Comprehensive analysis of Japanese archipelago population history by detecting ancestry-marker polymorphisms without using ancient DNA data

Yusuke Watanabe[1,2], Jun Ohashi[1*]

[1]Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo 113-0033, Japan

[2]Present Adress: Genome Medical Science Project Toyama Project, National Center for Global Health and Medicine, Tokyo 162-8655, Japan.

*Correspondence: juno-tky@umin.ac.jp

Abstract

Modern Japanese have two major ancestral populations: the indigenous Jomon people and immigrants from continental East Asia. To figure out the population history in Japanese archipelago, we developed a reference-free detection method of genetic components from ancestral populations using a summary statistic, the ancestry-marker index (*AMI*). We applied the *AMI* to modern Japanese samples and identified 208,648 SNPs that were likely derived from the Jomon people (Jomon-derived SNPs). The analysis of Jomon-derived SNPs in 10,842 modern Japanese individuals recruited from all the 47 prefectures of Japan showed that the genetic differences among the prefectures were mainly caused by differences in the admixture proportion of the Jomon people and the population size of immigrants varied between regions in mainland Japan. We also estimated the migration route of the ancestral Jomon population to Japanese archipelago and their phenotype frequencies based on the haplotype structures of modern Japanese composed of Jomon-derived SNPs.

Introduction

Modern Japanese populations are divided into three main populations: the Ainu, who live mainly in Hokkaido; the Ryukyuan, who live mainly in Okinawa; and mainland Japanese, who live in Honshu, Shikoku, and Kyushu (Fig. 1). As a powerful hypothesis of the formation processes of Japanese populations, a dual structure model[1] was proposed based on morphology. This model assumes that Japanese originated through a mixture of the Jomon people, who settled in the Japanese archipelago during the Jomon period (from 16,500 YBP to 2,800 YBP)[2–4], and the immigrants came to the Japanese archipelago from continental East Asia around the beginning of the Yayoi period (around 2,800 YBP)[4]. According to this model, compared to mainland Japanese, the Ainu and the Ryukyuan were genetically less influenced by immigrants. Findings from genetical studies not only support the dual structure model, but also reveal the detailed population history of Japanese archipelago[5–11]. Whole-genome analyses extracted from the remains of the Jomon people have suggested that the Jomon were highly differentiated from other East Asians, forming a basal lineage to the East and Northeast Asians[8,10,11]. The genetic relationship between a Jomon individual and other East Asians suggested that the ancestral population of the Jomon people is one of the earliest-wave migrants who might have taken a coastal route on the way from Southeast Asia toward East Asia[11]. It was also revealed that the Jomon people were genetically closely related to the Ainu/Ryukyuan, and that 10–20% of genomic components found in mainland Japanese are derived from the Jomon people[8,10]. Altogether, though some

previous studies on ancient Jomon genomes have referred to the population history of Japanese archipelago, these studies were based on a single Jomon individual. To understand the population-level characteristics, genetic information for multiple individuals is essentially required. The fact that the genomes of modern Japanese have inherited Jomon-derived genomic components means that the genomes of the modern Japanese contains equivalent genetic information of multiple individuals of the Jomon people. With such information, we can clarify, for example, how they admixed in the Japanese Archipelago. From regional differences in the Jomon-derived genomic components of modern Japanese, we can infer the admixture process between the Jomon people and continental East Asians in Japanese archipelago. It will also provide more reliable results on the migration route of the ancestral Jomon population to the Japanese archipelago, which has been discussed based on the genome of a single Jomon individual[10,11]. Moreover, while the previous study have referred to the phenotype of a single Jomon individual[10], genomic information of modern Japanese will reveal the phenotypic characteristics of the Jomon people as a population (i.e. phenotype frequency). In order to achieve the above objectives, the present study attempted to detect Jomon-derived genomic components of modern Japanese.

In populations derived from a mixture of two source populations, recombination between haplotypes from different source populations inevitably occurs after the admixture event. As a result, haplotypes from two ancestral populations are patchily present in the chromosomes of admixed population,

and the alleles in the haplotypes from each ancestral population are in linkage disequilibrium (LD) with each other (Supplementary Fig. 1). Most existing methods[12–16] for estimating the local ancestry of genomes using the LD state require genome information of ancestral populations or of modern populations as proxies of ancestral populations. In the Japanese population, only a few Jomon individuals[8–11] have been sequenced so far. Furthermore, the sequence depths of these samples are low, except for the Funadomari Jomon[10] people excavated from Rebun Island in Hokkaido, making it difficult to obtain sufficient genome information to accurately estimate the local ancestry of modern Japanese (i.e., the genomes of many Jomon individuals are required to perform highly accurate studies). Therefore, at present, it is not possible to estimate the local ancestry of modern mainland Japanese using previous methods. In this study, we developed a method using a summary statistic, the Ancestry-Marker Index (*AMI*), to detect ancestry-marker SNPs derived from the Jomon people in modern mainland Japanese that does not require genomes obtained from Jomon skeletal specimens. The *AMI* was developed with inspiration from S*, detecting archaic-hominin-derived haplotypes using the specific SNPs in the out-of-Africa population which assumed to be originate from admixture of archaic hominin and early Eurasians[17–20]. Since the Jomon people have been found to be highly differentiated from other East Asians[8,10], they are expected to have had specific variants that were not found in other East Asians. Thus, the modern mainland Japanese also likely have specific variants derived from the Jomon people. The *AMI* detects the Jomon-derived variants based on the LD between Japanese specific variants. Based on the *AMI*, we could successfully extract the

Jomon-derived variants from real genomic data of the Japanese. Using Jomon-derived variants, we conducted comprehensive analysis to elucidate the population history of the Japanese archipelago population; for example, regional differences, the migration route to Japanese archipelago and phenotypes of the Jomon people. Based on these results, we propose a model of admixture between the Jomon people and immigrants from continental East Asia in the Japanese Archipelago.

Results

Performance of *AMI*

To confirm the usefulness of the *AMI*, we performed a coalescent simulation assuming a mixture of the Jomon people and continental East Asians (Supplementary Fig. 2). The preliminary 10 Mb simulation suggested that modern Japanese still have haplotypes of several megabases that are derived from the Jomon people (see Supplementary Fig. 3 and Supplementary Note 1).

There are three types of Japanese specific variants: (type 1) Jomon-derived variants, which appeared in the Jomon lineage before the admixture event; (type 2) variants derived from continental East Asians, which appeared in the continental East Asian populations and were moved into the Japanese lineages through the admixture, but were eventually lost in the East Asian population; and (type 3) novel variants that appeared only in Japanese lineages after the admixture (Supplementary Fig. 4). Our aim is to distinguish Jomon-derived variants (type 1) from (type 2) and (type 3). Of these Japanese specific variants, the Jomon-derived variants (type 1) are considered to be accumulated on the same haplotype or to be in strong LD with each other (Supplementary Fig. 4 (b)). In the subsequent 1 Mb simulation, Japanese specific variants (types 1, 2, and 3) were extracted from each genealogy. The distributions of *AMI* for Jomon-derived variants (type 1) and other variant types (types 2 and 3) are shown in Fig. 2 (a). *AMI* is determined by first calculating the linkage

disequilibrium coefficient r^2 between Japanese specific variant pairs, and then counting the number of variants with $r^2 > 0.01$ for a focal Japanese specific variant. It was found that Jomon-derived variants (type 1) had larger *AMI* values than the other Japanese specific variants (types 2 and 3). Fig. 2 (a) indicates that although variant pairs with linkage disequilibrium coefficient $r^2 > 0.01$ can occur by chance, the number of variant pairs with $r^2 > 0.01$ is much larger in Jomon-derived variants (type 1) than in other Japanese specific variants (type 2 and 3). The receiver operating characteristic (ROC) analysis showed that Jomon-derived variants (type 1) could be distinguished from the other Japanese specific variants (types 2 and 3) by the *AMI* (area under the curve [AUC] = 0.91; Fig. 2 (b). AUC takes a value from 0 to 1, and the closer the value is to 1, the better to distinguish positive from negative.). The Youden index, a measure of the cutoff value, was 28.0374. We performed further simulations, varying the split time between the Jomon people and continental East Asians or the effective population size in simulation, to confirm robustness of *AMI* to different population history. Though the value of the Youden index varied depending on the population history assumed, the Jomon-derived variants could be accurately detected (Supplementary Fig. 5). According to Fig. 2, we set the threshold for detecting Jomon-derived variants at *AMI* > 28.0374.

We also attempted to detect the Jomon-derived genomic components by S*[17,18], a reference-free method previously proposed, and found that S* was unable to detect the Jomon-derived genomic components, perhaps due to an

159 insufficient number of Jomon-derived specific variants of mainland Japanese
160 (Supplementary Fig. 6 and Supplementary Note 2).
161
162 **Detection of Jomon variants in real data**
163
164 Using the data set of 87 KPGP Koreans[21] and 26 global populations of
165 1KG[22], approximately 1.7 million SNPs were found to be specific to mainland
166 Japanese (1KG JPT). Of these 1.7 million SNPs, 208,648 SNPs exceeding the
167 threshold of *AMI* were regarded as Jomon-derived SNPs. Jomon-derived SNPs
168 were distributed throughout the autosomal genome (Supplementary Fig. 7).
169 Hereafter, at the Jomon-derived SNPs, an allele only found in the mainland
170 Japanese population is called a “Jomon allele.”
171
172 To examine the detection accuracy of Jomon-derived SNPs, we calculated
173 the *JAS* for the Ikawazu[9,11]/Funadomari[10] Jomon individuals and mainland
174 Japanese. If Jomon-derived SNPs were properly detected by the *AMI*, the *JAS*
175 of the Ikawazu or Funadomari Jomon were expected to be higher than those of
176 mainland Japanese. Of the JPT mainland Japanese, NA18976 was genetically
177 close to continental East Asians in PCA (Supplementary Fig. 8) and was
178 expected to have a lower *JAS*. The distribution of the *JAS* is shown in
179 Supplementary Fig. 9. The mean *JAS* of 103 mainland Japanese individuals,
180 excluding NA18976, was 0.0164. As expected, NA18976 had the lowest *JAS*,
181 0.00269, which was much lower than that of the other mainland Japanese. The
182 *JAS* in the Ikawazu/Funadomari Jomon were 0.0523 and 0.0555, respectively,

183 indicating that the Jomon alleles were found more frequently in Jomon people
184 than in the modern mainland Japanese. These results suggest that the *AMI*
185 could detect SNPs derived from the Jomon people. It should also be noted that
186 the *JAS* values were only a few % for both Jomon individuals, which suggests
187 that the number of Jomon-specific variants obtained from *AMI* analyses of
188 modern Japanese were several tens of times greater than that obtained from the
189 whole genome sequence of a single Jomon individual.
190
191 **Detection of regional genetic differences in mainland Japanese by**
192 **Jomon-derived SNPs**
193
194 ***JAS* by region and prefecture**
195 Previous prefecture-scale population studies showed that the Tohoku,
196 Kanto, and Kyushu populations (Fig. 1) are genetically more closely related to
197 the Ryukyuans, while the Kinki and Shikoku populations are more closely related
198 to continental East Asians[23,24]. Based on these facts, we hypothesized that the
199 genetic regional differences among the modern mainland Japanese are caused
200 by regional geographical differences in the admixture proportion of the Jomon
201 and immigrants from continental East Asia. To verify this, we calculated the
202 average *JAS* for each geographic region and prefecture from imputed genotypes
203 of 3,917 Jomon-derived SNPs of 10,842 Japanese individuals previously used
204 for regional population genetic analysis[24]. We removed the Hokkaido samples,
205 which were largely affected by the immigration of Japanese after the Meiji period,
206 for subsequent analysis and a total of 10,412 samples were used. The samples

of each prefecture except for Hokkaido were divided into ten regions: Tohoku, Kanto, Hokuriku, Chubu, Tokai, Kinki, Chugoku, Shikoku, Kyushu, and Okinawa in accordance with a previous study[25] (Fig. 1 and Supplementary Table 1). The *JAS*s in these ten geographical regions are presented in Fig. 3 (a) and Supplementary Table 2. We found that the *JAS* was the highest in Okinawa (0.0255), followed by Tohoku (0.0189) and Kanto (0.018), and the lowest in Kinki (0.0163), followed by Shikoku (0.016). In prefecture scale, the average *JAS* in mainland Japan tended to be higher in prefectures located in the northernmost and southernmost parts of mainland Japan (Fig. 3 (b) and Supplementary Table 3). The *JAS* was especially high in Aomori (0.0192), Iwate (0.0195), Fukushima (0.0187), and Akita (0.0186) prefectures of the Tohoku region, as well as Kagoshima Prefecture (0.0186) in Kyushu. Japanese individuals in these prefectures are considered to possess more Jomon-derived genomic components than those in other prefectures. Prefectures with lower *JAS*s were in the Kinki and Shikoku regions, including Wakayama (0.0157), Nara (0.0156), Kochi (0.016), Tokushima (0.0161), and Mie (0.0161). These populations are considered to have more genomic components derived from continental East Asians. The *JAS* of each prefecture and the principal component 1 (PC1) value, which was obtained from the principal component analysis (PCA) of a previous study by the allele frequency of autosomal 183,708 SNPs in each prefecture[24] are plotted in Fig. 3 (c). The *JAS* was strongly correlated with PC1 (R = 0.91, two-sided t-test $P = 2.2 \times 10^{-16}$). The geographic distribution was not changed by tighter cutoff values (AMI > 100) for the detection of Jomon-derived SNPs by *AMI* (Supplementary Fig. 10 and Supplementary Note 3).

To confirm the results from *JAS*, *f3* statistic[26] was calculated for a single Jomon individual[10,11] (Supplementary Fig. 11 and Supplementary Table 4). The distribution of *f3*(each prefecture; Jomon individual, CHB) (Supplementary Fig. 11) was similar to the distribution of the *JAS*s shown in Fig. 3(b). The *f3* was generally small in prefectures with high *JAS*s, such as Kagoshima Prefecture in Kyushu and Aomori and Iwate Prefectures in Tohoku, and was large in the prefectures of regions with low *JAS*s, such as Kinki and Shikoku. Based on these two findings: a correlation between the *JAS* and the PC1 of Fig. 3 (c) and the concordance between the geographical distribution of *f3* and *JAS*, it is strongly suggested that the genetic regional differences of modern Japanese can be explained mainly by regional geographical differences in the admixture proportions of the Jomon people. It should be emphasized that the admixture, although to varying degrees, widely occurred throughout the Japanese archipelago. Notably, prefectures in the Tohoku region showed higher *JAS*s than those in the Kyushu region. However, *f3* values were lower in Kyushu than in Tohoku. Since the Jomon-derived SNPs were detected in 1KG JPT (Japanese living in Tokyo), the specific variants possessed by the Jomon people of the Kyushu region may not have been detected, and thus the *JAS* in Kyushu may have been underestimated. In other words, these results could reflect differences in the genetic background of Jomon people in Tohoku and Kyushu. Overall, the geographical gradient of *f3* in mainland Japan was more consistent with the *JAS* than the distances from the locations (Funadomari and Ikawazu) where Jomon samples were taken.

We assumed that the regional differences in the *JAS* were related to regional differences in population size during the Jomon period. Therefore, we examined the correlation between *JAS* and three indexes related to the Jomon population size. The *JAS* of each prefecture was significantly correlated with the number of archeological sites from the Jomon period (R = 0.69, two-sided t-test $P = 1.27\times10^{-7}$; Fig. 4 (a)). The *JAS* of each region also correlated with the population size estimated from the number of archeological sites in the Late Jomon period (R = 0.7, two-sided t-test $P = 3.6\times10^{-2}$; Fig. 4 (b)). Moreover, the *JAS* of each prefecture was strongly correlated with $\log_{10}$(number of archeological sites in the Yayoi period/number of archeological sites in the Late Jomon period) (R = -0.64, two-sided t-test $P = 2.08\times10^{-6}$; Fig. 4 (c)). It is considered that Figs. 4 (a) and (b) correspond to Jomon population size in the Jomon period and Fig. 4 (c) corresponds to the population growth rate occurring from the Late Jomon period to the Yayoi period. The correlation between *JAS* and population size in each region suggests that the smaller the population size in the Jomon period, the lower *JAS* in modern mainland Japan (i.e., the higher contribution of genomic components of immigrants from continental East Asia).

To summarize the above results, we can conclude that genetic differences among the regions of the modern Japanese population were mainly caused by differences in the admixture proportion of the Jomon people and that differences in the admixture proportion were caused by differences in the population sizes in each region during the Final Jomon period. Regarding these, previous

morphological analyses showed that, of several populations in Japan, the Hokkaido Ainu and contemporary Kinki populations had contrasting cranial morphologies, while other modern regional populations were intermediate, with the Tohoku population being relatively similar in morphology to the Hokkaido Ainu[1,27]. Archeological evidence suggests that immigrants from continental East Asia first reached northern Kyushu[2], which seems contradictory considering that the *JAS* was lower in the Kinki and Shikoku regions than in northern Kyushu. The reason for this could be that the Kinki and Shikoku regions had a smaller population size during the Final Jomon period (Fig. 4), and thus, the proportion of genomic components derived from the immigrants became larger than in the other regions. In this study, we could clearly evaluate the similarity of local populations to the Jomon people using the Jomon-derived SNPs, and could clarify the main cause of genetic differences among the regional populations in mainland Japan.

PCA of prefectures according to Jomon allele frequency

A PCA was conducted for 46 Japanese prefectures using Jomon allele frequencies (Supplementary Fig. 12 (a)). This PCA demonstrated that Okinawa Prefecture was separated from the other prefectures by PC1. Next, 45 prefectures in mainland Japan (Okinawa was excluded from further analyses based on Supplementary Fig. 12 (a)) were analyzed in PCA (Supplementary Fig. 12 (b)). The PC1 showed that prefectures in the Tohoku and Kanto regions, where higher *JAS*s were observed among Japanese prefectures, were greatly

differentiated from prefectures in Kinki and Shikoku, where lower *JAS*s were observed (Fig. 3). The *JAS* was strongly correlated with the PC1 of Supplementary Fig. 12 (b) (R = -0.94, $P < 2.2\times10^{-16}$; Supplementary Fig. 13), which shows that PC1 reflects the ancestry proportion of Jomon people in each prefecture. In contrast, the PC2 of Supplementary Fig. 12 (b) was strongly correlated with both the latitude and longitude of each prefecture (latitude: R = 0.78, two-sided t-test $P = 2.30\times10^{-10}$, longitude: R = 0.66, two-sided t-test $P = 9.31\times10^{-7}$; Supplementary Fig. 14 (a) and (b)). These results indicate that the PC2 is determined by the geographical location of each prefecture, which might reflect that the genetic background of the Jomon people may differ according to the geographical locations in the Japanese archipelago. Previous studies have shown regional differences in the skeletal morphology of the Jomon people in the Japanese archipelago[28–32]. For example, Kondo et al. suggested that Jomon craniofacial morphology, especially in the neurocranium, exhibit a northeast-to-southwest geographical cline across the Japanese archipelago[32]. To the best of our knowledge, including studies with ancient Jomon genomes, this is the first genome-wide study to refer to the genetic regional differences among the Jomon people in the Japanese archipelago.

Genetic relationships between the Jomon people and other East Asians

To clarify the genetic relationship between the Jomon people and other East Asian populations, we estimated allele frequencies of genome-wide SNPs in the Jomon people based on the haplotype structures composed of Jomon-derived

SNPs in the modern Japanese (Supplementary Fig. 15, Methods in detail). The *f3*(Onge; Estimated Jomon frequencies, X) was calculated using Onge as the outgroup of East Asians in line with a previous study[11]. The *f*3 values for various East Asians used as X are shown in Fig. 5. The maximum value of *f3* was obtained when the Ikawazu Jomon was used as X (f3=0.0656), suggesting that our estimation reflects the allele frequencies of the Jomon people. In a previous study, a particularly strong genetic affinity was found between the Ikawazu Jomon individual and Taiwan aborigines, suggesting that the ancestral population of the Jomon people migrated through the coastal areas of East Asia (=coastal route). However, in this study, the *f3* values for Taiwan aborigines (Ami and Atayal) were lower than Tujia and Miao. In other words, our study failed to replicate a particularly strong genetic affinity between the Taiwan aborigines and the Jomon people. For comparison, we calculated *f3*(Onge; Ikawazu Jomon or Funadomari Jomon, X) using the Ikawazu or Funadomari Jomon individual whose genome had been sequenced in previous studies (Supplementary Fig 16). In the case of Ikawazu Jomon, a strong genetic affinity was observed with Ami and Atayal, while in the Funadomari Jomon, the affinity was similar to or lower than that of Dai, Tujia and Miao. These results do not strongly support the coastal route obtained in the previous study of the single ikawazu Jomon genome, which seems not precisely reflect the population history of the Jomon people due to a sampling bias. Although higher *f3* values were observed for Ami and Atayal among East Asians in the previous study of the Funadomari Jomon individual[10], the previous *f3* analysis differs from the present study in that Mbuti (an African population) was used as the outgroup. The results of *f3* analysis

would vary depending on the outgroup setting. However, it is common between our estimated allele frequency of the Jomon people, Ikawazu and Funadomari that *f3* was lower when Southeast Asian populations (Burmese, Thai, and Cambodian) and Tibetan populations (Tibetan, Sherpa, and Chokhopani (an iron age individual (3.0–2.4 kya)) were used as X (Fig. 5 and Supplementary Fig. 16). The previous study showed that the ancestral population of the Jomon people migrated to East Eurasia through the southern side of the Himalayas (=southern route)[11]. A previous study of modern and ancient East Asians including the seven Jomon individuals suggest that the Jomon people have two ancestry components from ancient East Asians, i.e. the Interior South ancestry and the Coastal ancestry[33]. Considering the results of this study and previous studies, at least the ancestral population of Jomon people seem to have migrated to the Japanese archipelago from the south of East Eurasia (southern parts of the Himalayas) via the southern of China, although it cannot be concluded whether they passed through coastal or inland areas.

Haplotype structures composed of Jomon-derived SNPs in genes associated with characteristic phenotypes of East Asians

To estimate the phenotype frequencies in the Jomon people, we investigated the haplotype structures of four genes (*ABCC11*, *EDAR*, *ALDH2*, and *ADH1B*), each having a nonsynonymous SNP associated with characteristic phenotypes of East Asians[34–39]. The derived alleles of these four nonsynonymous SNPs are associated with the following phenotypes: *ABCC11*

375 rs17822931: dry ear wax[34], *EDAR* rs3827760: thicker hair[35] and shovel-shaped
376 incisors[36], and *ALDH2* rs671 and *ADH1B* rs1229984: lower alcohol tolerance[37–
377 39]. Haplotype structures composed of Jomon-derived SNPs are shown in Fig. 6.
378 The haplotypes in each region could be classified into four types according to
379 the presence or absence of the derived allele associated with the phenotype,
380 and the composition of Jomon-derived alleles. The frequency of each haplotype
381 in the Japanese population is presented in Table 1. Here, the haplotypes
382 containing the Jomon-derived SNPs are called “Jomon-derived haplotypes.” For
383 *ABCC11* and *EDAR*, Jomon-derived haplotypes were observed for both
384 ancestral and derived alleles of the phenotype-associated SNPs. In *ABCC11*
385 (Fig. 6 (a)), the frequencies of the Jomon-derived haplotypes in mainland
386 Japanese were 10.6% for the ancestral allele (wet ear wax) and 24% for the
387 derived allele (dry ear wax). In *EDAR* (Fig. 6(b)), the Jomon-derived haplotype
388 frequencies were 14.9% for the ancestral allele (thinner hair and
389 non-shovel-shaped incisors) and 17.3% for the derived allele (thicker hair and
390 shovel-shaped incisors) in mainland Japanese. The haplotypes containing
391 ancestral and derived alleles of the phenotype-associated SNPs had different
392 Jomon alleles. Thus, it is unlikely that these haplotypes were generated by
393 recombination in the Japanese population after the admixture between the
394 Jomon and immigrants from continental East Asia. The present results suggest
395 that modern Japanese have derived alleles from both the Jomon people and
396 immigrants from continental East Asia in *EDAR* and *ABCC11*. Previous studies
397 examining ancient DNA of the Hokkaido Jomon population obtained from
398 archeological sites showed that the derived allele (dry ear wax) of *ABCC11* was

present in the Hokkaido Jomon population at a frequency of 47.6%[40,41]. As for *EDAR*, although the frequency of the derived allele (shovel-shaped incisors and thicker hair) in the Jomon people has not been estimated, it has been shown that shovel-shaped incisors were found at a frequency of 68.9% in the Jomon people[42]. The results of these previous studies are consistent with our results. In *ALDH2* rs671 (Fig. 6 (c)), the Jomon-derived haplotypes containing the derived allele of phenotype-associated SNPs were found to be rare (2.4%) in modern Japanese, and the number of Jomon alleles per Jomon-derived haplotype containing the derived allele was very small. This suggests that the Jomon people had few derived alleles (lower alcohol tolerance) of rs671, and most of the derived alleles found in modern Japanese originated from continental East Asians. In *ADH1B* rs1229984 (Fig. 6 (d)), the total frequency of the Jomon-derived haplotypes was relatively lower than that of the other three genes. The frequencies of the Jomon-derived haplotype were 3.8% for the ancestral allele (higher alcohol tolerance) and 4.8% for the derived allele (lower alcohol tolerance). In addition, when we calculated the number of Jomon-derived SNPs per 1 Mb at the genome-wide scale (Supplementary Fig. 17), we found that the number of Jomon-derived SNPs was especially small in the region around *ADH1B* (red dashed line). Therefore, regarding *ADH1B*, it is possible that the Jomon people possessed the derived allele of rs1229984 at a higher frequency compared to that of *ALDH2,* but both the Jomon-derived haplotypes with ancestral and derived alleles may have been lost after the admixture in Japanese. Koganebuchi et al., [43] previously estimated that most of the derived alleles in *ALDH2* originated from immigrants from continental East

Asia, which agrees with our results, while they concluded that the genetic contribution of immigrants was small for *ADH1B*, which contradicts the results of this study. Their study[43] assumed that, among mainland Japanese, the population in northern Kyushu had a relatively large genetic contribution from immigrants, but this assumption is inconsistent with the *JAS* estimated in the present study (Fig. 3 (b)). Thus, it is more plausible that *ADH1B* haplotypes of mainland Japanese were introduced mainly by immigrants from continental East Asia, regardless of the allelic status (ancestral or derived) of rs1229984.

Discussion

In this study, we developed the *AMI* as a summary statistic to detect the Jomon-derived variants in modern Japanese without requiring any genomic sequences from the former. The computer simulation showed that *AMI* can detect ancestral variants with high accuracy, even in an admixed population whose source populations diverged tens of thousands of years ago. Since we were able to detect Jomon-derived SNPs by the *AMI* even changing the population history in the simulations, the present approach using the *AMI* is likely to be applied to other admixed populations which source population whose source populations diverged relatively recently. Potential applications include the population of Madagascar[44,45] and the current Polynesian population[46,47], which were formed around hundreds to thousands years ago by population admixture. As exemplified by these cases, the genetic diversity of modern humans has been greatly influenced by population admixture events[48,49]. The *AMI* will be a powerful tool for clarifying the population history of not only the Japanese but also other admixed populations. It should be noted that the threshold of the *AMI* was determined by the Youden index calculated based on coalescent simulations in this study, but one may set the threshold according to one's own research purpose; if one wants to reduce false positives (i.e., variants derived from ancestral admixture), one can set the threshold strictly; if one wants to reduce false negatives, one can just set the threshold loosely. Practically, the *AMI* threshold does not necessarily have to be set based on simulations that assume a population history. In this study, our main aim is to extract

455 Jomon-derived variants from the whole-genome to determine the prevalence of
456 each prefecture, so the threshold was set to pick up as many Jomon-derived
457 SNPs as possible in order to grasp the trend of the entire genome in each
458 Japanese prefectural population.
459
460 As for the process of population formation in the Japanese archipelago from
461 the Late Jomon period to the present, we propose a model, which is shown in
462 Fig. 7. From the Late to Final Jomon period, the Jomon people settled down in
463 mainland Japan, and the population size or the population density of the Jomon
464 people varied among regions. According to Koyama 1979[25], based on the
465 number of archeological sites, it was estimated that the population sizes in the
466 Tohoku and Kanto regions were relatively large at 43,800 and 52,100,
467 respectively, while those in the Kinki and Shikoku regions were relatively small at
468 4,400 and 2,700, respectively. Thus, in the Kinki and Shikoku regions, modern
469 Japanese have lower degrees of genomic components derived from the Jomon
470 people. In the Final Jomon period, continental East Asians arrived in northern
471 Kyushu and started to admix with the Jomon people in all regions of mainland
472 Japan. During the Yayoi period, the population size of immigrants was relatively
473 increased in the Kinki and Shikoku regions, where the populations were small at
474 the end of the Jomon period. Further analyses of ancient human DNA from the
475 Final Jomon period to the Yayoi period will allow the verification of the Japanese
476 population history model proposed in this study.
477

We proposed a method to estimate allele frequencies in ancestral populations by classifying haplotypes of the current population according to their origin using SNPs derived from the ancestral population as markers. Even for admixed populations for which ancient DNA analysis cannot be performed, the same approach as in this study will help to infer the population history reflecting the genetic information of a large part of individuals (namely, allele frequency) of the ancestral population and to clarify their phenotype frequencies from the estimated genotype frequencies. For example, by calculating the allele frequencies of ancestral population of modern Ryukyuans, which seemed to have the same roots as the Jomon people in mainland Japanese, and/or of the Jomon people in each region in the mainland Japanese as this study, we can clarify genetic regional differences among the ancient Jomon populations. Ryukyu islands have an extremely hot and humid environment, and the amount of DNA from human remains is small, making ancient DNA analysis difficult[50]. This method allows us to obtain genetic information equivalent to multiple ancient individuals in such hot and humid regions. The use of modern human populations from various regions of the Japanese archipelago will further clarify the migration history of the Jomon people in future studies.

Methods

Coalescent simulation by Msprime

To investigate the characteristics of the Jomon-derived autosomal genomic components of mainland Japanese, we conducted a coalescent simulation assuming the admixture of the Jomon and continental East Asians using msprime[51] (Supplementary Fig. 2). A remarkable feature of the msprime program is that it specifies the time and population where the mutation and coalescence events occurred. Our simulation code was made with reference to a previous study[52]. The split between the Jomon ancestors and continental East Asians was set to 1,200 generations ago (30,000 YBP), according to the divergence time estimated in Kanzawa-Kiriyama et al.,[10] (between 18,000 YBP and 38,000 YBP) and the beginning of the Jomon period (around 16,000 YBP)[2]. Migration from continental East Asia to mainland Japan was set between 120 and 80 generations ago, with reference to the beginning of the Yayoi period, around 2,800 years ago[4]. The total admixture proportion of the Jomon people in the modern mainland Japanese was set to 12%[8]. The effective population size was set to 5,000 for both populations. The mutation rate and recombination rate were set to 1.2×10^{-8} per bp per generation and 1.3×10^{-8} per bp per generation, respectively[53–56].

This study aimed to detect Jomon-derived variants based on LD among Japanese specific variants. There are three types of Japanese specific variants:

(type 1) Jomon-derived variants; (type 2) variants derived from continental East Asians; and (type 3) novel variants (Supplementary Fig. 4 (a) and (b)). It should be noted that Japanese specific variants generated earlier than the split time of the Jomon people and the continental East Asians were classified as Jomon-derived variants (type 1). We compared the LD status of three types of Japanese specific variants in the coalescent simulations. The origin of each haplotype of mainland Japanese can be estimated from coalescent time to the haplotypes of the Jomon or continental East Asians. That is, if a haplotype of a mainland Japanese sample coalesced with haplotypes of Jomon samples earlier than the admixture of the Jomon people and continental East Asians, the haplotype is inferred to be derived from Jomon. To extract the three types of Japanese specific variants (i.e., variants not found in samples from continental East Asians), 3,000 replicates of 1 Mb simulations were performed. We sampled 200 haplotypes from each of the four populations (modern mainland Japanese, modern continental East Asians, Jomon people 120 generations ago, and continental East Asians 120 generations ago) to detect variants observed in modern mainland Japanese but not seen in continental East Asians. Each Japanese specific variant was classified into (type 1) the Jomon-derived variant, (type 2) the continental East Asian-derived variant, and (type 3) the novel variant based on when and in which lineage the mutation occurred (Supplementary Fig. 4 (a)). To calculate the ancestry marker index (*AMI*), we first calculate the linkage disequilibrium coefficient r^2 between Japanese specific variant pairs within each 1Mb bin. For each type of the Japanese specific variant, the *AMI* was calculated as:

$$AMI = \frac{\{Number\ of\ variants\ with\ linkage\ disequilibrium\ coefficients\ (r^2)\ > 0.01\}}{(Number\ of\ Japanese\ specific\ variants\ per\ KB)}$$

Jomon-derived variants are expected to have higher *AMI* values. The performance of the *AMI* was verified by receiver operating characteristic (ROC) analysis using the ROCR package in R. The threshold to detect Jomon-derived variants was determined based on the Youden Index.

Detection and verification of Jomon-derived SNPs on autosomes using real data

Detection of Jomon-derived variants in real data

Jomon-derived SNPs were inferred from the whole genome sequence data from 26 populations from different parts of the world, including mainland Japanese (JPT) and four continental East Asian populations (CHB, CHS, CDX, and KHV), obtained from the 1000 Genomes Project Phase III (1KG)[22], and 87 individuals from the Korean Personal Genome Project[21]. In this study, only biallelic SNPs were used. Prior to extracting the Jomon-derived SNPs, we performed a principal component analysis (PCA) in PLINK (version 1.9)[57] using 1KG mainland Japanese (JPT) and Han Chinese (CHB) data. During this analysis, we found that one JPT individual (NA18976) was close to the continental East Asians (Supplementary Fig. 8), so NA18976 was excluded from subsequent analyses. First, 1,784,634 SNPs specific to 1KG JPT were detected using VCFtools v0.1.13[58]. Next, LD coefficients (r^2) were calculated between the Japanese specific SNPs located within 1 Mb from each other with the --hap-r2

option of VCFtools in combination with the --ld-window-bp option. The number of SNPs with $r^2 > 0.01$ was counted for each Japanese specific SNP. The density of Japanese specific variants per 1 kb of each chromosome was calculated using the --SNPdensity option of VCFtools, and the *AMI* was calculated for each Japanese specific SNP. To eliminate the possibility of sequence errors, regions with a density of Japanese specific variants per kb below a mean of - 1sd of each chromosome were excluded from the analysis. In this analysis, we assumed that the number of Japanese specific variants per kb, which is the denominator of the *AMI*, is constant for each chromosome (i.e., the numerator of the *AMI* was normalized for each chromosome). Based on the threshold set by the ROC analysis of simulated Japanese specific variants, we inferred variants originating from the Jomon people.

Verification of Jomon-derived SNPs based on whole-genome sequence data from Jomon remains

For the verification of Jomon-derived SNPs based on the whole genome sequence data, the "Jomon allele score" (*JAS*) was calculated for the Ikawazu[9,11] and Funadomari[10] Jomon, as well as for 104 individuals from the 1KG JPT. The *JAS* was calculated using the following formula:

$$JAS = \frac{(Jomon\ derived\ allele\ count)}{2*(\text{Number of total Jomon}-\text{derived SNPs})}.$$

The BAM file of the Ikawazu Jomon was provided by Hiroki Ota of Tokyo University, Tokyo, Japan, and Takashi Gakuhari of Kanazawa University, Ishikawa, Japan. The BAM file of the Funadomari Jomon was provided by Naruya Saito from the National Institute of Genetics, Shizuoka, Japan, and

Hideaki Kanzawa-Kiriyama from the National Museum of Nature and Science, Tokyo, Japan. The genotypes of Ikawazu Jomon and Funadomari Jomon samples were called by the UnifiedGenotyper tool in the GenomeAnalysisToolkit version 3.6[59]. For the Ikawazu Jomon, the --mbq 30 --ploidy 2 --output_mode EMIT_ALL_CONFIDENT_SITES options were specified. For the Funadomari Jomon, the options described in the original paper were specified. Jomon SNPs were subjected to LD pruning by the --indep-pairwise command of PLINK (--indep-pairwise 1000 200 0.8). In addition, only the Jomon-derived SNPs of depth ≥ 6 in the Ikawazu and Funadomari Jomon were used for the calculation of the *JAS*. As a result, 4,458 SNPs were used to calculate *JAS*.

Detection of regional genetic differences in mainland Japanese by Jomon-derived SNPs

Sample data

We used 183,708 SNPs from 10,842 individuals from the Japanese archipelago published by Watanabe et al.[24]. All the individuals investigated in this study were customers of the Japanese Direct to Consumer (DTC) genetic-testing service, HealthData Lab (Yahoo! Japan Corporation, Tokyo, Japan). They were provided an agreement, and informed consent was obtained for their data to be used for research. In this study, the Japanese archipelago was divided into eleven regions (Fig. 1 and Supplementary Table 1): Hokkaido (430 individuals), Tohoku (746 individuals), Kanto (3,990 individuals), Hokuriku (431 individuals), Chubu (410 individuals), Tokai (933 individuals), Kinki (1,861

614 individuals), Chugoku (600 individuals), Shikoku (314 individuals), Kyushu
615 (1,016 individuals), and Okinawa (111 individuals). All statistical analyses were
616 conducted at the Yahoo! Japan Corporation, with personal information of the
617 customers completely hidden. We obtained approval from the Ethics Committee
618 of the Yahoo! Japan Corporation.
619
620 **Imputation of genotypes of Jomon-derived SNPs**
621 Haplotype phasing and genotype imputation were performed using
622 EAGLE2[60] and Minimac3[61], respectively, with whole genome sequence data of
623 413 mainland Japanese[62] phased by SHAPEIT2[63]. After the imputation,
624 Jomon-derived SNPs with high imputation quality ($R^2 > 0.8$) were extracted. Also,
625 LD pruning was performed with PLINK (--indep-pairwise 1000 200 0.1), and a
626 total of 3,917 Jomon-derived SNPs were used for the analysis.
627
628 **Geographical distribution of the Jomon allele score**
629 In subsequent analyses, individuals from Hokkaido that were largely
630 affected by immigration after the Meiji period were excluded. Using 3,917
631 Jomon-derived SNPs, we calculated the *JAS* for individuals of each prefecture
632 and compared them between regions and prefectures.
633
634 ***f3*-testing of prefectural populations in Japan**
635 The *f3*-test[26] was carried out in order to examine the relatedness between
636 contemporary populations of each prefecture in Japan and the Funadomari
637 Jomon or Ikawazu Jomon. Each Jomon sample and the 1KG CHB were set as

638 the source populations of admixture, and each prefecture of Japan was set as
639 the target population, (described as *f3*(each prefecture; Jomon, CHB)). LD
640 pruning was carried out on whole genome SNPs common in the Japanese, 1KG
641 CHB, Funadomari Jomon, and Ikawazu Jomon by PLINK (--indep-pairwise 1000
642 200 0.1), with 17,492 SNPs being used for subsequent analyses. For the
643 Funadomari and Ikawazu Jomon people, VCF files were converted to PED files
644 using VCFtools and combined with the PED file of the Japanese. The PED files
645 were converted to eigenstrat format with the Admixtools[26] convertf command,
646 and then the *f3*-test was conducted with the Admixtools qp3Pop command.
647
648 **Examination of correlations between population size during the Jomon**
649 **period and *JAS* in each prefecture**
650 We compared the population size estimated from the number of
651 archeological sites in each prefecture, assuming that the population size per
652 archeological site was constant in each prefecture during the same period. We
653 examined the correlations between (a) the average *JAS* in each prefecture and
654 the number of archeological sites from the Jomon period (obtained from the
655 Statistical report of buried cultural properties, Agency of Cultural Affairs, Japan;
656 https://www.bunka.go.jp/seisaku/bunkazai/shokai/pdf/h29_03_maizotokei.pdf),
657 (b) the average *JAS* in each region and the population size estimated from the
658 number of archeological sites in the Late Jomon period[25], and (c) the average
659 *JAS* in each prefecture and the $\log_{10}$(number of archeological sites in the Yayoi
660 period/number of archeological sites in the Late Jomon period)[25]. Finally, (a) and
661 (c) were plotted for each prefecture, while (b) was plotted for each region

because data for each prefecture were not available. Correlation test was conducted by R cor.test function (df = 43).

PCA based on the Jomon allele frequencies of each prefecture

The Jomon allele frequency was calculated for 50 randomly sampled individuals from each prefecture using VCFtools version 0.1.13. A PCA was performed based on the Jomon allele frequency using R version 3.6.0. Correlation test between PC1 and JAS, and PC2 and longitude/latitude were conducted by R cor.test function (df = 43).

Estimation of the allele frequencies of genome-wide SNPs of the Jomon people based on the haplotype structure composed of the Jomon-derived SNPs

We calculated the allele frequencies of genome-wide SNPs in Jomon people based on the haplotype structure composed of the Jomon-derived SNPs (Supplementary Fig. 15). First, haplotypes surrounding a focal bi-allelic SNP with alleles 1 and 2 are classified into Jomon-derived haplotypes and Continental-East-Asian-derived haplotype, based on the presence or absence of Jomon alleles in the 500 kb upstream and downstream (1Mb in total) of the focal SNP. The frequency of the allele 1 of the focal SNP in Jomon people is expressed as follows.

$$Frequency\ of\ the\ allele\ 1\ in\ the\ Jomon\ people = \frac{Number\ of\ Jomon\ derived\ haplotypes\ containing\ the\ allele\ 1\ of\ the\ focal\ SNP}{Total\ number\ of\ Jomon\ derived\ haplotypes\ surrounding\ the\ focal\ SNP}$$

In the example of Supplementary Fig. 15, the frequency of the allele 1 in the Jomon people is 1/3. We calculated the allele frequency in the Jomon people

based on the phased-genotypes of 104 modern Japanese in 1KG for 5,316,769 SNPs in the whole genome. The --IMPUTE option of VCFtools was used to extract haplotypes composed of Jomon-derived SNPs in regions 500 kb upstream and downstream (1 Mb in total) of focal SNPs. Based on the estimated frequencies of genome-wide SNPs, we estimated the genetic relationship between the Jomon people and other East Asians. The population genotype data set (Panel 2240K) used in Gakuhari et al 2020[11] was kindly provided by Takashi Gakuhari of Kanazawa University, Ishikawa, Japan, and 81 individuals including the Ikawazu Jomon[9,11] of 38 East Asian populations were extracted. Nivkh and Ulchi were excluded from this analysis because it has been suggested that they were recently admixed with Jomon lineage population (likely Ainu, who have a genetic background of the Jomon people)[10,11]. Among the genome-wide SNPs for which we estimated the Jomon allele frequencies, we used 20,053 SNPs with genotype information was available in all East Asian individuals of 2240K. We then calculated *f3*(Onge ; Estimated Jomon frequencies, X) for testing the genetic relationship between the Jomon people and each test population X, using Onge as the outgroup of East Asians[11]. In addition, we focused on the haplotype structures composed of Jomon-derived SNPs in the regions surrounding SNPs associated with characteristic phenotypes of East Asians (*ABCC11*: rs17822931[34], *EDAR*: rs3827760[35,36], *ALDH2*: rs671[37,43,64–66], and *ADH1B*: rs1229984[38,67–70]), and estimated the frequencies of these phenotypes in the Jomon people.

Statistics and reproducibility.

Coalescent simulations of this study using msprime can be reproduced by specifying seeds in our code described in supplementary note 4. Statistical analyses were done using publicly available packages, so reproducibility can be accomplished using parameters described in Methods sections.

Data and code availability

The individual genotypes of 10,842 Japanese analyzed in this study are not available to avoid personal identification. The list of Jomon-derived SNPs detected in this study, and the allele frequencies of Jomon-derived SNPs in each Japanese prefecture are available from the corresponding author upon request. Our custom code for msprime simulation was described in Supplementary Note 4.

724 **References**

725

726 1. Hanihara, K. Dual structure model for the population history of the
727 Japanese. *Japan Rev.* **2**, 1–33 (1991).
728 2. Habu, J. *Ancient Jomon of Japan*. (Cambridge University Press, 2004).
729 3. Mizoguchi, K. *The archaeology of Japan : from the earliest rice farming*
730 *villages to the rise of the state*. (Cambrige University Press, 2013).
731 4. Fujio, S. *History of Yayoi Period*. (Kodansha, 2015).
732 5. Jinam, T. *et al.* The history of human populations in the Japanese
733 Archipelago inferred from genome-wide SNP data with a special reference
734 to the Ainu and the Ryukyuan populations. *J. Hum. Genet.* **57**, 787–795
735 (2012).
736 6. Jinam, T. A. *et al.* Unique characteristics of the Ainu population in Northern
737 Japan. *J. Hum. Genet.* **60**, 565–571 (2015).
738 7. Nakagome, S. *et al.* Model-based verification of hypotheses on the origin
739 of modern Japanese revisited by Bayesian inference based on
740 genome-wide SNP data. *Mol. Biol. Evol.* **32**, 1533–1543 (2015).
741 8. Kanzawa-Kiriyama, H. *et al.* A partial nuclear genome of the Jomons who
742 lived 3000 years ago in Fukushima, Japan. *J. Hum. Genet.* **62**, 213–221
743 (2017).
744 9. Mccoll, H. *et al.* The prehistoric peopling of Southeast Asia. *Science (80-.).*
745 **361**, 88–92 (2018).

746 10. Kanzawa-Kiriyama, H. *et al.* Late Jomon male and female genome
747 sequences from the Funadomari site in Hokkaido , Japan. *Anthropol. Sci.*
748 **127**, 83–108 (2019).
749 11. Gakuhari, T. *et al.* Ancient Jomon genome sequence analysis sheds light
750 on migration patterns of early East Asian. *Commun. Biol.* **3,** 437, (2020).
751 https://doi.org/10.1038/s42003-020-01162-2
752 12. Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct
753 ancestry in admixed populations. *PLoS Genet.* **5**, e1000519 (2009).
754 13. Baran, Y. *et al.* Fast and accurate inference of local ancestry in Latino
755 populations. *Bioinformatics* **28**, 1359–1367 (2012).
756 14. Churchhouse, C. & Marchini, J. Multiway Admixture Deconvolution Using
757 Phased or Unphased Ancestral Panels. *Genet. Epidemiol.* **37**, 1–12
758 (2013).
759 15. Guan, Y. Detecting structure of haplotypes and local ancestry. *Genetics*
760 **196**, 625–642 (2014).
761 16. Hui, D. *et al.* LAIT: A local ancestry inference toolkit. *BMC Genet.* **18**, 1–5
762 (2017).
763 17. Plagnol, V. & Wall, J. D. Possible ancestral structure in human populations.
764 *PLoS Genet.* **2**, e105 (2006).
765 18. Vernot, B. & Akey, J. M. Resurrecting Surviving Neandeltal Linages from
766 Modern Human Genomes. *Science (80-.).* **343**, 1017–1021 (2014).
767 19. Vernot, B. *et al.* Excavating Neandertal and Denisovan DNA from the
768 genomes of Melanesian individuals. *Science (80-.).* **352**, 235–239 (2016).

769 20. Jacobs, G. S. *et al.* Multiple Deeply Divergent Denisovan Ancestries in
770 Papuans. *Cell* **177**, 1010-1021.e1-e24 (2019).
771 21. Zhang, W. *et al.* Whole genome sequencing of 35 individuals provides
772 insights into the genetic architecture of Korean population. *BMC*
773 *Bioinformatics* **15**, 1–13 (2014).
774 22. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**,
775 68–74 (2015).
776 23. Yamaguchi-Kabata, Y. *et al.* Japanese Population Structure, Based on
777 SNP Genotypes from 7003 Individuals Compared to Other Ethnic Groups:
778 Effects on Population-Based Association Studies. *Am. J. Hum. Genet.* **83**,
779 445–456 (2008).
780 24. Watanabe, Y., Isshiki, M. & Ohashi, J. Prefecture-level population
781 structure of the Japanese based on SNP genotypes of 11,069 individuals.
782 *J. Hum. Genet.* **In press**, (2020).
783 25. Koyama, S. Jomon Subsistence and Population. *Senri Ethnol. Stud.* **2**, 1–
784 65 (1979).
785 26. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**,
786 1065–1093 (2012).
787 27. Hanihara, K. Geographic variation of modern Japanese crania and its
788 relationship to the origin of Japanese. *Homo* **36**, 1–10 (1985).
789 28. Kaifu, Y. Regional variation in mandibular morphology of the Jomon
790 people. *Bull Natl Sci Mus* **21**, 37–50 (1995).

29. Maeda, T. Mandibular ramus morphology of the Jomon people from Hokkaido. Anthropological. *Anthropol. Sci. (Japanese Ser.* **110**, 27–40 (2002).

30. Takigawa, W. Inter-regional variation of metric traits of limb bones in Jomon and modern Japanese. Anthropological. *Anthropol. Sci. (Japanese Ser.* **114**, 101–129 (2006).

31. Fukase, H. *et al.* Geographic variation in body form of prehistoric Jomon males in the Japanese archipelago: Its ecogeographic implication. *Am. J. Phys. Anthropol.* **149**, 125–135 (2012).

32. Kondo, O., Fukase, H. & Fukumoto, T. Regional variations in the Jomon population revisited on craniofacial morphology. *Anthropol. Sci.* **125**, 85–100 (2017).

33. Wang, C.-C. *et al.* Genomic Insights into the Formation of Human Populations in East Asia. *Nature* (2021). doi:https://doi.org/10.1038/s41586-021-03336-2

34. Yoshiura, K. I. *et al.* A SNP in the ABCC11 gene is the determinant of human earwax type. *Nat. Genet.* **38**, 324–330 (2006).

35. Fujimoto, A. *et al.* A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum. Mol. Genet.* **17**, 835–843 (2008).

36. Kimura, R. *et al.* A Common Variation in EDAR Is a Genetic Determinant of Shovel-Shaped Incisors. *Am. J. Hum. Genet.* **85**, 528–535 (2009).

813 37. Harada, S., Agarwal, D. P. & Goedde, H. W. Aldehyde Dehydrogenase
814 Deficiency As Cause of Facial Flushing Reaction To Alcohol in Japanese.
815 *Lancet* **318**, 982 (1981).
816 38. Edenberg H.J. & Bosron W.F. *Alcohol dehydrogenases. In:*
817 *Comprehensive Toxicology, Vol. 3: Biotransformation*. (Pergamon Elsevier
818 Science, 1997).
819 39. Edenberg, H. J. & McClintick, J. N. Alcohol dehydrogenases, aldehyde
820 dehydrogenases and alcohol use disorders: a critical review. *Alcohol Clin*
821 *Exp Res* **42**, 2281–2297 (2018).
822 40. Sato, T. *et al.* Allele frequencies of the ABCC11 gene for earwax
823 phenotypes among ancient populations of Hokkaido, Japan. *J. Hum.*
824 *Genet.* **54**, 409–413 (2009).
825 41. Kazuta, H. *et al.* Genotype frequencies of the ABCC11 gene in
826 2000-3000-year-old human bones from the Epi-Jomon and Jomon sites in
827 Hokkaido, Japan. *Anthropol. Sci.* **119**, 81–86 (2011).
828 42. Matsumura, H. A Microevolutional History of the Japanese People from a
829 Dental Characteristics Perspective. *Anthropol. Sci.* **102**, 93–118 (1994).
830 43. Koganebuchi, K. *et al.* The allele frequency of ALDH2*Glu504Lys and
831 ADH1B*Arg47His for the Ryukyu islanders and their history of expansion
832 among East Asians. *Am. J. Hum. Biol.* **29**, 1–14 (2017).
833 44. Pierron, D. *et al.* Genomic landscape of human diversity across
834 Madagascar. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E6498–E6506 (2017).

45. Pierron, D. *et al.* Strong selection during the last millennium for African ancestry in the admixed population of Madagascar. *Nat. Commun.* **9**, 1–9 (2018).

46. Skoglund, P. *et al.* Genomic insights into the peopling of the Southwest Pacific. *Nature* **538**, 510–513 (2016).

47. Pugach, I. *et al.* The gateway from near into remote oceania: New insights from genome-wide data. *Mol. Biol. Evol.* **35**, 871–886 (2018).

48. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).

49. Nielsen, R. *et al.* Tracing the peopling of the world through genomics. *Nature* **541**, 302–310 (2017).

50. Hofreiter, M. *et al.* The future of ancient DNA: Technical advances and conceptual shifts. *BioEssays* **37**, 284–293 (2015).

51. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Comput. Biol.* **12**, 1–22 (2016).

52. Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S. & Akey, J. M. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell* **173**, 1–9 (2018).

53. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome International Human Genome. *Nature* **409**, 860–921 (2001).

54. Altshuler, D. L. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

859 55. Scally, A. & Durbin, R. Revising the human mutation rate: Implications for
860 understand ing human evolution. *Nat. Rev. Genet.* **13**, 745–753 (2012).
861 56. Kong, A. *et al.* Rate of de novo mutations and the importance of father-s
862 age to disease risk. *Nature* **488**, 471–475 (2012).
863 57. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and
864 population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575
865 (2007).
866 58. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**,
867 2156–2158 (2011).
868 59. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework
869 for analyzing next-generation DNA sequencing data. *Genome Res.* **20**,
870 1297–1303 (2010).
871 60. Loh, P. R. *et al.* Reference-based phasing using the Haplotype Reference
872 Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
873 61. Das, S. *et al.* Next-generation genotype imputation service and methods.
874 *Nat. Genet.* **48**, 1284–1287 (2016).
875 62. Watanabe, Y. *et al.* Analysis of whole Y-chromosome sequences reveals
876 the Japanese population history in the Jomon period. *Sci. Rep.* **9**, 8556
877 (2019).
878 63. Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome
879 phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6
880 (2013).

64. Oota, H. *et al.* The evolution and population genetics of the ALDH2 locus: Random genetic drift, selection, and low levels of recombination. *Ann. Hum. Genet.* **68**, 93–109 (2004).

65. Li, H. *et al.* Refined geographic distribution of the oriental ALDH2* 504Lys (nee 487Lys) variant. *Ann. Hum. Genet.* **73**, 335–345 (2009).

66. Luo, H. R. *et al.* Origin and dispersal of atypical aldehyde dehydrogenase ALDH2*487Lys. *Gene* **435**, 96–103 (2009).

67. Osier, M. *et al.* Linkage disequilibrium at the ADH2 and ADH3 loci and risk of alcoholism. *Am. J. Hum. Genet.* **64**, 1147–1157 (1999).

68. Osier, M. V. *et al.* A global perspective on genetic variation at the ADH genes reveals unusual patterns of linkage disequilibrium and diversity. *Am. J. Hum. Genet.* **71**, 84–99 (2002).

69. Han, Y. *et al.* Evidence of positive selection on a class I ADH locus. *Am. J. Hum. Genet.* **80**, 441–456 (2007).

70. Li, H. *et al.* Geographically separate increases in the frequency of the derived ADH1B*47His allele in eastern and western Asia. *Am. J. Hum. Genet.* **81**, 842–846 (2007).

Acknowledgments

We are grateful to the individuals who participated in the study. We would like to express our deepest gratitude to Mr. Masahiro Inoue, Shota Arichi, and Akito Tabira who obtained the genotype data and provided the technical environment for analyzing them. We would like to thank Dr. Hiroki Ota of Tokyo University, Tokyo, Japan, and Dr. Takashi Gakuhari of Kanazawa University, Ishikawa, Japan, for providing us the BAM file of the IKawazu Jomon. Dr. Takashi Gakuhari also gave us the population genotype data set (Panel 2240K) used in Gakuhari et al 2020. We also thank Dr. Naruya Saito from the National Institute of Genetics, Shizuoka, Japan, and Dr. Hideaki Kanzawa-Kiriyama from the National Museum of Nature and Science, Tokyo, Japan for providing us the BAM file of the Funadomari Jomon. This study was partly supported by Grant-in-Aid for Scientific Research (B) (18H02514) and Grant-in-Aid for Scientific Research on Innovative Areas (19H05341) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

Author contributions

Y.W. and J.O. conceived the study. Y.W. designed and conducted the data analyses. Y.W. performed the computer simulations. Y.W. wrote the manuscript with support from J.O. J.O. supervised the project. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

925 **Fig. legends**

926 **Fig. 1 Map of the Japanese prefectures.** The prefectures of Japan are divided
927 into eleven regions. The prefecture numbers in Supplementary Table 1 are
928 indicated (the corresponding prefecture names are given in Supplementary
929 Table 1). In this study, “mainland Japanese” means the Japanese people except
930 for individuals from Hokkaido and Okinawa.

931

932 **Fig. 2 The performance of the *AMI* for the detection of the Jomon-derived**
933 **SNPs**

934 (a) Distribution of *AMI* simulated by msprime. The histogram of *AMI* for the
935 Jomon-derived variants (type 1) and the other variants (types 2 and 3) are
936 shown. The red dashed line indicates the threshold of *AMI* (28.0374) obtained
937 from ROC analysis for the detection of the Jomon-derived variants (type 1).
938 (b) ROC curve illustrating the performance of the *AMI* for the detection of the
939 Jomon-derived SNPs. The ROC curve was drawn based on the simulated data
940 shown in Fig. 2 (a). The *AMI* showed high accuracy (AUC = 0.91) for
941 discriminating the Jomon-derived variants (type 1) from the other variants (types
942 2 and 3).

943

944 **Fig. 3 JAS of each Japanese region.**

945 (a) Distribution of *JAS* in ten regions. The boxplot of the *JAS* is presented for
946 each of the ten regions, excluding Hokkaido.

(b) *JAS* of each prefecture in Mainland Japan. The average *JAS* by prefecture was calculated. Hokkaido and Okinawa prefectures are not illustrated. The prefecture with the higher average *JAS* is illustrated by the darker color.
(c) Relationship between the PC1 of the PCA performed in a previous study by the allele frequency of autosomal 183,708 SNPs in each prefecture and average *JAS*. Each prefecture was colored according to the region of Japan in Fig. 1. Horizontal axe: PC1, vertical axe: average *JAS*. Pearson's correlation coefficients (*R*), *P* values, regression lines and 95% CI are shown.

Fig. 4 Relationship between the *JAS* and values associated with the population size of prefectures in the Jomon to Yayoi periods.
The horizontal axis shows the average *JAS* and the vertical axis shows the (a) number of archaeological sites of the whole Jomon period, (b) population size in the Late Jomon period, and (c) $\log_{10}$ (number of archaeological sites in the Yayoi period/number of archaeological sites in the Late Jomon period). Pearson's correlation coefficients (*R*), *P* values, regression lines and 95% CI are shown in each figure. Each prefecture is colored according to the region in Fig. 1.

Fig. 5 f3-statistics between East Asians and estimated allele frequencies of the Jomon people.
f3(Onge ; Estimated Jomon frequencies, X), where X are the East Asians including 37 present day populations and one ancient population (Chokhopani, an iron age individual). Darker plots (higher *f3* value) indicates higher genetic affinity with the Jomon people.

971

972 **Fig. 6 Haplotype structures composed of Jomon-derived SNPs in four**
973 **genes associated with the characteristic phenotypes of East Asians.**
974 The haplotype structures surrounding four nonsynonymous SNPs, (a)
975 rs17822931 in *ABCC11* (associated with ear wax type), (b) rs3827760 in *EDAR*
976 (associated with hair thickness and shovel-shaped incisors), (c) rs671 in *ALDH2*
977 (associated with alcohol tolerance), and (d) rs1229984 in *ADH1B* (associated
978 with alcohol tolerance) are illustrated. Each horizontal line represents each
979 haplotype, and each vertical line represents each of the Jomon-derived SNP or
980 phenotype-associated SNP. The derived alleles of the SNPs associated with
981 phenotypes are shown in orange, and the ancestral alleles are shown in blue.
982 The red color represents the Jomon allele, i.e., horizontal lines containing the
983 red colored grid indicate the Jomon-derived haplotypes.
984

985 **Fig. 7 The formation process of the Japanese population from the Late**
986 **Jomon period to the present.**
987 From the Late to the Final Jomon period, the Jomon people settled down in
988 Mainland Japan, and the population size varied between regions. In the Final
989 Jomon period, the continental East Asians arrived in northern Kyushu and then
990 admixed with Jomon people in all the regions of Mainland Japan. In regions such
991 as Kinki and Shikoku, where the population size was smaller at the end of the
992 Jomon period, modern Japanese have lower degrees of genome components
993 derived from the Jomon people.

Table 1 Frequencies of the Jomon-derived haplotypes specified by phenotypes associated SNPs in *ABCC11*, *EDAR*, *ALDH2*, and *ADH1B*.

(a) *ABCC11*	rs17822931		
	Ancestral	Derived	Total
Jomon-derived haplotype	0.106	0.240	0.346
Non-Jomon–derived haplotype	0.014	0.639	0.653
Total	0.120	0.879	

(b) *EDAR*	rs3827760		
	Ancestral	Derived	Total
Jomon-derived haplotype	0.149	0.173	0.322
Non-Jomon–derived haplotype	0.048	0.630	0.678
Total	0.197	0.803	

(c) *ALDH2*	rs671		
	Ancestral	Derived	Total
Jomon-derived haplotype	0.125	0.024	0.149
Non-Jomon–derived haplotype	0.635	0.216	0.851
Total	0.760	0.240	

(d) *ADH1B*	rs1229984		
	Ancestral	Derived	Total
Jomon-derived haplotype	0.038	0.048	0.086
Non-Jomon–derived haplotype	0.231	0.683	0.914
Total	0.269	0.731	

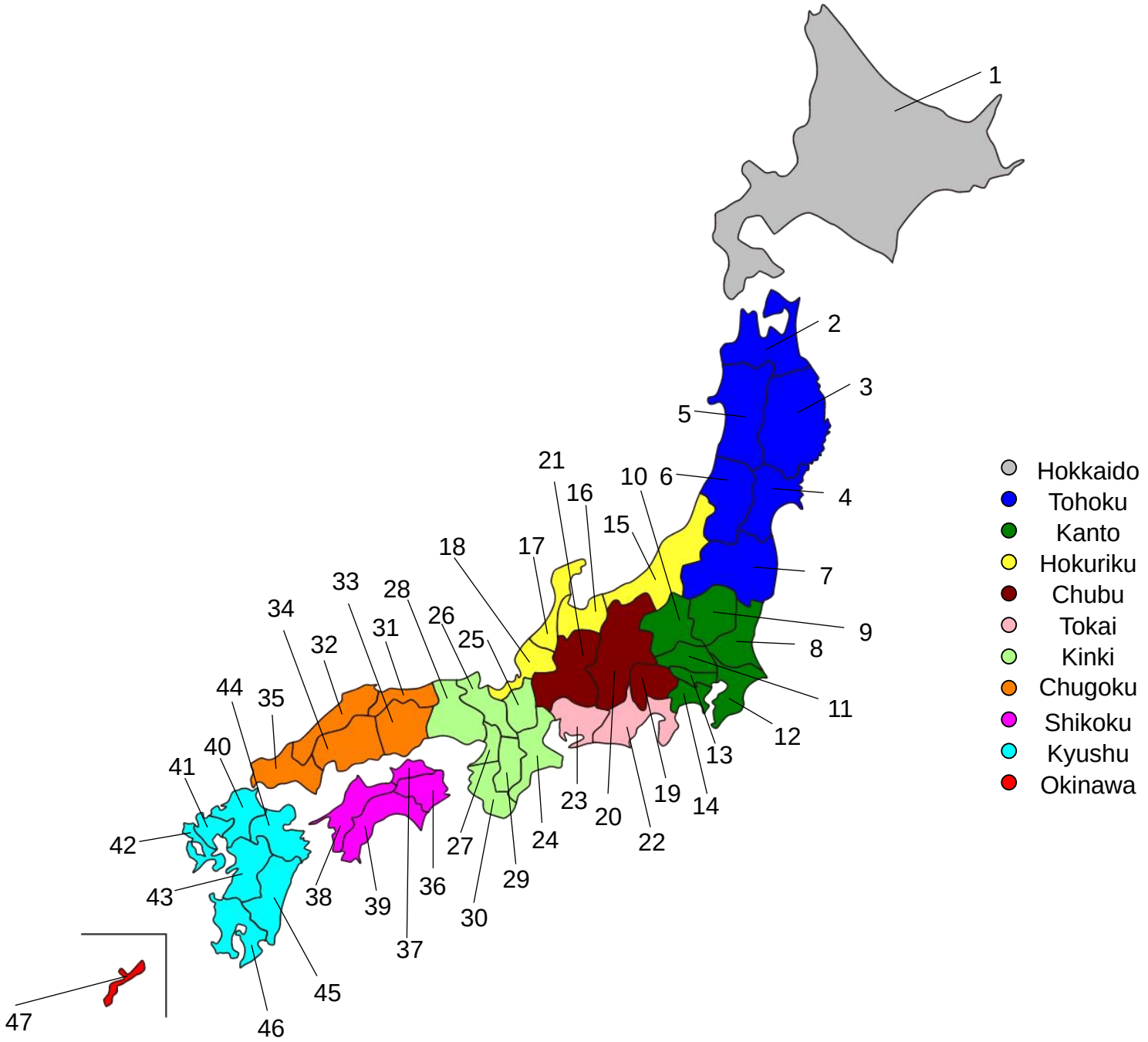
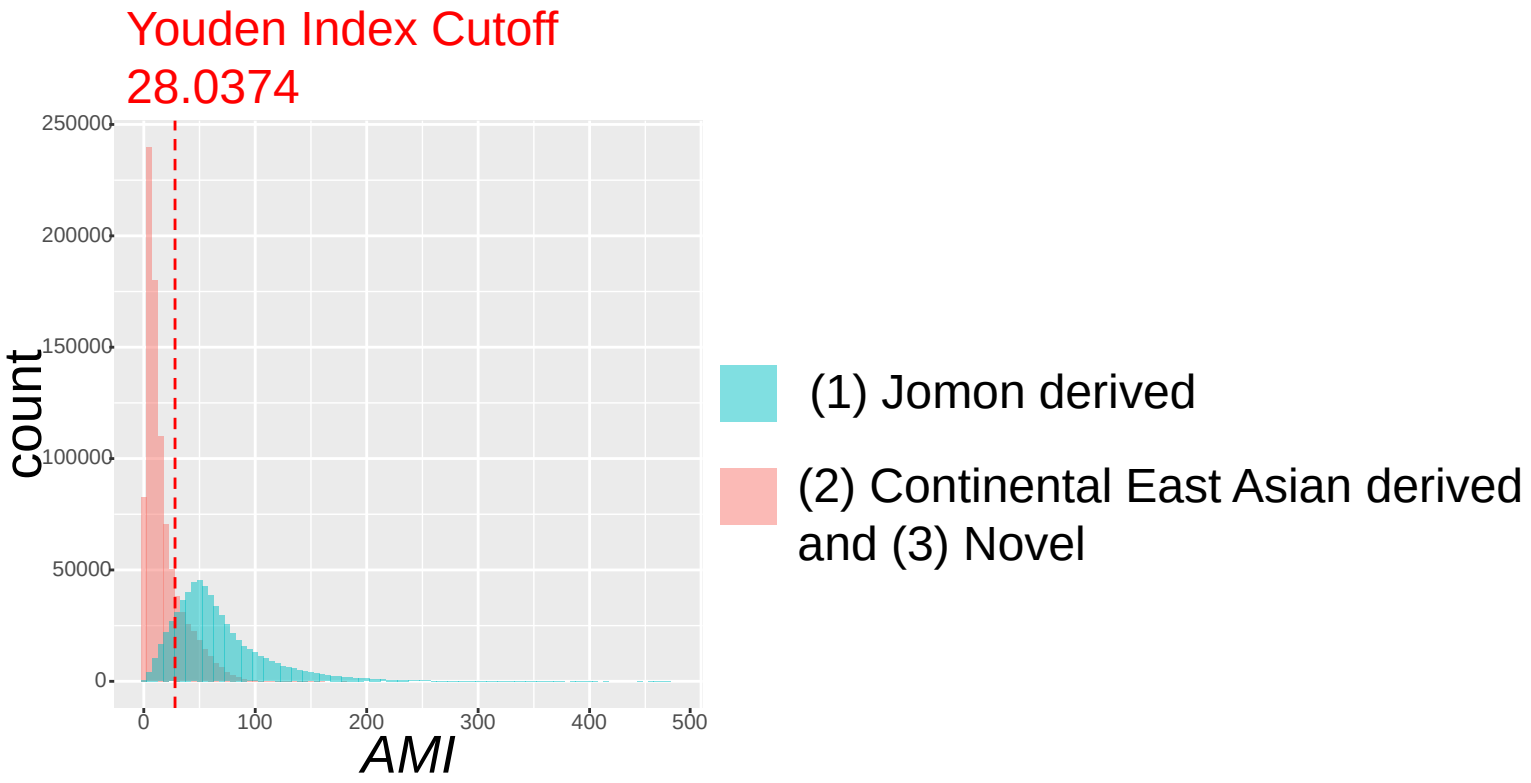

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
Hokkaido
Tohoku
Kanto
Hokuriku
Chubu
Tokai
Kinki
Chugoku
Shikoku
Kyushu
Okinawa

Fig. 1

(a)


Youden Index Cutoff
28.0374
250000
200000
150000
100000
50000
0
count
0
100
200
300
400
500
AMI
(1) Jomon derived
(2) Continental East Asian derived and (3) Novel

(b)

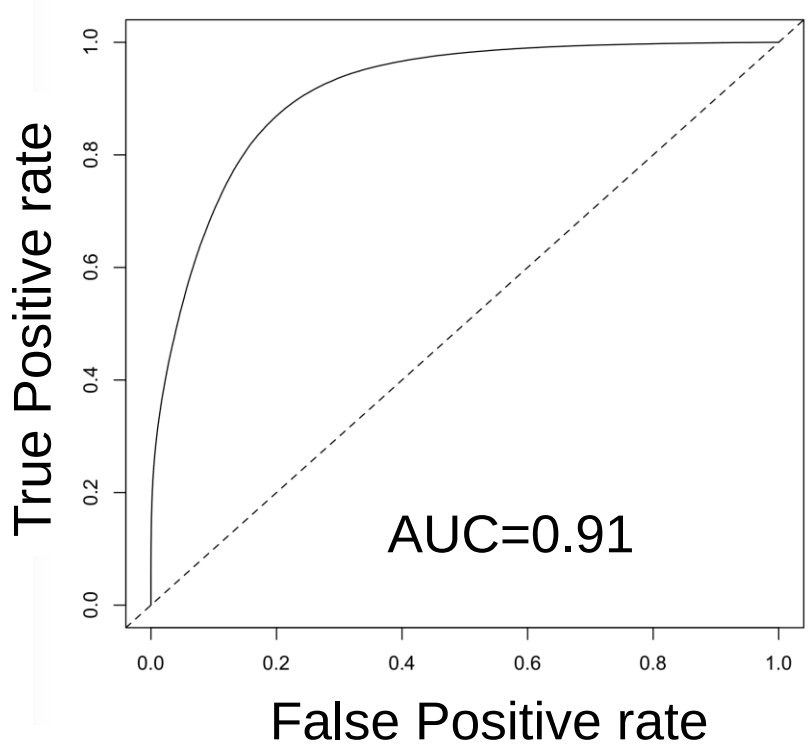
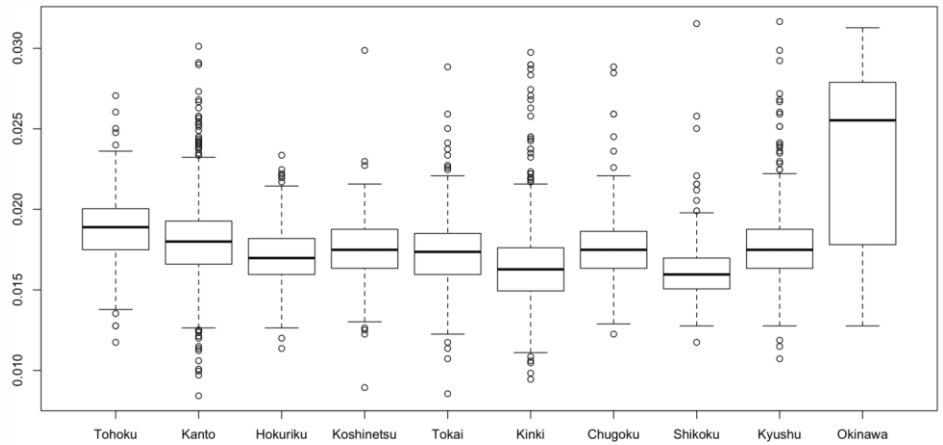

True Positive rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC=0.91
False Positive rate

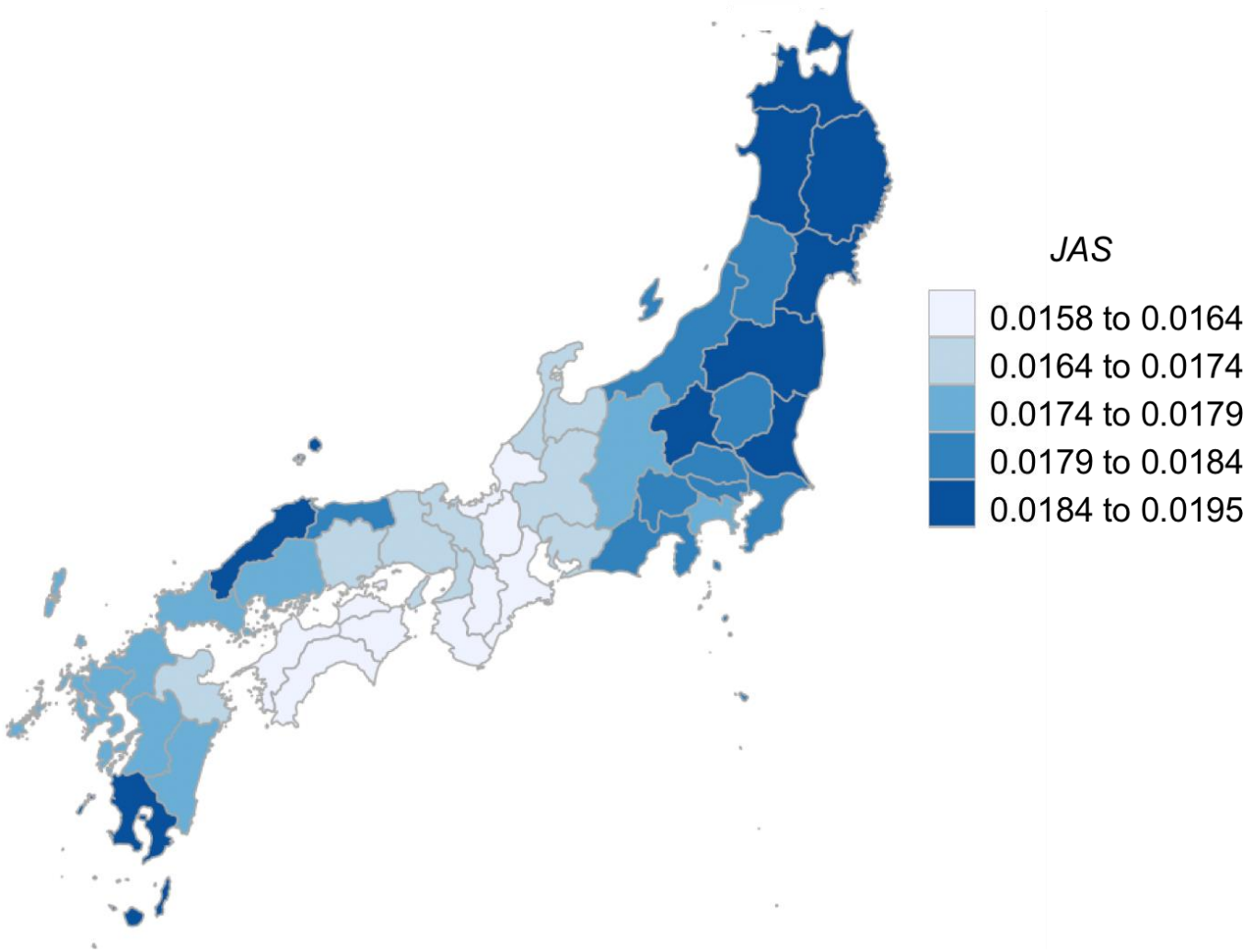
Fig. 2

(a)



	Tohoku	Kanto	Hokuriku	Koshinetsu	Tokai	Kinki	Chugoku	Shikoku	Kyushu	Okinawa
median	0.0189	0.018	0.017	0.0175	0.0174	0.0163	0.0175	0.016	0.0175	0.0255

(b)



(c)

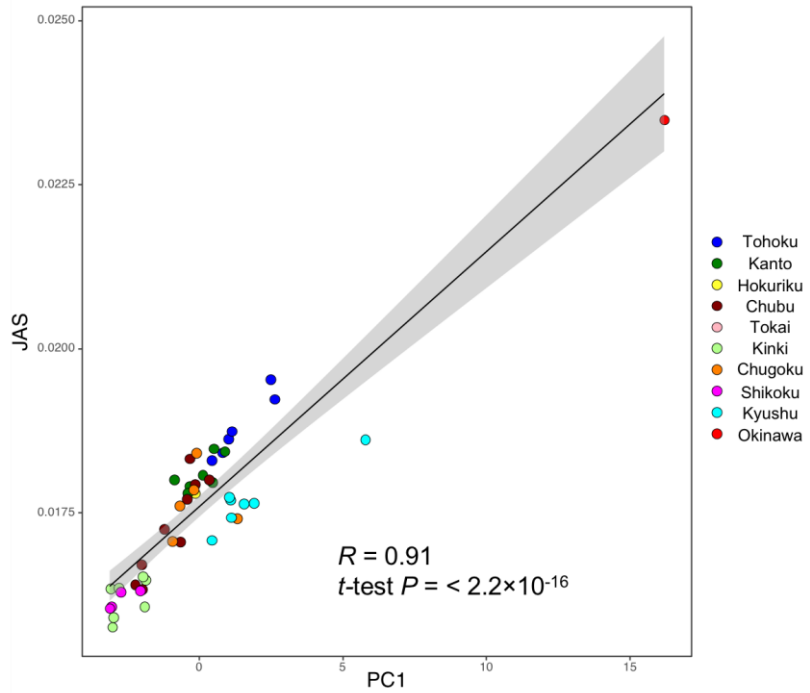


Fig. 3

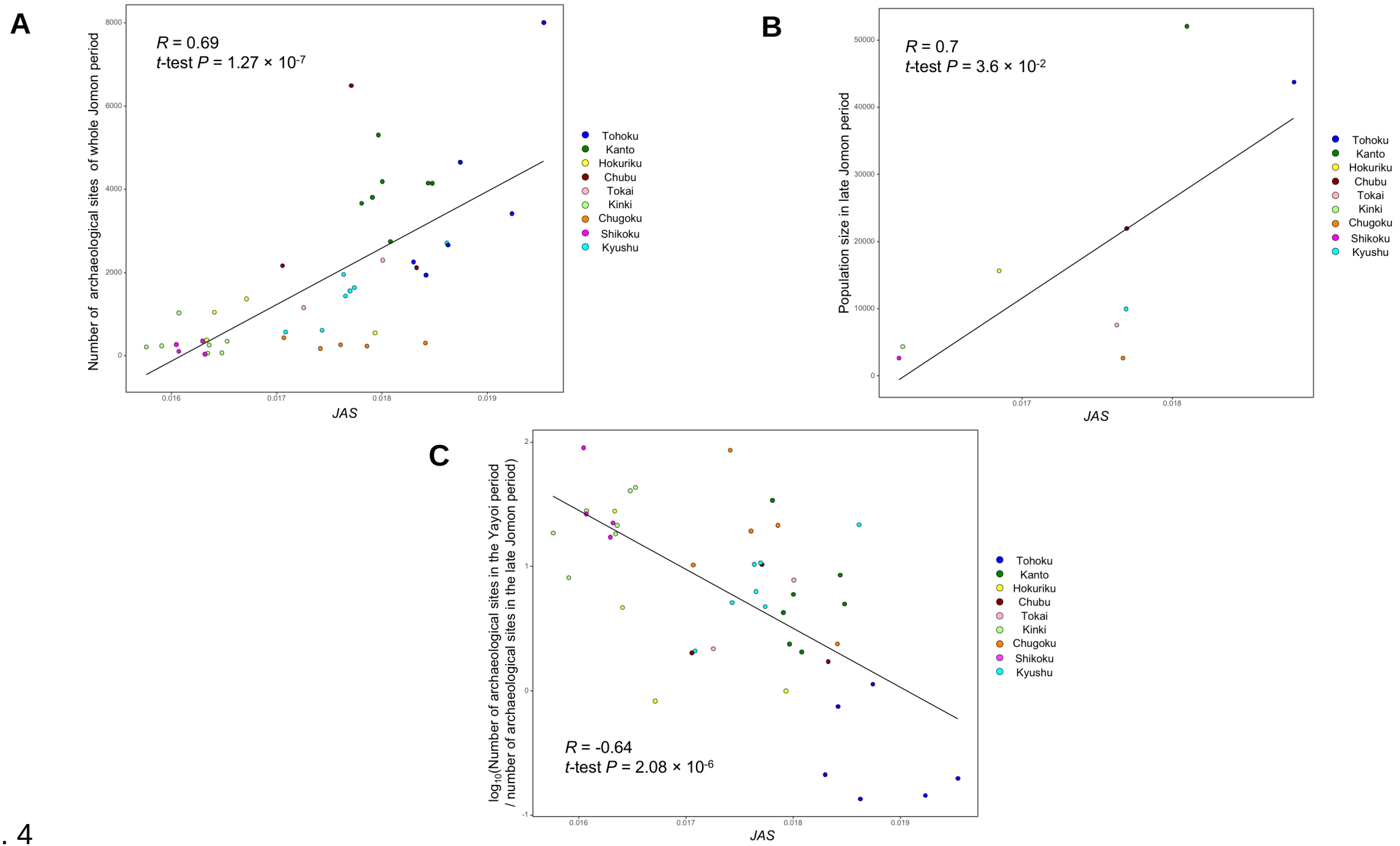

A
R = 0.69
t-test P = 1.27 × 10-7
Number of archaeological sites of whole Jomon period
JAS
B
R = 0.7
t-test P = 3.6 × 10-2
Population size in late Jomon period
JAS
C
R = -0.64
t-test P = 2.08 × 10-6
log10(Number of archaeological sites in the Yayoi period / number of archaeological sites in the late Jomon period)
JAS
Tohoku
Kanto
Hokuriku
Chubu
Tokai
Kinki
Chugoku
Shikoku
Kyushu

Fig. 4

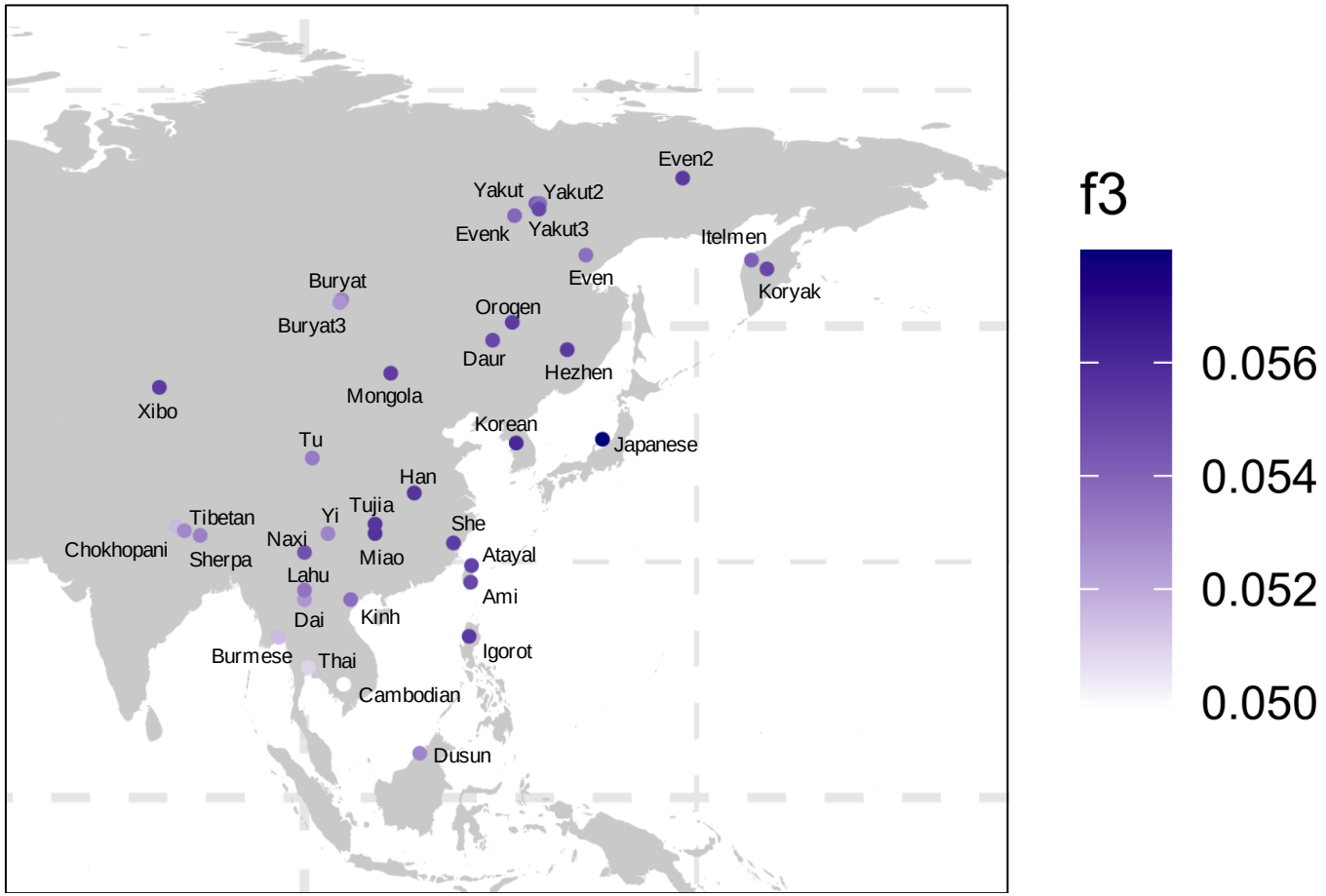

Even2
Yakut
Yakut2
Evenk
Yakut3
Itelmen
Even
Koryak
Buryat
Orogen
Buryat3
Daur
Hezhen
Mongola
Xibo
Korean
Japanese
Tu
Han
Tujia
Tibetan
Yi
She
Chokhopani
Naxi
Atayal
Sherpa
Miao
Lahu
Ami
Dai
Kinh
Igorot
Burmese
Thai
Cambodian
Dusun
f3
0.056
0.054
0.052
0.050

Fig. 5

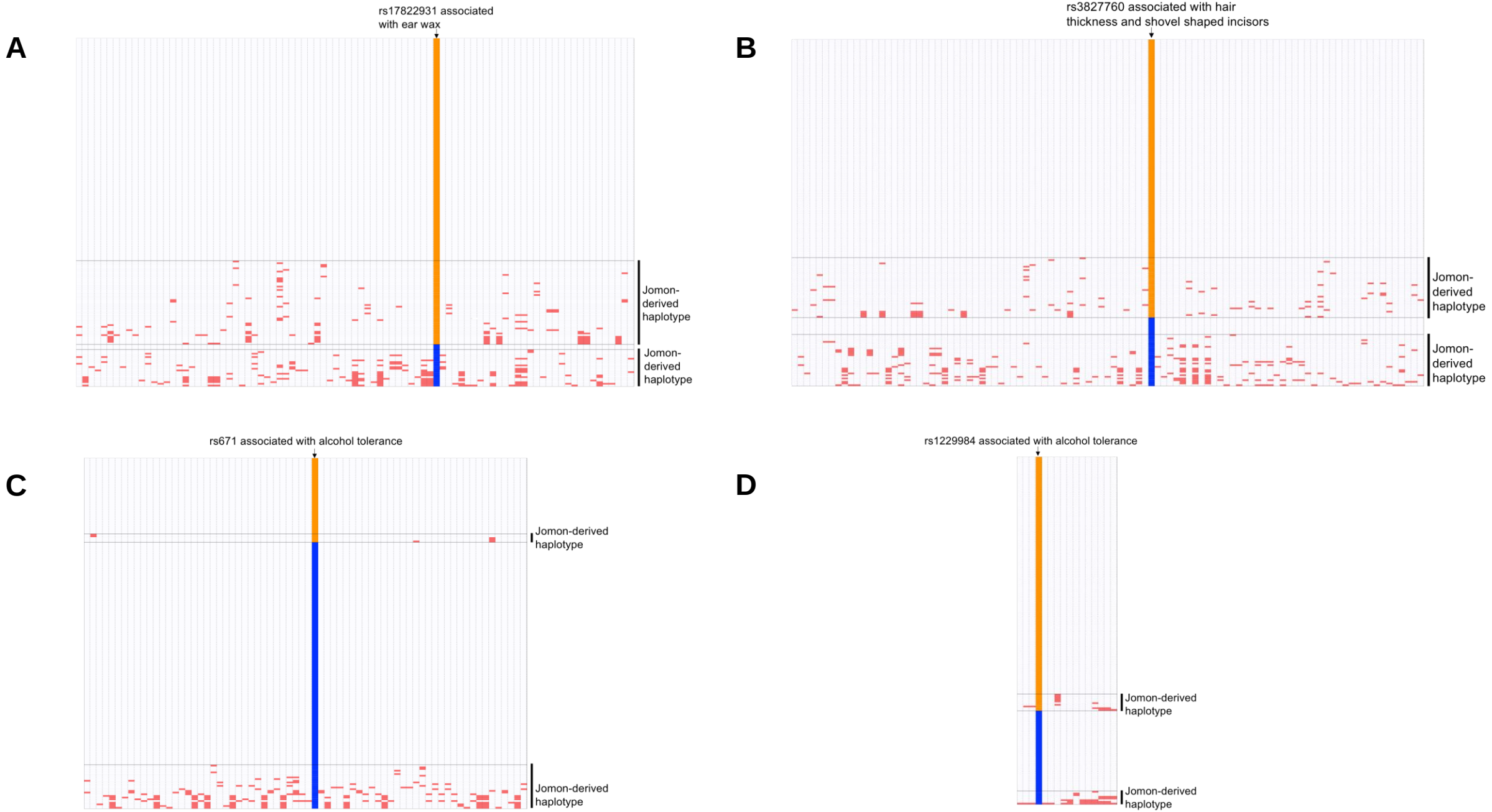

A
rs17822931 associated with ear wax
Jomon-derived haplotype
Jomon-derived haplotype
B
rs3827760 associated with hair thickness and shovel shaped incisors
Jomon-derived haplotype
Jomon-derived haplotype
C
rs671 associated with alcohol tolerance
Jomon-derived haplotype
Jomon-derived haplotype
D
rs1229984 associated with alcohol tolerance
Jomon-derived haplotype
Jomon-derived haplotype

Fig. 6

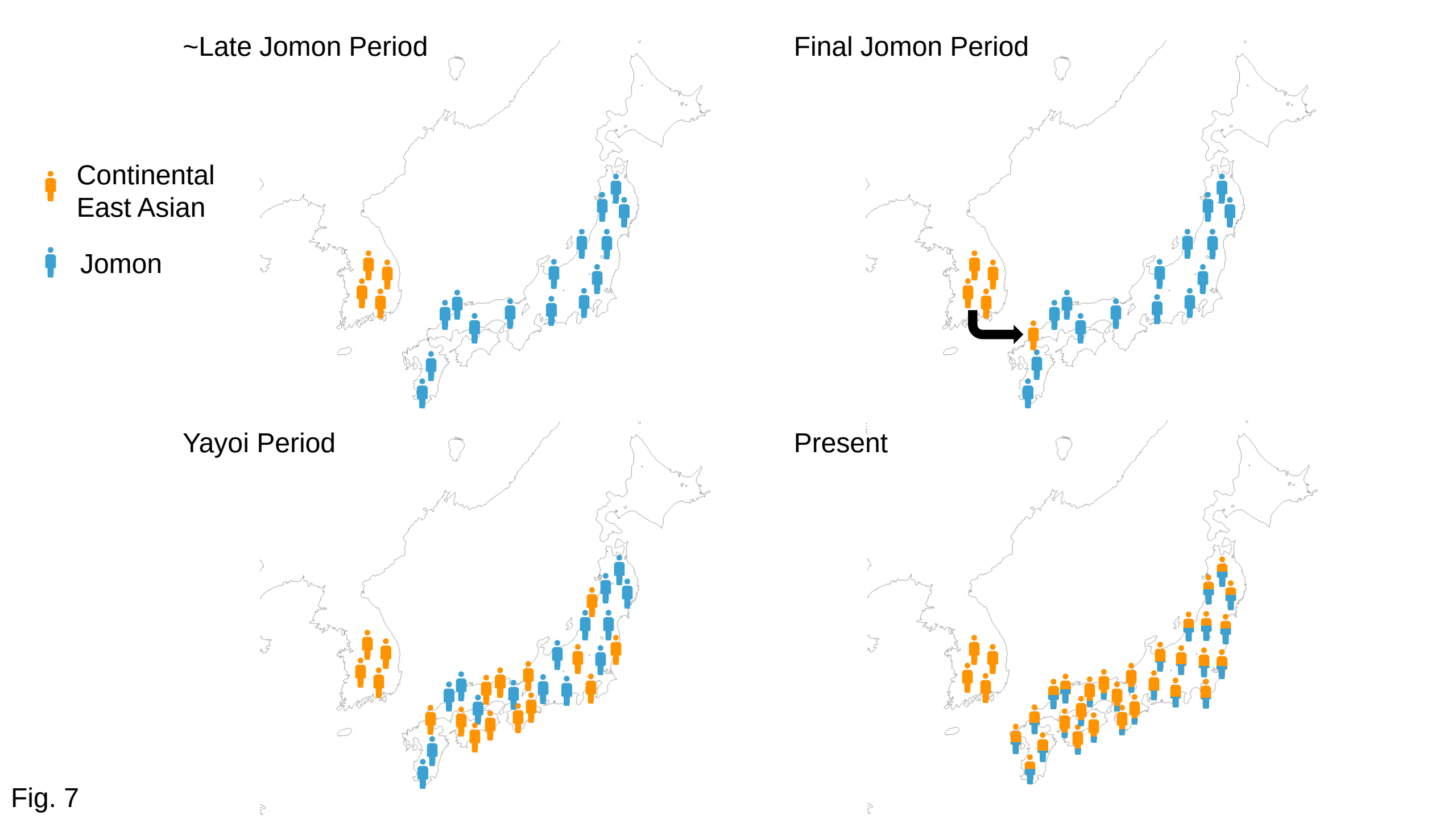

~Late Jomon Period
Continental East Asian
Jomon
Final Jomon Period
Yayoi Period
Present

Fig. 7