Insights into *Mus musculus* subspecies population structure across Eurasia revealed by whole-genome sequence analysis

Kazumichi Fujiwara[1,2], Yosuke Kawai[3], Kazuo Moriwaki[4,5,†], Toyoyuki Takada[6], Toshihiko Shiroishi[4], Naruya Saitou[7,8], Hitoshi Suzuki[9,] Naoki Osada[1,2,*]

[1]Graduate School of Information Science and Technology, Hokkaido University, Sapporo, 060-0814, Japan

[2]Global Station for Big Data and Cybersecurity, GI-CoRE, Hokkaido University, Sapporo, 060-0814, Japan

[3]Genome Medical Science Project (Toyama), National Center for Global Health and Medicine (NCGM), Tokyo, 162-8655, Japan

[4]RIKEN BioResource Research Center, 3-1-1 Koyadai, Tsukuba, Ibaraki, 305-0074, Japan

[5]Mammalian Genetics Laboratory, National Institute of Genetics, Mishima, 411-8540, Japan

[6]Integrated BioResource Information Division, RIKEN BioResource Research Center, 3-1-1 Koyadai, Tsukuba, Ibaraki, 305-0074, Japan

[7]Population Genetics Laboratory, National Institute of Genetics, Mishima, 411-8540, Japan

[8]Faculty of Medicine, University of the Ryukyus, Nishihara, Okinawa, 903-0215, Japan

[9]Graduate School of Environmental Science, Hokkaido University, Sapporo, 060-0810, Japan

† Deceased

*correspondence to:

Naoki Osada, nosada@ist.hokudai.ac.jp

Keywords:

- ✓ Animal migration
- ✓ Ecology
- ✓ Evolution
- ✓ Evolutional Ecology
- ✓ Population Genetics

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 molecular phylogenetic studies suggested that three major subspecies originated in South Asia and spread across the world with the migration of modern humans in prehistoric and historic times. Relatively recent migration events (~10,000 years ago to present) have complicated the genetic landscape of wild house mice. In this study, we sequenced the whole genomes of 98 wild house mice collected from Eurasia, particularly East Asia, Southeast Asia, and South Asia. We found that although wild house mice consist of three major genetic groups corresponding to the three major subspecies, individuals representing admixture between subspecies are much more ubiquitous than previously recognized. Furthermore, several samples showed an incongruent pattern of genealogies between mitochondrial and autosomal genomes. Using samples likely retaining the original genetic components of subspecies, we estimated the pattern and timing of divergence among the subspecies. Our results provide critical understanding of the genetic diversity of wild house mice at the global level, which may be particularly useful for future biomedical and evolutionary research.

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 three wild subspecies (Bonhomme et al. 1987; Didion and Villena 2013; Moriwaki et al. 1984; Yang et al. 2007; Yonekawa et al. 1980, 1981, 1982): North Eurasian *Mus musculus musculus* (MUS), South Asian *Mus musculus castaneus* (CAS), and western European *Mus musculus domesticus* (DOM). The first mouse reference genome sequence was built using a classical inbred strain, C57BL/6J (Chinwalla et al. 2002). Subsequently, dozens of whole-genome sequences of laboratory mouse strains have since been published (https://www.sanger.ac.uk/data/mouse-genomes-project/). The C57BL/6J genome is approximately 68%, 6%, and 3% derived from DOM, MUS, and CAS, respectively, while the mitochondrial genome indicates a single lineage belonging to DOM (Frazer et al. 2007). Moreover, Frazer et al. estimated that 10% of the C57BL/6J genome is derived from *Mus musculus molossinus*, which may have resulted from hybridization between MUS and CAS (Takada et al. 2013; Yonekawa et al. 1988).

Despite intensive effort to sequence the genomes of laboratory mice, genetic diversity of wild house mice has not been thoroughly investigated. Previous studies found that wild house mice are highly genetically diverse and have an estimated effective population size on the order of 10^5 (Baines and Harr 2007; Geraldes et al. 2008, 2011; Halligan et al. 2010, 2013). Classical inbred strains represent only a small fraction of the genetic diversity of wild house mice (Salcedo et al. 2007). Recent studies revealed a genome-wide pattern of polymorphisms in wild house mice mostly focusing 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 would improve our understanding of the worldwide genetic diversity of house mice.

Wild house mice are distributed over most of the world, including remote islands. CAS inhabits the present-day Indian subcontinent and Southeast Asia. DOM inhabits North and South America, Africa, the Middle East, Australia, Southwestern Europe, and many surrounding and remote islands. MUS inhabits Siberia, Central Asia, East Asia, and Northeastern Europe. The homeland of *Mus musculus* was proposed to be in 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 Villena 2013), approximately 350,000–500,000 years ago (Boursot et al. 1996; Geraldes et al. 2008, 2011; Salcedo et al. 2007; Suzuki et al. 2004).

Throughout prehistoric human migration, the wild house mice migrated and lived commensally with humans. With prehistoric and historic long-distance migration, house mice, whose staple foods are grains, expanded their range with the development of agriculture and cultural exchange (Bonhomme et al. 2010; Gabriel et al. 2011; Jones et al. 2013; Moriwaki et al. 1986; Sage 1981). Movement with humans brought genetically diverse subspecies into the secondary contact (Boursot et al. 1993; Duvaux et al. 2011), allowing admixture of their genetics (Bonhomme et al. 2007; Liu et al. 2015; Staubach et al. 2012) despite partial reproductive isolation between subspecies, such as DOM and MUS (White et al. 2011). DOM and MUS come into contact along a narrow hybrid zone in Europe, while CAS and MUS seem to have a broader hybrid zone across Central and Eastern Asia (Boursot et al. 1993). Previous phylogenetic and phylogeographic studies analyzed mitochondrial DNA sequences and limited single nucleotide variant (SNV) data from house mice (e.g., (Liu et al. 2008)). However, the prevalence of hybridization between subspecies in the global population has not been well studied using genome-wide sequence data. In the pre-genomic era, Boursot et al. recognized that genetic and phenotypic boundaries between subspecies were obscure due to high variability within subspecies, except the boundary between DOM and MUS in western and central Europe (Boursot et al. 1993). The reason for this obscurity may be that only a limited number of autosomal loci had been analyzed at the time.

Rapid advancement in sequencing technologies has enabled the use of whole-genome data to estimate population structure and phylogenetic histories. In this study, we sequenced the whole genomes of 98 wild house mice previously collected from across the Eurasian continent and Southeast Asian islands, focusing on East, Southeast, and South Asian mouse samples covering 16 countries. Our analysis shows a prevalence of hybridization between subspecies beyond the hybrid zones that is greater than previously thought, particularly between CAS and MUS in East Asia. Our results provide key understanding of the genetic diversity of house mice at the global level for future biomedical and evolutionary research.

Results

Genetic Diversity of Mus musculus

In this study, we analyzed 141 whole-genome sequencing samples of *M. musculus* and *M. spretus*. These included 98 newly sequenced *M. musculus* samples from Eurasia, including those from Estonia, 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) (Figure S1 and Table S1). After pruning individuals with equal to or greater than third-degree kinship, which was inferred using the kinship coefficient, 128 samples were used for further analysis (94 of our newly sequenced samples were retained). The gender of our newly sequenced 94 samples are summarized in Table S2. After the variants were filtered, we obtained 101,133,391 SNVs and 41,654,670 indels for the dataset including 128 samples of *M. spretus* and *musculus*, and 108,995,343 SNVs and 31,295,428 indels for the dataset including 121 samples of *M. musculus*. The overall transition/transversion ratio of our samples was 2.18. The Method section describes the detailed filtering process.

As in Li et al. 2020, mitochondrial genome sequences were clustered into four distinct clades, three of which presumably correspond to CAS, DOM, and MUS subspecies (Li et al. 2020). The mitochondrial genomes of samples from Nepal (NPL01 and NPL02) diverged before the split between CAS and DOM clades. When we classified the samples according to mitochondrial haplogroups, CAS and MUS showed the highest and lowest per-sample nucleotide diversity (heterozygosity): CAS, 0.0052; MUS, 0.0025; DOM, 0.0043. The ratio of nonsynonymous to synonymous polymorphic sites for each sample was 0.415 on average and lower in CAS, which had the largest genetic diversity Table S1 summarizes the basic statistics for each individual, including per-sample nucleotide diversity.

Genetic structure of wild house mice

We first performed PCA using autosomal SNVs of all 128 samples, including those from *Mus spretus* (Figure S2). As shown in Figure S2, all *M. musculus* and *M. spretus* (SPR) were clearly differentiated along principal component 1 (PC1). Principal component 2 (PC2) corresponds to variation within *M. musculus* subspecies. In PC2, *M. musculus* subspecies are differentiated into two clusters with some intermediate samples. Mitochondrial haplotypes and sampling locations indicate that these two clusters largely correspond to the DOM and CAS-

MUS groups.

We subsequently excluded SPR and replotted PCA using 85,027,131 autosomal SNVs, as shown in Figure S3. Figure 1a shows the locations of samples colored according to the three genetic components identified by the PCA plot (Figure 1b). Considering the plot in Figure S3, these three clusters correspond to the CAS, DOM, and MUS subspecies, whereas a wide range of admixture between CAS and MUS clusters was observed. In the PCA plot, Nepalese samples with distinct mitochondrial haplogroups were clustered with CAS samples. PC1 shows the genetic differences between CAS and DOM, while PC2 shows the genetic differences between CAS and MUS. As shown in the plot, the prevalence of hybrid individuals is evident, particularly between CAS and MUS. The Chinese samples are widely scattered throughout a wide range of the CAS-MUS cline. The Japanese samples were also scattered along the CAS-MUS cline, although the range was narrower than in Chinese samples. We also performed PCA using only X-chromosomal SNVs (Figure S4). This pattern was the same as that using autosomal SNVs; however, the plots were more tightly clustered at each vertex, indicating that admixture was less pronounced on the X chromosomes.

As shown in Figure 1b, samples from Germany, Bangladesh, and Korea are located at the vertices of the triangle. However, this pattern does not necessarily imply that these samples are representatives of each of the three major subspecies. The pattern may rather be generated by the strong genetic drift in the three populations. In order to identify samples that represented the smallest amount of gene flow between subspecies, we computed f_3 and f_4 statistics among different individuals (Table S3–S5 and Table S6–S11). We found that Indian CAS samples from mountainous regions, western European DOM samples, and Korean MUS samples showed the smallest amount of gene flow with other subspecies. Therefore, we selected an Indian sample (IND04), a Korean sample (KOR01), and a German sample (DEU01) as our reference samples of CAS, MUS, and DOM subspecies, respectively. We also found that one sample collected from Kathmandu in Nepal (NPL02) was the most distantly related to CAS, MUS, and DOM samples. The other sample from Nepal (NPL01) showed a similar pattern, but NPL01 was more closely related to CAS than NPL02 was.

Except for these populations, most CAS-MUS samples exhibited some extent of admixture between subspecies. For example, an Indian sample from Delhi (IND02) was slightly genetically closer to MUS (KOR01) than IND04. The Z score of f_4 (SPR, KOR01; IND04, IND02) was 2.82 (Table S12). Likewise, other

CAS samples from neighboring regions, such as Pakistan and Bangladesh, were more similar to MUS (KOR01) than IND03, IND04, and IND07 individuals (Table S10 and Table S11). CAS samples from East and Southeast Asian regions also showed various levels of admixture with MUS genomes. Likewise, all MUS samples in Eastern Europe showed 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 14.72 and 11.105, respectively (Table S9). Most MUS samples, particularly samples from Northern China and Japan, showed a high level of admixture with CAS, while a limited number of MUS samples showed a low level of admixture with CAS genomes. Among DOM samples, those from outside of western Europe, such as Iranian and Russian samples, showed slightly but significantly closer affinity to MUS and CAS samples than the western European samples did (Table S8 and Table S9).

The PCA plot inferred using nuclear genome data was labeled with the four mitochondrial haplogroups (CAS, DOM, MUS, and NPL; Table S1) and is shown in Figure 1c. Most individuals showed congruent patterns in mitochondrial and nuclear genomes, but some samples showed incongruent patterns. Particularly, in the PCA plot, the mitochondria–nuclear 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 (RUS11) had a DOM-type mitochondrial genome but a nuclear genome that was highly similar to MUS.

We also performed an ADMIXTURE analysis to infer the genomic ancestries of *M. musculus* genomes (Figure 2). We analyzed 121 *M. musculus* samples by selecting K values from 1 to 7, where K is a predefined number of clusters. We also performed cross-validation to infer the most suitable K value (Figure S5), and we 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. Most Japanese and Chinese samples showed a hybrid pattern with CAS and MUS components. At $K = 4$, another genetic component represented the specific genetic features of Japanese and Korean mice. This may have been attributable to strong genetic drift in the Korean samples. This cluster corresponds to *M. musculus* ssp. *molossinus*. The ADMIXTURE results using X-chromosomal SNVs were essentially the same as those using autosomal SNVs (Figure S6).

Inference of past demography using PSMC

To estimate the past demographic pattern of wild house mice, we performed PSMC analysis. Figure S7

presents the PSMC plots of all samples. In Figure 3, we show the results of two Indian samples (IND03, IND04) and one Nepalese sample (NPL02) for CAS, seven German samples for DOM (DEU01–07), and four Korean samples for MUS (KOR01–03 and 05). We also show the PSMC plots of seven SPR samples from Spain. One Korean sample (KOR04) was excluded from the figure because the sample showed a substantially different pattern from the other Korean samples (Figure S7). Trajectories of past population size change were overlaid for all selected samples with the 100 replications by bootstrap method. We assumed a mutation rate of 5.7×10^{-9} per base pair per generation (Milholland et al. 2017). Although some of the samples showed unusual patterns, the trajectories of most samples were largely classified into three categories, each representing CAS, DOM, and MUS. Assuming a generation time of 1 year, the *M. musculus* subspecies started to diverge prior to 500,000 years ago. Note that the divergence time inferred from the PSMC plot gives only the lower bound. After the initial divergence, three subspecies followed different trajectories of population size change. The effective population size inferred from IDN03 and IND04 reached a maximum around 100,000 years ago and later continued to decrease. On the other hand, the population size inferred from NPL02 showed two major peaks in population size, 200,000 years ago and 50,000–60,000 years ago. The Indian and Nepalese samples differentiated prior to 500,000 years ago. In contrast, both MUS and DOM experienced population shrinkage after 500,000 years ago. The trajectories of population sizes in DOM and MUS were very similar prior to 100,000 years ago. DOM later experienced two rounds of population size expansion, around 50,000 and 5,000 years ago. Although Korean MUS experienced small population expansion events, their population size continued to decrease after 500,000 years ago. The trajectories of Korean MUS suggest a very recent population size expansion, probably around 1,000 years ago. However, the pattern was neither reliable nor reproducible among different samples. In contrast, MUS samples from Western China, such as CHN02 from Hotan, showed the signature of population expansion around 10,000 years ago, which was only faintly observed in the Korean and East European MUS samples (Figure S7). Note that our estimation strongly relied on assumptions of generation time and mutation rate, which could be variable on the order of two- to threefold.

Genetic relationship among three subspecies

Previous phylogenetic studies using partial or complete mitochondrial genome sequences showed a sister relationship between CAS-DOM (Li et al. 2020; Geraldes et al. 2008) and CAS-MUS (Jing et al. 2014), while

genome sequencing studies of wild-derived mouse strains suggested a CAS-MUS clade (White et al. 2009; Keane et al. 2011). First, we performed *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, making SPR an outgroup (Figure 4). A larger f_3 statistic value indicates that the two subspecies share a larger amount of genetic drift, which means a closer relationship between the two subspecies. The CAS-MUS pair showed an f_3 value that was statistically larger than 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 confirm this result, we performed four population tests using f_4 statistics. The Z scores for f_4 (SPR, DEU01; IND04, KOR01), f_4 (SPR, IND04; DEU01, KOR01), and f_4 (SPR, KOR01; DEU01, IND04) values were 1.97, 28.58, and 16.10, respectively (Table S8, Table S6, Table S11). These results support a close genetic relationship between CAS and MUS. Table S13–S15 and Table S16–S21 present the f_3 statistics and four f_4 statistics for chromosome X, respectively.

However, the close genetic relationship between CAS and MUS does not necessarily mean that they are sister subspecies. In fact, we found that historical gene flow between CAS and MUS has been much stronger than that between other pairs of subspecies. We subsequently constructed a neighbor-joining tree using all pairwise genetic distances between samples and SPR as the outgroup (Figure S8 and S9). The tree implies a sister relationship between DOM and MUS, which is contradictory to the f_3 and f_4 statistics.

We further evaluated demographic models using the FastSimCoal2 coalescent simulator. We assumed an isolation-with-migration model with three populations of piecewise-constant sizes and asymmetric migration rates (Figure 5). We used two CAS samples (IND03 and IND04), seven DOM samples (DEU01–DEU07), and four MUS samples (KOR01, KOR02, KOR03, and KOR05) for the analysis.

We performed parameter optimizations for all three subspecies topologies including CAS-MUS, CAS-DOM, and DOM-MUS clades. We found that the DOM-MUS clade model had the highest log-likelihood value among the three models (Table 1). Assuming a generation time of 1 year, we estimated the divergence between DOM and MUS to have occurred around 250,000 years ago (247,347–260,989), while the divergence of the subspecies started around 410,000 years ago (396,670–426,357). In the CAS-DOM and CAS-MUS clade models, two divergence events were estimated to have occurred in short time intervals around 300,000 years ago. The

estimated population sizes were consistent among the three models that is, ~500,000 for CAS, ~150,000 for DOM, and ~60,000 for MUS (Table 1). On the other hand, the estimated ancestral population sizes were model dependent. Specifically, the ancestral population sizes of the three subspecies were estimated to be ~150,000 in the DOM-MUS clade model and ~500,000 in the CAS-DOM and CAS-MUS clade models.

Genetic features of the Nepalese population

For PCA and ADMIXTURE plots, the samples collected from Nepal (NPL01 and NPL02) were assigned to the CAS group with a minor contribution from MUS. However, previous studies showed that the Nepalese samples carried deeply divergent mitochondrial genomes (Terashima et al. 2006; Li et al. 2020). Our f_3 analysis showed that NPL02 was the most distantly related to IND04, KOR01, and DEU01 samples. In order to clarify the genetic relationship of Nepalese sample to the other samples, we reconstructed admixture graphs using qpGraph software assuming the DOM-MUS clade. We tested two models where NPL02 was modeled as a sister of CAS or deeply diverged lineage ancestral to all three subspecies (Figure S10 and S11), both of which did not show significant deviation from the models and were equally likely. However, in the both models, we had to assume at least three admixture events on the graphs, indicating that the genetic relationship among three subspecies and the Nepalese population would be highly complex and cannot be represented by a simple tree structure.

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 analyzed genome sequences focusing on western European (DOM), Middle Eastern (CAS), and North American (DOM) samples (Harr et al. 2016; Mack et al. 2018; Staubach et al. 2012). This is the first genome-wide study of wild house mice including three subspecies that focuses on genetic diversity across the Eurasian continent and Southeast Asian islands. This study clarifies the present and ancestral population structure of the species. Wild house mice have a much larger effective population size than the human population (Geraldes et al. 2008, 2011; Halligan et al. 2010, 2013), and our data confirmed that nucleotide diversity of wild house mice is much higher than that of humans. The human average is 0.08% to 0.12% (Arbiza et al. 2014; Prado-Martinez et al. 2013; Perry et al. 2012), while *M. musculus* is in a range between 0.036% and 1.083%. Specifically, the CAS subspecies showed the highest per-sample nucleotide diversity, consistent with previous studies (Harr et al. 2016; Geraldes et al. 2008; Phifer-Rixey et al. 2014).

PCA results in Figure 1b show a wide spectrum of samples within CAS and MUS genetic clines. In particular, the Japanese and Chinese samples exhibit a wide range of distribution along the PCA plot. Even samples from other locations, such as Southeast Asia, showed an admixture signature to some extent. ADMIXTURE results (Figure 2) largely agreed with PCA results. Our analysis using f_3 and f_4 statistics showed that samples from India, Germany, and Korea had the smallest genetic component derived from different subspecies. However, these results do not necessarily mean that the samples represent genuine subspecies with pure ancestry and without admixture between subspecies. This could be clarified by further sampling of wild mice in the Eurasian continent.

Our results may substantially change the simple trinity view of *M. musculus* subspecies. The observed pattern implies that admixture between CAS and MUS has continued since 10,000 years ago in Asia and that many "MUS-like CAS" and "CAS-like MUS" samples exist. This complex pattern has not been captured by mitochondrial phylogeny. Indeed, we observed many cases with incongruence between mitochondrial genotypes and autosomal genotypes. In particular, many Japanese samples harbor CAS-type mitochondrial genomes and highly MUS-like nuclear genomes (Figure 1c). A recent study of whole mitochondrial genome sequences suggested that the CAS-type mitochondrial genome was introduced to the Japanese archipelago in

the late Neolithic period (~3800–3200 years ago) and that the MUS-type mitochondrial genome migrated later (~2700 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. 2020). Such a pattern of nuclear–mitochondrial genotype mismatching was also reported in New Zealand house mice (Veale et al. 2018). Mechanisms that explain the incongruence between mitochondrial and nuclear genotypes are largely unknown; however, sex-biased migration and/or natural selection may be underlying mechanisms.

This study clarifies 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 a study using dense genome-wide SNP data from wild-derived inbred mouse strains (White et al. 2009). PCMC results showed that the initial divergence of *M. musculus* subspecies occurred more than 400,000 years ago. As shown in Figure 3, the lower bound is obvious from the results of PSMC, even though the PSMC population size estimation was somewhat unreliable. Marine isotope stage (MIS) 12, which was the coldest period in the last 2 million years, started 478,000 years ago and may correspond to the isolation of subspecies during the glacial period (Lisiecki and Raymo 2005).

In the presence of rampant gene flow between populations, inference of population splitting is extremely difficult. Although f_3 statistics showed that CAS and MUS shared the highest amount of genetic drift, our neighbor-joining tree with no migration between subspecies and coalescent simulations allowing gene flow between subspecies supported the model of DOM-MUS clade. In the coalescent study, the close genetic relationship between CAS and MUS was explained by a substantially higher migration rate rather than recent divergence. Nevertheless, admixture graphs under the three different models were highly complicated, requiring caution in the interpretation of results (Figure S10 and S11).

Altogether, our analyses support a DOM-MUS clade rather than a CAS-MUS clade. However, our demographic models were very simple, and the actual history of the three subspecies is more complex than these models. Future studies using a larger number of samples with more complex models may be necessary to clarify these results.

Mus musculus has been widely used as an animal model of evolutionary genetic and biomedical research. Revealing the genetic background and evolutionary history of this species substantially contributes to the

knowledge of these models among research communities. In this study, we report 98 novel whole-genome sequences of wild house mice, which have been collected from a range of regions focusing on the Eurasian continent and surrounding islands. Our analysis captured the genetic diversity of wild house mice on the Eurasian continent, which has not been well studied, and revealed a complex pattern of admixture among the three major subspecies. In particular, our population genetics analyses suggest that the DOM and MUS subspecies pair diverged most recently, while the evidence of substantial admixture among subspecies was observed. The high-quality whole-genome sequencing data presented in this study has important research implications for future studies.

Methods

Materials

We collected 98 wild house mouse samples, including "Moriwaki battery" samples, from a wide range of the Eurasian continent and remote islands. These samples are identical to those studied by Li et al. 2020, who reported and analyzed whole mitochondrial genome sequences. In addition, whole-genome sequencing data from 35 wild house mice and 8 western Mediterranean mice (*Mus spretus*, SPR), reported by Harr et al. 2016, were downloaded from the European Nucleotide Archive (PRJEB9450, PRJEB11742, PRJEB14167, PRJEB2176, and PRJEB11897). Supplementary Table S1 lists the detailed information regarding the samples, and Figure S1 illustrates their geographical sampling locations.

Mapping genomic read pairs and single-nucleotide variant calling

For the 98 wild house mouse samples, paired-end sequences 100-bp in length were determined using the BGISEQ-500 platform (Li et al. 2020). The quality of reads was checked and visualized using FastQC (Andrews 2010) and MultiQC (Ewels et al. 2016) programs.

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 was used for marking PCR duplicate reads (Faust and Hall 2014). For publicly available datasets, we used 43 samples with a median coverage >20. Raw SNV and insertions/deletion (indel) calls were performed using the GATK4 HaplotypeCaller program with the "-ERC GVCF" option (McKenna et al. 2010). All genomic variant call format (gVCF) files were merged using the GenomicDBImport function, and the variants of all samples were jointly called using the GenotypeGVCFs function.

Raw SNVs and indels were processed using the GATK4 Variant Quality Score Recalibration (VQSR). VQSR is a machine learning process that uses known variants as a training dataset and predicts whether a new variant would be a true positive or a false positive. In order to run GATK4 VQSR, we used the "mgp.v3.snps.rsIDdbSNPv137.vcf.gz" and "mgp.v3.indels.rsIDdbSNPv137.vcf.gz" files downloaded from the webserver 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 performed using the following parameters: QD < 2.0, FS > 60.0, MQ <

40.0, MQRankSum < −12.5, and ReadPosRankSum < −8.0. Moreover, 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 (90% acceptance in all reliable training SNVs datasets) were true positive SNVs and indels, and we used them for downstream analyses. VQSR-passed SNVs were further filtered according to mappability to the *Mus musculus* reference genome, computed using the GenMap software (Pockrandt et al. 2020). We computed mappability scores using the "-K 30" and "-E 2" options and analyzed sites with mappability values of 1. Mappability filtering retains highly unique regions in the reference genome.

Because we could not reliably distinguish the sexes of some of our samples from morphological records, we assigned the sexes of samples based on the read depth coverages on the X and Y chromosomes (Table S2). We used Samtools Depth to calculate the coverage of each sample in the non-pseudoautosomal regions of the sex chromosomes that passed the mappability filter. The ratios of average X-chromosomal to Y-chromosomal coverages showed a clear bimodal distribution, 0.96–1.15 and 7.43–195.20, where samples within the former and latter ranges likely represent male and female samples, respectively.

Kinship inference among samples was performed using KING software with the "—kinship" option (Manichaikul et al. 2010). According to KING software, the expected ranges of kinship coefficients are >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. The 13 samples (12 *Mus musculus* samples and 1 *Mus spretus* sample) were excluded with relationships closer than third-degree. In total, 128 samples (94 of our samples and 34 public samples) were kept after filtering process.

Synonymous and nonsynonymous variants were assigned using the SnpEff (Cingolani et al. 2012b) and SnpSift (Cingolani et al. 2012a) programs with house mouse gene annotation data version "GRCm38.101 (ftp://ftp.ensembl.org/pub/release-101/gtf/mus_musculus/)."

Population structure analysis

Using VCFtools (Danecek et al. 2011), SNVs were further filtered and converted to PLINK format (Purcell et al. 2007) containing only biallelic autosomal SNVs, retaining sites successfully genotyped in all samples.

Typically, SNVs in linkage disequilibrium are excluded from analysis of population structure; however, we did not eliminate these SNVs because our mouse samples were highly structured at the subspecies level and eliminating these SNVs would have left too few SNVs for analysis. Principal component analysis (PCA) was performed using the smartpca program (Patterson et al. 2006). Default parameter settings were used with the exception of declining to remove outlier samples. The color codes in Figure 1A and 1B were assigned according to Maxwell's color triangle. Pairwise F_{st} values between all pairs of populations were calculated using smartpca. Outgroup f_3 statistics were computed using Admixtools, employing the SPR population as an outgroup with 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). ADMIXTURE (Alexander et al. 2009) software was used for population stratification. We computed cross-validation error values from $K = 1$ to $K = 7$ 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 software

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

Demographic modeling using FastSimCoal2

Demographic parameter estimation and model selection were performed using the composite likelihood method implemented in FastSimCoal2 (Excoffier et al. 2013). Autosomal biallelic SNVs were used for the analysis. To reduce the effect of natural selection, exonic and flanking (within 10kb from exons) sites were filtered from the analysis, yielding a total of 563,035,033 sites. As *Mus spretus* is genetically divergent from house mice, we used site frequency spectrums of minor allele frequencies jointly determined among samples.

For each model, we generated 100,000 genealogies and ran 40 expectation maximization cycles. We repeated the above process 300 times with different seed values and adopted the demographic parameters with the highest log-likelihood.

Admixture graph construction

We constructed admixture graphs using qpGraph software and the default parameter settings (Lipson and Reich 2017). When we added migration edges, we assumed that the migration occurred from a population with a larger population size to that with a smaller population size. The migration edges were added until the model showed no deviation in terms of f_4 statistics.

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 are deposited to Dryad digital repository (https://datadryad.org/) under doi (10.5061/dryad.66t1g1k1j).

Competing interest statement

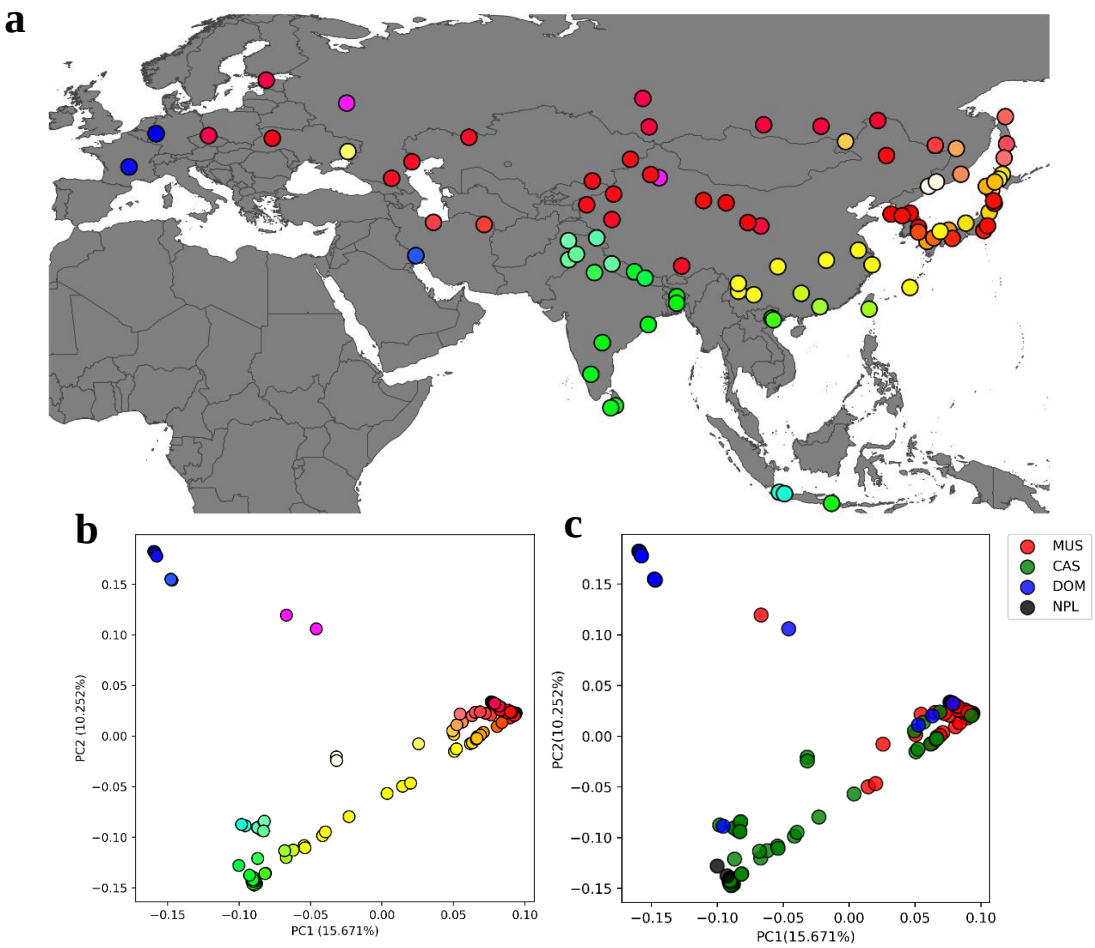
The authors declare no competing interests.

Acknowledgments

This work was partly supported by MEXT KAKENHI (grant 18H05511 to N.O.) and Cooperative Research of the Inter-University Research Institute Corporation. This research was supported by Global Station for Big Data and Cybersecurity, a project of Global Institution for Collaborative Research and Education at Hokkaido University.

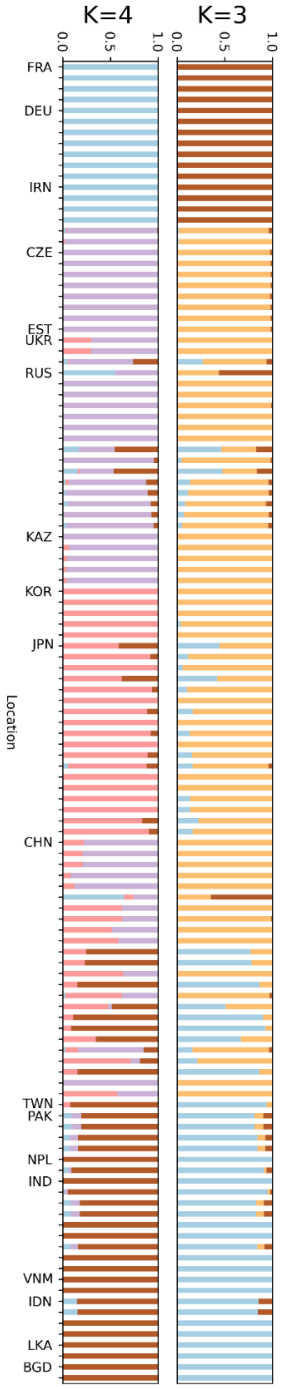
Figures

Figure 1



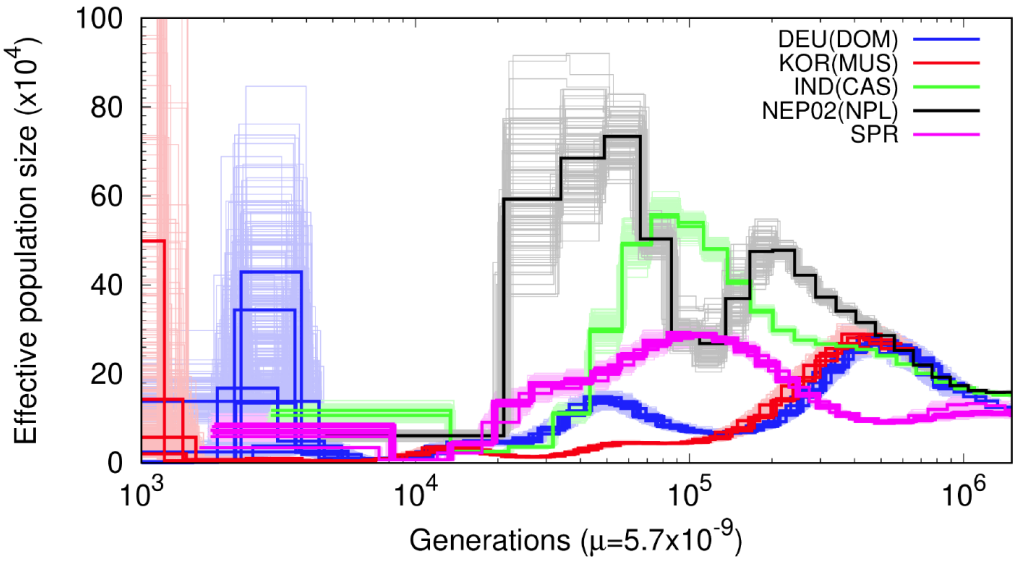
PCA results in wild house mice. (a) A geographical map of sampling locations. The colors of circles correspond to the eigenvalues in panel B. Figure S1 shows the detailed names of sample collection sites. (b) PCA plot of wild house mice using autosomal SNVs. The *x* and *y* axes represent the eigenvalues of PC1 and PC2, respectively. The circles are colored according to "Maxwell's color triangle," assigning three vertices to the RGB colors. The red, green, and blue color intensities correspond to MUS, CAS, and DOM genetic components, respectively.-The proportion of variance for each eigenvalue is shown in parentheses on the axis label. (C) PCA plot of wild house mice using autosomal SNVs labeled with the mitochondrial genome haplogroup of each sample. The proportion of variance for each eigenvalue is shown in parentheses on the axis label.

Figure 2



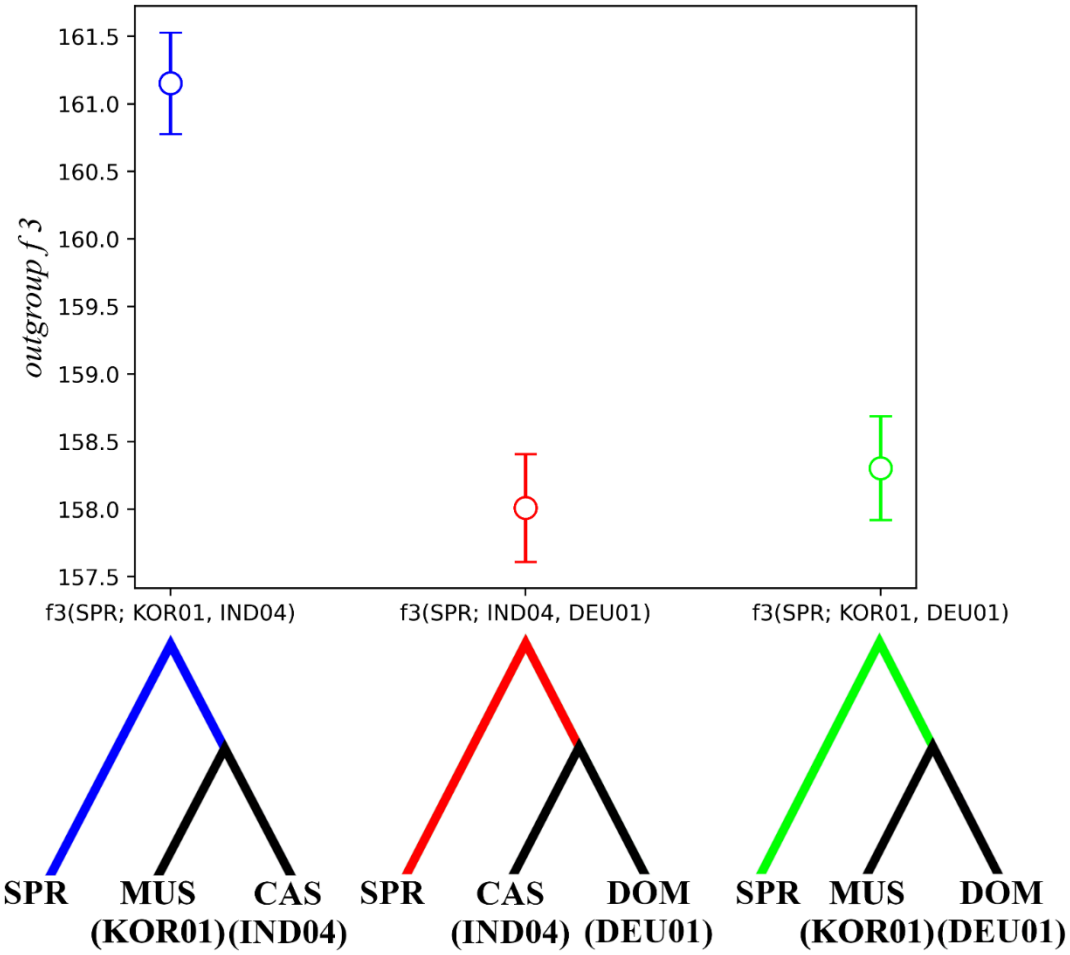
ADMIXTURE plot of autosomes showing the proportion of estimated subspecies genetic components. Results for K = 3 and K = 4 are presented. Samples with the same country code are ordered according to sampling site from west to east and north to south.

Figure 3



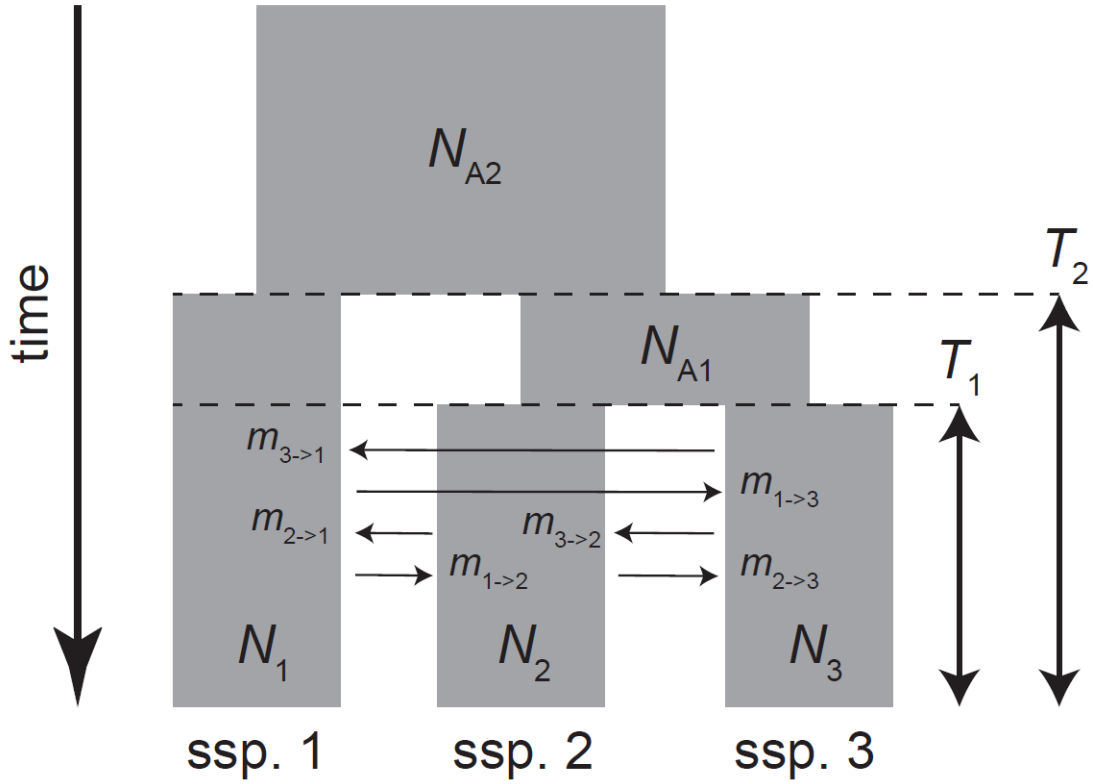
Inferred population sizes as determined using PSMC analysis. The *x* axis represents time before present scaled by the mutation rate of 0.57×10^{-8} per site per generation, and the *y* axis represents the effective population size. The red, green, blue, magenta, and black lines represent past population sizes of Korean (MUS), Indian (CAS), German (DOM), *M. spretus* (SPR), and Nepalese (NPL) samples, respectively. The lines with lighter color represent the bootstrap results of 100 replications.

Figure 4



Genetic distances between subspecies using outgroup f_3 statistics. The *y* axis indicates the outgroup f_3 statistics using SPR 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 experienced a larger amount of gene flow between populations.

Figure 5



A demographic model of three populations used for estimating population genetics parameters and model selection. We assumed that ssp.2 and ssp.3 diverged most recently. Migration rates (*m*, probability of gene migration) between populations are asymmetric. Population sizes changed after the isolation of populations but remained constant after the population split. Five population size parameters, two divergence time parameters, and six migration rate parameters were estimated.

Tables

Table 1. Estimation of demographic parameters*

(ssp1, ssp2), ssp3	(CAS, MUS), DOM	(CAS, DOM), MUS	(DOM, MUS), CAS
N_{CAS}	507,325 (504,593–522,052)	491,173 (486,261–506,695)	437,774 (437,729–451,133)
N_{DOM}	148,233 (148,044–152,706)	150,495 (150,463–154,080)	148,804 (147,300–152,410)
N_{MUS}	58,630 (58,088–60,824)	56,181 (56,049–57,987)	53,991 (53,566–55,669)
N_{A1}	131,025 (23,335–112,054)	101,281 (26,904–103,028)	263,065 (237,986–286,595)
N_{A2}	504,864 (503,815–519415)	503,517 (502,054–516,193)	153,653 (149,912–161,445)
T_1^{**}	304,562 (300,185–317,221)	322,140 (320,552–334,166)	253,358 (247,347–260,989)
T_2^{**}	304,757 (300651–317401)	322,335 (320,862–334,607)	413,019 (396,670–426,357)
*ln*L	−84,838,731	−84,838,737	−84,834,056

*All estimated parameters including migration rates are presented in Table S22, and 95% bootstrap confidence intervals are shown in parentheses.

**number of generations

Reference

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**: 1655–64.

Andrews S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data. *undefined*. 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 Genetics* **94**: 827–844.

Baines JF, Harr B. 2007. Reduced X-Linked Diversity in Derived Populations of House Mice. *Genetics* **175**: 1911–1921.

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, Orth A, Cucchi T, Rajabi-Maham H, Catalan J, Boursot P, Auffray J-C, Britton-Davidian J. 2010. Genetic differentiation of the house mouse around the Mediterranean basin: matrilineal footprints of early and late colonization. *Proc Biological Sci Royal Soc* **278**: 1034–43.

Bonhomme F, Rivals E, Orth A, Grant GR, Jeffreys AJ, Bois PR. 2007. Species-wide distribution of highly polymorphic minisatellite markers suggests past and present genetic exchanges among house mouse subspecies. *Genome Biol* **8**: R80.

Boursot P, Auffray J-C, Britton-Davidian J, Bonhomme F. 1993. The Evolution of House Mice. *Annu Rev Ecol Syst* **24**: 119–152.

Boursot P, Din W, Anand R, Darviche D, Dod B, Deimling FV, Talwar GP, Bonhomme F. 1996. Origin and radiation of the house mouse: mitochondrial DNA phylogeny. *J Evolution Biol* **9**: 391–415.

Chinwalla AT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, Graves TA, Hillier LW, Mardis ER, McPherson JD, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–562.

Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X. 2012a. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Frontiers Genetics* **3**: 35.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012b. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**: 80–92.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinform Oxf Engl* **27**: 2156–8.

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

Din W, Anand R, Boursot P, Darviche D, Dod B, Jouvin-Marche E, Orth A, Talwar GP, Cazenave P-A, Bonhomme F. 1996. Origin and radiation of the house mouse: clues from nuclear genes. *J Evolution Biol* **9**: 519–539.

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. *Bioinform Oxf Engl* **32**: 3047–8.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust Demographic Inference from Genomic and SNP Data. *Plos Genet* **9**: e1003905.

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

Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, Gupta RV, Montgomery J, Morenzoni MM, Nilsen GB, et al. 2007. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **448**.

Gabriel SI, Stevens MI, Mathias M da L, Searle JB. 2011. Of Mice and 'Convicts': Origin of the Australian House Mouse, Mus musculus. *Plos One* **6**: e28622.

Geraldes A, Basset P, Gibson B, Smith KL, Harr B, Yu H-T, Bulatova N, Ziv Y, Nachman MW. 2008. Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. *Mol Ecol* **17**: 5349–63.

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: RECOMBINATION AND SPECIATION IN MICE. *Mol Ecol* **20**: 4722–4736.

Halligan DL, Kousathanas A, Ness RW, Harr B, Eöry L, Keane TM, Adams DJ, Keightley PD. 2013. Contributions of Protein-Coding and Regulatory Change to Adaptive Molecular Evolution in Murid Rodents. *PLoS Genetics* **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.

Harr B, Karakoc E, Neme R, Teschke M, Pfeifle C, Pezer Ž, Babiker H, Linnenbrink M, Montero I, Scavetta R, et al. 2016. Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus. *Scientific Data* **3**.

Jing M, Yu H, Bi X, Lai Y, Jiang W, Huang L. 2014. Phylogeography of Chinese house mice (Mus musculus musculus/castaneus): distribution, routes of colonization and geographic regions of hybridization. *Mol Ecol* **23**: 4387–4405.

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

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, et al. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**: 289–294.

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**.

Li Y, Fujiwara K, Osada N, Kawai Y, Takada T, Kryukov AP, Abe K, Yonekawa H, Shiroishi T, Moriwaki K, et al. 2020. House mouse Mus musculus dispersal in East Eurasia inferred from 98 newly determined complete mitochondrial genome sequences. *Heredity* 1–16.

Lipson M, Reich D. 2017. A Working Model of the Deep Relationships of Diverse Modern Human Genetic Lineages Outside of Africa. *Mol Biol Evol* msw293.

Lisiecki LE, Raymo ME. 2005. A Pliocene-Pleistocene stack of 57 globally distributed benthic δ18O records. *Paleoceanography* **20**: n/a-n/a.

Liu KJ, Steinberg E, Yozzo A, Song Y, Kohn MH, Nakhleh L. 2015. Interspecific introgressive origin of genomic diversity in the house mouse. *P Natl Acad Sci Usa* **112**: 196–201.

Liu Y-H, Takahashi A, Kitano T, Koide T, Shiroishi T, Moriwaki K, Saitou N. 2008. Mosaic genealogy of the Mus musculus genome revealed by 21 nuclear genes from its three subspecies. *Genes Genet Syst* **83**: 77–88.

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

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. 2010. Robust relationship inference in genome-wide association studies. *Bioinform Oxf Engl* **26**: 2867–73.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–1303.

Milholland B, Dong X, Zhang L, Hao X, Suh Y, Vijg J. 2017. Differences between germline and somatic mutation rates in humans and mice. *Nat Commun* **8**: 15183.

Moriwaki K, Miyashita N, Suzuki H, Kurihara Y, Yonekawa H. 1986. Genetic Features of Major Geographical Isolates of Mus musculus. In *undefined* (eds. ["Michael Potter", "Joseph H. Nadeau", and "Michael P. Cancro"]), pp. 55–61, Springer Berlin Heidelberg, Berlin, Heidelberg.

Moriwaki K, Yonekawa H, Gotoh O, Minezawa M, Winking H, Gropp A. 1984. Implications of the genetic divergence between European wild mice with Robertsonian translocations from the viewpoint of mitochondrial DNA. *Genet Res* **43**: 277–287.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**: 289–290.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 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.

Perry GH, Melsted P, Marioni JC, Wang Y, Bainer R, Pickrell JK, Michelini K, Zehr S, Yoder AD, Stephens M, et al. 2012. Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Res* **22**: 602–610.

Phifer-Rixey M, Bi K, Ferris KG, Sheehan MJ, Lin D, Mack KL, Keeble SM, Suzuki TA, Good JM, Nachman MW. 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.

Pockrandt C, Alzamel M, Iliopoulos CS, Reinert K. 2020. GenMap: Ultra-fast Computation of Genome Mappability. *Bioinformatics* **36**: 3687–3692.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al. 2013. Great ape genetic diversity and population history. *Nature* **499**: 471–475.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, Bakker PIW de, Daly MJ, et al. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genetics* **81**: 559–575.

Sage RD. 1981. Wild mice. In *The mouse in biomedical research* (eds. H.L. Forester, J.D. Small, and J.G. Fox), Vol. 1 of, pp. 40–90, Academic Press, New York.

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.

Staubach F, Lorenc A, Messer PW, Tang K, Petrov DA, Tautz D. 2012. Genome Patterns of Selection and Introgression of Haplotypes in Natural Populations of the House Mouse (Mus musculus). *Plos Genet* **8**: e1002891.

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, Ebata T, Noguchi H, Keane TM, Adams DJ, Narita T, Shin-I T, Fujisawa H, Toyoda A, Abe K, et al. 2013. The ancestor of extant Japanese fancy mice contributed to the mosaic genomes of classical inbred strains. *Genome Res* **23**: 1329–38.

Terashima M, Furusawa S, Hanzawa N, Tsuchiya K, Suyanto A, Moriwaki K, Yonekawa H, Suzuki H. 2006. Phylogeographic origin of Hokkaido house mice (Mus musculus) as indicated by genetic markers with maternal, paternal and biparental inheritance. *Heredity* **96**: 128–138.

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.

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

Yonekawa H, Moriwaki K, Gotoh O, Hayashi J, Watanabe J, Miyashita N, Petras ML, Tagashira Y. 1981. Evolutionary relationships among five subspecies of Mus musculus based on restriction enzyme cleavage patterns of mitochondrial DNA. *Genetics* **98**: 801–16.

Yonekawa H, Moriwaki K, Gotoh O, Miyashita N, Matsushima Y, Shi L, Cho WS, Zhen X-L, Tagashira Y. 1988. Hybrid origin of Japanese mice “Mus musculus molossinus”: evidence from restriction analysis of mitochondrial DNA. *Mol Biol Evol* **5**: 63–78.

Yonekawa H, Moriwaki K, Gotoh O, Miyashita N, Migita S, Bonhomme F, Hjorth JP, Petras ML, Tagashira Y. 1982. Origins of Laboratory Mice Deduced from Restriction Patterns of Mitochondrial DNA. *Differentiation* **22**: 222–226.

Yonekawa H, Moriwaki K, Gotoh O, Watanabe J, Hayashi J, Miyashita N, Petras ML, Tagashira Y. 1980. Relationship Between Laboratory Mice and the Subspecies Mus musculus domesticus Based on Restriction Endonuclease Cleavage Patterns of Mitochondrial DNA. *Jpn J Genetics* **55**: 289–296.