Model-Based Verification of Hypotheses on the Origin of Modern Japanese Revisited by Bayesian Inference Based on Genome-Wide SNP Data

Shigeki Nakagome,1 Takehiro Sato,2 Hajime Ishida,2 Tsunehiko Hanihara,3 Tetsutaro Yamaguchi,4 Ryosuke Kimura,2 Shuhei Mano,*1 Hiroki Oota,*3 and The Asian DNA Repository Consortium

1Department of Mathematical Analysis and Statistical Inference, The Institute of Statistical Mathematics, Tokyo, Japan
2Department of Human Biology and Anatomy, Graduate School of Medicine, University of the Ryukyus, Okinawa, Japan
3Department of Anatomy, Kitasato University School of Medicine, Kanagawa, Japan
4Department of Orthodontics, School of Dentistry, Showa University, Tokyo, Japan

*Corresponding author: E-mail: smano@ism.ac.jp; hiroki_oota@med.kitasato-u.ac.jp.

Associate editor: Sohini Ramachandran

Abstract

Various hypotheses for the peopling of the Japanese archipelago have been proposed, which can be classified into three models: transformation, replacement, and hybridization. In recent years, one of the hybridization models ("dual-structure model") has been widely accepted. According to this model, Neolithic hunter-gatherers known as Jomon, who are assumed to have originated in southeast Asia and lived in the Japanese archipelago greater than 10,000 years ago, admixed with an agricultural people known as Yayoi, whom were migrants from the East Asian continent 2,000–3,000 years ago. Meanwhile, some anthropologists propose that rather, morphological differences between the Jomon and Yayoi people can be explained by microevolution following the lifestyle change. To resolve this controversy, we compared three demographic models by approximate Bayesian computation using genome-wide single nucleotide polymorphism (gwSNP) data from the Ainu people who are thought to be direct descendants of indigenous Jomon. If we assume Chinese people sampled in Beijing from HapMap have the same ancestry as Yayoi, then the hybridization model is predicted to be between 29 and 63 times more likely than the replacement and transformation models, respectively. Furthermore, our data provide strong support for a model in which the Jomon lineages had population structure diversified in local areas before the admixture event. Initial divergence between the Jomon and Yayoi ancestries was dated to late Pleistocene, followed by the divergence of Jomon lineages at early Holocene. These results suggest gwSNP data provides a detailed picture of the complex hybridization model for Japanese population history.

Key words: origin of modern Japanese, Bayesian inference, genome-wide SNP data.

Introduction

The prehistory of Japan is stratified by two unique cultures after the late-Paleolithic period: the Neolithic Jomon and the post-Neolithic Yayoi. The Neolithic Jomon culture began more than 10,000 years ago and featured a hunter-gatherer lifestyle that included some of the earliest pottery in the world, whereas the post-Neolithic Yayoi culture began 2,000–3,000 years ago and included weaving, metalworking, and large-scale rice cultivation expanding to include much of the Japanese archipelago (Habu 2004). However, it has been controversial whether people from the same lineage produced those cultures, as morphological traits of skeletal remains have suggested remarkable differences between individuals excavated from various Jomon and the Yayoi sites (Yamaguchi 1982; Imamura 1996).

Physical anthropologists, therefore, have proposed various hypotheses of population history in the Japanese archipelago that are classified into three models known as 1) transformation, 2) replacement, and 3) hybridization. The first model proposes modern Japanese people evolved from a single ancestral population (Hasebe 1940; Suzuki 1969). The Jomon people, who dwelt throughout the Japanese archipelago from 15,000 to 3,000 years ago, are thought to have gradually transformed their morphological features into those typical of modern Japanese (Suzuki 1981; Mizoguchi 1986). In this model, migrants after the Yayoi period may have contributed culturally but not genetically to the formation of Japanese, that is, the genetic contribution of migrants from the East Asian continent from around 2,000 years ago should be almost zero to the modern Japanese people. The second model involves complete replacement of indigenous Jomon people by the Yayoi (Howells 1966; Tuner 1976). Therefore, this model suggests modern Japanese are the direct descendants of the Yayoi people, that is, the genetic contribution of indigenous Jomon to the modern Japanese people should be zero. The hybridization model claims that modern Japanese are an admixed population between the Jomon and the Yayoi people (Kiyono 1949).
The dual structure model for the population history of the Japanese people is a hybridization model that Hanihara (1991) proposed concerning the demographic history of pre-historic Japanese based on morphological analyses of crania and teeth (Hanihara 1991). The dual structure model proposes two main events of (I) migration and (II) admixture: (I) the Jomon people originated somewhere in Southeast Asia and came to the Japanese archipelago around 12,000 years ago, whereas the Yayoi people came from northeast Asia around 2,000 years ago and entered the northern part of Kyushu, which is the western island of Honshu (main-island Japan), probably via the Korean peninsula. Then, (II) the Yayoi migrants admixed with the indigenous Jomon people gradually from west to east. Because the degree of Jomon or Yayoi ancestry may vary among local populations, the Japanese populations can be characterized by the dual structure of ancestral populations: people in main-island Japan have a high degree of genetic contribution from the Yayoi migrants, whereas the Ainu on Hokkaido (the northernmost island) and the Ryukyuans in Okinawa (the southernmost islands of the Japanese archipelago) have a higher degree of genetic contribution from the indigenous Jomon. Meanwhile, recent morphological studies have reported that Jomon skeletal series vary among geographic regions and that cranial variation is clearly distinct between the northeast and the southwest portions of the Japanese archipelago (Hanihara and Ishida 2009; Nakashima et al. 2010; Fukase et al. 2012a, 2012b). Therefore, the Jomon people are likely to have already diversified among the local regions before the Yayoi migration happened; thus, the population history in the Japanese archipelago may be complicated with the population structure of the Jomon.

The three models of the origins of modern Japanese can be interpreted as the question of whether the agricultural Yayoi people expanded into the Japanese archipelago by “cultural” or “demic” diffusions (Ammerman and Cavalli-Sforza 1984). Cultural diffusion happens if the knowledge and technology of agricultural production spreads without migration, which is equivalent to the transformation model. In contrast, the replacement model is the result of demic diffusion, assuming that the migrating farmers brought agriculture along with them. The hybridization model is the intermediate scenario between the two contrasting diffusions. The process of cultural diffusion is thought to be quick and to rarely have demographic consequences, whereas the spread of technologies by demic diffusion involves a massive movement of people (Cavalli-Sforza 2000). Distinguishing between the models of the origin of modern Japanese will help us to understand the onset and process of transition to agriculture in Japan. Some early genetic studies support the transformation (Nei 1995) and the replacement (Cavalli-Sforza et al. 1994) models. However, most studies based on mitochondrial DNA (mtDNA), classical markers in autosomes, and the Y chromosome have reinforced the part of the dual structure model concerning “hybridization” (Hammer and Horai 1995; Horai et al. 1996; Omoto and Saitou 1997; Tanaka et al. 2004; Hammer et al. 2006; Matsukusa et al. 2010; Koganebuchi et al. 2012) although no genetic evidence has been observed to support the idea that the Jomon originated in Southeast Asia (assuming the Ainu and the Ryukyuans are the direct descendants of the Neolithic Jomon [Omoto and Saitou 1997; Tajima et al. 2004]). Sequence data from the mtDNA hypervariable region have shown approximately 65% of the gene pool in Honshu Japan is derived from the gene flow from the East Asian continent after the Yayoi period (Horai et al. 1996). A phylogenetic analysis based on classical genetic markers has shown that the Ainu and the Ryukyuans have clustered separately from East Asian populations (Omoto and Saitou 1997). Haplogroup analyses based on single nucleotide polymorphisms (SNPs) and short tandem repeat polymorphisms of the Y chromosome have also indicated a close relationship between the Ainu and the Ryukyuans (Hammer et al. 2006; Koganebuchi et al. 2012).

In more recent years, significant amounts of genome-wide SNP (gwSNP) data from East and Southeast Asian populations have been reported (Yamaguchi-Kabata et al. 2008; Abdulla et al. 2009; Jinam et al. 2012). For modern Japanese, more than 140,000 SNPs in 7,003 individuals indicated significant genetic differentiation between the Ryukyuans and non-Ryukyuan Japanese (minus the Ainu), using principle component analyses (Yamaguchi-Kabata et al. 2008). Furthermore, nearly 1 million SNPs in 36 Ainu and 38 Ryukyuans showed that the two populations were closely clustered together in the phylogenetic tree, and the ancestral lineage was the closest to the Honshu Japanese, supporting admixture between the Jomon and Yayoi which is primarily claimed by the dual structure model (Jinam et al. 2012).

In spite of abundant gwSNP data, there have been few studies concerning demographic inference by estimating evolutionary parameters. In particular, because the transformation and replacement models are nested by the hybridization model with few (cultural diffusion) or many dispersing (demic diffusion) farmers, the admixture proportion can be a crucial parameter in distinguishing the three models. Therefore, further studies are needed in order to reconcile Japanese demographic history. Estimating demographic parameters based on genome-wide data can provide detailed insights into the demographic history of Japanese populations. In particular, large genetic data sets enable us to reconstruct the population history at finer scales, shedding light on evolutionary processes which have shaped patterns of genetic variation under complex demographic histories. One of the largest publicly available databases is HapMap, which includes two of the major East Asian populations: the Japanese from Tokyo (JPT) and the Chinese from Beijing (CHB) (Frazer et al. 2007). Here, we primarily use the Ainu gwSNP data reported in Jinam et al. (2012), making the explicit assumption that they are the direct descendants of the Neolithic Jomon, and that JPT are representative of modern Japanese. Furthermore, we presume CHB to be a remnant population of Yayoi migrants who are supposed to have originated in the northeast Asia (Hanihara 1991). Based on these gwSNP data, we apply Bayesian methodology to compare three alternative models for the origins of Japanese and estimate demographic parameters.
Results

Model Selection for the Three Hypotheses

The demographic models (transformation, replacement, and hybridization) are illustrated in figure 1, including three parameters common to the three models ($N_{Ainu}$, population size in Ainu; $N_{JPT}$, population size in JPT; $T_{AIN-CHB}$, time of population divergence between Ainu and CHB) and four parameters specific to each model ($T_{AIN-JPT}$, time of population divergence between Ainu and JPT; $T_{JPT-CHB}$, time of population divergence between JPT and CHB; $T_{ADMMIX}$, time of admixture between Ainu and CHB; $P$, the admixture proportion from the Jomon ancestry). A migration parameter ($m = 1.0 \times 10^{-5}$) after the admixture event was incorporated between Ainu and JPT to evaluate this effect on the model selection. The transformation model was characterized as the modern Japanese as direct descendants of the Jomon people. Next, we considered that the Ainu split from CHB at $T_{AIN-CHB}$ and that JPT diverged from Ainu at $T_{AIN-JPT}$. In contrast, the replacement model stated that the modern Japanese were of Yayoi ancestry and we postulated that JPT branched off from CHB at $T_{JPT-CHB}$. The hybridization model was expressed as the admixture between Ainu and CHB that happened at $T_{ADMMIX}$ subsequent to the divergence between them at $T_{AIN-CHB}$. Because complete isolation between Ainu and JPT seems to be an extreme assumption, we also incorporated a continuous gene flow between these two geographically proximate populations into the three models (fig. 1). Then, we tested the model selection with or without migration. We used kernel-approximate Bayesian computation (kernel-ABC), which is a likelihood-free approach for Bayesian inferences, to incorporate high dimensional summary statistics and improve approximation of posterior estimates given the data (Fukumizu et al. 2013; Nakagome, Fukumizu, et al. 2013; Nakagome, Mano, et al. 2013; Osada et al. 2013; Sato et al. 2014). Further, we applied the algorithm of kernel-ABC to the model selection framework based on approximated Bayes factor (aBF), which gives a summary of the evidence provided by the data in favor of one model as opposed to another model (Osada et al. 2013). In this study, we selected the model that best fit for the gwSNP data by aBF and estimated demographic parameters using kernel-ABC.

The gwSNP data were classified into haplotype frequency spectra (HFS) consisting of the haplotype frequency, $\delta_j$, which is the number of haplotypes observed in $j$ haplotypes in the samples. Patterns of haplotype frequency can be affected by both recombination and demographic history; the joint distribution of $\delta_j$ across genomic regions is a useful statistic for inference (Lohmueller et al. 2009). A recent study has theoretically demonstrated that a recombination effective population size is more sensitive to the change in population sizes than a coalescent effective population size based on nucleotide diversity or average pairwise nucleotide differences (Mano 2007). Furthermore, sharing of haplotypes between populations can reflect the time since an ancestral population diverged into subpopulations. Here, we utilized three-dimensional (3D)-HFS which is a 3D-array ($\delta_{ijh}$) representing the number of haplotypes that are observed on the $i$, $j$, and $h$ chromosomes in Ainu, JPT, and CHB (see Materials and Methods for more detail). The gwSNP data were divided into subsets of SNPs in which the genetic distance between the first and last SNP sites in the order of their physical positions was 0.15 cM. Each of the subsets represented haplotype variation in a genomic segment; the total segments (i.e., subsets) numbered 18,026. For each segment, the makeup of haplotypes among the populations was summarized into a 3D-HFS; however, the array of 3D-HFS was sparse (including “0”) and the matrix computation with them became unstable. Therefore, we merged the arrays from 100 segments into a set of 3D-HFS; the total number of 3D-HFS was 180 sets. Finally, we used the resulting 180 arrays as observed data in the kernel-ABC. Detailed procedures of our computational analyses are given in Materials and Methods.

![Figure 1](https://example.com/fig1.png)

**Fig. 1.** Three demographic models for the origins of mainland Japanese: (a) transformation, (b) replacement, and (c) hybridization. In all models, the population divergence initially occurs at $T_{AIN-CHB}$ between the Ainu and CHB. The subsequent events are different among the models. The transformation model assumes the divergence at $T_{AIN-JPT}$ between the Ainu and JPT, whereas the divergence between the JPT and CHB occurs at $T_{JPT-CHB}$ in the replacement model. The hybridization model shows that JPT originates from the admixture at $T_{ADMMIX}$ between the Ainu and CHB. Migration began between the Ainu and JPT after the admixture event.
Table 1. aMLs for the Three Models.

<table>
<thead>
<tr>
<th>Model</th>
<th>aML in loge-scaled</th>
<th>aBFa</th>
<th>aML in loge-scaled</th>
<th>aBFb</th>
</tr>
</thead>
<tbody>
<tr>
<td>No migration</td>
<td>−213.401</td>
<td>1/29.273</td>
<td>−216.175</td>
<td>1/63.463</td>
</tr>
<tr>
<td>Migration</td>
<td>−214.401</td>
<td>1/34.853</td>
<td>−214.780</td>
<td>1/17.537</td>
</tr>
</tbody>
</table>

aBF is calculated as the aML ratio of Model A or Model B to Model C.

A migration parameter between Ainu and JPT (m = 1.0E−5) is incorporated in the model after the divergence of Ainu and JPT (Model A) or after the divergence of JPT and CHB (Model B) and after the admixture between Ainu and CHB (Model C).

To select the best model among these hypotheses (fig. 1), we computed the approximated marginal likelihood (aML) in each model based on a kernel density estimate. As shown in table 1, the hybridization model showed the highest aML as ln(aML) = −212.025 (no migration) and −211.916 (migration), compared with the transformation model with ln(aML) = −215.401 and −215.467 and the replacement model with ln(aML) = −216.175 and −214.780. Next, we calculated the aBFs of the hybridization model with no migration and migration as 29.273 and 34.853 against the replacement model. These likelihoods were significantly higher in the models.

The choice of remnant population for Yayoi migrants may change the results of the model selection. The dual structure model hypothesized that the Yayoi came through the Korean Peninsula (Hanihara 1991). Indeed, Sato et al. (2014) showed that both the principal component analysis as well as the phylogenetic tree supported a closer relationship of Koreans to JPT than CHB. We replaced the gwSNP data from CHB with those from Koreans that was imputed, which increased the number of SNP sites which overlapped with the Ainu data (see Materials and Methods). The hybridization model best fits with the observed data under either of the no migration with ln(aML) = −213.646 and migration with ln(aML) = −213.614 models. These likelihoods were 44,547 and 54,664 times or 11.868 and 5.126 times higher than the transformation or replacement models. The aBF value was less significant under the replacement with migration model. This can be explained by the strong genetic similarity of Koreans with JPT that is likely to increase the admixture proportion from Koreans and to make the hybridization model indistinguishable from the replacement model.

We further investigated the hypothesis that the hybridization model is the best fit for the gwSNP data.

Parameter Estimation under the Complex Hybridization Models

The simplified hybridization model chosen from our model selection can be used as a starting point for further investigation into more realistic scenarios consisting of various prior conditions. Here, we hypothesized several complex hybridization models by including different sets of priors for three parameters (fig 2): 1) symmetric beta distributions, which is B(α, α), for the prior of admixture proportion (P) (supplementary fig S2, Supplementary Material online); 2) ancestral population sizes before the divergence between Ainu and CHB (NANC = 1,000, 2,500, 5,000, or NAIN which follows log-normal (LN; 5,000, 5,000²)); and 3) migration rates between Ainu and JPT (m = 0.0, 1.0E−5, 1.0E−3, or 1.0E−1). First, the hybridization model is characterized by the admixture event between Ainu and CHB; hence, the choice of hyper parameters for the prior of P can be crucial in estimating the posterior distribution. We compared MLs of six hyper parameters (α = 0.1, 0.25, 0.5, 1.0, 2.0, and 5.0) under NANC = NAIN and m = 0.0 (supplementary table S2, Supplementary Material online). Although differences in their aMLs were less significant in terms of aBFs, the U-shaped prior with α = 0.5 exhibited the highest likelihood among the hyper parameters. Second, NANC is likely to greatly contribute to shaping the amount of genetic variation among the three populations. We tested four conditions in which NANC was the same as NAIN or fixed at 1,000, 2,500, and 5,000 under m = 0.0 (supplementary table S3, Supplementary Material online). The fixed NANC = 1,000 showed a significantly higher likelihood than the other conditions. Third, continuous migration between Ainu and JPT after the admixture event can be important to defining shared genetic variation (supplementary table S4, Supplementary Material online). The likelihood of m = 1.0E−5 was significantly higher than m = 1.0E−3 and 1.0E−1. Meanwhile, the difference in the likelihoods between m = 0.0 and 1.0E−5 was less significant, as indicated in table 1.

Additional parameters of NANC and m jointly contribute to fitting the models to our observation. We evaluated the MLs of complex hybridization models based on three-by-three possible scenarios: NANC = NAIN, 1,000, or 2,500 and m = 0.0, 1.0E−5, 1.0E−3 (supplementary table S5, Supplementary Material online). The likelihoods were significantly higher in the model in which NANC = 1,000 and m = 0.0 than the other models, except for the model in which NANC = 1,000 and m = 1.0E−5. These results suggest that the models conditional on P ~ B(0.5, 0.5), NANC = 1,000, and m = 0.0 or 1.0E−5 are likely to be the best fit for the observed gwSNP data. The cranial variation of the Jomon skeletal series between northeast and southwest of the Japanese archipelago has been reported (Hanihara and Ishida 2009; Nakashima et al. 2010; Fukase et al. 2012a, 2012b). Here, we...
tested models postulating that the Jomon lineage diverged into two lineages at $T_{JOMON}$ generations ago: one lead to the Ainu population and the other experienced admixture resulting in a greater percentage of Yayoi ancestry (fig. 3). Furthermore, a significant gene flow ($m = 1.0E^{-5}$) between the Ainu and the JPT within the past 20 generations was incorporated into the models. Although there was no significant difference in MLs, all of the models with the population structure of the Jomon showed higher likelihoods than the models without the Jomon structure (table 2).

The highest likelihood model indicated an old divergence of Ainu and CHB populations an estimated 736 generations ago (95% confidence interval: 607–873), which was translated into 18,400 years ago based on the assumption of 25 years per generation. This model predicts the divergence of Jomon lineages occurred 598 generations ago (14,950 years ago) with admixture occurring 229 generations ago (5,725 years ago) with 36.0% (22.2–53.3%) of the admixture proportion from Jomon ancestry to the mainland Japanese.

NAIN was estimated at 6,219 (4,967–7,395) which is approximately a half of NJPT (10,207 (7,366–12,036). In addition, the other models, either with or without assuming the population structure of the Jomon, indicated that $T_{AIN-CHB}$ was old and $N_{AIN}$ was a half or a third of $N_{JPT}$. Meanwhile, the posterior estimates of $T_{ADMIX}$ are still older than 120 generations ago (3,000 years ago), which archaeological and anthropological studies have traditionally proposed. Although estimating the parameters by fixing $T_{ADMIX}$ at 120 revealed that the other parameters of $N_{AIN}$, $N_{JPT}$, and $P$ were scaled down to half of those estimates from the models with unfixed $T_{ADMIX}$, the likelihoods were similar between the models with unfixed and fixed $T_{ADMIX}$ (table 2 and supplementary table S6, Supplementary Material online). These results suggest that the estimates of $T_{ADMIX}$ strongly depend on the other parameters, implying the difficulty in estimating $T_{ADMIX}$ under the complex hybridization models. Our combined approaches of parameter estimation with model selection support the new model that the Jomon people were already diversified in the Japanese archipelago before the admixture event with the Yayoi occurred, and one of the Jomon lineages is a direct ancestor of the modern Japanese, yet the other lineage leads to the contemporary Ainu.

**Discussion**

The demographic inference based on the gwSNP data enables us to depict the peopling of Japan by combining our results with previous genetic, anthropological and archaeological studies. Our Bayesian inference based on genomic-scaled data provides strong evidence to support the hybridization model as the best fit model of Japanese demographic history.
In this study, we assume that the transformation and replacement models are nested under the hybridization model by restricting the ancestry proportion from the Jomon to “1” or “0” (fig. 1). Although the three models are not independent of each other, the relationship can be complicated because of the different degrees of freedom among the models. Model selection based on aBF is one of the most useful ways to distinguish between models based on the observed data. Although the estimates of admixture proportion depend on the complexity in the models (table 2), our results suggest that the patterns of genetic variation in gwSNPs are unlikely to be explained by those exceptional cases of the hybridization model.

The dual structure of Japanese populations as proposed in Hanihara (1991) hypothesized that the ancestries of the Jomon and the Yayoi people admixed within the Japanese archipelago, and further, the degree of admixture differs among local populations. There is archaeological evidence (Paleolithic artifacts) for an early wave of migration that brought the Jomon culture to Japan more than 15,000 years ago (Habu 2004; Craig et al. 2013). Our estimates of $T_{AIN-CHB}$ > 18,400 years ago are likely to be consistent with this date regardless of whether the models do or do not incorporate the Jomon structure and/or migration between the Ainu and JPT (table 2). The sea level gradually increased and the land bridges disappeared...

Table 2. Estimated Demographic Parameters under the Complex Hybridization Models.

<table>
<thead>
<tr>
<th>Parameter and Prior Distributiona</th>
<th>$N_{ANC} = 1,000$</th>
<th>$N_{ANC} = 1,000$; $T_{JOMON}$ is Incorporated into the Modelb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$m = 0.0$</td>
<td>$m = 1.0E - 5$</td>
</tr>
<tr>
<td>$N_{AIN}$ LN(5000, 5000$^a$)</td>
<td>6,472 (5,337–7,556)</td>
<td>6,364 (5,035–7,570)</td>
</tr>
<tr>
<td>$N_{JPT}$ LN(5000, 5000$^a$)</td>
<td>12,824 (9,628–15,398)</td>
<td>12,799 (9,560–15,778)</td>
</tr>
<tr>
<td>$T_{ADMX}$ LN(200, 200$^a$)</td>
<td>298 (226–358)</td>
<td>304 (230–374)</td>
</tr>
<tr>
<td>$T_{JOMON}$ LN(300, 300$^a$)</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>$T_{AIN-CHB}$ LN(400, 400$^a$)</td>
<td>748 (627–884)</td>
<td>758 (629–893)</td>
</tr>
<tr>
<td>$P$ B(0.5, 0.5)</td>
<td>36.1% (27.1–46.3%)</td>
<td>34.7% (25.7–44.9%)</td>
</tr>
<tr>
<td>ln(aML)</td>
<td>−209.374</td>
<td>−209.401</td>
</tr>
</tbody>
</table>

---

$^a$LN indicates a log normal distribution with a mean ($\mu$) and variance ($\mu^2$), and $^b$B indicates a beta distribution under a given parameter.

Genome-wide means of posterior estimates with 95% confidence intervals.
at the end of the last glacial maximum which was around 12,000 years ago (Ono 1995). Therefore, Jomon people have been isolated in the Japanese archipelago since the late Pleistocene.

The effective population size estimated for the Ainu is a half of or a third of that for the JPT (table 2). This could be related to the hunter-gatherer lifestyle of the Ainu having resulted in little to no population growth (Cavalli-Sforza et al. 1994; Watson et al. 1996). Therefore, substantial effects of genetic drift must have shaped the genetic diversity in the Ainu. For the admixture event, our direct estimates of $P$ between 35% and 45% are close to the previous estimate for maternal mtDNA contribution ($P = 35\%$) and autosomal data ($P = 23.1–39.5\%$) (Horai et al. 1996; He et al. 2012). Taking into account the small population size of the Ainu and the large size of the CHB population ($N_{CHB} = 29,204$), the lifestyle of the Yayoi, who were agricultural people, could reflect demographic characteristics. Our estimations of effective population size are compatible with the contrast between Jomon and Yayoi cultures of subsistence.

Our estimates of $T_{ADMIX}$ (table 2) are 2,000–3,000 years older than the age traditionally proposed by archaeological records and anthropological studies as the beginning of the Yayoi period. To further investigate this discrepancy, we assume that the Jomon people exhibited regional genetic heterogeneity (Hanihara and Ishida 2009; Nakashima et al. 2010; Fukase et al. 2012a, 2012b). In terms of MLs, the complex hybridization models that include the population structure of the Jomon are a better fit for the gwSNP data than the other models (table 2). In addition, the estimates of $T_{ADMIX}$ become younger if we take account of the effect of genetic heterogeneity in the Jomon lineages. Furthermore, a weak migration ($m = 1.0E–5$) followed by a strong migration in the past 20 generations ($m = 1.0E–3$) is of higher likelihood than the no migration model. Meanwhile, the pattern of migration did not change the likelihood of the model postulating a single Jomon lineage (data not shown). Therefore, the joint effects of the population structure of the Jomon and migration seem to partially improve the fitting of models to the gwSNP data. Another possibility that we have to consider is an admixture of Jomon with non-Yayoi populations prior and/or subsequent to the Yayoi admixture; because we assume the Jomon people have been isolated in the Japanese archipelago without any gene flow from the continent, a plausible admixture between the Ainu and the Okhotsk or another unknown people would makes the time estimate older (Tajima et al. 2004; Sato et al. 2007; Hanihara and Ishida 2009; Kaburagi et al. 2010; Jinam et al. 2012). Alternatively, gene flow between the Jomon and Yayoi people prior to cultural diffusion might also offer an explanation of the time estimate. This is a startling interpretation when compared with conventional anthropological and archaeological interpretations, but it might be worth considering a possible time lag between the Yayoi migration and the Yayoi culture.

Although the origin and the demographic history of the Ainu remain unclear, unique genetic characteristics of the Ainu might represent ancestry-specific portion derived from the Jomon people (Tajima et al. 2004; Sato et al. 2007; Kaburagi et al. 2010; Jinam et al. 2012). Our inference shows that the Ainu and the CHB share an ancestral population that dates to the late Pleistocene; archaeological evidence shows the peopling of the Japanese archipelago began during this time frame. From these results we conclude that the ancestral people of the Ainu (or some of their ancestors) established the Jomon culture sometime at the end of Pleistocene, while a part of the ancestral population of CHB migrated to Japan and started admixing with the Jomon people about 5,000–6,000 years ago.

Our results strongly support the dual structure model for the peopling in the Japanese archipelago produced through hybridization between the Jomon and Yayoi people. Furthermore, we propose a more complicated model in which the Jomon people have diversified on the Japanese archipelago because they were isolated from the continent around 18,000 years ago. Admixture did occur between the Jomon and Yayoi people, yet some of the Jomon people still kept their own lineages without being strongly affected by the gene flow from the Yayoi people. The older divergence time of the Jomon and Yayoi lineages is compatible with the Hanihara’s hypothesis (1991), which claims the Jomon people are direct descendants of late-Paleolithic populations somewhere in Southeast Asia, whereas the Yayoi people were a population adapted to a cold climate in Northeast Asia. As the first step for resolving the peopling of Japan statistically and quantitatively, we simplified the models to make it possible to test the three hypotheses established from the previous studies. Then, we tested complex scenarios under the hybridization model to fit them with the gwSNP data. However, there should be other possible scenarios that we have to consider, and further studies taking account of realistic scenarios should be necessary to untangle a true evolutionary history of Japanese. To this end, our results can be used as a starting point to test more complicated hypotheses using a genomic-scaled data. In particular, the origins of the Jomon and Yayoi people and their demographic history, including the source of gene flow to Ainu, still remain unresolved. Future studies, including ancient genome analysis of Jomon and Yayoi skeletal remains as well as population-scaled gwSNP and genomic sequence analysis for east and northeast Asians will bring new insights into the complex demography in Japanese populations.

Materials and Methods

gwSNP Data

We used the same data as Jinam et al. (2012), which include gwSNPs in the Ainu ($2n = 72$), JPT ($2n = 90$), and CHB ($2n = 84$) groups. Briefly, gwSNPs were filtered with a confidence score $>0.008$, a call rate $<95\%$, and a deviation from Hardy–Weinberg equilibrium ($P < 0.001$), which may be caused by genotyping errors. A total of 641,314 SNPs passed data filtering and quality checks. A detailed procedure is described in Jinam et al. (2012).

We also used the gwSNP data of 221 Korean individuals (Haga et al. 2013) as an example of a remnant population of Yayoi migrants, instead of CHB. Genotype imputation was
performed for the Koreans with BEAGLE 3.0 (Browning and Browning 2009) using HapMap JPT and CHB as the reference panels because Korean samples were genotyped with an Illumina HumanOmniExpress BeadChip; in contrast, the Ainu samples were genotyped with an Affymetrix genome-wide Human SNP Array 6.0. The imputation accuracy was confirmed by the following procedures. From the Korean data determined using the Illumina platform, approximately 5,000 SNPs were chosen at random and were set to missing. We applied genotype imputation to these missing data and compared imputed genotypes to experimentally determined genotypes. As a result, approximately 98% of the SNPs showed an imputation accuracy rate of greater than 95%.

Calculating 3D-HFS

The data used to calculate summary statistics in kernel-ABC were 460,013 SNPs that were polymorphic in all three populations: Ainu, JPT, and CHB; these were defined as "shared SNPs." Next, we estimated haplotypes to summarize the data as HFS because this can account for a pattern of recombination at a given genomic segment. We calculated summary statistics for the gwSNP data as follows: first, the gwSNP data were divided into subsets of SNPs in which the genetic distance between the first and last SNPs in terms of their physical positions was 0.15 cM as calculated from the Haldane’s mapping function, \(0.5 \times (1 - e^{-2x})\) where \(x\) is the map distance in centimorgans. The HapMap Release 22 (NCBI 36) recombination map was downloaded from the HapMap database (http://hapmap.ncbi.nlm.nih.gov/, last accessed February 2013; Frazer et al. 2007). Each subset was used to represent the haplotype diversity in a given genomic segment among the three populations. The total number of segments (i.e., subsets) was 18,026, where the mean number of SNPs in a segment was 26 with a standard deviation (SD) of 19 SNPs; the segment interval had a mean of 138,084 bp with an SD of 271,800 bp. Second, we estimated haplotypes from each subset of SNPs for the Ainu, JPT, or CHB groups, respectively, using the PHASE v.2.1 program (Stephens et al. 2001). Third, based on the makeup of haplotypes among the Ainu, JPT, and CHB we generated a 3D-HFS (\(5 \times 5 \times 5\) array) in which each element \((\delta_{ij})\) represented the number of haplotypes that were observed at \(i, j, h\) chromosomes in the Ainu, JPT, and CHB groups. The bin size of haplotype frequency was defined as "10" and the interval of the last bin ranged from 30 to the maximum number of haplotypes in a population as follows: "0, 1–10, 11–20, 21–30, 31–72" in Ainu, "0, 1–10, 11–20, 21–30, 31–90" in JPT, and "0, 1–10, 11–20, 21–30, 31–84" in CHB (supplementary fig. S1, Supplementary Material online). Each segment represented a sparse array of 3D-HFS; we then merged 100 arrays into one 3D-HFS (\(S_{3D-HFS}\)) by randomly choosing 100 segments. Then, we obtained 180 summary statistics \(S_{3D-HFS}(1), S_{3D-HFS}(2)\ldots S_{3D-HFS}(180)\) from the 18,026 segments and estimated posterior means of parameters using kernel-ABC based on each of \(S_{3D-HFS}(i)\) \((i = 1, 2, \ldots, 180)\).

Similarly, we calculated summary statistics for the gwSNP data from the Korean group (Sato et al. 2014). First, we randomly chose the same number of samples as CHB \((n = 42)\) from the data. The total number of SNPs that were polymorphic in all the three populations was 431,270. Second, we divided the gwSNP data into 17,447 windows with a genetic distance of 0.15 cM. Then, we phased the SNPs included in each window and calculated 3D-HFS for the window. Finally, we combined the summary statistics across 100 randomly chosen windows into one 3D-HFS, so the total sets were 174 \(S_{3D-HFS}\) in Koreans \((j = 1, 2, \ldots, 174)\).

Generating Simulation Data

We generated three sets of simulation data under different demographic models. The first set was used to choose the best fitting model from the (a) transformation, (b) replacement, and (c) hybridization models (fig. 1). Parameters were the effective population sizes in Ainu \(N_{AIN}\) and JPT \(N_{JPT}\), divergence times between Ainu and CHB \(T_{AIN-CHB}\), between Ainu and JPT \(T_{AIN-JPT}\), and between JPT and CHB \(T_{JPT-CHB}\), and the admixture proportions from the Jomon ancestry \(P\). Three parameters \(N_{AIN}, N_{JPT}, T_{JPT-CHB}\) were common to the three models, while the remaining parameters were specific to each model \(T_{AIN-JPT}\) in the transformation model; \(T_{JPT-CHB}\) in the replacement model; \(T_{ADMIX}\) and \(P\) in the hybridization model. We assumed that the ancestral population size of Ainu and JPT or among three populations was equal to \(N_{AIN}\) and that the initial population size in CHB was 11,154 \(N_{CHB}\), which expanded to 29,204 \(N_{CHB-exp}\) at \(T_{AIN-JPT}\), \(T_{JPT-CHB}\), or \(T_{ADMIX}\) respectively (Gravel et al. 2011). Furthermore, we hypothesized migration between Ainu and JPT \((m = 1.0E-5)\) that started from \(T_{AIN-JPT}\), \(T_{JPT-CHB}\), or \(T_{ADMIX}\) and continued to the present. Prior distributions were given as a symmetric beta distribution for the ancestry proportion from Jomon, \(P \sim B(1,0,1,0)\), and an LN distribution for the population sizes and times in which variance was a square of the mean \((\mu)\) as follows: \(N_{AIN}\) or \(N_{JPT} \sim LN(\mu = 5,000, \mu^2 = 5,000^2)\), \(T_{ADMIX} \sim LN(400, 400^2)\), \(T_{JPT-CHB} \sim LN(200, 200^2)\).

The second set was generated under the complex hybridization models including different hyper parameters \(\alpha\) of priors for \(P\), ancestral population sizes of all three populations \(N_{ANCE}\), and migration rates \(m\) between the Ainu and JPT (fig. 2). The prior conditions for the other parameters were the same as the hybridization model described in figure 1. The symmetric beta distributions were set as \(\alpha = 0.1, 0.25, 0.5, 1.0, 2.0, \) or 5.0 (supplementary fig. S2, Supplementary Material online). We also assumed that \(N_{ANCE}\) was fixed at 1,000, 2,500, 5,000, or the same as \(N_{AIN}\) following \(LN(5,000, 5,000^2)\). Various migration rates between the Ainu and JPT \((m = 1.0E-5, 1.0E-3, \) and \(1.0E-1)\) were incorporated into the models after the admixture event.

The third set was based on the model that included the additional parameter, \(J_{JOMON}\), that defined the time of divergence in Jomon lineages (fig. 3). In this model, we assumed that Jomon people genetically differentiated within the Japanese archipelago and diverged into two lineages: one led to Ainu and the other contributed to the admixture as
the Jomon ancestry. This event happened between \( T_{\text{AIN-CHB}} \) and \( T_{\text{ADMix}} \), so \( T_{\text{Jomon}} \) was sampled from \( LN(300, 300) \). We also incorporated two different patterns of migration between Ainu and JPT into this model: one was that a weak migration \( (m = 1.0E-5) \) occurred from \( T_{\text{ADMix}} \) to the present, and the other was that the migration rate was assumed to be \( 1.0E-5 \) but the rate changed 20 generations ago to \( m = 1.0E-3 \).

We performed coalescent simulations under all sets of demographic models. Simulations were performed using the program package ms, which generates samples from the coalescent model (Hudson 2002). Because our gwSNP data were ascertained by their frequencies, we corrected the ascertainment bias based on observed site frequency spectra (SFS) without taking into account the ancestral/derived state. The folded SFS in the gwSNP data included fewer rare SNPs (minor allele frequency, MAF \( \leq 5\% \)) and more common SNPs (MAF \( > 5\% \)) than the folded SFS in simulation data which showed more rare SNPs and fewer common SNPs as we expected under neutral conditions. We generated the 3D-folded SFS (\( 6 \times 6 \times 6 \) array) from 460,013 SNPs, which represented the frequency of shared SNPs among the Ainu, JPT, and CHB. Bin sizes for six bins were as follows: \( \text{Ainu: } "1-2 \text{ and } 70-71", "3-4 \text{ and } 68-69", "5-6 \text{ and } 66-67", "7-8 \text{ and } 64-65", "9-10 \text{ and } 62-63", \text{ and } "11-61"; \) JPT: \( "1-2 \text{ and } 88-89", "3-4 \text{ and } 86-87", "5-6 \text{ and } 84-85", "7-8 \text{ and } 82-83", "9-10 \text{ and } 80-81", \text{ and } "11-79"; \) CHB: \( "1-2 \text{ and } 82-83", "3-4 \text{ and } 80-81", "5-6 \text{ and } 78-79", "7-8 \text{ and } 76-77", "9-10 \) and 74-75, and "11-73". Then, we used the observed 3D-folded SFS as a filter to fit the simulated data with the observed data. Our simulation procedures were as follows:

for \( i = 1 \) to \( n \) \((\text{the total number of simulations})\) do

1. Sample the parameters \( \theta_i \) from the prior distributions.

for \( j = 1 \) to 100 \((\text{the number of genomic segments})\) do

2. Sample a genomic size \( (G_j) \) of a region from \( LN(138084, 2718002) \) whose mean and variance were calculated from the observed gwSNP data. 

3. Generate simulated data using \( \theta_i \) and \( G_j \). The population mutation rate \( (\theta) \) and recombination rate \( (\rho) \) are given as \( \theta = 4 \times N_{\text{AIN}} \times G_j \times 2.5 \times 10^{-8} \text{ bp/site/} \text{ generation} \text{ (Nachman and Crowell 2000)} \text{ and } \rho = 4 \times N_{\text{AIN}} \times 0.0015 \).

4. Calculate the 3D-folded SFS for the simulated data and correct the spectrum based on the observed 3D-folded SFS (see above). Since there was no ascertainment bias in the simulation data, we randomly chose a given number of SNPs in each bin to fit the ratios of SNPs in the simulated 3D-folded SFS with those in the observed 3D-folded SFS.

5. Calculate summary statistics (3D-HFS) for the corrected simulated data.

6. Go to step 2 keeping \( j = j \) if the number of common SNPs among the Ainu, JPT, and CHB is "0" or "1". Otherwise, go to the next loop \( (j = j + 1) \).

end for

We generated \( n = 1,050,000 \) simulation samples for the first set of models and \( n = 100,000 \) samples for the second and third sets (figs. 2 and 3). These simulation samples were used in the subsequent Bayesian inference.

Model Selection by aBF

Based on Osada et al. (2013), we selected the model that best fit the gwSNP data. We chose the bandwidth of the kernel using cross-validation based on 50,000 simulation samples in the first set (Osada et al. 2013). Then, we computed kernel density estimates of aMLs using the remaining one million samples in the first set (fig. 1) and 100,000 samples in the second and third sets (figs. 2 and 3). Based on the observed summary statistics \((S_{3D-HFS}^{(1)}, S_{3D-HFS}^{(2)}, \ldots S_{3D-HFS}^{(180)})\), we estimated \((P(S_{3D-HFS}^{(1})|M), P(S_{3D-HFS}^{(2})|M), \ldots P(S_{3D-HFS}^{(180})|M), \) and averaged them.

Parameter Estimation by Kernel-ABC

We estimated the posterior means of parameters using kernel-ABC with the observed summary statistics \((S_{3D-HFS}^{(1)}, S_{3D-HFS}^{(2)}, \ldots S_{3D-HFS}^{(180)})\) and the simulated summary statistics \((S_{3D-HFS}^{(i})|i = 1, \ldots, n)\). Details of this procedure are given in the previous study (Nakagome, Fukumizu, et al. 2013). Genome-wide distributions of posterior means are shown in table 2.

Supplementary Material

Supplementary figures S1 and S2 and tables S1–S6 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank two anonymous reviewers for valuable comments. S.N. and T.S were supported as a Grant-in-Aid for the Japan Society for the Promotion of Science (JSPS) Research Fellow (24–3234) and as a Grant-in-Aid for the JSPS Research Fellow (25–2544). This study was supported in part by Grant-in-Aid for the JSPS Nos. 21370108 and 24370099 to H.O., No. 22687023 to R.K., No. 22370087 to J.O., and No. 22470099 to T.H. The member list of the Asian DNA Repository Consortium: Keichi Omoto, Katsushi Tokunaga, Naruya Saitou, Shoji Kawamura, Hideyuki Tanabe, Kazuo Uematsu, Atsushi Tajima, Shuhei Mano, Hiroki Oota, Naoki Osada, Ryosuke Kimura, Toshimitsu Yamamoto, Jun Ohashi, Yumiko Suto, Nao Nishida, Ryuichi Sakate, Hiromi Sawai, Shigeki Nakagome (March 8, 2014).

References


