration from North America [34–37]. Based on the typology of elements of the harpoon complex, the distinctive features of the Eskimo culture are not seen on Chukotka until around 2,200 ya and the earliest Eskimo sites along the coasts are dated to around 2,000 ya [38]. Radiocarbon dates, however, put the earliest Eskimo cultures on Chukotka (Old Bering and Okvik cultures) at ca. 1700-1500 ya [39]. Our results support an origin of the Naukan Yupik in North America and their subsequent back-migration to Siberia. Furthermore, the dates of reciprocal admixture estimated for the Naukan Yupik and the Chukchi of ca. 1,500 ya are in good agreement with the time of the first appearance of the Neo-Eskimo culture in Siberia.

General comments

For many of the Siberian populations, we can identify and distinguish signals of admixture events shared between populations or population subgroups (e.g. the signal of European admixture shared by all the Turkic-speaking groups) from local differential gene flow (e.g. the

recent Yakut admixture which affected some, but not all, Even subgroups). We observe what might be the genetic impact of the powerful Xiongnu empire of ca. 2,000 years ago on the surrounding populations. The Xiongnu empire was centred around today's Mongolia, but stretched far beyond its current borders. The early dates of Asian admixture that we infer for the populations of the Altai and Sayan Mountains (Altaians and Tuvans) as well as the populations which trace their origin to lake Baikal (Buryats, Yakuts, and ultimately Dolgans) of this powerful empire, rather than reflecting the impact of the much more recent Mongol empire under Gengis Khan and his successors. Similar early Asian admixture that might reflect Xiongnu influence is also observed in all the Tungusic-speaking populations; the fact that this ancestry is shared by all the Tungusic peoples suggests that the gene flow must have happened before Evens and Evenks (and Oroqen) diverged and expanded across Siberia. In contrast, the Mongol empire founded by Gengis Khan, which lasted from 1206 to1368 , which was the largest empire in the world and stretched from central Europe to the Sea of Japan and from southern Siberia to India and Indochina, appears to have had a genetic impact only on the Buryats - a population in immediate proximity to Mongolia. This is consistent with the idea that the Mongol empire was maintained via appointed local viceroys with little if any physical presence of the conquerors. In fact, Russian territories (Rus') were in subjection to the Golden Horde (part of the Mongol empire and its successors) for nearly 300 years, yet no evidence of this was detected in Y chromosome analyses of Russians [6], and the absence of this signal is confirmed here with dense autosomal data (see above for the analyses and discussion of the admixture in Russians). Interestingly, it was suggested that the expansion of the Yakuts from South Siberia [21,40] was precipitated by the advancing armies of Genghis Khan [41]. Our results are consistent with this hypothesis: unlike the Turkic-speaking populations of southern Siberia (as well as the Mongolic-speaking Buryats), the Turkic speakers of northern Siberia show a signal of earlier Asian gene flow (possibly associated with the Xiongnu), but not of the later Asian gene flow that could be associated with the Mongol expansion under Genghis Khan. The same pattern is observed in all the Evenk and Even subgroups. Since some researchers suggest that the separation of the Evenks and Evens from the South Tungusic-speakers (e.g. Hezhen) occurred only ca. 800 years ago [42], it is possible that the invading Mongolian armies also forced the Tungusic expansion north and across Siberia (as the vast territorries south and north of the Amur river were conquered and became part of the Mongol empire).

References

1. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007 Oct 18;449(7164):851–61.

2. Pugach I, Matveyev R, Wollstein A, Kayser M, Stoneking M. Dating the age of admixture via wavelet transform analysis of genome-wide data. Genome Biol. 2011 Feb 25;12(2):R19.

3. Tang H, Coram M, Wang P, Zhu X, Risch N. Reconstructing genetic ancestry blocks in admixed individuals. Am J Hum Genet. 2006 Jul;79(1):1–12.

4. Andrés AM, Clark AG, Shimmin L, Boerwinkle E, Sing CF, Hixson JE. Understanding the accuracy of statistical haplotype inference with sequence data of known phase. Genet Epidemiol. 2007 Nov;31(7):659–71.

5. Brisbin A, Bryc K, Byrnes J, Zakharia F, Omberg L, Degenhardt J, et al. PCAdmix: Principal Components-Based Assignment of Ancestry Along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. Hum Biol. 2012 Aug 1;84(4):343–64.

6. Balanovsky O, Rootsi S, Pshenichnov A, Kivisild T, Churnosov M, Evseeva I, et al. Two Sources of the Russian Patrilineal Heritage in Their Eurasian Context. Am J Hum Genet. 2008 Oct 1;82(1):236–50.

7. Khrunin AV, Khokhrin DV, Filippova IN, Esko T, Nelis M, Bebyakova NA, et al. A genome-wide analysis of populations from European Russia reveals a new pole of genetic diversity in northern Europe. PloS One. 2013;8(3):e58552.

8. Alekseeva TI, editor. Vostochnye slavjane: antropologicheskaja i etnicheskaja istorija [Eastern Slavs: anthropological and ethnic history]. Moscow: Nauchnyj mir; 2002.

9. Delaneau O, Coulonges C, Zagury J-F. Shape-IT: new rapid and accurate algorithm for haplotype inference. BMC Bioinformatics. 2008;9:540.

10. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007 Nov;81(5):1084–97.

11. Naumov IV. The History of Siberia. Taylor & Francis; 2006. 257 p.

12. Solodovnikov KN. Materialy k antropologii afanasievskoj kultury [Materials to anthropology of Afanasevo culture]. Drevnosti Altaja Antiq Altai. 2003;10:3–27.

13. Savinov DG, Podolskiy ML, editors. Okunevskij Sbornik: Kultura i ee okruzhenie 2 [Okunev collection: Culture and its surroundings 2]. Saint Petersburg: SPbGU [Publisher of the Saint Petersburg State University]; 2006.

14. Sokolova LA. Okunev cultural tradition in the stratigraphic aspect. Archaeol Ethnol Anthropol Eurasia. 2007;30(1):41–51.

15. The Cambridge History of Early Inner Asia [Internet]. Cambridge University Press. [cited 2015 Apr 16]. Available from: http://www.cambridge.org/us/academic/subjects/history/middle-east-history/cambridge-history-early-inner-asia

16. Mooder KP, Schurr TG, Bamforth FJ, Bazaliiski VI, Savel'ev NA. Population affinities of Neolithic Siberians: a snapshot from prehistoric Lake Baikal. Am J Phys Anthropol. 2006 Mar;129(3):349–61.

17. Nikolaev RV. Prehistoric cultural links along the Yenisei: Revelations from a bronze idol. In: Vajda EJ, editor. Languages and Prehistory of Central Siberia. John Benjamins Publishing Company; 2004. p. 257–62.

18. Kim HJ. The Huns, Rome and the Birth of Europe. Cambridge : New York: Cambridge University Press; 2013. 348 p.

19. Tumen D. Anthropology of Archaeological populations from Inner Asia. Mong J Anthropol Ethnol. 2008;4(1(312)):162–83.

20. Keyser-Tracqui C, Crubezy E, Ludes B. Nuclear and Mitochondrial DNA Analysis of a 2,000-Year-Old Necropolis in the Egyin Gol Valley of Mongolia. Am J Hum Genet. 2003 Aug;73(2):247–60.

21. Crubézy E, Amory S, Keyser C, Bouakaze C, Bodner M, Gibert M, et al. Human evolution in Siberia: from frozen bodies to ancient DNA. BMC Evol Biol. 2010 Jan 25;10(1):25.

22. González-Ruiz M, Santos C, Jordana X, Simón M, Lalueza-Fox C, Gigli E, et al. Tracing the Origin of the East-West Population Admixture in the Altai Region (Central Asia). PLoS ONE. 2012 Nov 9;7(11):e48904.

23. Li C, Li H, Cui Y, Xie C, Cai D, Li W, et al. Evidence that a West-East admixed population lived in the Tarim Basin as early as the early Bronze Age. BMC Biol. 2010 Feb 17;8:15.

24. Derevianko AP, Clark GH. New Data on the Archaeology of the Middle Amur. Arct Anthropol. 1965 Jan 1;3(1):136–41.

25. Barfield TJ. The shadow empires: imperial state formation along the Chinese-Nomad frontier. In: Alcock S, D'Altroy TN, Morrison KD, Sinopoli CM, editors. Empires: Perspectives from Archaeology and History. Cambridge: Cambridge University Press; 2001. p. 10–41.

26. Pletneva LM, editor. Smena kultur i migraciii v Zapadnoj Sibiri [Replacement of cultures and migrations in Western Siberia] [Internet]. Tomsk: Izdatelstvo TGU (Tomsk State University Publisher); 1987 [cited 2015 Jan 16]. Available from: http://arheologija.ru/smena-kultur-i-migratsii-v-zapadnoy-sibiri/

27. Larsen LB, Vinther BM, Briffa KR, Melvin TM, Clausen HB, Jones PD, et al. New ice core evidence for a volcanic cause of the A.D. 536 dust veil. Geophys Res Lett. 2008 Feb 1;35(4):L04708.

28. Vasiliev VI. K probleme etnogeneza severosamodijskih narodov [On the question of ethnogenesis of the Samoyedic peoples of the North]. In: Gurvich IS, editor. Socialnaja organizacija i kultura narodov severa [Social organization and culture of the peoples of the North]. Moscow: Nauka; 1974. p. 133–75.

29. Vasil'ev SA, Kuzmin YV, Orlova LA, Dementiev VN. Radiocarbon-based chronology of the Paleolithic in Siberia and its relevance to the peopling of the New World. Radiocarbon. 2002;44(2):503–30.

30. Goebel T, Slobodin SB, Waters MR. New dates from Ushki-1, Kamchatka, confirm 13,000 cal BP age for earliest Paleolithic occupation. J Archaeol Sci. 2010 Oct;37(10):2640–9.

31. Dikov NN, Clark GH. The Stone Age of Kamchatka and the Chukchi Peninsula in the Light of New Archaeological Data. Arct Anthropol. 1965 Jan 1;3(1):10–25.

32. Derenko M, Malyarchuk B, Grzybowski T, Denisova G, Dambueva I, Perkova M, et al. Phylogeographic analysis of mitochondrial DNA in northern Asian populations. Am J Hum Genet. 2007 Nov;81(5):1025–41.

33. Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature. 2010 Feb 11;463(7282):757–62.

34. Raghavan M, DeGiorgio M, Albrechtsen A, Moltke I, Skoglund P, Korneliussen TS, et al. The genetic prehistory of the New World Arctic. Science. 2014 Aug 29;345(6200):1255832.

35. Tamm E, Kivisild T, Reidla M, Metspalu M, Smith DG, Mulligan CJ, et al. Beringian Standstill and Spread of Native American Founders. PLoS ONE. 2007 Sep 5;2(9):e829.

36. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, et al. Reconstructing Native American population history. Nature. 2012 Aug 16;488(7411):370–4.

37. Achilli A, Perego UA, Lancioni H, Olivieri A, Gandini F, Hooshiar Kashani B, et al. Reconciling migration models to the Americas with the variation of North American native mitogenomes. Proc Natl Acad Sci U S A. 2013 Aug 27;110(35):14308–13.

38. Dikov NN. Arheologicheskije pamjatniki Kamchatki, Chukotki i verhnej Kolymy [Archaeological Sites of Kamchatka, Chukotka and the Upper Kolyma]. Moscow: Nauka; 1977.

39. Rainey F, Ralph E. Radiocarbon Dating in the Arctic. Am Antiq. 1959 Apr 1;24(4):365–74.

40. Tokarev SA, Gurvich IS. Yakuty [Yakuts]. In: Levin MG, Potapov LP, editors. Narody Sibiri [Peoples of Siberia]. Moscow: Publisher of the Academy of Sciences of the USSR; 1956.

41. Pakendorf B, Novgorodov IN, Osakovskij VL, Danilova AP, Protod'jakonov AP, Stoneking M. Investigating the effects of prehistoric migrations in Siberia: genetic variation and the origins of Yakuts. Hum Genet. 2006 Oct;120(3):334–53.

42. Tugolukov VA. Ėtničeskie korni tungusov [Ethnic roots of the Tungus people]. In: Gurvich IS, editor. Ėtnogenez narodov severa [Ethnogenesis of the peoples of the North]. Moscow: Nauka; 1980. p. 152–76.

Table S1 **Populations included in the study**

Populations are grouped by linguistic affiliation

population	subgroup	N	subsistence	provenance	latitude and longitude coords
Turkic					
Altaians		12	pastoralist (horse/cattle)	Rasmussen et al 2010	51.57N 85.59E (Gorno-Altaisk)
Tuvans		16	pastoralist (horse/cattle)	Rasmussen et al 2010	51.43N 94.27E (Kyzyl)
Yakuts			pastoralist (horse/cattle)		
	CEPH	24		Li et al 2008	62.98N 129.50E
	Verhoyansk	8		this study	68.40N 136.14E
	not specified	13		Reich et al 2012	
Dolgans			H-G/ reindeer herder		
	Taimyr	10		this study	69.21N 86.13E
	Anabar	10		this study	72.48N 113.14E
Tungusic-Manchu					
N.Tungusic					
Evens			H-G/ reindeer herder	this study	
	Sakkyryyr	7			67.48N 130.24E
	Sebjan	10			64.24N 126.16E
	Tompo	10			64.39N 135.57E
	Berezovka	6			67.30N 154.40E
	Kamchatka	5			55.55N 158.41E (Esso)
Evenks			H-G/ reindeer herder	this study	
	Taimyr	5			68.18N

					88.45E (Khantayskoye Lake)
	Stony-Tunguska	6			61.47N 93.35E
Oroqen		10		Li et al 2008	50.43N 126.50E
S.Tungusic					
Hezhen		9		Li et al 2008	47.50N 133.50E
Manchu					
Xibo		9		Li et al 2008	43.50N 81.50E
Mongolic					
Mongolians		10	pastoralist (horse/cattle)	Li et al 2008	45N 111E
Buryats		19	pastoralist (horse/cattle)	Rasmussen et al 2010	51.50N 107.36E
Daur		9		Li et al 2008	48.50N 124.00E
Tu Mongolians		10		Li et al 2008	36N 101E
Uralic					
N.Samoyedic					
Nganasan			H-G		
		10		Rasmussen et al 2010	71.60N 92.49E (Ust-Avam)
		14		Reich et al 2012	n/a
Nenets			reindeer herder		
	Taimyr	6		this study	69.21N 86.13E
	Yamal	6		this study	69.42N 70.36E
	not specified	3		Reich et al 2012	n/a
S.Samoyedic					
Selkup		9	H-G/ reindeer herder	Rasmussen et al 2010	65.42N 82.27E (Krasnoselkup)
Ugric					
Khanty		35		Reich et al	61N

				2012	69E (Khanty-Mansiysk, Ugry)
Chukotko-Kamchatkan					
Chukchi			H-G/ reindeer herder		
		9		Rasmussen et al 2010	64.44N 177.30E (Anadyr)
		19		Reich et al 2012	n/a
Koryaks			H-G/ reindeer herder	Rasmussen et al 2010	55.55N 158.41E (Palana)
Eskimo-Aleut					
Yupik					
Naukan Yupik		16		Reich et al 2012	64.30N -173.01E (Novoe Chaplino)
Inuit					
Greenland				Reich et al 2012	
	east	7			65.38N -37.40E (Tasiliak)
	west	8			64.10N -51.43E (Nuuk)
Aleut					
Aleuts		8		Reich et al 2012	53.40N -166.38E (Unalaska island)
isolate language					
Yukaghirs		13		Reich et al 2012	68.52N 161.27E (upper Kolyma River)
Yeniseic					
Kets		2		Rasmussen et al 2010	62.29N 86.17E (Kellog)
Indo-European					

Slavic					
Russians		25		Li et al 2008	61N 40E
Romance					
French		29		Li et al 2008	46N 2E
Italian				Li et al 2008	
	Bergamo	13			46N 10E
	Tuscan	8			43.48N 11.14E
Sino-Tibetan					
Han Chinese		44		Li et al 2008	32.27N 114.00E
Japonic					
Japanese		28		Li et al 2008	38N 138E

Table S2. **Results of the AHG test**

	A/B,C	B/C,A	A/C,B
Tungusic speakers:			
Evenks_StonyTunguska	0.0143	-0.0168	-0.0026
Evenks_Taimyr	0.01	-0.0091	0.0007
Evens_Sakkyryyr	-0.1127	0.1285	0.0159
Evens_Tompo	-0.044	0.0494	0.0054
Evens_Sebjan	-0.1473	0.1676	0.0203
Eastern Evens (Evens_Kamchatka and Evens Berezovka)	-0.0016	-0.0072	-0.0088
Oroqen	0.001	0.2375	0.2386
Hezhen	-0.0864	0.3882	0.3018
	A/B,C	B/C,A	A/C,B
Turkic speakers:			
Yakuts_CEPH	-0.002	0.0076	0.0055
Yakuts_Verhoyansk	0.0026	-0.0111	-0.0084
Dolgans_Anabar	-0.0304	0.0054	-0.025
Dolgans_Taimyr	-0.0302	0.0077	-0.0224

	A/B,C	B/C,A	A/C,B									
Samoyedic speakers												
Selkup and Nenets_Yamal	-0.0065	-0.1461	-0.1526									
	A/B,C	B/C,A	A/C,B	A/B,D	B/D,A	A/D,B	A/D,C	D/C,A	A/C,D	B/C,D	D/C,B	D/B,C
South Siberia												
Buryats	0.0008	0.0019	0.0027									
Tuvans	0.0009	0.006	0.007	0.0026	-0.00221	0.0004	0.008	0.0082	0.0162	0.0136	-0.0065	0.0071
Altaians	0.0042	-0.007	-0.0027	-0.0161	0.0037	-0.0125	-0.0131	-0.0106	0.0025	0.0186	-0.0098	0.0088

* Eastern Evens are pooled because of small sample sizes

To work out the order of admixture events in populations with more than three sources of ancestry, we consider each trio of ancestries separately. For example, for the Altaians we first try to resolve the ordering of A,B and C ancestries. However the results of the three covariance tests for all possible configurations of A,B and C are inconclusive, hence this ordering cannot be resolved. The next trio we consider is A,B and D. Here, the smallest covariance is observed between the ratio of B/D and A, we can therefore conclude that A is the most recent of these three ancestries. The next trio is A,D and C. For this trio the smallest absolute value of covariance is observed between A/C and D, hence A ancestry must have mixed with C ancestry before mixing with D ancestry. The covariance tests for the last trio B,C and D are again inconclusive, yet the only possible configuration for the entire admixture graph given these results is (BD)(AC).

Figure S1

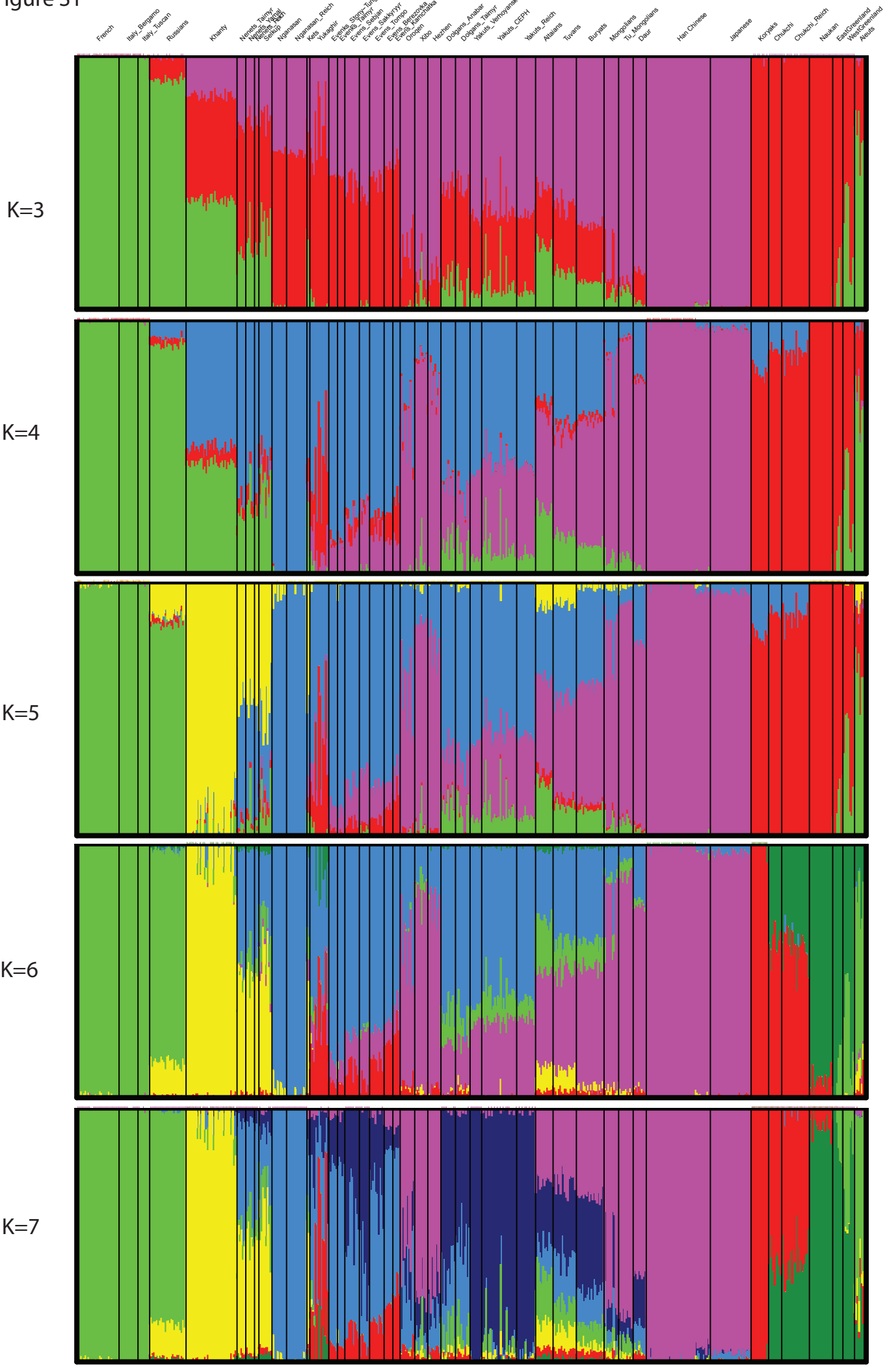


FIGURE S1. Results of the ADMIXTURE analysis for K = 3 through K = 7.

FIGURE S2

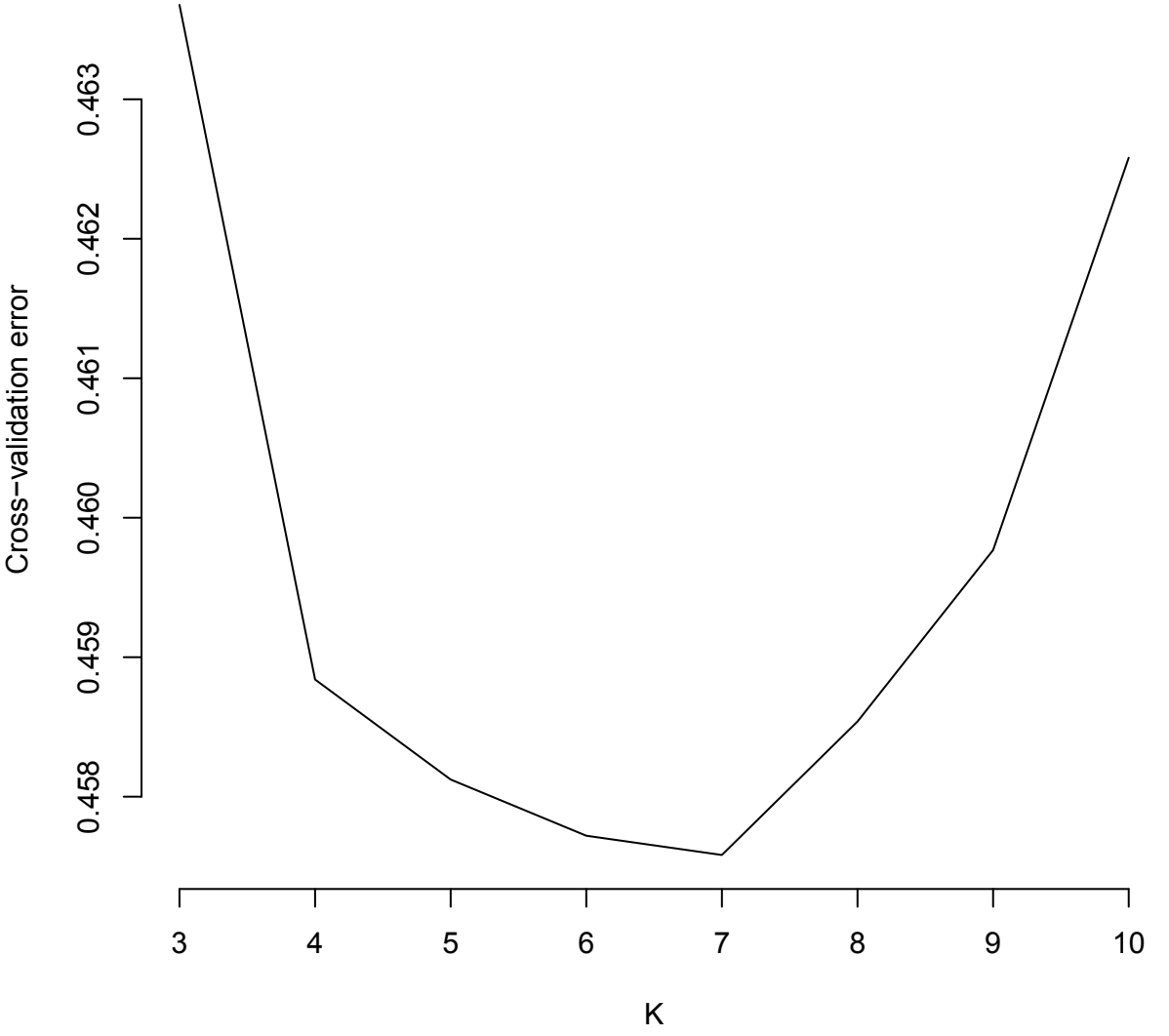


FIGURE S2. Estimated cross-validation error for the ADMIXTURE runs for K=3 through K=10.

Figure S3

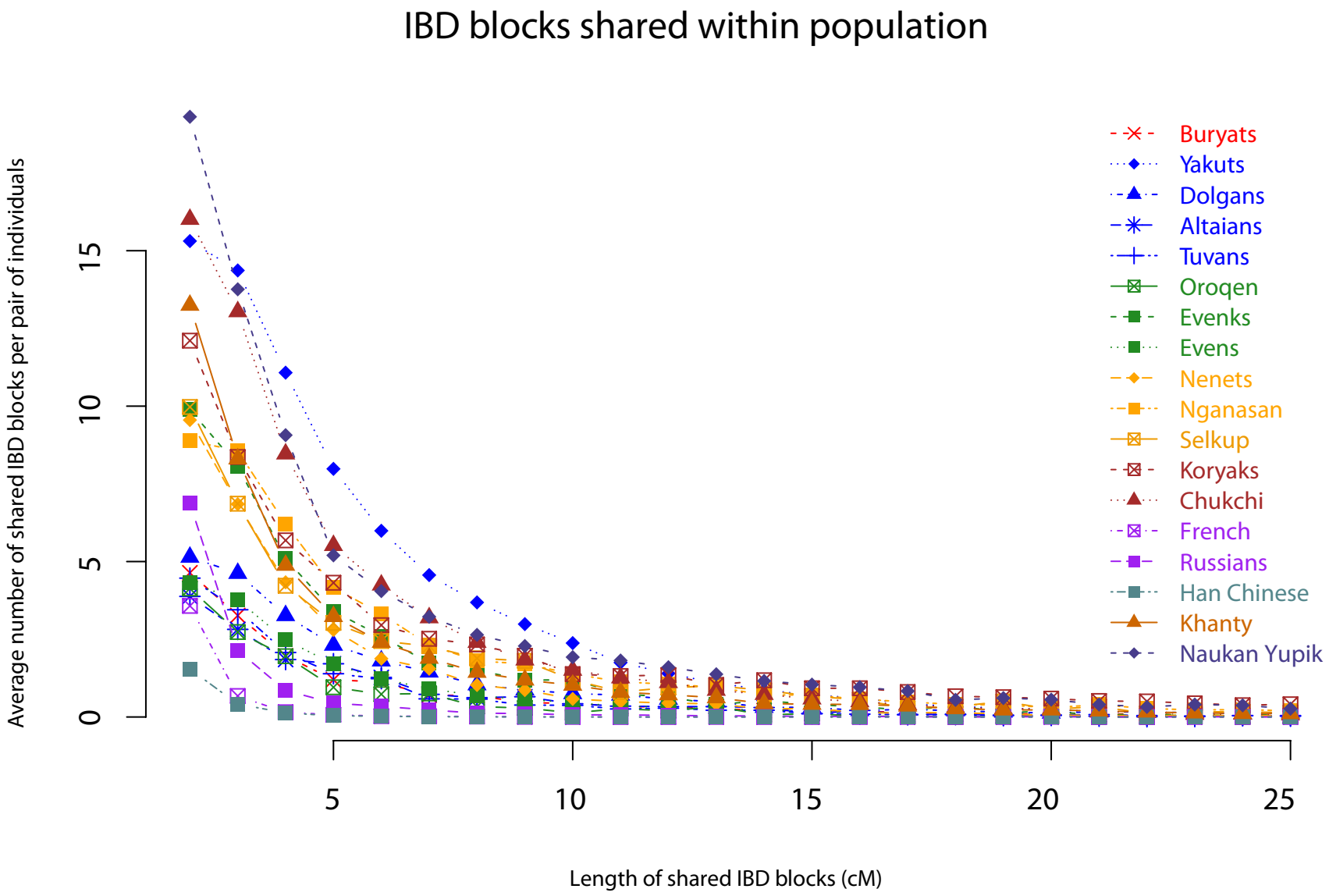


FIGURE S3. Analysis of IBD blocks shared within populations. The average number of IBD blocks shared per pair of individuals belonging to the same population is plotted against the genetic length of such blocks (blocks smaller than 2 cM were excluded from the analysis). Populations are coloured according to their linguistic affiliation, and the colouring scheme is described in the legend to Figure 2 of the Main Text.

Figure S4

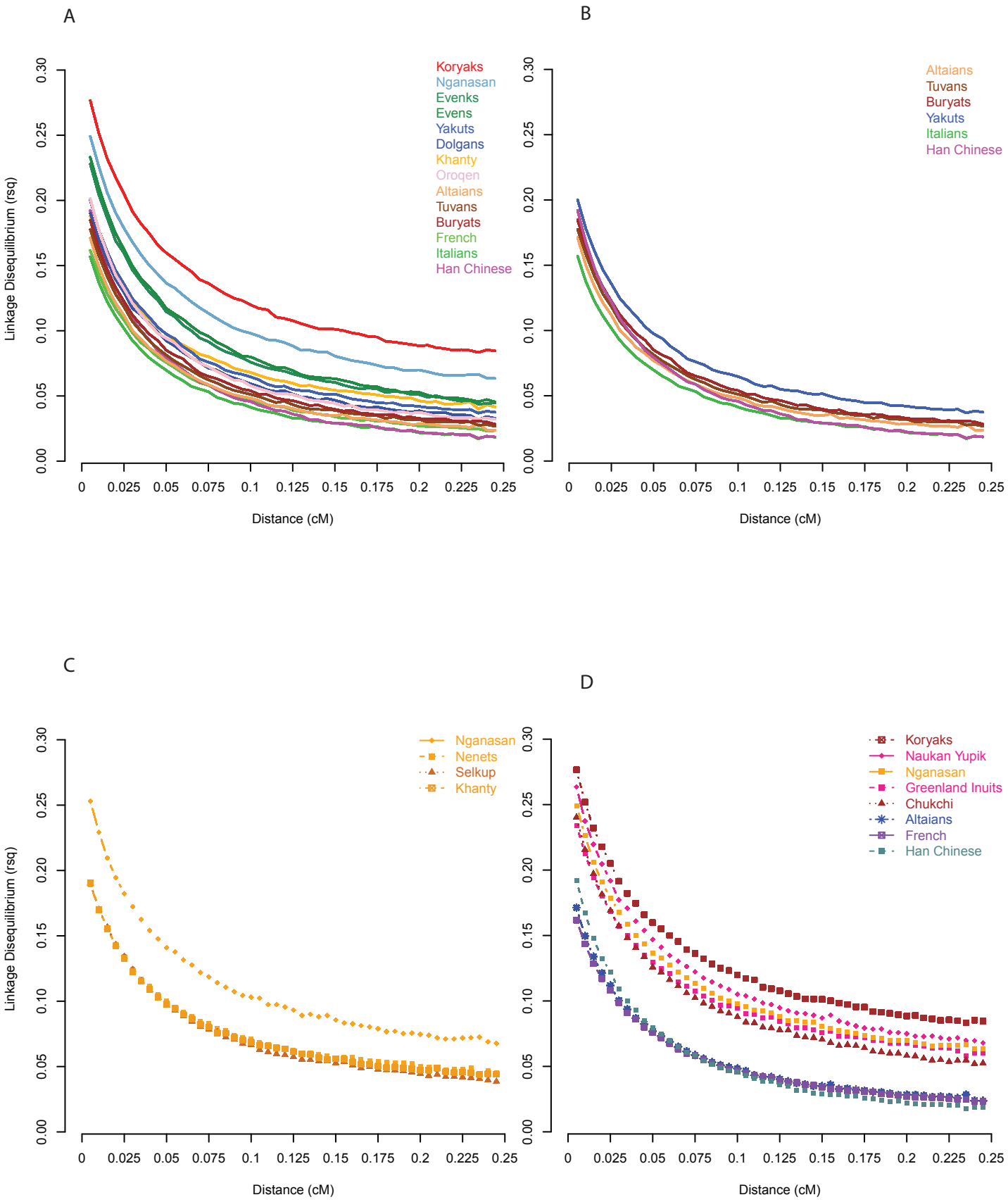


FIGURE S4. LD patterns in populations of Siberia. Linkage disequilibrium (LD) is measured for each population and each pair of SNPs within 50 evenly spaced recombination distance categories. The shortest genetic distances between the SNPs are represented on the left and the largest genetic distances on the right. (A) LD patterns in a representative subset of populations. (B) LD pattern in the Yakuts in comparison to Europeans, East Asians and populations of South Siberia. (C) LD pattern in the Nganasan in comparison to the other Samoyedic-speakers and the Ugric-speaking Khanty. (D) LD pattern in the populations that live on the margins of the Siberian landmass in comparison to Europeans, East Asians and Altaians of South Siberia.

FIGURE S5

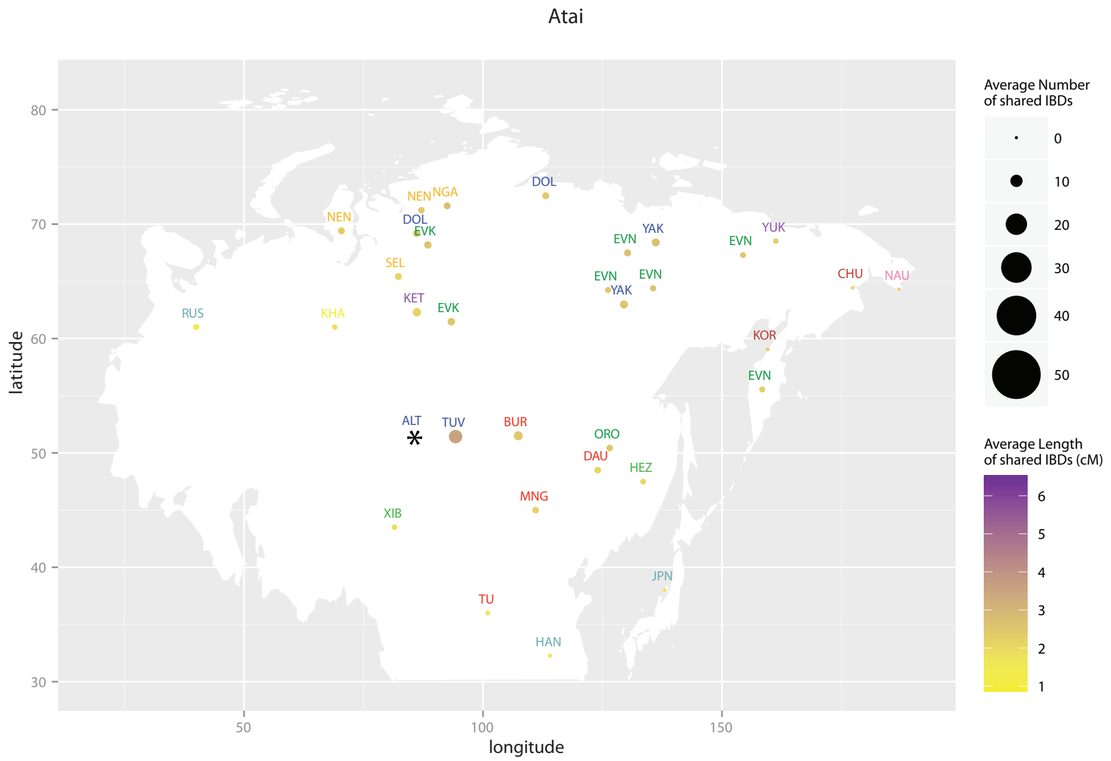
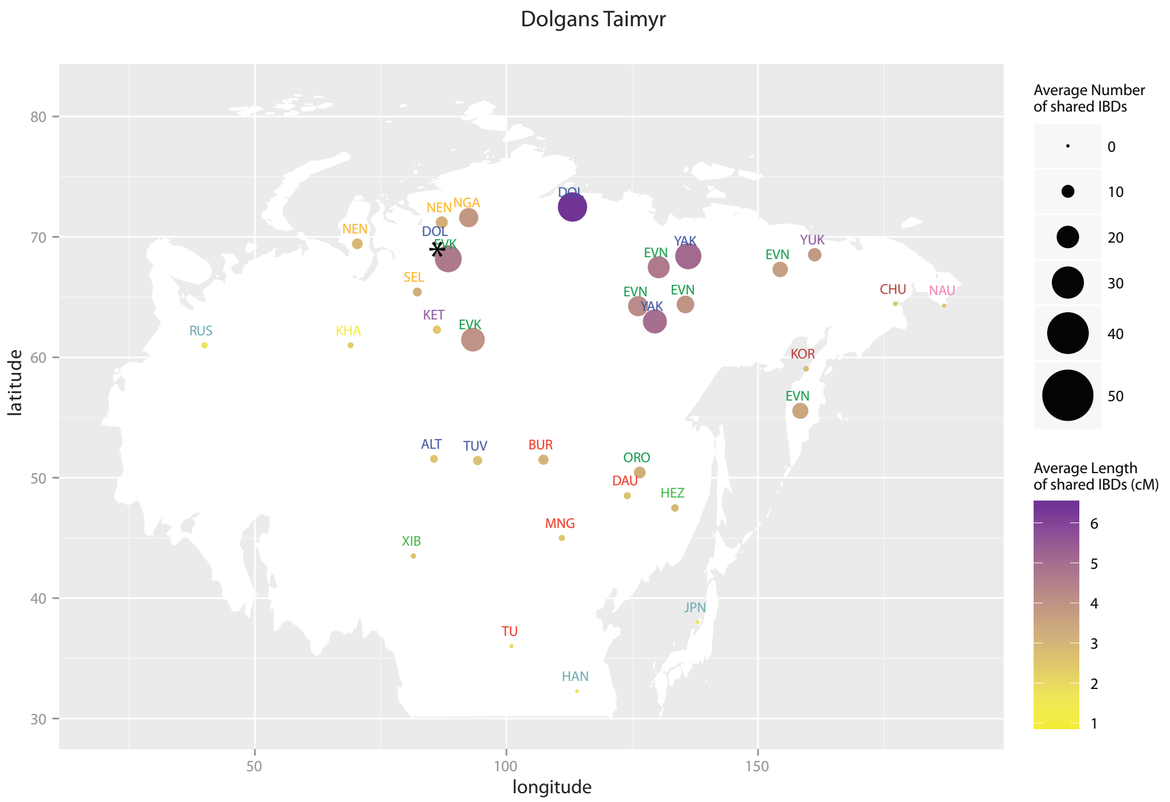
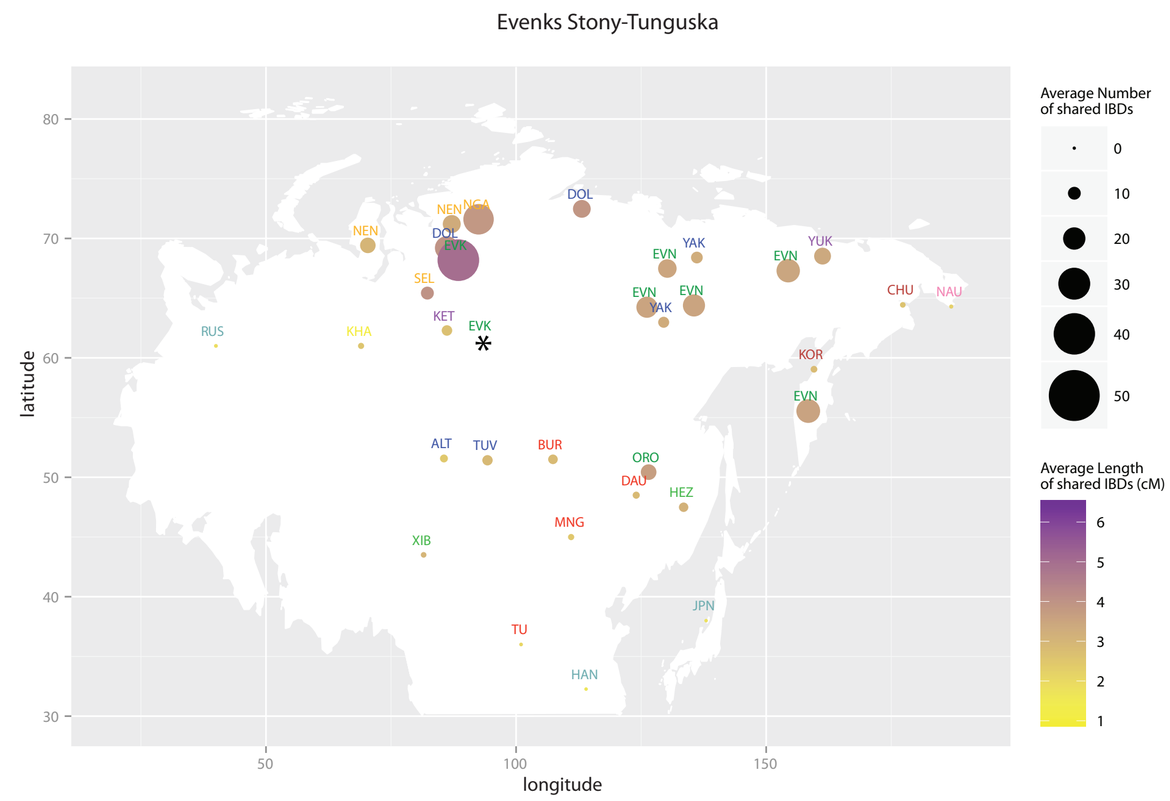
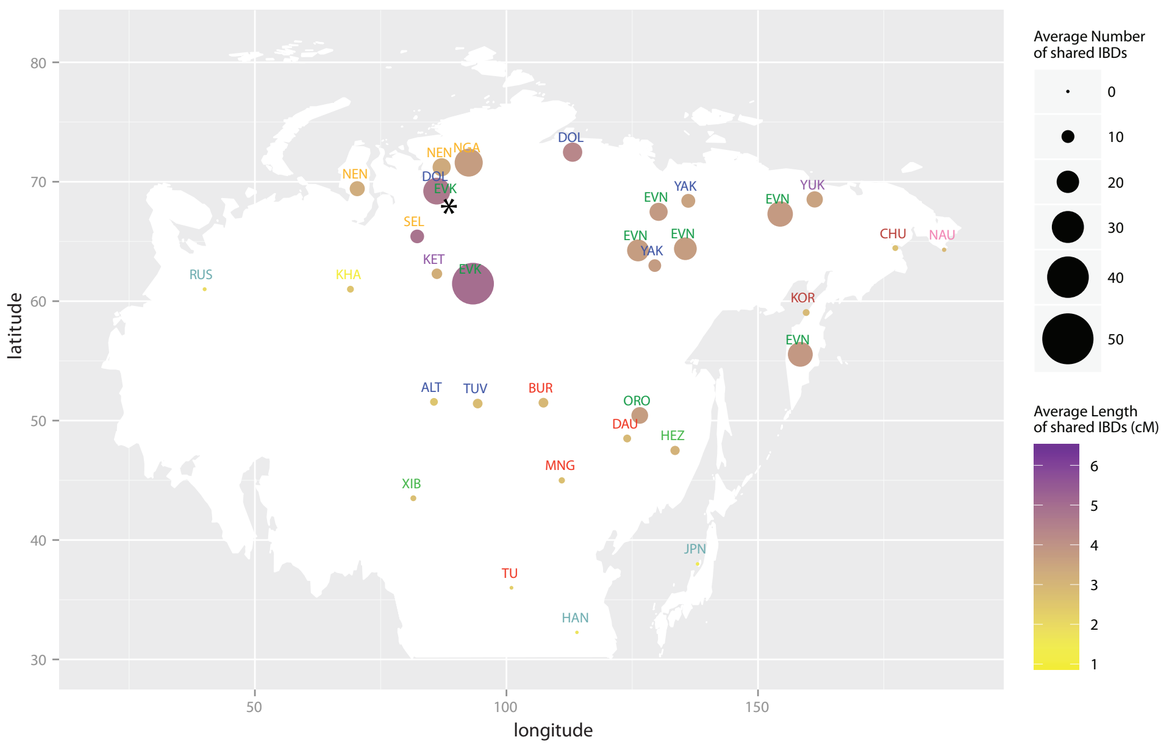


FIGURE S5. Recent relatedness based on IBD blocks. Populations are arranged in alphabetical order. Each data point represents the results for the comparison of the population marked with an asterisk to each of the other populations in the dataset. Data points are placed on the map according to the sampling location of each population (geographic coordinates are listed in Table S1). Population labels are abbreviated to the first 3 letters of the population name, except EVN = Even, EVK = Evenk, MNG = Mongolian, JPN = Japanese. Each label is color coded according to the population's linguistic affiliation as described in the legend to Figure 2. The size of each circle is proportional to the mean number of IBD segments shared between the population marked with an asterisk and the population named in the label. The color intensity is proportional to the mean length of such shared IBD segments.

FIGURE S5. Recent relatedness based on IBD blocks (continued).







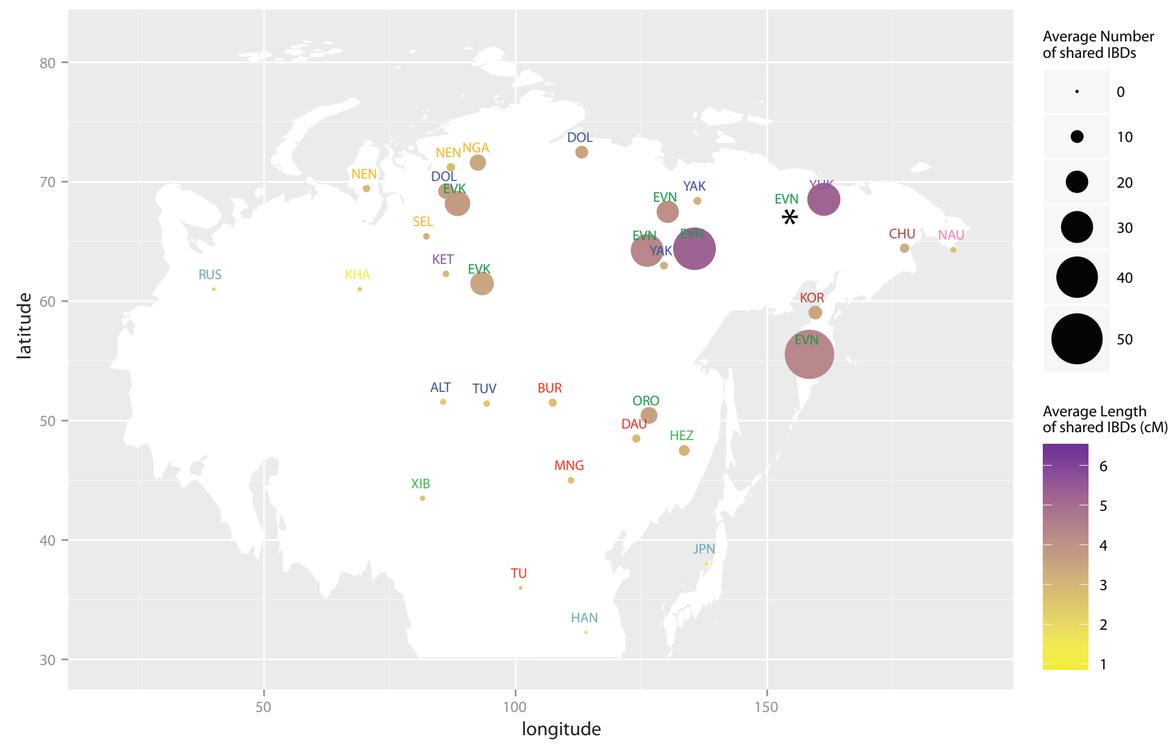
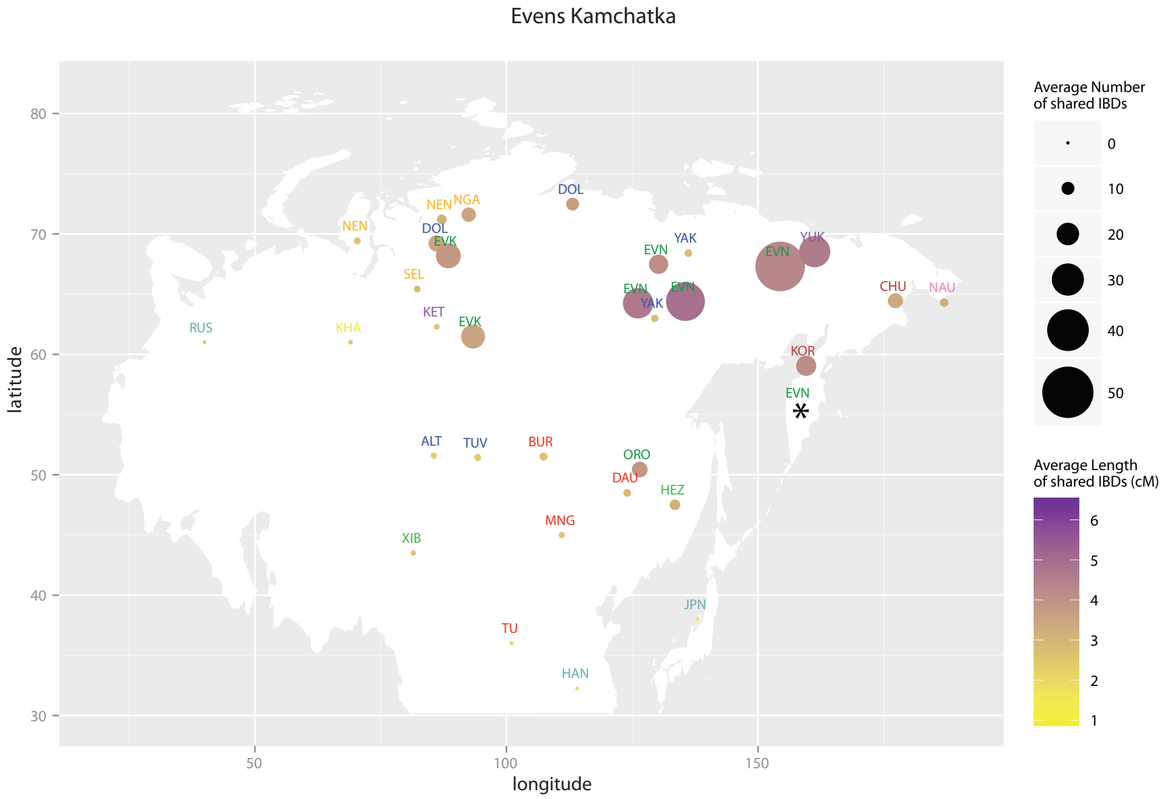
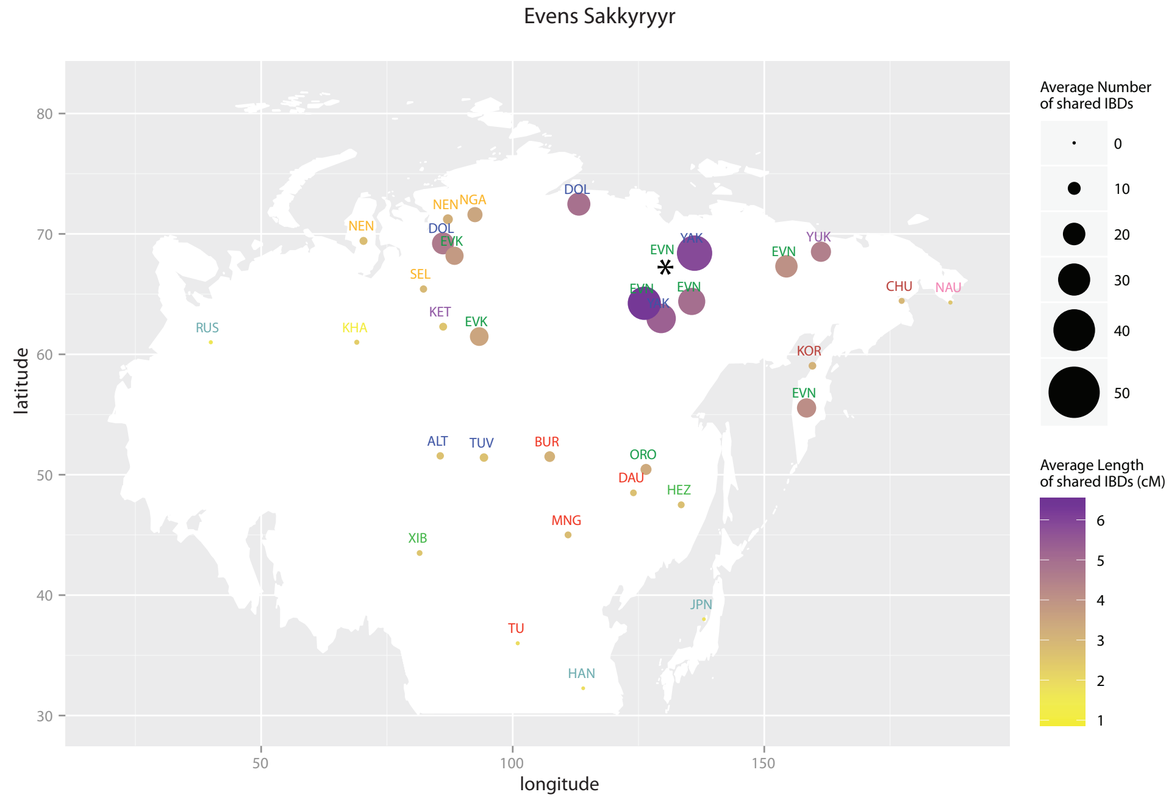
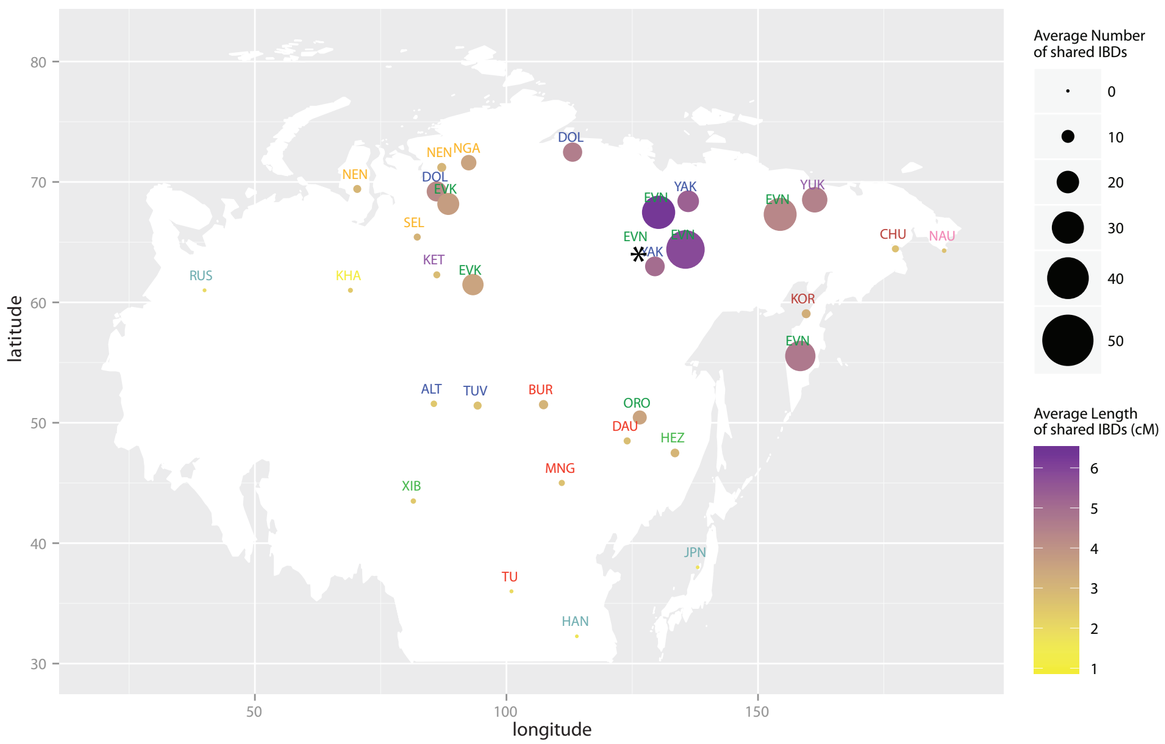


FIGURE S5. Recent relatedness based on IBD blocks (continued).







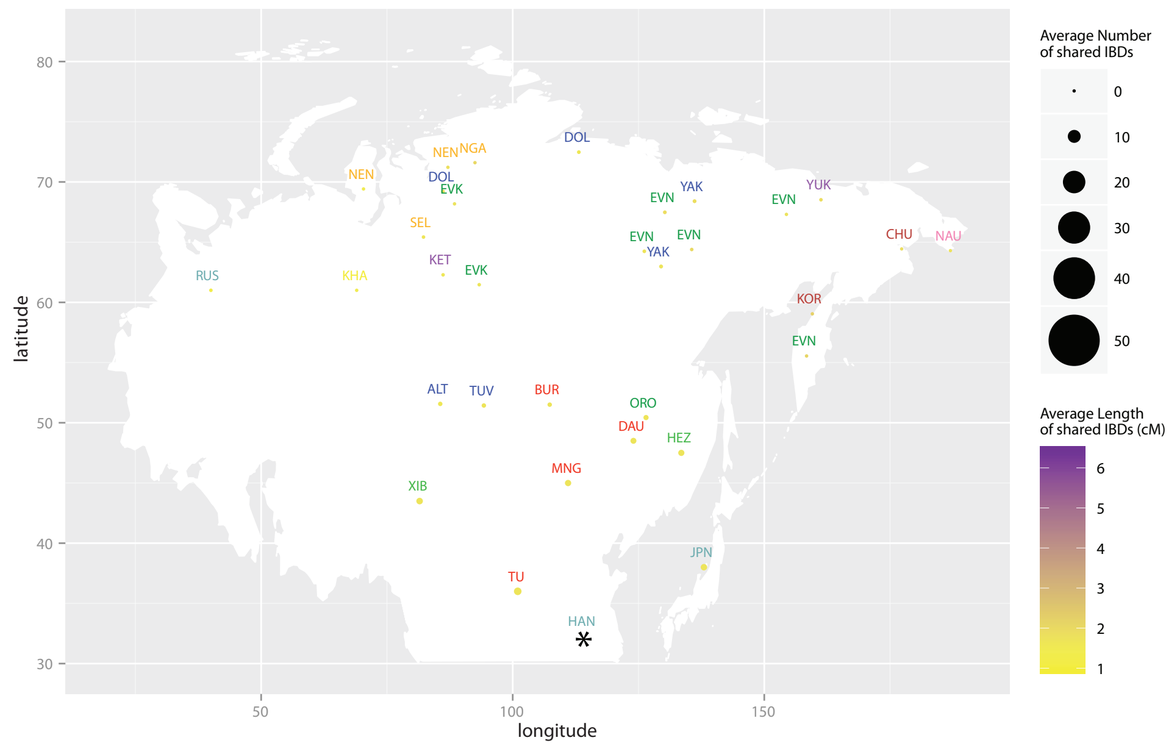
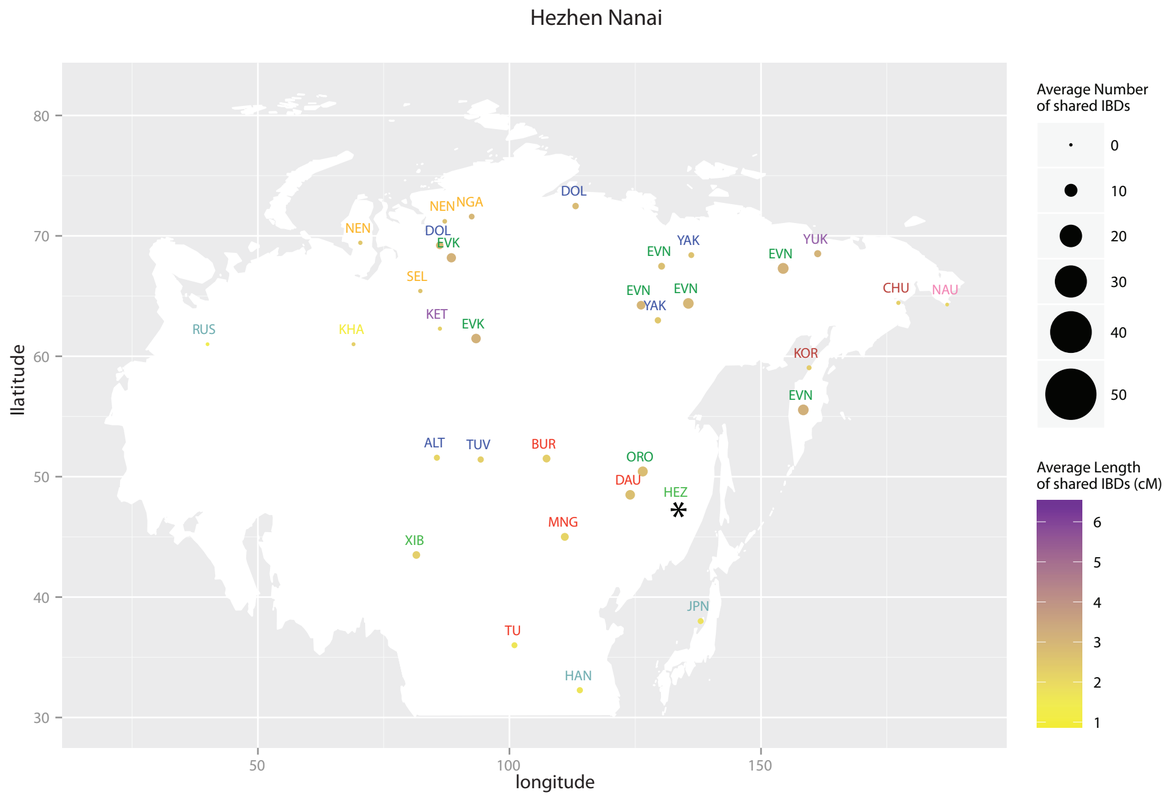
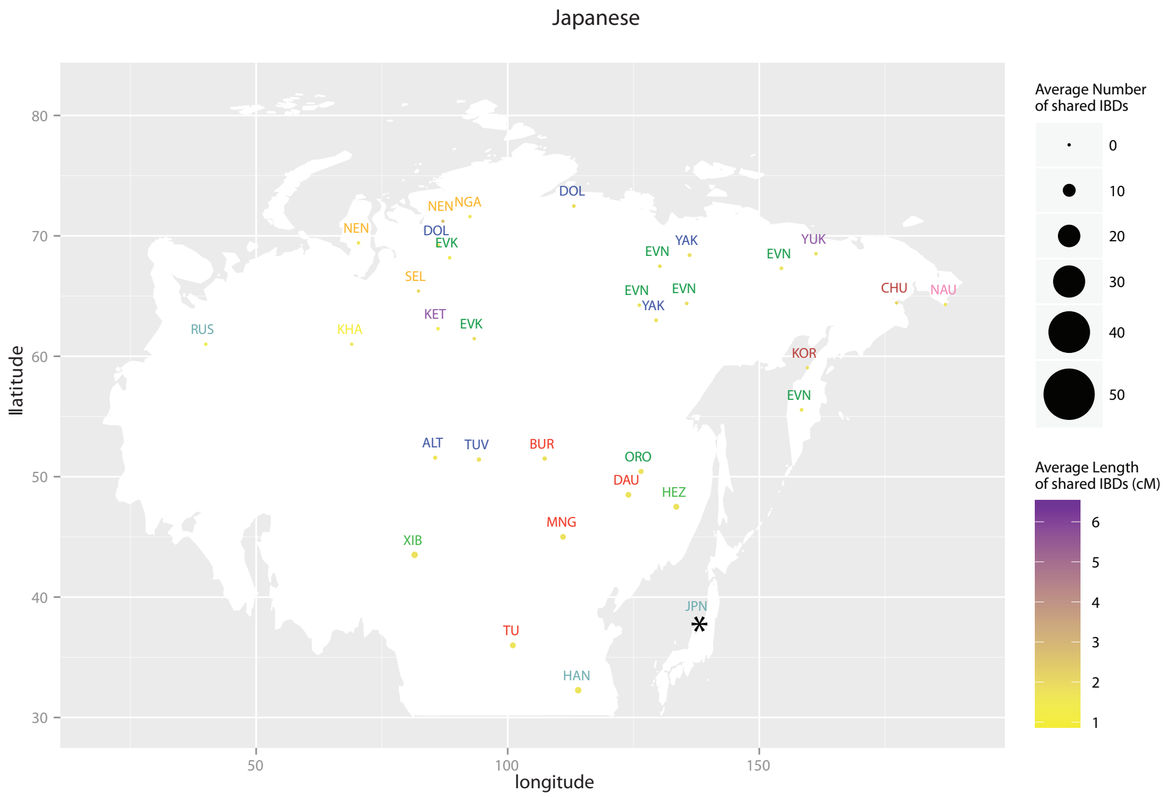
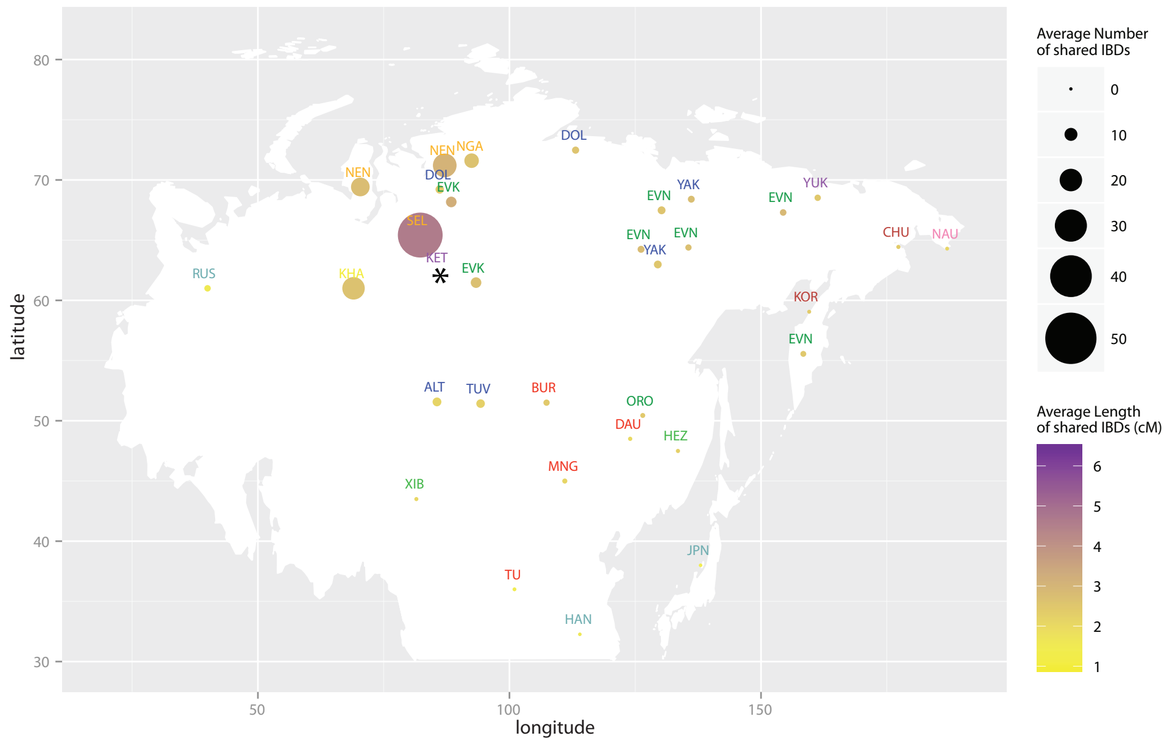


FIGURE S5. Recent relatedness based on IBD blocks (continued).







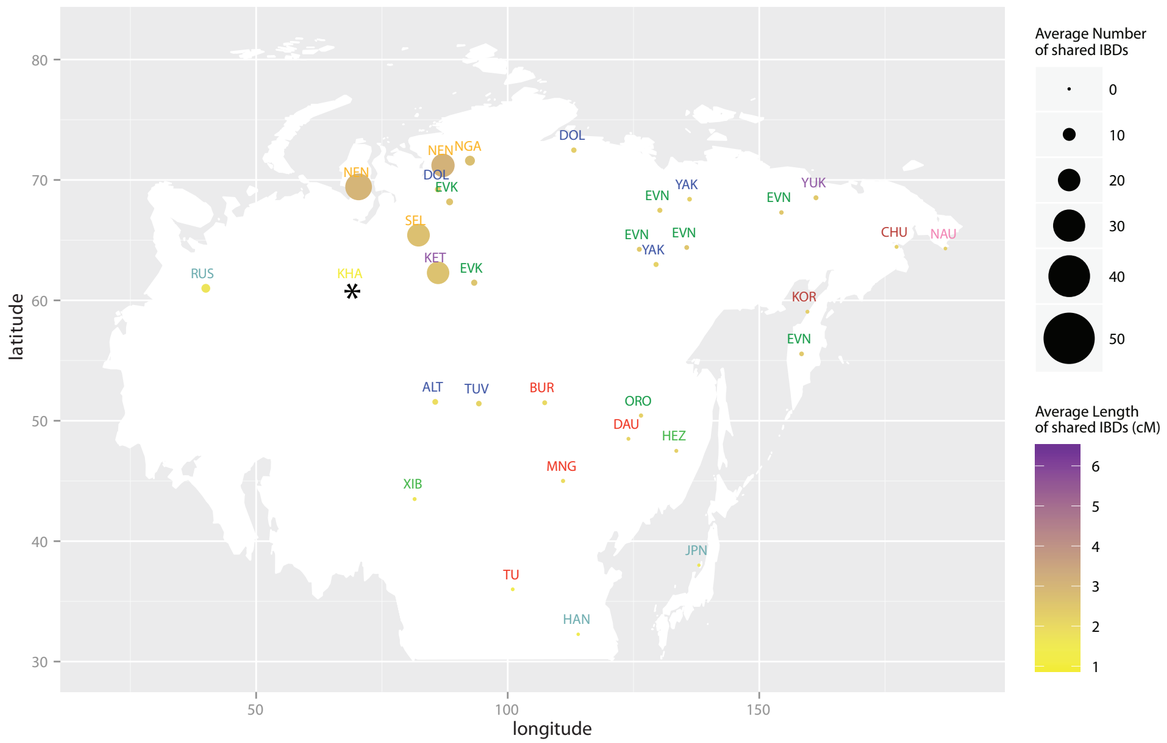
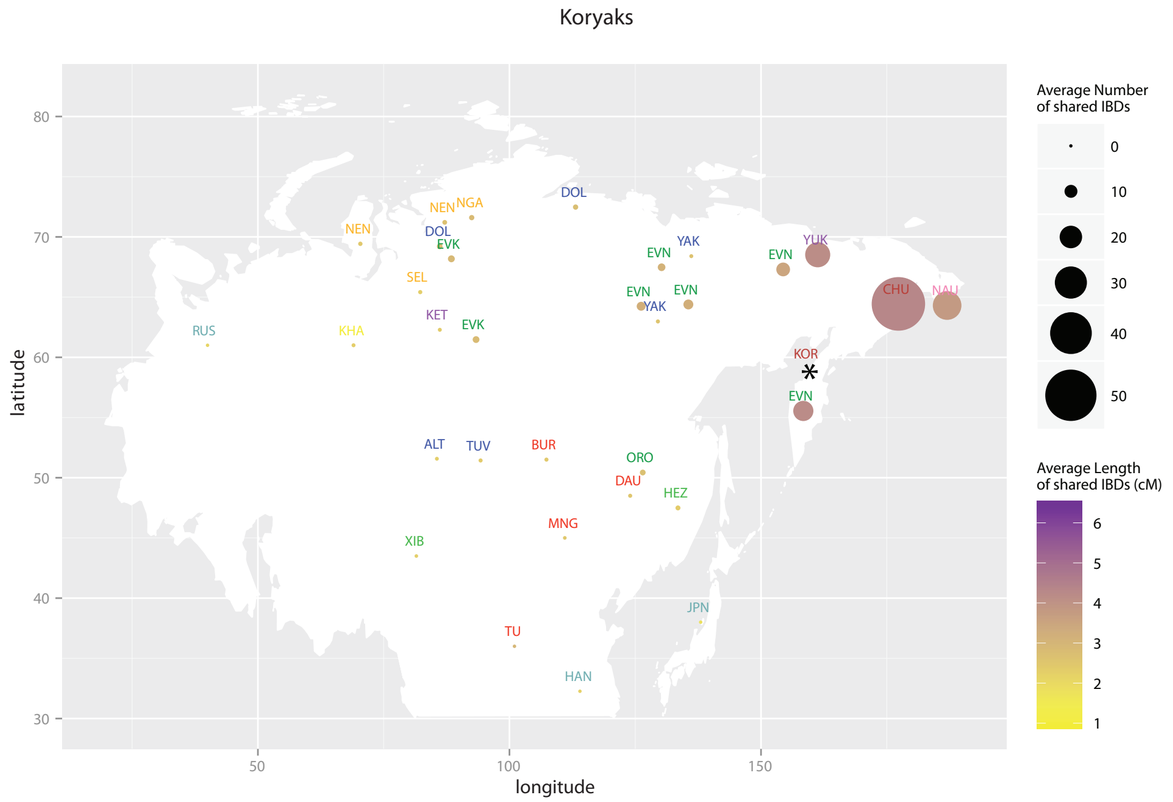
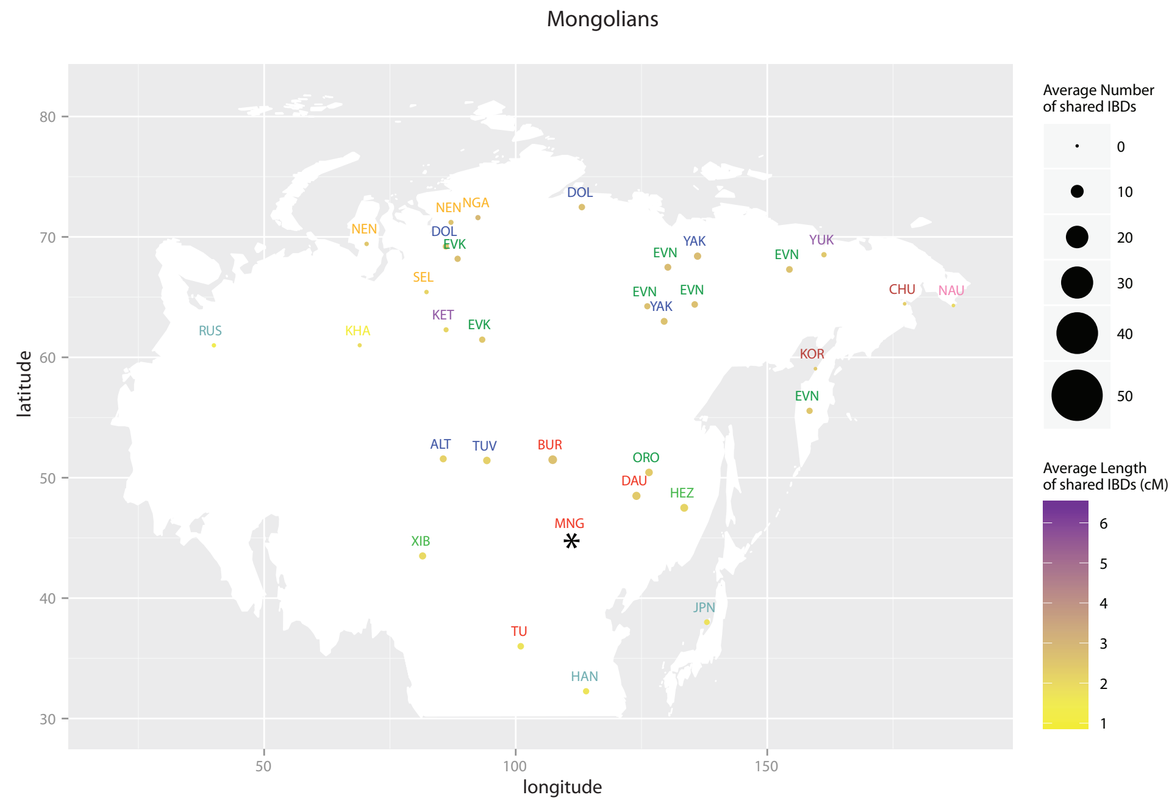
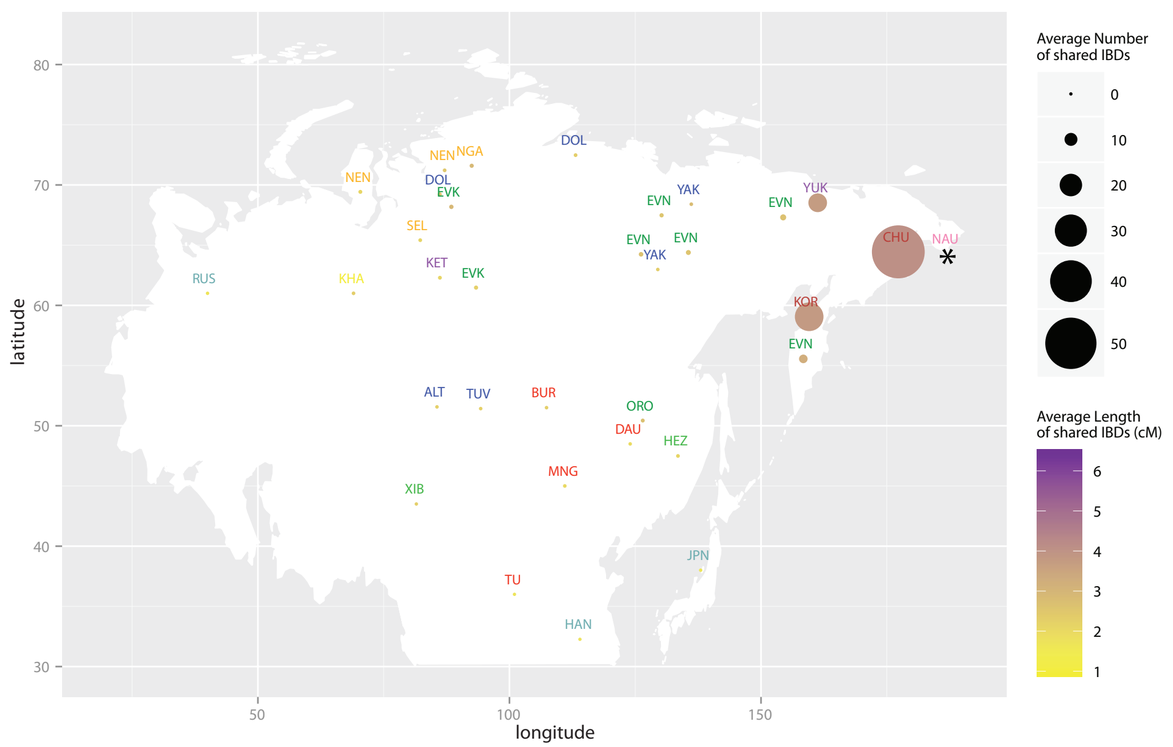


FIGURE S5. Recent relatedness based on IBD blocks (continued).







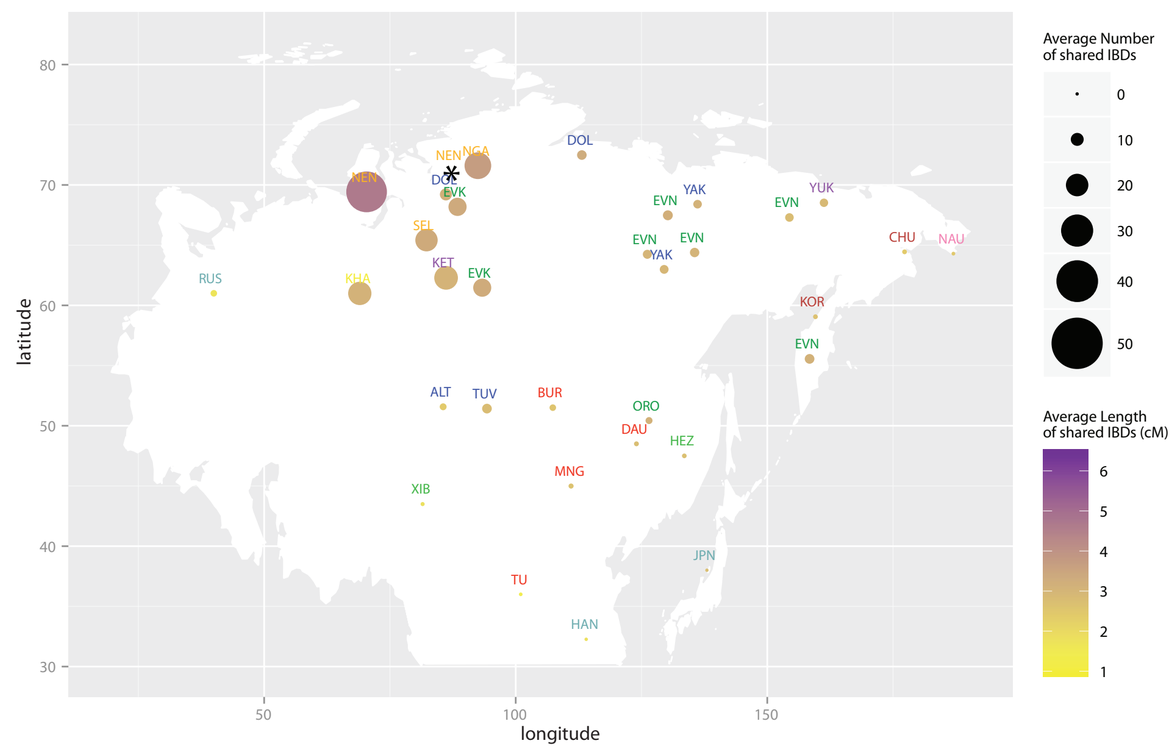
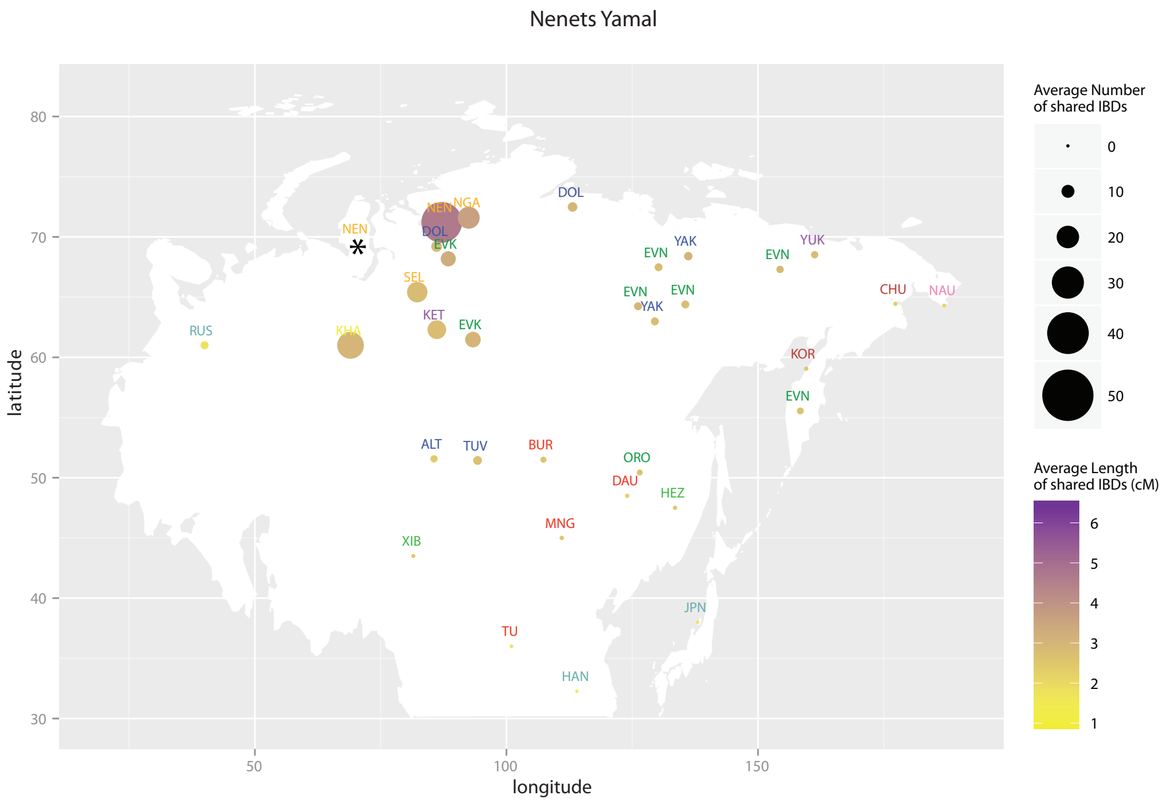
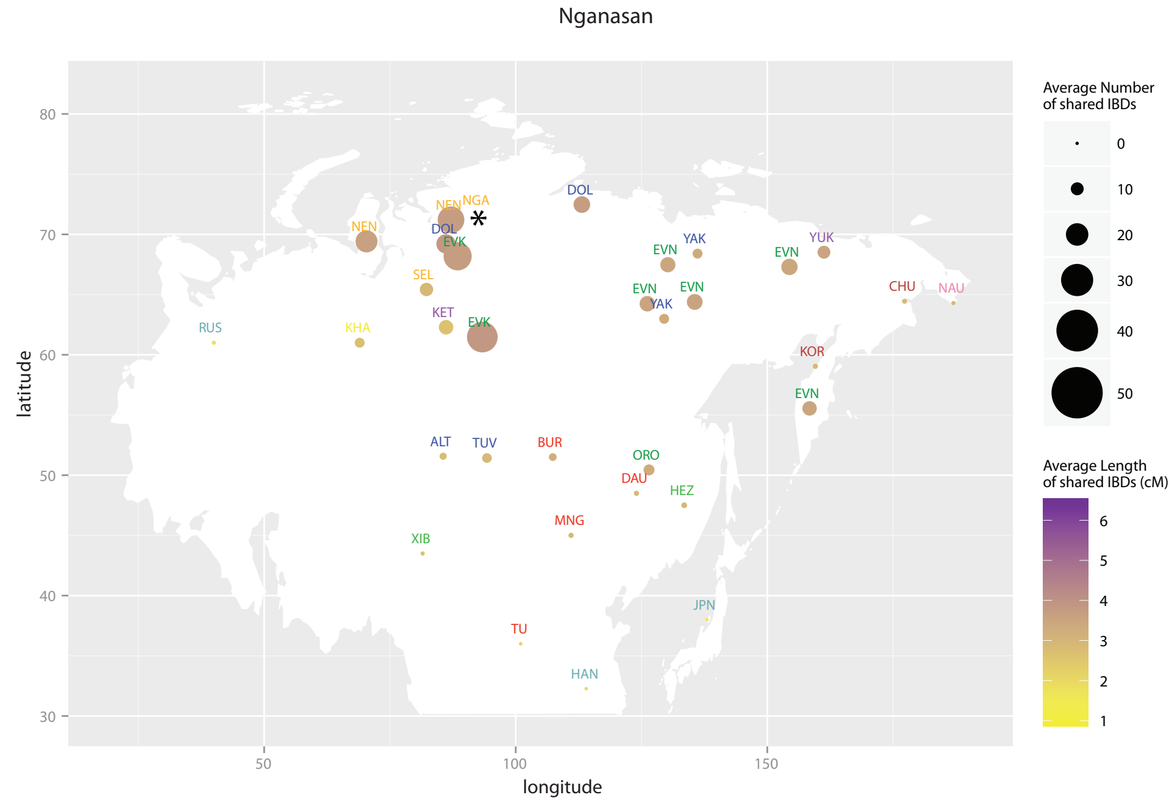
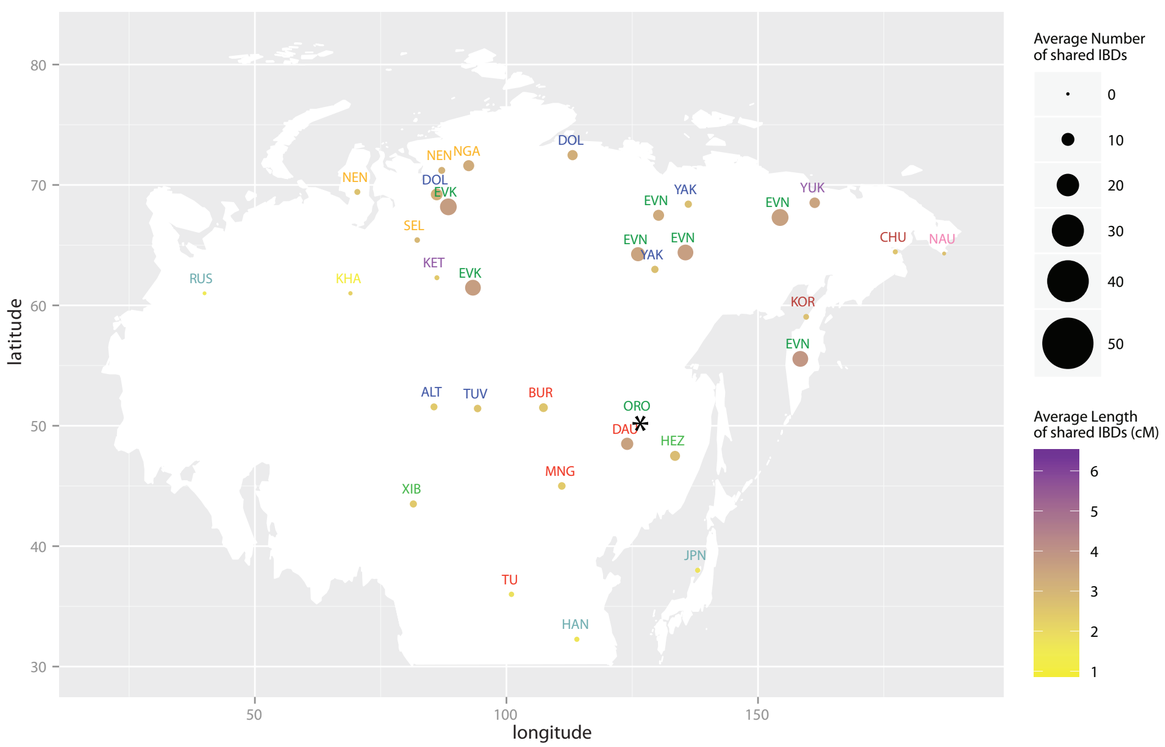


FIGURE S5. Recent relatedness based on IBD blocks (continued).







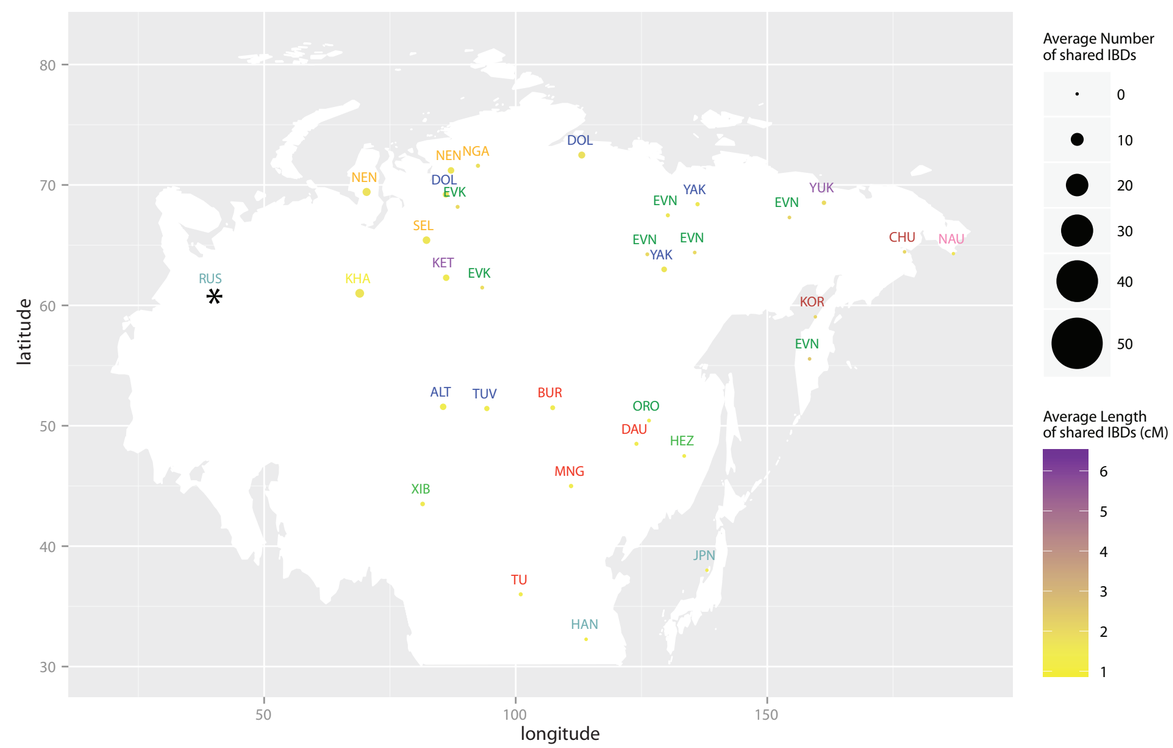
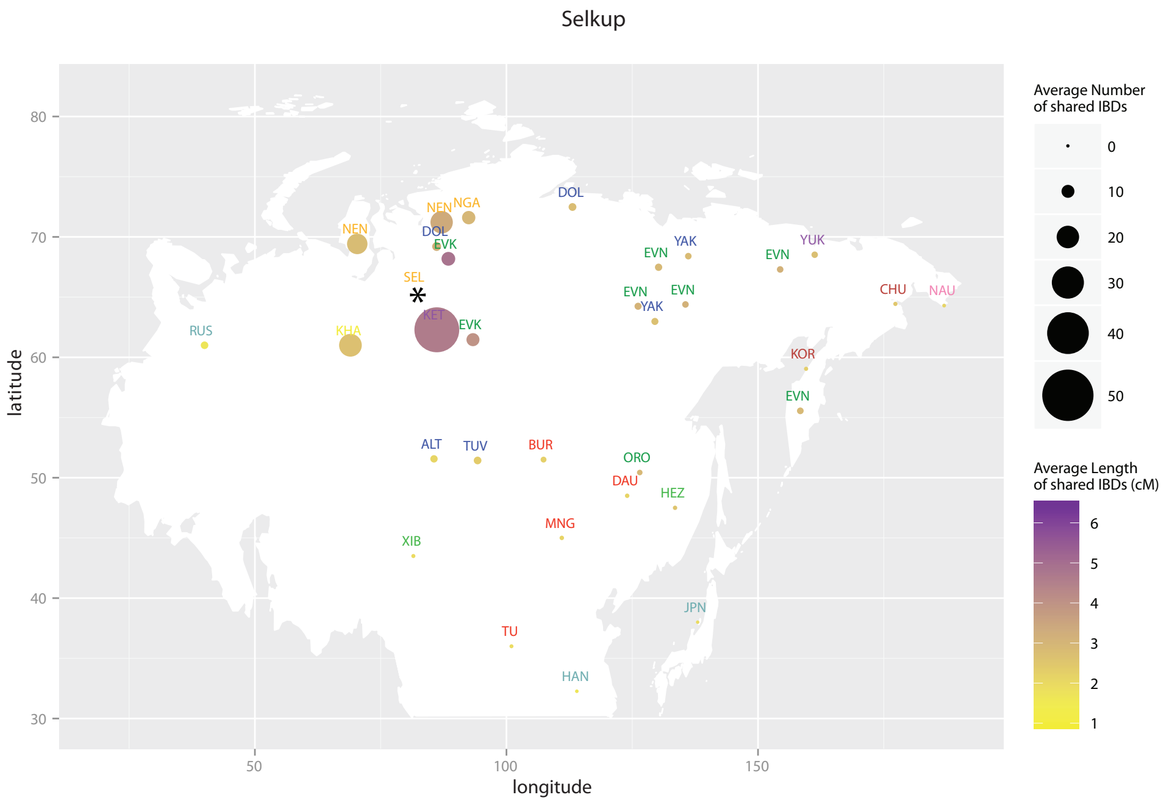
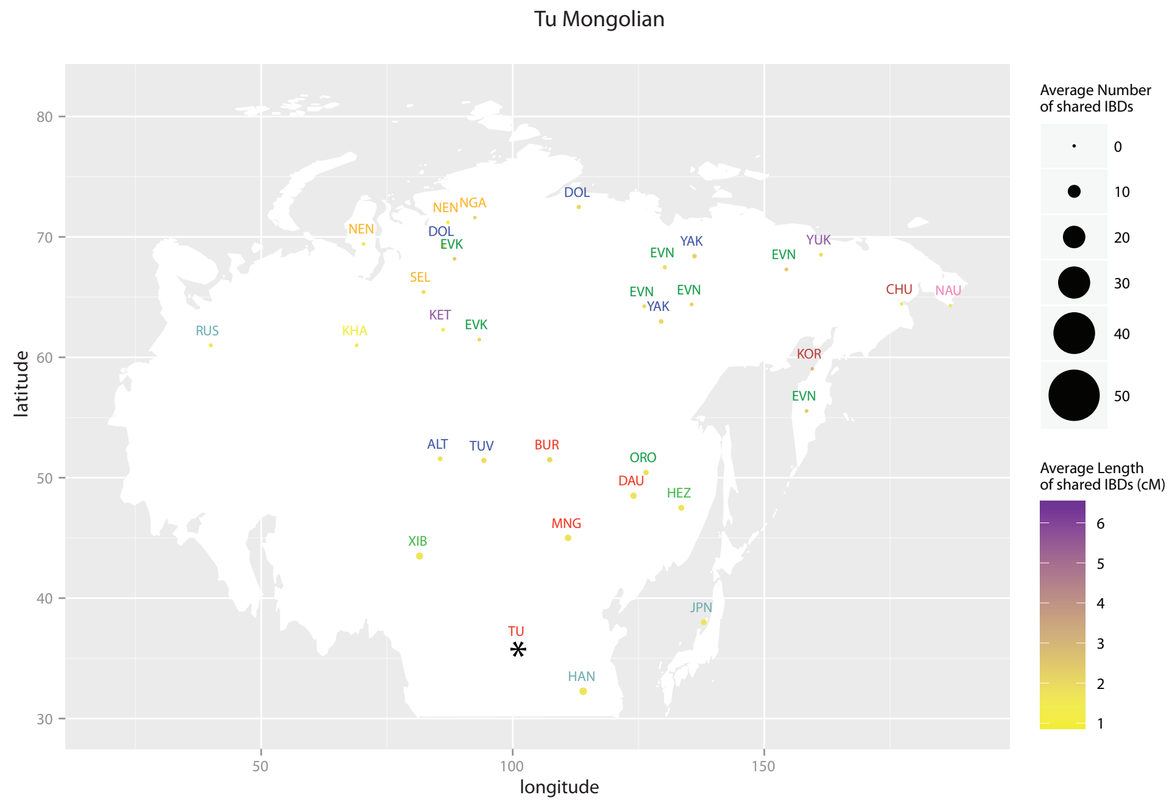
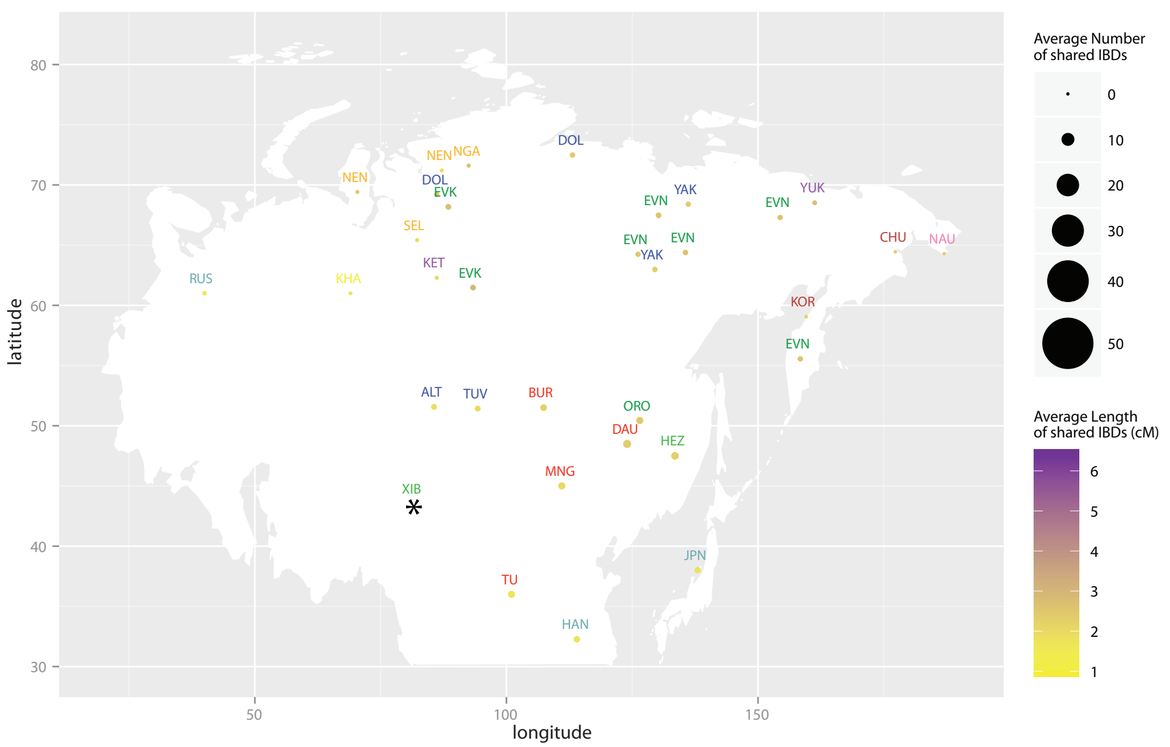


FIGURE S5. Recent relatedness based on IBD blocks (continued).







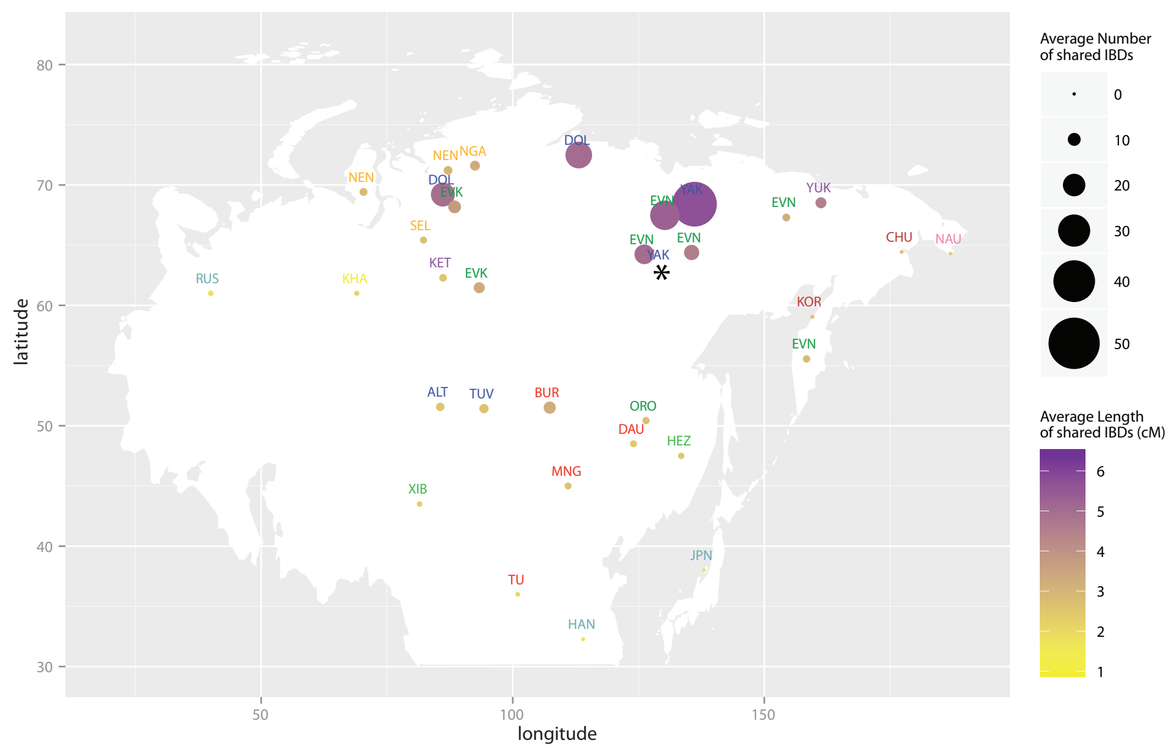


FIGURE S5. Recent relatedness based on IBD blocks (continued).

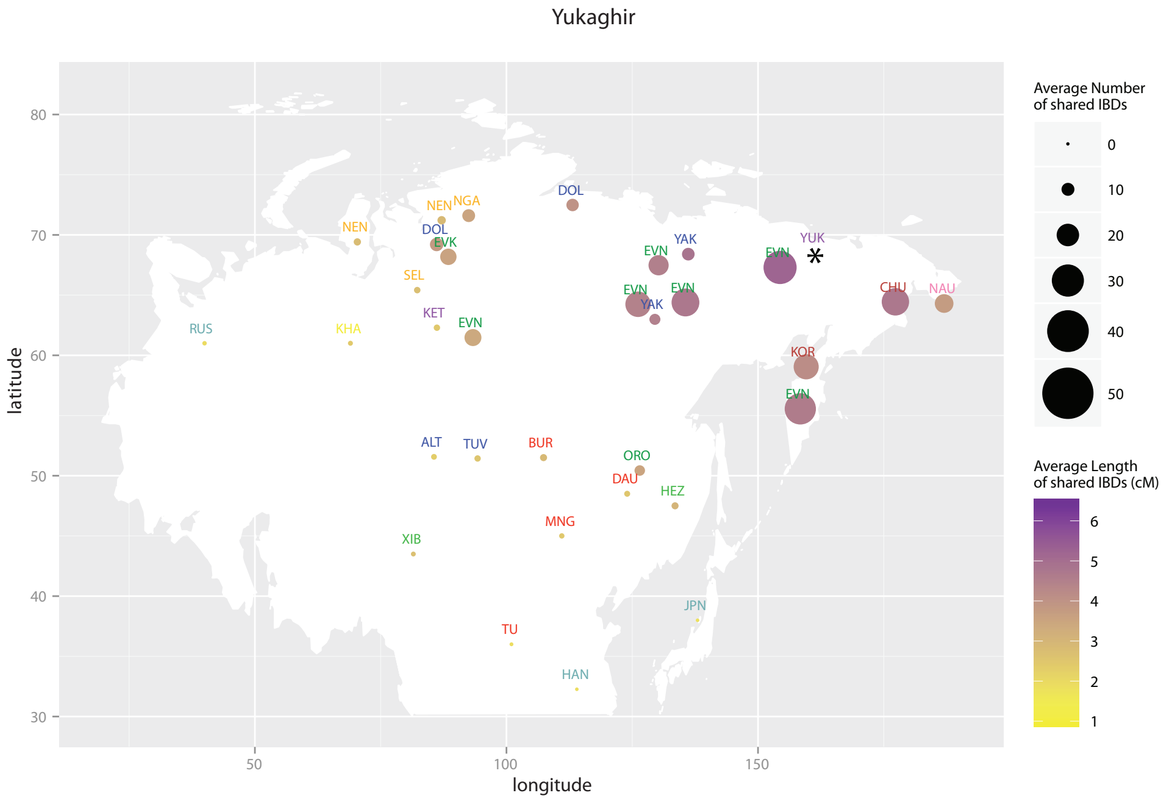


Figure S6

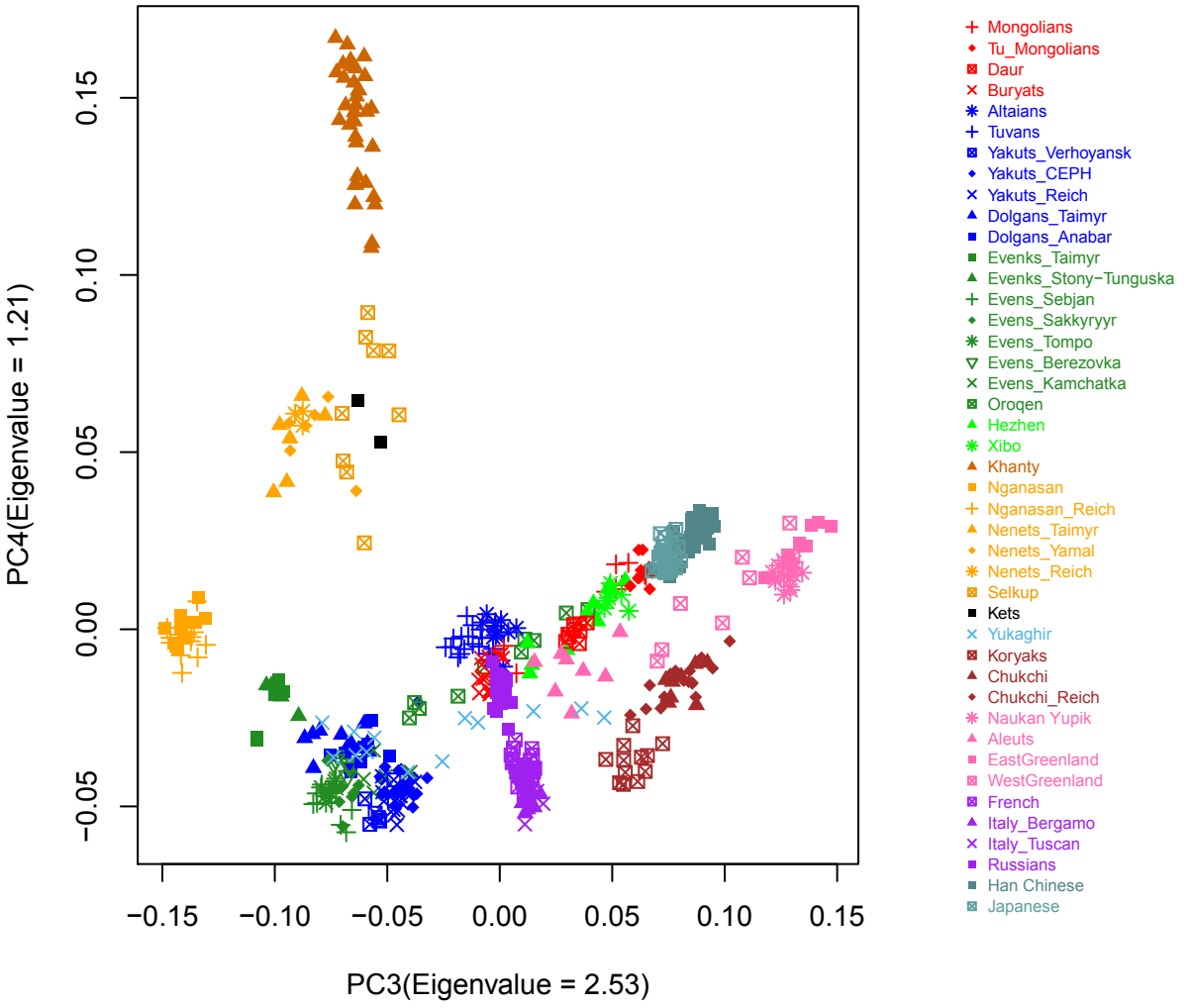


FIGURE S6. Results of the PC analysis: PC3 vs PC4. Each coloured label represents an individual, and individuals are coloured according to their linguistic affiliation, see legend to Figure 2 of the Main Text for details.

Figure S7

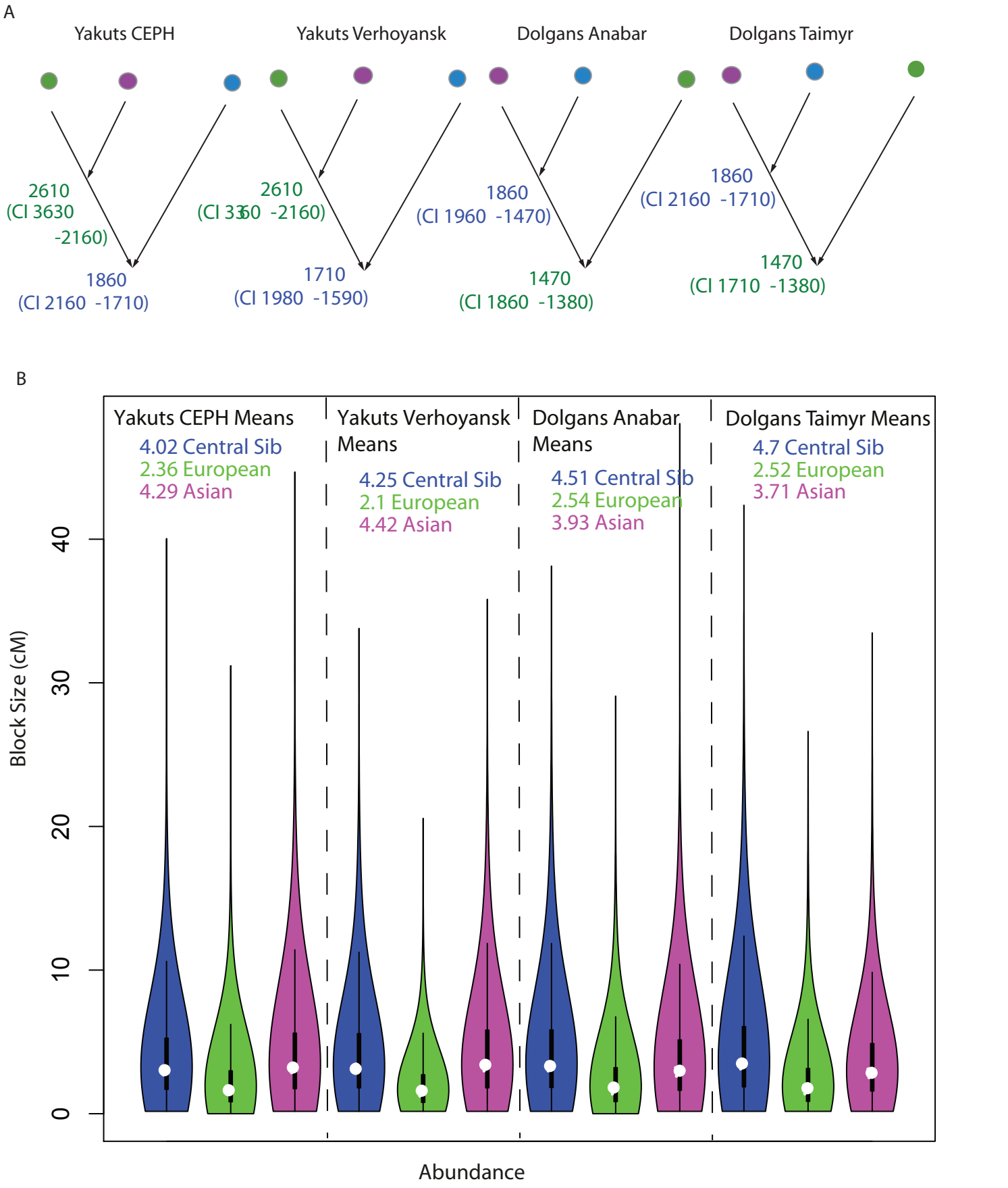


FIGURE S7. Admixture profiles for the northern Siberian Turkic-speaking populations: Yakuts and Dolgans. (A) Admixture history graphs and admixture dates inferred for each population and each admixture episode. Proxy parental populations for the different ancestral components (represented as circles) were as follows: European (green) = Italians, Central Siberian (blue) = Nganasan, Asian (pink) = Han Chinese. (B) Cumulative distribution of all ancestry blocks. For each population the plot captures the total abundance of blocks of each ancestry (x axis) of different genetic lengths in cM (y axis); the average width of the blocks of each ancestry and variance around the mean are also shown.

FIGURE S8

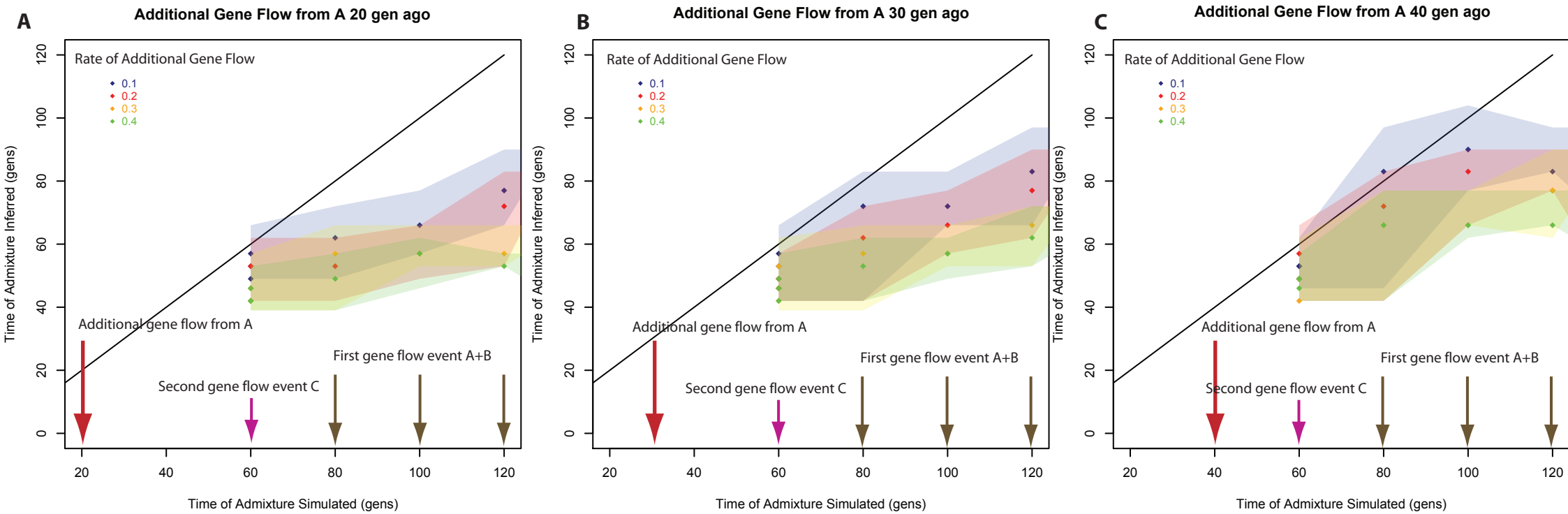


FIGURE S8. Admixture date estimates for artificially constructed admixed populations, where admixture history involved one ancestry which was contributed twice. Each panel shows a comparison of simulated vs. estimated dates of admixture for 12 sets of simulated populations (each with the same recent episode of admixture 60 generations ago, and a different earlier episode of admixture 120, 100 or 80 generations ago). Additional gene flow with four different admixture rates (? = 0.1, 0.2, 0.3 or 0.4) from a population which had already contributed to an earlier episode was simulated for each of these populations. The shaded area denotes the 95% confidence interval for the inferred dates of admixture. Additional gene flow was introduced at (A) 20 (B) 30 or (C) 40 generations ago.

Figure S9

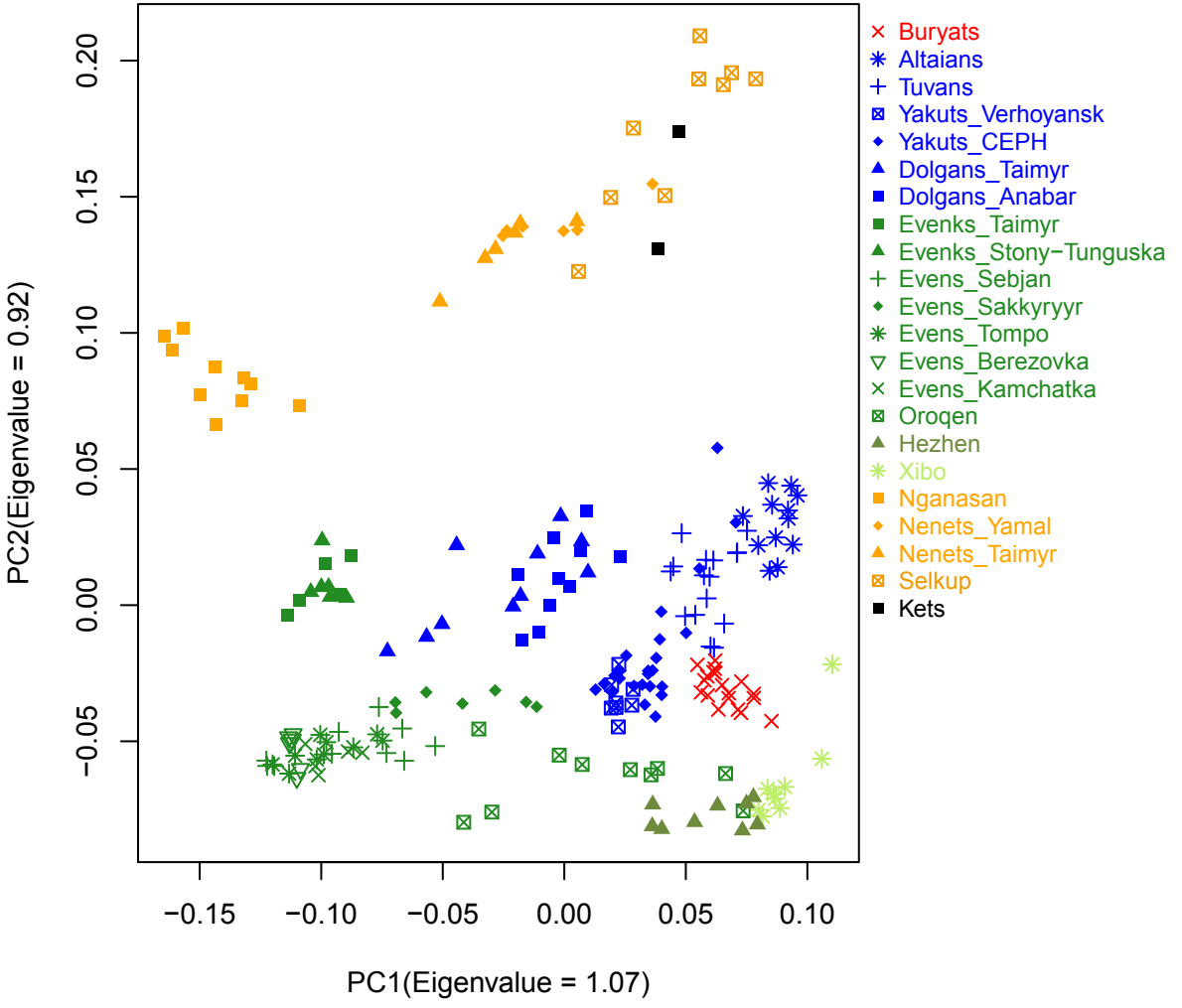


FIGURE S9. Results of the PC analysis for a subset of the dataset. To better understand the relationships between closely related Siberian populations, the PC analysis was performed excluding those populations driving the main axes of variation in the PCA for the entire dataset (all non-Siberian populations as well as the populations of the Russian Far East , Khanty and Kets). The heavily admixed Yukaghirs were also excluded from this analysis.

Figure S10A

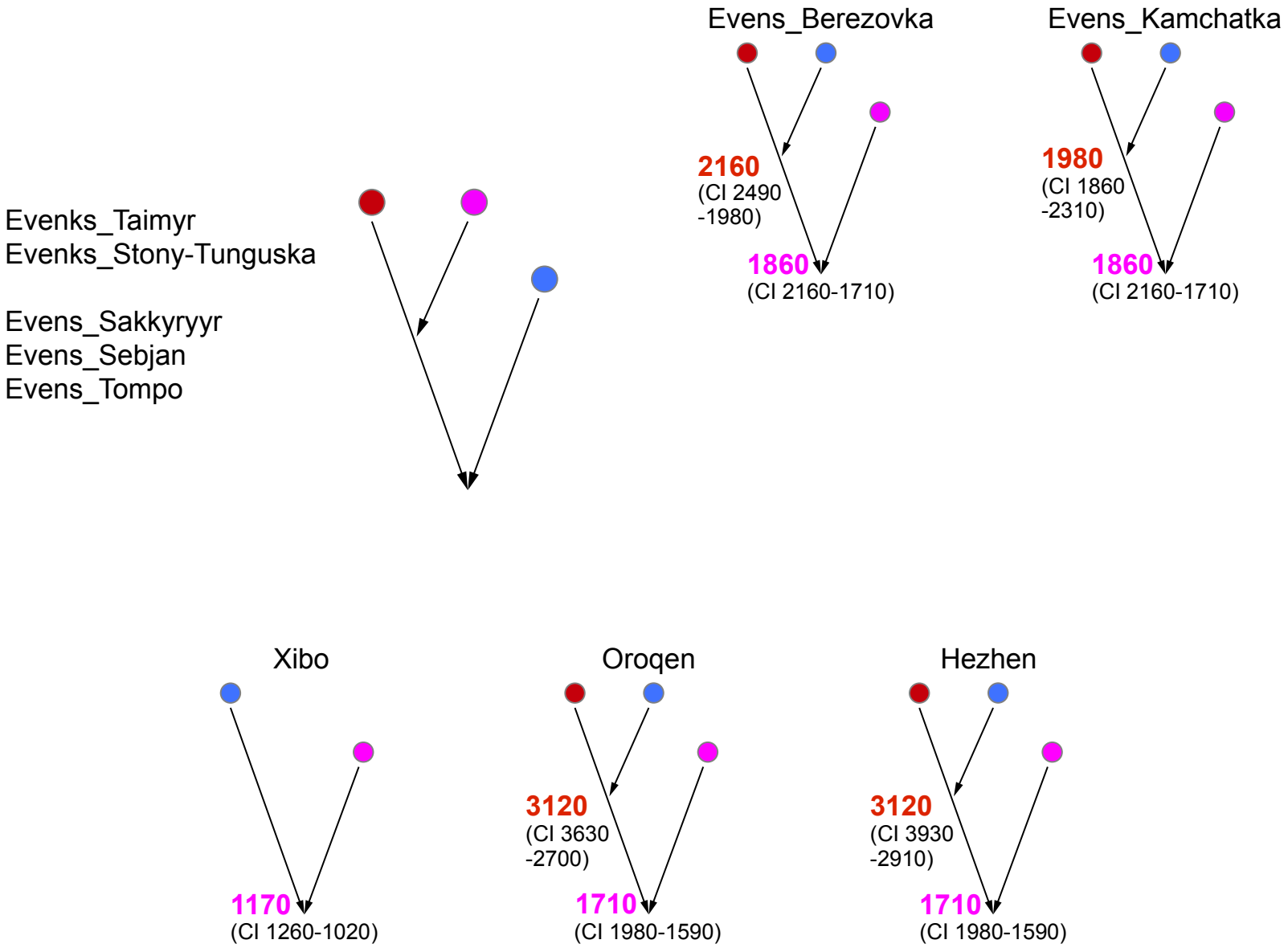


FIGURE S10. Admixture profiles for the Tungusic-speaking populations: Evenks, Evens, Oroqen, Hezhen and Xibo. (A) Admixture history graphs and admixture dates inferred for each population and each admixture episode. Proxy parental populations for the different ancestral components (represented as circles) were as follows: Central Siberian (blue) = Nganasan, Far Eastern (red) = Koryaks, Asian (pink) = Han Chinese.

Figure S10B

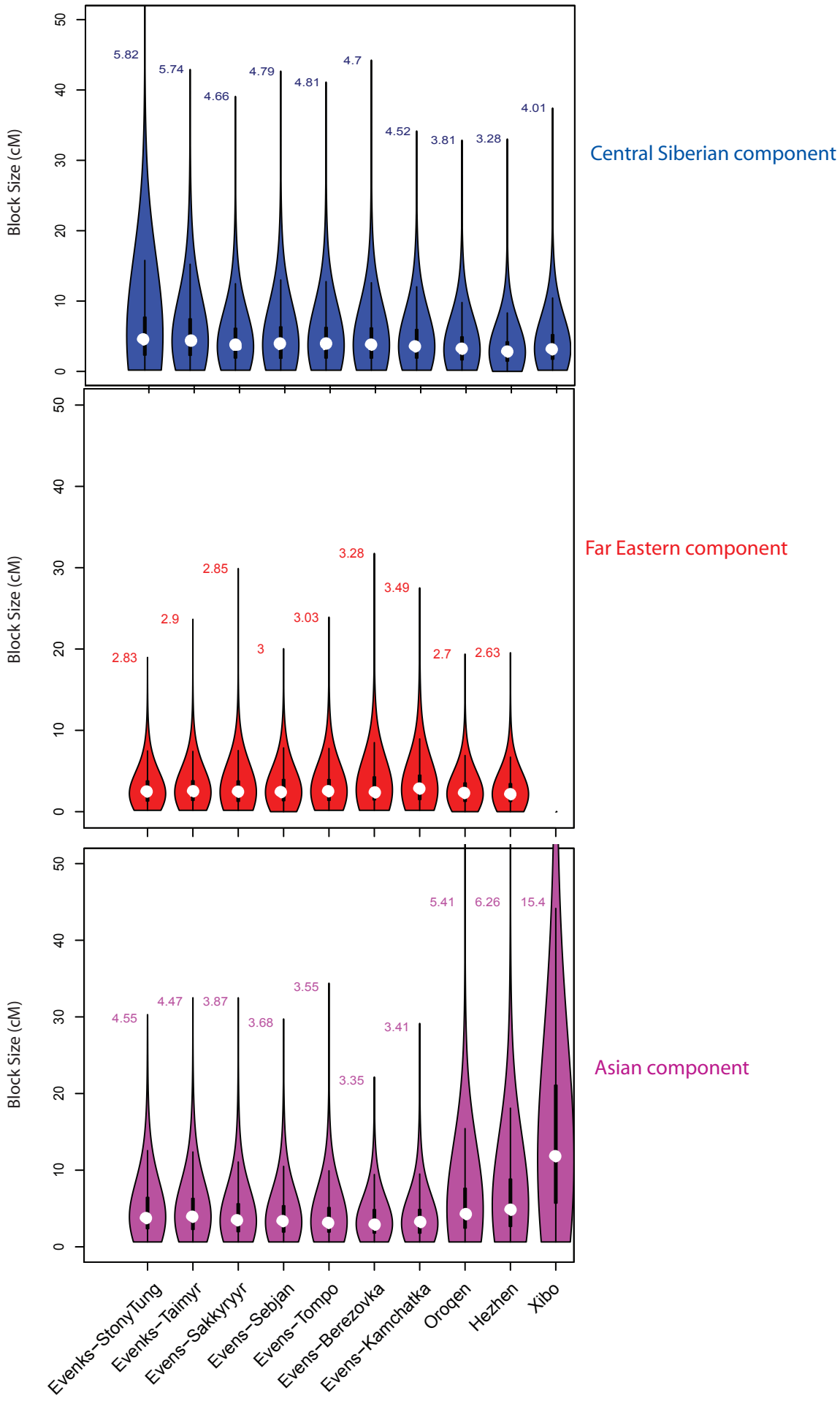


FIGURE S10. Admixture profiles for the Tungusic-speaking populations: Evenks, Evens, Oroqen, Hezhen and Xibo. (B) Cumulative distribution of all ancestry blocks. For each population the plot captures the total abundance of blocks of each ancestry (x axis) of different genetic lengths in cM (y axis); the average width of the blocks of each ancestry and variance

Figure S11

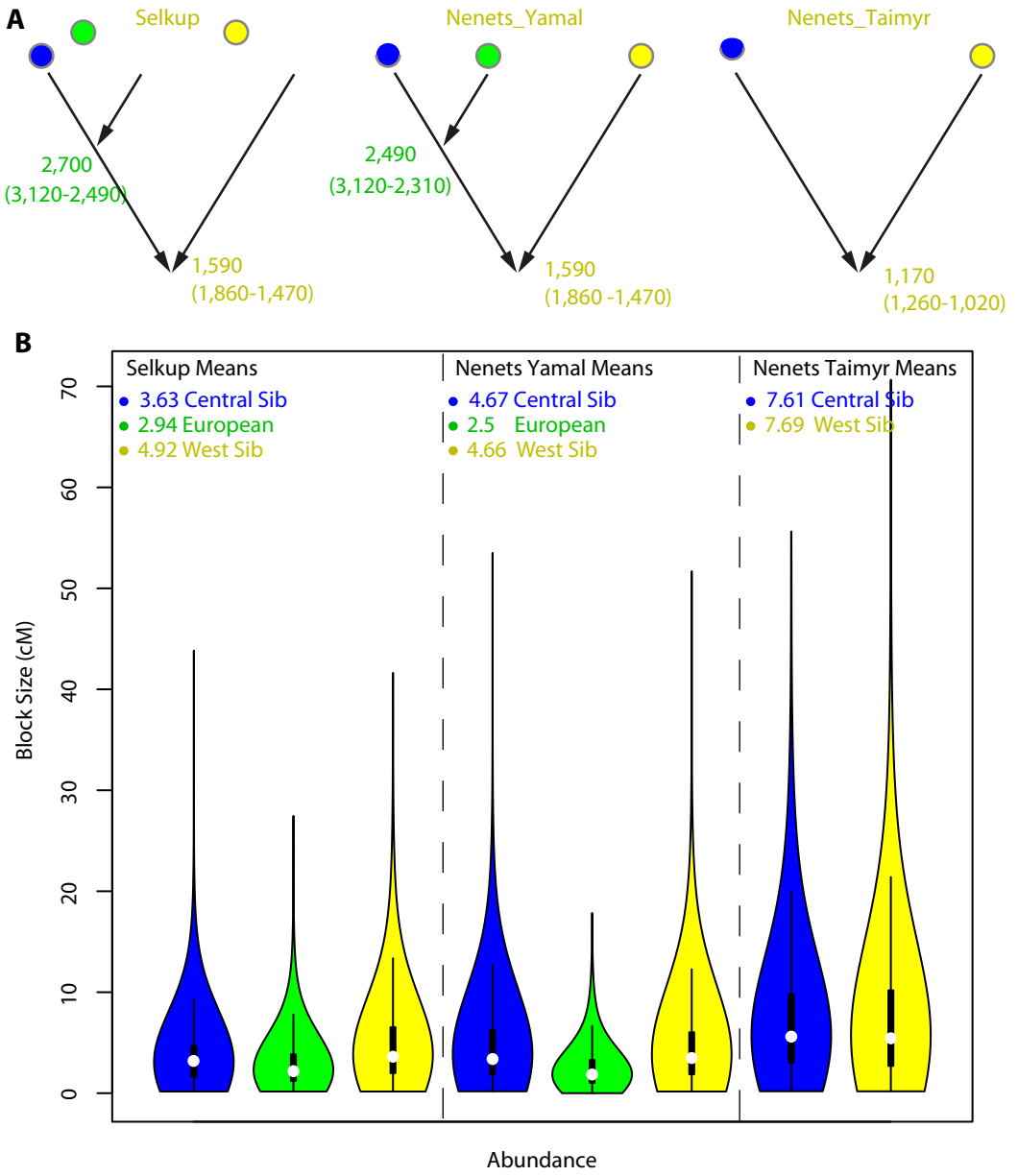


FIGURE S11 Admixture profiles for the Samoyedic-speaking populations: Selkup and Nenets. (A) Admixture history graphs and admixture dates inferred for each population and each admixture episode. Proxy parental populations for the different ancestral components (represented as circles) were as follows: Central Siberian (blue) = Nganasan, Western Siberian (yellow) = Khanty, European (green) = Italians. (B) Cumulative distribution of all ancestry blocks. For each population the plot captures the total abundance of blocks of each ancestry (x axis) of different genetic lengths in cM (y axis); the average width of the blocks of each ancestry and variance around the mean are also shown.

Figure S12A

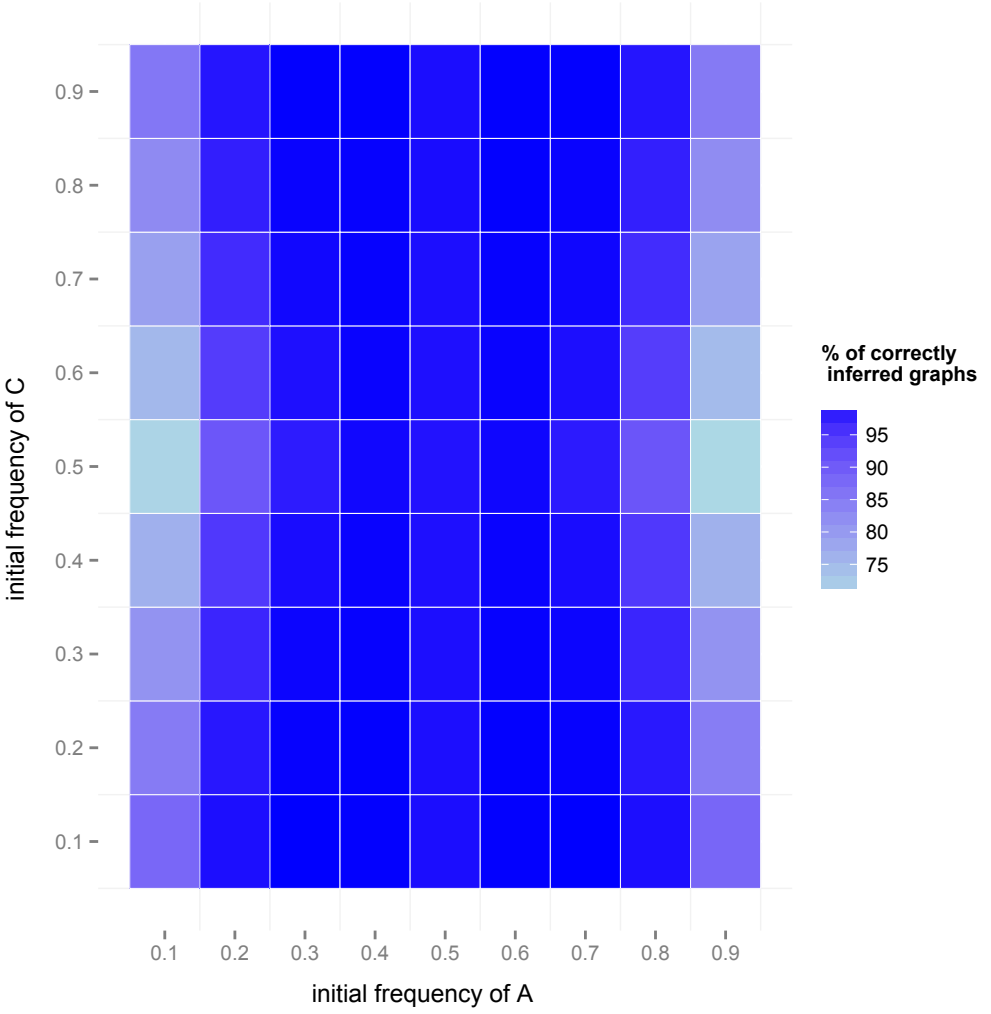


FIGURE S12. Results of simulations to validate the Admixture History Graph method. (A) The simulated history is ((AB)C), where ancestries A and B were already present in a population at the time of admixture with C. Each rectangle summarizes the results for 100,000 simulations performed for a given set of A and C values, and the colour intensity reflects the number of simulations in which the tree was inferred correctly. The lowest value of 70% correctly inferred graphs is observed when the final frequency of A (or B) is below 10%, and the frequency of C is around 50%.

Figure S10B

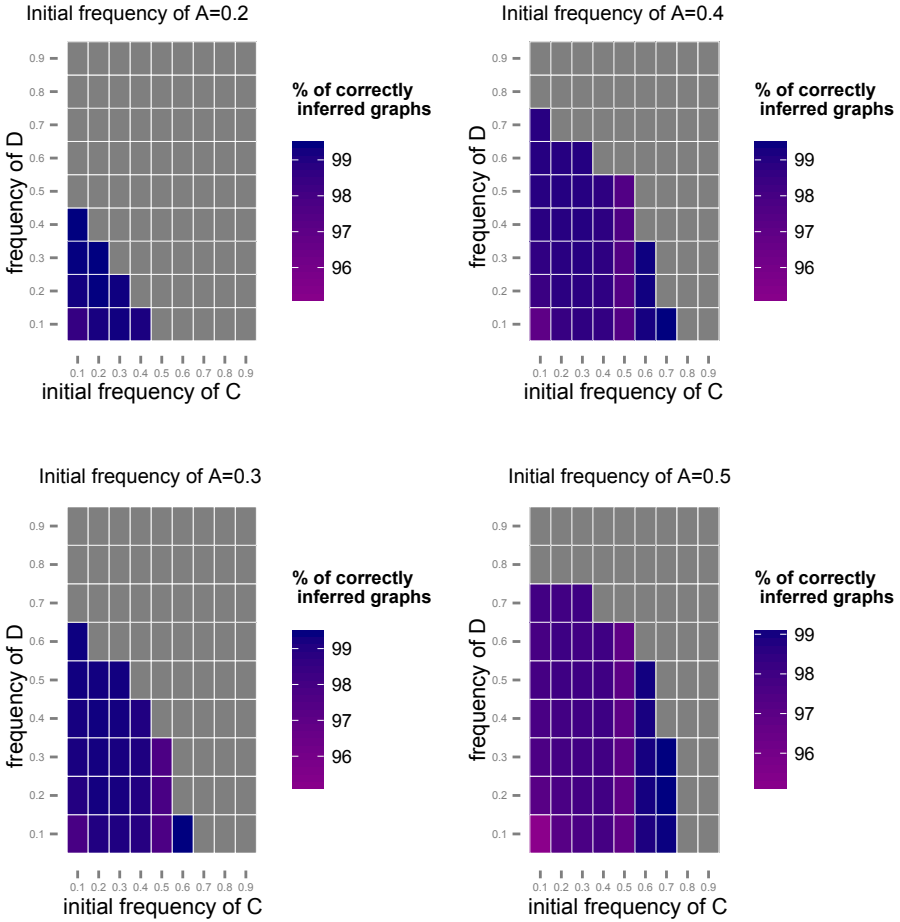


FIGURE S12. Results of simulations to validate the Admixture History Graph method. (B) The simulated history is (((AB)C)D), where an already admixed population with two ancestry components A and B is first admixed with C, which is then followed by another admixture with D. Based on the simulation results of ((AB)C) admixture histories (previous panel), the scenarios where the final frequency of A (or B) dropped to below 10% were excluded. In each panel, each rectangle summarizes the results for 100,000 simulations performed for a given set of C and D values, with the value of A given in the panel's title. The color intensity reflects the number of simulations in which the tree was inferred correctly. Grey rectangles denote impossible combinations of A, C and D. For example, if the initial frequency of A is 0.2, that of C 0.5, and that of D 0.2, then addition of C would reduce the frequency of A to 0.1, while subsequent influx of ancestry D would bring its final frequency to 0.08 (see formulae given in Figure 7), which is too low given the criteria we set for these simulations.

Figure S10C

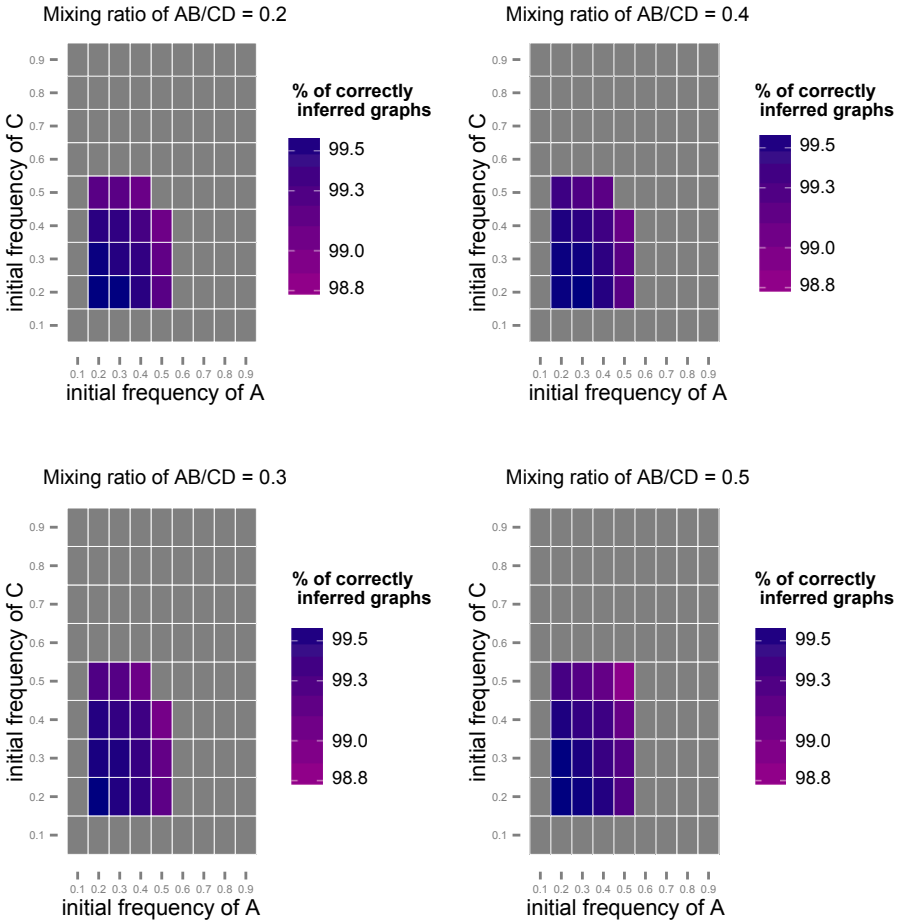


FIGURE S12. Results of simulations to validate the Admixture History Graph method. (C) The simulated history is ((AB)(CD)), where admixture occurs between two already admixed population. In each panel, each rectangle summarizes the results for 100,000 simulations performed for a given set of A and C values, with the value given in the panel's title denoting the mixing ratio between the AB and CD populations. The color intensity reflects the number of simulations in which the tree was inferred correctly, and grey rectangles denote impossible combinations of A, C and D.

FIGURE S13

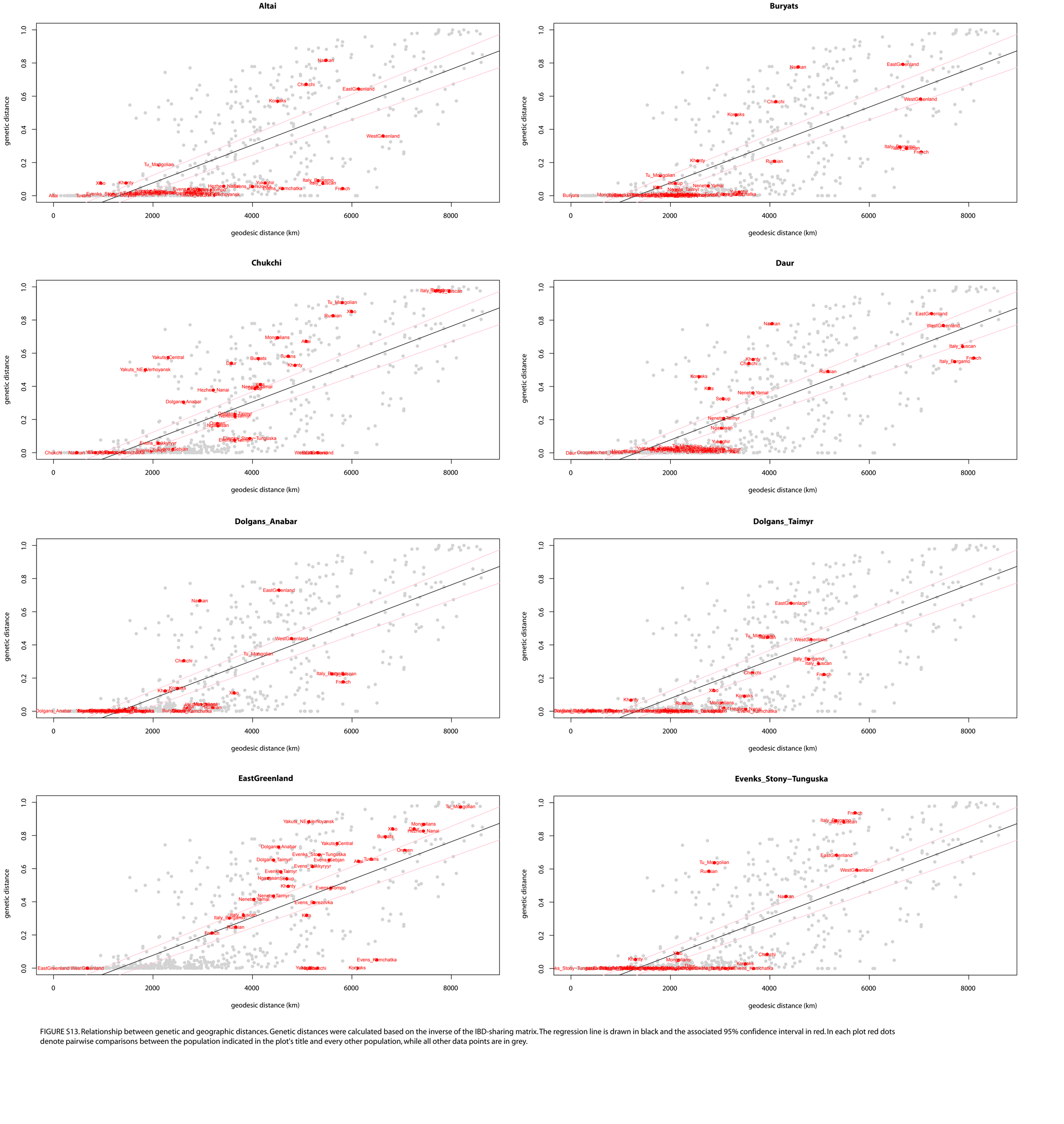
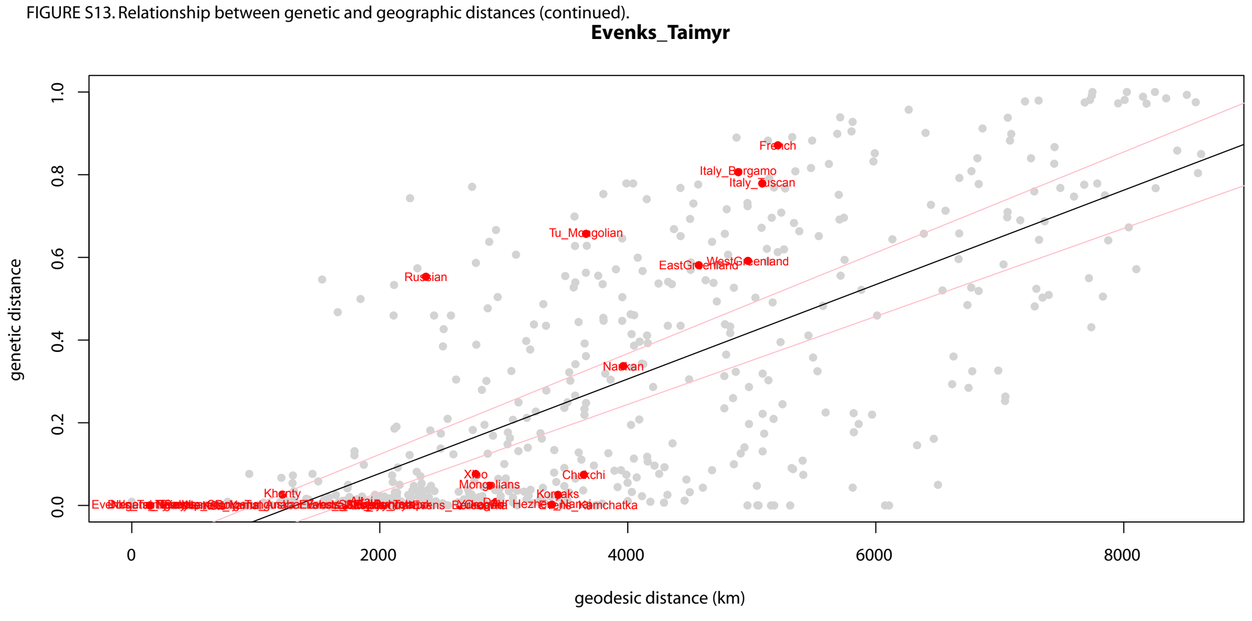
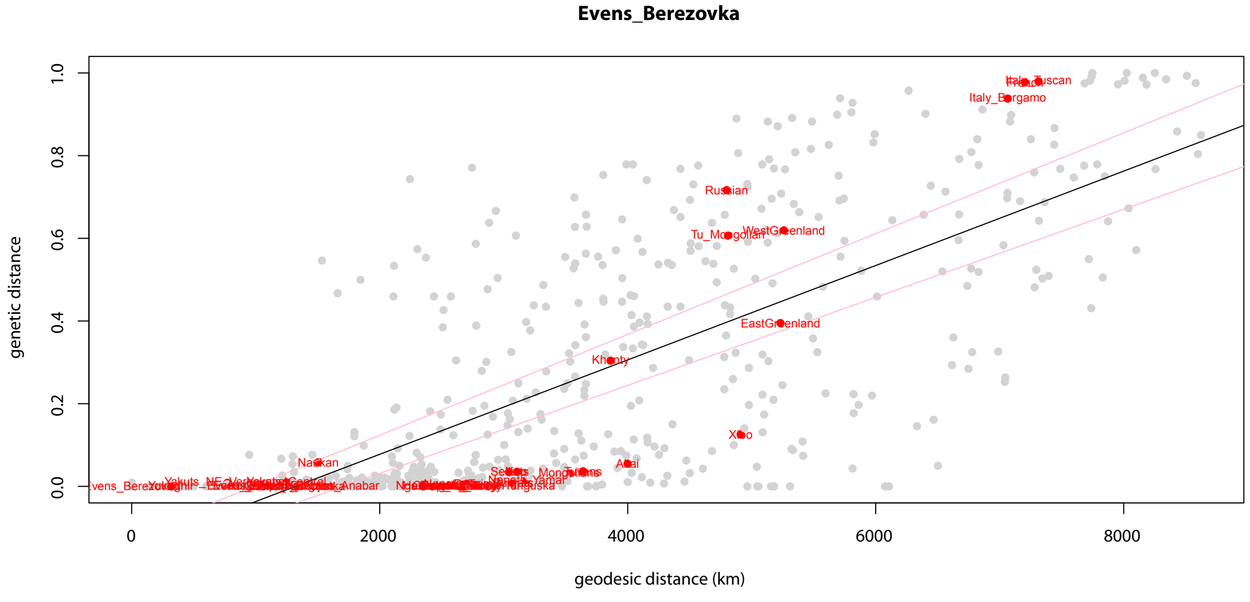
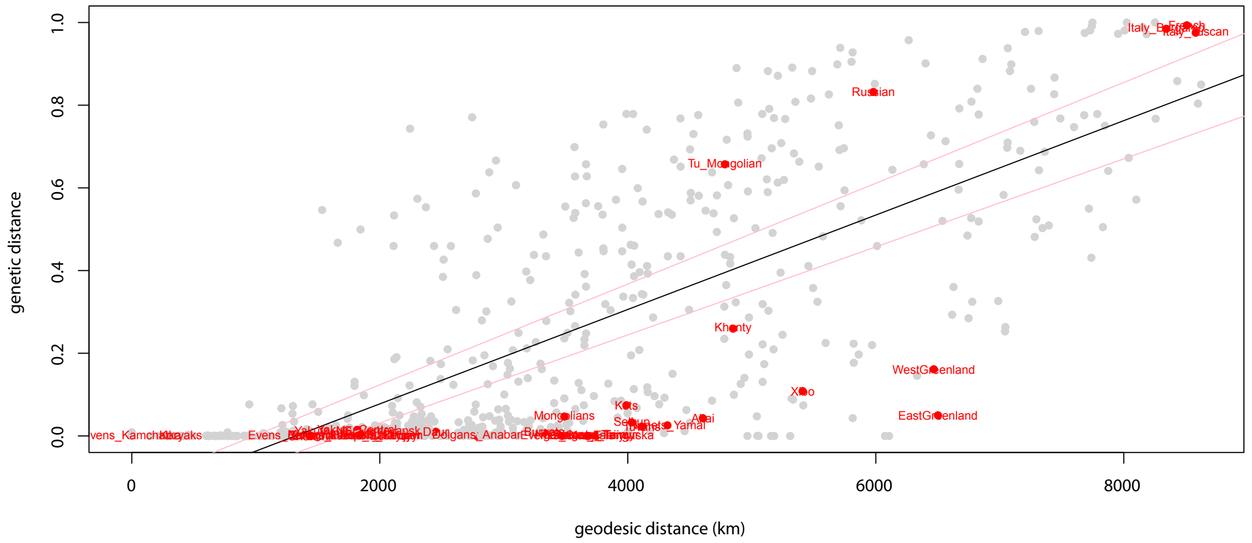


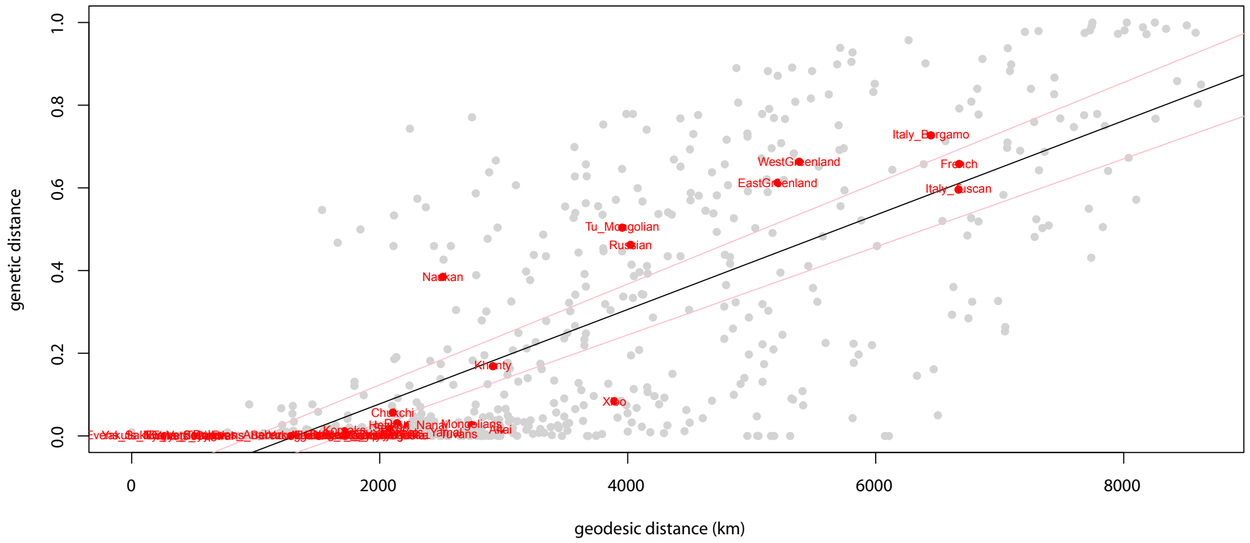
FIGURE S13. Relationship between genetic and geographic distances. Genetic distances were calculated based on the inverse of the IBD-sharing matrix. The regression line is drawn in black and the associated 95% confidence interval in red. In each plot red dots denote pairwise comparisons between the population indicated in the plot's title and every other population, while all other data points are in grey.

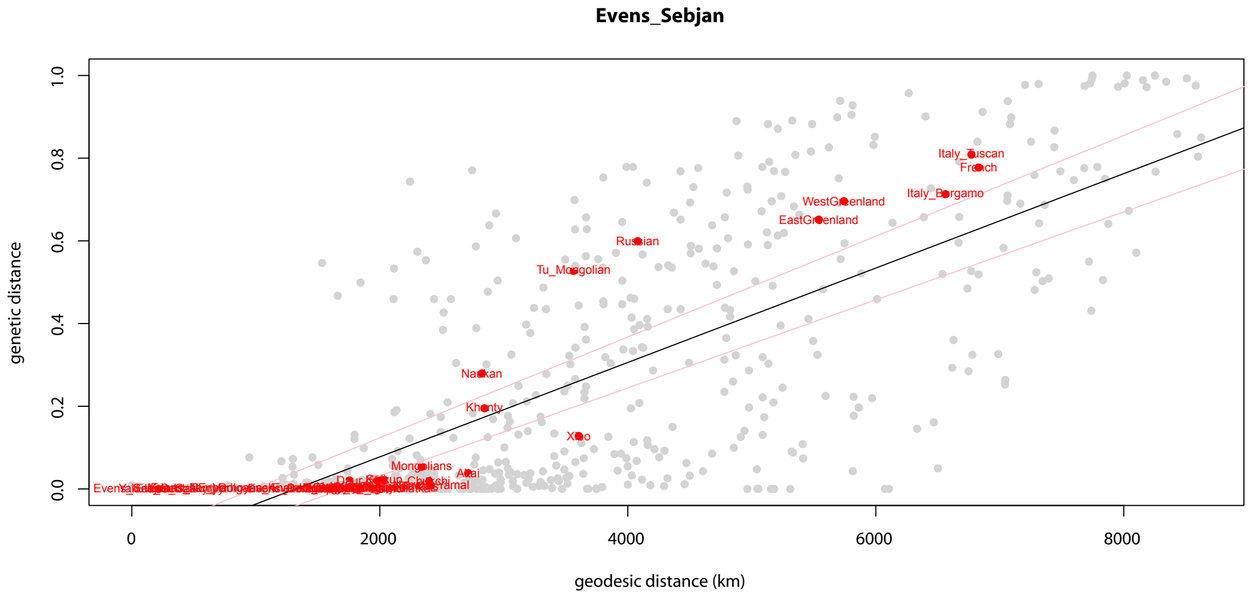
FIGURE S13. Relationship between genetic and geographic distances (continued).

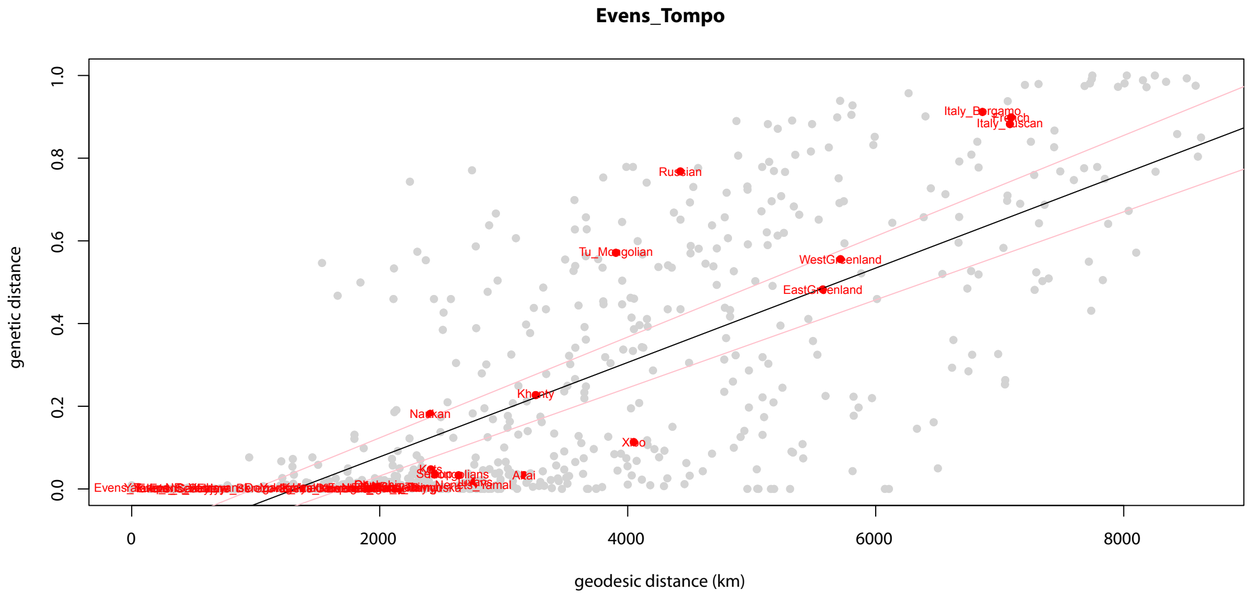


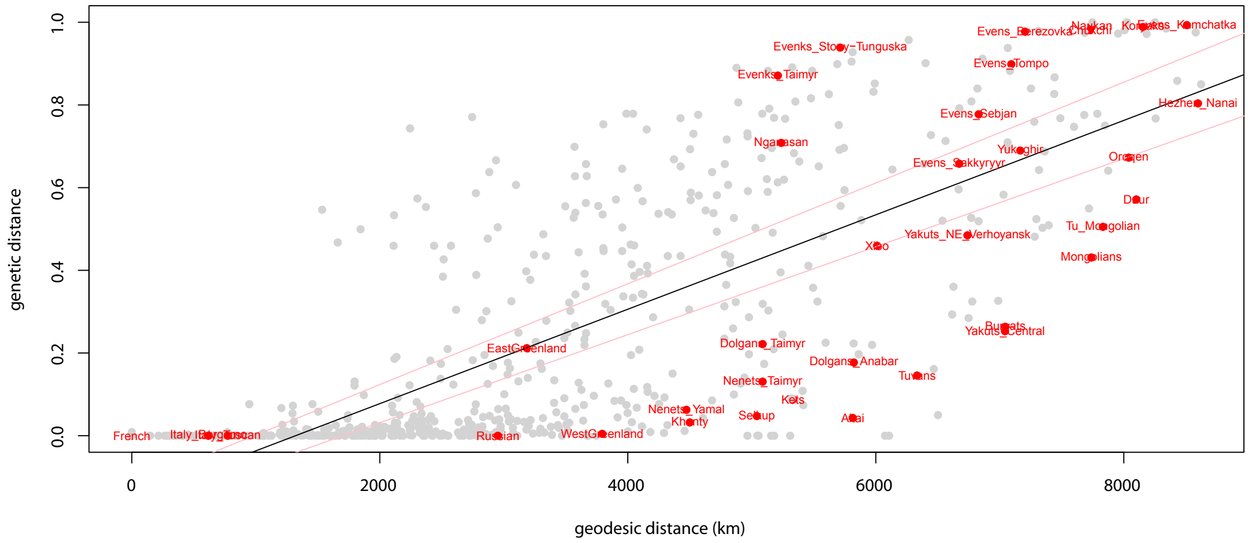












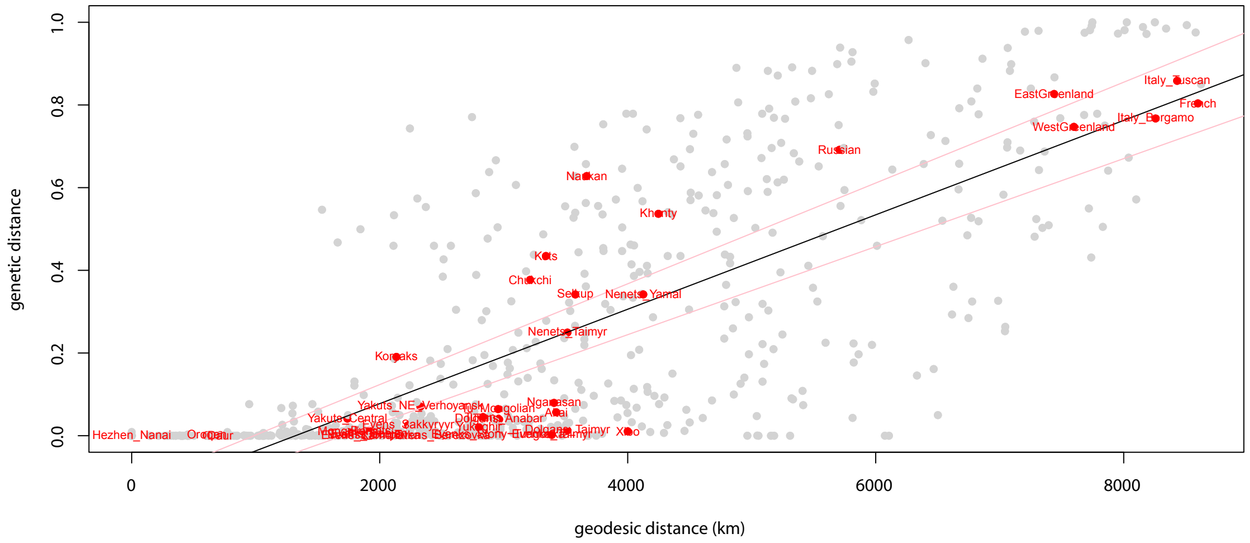
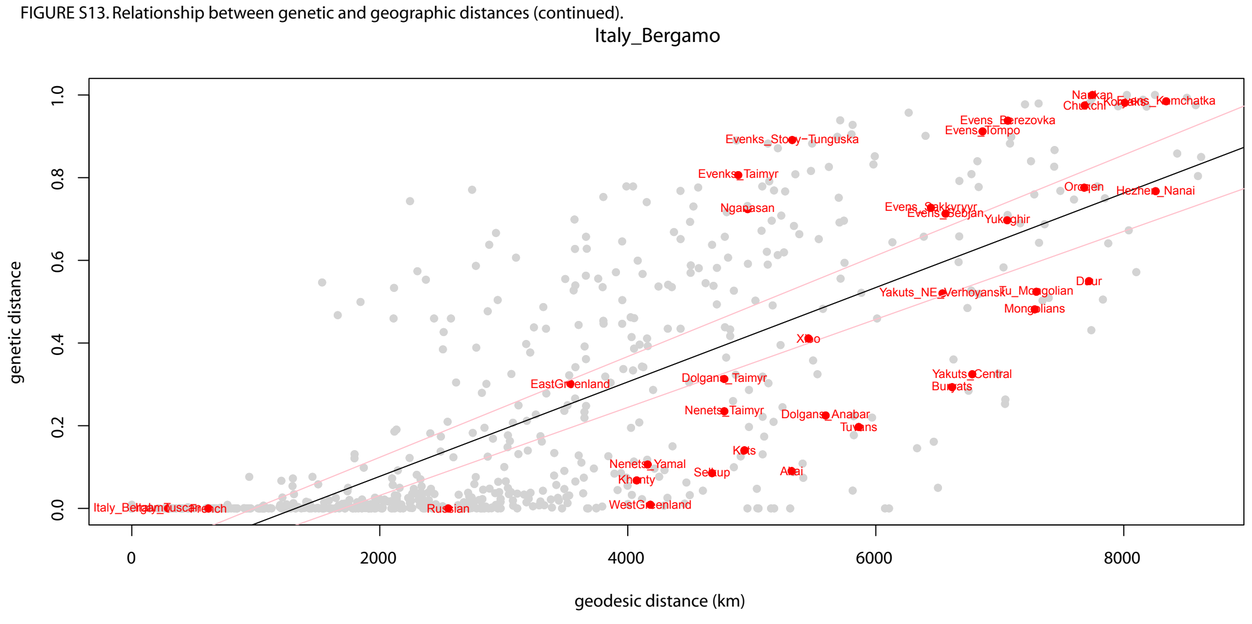
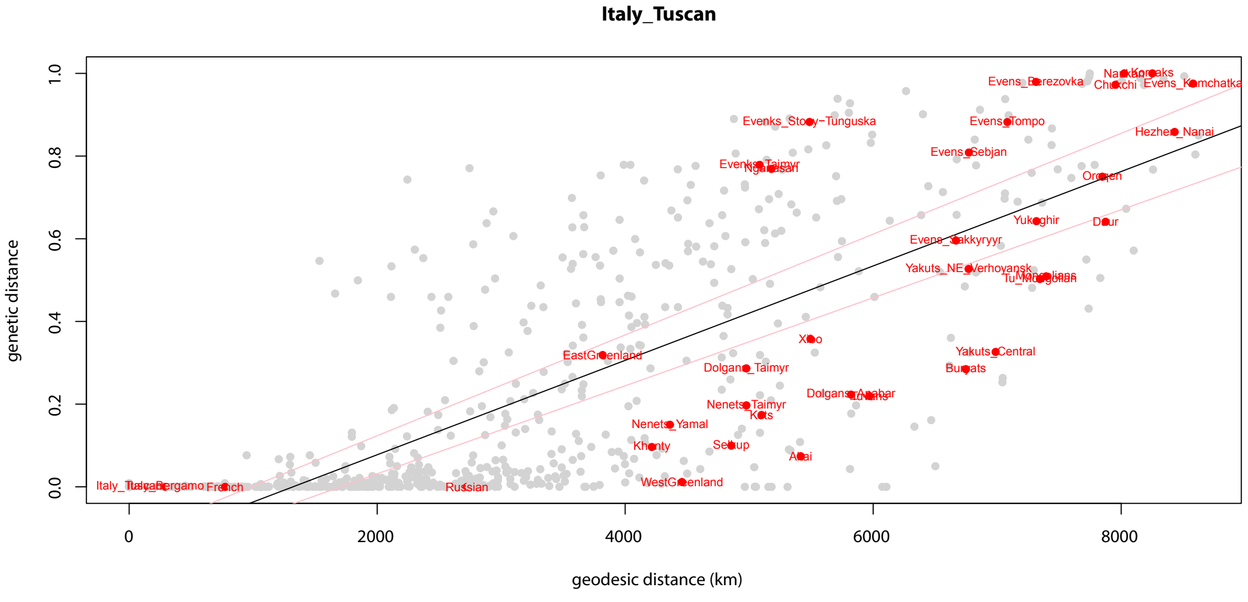
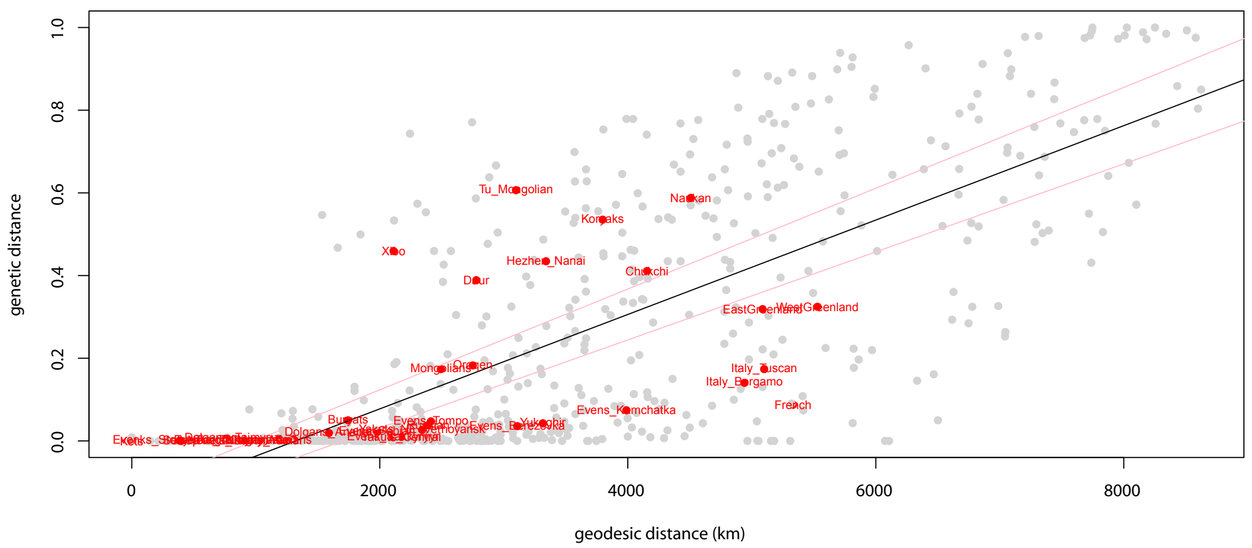
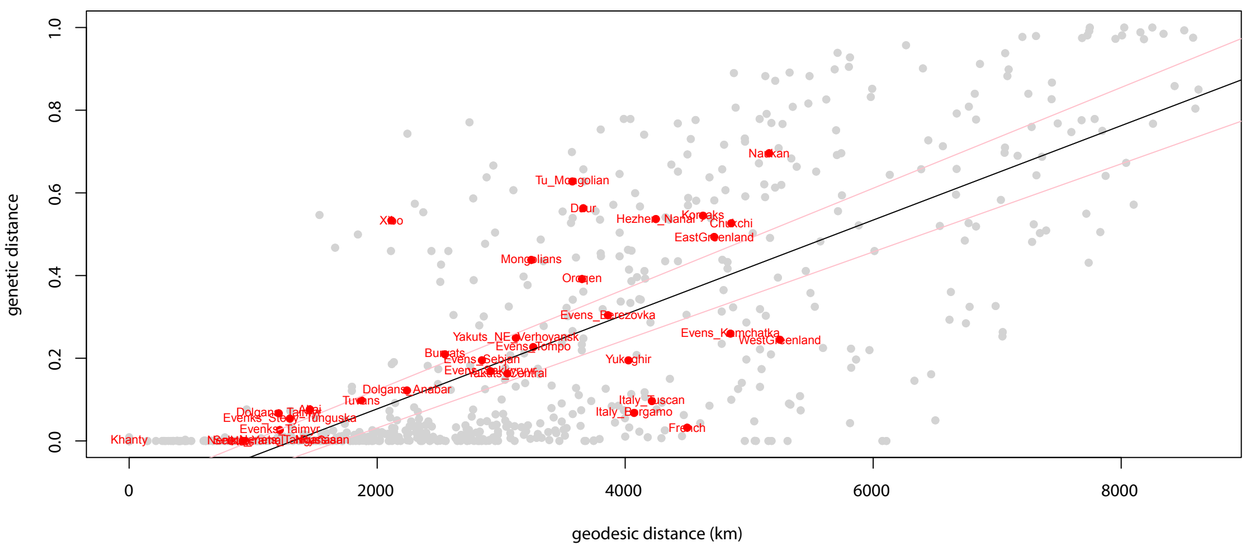


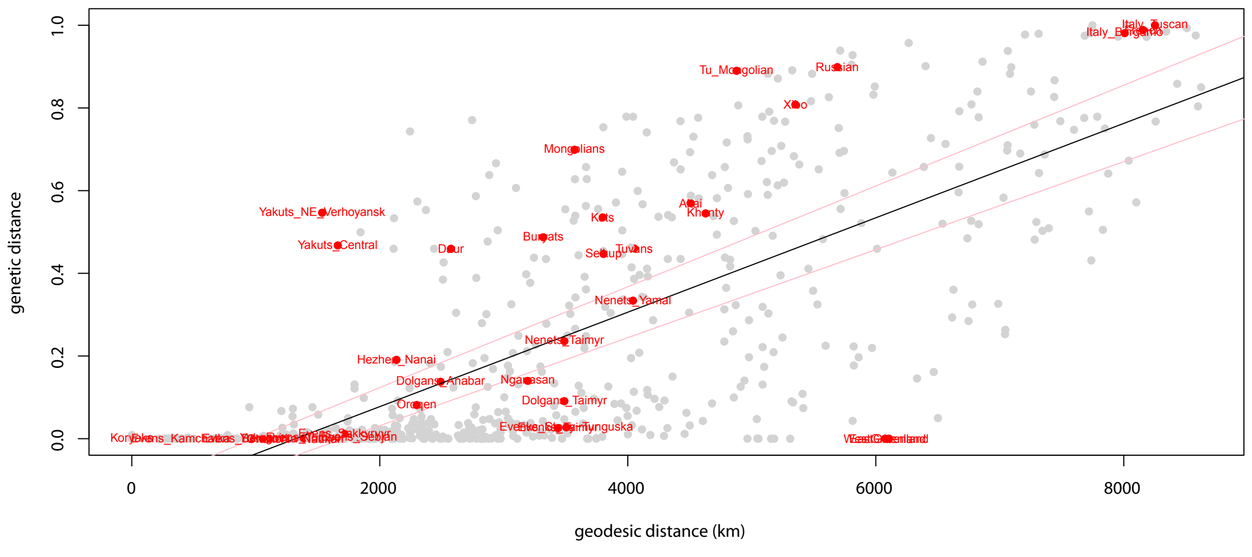
FIGURE S13. Relationship between genetic and geographic distances (continued).

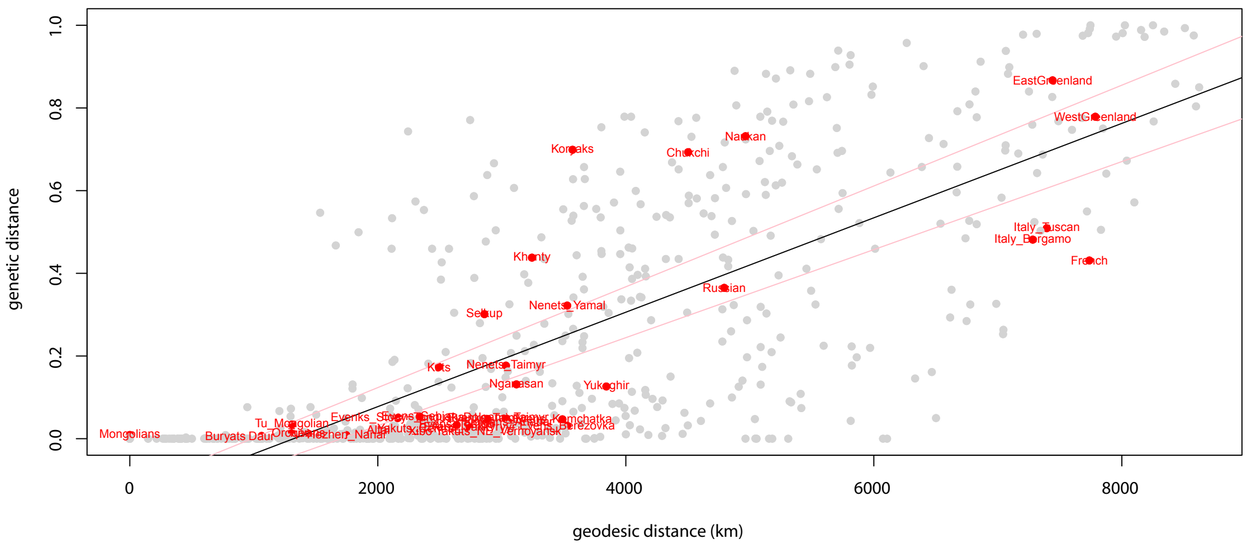


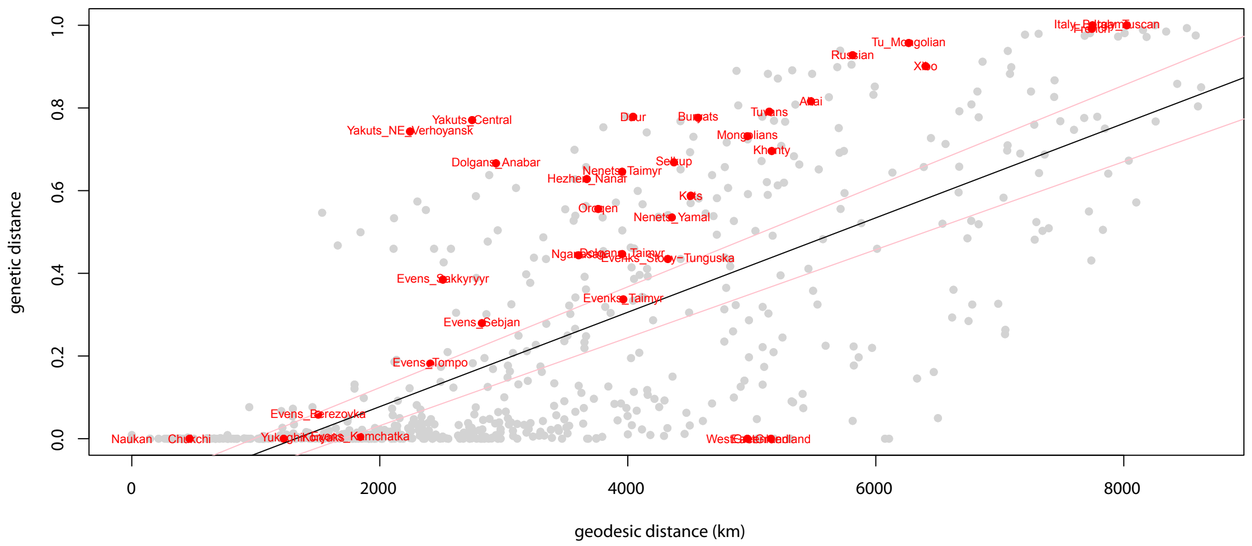












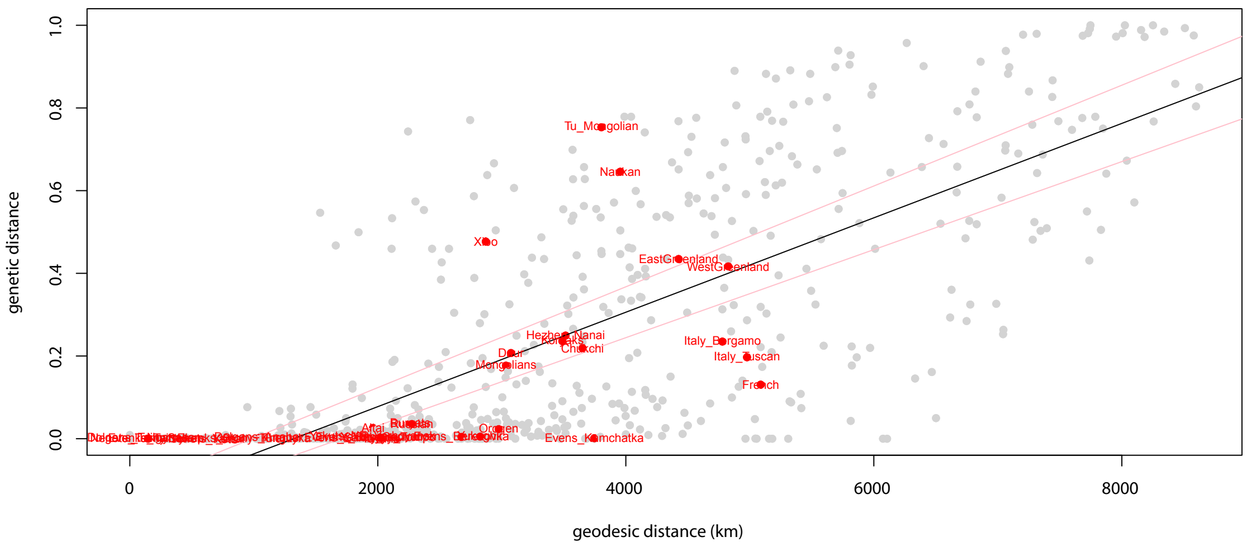
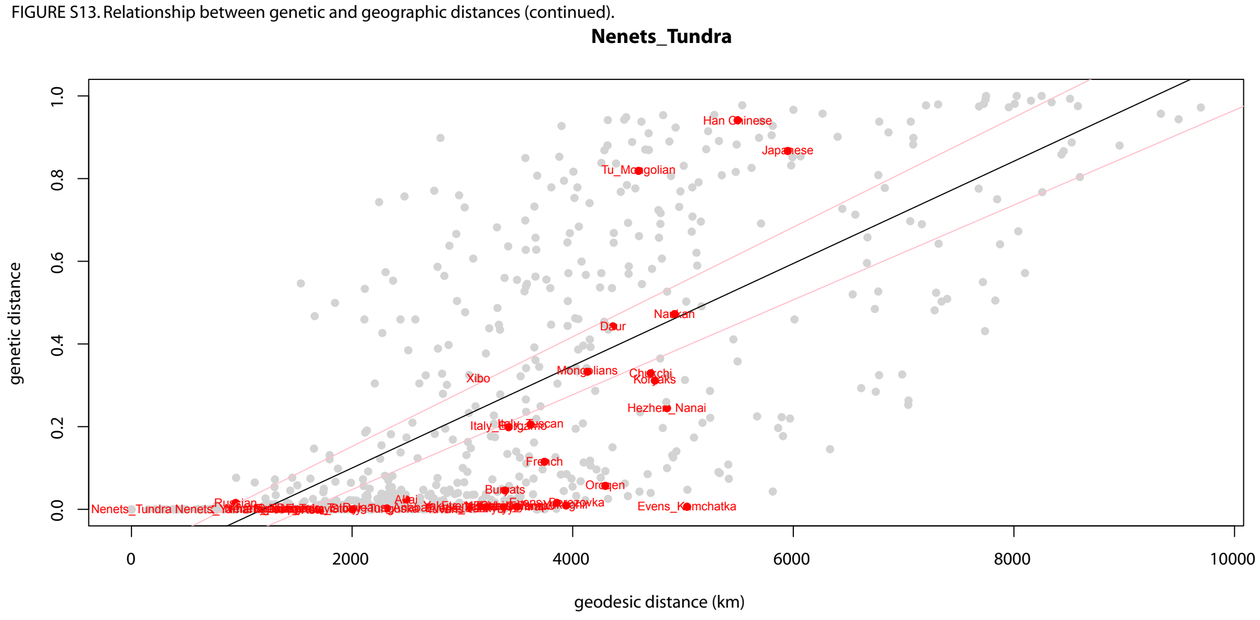
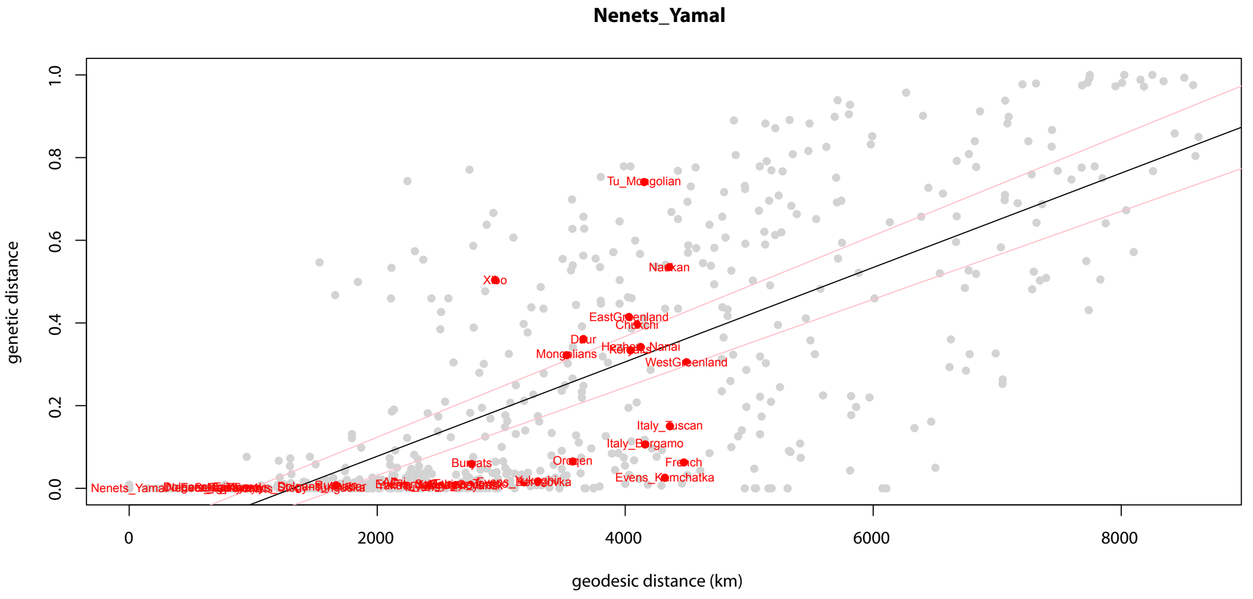
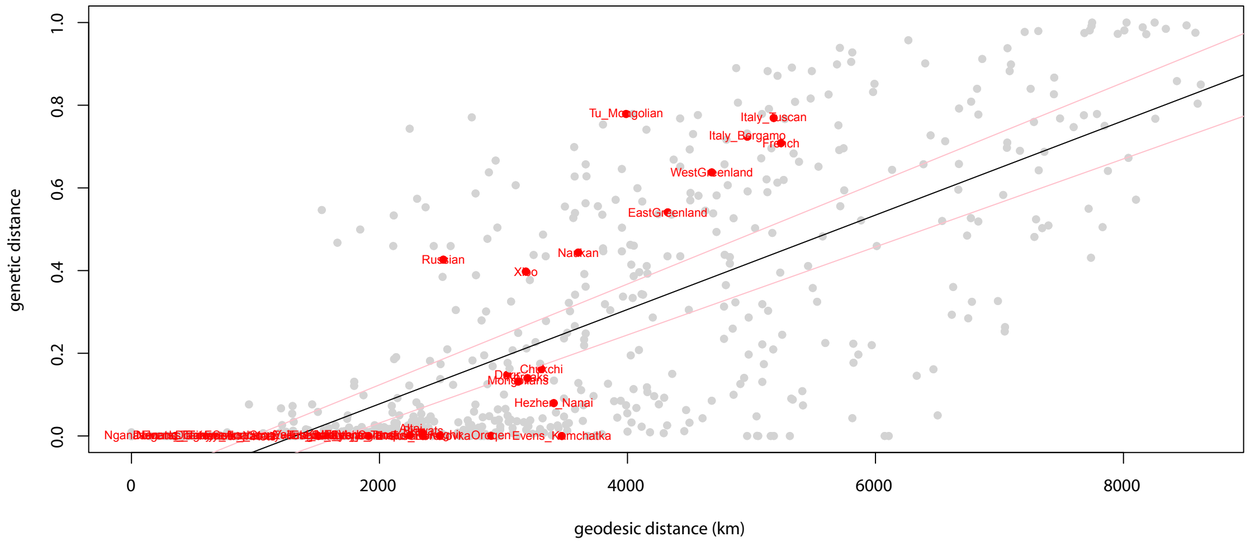
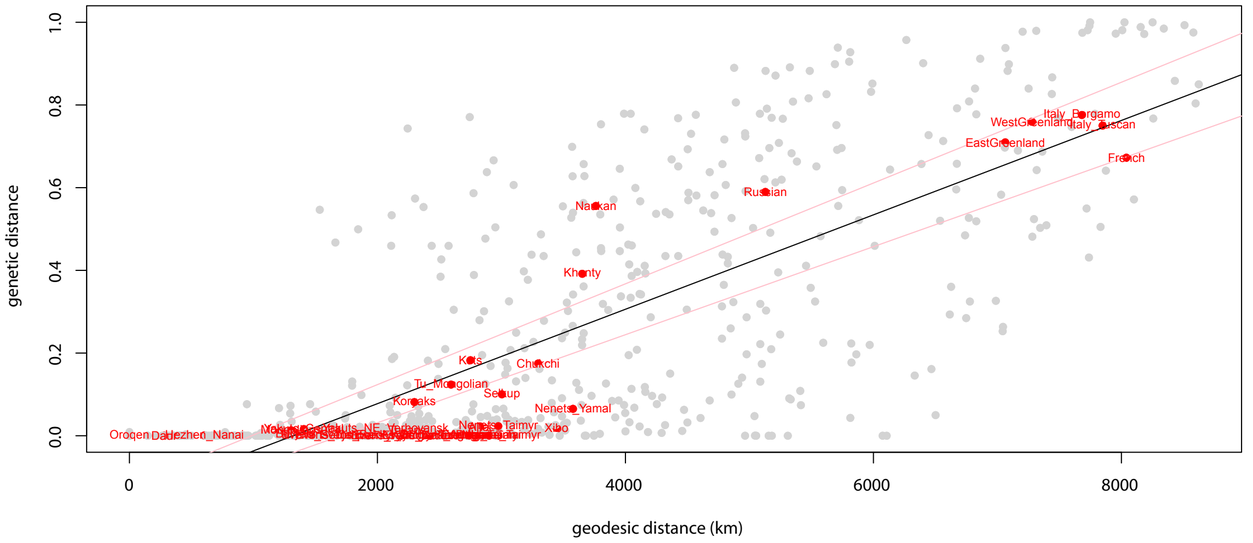


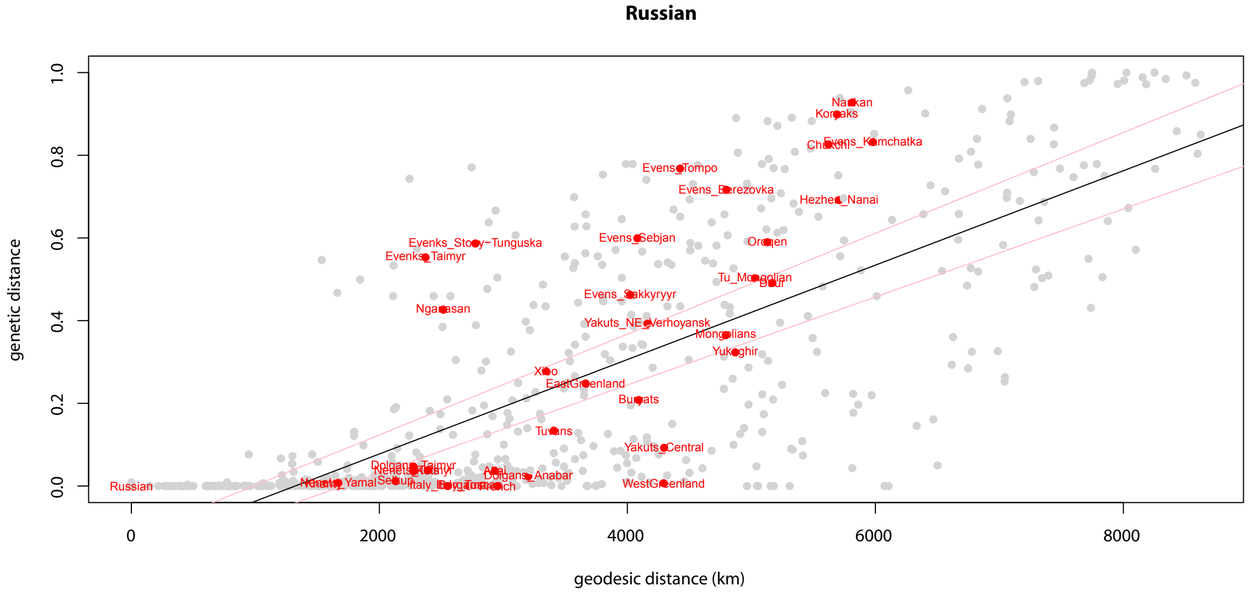
FIGURE S13. Relationship between genetic and geographic distances (continued).

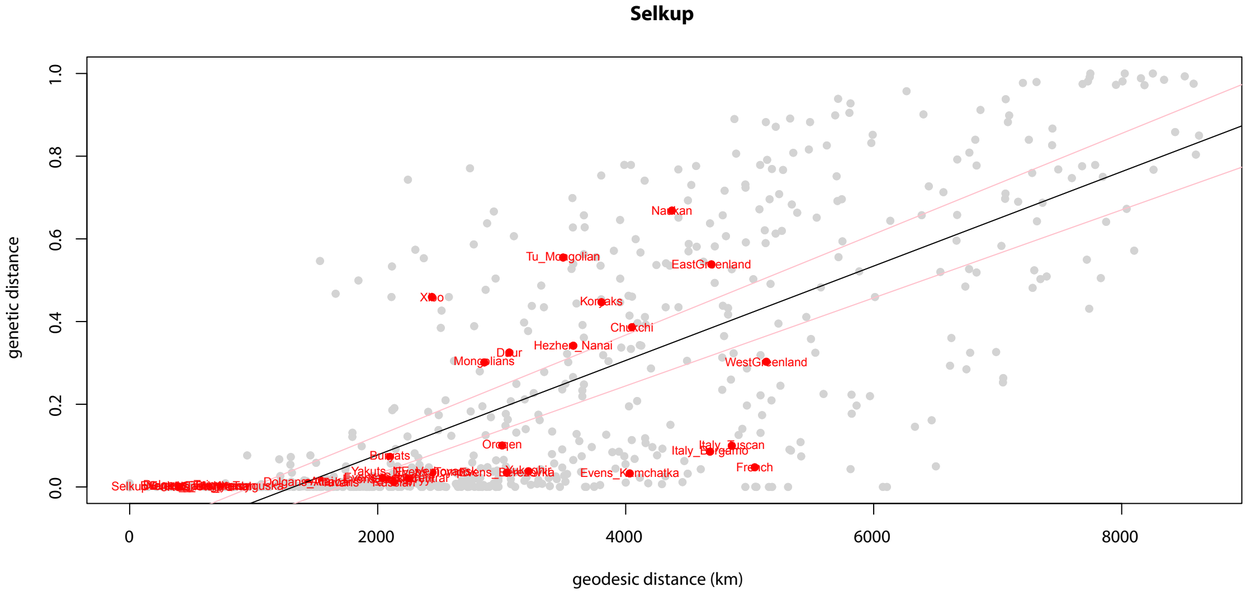


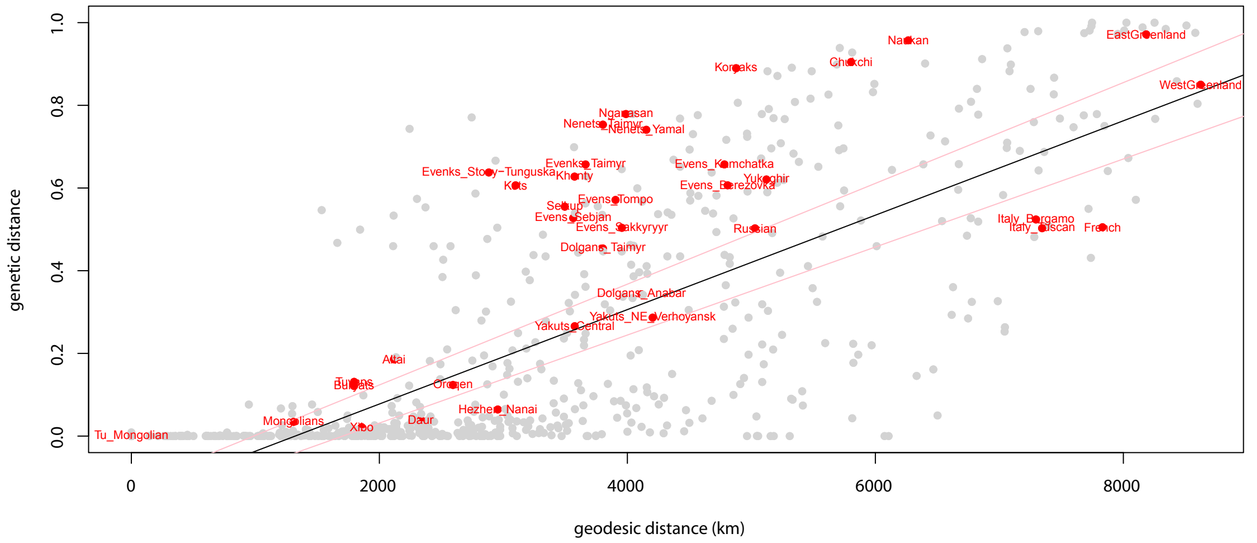












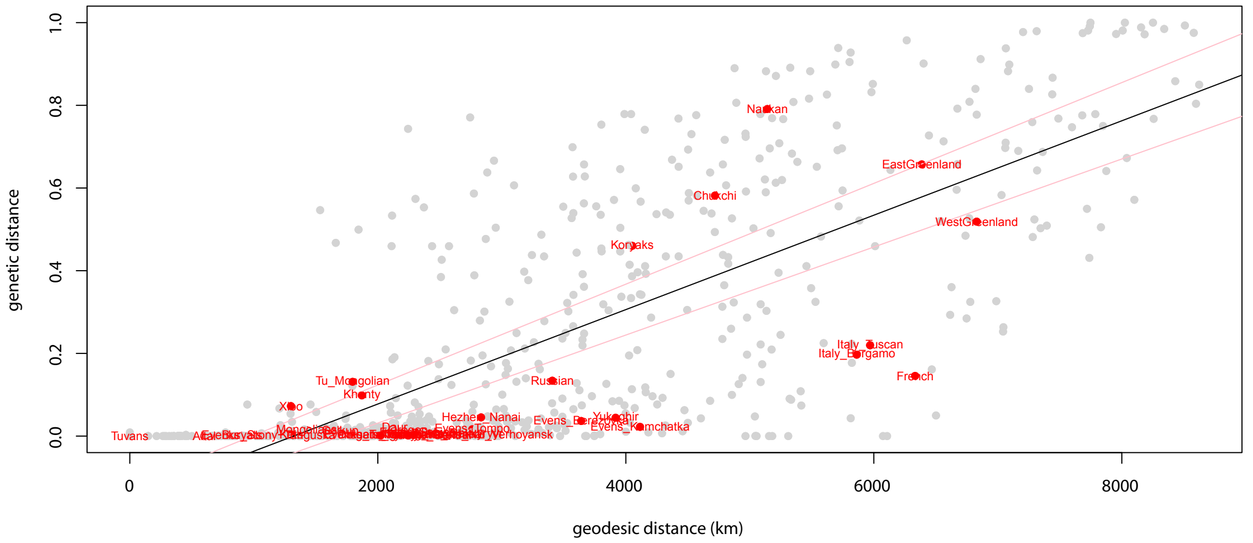
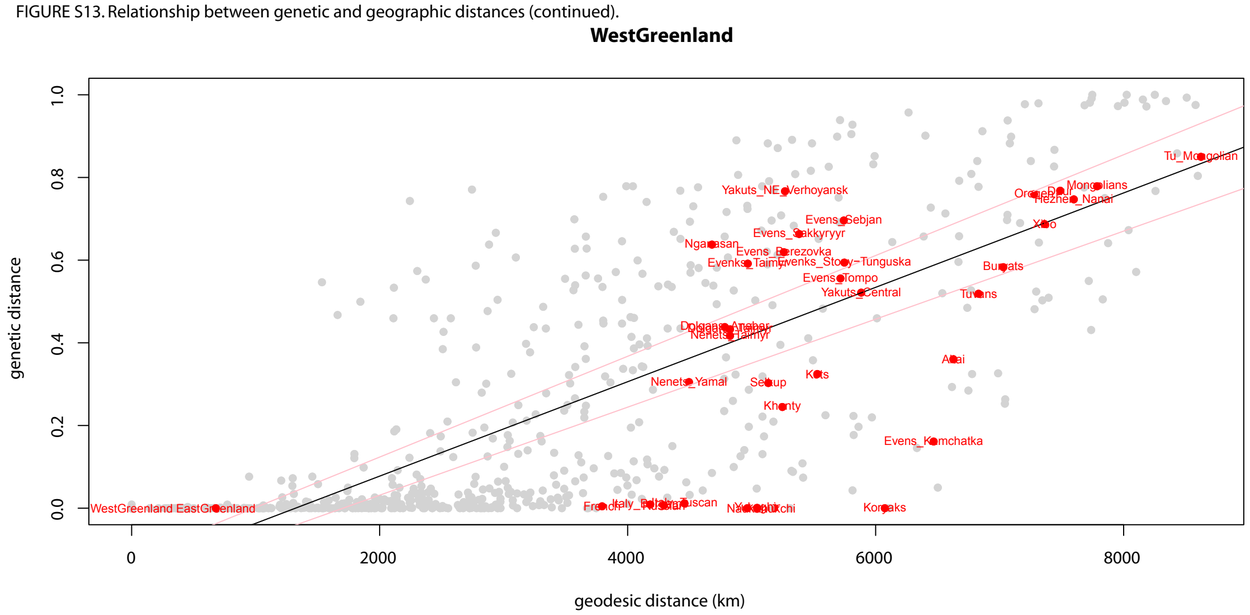
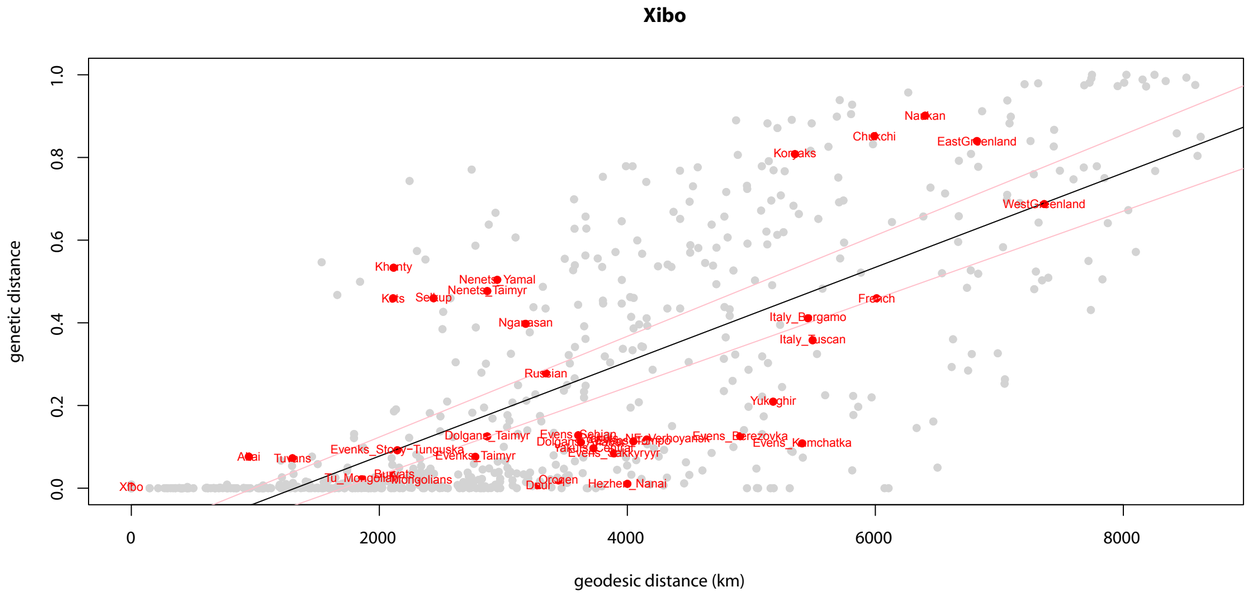
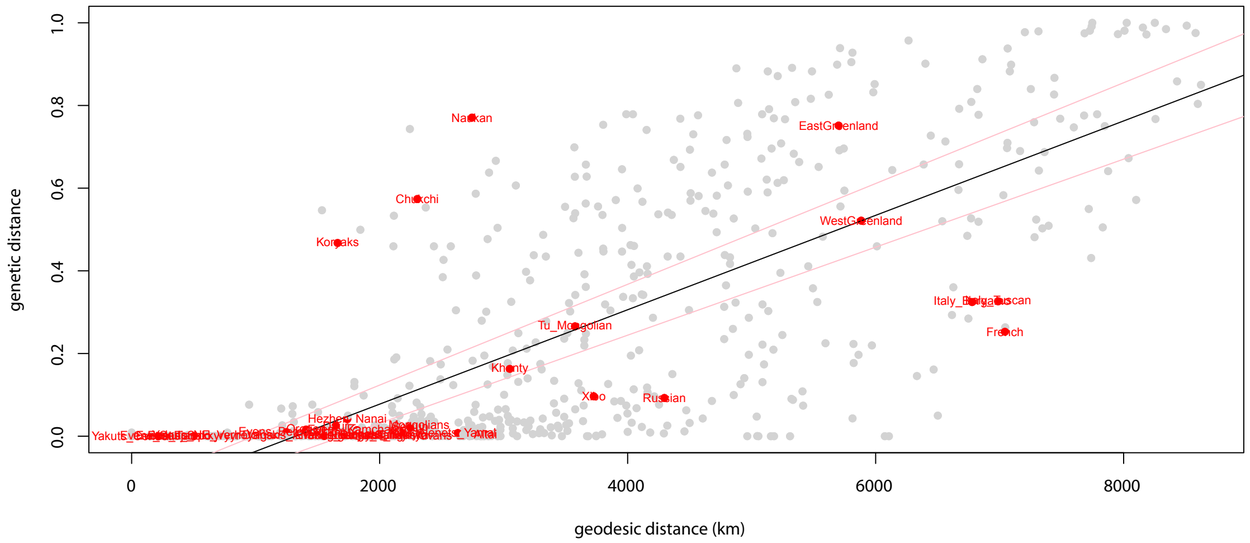
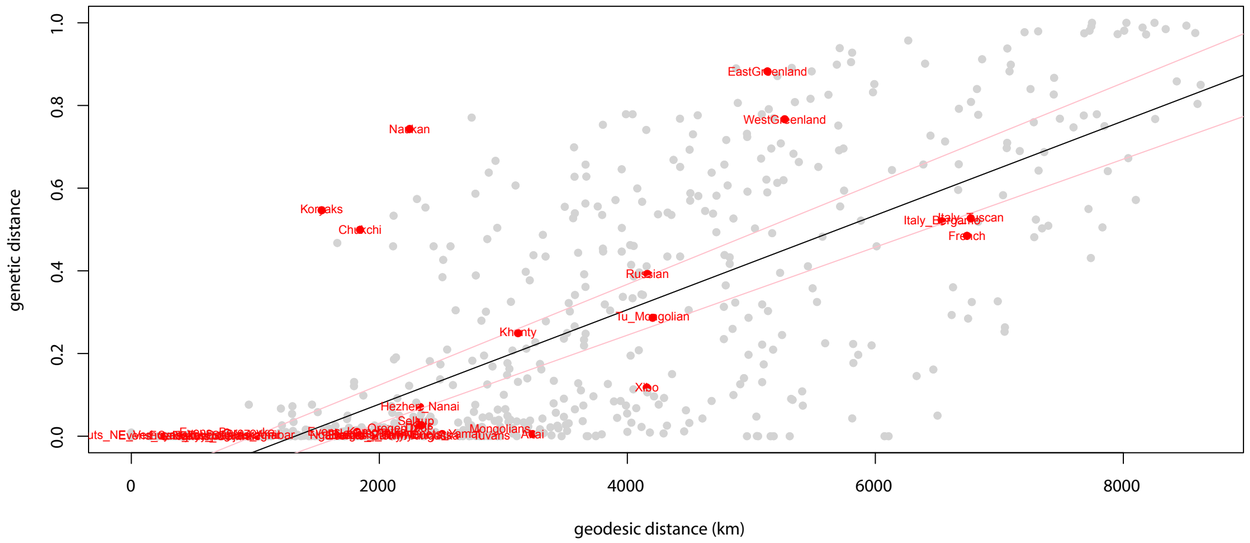


FIGURE S13. Relationship between genetic and geographic distances (continued).









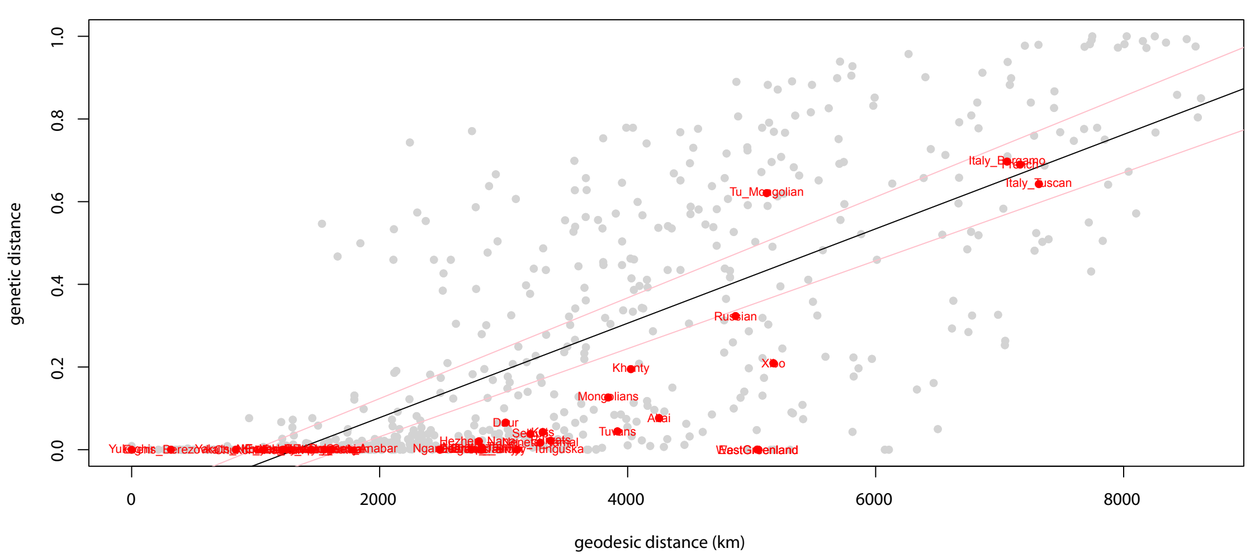


Figure S14

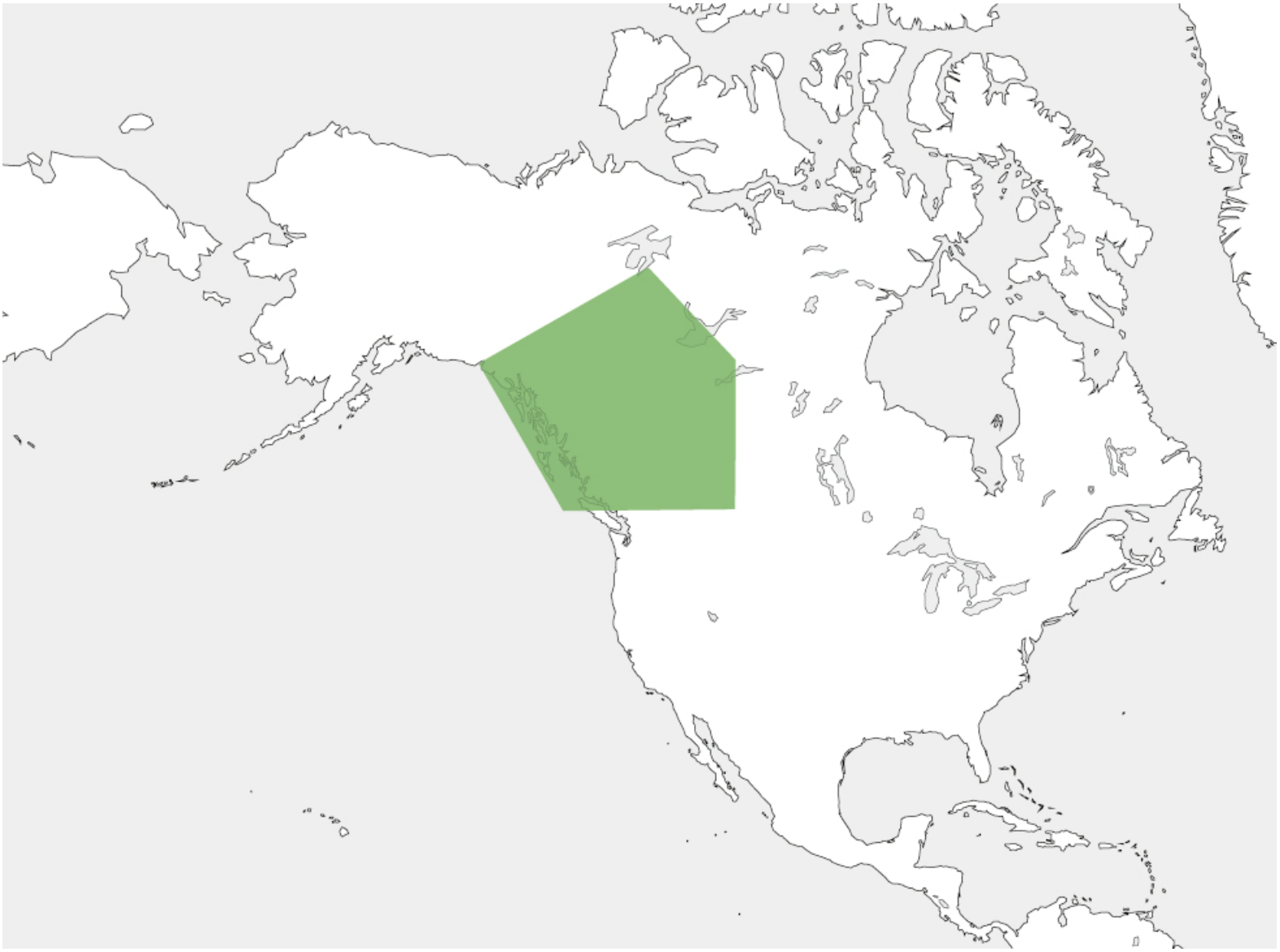


FIGURE S14. Extrapolated origins of the Siberian Naukan Yupik. The extrapolation is based on the results of the regression analysis of the genetic and geographic distances (Figure S13), which revealed that all the genetic distances between the Naukan Yupik and other populations in Siberia are elevated given the geographic distance between them. In order to improve the fit between the genetics and geography, the Naukan Yupik were assigned different sets of geographic coordinates in Alaska, Canada and Greenland, each time recalculating the regression coefficient. The best improvement was achieved when the Naukan Yupik were assigned any set of coordinates in the geographic region of Canada circumscribed here by the shaded area.

Figure S15

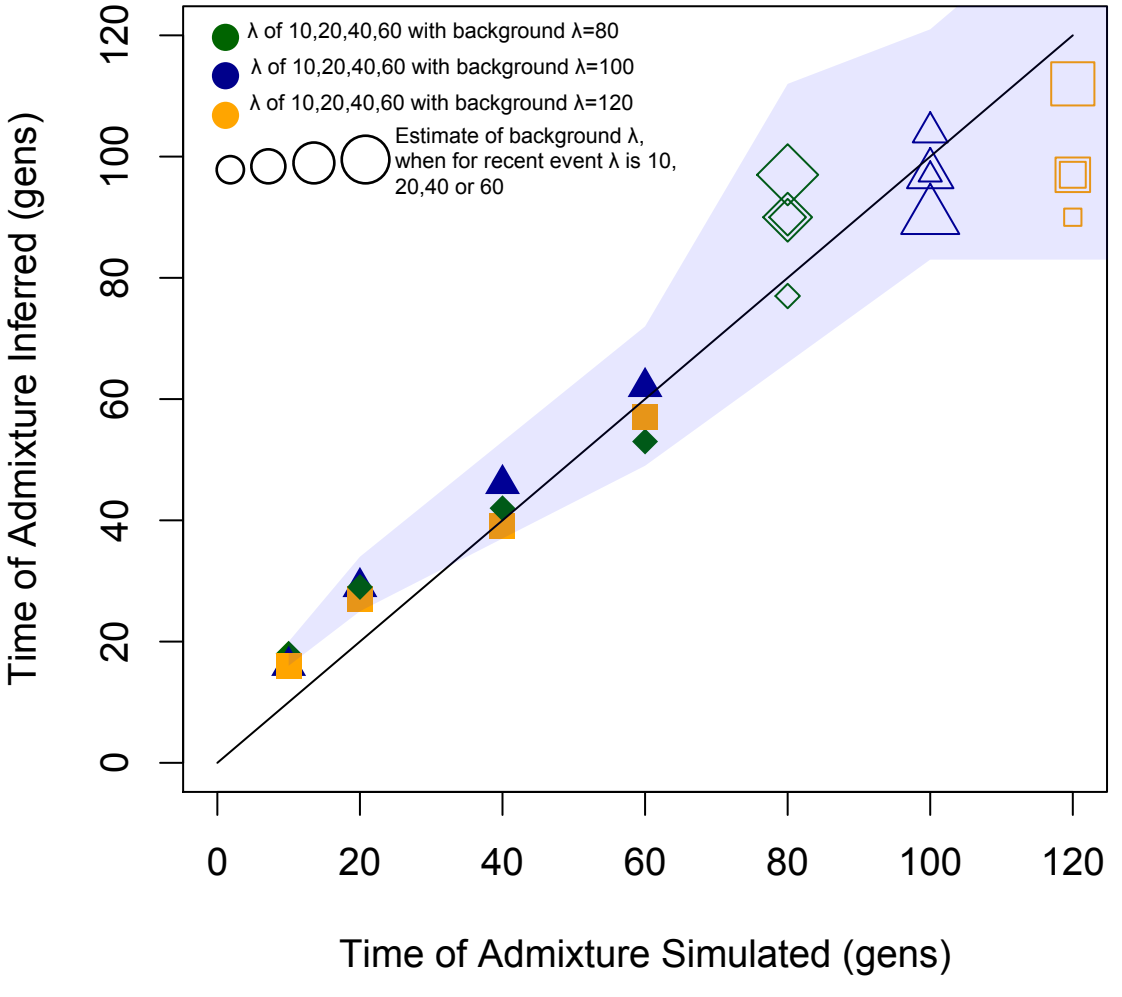


FIGURE S15. Performance of the wavelet transform analysis (adapted to include complex admixture scenarios) in recovering the simulated dates of admixture. The method was applied to 12 artificial datasets with different admixture histories. Each recent admixture time point (? = 10, 20, 40 and 60) was simulated three times, each time using a different admixture background (? = 80, 100 or 120). The shaded area denotes the 95% confidence interval for the inferred dates of admixture.

Figure S16

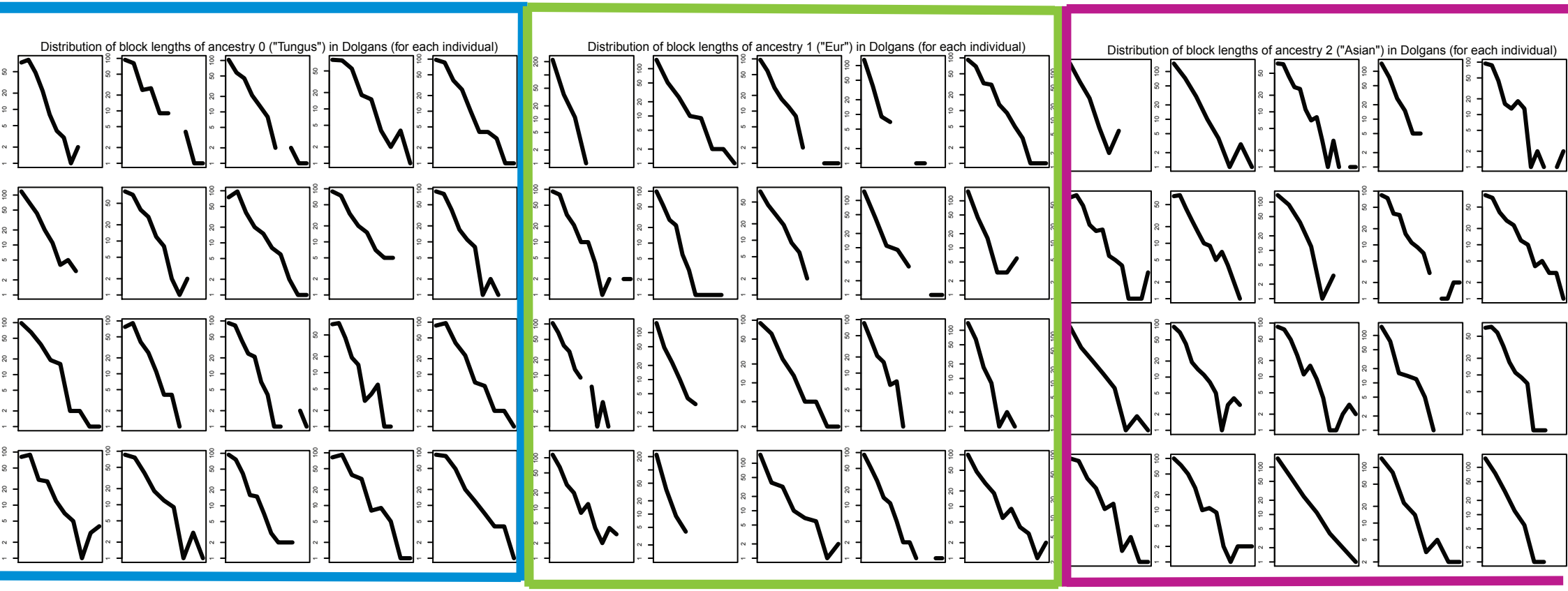


FIGURE S16. Test for excess of small-sized ancestry blocks as inferred by PCAdmix, with the expectation that an excess of short ancestral blocks would indicate erroneous insertions by PCAdmix. Shown is the genomewide distribution of block widths of each ancestry inferred for each Anabar Dolgan individual. Block width in (cM) is shown on the x-axis, while abundance is represented on the y-axis (note that the scale here is logarithmic).

Figure S17

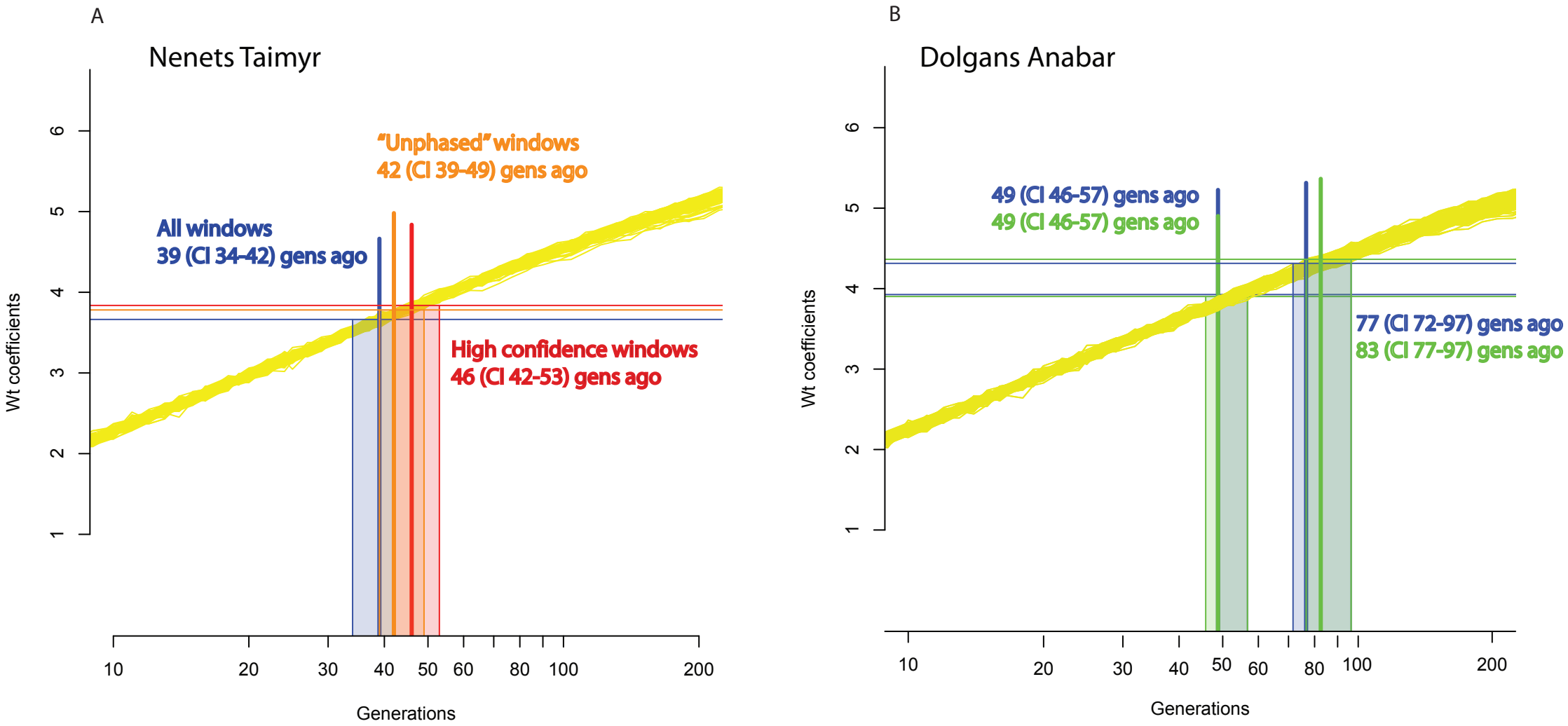


FIGURE S17 (A). Comparison of dates of admixture inferred using a full set of windows (blue) vs. only high confidence windows (red), and on artificially "unphased" data (orange). Dating is based on simulated data from 100 simulations with a 30% migration rate. Each curve represents a single admixed population. Average WT centers calculated for 100 chromosomes drawn at random from each population at exponentially growing time points are plotted as a function of time. Vertical lines indicate the time estimate, and shaded boxes define the confidence intervals. Time estimates are based on the entire unfiltered sample of Taimyr Nenents, and the admixture rate for the simulated data is chosen so as to match the empirical data. (B). Admixture time estimates (for two episodes of admixture) based on data phased with either BEAGLE or SHAPEit. Measurements obtained for a more recent vs. an older admixture episode using the BEAGLE-phased and SHAPEit-phased data are shown by blue and green horizontal lines, respectively. Vertical lines indicate the time estimate (blue font for the BEAGLE-phased, and green font for the SHAPE-it phased data), and shaded boxes define the confidence intervals. Time estimates are based on the entire unfiltered sample of Anabar Dolgans, and the admixture rate for the simulated data is chosen so as to match the empirical data.