Research Article

Evolutionary and dispersal history of Eurasian house mice *Mus musculus* clarified by more extensive geographic sampling of mitochondrial DNA

Hitoshi Suzuki[1*], Mitsuo Nunome[1], Gohta Kinoshita[1], Ken P. Aplin[2], Peter Vogel[3], Alexey P. Kryukov[4], Mei-Lei Jin[5], Sang-Hoon Han[6], Ibnu Maryanto[7], Kimiyuki Tsuchiya[8], Hidetoshi Ikeda[9], Toshihiko Shiroishi[10], Hiromichi Yonekawa[11], and Kazuo Moriwaki[12]

[1]Laboratory of Ecology and Genetics, Graduate School of Environmental Earth Science, Hokkaido University, Sapporo 060-0810, Japan;

[2]Division of Mammals, National Museum of Natural History, Smithsonian Institution, Washington D.C., 20013-7012, U.S.A.;

[3]Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland;

[4]Institute of Biology and Soil Science, Russian Academy of Sciences, Vladivostok 690022, Russia;

[5]Shanghai Research Center of Biotechnology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200233, China;

[6]National Institute of Biological Resources, Environmental Research Complex, Incheon 404-170, Korea;

[7]Museum Zoologicum Bogoriense, Indonesian Institute of Sciences, Jl Raya Jakarta

Km 46, Cibinong 16911, Indonesia;

[8]Laboratory of Bioresources, Applied Biology Co., Ltd., Minami-Aoyama, Minato-ku, Tokyo 107-0062, Japan;

[9]Department of Veterinary Public Health, Nippon Veterinary and Animal Science University, Musashino, Tokyo 180-8602, Japan;

[10]Mammalian Genetics Laboratory, National Institute of Genetics, Research Organization of Information and Systems, Mishima 411-8540, Japan;

[11]Department of Laboratory Animal Science, Tokyo Metropolitan Institute of Medical Science, Tokyo 156-8506, Japan;

[12]RIKEN, Bioresource Center, Tsukuba 305-0074, Japan;

*Correspondence should be addressed: Hitoshi Suzuki, Fax +81-11-706-2279

htsuzuki@ees.hokudai.ac.jp

Abstract

We examined sequence variation of mitochondrial DNA control region and cytochrome *b* gene of the house mouse (*Mus musculus sensu lato*) drawn from ca. 200 localities, with 286 new samples drawn primarily from previously unsampled portions of their Eurasian distribution and with the objective of further clarifying evolutionary episodes of this species before and after the onset of human-mediated long-distance dispersals. Phylogenetic analysis of the expanded data detected five equally distinct clades, with geographic ranges of northern Eurasia (*musculus*, MUS), India and Southeast Asia (*castaneus*, CAS), Nepal (unspecified, NEP), western Europe (*domesticus*, DOM), and Yemen (*gentilulus*). Our results confirm previous suggestions of Southwestern Asia as the likely place of origin of *M. musculus* and the region of Iran, Afghanistan, Pakistan, and northern India, specifically as the ancestral homeland of CAS. The divergence of the subspecies lineages and of internal sublineage differentiation within CAS were estimated to be 0.37-0.47 and 0.14-0.23 million years ago (mya), respectively, assuming a split of *M. musculus* and *Mus spretus* at 1.7 mya. Of four CAS sublineages detected, only one extends to eastern parts of India, Southeast Asia, Indonesia, Philippines, South China, Northeast China, Primorye, Sakhalin and Japan, implying a dramatic range expansion of CAS out of its homeland during an evolutionary short time, perhaps associated with the spread of agricultural practices. Multiple and non-coincident eastward dispersal events of MUS sublineages to distant geographic areas, such as northern China, Russia, and Korea, are inferred, with the possibility of several different routes.

Key words: mitochondrial DNA; cytochrome *b*; control region; phylogeography; wild house mouse

Running head: mtDNA variation in Eurasian House mice

Introduction

Despite the rapid rise of polygenic and genomic approaches to the analysis of population history (e.g. Abe et al., 2004; Stoneking and Delfin, 2010; Yang et al., 2011), the study of mitochondrial DNA (mtDNA) continues to play a significant role in the investigation of many species. In the case of the house mouse complex (*Mus musculus* Complex), the availability of large numbers of mtDNA sequences derived from European and other populations has facilitated detailed analysis of both prehistoric and historic range expansions (Rajabi-Maham et al., 2008; Gabriel et al., 2010, 2011; Bonhomme et al., 2011; Jones et al., 2010), often with significant implications for human history. By contrast, the other major lineages of the house mouse are known from far fewer sequences and this has hindered progress on even some of the most basic questions of phylogeography, such as their likely places of origin and the timing and routes of major dispersal episodes.

Early investigations of house mouse mtDNA, using the method of Restriction Fragment Length Polymorphism (RFLP; e.g. Yonekawa et al., 1981, 1986), identified three major haplogroups among wild house mice. These appeared to be associated with recognized subspecies and were designated accordingly: a DOM haplogroup in *M. m. domesticus* from western Europe and North Africa (also southern Africa, Australia and the Americas, all as historical introductions); a MUS haplogroup in *M. m. musculus* from the northern part of Eurasia excluding western Europe; and a CAS haplogroup in *M. m. castaneus* from Southeast Asia. Later studies of mtDNA suggested a number of other possible divergent lineages: a BAC haplogroup in *M. m. bactrianus* from Afghanistan and Pakistan (Boursot et al., 1993, 1996; Yonekawa et al., 1994); a GEN

haplogroup in *M. m. gentilulus* from Yemen (Prager et al., 1998) and Madagascar (Duplantier et al., 2002); and most recently, another divergent but as yet unnamed haplogroup from Nepal (Terashima et al., 2006). Broader genomic comparisons using microsatellites (Sakai et al., 2005), single nucleotide polymorphic sites (Abe et al., 2004), and whole-genome sequences (Frazer et al., 2007) support the notion that each of the MUS, CAS and DOM mtDNA haplogroups represents a longstanding evolutionary lineage. However, the remaining mtDNA haplogroups (BAC and GEN) have not been subject to the same level of scrutiny, hence their status remains uncertain.

In this paper we fill a number of the remaining gaps in geographic mtDNA coverage for the house mouse, with a particular emphasis on the Indian sub-continent, China, and far eastern Russia. Addition of mtDNA sequences from these key areas sheds light on several issues, including 1) the likely ancestral range of each of the major evolutionary lineages; and 2) the direction and timing of range expansions, with a particular focus on East Asia, China and Japan, where multiple mtDNA lineages are known to regionally co-occur (Moriwaki et al., 1984; Yonekawa et al., 1986, 2003; Terashima et al., 2006; Nunome et al., 2010a).

Materials and methods

Materials

Our new sequencing effort is based chiefly on samples of House mouse genomic DNA stored in the National Institute of Genetics, Mishima, Japan. These were collected in China, India, Russia, and a variety of other countries, on expeditions organized by KM

during 1983-2003 (MG series, stored in the National Institute of Genetics; and BRC Series, stored in the RIKEN Bio-Resource Center), and by HI and KT during 1989-1992 (HI series, stored in Hokkaido University). We also used DNA samples stored at Hokkaido University (HS series), including mice collected by PV (IZEA series, vouchered in the Institut de Zoologie et d'Ecologie Animale of Lausanne University) and KA (ANWC series, vouchered in the Australian National Wildlife Collection). Some of the same samples have been used in previous studies (e.g. Yonekawa et al., 1988, 1994, 2003; Miyashita et al., 1994; Nagamine et al., 1994; Tsuchiya et al., 1994; Munclinger et al., 2002; Spiridonova et al., 2004).

New sequences were generated for mtDNA control region (CR) from 212 House mouse individuals from 137 localities; the cytochrome *b* gene (*Cytb*) was also sequenced from a subset of 167 individuals from 106 localities (Table S1). In our sampling we strived to achieve maximum geographic coverage, at the cost of small samples sizes (frequently just one) for each locality. While this reduced the scope for sophisticated analysis of population expansion scenarios, it increased the likelihood of detecting previously undiscovered components of mtDNA diversity. The geographic distribution of the new sequences is shown in Figure 1.

We downloaded a further 571 CR sequences and 41 *Cytb* sequences of *M. musculus* from public databases, drawn primarily from the work of Prager et al. (1996, 1998), Gündüz et al. (2005), Rajabi-Maham et al. (2008) and Bonhomme et al. (2011), along with representative sequences of closely related species for use as outgroups. Our sequence alignments for each mtDNA region are provided in Appendix A and B.

Sequence analyses

The PCR and direct sequencing of the CR (around 800-bp; Yasuda et al., 2005) and *Cytb* (1140 bp; Suzuki et al., 2004) were performed according to previously described methods. Two primers were used for sequence determination of CR in *M. musculus*; CR1: 5'-CATGCCTTGACGGCTATGTT-3' and CR2: 5'-ATCGCCCATACGTTCCCCTT-3'. The double-stranded PCR product was sequenced utilizing the PRISM Ready Reaction DyeDeoxy Terminator Cycle Sequencing Kit (ABI) and an ABI3130 automated sequencer. Sequences of *M. cypriacus, M. macedonicus, M. spicilegus,* and *M. spretus* were obtained from the databases and used as outgroups in phylogenetic inference.

Phylogeny and divergence time estimation

Sequences were aligned by eye using MEGA5 (Tamura et al., 2011). Prior to further analyses, we deleted tandem repeat sequences of 75-76 bp in CR of some MUS and some CAS haplotypes (identified by Prager et al., 1996, 1998) and an 11-bp insertion in CR of some DOM sequences, while encoding the occurrence of these repeats into the taxon name to check for conformation with phyletic lineages.

To obtain a general impression of clustering topology we constructed Neighbor-Net (NN) networks for reduced datasets of 399 unique CR haplotypes and 98 unique *Cytb* haplotypes, and using the default parameters of uncorrected *P* distance and the EqualAngle algorithm, as implemented in SplitsTree 4.10 software (Bryant and Moulton, 2004). The principal advantage of this hypothesis-poor method over others that generate dichotomous branching networks or trees is that NN networks illustrates

all potentially supported splits among a group of sequences as a reticulation. The potential complexity of a dataset is thus represented rather than reduced by this method, while any predominant network topology remains visible. Further insights into the structure of each of the CAS, MUS and DOM mtDNA lineages was obtained by constructing NN networks, together with Median-Joining (MJ) networks (Bandelt et al. 1999), as implemented in SplitsTree 4.10.

Maximum likelihood (ML) phylogenies were constructed for each of the CR and *Cytb* datasets and for a concatenate dataset for 30 individuals. We used the PhylML algorithm (Guindon and Gascuel, 2003) with the HKY substitution model, as implemented on the ATGC website (http://www.atgc-montpellier.fr/). A maximum parsimony (MP) method and the neighbor-joining (NJ; Saitou and Nei, 1987) method were taken for phylogenetic inference with concatenate sequences using PAUP 4.0b10 (Swofford, 2001). Bootstrap analysis was carried out with 1000 pseudoreplicates in the ML and NJ analyses, and 100 pseudoreplicates in the MP analysis. Sub-groups are designated within each of the major mtDNA lineages only if there was moderate to good bootstrap support (BSS = 0.7 - 0.9) from the ML analysis, combined with concordant structure in the NN networks.

We estimated the age of most recent common ancestors (TMRCAs) for mtDNA clades using the *Cytb* sequences and a relaxed Bayesian molecular clock with uncorrelated rates (BEAST v1.6.1, Drummond and Rambaut, 2007), as described previously (Nunome et al., 2010b). For this analysis we used *M. cypriacus*, *M. macedonicus*, *M. spicilegus*, and *M. spretus*, the remaining members of the *Mus musculus* Species Group, as outgroup taxa. For the root node of the *Mus musculus*

Species Group we assigned a prior value of 1.7 mya (95% HPD: 1.45 - 1.95), which is a based on molecular divergences of single copy nuclear gene sequences (*Irbp* and *Rag1*), calibrated against the known fossil record of the genus *Mus* and other Murinae (Suzuki et al., 2004; Shimada et al., 2010). The monophyletic setting was applied for clades of the lineages of the four subspecies groups (CAS, MUS, DOM, and NEP) and *M. m. subspecies*. Then TMRCAs were estimated by the Bayesian Markov-chain Monte-Carlo (MCMC) method, using the HKY substitution model as selected under the Akaike Information Criterion in MrModeltest version 2.2 (Nylander, 2004). Analyses were run for 50 million generations from a UPGMA starting tree with sampling at every 5000 generations following 5 million burn-in generations. The convergence of MCMC chains and the effective sample size (ESS) values exceeding 200 for all parameters were assessed using the software Tracer version 1.5 (Rambaut and Drummond, 2009). BEAST analysis was not performed with the CR sequences due to the greater inequality in branch lengths observed on the CR ML trees, compared with the *Cytb* ML trees, suggestive of less regular substitution fixation rates over evolutionary time.

Assessment of historical demographical processes

The DnaSP programme, version 5.00.7 (Librado and Rozas, 2009), was used to estimate haplotype diversity (*Hd*), nucleotide diversity (π), mean number of pairwise differences among sequences (k), and Tajima's *D* value. The same software was used for the analysis of mtDNA sequence mismatch distributions, measured as substitutional differences between pairs of haplotypes. Estimates of the expansion parameter tau (τ)

were calculated using Arlequin version 3.5 (Excoffier and Lischer, 2010). Population expansion times were estimated under the assumption of a constant molecular clock and using mutation rates from 2.5%, 10% and 20% using the online tool developed by Schenekar and Weiss (2011; available at http://www.uni-graz.at/zoowww/mismatchcalc/mmc1.php). The goodness-of-fit of the observed distribution to the expected distribution under the sudden-expansion model (Rogers, 1995) was tested by computing the sum of squares deviation (SSD).

Results

Characteristics of major haplogroups

The NN network generated from the *Cytb* dataset features four well-differentiated haplogroups (Fig. 2a), three of correspond to the previously identified DOM, CAS and MUS lineages. The fourth haplogroup includes two sequences derived from Nepalese mice, one reported previously by Terashima et al. (2006; HS1467) and one new to this study (HS1523). For convenience, this haplogroup is herein labeled NEP to indicate its geographic origin. No *Cytb* sequences are available for the GEN haplogroup. The four *Cytb* haplogroups are equally divergent from a common central hub and outgroups either join this central hub or have an affinity with the MUS haplogroup. The CAS haplogroup appears to contain deeper lineage diversity than either the MUS or DOM haplogroups, both of which have distinctly brush-like terminal segments. Overall, the topology of the NN network for the *Cytb* dataset is suggestive of a more or less simultaneous diversification of an ancestral *Mus musculus* stock into multiple evolutionary lineages, and also indicative of much recent diversification in each of the

DOM and MUS haplogroups.

The ML phylogeny generated from the *Cytb* dataset also features the same four clades with support values between 99% (DOM) and 100% (CAS) (Fig. 3a). Monophyly of *Mus musculus* (*sensu lato*) is well-supported relative to the outgroups but there is no support for any special relationships among the four haplogroups.

The NN network generated from the CR dataset (Fig. 2b) shows a well-differentiated cluster of DOM sequences but less marked segregation among the other groups which now includes GEN. A network generated without DOM (Fig. 2c) shows well-differentiated haplogroups for GEN and MUS, and a less cohesive cluster of 5-6 haplogroups that includes 4-5 that might be regarded as 'CAS' (CAS-1, CAS-2, CAS-3, CAS-4, and AF074526) and one that includes the two NEP haplotypes (HS1467, Tukuche; HS1523, Kathmandu) as well as 'CAS' types 13 (AF074524, Kathmandu) and 14 (AF074525, Nuwakot) from Prager et al. (1998). The GEN and some CAS haplogroups are more divergent from the central hub than other haplogroups but this may be due in part to missing data in some sequences obtained by Prager et al. (1998) from museum skins.

The ML phylogeny generated from the CR dataset features five major clades with support values between 86% (CAS) and 100% (DOM) (Fig. 3b). Monophyly of *Mus musculus* (*sensu lato*) is well-supported but there is no strong support for any special relationships among the haplogroups, as well as in the *Cytb* dataset.

To further explore the phylogenetic relationships among the haplogroups, we constructed an ML phylogeny using concatenate sequences (CR+*Cytb*) for the 30 individuals represented in both datasets. The resultant trees remain ambiguous for

branching order among the four major lineages of CAS, DOM, NEP and MUS (Fig. 4).

Genetic diversity in each of the main haplogroups is summarized according to a variety of standard parameters in Table 1. Excluding NEP where n=2, for *Cytb* the highest nucleotide diversity (*Pi*) is observed in DOM, followed by CAS and MUS; while for CR nucleotide diversity is highest in CAS, followed by DOM and GEN, with MUS once again lowest. The average number of nucleotide differences (*k*) is highest for *Cytb* in DOM, followed by CAS and MUS; and highest for CR in DOM, followed by CAS, GEN, and MUS. The number of distinct haplotypes (H) and number of polymorphic sites (S) both are clearly correlated with the total number of samples (N) in each of the *Cytb* and CR datasets (Table 1).

Geographic distribution of major haplogroups

The newly determined haplotypes show geographic distributions largely consistent with expectation based on previous findings (Fig. 1a). DOM haplotypes are concentrated around the Mediterranean region but show numerous widely dispersed outliers including localities within MUS territory in western and northeastern Russia and in China, and within CAS territory in the Philippines and Indonesia; MUS haplotypes are predominant in northern part of Eurasia excluding western Europe; and CAS haplotypes are predominant across South and Southeast Asia but with outliers in Japan, the Middle East and eastern Russia.

A more detailed mapping of new and previously published sequences from South Asia through to the Middle East illustrates the concentration of mtDNA diversity in southwestern Asia for *M. musculus* as a whole and additionally for subgroups within

CAS (Fig. 1b; identity of sub-groups discussed below).

Genetic and phylogeographic structure of individual haplogroups

CAS haplogroup

Four well-differentiated sub-groups within CAS are clearly depicted in the NN network for the CR dataset (Fig. 2c) and they are also evident in the NN for the smaller *Cytb* dataset (Fig. 2a) and in the ML tree for the concatenated dataset (Fig. 4); they are herein designated as CAS-1, CAS-2, CAS-3 and CAS-4, as mentioned above. Two of these sub-groups were identified by Terashima et al. (2006) and labeled CAS-II (= CAS-1 of this study) and CAS-I (= CAS-2 of this study). An outlier CR sequence (AF074526 = CAS type 15 of Prager et al., 1998, from Ilam, western Nepal) may represent a fifth sub-group (Fig. 2c) but this requires confirmation as it was obtained from a museum skin and contains several gaps. The CAS sub-groups emerge from a central hub on the NN networks and, with the exception of AF074526, show approximately equivalent degrees of divergence. Each of the main sub-groups also shows relatively deep haplotype diversity; uniquely in CAS-1, this includes a brush-like structure suggestive of recent radiation from a common ancestral haplotype.

The ML phylogeny for the concatenated dataset (Fig. 4) recovered monophyletic clades with good support (> 90%) for CAS-1, CAS- 2 and CAS-3, indicated a close relationships between CAS-3 and CAS-4 (BRC3025) with low or moderate support (> 50%), and suggested a basal derivation of CAS-2 with low or moderate support (>50%).

The phylogeographic pattern for the CAS haplogroup appears relatively

uncomplicated. The greatest haplotype diversity is observed in Pakistan and northern India where all four sub-groups are present but CAS-2 and CAS-3 are dominant (Fig. 1b). Approximately half of the sequences of CAS-2 share a 76 bp tandem repeat (reported by Prager et al., 1996, 1998) which further supports the monophyly of this sub-group; these include mice from Taitung in Taiwan and Hanoi in Vietnam (Fig. 2d), constituting the only occurrences of the CAS-2 haplogroup outside of India and Pakistan.

Subgroups CAS-1 to CAS-3 are represented in central India but CAS-1 alone is more widely distributed, with representation in southern India and Sri Lanka, and also across southeast Asia, China and eastern Russia to Japan (Fig. 1). A NN analysis of using concatenate sequences (CR+*Cytb*) from 40 individuals of CAS-1 (Fig. 5a) suggested the presence of a further sub-division that we recognize as CAS-1a and CAS-1b. CAS-1a haplotypes come from two localities in southern China (Guilin and Kunming), from northern Japan (northern Honshu and Hokkaido), and from southern Sakhalin. CAS-1b haplotypes come from a wider geographic area including several parts of India, Bangladesh, Sri Lanka, Myanmar, southern China, Hainan Island, southern Sakhalin and Primorye, eastern Indonesia, and Morocco.

MUS haplogroup

The MUS haplogroup appears to be comprised of two main sub-groups which are herein designated MUS-1 and MUS-2. These are most clearly expressed in the NN network based on concatenated CR and *Cytb* data from 38 individuals (Fig. 5b) but they are also evident in the networks generated from the individual data sets (CR,

Fig. 2e; *Cytb*, Fig. 2f).

A total of seven clusters were identified within MUS-1 on the CR NN network (labeled i-vii on Fig. 2e); the majority of these clusters show a high level of geographic fidelity. Based on relationships observed in the NN networks for *Cytb* (Fig. 2f) and the concatenated data set (Fig. 5b), we suggest that these CR phyletic groups can be revolved into three phyletic lineages that we herein designate as MUS-1a (CR clusters i, iii, v, vii), MUS-1b (CR clusters iii, v), and MUS-1c (CR cluster vi).

Sub-group MUS-1 as a whole is represented across the entire geographic range of MUS. However, its components show high fidelity to discrete geographic areas: MUS-1a is largely confined to eastern Europe (Ukraine, Moldova, south Siberia, and Primorye in the Russian Far East; see Fig. 1 for the geographic distribution); MUS-1b is predominantly Chinese, being represented at multiple localities spanning the entire breadth of China, from Xanjiang Uyghur Autonomous Region in the northwest to Shandong Province on the eastern seaboard, though there are several occurrences of in Transcaucasia, in Iran, in eastern Europe, and in Russia adjacent to China; and MUS-1c is geographically restricted to far northeast China (Tumen, Qiqihar), Korea and Japan, with one outlier recorded from the coastal city of Kraskino, near the Russian-Korean border (Fig. 1c).

The MUS-2 sub-group is distributed across the eastern half of the range of MUS, with representation to the north and east of the Caspian Sea (Kazakhstan and Turkmenistan, respectively), south Siberia in the Altai Mountains, Novosibirsk and Irkutsk, Primorye, and across China including localities in the far northwest (Ili Khazakh Autonomous Prefecture), the central region (Ningxia Hui Autonomous

Region), the far north (Manasi), the Tibetan Plateau (Lhasa), and Shandong Province in the east (Liyang). Most of the MUS-2 haplotypes recorded to date are known from single localities and many differ by two or more nucleotide substitutions from the closest sequences (Fig. 2e, f). Only two MUS-2 haplotypes were detected at multiple localities and neither appears to be an ancestral haplotype. In each case, the shared haplotypes are recorded from widely separated localities, suggestive of recent long-distance dispersal or translocation.

DOM haplogroup

The NN network for DOM CR sequences is an explosively radiating structure, likened by Bonhomme et al. (2011: Fig. 1b) to a "multiple-armed sea star" (Fig. 2g). The additional 23 CR sequences added in this study do not disrupt the primary structure of the NN network with eleven haplogroups (HGs), though HGs 1 and 2 appear somewhat more mixed than in the presentation of Bonhomme et al. (2011: Fig. 1b) and the small HG9 appears to have disaggregated into basal positions within HGs 1, 2 and 7. Most of our new sequences fall into HG11 which corresponds with Clade F of Jones et al. (2010), including sequences from the novel (outlier) localities of Somalia (HS3700), central China (MG509, MG566), and Java, Indonesia (HS2322).

Most of the HGs are also evident in an ML tree (not shown) though supporting values were low (65% for HG5 and less than 50% for others). However, HG9 occupies a more diffuse central position consistent with its lack of unity in the NN network, and HGs 1 and 2 were not supported, though most members of these putative HGs associated correctly in smaller clades. An aggregation of HG7 with HG8,

is also evident in both the NN network and the ML tree, albeit with no substantial support in the ML analysis.

The smaller *Cytb* dataset presents a simpler picture (Fig. 2h). The NN network shows 12 clusters, some of which are represented by single sequences. Six of the clusters can be correlated to CR HGs based on the subset of individuals represented in both datasets. A MJ network (not shown) shows a completely stellar arrangement with minimal reticulation and with all terminal haplotypes similarly divergent (3-6 nucleotide substitutions) from a central node. This putative ancestral haplotype has not been detected. HGs 8 and 9 derive from a common primary branch on the MJ network.

The various analyses performed on DOM sequences do not suggest any grounds for its formal sub-division, as was suggested above for each of CAS and MUS. Rather, the topology appears to be genuinely explosive, involving differentiation of multiple, regionally-based matrilines, as concluded also by Rajabi-Maham et al. (2008, 2012) and Bonhomme et al. (2011).

Divergence time among and within the haplogroups

Divergence estimates generated by BEAST for each of the four major haplogroups have central values that range between 0.37-0.46 mya (Fig. 6); in each case the 95% highest probability density values have spans of around ± 0.16 mya (Table 2). The TMRCA of subgroup diversification within each of the CAS, MUS, DOM and NEP haplogroups was estimated at 0.22 ± 0.08, 0.15 ± 0.07, 0.13 ± 0.04 and 0.14 ± 0.06 mya, respectively (Table 2).

The Tajima's *D* values for all haplogroups and subgroups were significantly

negative, indicating various phases of rapid population growth involving mice with matrilines in CAS-1, CAS-1b, MUS-1, MUS-1b, MUS-1c and DOM (Table 3). We estimated the age of population growth in each of the phyletic groups under four different mutation rates of *Cytb*; 2.5%, 10% and 20% (Table 3).

The mismatch distribution for the CAS-1 *Cytb* dataset (Supplementary Fig. S1) shows a multi-peaked distribution which is consistent with the notion of CAS-1 as a well-structured haplogroup (Fig. 5a). CAS-1b shows a good mismatch conformation to a model of recent population growth, with further support coming from a statistically significant negative value for Tajima's *D* (Table 3). Tajima's *D* was negative but not statistically significant for CAS-1a. In MUS there is support for recent population expansion of MUS-1 as a whole and for each of MUS-1b and MUS-1c, each backed up by statistically significant negative values for Tajima's *D*, though the SSD value for MUS-1 was significant ($P < 0.01$), as evidence for departure from the estimated model of population expansion (Table 3). In contrast, there is no support for recent population expansion of either MUS-1a and MUS-2. The mismatch distribution for DOM both CR (data not shown) and *Cytb* datasets (Fig. S1) shows near perfect conformation with the population growth and decline model provided by DNASP. Neutrality test statistics also point to a significant phase of population expansion in the recent history of DOM (Tajima's $D = -2.02$, $P < 0.05$), as concluded previously by others (Rajabi-Maham et al., 2008; Bonhomme et al., 2011), though the SSD value was significant ($P = 0.024$).

Discussion

Much of our current understanding of *Mus musculus* phylogeography remains little

modified from the conclusions of early studies of mtDNA (e.g. Boursot et al., 1993, 1996; Yonekawa et al., 1994; Prager et al., 1996, 1998; Boissinot and Boursot, 1997) and of allozymes and other nuclear markers (e.g. Bonhomme et al., 1984; Miyashita et al., 1994; Din et al., 1996). Three of the most persistent notions to emerge from these early studies are: 1) the understanding that the common ancestor of all of the major *M. musculus* haplogroups arose in the region of western to central Eurasia, either somewhere in the mountainous terrain that extends from Transcaucasia through to northwest India (Boursot et al., 1993, 1996; Din et al., 1996) or possibly in the low-lying region of Mesopotamia (Prager et al., 1993, 1996); 2) the belief that the broader distribution of all major haplogroups is due to range expansions that occurred following the development of commensalism and thus within the last 10,000 years; and 3) the conclusion that the CAS lineage is genetically more diverse and probably older than either of DOM or MUS, with MUS probably trailing DOM in this regard.

To date these notions have been subject to detailed scrutiny only for the DOM haplogroup (Gündüz et al., 2005; Darvish et al., 2006; Rajabi-Maham et al., 2008; Bonhomme et al., 2011; Duvaux et al., 2011). In this case, the majority of results uphold the general assumptions as outlined above.

For each of the MUS and CAS haplogroups the most comprehensive phylogeographic analyses prior to this study were contained in the work of Prager et al. (1996, 1998). For their initial study of the *musculus* and *domesticus* lineages geographic sampling was heavily biased toward Europe, with only a smattering of samples derived from the eastern range of *musculus*. In the later study this was partially rectified through the laborious extraction of DNA from museum skin samples from

eastern populations of *musculus* and *castaneus*. Despite this remarkable effort, major geographic gaps in sampling remained; and with such large geographic areas to cover, sample sizes were small for all regions.

Our sampling has filled many of the gaps in geographic coverage, especially for the Indian subcontinent, Indochina and the Far East. However, the issue of small sample sizes remains and will not be solved without further field collecting on a multi-regional scale. Nevertheless, our broader sampling produces new insights into the phylogeography of each of the CAS and MUS groups, and allows us to challenge several key aspects of the current understanding.

A homeland for *M. musculus* in southwestern Asia

The ancestral homeland of *Mus musculus* is most likely to coincide with a broad region of co-occurrence of the various phylogroups and it should encompass or abut the geographic range of the most restricted phylogroups, namely GEN and NEP of the Arabian Peninsula and Himalayan region, respectively. Under these criteria, the region of southwestern Asia, encompassing modern day Iraq, Iran, Afghanistan, Pakistan, and northwestern India stands out as the most likely candidate area. Bonhomme et al. (1984) reached the same conclusion on different evidence, namely the higher levels of variation in nuclear genes among mice in this area compared with peripheral regions (see also Suzuki et al., 1986; Boursot et al., 1996; Boissinot and Boursot, 1997; Prager et al., 1998; Darvish et al., 2006; Duvaux et al., 2011). As discussed at length by Prager et al. (1998), mtDNA lineage boundaries in this area show general association with major geographic barriers (see also Duvaux et al., 2011). In particular, the Zagros

Mountains divide DOM in the west from CAS in the east, while the Elburz Mountains divide MUS in the north from CAS in the south. Similarly, the mountain chains of the Hindu Kush separate populations of MUS and CAS in northern Afghanistan, though the present day distribution of the mtDNA haplotypes is not always associated with the mountainous range (e.g., MUS in Kabul, Afghanistan, Fig. 1a). A process of allopatric differentiation is indicated, as suggested also by the lack of overt ecological differentiation among the divergent populations.

Our divergence estimates of 0.37-0.47 mya (see Table 2 for confidence interval) for the major mitochondrial phylogroups are in good accord with previous determinations (Rajabi-Maham et al., 2008; Terashima et al., 2006). Initial lineage diversification evidently predates the dispersal of modern humans out of Africa, hence it is likely that initial phases of range expansion were not mediated by human activity, unless of course the impact of early human populations on the environment was much greater than currently understood.

An interesting biogeographic observation is that the inferred place of origin of *M. musculus* southwestern Asia lacks any other co-occurring mouse species belong to subgenus *Mus*. In this regard, it differs from each of peninsular India, where *M. booduga* and *M. terricolor* of the *M. booduga* Species Groups are both found (Musser and Carleton, 2005); Indochina, that hosts a variety of species in the *M. booduga* and *M. cervicolor* Species Groups (Suzuki and Aplin, 2012); and eastern Europe, where other species of the *Mus musculus* species group are present. It is tempting to speculate that the presence of these ecologically similar native species in surrounding areas formerly served to constrain the geographic distribution of *M. musculus*.

The distribution and ecology of contemporary *castaneus* populations in Asia provides further clues to its regional history. As summarized by Marshall (1977: 205-206) the only 'outdoor commensal' (i.e. agricultural field) populations of CAS mice are found in the semi-arid habitats of Pakistan (the *bactrianus* morphotype). Elsewhere on the Indian subcontinent and through into Southeast Asia, house mice are found only as 'indoor commensals'; furthermore, across Southeast Asia, they are generally confined to larger towns and absent in rural villages (see also Aplin et al., 2006). Marshall (1977) attributed the absence of house mice from agricultural contexts in these areas to competitive exclusion by other species of *Mus*, notably members of the *Mus booduga* Species Group on the Indian subcontinent and members of the *Mus cervicolor* Species Group in South East Asia; and he attributed the absence of house mice in rural villages in Southeast Asia to the presence of commensal species of *Rattus* such as *R. exulans.*

Phylogeography of the CAS lineage

The CAS lineage has been subject to two different phylogeographic interpretations. Boursot et al. (1993, 1996) proposed that the northern Indian subcontinent was both the place of origin of *Mus musculus* and the cradle of genetic diversity within this group. This model was based on the discovery in this area of numerous highly divergent mtDNA lineages (Boursot et al., 1996) and levels of nuclear diversity (as determined by allozyme electrophoresis) that exceeded those found in European populations of *domesticus* and *musculus* (Din et al., 1996). To explain these dual observations Boursot et al. (1993, 1996) proposed a 'centrifugal' model of differentiation in which the

ancestors of each of the *domesticus*, *musculus* and *castaneus* lineages dispersed to the west, east and north, each carrying a subset of the mtDNA and nuclear diversity, and subsequently undergoing local differentiation. They referred to populations in the ancestral area as "*Mus musculus* subsp." and restricted use of the name *castaneus* to populations in southern India, southern China and Indochina. Boursot et al. (1996) referred to the northern Indian and Pakistani populations as an 'oriental group', while Yonekawa et al. (1994) subsequently applied the existing name *bactrianus* to these populations.

Prager et al.'s (1998) version of CAS phylogeography is based primarily on interpretation of mtDNA phylogeny. While confirming a high diversity of mtDNA types in northern India and Pakistan, they regarded these to be part of a monophyletic *castaneus* lineage distinct from each of *domesticus*, *musculus* and the newly recognized *gentilulus* lineage of the Arabian Peninsula. Prager et al. (1998) developed a model of 'sequential' derivation of the lineages to reflect their phylogenetic branching order – *domesticus* being the oldest branch, followed by *gentilulus*, *castaneus* and *musculus*. They preferred to locate the ancestral pre-*domesticus* stock in the Near East, within the current range of *domesticus*, and regarded the progressive derivation of other lineages as a consequence of sequential dispersal events that took house mice south onto the Arabian Penisula, then east onto the Indian subcontinent, and finally, north through the mountains of northwest India and Pakistan to occupy the great Eurasian steppe. Within CAS, Prager et al. (1998:858) suggest a relatively long phase of regional diversification on the Indian subcontinent, followed by a 'more recent' dispersal into the 'humid lowlands of Southeast Asia'.

Our sampling for CAS is relatively extensive and the results go far towards illuminating the historical phylogeography of this haplogroup. In keeping with the findings of Prager et al. (1998) and contrary to the predictions of Boursot et al. (1993, 1996), we recovered reciprocal monophyly with good to excellent support among all of the major haplogroups, including CAS. While Prager et al. (1998) considered the branching order among the major haplogroups to be resolved, our larger CR and *Cytb* dataset fails to provide a robust phylogenetic structure at this level, although there is a suggestion of special affinity between CAS and MUS, and between DOM and NEP. Like both groups of previous researchers, we found the highest mtDNA diversity and depth in CAS populations inhabiting the mountainous region of northwest India and Pakistan, with a loss of haplotype lineage diversity from north to south on the Indian subcontinent (Boursot et al., 1996), and from west to east into Southeast Asia (Prager et al., 1998). Despite this general agreement, Boursot et al (1996) clearly regard *castaneus* in their restricted application of the name to be a long-term resident of Southeast Asia, while Prager et al. (1998) portray this as a relatively recent phase of dispersal of *castaneus* though without specifying any time frame.

Low nucleotide diversity in the widely distributed CAS-1 sub-group is evident in this study. This is consistent with that observed by the recent work on the *castaneus* subspecies group done by Rajabi-Maham et al. (2012; see also Bonhomme and Searle, 2012). Our results are suggestive of a relatively recent range expansion of CAS-1 to a large geographic areas covering the south and east Indian subcontinent, Southeast Asia, Indonesia, South China, Northeast China and the Russian Far East (Fig. 5). On the other hand, the presence of the locally restricted phyletic group, CAS-1a is suggestive

of stepwise historical range expansion of CAS-1. Haplotype diversity within CAS-1a, the sub-group found in mice from South China (Kunming and Guilin), northern Honshu, Hokkaido, and South Sakhalin, was most likely produced by subsequent dispersal and is suggestive of several thousands of years of *in situ* evolution. Furthermore, the location of the Japanese cluster at the far eastern periphery of the CAS distribution implies a significantly earlier onset for dispersal onto the Indian subcontinent and thence through to East Asia.

We suspect that the dispersal of CAS-1 mice occurred in response to ecological transformation of the landscape by early agriculturalists and the emergence of urban centers and trade networks. As has been postulated for the Middle East (Auffray et al., 1990; Cucchi and Vigne, 2006), South Asian populations of *Mus musculus* are likely to have benefited from the creation of new agricultural landscapes, and the common practice of storing harvested grain inside villages and even inside houses provided the context for development of commensalism. Long-distance dispersal is part and parcel of commensalism, with mice being carried as stowaways during transport of grain, building materials, clothing and bedding (Pocock et al. 2005).

Although the archaeological record of agriculture is less comprehensive for Asia than for the Middle East and Europe, there is good evidence for domestication of cereal crops including rice and millet by about 9,000 years ago in several parts of South and East Asia (Khush 1997; Londo et al., 2006; Zheng et al., 2009; Molina et al., 2011) and even earlier evidence for long distance overland and maritime trade (Oka and Kusimba, 2008). Assuming that populations experienced a sudden or exponential growth, we calculated τ values from the *Cyt-b* sequences and estimated times since the

onset of population expansions. Higher rates of mutation (e.g. 10% or 20% per million years per lineage) rather than lower rates (e.g. 2.5%) are considered to be realistic for assessing rather recent diversifying events (Ho et al., 2005). We obtained a τ value of 1.7 for CAS-1b (n=17) which under mutation rates (per million years per lineage) of 10% and 20%, gives expansion times of 7,600 and 3,800 years, respectively (Table 3). A τ value was not calculated for CAS-1a but it too is likely to have commenced its dispersal and diversification in China, the Russian Far East and Japan in prehistoric times. In this regard, it is of interest to note archaeological evidence for rice cultivation along the upper Yangtze river (e.g. Yunnan province, here represented by mice from Kunming) at 4500 years ago (Fuller et al., 2010) and recent genetic evidence from an intensive genome survey on wild and cultivated rice, suggesting the Pearl River (Guangxi province, here represented by mice from Guilin) in southern China is the place of the first development of cultivated rice (Bonhomme and Searle, 2012; Huang et al., 2012).

Comparatively recent long-distance dispersal of CAS mice most likely explains the detection of CAS-2 haplotypes at isolated localities in Taiwan and Vietnam, though we could not exclude out the possibility that these are relictual haplotypes, either rare survivors of an earlier dispersal of CAS-2 mice out of India that was swamped by a later CAS-1 dispersal, or the last remnants of incomplete lineage sorting of an immigrant population with a mixture of CAS-1 and CAS-2 haplotypes. In the case of the individual from Taiwan, the fact that its nuclear genetic profile is fully consistent with other East Asian populations of CAS and differ from mice from India

and Pakistan with the CAS-2 mtDNA haplotypes (Nunome et al., 2010a; Kodama et al., unpublished) suggests that we are not dealing with a novel invader but perhaps with a product of mtDNA introgression.

Phylogeography of the MUS lineage

Previous phylogeographic interpretations of the house mouse group do not vary much in regard to the geographic origin of the MUS haplogroup. Boursot et al. (1993: 406) speculated that "the cradle of *M. m. musculus* could be in Transcaucasia or east of the Caspian Sea", while Prager et al. (1993, 1996) saw the origin of MUS as the product of northward dispersal from a proto-CAS population occupying the region east of the Caspian Sea, followed by range expansion. Both groups of researchers also agree that MUS populations subsequently dispersed west into central Europe and east into China and Japan, and this scenario has been adopted as paradigmatic by Japanese researchers interested in the origin of the indigenous *molossinus* population (Yonekawa et al., 1988; Terashima et al., 2006; Nunome et al., 2010a). Yonekawa et al. (1988) postulated that MUS populations relatively recently expanded into China where CAS populations had already colonized but few other workers have expressed an opinion on the earlier timing of the remarkable eastward expansion of MUS. Nunome et al. (2010a) suggested a latitudinal division within MUS between northern (MUS-I) and southern (MUS-II) groups, based on phylogeographic analyses of nuclear gene sequences, and posited that range expansion of the MUS haplogroup from west to east across continental Eurasia followed separate northern and southern dispersal routes, with separate expansion again into eastern Europe.

Much of the interest in the geographic distribution of MUS has focused on its genetic interaction with mice of other haplogroups. In the European context numerous studies have examined the evolutionary dynamics of a narrow hybrid zone with DOM that runs from Norway through Denmark, Germany and Austria to eastern Bulgaria (Hunt and Selander, 1973; Sage et al., 1993; Boursot et al., 1993; Jones et al., 2010). There are grounds to believe that initial contact may have occurred further west in Europe with the current position stabilizing after a period of eastward retreat of *musculus* (Gyllensten and Wilson, 1987). Whatever the case, the age of the contact zone is constrained by the timing of the DOM migrations along the shores of the Mediterranean, an event that is thought to date to within the last 2-3,000 years (Cucchi et al., 2005).

In Transcaucasia, gene flow between complexly parapatric populations of MUS and DOM is thought to explain a 300 km wide zone of genetic admixture (Mezhzherin et al., 1989; Frisman et al., 1990; Milishnikov et al., 1990); however, an alternative interpretation attributes the genetic diversity to a high level of ancestral polymorphism in the regional MUS population (Milishnikov et al., 2004), equivalent to that observed among the 'oriental group' of mice in northern India and Pakistan (Boursot et al., 1993, 1996; Din et al., 1996). This would be consistent with long residency of the MUS population in this area. MUS and CAS populations also come into secondary contact in China (Moriwaki et al., 1994); however, both the geography and the genetic outcome of these interactions remain poorly documented.

Our expanded sampling among eastern House mouse populations sheds significant new light on the evolutionary history of the MUS haplogroup. We identify

two major sub-groups within MUS – MUS-1 and MUS-2 – and a total of three phylogeographic components within MUS-1: MUS-1a in Moldova, Ukraine, N Caspian Sea and Russian Siberia; MUS-1b in East Europe, Kazakhstan and China; and MUS-1c in Korea and Japan. The origin of the MUS-1 and MUS-2 sub-groups is ancient, with a divergence estimate from BEAST of 150,000 ± 13,000 years (Fig. 6, Table 2). Both sub-groups are represented in the area around the Caspian Sea and it seems likely that both matrilines originated within this ancestral geographic area.

Rapid population expansion was inferred for each of MUS-1b and MUS-1c (Table 3). Estimates of expansion times for these lineages (Table 3) suggest an early expansion of MUS-1b in northern China (τ = 4.9 CI: 2.9 – 6.5; e.g., 21,000 and 10,800 years ago, with an assumption of the mutation rate of 10% and 20% per million years per lineage, respectively), followed by a later expansion of MUS-1c in northeastern Russia, Korea and Japan at (τ = 1.5 CI: 0.7 – 2.5; e.g. 6,600 and 3,300 years ago).

The notion of ancient population expansions in eastern Eurasia is clearly at odds with the conventional notion of a recent west to east dispersal of the MUS haplogroup. However, other lines of genetic evidence similarly point to a long residency of the MUS haplogroup in central Russia and the Far East. For example, the beta-hemoglobin gene (*Hbb*) shows contrasting predominant alleles in the lower Yellow River basin and in the remaining western portion of northern China (Hbb^{p} and Hbb^{w1}, respectively; Miyashita et al., 1994; see also Moriwaki, 1994); and mice from the eastern part of China are known to have relatively longer tails (tail ratio: ~93%) than those from the rest of MUS territory in China (81%; Tsuchiya et al., 1994).

Finally, we note that the area in which MUS-1c is predominant – the Korean

Peninsula and nearby continental area – harbors unique genetic components in both Y-specific gene sequences and nuclear gene sequences (e.g. Nagamine et al., 1994; Terashima et al., 2006; Nunome et al., 2010a). Under the existing paradigm of west to east dispersal, these phylogeographic patterns might be attributed either to genetic drift following migration of ancestral populations with diverse genetic components or to multiple migration events by mice carrying different genetic components, perhaps by different routes. However, neither of these scenarios can readily account for the evidence of ancient population expansions within geographically restricted matrilines. Accordingly, we favor the alternative model of regional differentiation within a long-term resident population.

The fossil record should be able to arbitrate this issue and it is of great interest to note that paleontologists have long recognized *Mus musculus* as a component of the Chinese mammal fauna since the middle part of the Middle Pleistocene (i.e. c. 500,000 years ago); e.g. Zheng et al. (1997 and references cited therein). While the taxonomic identification of the fossils might be challenged, the determination is at least plausible given the molecular evidence for early diversification among Chinese *musculus* populations. However, there is a risk of circularity in such arguments and an urgent need for critical appraisal of the relevant fossils.

The European sub-group MUS-1a contains substantial haplotype diversity including persistent ancestral haplotypes and two deeply divided haplotype series, each of which contains relatively shallow stellar clusters derived from populations near the western limit of the MUS geographic range. This pattern is suggestive of a broad westward expansion of a MUS population into eastern Europe, with limited filtering of

haplotype diversity. As summarized by Auffray et al. (1990), the long history of *Mus musculus* in Europe is dominated by large expansions and contractions of range driven by glacial cycles. At the height of the last glaciation *M. musculus* was rare or absent across most of eastern Europe which supported a mosaic of periglacial forest-steppe, steppe and semi-desert habitats (Markova et al., 2009). Refugial forest habitats were restricted to small patches in the Crimea, in the Transcarpathian region, and in the Caucasus (Markova et al., 2009) and it is of interest to note fossil occurrences of *M. musculus* in the Carpathian-Balkan region during the warm interval (33-24,000 years ago) immediately prior to the last glacial maximum (Markova, 2010). However, in view of the high level of genetic diversity within MUS-1a and the lack of a strong signal of recent population expansion, it seems likely that mice persisted in multiple localities, perhaps including both forest and semi-desert habitats. This issue warrants further consideration.

MUS-1a contains a discrete lineage characterized by a 75-bp duplication, first detected by Prager et al. (1998) in a mouse from Kishinev in Moldova. We found closely related haplotypes at low frequency in mice from eastern Europe (e.g. Donetsk, Ukraine) and also from Khasan in Primorye, Russia (Fig. 5b). Given the other evidence of regional differentiation of mtDNA within MUS, we are inclined to view MUS-1a as originally restricted to eastern Europe (Ukraine), with its more easterly occurrences being a product of long-range transport by modern means. The locality of Novosibirsk, for example, is sited on the Turkestan-Siberia Railway that was built in the early 20th Century (in 1930) and connects the Caspian Sea to localities in Central Asia. The link to the Primorye region of the Russian Far East is less readily accounted for by overland

transportation but might be explained by the activities of the Russian government to introduce kazak and peasants to the Russian Far East in the late 19th Century; upwards of 90,000 people (and perhaps a few mice) from Odessa in the Ukraine settled in the Ussuri Region of Primorye (http://www.fegi.ru/prim/geografy/etap.htm).

Phylogeography of the DOM lineage

Our small number of new DOM sequences contributes only a few insights into the history of this well-studied lineage (Gabriel et al., 2011; Bonhomme et al., 2011; Jones et al., 2010). The onset of the expansion ~~of~~ is estimated to be 12,000 years ago as the youngest timing, assuming the mutation rate of 20% per million years per lineage (Table 3), which is harmonious with the recent arguments based on zooarcheological records (Cucchi et al., 2005; Rajabi-Maham et al., 2008; Bonhomme and Searle, 2012).

We recovered the expected "Clade F" haplotypes from mice collected in North America, Australia and Africa (Senegal, Somalia) but also detected them in mice from several localities in Asia, namely Lanzhou and Xining in China, and Bogor on Java in Indonesia. At Bogor, CAS and DOM mtDNA haplotypes were found to co-occur in one population.

A high frequency of DOM haplotypes was also detected in the Russian Far East, thereby supporting previous claims of DOM-MUS-CAS interactions in this area based on studies of chromosomes, allozymes and RAPDs (Frisman et al., 2011; Spiridonova et al., 2011). Interestingly though, the DOM haplotypes recovered at Primorye (HS1466) and Sakhalin (HS3606, HS3607) are not "Clade F" but are related specifically to haplotypes from Cameroon (e.g. AFWCMR41; Bonhomme et al., 2011).

This connection is very likely explained by long-distance dispersals associated with human activities in modern times.

The detection of DOM haplotypes in numerous corners of the world is testimony to the ongoing dispersal of *M. musculus*, and encourages further study of the impact of occasional arrival of 'exotic' mice on the genetic constitution of pre-established mouse populations (Rajabi-Maham et al., 2008; Searle et al., 2009a, b; Gabriel et al., 2010; Bonhomme et al., 2011). To further illustrate this point, mice with both DOM and CAS mtDNA haplotypes have been captured in Japanese international ports (Tsuda et al., 2007) and Nunome et al. (2010a) provided robust evidence from their nuclear haplotype analysis of genetic introgression by DOM components of Japanese house mice. The extent to which genetic introgression may now be shaping the future evolution of the house mouse is an interesting topic – one that has bearing on other commensal mammals including the black rat *Rattus rattus* which also displays comparable signals of former geographic subdivision and recent intermingling as a consequence of commensalism and human-assisted dispersal (Chinen et al., 2005; Aplin et al., 2011; Bastos et al., 2011; Lack et al., 2012).

Concluding remarks

The expanded mtDNA dataset raises a number of important new issues regarding the prehistory of the house mouse. Most significantly, it has identified one particular CAS sub-group (CAS-1) that has expanded into southern India, Southeast and East Asia, and raised the possibility that this expansion is linked to the emergence of agricultural lifestyles and of Asian civilizations. Also of significance is our suggestion that MUS

populations have a long history of residency in eastern Russia and China, contrary to the existing paradigm of recent expansion from west to east. Finally, our results emphasize the role of long-distance dispersal in shaping contemporary pattern of distribution and opportunities for interaction between each of the major lineages within *Mus musculus*.

Our study also demonstrates the value of continuing efforts to fill gaps in geographic coverage of *Mus musculus* mtDNA. Moreover, it highlights the need for ongoing field collecting to increase local sampling and the need for more comprehensive assessments of population genetic history using nuclear markers. From our preliminary work with nuclear genes on this group, it is clear that much deeper divergence between subspecies groups is observed in some regions of the genome than in others (e.g., Suzuki et al., 2004; Nunome et al., 2010a), and also evident that different markers can yield strongly contrasting phylogeographic structure, such as in southern China, where CAS mtDNA dominates but both CAS and MUS components are detected in nuclear genes (e.g., Nunome et al., 2010a). Finally, it is worth mentioning the as yet unexplored potential for detailed study of Central and East Asian house mouse populations to reveal important new aspects of human history, including the emergence of agricultural lifestyles and of regional trade networks.

DATA ARCHIVING

The nucleotide sequences reported in this paper appear in the DDBJ, EMBL, and GenBank nucleotide sequence databases under the following accession numbers AB649455–AB649770, AB819902-AB819920 and AB820897-AB820942 (Table S1).

Sequence data files in nexus file format, together with Supplementary Information files are stored at Dryad repository: doi:10.5061/dryad.rf161.

ACKNOWLEDGEMENTS

We wish to express our appreciation to Kuniya Abe, Masahiro A. Iwasa, Martua H. Sinaga, and Shumpei P. Yasuda, for their valuable advice in this study. We thank Francois Catzeflis, Angela Frost, Naoto Hanzawa, Hideo Igawa, Oleg E. Lopatin, Hiromi Okamura, Yoshifumi Matsushima, Hidetoshi Matsuzawa, Natan Mise, Nobumoto Miyashita, Pavel Munclinger, Robert Palmer, Kenkichi Sasaki, Hironori Ueda, Keiichi Yokoyama and numerous other collectors of mice for kind help in supplying the valuable samples used in this study. This study was in part supported by Grant-in-Aid for Scientific Research (C) from Japan Society for the Promotion of Science (JSPS, 23570101). We would like to thank Heiwa Nakajima Foundation for its generous support.

References

Abe K, Noguchi H, Tagawa K, Yuzuriha M, Toyoda A, Kojima T *et al*. (2004). Contribution of Asian mouse subspecies *Mus musculus molossinus* to genomic constitution of strain C57BL/6J as defined by BAC-end sequence-SKIP analysis. *Genom Res* **14:** 2439–2447.

Aplin KP, Brown PR, Singleton GR, Douangboupha B, Khamphoukheo K (2006) Rodents in the rice environments of Laos. In: Schiller JM, Chanphengxay MB, Linquist B, Appa Rao S (eds). *Rice in Laos*. International Rice Research Institute: Los Banos, pp 291–308.

Aplin K, Suzuki H, Chinen AA, Chesser RT, ten Have J, Donnellan SC *et al*. (2011). Multiple geographic origins of commensalism and complex dispersal history of black rats. *PLoS ONE* **6:** e26357.

Auffray J-C, Vanlerberghe F, Britton-Davidian J. (1990). The house mouse progression in Eurasia: a palaeontological and archaeozoological approach. *Biol J Linne Soc* **41:** 13–25.

Bandelt HJ, Forster P, Rohl A (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16:** 37–48.

Bastos A, Nair D, Taylor P, Brettschneider H, Kirsten F *et al*. (2011). Genetic monitoring detects an overlooked cryptic species and reveals the diversity and distribution of three invasive *Rattus* congeners in south Africa. *BMC Genet* **12:** 26.

Bryant D, Moulton V (2004). Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol* **21:** 255–265.

Bonhomme F, Anand R, Darviche D, Din W, Boursot P (1994). The house mouse as a ring species? In: Moriwaki K, Shiroishi T, Yonekawa H (eds). *Genetics in Wild Mice*, Japan Scientific Society Press: Tokyo/S Karger Basal, pp 13–23.

Bonhomme F, Catalan J, Britton-Davidson J, Chapman VM, Moriwaki K, Nevo E *et al*. (1984). Biochemical diversity and evolution in the genus *Mus*. *Biochem Genet* **22:** 275–303.

Bonhomme F, Miyashita N, Boursot P, Catalan J, Moriwaki K (1989). Genetical variation and polyphyletic origin in Japanese *Mus musculus*. *Heredity* **63:** 299–308.

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. *Genom Biol* **8:** R80.

Bonhomme F, Orth A, Cucchi T, Rajabi-Maham H, Catalan J, Boursot P *et al*. (2011). Genetic differentiation of the house mouse around the Mediterranean basin: matrilineal footprints of early and late colonization. *Proc R Soc B* **278:** 1034–1043.

Bonhomme F, Searle JB (2012). House mouse phylogeography. In: Macholán M, Baird SJE, Munclinger P, Piálek J (eds). *Evolution of the house mouse* (*Cambridge series in morphology and molecules*), Cambridge: Cambridge University Press, pp 278–296.

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

Boissinot S, Boursot P (1997). Discordant phylogeographic patterns between the Y chromosome and mitochondrial DNA in the house mouse – selection on the Y chromosome? *Genetics* **146:** 1019–1034.

Boursot P, Din W, Anand R, Darviche D, Dod B, von Deimling F *et al.* (1996). Origin and radiation of the house mouse: mitochondrial DNA phylogeny. *J Evol Biol* **9:** 391–415.

Chinen AA, Suzuki H, Aplin KP, Tsuchiya K, Suzuki S (2005). Preliminary genetic characterization of two lineages of black rats (*Rattus rattus* sensu lato) in Japan with evidence for introgression at several localities. *Gene Genet Syst* **80:** 367–375.

Cucchi T, Vigne JD, Auffray JC (2005). First occurrence of the house mouse (*Mus musculus domesticus*, Schwarz and Schwarz 1943) in the Western Mediterranean: a zooarchaeological revision of subfossil occurrences. *Biol J Linn Soc* **84:** 429–445

Cucchi T, Vigne JD (2006). Origin and diffusion of the house mouse in the Mediterranean. *Hum Evol* **21:** 95–106.

Darvish J, Orth A, Bonhomme F (2006). Genetic transition in the house mouse, *Mus musculus* of Eastern Iranian Plateau. *Folia Zool* **55:** 349–357.

Din W, Anand R, Boursot P, Darviche D, Dod B, Jouvin-Marche E *et al*. (1996). Origin and radiation of the house mouse: clues from nuclear genes. *J Evol Biol* **9:** 519–539.

Drummond AJ, Rambaut A (2007). Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7:** 214.

Duplantier JM, Orth A, Catalan J, Bonhomme F (2002). Evidence for a mitochondrial lineage originating from the Arabian peninsula in the Madagascar House Mouse. *Heredity* **89:** 154–158.

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.

Excoffier L, Lischer HEL (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* **10:** 564–567.

Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ *et al*. (2007). A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **448:** 1050–1053.

Frisman LV, Korobitsyna KV, Yakimenko LV, Vorontsov NN (1990). Biochemical Groups of House Mice Inhabiting the Soviet Union, in *Evolyutsionnye geneticheskie issledovaniya mlekopitayushchikh: Tezisy dokladov* (Evolutionary Genetic Studies in Mammals: Proc. Conf.), Vladivostok: Dal'nevost. Otd. Akad. Nauk SSSR, part 1, pp 35–54.

Frisman LV, Korobitsyna KV, Yakimenko LV, Munteanu AI, Moriwaki K (2011). Genetic variability and the origin of house mouse from the territory of Russia and neighboring countries. *Russ J Genet* **47:** 590–602.

Fuller DQ, Sato YI, Castillo C, Qin L, Weisskopf A, Kingwell-Banham E *et al*. (2010). Consilience of genetics and archaeobotany in the entangled history of rice. *Archaeol Anthropol Sci* **2:** 115–131

1 Gabriel SI, Jóhannesdóttir F, Jones EP, Searle JB (2010) Colonization, mouse-style.
2 *BMC Biol* **8:**131.
3 Gabriel SI, Stevens MI, Mathias ML, Searle JB (2011). Of mice and 'convicts': Origin
4 of the Australian house mouse, *Mus musculus. PLoS ONE* **6:** et622.
5 Guindon S, Gascuel O (2003). A simple, fast, and accurate algorithm to estimate large
6 phylogenies by maximum likelihood. *Syst Biol* **52:** 696–704.
7 Gündüz İ, Rambau RV, Tez C, Searle JB (2005). Mitochondrial DNA variation in the
8 western house mouse (*Mus musculus domesticus*) close to its site of origin:
9 studies in Turkey. *Biol J Linn Soc* **84:** 473–485.
10 Gyllensten U, Wilson AC (1987). Interspecific mitochondrial DNA transfer and the
11 colonization of Scandinavia by mice. *Genet Res* **49:** 25–29.
12 Ho SYW, Phillips MJ, Cooper A, Drummond AJ (2005). Time dependency of
13 molecular rate estimates and systematic overestimation of recent divergence
14 times. *Mol Biol Evol* **22:** 1561–1568.
15 Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q *et al.* (2012). A map of rice
16 genome variation reveals the origin of cultivated rice. *Nature* **490:** 497–501.
17 Hunt WG, Selander RK (1973). Biochemical genetics of hybridization in European
18 house mouse, *Heredity* **31:** 11–33.
19 Jones EP, van der Kooij J, Solheim R, Searle JB (2010). Norwegian house mice (*Mus*
20 *musculus musculus*/*domesticus*): Distributions routes of colonization and
21 patterns of hybridization. *Mol Ecol* **19:** 5252–5264.
22 Jones EP, Skirnisson K, McGovern TH, Gilbert MTP, Willerslev E, Searle JB (2012).
23 Fellow travellers: a concordance of colonization patterns between mice and

men in the North Atlantic region. *BMC Evol Biol* **12:** 35.

Khush GS (1997). Origin dispersal cultivation and variation of rice. *Plant Mol Biol* **35:** 25–34.

Lack JB, Greene DU, Conroy CJ, Hamilton MJ, Braun JK, Mares MA *et al*. (2012). Invasion facilitates hybridization with introgression in the *Rattus rattus* species complex. *Mol Ecol* **21:** 3545–3561.

Librado P, Rozas J (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25:** 1451–1452.

Londo JP, Chiang YC, Hung KH, Chiang TY, Schaal BA (2006). Phylogeography of Asian wild rice *Oryza rufipogon* reveals multiple independent domestications of cultivated rice *Oryza sativa*. *Proc Natl Acad Sci USA* **103:** 9578–9583.

Markova AK, Simakova AN, Puzachenko AY (2009). Ecosystems of Eastern Europe at the time of maximum cooling of the Valdai glaciation (24–18 kyr BP) inferred from data on plant communities and mammal assemblages. *Quat Internat* **201:** 53–59.

Markova AK, Puzachenko AY, van Kolfschoten T (2010). The North Eurasian mammal assemblages during the end of MIS 3 (Brianskian–Late Karginian–Denekamp Interstadial). *Quat Internat* **212:** 149–158

Marshall JT (1977). A synopsis of Asian species of *Mus* (Rodentia, Muridae). *Bull Am Mus Nat Hist* ***158*:** 173–220.

Mezhzherin SV, Kotenkova EV, Mikhailenko AG (1998). The house mice, *Mus musculus s.l.*, hybrid zone of Trans- Caucasus. Zeitschrift für Säugetierkunde. **63:** 154–168.

Milishnikov AN, Lavrenchenko LA, Lavrenchenko LA, Orlov VN (1990). High-level introgression of *Mus domesticus* genes in Transcaucasian *Mus musculus* s. str. Populations. *Dokl. Akad. Nauk SSSR* **311:** 764–768.

Milishnikov AN, Lavrenchenko LA, Lebedev VS (2004). Origin of the house mice (superspecies complex *Mus musculus sensu lato*) from the Transcaucasian region: A new look at dispersal routes and evolution. *Genetika* **40:** 1234–1250.

Miyashita N, Kawashima T, Wang CH, Jin ML, Wang F, Gotoh H *et al*. (1994). Genetic polymorphisms of *Hbb* haplotypes in wild mice. In: Moriwaki K, Shiroishi T, Yonekawa H (eds) *Genetics in Wild Mice*, Japan Scientific Society Press: Tokyo/S Karger Basal, pp 85–93.

Molina J, Sikora M, Garud N, Flowers JM, Rubinsteina S, Reynoldsb A *et al.* (2011). Molecular evidence for a single evolutionary origin of domesticated rice. *Proc Natl Acad Sci USA* **108**: 8351–8356.

Moriwaki K (1994). Wild mouse from geneticist's viewpoint. In: Moriwaki K, Shiroishi T, Yonekawa H (eds) *Genetics in Wild Mice*. Japan Scientific Society Press: Tokyo/S Karger Basal, pp xiii–xxiv.

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

Moriwaki K, Shiroishi T, Yonekawa H (1994). Genetics in Wild Mice. It's application to Biomedical Research. Japan Scientific Societies Press: Tokyo/S Karger Basal.

Munclinger P, Bozıkova E, Sugerkova M, Pialek J, Macholan M (2002). Genetic variation in house mice (*Mus* Muridae Rodentia) from the Czech and Slovak Republics. *Folia Zool* **51:** 81–92.

Musser GG, Carleton MD, (2005). Family Muridae. In: Wilson DE, Reeder DM (eds). *Mammal Species of the World*, 3rd edn. The John Hopkins University Press: Baltimore, pp 894–1531.

Nagamine CM, Shiroishi T, Miyashita N, Tsuchiya K, Ikeda H, Namikawa T *et al*. (1994). Distribution of the Molecularossinus allele of *Sry* the testis-determining gene in wild mouse. *Mol Biol Evol* **11:** 864–874.

Nunome M, Ishimori C, Aplin KP, Yonekawa H, Moriwaki K, Suzuki H (2010a). Detection of recombinant haplotypes in wild mice (*Mus musculus*) provides new insights into the origin of Japanese mice. *Mol Ecol* **19:** 2474–2489.

Nunome M, Torii H, Matsuki R, Kinoshita G, Suzuki H (2010b). The influence of Pleistocene refugia on the evolutionary history of the Japanese hare *Lepus brachyurus. Zool Sci* **27:** 7469–754.

Nylander JAA (2004). MrModeltest v2. Program distributed by author. Evolutionary Biology Centre, Uppsala University, Uppsala.

Oka R, Kusimba C (2008). The Archaeology of Trading Systems Part 1: Towards a New Trade Synthesis. *J Archaeol Res* **16:** 339–395.

Pocock Pocock MJO, Hauffe HC, Searle JB (2005). Dispersal in house mice. *Biol J Linn Soc* **84:** 565–583.

Prager EM, Sage RD, Gyllensten ULF, Thomas WK, Huebner R, Jones CS *et al*. (1993). Mitochondrial DNA sequence diversity and the colonization of Scandinavia by house mice from East Holstein. *Biol J Linn Soc* **50:** 85–122.

Prager EM, Tichy H, Sage RD (1996). Mitochondrial DNA sequence variation in the eastern house mouse *Mus musculus*: comparison with other house mice and report of a 75-bp tandem repeat. *Genetics* **143:** 427–446.

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

Rajabi-Maham H, Orth A, Bonhomme F (2008). Phylogeography and postglacial expansion of *Mus musculus domesticus* inferred from mitochondrial DNA coalescent from Iran to Europe. *Mol Ecol* **17:** 627–641.

Rajabi-Maham H, Orth A, Siahsarvie R, Boursot P, Darvish J, Bonhomme F (2012). The south-eastern house mouse *Mus musculus castaneus* (Rodentia: Muridae) is a polytypic subspecies. *Biol J Linn Soc* **107:** 295–306.

Rambaut A, Drummond AJ (2009). Tracer v1.5. Available from http://beast.bio.ed.ac.uk/Tracer.

Sage RD, Atchley WR, Capanna E (1993). House mice as models in systematic biology. *Syst Biol* **42:** 523–561.

Saitou N, Nei M (1987). The neighbor-joining method: a mew method for reconstructing phylogenetic trees. *Mol Biol Evol* **4:** 406–425.

Sakai T, Kikkawa Y, Miura I, Inoue T, Moriwaki K, Shiroishi T *et al*. (2005). Origins of mouse inbred strains deduced from whole-genome scanning by polymorphic

microsatellite loci. *Mammal Genome* **16:** 11–19.

Schenekar T, Weiss S (2011). High rate of calculation errors in mismatch distribution analysis results in numerous false inferences of biological importance. *Heredity* **107:** 511–512.

Searle JB, Jamieson PM, Gündüz İ, Stevens MI, Jones EP, Gemmill CEC et al. (2009a). The diverse origins of New Zealand house mice. *Proc R Soc B* **276:** 209–217

Searle JB, Jones CS, Gündüz İ, Scascitelli M, Jones EP, Herman JS et al. (2009b). Of mice and (Viking?) men: phylogeography of British and Irish house mice. *Proc R Soc B* **276:** 201–207.

Shimada T, Aplin KP, Suzuki H (2010). *Mus lepidoides* (Muridae Rodentia) of Central Burma is a distinct species of potentially great evolutionary and biogeographic significance. *Zool Sci* **27:** 449–459.

Spiridonova LN, Chelomina GN, Moriwaki K, Yonekawa H, Bogdanov AS (2004). Genetic and taxonomic diversity of the house mouse *Mus musculus* from the Asian part of the former Soviet Union. *Russ J Genet* **40:** 1378–1388.

Spiridonova LN, Kiselev KV, Korobitsyna KV (2011). Discordance in the distribution of markers of different inheritance systems (nDNA mtDNA and chromosomes) in the superspecies complex *Mus musculus* as a result of extensive hybridization in Primorye. *Russ J Genet* **47:** 100–109.

Stoneking M, Delfin F (2010). The human genetic history of East Asia: weaving a complex tapestry. *Curr Biol* **20:** R188–193.

Suzuki H, Miyashita N, Moriwaki K, Kominami R, Muramatsu M, Kanehisa T *et al.* (1986). Evolutionary implication of heterogeneity of the nontranscribed spacer

region of ribosomal DNA repeating units in various subspecies of *Mus musculus*. *Mol Biol Evol* **3:** 126–137.

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.

Suzuki H, Aplin KP (2012). Phylogeny and biogeography of the genus *Mus* in Eurasia. In: Macholán M, Baird SJE, Munclinger P, Piálek J (eds). *Evolution of the house mouse* (*Cambridge series in morphology and molecules*), Cambridge: Cambridge University Press, pp 35–64.

Swofford DL (2001). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28:** 2731–2739.

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

Tsuchiya K, Miyashita N, Wang CH, Wu XL, He XQ, Jin ML *et al*. (1994). Taxonomic study of the genus *Mus* in China, Korea and Japan–Morphologic identification. In: Moriwaki K, Shiroishi T, Yonekawa H (eds). *Genetics in Wild Mice*, Japan Scientific Society Press: Tokyo/S Karger Basal, pp 3–12.

Tsuda K, Tsuchiya K, Aoki H, Iizuka S, Shimamura H, Suzuki S *et al*. (2007). Risk of

accidental invasion and expansion of allochthonous mice in Tokyo metropolitan coastal areas in Japan. *Genes Genet Syst* **82:** 421–428.

Yang H, Wang JR, Didion JP, Buus RJ, Bell TA, Welsh CE *et al* (2011). Subspecific origin and haplotype diversity in the laboratory mouse. *Nat Genet* **43:** 648–655.

Yasuda SP, Vogel P, Tsuchiya K, Han SH, Lin LK, Suzuki H (2005). Phylogeographic patterning of mtDNA in the widely distribtuted harvest mouse (*Micromys minutus*) suggests dramatic cycles of range contraction and expansion. *Can J Zool* **83:** 1411–1420.

Yonekawa H, Gotoh O, Tagashira Y, Matsushima N, Shi L, Cho WS *et al*. (1986). A hybrid origin of Japanese mice "*Mus musculus molossinus*". *Cur Topics Microbiol Immunol* **127:** 62–67.

Yonekawa H, Moriwaki K, Gotoh O, Miyashita N, Matsushima N, Shi LM *et al*. (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, Hayashi JI, Watanebe J, Miyashita N *et al*. (1981). Evolutionary relationships among five subspecies *Mus musculus* based on restriction enzyme cleavage patterns of mitochondrial DNA. *Genetics* **98:** 801–816.

Yonekawa H, Takahama S, Gotoh O, Miyashita N, Moriwaki K (1994). Genetic diversity and geographic distribution of *Mus musculus* subspecies based on the polymorphism of mitochondrial DNA. In: Moriwaki K, Shiroishi T, Yonekawa H (eds). *Genetics in Wild Mice*. Japan Scientific Societies Press: Tokyo/S Karger Basal, pp 25–40.

Yonekawa H, Tsuda K, Tsuchiya K, Yakimenko L, Korobitsyna K, Chelomina GN *et al*. (2003). Genetic diversity geographic distribution and evolutionary relationships of *Mus musculus* subspecies based on polymorphisms of mitochondrial DNA. In: Kryukov A, Yakimenko L (eds) *Problems of evolution*, Dalnauka: Vladivostok. Vol 5, pp. 90–108.

Zheng SH, Zhang ZQ, Liu LP (1997). Pleistocene mammals from fissure-fillings of Sunjiashan hill, Shandong, China. *Vertebrata PalAsiatica* **35:** 201–216. (In Chinese with English summary).

Zheng YF, Sun GP, Qin L, Li C, Wu X, Chen X (2009). Rice fields and modes of rice cultivation between 5000 and 2500 BC in east China. *J Archaeol Sci* **36:** 2609–2616.

Figure legends

Figure 1. Collection localities and mitochondrial genotypes in Eurasia of *Mus musculus* samples examined in this study (**a**). New samples genotyped for this study are shown. Detailed locality names and sample codes are listed in Supplementary Table 1. Five major mitochondrial groups representing five subspecies groups, *M. m. musculus* (blue: MUS), *M. m. domesticus* (red: DOM), and *M. m. castaneus* (yellow: CAS), *M. m. gentilulus* (white: GEN), and the divergent lineage occurring in Nepal (orange: NEP) are differentially shown. The specific haplotype group of DOM that broadly dispersed to a variety of countries (Australia, Canada, China, Germany, Indonesia, Senegal, Somalia) are marked with arrowheads. Together with those from Prager et al. (1998), spatial patterns for the mitochondrial genotypes are shown for mice from Central Asia based on combination of new and previously published sequences (sources) (**b**), where further subdivision of the CAS lineage into four (CAS-1, CAS-2, CAS-3, CAS-4) are detected. The types of the four subgroups of CAS are shown in circle with numerical numbers (black, Prager et al., 1998; red, in this study). Further subdivision of the MUS lineages into two, MUS-1 (light blue) and MUS-2 (dark blue), and the MUS-1 sublineage into three (MUS-1a, MUS-1b, MUS-1c) is suggested in this study (**a**, **c**).

Figure 2. Neighbor-Net networks tree based on the cytochrome *b* gene (*Cytb*; **a**, **f**, **h**) and control region (CR; **b**, **c**, **d**, **e**, **g**) of the mitochondrial DNA, with tip labels for the three major subspecies groups, *M. m. musculus* (MUS), *M. m. castaneus* (CAS) and *M.*

m. domesticus (DOM) and two rather geographically confined groups of *M. m. gentilulus* (GEN) and Nepalease mice (NEP). The portion of the CR network was enlarged to show the details of the branching patterns for CAS-2, in which most of members possess a 75-bp repeat (**d**). The codes for the haplogroups (HGs) in the CR (**g**) and *Cytb* (**h**) network for DOM were taken from those used in Bonhomme et al. (2011).

Figure 3. ML trees for mitochondrial DNA sequences of the cytochrome *b* gene (**a**) and control region (**b**). The PhylML algorithm (Guindon and Gascuel, 2003) was used for the tree reconstruction and bootstrap analysis (100 replications). Bootstrap values (>50%) are shown under basal branches.

Figure 4. ML tree for concatenated mitochondrial DNA haplotypes (control region and cytochrome *b* gene) using representatives for the four major haplogroups of *Mus musculus* and *M. macedonicus* as outgroup. Bootstrap values (>50%) are shown under basal branches (ML/MP/NJ).

Figure 5. Neighbor-Net networks of concatenate sequences of control region and cytochrome *b* gene (ca. 2020 bp) from individuals representing the sublineage of CAS, CAS-1 (**a**) and MUS (**b**). Prominent subgroups appeared in the networks are indicated.

Figure 6. Divergence time estimates (million years ago, mya) of *Mus musculus* phylogroups and its closely related species, based on a Bayesian relaxed molecular

clock applied to the mitochondrial cytochrome *b* sequences (1140 bp). The posterior probability and 95% HPD intervals of node ages in mya (gray bars) are shown in particular nodes with ancient divergent. The time estimates of 1.7 mya for the root node of the divergence of *M. spretus* and the other species of *M. musculus* Species Group (Suzuki et al., 2004) was used as calibration point. Sequences obtained from the databases are marked with their accession numbers and asterisks.