<table>
<thead>
<tr>
<th>Title</th>
<th>Ancient onset of geographical divergence, interpopulation genetic exchange, and natural selection on the Mc1r coat colour gene in the house mouse (Mus musculus)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Kodama, Sayaka; Nunome, Mitsuo; Moriwaki, Kazuo; Suzuki, Hitoshi</td>
</tr>
<tr>
<td>Citation</td>
<td>Biological journal of the Linnean Society, 114(4), 778-794</td>
</tr>
<tr>
<td>Issue Date</td>
<td>2015-04-14</td>
</tr>
<tr>
<td>Doc URL</td>
<td><a href="http://hdl.handle.net/2115/61183">http://hdl.handle.net/2115/61183</a></td>
</tr>
<tr>
<td>Type</td>
<td>article</td>
</tr>
<tr>
<td>File Information</td>
<td>Ancient onset of geographic divergence, interpopulation genetic exchange, and natural selection on the Mc1r coat colour gene in the house mouse (Mus musculus).pdf</td>
</tr>
</tbody>
</table>
Ancient onset of geographic divergence, interpopulation genetic exchange, and natural selection on the McIr coat-colour gene in the house mouse (Mus musculus)β

SAYAKA KODAMA¹, MITSUO NUNOME², KAZUO MORIWAKI³† and HITOSHI SUZUKI¹*

¹Laboratory of Ecology and Genetics, Graduate School of Environmental Earth Science, Hokkaido University, Kita-ku, Sapporo 060-0810, Japan
²Laboratory of Animal Genetics, Department of Applied Molecular Biosciences, Graduate School of Bioagricultural Sciences, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
³RIKEN, Bioresource Center, Tsukuba, Ibaraki 305-0074, Japan

*Corresponding author. E-mail: htsuzuki@ees.hokudai.ac.jp
†Deceased

Running head: HYBRID ORIGIN OF HOUSE MICE IN THEIR HOMELAND

We examined genetic variation in house mice from India and Pakistan, a predominant part of the predicted homeland of this species and the territory of the subspecies *Mus musculus castaneus* (CAS), using a nuclear marker for seven tandemly arranged genes (*Fanca–Spire2–Tcf25–Mc1r–Def8–Af3l1–Dbndd1*) and compared them with those previously determined for mice from other parts of Eurasia. Construction of a network with the concatenate sequences yielded three distinct clusters representing the three major subspecies groups: CAS, *M. m. domesticus* (DOM), and *M. m. musculus* (MUS). STRUCTURE analysis provided evidence for further subdivision of CAS into two main haplogroups within the Indian subcontinent. Single-gene networks revealed not only gene-specific architecture for subgrouping in CAS, but also allelic exchange among subspecies. These results suggest the earlier onset of allopatric divergence in the predicted homeland (the Middle East and Indian subcontinent) and subsequent intermittent admixing via gene flow across the CAS haplogroups and among the three subspecies groups. A comparison of the levels of nucleotide diversity among the gene regions revealed a less divergent state in the chromosome region containing *Mc1r* and its adjacent genes, indicative of a selective sweep, suggesting the involvement of natural selection in the *Mc1r* allelic variation.
INTRODUCTION

*Mus musculus* originated from south-western Asia and extended its range to all of Eurasia in association with prehistoric human movements. The species is now a well-known model organism and a large number of strains with different geographic origins and hence different genetic backgrounds are used for biomedical research (e.g. Takada *et al.*, 2013). Therefore, knowing the genetic background of the species throughout its range would be useful. In addition, tracking the dispersal of mouse lineages should give insight into the prehistoric movements of the humans that the mice accompanied (e.g. Boursot *et al.*, 1993; Bonhomme & Searle, 2012; Cucchi, Auffray & Vigne, 2012; Gabriel, Mathias & Searle, 2013; Jones *et al.*, 2013; Jing *et al.* 2014). The origin of *M. musculus* before its human-mediated dispersal is most commonly identified as either India (Boursot *et al.*, 1993, 1996; Din *et al.*, 1996) or the Middle East (Prager, Orrego & Sage, 1998; Suzuki *et al.*, 2013), based on various molecular markers, including mitochondrial DNA, allozymes, and nuclear intron sequences. Under these scenarios, the ancestral stock gave rise to three dispersing lineages, *M. m. castaneus* (CAS; southern Asia, Southeast Asia, southern China), *M. m. domesticus* (DOM; Western Europe), and *M. m. musculus* (MUS; northern Eurasia excluding Western Europe), which made secondary contacts in Eurasia (Boursot *et al.*, 1993). However, difficulty has arisen in assessing the evolutionary history in this species, even after intensive molecular work, for various reasons, including intra- and interspecies introgression events and intricate human-mediated propagation in both early prehistory and recent times. Regardless, no intensive genetic analyses of mice have been performed in their original range.

Mitochondrial DNA (mtDNA) analyses have identified five distinct lineages that diverged roughly a half million years ago (mya): those representing the three major subspecies CAS, DOM, and MUS and two local lineages confined to Yemen (*M. m. gentilulus*) and Nepal (Prager *et al.*, 1998; Suzuki *et al.*, 2013). In contrast, studies using nuclear markers provide substantially different spatiotemporal views of the evolution of *M. musculus*. Based on isozyme analyses, mice from southern Asia, the hypothesised origin of CAS, are separated into two geographic groups: mice from Iran, Afghanistan, Pakistan, and northern India, called “*Mus bactrianus*” (or *M. m. bactrianus*) and mice from the rest of Eurasia, including southern Asia, called *M. m. castaneus* (Bonhomme *et al.*, 1984). In addition, sequences of nuclear genes such as *Irbp* and *Rag1* suggest ancient divergence dating back to that between *M. musculus* and closely related species such as *M. spretus* and *M. spicilegus* (e.g. 1.7 mya; Suzuki *et al.*, 2004). In comparison,
the mtDNA analyses tend to show lower levels of divergence among the geographic groups, including those from the territory of “M. bactrianus” (0.1–0.2 mya; Suzuki et al., 2013). Therefore, the genetic architecture of CAS remains unclear.

Organisms in the wild are subject to a specific evolutionary mode, called reticulate evolution, in which genetic introgression and hybridisation occur intermittently among closely related species or geographically situated conspecific populations, as is documented in M. musculus (e.g. Bonhomme et al., 2007; Nunome et al., 2010; Rajabi-Maham et al., 2012). Analyses of several different phylogenetic markers are essential to elucidate the evolutionary histories of organisms in which multiple hybridisation events are suspected over the long term. Haplotype structure analysis, which monitors recombination events focused on variation in a chromosome region using representative linked gene markers, is effective for assessing the time span after detecting recombinant haplotypes and can be used to time ongoing hybridisation events (Martinsen et al., 2001; Koopman et al., 2007; Nunome et al., 2010). Since the density of recombination points reflects the number of generations since the hybridisation event, assessment of haplotype structure, which is defined by a combination of alleles in an appropriate length of chromosome, allows us to consider spatiotemporal aspects of hybridisation.

Nunome et al. (2010) adapted haplotype analyses to wild mice mainly from East Asia using eight linked gene markers. The aligned set of sequences has proven to be useful as a biparental genealogical marker in inferring population genetic structure. Moreover, by showing the recombinant haplotypes, it clarified the presence of historical hybridisation between CAS and MUS in northern Japan, including Hokkaido and northern Honshu. However, this study drew attention to several unresolved issues related to the evolution of Eurasian wild mice. First, MUS is further subdivided in northern Eurasia into northern (MUS-I; Eastern Europe, Siberia, and Primorye) and southern (MUS-II; Kazakhstan, Uzbekistan, northern China, Korea, and Japan) phylogroups, although little is known about this evolutionary episode. Second, the recombinant haplotypes from northern Japan involve components of MUS, CAS, and DOM. The presence of a short DOM segment (20–30 kb) was unexpected because it contradicts the expectation that Japanese wild mice possess ancient MUS and CAS DNA, as the major and minor components, respectively (Terashima et al., 2006). Third, the level of variability differs markedly from locus to locus, even within the same subspecies (Nunome et al., 2010). For example, the gene haplotypes assigned to MUS showed diverse patterns, ranging from one (Fanca) or two (Spire2) to 10 (Dbndd1) or
23 (Afg3ll) haplotypes. Finally, even CAS and DOM possess their own sequences over the loci examined, except for Mc1r (409-bp analysis), for which CAS and DOM share the same Mc1r allele, while subspecies-specific sequences are detected in both neighbouring genes.

Haplotype structure analysis also offers the prospect of examining the presence of positive natural selection on specific functional genes through an investigation of the reduction in nucleotide diversity, a sign of a selective sweep (e.g. Nielsen et al., 2005). This phenomenon is characterised by extensive linkage disequilibrium and limited haplotype diversity and is seen in a variety of organisms, including North American grey wolves (Canis lupus; Anderson et al., 2009) and wild mice (M. spretus and M. musculus; Song et al., 2011). Among rapidly evolving phenotypes, coat colour variation is an ideal character to examine. In Asia, the dorsal coat colour variation in M. musculus due to melanin synthesis is significantly correlated with precipitation: dark coats are observed in more humid habitats and pale coats in drier habitats (Lai et al., 2008). This implies a role of concealment as a selective force affecting dorsal coat colour in house mice (Lai et al., 2008). Two genes, agouti signalling protein (Asip) and the melanocortin-1 receptor (Mc1r), effectively drive the evolution of coat colour via phenotypic change arising from single mutations (e.g. Hubbard et al., 2010; Suzuki, 2013). For example, the effect of natural selection on the coat colour gene Asip has been documented in the deer mouse Peromyscus maniculatus (Linnen et al., 2009; Vignieri, Larson & Hoekstra, 2010). In Peromyscus polionotus, another example, amino acid changes in Mc1r are associated with an adaptive coat colour phenotype, which functions as camouflage against pale sand dunes (Mullen et al., 2009). The Mc1r sequences show different evolutionary rates of amino acid replacement in the phylogeny of Mus, possibly associated with coat colour changes for local adaptation (Shimada et al., 2009). The sequence variation in Mc1r and its neighbouring genes lying in a 200-kb stretch has been analysed in wild mice collected from throughout Eurasia and a broad area of the Japanese islands (Nunome et al., 2010). Investigation of this chromosome region with sufficient specimens should help us to determine the impact of natural selection on coat colour genes.

In this study, our basic objectives were to explore various evolutionary issues related to M. musculus, specifically its genetic state in its homeland before it underwent the long-range dispersal events associated with prehistoric human movements. Knowledge of the genetic architecture of the homeland area would help to identify the exact source areas of the lineages that dispersed with human activities in the past and
recently. Here, we focus on the CAS lineages in Southeast Asia, southern China, and Japan. We also focused on the involvement of natural selection on the Mc1r gene in the process of the genetic differentiation of the subspecies groups. We determined the sequences of the seven genes located in the chromosome region near Mc1r in wild mice collected from India and Pakistan and conducted phylogeographic analyses making use of available sequences representing other geographic lineages of Eurasian wild mice (Nunome et al., 2010).

MATERIALS AND METHODS

Specimens
We examined the genetic variation in nearly 40 wild-captured mice from 18 localities in India and Pakistan (Table 1; Fig. 1), most of which were collected by H. Ikeda and K. Tsuchiya in 1989–1992 (Yonekawa et al., 2003; Suzuki et al., 2013). The 54 wild-captured mice previously examined are those from 52 localities throughout Eurasia used in our previous study (Nunome et al., 2010). We determined gene sequences in five closely related species, three from the same species group, Mus spretus (strain SEG), M. spicilegus (HS3212), and M. macedonicus (HS537), and two relatively divergent taxa, M. caroli (HS1469) and M. terricolor (HS3962) (see Suzuki et al., 2004; Suzuki & Aplin, 2012 for their inferred phylogenetic relationships).

Sequence analyses
Previously, we examined eight genes, which are arranged at 20–30-kb intervals in a 200-kb portion of mouse chromosome 8 (Nunome et al., 2010). Here, we re-examined the genes, excluding Tubb3, which has less pronounced allelic diversity. We amplified segments of Fanca (448 bp), Spire2 (436 bp), Tcf25 (613 bp), Mc1r (409 bp), Def8 (370 bp), Afg3ll (510 bp), and Dbndd1 (550 bp) (see Fig. 2 of Nunome et al., 2010). The primers used in the polymerase chain reaction (PCR) analyses included those used by Nunome et al. (2010). The sequence of the entire coding region (948 bp) of mouse Mc1r was determined using the primers previously described (Shimada et al., 2009). We determined sequences of the non-coding region (Mc1r-NC) immediately upstream of the Mc1r coding region, with primers newly designed (Mc1r_ NCF: 5’-TCCCTTCTTCTCCAGAGTCC-3’; Mc1r_ NCR: 5’-GCACGATGCTGACACTTACC-3’) using the Ensemble Mouse Genome Database...
The amplifications were carried out for 35 cycles each consisting of 30 s at 96°C for denaturation, 30 s at 55–64°C for annealing, and 30 s at 72°C for extension. The reaction mixtures (20 μl) contained 2.5 mM MgCl₂. Master Mix 360 DNA Polymerase (Applied Biosystems, Foster City, CA, USA) was used in the PCR. The first PCR cycle was preceded by 10 min at 95°C to activate the polymerase. The double-standard PCR product was purified using a 20% polyethylene glycol–2.5 M NaCl precipitation method, sequenced using the PRISM Ready Reaction Dye Deoxy Terminator Cycle Sequencing Kit (Applied Biosystems), and run on an ABI3130 automated sequencer (Applied Biosystems). Both strands were sequenced using either the forward or reverse primer used for PCR.

**HAPLOTYP ASSSESSMENT FROM ALLELIC COMBINATIONS OF THE SEVEN LOCI**

We used PHASE v2.1 (Stephens, Smith & Donnelly, 2001; Stephens & Scheet, 2005) with the default settings to determine unique allele arrangements in each of the gene regions examined. Networks of the single-gene sequences and concatenated gene sequences for Fanca, Spire2, Tcf25, Mc1r, Def8, Afg3l1, and Dbndl1 were constructed using the Neighbour-Net (NN) method, as implemented in SPLITS TREE v4.11.3 (Huson & Bryant, 2006). The number of pairwise differences, nucleotide diversity (π), neutrality test, and Tajima’s D were calculated using ARLEQUIN (ver. 3.5) (Excoffier & Lischer, 2010).

Population genetic structure was inferred by Bayesian clustering analysis using STRUCTURE, version 2.3 (Pritchard, Stephens & Donnelly, 2000) with concatenated sequences for the seven loci. The dataset used in the clustering analysis consisted of a total of 208 SNPs. K was estimated from 5 000 000 Markov chain Monte Carlo (MCMC) generations, sampled every 10 000 generations after a burn-in period of 2 000 000 generations. To determine an appropriate K, we used Evanno’s method in STRUCTURE HARVESTER (Earl & vonHoldt, 2012).

Considering the lengths of exogenous segments, one can predict the date when introgression took place (Nunome et al., 2010). A general indication of the time frames of meiotic recombination events can be inferred from the formula \( P = (1 - r)^G \) (Stephens et al., 1998), where \( 1 - P \) is the probability of recombination, \( G \) is generations, and \( r \) is the recombination rate (neglecting mutation). Transforming this equation, \( G \) can be estimated with \( G = -\ln(P)/r \). The average recombination rate within the Mus genome was estimated to be 0.52 cM/Mb (Jensen-Seaman et al., 2004). For example, a 10-kb segment carries an estimate of \( 5.2 \times 10^{-5} \) for \( r \). The estimate of \( G \) is the sum of...
generations for exogenous segments in the heterozygous state and does not account for
generations in the homozygous state, and is consequently the minimum value for the
generations elapsed since the introgression event.

**ESTIMATION OF TIMES OF DIVERGENCE**

BEAUti and BEAST v1.7.5 (Drummond et al., 2012) was used for the Bayesian
Markov-chain Monte-Carlo (MCMC) analyses of the nucleotide sequence data
(Drummond et al., 2005) to estimate the time to the most recent common ancestor
(TMRCA) and 95% highest posterior density (HPD). As outgroup taxa, we used *Mus
caroli, M. terricolor, M. spretus, M. spicilegus*, and *M. macedonicus* and the predicted
divergence time of 1.7 mya was used as a prior for the node separating *M. spretus* and
*M. musculus* (Suzuki et al., 2004). For the MCMC posterior analyses, the chain length
was 10,000,000. After a 1000-tree burn-in, 10,000 trees were used for the analyses. The
Bayesian tree was generated under the relaxed clock model (HKY substitution) and the
best-fit model was inferred in MEGA v5.0 (Tamura et al., 2011). The convergence of
the MCMC chains and the effective sample size (ESS) values exceeding 200 for all
parameters were assessed using Tracer v1.5 (Rambaut & Drummond, 2009).

**DATA ARCHIVING**

The nucleotide sequences reported in this paper appear in the DDBJ, EMBL, and
GenBank nucleotide sequence databases under the accession numbers
AB909504–AB909921. Sequence data files in nexus file format, together with
Supplementary Information files, will be stored in the Dryad repository.

**RESULTS**

**PHYLOGENIES OF SINGLE GENE SEQUENCES**

We determined the sequences of seven markers used previously (Fanca, Spire2, Tcf25,
Mcl1r, Def8, Afg3II, and Dbnndl; Nunome et al., 2010) in 39 mice collected from India
and Pakistan covering a major portion of the predicted homeland of CAS (Table 1; Fig.
1) and separated allelic sequences with a Bayesian method using PHASE (Stephens et
al., 2001; Stephens & Scheet, 2005). Neighbour-net (NN) networks (Fig. 2) were
constructed for the seven markers with the newly determined sequences, together with
those previously determined (59 mice) assigned to CAS, DOM, and MUS (Nunome et
This exemplified a higher level of allelic diversity in CAS mice from Pakistan and India, integrating the sequences of mice from India and Pakistan into two or more divergent clusters in six gene-sequence sets, excluding that of \( Mc1r \). Mice from the DOM territory formed a single cluster for each of the seven genes. For mice from the MUS territory, the sequences formed single clusters for the first five loci (\( Fanca, Spire2, Tcf25, Mc1r, \) and \( Def8 \)), while they formed three distinct clusters for the last two loci, \( Afg3l1 \) and \( Dbndd1 \).

The NN networks show the genetic exchange between the subspecies. The majority of CAS sequences of \( Mc1r \) (409 bp) from India and Pakistan (81\%) was assigned to a single genotype identical to that of the DOM mice. For \( Def8 \), several mice from northern India (DNA codes 74, 75) and northern Pakistan (60, 62, 63) possessed a sequence assigned to DOM (Fig. 2A). The allelic sequences of mice from Japan were overwhelmingly derived from MUS, although some from northern Japan (northeast Honshu and Hokkaido) were non-MUS sequences, showing high affinity with CAS sequences (\( Spire2, Tcf25, Mc1r, \) and \( Dbndd1 \)) and DOM (\( Def8 \) and \( Afg3l1 \)), as previously observed (Nunome et al., 2010). The non-MUS \( Def8 \) sequence from northern Japan was identical to that in DOM mice and the above-mentioned mice from the CAS territory (Fig. 2A, arrowhead). The \( Afg3l1 \) sequence (510 bp) from northern Honshu and Hokkaido differed from those representing DOM by one base (site 445), so that they were near the DOM cluster in the NN network and differed from any allele from the CAS territory (Fig. 2A, arrowhead).

**NETWORK OF CONCATENATE SEQUENCES**

Combinations of the seven linked gene regions (\( Fanca, Spire2, Tcf25, Mc1r, Def8, Afg3l1, \) and \( Dbndd1 \)) were assessed with PHASE (Stephens et al., 2001; Stephens & Scheet, 2005). The NN network of the haplotypes of the concatenated sequences formed several distinct clusters (Fig. 3A). We assessed the genetic ancestry of subspecies, based on the clustering patterns perceived in the networks for the single-gene sequences (Fig. 2A). This allowed us to recognise three subspecies clusters representing the three subspecies groups (CAS, DOM, and MUS), while they simultaneously exhibited a number of recombinant haplotypes or haplogroups. Notably, seven divergent clusters were tentatively seen within the CAS cluster. The haplogroups showed rough geographic distributions in northern India and Pakistan (CAS-A), southern India (CAS-B), northern and central India (CAS-C), Southeast Asia (CAS-D), and India and Pakistan (CAS-E) (Fig. 3C). Notably, CAS-D comprised the haplotypes recovered from...
a large geographic area of Southeast Asia, South China and Indonesia and can be characterized as the lineage dispersed with the prehistorical human movement.

One of the recombinant haplogroups seen in the concatenate NN network (Fig. 3A), here designated Re-1[C, D, M], was recovered from northern Japan, and possessed three components that could be assigned to CAS, DOM, and MUS. Another recombinant haplogroup (Re-10[C, D, M]) consisted of haplotypes of a mouse (74-H1184 in Fig. 1) from Delhi, India, in which the alleles of Tcf25 and Def8 were closely related to MUS and DOM alleles, respectively. The specific CAS haplotypes from India and Pakistan, in which an allele assigned to DOM was embedded, formed a scattered pattern in the CAS cluster.

The MUS mice tended to form two clusters, with the groups corresponding to geographic regions, i.e., a ‘northern’ (MUS-I) subgroup representing mice from Eastern Europe, Siberia, and Primorye, and a ‘southern’ (MUS-II) subgroup from Kazakhstan, Uzbekistan, northern China, Korea, and Japan (Nunome et al., 2010). In contrast, the DOM group formed a tight-knit cluster on the network tree, with much less internal divergence than within either CAS or MUS.

The Bayesian clustering analysis (Drummond et al., 2012; Earl & vonHoldt, 2012) with the combined dataset for the seven gene markers indicated that the wild mouse population is composed of four genetic clusters (K =4: CAS-I, CAS-II, DOM, and MUS), the first two of which are represented by the abovementioned haplogroup A and a haplogroup set of B, D, and E (Fig. 3A), with apparent geographic subdivision into northern (Himalayas) and southern parts of the Indian subcontinent, respectively (Fig. 3C). From the visualisation of K = 4, most of the individuals (codes 57, 73, 83, 84) belonging to haplogroup C were assigned to the DOM cluster in addition to the CAS related clusters CAS-I and CAS-II.

We performed a phylogenetic analysis using the concatenated sequences (n = 34; 3402 bp) using BEAST v1.7.5 to assess the phylogenetic relationships and the temporal aspects of the divergence of haplotypes representing the geographic groups. We removed the haplotypes that were assigned to be of hybrid origin deduced from the network and STRUCTURE analyses in order to clarify the temporal patterns of ancient divergences. We used five outgroup taxa with a prior of 1.7 mya for the divergence of M. spretus from the other members of the M. musculus species group. The phylogenetic analysis recovered the four monophyletic groups predicted by the STRUCTURE analysis (Fig. 4). TMRCAs for CAS-I and CAS-II, and DOM and MUS were calculated
to be 1.05 mya (95% HPD = 0.85–1.27 mya) and 0.91 mya (95% HPD = 0.70–1.16 mya), respectively.

**SEQUENCE VARIABILITY OF Mc1r AND ITS FLANKING GENES**

We determined the entire coding region of Mc1r (948 bp) from 84 mice including 45 from India and Pakistan to get a finer view of the evolutionary dynamics of single-nucleotide polymorphisms (SNPs) across the subspecies, revealing 28 alleles (Fig. 5A). The coding sequences featured a total of 28 variable sites, with 19 nonsynonymous and nine synonymous mutations. The network tree revealed the presence of a predominant allele (allele 5) representing certain portions of the DOM and CAS mice examined. The prevalent allele possessed a C to T mutation at site 302 changing the amino acid at codon position 101 from alanine to valine (Fig. 5B; Table 1).

The sequence analysis of the entire Mc1r coding region revealed another prevalent mutation (A218G) with a predicted amino acid change from tyrosine to cysteine at codon position 73, in mice from Pakistan and northern India (Table 1).

To further address the divergence of Mc1r sequences in CAS and DOM, we determined the sequences of the 377-bp Mc1r non-coding region, Mc1r-NC, immediately upstream from the Mc1r coding region in 81 mice and constructed a network with the sequences obtained (Fig. 5A). DOM mice were shown to have a single unique NC sequence (red circle). Eleven alleles were identified among the sequences of mice from India and Pakistan (orange circles), but none of them were shared with the allele exclusive to DOM mice. The mice from northern Japan that had the Mc1r coding sequence (allele 5) shared by CAS and DOM had the predominant CAS allele in Mc1r-NC, differing from the NC allele unique to DOM by one base, providing a evidence for the subspecies origin of the coding sequence.

To examine the spectrum of nucleotide diversity ($\pi$) across the chromosome region examined in each of the three subspecies groups, we choose non-recombinant haplotypes, accounting for the patterns shown in the network and STRUCTURE analyses (Fig. 3). In total, 172 haplotypes that were assigned as CAS (n = 91), DOM (n = 15), or MUS (n = 66) were used to compare $\pi$ among the seven gene regions, together with those for Mc1r-NC upstream from the coding region (Fig. 2B; Table S1). In MUS, an apparent reduction in nucleotide diversity was seen in the Fanca, Spire2, Tcf25, Mc1r, and Def8 gene loci, while the remaining two gene loci were markedly divergent, suggestive of recent introgression between the haplogroups MUS-I and MUS-II. In
CAS and DOM, a substantial reduction in nucleotide diversity was observed in Mc1r and the adjacent loci at the central portion near Mc1r and Mc1r-NC, while the levels of the nucleotide diversity were relatively high in genes in the flanking regions on both sides. We calculated Tajima’s $D$ for each of the three subspecies groups using the data for the eight gene regions examined, including Mc1r-NC. Negative values were obtained for Mc1r for CAS and DOM, although the values were not significant (Table S1).

**DISCUSSION**

ASSESSING THE ANCIENT GENETIC STRUCTURING OF *M. MUSCULUS* IN ITS HOMELAND

We assessed the ancient state of genetic structure of *M. musculus* in its presumed species homeland and the dispersal events associated with prehistoric human movements, which led to the current occurrence of the three major subspecies groups in northern Eurasia (MUS), southern Asia (CAS), and westernmost Eurasia (DOM). This study highlights the intricate phylogenetic relationships of the subspecies groups and unveiled the previously concealed evolutionary history of *M. musculus*.

The phylogenetic patterns of concatenated sequences give a general view of the genetic structure of *M. musculus* (Fig. 3). The NN network disclosed three genetically distinct haplogroups representing the three major subspecies groups: CAS, DOM, and MUS. STRUCTURE analysis further suggested the subdivision of CAS into two clusters, here designated CAS-I and CAS-II, with apparent geographic affinity in the northern and southern parts of the Indian subcontinent, respectively (Fig. 3). In contrast to the shallow divergence of mtDNA of *M. musculus* (0.35–0.5 mya, Suzuki et al., 2013), the divergence of these geographic groups is likely as deep as those between closely related species (i.e., *M. macedonicus*, *M. spretus*, and *M. spicilegus*; Fig. 4). This is in good accord with molecular phylogenetic studies on nuclear gene sequences that observed higher levels of divergence (e.g. Suzuki et al., 2004; Bonhomme et al., 2007). These considerations thus imply that *M. musculus* has a long evolutionary history with several distinct geographic groups, perhaps arising 1–2 mya in a wide area of its ancient territory, i.e., Iran, Afghanistan, Pakistan, and northern India.

The haplotypes of mice from the CAS territory were substantially diverse (Figs. 3, 4). Our current phylogenetic work portrayed spatial structuring within the Indian subcontinent. A trend in the basal divergences of the lineages recovered from northern India and Pakistan is present along with clear genetic differentiation in other
haplogroups represented by mice from certain geographic areas, such as CAS-B from the southern part of the subcontinent, with an estimated divergence time of ca. 0.5 mya (Fig. 4). This implies the primary occurrence of CAS to be in Pakistan and northern India and an early secondary range expansion to the southern part of the Indian subcontinent. This contradicts the previous view from mtDNA analyses in which mice from the eastern and southern parts of the Indian subcontinent share less divergent mtDNA sequences from a recent expansion within the last 10,000 years (Suzuki et al., 2013; see Fig. 3C). It is thus suggested that mice occurring in eastern and southern India experienced recent intensive mtDNA introgression, while maintaining an ancient divergence in the nuclear genome.

ANCIENT INTROGRESSION AMONG GEOGRAPHIC GROUPS

The genetic exchanges among the original homeland geographic groups before the long-range dispersal events are poorly documented. However, our results shed light on the genetic exchanges among the geographic groups of *M. musculus* occurring in India and Pakistan. The historical genetic exchanges among the geographic groups are evident in the single-gene network (Fig. 2); the DOM-assigned sequence of Def8 is recovered from mice from the CAS territory, including Islamabad (code 60), Lahore (codes 62, 63), and Delhi (codes 74, 75). Notably, some CAS mice (e.g. codes 60, 89) formed a cluster separate from the rest of the CAS-assigned alleles and localised near the DOM-assigned cluster for *Spire2* (Fig. 2B). These results imply that intersubspecies introgression happened from time to time.

The sizes of present-day DOM segments incorporated in mice from the CAS territory are likely short, *i.e.*, the size of single genes or 10–20 kb. In the case of Def8, for example, the mice from Pakistan and northern India carrying the DOM alleles possess CAS alleles at the immediately adjacent gene loci. A parsimonious explanation for the shortened DOM segment is that introgression event(s) happened at some ancient time, followed by long-term backcrossing of the DOM segments against the CAS background genome. Haplotype breakdown arising through recombination in a given 10-kb stretch is expected to be seen with >50% probability after 14,000 generations (1–3 generations per year), assuming that the recombination rate is 0.52 cM/Mb (Jensen-Seaman et al., 2004). These considerations suggest that substantial time has passed since the introgression occurred. Therefore, the gene introgression between the subspecies may have manipulated such gene (or gene segment)-specific incorporation of
exogenous genes in recipient subspecies genomes, facilitating the maintenance of genetic diversity.

The possible occurrence of an introgression event involving genes of DOM origin into the CAS background mice in the homeland area, the Indian subcontinent, is a clue to the enigmatic issue of the CAS segment harbouring DOM-related genes (Def8 and Afg3l1) in the northern Japanese mouse genome. Given that ancient introgression events involving Def8 are evident in the homeland area, the presence of a rather long DOM segment (20–30 kb, i.e. Def8 plus Afg3l1) in the northern Japanese mouse population might be attributable to ancient introgression from DOM to CAS and to early isolation by dispersal to a place where conspecific populations were absent, followed by the acquisition of a homogeneous state for the exogenous segment of CAS and DOM in the MUS background mice via a founder effect.

The genetic structuring for the three CAS subgroups in the Indian subcontinent seen in the NN network using the concatenate sequences was not reproduced in the single-gene trees (Fig. 2). This implies that substantial gene flow occurred among the geographic groups during the course of evolution and contributed to generating the diversified haplotypes of the concatenate sequences of the CAS groups. The STRUCTURE analysis supports this view (Fig. 3B). Overall, one can reasonably presume that the evolutionary state of *M. musculus* is a typical example of reticulate evolution.

**Genealogical Assignment of the Human-Related Dispersal in Asia**

Although prehistoric human movements across Eurasia are known to have assisted the long-distance dispersal of house mice from its homeland (Boursot *et al.*, 1993; Prager *et al.*, 1998), the exact source areas are vague. Our previous mtDNA study showed that the eastward movements of CAS mice brought one of the four mtDNA subtypes, CAS-1, to a broad area of southern China, continental Southeast Asia, the Philippines, Indonesia, Japan, and Sakhalin (Suzuki *et al.*, 2013). However, the magnitude of the differences among the CAS-1 sequences from the eastern part of the Indian subcontinent is limited, perhaps due to overlaying of part of the homeland in the Indian subcontinent by the newly expanded mtDNA subtype; therefore, we are unable to identify a dispersal track from the mtDNA data. In this study, the majority of the nuclear haplotypes of the dispersed lineages (CAS-D) was found to be closely related to one another and shown to be nearly identical to the haplotype of the mouse from Pune, central India, while differing from the other haplotypes CAS-A, B, C, and E recovered from India (see Fig.
These results suggest that Southeast Asia, southern China, and Indonesia are representative areas to which *M. musculus* expanded efficiently, mediated by the historical human eastward movement, and that its source area is confined to a geographic locality, perhaps somewhere in India with the haplotypes belonging to CAS-D.

The CAS mice appear to have expanded their range to the northern tip of the Japanese islands, where the majority of the present-day genetic component is that of MUS. Accordingly, the 200-kb segments represent recombinant haplotypes between MUS and non-MUS, as seen in mice from northern Japan; this recombinant is $\text{Fanca}^{\text{MUS}} - \text{Spire2}^{\text{CAS}} - \text{Tcf25}^{\text{CAS}} - \text{Mc1r}^{\text{CAS}} - \text{Def8}^{\text{DOM}} - \text{Afg3}^{\text{DOM}} - \text{Dbnnd}^{\text{CAS}}$ (Nunome *et al.*, 2010). If we consider that the CAS fragment is 200 kb in size, the recombinant segments comprise components of CAS and DOM and the time of the hybridisation is estimated to be around 700 generations ago (Nunome *et al.*, 2010). The non-MUS Def8 and Afg3ll sequences are assigned as DOM, and the unique combination of CAS and DOM components is thought to have been generated in the source area, as mentioned above. However, no such combination of CAS and DOM has yet been found in the predicted source area, perhaps due to intensive backcrossing of the introgressed DOM segments with CAS mice.

The analysis of linked nuclear gene sequences is useful for phylogeographic studies for several reasons. First, a set of concatenated sequences is helpful for reconstructing a reliable population genetic structure without making any particular assumptions. Second, in a species that has experienced dispersal and secondary contact, the method used here can help to identify the source locality and contact areas. By setting the intervals of gene loci properly, one can infer the time span of hybridisation. Moreover, if a gene in the central part of the chromosome region used in the haplotype analyses has undergone adaptive evolution, we can test this possibility by measuring the reduced level of nucleotide diversity, as discussed below.

### Possible involvement of natural selection acting on the coat colour gene 

*Mc1r*

The prominent features of the genetic diversity of *M. musculus* are that both the degree of nucleotide diversity and the phylogenetic patterns differ substantially among the genes examined, despite the fact that they are arranged tightly at approximately 20-kb intervals. This implies that *M. musculus* has experienced gene flow within and between geographic groups and the subsequent geographic allocation of specific alleles across
geographic barriers in a gene-specific manner. Concerning the ancient gene flow across the geographic groups in its homeland, identifying the factors shaping the geographic permeability of specific alleles is of interest. Mc1r is one of the best-known genes promoting the evolution of pigmentation in birds and mammals (e.g. Klungland et al., 1995; MacDougall-Shackleton, Blanchard & Gibbs, 2003; Hoekstra & Price, 2004; Hoekstra, 2006; Hubbard et al., 2010; Manceau et al. 2010; Suzuki, 2013). Our data include evidence for the inclusion of natural selection in shaping the patterns with respect to the coat colour gene Mc1r.

Comparing the nucleotide diversity across the chromosome region, V- and U-shaped graphs are detected in CAS and DOM, respectively (Fig. 2B; Table S1). This feature, known as selective sweep, is considered a footprint of positive selection (Maynard Smith & Haigh, 1974; Teschke, Büntge & Tautz, 2012) and the cases accompanied by gene flow across different taxa are known as selective gene flow or adaptive introgression (Martinsen et al., 2001; Bull et al., 2006; Anderson et al., 2009; Song et al., 2011). The trend was confirmed in the adjacent upstream non-coding region, Mc1r-NC, suggesting that the lower diversity of Mc1r is not due to a comparison of a coding sequence with non-coding (e.g. intron) sequences. The V-shaped pattern in CAS is attributable to long-term processing of the polymorphism. The U-shaped pattern in DOM indicates a greater level of linkage disequilibrium affecting Mc1r and the adjacent region, indicating the relatively recent involvement of natural selection (Maynard Smith & Haigh, 1974). Accordingly, the values of Tajima’s D were negative for Mc1r in CAS and DOM, although not statistically significant (Table S1). These results support the idea that natural selection has acted on the Mc1r gene or the adjacent region in the homeland of CAS and DOM during the course of evolution.

If natural selection is responsible for the specific population genetic patterns of Mc1r in CAS and DOM, certain Mc1r allele(s) should involve a functional change, namely an amino acid replacement(s) altering coat colour. A candidate is allele 5 of the Mc1r-coding that is shared by CAS (61%) and DOM (30%) (Fig. 5), whose state could be explained either by ancestral polymorphism or introgression. The allele shows two diagnostic derived mutations at sites 51 (C to T) and 302 (C to T) (Fig. 5B), which are synonymous and non-synonymous mutations, respectively. Hence, the latter is a candidate as the responsible mutation, as it is predicted to replace the amino acid alanine with valine at codon position 101 (A101V). The site of the candidate mutation at codon 101 corresponds to the first extracellular loop (LC1), where the reverse agonist of agouti signalling protein (ASIP) makes contact (Patel et al., 2010) and is predicted to be important for the adaptive evolution of coat colour (Eizirik et al., 2003). Considering coat colour (Marshall, 1998), mice occurring in Pakistan and western and central India have a straw- or golden-coloured pelage dorsally, and have been classified as specific
taxa, such as “bactrianus” and “tytleri”, while mice from adjacent northern (Nepal) and southern (south India) areas have dark pelage dorsally and are traditionally classified as “homoulus” and “castaneus”, respectively (Marshall, 1998). A future study focusing on the relationships between the Mc1r SNP and coat colour variation would illuminate the taxonomic issue of the traditional subspecies grouping.

CONCLUSIONS

Our results demonstrate that M. musculus has a long evolutionary history since the establishment of the geographic groups leading to the major subspecies CAS, DOM, and MUS. A deep phylogenetic subdivision within CAS is evident, involving two distinct phylogroups with predicted origins on the Indian subcontinent. Apparent signals of gene flow exist among the three subspecies, which enabled them to mediate allelic exchanges. Notably, natural selection appears to have been involved in the intensive geographic expansion of Mc1r alleles during the course of evolution. The wild population of house mice, with its extraordinary genetic diversity and geographically separated groups, is an ideal subject for studies on evolutionary biology, in which natural selection is one of the major concerns (Teschke et al., 2012). The method we adopted in this study (i.e. the analysis of linked nuclear multilocus markers) has been proven useful for assessing population genetic structure and genetic interaction in areas of secondary contacts.

ACKNOWLEDGEMENTS

We wish to express our appreciation to K. Abe, K. P. Aplin, T. Takada, T. Shiroishi, K. Tsuchiya and H. Yonekawa for their valuable advice with this study. We also thank G. N. Chelomina, L. N. E. Dokuchaev, V. Frisman, A. Frost, S.-H. Han, N. Hanzawa, T. Hosoda, H. Igawa, H. Ikeda, M. A. Iwasa, M.-L. Jin, I. Kartavtseva, V. P. Korablev, K. K. V. Korobisyna, M. Kusayama, A. P. Kryukov, O. E. Lopatin, I. Maryanto, Y. Matsuda, H. Matsuzawa, G. Mise, N. Miyashita, P. Munclinger, I. Munechika, N. Nakajima, E. Nevo, H. Okamura, R. Palmer, M. Pavlenko, M. Sakaizumi, K. Sasaki, J. J. Sato, T. Shimada, M. H. Sinaga, A. Suyanto, S. Suzuki, M. Terashima, M. Tomozawa, H. Ueda, P. Vogel, S. P. Yasuda, K. Yokoyama, and S. Wakana for their help in collecting the animals. We are also grateful for extensive comments from anonymous reviewers. This study was funded by a grant-in-aid for Scientific Research (B) to HS (24405013) from the Japan Society for the Promotion of Science (JSPS). We thank the Heiwa Nakajima Foundation for its generous support.
REFERENCES


Patel MP, Cribb Fabersunne CS, Yang YK, Kaelin CB, Barsh GS, Millhauser GL. 2010. Loop-swapped chimeras of the agouti-related protein and the agouti signaling


Figure legends

Figure 1. The localities where the wild mice used in this study were collected together with the specimen codes for those previously analysed (1–59; see Nunome et al., 2010 for further details) and those newly determined (60–98). The detailed locality names and sample codes are listed in Table 1. Localities with SNP C→T at site 302 in the Mc1r coding region are represented by bold circles. Rough geographic ranges for the three major subspecies (CAS: M. m. castaneus, DOM: M. m. domesticus, and MUS: M. m. musculus) and the two previously predicted MUS groups (Nunome et al., 2010) are shown.

Figure 2. Genetic variation of the seven gene markers. A, Neighbour-Net networks derived from single gene sequence data sets for Fanca, Spire2, Tcf25, Mc1r, Def8, Afg3l1, and Dbndl in Mus musculus and five species as the outgroup: M. caroli (1), M. terricolor (2), M. spreus (3), M. spicilegus (4), and M. macedonicus (5). The subspecies groups M. m. domesticus (DOM), M. m. castaneus (CAS), and M. m. musculus (MUS) were assessed by referring to previously obtained data (Nunome et al., 2010). The newly determined allele sequences originating from India and Pakistan were assigned as CAS, although several Def8 sequences were integrated into the cluster assigned as DOM. The positions of non-MUS sequences from northern Japan are marked with arrowheads. Genetic variation of the six gene markers. B, comparison of the nucleotide diversities (%) using sequences assigned for the three subspecies groups. The non-coding region (Mc1r-NC) immediately upstream of the Mc1r (melanocortin-1 receptor gene) coding region was included.

Figure 3. Assessment of population genetic structure using the dataset of the seven nuclear genes. A, Neighbour-Net network with the concatenated sequences from 98 Mus musculus with subspecies groups M. m. domesticus (DOM), M. m. castaneus (CAS), and M. m. musculus (MUS) and five outgroup species. Seven haplogroups seen in the CAS cluster (CAS- A–E) are tentatively indicated. Those apparently recognised as recombinant haplotypes (Re- 1~ 8: Nunome et al., 2010; Re-9~11: this study) are indicated by subspecies components (C, castaneus; D, domesticus; M, musculus). B, Population genetic structure was assessed using STRUCTURE, version 2.3. Genetic component proportions of individuals are represented by thin horizontal bars composed of four clusters. C, Distribution of the seven CAS haplogroups tentatively observed in
the network analysis. Representative haplotypes related to the seven gene markers are schematically shown, exhibiting DOM components embedded in CAS chromosome segments in wild mice from northern Japan. Mitochondrial DNA types are shown (Suzuki et al., 2013), emphasizing the CAS subtypes (CAS 1-4).

**Figure 4.** Phylogenetic tree of *Mus musculus* constructed using the Bayesian method with concatenate sequences (seven genes; 3402 bp). Statistical support is indicated by Bayesian posterior probabilities. The times of the most common recent ancestors and 95% highest posterior density estimated by the Bayesian analysis are shown for each major calibration point. The numbers below the lines at the nodes are the Bayesian posterior probabilities. See Figure 3 for the four CAS haplogroups.

**Figure 5.** A, Median Joining (MJ) networks constructed with the entire coding region of the melanocortin-1 receptor gene (*Mc1r*, 948 bp; right) and the immediate upstream non-coding region (*Mc1r*-NC, 377 bp; left), showing the subspecies groups *M. m. domesticus* (DOM), *M. m. castaneus* (CAS), and *M. m. musculus* (MUS) based on geographic origin. Each circle represents an allele and its size reflects the allele frequency. Numbers along branches and adjacent to alleles indicate nucleotide positions and codes of alleles (alleles 1-5) of interest. Allele 1 is predicted to be the prototype from the comparison with outgroup sequences. Allele B, Schematic representation of the gene structure of *Mc1r*. Presumed derived mutations of nucleotides and amino acids, differing from the predicted original haplotype (hap 1), are shown in haplotypes representing the subspecies groups (hap 2-6).

**SUPPORTING INFORMATION**

Additional Supporting Information may be found in the online version of this article at the publisher’s website:

Table S1. Genetic diversity indices of the eight nuclear gene markers. The number of segregating sites (S), mean number of pairwise differences (π), standard error (SE), and Tajima’s D were calculated using ARLEQUIN 3.1 (Excoffier & Lischer, 2010).
Kodama et al. Fig. 4
**A**

Non-coding (NC) region

Coding region

- **DOM**
- **MUS**
- **CAS: SE Asia + S China**
- **CAS: India + Pakistan**

**B**

- **Mc1r**
  - 5' UTR
  - Amino-acid coding region

- **Haplotypes**
  - **Hap 1**
    - G
    - G
    - C
    - Tyr
    - Ala
    - G
  - Prototype: 1
  - Subspecies: MUS

  - **Hap 2**
    - A
    - C
    - Cys
  - Prototype: 2
  - Subspecies: MUS

  - **Hap 3**
    - T
    - Val
  - Prototype: 3
  - Subspecies: CAS

  - **Hap 4**
    - T
    - Val
  - Prototype: 4
  - Subspecies: CAS

  - **Hap 5**
    - T
    - Val
  - Prototype: 5
  - Subspecies: CAS

  - **Hap 6**
    - A
    - T
    - Val
  - Prototype: 5
  - Subspecies: DOM

*Kodama et al. Fig. 5*
<table>
<thead>
<tr>
<th>Specimen Code</th>
<th>Locality</th>
<th>Haplogroup</th>
<th>Haplogroup</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 PGN2.HS3949</td>
<td>Canada:</td>
<td>Peking</td>
<td>DOM</td>
</tr>
<tr>
<td>2 MBi6.HS3958</td>
<td>Germany:</td>
<td>Weidesgrun</td>
<td>DOM</td>
</tr>
<tr>
<td>3 MB9.HS3959</td>
<td>Czech Rep.:</td>
<td>Kubelhof</td>
<td>DOM</td>
</tr>
<tr>
<td>4 BFM2.HS3947</td>
<td>Czech Rep.:</td>
<td>Montpellier</td>
<td>DOM</td>
</tr>
<tr>
<td>5 PWK.HS3948</td>
<td>Czech Rep.:</td>
<td>Lhotta</td>
<td>MUS-I</td>
</tr>
<tr>
<td>6 BLG2.HS3951</td>
<td>Blugaria:</td>
<td>Toshovo</td>
<td>MUS-I</td>
</tr>
<tr>
<td>7 HS589</td>
<td>Italy:</td>
<td>Aosta</td>
<td>MUS-I</td>
</tr>
<tr>
<td>8 HZ788</td>
<td>Czech Rep.:</td>
<td>Liborezy</td>
<td>MUS-I</td>
</tr>
<tr>
<td>9 MG3044</td>
<td>Estonia:</td>
<td>Tallinn</td>
<td>MUS-I</td>
</tr>
<tr>
<td>10 MG3065</td>
<td>Ukraine:</td>
<td>Donetsk</td>
<td>MUS-I, MUS-II</td>
</tr>
<tr>
<td>11 HS1338</td>
<td>Uzbekistan:</td>
<td>Tashkent</td>
<td>MUS-II</td>
</tr>
<tr>
<td>12 HS1464</td>
<td>Kazakhstan:</td>
<td>Aktubinsk</td>
<td>MUS-I</td>
</tr>
<tr>
<td>13 HS3602</td>
<td>Kyrgyzstan:</td>
<td>Bishkek</td>
<td>MUS-II</td>
</tr>
<tr>
<td>14 NIG934</td>
<td>Iran:</td>
<td>Alvaz</td>
<td>Re-[4D, M]</td>
</tr>
<tr>
<td>15 NIG935</td>
<td>Uzbekistan:</td>
<td>Nowshahr</td>
<td>MUS-I</td>
</tr>
<tr>
<td>16 MG3055</td>
<td>Russia:</td>
<td>Moscow</td>
<td>DOM</td>
</tr>
<tr>
<td>17 HS3612</td>
<td>Astrakhan:</td>
<td>MUS-I</td>
<td></td>
</tr>
<tr>
<td>18 HS3605</td>
<td>Gorno-Altay:</td>
<td>MUS-I</td>
<td></td>
</tr>
<tr>
<td>19 HS3608</td>
<td>Irkutsk:</td>
<td>MUS-I</td>
<td></td>
</tr>
<tr>
<td>20 HS3604</td>
<td>Tomsk:</td>
<td>MUS-I</td>
<td></td>
</tr>
<tr>
<td>21 HS1845</td>
<td>Magadan:</td>
<td>MUS-I, Re-[6D, M]</td>
<td></td>
</tr>
<tr>
<td>22 HS1461</td>
<td>Khabarovsky:</td>
<td>MUS-I, MUS-I</td>
<td></td>
</tr>
<tr>
<td>23 HS1466</td>
<td>Amurskii:</td>
<td>MUS-I</td>
<td></td>
</tr>
<tr>
<td>24 HS1412</td>
<td>Krasnoi:</td>
<td>MUS-I</td>
<td></td>
</tr>
<tr>
<td>25 HS1335</td>
<td>Khasan:</td>
<td>MUS-I</td>
<td></td>
</tr>
<tr>
<td>26 HS3606</td>
<td>Okha:</td>
<td>Re-[2C, D, M]</td>
<td></td>
</tr>
<tr>
<td>27 HS3845</td>
<td>Okha:</td>
<td>Re-[2C, D, M]</td>
<td></td>
</tr>
<tr>
<td>28 HS945</td>
<td>Tomari:</td>
<td>MUS-I, Re-[3C, D, M]</td>
<td></td>
</tr>
<tr>
<td>29 HS1946</td>
<td>Japan:</td>
<td>Kushiro</td>
<td>DOM, MUS-II</td>
</tr>
<tr>
<td>30 HS1947</td>
<td>Asahikawa:</td>
<td>MUS-II, Re-[1C, D, M]</td>
<td></td>
</tr>
<tr>
<td>31 HS2781</td>
<td>Obhiro:</td>
<td>MUS-II</td>
<td></td>
</tr>
<tr>
<td>32 HS2340</td>
<td>Otaru:</td>
<td>Re-[5C, D, M]</td>
<td></td>
</tr>
<tr>
<td>33 HS2446</td>
<td>Nagamuma:</td>
<td>MUS-II</td>
<td></td>
</tr>
<tr>
<td>34 HS2451</td>
<td>Date:</td>
<td>MUS-II</td>
<td></td>
</tr>
<tr>
<td>35 HS2779</td>
<td>Kyowa:</td>
<td>MUS-II</td>
<td></td>
</tr>
<tr>
<td>36 HS3834</td>
<td>Rankoshi:</td>
<td>Re-[1C, D, M]</td>
<td></td>
</tr>
<tr>
<td>37 HS394</td>
<td>Onuma:</td>
<td>MUS-II, Re-[5C, D, M]</td>
<td></td>
</tr>
<tr>
<td>38 HS2461</td>
<td>Sakekawa:</td>
<td>Re-[1C, D, M]</td>
<td></td>
</tr>
<tr>
<td>39 HS2458</td>
<td>Tendo:</td>
<td>MUS-II, Re-[1C, D, M]</td>
<td></td>
</tr>
<tr>
<td>40 HS2454</td>
<td>Ototsu:</td>
<td>MUS-II</td>
<td></td>
</tr>
<tr>
<td>41 HS2456</td>
<td>Sendai:</td>
<td>MUS-II</td>
<td></td>
</tr>
<tr>
<td>42 HS2457</td>
<td>Sakata:</td>
<td>MUS-II, Re-[1C, D, M]</td>
<td></td>
</tr>
<tr>
<td>43 HS2468</td>
<td>Chiba:</td>
<td>MUS-II</td>
<td></td>
</tr>
<tr>
<td>44 HS3839</td>
<td>Atsugi:</td>
<td>DOM, MUS-II</td>
<td></td>
</tr>
<tr>
<td>45 MSM.HS3950</td>
<td>Mishima:</td>
<td>MUS-II</td>
<td></td>
</tr>
<tr>
<td>46 HS2472</td>
<td>Miyazaki:</td>
<td>MUS-II, Re-[7C, M]</td>
<td></td>
</tr>
<tr>
<td>47 HS3603</td>
<td>Kagoshima:</td>
<td>MUS-II, Re-[7C, M]</td>
<td></td>
</tr>
<tr>
<td>48 HS1368</td>
<td>Korea:</td>
<td>Gyeongju</td>
<td>MUS-II</td>
</tr>
<tr>
<td>49 HS3540</td>
<td>Bagan:</td>
<td>MUS-II</td>
<td></td>
</tr>
<tr>
<td>50 MG5065</td>
<td>China:</td>
<td>Beijing</td>
<td>MUS-II</td>
</tr>
</tbody>
</table>

Individuals with codes 1-59 were those analyzed in previous study (Nunome et al., 2010). These DNA samples are stored in National Institute of Genetics (NIG) and the RIKEN Bio-Resource Center (BRC), and personally by HS at Hokkaido University (HS, HL, originally collected by H. Ikeda).

Phylogroups are taken from Fig. 3 and recombinant haplogroups are indicated with its subspecies component (C, d; D, e). Individuals with codes 1-59 were those analyzed in previous study (Nunome et al., 2010). These DNA samples are stored in National Institute of Genetics (NIG) and the RIKEN Bio-Resource Center (BRC, and personally by HS at Hokkaido University (HS, HL, originally collected by H. Ikeda).
Table S1
Genetic diversity indices of the eight nuclear gene markers. The number of segregating sites (S), mean number of pairwise difference (π), its standard error, and Tajima’s D were calculated using ARLEQUIN 3.1 (Excoffier and Lischer, 2010).

<table>
<thead>
<tr>
<th>Subspecies</th>
<th>Fanca</th>
<th>Spire2</th>
<th>Tcf25</th>
<th>Mc1r upst.</th>
<th>Mc1r</th>
<th>Def8</th>
<th>Afg3ll</th>
<th>Dbnal1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CAS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>85</td>
<td>91</td>
<td>73</td>
<td>92</td>
<td>91</td>
<td>89</td>
<td>87</td>
<td>76</td>
</tr>
<tr>
<td>S</td>
<td>25</td>
<td>21</td>
<td>36</td>
<td>10</td>
<td>5</td>
<td>14</td>
<td>20</td>
<td>23</td>
</tr>
<tr>
<td>π</td>
<td>0.015624</td>
<td>0.013251</td>
<td>0.017766</td>
<td>0.004099</td>
<td>0.001137</td>
<td>0.010405</td>
<td>0.013933</td>
<td>0.014696</td>
</tr>
<tr>
<td>SE</td>
<td>+/- 0.008213</td>
<td>+/- 0.007089</td>
<td>+/- 0.009058</td>
<td>+/- 0.002743</td>
<td>+/- 0.001098</td>
<td>+/- 0.005920</td>
<td>+/- 0.001098</td>
<td>+/- 0.007645</td>
</tr>
<tr>
<td>Tajima’s D</td>
<td>-0.09297</td>
<td>0.85586</td>
<td>1.30983</td>
<td>-0.60743</td>
<td>-1.10120</td>
<td>-0.09957</td>
<td>0.90854</td>
<td>0.29922</td>
</tr>
<tr>
<td><strong>DOM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>S</td>
<td>8</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>π</td>
<td>0.009418</td>
<td>0.002058</td>
<td>0.003016</td>
<td>0</td>
<td>0.000314</td>
<td>0.001008</td>
<td>0.001272</td>
<td>0.005022</td>
</tr>
<tr>
<td>SE</td>
<td>+/- 0.00555</td>
<td>+/- 0.0017</td>
<td>+/- 0.002078</td>
<td>0</td>
<td>+/- 0.000554</td>
<td>+/- 0.00118</td>
<td>+/- 0.00117</td>
<td>+/- 0.003156</td>
</tr>
<tr>
<td>Tajima’s D</td>
<td>0.65091</td>
<td>1.19844</td>
<td>0.4709</td>
<td>0</td>
<td>-1.15945</td>
<td>0.23502</td>
<td>-1.51811*</td>
<td>-0.38804</td>
</tr>
<tr>
<td><strong>MUS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>65</td>
<td>66</td>
<td>66</td>
<td>54</td>
<td>66</td>
<td>66</td>
<td>66</td>
<td>66</td>
</tr>
<tr>
<td>S</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>π</td>
<td>0</td>
<td>0.000139</td>
<td>0.00101</td>
<td>0.002658</td>
<td>0.001926</td>
<td>0.001146</td>
<td>0.007099</td>
<td>0.010664</td>
</tr>
<tr>
<td>SE</td>
<td>0</td>
<td>+/- 0.000338</td>
<td>+/- 0.000887</td>
<td>+/- 0.002016</td>
<td>+/- 0.001926</td>
<td>+/- 0.001208</td>
<td>+/- 0.004042</td>
<td>+/- 0.005721</td>
</tr>
<tr>
<td>Tajima’s D</td>
<td>0</td>
<td>-1.43204*</td>
<td>-1.4308</td>
<td>2.25873</td>
<td>-0.61065</td>
<td>-1.44214*</td>
<td>0.22811</td>
<td>2.20778</td>
</tr>
</tbody>
</table>

n: the number of sequences examined (n). The number of segregating sites (S), mean number of pairwise difference (π), its standard error, and Tajima’s D

*P < 0.05.