Detection of ancestry-derived variants of modern Japanese revealed the formation process of regional gradations of the current Japanese archipelago population

Authors

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

Affiliations

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

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

Address correspondence to Jun Ohashi

Email: jun_ohashi@bs.s.u-tokyo.ac.jp

ABSTRACT

Modern Japanese have two major ancestral populations: the indigenous Jomon people and continental immigrants. To figure out the formation process of current Japanese population, we developed a reference-free detection method of variants derived from ancestral populations using a summary statistic, the ancestry-marker index (*AMI*), and identified 208,648 SNPs that appear to have been derived from the Jomon people (Jomon-derived variants). The analysis of Jomon-derived variants in 10,842 modern Japanese individuals recruited from all over Japan revealed that the admixture proportions of the Jomon people varied between prefectures. The estimated allele frequencies of genome-

wide SNPs in the ancestral populations of modern Japanese suggested their phenotypic characteristics possibly for adaptation to their respective livelihoods. According to our findings, we propose a formation model of modern Japanese population; regional variations in admixture proportions of the Jomon people and continental East Asians formed genotypic and phenotypic gradations of current Japanese archipelago populations.

Main text

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 (Extended Data Fig. 1). As a powerful hypothesis of the formation processes of Japanese populations, a dual structure model[1] was proposed based on morphological findings. This model assumes that Japanese originated through a mixture of the Jomon people, the Neolithic hunter-gatherers 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 with their rice-farming technology from continental East Asia around the beginning of the Yayoi period (around 2,800 YBP)[4]. The rice farming of continental immigrants subsequently spread throughout Japan and brought about a major transformation in ancient Japanese society. According to the dual structure model, compared to mainland Japanese, the Ainu and the Ryukyuan were genetically less influenced by immigrants. Genetic studies not only supported the dual structure model, but also revealed the detailed population history of Japanese archipelago[5–11]. Whole-genome analyses extracted from the remains of the Jomon people showed that the Jomon people 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]. Recent studies found that, in addition to the "East Asian" population which is closely related to modern Han Chinese, the "Northeast Asian" population also contributed to the ancestry of modern Japanese[12,13]. Cooke et al 2021 showed the deep divergence of the Jomon people from continental populations including the "East Asians" and "Northeast Asians", so it can be concluded that the modern mainland Japanese are a population with genomic components derived from a basal East Asian lineage (i.e., the Jomon people) and from continental East Asians. We collectively refer the two continental ancestral populations pointed out in Cooke et al. as "the continental East Asians" in this article, unless a distinction is necessary.

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 expected to be in linkage disequilibrium (LD) with each other (Fig. S1). In this study, we developed a method using a summary statistic, the Ancestry-Marker Index (*AMI*), to detect ancestry-marker variants derived from the Jomon people in modern mainland Japanese. A key feature of the *AMI* is that it 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 originated from admixture of archaic hominins and early Eurasians[14–17]. Since the Jomon people were highly differentiated from other East Asians[8,10], they are expected to have had specific

variants that are not found in present-day East Asian populations other than the Japanese. Thus, it is likely that modern mainland Japanese also have specific variants derived from the Jomon people. The *AMI* detects the Jomon-derived variants based on the LD between Japanese specific variants. We could successfully extract the Jomon-derived variants from the real genomic data of the Japanese. We conducted comprehensive analyses using Jomon-derived variants, which enabled us to infer the phenotypic characteristics that have not appeared in skeletal morphology and ancient individual genomes from the Jomon remains, and to elucidate the mechanisms by which genetic and phenotypic regional gradations among the Japanese arose. Based on our findings, we propose a model of formation process of current regional populations in the Japanese archipelago.

Results

Development of the *AMI*, a summary statistic to detect ancestry-derived variants

First of all, we attempted to detect the Jomon-derived genomic components by S*[14,15], a reference-free method previously proposed. We found that S* was unable to detect the Jomon-derived genomic components by a coalescent simulation assuming population history of modern Japanese (Supplementary Text 1 and Figs. S2 and S3). The Jomon population has been isolated in the Japanese archipelago for a much shorter period since its divergence from other Asian populations than the divergence time of archaic Hominin and modern humans, which is the main target of S*, so there seem no haplotypes in which Jomon-specific variation is in proximity being enough to S* analysis. We therefore developed a new summary statistic, the *AMI*. The *AMI* was developed based on the concept that the Jomon people, a part of the ancestral population of modern Japanese,

were highly genetically differentiated from other continental East Asian populations, and modern Japanese have inherited specific variants accumulated in Jomon lineages which are not observed in other East Asian populations (type1; Jomon-derived variants in Fig. 1 and Fig. S4). Our first step is extracting specific variants of modern Japanese to remove variants that have emerged in the non-Jomon ancestry lineage after the divergence of the Jomon people and the non-Jomon ancestry (i.e., orange triangle in Fig. 1). In the next step, we calculate *AMI* to distinguish Jomon-derived variants from other types of Japanese specific variants. There are two types of Japanese specific variants other than (type 1) Jomon-derived variants: (type 2) variants derived from continental East Asians, which emerged in the continental East Asian lineages 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 emerged only in Japanese lineages after the admixture (Fig. 1 and Fig. S4). Of these Japanese specific variants, the Jomon-derived variants (type 1) are considered to be accumulated on the same haplotype or to be in LD with each other (Fig. S4 (B)). In other words, variants belonging to (type 1) are expected to have more variant pairs in LD compared to (type 2 and 3) (Fig. S4 (B)). To calculate the *AMI,* we first compute the LD coefficient r^2 between Japanese specific variant pairs. The *AMI* is the count of variants where r^2 exceeds the cutoff for a focal Japanese specific variant, divided by the density of specific variants per kb. Calculating the *AMI*, those in LD with more Japanese specific variants will be distinguished as Jomon-derived variants (type 1).

To confirm the usefulness of the *AMI*, we performed a coalescent simulation assuming a mixture of the Jomon people and continental East Asians (Extended Data Fig. 2). In the 1 Mb simulation, Japanese specific variants (types 1, 2, and 3) were

extracted from each genealogy. First, we calculated r^2 values between Jomon-derived variants (type 1) or between other types of Japanese specific variants (type 2 and 3 together), suggesting that Jomon-derived variants (type 1) had more variant pairs of $r^2 > 0.01$ than (types 2 and 3) (Extended Data Fig. 3). We then defined the r^2 threshold value for the *AMI* to 0.01, and calculated the *AMI* for Jomon-derived variants (type 1) and other variant types (types 2 and 3). The distributions of *AMI* showed that Jomon-derived variants (type 1) had larger *AMI* values than the other Japanese specific variants (types 2 and 3) (Fig. 2 (A)). 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)). 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 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 (Fig. S5).

Detection of Jomon variants in real data

We focused only on bi-allelic SNPs to detect Jomon-derived variants. Using the dataset of 87 KPGP Koreans[18] and 26 global populations of 1KG[19], approximately 1.7 million SNPs were found to be specific to mainland Japanese (1KG JPT). Of these 1.7 million SNPs, 208,648 SNPs exceeding the threshold of *AMI* were regarded as Jomon-derived variants. Jomon-derived variants were distributed throughout the autosomal genome (Extended Data Fig. 4).

To examine the detection accuracy of Jomon-derived variants, we calculated the Jomon allele score (*JAS*) for two Jomon individuals, Ikawazu[9,11] and Funadomari[10], and 104 mainland Japanese. If Jomon-derived variants were properly detected by the *AMI*, the *JAS* of the Ikawazu or Funadomari Jomon were expected to be higher than those of mainland Japanese. Of the JPT mainland Japanese, NA18976 was genetically close to continental East Asians in PCA (Fig. S6) and was expected to have a lower *JAS*. The distribution of the *JAS* is shown in Extended Data Fig. 5. The mean *JAS* of 103 mainland Japanese individuals, excluding NA18976, was 0.0164. As expected, NA18976 had the lowest *JAS*, 0.00269, which was much lower than that of the other mainland Japanese. The *JAS* of Ikawazu Jomon and Funadomari Jomon were 0.0523 and 0.0555, respectively, indicating that the Jomon-derived variants were found more frequently in Jomon people than in the modern mainland Japanese. These results suggest that the *AMI* could detect SNPs derived from the Jomon people. It should also be noted that the *JAS* values were only a few percent even for both Jomon individuals, which suggests that the number of Jomon-specific variants obtained from *AMI* analyses of modern Japanese were several tens of times greater than that obtained from the whole genome sequence of a single Jomon individual.

Detection of regional genetic differences in mainland Japanese by Jomon-derived variants *JAS* by region and prefecture

Previous prefecture-scale population studies showed that the Tohoku, Kanto, and Kyushu populations (Extended Data Fig. 1) are genetically more closely related to the Ryukyuans, while the Kinki and Shikoku populations are more closely related to

continental East Asians[20,21]. Based on these facts, we hypothesized that the genetic regional differences among the modern mainland Japanese are caused by regional geographical differences in the admixture proportion of the Jomon and immigrants from continental East Asia. To verify this, we calculated the average *JAS* for each geographic region and prefecture from imputed genotypes of 3,917 Jomon-derived variants of 10,842 Japanese individuals previously used for regional population genetic analysis[21]. We removed the Hokkaido samples, which were largely affected by the immigration of Japanese after the Meiji period (1,868~), 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[22] (Extended Data Fig. 1 and Table S1). The *JAS*s in these ten geographical regions are presented in Fig. 3 (A) and Table S2. 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 Table S3). 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. Interestingly, the *JAS* in Shimane Prefecture (0.0184) was on the same level as the prefectures in Tohoku and Kagoshima in Kyushu, being consistent with the genetic affinity in the Izumo individuals to Okinawa and Kyushu individuals in Jinam et al[23]. 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[21] 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 variants by *AMI* (Fig. S7 and Supplementary Text 2).

We assumed that the regional differences in the *JAS* were related to regional differences in population size during the Jomon period, so 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). In summary, we can conclude that genetic gradations among the regional population of modern Japanese were mainly caused by differences in the admixture proportion of the Jomon people, perhaps due to population size differences in each region during the Final Jomon to the Yayoi period.

Allele frequencies estimation by Jomon-derived haplotypes reveals the phenotypic characteristics of the Jomon and continental East Asian ancestry

We devised an estimation method of the allele frequencies of genome-wide SNPs in the Jomon people, the ancestors of modern Japanese, prior to their admixture with continental populations without using Jomon individual genomes. Modern Japanese haplotypes surrounding a focal SNP can be classified into "Jomon-derived haplotypes" and "continental haplotypes" according to the presence Jomon-derived variants (Extended Data Fig. 6). The allele frequency of the Jomon people in the focal SNP can be estimated by the proportion of each allele within Jomon-derived haplotypes (Extended Data Fig. 6). The proportion of each allele within continental haplotypes is the allele frequency of the continental ancestry of modern Japanese. Using 419 modern Japanese genomes (THC dataset)[24], we applied our method for 6,481,773 SNPs in the Jomon and continental ancestries of modern Japanese ("THC Jomon ancestry" and "THC continental ancestry") to estimate the allele frequencies. The allele frequencies of THC ancestries were verified by ancient genomes previously reported (11 Jomon and three Kofun individuals[9–11,13]) and genomes of THC modern Japanese and Han Chinese (CHB)[19] (Fig. 5). Kofun individuals, who were excavated from mainland Japan 1,300-1,400 years ago, have similar Jomon ancestry proportion of modern Japanese[13].

First, a pairwise *f3* assuming the Yoruba people as an outgroup was used to test whether the allele frequencies of THC Jomon ancestry resembles that of the actual Jomon individuals (Fig. 5). The *f3* value of the THC Jomon ancestry with each Jomon individual from previous studies took higher values ($f3 > 0.04$) than that of with Kofun, modern Japanese and modern Chinese. The *f3* values calculated between the THC Jomon ancestry and the Jomon individuals were comparable to that of the Jomon individual pairs (Fig. 5). These results indicate that the THC Jomon ancestry were genetically close to the actual Jomon individuals, and that we can successfully infer allele frequencies of the Jomon people by the Jomon-derived haplotypes found in modern Japanese. These results also provide strong assurance that the Jomon-derived variants of modern Japanese we extracted were indeed originated through the Jomon people.

To estimate the phenotypic characteristics of the Japanese ancestral populations, we combined previous GWAS results[25–27] with allele frequencies of genome-wide SNPs of THC modern Japanese, THC Jomon and continental ancestries. We calculated the mean $2\beta f$ value for each population allele frequencies. The $2\beta f$ is the 2 × (effect allele frequency) multiplied by (effect size) for each SNP associated with a focal trait, and the $\text{mean } 2\beta f$ over genome-wide SNPs allows us to evaluate the average phenotype in a given population for the trait. The $\text{mean } 2\beta f$ corresponds to the average polygenic score (PS)[28] within a focal population. We calculated the $\text{mean } 2\beta f$ values for 60 traits in the previous QTL GWAS (Table S4). First focusing on THC modern Japanese, the $\text{mean } 2\beta f$ values for all 60 traits in the modern Japanese were close to the continental ancestry. Deviation of the population frequency

from the ancestry proportion in an admixed population is known to be a signal of positive natural selection after the admixture event[29,30]. The mean $2\beta f$ values of THC modern Japanese reflected the high percentage of continental ancestry in the Japanese population (80~90%), and it seems unlikely that phenotypes of the Jomon ancestry excessed in the modern Japanese due to natural selection. Next, to identify traits with particularly large differences between Jomon and continental ancestries, we calculated *D* statistics based on the null distribution obtained from simulations. We can infer that the larger the absolute value of *D*, the more significant the phenotypic difference between Jomon and continental ancestries were. *D* takes a positive value when the mean $2\beta f$ is greater for THC Jomon ancestries than for THC continental ancestries. The *D* values for each trait were shown in Fig. 6 A and Table S4. Traits with significantly large absolute *D* values included height (trait ID: Height, *D* = -17.5), eosinophil count (trait ID: Eosino, *D* = -14.8), C-reactive protein (trait ID: CRP, *D* = -11.9) and blood sugar (trait ID: BS, *D* = 9.9). Among these traits, we focused on height, eosinophil count and blood sugar. Eosinophils are a variety of white blood cells and one of the immune system components that play an important role in the response to helminth infection[31]. Figs. 6 B, C and D showed the histograms of the mean $2\beta f$ obtained by simulations for height, eosinophil count and blood sugar and the mean $2\beta f$ values for THC modern Japanese, THC Jomon and continental ancestries. Figs. 6 A - D indicated that the Jomon ancestry had genetically shorter stature and higher blood sugar, while the continental ancestry had genetically taller stature and higher eosinophil counts. With regard to height, our results were very convincing because several previous morphological studies suggested that the Jomon people had shorter statue than the people after the migration from continental East Asia, such as

the people of the Yayoi and the Kofun period[32–34]. Concerning the blood sugar and the eosinophil counts, it seemed to have been genetic adaptations to their respective livelihoods; the Jomon people may have needed to maintain high blood sugar levels in their hunter-gatherer lifestyle, while continental East Asian populations may have needed to increase their resistance to helminth infections during their agricultural lifestyles. We will provide more detailed description in the Discussion section.

Isshiki et al. suggested that PS for height correlated with the mean height of each prefectural population in Japan and those genetically closely related to the Han Chinese have greater PS for height among regional populations in modern Japan[35]. Combining Isshiki et al. with our finding, we can strongly argue that the regional variation in height among the Japanese archipelago was caused by differences in the ancestry proportions of the Jomon people, who had genetic factors for shorter statue, and continental East Asians, who had genetic factors for taller statue. In addition to modern Japanese height, we would like to report a case in which regional differences in ancestry proportions affected the regional diversity in a modern Japanese phenotype in relation to eosinophil counts. Eosinophils had strongly association with allergic inflammation such as asthma in contemporary human populations[36]. According to the previous GWAS on eosinophil counts in the Japanese population, eosinophil counts had a significant genetic correlation with the asthma risk[26]. We focused on regional differences in the incidence rates of asthma exacerbation, which is related to eosinophils[37], among mainland Japanese populations. Fig. 6 E plotted mean *JAS* values on the x-axis and the incidence rate of asthma exacerbations on the y-axis for each prefecture[38]. We observed a significant inverse correlation between the mean *JAS*

values and the incidence rate of asthma exacerbations for prefectures in mainland Japan (R = -0.41, P = 0.005), indicating that the risk of asthma exacerbations was higher for regional populations in mainland Japan that has a higher ancestry proportion of continental East Asians.

Discussion

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 variants 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. The genetic diversity of modern humans has been greatly influenced by population admixture events[29,39–42]. The *AMI* will be a powerful tool for the population history of not only the Japanese but also other admixed populations. It should be noted that we determined the threshold of the *AMI* by the Youden index based on coalescent simulations in present study, but one may set the threshold according to one's own research purpose; if one wants to reduce false positives, one can set the threshold strictly; if one wants to reduce false negatives, one can set the threshold loosely. Practically, the *AMI* threshold does not necessarily have to be set based on simulations that assume a population history. In the regional comparison of modern Japanese, the threshold was set loosely to pick up as many Jomon-derived variants as possible in order to grasp the trend of the whole genome in

each Japanese prefectural population. On the other hand, in allele frequency estimation of the Jomon people by the Jomon-derived haplotypes of modern Japanese, the threshold was set strictly because false positives of Jomon-derived variants may lead to incorrect estimation of Jomon-derived haplotype frequencies.

We proposed an allele frequency estimation method in ancestral populations by classifying haplotypes of the current population based on their origin. Combining allele frequencies from THC Jomon and continental ancestries with previous GWAS results, we reported three traits for which the Japanese ancestral populations were presumed to have exhibited a characteristic phenotype: height, blood sugar, and eosinophil count. Regarding height, several morphological analyses suggested the short stature of the Jomon people[32–34], and Isshiki et al. found that among regional populations in the Japanese archipelago, those genetically closely related to the Han Chinese had greater PS for height[35]. These previous studies gave a strong assurance of our finding that the Jomon people were genetically shorter than the continental ancestry. As for blood sugar, we inferred genetically higher blood sugar for the Jomon ancestry compared to the continental ancestry. The diet of the prehistoric and modern population of the Japanese archipelago has been investigated based on human and faunal bone collagen and human scalp hair[43–46], suggesting that the Jomon hunter-gatherers had wide range dietary customs for adaptations to a variety of environments of the Japanese archipelago compared to modern Japanese. The diet of the Japanese archipelago population seemed to have changed significantly to become more dependent on agricultural products with the introduction of rice cultivation by continental East Asians after the Late Jomon period. In the Yayoi period, carbohydrate intake from crops increased, which affected the

Japanese archipelago population in the Yayoi period, including a higher incidence of dental caries[47,48]. The Jomon ancestry of modern Japanese may have had more difficulty maintaining blood sugar levels with food foraging than continental East Asian ancestry with rice farming, and thus genetic factors for higher blood sugar levels would have been helpful for them. Concerning eosinophil counts, we found the continental ancestry to have genetically higher eosinophil counts than the Jomon ancestry. Agricultural populations are generally more exposed to parasitic organisms than hunter-gatherer populations and have a higher frequency of helminth infections[49,50]. We speculate that the continental East Asian ancestry of modern Japanese, having begun rice farming prior to their migration to the Japanese archipelago, needed to increase their resistance to helminths compared to the Jomon ancestry, and acquired genetic factors for increasing their eosinophil counts. This view is supported by a previous study of the mtDNA of the blood fluke *Schistosoma japonicum*[51]. The phylogenetic analysis indicated the dispersal of *Schistosoma japonicum* radiated from the middle and lower reaches of the Yangtze River, where rice farming originated, to various parts of East Asia occurred after 10,000 years ago with the spread of rice farming culture[51]. Furthermore, Kanehara and Kanehara found no roundworm eggs in soil samples at Sannai-Maruyama site despite the detection of many whipworm eggs[52], so Matsui et al. argued that the prevalence of roundworm infection occurred after the Yayoi Period with the beginning of rice agriculture[53]. Following these findings, the next question is which species of the helminths contributed to the elevated eosinophil counts in the continental ancestry. Previous studies with murine models including eosinophil-free transgenic mice have indicated that the impact of eosinophils on helminth infection is not uniform; there were both evidence for eosinophil-mediated clearance of worms and increase in worm numbers[54]. Mitre and Klion mentioned, “This is due to a

number of factors, including the heterogeneity of the many helminths included in the designation "helminth," the complexity and redundancy in the host immune response to helminths, and the pleiotropic functions of eosinophils themselves"[54]. In future, we expect to be clarified above question by helminth eggs and helminth-derived DNA in coprolites at archaeological sites in continental East Asia and the Japanese archipelago. We also found that differences in ancestry proportions had a significant influence on regional variation in the incidence rate of asthma exacerbations in modern Japan. As for eosinophil counts and modern regional variations in asthma exacerbations, possible scenarios are as follows. For continental East Asian farmers before the migration to the Japanese archipelago, increased eosinophil counts were important as a protection against helminth infections. Continental East Asian farmers migrated into the Japanese archipelago during the Late Jomon and Yayoi periods with their rice-farming techniques, resulting in large-scale interbreeding with the indigenous Jomon hunter-gatherers. The regional populations of the current Japanese archipelago with a high ancestry proportion of continental East Asians have retained genetic factors that increase eosinophil counts, which resulted in a higher risk of asthma exacerbation. As a result, regional gradations in the incidence rate of asthma exacerbation among modern Japanese were caused by regional differences in the ancestry proportion of the continental East Asians. Overall, according to our findings with genome-wide allele frequencies and mean $2\beta f$ of the Jomon and continental ancestries, we can conclude that (1) some phenotypes of Jomon people and continental East Asians were highly diverged at the genome-wide scale (2) some phenotypic differences, especially in blood sugar and eosinophil counts, may have been the result of genetic adaptations to the respective livelihoods of the Jomon people and continental East Asians (3) regional variations in admixture proportions of the Jomon

people and continental East Asians formed phenotypic gradations of current Japanese archipelago populations. Integrating future GWAS of other traits in modern Japanese with allele frequencies on the Jomon ancestry would reveal their phenotypic characteristics that do not appear in excavated skeletal morphology. It is also possible, perhaps, to discover further regional gradations in phenotypes of modern Japanese caused by regional differences in admixture proportions, as in the case of height and asthma exacerbation.

As for the process of population formation in the Japanese archipelago from the Late Jomon period to the present, we propose a model, which is shown in Fig. 7. From the Late to Final Jomon period, the Jomon hunter-gatherers settled down in mainland Japan. They were a relatively short statured population with genetic factors to adapt to their hunter-gatherer lifestyle, such as high blood sugar for resistance to starvation. The population size or the population density of the Jomon people varied among regions. Koyama 1979[22] estimated that the population sizes in the Tohoku and Kanto regions were relatively large at 43,800 and 52,100, respectively, while those in the Kinki and Shikoku regions were relatively small at 4,400 and 2,700, respectively. At the same time, rice-farming populations lived in continental East Asia; they had relatively tall stature and genetically adapted to their livelihood with higher eosinophil counts for protecting against helminths. In the Final Jomon period, the continental East Asians arrived in northern Kyushu with their rice-farming technique and then admixed with the Jomon people in all the regions of mainland Japan. During the Yayoi period, the population size of immigrants was relatively increased in the Kinki and Shikoku regions, where the populations were small at the end of the Jomon period. Regional differences in population size from the Final Jomon to the Yayoi period made a variation in admixture proportions of the Jomon

people and continental East Asians among regions in modern Japanese archipelago. Regional variations in admixture proportion have resulted in the geographic gradations of today's Japanese genotypes and phenotypes.

Materials and 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[55] (Extended Data 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[56]. Our custom code for msprime simulation was described in Supplementary Text 3. 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 (between 18,000 YBP and 38,000 YBP) estimated in Kanzawa-Kiriyama et al.[10] 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[57–60].

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 (Fig. 1, Fig. S4 (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 (Fig. 1 and Fig. S4). To calculate the ancestry marker index (*AMI*), we first calculate the LD 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 variants on autosomes using real data

Detection of Jomon-derived variants in real data

Jomon-derived variants 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)[19], and 87 individuals from the Korean Personal Genome Project[61]. In this study, only biallelic SNPs were used. Prior to extracting the Jomon-derived variants, we performed a principal component analysis (PCA) in PLINK (version 1.9)[62] 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 (Fig. S6), so NA18976 was excluded from subsequent analyses. First, 1,784,634 SNPs specific to 1KG JPT were detected using VCFtools v0.1.13[63]. 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 SNPs originating from the Jomon people in the real data.

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

For the verification of Jomon-derived variants based on the whole genome sequence data, we devised the "Jomon allele score" (*JAS*). The *JAS* was calculated using the following formula:

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

We calculated *JAS* for the Ikawazu[9,11] and Funadomari[10] Jomon, as well as for 104 individuals from the 1KG JPT. 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[64]. 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-derived variants were subjected to LD pruning by the --

indep-pairwise command of PLINK (--indep-pairwise 1000 200 0.8). In addition, only the Jomon-derived variants 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 variants

Sample data

We used 183,708 SNPs from 10,842 individuals from the Japanese archipelago published by Watanabe et al. 2021[21]. 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 (Extended Data Fig. 1 and Table S1): Hokkaido (430 individuals), Tohoku (746 individuals), Kanto (3,990 individuals), Hokuriku (431 individuals), Chubu (410 individuals), Tokai (933 individuals), Kinki (1,861 individuals), Chugoku (600 individuals), Shikoku (314 individuals), Kyushu (1,016 individuals), and Okinawa (111 individuals). All statistical analyses were conducted at the Yahoo! Japan Corporation, with personal information of the customers completely hidden. We obtained approval from the Ethics Committee of the Yahoo! Japan Corporation.

Imputation of genotypes of Jomon-derived variants

Haplotype phasing and genotype imputation were performed using EAGLE2[65] and

Minimac3[66], respectively, with whole genome sequence data of 413 mainland Japanese[24] phased by SHAPEIT2[67]. After the imputation, Jomon-derived variants with high imputation quality ($R^2 > 0.8$) were extracted. Also, LD pruning was performed with PLINK (--indep-pairwise 1000 200 0.1), and a total of 3,917 Jomon-derived variants were used for the analysis.

Geographical distribution of the Jomon allele score

In subsequent analyses, individuals from Hokkaido that were largely affected by immigration after the Meiji period were excluded. Using 3,917 Jomon-derived variants, we calculated the *JAS* for individuals of each prefecture and compared them between regions and prefectures.

We compared the population size estimated from the number of archeological sites in each prefecture, assuming that the population size per archeological site was constant in each prefecture during the same period. We examined the correlations between (A) the average *JAS* in each prefecture and the number of archeological sites from the Jomon period (obtained from the Statistical report of buried cultural properties, Agency of Cultural Affairs, Japan; https://www.bunka.go.jp/seisaku/bunkazai/shokai/pdf/h29_03_maizotokei.pdf), (B) the average *JAS* in each region and the population size estimated from the number of archeological sites in the Late Jomon period[22], and (C) the average *JAS* in each prefecture and the $\log_{10}$(number of archeological sites in the Yayoi period/number of archeological sites in the Late Jomon period)[22]. Finally, (A) and (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).

Inference of phenotypic characteristics of the Jomon and continental East Asian ancestries by estimating their allele frequencies of genome-wide SNPs by Jomon-derived haplotypes

First, we applied the *AMI* to detect Jomon-derived variants in the whole genomes of 419 modern Japanese (THC dataset, *AMI* threshold = 100). The haplotypes of the regions around a focal SNP site (circles in Extended Data Fig. 6) were classified as "Jomon-derived haplotypes" or "continental haplotypes", depending on the presence of Jomon-derived variants in the 10 kb region upstream or downstream of the focal SNP. We extracted 67,607 Jomon-derived variants for estimating allele frequencies in THC Jomon ancestry and THC continental ancestry. We inferred allele frequencies of 6,481,773 genome-wide SNPs of minor allele frequencies > 1% in modern Japanese. The allele frequency estimation was performed by a custom script in R.

The estimated allele frequencies in the THC ancestries were compared with those in the 11 Jomon[9–11,13], three Kofun[13], THC 419 modern Japanese[24] and 103 Han Chinese (CHB from 1000 Genome Project) individuals[19]. In addition to the already genotyped Ikawazu and Funadomari Jomon individuals, bam files of nine Jomon and three Kofun individuals were provided by Dr. Nakagome of the School of Medicine, Trinity College Dublin, Dublin, Ireland, and the genotypes were determined by GATK UnifiedGenotyper. In the pairwise f3(A, B, Yoruba (= YRI in 1000 Genomes Project)) test, we used 32,143 SNPs that were genotyped in all the individuals of Jomon, Kofun and modern individuals.

To estimate the phenotypic characteristics of the Japanese ancestral population, we combined genome-wide allele frequencies from THC Jomon and continental ancestries with GWAS for 60 quantitative traits in modern Japanese population(Table S1)[25–27]. First, LD pruning was performed using the --indep-pairwise 100 200 0.1 option in plink to extract 1,594,120 SNPs with no LD in modern Japanese out of approximately 6 million SNPs. Referring to the QTL GWAS results for Japanese from previous studies, n SNPs with $P < 0.01$ for each trait were used to calculate the following index:

$$\text{mean } 2\beta f = \frac{\sum_{i=1}^{n} 2\beta_i f_i}{n}$$

Here β_i represents the effect of the i-th SNP of previous QTL GWAS results, and f_i represents the effect allele frequency for the i-th SNP. The $\text{mean } 2\beta f$ represents the average phenotype of a focal population, and the phenotype of each population can be compared by the magnitude of this value. For each trait, we calculated the $\text{mean } 2\beta f$ values for THC modern Japanese, Jomon and continental ancestries. The next step was to identify traits with significantly large differences in $\text{mean } 2\beta f$ between THC Jomon and continuous ancestries. For each trait, the null distribution of $\text{mean } 2\beta f$ was estimated by 1,000 simulations, where the allele frequency f_i for each SNP was randomly switched to THC Jomon or continental ancestries and calculated $\text{mean } 2\beta f$ values. Based on the 97.5th and 2.5th percentile of the null distribution of $\text{mean } 2\beta f$, the following index D was calculated for 60 traits to verify the magnitude of phenotypic differences between THC Jomon and continental ancestries:

$$D = \frac{(\text{mean } 2\beta f)_{Jomon} - (\text{mean } 2\beta f)_{Continental}}{\text{97.5th percentile of } (\text{mean } 2\beta f)_{\text{Permutation}} - \text{2.5th percentile of } (\text{mean } 2\beta f)_{\text{Permutation}}}$$

The larger the absolute value of D, the more different between Jomon and continental

ancestries is for a focal trait.

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 variants detected in this study, and the allele frequencies of Jomon-derived variants in each Japanese prefecture are available from the corresponding author upon request. Our custom code for msprime simulation was described in Supplementary Text 3.

References

1. Hanihara, K. Dual structure model for the population history of the Japanese. *Japan Rev.* **2**, 1–33 (1991).
2. Habu, J. *Ancient Jomon of Japan*. (Cambridge University Press, 2004).
3. Mizoguchi, K. *The archaeology of Japan : from the earliest rice farming villages to the rise of the state*. (Cambrige University Press, 2013).
4. Fujio, S. *History of Yayoi Period*. (Kodansha, 2015).
5. Jinam, T. *et al.* 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 (2012).
6. Jinam, T. A. *et al.* Unique characteristics of the Ainu population in Northern Japan. *J. Hum. Genet.* **60**, 565–571 (2015).
7. Nakagome, S. *et al.* 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 (2015).

8. Kanzawa-Kiriyama, H. *et al.* A partial nuclear genome of the Jomons who lived 3000 years ago in Fukushima, Japan. *J. Hum. Genet.* **62**, 213–221 (2017).

9. Mccoll, H. *et al.* The prehistoric peopling of Southeast Asia. *Science (80-.).* **361**, 88–92 (2018).

10. Kanzawa-Kiriyama, H. *et al.* Late Jomon male and female genome sequences from the Funadomari site in Hokkaido , Japan. *Anthropol. Sci.* **127**, 83–108 (2019).

11. Gakuhari, T. *et al.* Ancient Jomon genome sequence analysis sheds light on migration patterns of early East Asian. *Commun. Biol.* **3 437**, (2020).

12. Osada, N. & Kawai, Y. Exploring models of human migration to the japanese archipelago using genome-wide genetic data. *Anthropol. Sci.* **129**, 45–58 (2021).

13. Cooke, N. P. *et al.* Ancient genomics reveals tripartite origins of Japanese populations. *Sci. Adv.* **7**, 1–16 (2021).

14. Plagnol, V. & Wall, J. D. Possible ancestral structure in human populations. *PLoS Genet.* **2**, e105 (2006).

15. Vernot, B. & Akey, J. M. Resurrecting Surviving Neandeltal Linages from Modern Human Genomes. *Science (80-.).* **343**, 1017–1021 (2014).

16. Vernot, B. *et al.* Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science (80-.).* **352**, 235–239 (2016).

17. Jacobs, G. S. *et al.* Multiple Deeply Divergent Denisovan Ancestries in Papuans. *Cell* **177**, 1010-1021.e1-e24 (2019).

18. Kim, J. *et al.* The Origin and Composition of Korean Ethnicity Analyzed by Ancient and Present-Day Genome Sequences. *Genome Biol. Evol.* **12**, 553–565 (2020).

19. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

20. Yamaguchi-Kabata, Y. *et al.* Japanese Population Structure, Based on SNP Genotypes from 7003 Individuals Compared to Other Ethnic Groups: Effects on Population-Based Association Studies. *Am. J. Hum. Genet.* **83**, 445–456 (2008).

21. Watanabe, Y., Isshiki, M. & Ohashi, J. Prefecture-level population structure of the Japanese based on SNP genotypes of 11,069 individuals. *J. Hum. Genet.* **66**, 431–437 (2021).

22. Koyama, S. Jomon Subsistence and Population. *Senri Ethnol. Stud.* **2**, 1–65 (1979).

23. Jinam, T. *et al.* Genome-wide SNP data of Izumo and Makurazaki populations support inner-dual structure model for origin of Yamato people. *J. Hum. Genet.* **66**, 681–687 (2021).

24. Watanabe, Y. *et al.* Analysis of whole Y-chromosome sequences reveals the Japanese population history in the Jomon period. *Sci. Rep.* **9**, 8556 (2019).

25. Akiyama, M. *et al.* Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat. Genet.* **49**, 1458–1467 (2017).

26. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).

27. Akiyama, M. *et al.* Characterizing rare and low-frequency height-associated variants in the Japanese population. *Nat. Commun.* **10**, (2019).

28. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet.* **9**, (2013).

29. 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).

30. Isshiki, M. *et al.* Admixture and natural selection shaped genomes of an Austronesian-speaking population in the Solomon Islands. *Sci. Rep.* **10**, 1–9 (2020).

31. Ravin, K. A. & Loy, M. The eosinophil in infection. *Clin. Rev. Allergy Immunol.* **50**, 214–227 (2015).

32. Hiramoto, Y. Secular Change of Estimated Stature of Japanese in Kanto District from the Prehistoric Age to the Present Day. *J. Anthropol. Soc. Nippon* **80**, 221–236 (1972).

33. Kaifu, Y. Human Skeletal Remains of the Yayoi Period from the Iwatusbo Cave Site in Gunma Prefecture, Kanto District. *J. Anthropol. Soc. Nippon* **100**, 449–483 (1992).

34. Wada, Y. & Motomura, H. Temporal changes in stature of Western Japanese based on limb characteristics. *Anthropol. Sci.* **108**, 147–168 (2000).

35. Isshiki, M., Watanabe, Y. & Ohashi, J. Geographic variation in the polygenic score of height in Japan. *Hum. Genet.* **140**, 1097–1108 (2021).

36. Lambrecht, B. N. & Hammad, H. The immunology of asthma. *Nat. Immunol.* **16**, 45–56 (2015).

37. Nakagome, K. & Nagata, M. Involvement and Possible Role of Eosinophils in Asthma Exacerbation. *Front. Immunol.* **9**, 1–8 (2018).

38. Yokoyama, A. *et al.* Regional differences in the incidence of asthma exacerbations in Japan: A heat map analysis of healthcare insurance claims data. *Allergol. Int.* **71**, 47–54 (2022).

39. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations

for present-day Europeans. *Nature* **513**, 409–413 (2014).

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

41. Pierron, D. *et al.* Genomic landscape of human diversity across Madagascar. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E6498–E6506 (2017).

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

43. Minagawa, M. Dietary pattern of prehistoric Japanese populations inferred from stable carbon and nitrogen isotopes in bone protein. *Bull. Natn. Mus. Jap. Hist.* **86**, 333–357, (in Japanese) (2001).

44. Yoneda, M. *et al.* Isotopic evidence of inland-water fishing by a Jomon population excavated from the Boji site, Nagano, Japan. *J. Archaeol. Sci.* **31**, 97–107 (2004).

45. Kusaka, S., Hyodo, F., Yumoto, T. & Nakatsukasa, M. Carbon and nitrogen stable isotope analysis on the diet of Jomon populations from two coastal regions of Japan. *J. Archaeol. Sci.* **37**, 1968–1977 (2010).

46. Kusaka, S. *et al.* Homogeneous diet of contemporary Japanese inferred from stable isotope ratios of hair. *Sci. Rep.* **6**, 1–11 (2016).

47. Temple, D. H. & Larsen, C. S. Dental Caries Prevalence as Evidence for Agriculture and Subsistence Variation During the Yayoi Period in Prehistoric Japan : Biocultural Interpretations of an Economy in Transition. **512**, 501–512 (2007).

48. Fujita, H. Dental Caries in Japanese Human Skeletal Remains. *J. Oral Biosci.* **51**, 105–114 (2009).

49. Reinhard, K. Cultural ecology of prehistoric parasitism on the Colorado Plateau as evidenced by coprology. *Am. J. Phys. Anthropol.* **77**, 355–356 (1988).

50. London, D. & Hruschka, D. Helminths and human ancestral immune ecology: What is the evidence for high helminth loads among foragers? *Am. J. Hum. Biol.* **26**, 124–129 (2014).

51. Yin, M. *et al.* Co-dispersal of the blood fluke Schistosoma japonicum and Homo sapiens in the Neolithic Age. *Sci. Rep.* **5**, 1–8 (2015).

52. Kanehara, M. & Kanehara, M. In the forefront of research, Jomon people suffered from parasites. *Gekkan Rekishi Kaido No. 84 (Monthly Mag. Histrical Trailways), Tokyo, Japan.* (1995).

53. Matsui, A., Kanehara, M. & Kanehara, M. Palaeoparasitology in Japan - Discovery of Toilet Features. *Mem. Inst. Oswaldo Cruz* **98**, 127–136 (2003).

54. Mitre, E. & Klion, A. D. Eosinophils and helminth infection: protective or pathogenic? *Semin. Immunopathol.* **43**, 363–381 (2021).

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

56. 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).

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

58. Altshuler, D. L. *et al.* A map of human genome variation from population-scale

sequencing. *Nature* **467**, 1061–1073 (2010).

59. Scally, A. & Durbin, R. Revising the human mutation rate: Implications for understand ing human evolution. *Nat. Rev. Genet.* **13**, 745–753 (2012).

60. Kong, A. *et al.* Rate of de novo mutations and the importance of father-s age to disease risk. *Nature* **488**, 471–475 (2012).

61. Zhang, W. *et al.* Whole genome sequencing of 35 individuals provides insights into the genetic architecture of Korean population. *BMC Bioinformatics* **15**, 1–13 (2014).

62. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

63. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

64. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

65. Loh, P. R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

66. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

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

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. 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. We also thank Dr. Nakagome Shigeki from the the School of Medicine, Trinity College Dublin, Dublin, Ireland for providing us the BAM file of the nine Jomon and three Kofun individuals. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

Funding: 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 and 21H00336) and Grant-in-Aid for Early-Career Scientists (21K15175) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

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.

Figures and Tables

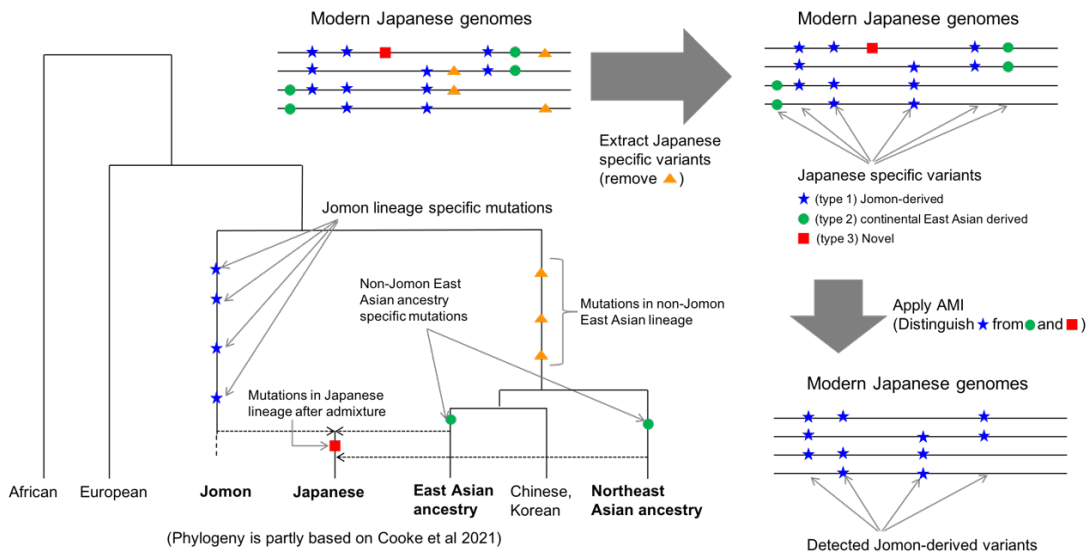


Fig. 1 Overview of our *AMI*-based detection method of Jomon-derived variants of modern Japanese

Phylogenetic relationships between modern Japanese, the Jomon people and continental East Asians[8,10,11,13]. Modern Japanese genomes contain a mixture of variants that occurred on each ancestral lineage of modern Japanese (Jomon and continental East Asian ancestries including "East Asians" and "Northeast Asians" in Cooke et al., 2021). The Jomon people were highly differentiated from continental East Asians, so modern Japanese genomes have Jomon-lineage specific variants (blue stars in the phylogeny) that are not found in other modern continental East Asians (Chinese, Koreans and so on); *AMI* is an index to distinguish Jomon-derived variants from other Japanese-specific variants based on the linkage disequilibrium status.

Fig. 2 Performance of the *AMI* for the detection of the Jomon-derived variants

(A) Distribution of *AMI* simulated by msprime. The histogram of *AMI* for the Jomon-derived variants (type 1) and the other variants (types 2 and 3) are shown. The red dashed line indicates the threshold of *AMI* (28.0374) obtained from ROC analysis for the detection of the Jomon-derived variants (type 1).

(B) ROC curve illustrating the performance of the *AMI* for the detection of the Jomon-derived variants. The ROC curve was drawn based on the simulated data shown in Fig. 2 (A). The *AMI* showed high accuracy (AUC = 0.91) for discriminating the Jomon-derived variants (type 1) from the other variants (types 2 and 3).

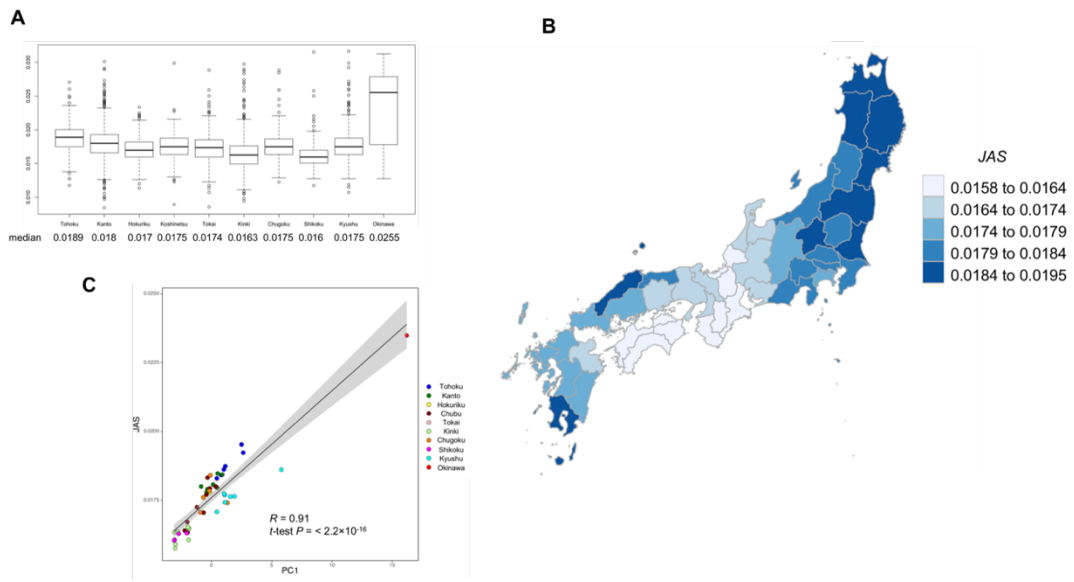


Fig. 3 JAS of each Japanese region.

(A) Distribution of *JAS* in ten regions. The boxplot of the *JAS* is presented for 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 Extended Data Fig. 1. Horizontal axe: PC1, vertical axe: average *JAS*. Pearson's correlation coefficients (*R*), *P* values, regression lines and 95% CI are shown.

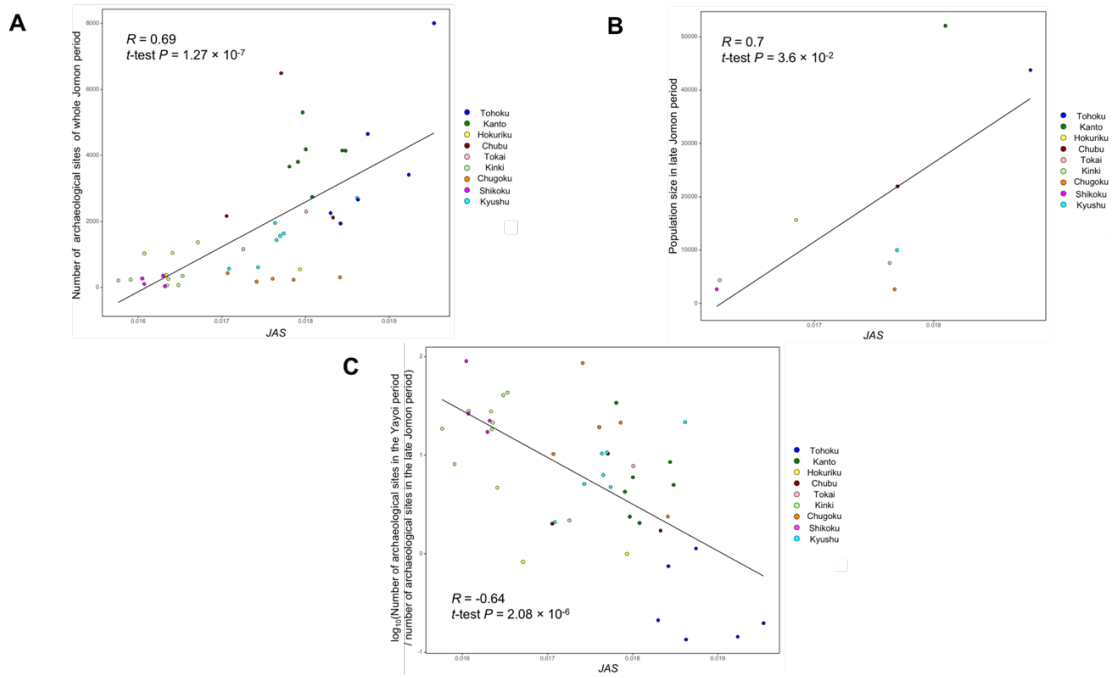


Fig. 4 Relationship between the *JAS* and values associated with the population size of prefectures in the Jomon to Yayoi periods. The horizontal axis shows the average *JAS* and the vertical axis shows the (A) number of archaeological sites of the whole Jomon period, (B) population size in the Late 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 are shown in each figure. Each prefecture is colored according to the region in Extended Data Fig. 1.

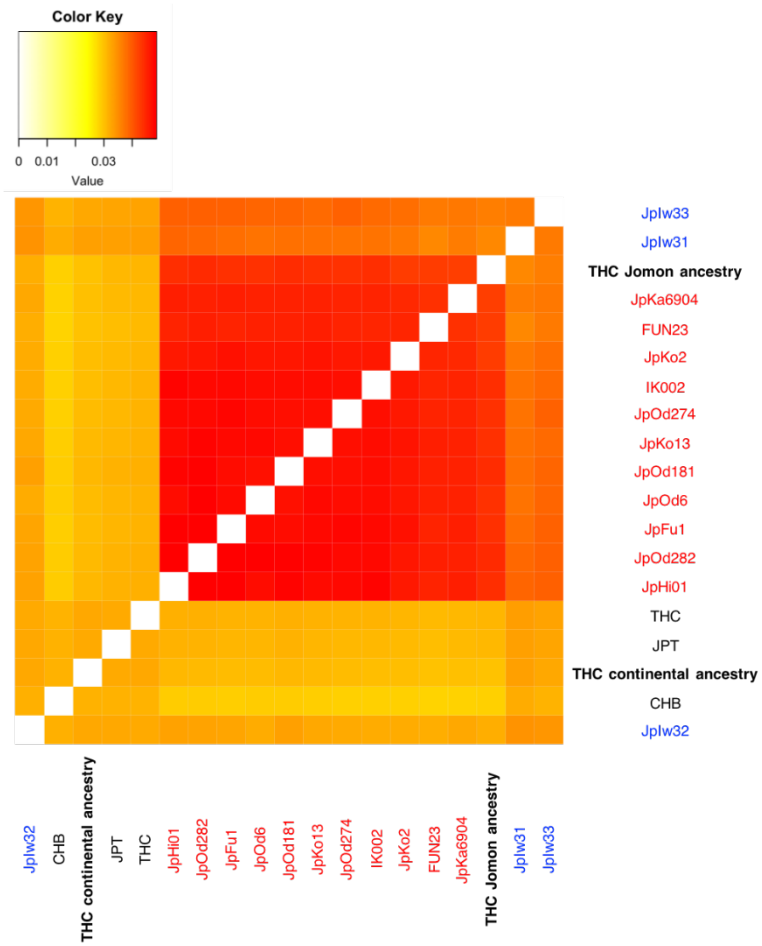


Fig. 5 Accuracy of the allele frequency estimation in the Jomon people from Jomon-derived haplotypes of modern Japanese

A heatmap of pairwise f3 statistics between the actual Jomon (red) and Kofun (blue) individuals, modern Japanese (JPT and THC), modern Han Chinese (CHB) and Jomon/continental ancestries of modern Japanese inferred from Jomon-derived haplotypes of modern Japanese.

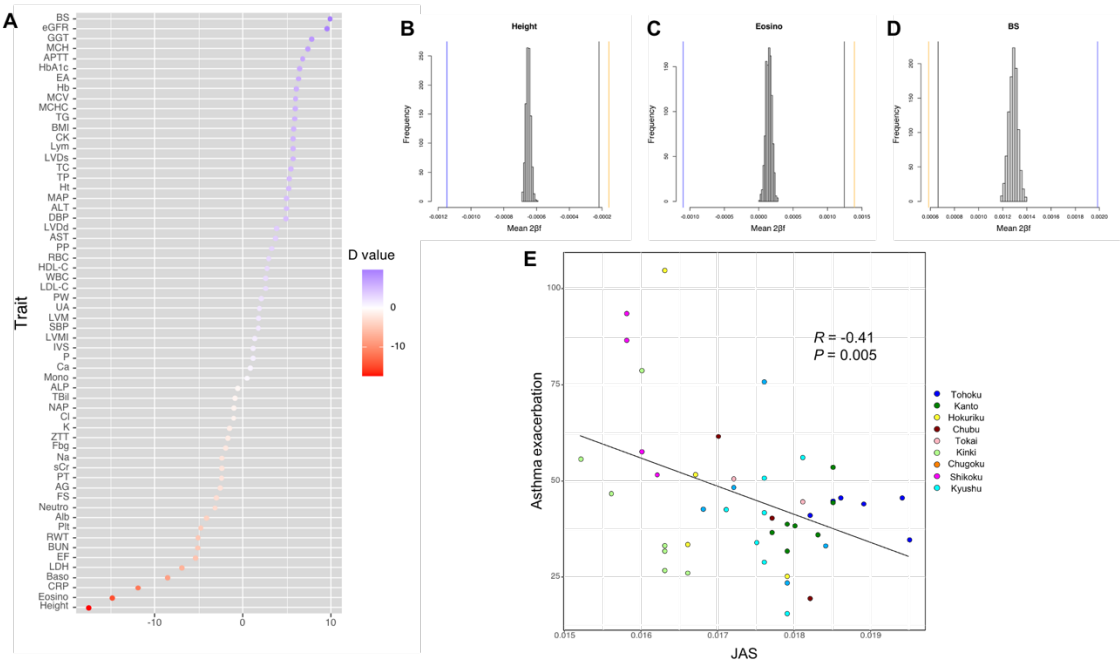


Fig. 6 Inference of phenotypic characteristics of the Jomon and continental ancestries and impact of regional admixture proportion differences on regional phenotypic variations of modern Japanese.

(A) *D* value for each phenotype, plotted in dark purple when the Jomon ancestry had higher mean $2\beta f$ values than the continental ancestry, and dark red vice versa. (B) - (D) Histograms of the mean $2\beta f$ obtained by simulations for height, eosinophil count and blood sugar. Mean $2\beta f$ was shown for THC modern Japanese (a black line), THC Jomon (a blue line) and continental ancestries (an orange line) for each phenotype. (E) Relationship between the *JAS* and incidence rates of asthma exacerbation of prefectural populations of modern Japanese. Pearson's correlation coefficients (*R*), *P* values are shown in each figure. Each prefecture is colored according to the region in Extended Data Fig. 1.

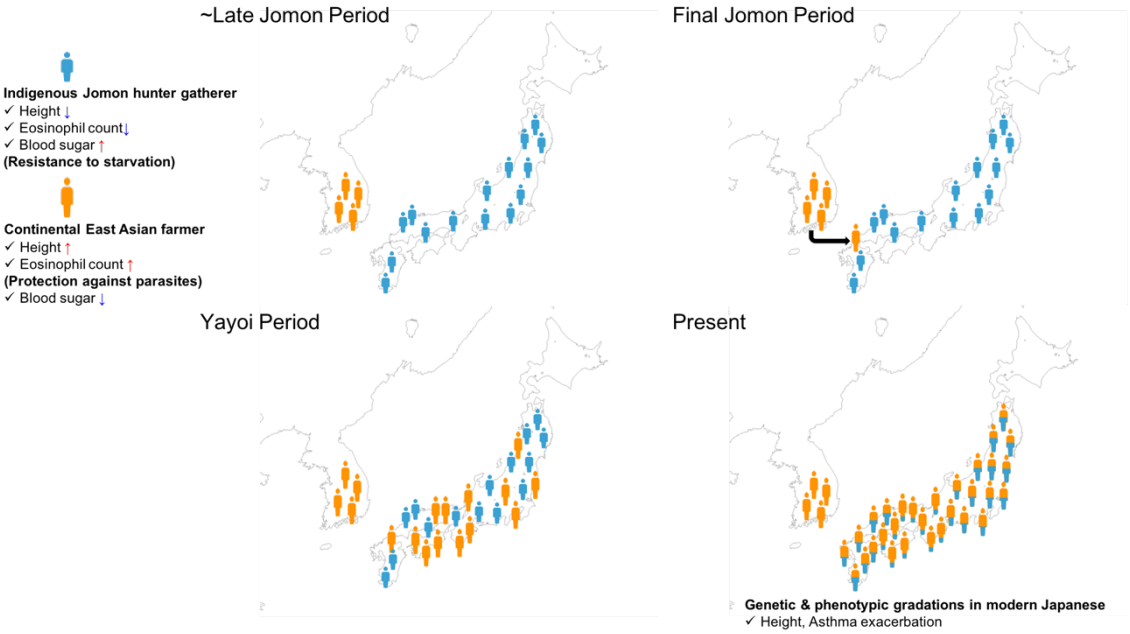
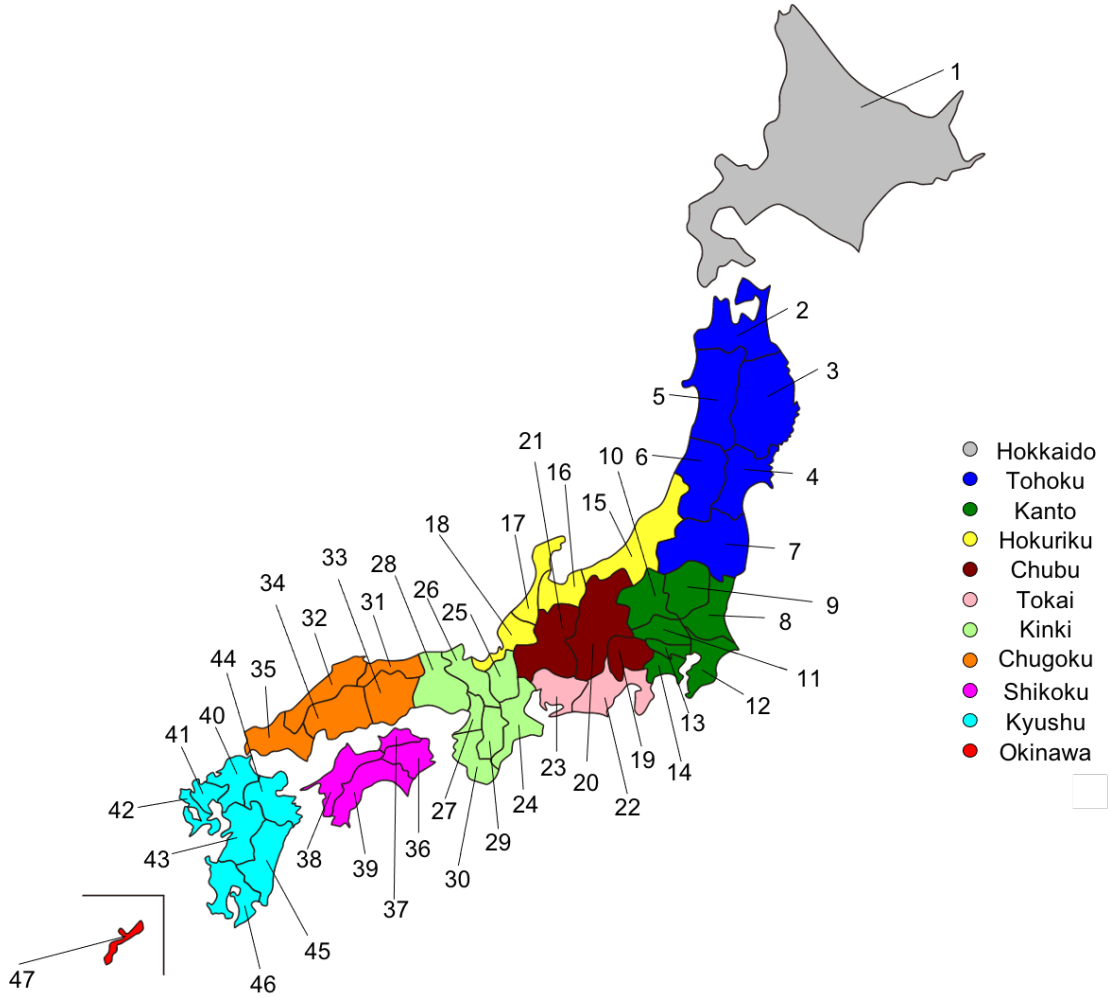
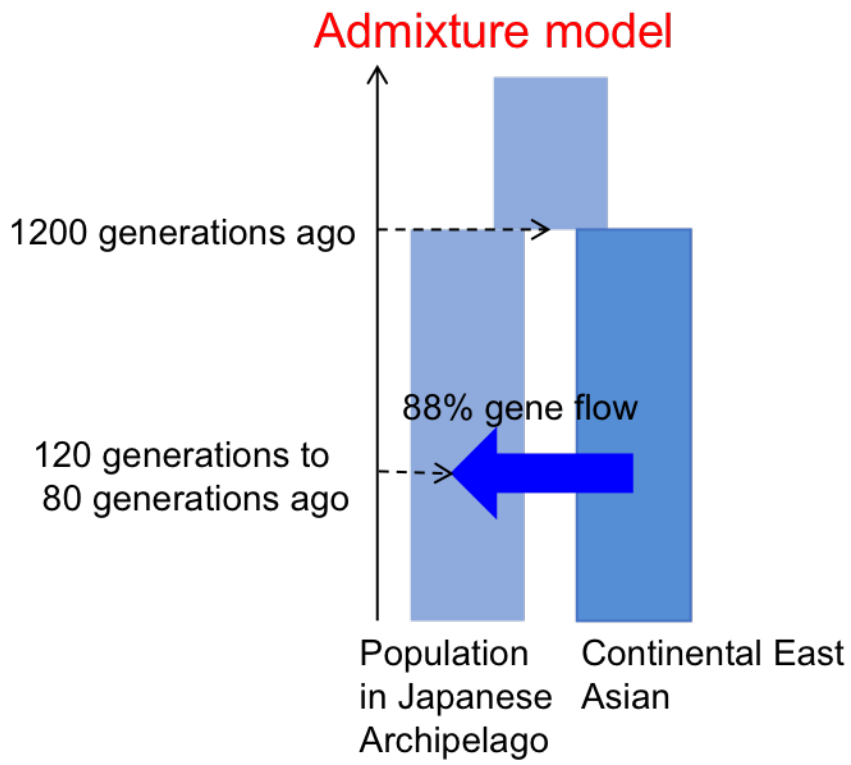


Fig. 7 Formation process of regional population of the Japanese archipelago.

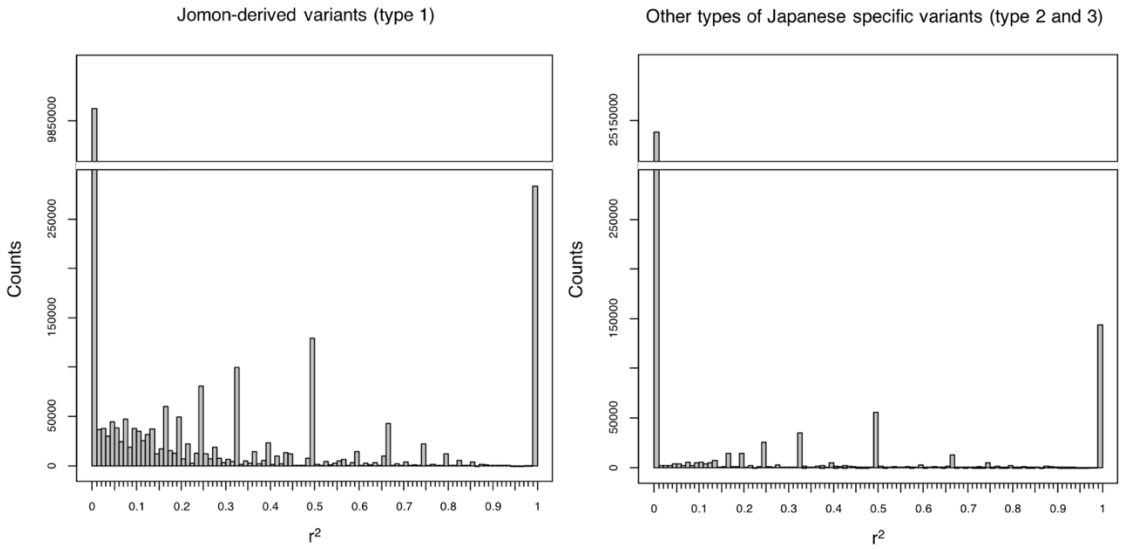
Extended Data Figures



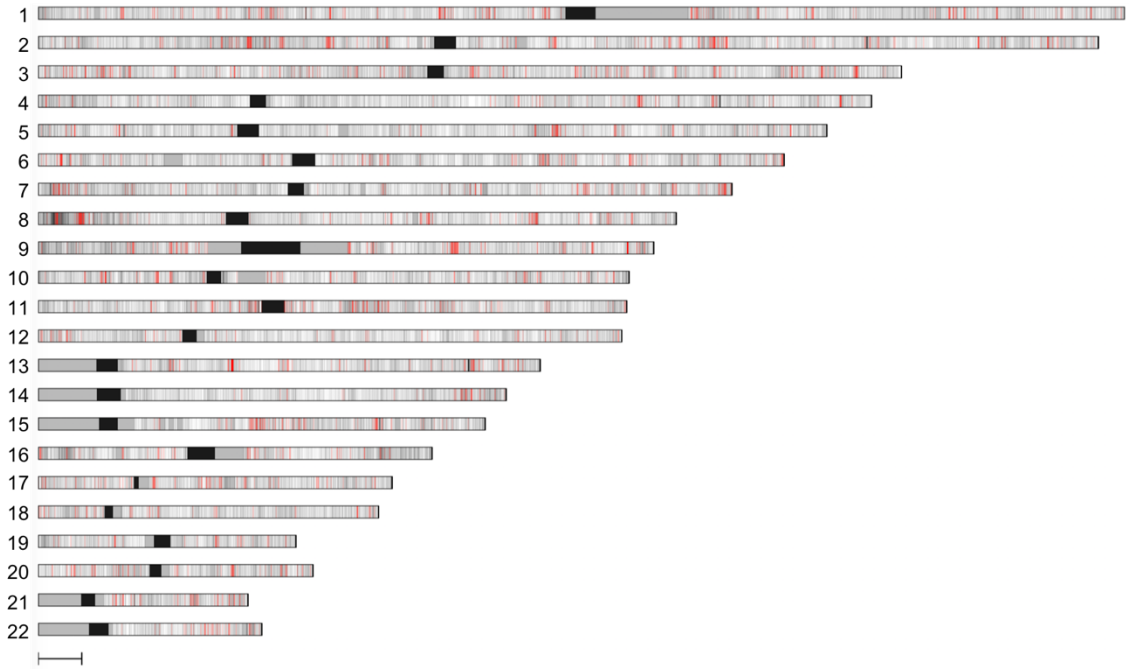
Extended Data Fig. 1 Map of the Japanese prefectures. The prefectures of Japan are divided into eleven regions. The prefecture numbers in Table S1 are indicated (the corresponding prefecture names are given in Table S1). In this study, "mainland Japanese" means the Japanese people except for individuals from Hokkaido and Okinawa.



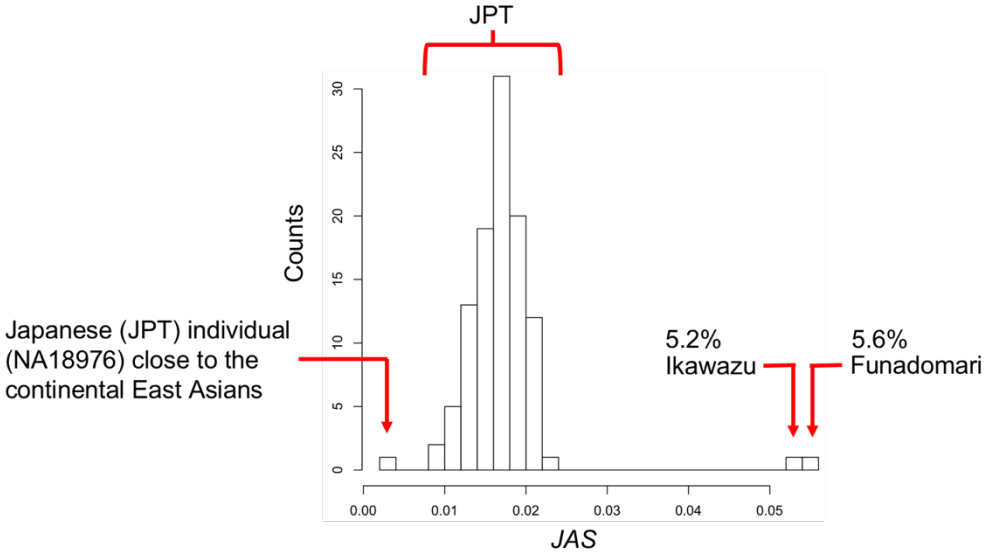
Extended Data Fig. 2 Basic demographic model of the Japanese population assumed in my coalescent simulations. The split between the Jomon ancestors and the continental East Asians was set to 1,200 generations ago (30,000 YBP). The continuous migration from the continental East Asian to the Jomon was set between 120 and 80 generations ago (3,000 ~2,000 YBP). The total admixture proportion of the Jomon people in the modern Mainland Japanese was set to 12%. The effective population size was 5,000 for each population. Some parameters were varied as necessary for the analysis of Fig. S2, S3, and S5 to validate the performance of S* and the *AMI*.



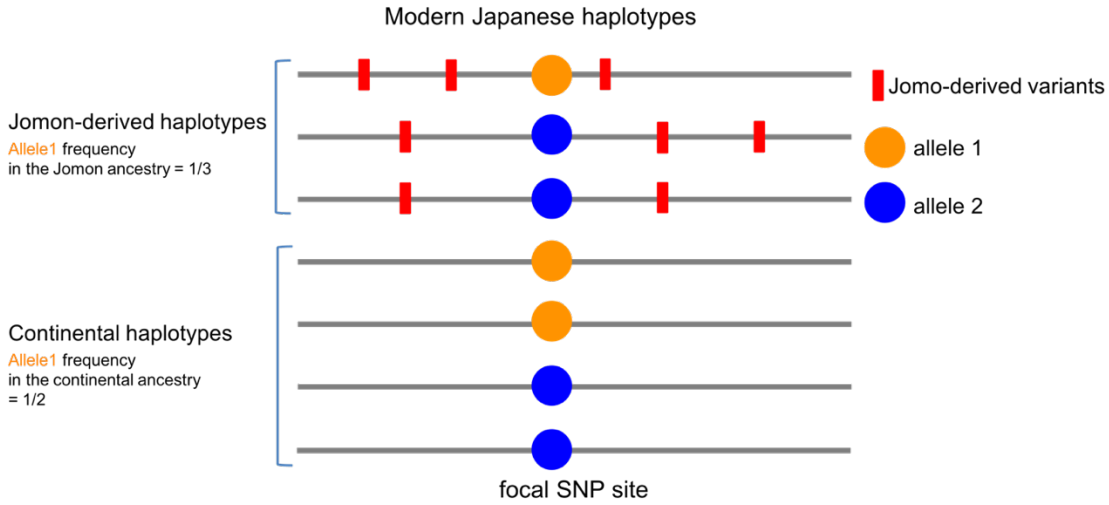
Extended Data Fig. 3 Histograms of r^2 value between Jomon-derived variants (type 1) and between other types of Japanese specific variants (type 2 and 3 together). A large gap existed between the counts of SNP pairs with $r^2 < 0.01$ and those with $r^2 > 0.01$, so the y-axis values that exceed 300,000 counts were partly omitted to emphasize the difference of SNP pair counts with $r^2 > 0.01$ between (type 1) and (type 2 and 3).



Extended Data Fig. 4 Distribution of Jomon-derived variants in the Japanese genomes. The vertical lines represent the Jomon-derived variants, with gray representing SNPs with a Jomon allele frequency of less than 5% and red representing SNPs with a Jomon allele frequency of more than 5%. The black square shows the centromere. The gray square shows the regions with a density of Japanese specific variants below a mean - 1sd of each chromosome (referred to in "Detection of the Jomon derived variants in real data" of the Materials and Methods section). The 10Mb scale bar is shown under chromosome 22.



Extended Data Fig. 5 Distribution of the *JAS* of the Mainland Japanese and Jomon individuals. The *JAS* were calculated for 104 individuals of JPT (Japanese in Tokyo, Japan) and two Jomon samples (Funadomari and Ikawazu).



Extended Data Fig. 6 Estimation method of the allele frequencies of genome-wide SNPs in the Jomon and continental ancestries of modern Japanese. The modern Japanese haplotypes surrounding a focal variant site can be classified into "Jomon-derived haplotypes" and "continental haplotypes" according to the presence Jomon alleles of Jomon-derived variants (red bar). The allele frequency of the Jomon people in the focal variant site can be estimated by the proportion of each allele within Jomon-derived haplotypes.