Deep History of East Asian Populations Revealed Through Genetic Analysis of the Ainu

Choongwon Jeong,*[,1] Shigeki Nakagome,*[,†,1,2] and Anna Di Rienzo*[,2]
*Department of Human Genetics, University of Chicago, Chicago, Illinois 60615, and †Department of Mathematical Analysis and Statistical Inference, The Institute of Statistical Mathematics, Tokyo, Japan, 190-8562

ABSTRACT Despite recent advances in population genomics, much remains to be elucidated with regard to East Asian population history. The Ainu, a hunter–gatherer population of northern Japan and Sakhalin island of Russia, are thought to be key to elucidating the prehistory of Japan and the peopling of East Asia. Here, we study the genetic relationship of the Ainu with other East Asian and Siberian populations outside the Japanese archipelago using genome-wide genotyping data. We find that the Ainu represent a deep branch of East Asian diversity more basal than all present-day East Asian farmers. However, we did not find a genetic connection between the Ainu and populations of the Tibetan plateau, rejecting their long-held hypothetical connection based on Y chromosome data. Unlike all other East Asian populations investigated, the Ainu have a closer genetic relationship with northeast Siberians than with central Siberians, suggesting ancient connections among populations around the Sea of Okhotsk. We also detect a recent genetic contribution of the Ainu to nearby populations, but no evidence for reciprocal recent gene flow is observed. Whole genome sequencing of contemporary and ancient Ainu individuals will be helpful to understand the details of the deep history of East Asians.

KEYWORDS Jomon; Tibet; migration; demography

THE rapid development of genomic technologies has greatly enhanced our understanding of the history of modern human dispersal out of Africa and the peopling of different continents (Li *et al.* 2008; Green *et al.* 2010; Reich *et al.* 2010). However, the history of populations in East Asia, including Siberia, remains poorly understood even though they account for a large fraction of the human population (Stoneking and Delfin 2010; Oota and Stoneking 2011). There is still no clear consensus regarding basic questions such as when, where, and how many times modern humans migrated into East Asia and Siberia. For example, several previous studies based on contemporary samples concluded that one migration from south to north could explain the genetic structure of East Asians (Li *et al.* 2008; HUGO Pan-Asian SNP Consortium 2009). However, recent studies indicate that this scenario is too simplistic: a recent study detected western Eurasian ancestry in an individual in southern Siberia 24,000 years ago as well as a substantial contribution of this ancestry to the gene pool of Native Americans (Raghavan *et al.* 2014), and several studies detected a clear genetic differentiation between northeast Siberians and central-south Siberians, with the latter being more closely related to northeast Asians (Rasmussen *et al.* 2010; Fedorova *et al.* 2013). Indigenous high-altitude populations of the Tibetan plateau provide another example of East Asian populations that do not fit into the simple one migration hypothesis. Tibetans and Sherpa show a divergent history from lowland East Asian populations such as Han Chinese (Jeong *et al.* 2014), and their adaptive haplotype spanning the *EPAS1* (endothelial PAS domain-containing protein 1) gene shares its ancestry with that of an archaic hominin (Huerta-Sánchez *et al.* 2014).

Considering robust evidence of human habitation in Arctic Siberia before the last glacial maximum (LGM) (Pitulko *et al.* 2004), it is possible that there were multiple expansions (from south to north) and contractions (from north to south) of human populations in mainland East Asia and Siberia over a long period of time, generating a complex pattern of genetic

doi: 10.1534/genetics.115.178673
Manuscript received May 27, 2015; accepted for publication October 19, 2015; published Early Online October 21, 2015.
Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.178673/-/DC1.
[1] These authors contributed equally to this work.
[2] Corresponding authors: Department of Human Genetics, Cumming Life Science Center, 920 East 58th St., Room 424, University of Chicago, Chicago, IL 60615. E-mail: dirienzo@bsd.uchicago.edu; and Department of Mathematical Analysis and Statistical Inference, The Institute of Statistical Mathematics, Tokyo, Japan, 190-8562. E-mail: snakagome@bsd.uchicago.edu

relationships among contemporary populations. Population isolates frequently provide critical information to understand the genetic structure of surrounding populations with more complex histories. For example, it has been proposed that Sardinians are key to understanding the changes in population structure in mainland Europe with the arrival of Neolithic farmers (Keller *et al.* 2012; Skoglund *et al.* 2012). The Onge people from the Andaman Islands have also been proposed as the best contemporary representatives of the "Ancestral South Indian" ancestry (Reich *et al.* 2009; Moorjani *et al.* 2013). Therefore, unraveling the genetic history of population isolates in East Asia and Siberia may provide new insights into the initial colonization of these regions.

The Ainu people are an indigenous population of Hokkaido, a northern island in the Japanese archipelago, and of the southern part of Sakhalin islands (Figure 1). They have been proposed by archaeologists, linguists, and geneticists to be the direct descendants of prehistoric Japanese hunter–gatherers, associated with the Jomon pottery culture, dating back to 16,500 years before the present (Hanihara 1991; Habu 2004). The dual structure model for the Japanese population (Hanihara 1991) envisions that the contemporary Japanese are a mixture of two distinct genetic sources, one from the indigenous Jomon hunter–gatherers and the other from East Asian rice farmers who first migrated into the archipelago ~2,300 years ago ("Yayoi culture"). Genetic data clearly support this model in two main regards. First, the genetic profile of the mainland Japanese reveals a strong signature of admixture, best modeled as a mixture of Ainu-related ancestry and continental East Asians (Jinam *et al.* 2012; Nakagome *et al.* 2015). Second, the Ainu are genetically closer to the Ryukyuan, who live in the southernmost islands of the Japanese archipelago, than to the mainland Japanese (Tajima *et al.* 2002; Matsukusa *et al.* 2010; Jinam *et al.* 2012; Koganebuchi *et al.* 2012). This suggests that inhabitants of the northern and southernmost parts of the archipelago were genetically most isolated from the incoming farmers, who first arrived in the central part of the archipelago. However, the origin of the Ainu people in the context of eastern Eurasian population history has not been thoroughly investigated using genome-scale variation data.

In this study, we investigated the genetic relationship of the Ainu with surrounding East Asian and Siberian populations, using genome-wide genotype data, and found that the Ainu gene pool is basal to all other East Asian populations. In addition, the Ainu show unusual patterns of excess genetic affinity with low-altitude East Asians and northeast Siberians, as well as signatures of genetic adaptations to their local environments and maritime hunter–gatherer life style.

Materials and Methods

Genotype data

We obtained genome-wide genotyping data of worldwide populations from several previous publications. First, we obtained genotype data of 36 Ainu individuals from a previous study, typed on the Affymetrix Genome-Wide Human SNP 6.0 array (Jinam *et al.* 2012). We estimated genetic relatedness for all pairs of Ainu individuals using PLINK v1.07 (Purcell *et al.* 2007), with 396,552 autosomal SNPs with minor allele frequency ≥0.05. We randomly removed one individual from each pair with coefficient of relationship (r) > 0.1875, which corresponds to relatedness between first cousins ($r = 0.125$) and half siblings ($r = 0.25$). This step removed 11 individuals, involved in 17 of 630 pairs, some of which correspond to first degree relatives (Supporting Information, Figure S1). Second, we used genotype data of 1963 individuals from 183 worldwide populations, "Affymetrix Human Origins Fully Public Dataset" (Lazaridis *et al.* 2014), typed on the Affymetrix Axiom Genome-Wide Human Origins 1 array (Patterson *et al.* 2012). We removed three individuals with genotype missing rate >5% and kept SNPs with missing rate not exceeding our criteria in all 152 populations with ≥5 individuals. Specifically, we allowed one missing genotype in populations with <20 individuals, and two missing genotypes in populations with ≥20 individuals. After this filtering, an overlap of 103,218 SNPs between the Ainu and Human Origins datasets was used for the majority of analyses ("WA" dataset, for worldwide and Ainu data). Third, we overlapped the "WA" dataset with genotype data for the Sherpa and Tibetans genotyped on Illumina arrays. Specifically, we took 21 Sherpa individuals described as the "high-altitude proxy" samples in a previous study (Jeong *et al.* 2014) and 30 Tibetans from near Lhasa, Tibet Autonomous Region in China (Wang *et al.* 2011). We used the overlapping 45,513 SNPs ("WHA" dataset, for worldwide, high-altitude and Ainu data) for most of the analyses along with the WA dataset to investigate the genetic relationships of the Ainu and the high-altitude East Asians. Fourth, we overlapped the WHA dataset with the genotype data of two Nivkh individuals (Fedorova *et al.* 2013) for additional genetic clustering analysis. Fifth, for genome scans of positive selection in the Ainu, we overlapped the Ainu genotype data with the 1000 Genomes Project phase 3 dataset, downloaded from https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference. This includes 540,304 SNPs ("1KG-Ainu" dataset). Finally, we overlapped the 1KG-Ainu dataset with available high-coverage Illumina whole genome sequences of contemporary humans and archaic hominins. For this, we retrieved genotype calls of high-coverage Denisovan (Meyer *et al.* 2012) and Altai Neandertal (Prüfer *et al.* 2014), in variant call format (VCF). Chimpanzee alleles were extracted from the chimpanzee genome assembly Pan_troglodytes-2.1.4 (Pantro4), using LiftOver tool (http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/liftOver) to convert coordinates between the human reference sequence (GRCh37) and Pantro4. We also obtained short read data of 13 individuals (1–2 individuals from Han, Dai, Sherpa, Yoruba, Karitiana, Sardinian, Papuan, and Australian aborigine) from previous studies (Meyer *et al.* 2012; Jeong *et al.* 2014; Prüfer *et al.* 2014). Short reads were aligned to the human reference (GRCh37) using BWA backtrack 0.7.4-r385 (Li and

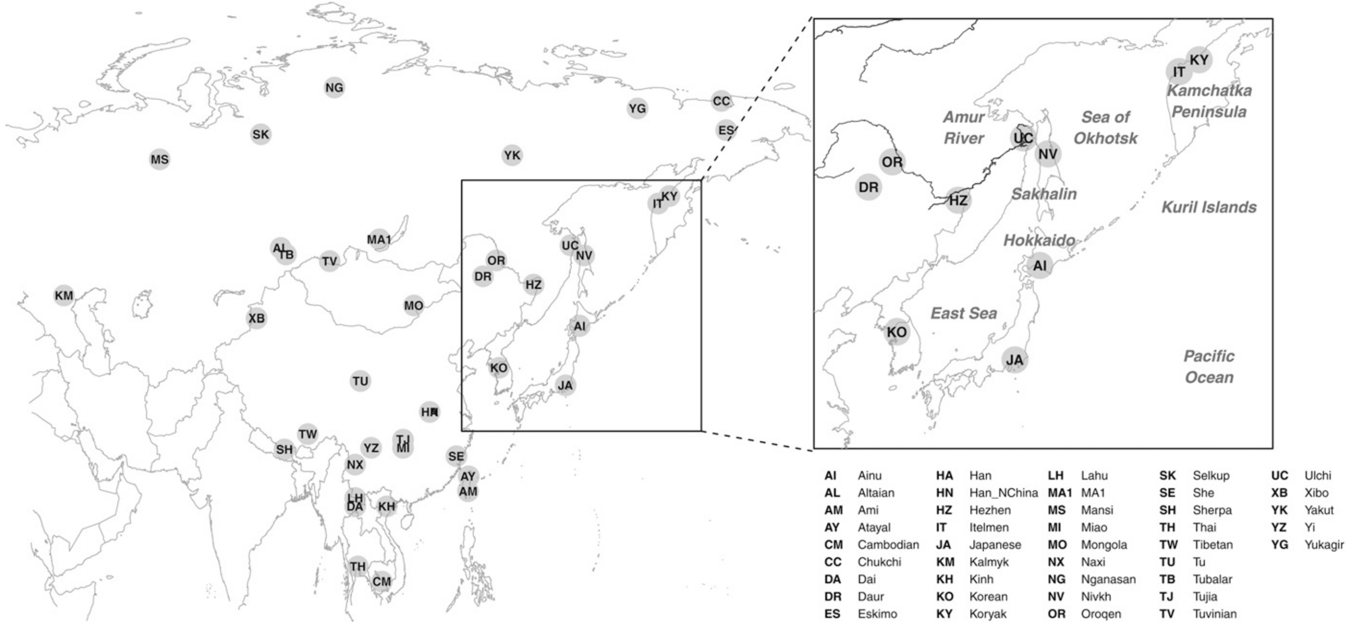


Figure 1 Geographic location of East Asian and Siberian population samples used in this study. The zoom-in plot highlights the region around the Japanese archipelago and the Sea of Okhotsk.

Durbin 2009) with "-q 15" option, duplicate removed with Picard tool v1.98 (http://broadinstitute.github.io/picard/), locally realigned around indels, and base quality score recalibrated using the Genome Analysis Toolkit (GATK) v2.8-1 (McKenna *et al.* 2010; Depristo *et al.* 2011) following the "best practice workflows" from GATK (Auwera *et al.* 2013). We kept only properly paired nonduplicate reads with phred-scaled mapping quality ≥30 using Samtools v1.2 (Li *et al.* 2009). For each sample, we called genotypes across all sites using the GATK UnifiedGenotyper module, based on bases with phred-scaled quality score ≥30, and kept sites only if phred-scaled quality score was ≥50. We combined genotype calls of all modern and archaic individuals using bcftools v1.1 (Li *et al.* 2009), removed sites if they had any missing genotype, showed more than one alternative allele (including 1KG data and chimpanzee alleles), or located within either human CpG islands (Wu *et al.* 2010) or human/chimpanzee repeat regions. Finally, we intersected these data with the Ainu genotype data and removed strand ambiguous (A/T or G/C) SNPs, leaving a total of 322,011 SNPs ("CND-1KG-Ainu" dataset, CND for Chimpanzee-Neandertal-Denisova).

Assessment of genetic homogeneity within the Ainu individuals

To determine if 25 unrelated Ainu individuals represent a homogenous gene pool, we performed a model-based genetic clustering analysis using ADMIXTURE v1.22 (Alexander *et al.* 2009). For this analysis, we included 25 unrelated Ainu individuals and sets of 30 randomly chosen individuals per each 1KG East Asian population from the 1KG-Ainu dataset. We removed SNPs with minor allele frequency of <0.01 and randomly removed one from each pair of SNPs with $r^2 > 0.2$ ("–indep-pairwise 200 25 0.2" option in PLINK), leaving 84,462 SNPs for the analysis. We ran 50 replicates with random seeds for the number of clusters (K) from 2 to 6 and chose a run with the maximum log likelihood for each K. The optimal value $K = 2$ was chosen based on its lowest fivefold cross validation error. We also performed a principal component analysis (PCA) of 1KG East Asians and the Ainu individuals, as implemented in the smartpca program in the EIGENSOFT package v4.1 (Patterson *et al.* 2006; Price *et al.* 2006).

We further estimated admixture time in the Ainu, using a decay of weighted admixture linkage disequilibrium (LD) as implemented in ALDER v1.03 (Loh *et al.* 2013). For this, we performed a two-reference ALDER analysis, using 1KG JPT (Japanese in Tokyo, Japan) and 12 Ainu with 100% Ainu ancestry in the ADMIXTURE analysis ("Ainu" in Figure S2A) as references and the other 10 Ainu as the target ("Ainu2" in Figure S2A). Three individuals who clustered together with mainland Japanese ("Ainu3" in Figure S2) were excluded from the analysis. We also ran ALDER with all 22 Ainu individuals as the target population, with SNP loadings for the Ainu–Japanese cline in PCA (PC1 in Figure S2B) as a weight function instead of specifying reference populations, to check if our split of Ainu individuals into two groups generates a bias in estimation. For both analyses, we applied bin size of 0.025 cM.

Because both analyses above suggested a recent admixture (12.3 and 11.6 generations, respectively) and we are interested in investigating the ancient history of the Ainu (Figure S3), we focused on the 12 individuals with no mainland Japanese ancestry (Ainu in Figure S2).

Population clustering and TreeMix analyses

We conducted a genetic clustering analysis of worldwide populations, with a subset of the WHA dataset, including all East Asian and Siberian individuals, using ADMIXTURE v1.22. For each dataset, we ran 50 replicates with random seeds for the numbers of clusters (K) from 2 to 9 and chose a run with the maximum log likelihood for each K. In all analyses, 5–10 best runs for each K had log likelihoods ranging within a difference of 1, supporting a convergence of the best runs. We chose the optimal K for each dataset by taking one with the lowest fivefold cross-validation error.

Then, we built a consensus tree of 15 worldwide populations representing major ancestry components inferred from the ADMIXTURE analysis. For this, we first removed all populations showing a negative three-population (f_3) statistic (Reich *et al.* 2009; Patterson *et al.* 2012) to exclude populations that experienced recent admixture. Two additional populations were excluded because they showed significant evidence for admixture using ALDER: the Ulchi (using Korean and Itelmen as references; $Z = 4.65$ and $P = 3.3 \times 10^{-6}$) and the Eskimos (using Surui and the Chukchi as references; $Z = 4.19$ and $P = 2.8 \times 10^{-5}$). Second, for populations outside East Asia and Siberia, we arbitrarily chose Ju_hoan_North and Mandenka for Africans, Sardinian and Basque for Europeans, Papuan for Oceania, and Karitiana and Surui for Native Americans. These populations represent major branches of human continental diversity and have been used as such representatives in many population genetic studies (Li *et al.* 2008; Reich *et al.* 2011; Keller *et al.* 2012; Pickrell *et al.* 2012; Skoglund *et al.* 2015). Last, we removed the She from the analysis because of its unstable position in the population trees generated by TreeMix (Pickrell and Pritchard 2012). The resulting set of 15 populations covers well all ancestry components inferred from the ADMIXTURE analysis (Figure 2). Five hundred bootstrap replicates of the maximum-likelihood tree were generated for the final 15 populations by TreeMix, with 50 SNPs per block ("-k 50"). A majority consensus tree was inferred using the R package "ape" (Paradis *et al.* 2004). We also performed TreeMix analysis with 1 to 5 migration edges allowed ("-m 1" to "-m 5") to detect major patterns of extra population affinity not captured by the tree in Figure 3. One hundred bootstrap replicates were conducted for each m value.

Formal tests of admixture

We calculated three-population (f_3) and Patterson's D statistics (Green *et al.* 2010) for all combinations of 71 worldwide populations (Table S1), using the qp3Pop and qpDstat programs in the ADMIXTOOLS v1.1 package (Patterson *et al.* 2012). All contemporary populations from the human genome diversity panel were included. In addition, all the other East Asian and Siberian populations were included if they had sample size ≥ 5. Four Yukagir individuals were removed because they show a large proportion of European-related ancestry, likely due to recent admixture.

We used the admixture LD decay-based method implemented in ALDER to provide additional evidence for admixture as well as an estimate of time since admixture, assuming a single pulse of admixture. We ran ALDER with two reference populations chosen based on the three-population test results. We applied bin size of 0.025 cM and required a minimum genetic distance of 0.5 cM between bins.

Frequency of derived alleles with selection signals in East Asians

We interrogated three variants, *EDAR V370A* (rs3827760), *OCA2 H615R* (rs1800414), and a noncoding SNP rs3811801 in the *ADH* gene cluster, which carry a selective sweep signal in East Asians. Because rs3827760 and rs3811801 were not included in the Ainu data, we imputed them using IMPUTE2 (Howie *et al.* 2009) with 1KG phase 3 dataset as a reference. Rs3827760 was imputed with high confidence (genotype posterior probability >0.98) for all 12 individuals. Except for two individuals who were omitted from further analysis, all imputed genotypes at rs3811801 had posterior probability ≥ 0.89. Therefore, although imputation in an isolated population may have reduced accuracy, genotypes were imputed with high confidence in our samples.

Genome scans of recent positive selection in the Ainu

We used autosomal SNPs from the 1KG-Ainu dataset to detect genomic regions showing signatures of recent positive selection in the Ainu. For this, we calculated the cross-population extended haplotype homozygosity (XP-EHH) statistic (Sabeti *et al.* 2007) against 1KG phase 3 CHB (Han Chinese in Beijing, China) and the population branch statistic (PBS) (Yi *et al.* 2010) using CHB as a comparison group and 1KG phase 3 CEU (CEPH Utah residents with northern and western European ancestry) as an outgroup. We removed strand ambiguous (G/C and A/T) SNPs from the analysis. To perform the XP-EHH analysis, we first phased the Ainu genotype data using SHAPEIT2 v2.r790 (Delaneau *et al.* 2013) with 1KG phase 3 data as a reference. The most likely haplotypes were chosen after running SHAPEIT2 with default parameter values. We summarized signals for each of 500-kb windows sliding by 50 kb. First, we counted the number of total SNPs (n_{total}) and top 1% signal SNPs (n_{top}) for each window and each test. Second, we calculated a simple binomial probability $P(X \geq n_{\text{top}})$ assuming a binomial distribution with success probability 1%, *i.e.*, $X \sim B(n_{\text{total}}, 0.01)$. Probability of 1 was assigned to windows with $n_{\text{total}} < 20$. Then, we prioritized windows with binomial $P \leq 0.01$ for both XP-EHH and PBS and merged adjacent windows. Finally, we narrowed down the peaks by removing 50-kb windows that do not harbor any top 1% SNP, resulting in 66 signal peaks for further analysis.

We used a simulation-based approach to test if our selection scans have enough power to distinguish loci under positive selection from neutrally evolving ones, given the small effective population size of the Ainu and our small sample size. For this purpose, we focused on the top 10 regions among the above 66, which have the highest PBS statistics. In

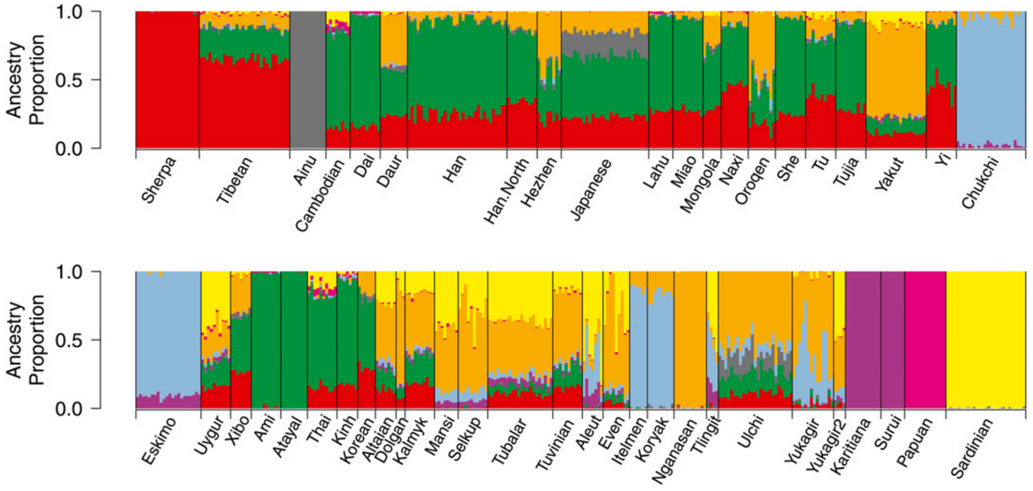


Figure 2 ADMIXTURE analysis of East Asian and Siberian populations with $K = 8$. Ainu individuals are assigned to a distinct ancestry component (dark gray), which is present also in Japanese and Ulchi individuals. Siberian ancestry is divided into two major components, one for northeast Siberians (skyblue in Chukchi, Eskimo, Itelmen, and Koryak) and the other for central Siberians (orange in Nganasan and other Siberian and northeast Asian populations). Four Yukagir individuals harboring a large proportion of European ancestry were labeled as "Yukagir2."

our procedure, we first simulated neutral trajectories of derived alleles under the Wright–Fisher model with a constant size of $N_e = 2{,}219$, which we estimated from LD decay. Specifically, we fit a nonlinear regression model with the equation $E(r^2) = 1/(1 + 4N_e c) + 1/n$, where c is a genetic distance in morgans and n is the number of sampled chromosomes (Tenesa *et al.* 2007). We also repeated our simulations with a more conservative estimate $N_e = 1000$. Although it is hard to model accurately demographic history using array genotyping data, the range of N_e values we used is likely to span a range of relevant demographic scenarios. The time of divergence between Ainu and the other East Asians was assumed to be 800 or 1000 generations ago, and the frequency in Ainu at the time of the split was taken from the current frequency in the other East Asians. At the end of each simulation, we took the frequency (f_{present}) and sampled 24 alleles following a binomial distribution of probability f_{present}. We repeated this process 10,000 times and calculated an empirical probability of our data by taking the fraction of simulations where the counts of simulated-derived alleles are greater than those observed in the Ainu sample.

Results

A subset of the Ainu samples is the result of recent admixture

We first investigated if the Ainu samples in our study represent a homogenous gene pool. Both a PCA and a genetic clustering analysis showed that the Ainu samples are genetically heterogeneous and form a few distinct clusters (Figure S2). Based on these analyses, three individuals labeled as Ainu were indistinguishable from the mainland Japanese samples (Figure S2); therefore, they were removed from the analysis. The same analyses also identified 10 Ainu individuals with substantial non-Ainu ancestry (>11%). Using ALDER, which is based on weighted admixture LD decay, we estimated the admixture time for these individuals to be 12.3 generations ago (Figure S3), further supporting a recent mixture ($P = 1.1 \times 10^{-6}$). If we included the entire sample of unrelated Ainu, a similar analysis with SNP loading on PC1 as a weight vector, without specifying reference populations, returned a similarly recent estimate (11.6 generations ago). Furthermore, to explore more complex admixture scenarios, we also fit data to a two-pulse admixture model, resulting in a combination of a younger (5.2 ± 2.1 generations ago) and an older (40.7 ± 8.9 generations ago) admixture event (Figure S3). When using a more stringent cutoff of 2.0 cM for the minimum genetic distance between markers, estimates were younger: 8.8 ± 2.1 for a single-pulse model and 4.6 ± 2.0 and 30.4 ± 10.1 for a two-pulse model. Even the older estimates of 30–40 generations ago from the two-pulse model, which set an upper bound of a continued gene flow, are too recent to be consistent with the initial expansion of the Yayoi culture into the Japanese archipelago (Jinam *et al.* 2012; Nakagome *et al.* 2015) or with the hypothesized gene flow from the so called "Okhotsk culture," which spread throughout the Sakhalin and Hokkaido islands between the 5th and 11th centuries (Befu and Chard 1964; Ohyi 1975). Because we are primarily interested in the ancient history of the Ainu gene pool and its relationship with worldwide populations, we decided to primarily use 12 individuals with 100% Ainu ancestry (Ainu in Figure S2) as representative of the original Ainu gene pool in the following analyses. However, we also performed several analyses with all 22 Ainu individuals in parallel and obtained comparable results, as presented below.

The Ainu form an outgroup to all East Asian farmers including Tibetan populations

To infer the genetic relationship of the Ainu gene pool with worldwide populations, we compiled genotype data of the Ainu, the Sherpa, Tibetans, and 183 populations around the world as described in the *Materials and Methods* section (WHA dataset). With this dataset, we first characterized the Ainu ancestry in the context of East Asian genetic diversity, by performing a model-based unsupervised genetic clustering as implemented in the program ADMIXTURE. With the optimal number of ancestral components ($K = 8$), the Ainu

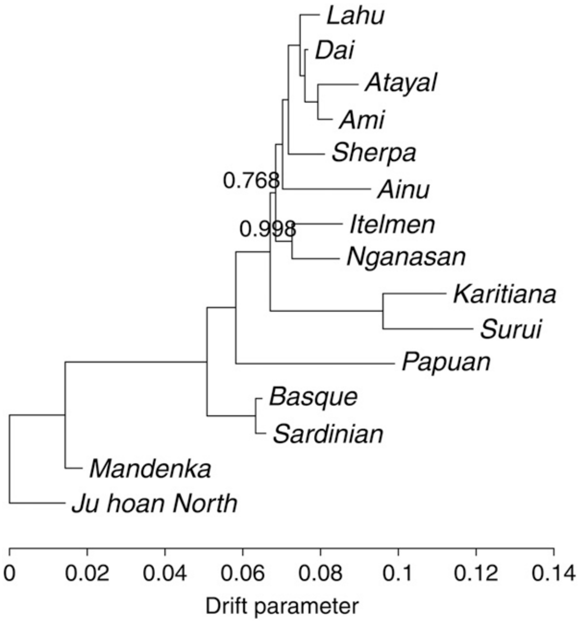


Figure 3 A consensus tree of 15 worldwide populations inferred from 500 bootstrap replicates of maximum likelihood trees using TreeMix. Numbers show the bootstrap support on the corresponding nodes. Nodes with no number were supported in 100% of the bootstrap replicates.

individuals are assigned to a distinct ancestry (Figure 2). In suboptimal runs with fewer ancestral components ($K \leq 6$), the Ainu, as most other East Asians, are modeled as a mixture of Siberian and East Asian ancestries, with limited contributions from other populations (Figure S4). Interestingly, Ainu-related ancestry is present in the Japanese and the Ulchi people from the lower basin of the Amur River (17.8 and 13.5% mean ancestry in Figure 2, respectively), as well as in two Nivkh individuals, an indigenous population from the Sakhalin island, in a similar analysis with additional samples (27.2% mean ancestry; Figure S5). This suggests a potential gene flow from an Ainu-related gene pool into these surrounding populations.

Then, we further explored the genetic relationships among the ancestry components inferred from the ADMIXTURE analysis. To do this, we aimed to choose population samples that were good representatives of the global population structure (Figure 2). First, we removed populations with signatures of recent admixture not involving the Ainu ancestry, suggested either by negative values of the three-population (f_3) statistic or by significant decay of weighted admixture LD (see *Materials and Methods*). To simplify the analysis, we arbitrarily chose one or two populations to represent each of the major continental groups outside of East Asia, but kept all East Asian and Siberian populations with no signal of recent admixture. This process resulted in a total of 15 populations (Figure 3), which cover all ancestry components identified by ADMIXTURE.

Using a maximum likelihood-based algorithm implemented in TreeMix, we found that the Ainu can be modeled as an outgroup to all East Asian farmers (Ami, Atayal, Dai, Lahu, and the Sherpa; Figure 3) in all 500 bootstrap replicates. The long terminal branch leading to the Ainu (Figure 3), as well as slow decay of LD (Figure S6), suggests a strong genetic drift, expected for a small population isolate. The Ainu's position as an East Asian outgroup in the tree is unlikely due to gene flow from outside East Asia, as no significant results were found with Patterson's D statistic in the form of D(African, non-African outgroup; Ainu, East Asian farmer) (Table S2). We further investigated this point by using an additional dataset of 320K SNPs in a smaller set of populations including also the Neandertal and Denisova data ("CND-1KG-Ainu" dataset). Consistent with the TreeMix analysis, the Ainu form a clade with East Asian populations relative to Europeans or South Asians (Table S3 and Figure S7). Additionally, at the resolution provided by this data set, the Ainu are not inferred to contain more archaic ancestry than other East Asian populations. The TreeMix result is unlikely to be an artifact of either genetic drift or variant ascertainment bias: this algorithm was shown to work well even if population-specific variants are common (extreme genetic drift) or if all variants are ascertained in a single population (Pickrell and Pritchard 2012). In addition, the allele frequency distribution of the SNPs included in the analysis was similar across the Ainu and the other East Asian, Siberian, and Native American populations (Figure S8) and it did not vary substantially between different intersected sets of SNPs (Figure S9). The Sherpa formed an outgroup to the lowland East Asian farmers, consistent with our previous study showing a deep split between high- and low-altitude East Asians (Jeong *et al.* 2014). Siberian populations (Nganasan and Itelmen) were modeled either as a sister group of all East Asians including the Ainu (76.8%; Figure 3) or as a sister group of Native Americans, Karitiana, and Surui (23.2%; Figure S10). When we allowed for migration edges in the TreeMix analysis, gene flow events between Europeans and Native Americans or Siberians were robustly inferred in all bootstrap replicates (Figure S11 and Table S4). This pattern is in agreement with the findings of previous studies on the genetic history of Native Americans and Siberian populations (Rasmussen *et al.* 2010; Fedorova *et al.* 2013; Raghavan *et al.* 2014).

The Ainu share more ancestry with low-altitude than with high-altitude East Asians

Human populations often have a complicated genetic history, which cannot be fully captured by a simple tree-based model. The distribution of residual covariances from the maximum likelihood trees indeed suggests that the consensus tree cannot fully explain the data for many of the populations, including the Ainu (Figure S12). Therefore, as a next step, we investigated if the Ainu have extra affinity with other populations beyond what could be inferred in a bifurcating tree. First, we tested if East Asian farmer populations are symmetric to each other in terms of their relationship with the Ainu, as suggested by the Ainu position as an outgroup to these populations in the consensus population tree (Figure 3 and

Figure S13A). If the population relationships in the consensus tree hold, two East Asian farmer populations should be equally close to the Ainu. However, the D statistics in the form of D(outgroup, Ainu; Sherpa, other East Asian) consistently showed significantly positive values ($D > +2.9$ SD; Table S5), suggesting gene flow between the Ainu and lowland East Asian populations after their split from the high-altitude populations. Unsurprisingly, the inclusion of the 10 recently admixed Ainu individuals further strengthened this pattern ($D = +3.0$ to $+8.6$ SD for the same set of tests). Genetic affinity tests of East Asian populations with the Ainu, assessed by the outgroup f_3 statistic (Raghavan *et al.* 2014), also supported a closer relationship between the Ainu and lowland East Asians than between the Ainu and the Sherpa (Figure S14). TreeMix analysis allowing migration edges also detected a similar relationship: an edge between the Ainu and lowland East Asians was inferred for 59–88% of bootstrap replicates, when three or more migration edges were allowed (Figure S11 and Table S4).

The Ainu share ancestry with northeast Siberians but not with central Siberians

Previous genetic studies of Siberian populations clearly demonstrated genetic differentiation between northeast Siberians and the rest of the Siberian populations (Volodko *et al.* 2008; Rasmussen *et al.* 2010; Fedorova *et al.* 2013; Raghavan *et al.* 2014). Our genetic clustering analysis recapitulates this observation, by separating the northeast Siberian ancestry (skyblue in Figure 2; concentrated in Eskimo, Chukchi, Itelmen, and Koryak) from central Siberian ancestry (orange in Figure 2; most prevalent in the Nganasan and present in southern Siberians and northeast Asians). Even though the Itelmen and the Nganasan cluster together in our population tree (Figure 3 and Figure S10, Figure S13B), most East Asian populations are genetically closer to the Nganasan than to the Itelmen, as shown by their negative D(African, East Asian; Nganasan, Itelmen) statistic (-2.4 to -12.0 SD; Figure 4 and Figure S15). Interestingly, the Ainu were the only East Asian population showing a closer affinity to the Itelmen than to the Nganasan, although the observed negative D statistic was within statistical noise ($+1.5$ SD; Figure 4). This pattern was robust to the inclusion of the 10 recently admixed Ainu individuals, which, as expected, dampens the signal due to the presence of mainland Japanese ancestry ($+1.0$ SD; Figure S16). Consistent with this result, a D test in the form D(African, Siberian, Ainu, East Asian) showed that the Itelmen are genetically closer to the Ainu than to East Asian farmer populations, but the Nganasan are symmetric in their relationship to the Ainu and the East Asian farmers (Table S6). Northeast Asian or southern Siberian populations could not be directly compared in this way because of the shared ancestry between the Itelmen and the Nganasan. We obtained qualitatively similar results when we replaced the Itelmen with the Chukchi (Figure S17 and Table S6). The TreeMix analysis also detected migration edges between the Itelmen and the Ainu when four or five migration edges were allowed, but only in 13–23% of bootstrap replicates (Table S4).

An Ainu-related ancestry was introduced Into nearby populations

Our genetic clustering analysis strongly suggests that the Ainu-related ancestry substantially contributed to the gene pools of nearby populations, such as the Japanese or the Ulchi (Figure 2 and Figure S4, Figure S5). However, strong genetic drift in the Ainu may artificially generate such a signal. Therefore, we applied two formal tests of admixture, which use different aspects of genetic variation data, to test for Ainu-related admixture in these populations. First, three-population test statistics were significantly negative when the Ainu were used as a reference: -22.2 SD for the Japanese (using the Ainu and Han as references) and -3.9 SD for the Ulchi (using the Ainu and Nganasan as references). Second, admixture LD decay was clear in both populations ($P = 3.7 \times 10^{-26}$ and $P = 8.7 \times 10^{-9}$ for the Japanese and the Ulchi, respectively), with estimates of admixture time ~70 and 22 generations ago for the Japanese and the Ulchi, respectively (Figure S18, Figure S19).

The Ainu genome harbors shared and unique signatures of adaptations

To investigate adaptive evolution occurring in the Ainu, we analyzed the allele frequencies of variants known to be swept to high frequency in East Asians and performed genomic scans of recent positive selection across the Ainu genome.

A nonsynonymous *V370A* (rs3827760) mutation in the *EDAR* (ectodysplasin A receptor) gene harbors a strong selective sweep signal shared among low- and high-altitude East Asians and Native Americans, which is not present in contemporary western Eurasian populations (Kamberov *et al.* 2013). In addition, the derived allele is associated with "East Asian phenotypes," such as shovel shaped incisors (Kimura *et al.* 2009). In sharp contrast to surrounding populations, this allele occurs at only 25% (6 of 24) frequency in the Ainu (Table S7). Consistent with this finding, the Ainu are also reported to have the sundadont dental pattern, even though the sinodont pattern, which is associated with shovel shaped incisors, is the dominant one in northeast Asia (Howells 1997). This suggests that the Ainu ancestors may not have shared the selective pressures for *EDAR V370A* with other ancestral East Asian and Native American populations.

In contrast to *EDAR V370A*, two variants occurring at high frequency in East Asia and virtually absent elsewhere, rs1800414 (*H615R*) in the *OCA2* gene (Hider *et al.* 2013) and rs3811801 in the *ADH* gene cluster (Li *et al.* 2011), have high frequency in the Ainu (Table S7). The onset of positive selection on these two variants was estimated to have occurred <11,000 years ago (Peng *et al.* 2010; Li *et al.* 2011; Chen *et al.* 2015). Therefore, the Ainu ancestors may have shared Holocene environmental factors favoring these variants with other East Asians, although gene flow between the Ainu and other East Asians, as we inferred from genome-wide

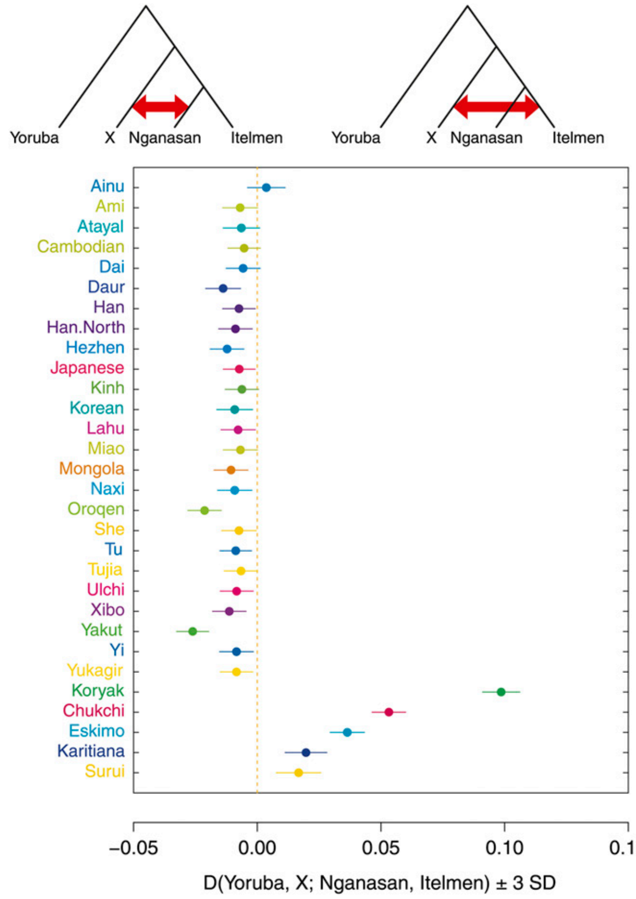


Figure 4 The genetic affinity of East Asian and Siberian populations to the Nganasan and the Itelmen, respectively, measured by Patterson's *D*(Yoruba, X; Nganasan, Itelmen). Horizontal bars around the value represent ±3 SD.

SNP data, may also have contributed to their high frequencies in the Ainu. Interestingly, they occur at low frequency, ~10% in the Sherpa and Tibetans, raising the possibility that selective pressure on these variants was different in the high-altitude environments.

To find genomic loci involved in local adaptations in the Ainu, we performed a LD-based test of recent positive selection (XP-EHH), and an allele frequency-based test (PBS), in the Ainu against CHB from the 1KG phase 3 dataset. We found a total of 66 genomic regions containing excess SNPs with extreme values (top 1%), defined by a binomial probability ≤0.01 of having the number of extreme SNPs equal to or greater than the observed one, of both XP-EHH and PBS statistics (Table S8). One such region on chromosome 11 spans the Apolipoprotein (*APO*) gene cluster, including the *APOA5*, *APOA4*, *APOC3*, and *APOA1* genes. This region contains a noncoding SNP rs964184 associated with levels of blood triglycerides and LDL cholesterol, HDL cholesterol, and risk of coronary artery disease and ischemic stroke (Global Lipid Genetics Consortium 2013; Dichgans *et al.* 2014). Interestingly, the risk allele has much higher frequency in the Ainu, reaching 75% in comparison to 22% in 1KG CHB, and is found on a haplotype with extended LD (Figure S20).

In addition, multiple loci among the above 66 regions include SNPs showing extreme allele frequency differentiation between the Ainu and 1KG East Asians excluding JPT (Table S9). Several genes in these loci have been reported to be associated with other metabolic traits, such as body mass index, glomerular filtration rate, and serum metabolite levels (Table S9). Because the Ainu sample size is small ($n = 12$) and the Ainu have had historically small effective population size (Figure S6), high allele frequency divergence may not necessarily be due to local adaptations. Therefore, we relied on simulations to test if the observed allele frequency difference is greater than expected by chance for a small population and a sample size as small as ours. We found that for all of the 10 top PBS regions, the difference in allele frequency is unlikely to be due to chance (empirical P-value ≤0.01), based on neutral simulations of population of constant effective population size (N_e) of 2219, as estimated using an LD decay method (Table S9). We obtained similar results (9 of 10 top PBS regions with empirical P-value ≤0.05) (Table S9), when we used a smaller, more conservative estimate of $N_e = 1000$.

Discussion

Our genome-wide analysis shows that the Ainu are one of the deepest branches of East Asian diversity, forming an outgroup to all present-day East Asian farmers, including high-altitude populations (Figure 3). The deep history of the Ainu is consistent with the archaeological record for the Jomon culture in Japan starting 16,500 years ago (Habu 2004) as well as their hunter–gatherer life style. Therefore, the ancestors of the Ainu are likely to have reached the Japanese archipelago in an early migration event distinct from the spread of farmer populations across East Asia.

Interestingly, we find evidence for extra genetic affinity between the Ainu and northeast Siberians (Itelmen and Chukchi), who share ancestry with Native Americans. This finding coupled with the ancient origin of the Ainu raises the possibility that the same migration event led to the settlement of Jomon hunter–gatherers and to the initial dispersal of Native American ancestors. If this is the case, this first northward migration took place before the LGM ("scenario 1" in Figure 5A). This proposal is consistent with previous studies that suggested a connection between Jomon or Ainu people and Native Americans based on morphological and genetic evidence (Tokunaga *et al.* 2001; Adachi *et al.* 2009; Owsley and Jantz 2014) (Figure 4 and Table S6). Under this scenario, the split between the Ainu and Native American ancestors is likely to have occurred earlier than the gene flow of western Eurasian ancestry into the Native American ancestors (Raghavan *et al.* 2014), because the Ainu and other East Asians are symmetrically related to contemporary Europeans and to the ancient MA1 sample (Table S2). However, a recent study reported no genetic affinity between the Ainu and the Kennewick man (Rasmussen *et al.* 2015). An alternative to

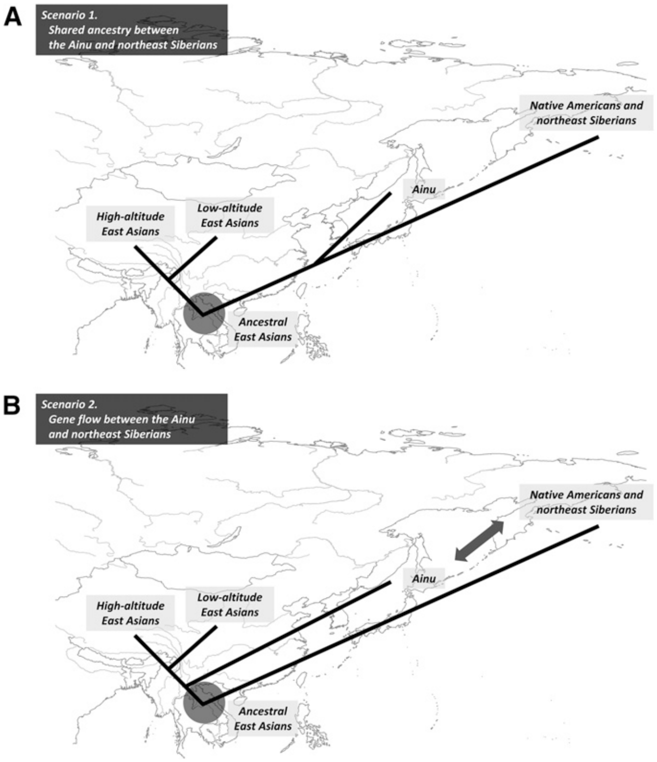


Figure 5 A summary of competing scenarios for the observed excess affinity of the Ainu with northeast Siberians. (A) "Scenario 1" proposes a shared ancestry between the Ainu and northeast Siberians. (B) "Scenario 2" proposes a later gene flow between the Ainu and northeast Siberians.

this scenario is that there may have been more recent gene flow between the Ainu and northeast Siberian populations ("scenario 2" in Figure 5B). Our TreeMix analysis with migration edges suggests a gene flow from the northeast Siberians to the Ainu, although both the pattern itself and its direction are not robust (Table S4). As explained above, the spread of the Okhotsk culture is unlikely to account for this finding, although such a contact across the Sea of Okhotsk may have happened earlier than the Okhotsk culture. Whole genome sequence data of the Ainu, ancient Jomon samples, and northeast Siberians will shed more light on the details of this history (Li and Durbin 2011; Schiffels and Durbin 2014).

Surprisingly, we also find extra genetic affinity between the Ainu and lowland farmer populations in comparison to the Sherpa (Figure S14 and Table S5), indicating gene flow between these two groups of populations. A long-standing hypothesis posits that Ainu and Tibetans share a part of their ancestry that is not present in other East Asian populations based on patterns of Y chromosome variation. The Y chromosome haplogroup D-M174 shows a striking pattern of geographic distribution: it is highly prevalent in Tibetans and Japanese (especially in the Ainu) and virtually absent everywhere else in Eurasia (Hammer *et al.* 2006; Shi *et al.* 2008; Chiaroni *et al.* 2009; Stoneking and Delfin 2010). A possible explanation for the Y chromosome data is that the Tibetan and the Japanese branches of this haplogroup have deep coalescence times, *i.e.*, older than 30,000 years before the present (Shi *et al.* 2008). Therefore, even if the presence of the D-M174 haplogroup in the Ainu and Tibetans is due to shared ancestry, the shared history of these populations was short and left only a weak genome-wide signature of shared variation in their gene pools.

Even though we find strong evidence of gene flow between the Ainu and lowland East Asian farmers, it is hard to establish whether migrations were mainly unidirectional and, if so, which direction was predominant. One possibility is that Ainu-related populations, probably hunter–gatherers, once occupied mainland East Asia preceding the expansion of farmers and that they contributed to the gene pool of the latter. The observation of Ainu-related ancestry in the Ulchi from the lower Amur River basin (Figure 2 and Figure S4, Figure S5) is consistent with the presence of such an Ainu-related population in mainland northeast Asia. This model of gene flow is not expected to generate a signature of admixture in the Ainu. Consistent with this scenario, we fail to detect an admixture signal in the Ainu beyond the 10 recently admixed individuals (Figure S2, Figure S3): three-population statistics are strongly positive for all combinations of reference populations listed in Table S1 ($> +25$ SD). Extended LD in the Ainu and the lack of a reference population representing the Jomon ancestry made it difficult to test for admixture in the Ainu using ALDER: indeed, decay constant and amplitude parameter estimates from one-reference ALDER analysis did not change regardless of our choice of reference population (decay constant = 9.0–9.8 generations ago, amplitude = 1.6 to 2.3×10^{-4} across all 26 1KG populations as a reference). This pattern was reported to be a likely false positive in the ALDER analysis, which occurs when the target population experienced strong genetic drift (Loh *et al.* 2013). Therefore, we probably did not have enough power to accurately infer the timing and direction of ancient gene flow events, such as those we found between the Ainu and lowland East Asians or northeast Siberians. Genetic analysis of ancient Jomon and Ainu samples over a range of time periods will be critical to distinguish among these hypotheses.

While we confirm the evidence first reported by Jinam *et al.* (2012) for an admixture event between Yayoi farmers and Ainu ancestors, we do not find evidence supporting a claim for a gene flow into the Ainu from an unknown population. This claim was based on a group of five Ainu individuals clustering away from the rest in PCA plots (Jinam *et al.* 2012). We think this is an artifact of including close relatives in PCA, a well-known phenomenon. Indeed, a recent reanalysis by Jinam *et al.* (2015) also found that exclusion of close relatives from the analysis removed this clustering pattern (Jinam *et al.* 2015). We discuss this issue in more detail in Text S1.

We took a cautious approach in performing and interpreting genome scans of recent positive selection in the Ainu, because of their small effective population size and our small sample size. For example, we decided not to use the integrated haplotype score approach (Voight *et al.* 2006) because it requires substantial numbers of haplotypes harboring both ancestral and derived alleles at the focal variant. Considering this limitation, it is particularly encouraging that we find several loci harboring extreme allele frequency differentiation in the Ainu (greater than expected under neutrality based on simulations) in comparison to 1KG CHB data (Table S9). In particular, lipid metabolism may have been a key process for local adaptation in the Ainu, as suggested by the selection signature around the *APO* gene cluster and their historical heavy dependence on a marine subsistence. Archaeological evidence for heavy reliance of the prehistoric Jomon culture, particularly in the northeastern part of Japan, on marine mammals and fish (Chisholm and Koike 1999; Yoneda *et al.* 2002) may provide a plausible link between our findings and local adaptations of the Ainu and Jomon people.

Acknowledgments

We thank the Asian DNA Repository Consortium for sharing the Ainu genotype data for this study. We are grateful to Hiroki Oota, Shuhei Mano, John Novembre, Matthew Stephens, and David Witonsky for helpful discussions and advice on data analysis methods. Computational resources for data analysis were provided by Tadashi Imanishi in the Biomedical Informatics Laboratory, at Tokai University School of Medicine, the Center for Research Informatics, at the University of Chicago, and the Beagle supercomputer provided by the Computation Institute and the Biological Science Division of the University of Chicago and Argonne National Laboratory, under grant 1S10OD018495-01. This work was supported in part by the National Institutes of Health grant R01HL119577 to A.D. C.J. was supported by a Samsung scholarship. S.N. was supported by a postdoctoral fellowship for research abroad from the Japan Society for the Promotion of Science (26-541).

Literature Cited

Adachi, N., K. i. Shinoda, K. Umetsu, and H. Matsumura, 2009 Mitochondrial DNA analysis of Jomon skeletons from the Funadomari site, Hokkaido, and its implication for the origins of Native American. Am. J. Phys. Anthropol. 138: 255–265.

Alexander, D. H., J. Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19: 1655–1664.

Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel *et al.*, 2013 From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr. Protoc. Bioinformatics 43: 11.10.11–11.10.33.

Befu, H., and C. S. Chard, 1964 A prehistoric maritime culture of the Okhotsk Sea. Am. Antiq. 30: 1–18.

Chen, H., J. Hey, and M. Slatkin, 2015 A hidden Markov model for investigating recent positive selection through haplotype structure. Theor. Popul. Biol. 99: 18–30.

Chiaroni, J., P. A. Underhill, and L. L. Cavalli-Sforza, 2009 Y chromosome diversity, human expansion, drift, and cultural evolution. Proc. Natl. Acad. Sci. USA 106: 20174–20179.

Chisholm, B., and H. Koike, 1999 Reconstructing prehistoric Japanese diet using stable isotopic analysis, pp. 199–222 in *Interdisciplinary Perspectives on the Origins of the Japanese*, edited by K. . Omoto International Research Center for Japanese Studies, Kyoto, Japan.

Delaneau, O., J.-F. Zagury, and J. Marchini, 2013 Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods 10: 5–6.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43: 491–498.

Dichgans, M., R. Malik, I. R. König, J. Rosand, R. Clarke *et al.*, 2014 Shared genetic susceptibility to ischemic stroke and coronary artery disease A genome-wide analysis of common variants. Stroke 45: 24–36.

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

Global Lipid Genetics Consortium, 2013 Discovery and refinement of loci associated with lipid levels. Nat. Genet. 45: 1274–1283.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel *et al.*, 2010 A draft sequence of the Neandertal genome. Science 328: 710–722.

Habu, J., 2004 *Ancient Jomon of Japan*. Cambridge University Press, Cambridge, UK.

Hammer, M. F., T. M. Karafet, H. Park, K. Omoto, S. Harihara *et al.*, 2006 Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. J. Hum. Genet. 51: 47–58.

Hanihara, K., 1991 Dual structure model for the population history of the Japanese. Japan Review 2: 1–33.

Hider, J. L., R. M. Gittelman, T. Shah, M. Edwards, A. Rosenbloom *et al.*, 2013 Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry. BMC Evol. Biol. 13: 150.

Howells, W. W., 1997 *Getting Here: The story of Human Evolution*. Compass Press, Washington, USA.

Howie, B. N., P. Donnelly, and J. Marchini, 2009 A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5: e1000529.

Huerta-Sánchez, E., X. Jin, Z. Bianba, B. M. Peter, N. Vinckenbosch *et al.*, 2014 Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature 512: 194–197.

HUGO Pan-Asian SNP Consortium, 2009 Mapping human genetic diversity in Asia. Science 326: 1541–1545.

Jeong, C., G. Alkorta-Aranburu, B. Basnyat, M. Neupane, D. B. Witonsky *et al.*, 2014 Admixture facilitates genetic adaptations to high altitude in Tibet. Nat. Commun. 5: 3281.

Jinam, T., N. Nishida, M. Hirai, S. Kawamura, H. Oota *et al.*, 2012 The history of human populations in the Japanese Archipelago inferred from genome-wide SNP data with a special reference to the Ainu and the Ryukyuan populations. J. Hum. Genet. 57: 787–795.

Jinam, T. A., H. Kanzawa-Kiriyama, I. Inoue, K. Tokunaga, K. Omoto *et al.*, 2015 Unique characteristics of the Ainu population in Northern Japan. J. Hum. Genet. 60: 565–571.

Kamberov, Y. G., S. Wang, J. Tan, P. Gerbault, A. Wark *et al.*, 2013 Modeling recent human evolution in mice by expression of a selected *EDAR* variant. Cell 152: 691–702.

Keller, A., A. Graefen, M. Ball, M. Matzas, V. Boisguerin *et al.*, 2012 New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat. Commun. 3: 698.

Kimura, R., T. Yamaguchi, M. Takeda, O. Kondo, T. Toma *et al.*, 2009 A common variation in *EDAR* is a genetic determinant of shovel-shaped incisors. Am. J. Hum. Genet. 85: 528–535.

Koganebuchi, K., T. Katsumura, S. Nakagome, H. Ishida, S. Kawamura *et al.*, 2012 Autosomal and Y-chromosomal STR markers reveal a close relationship between Hokkaido Ainu and Ryukyu islanders. Anthropol. Sci. 120: 199–208.

Lazaridis, I., N. Patterson, A. Mittnik, G. Renaud, S. Mallick *et al.*, 2014 Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513: 409–413.

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754–1760.

Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature 475: 493–496.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079.

Li, H., S. Gu, Y. Han, Z. Xu, A. J. Pakstis *et al.*, 2011 Diversification of the *ADH1B* gene during expansion of modern humans. Ann. Hum. Genet. 75: 497–507.

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

Loh, P.-R., M. Lipson, N. Patterson, P. Moorjani, J. K. Pickrell *et al.*, 2013 Inferring admixture histories of human populations using linkage disequilibrium. Genetics 193: 1233–1254.

Matsukusa, H., H. Oota, K. Haneji, T. Toma, S. Kawamura *et al.*, 2010 A genetic analysis of the Sakishima islanders reveals no relationship with Taiwan aborigines but shared ancestry with Ainu and main-island Japanese. Am. J. Phys. Anthropol. 142: 211–223.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297–1303.

Meyer, M., M. Kircher, M.-T. Gansauge, H. Li, F. Racimo *et al.*, 2012 A high-coverage genome sequence from an archaic Denisovan individual. Science 338: 222–226.

Moorjani, P., K. Thangaraj, N. Patterson, M. Lipson, P.-R. Loh *et al.*, 2013 Genetic evidence for recent population mixture in India. Am. J. Hum. Genet. 93: 422–438.

Nakagome, S., T. Sato, H. Ishida, T. Hanihara, T. Yamaguchi *et al.*, 2015 Model-based verification of hypotheses on the origin of modern Japanese revisited by Bayesian inference based on genome-wide SNP data. Mol. Biol. Evol. 32: 1533–1543.

Ohyi, H., 1975 The Okhotsk culture, a maritime culture of the southern Okhotsk Sea region, pp. 123–158 in *Prehistoric Maritime Adaptations of the Circumpolar Zone*, edited by W. Fitzhugh. Aldine Publishing Company, Chicago.

Oota, H., and M. Stoneking, 2011 Effect of human migration on genome diversity in East Asia, pp. 173–187 in *Racial Representations in Asia*. Kyoto University Press, Kyoto.

Owsley, D. W., and R. L. Jantz, 2014 *Kennewick Man: The Scientific Investigation of an Ancient American Skeleton*, Texas A&M University Press, College Station, TX, USA.

Paradis, E., J. Claude, and K. Strimmer, 2004 APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289–290.

Patterson, N., A. L. Price, and D. Reich, 2006 Population structure and eigenanalysis. PLoS Genet. 2: e190.

Patterson, N., P. Moorjani, Y. Luo, S. Mallick, N. Rohland *et al.*, 2012 Ancient admixture in human history. Genetics 192: 1065–1093.

Peng, Y., H. Shi, X.-b. Qi, C.-j. Xiao, H. Zhong *et al.*, 2010 The *ADH1B* Arg47His polymorphism in East Asian populations and expansion of rice domestication in history. BMC Evol. Biol. 10: 15.

Pickrell, J. K., and J. K. Pritchard, 2012 Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8: e1002967.

Pickrell, J. K., N. Patterson, C. Barbieri, F. Berthold, L. Gerlach *et al.*, 2012 The genetic prehistory of southern Africa. Nat. Commun. 3: 1143.

Pitulko, V. V., P. A. Nikolsky, E. Y. Girya, A. E. Basilyan, V. E. Tumskoy *et al.*, 2004 The Yana RHS site: humans in the Arctic before the last glacial maximum. Science 303: 52–56.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38: 904–909.

Prüfer, K., F. Racimo, N. Patterson, F. Jay, S. Sankararaman *et al.*, 2014 The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505: 43–49.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559–575.

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

Rasmussen, M., Y. Li, S. Lindgreen, J. S. Pedersen, A. Albrechtsen *et al.*, 2010 Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 463: 757–762.

Rasmussen, M., M. Sikora, A. Albrechtsen, T. S. Korneliussen, J. V. Moreno-Mayar *et al.*, 2015 The ancestry and affiliations of Kennewick Man. Nature 523: 455–458.

Reich, D., K. Thangaraj, N. Patterson, A. L. Price, and L. Singh, 2009 Reconstructing Indian population history. Nature 461: 489–494.

Reich, D., R. E. Green, M. Kircher, J. Krause, N. Patterson *et al.*, 2010 Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468: 1053–1060.

Reich, D., N. Patterson, M. Kircher, F. Delfin, M. R. Nandineni *et al.*, 2011 Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am. J. Hum. Genet. 89: 516–528.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918.

Schiffels, S., and R. Durbin, 2014 Inferring human population size and separation history from multiple genome sequences. Nat. Genet. 46: 919–925.

Shi, H., H. Zhong, Y. Peng, Y.-L. Dong, X.-B. Qi *et al.*, 2008 Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. BMC Biol. 6: 45.

Skoglund, P., H. Malmström, M. Raghavan, J. Storå, P. Hall *et al.*, 2012 Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. Science 336: 466–469.

Skoglund, P., S. Mallick, M. C. Bortolini, N. Chennagiri, T. Hunemeier *et al.*, 2015 Genetic evidence for two founding populations of the Americas. Nature 525: 104–108.

Stoneking, M., and F. Delfin, 2010 The human genetic history of East Asia: weaving a complex tapestry. Curr. Biol. 20: R188–R193.

Tajima, A., I.-H. Pan, G. Fucharoen, S. Fucharoen, M. Matsuo *et al.*, 2002 Three major lineages of Asian Y chromosomes: implications for the peopling of east and southeast Asia. Hum. Genet. 110: 80–88.

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke *et al.*, 2007 Recent human effective population size estimated from linkage disequilibrium. Genome Res. 17: 520–526.

Tokunaga, K., J. Ohashi, M. Bannai, and T. Juji, 2001 Genetic link between Asians and native Americans: evidence from HLA genes and haplotypes. Hum. Immunol. 62: 1001–1008.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS Biol. 4: e72.

Volodko, N. V., E. B. Starikovskaya, I. O. Mazunin, N. P. Eltsov, P. V. Naidenko *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.

Wang, B., Y.-B. Zhang, F. Zhang, H. Lin, X. Wang *et al.*, 2011 On the origin of Tibetans and their genetic basis in adapting high-altitude environments. PLoS One 6: e17002.

Wu, H., B. Caffo, H. A. Jaffee, R. A. Irizarry, and A. P. Feinberg, 2010 Redefining CpG islands using hidden Markov models. Biostatistics 11: 499–514.

Yi, X., Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo *et al.*, 2010 Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329: 75–78.

Yoneda, M., A. Tanaka, Y. Shibata, M. Morita, K. Uzawa *et al.*, 2002 Radiocarbon marine reservoir effect in human remains from the Kitakogane site, Hokkaido, Japan. J. Archaeol. Sci. 29: 529–536.

Communicating editor: S. Ramachandran

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.178673/-/DC1

Deep History of East Asian Populations Revealed Through Genetic Analysis of the Ainu

Choongwon Jeong, Shigeki Nakagome, and Anna Di Rienzo

DOI: 10.1534/genetics.115.178673

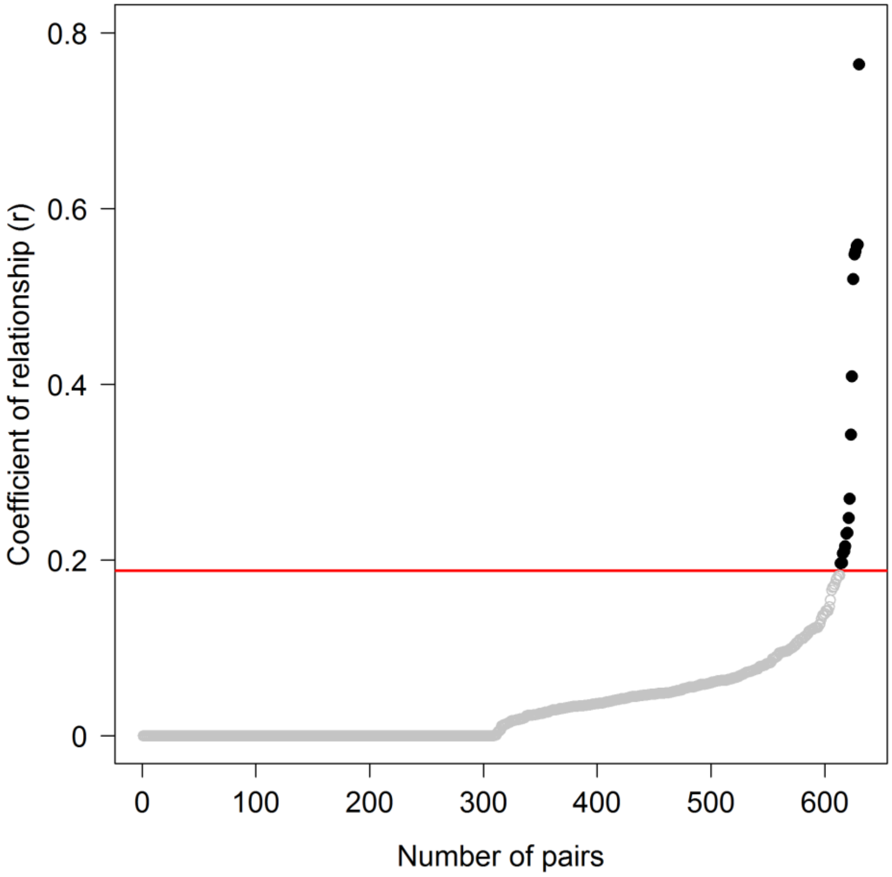


Figure S1 Cumulative distribution of coefficient of relationship (r) between pairs of Ainu individuals. 11 individuals from 17 pairs with r > 0.1875 (black filled dots above the red line for r = 0.1875) were removed from the downstream population genetic analysis.

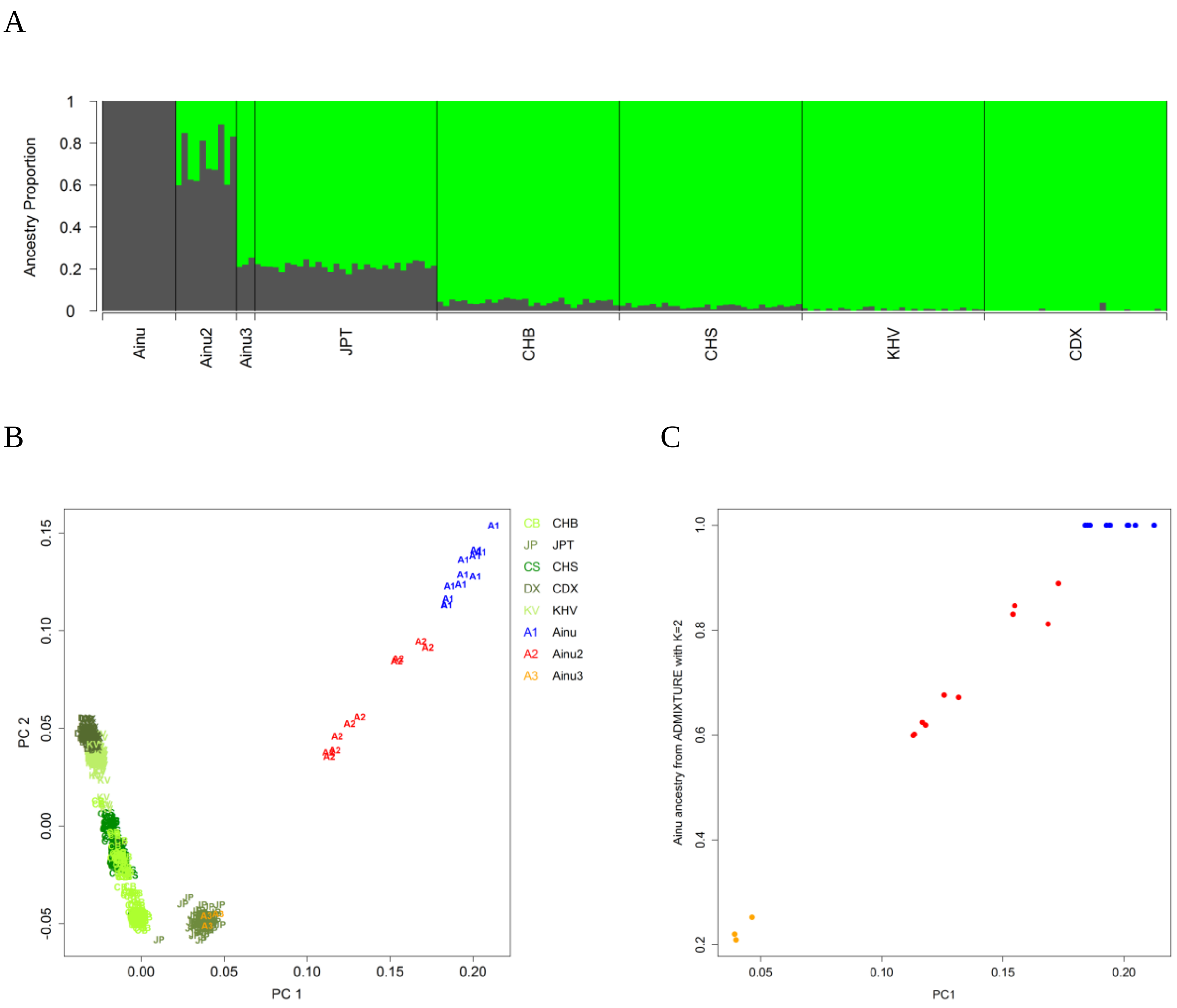


Figure S2 Identification of Ainu individuals with recent mainland Japanese ancestors. (A) *ADMIXTURE* analysis with K=2 and (B) PCA of 1KG East Asian and Ainu individuals identified 12 Ainu individuals with the highest Ainu ancestry (dark grey in A and PC 1 in B). Another 10 Ainu individuals ("Ainu2") formed two discrete clusters between the "Ainu" cluster and mainland Japanese (JPT). Three Ainu individuals clustered closely with mainland Japanese ("Ainu3"). (C) *ADMIXTURE* with K=2 and PCA results closely match, detecting the same four clusters, including those with 100% Ainu ancestry.

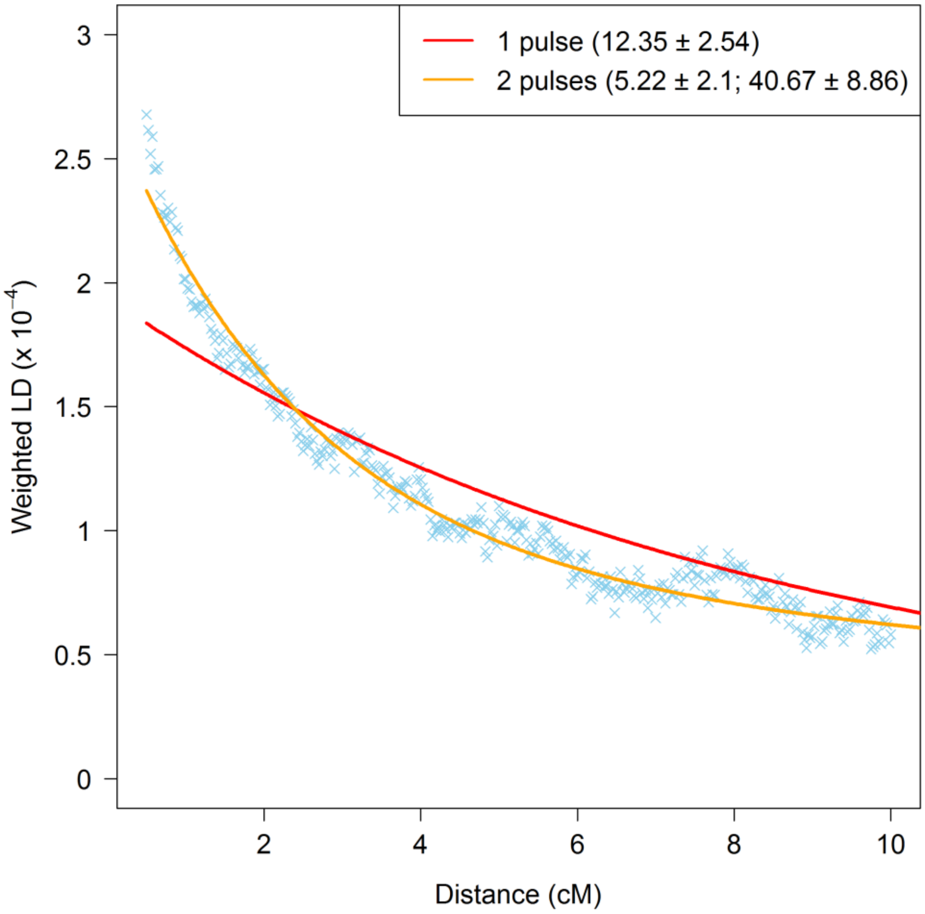


Figure S3 Weighted admixture LD decay in the 10 admixed Ainu with the unadmixed Ainu and 1KG JPT as references. The estimated times (in generations) of the inferred admixture events and their standard deviations are shown in parenthesis.

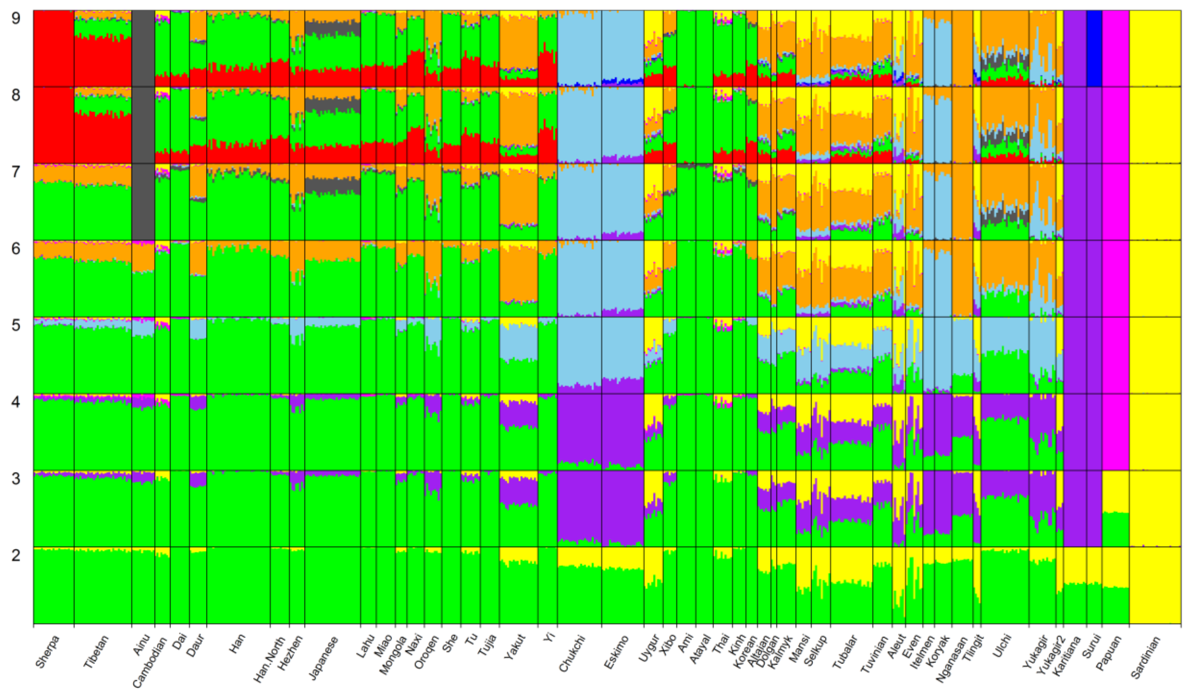


Figure S4 *ADMIXTURE* analysis of East Asian and Siberian populations with K = 2 to 9. Ainu individuals are assigned to their own ancestry (dark grey) with K = 7, following separation of ancestry components concentrated in Sardinians (yellow), Native Americans (Karitiana and Surui; purple), Papuans (magenta), northeast Siberians (skyblue) and central Siberians and northeast Asians (orange; most concentrated in the Nganasan).

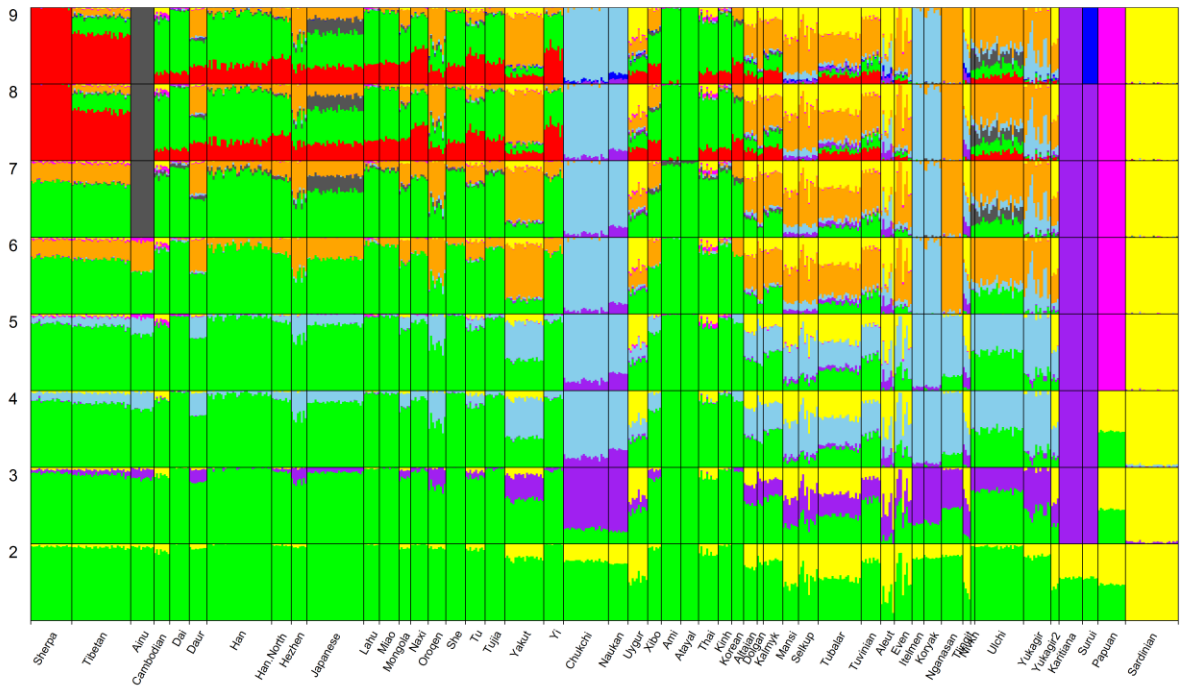


Figure S5 *ADMIXTURE* analysis of East Asian and Siberian populations with K = 2 to 9. This analysis includes two Nivkh individuals, showing a substantial proportion of ancestry shared with the Ainu with K ≥ 7 (dark grey).

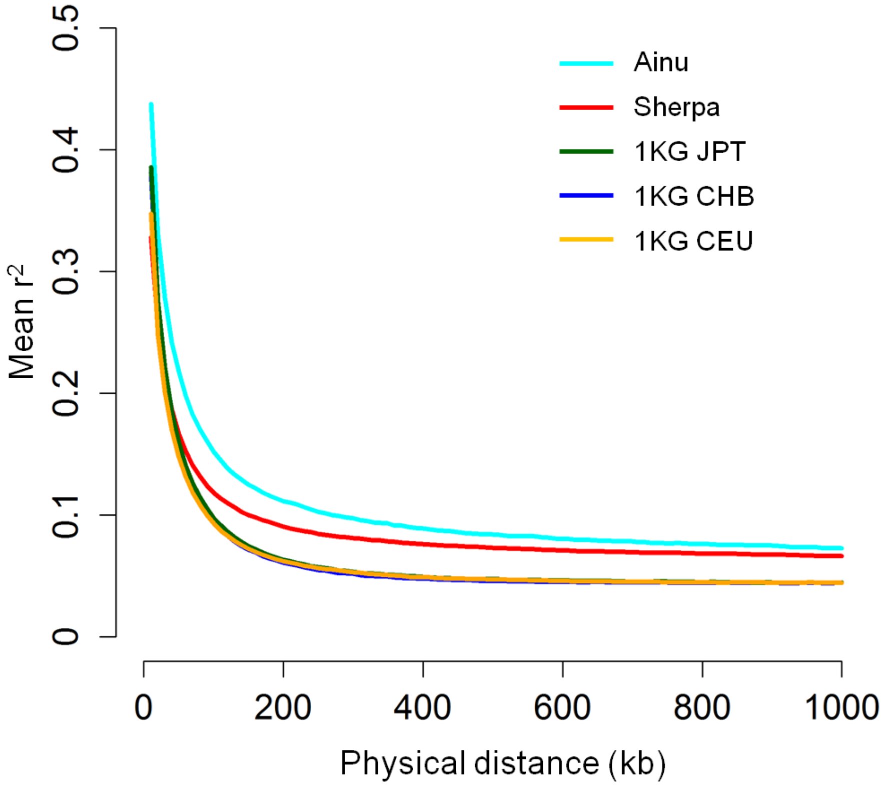
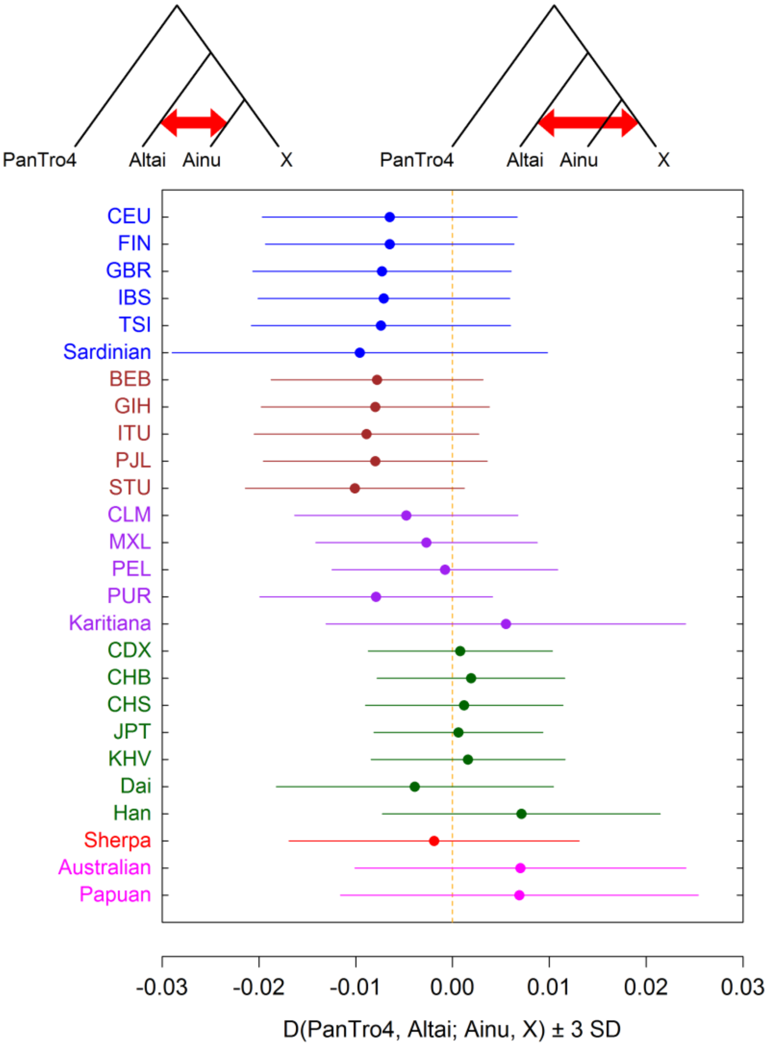


Figure S6 LD decay across physical distance in the Ainu, Sherpa and 1KG populations. Mean r^2 values were calculated for all pairs of SNPs for 10 kb distance bins. For each population, we randomly chose 12 individuals and calculated LD to match the Ainu sample size.

A


PanTro4
Altai
Ainu
X
PanTro4
Altai
Ainu
X
CEU
FIN
GBR
IBS
TSI
Sardinian
BEB
GIH
ITU
PJL
STU
CLM
MXL
PEL
PUR
Karitiana
CDX
CHB
CHS
JPT
KHV
Dai
Han
Sherpa
Australian
Papuan
-0.03
-0.02
-0.01
0.00
0.01
0.02
0.03
D(PanTro4, Altai; Ainu, X) ± 3 SD

B

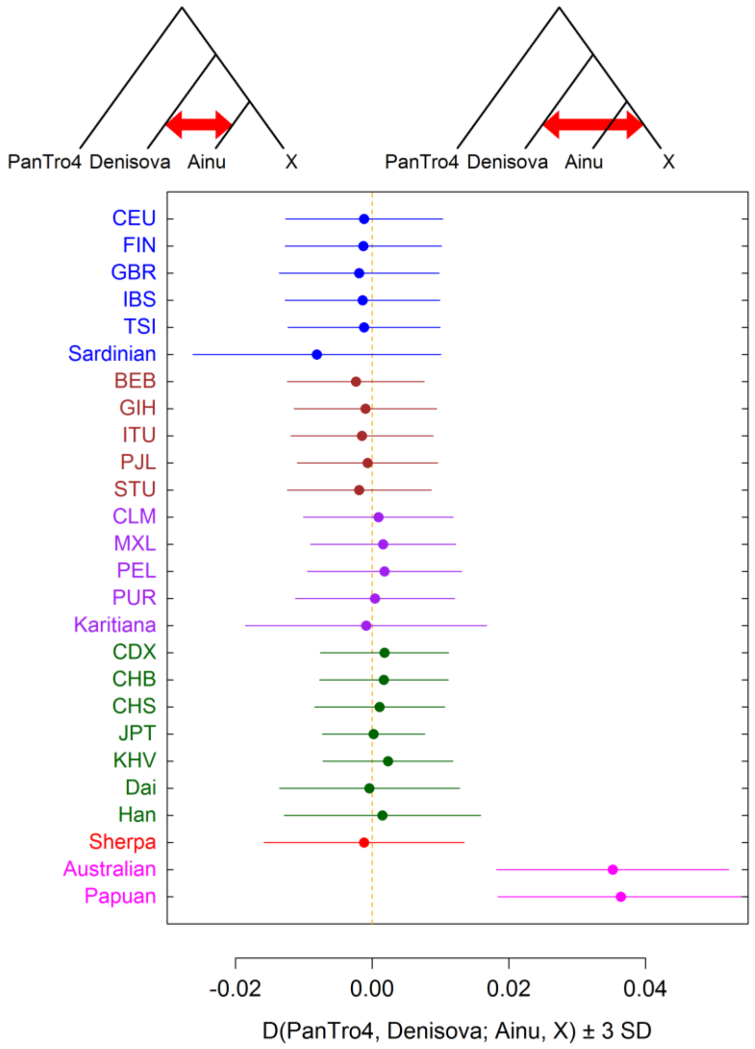
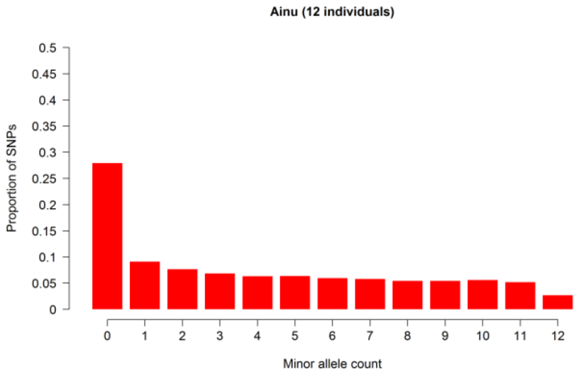
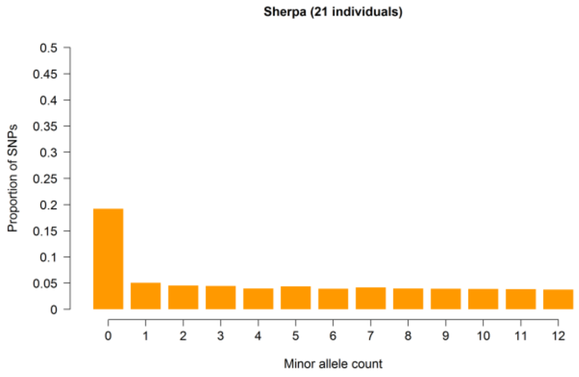


Figure S7 Genetic affinity of the Ainu and other non-African populations to archaic hominins, (A) Altai Neandertal and (B) Denisovan, measured by Patterson's D(YRI, Archaic; Ainu, X). The Ainu show a similar level of archaic ancestry with the other East Asian populations (|D| < 1.5 SD). The "CND-1KG-Ainu" data set was used for this analysis. Horizontal bars around the value represent ± 3 standard deviations.

A

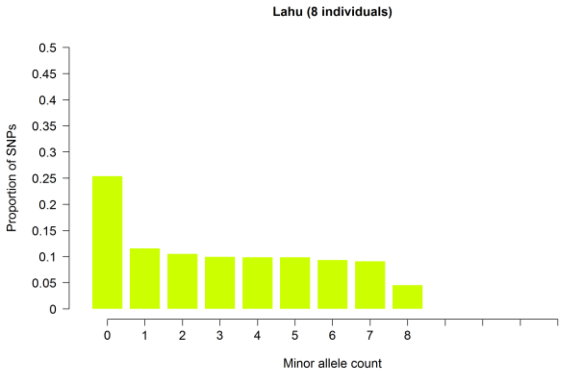
B

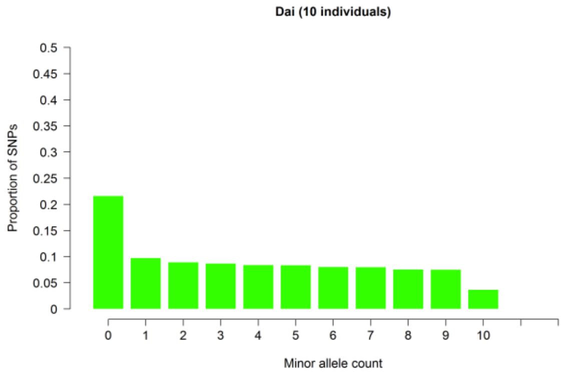

Ainu (12 individuals)
Proportion of SNPs
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9 10 11 12
Minor allele count


Sherpa (21 individuals)
Proportion of SNPs
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9 10 11 12
Minor allele count

C

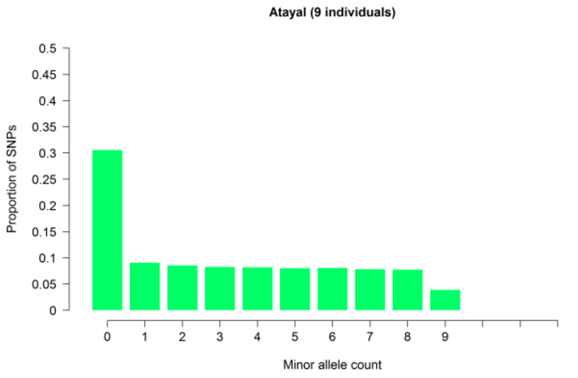
D

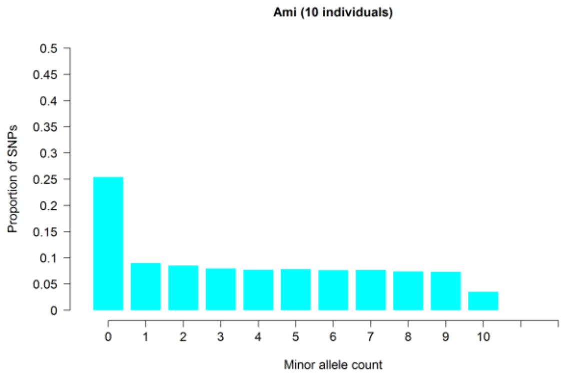

Lahu (8 individuals)
Proportion of SNPs
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8
Minor allele count


Dai (10 individuals)
Proportion of SNPs
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9 10
Minor allele count

E

F


Atayal (9 individuals)
Proportion of SNPs
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9
Minor allele count


Ami (10 individuals)
Proportion of SNPs
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9 10
Minor allele count

G

H

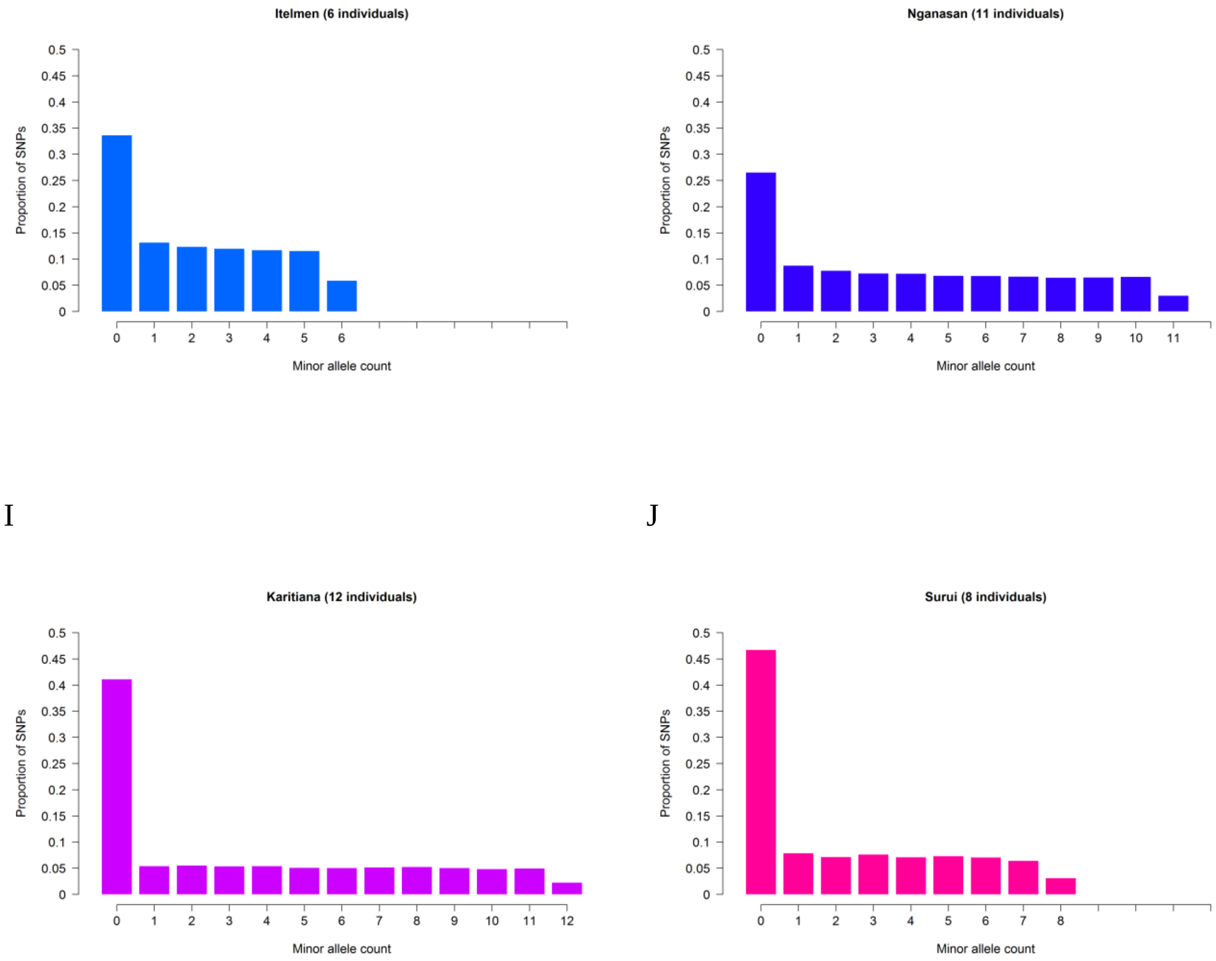


Figure S8 Minor allele count distribution of SNPs in the "WHA" data set in (A) Ainu, (B) Sherpa, (C) Lahu, (D) Dai, (E) Atayal, (F) Ami, (G) Itelmen, (H) Nganasan, (I) Karitiana and (J) Surui. In all populations, minor allele count distribution was flat except for high numbers among the fixed SNPs (19% in the Sherpa to 47% in Surui). The Ainu do not have a particularly high proportion of fixed SNPs (28%) in comparison to the other East Asian, Siberian or Native American populations.

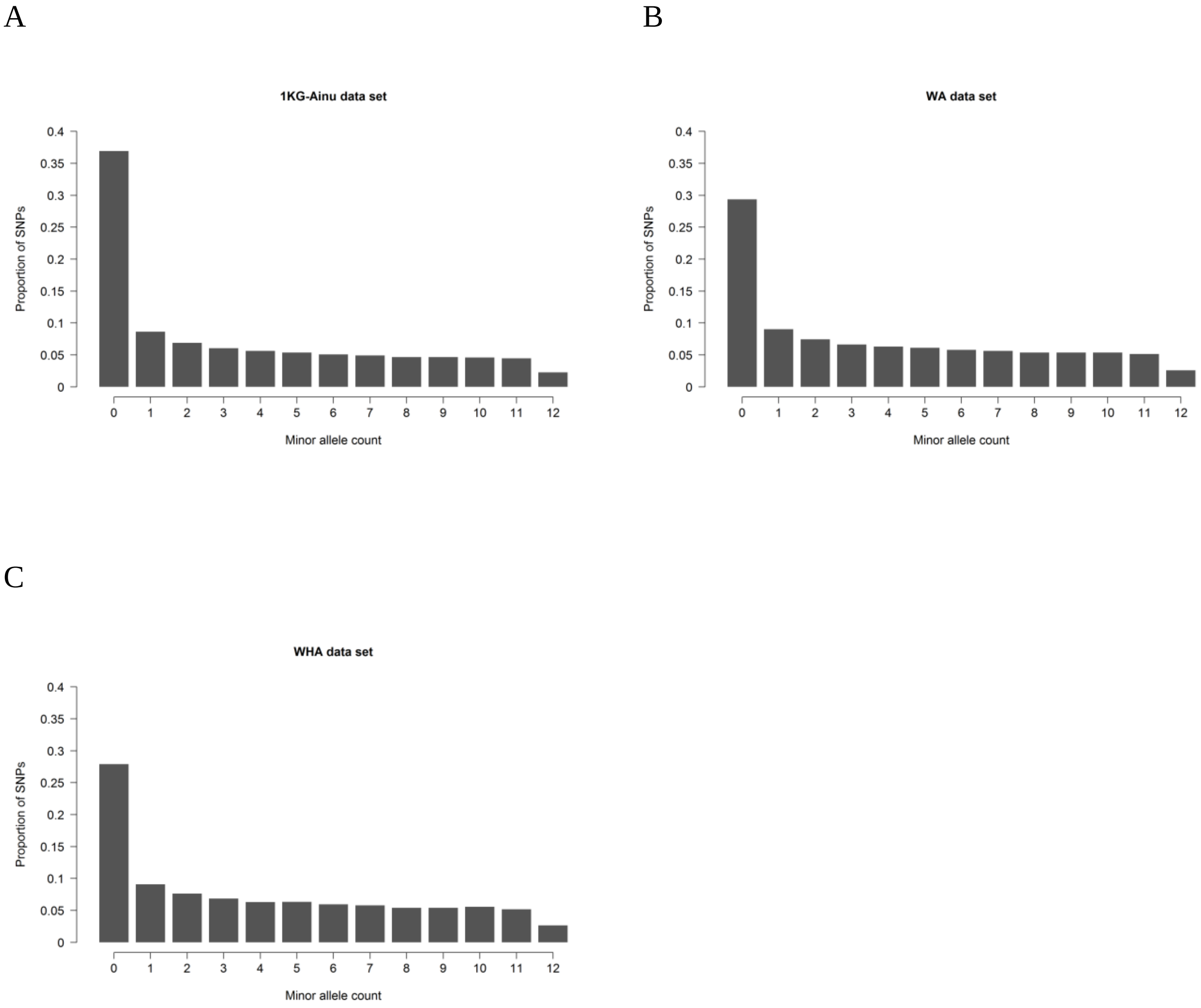


Figure S9 Minor allele count distribution in the 12 Ainu individuals in (A) “1KG-Ainu” (540,304 SNPs), (B) “WA” (103,218 SNPs) and (C) “WHA” (45,513 SNPs) data sets.

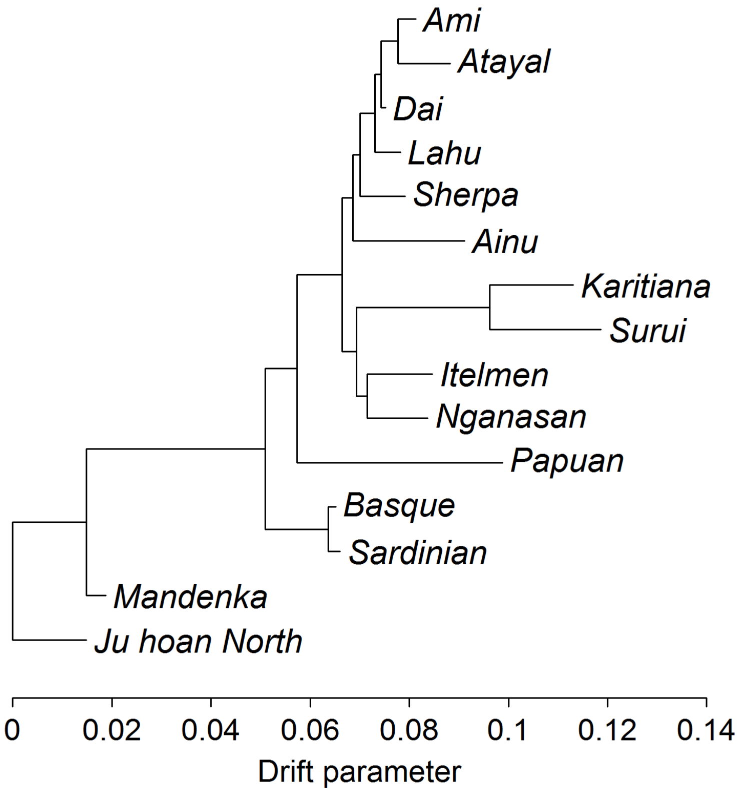
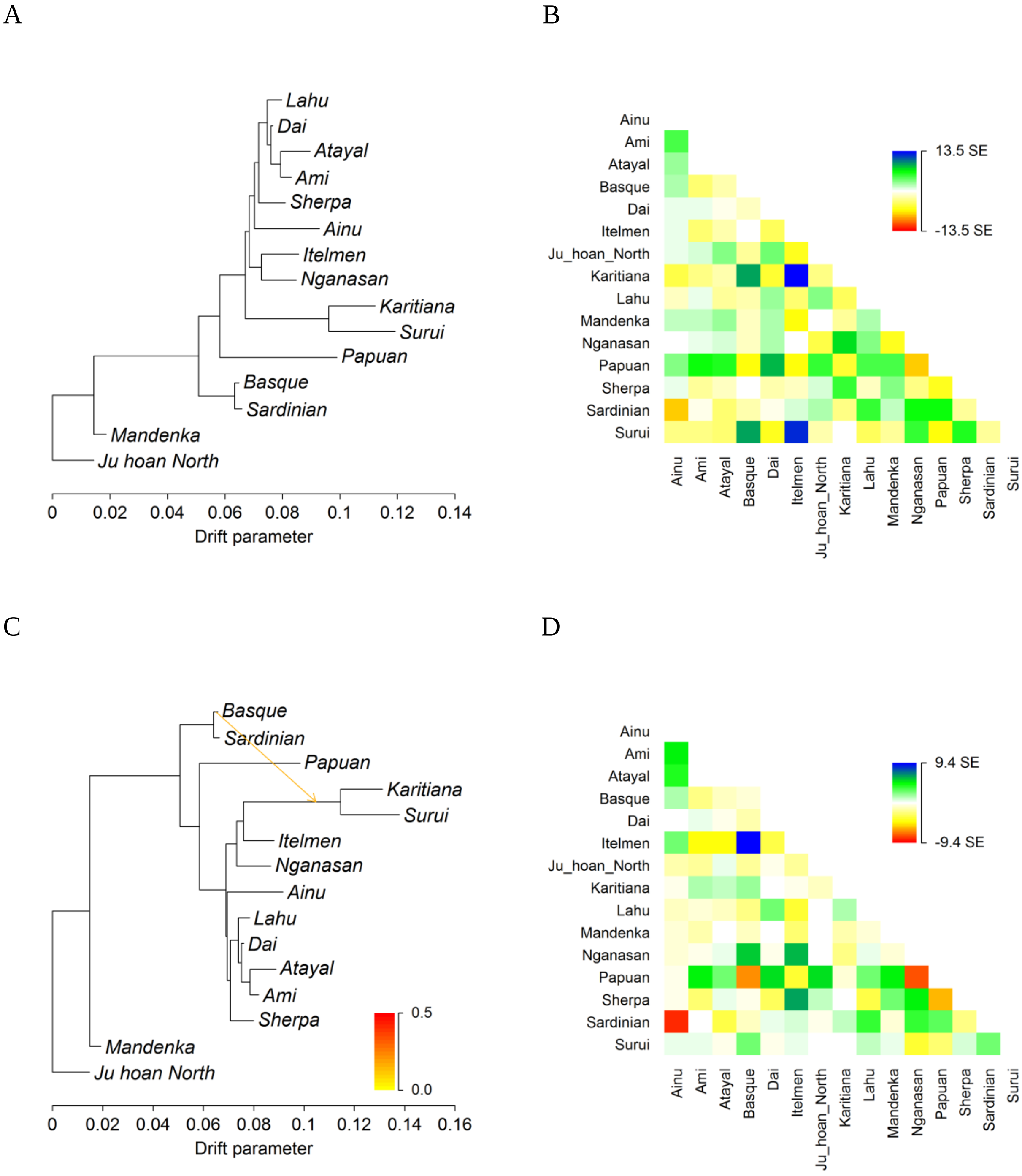
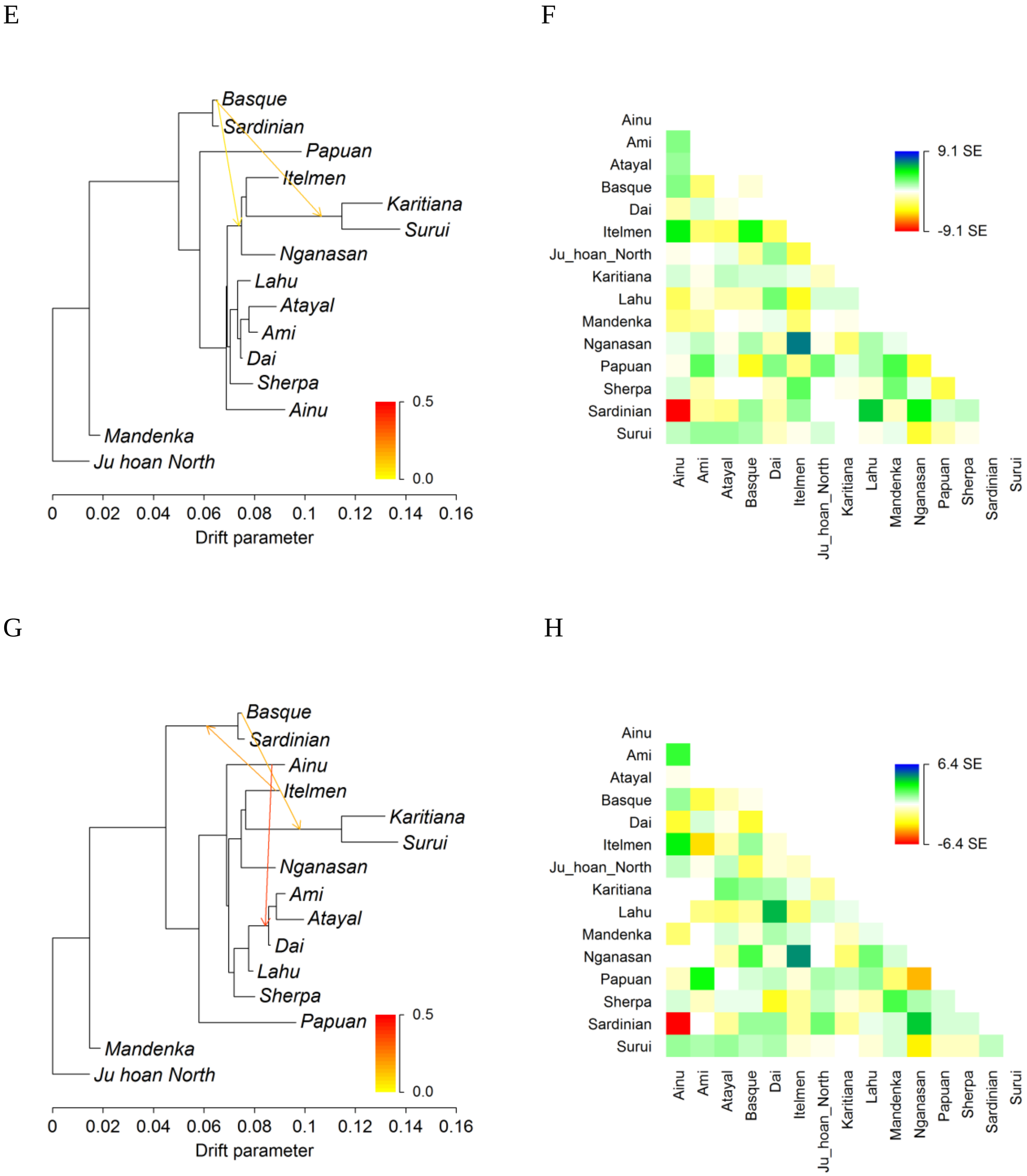


Figure S10 A tree with minority topology of 15 world-wide populations inferred from 500 bootstrap replicates of maximum likelihood trees using *TreeMix*. This topology was supported in 23.2% of replicates. The only difference from a consensus tree in Figure 3 is the position of (Itelmen, Nganasan) clade as a sister group to Native Americans.


A
B
Lahu
Dai
Atayal
Ami
Sherpa
Ainu
Itelmen
Nganasan
Karitiana
Surui
Papuan
Basque
Sardinian
Mandenka
Ju hoan North
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
Drift parameter
Ainu
Ami
Atayal
Basque
Dai
Itelmen
Ju_hoan_North
Karitiana
Lahu
Mandenka
Nganasan
Papuan
Sherpa
Sardinian
Surui
13.5 SE
-13.5 SE
C
D
Basque
Sardinian
Papuan
Karitiana
Surui
Itelmen
Nganasan
Ainu
Lahu
Dai
Atayal
Ami
Sherpa
Mandenka
Ju hoan North
0.5
0.0
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
Drift parameter
Ainu
Ami
Atayal
Basque
Dai
Itelmen
Ju_hoan_North
Karitiana
Lahu
Mandenka
Nganasan
Papuan
Sherpa
Sardinian
Surui
9.4 SE
-9.4 SE


E
F
Basque
Sardinian
Papuan
Itelmen
Karitiana
Surui
Nganasan
Lahu
Atayal
Ami
Dai
Sherpa
Ainu
Mandenka
Ju hoan North
0.5
0.0
0 0.02 0.04 0.06 0.08 0.1 0.12 0.14 0.16
Drift parameter
Ainu
Ami
Atayal
Basque
Dai
Itelmen
Ju_hoan_North
Karitiana
Lahu
Mandenka
Nganasan
Papuan
Sherpa
Sardinian
Surui
9.1 SE
-9.1 SE
G
H
Basque
Sardinian
Ainu
Itelmen
Karitiana
Surui
Nganasan
Ami
Atayal
Dai
Lahu
Sherpa
Papuan
Mandenka
Ju hoan North
0.5
0.0
0 0.02 0.04 0.06 0.08 0.1 0.12 0.14 0.16
Drift parameter
Ainu
Ami
Atayal
Basque
Dai
Itelmen
Ju_hoan_North
Karitiana
Lahu
Mandenka
Nganasan
Papuan
Sherpa
Sardinian
Surui
6.4 SE
-6.4 SE

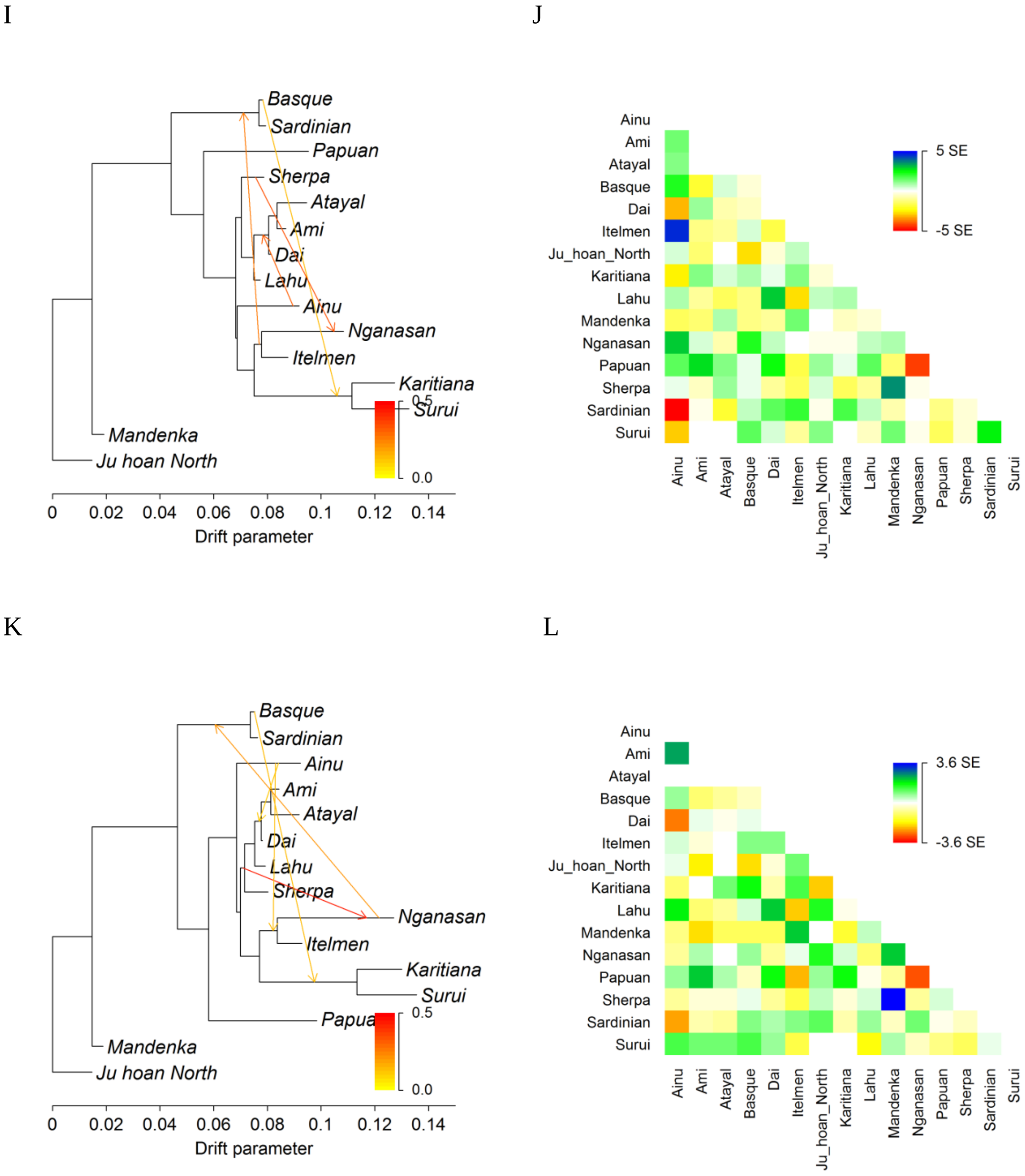


Figure S11 *TreeMix* results with 0 to 5 migration edges. A single representative run was chosen from 100 bootstrap replicates. (A, C, E, G, I, K) Maximum likelihood tree with 0 to 5 migration edges. (B, D, F, H, J, L) Residual covariance matrices for the corresponding trees.

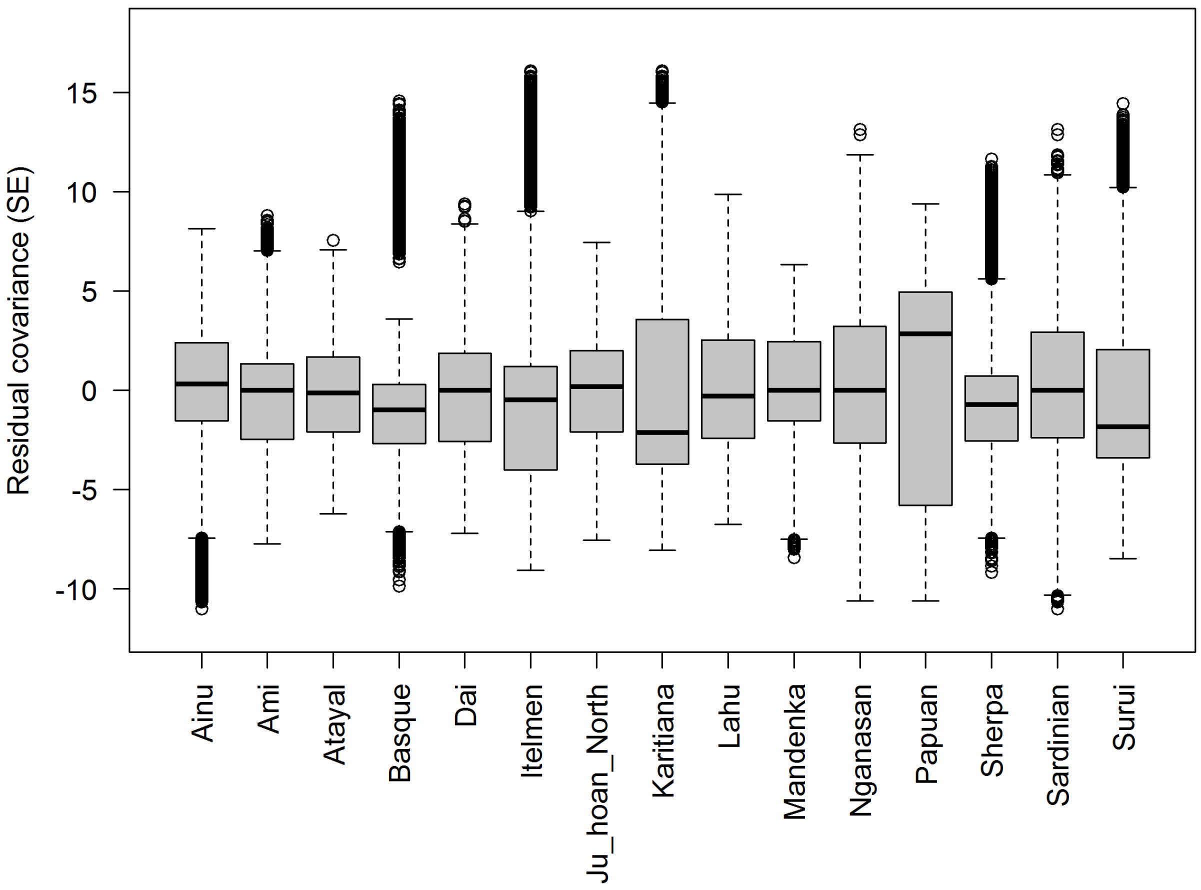


Figure S12 Distribution of residual covariance (in standard error, SE) for each population across 500 bootstrap replicates of *TreeMix* with no migration edge allowed. All populations show small mean deviations from zero, suggesting that a population tree without migration edges does not fully explain the data. However, the Ainu show a similar level of deviation from zero, suggesting that they are not an unusual outlier in this analysis.

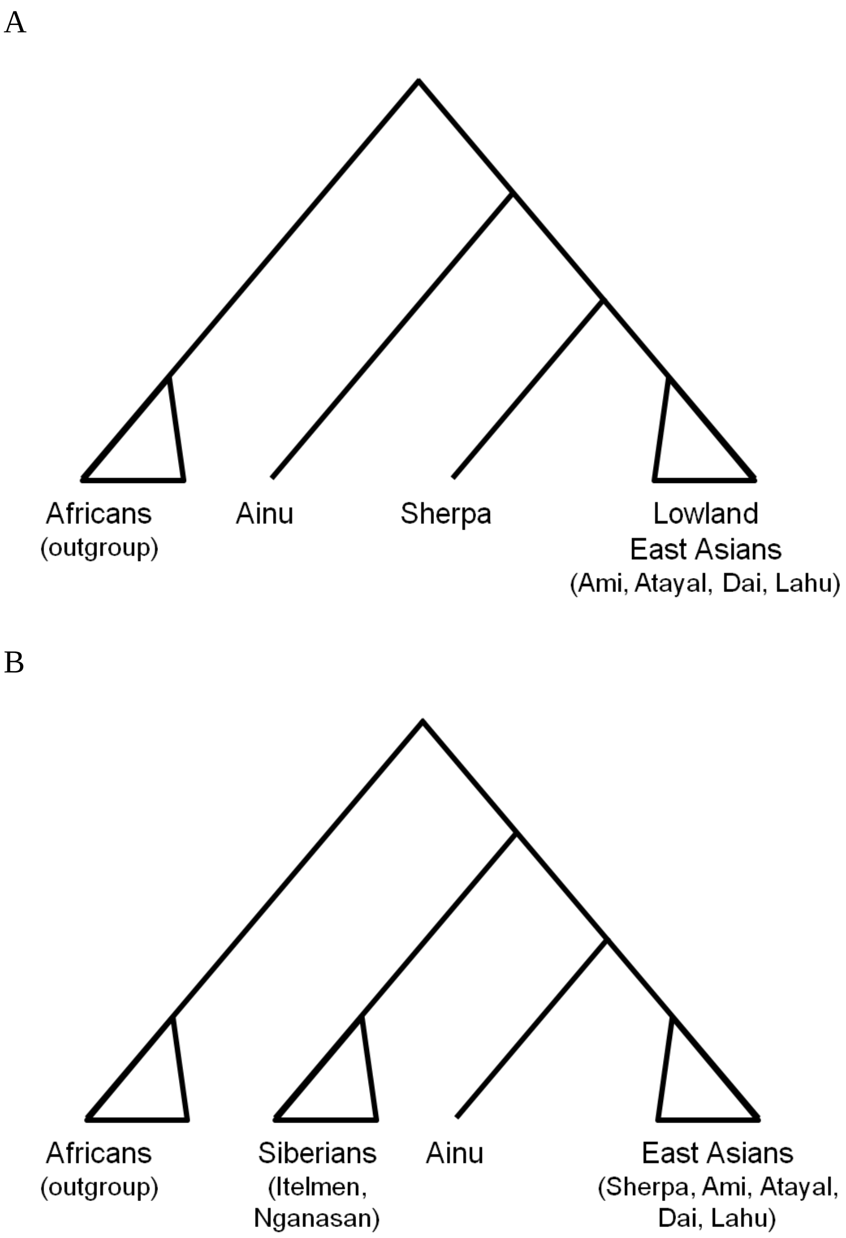


Figure S13 Two hypothetical scenarios of population relationships based on the population trees. (A) The Sherpa and lowland East Asian populations (Ami, Atayal, Dai and Lahu) are equally related to the Ainu, their shared outgroup. (B) The Ainu and other East Asian populations including the Sherpa are equally related to Siberian populations, their shared outgroup.

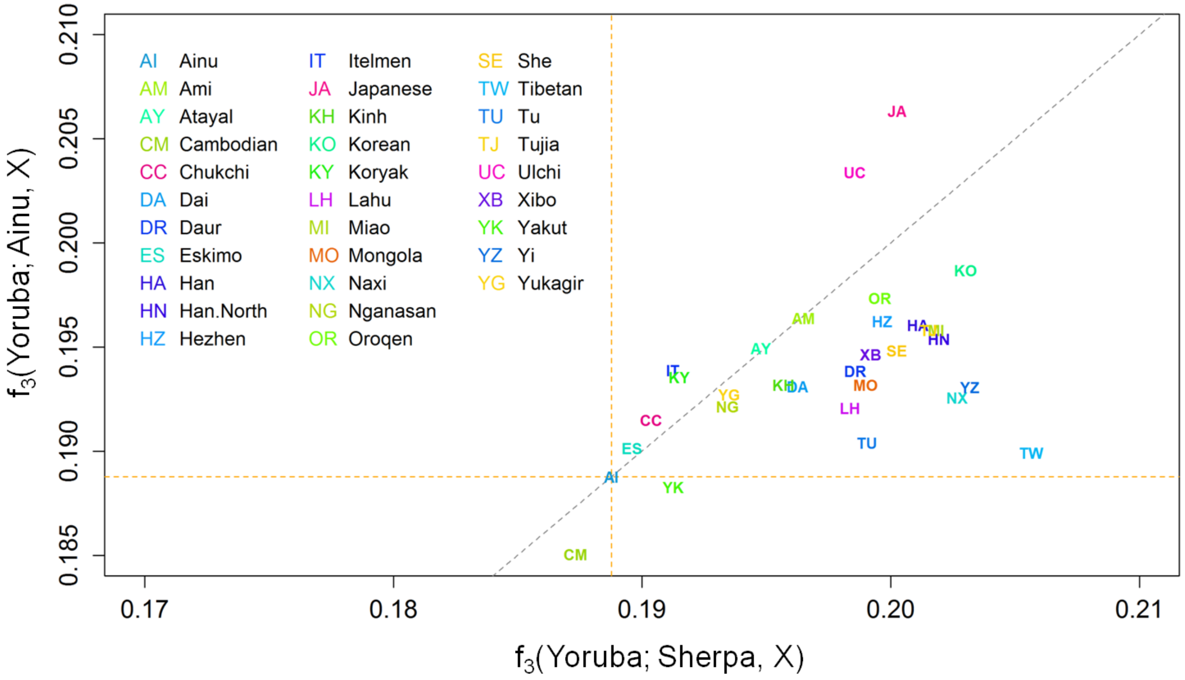


Figure S14 Genetic affinity of East Asian and Siberian populations with the Ainu and the Sherpa measured by outgroup f_3 statistic. Most East Asian populations are closer to the Sherpa than to the Ainu, except for Japanese (JA), Ulchi (UC) and northeast Siberians (IT, KY, CC and ES). Sherpa and East Asians as well as Ainu and East Asians are in general closer to each other than Sherpa and Ainu (marked by dotted orange lines). The dotted grey line marks a line with slope of 1.

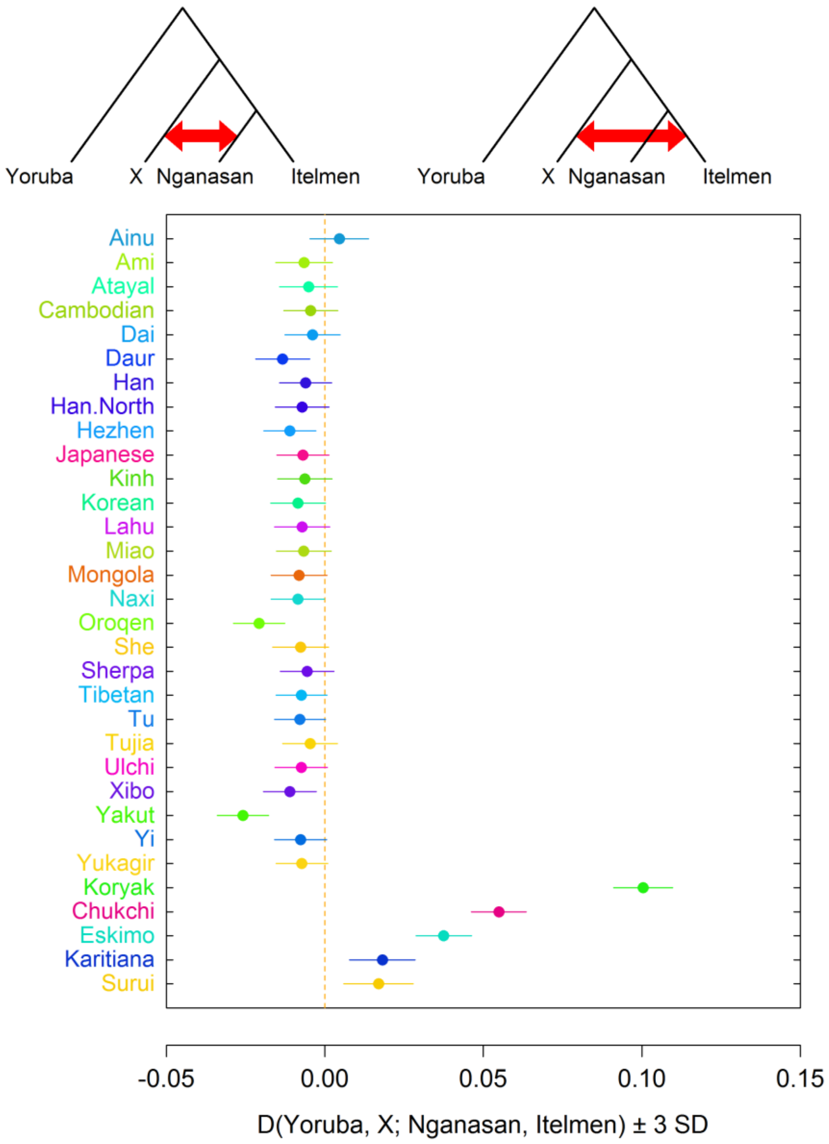


Figure S15 The genetic affinity of East Asian and Siberian populations to Nganasan and Itelmen measured by Patterson's D(Yoruba, X; Nganasan, Itelmen). In this analysis, we used only the 12 Ainu individuals without non-Ainu ancestry. Both the Sherpa and Tibetans are closer to the Nganasan than to the Itelmen as the other East Asians are. The "WHA" data set used for this analysis includes the Sherpa and Tibetan samples. Horizontal bars around the value represent ± 3 standard deviations.

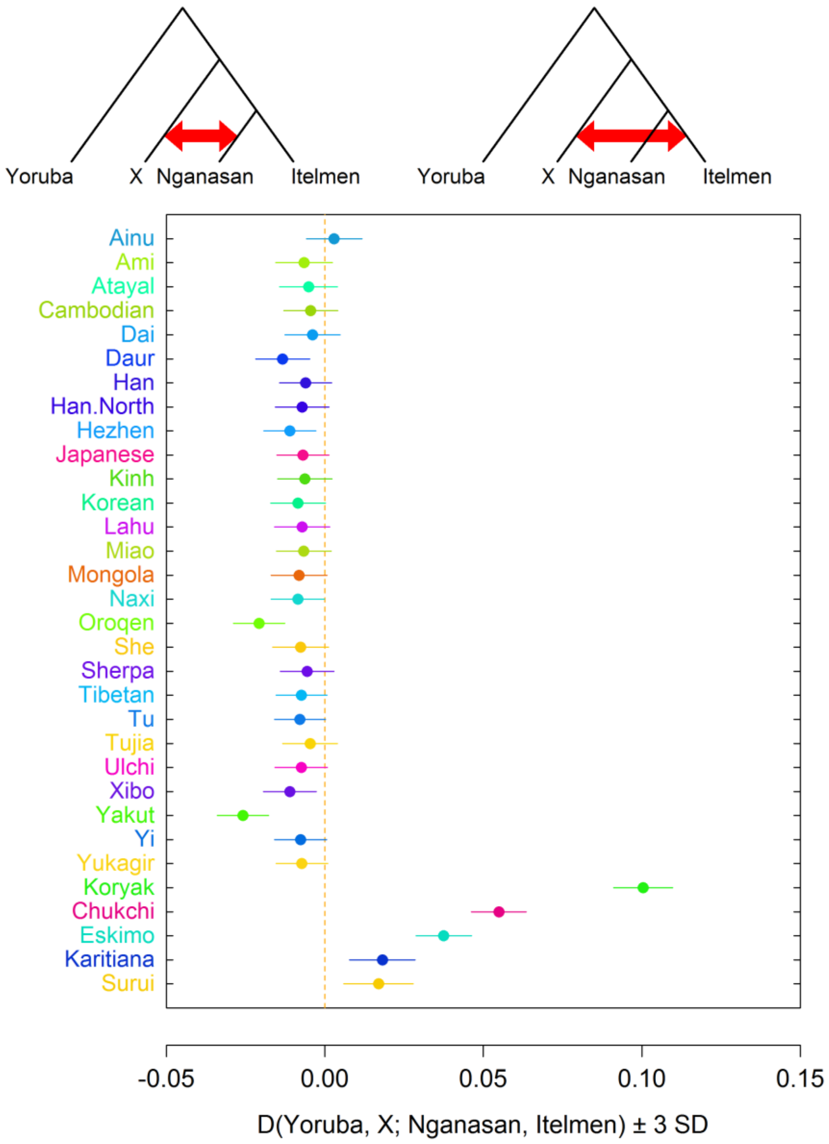


Figure S16 Genetic affinity of East Asian and Siberian populations to Nganasan and Itelmen measured by Patterson's D(Yoruba, X; Nganasan, Itelmen). In this analysis, we used the 22 unrelated Ainu individuals including the recently admixed ones. Comparing these results to those in Figure S15, it is clear that the conclusions are not sensitive to the omission of the recently admixed Ainu. The "WHA" data set used for this analysis includes the Sherpa and Tibetan samples. Horizontal bars around the value represent ± 3 standard deviations.

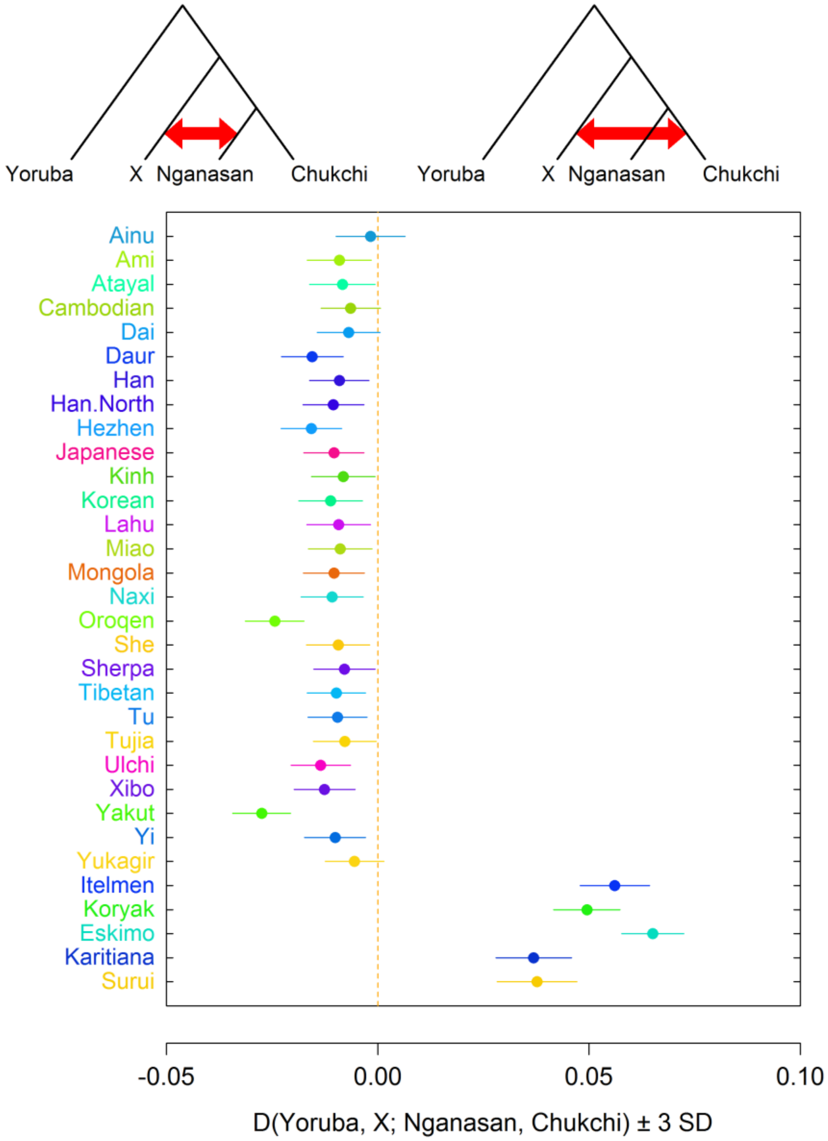


Figure S17 Genetic affinity of East Asian and Siberian populations to Nganasan and Chukchi measured by Patterson's D(Yoruba, X; Nganasan, Chukchi). All East Asian populations except for the Ainu are significantly closer to the Nganasan. Native Americans (Surui and Karitiana) and northeast Siberians (Eskimo, Itelmen and Koryak) are significantly closer to the Chukchi. Horizontal bars around the value represent ± 3 standard deviations.

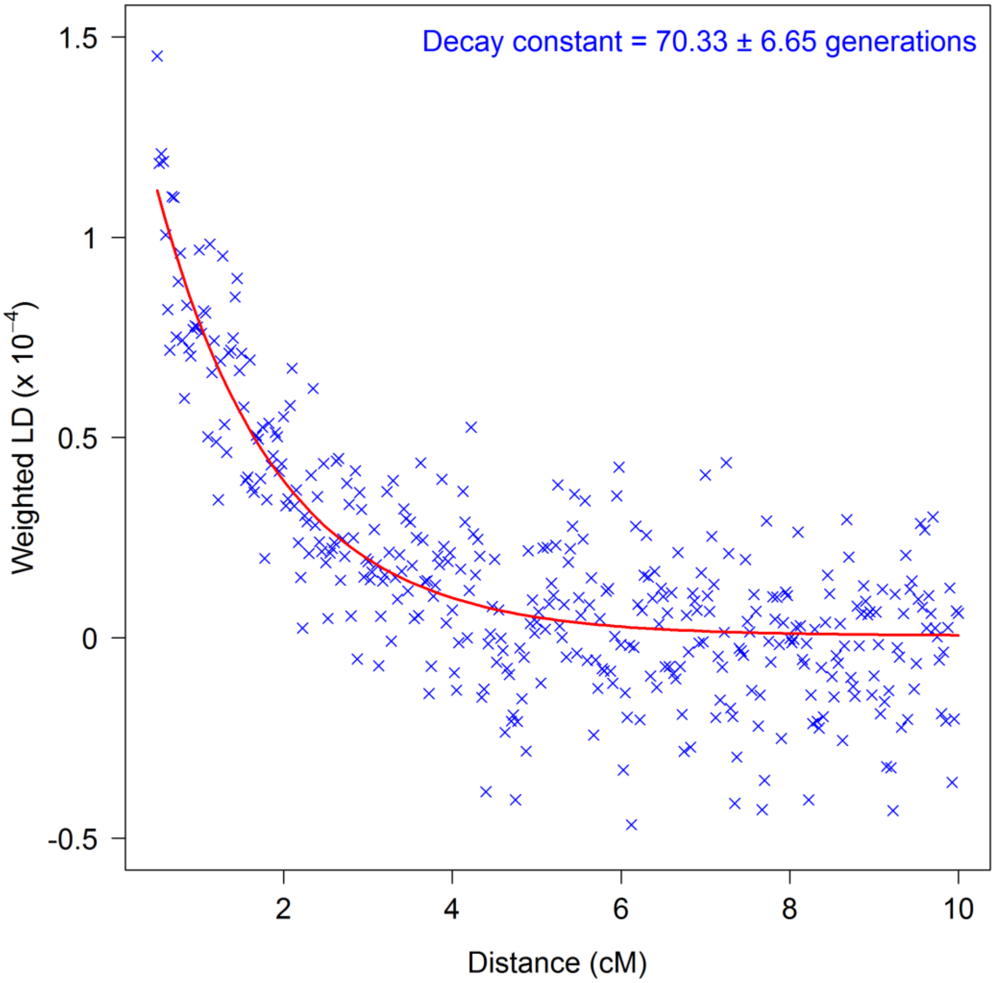


Figure S18 Weighted admixture LD decay in the Japanese with the Ainu and the Han as references.

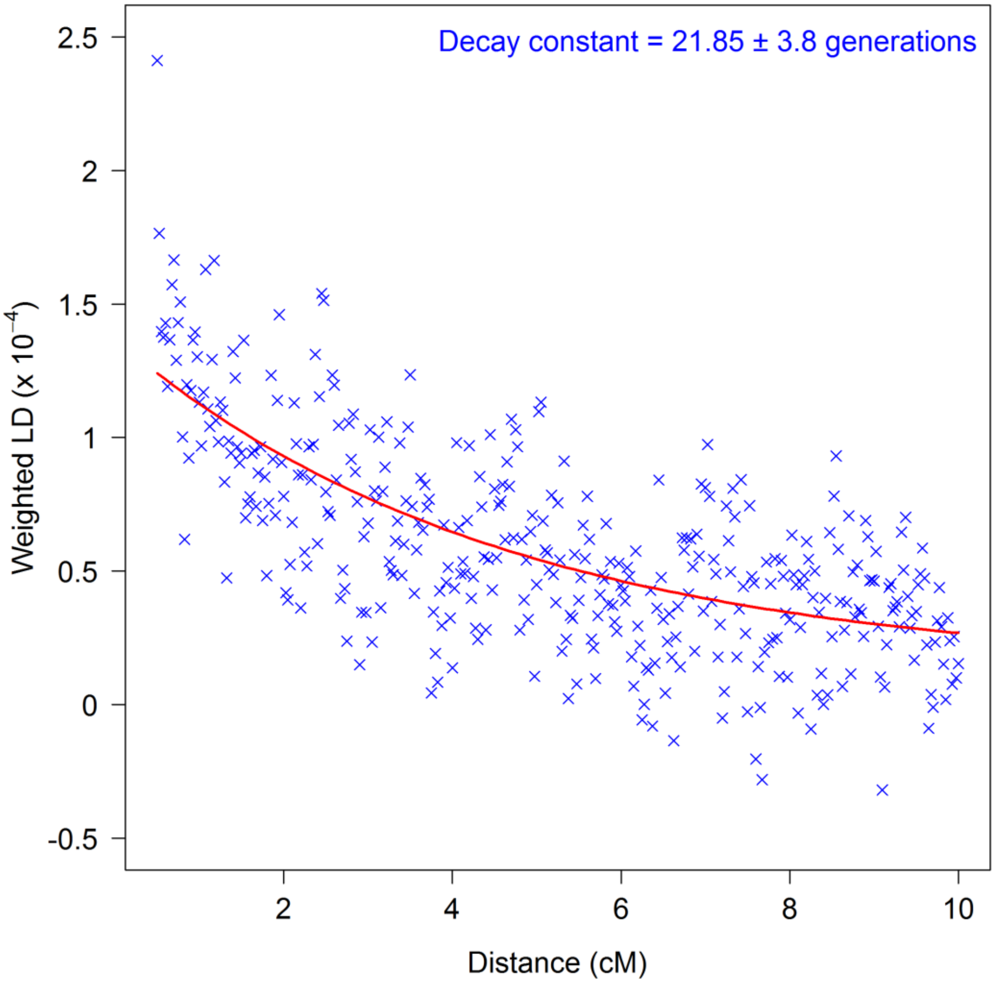
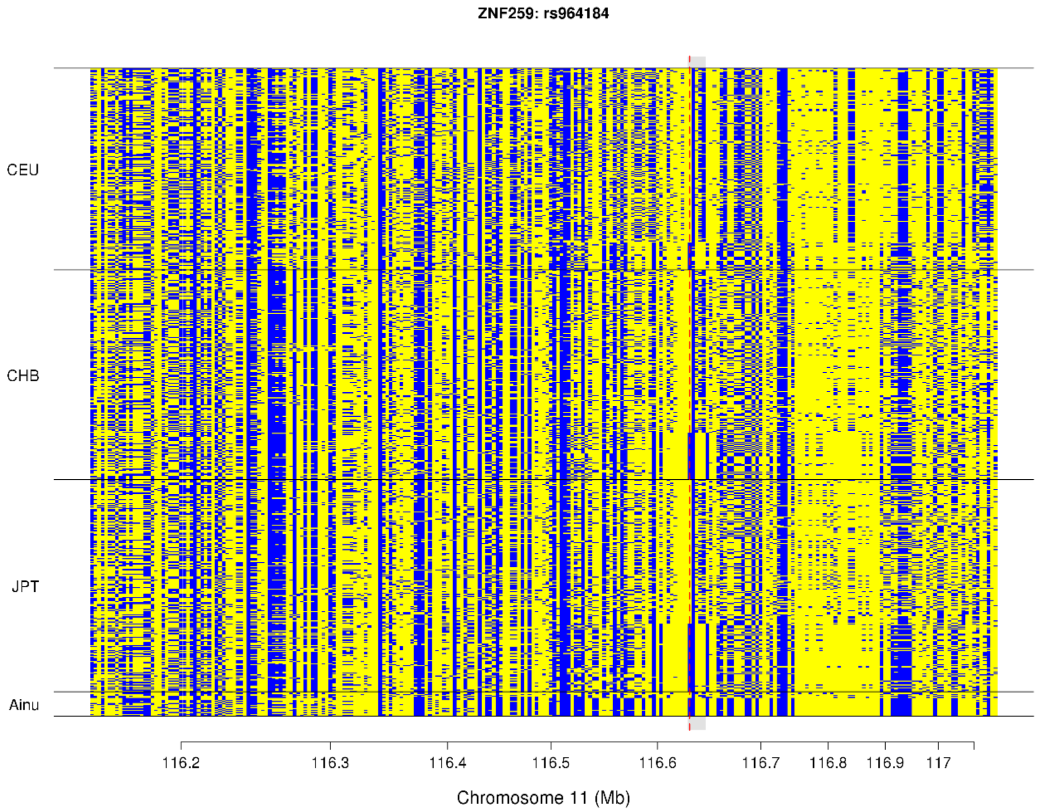
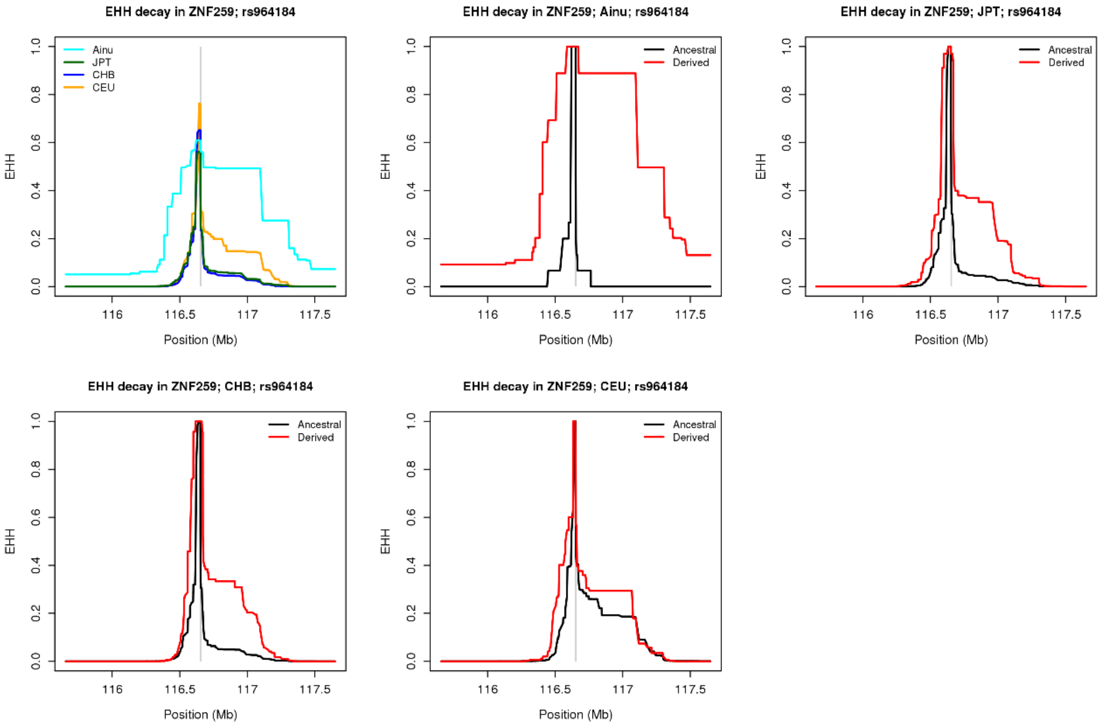


Figure S19 Weighted admixture LD decay in Ulchi population with the Ainu and the Nganasan as references.


ZNF259: rs964184
CEU
CHB
JPT
Ainu
116.2
116.3
116.4
116.5
116.6
116.7
116.8
116.9
117
Chromosome 11 (Mb)

A


EHH decay in ZNF259; rs964184
Ainu
JPT
CHB
CEU
EHH decay in ZNF259; Ainu; rs964184
EHH decay in ZNF259; JPT; rs964184
EHH decay in ZNF259; CHB; rs964184
EHH decay in ZNF259; CEU; rs964184
Ancestral
Derived
EHH
0.0
0.2
0.4
0.6
0.8
1.0
116
116.5
117
117.5
Position (Mb)

B

Figure S20 Haplotype structure and EHH decay around rs964184 near the *APOA1* gene. (A) Haplotype structure around rs964184 in the Ainu and 1KG phase 3 JPT, CHB and CEU populations. Each column represents a variant and each row represents a phased haplotype. Yellow and blue colors represent ancestral and derived alleles, respectively. (B) EHH decay around rs964184. Ainu haplotypes harboring a derived allele at rs964184 show extended LD.

Table S1 A list of 71 populations used for calculating three-population (f_3) and Patterson’s D statistics.

Chimp	Altai	Denisovan	LBK	Loschbour	MA1
BantuKenya	Biaka	Mandenka	Mbuti	Ju_hoan_North	Yoruba
BantuSA	Adygei	Basque	Bergamo	French	Orcadian
Russian	Sardinian	Tuscan	BedouinB	Druze	Mozabite
Palestinian	Balochi	Brahui	Kalash	Xibo	Cambodian
Dai	Daur	Han	Han_NChina	Hezhen	Japanese
Lahu	Miao	Mongola	Naxi	Oroqen	She
Tu	Tujia	Yakut	Yi	Chukchi	Eskimo
Sherpa	Tibetan	Ainu	Ami	Atayal	Kinh
Thai	Korean	Ulchi	Itelmen	Kalmyk	Koryak
Yukagir	Nganasan	Altaian	Mansi	Selkup	Tubalar
Tuvinian	Karitiana	Surui	Bougainville	Papuan	

Table S2 Populations outside of East Asia have a symmetric relationship with the Ainu and East Asian farmer populations (Ami, Atayal, Dai, Lahu and Sherpa), suggesting that the Ainu do not harbor substantial non-East Asian ancestry. Patterson's D statistic was calculated in the form of D(Pop1, Pop2; Pop3, Pop4), using Yoruba as an outgroup (Pop1).

Pop1	Pop2	Pop3	Pop4	D	Z
Yoruba	Sardinian	Ainu	Ami	-0.0035	-1.331
			Atayal	-0.0043	-1.481
			Dai	-0.0057	-2.305
			Lahu	-0.0049	-1.907
			Sherpa	-0.0031	-1.249
	MA1		Ami	-0.0039	-0.714
			Atayal	-0.0028	-0.469
			Dai	-0.0009	-0.173
			Lahu	-0.0023	-0.423
			Sherpa	0.0037	0.696
	Karitiana		Ami	0.0045	1.269
			Atayal	0.0039	1.007
			Dai	0.0007	0.194
			Lahu	0.0026	0.704
			Sherpa	0.0017	0.480
	Papuan		Ami	0.0028	0.816
			Atayal	-0.0021	-0.597
			Dai	-0.0009	-0.286
			Lahu	-0.0021	-0.611
			Sherpa	-0.0033	-1.032

Table S3 The Ainu form a clade with East Asian populations in comparison to contemporary populations outside of East Asia as well as to archaic hominins (|D| < 3 SD), which suggests that the Ainu do not harbor a substantial amount of non-East Asian ancestry. Based on the "CND-1KG-Ainu" data set, Patterson's D statistic was calculated in the form of D(Pop1, Pop2; Pop3, Pop4), using 1KG YRI as an outgroup (Pop1).

Pop1	Pop2	Pop3	Pop4	D	Z
YRI	Altai Neandertal	Ainu	JPT	-0.0011	-0.491
			CHB	0.0000	-0.016
			CHS	-0.0002	-0.074
			KHV	-0.0006	-0.223
			CDX	-0.0013	-0.520
			Sherpa	0.0003	0.068
	Denisova		JPT	-0.0014	-0.679
			CHB	-0.0003	-0.115
			CHS	-0.0003	-0.124
			KHV	-0.0001	-0.064
			CDX	-0.0007	-0.322
			Sherpa	0.0008	0.219
	CEU		JPT	-0.0024	-1.539
			CHB	-0.0016	-0.933
			CHS	-0.0034	-1.993
			KHV	-0.0036	-2.108
			CDX	-0.0048	-2.819
			Sherpa	-0.0012	-0.488
	GIH		JPT	-0.0025	-1.749
			CHB	-0.0025	-1.593
			CHS	-0.0030	-1.891
			KHV	-0.0027	-1.757
			CDX	-0.0038	-2.457
			Sherpa	-0.0027	-1.275
	Papuan		JPT	-0.0010	-0.434
			CHB	-0.0020	-0.805
			CHS	-0.0009	-0.354
			KHV	-0.0013	-0.505
			CDX	-0.0015	-0.562
			Sherpa	-0.0075	-1.949

Table S4 Major migration edges inferred from the *TreeMix* analyses, allowing 1 to 5 migration edges (m), suggest gene flow events between Europeans and Native Americans and/or Siberians, between the Ainu and lowland East Asian farmers, and to a lesser degree, between the Ainu and the Itelmen. Migration edges are counted across 100 bootstrap replicates. Here we tabulated migration edges inferred in more than 5% of replicates. When multiple populations are listed, the value refers to an internal branch that is the most recent common ancestor of all listed populations. Ainu related migration edges are highlighted in blue font. EA = East Asians, SB = Siberians, NA = Native Americans.

m	Source	Target	Count
1	Basque	(Karitiana, Surui)	77
	(Karitiana, Surui)	Basque	23
2	Basque or (Basque, Sardinian)	(Karitiana, Surui)	67
	(Karitiana, Surui)	(Basque, Sardinian)	33
	Itelmen	(Basque, Sardinian)	32
	Basque	(SB, NA)	31
	(EA, SB, NA)	(Karitiana, Surui)	10
	Sherpa	Nganasan	8
	Papuan	EA	14
3	Basque or (Basque, Sardinian)	(Karitiana, Surui)	74
	(Karitiana, Surui)	(Basque, Sardinian)	24
	Itelmen or (Itelmen, Nganasan)	(Basque, Sardinian)	41
	Basque	(SB, NA)	26
	Ainu	**(Ami, Atayal) or (Ami, Atayal, Dai)**	**43**
	Ami or (Ami, Atayal)	**Ainu**	**15**
	Sherpa	Nganasan	10
	Papuan	EA	13
	Ju_hoan_North	Papuan	12

(Continued in the next page)

Table S4 *(Continued from the previous page)*

m	Source	Target	Count
4	Basque or (Basque, Sardinian)	(Karitiana, Surui)	68
	(Karitiana, Surui)	(Basque, Sardinian)	28
	Itelmen, Nganasan or SB	(Basque, Sardinian)	36
	Basque	(SB, NA)	31
	Itelmen	Nganasan	9
	Ngansan	Itelmen	7
	Ainu	**(Ami, Atayal) or (Ami, Atayal, Dai)**	**52**
	Ami or (Ami, Atayal)	**Ainu**	**14**
	Itelmen	**Ainu**	**12**
	(EA, SB, NA)	(Karitiana, Surui)	14
	Sherpa	Nganasan	24
	(Ami, Atayal, Dai, Lahu, Sherpa)	Nganasan	7
	Papuan	EA	9
	Ju_hoan_North	Papuan	16
5	Basque or (Basque, Sardinian)	(Karitiana, Surui)	55
	(Karitiana, Surui)	Basque or (Basque, Sardinian)	46
	Itelmen, Nganasan or SB	(Basque, Sardinian)	39
	Basque	(SB, NA)	14
	Itelmen	Nganasan	8
	Ainu	**(Ami, Atayal) or (Ami, Atayal, Dai)**	**72**
	Ami or (Ami, Atayal)	**Ainu**	**17**
	Itelmen	**Ainu**	**17**
	Ainu	**Itelmen**	**6**
	Ainu	**(Itelmen, Nganasan)**	**12**
	(EA, SB, NA)	(Karitiana, Surui)	24
	Sherpa	Nganasan	24
	(Ami, Atayal, Dai, Lahu, Sherpa)	Nganasan	15
	Papuan	EA or (EA but Ainu)	18
	(Ami, Atayal) or Dai	Papuan	18
	Ju_hoan_North	Papuan	14

Table S5 The Ainu are more closely related to lowland East Asian farmer populations (Ami, Atayal, Dai and Lahu) than to the Sherpa or to Tibetans, suggesting gene flow between the two groups after lowland East Asians split from the high-altitude East Asians. Choice of outgroup population did not affect results, suggesting that significant positive D statistic is not due to gene flow between the Sherpa or Tibetans and non-East Asian populations.

Pop1	Pop2	Pop3	D(Pop1, Pop2; Pop3, X) (Z)			
			Ami	Atayal	Dai	Lahu
Yoruba	Ainu	Sherpa	0.0192 (7.465)	0.0155 (5.532)	0.0109 (4.855)	0.0084 (3.334)
		Tibetan	0.0163 (7.064)	0.0126 (4.933)	0.0080 (4.051)	0.0054 (2.427)
Sardinian		Sherpa	0.0201 (8.473)	0.0170 (6.596)	0.0138 (6.449)	0.0104 (4.597)
		Tibetan	0.0158 (7.411)	0.0128 (5.342)	0.0095 (5.071)	0.0061 (3.007)
MA1		Sherpa	0.0257 (5.907)	0.0199 (3.903)	0.0157 (3.742)	0.0147 (3.444)
		Tibetan	0.0202 (5.097)	0.0144 (3.503)	0.0101 (2.835)	0.0091 (2.335)
Karitiana		Sherpa	0.0175 (6.112)	0.0143 (4.432)	0.0128 (4.835)	0.0080 (2.897)
		Tibetan	0.0144 (5.724)	0.0112 (3.817)	0.0097 (4.247)	0.0049 (1.943)
Papuan		Sherpa	0.0139 (4.933)	0.0150 (4.908)	0.0090 (3.578)	0.0075 (2.920)
		Tibetan	0.0093 (3.687)	0.0105 (3.825)	0.0044 (2.091)	0.0029 (1.264)

Table S6 Northeast Siberians (Itelmen and Chukchi) are more closely related to the Ainu than to the other East Asians (Ami, Atayal, Dai, Lahu and the Sherpa). The Nganasan, a central Siberian population, do not show such an asymmetric relationship.

Pop1	Pop2	Pop3	Pop4	D	Z
Yoruba	Nganasan	Ainu	Ami	0.0021	0.670
			Atayal	-0.0008	-0.236
			Dai	-0.0021	-0.686
			Lahu	0.0004	0.120
			Sherpa	0.0032	1.032
	Itelmen		Ami	-0.0087	-2.710
			Atayal	-0.0102	-2.876
			Dai	-0.0102	-3.379
			Lahu	-0.0109	-3.318
			Sherpa	-0.0065	-2.060
	Chukchi		Ami	-0.0051	-1.639
			Atayal	-0.0074	-2.257
			Dai	-0.0072	-2.455
			Lahu	-0.0070	-2.222
			Sherpa	-0.0028	-0.955

Table S7 Allele frequencies of three SNPs with selection signals in East Asians.

		Allele		Derived allele frequency					
Gene	SNP	Anc	Der	Ainu	JPT	CHB	AMR	EUR	AFR
EDAR	Rs3827760	A	G	0.250	0.803	0.937	0.392	0.011	0.003
OCA2	Rs1800414	T	C	0.875	0.572	0.592	0.000	0.000	0.001
ADH	Rs3811801	G	A	0.500	0.702	0.592	0.000	0.000	0.000

Anc = ancestral allele; Der = derived allele; AMR = 1KG phase 3 American populations; EUR = 1KG phase 3 European populations; AFR = 1KG phase 3 African populations

Table S8 The list of 66 genomic regions harboring both XP-EHH and PBS signals in the Ainu, including top signal SNPs and the closest genes.

Region	XP-EHH				PBS			
	Top SNP	Top signal	Gene	Distance (kb)	Top SNP	Top signal	Gene	Distance (kb)
chr1:30817753-31217752	rs1188387	3.442	*MATN1*	34	rs4949290	1.470	*LAPTM5*	1
chr1:54467753-54967752	rs12734042	2.848	*SSBP3*	34	rs3753405	1.798	*SSBP3*	0
chr1:162017753-163067752	rs3927641	3.743	*NOS1AP*	0	rs10919316	1.280	*C1orf226*	3
chr1:175517753-175717752	rs10913052	3.039	*TNR*	0	rs12049604	1.065	*TNR*	0
chr1:232317753-232817752	rs12731562	2.990	*SIPA1L2*	0	rs10910538	1.219	*SIPA1L2*	0
chr2:13665703-13765702	rs2140525	2.852	*TRIB2*	794	rs6432398	1.108	*TRIB2*	814
chr2:48915703-49265702	rs4953650	3.031	*FSHR*	0	rs2268359	0.858	*FSHR*	0
chr2:133465703-133765702	rs7606532	3.125	*NCKAP5*	0	rs16841046	2.091	*NCKAP5*	0
chr2:134115703-134615702	rs1004045	3.316	*NCKAP5*	277	rs2012254	1.063	*NCKAP5*	0
chr2:139365703-139665702	rs12612135	3.312	*NXPH2*	107	rs10172094	1.110	*NXPH2*	113
chr2:150815703-151615702	rs2879927	2.945	*RND3*	129	rs16827946	1.128	*RND3*	287
chr3:11619244-11869243	rs301551	3.200	*TAMM41*	34	rs7618099	1.434	*TAMM41*	0
chr3:71269244-71369243	rs1522174	2.748	*FOXP1*	0	rs2037477	1.188	*FOXP1*	0
chr3:76769244-77219243	rs774590	3.290	*ROBO2*	268	rs1166980	1.150	*ROBO2*	160
chr3:77969244-78719243	rs1495598	3.624	*ROBO2*	407	rs4680999	2.217	*ROBO1*	33
chr3:80919244-81419243	rs17018503	3.151	*GBE1*	395	rs11714085	1.110	*GBE1*	189
chr3:95219244-96119243	rs6767219	3.676	*EPHA6*	439	rs9826768	2.510	*EPHA6*	906
chr3:136969244-137319243	rs1542535	2.632	*IL20RB*	336	rs12490164	1.566	*SOX14*	287
chr4:36418911-36518910	rs10001470	2.724	*DTHD1*	115	rs6531440	0.777	*DTHD1*	89
chr4:86718911-87218910	rs3796625	2.862	*ARHGAP24*	0	rs345326	1.910	*ARHGAP24*	0
chr4:111768911-112268910	rs12642421	2.965	*PITX2*	682	rs17042632	1.427	*PITX2*	352
chr4:188818911-189118910	rs7375901	3.959	*TRIML1*	13	rs12500011	1.102	*TRIML2*	0

(Continued in the next page)

Table S8 *(Continued from the previous page)*

Region	XP-EHH				PBS			
	Top SNP	Top signal	Gene	Distance (kb)	Top SNP	Top signal	Gene	Distance (kb)
chr5:17886344-18786343	rs6876500	2.789	*BASP1*	694	rs17631488	1.311	*CDH18*	695
chr5:77636344-79536343	rs784589	3.872	*ARSB*	0	rs16877259	1.451	*CMYA5*	44
chr6:4542106-5642105	rs12190412	3.696	*LYRM4*	0	rs7767658	1.756	*CDYL*	56
chr6:6792106-7592105	rs2764092	3.396	*SSR1*	0	rs9328401	1.753	*RREB1*	0
chr6:96142106-96692105	rs4839726	2.902	*MANEA*	106	rs4840245	1.594	*FUT9*	0
chr6:139742106-139992105	rs17406820	2.928	*CITED2*	74	rs9495561	1.430	*CITED2*	290
chr7:11293259-11893258	rs7782151	3.009	*THSD7A*	0	rs17165101	1.098	*THSD7A*	0
chr7:125743259-126093258	rs17149771	2.766	*GRM8*	2	rs671524	1.070	*GRM8*	280
chr7:134243259-134793258	rs6955887	3.554	*AKR1B15*	0	rs12666741	0.834	*BPGM*	9
chr7:152993259-153393258	rs84034	2.789	*ACTR3B*	475	rs6958821	1.259	*DPP6*	331
chr8:77365982-77515981	rs13259565	2.857	*ZFHX4*	151	rs16939290	1.083	*ZFHX4*	156
chr9:7646587-8196586	rs4742455	3.164	*PTPRD*	196	rs2026463	1.078	*C9orf123*	48
chr9:71946587-72596586	rs11139898	3.103	*PTAR1*	9	rs1493048	1.160	*C9orf135*	42
chr9:93596587-93946586	rs192703	2.988	*SYK*	69	rs16907244	1.594	*AUH*	30
chr9:106346587-106596586	rs7034565	2.758	*SMC2*	326	rs10990903	1.741	*SMC2*	402
chr9:114246587-114896586	rs7030655	2.722	*C9orf84*	0	rs10981136	1.039	*UGCG*	14
chr9:121296587-121446586	rs7871273	3.055	*DBC1*	535	rs7024908	1.328	*DBC1*	585
chr9:133296587-133696586	rs1215988	2.796	*ASS1*	0	rs476067	1.497	*ASS1*	0
chr9:134596587-134996586	rs3012755	4.001	*MED27*	55	rs2987405	0.935	*MED27*	0
chr10:1604427-2054426	rs4457675	2.985	*ADARB2*	52	rs17156778	1.811	*ADARB2*	0
chr11:70798510-71548509	rs1792287	3.075	*DHCR7*	88	rs7125171	1.215	*DHCR7*	28
chr11:72198510-72698509	rs341078	2.692	*PDE2A*	30	rs11235622	1.415	*FCHSD2*	0

(Continued in the next page)

Table S8 *(Continued from the previous page)*

Region	XP-EHH				PBS			
	Top SNP	Top signal	Gene	Distance (kb)	Top SNP	Top signal	Gene	Distance (kb)
chr11:74498510-74998509	rs1111425	4.036	*XRRA1*	0	rs3824903	1.119	*SLCO2B1*	0
chr11:116498510-117048509	rs7123583	3.777	*BUD13 / APOA5*	19 / 60	rs2186670	1.697	*BUD13 / APOA5*	87 / 128
chr11:124598510-125198509	rs600702	3.077	*PKNOX2*	0	rs7129737	0.960	*ROBO4*	0
chr12:341619-1091618	rs524468	3.675	*SLC6A13*	0	rs2305164	2.294	*SLC6A13*	0
chr12:3641619-4041618	rs12370980	3.012	*EFCAB4B*	0	rs11062788	1.290	*EFCAB4B*	0
chr12:51791619-52641618	rs7974792	3.104	*SLC4A8*	0	rs17126441	0.841	*LOC283403*	0
chr12:52991619-53341618	rs694714	2.706	*KRT72*	0	rs17116681	2.050	*KRT72*	0
chr12:70991619-71491618	rs11178602	3.162	*TSPAN8*	27	rs2717420	1.090	*PTPRB*	0
chr12:108441619-108941618	rs1515633	2.670	*CMKLR1*	42	rs4964714	1.168	*FICD*	36
chr14:84274003-84674002	rs17119226	3.274	*FLRT2*	1639	rs4904162	1.168	*FLRT2*	1382
chr14:99624003-99924002	rs807734	2.791	*BCL11B*	0	rs17098384	1.503	*BCL11B*	0
chr15:28021673-28221672	rs4778192	2.629	*OCA2*	0	rs1800414	1.379	*OCA2*	0
chr15:34021673-34271672	rs1036004	2.720	*RYR3*	0	rs7496144	1.237	*AVEN*	0
chr15:50821673-51771672	rs17647084	3.183	*TNFAIP8L3*	0	rs2899473	1.444	*CYP19A1*	0
chr15:53571673-54121672	rs4776161	2.986	*WDR72*	0	rs518263	1.962	*WDR72*	0
chr15:92271673-92771672	rs17696427	3.387	*SLCO3A1*	12	rs8032332	1.006	*SLCO3A1*	75
chr16:51945481-52445480	rs2010842	3.459	*TOX3*	151	rs12597728	1.598	*TOX3*	119
chr16:84745481-85245480	rs8058723	3.215	*FAM92B*	99	rs16974564	0.999	*USP10*	0
chr18:49836305-50186304	rs12607660	3.027	*DCC*	0	rs919634	1.378	*DCC*	9
chr18:59486305-59586304	rs7505697	3.428	*RNF152*	0	rs12607798	1.123	*RNF152*	5
chr20:2961795-3461794	rs6133002	3.133	*PTPRA*	0	rs6107292	1.474	*ATRN*	11
chr20:43961795-44611794	rs459681	2.861	*SPINT4*	18	rs6032336	2.468	*WFDC8*	0

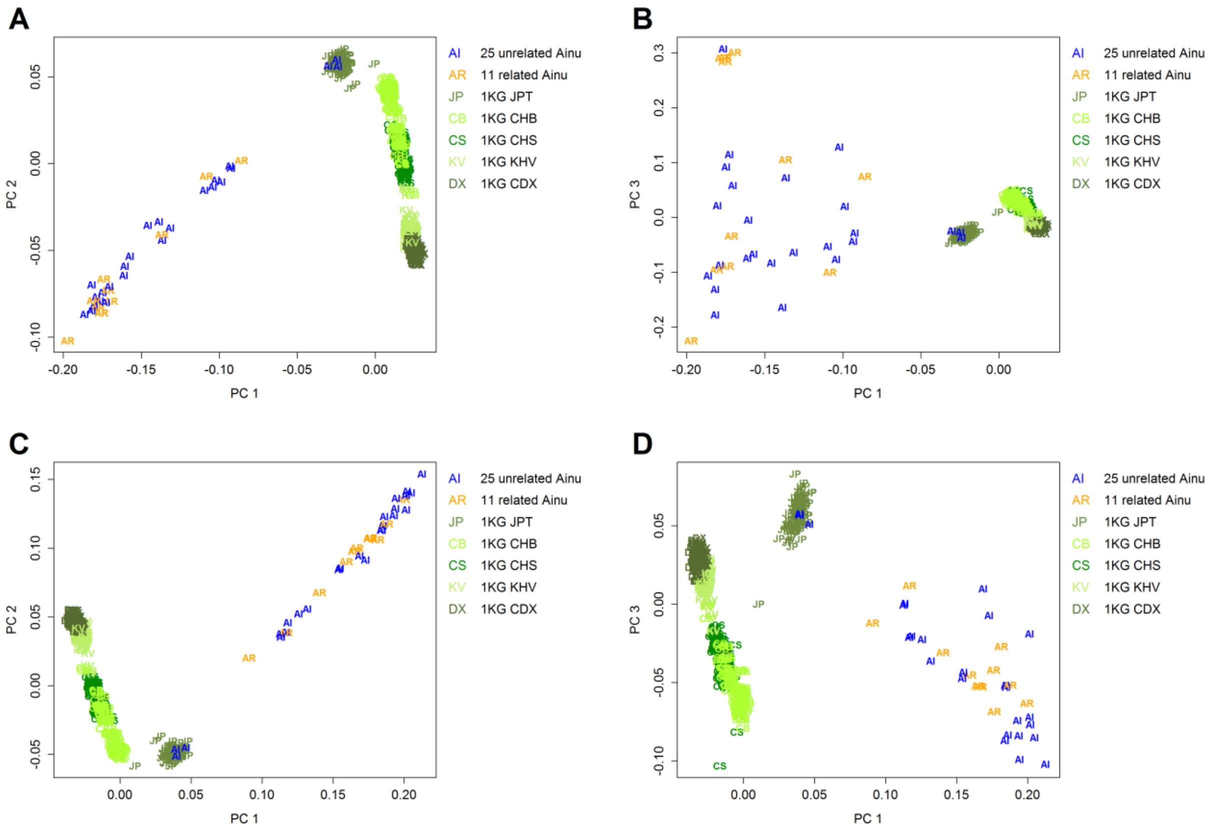
Table S9 Top 10 genomic regions harboring extreme PBS signal in the Ainu in comparison to CHB.

Region	Top SNP	Anc[1]	Der[2]	Derived allele frequency		SNP location	GWAS catalogue[4] (based on the name of gene)	Empirical P-value[5]			
								N_e=2,219		N_e=1,000	
				EA[3]	Ainu			T=800	T=1,000	T=800	T=1,000
Chr1: 54467753−54967752	rs3753405	T	C	1.1%	54.2%	*SSBP3*	None	0.004	0.010	0.066	0.112
Chr2: 133465703-133765702	rs16841046	T	C	15.0%	87.5%	*NCKAP5*	Hypersomnia, glaucoma, and height	0.000	0.001	0.015	0.031
Chr3: 77969244-78719243	rs4680999	G	T	1.0%	70.8%	Upstream of *ROBO1*	Aspartate transaminase in liver, brain activity in schizophrenia patients	0.000	0.002	0.020	0.046
Chr3: 95219244-96119243	rs9826768	G	A	0.1%	70.8%	Upstream of *EPHA6*	Blood trace element (Cu and Zn levels), serum free IGF-1 level (obesity related)	0.000	0.001	0.017	0.040
Chr4: 86718911-87218910	rs345326	G	A	10.1%	75.0%	*ARHGAP24*	PR interval in electrocardiogram	0.002	0.004	0.028	0.050
Chr10: 1604427-2054426	rs17156778	G	A	12.2%	75.0%	*ADARB2*	Radiation response in LCL, Emphysema, BMI, % body fat	0.002	0.005	0.036	0.063
Chr12: 341619-1091618	rs2305164	C	T	2.8%	83.3%	*SLC6A13*	Serum metabolite level (pyroglutamine, deoxycarnitine, betaine,), glomerular filtration rate	0.000	0.000	0.008	0.020
Chr12: 52991619-53341618	rs17116681	G	A	18.4%	91.7%	*KRT72*	None	0.001	0.001	0.013	0.023
Chr15: 53571673-54121672	rs518263	T	C	1.6%	62.5%	*WDR72*	Blood urea nitrogen level, serum creatinine level, glomerular filtration rate, longevity, cognitive function	0.001	0.005	0.038	0.076
Chr20: 43961795-44611794	rs6032336	A	G	95.4%	20.8%	*WFDC8*	None	0.000	0.000	0.012	0.030

[1] Ancestral allele; [2] Derived allele; [3] 1KG phase 3 East Asians (excluding JPT); [4] Last date accessed to the database (https://www.genome.gov/26525384) is April 8th, 2015; [5] Empirical *p*-values were calculated as the proportion of simulations with their simulated frequency of the derived allele equal to or greater than the observed frequency in Ainu.

File S1 Inclusion of close relatives in PCA generates artificial clusters in the Ainu

In a previous study of the same Ainu genotype data set, it was claimed that the Ainu received gene flow from mainland Japanese and from an unknown population (Jinam *et al.* 2012). These gene flow events were suggested based on their principal component analysis (PCA) using HapMap East Asians, mainland Japanese, Ryukyuans and all 36 Ainu individuals. More specifically, a scattered distribution of the Ainu individuals across their first PC, which separates the Ainu from other East Asians, was interpreted as evidence for admixture with mainland Japanese. Likewise, a cluster of five Ainu individuals apart from the others along their second PC was interpreted as evidence for another admixture event with the unknown second source population. Analyzing the same Ainu data, we clearly detected a recent gene flow from mainland Japanese into the Ainu (Figures S2-S3), confirming Jinam *et al*'s first finding. However, we did not find supporting evidence for the second admixture event from an unknown source in our PCA results, in which we used 25 unrelated Ainu individuals (Figure S2B). Our additional analysis strongly suggests that this inconsistency is easily explained by the fact that we omitted 11 closely related individuals from the analysis (Figure S1). It is well known that relatedness may affect PCA results whereby related individuals tend to cluster closely in PCA plots. Here, we provide PCA plots including related individuals (panels A and B), which generate clusters similar to those in Jinam *et al* but not present in PCA plots without related individuals (panels C and D). Specifically, panel B recapitulates Figure 1a of Jinam *et al*, which reported a cluster of five Ainu individuals along their second PC (upper left side in the plot). We found that they are highly related to each other. So, we think this difference in the data is a major contributor to the apparent discrepancy. For running PCA, we used 504 1KG phase 3 East Asians and the Ainu individuals, either all 36 (for panels A and B) or 25 unrelated ones (for panels C and D). For panels C and D, 11 related Ainu individuals were projected onto PC planes.



Text S1. Literature Cited

Jinam, T., N. Nishida, M. Hirai, S. Kawamura, H. Oota *et al.*, 2012 The history of human populations in the Japanese Archipelago inferred from genome-wide SNP data with a special reference to the Ainu and the Ryukyuan populations. J. Hum. Genet. 57: 787-795.