Insights into *Mus musculus* Population Structure across Eurasia Revealed by Whole-Genome Analysis

Kazumichi Fujiwara [1,2], Yosuke Kawai [3], Toyoyuki Takada[4], Toshihiko Shiroishi[5], Naruya Saitou[6], Hitoshi Suzuki [7], and Naoki Osada [1,2,*]

[1]Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan

[2]Global Station for Big Data and Cybersecurity, GI-CoRE, Hokkaido University, Sapporo, Japan

[3]Genome Medical Science Project (Toyama), National Center for Global Health and Medicine (NCGM), Tokyo, Japan

[4]Integrated BioResource Information Division, RIKEN BioResource Research Center, Tsukuba, Japan

[5]RIKEN BioResource Research Center, Tsukuba, Japan

[6]National Institute of Genetics, Mishima, Japan

[7]Graduate School of Environmental Science, Hokkaido University, Sapporo, Japan

*Corresponding author: Email: nosada@ist.hokudai.ac.jp.

Accepted: 03 May 2022

Abstract

For more than 100 years, house mice (*Mus musculus*) have been used as a key animal model in biomedical research. House mice are genetically diverse, yet their genetic background at the global level has not been fully understood. Previous studies have suggested that they originated in South Asia and diverged into three major subspecies, almost simultaneously, approximately 110,000–500,000 years ago; however, they have spread across the world with the migration of modern humans in prehistoric and historic times (∼10,000 years ago to the present day) and have undergone secondary contact, which has complicated the genetic landscape of wild house mice. In this study, we sequenced the whole-genome sequences of 98 wild house mice collected from Eurasia, particularly East Asia, Southeast Asia, and South Asia. Although wild house mice were found to consist of three major genetic groups corresponding to the three major subspecies, individuals representing admixtures between subspecies were more prevalent in East Asia than has been previously recognized. Furthermore, several samples exhibited an incongruent pattern of genealogies between mitochondrial and autosomal genomes. Using samples that likely retained the original genetic components of subspecies with the least admixture, we estimated the pattern and timing of divergence among the subspecies. The estimated divergence time of the three subspecies was 187,000–226,000 years ago. These results will help us to understand the genetic diversity of wild mice on a global scale, and the findings will be particularly useful in future biomedical and evolutionary studies involving laboratory mice established from such wild mice.

Key words: *Mus musculus*, house mouse, population genetics, population demography, genetic diversity.

Significance

Because the house mouse (*Mus musculus*) is widely used in genetics and biomedical research, it is important to understand the genetic status of wild house mice from which research strains are derived. However, the global genetic diversity of wild house mice is not well understood. In this study, we investigated the genetic landscape of wild house mice using the samples collected from across the Eurasian continent and Southeast Asian islands, particularly East, Southeast, and South Asia. The genetic resources provided here are expected to facilitate future research involving house mice.

© The Author(s) 2022. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Introduction

The house mouse (*Mus musculus*) has been an important animal model in biomedical research for more than 100 years, and many inbred strains are currently available for research. Inbred laboratory strains are genetically diverse, originating from at least three wild subspecies (Yonekawa et al. 1980, 1982, 1981; Moriwaki et al. 1984; Bonhomme et al. 1987; Yang et al. 2007; Didion and de Villena 2013): *M. m. musculus* (MUS) from northern Eurasia, *M. m. castaneus* (CAS) from southern Asia, and *M. m. domesticus* (DOM) from western Europe. Hereafter we use these abbreviations to represent the subspecies. The first mouse reference genome sequence was created using the classical inbred strain C57BL/6J (Mouse Genome Sequencing Consortium 2002), and dozens of whole-genome sequences of laboratory mouse strains have been published ever since (Keane et al. 2011). The genome of the classical inbred strain is derived from approximately 94.3% DOM, 5.4% MUS, and 0.3% CAS (Keane et al. 2011), whereas the mitochondrial genome is that of DOM (Frazer et al. 2007). In addition, Frazer et al. 2007 estimated that 10% of the classical inbred strain genome is derived from *M. m. molossinus* (MOL), which is thought to have arisen from hybridization between MUS and CAS found in Japan (Yonekawa et al. 1988; Takada et al. 2013). Various strains of laboratory mice have been investigated, with some researchers analyzing the sequences of diverse mouse strains focusing on the origin of subspecies (e.g., Yang et al. 2011).

Despite intensive effort to sequence the genomes of laboratory mice, the genetic diversity of wild house mice has yet to be thoroughly investigated. Previous studies found that wild house mice are highly genetically diverse and have an estimated effective population size of around 10^5 (Baines and Harr 2007; Geraldes et al. 2008, 2011; Halligan et al. 2010, 2013). Classical inbred strains represent a small fraction of the genetic diversity of wild house mice (Salcedo et al. 2007). Previous studies have demonstrated a genome-wide pattern of polymorphisms in wild house mice, but such studies have mostly focused on CAS or DOM within limited geographic ranges (Halligan et al. 2013; Harr et al. 2016; Phifer-Rixey et al. 2018). Therefore, a large-scale genome sequencing study covering the Eurasian continent and surrounding islands will substaintially improve our understanding of the worldwide genetic diversity of house mice.

Wild house mice are distributed almost worldwide, including on remote islands: CAS inhabits the Indian subcontinent and Southeast Asia; DOM inhabits North and South America, Africa, the Middle East, Australia, Southwestern Europe, and many surrounding and remote islands; and MUS inhabits Siberia, Central Asia, East Asia, and Northeastern Europe. In addition, more subspecies have been proposed by previous studies. For example, wild house mice found in Japan are considered to be a hybrid between CAS and MUS, forming the independent subspecies MOL (Yonekawa et al. 1988; Takada et al. 2013). Furthermore, mitochondrial phylogeny suggests that the subspecies *M. m. gentilulus* exists in the Arabian Peninsula (Prager et al. 1998) and that an unspecified potential subspecies exists in Nepal (Suzuki et al. 2013; Hardouin et al. 2015). Moreover, highly diversified nuclear genomes and mitochondrial lineages of *M. musculus* have been reported in Iran (Rajabi-Maham et al. 2012; Hardouin et al. 2015). The original homeland of *M. musculus* has been proposed as the northern part of the Indian subcontinent (Boursot et al. 1993; Din et al. 1996), and its common ancestors diverged into the three subspecies, almost simultaneously (Didion and de Villena 2013), approximately 110,000–500,000 years ago (Boursot et al. 1996; Suzuki et al. 2004; Salcedo et al. 2007; Geraldes et al. 2008, 2011; Bonhomme and Searle 2012; Phifer-Rixey et al. 2020).

Wild house mice migrated and lived commensally with humans. With prehistoric and historic long-distance migration of humans, house mice, for which the staple food is grain, expanded their range with the development of agriculture and cultural exchange (Sage 1981; Moriwaki et al. 1986; Bonhomme et al. 2010; Gabriel et al. 2011; Jones et al. 2013). As the mice moved with humans, genetically diverse subspecies were brought into secondary contact (Boursot et al. 1993; Duvaux et al. 2011), which allowed admixture of their genomes (Bonhomme et al. 2007; Staubach et al. 2012; Liu et al. 2015), despite partial reproductive isolation between subspecies, for example, DOM and MUS (White et al. 2011). DOM and MUS make contact along a narrow hybrid zone in Europe, whereas CAS and MUS seem to have a broader hybrid zone (e.g., Vanlerberghe et al. 1986; Payseur et al. 2004; Dod et al. 2005; Raufaste et al. 2005; Macholn et al. 2007; Teeter et al. 2010, 2008; Jones et al. 2011; Wang et al. 2011; Ďureje et al. 2012) across Central and Eastern Asia (Boursot et al. 1993; Jing et al. 2014). Previous phylogenetic and phylogeographic studies have analyzed mitochondrial DNA sequences and limited nuclear gene sequence data from house mice (e.g., Liu et al. 2008). However, the prevalence of hybridization between subspecies in the global population has yet to be elucidated using genome-wide sequence data. In the pregenomic era, it was recognized that the genetic and phenotypic boundaries between subspecies, except for the boundary between DOM and MUS in western and central Europe, were obscure due to high variability within subspecies (Boursot et al. 1993). This obscurity may have been caused by the limited number of autosomal loci analyzed at that time.

The rapid advancement in sequencing technologies has enabled the use of whole-genome data to estimate population structures and phylogenetic histories. In this study, we sequenced the whole genomes of 98 wild house mice that had previously been collected from across the Eurasian continent and Southeast Asian islands, particularly mouse samples from East, Southeast, and South Asia. Our analysis revealed that the hybridization between subspecies is prevalent beyond the hybrid zones at a higher level than was previously estimated, particularly that between CAS and MUS in East Asia. Moreover, we estimated the past population size of all individuals used in this study, which allows us to infer the population history of wild house mice that are commensal to humans with agricultural culture. These results improve our understanding of the genetic diversity of house mice at the global level and will facilitate future biomedical and evolutionary research.

Results

Genetic Diversity of *Mus musculus*

In this study, we analyzed 141 whole-genome-sequenced samples of *M. musculus* and *Mus spretus* (SPR). These included 98 newly sequenced *M. musculus* samples from Eurasia, including samples from Estonia (EST), Ukraine (UKR), Russia (RUS), Iran (IRN), Kazakhstan (KAZ), Pakistan (PAK), India (IND), Sri Lanka (LKA), Nepal (NPL), China (CHN), Vietnam (VNM), Indonesia (IDN), Taiwan (TWN), Korea (KOR), and Japan (JPN) (supplementary fig. S1, Supplementary Material online). Data for the other 35 *M. musculus* and 8 SPR samples were previously published by Harr et al. (2016) including *M. musculus* samples from Germany (DEU), France (FRA), Iran (IRN), Czech Republic (CZE), Kazakhstan (KAZ), India (IND), and SPR samples from Spain. After filtering individuals with equal to or greater than third-degree kinship, which was inferred using the kinship coefficient, 128 samples were retained for further analysis, including 94 of our newly sequenced samples. The samples used in this study are listed in supplementary table S1, Supplementary Material online. Of the 94 newly sequenced samples, 61 were males and 33 were females. The sex of each of these samples is summarized in supplementary table S2, Supplementary Material online. After the variants were filtered, we obtained 134,030,288 SNVs and 31,618,947 indels for the dataset containing 128 samples of SPR and *M. musculus*, and 107,337,961 SNVs and 25,875,963 indels for the dataset containing 121 samples of *M. musculus*. The overall transition/transversion ratio of our samples was 2.23. The detailed filtering process is described in the Materials and Methods section.

As reported by Li et al. (2021), mitochondrial genome sequences were clustered into four distinct clades, three of which presumably corresponded to CAS, DOM, and MUS subspecies. The mitochondrial genomes of samples from Nepal (NPL01 and NPL02) diverged before the split between the CAS and DOM clades. When we classified our 94 newly sequenced samples according to mitochondrial haplogroups, the per-sample nucleotide diversity (heterozygosity) of the three subspecies was as follows: 0.00006–0.00757 for CAS (including the NEP mitochondrial haplogroup), 0.00003–0.00450 for MUS, and 0.00040–0.00561 for DOM. The average ratio of nonsynonymous to synonymous polymorphic sites for each sample was 0.415. The basic statistics for each individual, including per-sample nucleotide diversity for all 94 newly sequenced samples, are presented in supplementary table S3, Supplementary Material online.

Genetic Structure of Wild House Mice

We conducted principal component analysis (PCA) of all 128 samples using 100,832,598 autosomal SNVs, including those from SPR (supplementary fig. S2, Supplementary Material online). All *M. musculus* and *M. spretus* (SPR) were clearly differentiated along principal component 1 (PC1); principal component 2 (PC2) corresponded to variation within the *M. musculus* subspecies (supplementary fig. S2, Supplementary Material online). In PC2, the *M. musculus* subspecies were differentiated into two clusters with some intermediate samples. Mitochondrial haplotypes and sampling locations indicated that these two clusters corresponded to the DOM and CAS–MUS groups.

We subsequently excluded SPR and replotted the PCA using 84,744,729 autosomal SNVs. Figure 1*a* presents the locations of samples colored according to the three genetic components identified in the PCA plot (fig. 1*b*). The PC score for each sample is presented in supplementary table S4, Supplementary Material online. Considering the sampling locations, these three clusters corresponded to the CAS, DOM, and MUS subspecies, whereas a wide range of admixture between the CAS and MUS clusters was observed (fig. 1*b*). In the PCA plot, Nepalese samples with distinct mitochondrial haplogroups were clustered with CAS samples. PC1 shows the genetic differences between MUS and CAS–DOM, whereas PC2 shows the genetic differences between DOM and CAS–MUS. As presented in figure 1*a* and *b*, hybrid individuals were prevalent, particularly between CAS and MUS. The Chinese samples were scattered throughout a wide range of the CAS–MUS cline, and the Japanese samples were also scattered along this cline, although the range was narrower than that in the Chinese samples and skewed toward the MUS cluster. We also conducted PCA using only X-chromosomal SNVs (supplementary fig. S3, Supplementary Material online). This pattern was the same as that obtained using autosomal SNVs; however, the plots were more tightly clustered at each vertex, indicating that admixture was less pronounced on the X chromosomes. For the X chromosomes,

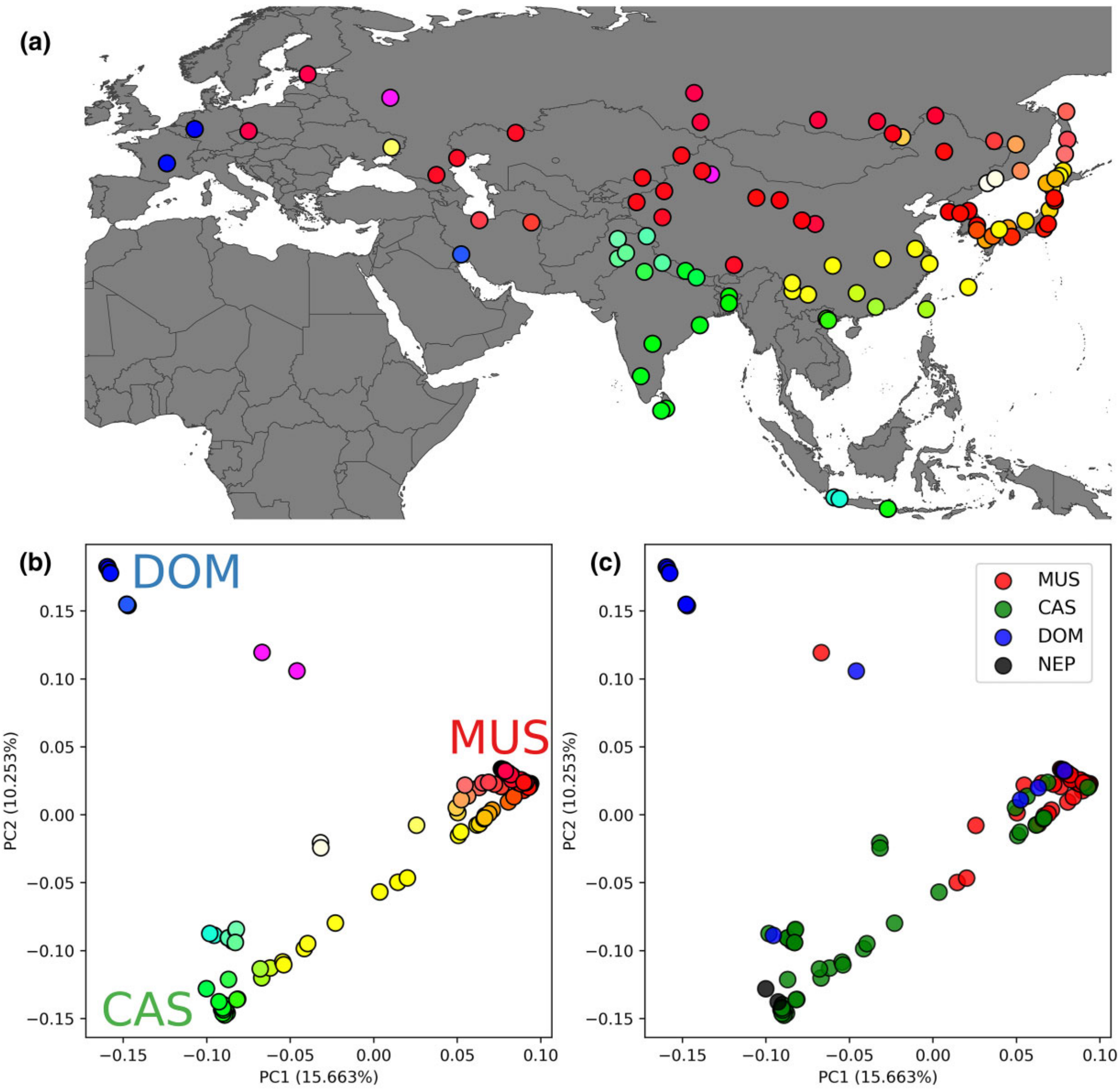

Fig. 1.—PCA results in wild house mice. (*a*) Geographical map of sampling locations. Circles represent each individual, and the assigned color is the same as that in panel (*b*). Supplementary fig. S1, Supplementary Material online presents the detailed names of sample collection sites. (*b*) PCA plot of wild house mice produced using autosomal SNVs. The *x* and *y* axes represent PC1 and PC2, respectively. Circles are colored according to "Maxwell's color triangle," assigning three vertices to the RGB colors. Red, green, and blue color intensities correspond to the MUS, CAS, and DOM genetic components, respectively. The proportion of variance for each PC is shown in parentheses on the axis label. (*c*) PCA plot of wild house mice produced using autosomal SNVs labeled with the mitochondrial genome haplogroup of each sample. NEP represents the mitochondrial haplogroup of Nepalese origin. The proportion of variance for each PC is shown in parentheses on the axis label.

the PC scores for each sample are presented in supplementary table S4, Supplementary Material online. PCA plots produced using autosomal and X-chromosomal data without linkage disequilibrium (LD) are presented in supplementary figs. S4 and S5, Supplementary Material online.

In figure 1*b*, samples from Germany, Bangladesh, and Korea are located at the vertices of a triangle (supplementary table S4, Supplementary Material online). However, this pattern does not necessarily imply that these samples are representatives of each of the three major subspecies; instead, the pattern may have been generated by the strong genetic drift in the three populations. To identify samples representing the lowest level of gene flow between subspecies, we computed the f_3 and f_4 statistics among different individuals (supplementary tables S5–S7 and S8–S13, Supplementary Material online, respectively). Indian CAS samples from mountainous regions, western

European DOM samples, and Korean MUS samples exhibited the lowest levels of gene flow compared with the other subspecies. Therefore, we selected an Indian sample (IND04), a Korean sample (KOR01), and a German sample (DEU01) as our reference samples of the CAS, MUS, and DOM subspecies, respectively.

Except for the aforementioned samples, most CAS–MUS samples exhibited some extent of admixture between subspecies. For example, compared with IND04, another Indian sample from Delhi (IND02) was slightly genetically closer to MUS (KOR01). The Z score of f_4 (SPR, KOR01; IND04, IND02) was 2.814 (supplementary table S14, Supplementary Material online). Compared with the IND03, IND04, and IND07 samples, other CAS samples from neighboring regions, such as Pakistan and Bangladesh, were more similar to MUS (KOR01) (supplementary tables S11 and S12, Supplementary Material online). CAS samples from East and Southeast Asian regions also exhibited various levels of admixture with MUS genomes. Likewise, all MUS samples from Eastern Europe had a significantly closer relationship to DOM than the other Asian MUS samples. For example, the Z scores of f_4 (SPR, DEU01; KOR01, CZE01) and f_4 (SPR, DEU01; KOR01, KAZ01) were 14.724 and 11.104, respectively (supplementary table S11, Supplementary Material online). The majority of MUS samples from East Asia, particularly those from Northern China and Japan, exhibited a high level of admixture with CAS. Compared with DOM samples from Western Europe, those from outside Western Europe, such as the Iranian and Russian samples, exhibited slightly but significantly closer affinity to MUS and CAS samples (supplementary tables S10 and S11, Supplementary Material online). One sample collected from Kathmandu in Nepal (NPL02) was the most distantly related to the CAS, MUS, and DOM samples. Another sample from Nepal (NPL01) exhibited a similar pattern; however, compared with NPL02, NPL01 was more closely related to CAS.

The PCA plot produced using nuclear genome data was labeled with the four mitochondrial haplogroups (CAS, DOM, MUS, and NEP; supplementary table S3, Supplementary Material online) and is presented in figure 1c. Most samples exhibited congruent patterns in their mitochondrial and nuclear genomes, although the patterns were incongruent in some samples. As presented in the PCA plot (fig. 1b and c), mitochondria–nuclear genome incongruence was more commonly observed in the MUS cluster than in the CAS or DOM cluster. For example, a sample from Khabarovsk in the Russian Far East (RUS13) had a DOM-type mitochondrial genome but a nuclear genome that was highly similar to that of MUS.

To infer the genomic ancestries of *M. musculus* genomes, we also conducted ADMIXTURE analysis (fig. 2). We analyzed 121 *M. musculus* samples by selecting *K*

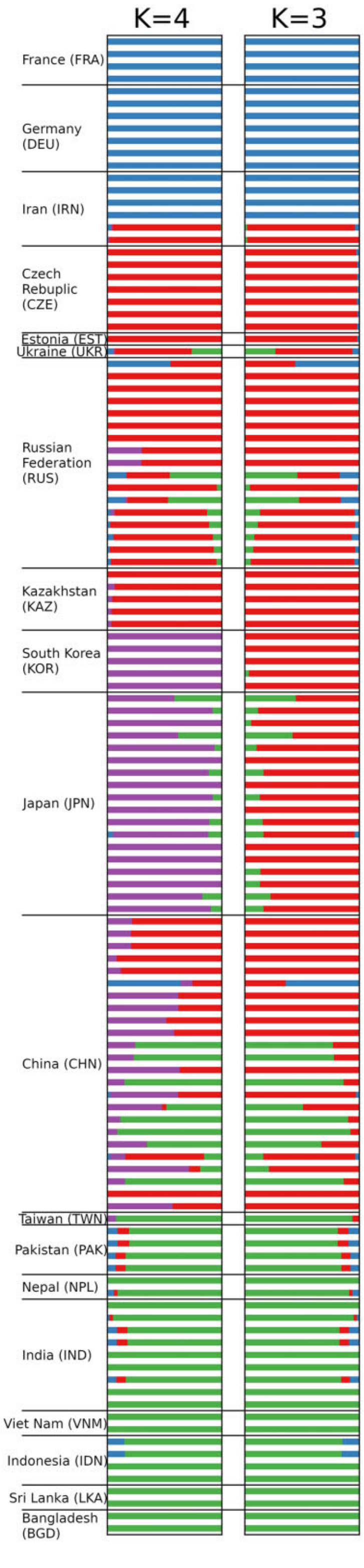

FIG. 2.—ADMIXTURE plot using autosomal data. Plot showing the proportion of estimated subspecies genetic components. Results for cluster $K = 3$ and $K = 4$ are presented. Samples with the same country code are ordered according to the sampling site from west to east and north to south.

values from 1 to 5, where K is a predefined number of ancient components. We also performed cross-validation to infer the most suitable K value (supplementary fig. S6, Supplementary Material online) and found that the cross-validation error rate was the lowest at $K = 4$. At $K = 3$, we confirmed the presence of three genetic components corresponding to CAS, DOM, and MUS, respectively. Most Japanese and Chinese samples exhibited a hybrid pattern with the CAS and MUS components. At $K = 4$, another genetic component represented the specific genetic features of Japanese and Korean mice. The results of the ADMIXTURE analysis using X-chromosomal SNVs were largely the same as those obtained using autosomal SNVs (supplementary fig. S7, Supplementary Material online). The ADMIXTURE plots of autosomes and X chromosomes without LD are presented in supplementary figs. S8 and S9, Supplementary Material online.

Inference of Past Demography Using PSMC and MSMC

To estimate the past demographic pattern of wild house mice, we conducted pairwise sequentially Markovian coalescent (PSMC) analysis for each individual. Supplementary fig. S10, Supplementary Material online presents the PSMC plots of all individual samples. Although some of the samples exhibited unusual patterns, the trajectories of most samples could be largely classified into three categories representing CAS, DOM, and MUS (supplementary fig. S10, Supplementary Material online).

We investigated past demography in more recent years by performing multiple sequentially Markovian coalescent (MSMC) analysis using the MSMC/MSMC2 software, which uses information on multiple haplotypes from each population. Figure 3 presents the results of four haplotypes from two Indian samples (IND03 and IND04) for CAS, eight haplotypes from four German samples (DEU01, DEU03, DEU04, and DEU06) for DOM, and eight haplotypes from four Korean samples (KOR01–03 and KOR05) for MUS. Because the number of analyzed haplotypes was smaller in CAS than in DOM and MUS, the demography of CAS in recent years (i.e., after ~10,000 years ago) may not be reliably estimated. To convert the generation number into years, we assumed a mutation rate of 5.7×10^{-9} per base pair per generation (Milholland et al. 2017) and a generation time of 1 year. Three subspecies followed different trajectories of population size change. The effective population size of CAS increased in an ancient time period around 100,000 years ago and later continued to decrease.

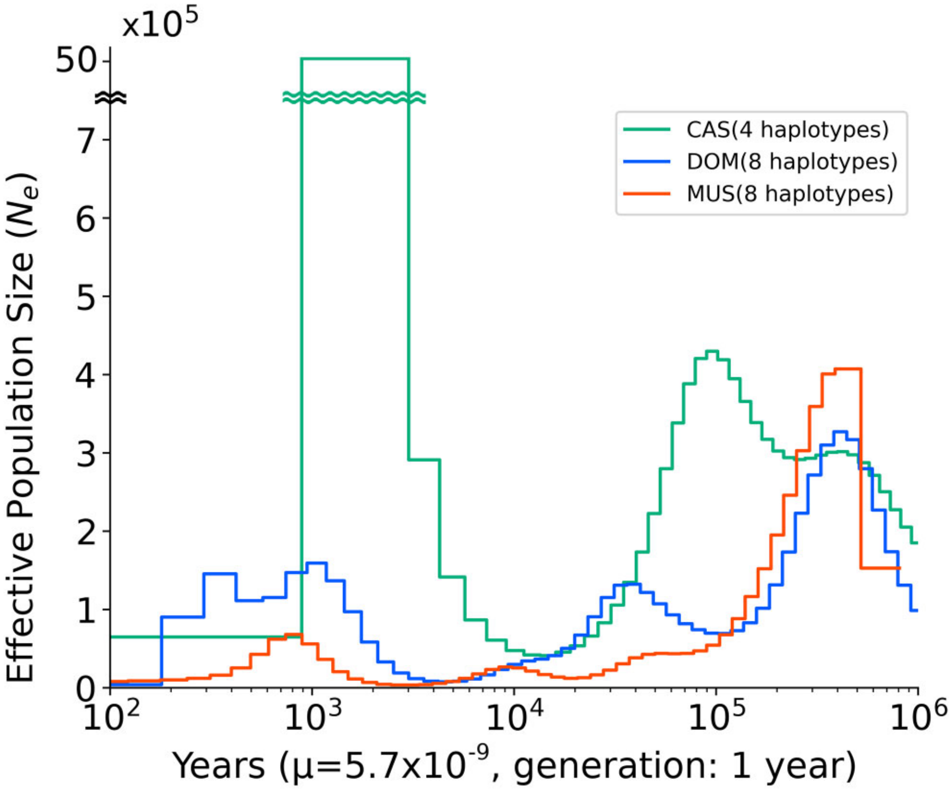

Fig. 3.—Inferred population sizes determined using MSMC analysis. x axis represents time before the present assuming a mutation rate of 0.57×10^{-8} per site per generation and a generation time of 1 year. y axis represents the effective population size. The red, green, and blue lines represent past population sizes of Korean (MUS), Indian (CAS), and German (DOM) samples, respectively.

After the decline, the effective population size of CAS reached its maximum level around 1,000–3,000 years ago. In contrast, both MUS and DOM genomes showed the signs of population shrinkage roughly 200,000 years ago. The trajectories of population sizes in DOM and MUS were similar prior to 100,000 years ago. DOM later experienced two rounds of population bottleneck and expansion, around 50,000 and 5,000 years ago. The trajectories of Korean MUS showed the signature of recent population bottleneck and expansion, probably between 2,000 and 4,000 years ago.

Genetic Relationship among Three Subspecies

Previous phylogenetic studies, in which partial or complete mitochondrial genome sequences were used, demonstrated a sister relationship between CAS–DOM (Geraldes et al. 2008; Li et al. 2021) and CAS–MUS (Jing et al. 2014), and genome sequencing studies of wild-derived mouse strains suggested a CAS–MUS clade (White et al. 2009; Keane et al. 2011; Phifer-Rixey et al. 2020). In this study, we conducted *f*-statistics analysis using representative samples of each subspecies, that is, IND04 for CAS, DEU01 for DOM, and KOR01 for MUS. We calculated outgroup f_3 statistics for all pairs among the three target samples with SPR used as an outgroup (fig. 4). A larger f_3 statistic value indicates that two subspecies share a larger amount of genetic drift, that is, a closer relationship exists between the two subspecies. The f_3 value of the CAS–MUS pair was statistically larger than that of the other comparisons between subspecies ($P = 3.12 \times 10^{-8}$ between CAS–MUS and CAS–DOM, $P = 6.22 \times 10^{-7}$ between CAS–MUS and DOM–MUS; Tukey's test). To validate this result, we conducted four population tests using f_4 statistics. The Z scores for the f_4 (SPR, DEU01; IND04, KOR01), f_4 (SPR, IND04; DEU01, KOR01), and f_4 (SPR, KOR01; DEU01, IND04) values were 2.003, 28.518, and 16.001, respectively (supplementary tables S10, S8, and S13, Supplementary Material online, respectively). These results support a close genetic relationship between CAS and MUS. Supplementary tables S15–S17 and S18–S23, Supplementary Material online show the f_3 statistics and four f_4 statistics for chromosome X, respectively.

We also constructed a neighbor-joining tree using pairwise genetic distances between samples, initially with all samples including SPR as the outgroup (supplementary figs. S11 and S12, Supplementary Material online). The initial tree showed a sister relationship between DOM and MUS with an exceptionally longer branch of the DOM clade, which contradicted the f_3 and f_4 statistics. However, two particular samples, CHN06 from Urumqi and RUS01 from Moscow, which exhibited a strong hybridization signature between MUS and DOM, potentially distorted the pattern. To solve this problem, we reconstructed a tree excluding all potential hybrid individuals, that is, a phylogenetic tree

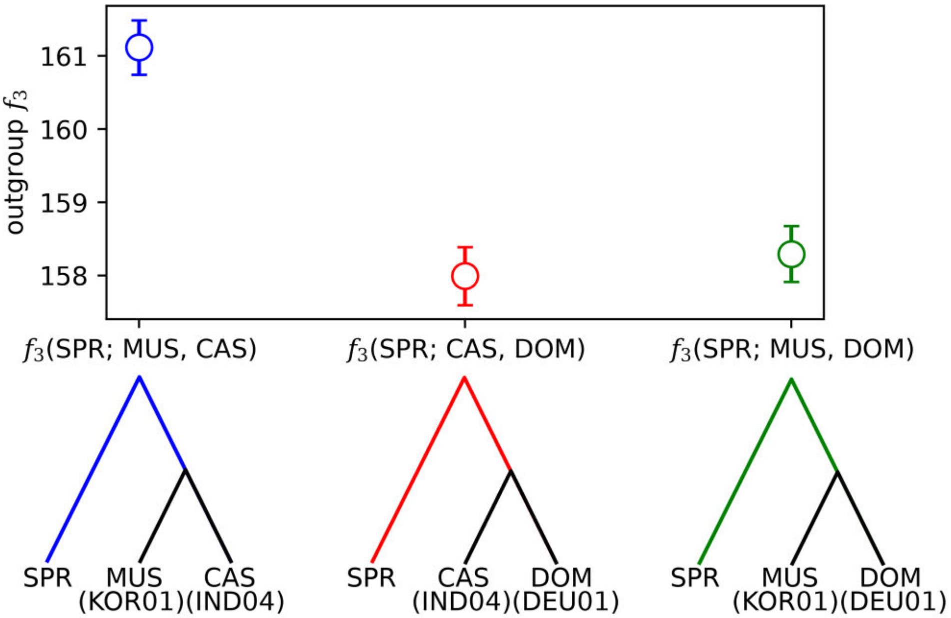

FIG. 4.—Genetic distances between subspecies calculated using outgroup f_3 statistics. *y* axis indicates the outgroup f_3 statistics with SPR used as the outgroup. Higher f_3 values indicate that the two target populations shared more genetic drift, implying that the two populations diverged more recently or that a larger amount of gene flow had occurred between the populations.

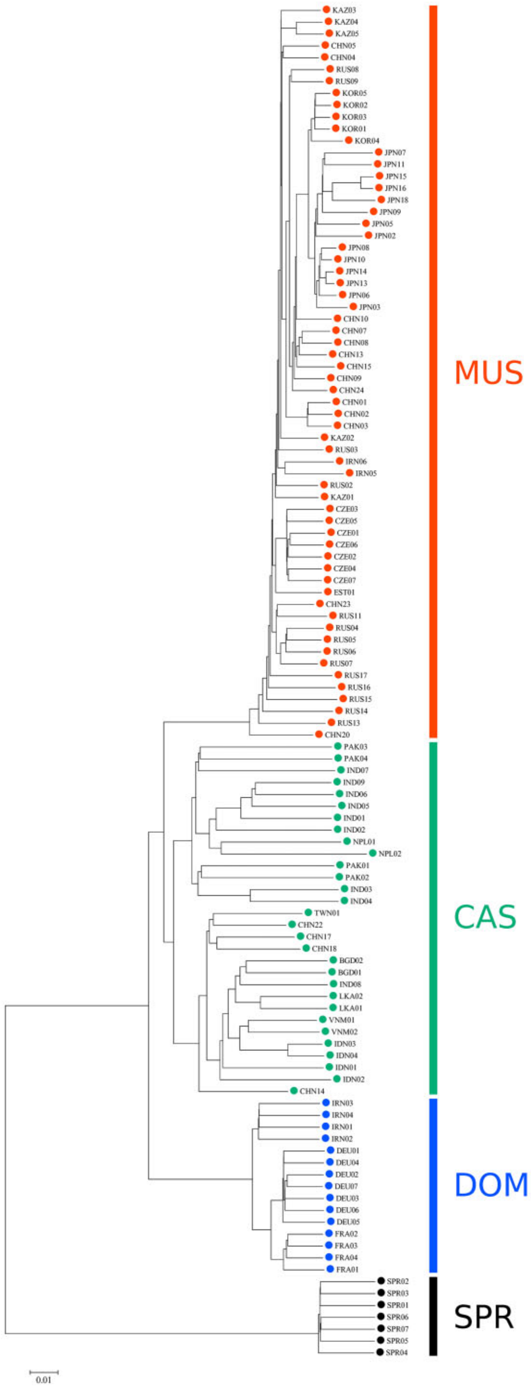

Fig. 5.—Neighbor-joining tree of the *M. musculus* subspecies. Hybrid individuals were excluded from the reconstruction. Red, green, and blue correspond to MUS, CAS, and DOM, respectively. SPR was used as the outgroup of *M. musculus*.

constructed using only individuals with >80% signal for one subspecies ancestry from the ADMIXTURE results; consequently, we obtained the pattern of the CAS–MUS clade (fig. 5).

Estimation of Divergence Times between Subspecies

To estimate the divergence timing of the three subspecies, we conducted cross-subspecific MSMC analysis (fig. 6). Initially, we used Indian, German, and Korean samples as the representative samples for each subspecies, but we found that the strong population bottleneck that occurred in the ancestors of Korean samples made it difficult to accurately infer the population history before the bottleneck. Therefore, we used Kazakhstan samples as representatives of MUS. Although Kazakhstan samples exhibited some level of admixture with DOM, this would not affect the estimated subspecies divergence time if the admixture occurred relatively recently (i.e., after 10,000 years ago). The divergence times between CAS and DOM, CAS and MUS, and DOM and MUS were separately estimated using time points when the relative cross-coalescent rate (rCCR) was 0.5. The divergence between CAS and MUS was most recent (95% CI: 187,365–188,647), consistent with the earlier analyses. The divergence times between CAS and DOM and DOM and MUS were almost equal, that is, 223,614–225,306 and 245,411–247,175 years ago (95% CIs), respectively.

Discussion

Although house mice have been widely used in biomedical research, the global genetic landscape of wild house mice has not been clear. Previous studies have included analysis of the genome sequences of western European (DOM), Middle Eastern (CAS), and North American (DOM) samples (Staubach et al. 2012; Harr et al. 2016; Mack et al. 2018). This is the first genome-wide study of wild house mice that includes all three subspecies and focuses on the genetic diversity across the Eurasian continent and Southeast Asian islands. Furthermore, we elucidated the present and ancestral population structure of the species.

Wild house mice have a much larger effective population size than humans (Geraldes et al. 2008, 2011; Halligan et al. 2010, 2013). Based on a phylogenetic tree excluding hybrid individuals (fig. 5), we calculated the nucleotide diversity in each subspecies: 0.527% for CAS, 0.244% for DOM, and 0.225% for MUS. Given that the average human nucleotide diversity is 0.08–0.12%, our results confirm that the nucleotide diversity of wild house mice is much higher than that of humans (Perry et al. 2012; Prado-Martinez et al. 2013; Arbiza et al. 2014). These values are consistent with those in previous studies (Geraldes et al. 2008; Phifer-Rixey et al. 2014; Harr et al. 2016), which showed that CAS has the highest diversity in all subspecies.

The PCA results presented in figure 1*b* revealed a wide spectrum of samples within the CAS and MUS genetic clines. In particular, the Japanese and Chinese samples were widely distributed along the PCA plot. Even samples

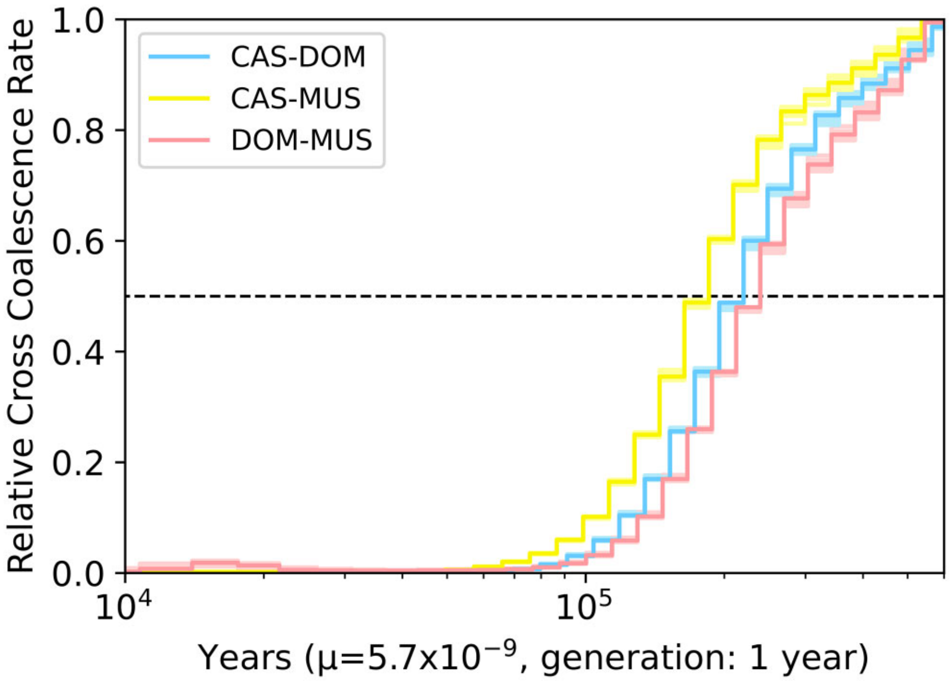

FIG. 6.—Multiple sequentially Markovian coalescent (MSMC) plot of each subspecies. Diagram showing the divergence process of the *M. musculus* subspecies by the cross-coalescence rate using to the MSMC method. *x* axis represents the time before the present assuming a mutation rate of 0.57 × 10^{-8} per site per generation and a generation time of 1 year. *y* axis represents the relative cross-coalescence rates (rCCRs). The magenta, cyan, and yellow lines correspond to the DOM–MUS, CAS–DOM, and CAS–MUS rCCR changes, respectively. The dotted line shows the rCCR = 0.5 point, which is heuristically identified as the estimated time at which the two populations split.

from other locations, such as Southeast Asia, demonstrated an admixture signature to some extent. The ADMIXTURE results (fig. 2) were largely in agreement with the PCA results. Our analysis using the f_3 and f_4 statistics showed that samples from India, Germany, and Korea had the smallest genetic component derived from different subspecies. Although these samples do not necessarily represent genuine subspecies with pure ancestry and without admixture between subspecies, this could be verified by additional sampling of wild mice in the Eurasian continent.

Our PSMC plots can be classified into three categories representing three subspecies, but relatively recent effective population size varied from sample to sample (supplementary fig. S10, Supplementary Material online). For example, in MUS, comparing CHN03 (China: Aksu), EST01 (Estonia: Tallinn), and IRN06 (Iran: Mashhad), all showed similar population size changes until ~50,000 years ago, but the Chinese Aksu population more recently experienced a strong bottleneck 20,000–30,000 years ago and then an increase in population size about 10,000 years ago (supplementary fig. S10, Supplementary Material online). A population size increase about 10,000 years ago was also observed in the Iranian Mashhad population, but the magnitude of population bottleneck and expansion was milder than that in the Chinese Aksu population. In contrast, the Estonian Tallinn population exhibited little increase in population size around 10,000 years ago; thus, the three populations experienced distinctly different region-dependent population size changes. As aforementioned, wild house mice in certain regions, such as Russia, the Japanese archipelago, and the Korean peninsula, were subject to an extreme population bottleneck, making it difficult to follow the past genetic demography of some individuals due to the substantial loss of polymorphic markers. For example, it was impossible to trace the population history of the RUS06 (Russia: Irkutsk), JPN13 (Japan: Ashiro), and KOR01 (Korea: Baengnyeong Island) samples before 100,000 years ago due to the strong bottleneck effect (supplementary fig. S10, Supplementary Material online).

Our results could substantially alter the simple trinity view of *M. musculus* subspecies. The observed pattern implies that the admixture between CAS and MUS has continued to occur for 10,000 years in Asia and that many "MUS-like CAS" and "CAS-like MUS" samples exist. This complex pattern has not previously been captured using mitochondrial phylogeny. Indeed, we observed many cases with incongruence between mitochondrial and autosomal genotypes. In particular, many Japanese samples harbor CAS-type mitochondrial and MUS-like nuclear genomes (fig. 1c). Based on a study of whole mitochondrial genome sequences, Li et al. (2021) suggested that the CAS-type mitochondrial genome was introduced to the Japanese archipelago in the late Neolithic period (~3,500 years ago) and that the MUS-type mitochondrial genome migrated later (~2,700 years ago) and quickly spread to the archipelago. The latter migration of house mice may coincide with the introduction of intense rice farming to the Japanese archipelago (Li et al. 2021). Such a pattern of nuclear–mitochondrial genotype mismatching has also been reported in house mice from New Zealand (Veale et al. 2018), Mainland China and Taiwan (Geraldes et al. 2008), and Madagascar (Fujiwara et al. 2021).

Interestingly, the incongruence between mitochondrial and autosomal genotypes was particularly prevalent among MUS. In samples with autosomal MUS genotypes, mitochondrial haplotypes corresponding to CAS, MUS, and DOM were identified. Previous studies reported a similar bias in the European hybrid zone of DOM–MUS, in which DOM-type haplotypes were observed in MUS, but MUS-type haplotypes were not observed in DOM (Božíková et al. 2005). Hybrid incompatibility associated with cytonuclear incompatibility is one possible explanation (Johnson 2010), but a reciprocal mtDNA transplant experiment revealed the opposite result, that is, that MUS-type (PWD) embryos with DOM-type (B6) mitochondria are more likely to be lethal than reciprocally transplanted embryos (Ma et al. 2016). Thus, additional ecological and experimental studies are required to reveal the cause of this bias.

Our analysis of past demography revealed different trajectories of population sizes in the three subspecies. Previous studies have suggested that the divergence of the three subspecies of *M. musculus* occurred 250,000 years ago for the substitution at the third codon position in mitochondrial Cyt b (Bonhomme and Searle 2012), 110,000–320,000 years ago for the isolation-with-migration model using whole-genome SNPs (Phifer-Rixey et al. 2020). Summarizing previous studies, the estimated divergence times fall within the range of 110,000–500,000 years ago (e.g., Boursot et al. 1996; Suzuki et al. 2004; Salcedo et al. 2007; Duvaux et al. 2011; Geraldes et al. 2011, 2008; Bonhomme and Searle 2012; Suzuki et al. 2013). In our analysis based on the MSMC results, we estimated that the divergence of the three subspecies occurred 187,000–247,000 years ago and the most recent subspecies divergence, that of CAS and MUS, occurred ~188,000 years ago, which are within the range of previous estimates. It should be noted that these estimates of divergence timing are highly dependent on the germline mutation rate and generation time used in our study.

Interestingly, both DOM and MUS experienced a recent strong population bottleneck and expansion, which was likely associated with the spread of agriculture. Entries into the expansion phase were much earlier in German DOM (4,000–6,000 years ago) than in Korean MUS (2,000–4,000 years ago), and this difference may reflect the different histories of agriculture in the two regions. In contrast, the population size of CAS was smallest at ~10,000–20,000 years ago, but the reduction was much lower than that in the other subspecies. A less severe population bottleneck would explain the higher genetic diversity of CAS.

This study elucidates the genome-wide relationship among *M. musculus* subspecies. Our f_3 and f_4 statistics support the close relationship between CAS and MUS, which is consistent with the study of White et al. (2009), in which dense genome-wide SNP data from wild-derived inbred mouse strains was used. In the presence of rampant gene flow between populations, it is difficult to infer population splitting. In particular, we found that the removal of obvious hybrid samples drastically altered the topology of the evolutionary relationship of subspecies (fig. 5 and supplementary fig. S12, Supplementary Material online). Additional analyses using different data types (f_3 statistics and MSMC; figs. 4 and 6) indicated that the CAS–MUS clade is most likely. As a note of caution, bifurcating tree construction methods that do not assume any gene flow among taxon could potentially lead to biased results if hybrid samples were included in the analysis.

M. musculus has been widely used as an animal model in evolutionary genetics and biomedical research. Revealing the genetic background and evolutionary history of this species will significantly contribute to the understanding of these models among research communities. In this study, we reported 98 novel whole-genome sequences of wild house mice, which were collected from a range of regions in the Eurasian continent and surrounding islands. Our analysis captured the genetic diversity of wild house mice on this continent, which has not previously been well studied, and revealed a complex pattern of admixture among three major subspecies. Indeed, the extent and geographic range of admixture was greater than previously thought, particularly between CAS and MUS; most of the samples from Southeast and East Asia showed some level of gene flow from different subspecies. The high-quality whole-genome sequencing data presented in this study will be important for future research on evolutionary ecology, population dynamics, and natural selection by introgression among subspecies of wild house mice.

Materials and Methods

Materials

Ninety eight wild house mouse samples from across the Eurasian continent and surrounding remote islands were collected. These samples are identical to those used by Li et al. (2021), who reported and analyzed whole mitochondrial genome sequences. In addition, whole-genome sequencing data from 35 wild *M. musculus* and 8 SPRs, previously reported by Harr et al. (2016), were downloaded from the European Nucleotide Archive (PRJEB9450, PRJEB11742, PRJEB14167, PRJEB2176, and PRJEB11897). In this study, the mitochondrial haplogroup of each sample was determined using the data provided by Li et al. (2021). For the mitochondrial sequences of Harr et al. (2016) dataset, a maximum likelihood phylogenetic tree was created along with the mitochondrial sequences determined by Li et al. (2021) to assign mitochondrial haplogroups. Supplementary tables S1–S3, Supplementary Material online include detailed information on the samples, and the geographical sampling locations are presented in supplementary fig. S1, Supplementary Material online.

Mapping Genomic Read Pairs and Single-Nucleotide Variant Calling

For the 98 wild house mouse samples, paired-end sequences of 100 bp in length were determined using the BGISEQ-500 platform. The quality of reads was checked and visualized using FastQC (Andrews 2010) and MultiQC (Ewels et al. 2016).

All raw reads were mapped to the GRCm38 (mm10) house mouse reference genome sequence using the bwa-mem algorithm with the "-M" option (Li and Durbin 2009). The samblaster program with "-M" option were used to mark PCR duplicate reads (Faust and Hall 2014). For the Harr et al. (2016) dataset, we used 43 samples with a median coverage of >20. All of the newly sequenced samples had a median coverage of at least >25 (supplementary table S3, Supplementary Material online). Raw SNV and insertion/deletion (indel) calls were performed using GATK4 HaplotypeCaller with the "-ERC GVCF" option (McKenna et al. 2010). All genomic variant call format (gVCF) files were merged using the CombineGVCFs function, and the variants of all samples were jointly called using the GenotypeGVCFs function.

Raw SNVs and indels were processed using GATK4 Variant Quality Score Recalibration (VQSR), a machine learning process that uses known variants as a training dataset and predicts whether a new variant is a true positive or false positive. To run GATK4 VQSR, we used the files "mgp.v3.snps.rsIDdbSNPv137.vcf.gz" and "mgp.v3.indels.rsIDdbSNPv137.vcf.gz," which were downloaded from the web server of the Sanger Institute (ftp://ftp-mouse.sanger.ac.uk/REL-1303-SNPs_Indels-GRCm38/), as training datasets for SNVs and indels, respectively. We also included hard-filtered SNV data as a training dataset. A hard-filtering process for SNVs was conducted using the following parameters: QD < 2.0; FS > 60.0; MQ < 40.0; MQRankSum < −12.5; and ReadPosRankSum < −8.0. In contrast, a HARD-filtering process for indels was performed using the following parameters: QD < 2.0; FS > 200.0; InbreedingCoeff < −0.8; ReadPosRankSum < −20.0; and SOR > 10.0. We assumed that SNVs and indels within the 90% tranche (i.e., 90% acceptance in all reliable training SNV datasets) were true-positive SNVs and indels, and these were used for downstream analyses. VQSR-passed SNVs were further filtered according to their mappability to the *M. musculus* reference genome, which was achieved using GenMap (Pockrandt et al. 2020). Mappability filtering retains highly unique regions in the reference genome. We computed mappability scores using the "-K 30" and "-E 2" options and analyzed sites with mappability values of 1.

Because we could not reliably distinguish the sex of some of our samples based on morphological records, we assigned sex based on the read depth coverages on the X and Y chromosomes (supplementary table S2, Supplementary Material online). We used samtools "depth" to calculate the coverage of each sample in the nonpseudoautosomal regions of the sex chromosomes that passed through the mappability filter. The ratios of average X-chromosomal to Y-chromosomal coverages exhibited a clear bimodal distribution, 0.96–1.15 and 7.43–195.20, in which the samples within the ranges likely represented males and females, respectively.

Kinship inference among samples was performed using KING with the "–kinship" option (Manichaikul et al. 2010). According to KING, the expected ranges of kinship coefficients were >0.354 for duplicate/monozygotic twins, >0.177 and <0.354 for first-degree relationships, >0.0884 and <0.177 for second-degree relationships, >0.0442 and <0.0884 for third-degree relationships, and <0.0442 for unrelated individuals. We excluded 13 samples (12 *M. musculus* and 1 SPR) with relationships closer than third degree. In total, 128 samples (94 of our samples and 34 public samples) were retained after the filtering process.

Synonymous and nonsynonymous variants were assigned using SnpEff (Cingolani, Platts, et al. 2012) and SnpSift (Cingolani, Patel, et al. 2012) with house mouse gene annotation data version "GRCm38.101 (ftp://ftp.ensembl.org/pub/release-101/gtf/mus_musculus/)." This was calculated by counting the number of synonymous and nonsynonymous variants on a gene-by-gene basis.

Population Structure Analysis

Using VCFtools (Danecek et al. 2011), SNVs were further filtered and converted to the PLINK format (Purcell et al. 2007) containing only biallelic autosomal SNVs and retaining sites successfully genotyped in all samples. Typically, SNVs in LD are excluded from the population structure analysis; however, in our main analysis, we did not eliminate these SNVs as our mouse samples were highly structured at the subspecies level, and the elimination of these SNVs have removed too many SNVs and prevented analysis. We also created PCA and ADMIXTURE plots without LD (supplementary figs. S4, S5, S8, and S9, Supplementary Material online) by specifying the following parameters: window size = 50 kb, window step size = 5, and variance inflation factor = 2. PCA was conducted using the smartpca program from Eigensoft (Patterson et al. 2006); default parameter settings were used with the exception of declining to remove outlier samples. The color codes in figure 1*a* and *b* were assigned according to Maxwell's color triangle. Outgroup f_3 statistics were computed using Admixtools, with the SPR population used as an outgroup via the "outgroupmode" option (Patterson et al. 2012). The f_4 statistics were also computed using Admixtools with the "f4 mode" and "printsd" options (Patterson et al. 2012). In addition, the ADMIXTURE (Alexander et al. 2009) software was used for population stratification. We computed cross-validation error values (–cv option) from $K = 1$ to $K = 5$ for datasets that either included or excluded SPR. The identity-by-state (IBS) distance matrix between all pairs of individuals was calculated using the PLINK "–distance 1-ibs" option, and the matrix was used to construct a neighbor-joining tree (Saitou and Nei 1987) using the Ape package in R (Paradis et al. 2004).

Demographic Inference Using PSMC and MSMC/MSMC2

The sampled individuals were subjected to PSMC analysis (Li and Durbin 2011). To obtain the required input for PSMC, a consensus autosomal genome sequence for each individual, the "mpileup" samtools command was applied to the dataset using the "-C 50, -O, -D 60, -d 10" options. PSMC analysis options (-t and -p) were selected according to the default settings suggested for the PSMC software. The time interval parameters were set to "4 + 25 * 2 + 4 + 6" with 25 iterations. To show the standard errors of the population size (N_e) estimates, we conducted 100 replications using the bootstrap method for the representative samples of each subspecies.

MSMC (Schiffels and Durbin 2014) version 2 (MSMC2: https://github.com/stschiff/msmc2; Schiffels and Wang 2020; Malaspinas et al. 2016) was used to estimate changes in N_e as well as the divergence time of subspecies. During the MSMC/MSMC2 analysis, we performed estimations using phased haplotype sequences as input. We estimated phased haplotypes using the Shapelt4 software (Delaneau et al. 2019). The "Mapping Data for G2F1-Based Coordinates" from "Mouse Map Converter (http://cgd.jax.org/mousemapconverter/)" were used to provide the recombination rate input file for MSMC2. Mappability was taken considered, and nonunique sequence positions were not used for calculations. The time interval parameters were set to "1 * 2 + 50 * 1 + 1 * 2 + 1 * 3" with 20 iterations. To estimate the divergence time of subspecies, we used two haplotypes from each population (four haplotypes in total) as an input for MSMC2. According to Shiffels and Durbin (2014), the rCCR variable should be between 0 and 1 (however, the calculation is unavoidably >1 in some cases), with a value close to 1 indicating that two populations were one population at a specific point in time. Heuristically, rCCR = 0.5 is considered to indicate the estimated time at which the two populations separated. Bootstrap was conducted by cutting the original input data into 5-Mb blocks and randomly sampling such blocks to artificially create a 3-Gbp-length genome. Of these artificially created datasets, 20 were used for analysis. To estimate N_e over time, we used eight haplotypes from four samples of each MUS (KOR01–03 and 05) and DOM (DEU01, DEU03, DEU04, and DEU06) population and four haplotypes from two samples of CAS (IND03 and IND04). To estimate subspecies divergence, we used four haplotypes from two populations for each combination of CAS–MUS (IND04 and KAZ01), DOM–MUS (DEU07 and KAZ01), and CAS–DOM (IND04 and DEU07).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We would like to express our sincere appreciation to Dr. Kazuo Moriwaki for collecting the valuable wild house mouse samples. This work was supported by MEXT KAKENHI (grant 18H05511 to N.O.), Cooperative Research of the Inter-University Research Institute Corporation. This research was also supported by Global Station for Big Data and Cybersecurity, a project of the Global Institution for Collaborative Research and Education at Hokkaido University.

Data Availability

The short-read sequence data generated in this study have been submitted to the DDBJ BioProject database (https://www.ddbj.nig.ac.jp/bioproject/) under accession number PRJDB11027. The datasets, parameter setting files, and scripts required for reproducing the analysis have been deposited in the Dryad digital repository (https://datadryad.org/) under doi:10.5061/dryad.66t1g1k1j.

Literature Cited

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

Andrews S. Babraham bioinformatics - FastQC a quality control tool for high throughput sequence data [cited 2022 May 13]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Arbiza L, Gottipati S, Siepel A, Keinan A. 2014. Contrasting X-linked and autosomal diversity across 14 human populations. Am J Hum Genet. 94:827–844.

Baines JF, Harr B. 2007. Reduced X-linked diversity in derived populations of house mice. Genetics. 175:1911–1921.

Bonhomme F, et al. 2007. Species-wide distribution of highly polymorphic minisatellite markers suggests past and present genetic exchanges among house mouse subspecies. Genome Biol. 8:R80.

Bonhomme F, et al. 2010. Genetic differentiation of the house mouse around the Mediterranean basin: matrilineal footprints of early and late colonization. Proc Biol Sci R Soc. 278:1034–1043.

Bonhomme F, Guenet J-L, Dod B, Moriwaki K, Bulfield G. 1987. The polyphyletic origin of laboratory inbred mice and their rate of evolution. Biol J Linn Soc. 30:51–58.

Bonhomme F, Searle JB. 2012. House mouse phylogeography. In: Macholán M, Baird SJE, Munclinger P, Piálek J, editors. Evolution of the house mouse. Cambridge: Cambridge University Press. p. 278–298.

Boursot P, et al. 1996. Origin and radiation of the house mouse: mitochondrial DNA phylogeny. J Evol Biol. 9:391–415.

Boursot P, Auffray JC, Britton-Davidian J, Bonhomme F. 1993. The evolution of house mice. Annu Rev Ecol Syst. 24:119–152.

Božíková E, et al. 2005. Mitochondrial DNA in the hybrid zone between *Mus musculus musculus* and *Mus musculus domesticus*: a comparison of two transects. Biol J Linn Soc. 84:363–378.

Cingolani P, et al. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly. 6:80–92.

Cingolani P, et al. 2012. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program. SnpSift. Front Genet. 3:35.

Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics. 27:2156–2158.

Delaneau O, Zagury J-F, Robinson MR, Marchini JL, Dermitzakis ET. 2019. Accurate, scalable and integrative haplotype estimation. Nat Commun. 10:5436.

Didion JP, de Villena FP-M. 2013. Deconstructing *Mus gemischus*: advances in understanding ancestry, structure, and variation in the genome of the laboratory mouse. Mamm Genome. 24:1–20.

Din W, et al. 1996. Origin and radiation of the house mouse: clues from nuclear genes. J Evolution Biol. 9:519–539.

Dod B, Smadja C, Karn RC, Boursot P. 2005. Testing for selection on the androgen-binding protein in the Danish mouse hybrid zone. Biol J Linn Soc. 84:447–459.

Ďureje Ľ, Macholán M, Baird SJE, Piálek J. 2012. The mouse hybrid zone in Central Europe: from morphology to molecules. Folia Zool. 61:308–318.

Duvaux L, Belkhir K, Boulesteix M, Boursot P. 2011. Isolation and gene flow: inferring the speciation history of European house mice. Mol Ecol. 20:5248–5264.

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 32:3047–3048.

Faust GG, Hall IM. 2014. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics. 30:2503–2505.

Frazer KA, et al. 2007. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature. 448:1050–1053.

Fujiwara K, et al. 2021. Whole-genome sequencing analysis of wild--caught house mice *Mus musculus* from Madagascar. BioRxiv. doi:10.1101/2021.09.10.459745.

Gabriel SI, Stevens MI, Mathias MdL, Searle JB. 2011. Of mice and 'convicts': origin of the Australian house mouse, *Mus musculus*. Plos One. 6:e28622.

Geraldes A, et al. 2008. Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. Mol Ecol. 17:5349–5363.

Geraldes A, Basset P, Smith KL, Nachman MW. 2011. Higher differentiation among subspecies of the house mouse (*Mus musculus*) in genomic regions with low recombination. Mol Ecol. 20: 4722–4736.

Halligan DL, et al. 2013. Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. PLoS Genet. 9:e1003995.

Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. 2010. Evidence for pervasive adaptive protein evolution in wild mice. Plos Genet. 6:e1000825.

Hardouin EA, et al. 2015. Eurasian house mouse (*Mus musculus L.*) differentiation at microsatellite loci identifies the Iranian plateau as a phylogeographic hotspot. BMC Evol Biol. 15:26.

Harr B, et al. 2016. Genomic resources for wild populations of the house mouse, *Mus musculus* and its close relative *Mus spretus*. Sci Data. 3:160075.

Jing M, et al. 2014. Phylogeography of Chinese house mice (*Mus musculus musculus/castaneus*): distribution, routes of colonization and geographic regions of hybridization. Mol Ecol. 23:4387–4405.

Johnson NA. 2010. Hybrid incompatibility genes: remnants of a genomic battlefield? Trends Genet. 26:317–325.

Jones EP, Eager HM, Gabriel SI, Jóhannesdóttir F, Searle JB. 2013. Genetic tracking of mice and other bioproxies to infer human history. Trends Genet. 29:298–308.

Jones EP, Jóhannesdóttir F, Gündüz İ, Richards MB, Searle JB. 2011. The expansion of the house mouse into north-western Europe. J Zool. 283:257–268.

Keane TM, et al. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature. 477:289–294.

Li Y, et al. 2021. House mouse *Mus musculus* dispersal in East Eurasia inferred from 98 newly determined complete mitochondrial genome sequences. Heredity. 126:132–147.

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

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

Liu YH, et al. 2008. Mosaic genealogy of the *Mus musculus* genome revealed by 21 nuclear genes from its three subspecies. Genes Genet Syst. 83:77–88.

Liu KJ, et al. 2015. Interspecific introgressive origin of genomic diversity in the house mouse. Proc Natl Acad Sci USA. 112:196–201.

Ma H, et al. 2016. Incompatibility between nuclear and mitochondrial genomes contributes to an interspecies reproductive barrier. Cell Metab. 24:283–294.

Macholn M, et al. 2007. Genetic analysis of autosomal and X-linked markers across a mouse hybrid zone. Evolution. 61:746–771.

Mack KL, Ballinger MA, Phifer-Rixey M, Nachman MW. 2018. Gene regulation underlies environmental adaptation in house mice. Genome Res. 28:1636–1645.

Malaspinas A-S, et al. 2016. A genomic history of aboriginal Australia. Nature. 538:207–214.

Manichaikul A, et al. 2010. Robust relationship inference in genome-wide association studies. Bioinformatics. 26:2867–2873.

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

Milholland B, et al. 2017. Differences between germline and somatic mutation rates in humans and mice. Nat Commun. 8:15183.

Moriwaki K, et al. 1984. Implications of the genetic divergence between European wild mice with Robertsonian translocations from the viewpoint of mitochondrial DNA. Genet Res. 43:277–287.

Moriwaki K, Miyashita N, Suzuki H, Kurihara Y, Yonekawa H. 1986. Genetic features of major geographical isolates of *Mus musculus*. In: Potter M, Nadeau JH, Cancro MP, editors. The wild mouse in immunology. Heidelberg: Springer. p. 55–61.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420:520–562.

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

Patterson N, et al. 2012. Ancient admixture in human history. Genetics. 192:1065–1093.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. Plos Genet. 2:e190.

Payseur BA, Krenz JG, Nachman MW. 2004. Differential patterns of introgression across the X chromosome in a hybrid zone between two species of house mice. Evolution. 58:2064–2078.

Perry GH, et al. 2012. Comparative RNA sequencing reveals substantial genetic variation in endangered primates. Genome Res. 22:602–610.

Phifer-Rixey M, et al. 2018. The genomic basis of environmental adaptation in house mice. Plos Genet. 14:e1007672.

Phifer-Rixey M, Bomhoff M, Nachman MW. 2014. Genome-wide patterns of differentiation among house mouse subspecies. Genetics. 198:283–297.

Phifer-Rixey M, Harr B, Hey J. 2020. Further resolution of the house mouse (*Mus musculus*) phylogeny by integration over isolation-with-migration histories. BMC Evol Biol. 20:120.

Pockrandt C, Alzamel M, Iliopoulos CS, Reinert K. 2020. GenMap: ultra-fast computation of genome mappability. Bioinformatics. 36:3687–3692.

Prado-Martinez J, et al. 2013. Great ape genetic diversity and population history. Nature. 499:471–475.

Prager EM, Orrego C, Sage RD. 1998. Genetic variation and phylogeography of Central Asian and other house mice, including a major new mitochondrial lineage in Yemen. Genetics. 150:835–861.

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

Rajabi-Maham H, et al. 2012. The south-eastern house mouse *Mus musculus castaneus* (Rodentia: Muridae) is a polytypic subspecies. Biol J Linn Soc. 107:295–306.

Raufaste N, et al. 2005. Inferences of selection and migration in the Danish house mouse hybrid zone. Biol J Linn Soc. 84:593–616.

Sage RD. 1981. Wild mice. In: Forester HL, Small JD, Fox JG, editors. The mouse in biomedical research. Vol. 1. New York: Academic Press. p. 40–90.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–25.

Salcedo T, Geraldes A, Nachman MW. 2007. Nucleotide variation in wild and inbred mice. Genetics. 177:2277–2291.

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

Schiffels S, Wang K. 2020. MSMC and MSMC2: the multiple sequentially Markovian coalescent. In: Dutheil JY, editor. Statistical population genomics. New York: Humana Press. p. 147–166.

Staubach F, et al. 2012. Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (*Mus musculus*). Plos Genet. 8:e1002891.

Suzuki H, et al. 2013. Evolutionary and dispersal history of Eurasian house mice *Mus musculus* clarified by more extensive geographic sampling of mitochondrial DNA. Heredity. 111:375–390.

Suzuki H, Shimada T, Terashima M, Tsuchiya K, Aplin K. 2004. Temporal, spatial, and ecological modes of evolution of Eurasian *Mus* based on mitochondrial and nuclear gene sequences. Mol Phylogenet Evol. 33:626–646.

Takada T, et al. 2013. The ancestor of extant Japanese fancy mice contributed to the mosaic genomes of classical inbred strains. Genome Res. 23:1329–1338.

Teeter KC, et al. 2008. Genome-wide patterns of gene flow across a house mouse hybrid zone. Genome Res. 18:67–76.

Teeter KC, et al. 2010. The variable genomic architecture of isolation between hybridizing species of house mice. Evolution. 64: 472–485.

Vanlerberghe F, Dod B, Boursot P, Bellis M, Bonhomme F. 1986. Absence of Y-chromosome introgression across the hybrid zone between *Mus musculus domesticus* and *Mus musculus musculus*. Genet Res. 48:191–197.

Veale AJ, Russell JC, King CM. 2018. The genomic ancestry, landscape genetics and invasion history of introduced mice in New Zealand. Roy Soc Open Sci. 5:170879.

Wang L, et al. 2011. Measures of linkage disequilibrium among neighbouring SNPs indicate asymmetries across the house mouse hybrid zone. Mol Ecol. 20:2985–3000.

White MA, Ané C, Dewey CN, Larget BR, Payseur BA. 2009. Fine-scale phylogenetic discordance across the house mouse genome. Plos Genet. 5:e1000729.

White MA, Steffy B, Wiltshire T, Payseur BA. 2011. Genetic dissection of a key reproductive barrier between nascent species of house mice. Genetics. 189:289–304.

Yang H, et al. 2011. Subspecific origin and haplotype diversity in the laboratory mouse. Nat Genet. 43:648–655.

Yang H, Bell TA, Churchill GA, de Villena FP-M. 2007. On the subspecific origin of the laboratory mouse. Nat Genet. 39:1100–1107.

Yonekawa H, et al. 1980. Relationship between laboratory mice and the subspecies *Mus musculus domesticus* based on restriction endonuclease cleavage patterns of mitochondrial DNA. Jpn J Genet. 55:289–296.

Yonekawa H, et al. 1981. Evolutionary relationships among five subspecies of *Mus musculus* based on restriction enzyme cleavage patterns of mitochondrial DNA. Genetics. 98:801–816.

Yonekawa H, et al. 1982. Origins of laboratory mice deduced from restriction patterns of mitochondrial DNA. Differentiation. 22: 222–226.

Yonekawa H, et al. 1988. Hybrid origin of Japanese mice "*Mus musculus molossinus*": evidence from restriction analysis of mitochondrial DNA. Mol Biol Evol. 5:63–78.

Associate editor: David Enard