



| | |
|-------------------------------|--|
| Title | Evolutionary and dispersal history of Eurasian house mice <i>Mus musculus</i> clarified by more extensive geographic sampling of mitochondrial DNA |
| Author(s) | Suzuki, Hitoshi; Nunome, Mitsuo; Kinoshita, Gohta; Aplin, Ken P.; Vogel, Peter; Kryukov, Alexey P.; Jin, Mei-Lei; Han, Sang-Hoon; Maryanto, Ibnu; Tsuchiya, Kimiyuki; Ikeda, Hidetoshi; Shiroishi, Toshihiko; Yonekawa, Hiromichi; Moriwaki, Kazuo |
| Citation | Heredity, 111(5): 375-390 |
| Issue Date | 2013-11 |
| Doc URL | http://hdl.handle.net/2115/55342 |
| Type | article (author version) |
| Additional Information | There are other files related to this item in HUSCAP. Check the above URL. |
| File Information | Suzuki_Heredity_Huscup.pdf (本文) |



[Instructions for use](#)

1 **Research Article**

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

5

6 Hitoshi Suzuki^{1*}, Mitsuo Nunome¹, Gohta Kinoshita¹, Ken P. Aplin², Peter
7 Vogel³, Alexey P. Kryukov⁴, Mei-Lei Jin⁵, Sang-Hoon Han⁶, Ibnu
8 Maryanto⁷, Kimiyuki Tsuchiya⁸, Hidetoshi Ikeda⁹, Toshihiko Shiroishi¹⁰,
9 Hiromichi Yonekawa¹¹, and Kazuo Moriwaki¹²

10

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

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

15 ³Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne,
16 Switzerland;

17 ⁴Institute of Biology and Soil Science, Russian Academy of Sciences, Vladivostok
18 690022, Russia;

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

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

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

1 Km 46, Cibinong 16911, Indonesia;

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

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

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

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

10 ¹²RIKEN, Bioresource Center, Tsukuba 305-0074, Japan;

11

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

13 htsuzuki@ees.hokudai.ac.jp

14

1 **Abstract**

2 We examined sequence variation of mitochondrial DNA control region and cytochrome
3 *b* gene of the house mouse (*Mus musculus sensu lato*) drawn from ca. 200 localities,
4 with 286 new samples drawn primarily from previously unsampled portions of their
5 Eurasian distribution and with the objective of further clarifying evolutionary episodes
6 of this species before and after the onset of human-mediated long-distance dispersals.
7 Phylogenetic analysis of the expanded data detected five equally distinct clades, with
8 geographic ranges of northern Eurasia (*musculus*, MUS), India and Southeast Asia
9 (*castaneus*, CAS), Nepal (unspecified, NEP), western Europe (*domesticus*, DOM), and
10 Yemen (*gentilulus*). Our results confirm previous suggestions of Southwestern Asia as
11 the likely place of origin of *M. musculus* and the region of Iran, Afghanistan, Pakistan,
12 and northern India, specifically as the ancestral homeland of CAS. The divergence of
13 the subspecies lineages and of internal sublineage differentiation within CAS were
14 estimated to be 0.37-0.47 and 0.14-0.23 million years ago (mya), respectively,
15 assuming a split of *M. musculus* and *Mus spretus* at 1.7 mya. Of four CAS sublineages
16 detected, only one extends to eastern parts of India, Southeast Asia, Indonesia,
17 Philippines, South China, Northeast China, Primorye, Sakhalin and Japan, implying a
18 dramatic range expansion of CAS out of its homeland during an evolutionary short
19 time, perhaps associated with the spread of agricultural practices. Multiple and
20 non-coincident eastward dispersal events of MUS sublineages to distant geographic
21 areas, such as northern China, Russia, and Korea, are inferred, with the possibility of
22 several different routes.
23

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

2 house mouse

3

4 **Running head: mtDNA variation in Eurasian House mice**

5

1 **Introduction**

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

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

1 haplogroup in *M. m. gentilulus* from Yemen (Prager et al., 1998) and Madagascar
2 (Duplantier et al., 2002); and most recently, another divergent but as yet unnamed
3 haplogroup from Nepal (Terashima et al., 2006). Broader genomic comparisons using
4 microsatellites (Sakai et al., 2005), single nucleotide polymorphic sites (Abe et al.,
5 2004), and whole-genome sequences (Frazer et al., 2007) support the notion that each
6 of the MUS, CAS and DOM mtDNA haplogroups represents a longstanding
7 evolutionary lineage. However, the remaining mtDNA haplogroups (BAC and GEN)
8 have not been subject to the same level of scrutiny, hence their status remains
9 uncertain.

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

18

19 **Materials and methods**

20 **Materials**

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

1 during 1983-2003 (MG series, stored in the National Institute of Genetics; and BRC
2 Series, stored in the RIKEN Bio-Resource Center), and by HI and KT during
3 1989-1992 (HI series, stored in Hokkaido University). We also used DNA samples
4 stored at Hokkaido University (HS series), including mice collected by PV (IZEA
5 series, vouchered in the Institut de Zoologie et d'Ecologie Animale of Lausanne
6 University) and KA (ANWC series, vouchered in the Australian National Wildlife
7 Collection). Some of the same samples have been used in previous studies (e.g.
8 Yonekawa et al., 1988, 1994, 2003; Miyashita et al., 1994; Nagamine et al., 1994;
9 Tsuchiya et al., 1994; Munclinger et al., 2002; Spiridonova et al., 2004).

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

18 We downloaded a further 571 CR sequences and 41 *Cytb* sequences of *M.*
19 *musculus* from public databases, drawn primarily from the work of Prager et al. (1996,
20 1998), Gündüz et al. (2005), Rajabi-Maham et al. (2008) and Bonhomme et al. (2011),
21 along with representative sequences of closely related species for use as outgroups. Our
22 sequence alignments for each mtDNA region are provided in Appendix A and B.

23

1 **Sequence analyses**

2 The PCR and direct sequencing of the CR (around 800-bp; Yasuda et al.,
3 2005) and *Cytb* (1140 bp; Suzuki et al., 2004) were performed according to previously
4 described methods. Two primers were used for sequence determination of CR in *M.*
5 *musculus*; CR1: 5'-CATGCCTTGACGGCTATGTT-3' and CR2:
6 5'-ATCGCCCATACGTTCCCCTT-3'. The double-stranded PCR product was
7 sequenced utilizing the PRISM Ready Reaction DyeDeoxy Terminator Cycle
8 Sequencing Kit (ABI) and an ABI3130 automated sequencer. Sequences of *M.*
9 *cypriacus*, *M. macedonicus*, *M. spicilegus*, and *M. spretus* were obtained from the
10 databases and used as outgroups in phylogenetic inference.

11

12 **Phylogeny and divergence time estimation**

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

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

1 all potentially supported splits among a group of sequences as a reticulation. The
2 potential complexity of a dataset is thus represented rather than reduced by this method,
3 while any predominant network topology remains visible. Further insights into the
4 structure of each of the CAS, MUS and DOM mtDNA lineages was obtained by
5 constructing NN networks, together with Median-Joining (MJ) networks (Bandelt et al.
6 1999), as implemented in SplitsTree 4.10.

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

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

1 Species Group we assigned a prior value of 1.7 mya (95% HPD: 1.45 - 1.95), which is
2 a based on molecular divergences of single copy nuclear gene sequences (*Irbp* and
3 *Rag1*), calibrated against the known fossil record of the genus *Mus* and other Murinae
4 (Suzuki et al., 2004; Shimada et al., 2010). The monophyletic setting was applied for
5 clades of the lineages of the four subspecies groups (CAS, MUS, DOM, and NEP) and
6 *M. m. subspecies*. Then TMRCAs were estimated by the Bayesian Markov-chain
7 Monte-Carlo (MCMC) method, using the HKY substitution model as selected under
8 the Akaike Information Criterion in MrModeltest version 2.2 (Nylander, 2004).
9 Analyses were run for 50 million generations from a UPGMA starting tree with
10 sampling at every 5000 generations following 5 million burn-in generations. The
11 convergence of MCMC chains and the effective sample size (ESS) values exceeding
12 200 for all parameters were assessed using the software Tracer version 1.5 (Rambaut
13 and Drummond, 2009). BEAST analysis was not performed with the CR sequences due
14 to the greater inequality in branch lengths observed on the CR ML trees, compared
15 with the *Cytb* ML trees, suggestive of less regular substitution fixation rates over
16 evolutionary time.

17

18 **Assessment of historical demographical processes**

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

1 were calculated using Arlequin version 3.5 (Excoffier and Lischer, 2010). Population
2 expansion times were estimated under the assumption of a constant molecular clock
3 and using mutation rates from 2.5%, 10% and 20% using the online tool developed by
4 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
5 distribution under the sudden-expansion model (Rogers, 1995) was tested by
6 computing the sum of squares deviation (SSD).
7

8

9 **Results**

10 **Characteristics of major haplogroups**

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

1 DOM and MUS haplogroups.

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

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

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

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

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

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

11

12 **Geographic distribution of major haplogroups**

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

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

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

2

3 **Genetic and phylogeographic structure of individual haplogroups**

4 *CAS haplogroup*

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

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

23 The phylogeographic pattern for the CAS haplogroup appears relatively

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

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

18

19 *MUS haplogroup*

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

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

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

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

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

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

8

9 *DOM haplogroup*

10 The NN network for DOM CR sequences is an explosively radiating structure, likened
11 by Bonhomme et al. (2011: Fig. 1b) to a “multiple-armed sea star” (Fig. 2g). The
12 additional 23 CR sequences added in this study do not disrupt the primary structure of
13 the NN network with eleven haplogroups (HGs), though HGs 1 and 2 appear somewhat
14 more mixed than in the presentation of Bonhomme et al. (2011: Fig. 1b) and the small
15 HG9 appears to have disaggregated into basal positions within HGs 1, 2 and 7. Most of
16 our new sequences fall into HG11 which corresponds with Clade F of Jones et al.
17 (2010), including sequences from the novel (outlier) localities of Somalia (HS3700),
18 central China (MG509, MG566), and Java, Indonesia (HS2322).

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

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

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

10 The various analyses performed on DOM sequences do not suggest any
11 grounds for its formal sub-division, as was suggested above for each of CAS and MUS.
12 Rather, the topology appears to be genuinely explosive, involving differentiation of
13 multiple, regionally-based matriline, as concluded also by Rajabi-Maham et al. (2008,
14 2012) and Bonhomme et al. (2011).

15

16 **Divergence time among and within the haplogroups**

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

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

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

5 The mismatch distribution for the CAS-1 *Cytb* dataset (Supplementary Fig.
6 S1) shows a multi-peaked distribution which is consistent with the notion of CAS-1 as
7 a well-structured haplogroup (Fig. 5a). CAS-1b shows a good mismatch conformation
8 to a model of recent population growth, with further support coming from a statistically
9 significant negative value for Tajima's *D* (Table 3). Tajima's *D* was negative but not
10 statistically significant for CAS-1a. In MUS there is support for recent population
11 expansion of MUS-1 as a whole and for each of MUS-1b and MUS-1c, each backed up
12 by statistically significant negative values for Tajima's *D*, though the SSD value for
13 MUS-1 was significant ($P < 0.01$), as evidence for departure from the estimated model
14 of population expansion (Table 3). In contrast, there is no support for recent population
15 expansion of either MUS-1a and MUS-2. The mismatch distribution for DOM both CR
16 (data not shown) and *Cytb* datasets (Fig. S1) shows near perfect conformation with the
17 population growth and decline model provided by DNASP. Neutrality test statistics
18 also point to a significant phase of population expansion in the recent history of DOM
19 (Tajima's $D = -2.02$, $P < 0.05$), as concluded previously by others (Rajabi-Maham et al.,
20 2008; Bonhomme et al., 2011), though the SSD value was significant ($P = 0.024$).

21

22 **Discussion**

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

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

14 To date these notions have been subject to detailed scrutiny only for the
15 DOM haplogroup (Gündüz et al., 2005; Darvish et al., 2006; Rajabi-Maham et al.,
16 2008; Bonhomme et al., 2011; Duvaux et al., 2011). In this case, the majority of results
17 uphold the general assumptions as outlined above.

18 For each of the MUS and CAS haplogroups the most comprehensive
19 phylogeographic analyses prior to this study were contained in the work of Prager et al.
20 (1996, 1998). For their initial study of the *musculus* and *domesticus* lineages
21 geographic sampling was heavily biased toward Europe, with only a smattering of
22 samples derived from the eastern range of *musculus*. In the later study this was partially
23 rectified through the laborious extraction of DNA from museum skin samples from

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

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

10

11 **A homeland for *M. musculus* in southwestern Asia**

12 The ancestral homeland of *Mus musculus* is most likely to coincide with a broad region
13 of co-occurrence of the various phylogroups and it should encompass or abut the
14 geographic range of the most restricted phylogroups, namely GEN and NEP of the
15 Arabian Peninsula and Himalayan region, respectively. Under these criteria, the region
16 of southwestern Asia, encompassing modern day Iraq, Iran, Afghanistan, Pakistan, and
17 northwestern India stands out as the most likely candidate area. Bonhomme et al.
18 (1984) reached the same conclusion on different evidence, namely the higher levels of
19 variation in nuclear genes among mice in this area compared with peripheral regions
20 (see also Suzuki et al., 1986; Boursot et al., 1996; Boissinot and Boursot, 1997; Prager
21 et al., 1998; Darvish et al., 2006; Duvaux et al., 2011). As discussed at length by Prager
22 et al. (1998), mtDNA lineage boundaries in this area show general association with
23 major geographic barriers (see also Duvaux et al., 2011). In particular, the Zagros

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

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

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

1 The distribution and ecology of contemporary *castaneus* populations in Asia
2 provides further clues to its regional history. As summarized by Marshall (1977:
3 205-206) the only ‘outdoor commensal’ (i.e. agricultural field) populations of CAS
4 mice are found in the semi-arid habitats of Pakistan (the *bactrianus* morphotype).
5 Elsewhere on the Indian subcontinent and through into Southeast Asia, house mice are
6 found only as ‘indoor commensals’; furthermore, across Southeast Asia, they are
7 generally confined to larger towns and absent in rural villages (see also Aplin et al.,
8 2006). Marshall (1977) attributed the absence of house mice from agricultural contexts
9 in these areas to competitive exclusion by other species of *Mus*, notably members of
10 the *Mus booduga* Species Group on the Indian subcontinent and members of the *Mus*
11 *cervicolor* Species Group in South East Asia; and he attributed the absence of house
12 mice in rural villages in Southeast Asia to the presence of commensal species of *Rattus*
13 such as *R. exulans*.

14

15 **Phylogeography of the CAS lineage**

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

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

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

1 Our sampling for CAS is relatively extensive and the results go far towards
2 illuminating the historical phylogeography of this haplogroup. In keeping with the
3 findings of Prager et al. (1998) and contrary to the predictions of Boursot et al. (1993,
4 1996), we recovered reciprocal monophyly with good to excellent support among all of
5 the major haplogroups, including CAS. While Prager et al. (1998) considered the
6 branching order among the major haplogroups to be resolved, our larger CR and *Cytb*
7 dataset fails to provide a robust phylogenetic structure at this level, although there is a
8 suggestion of special affinity between CAS and MUS, and between DOM and NEP.
9 Like both groups of previous researchers, we found the highest mtDNA diversity and
10 depth in CAS populations inhabiting the mountainous region of northwest India and
11 Pakistan, with a loss of haplotype lineage diversity from north to south on the Indian
12 subcontinent (Boursot et al., 1996), and from west to east into Southeast Asia (Prager et
13 al., 1998). Despite this general agreement, Boursot et al (1996) clearly regard
14 *castaneus* in their restricted application of the name to be a long-term resident of
15 Southeast Asia, while Prager et al. (1998) portray this as a relatively recent phase of
16 dispersal of *castaneus* though without specifying any time frame.

17 Low nucleotide diversity in the widely distributed CAS-1 sub-group is evident
18 in this study. This is consistent with that observed by the recent work on the *castaneus*
19 subspecies group done by Rajabi-Maham et al. (2012; see also Bonhomme and Searle,
20 2012). Our results are suggestive of a relatively recent range expansion of CAS-1 to a
21 large geographic areas covering the south and east Indian subcontinent, Southeast Asia,
22 Indonesia, South China, Northeast China and the Russian Far East (Fig. 5). On the
23 other hand, the presence of the locally restricted phyletic group, CAS-1a is suggestive

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

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

17 Although the archaeological record of agriculture is less comprehensive for
18 Asia than for the Middle East and Europe, there is good evidence for domestication of
19 cereal crops including rice and millet by about 9,000 years ago in several parts of
20 South and East Asia (Khush 1997; Londo et al., 2006; Zheng et al., 2009; Molina et al.,
21 2011) and even earlier evidence for long distance overland and maritime trade (Oka
22 and Kusimba, 2008). Assuming that populations experienced a sudden or exponential
23 growth, we calculated τ values from the *Cyt-b* sequences and estimated times since the

1 onset of population expansions. Higher rates of mutation (e.g. 10% or 20% per million
2 years per lineage) rather than lower rates (e.g. 2.5%) are considered to be realistic for
3 assessing rather recent diversifying events (Ho et al., 2005). We obtained a τ value of
4 1.7 for CAS-1b (n=17) which under mutation rates (per million years per lineage) of
5 10% and 20%, gives expansion times of 7,600 and 3,800 years, respectively (Table 3).
6 A τ value was not calculated for CAS-1a but it too is likely to have commenced its
7 dispersal and diversification in China, the Russian Far East and Japan in prehistoric
8 times. In this regard, it is of interest to note archaeological evidence for rice cultivation
9 along the upper Yangtze river (e.g. Yunnan province, here represented by mice from
10 Kunming) at 4500 years ago (Fuller et al., 2010) and recent genetic evidence from an
11 intensive genome survey on wild and cultivated rice, suggesting the Pearl River
12 (Guangxi province, here represented by mice from Guilin) in southern China is the
13 place of the first development of cultivated rice (Bonhomme and Searle, 2012; Huang
14 et al., 2012).

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

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

4

5 **Phylogeography of the MUS lineage**

6 Previous phylogeographic interpretations of the house mouse group do not vary much
7 in regard to the geographic origin of the MUS haplogroup. Boursot et al. (1993: 406)
8 speculated that “the cradle of *M. m. musculus* could be in Transcaucasia or east of the
9 Caspian Sea”, while Prager et al. (1993, 1996) saw the origin of MUS as the product of
10 northward dispersal from a proto-CAS population occupying the region east of the
11 Caspian Sea, followed by range expansion. Both groups of researchers also agree that
12 MUS populations subsequently dispersed west into central Europe and east into China
13 and Japan, and this scenario has been adopted as paradigmatic by Japanese researchers
14 interested in the origin of the indigenous *molossinus* population (Yonekawa et al.,
15 1988; Terashima et al., 2006; Nunome et al., 2010a). Yonekawa et al. (1988) postulated
16 that MUS populations relatively recently expanded into China where CAS populations
17 had already colonized but few other workers have expressed an opinion on the earlier
18 timing of the remarkable eastward expansion of MUS. Nunome et al. (2010a)
19 suggested a latitudinal division within MUS between northern (MUS-I) and southern
20 (MUS-II) groups, based on phylogeographic analyses of nuclear gene sequences, and
21 posited that range expansion of the MUS haplogroup from west to east across
22 continental Eurasia followed separate northern and southern dispersal routes, with
23 separate expansion again into eastern Europe.

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

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

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

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

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

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

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

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

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

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

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

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

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

5

6 **Phylogeography of the DOM lineage**

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

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

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

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

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

17

18 **Concluding remarks**

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

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

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

19

20 **DATA ARCHIVING**

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

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

3

4 **ACKNOWLEDGEMENTS**

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

1 **References**

- 2 Abe K, Noguchi H, Tagawa K, Yuzuriha M, Toyoda A, Kojima T *et al.* (2004).
3 Contribution of Asian mouse subspecies *Mus musculus molossinus* to
4 genomic constitution of strain C57BL/6J as defined by BAC-end
5 sequence-SKIP analysis. *Genom Res* **14**: 2439–2447.
- 6 Aplin KP, Brown PR, Singleton GR, Douangboupha B, Khamphoukheo K (2006)
7 Rodents in the rice environments of Laos. In: Schiller JM, Chanphengxay
8 MB, Linquist B, Appa Rao S (eds). *Rice in Laos*. International Rice Research
9 Institute: Los Banos, pp 291–308.
- 10 Aplin K, Suzuki H, Chinen AA, Chesser RT, ten Have J, Donnellan SC *et al.* (2011).
11 Multiple geographic origins of commensalism and complex dispersal history
12 of black rats. *PLoS ONE* **6**: e26357.
- 13 Auffray J-C, Vanlerberghe F, Britton-Davidian J. (1990). The house mouse
14 progression in Eurasia: a palaeontological and archaeozoological approach.
15 *Biol J Linne Soc* **41**: 13–25.
- 16 Bandelt HJ, Forster P, Rohl A (1999). Median-joining networks for inferring
17 intraspecific phylogenies. *Mol Biol Evol* **16**: 37–48.
- 18 Bastos A, Nair D, Taylor P, Brettschneider H, Kirsten F *et al.* (2011). Genetic
19 monitoring detects an overlooked cryptic species and reveals the diversity
20 and distribution of three invasive *Rattus* congeners in south Africa. *BMC*
21 *Genet* **12**: 26.
- 22 Bryant D, Moulton V (2004). Neighbor-Net: an agglomerative method for the
23 construction of phylogenetic networks. *Mol Biol Evol* **21**: 255–265.

- 1 Bonhomme F, Anand R, Darviche D, Din W, Boursot P (1994). The house mouse as a
2 ring species? In: Moriwaki K, Shiroishi T, Yonekawa H (eds). *Genetics in*
3 *Wild Mice*, Japan Scientific Society Press: Tokyo/S Karger Basel, pp 13–23.
- 4 Bonhomme F, Catalan J, Britton-Davidson J, Chapman VM, Moriwaki K, Nevo E *et al.*
5 (1984). Biochemical diversity and evolution in the genus *Mus*. *Biochem*
6 *Genet* **22**: 275–303.
- 7 Bonhomme F, Miyashita N, Boursot P, Catalan J, Moriwaki K (1989). Genetical
8 variation and polyphyletic origin in Japanese *Mus musculus*. *Heredity* **63**:
9 299–308.
- 10 Bonhomme F, Rivals E, Orth A, Grant GR, Jeffreys AJ, Bois PR (2007). Species-wide
11 distribution of highly polymorphic minisatellite markers suggests past and
12 present genetic exchanges among house mouse subspecies. *Genom Biol* **8**:
13 R80.
- 14 Bonhomme F, Orth A, Cucchi T, Rajabi-Maham H, Catalan J, Boursot P *et al.* (2011).
15 Genetic differentiation of the house mouse around the Mediterranean basin:
16 matrilineal footprints of early and late colonization. *Proc R Soc B* **278**:
17 1034–1043.
- 18 Bonhomme F, Searle JB (2012). House mouse phylogeography. In: Macholán M, Baird
19 SJE, Munclinger P, Piálek J (eds). *Evolution of the house mouse (Cambridge*
20 *series in morphology and molecules)*, Cambridge: Cambridge University
21 Press, pp 278–296.
- 22 Boursot P, Auffray JC, Britton-Davidian J, Bonhomme F (1993). The evolution of
23 house mice. *Ann Rev Ecol Syst* **24**: 119–152.

- 1 Boissinot S, Boursot P (1997). Discordant phylogeographic patterns between the Y
2 chromosome and mitochondrial DNA in the house mouse – selection on the
3 Y chromosome? *Genetics* **146**: 1019–1034.
- 4 Boursot P, Din W, Anand R, Darviche D, Dod B, von Deimling F *et al.* (1996). Origin
5 and radiation of the house mouse: mitochondrial DNA phylogeny. *J Evol*
6 *Biol* **9**: 391–415.
- 7 Chinen AA, Suzuki H, Aplin KP, Tsuchiya K, Suzuki S (2005). Preliminary genetic
8 characterization of two lineages of black rats (*Rattus rattus* sensu lato) in
9 Japan with evidence for introgression at several localities. *Gene Genet Syst*
10 **80**: 367–375.
- 11 Cucchi T, Vigne JD, Auffray JC (2005). First occurrence of the house mouse (*Mus*
12 *musculus domesticus*, Schwarz and Schwarz 1943) in the Western
13 Mediterranean: a zooarchaeological revision of subfossil occurrences. *Biol J*
14 *Linn Soc* **84**: 429–445
- 15 Cucchi T, Vigne JD (2006). Origin and diffusion of the house mouse in the
16 Mediterranean. *Hum Evol* **21**: 95–106.
- 17 Darvish J, Orth A, Bonhomme F (2006). Genetic transition in the house mouse, *Mus*
18 *musculus* of Eastern Iranian Plateau. *Folia Zool* **55**: 349–357.
- 19 Din W, Anand R, Boursot P, Darviche D, Dod B, Jouvin-Marche E *et al.* (1996). Origin
20 and radiation of the house mouse: clues from nuclear genes. *J Evol Biol* **9**:
21 519–539.
- 22 Drummond AJ, Rambaut A (2007). Beast: Bayesian evolutionary analysis by sampling
23 trees. *BMC Evol Biol* **7**: 214.

- 1 Duplantier JM, Orth A, Catalan J, Bonhomme F (2002). Evidence for a mitochondrial
2 lineage originating from the Arabian peninsula in the Madagascar House
3 Mouse. *Heredity* **89**: 154–158.
- 4 Duvaux L, Belkhir K, Boulesteix M, Boursot P (2011). Isolation and gene flow:
5 inferring the speciation history of European house mice. *Mol Ecol* **20**:
6 5248–5264.
- 7 Excoffier L, Lischer HEL (2010). Arlequin suite ver 3.5: a new series of programs
8 to perform population genetics analyses under Linux and Windows. *Mol*
9 *Ecol Resour* **10**: 564–567.
- 10 Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ *et al.* (2007). A
11 sequence-based variation map of 8.27 million SNPs in inbred mouse strains.
12 *Nature* **448**: 1050–1053.
- 13 Frisman LV, Korobitsyna KV, Yakimenko LV, Vorontsov NN (1990). Biochemical
14 Groups of House Mice Inhabiting the Soviet Union, in *Evolyutsionnye*
15 *geneticheskie issledovaniya mlekopitayushchikh: Tezisy dokladov*
16 (Evolutionary Genetic Studies in Mammals: Proc. Conf.), Vladivostok:
17 Dal'nevost. Otd. Akad. Nauk SSSR, part 1, pp 35–54.
- 18 Frisman LV, Korobitsyna KV, Yakimenko LV, Munteanu AI, Moriwaki K (2011).
19 Genetic variability and the origin of house mouse from the territory of Russia
20 and neighboring countries. *Russ J Genet* **47**: 590–602.
- 21 Fuller DQ, Sato YI, Castillo C, Qin L, Weisskopf A, Kingwell-Banham E *et al.* (2010).
22 Consilience of genetics and archaeobotany in the entangled history of rice.
23 *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

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

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

3 25–34.

4 Lack JB, Greene DU, Conroy CJ, Hamilton MJ, Braun JK, Mares MA *et al.* (2012).

5 Invasion facilitates hybridization with introgression in the *Rattus rattus*

6 species complex. *Mol Ecol* **21**: 3545–3561.

7 Librado P, Rozas J (2009). DnaSP v5: a software for comprehensive analysis of DNA

8 polymorphism data. *Bioinformatics* **25**: 1451–1452.

9 Londo JP, Chiang YC, Hung KH, Chiang TY, Schaal BA (2006). Phylogeography of

10 Asian wild rice *Oryza rufipogon* reveals multiple independent domestications of

11 cultivated rice *Oryza sativa*. *Proc Natl Acad Sci USA* **103**: 9578–9583.

12 Markova AK, Simakova AN, Puzachenko AY (2009). Ecosystems of Eastern Europe

13 at the time of maximum cooling of the Valdai glaciation (24–18 kyr BP)

14 inferred from data on plant communities and mammal assemblages. *Quat*

15 *Internat* **201**: 53–59.

16 Markova AK, Puzachenko AY, van Kolfschoten T (2010). The North Eurasian

17 mammal assemblages during the end of MIS 3 (Brianskian–Late

18 Karginian–Denekamp Interstadial). *Quat Internat* **212**: 149–158

19 Marshall JT (1977). A synopsis of Asian species of *Mus* (Rodentia, Muridae). *Bull Am*

20 *Mus Nat Hist* **158**: 173–220.

21 Mezhzherin SV, Kotenkova EV, Mikhailenko AG (1998). The house mice, *Mus*

22 *musculus s.l.*, hybrid zone of Trans- Caucasus. *Zeitschrift für Säugetierkunde*.

23 **63**: 154–168.

- 1 Milishnikov AN, Lavrenchenko LA, Lavrenchenko LA, Orlov VN (1990). High-level
2 introgression of *Mus domesticus* genes in Transcaucasian *Mus musculus* s. str.
3 Populations. *Dokl. Akad. Nauk SSSR* **311**: 764–768.
- 4 Milishnikov AN, Lavrenchenko LA, Lebedev VS (2004). Origin of the house mice
5 (superspecies complex *Mus musculus sensu lato*) from the Transcaucasian
6 region: A new look at dispersal routes and evolution. *Genetika* **40**: 1234–1250.
- 7 Miyashita N, Kawashima T, Wang CH, Jin ML, Wang F, Gotoh H *et al.* (1994). Genetic
8 polymorphisms of *Hbb* haplotypes in wild mice. In: Moriwaki K, Shiroishi T,
9 Yonekawa H (eds) *Genetics in Wild Mice*, Japan Scientific Society Press:
10 Tokyo/S Karger Basal, pp 85–93.
- 11 Molina J, Sikora M, Garud N, Flowers JM, Rubinsteina S, Reynoldsb A *et al.* (2011).
12 Molecular evidence for a single evolutionary origin of domesticated rice.
13 *Proc Natl Acad Sci USA* **108**: 8351–8356.
- 14 Moriwaki K (1994). Wild mouse from geneticist's viewpoint. In: Moriwaki K,
15 Shiroishi T, Yonekawa H (eds) *Genetics in Wild Mice*. Japan Scientific
16 Society Press: Tokyo/S Karger Basal, pp xiii–xxiv.
- 17 Moriwaki K, Yonekawa H, Gotoh O, Minezawa M, Winking H, Gropp A (1984).
18 Implications of the genetic divergence between European wild mice with
19 Robertsonian translocations from the viewpoint of mitochondrial DNA. *Genet*
20 *Res* **43**: 277–287.
- 21 Moriwaki K, Shiroishi T, Yonekawa H (1994). *Genetics in Wild Mice*. It's application to
22 Biomedical Research. Japan Scientific Societies Press: Tokyo/S Karger Basal.

- 1 Munclinger P, Božíkova E, Sugerková M, Pialek J, Macholan M (2002). Genetic
2 variation in house mice (*Mus Muridae Rodentia*) from the Czech and Slovak
3 Republics. *Folia Zool* **51**: 81–92.
- 4 Musser GG, Carleton MD, (2005). Family Muridae. In: Wilson DE, Reeder DM (eds).
5 *Mammal Species of the World*, 3rd edn. The John Hopkins University Press:
6 Baltimore, pp 894–1531.
- 7 Nagamine CM, Shiroishi T, Miyashita N, Tsuchiya K, Ikeda H, Namikawa T *et al.*
8 (1994). Distribution of the Molecularossinus allele of *Sry* the
9 testis-determining gene in wild mouse. *Mol Biol Evol* **11**: 864–874.
- 10 Nunome M, Ishimori C, Aplin KP, Yonekawa H, Moriwaki K, Suzuki H (2010a).
11 Detection of recombinant haplotypes in wild mice (*Mus musculus*) provides
12 new insights into the origin of Japanese mice. *Mol Ecol* **19**: 2474–2489.
- 13 Nunome M, Torii H, Matsuki R, Kinoshita G, Suzuki H (2010b). The influence of
14 Pleistocene refugia on the evolutionary history of the Japanese hare *Lepus*
15 *brachyurus*. *Zool Sci* **27**: 7469–754.
- 16 Nylander JAA (2004). MrModeltest v2. Program distributed by author. Evolutionary
17 Biology Centre, Uppsala University, Uppsala.
- 18 Oka R, Kusimba C (2008). The Archaeology of Trading Systems Part 1: Towards a
19 New Trade Synthesis. *J Archaeol Res* **16**: 339–395.
- 20 Pocock Pocock MJO, Hauffe HC, Searle JB (2005). Dispersal in house mice. *Biol J*
21 *Linn Soc* **84**: 565–583.

- 1 Prager EM, Sage RD, Gyllensten ULF, Thomas WK, Huebner R, Jones CS *et al.*
2 (1993). Mitochondrial DNA sequence diversity and the colonization of
3 Scandinavia by house mice from East Holstein. *Biol J Linn Soc* **50**: 85–122.
- 4 Prager EM, Tichy H, Sage RD (1996). Mitochondrial DNA sequence variation in the
5 eastern house mouse *Mus musculus*: comparison with other house mice and
6 report of a 75-bp tandem repeat. *Genetics* **143**: 427–446.
- 7 Prager EM, Orrego C, Sage RD (1998). Genetic variation and phylogeography of
8 Central Asian and other house mice including a major new mitochondrial
9 lineage in Yemen. *Genetics* **150**: 835–861.
- 10 Rajabi-Maham H, Orth A, Bonhomme F (2008). Phylogeography and postglacial
11 expansion of *Mus musculus domesticus* inferred from mitochondrial DNA
12 coalescent from Iran to Europe. *Mol Ecol* **17**: 627–641.
- 13 Rajabi-Maham H, Orth A, Siahsarvie R, Boursot P, Darvish J, Bonhomme F (2012).
14 The south-eastern house mouse *Mus musculus castaneus* (Rodentia:
15 Muridae) is a polytypic subspecies. *Biol J Linn Soc* **107**: 295–306.
- 16 Rambaut A, Drummond AJ (2009). Tracer v1.5. Available from
17 <http://beast.bio.ed.ac.uk/Tracer>.
- 18 Sage RD, Atchley WR, Capanna E (1993). House mice as models in systematic
19 biology. *Syst Biol* **42**: 523–561.
- 20 Saitou N, Nei M (1987). The neighbor-joining method: a new method for
21 reconstructing phylogenetic trees. *Mol Biol Evol* **4**: 406–425.
- 22 Sakai T, Kikkawa Y, Miura I, Inoue T, Moriwaki K, Shiroishi T *et al.* (2005). Origins of
23 mouse inbred strains deduced from whole-genome scanning by polymorphic

- 1 microsatellite loci. *Mammal Genome* **16**: 11–19.
- 2 Schenekar T, Weiss S (2011). High rate of calculation errors in mismatch distribution
3 analysis results in numerous false inferences of biological importance.
4 *Heredity* **107**: 511–512.
- 5 Searle JB, Jamieson PM, Gündüz İ, Stevens MI, Jones EP, Gemmill CEC et al. (2009a).
6 The diverse origins of New Zealand house mice. *Proc R Soc B* **276**: 209–217
- 7 Searle JB, Jones CS, Gündüz İ, Scascitelli M, Jones EP, Herman JS et al. (2009b). Of
8 mice and (Viking?) men: phylogeography of British and Irish house mice.
9 *Proc R Soc B* **276**: 201–207.
- 10 Shimada T, Aplin KP, Suzuki H (2010). *Mus lepidoides* (Muridae Rodentia) of Central
11 Burma is a distinct species of potentially great evolutionary and
12 biogeographic significance. *Zool Sci* **27**: 449–459.
- 13 Spiridonova LN, Chelomina GN, Moriwaki K, Yonekawa H, Bogdanov AS (2004).
14 Genetic and taxonomic diversity of the house mouse *Mus musculus* from the
15 Asian part of the former Soviet Union. *Russ J Genet* **40**: 1378–1388.
- 16 Spiridonova LN, Kiselev KV, Korobitsyna KV (2011). Discordance in the distribution
17 of markers of different inheritance systems (nDNA mtDNA and
18 chromosomes) in the superspecies complex *Mus musculus* as a result of
19 extensive hybridization in Primorye. *Russ J Genet* **47**: 100–109.
- 20 Stoneking M, Delfin F (2010). The human genetic history of East Asia: weaving a
21 complex tapestry. *Curr Biol* **20**: R188–193.
- 22 Suzuki H, Miyashita N, Moriwaki K, Kominami R, Muramatsu M, Kanehisa T *et al.*
23 (1986). Evolutionary implication of heterogeneity of the nontranscribed spacer

- 1 region of ribosomal DNA repeating units in various subspecies of *Mus*
2 *musculus*. *Mol Biol Evol* **3**: 126–137.
- 3 Suzuki H, Shimada T, Terashima M, Tsuchiya K, Aplin K (2004). Temporal spatial and
4 ecological modes of evolution of Eurasian *Mus* based on mitochondrial and
5 nuclear gene sequences. *Mol Phylogenet Evol* **33**: 626–646.
- 6 Suzuki H, Aplin KP (2012). Phylogeny and biogeography of the genus *Mus* in Eurasia.
7 In: Macholán M, Baird SJE, Munclinger P, Piálek J (eds). *Evolution of the*
8 *house mouse (Cambridge series in morphology and molecules)*, Cambridge:
9 Cambridge University Press, pp 35–64.
- 10 Swofford DL (2001). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other
11 Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
- 12 Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011). MEGA5: molecular
13 evolutionary genetics analysis using maximum likelihood, evolutionary distance,
14 and maximum parsimony methods. *Mol Biol Evol* **28**: 2731–2739.
- 15 Terashima M, Furusawa S, Hanzawa N, Tsuchiya K, Suyanto A, Moriwaki K *et al.* (2006).
16 Phylogeographic origin of Hokkaido house mice (*Mus musculus*) as indicated by
17 genetic markers with maternal paternal and biparental inheritance. *Heredity* **96**:
18 128–138.
- 19 Tsuchiya K, Miyashita N, Wang CH, Wu XL, He XQ, Jin ML *et al.* (1994). Taxonomic
20 study of the genus *Mus* in China, Korea and Japan—Morphologic identification. In:
21 Moriwaki K, Shiroishi T, Yonekawa H (eds). *Genetics in Wild Mice*, Japan
22 Scientific Society Press: Tokyo/S Karger Basal, pp 3–12.
- 23 Tsuda K, Tsuchiya K, Aoki H, Iizuka S, Shimamura H, Suzuki S *et al.* (2007). Risk of

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

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

5 Yasuda SP, Vogel P, Tsuchiya K, Han SH, Lin LK, Suzuki H (2005). Phylogeographic
6 patterning of mtDNA in the widely distributed harvest mouse (*Micromys*
7 *minutus*) suggests dramatic cycles of range contraction and expansion. *Can J Zool*
8 **83**: 1411–1420.

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

12 Yonekawa H, Moriwaki K, Gotoh O, Miyashita N, Matsushima N, Shi LM *et al.* (1988).
13 Hybrid origin of Japanese mice “*Mus musculus molossinus*”: evidence from
14 restriction analysis of mitochondrial DNA. *Mol Biol Evol* **5**: 63–78.

15 Yonekawa H, Moriwaki K, Gotoh O, Hayashi JI, Watanebe J, Miyashita N *et al.* (1981).
16 Evolutionary relationships among five subspecies *Mus musculus* based on
17 restriction enzyme cleavage patterns of mitochondrial DNA. *Genetics* **98**:
18 801–816.

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

- 1 Yonekawa H, Tsuda K, Tsuchiya K, Yakimenko L, Korobitsyna K, Chelomina GN *et al.*
2 (2003). Genetic diversity geographic distribution and evolutionary
3 relationships of *Mus musculus* subspecies based on polymorphisms of
4 mitochondrial DNA. In: Kryukov A, Yakimenko L (eds) *Problems of evolution*,
5 Dalnauka: Vladivostok. Vol 5, pp. 90–108.
- 6 Zheng SH, Zhang ZQ, Liu LP (1997). Pleistocene mammals from fissure-fillings of
7 Sunjiashan hill, Shandong, China. *Vertebrata Palasiatica* **35**: 201–216. (In
8 Chinese with English summary).
- 9 Zheng YF, Sun GP, Qin L, Li C, Wu X, Chen X (2009). Rice fields and modes of rice
10 cultivation between 5000 and 2500 BC in east China. *J Archaeol Sci* **36**:
11 2609–2616.
12

1 **Figure legends**

2

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

20

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

1 *m. domesticus* (DOM) and two rather geographically confined groups of *M. m.*
2 *gentilulus* (GEN) and Nepalese mice (NEP). The portion of the CR network was
3 enlarged to show the details of the branching patterns for CAS-2, in which most of
4 members possess a 75-bp repeat (**d**). The codes for the haplogroups (HGs) in the CR
5 (**g**) and *Cytb* (**h**) network for DOM were taken from those used in Bonhomme et al.
6 (2011).

7

8 **Figure 3.** ML trees for mitochondrial DNA sequences of the cytochrome *b* gene (**a**)
9 and control region (**b**). The PhylML algorithm (Guindon and Gascuel, 2003) was used
10 for the tree reconstruction and bootstrap analysis (100 replications). Bootstrap values
11 (>50%) are shown under basal branches.

12

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

17

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

21

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

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