Introduction

The origin of the Japanese people—a brief introduction

The Japanese archipelago is at the eastern end of East Asia, and has a current population of approximately 127 million people. The history of Yaponesians, a term referring to people prehistorically and historically living in the archipelago (Yaponesian Genome Project Managing Group, 2020), is of long-standing interest among anthropologists, archeologists, linguists, and historians studying the history of Japan. While the ‘dual-structure’ model proposed by Hanihara in 1991 has been considered the primary working hypothesis for three decades, recent advances in DNA typing and sequencing technologies provide an unprecedented amount of present-day and ancient human nuclear genome data, which enable us to refine or extend the dual-structure model. In this review, we summarize recent genome sequencing efforts of present-day and ancient people in Asia, mostly focusing on East Asia, and we discuss the possible migration routes and admixture patterns of Japanese ancestors. We also report on a meta-analysis we performed by compiling publicly available datasets to clarify the genetic relationships of present-day and ancient Japanese populations with surrounding populations. Because the ancient genetic data from the Japanese archipelago have not yet been fully analyzed, we have to corroborate models of prehistoric human movement using not only new genetic data but also linguistic and archeological data to reconstruct a more comprehensive history of the Japanese people.

Key words: Japanese, East Asia, ancient genome, Jomon, Yayoi
For over three decades, revealing the history of human populations using genetic data relied mostly on uniparental genetic markers, i.e. mitochondrial genomes and Y chromosomes. However, recent advances in DNA typing and sequencing technologies have provided an unprecedented amount of human nuclear genome data. Because nuclear genomes recombine with each other and contain the information of multiple genealogies, we can obtain information for millions of genealogies from genome-wide genotyping data. The study of nuclear genomes is therefore able to reveal the history of populations on a granular level, even with a relatively small number of samples. In addition, it has become practical to analyze genome-wide polymorphism data of ancient people, even including archaic Homo species such as Neanderthal and Denisovan, providing many findings about prehistoric human migration (Green et al., 2010; Prüfer et al., 2014; Reich et al., 2010). Within just a few years, several key papers analyzed the genome-wide polymorphism data of ancient samples from Southeast, East, and Northeast Asia, including many Paleolithic and Neolithic specimens, and provided numerous novel insights into prehistoric human migration around the Japanese archipelago (Gakuhr et al., 2020; Kanzawa-Kiriyama et al., 2019; McColl et al., 2018; Ning et al., 2020; Wang et al., 2020a, b; Yang et al., 2020).

In this review, we concisely summarize recent genome sequencing studies of present-day and ancient people in Asia, mostly focusing on East Asia, and discuss the possible migration routes and admixture patterns of Japanese ancestors. We draw readers’ attention also to recent review papers from different perspectives by de Boar et al. (2020) and Zhang and Fu (2020) for further study. We note that the scenarios presented here are somewhat simplified. The actual migration history of humans would have been much more complex than can be described by a simple narrative.

In addition to the concise review, we report on a meta-analysis we performed by compiling publicly available datasets and we clarify the genetic relationships of present-day and ancient Japanese populations with surrounding populations. Because many ancient genome studies have used similar datasets (or reference panels) for their analysis, it is a good opportunity to compare analytical results using different datasets for the validation of previous findings. Here we compile high-coverage whole-genome sequencing (WGS) data of present-day humans and combine them with the variant information of previously published ancient genomes (see Methods). We focus on the analysis of autosomal genome-wide single nucleotide variant (SNV) datasets and do not discuss the details of research using mitochondrial and Y-chromosomal sequences except where necessary. We refer readers to recent papers summarizing the pattern of mitochondrial and Y-chromosomal diversity in Southeast, East, and Northeast Asia for related discussion (Kivisild et al., 2002; Stoneking and Delfin, 2010; Tanaka et al., 2004).

### Paleolithic population history in East Eurasia

How anatomically modern humans migrated from Africa to East Asia is a key question for understanding the establishment of present-day Asian populations. Studies of genome diversity in present-day and ancient Asians point to two major routes of prehistoric migration to East Eurasia: the southern and northern routes relative to the Himalayan mountains. Whether the migration through the southern route occurred once or more is under debate both in genetic and in archeological research (Stringer and Andrews, 1988; Macaulay et al., 2005; Malaspinas et al., 2016; Bae et al., 2017; Lipson and Reich, 2017). Along the northern route, it is likely that West and East Eurasian populations were well interconnected during the prehistoric and historic periods via the steppe regions and the Central Siberian plateau (Xu et al., 2008; de Barros Damgaard et al., 2018; Sikora et al., 2019; Wang et al., 2020a). In this section, we summarize how anatomically modern humans migrated out of Africa to East Eurasia and how Paleolithic population movement and admixture formed the current genetic architecture of East and Northeast Asians.

#### Upper Paleolithic southern migration route

Via the southern route, ancestors of current Asian populations reached Southeast Asia and a part of Oceania around 70000–50000 years ago, probably through a coastal dispersal route (Bae et al., 2017). At the time, the Eurasian continent was in the Würm glacial stage. The Indochina peninsu-
la, the Malay peninsula, and many islands, including Borneo, Java, Sumatra, and the Philippines, formed a landmass called Sundaland, which was connected to the Eurasian continent (Hall and Morley, 2004). Among descendants of the migrants, Papuans, aboriginal Australians, and Philippine Negritos show a larger amount of introgression of genomic fragments with Denisovan ancestry than the other East Asians, including Andamanese (Onge) and Malaysian Negritos (Jehai). This suggests that archaic humans related to the Denisovan lived somewhere in (or nearby) Southeast Asia (Reich et al., 2011; Jinam et al., 2013, 2017). When East Asian populations are clustered as a tree with no assumption of migration between populations, most East and Northeast Asian populations form a single cluster, with Papuans, Australians, and Onge people at root positions, suggesting that the Southern route predominantly shaped the genetic features of present-day East and Northeast Asian populations (HUGO Pan-Asian SNP Consortium, 2009; Mallick et al., 2016; Lipson and Reich, 2017;).

**Upper Paleolithic northern migration route**

The oldest samples providing the genetic evidence of the northern migration route come from a high-coverage genome sequence of individuals excavated from the Yana RHS site in northeastern Siberia (Figure 2), which is about 31600 years old (Sikora et al., 2019). A wide range of artifacts, including bonecrafts of wooly rhinoceros and mammoths, were excavated at the site (Pitulko et al., 2004). The analysis of genome sequences showed that the samples were deeply diverged from most present-day East Asians and more closely related to present-day Europeans, suggesting that the population reached the area through a route different to the southern route. A 24000-year-old individual excavated near Lake Baikal (Figure 2), also known as the Mal’ta boy, and 17000-year-old individuals from the Afontova Gora II site (Afontova Gora 2 and 3) showed similar genetic features to the Yana individuals (Raghavan et al., 2014; Fu et al., 2016; Sikora et al., 2019). Interestingly, genetic data suggested that Yana individuals received a large amount of gene flow from the East Asian lineage (Sikora et al., 2019; Yang et al., 2020). A similar tie between early Western Eurasians and East Asians was also reported by Lipson and Reich (Lipson and Reich, 2017). In addition, Sikora et al. (2019) used an admixture graph analysis to model the Mal’ta individual (MA-1) as descendants of a population split from the ancestral Yana, with a minor contribution from Late Paleolithic hunter-gatherers in Caucasus after the split (Sikora et al., 2019). Using our dataset, we confirmed that Yana individuals shared more genetic drift with East Asian populations than present-day Europeans do, and MA-1 shared more genetic drift with present-day Europeans than Yana individuals did (Supplementary Table 1). The populations of early migrants to Northeast Asia, represented by Yana, Mal’ta, and Afontova Gora individuals, are labeled as Ancient North Siberians (Sikora et al., 2019), although we do not yet have a complete picture of this population.

Around 20000 years ago (the time estimate has a broad range, probably between 25000 and 15000 years ago (Sikora et al., 2019))...
Ancient North Siberians admixed with the people who took the southern route, at least for the second time. The relatively high-coverage genome from a 9800-year-old Kolyma individual (Kolyma1, Figure 2) was closely related to present-day East and Northeast Asians, but also had a significant amount of gene flow from Ancient North Siberians. Interestingly, the Kolyma individual showed a stronger genetic affinity to the Mal’ta than to the Yana individuals. This observation is consistent with the archeological evidence that shows a discontinuity in human occupation at Yana for about 10000 years, meaning that genetic components specific to Yana did not directly contribute to any present-day populations. The Kolyma individual was genetically close to present-day ethnic groups in the Chukchi and Kamchatka regions and to Native Americans, but the shared ancestry with Ancient North Siberians was more pronounced in the Native Americans than in the Kolyma individual (Sikora et al. 2019). Using a maximum composite likelihood method, they estimated the divergence time of Kolyma and Native American lineages as around 24000 years ago and postulated that these lineages received gene flow from Ancient North Siberians after their split (Sikora et al., 2019).

The above observation indicates that the ancestral population of the Kolyma people and Native Americans independently received gene flow from Ancient North Siberians. When and how did the admixture take place? Studies of mitochondrial and autosomal genomes from ancient and present-day Alaskans and Native Americans suggest that the ancestral populations for all Native Americans were isolated in the Beringia regions during the Last Glacial Maximum (LGM) from 25000 to 15000 years ago and spread to the New World after that. This is referred to as the standstill model (Bonatto and Salzano, 1997; Tamm et al., 2007; Hoffecker et al., 2016; Moreno-Mayar et al., 2018). Under this model, we are able to assume that the admixture between a population from the southern route and Ancient North Siberians occurred around 25000 years ago. It might be reasonable to assume that the admixture of populations occurred at refugia during the LGM; however, genetic data alone do not indicate the location of any actual refugium. Sikora et al. (2019) performed a paleoclimate reconstruction to find potential regions for such refugia during the LGM and showed southern Beringia could potentially have been the one.

**Paleolithic population history in Southeast and East Asia**

After reaching Sundaland, some populations that had taken southern route migrated further northward. Fossil remains of anatomically modern *Homo sapiens* appeared as early as 40000 years ago in several places in China, such as Zhoukoudian Upper Cave and Tianyuan Cave (Figure 2) (Kaifu and Fujita, 2012; Shelach-Lavi, 2015). A partial genome sequence of the Tianyuan individual was determined by Fu et al. (2013). Studies showed the Tianyuan lineage at the root of all East and Northeast Asian populations (Fu et al., 2013; Yang et al., 2017); however, mixed results have been obtained when trying to determine whether the lineage of the Tianyuan split from ancestors was before or after the split of the early Asians such as Papuans and Negritos (Fu et al., 2013; Yang et al., 2017; McColl et al., 2018; Kim et al., 2020). We performed a meta-analysis with $f_{st}$ statistics to confirm the genetic relationship of the Tianyuan individual to other East Asian populations. $f_{st}$ statistics with a configuration of $f_{st}(W, X; Y, Z)$ measure the amount of shared genetic drift between W and Y, or X and Z. If we fix the population W as the outgroup population that did not experience recent gene flow with X, Y, and Z, then positive and negative $f_{st}$ values indicate the gene flow between X and Z, and between X and Y, respectively. Throughout this article, we fix the Yoruba population as the W. Here we present the results fixing the Dai population as the Z and selecting various ancient and present-day Southeast, East, and Northeast Asian populations as the Y in Supplementary Tables 2 and 3. Most Southeast, East, and Northeast Asian populations, including Jomon, are nearly equally distant from the Tianyuan individual, supporting the hypothesis that the Tianyuan population are diverged from the lineage basal to all East and Northeast Asians. The Tianyuan individual was significantly more closely related to the Dai population than to the Jehai, Kusunda, and Papuan populations (Supplementary Table 3). However, a further $f_{st}$ analysis showed that the Tianyuan individual would be placed to more basal position of Asians than the Kusunda and Jehai, as the deviation of $f_{st}$ statistics is considerably smaller in $f_{st}$(Yoruba, Tianyuan; Kusunda/Jehai, Dai) than in $f_{st}$(Yoruba, Kusunda/Jehai; Tianyuan, Dai) (Supplementary Table 3). The reason why the Tianyuan individual is more closely related to the Dai population than to the Kusunda/Jehai populations is not clear, but might be due to the Kusunda and Jehai genomes containing some genetic component derived from relatively distant populations. For example, we found that the Kusunda population is more closely related to Europeans (CEU, Utah residents with Northern and Western European ancestry) than the Dai population is ($Z = -3.06$, Supplementary Table 3). Such unknown bias would make it difficult to define basal populations by presenting a tree-shaped population relationship. In contrast, the Papuan population showed only marginal deviation of $f_{st}$(Yoruba, Papuan; Tianyuan, Dai) value ($Z = 2.26$), supporting the hypothesis that the population related to the Tianyuan individual diverged after the split of the Papuan population (Supplementary Table 3).

Interestingly, previous studies showed that the Tianyuan individual showed some genetic similarity to a 35000-year-old ancient West Eurasian individual (GoyetQ116-1) and some specific native South Americans (Surui) (Yang et al., 2017). In addition, Sikora et al. (2019) presented data showing that the Yana samples have substantially more genetic affinity to the Tianyuan individual than to other early Western hunter-gatherers (Sikora et al., 2019), indicating the early contact between the populations passing along the southern and northern routes. A similar pattern was also shown by Yang et al. (2020).

Several lineages, such as Jehai, Kusunda, and Jomon, form basal populations to all East and Northeast Asian populations. These populations and a population related to the Tianyuan, seem to have spread to widely across South, Southeast, and East Asia. In addition to these populations, a recent ancient-genome sequencing of Hoabinhian hunter-gatherers from Laos and Malaysia (~5800–2300 BCE) showed that these hunter-gatherers were genetically close to
Andamanese Negritos (Onge) and admixed with populations from southern China in the Neolithic era (McColl et al., 2018). In addition, the Hoabinhian individuals showed a statistically significant genetic relationship with a Jomon individual from Ikawazu demonstrating the southern-route origin of the Jomon people (McColl et al., 2018; Gakuhari et al., 2020; but also see de Boer et al., 2020 for a different result).

**Neolithic population history in East Eurasia**

**Neolithic population history in Southeast Asia**

As shown in the previous section, a population split from Tianyuan became the major ancestral root of all present-day populations in East and Northeast Asia. In a principal-component analysis (PCA) plot including a wide-range of Southeast, East, and Northeast Asian populations presented by Yang et al. (2020), two major genetic clines were observed: a south-to-north genetic cline and a Han-to-Tibetan genetic cline. Figure 3 shows our PCA plot of East and Northeast Asians with projections of several ancient individuals in East Asia. Note that we focus here on the East Asian populations and do not include Tibetan-related populations, because our primary dataset did not contain many Tibetan-related individuals. In the plot, we can see Southeast Asian populations such as Cambodians, Dai people, and Kinh Vietnamese are placed on the southernmost side of the cline. In southern China, Neolithic populations lived around 7500 BCE in the Middle and Lower Yangtze River basin and rice-farming became prevalent as early as 5000 BCE (Bellwood, 2004; Shelach-Lavi, 2015). Recent studies of ancient genomes mostly support the scenario where southward movements of rice-farming Neolithic people from southern China greatly influenced the current genetic architecture of Southeast Asians, Melanesians, Micronesians, and Polynesians (Lipson et al., 2018a; McColl et al., 2018; Yang et al., 2020). The pattern is also consistent with the distribution of diverse languages spoken in the area and provides examples of how human migration in the Late Neolithic era can explain the current distribution of various languages (Diamond and Bellwood, 2003; Bellwood, 2004).

The best-known example of migration patterns linking to linguistic patterns is the case of the Austronesian-speaking Neolithic population’s expansion to Micronesia, Melanesia, and Polynesia through the island of Taiwan, which has been supported by both archeological and linguistic data (Gray and Jordan, 2000; Diamond and Bellwood, 2003; Bellwood, 2004). The expansion was estimated to start around 3000 BCE, arrive at the Bismarck Archipelago around 1000 BCE (Sheppard et al., 2015), and finally reach the eastern part of the Polynesian islands by 1000 AD. The migration process is also well reflected by the genetic data of present-day Polynesians (Wollstein et al., 2010; Ko et al., 2014; Lipson et al., 2014). The ancient-genome sequences of Vanuatu (~1100–700 BCE) and Tonga (~700–300 BCE) populations revealed a detailed but complex migration history from New Guinea.
The peopling of the Tibetan Plateau is particularly important in the study of the origin of the Japanese people, because previous genetic studies highlighted a particular Y-chromosomal haplogroup (D-M174 or D1) that was shared among Jomon, Tibetan, and Andamanese people (Underhill et al., 2001; Thangaraj et al., 2003; Hammer et al., 2006; Shi et al., 2008). Studies of mitochondrial haplotypes proposed a dual-structure model of present-day Tibetans, parallel to the situation in the Japanese archipelago (Qin et al., 2010; Liu et al., 2020). Among the four centers, the people in the Amur River basin showed the smallest amount of change in their genetic components at a later stage with or without gene flow.

The origin of Tibet highlanders

The peopling of the Tibetan Plateau is particularly important in the study of the origin of the Japanese people, because previous genetic studies highlighted a particular Y-chromosomal haplogroup (D-M174 or D1) that was shared among Jomon, Tibetan, and Andamanese people (Underhill et al., 2001; Thangaraj et al., 2003; Hammer et al., 2006; Shi et al., 2008). Studies of mitochondrial haplotypes proposed a dual-structure model of present-day Tibetans, parallel to the situation in the Japanese archipelago (Qin et al., 2010; Li et al., 2015). Autosomal studies also indicated the Tibetan populations consist of the admixture of ancient Paleolithic people and new Neolithic migrants; in particular, recent ancient-genome sequencing showed that the Neolithic migrants were closely related to Neolithic northern East Asians in the Yellow River basin rather than the southern East Asians (Lu et al., 2016; Wang et al., 2020b; Yang et al., 2020). The pattern is consistent with a study proposing a northern Chinese origin of Sino-Tibetan languages from linguistic data (Zhang et al., 2019), although the inferred pattern of gene flow from northern East Asia might have been very complex, including multiple migration waves (Lu et al., 2016; Wang et al., 2020a).

The origin of the Paleolithic Tibetan population, however, remains unclear. Jeong et al. (2016) showed a deeply shared ancestry between Sherpa and Ainu people using single-nucleotide polymorphism (SNP) chip data, consistent with the Y-chromosomal pattern. In contrast, Kanzawa-Kiriyama et al. (2019) did not detect a significant genetic tie between Sherpa and Jomon people. A study by Yang et al. (2020) using qpAdm analysis showed that Late Neolithic (Chokhpani) and present-day Tibetans can be explained by the admixture between northern Japanese and Tibetans. A recent unpublished study, analyzing many present-day and ancient genome sequences in East Asia, suggested Hoabinhian/Onge-related ancestry of Paleolithic Tibetans (Wang et al., 2020a). Whether a basal population for present-day Tibetans harbored a shared lineage with the ancestors of Jomon people is not totally clear, but the above studies corroborate the finding that present-day Tibetans have genetic components of a deeply diverged basal Asian population.

Neolithic population history in East and Northeast Asia

The earliest complex civil communities in East Asia developed in present-day China. Recent studies of ancient DNA sequencing in Neolithic East and Northeast Asia revealed insights into how the population structure in this region shifted (Ning et al., 2020; Wang et al., 2020a; Yang et al., 2020). Across the Northeast to South China and to Primorsky in Russia, four major centers for Neolithic cultures are recognized: the Amur River basin, the West Liao River (a tributary of the Liao River) basin, the Yellow River basin, and the Yangtze River basin (Figure 2). Although these four centers developed with their own unique features, archaeological evidence implies that they were interconnected and influenced each other (Shelach-Lavi, 2015).

Among the four centers, the people in the Amur River basin showed the smallest amount of change in their genetic features during the Neolithic/Bronze Age/Iron Age periods. It has been shown that Early to Late Neolithic individuals in the inland portion of the Amur River basin, as well as individuals from the coastal Devil’s Gate (Figure 2), are genetically highly similar to present-day Tungusic-speaking people, such as Ulchi, Oroqen, and Hezhen (Nanai) people (Siska et al., 2017; Ning et al., 2020). However, the effect from Paleo-Siberians during the Paleolithic period is not negligible. Present-day people in the Amur River basin share slight but significant genetic drift with Paleo-Siberians. Using a composite likelihood estimation based on the joint site frequency spectrum, Sikora et al. (2019) reported that present-day Siberians (Even) have a 3.9% admixture from a lineage related to the Kolyma individual.

Unfortunately, ancient-genome sequences from the Yangtze River basin have not yet been reported. Instead of direct sampling, several ancient-genome sequences around South China have been analyzed, as described earlier (Yang et al., 2020). Recent ancient-genome studies in the West Liao River basin, the Yellow River basin, and South China showed a general trend indicating that the Early Neolithic northern and southern East Asians were more highly differ-
entiated from each other and the admixture between the two groups continued during and after the Neolithic period (Yang et al., 2020). Gene flow in both directions, north to south and south to north, was identified; in the study of Yang et al. (2020), the effect of north-to-south gene flow was not restricted to South China; it extended to Southeast Asia and the Tibetan Plateau. Yang et al. (2020) also showed genetic similarities among populations in coastal regions of East Asia, suggesting the importance of coastal migration over time. The pattern of admixture is also evident in the PCA plot (Figure 3); the Han Chinese from Beijing (CHB) have a broad distribution along the north-to-south genetic cline, and the Han Chinese from Hunan and Fujian (CHS) clustered toward the southern genetic components.

In considering the origin of present-day Japanese people, it is also important to summarize the population history of people in the West Liao River basin. The West Liao River basin lies at the root of the Korean peninsula. The similarity in pottery-making style between the Late Neolithic to Bronze Age cultures in the West Liao River basin and the Late Neolithic cultures in the Korean peninsula (Jeulmun pottery period) has frequently been acknowledged (Kim, 2015). The region currently consists of dry land but was suitable for millet farming in the Neolithic period (Shelach-Lavi, 2015; Li et al., 2020). A geological study showed that a climatic shift occurred around 2000 BCE in this region (Yang et al., 2015). The major Neolithic/Bronze Age cultures there were Xinglongwa (6200–5400 BCE), Zhaoabogou (5400–4500 BCE), Hongshan (4500–3000 BCE), Lower Xiajiadian (2200–1600 BCE), and Upper Xiajiadian (1200–600 BCE) cultures. Ning et al. (2020) showed that in this region, the genetic components of northern East Asians kept decreasing during the Middle to Late Neolithic periods but increased again in the Bronze Age Upper Xiajiadian culture. The increase of northern genetic components at the later stage was probably triggered by the climate change that occurred around 2000 BCE, which led to a lifestyle change from millet farming to pastoralism. This matches with the archeological differences between the Lower and Upper Xiajiadian cultures.

**Establishment of Japanese populations**

In this section, we summarize previous studies that help to elucidate the origin of people in the Japanese archipelago and present additional data analysis. Since genome-wide SNP studies revealed present-day Ainu and Ryukyu people to have specific genetic features compared with the mainland Japanese population (Yamaguchi-Kabata et al., 2008; Jinam et al., 2012, 2015a; Sato et al., 2014a; Sakaue et al., 2020), we focus here mostly on the migration and admixture process in mainland Japan.

**Origin of Jomon people**

One of the earliest signatures for the presence of humans in the Japanese archipelago is a finding of Paleolithic stone tools at several sites dating back to about 38000 BP in mainland Japan (Kudo and Kumon, 2012) and around 35000 BP in the Ryukyus (Fujita et al., 2016). In addition, several Pleistocene human fossil remains were identified in the Ryukyus (Baba and Narasaki, 1991; Mats’ura, 1999; Nakagawa et al., 2010). Any genetic connection between these Paleolithic people and Jomon people is largely unknown, but a recent genome-wide SNP study of present-day Ryukyuans, who have inherited stronger Jomon ancestry than the mainland Japanese, suggested that there is no major genetic contribution from ancient people to present-day people in the Ryukyus. The genetic divergence among different Ryukyuan islanders and divergence between Ryukyuans and Han Chinese population were too shallow for Ryukyuans to be descendants of the Paleolithic people there (Sato et al., 2014a).

In recent years, several genomes of approximately 2500- to 4000-year-old Jomon samples have been successfully sequenced (Kanzawa-Kiriyma et al., 2017, 2019; McColl et al., 2018; Gakuhari et al., 2020). All of the studies agreed that the Jomon people represent a population deeply diverged from other East Asian populations and that Ainu people are more closely related to Jomon people than present-day mainland Japanese people are (Kanzawa-Kiriyma et al., 2019; Gakuhari et al., 2020). Whether the Jomon people were genetically heterogeneous is a matter of debate, but so far, the Jomon genome sequence data do not show any systematic bias among the samples (Kanzawa-Kiriyma et al., 2019; Gakuhari et al., 2020). In order to test whether such heterogeneity is detectable, we applied a statistically rigorous test for the homogeneity of Jomon people using the qpWave algorithm (Haak et al., 2015). The algorithm tests whether two target populations are equally related to all outgroup populations. With this analysis, we could not reject the model where Ikawazu (IK002) and Funadomari (F23) individuals belonged to a single population with respect to the other populations ($\chi^2$, 0.88; df, 21) with a currently available dataset.

The Jomon people also have broad genetic affinity to coastal populations of East Asia, such as aboriginal Taiwanese (Ami and Atayal), Korean, Nivkh, and Ulchi people (McColl et al., 2018; Kanzawa-Kiriyma et al., 2019; Gakuhari et al., 2020). This pattern supports the hypothesis that the Jomon people entered the Japanese archipelago via a southern coastal route (Gakuhari et al., 2020), but is also consistent with admixture during the Neolithic (Jomon) period along the coastal regions of East Asia (Yang et al., 2020). The strong affinity of Nivkh and Ulchi to Jomon can also be explained by more recent gene flow through Ainu (Gakuhari et al., 2020). Gakuhari et al. (2020) reported supportive evidence for the southern coastal route hypothesis in archeological remains and morphological traits. We also examined the similarity of the Jomon people to present-day individuals using our dataset. Our results were generally consistent with previous findings (Supplementary Table 4); present-day Japanese Koreans and show statistically significant ties with a Jomon individual: $f_4$(Yoruba, Jomon; Japanese, Dai) = –0.0447 ($Z = –34.35$ and $f_4$(Yoruba, Korean; Japanese, Dai) = –0.0082 ($Z = –6.47$). However, the tie between Taiwan aboriginal Ami and Jomon was only marginally significant: $f_4$(Yoruba, Jomon; Ami, Dai) = –0.0090 ($Z = –2.51$). Considering the statistically significant genetic relationship between Jomon and Ami in the previous study (Gakuhari et al., 2020), the lower statistical support in this study may be due to the choice of the Dai population as a
Archeological evidence implies that a particular Paleolithic technology, microblades, spread through Northeastern Asia and northern East Asia, including the Japanese archipelago, around 25000–20000 years ago (Lee, 2009). The spread of microblades might have been related to the southward movement of the Ancient North Siberians or Paleosiberians during the LGM. Previous studies attempted to evaluate the genetic connection between the Jomon and Ancient North Siberians, using the Mal’ta individual (MA-1), but did not find strong support for the relationship (Kanzawa-Kiriyama et al., 2017, 2019; Gakuhari et al., 2020). However, Yang et al. (2020) reported that the genetic component of Paleosiberians (Kolyma) does exist in present-day people in northern and southern China, suggesting that the impact of Ancient North Siberian/Paleo-Siberian ancestry is more widespread than previously thought.

We show $f_4$(Yoruba, MA-1; Y, Dai) values in Figure 4A, using various Asian populations as the Y. The figure indicates that, while Northeast Asian populations showed significantly closer relationship with the Mal’ta individual ($Z < -3.0$) than the Dai population did, the other East and Southeast Asian populations did not show any significant association with the genome of the Mal’ta individual.

Here we also present an interesting yet mysterious relationship between Jomon and Yana individuals. We observed a marginally significant genetic similarity between Jomon and Yana individuals and between present-day Japanese and Yana individuals using $f_4$ statistics; the $f_4$(Yoruba, Yana; Jomon, Dai) value was $-0.0144$ ($Z = -2.720$) and the $f_4$(Yoruba, Yana; Japanese, Dai) value was $-0.0034$ ($Z = -2.905$) (Figure 4B). This pattern cannot be solely explained by the gene flow from the Ancient North Siberians/Paleo-Siberians to East Asians. For example, for all present-day Northeast Asian populations (population Y) we analyzed, $f_4$(Yoruba, MA-1; Y, Dai) values were consistently smaller than $f_4$(Yoruba, Yana; Y, Dai) values (Figure 4A, B), suggesting MA-1 rather than Yana individual was responsible for the Ancient North Siberian genetic component in the present-day Northeast Asians, as described above. In contrast, the $f_4$(Yoruba, Yana; Jomon/Japanese, Dai) value was smaller than the $f_4$(Yoruba, MA-1; Jomon/Japanese, Dai) value, which was opposite to the other cases except for the Tianyuan individual, who certainly did not accept any gene flow from the MA-1 individual, because the Tianyuan sample is ~40000 years old. This pattern could be explained by one of two scenarios. First, similar to the population related to Tianyuan, an ancient population closely related to Jomon could have been involved in the ancient contact between the people migrating via the southern and northern routes. A second possibility is that a population more closely related to Yana than to Mal’ta migrated south during the LGM and admixed with the ancestral Jomon lineage only in the Japanese archipelago. Although we currently do not have any genetic data that help to distinguish the likelihood of these two scenarios, it is difficult to imagine that descendants or populations closely related to Yana left the signature of admixture only to Jomon. We should note that the deviation of $f_4$ statistics showing the tie between Jomon and Yana was only marginally significant. Further validation studies using better-quality ancient samples would be necessary. Future
studies, particularly additional samples of Paleo-Siberians and ancient Central Asians, would help to elucidate the reason for the potential genetic connection between Yana and Jomon people.

**Origin of Yayoi people**

The PCA plot in Figure 3 clearly shows that the present-day population genetically closest to the mainland Japanese is Korean. Archeological evidence also supports the cultural similarity at the beginning of the Yayoi period (~1000 BCE) between the northern Kyushu and the southwest Korean peninsula (Fujio, 2017). An estimate of the split time between present-day mainland Japanese and Koreans using haplotype-phasing data mostly supports the hypothesis that the split occurred about 3000 years ago (Kim et al., 2020). However, the time estimate needs to be interpreted carefully in general, because like all estimates based on genetic data, it depends on assumptions of average generation time and mutation/recombination rate in genomes.

A remaining question is how these rice-farmers migrated to or were established in the Korean peninsula. When rice-farming was introduced to the northern Kyushu, the Korean peninsula was in the Mumun period, with a culture similar to the one in the northern Kyushu (Fujio, 2017). Studies of Y-chromosomal haplotypes revealed that the lineages of Y-chromosomal haplogroup O, which has the highest frequency in the mainland Japanese population, are also frequent in Korean and Southeast Asian populations (Hammer et al., 2006). The tie between Southeast Asia and Japan–Korea supports the hypothesis that rice-farming originated in the Yangtze River basin and was then introduced to the Korean peninsula and Japanese archipelago. We tested the hypothesis using available ancient and present-day genome data. Ancient-genome data were obtained from the studies of Ning et al. (2020) and Yang et al. (2020). We evaluated whether present-day Japanese and Korean people are genetically similar to any of the Neolithic populations linked to the West Liao River basin, the Yellow River basin, and the Yangtze River basin. Neolithic samples from the Yangtze River basin were not available, and therefore we used Tanshishan sites (~2500–2000 BCE) as proxy. The samples from the West Liao River basin were BLSM45 (~3300–3100 BCE) with the Hongshan culture and EDM176 (~2000–1600 BCE) with the Lower Xiajiadian culture. The sample names from the Yellow River basin were WD-WT1H16 (~2100–2000 BCE) and PLTM313 (~2100–1900 BCE) with the Henan Longshan culture, Bianbian (~7600–7500 BCE) with the Houli culture, and Boshan (~6400–6100 BCE). The $f_1$ statistics showed that both present-day Japanese and Korean populations were genetically similar to the samples from the Early Neolithic Yellow River basin (Boshan) (Supplementary Table 5). However, the statistical significance of the similarity to Boshan compared with CHB was marginal ($Z = 2.47$). On the other hand, we found Japanese–Korean migrants were significantly more closely related to CHB than Neolithic southern Chinese individuals (Supplementary Table 5), indicating that the people in the Yangtze River basin did not substantially contribute to the genetic basis of present-day Japanese and Korean populations. A parallel pattern to the $f_1$ statistics is observed in the PCA plot of Figure 3; the Early Neolithic Boshan and Bianbian individuals in the Lower Yellow River basin were plotted fairly close to present-day Korean individuals. Previous studies showed that the ancient samples from North China have more northern East Asian genetic components than the present-day Han Chinese. The results suggest that the basal population of present-day Japanese and Korean people was somewhat different genetically to the Late Neolithic populations in the Yellow River and West Liao River basins. We will know more when ancient-genome data of Neolithic people in the Korean peninsula become available.

If the dual-structure model generally holds, we would like to know how much Jomon ancestry the present-day mainland Japanese have. Previous studies proposed a wide range of Jomon ancestry fractions, from 9% to 45%, using different datasets, different models, and different assumptions (Horai et al., 1996; Nakagome et al., 2015; Kanzawa-Kiriyama et al., 2017, 2019; McColl et al., 2018). We first estimated the Jomon ancestry fraction using the $f_1$ ratio test in a similar way as Kanzawa-Kiriyama et al. (2019) did, assuming Jomon and present-day Korean are the source populations of mainland Japanese with Papuan and Dai as outgroup populations. With this assumption, the estimated values were highly similar to the previous one ($0.122 \pm 0.009$; standard error with the block jackknife method). Because the estimates of the admixture fraction are not robust for the choice of outgroups, we subsequently evaluated the results of a qpAdm analysis, using many different populations to compute the $f_1$ statistics (Haak et al., 2015).

Yang et al. (2020) suggested the present-day Japanese are best modeled as an admixture between the Paleolithic northern East Asian individuals (Boshan) and Jomon individuals (Ikawazu) with the Jomon ancestry ratio being 0.38. We tested the model using the same source populations (Boshan and Ikawazu) but obtained a much smaller Jomon ancestry fraction ($0.103 \pm 0.030$). The reason for the discrepancy is not clear, but it is likely because we used a different set of right (outgroup) populations to those used in the analysis by Yang et al. (2020). When we assumed that the present-day Japanese population was an admixture of present-day Korean and Jomon populations (represented by F23), the estimated Jomon ancestry by qpAdm was $0.102 \pm 0.005$, slightly smaller than, but similar to, the values we obtained with the $f_1$ ratio test. We further tested a three-way admixture model using the left populations (possible source populations) as Jomon, Korean, and Hezhen, and excluded Hezhen and Oroqen from the right populations. The three-way model (Korean, Hezhen, and Jomon, $\chi^2$: 45.4) fits significantly better than the two-way model (Jomon and Korean, $\chi^2$: 58.9). The ancestry fraction of Jomon, Korean, and Hezhen was estimated to be 0.096, 0.832, and 0.072, respectively, but the three-way model does not fully explain the data either. If we assume that there was no additional migration after the Yayoi period from the Amur River basin, the pattern would imply that Yayoi migrants might harbor more northern East Asian genetic components than present-day Koreans do, as shown in the previous analysis with $f_1$ statistics. When F23 was replaced to IK002, we obtained essentially the same results (data not shown). In conclusion, our analysis showed that Jomon an-
cerity in the present-day Japanese is around a lower bound of the previous estimates and does not overly exceed 10%.

The 10% Jomon ancestry of present-day Japanese nuclear genomes, however, seems to contradict the previous estimates made using mitochondrial and Y-chromosomal genotype data. Considering the genotype distribution of present-day Japanese and Jomon people, mitochondrial haplogroup M7a, N9b, and G1b, and Y-chromosome haplogroup D1a2a (M-55) are considered to be a typical Jomon allele (Tanaka et al., 2004; Hammer et al., 2006; Shi et al., 2008; Adachi et al., 2009; Sato et al., 2014b). The high frequency of the Y-chromosomal Jomon haplotype (~30%) clearly shows that Jomon ancestry in the present-day Japanese population is much stronger on the Y chromosomes than on autosomes (Sato et al., 2014b). In contrast, the Jomon ancestry proportion of mitochondrial genomes is less certain because the frequency of M7a and N9b haplogroups in Jomon people are somewhat variable across the Japanese archipelago (Adachi et al., 2009). If we assumed that the proportion of M7a and N9b haplogroups in Jomon was around 70%, the mitochondrial Jomon ancestry would be around 15% in present-day Japanese individuals. The observed imbalance of Jomon ancestry among autosomal, Y-chromosomal, and mitochondrial genomes, which we refer to as the ‘admixture paradox’, seems confusing but worthwhile to study further to elucidate the process of admixture of Jomon and Yayoi genetic components.

The above arguments, however, seem too simplified when we consider a recent ancient genome sequencing study of an approximately 2000-year-old Yayoi female from a typical migrant culture (Shinoda et al., 2020). The analysis showed that the genetic features of samples fell within the PCA plots of present-day Japanese individuals, indicating that they already had admixed genetic components of Jomon and continental ones. The finding further complicates the story of the origin of present-day Japanese. More ancient-genome sequencing of populations present c. 1000 BCE in both the Korean peninsula and Japanese archipelago could help to explain these findings.

Conclusions

As summarized in this review, recent studies of ancient-genome sequencing have provided us with a tremendously important insight into the movement of prehistoric people in Asia. Because the ancient genetic data in the Japanese archipelago have not yet been fully analyzed, we have to corroborate the prehistoric human movements using not only genetic but also linguistic and archeological data in order to reconstruct a more comprehensive history of the region. Compared with studies in China, the genome sequencing study of present-day and ancient people in the Japanese archipelago is not as well developed. Even ancestry information about present-day Japanese genome sequences has not been thoroughly investigated yet. Future studies using ancestry information of present-day Japanese samples and ancient samples across prehistoric and historic time periods would further elucidate the formation process of present-day Japanese populations in fine detail, and could finally answer the questions raised in Figure 1.

Methods

Genotyping of human genome sequences obtained from large whole-genome sequence panels

We collected modern human genome data distributed from public databases. We used samples from the 1000 Genome Project (1KGP) (1000 Genomes Project Consortium, 2015), which were reanalyzed by high-coverage WGS. In addition, the data analyzed by the Simons Genome Diversity Project (SGDP) (Mallick et al., 2016), the Human Genome Diversity Project (HGDP) (Bergström et al., 2020), and the Korean Personal Genome Project (KPGP) (Kim et al., 2018) were used. A list of project accession numbers in the public database is summarized in Table 1, and a list of individuals is shown in Supplementary Table 6. In this study, we used GRCh38 as a reference sequence for read mapping. For 1KGP, we retrieved the CRAM files mapped to GRCh38 and used them directly in this study. The BAM files of SGDP and HGDP, which were mapped to the GRCh37 reference sequence, were retrieved from databases and converted to fastq files. For KPGP, we retrieved the fastq files from the database. These fastq files were mapped to GRCh38 and the subsequent CRAM files were analyzed together with the 1KGP data.

We followed the workflow recommended by the GATK best practice (Van der Auwera et al., 2013). The read sequences in fastq format were mapped to GRCh38 using the bwa-mem algorithm (Li and Durbin, 2009) followed by removal of duplicate reads. The variants, including SNVs and insertion and deletion (indels), were discovered by the HaplotypeCaller algorithm implemented in GATK4.1. We performed the HaplotypeCaller with default options except for the ‘-ERC GVCF’ option to allow joint calling in the following process. Joint calling is a process of final genotyping, in which the genotypes of all samples, including the homoyzogous reference allele, can be determined simultaneously by analyzing information from multiple samples. We conducted the joint calling with the GenotypeGVCF algorithm implemented in sentieon v201911 (Sentieon Inc.) which is functionally equivalent to GATK’s joint calling procedures (GenomicDBImport followed by GenotypeGVCFs). To filter low-quality variants, we applied the Variant Quality Score Recalibration (VQSR) approach, which scores the quality of variant by a machine-learning method. We applied VQSR with the valCal and ApplyVarCal programs implemented in sentieon v201911 to the variant call data. The HapMap and 1000 Genomes Omni2.5 sites, the high-confidence SNPs of 1000G sites, and the dbSNP151 sites were used as true datasets, training datasets, and known datasets, respectively. SNPs that passed the 99.9% sensitivity filter by

| Table 1. The list of the accession numbers for whole-genome sequencing data |
|-----------------------------|------------------|
| Project name                 | Accession number  |
| 1000 Genome Project          | PRJEB31736        |
| Simon’s Genome Diversity Project | PRJEB9589, ERP010710 |
| Human Genome Diversity Project | PRJEB6463        |
| Korean Personal Genome Project | PRJNA284338      |
VQSR were used for subsequent analyses.

In order to focus on East Asian populations, samples in the following 24 present-day populations were extracted from the whole dataset: Yoruba, Luhya, Bengali, STU (Sri Lankan Tamil in the UK), CEU, GBR (British), Japanese, Korean, Hezhen, Oroqen, Daur, Mongolian, Yakut, Karitiana, CHB (Han Chinese in Beijing), CHS (Southern Han Chinese), She, Miao, Dai, Ami, Kinh, Cambodian, Kusunda, and Papuan. For the populations labeled with three capital letters, such as CHB, we solely used the dataset obtained from 1KGP, whereas we merged populations from different panels on the basis of the descriptions in Supplementary Table 6. Mongolian, Miao, and Cambodian were originally labeled as Mongola, Miaozu, and Cambodians in some datasets, respectively, but the labels were changed in the merging process. We performed PCA using smartpca for initial quality checking, and excluded the following samples from the analysis: NA18976 (Japanese), KPGP-00232 (Korean), KPGP-00364 (Korean), HGDP00720 (Cambodian), HGDP01219 (Daur), and LP6005443-DNA:D06 (Korean).

Genotyping of additional human genome sequences

The additional sequence data were downloaded from the public databases. Sample names and accession numbers are described in Supplementary Table 7. We used the F23 female as a representative of the Jomon population unless otherwise noted. All short reads were mapped to GRCh38 using the bwa-mem algorithm with a default parameter setting (Li and Durbin, 2009). For the genome sequences of relatively high coverage (Yana1, Yana2, Kolyma1, F23, JHF05, JHM06), we genotyped the samples using the standard GATK genotyping pipeline. For the other low-coverage genomes, we used the pileupCaller program (https://github.com/stschiff/sequenceTools) to randomly pick up one read for each variant site, restricting the sites to those with a quality score ≥30 and mapping quality ≥30. All of the genotype files were converted into the Eigenstrat file format and merged into a single set of files. The genotypes of ancient samples analyzed by Yang et al. (2020) (Boshan, Bianbian, Tanshishan, and Xitoucun) were downloaded from the National Genomics Data Center (BIG Data Center Members, 2018; Yang et al., 2020) as Eigenstrat genotype format files, and then merged to the other dataset.

Population structure analysis

We restricted our analysis to strict-masked sites from 1KGP, which are uniformly accessible for WGS (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/working/20160622_genome_mask_GRCh38/StrictMask/). We used only autosomal transversion SNPs for the further analysis in order to avoid the systematic biases caused by cytosine deamination in ancient samples. In total we obtained 48428968 variant sites. PCA was performed using the smartpca program (Patterson et al., 2006). Fourteen present-day Asian populations (Cambodian, Dai, Kinh, Ami, She, Miao, CHS, CHB, Korean, Japanese, Daur, Oroqen, Hezhen, and Yakut) were used to compute PCs and the ancient samples were projected onto the plot using the lsqprojection option in smartpca. AdmixTools was used to calculate $F_4$ statistics (Patterson et al., 2012). The admixture analysis using qpWave and qpAdm (Haak et al., 2015) was performed with 22 present-day populations as right (out-group) populations: Yoruba, Luhya, CEU, GBR, STU, Bengali, Papuan, Jehai, Kusunda, Cambodian, Kinh, Dai, She, Miao, Ami, CHS, CHB, Daur, Hezhen, Oroqen, Mongolian, and Karitiana. When Hezhen was used as a left (reference) population, Hezhen and Oroqen were removed from the right populations.

Acknowledgments

We would like to thank the two anonymous reviewers for helpful suggestions. This work was partly supported by MEXT KAKENHI (grant 18H05511 to N.O.). We also appreciate support from the Yaponesian Genome Project (Deciphering Origin and Establishment of Yaponesians mainly based on Genome Sequences). Computations were partially performed on the NIG supercomputer at National Institute of Genetics, ROIS. The 1KGP data were generated at the New York Genome Center with funds provided by NHGRI Grant 3UM1HG008901-03S1. The authors would like to thank Enago (www.enago.jp) for the English language review.

Authors’ contributions

N.O. and Y.K. analyzed the data and N.O. wrote the review part.

Conflict of interest statement

The authors declare no conflict of interest.

References


considered from the hypothesis of farmer/language spread. Evolutionary Human Sciences, 2: e13.


