Supplementary Material

mtDNA Data Indicates a Single Origin for Dogs South of Yangtze River, less than 16,300 Years Ago, from Numerous Wolves

Contents **Page**

The earliest archaeological evidence for dog

There are several difficulties with the interpretation of the archaeological record of early dogs. Firstly, there is great difficulty in discriminating between small wolves and domestic dogs. There are a number of osteological traits used to distinguish domesticated dogs from wolves, e.g. crowding of the teeth caused by shortening of the facial region (Clutton-Brock 1995). However, the traits are not totally exclusive but are sometimes found also among wolves (Musil 2000), and the extent of the variation in the ancient wolf populations is often not well studied. Mostly, only a few parts of the skeleton is preserved, giving only a few traits to study. It also seems probable that the earliest domestic dogs, early in the domestication process, were not fully differentiated from wolf, making things even more difficult. Secondly, the amount of archaeological work, and in particular systematic surveys of animal remains, varies considerably between different parts of the world. The efforts to actively search for the subtle evidence of early dogs are governed by the goal of the archaeological work, and possibly influenced by the interest in dogs in the region in question. Finally, dating of the canid samples, unless made by C-14 dating directly on the dog sample, may be a source of error. There are examples of dog remains given very early dates based on dating by the cultural period of the layer, or by C-14 dating of the layer, that have later been corrected to more recent dates by C-14 dating of the sample itself (Clutton-Brock 1995; Street 2002).

There are especially two canid remains that are often cited as evidence for the existence of domestic dog as early as 14,000 and 13,000-17,000 years ago (ya), that deserve special mentioning. (i) What is often claimed to be the oldest evidence for domestic dog is a mandible and a few other skeletal parts from a canid found in a human grave in Bonn-Oberkassel in Germany (Nobis (1979). The archaeological survey was performed in 1914 at which time the mandible was classified to come from wolf, but in 1979 it was re-classified as dog, and was dated at 14,000 BP based on the cultural context of the grave. However, in 1994 a C-14 dating of both the dog bone and the human remains in the grave showed the grave to be only 12,000 years old (Street 2002). Furthermore, it has been questioned that the remains can be firmly classified as dog based on the limited skeletal

material (Wang and Tedford 2008). (ii) Two canid skulls were found in western Russia, in layers C-14 dated at 13,000-17,000 BP (Sablin and Khlopachev 2002). However, while these canids have some dog like features, they also have several non-dog like features typical for wolf, so the morphological classification as dog does not seem positive (Wang and Tedford 2008).

What now seems to be the earliest reasonably firm evidence for domestic dog is a number of canids from the Natufian culture in today's Israel, including several canids intentionally buried together with humans (Davis and Valla 1978; Dayan 1994; Tchernov and Valla 1997). One buried puppy was C-14 dated at 11,500 BP (Davis and Valla 1978), and the other canids are from cultural layers of approximately the same age. The canids are small and have some dog like traits, which together with the seemingly close connection between canids and humans and the relatively large number of remains indicates that they are domestic dogs. However, even these canids show only some of the osteological traits that distinguish dogs from wolves (Tchernov and Valla 1997), and the possibility that they may be small wolves rather than dogs has been suggested as possible but unlikely (Dayan 1994).

In China, archaeological remains of animals have not been systematically surveyed (Underhill 1997), possibly explaining that that the oldest firm evidence for dog in China is only at least 7,100 years old. Evidence for dogs in North China by 9,700-10,800 ya has been reported (Underhill 1997; Jin and Xu 1992), but no details are given about the morphological details upon which this conclusion was based and the status of the evidence is therefore unclear. The mtDNA evidence presented in this study calls for systematic studies of dog remains in southern East Asia, in search for possible evidence for early domestic dogs.

Analysis of mtDNA from the ancient canid samples, if sufficiently well preserved, may be a method to establish if they are from dog or wolf. Together with C-14 dating of the samples this may potentially give a very precise date for the dog origins and, if applied to samples from across the world, describe the earliest global spread of dogs.

Attempts at dating dog origins based on mtDNA data

In an article by Vilà *et al*. (1997) it was suggested that mtDNA data indicates an origin of dogs from wolf >100,000 ya, much earlier than indicated from archaeological data. This was based on the fact that the largest phylogenetic clade, clade A (called clade I in Vilà *et al*. (1997)), when based on CR data is a dense clade with a great distance to the most recent common ancestor (MRCA). Based on the dense shape it was assumed that the dog haplotypes in clade A all originated from a single wolf founder haplotype, the MRCA. The mean distance of haplotypes to the MRCA is, for the data set in this study, 2.23 substitutions (s.e.m. = 0.0006), which corresponds to a minimum time to the MRCA of 88,800 years. However, as shown by the analysis of whole mtDNA genomes in this study (fig. 2a and b), "dog clade A" more probably derives from several different wolf haplotypes already forming a "wolf clade A". The subclades deriving from the different wolf founder haplotypes have much shorter distance to their respective MRCAs, and consequently lower age, than the entire clade A. We can therefore definitely dismiss the assertion (Vila et al. 1997) that mtDNA data indicates an origin of dogs much earlier than is indicated by the archaeological record, 10.000-15,000 years ago.

The problem with using CR data for dating the origin of dogs is that it is not perfectly suited for studies of the last 11,500 years (the time of origin indicated by archaeological data), since the mutation rate for the CR is at most one substitution per 40,000 years and lineage. Therefore, if "dog clade A" was formed only 11,500 ya, from several different wolf CR haplotypes differing by just a few mutations and already forming "wolf clade A", the founder haplotypes would today remain identical and the haplotypes of today's dogs would still look like a single clade, since there would not have been enough time for mutations to resolve it into well separated subclades. Thus, in the case the dog originated 11,500 ya or less it is not possible to identify the wolf founder haplotypes and therefore not possible to calculate the time of origin, based on the CR data. This is shown in this study, comparing the minimum-spanning networks for the CR and the whole genome mtDNA data (fig. 1b and 2b). What is, for the whole genome mtDNA data, distinctly

resolved subclades differing by at least 30 substitutions from each other (fig. 2b), is for the CR based network unresolved groups differing by a single substitution (fig. 1b).
Since clades B and C have shorter distances to their respective MRCAs (0.36 [s.e.m. = 0.0014] and 0.80 [0.0042] substitutions, respectively) than clade A, the hypothesis that clade A originated from a single wolf haplotype also implied that clades B and C must have originated at a separate time long after clade A, thus indicating multiple origins for the dog (Vilà 1997). However, in this study we show, by simulations of the proportions of clades A, B and C across regions, that it is very unlikely that clade A, B and C would have originated at different times.

Just as in the case of dogs, it has been noted for a number of other domestic animals that estimates of the time of domestication based on mtDNA data have given dates that are much earlier than indicated by the archaeological evidence (Ho and Larson 2006). These studies were all based on analysis of only the CR. In this study, we show that the number of founders for the dog is considerably underestimated when the CR is used for identification of founder haplotypes, but that analysis of mtDNA genomes gives the necessary resolution. Obviously, because of the short time since the origin of the domestic animals, analysis of merely the CR does not give the necessary resolution; the analysis fails to identify all founder haplotypes and therefore places the hypothetical wild founders too far back in time. It is therefore possible that analysis of mtDNA genomes may give more reliable estimates for the time of origin also for other domestic animals.

The origins of clades D, E and F

In contrast to clades A, B and C; clades D, E and F are not found universally, and E and F only at low frequencies (table 1, table S1 and Supplementary Dataset S1). The small number of individuals having these clades makes any conclusions about their origin difficult. However, the fact that the clades are found only regionally suggests that they derive from regional crossbreeding with wolf. Simulation analyses (See below in: “Population genetic simulation analysis of the number of origins in time and space for

clades A, B and C") show that haplotypes deriving from a second (in time) introduction from wolf have great difficulty spreading to other regions with already established populations. Furthermore, had the clades originated from domestication of wolf in a region without dogs, they would initially have had a frequency of 100% in that dog population, and it is not likely that the frequencies would decrease to the observed low values.

However, clade E (a single haplotype found in totally 10 dogs) is found in a relatively large region across East Asia (Southeast Asia, Korea and Japan) and it is therefore possible that it originated at the original domestication of dogs together with clades A, B and C, but did not (like some subclades of clade A) spread out of East Asia. Clade F (three different haplotypes, each found in a single dog) is found in only three dogs sampled in Japan, two dogs of Japanese breeds and one Siberian Husky, but sampled in Japan. Possibly these haplotypes derive from wolf-dog crossbreeding in Japan. Clade D seems to have two separate origins from wolf, since one subclade of sequences (haplotypes D1-D4 and D8) is found only in North Scandinavian dogs and the other subclade (haplotypes D5-D7) in SW Asia, North Africa and Iberia, and the distance to the MRCA is large (fig. 3a, table S3). In the North Scandinavian dogs the frequency is high for the clade D haplotypes (34%). For the other subclade, found in totally six dogs, the frequency is low everywhere (<3%).

Assortment of CR haplotypes into mtDNA genome derived subclades

CR sequences in clade A were assorted into subclades (fig. S1, table 2 and table S2) based on the following criteria:

a) Whole mtDNA genomes were sequenced for individuals representing almost all parts of the control region MS networks to cover the mtDNA diversity (fig. 1c). Special attention was paid to key haplotypes (A29, A44, A105, A97, A84, A141, A8, A92, A65, and A49) which may be central for subclades. The mtDNA genome sequencing resulted in the same cluster pattern as the control region data

except for A43 and A78, which would be assigned to a2 and a3 respectively based on control region data, but were assigned to a1 based on mt-genome sequencing.

b) Individuals not sequenced for the whole mtDNA genome were assigned to subclades a2-a5 if they had a mutation diagnostic for the clade and had no connection to haplotypes belonging to other subclades.

c) Diagnostic mutations were identified in the control region for subclades a2-a5. Subclade a6 did not have diagnostic mutations but the two haplotypes in a6 (A92 and A117) were confirmed by the mtDNA genome sequencing to form an independent subclade.

   For subclade a3, all haplotypes carry a "T" at position 187. Four haplotypes with this mutation: A42, A78, A128, A132, were not assigned to a3. A78 was assigned to subclade a1 based on the mt-genome sequencing. A42 and A132 are at least 2 steps from other haplotypes in a3 and may therefore form a separate subclade, and A128 is separated by a single step to both a3 and a1, and are therefore not assigned to any subclade.

   For subclade a2, all haplotypes carry an "A" in position 209. A58 has the same mutation but belongs clearly to subclade a5, which may be a result of homoplasy. A34, A43, and A49 have the same mutation, but were confirmed to belong to subclade a1 by mtDNA genome sequencing. A52 and A100 have the mutation but are connected to subclade a1 haplotypes, and therefore not assigned to any subclade.

   For subclade a4, all haplotypes carry a "T" at position 364. Outside a4, also A54 has the same mutation but it is several steps from a4.

   For subclade a5, all haplotypes except A141 carry a "T" at position 516. However, A141 was confirmed to belong to this subclade by mtDNA genome sequencing.

The rest of the haplotypes assigned to subclade a1 do not carry any of the diagnostic mutations. It is noticeable that two A9 sequences according to the CR fell into different subclades according to entire mtDNA genome sequences.

CR sequences in clade B were assumed to belong to subclade b1, except for those shown by sequencing of the entire mtDNA genome to belong to subclade b2, and haplotype B26 which is derived from B4 (shown to belong to b2). For CR sequences in clade C, haplotype C3 and haplotypes (except C1) connected by a single substitution to C3 were assumed to belong to subclade c2, and the rest to subclade c1.

Population genetic simulation analysis of the number of origins in time and space for clades A, B and C

We carried out population genetic simulations to test whether multiple origin hypotheses are compatible with the homogeneous distribution of clades A, B, and C across the world. We used a simple stepping-stone model, which consists of three major regions (Continental Europe, Southwest Asia, and China) and two peripheral regions (Britain and Japan) (fig. S5). The three major regions consist of 25 demes, while the two peripheral regions consist of a single deme. Each deme has an effective size of N, and exchanges $N*m$ migrants with the neighbouring demes ($1/2*N*m$ for each) per generation.

The actual dog population exhibits a considerably low level of differentiation in terms of the proportions of the three major clades A, B, and C. We simulated a variety of hypotheses on the origin of the three major clades and calculated the probability of obtaining the observed level of differentiation. If the probability is low under a certain hypothesis, we can conclude that the hypothesis is unlikely.

For a quantitative evaluation, we used a differentiation index (DI) which is equivalent to the F_{ST} defined by Nei (Nei 1977) in a haploid and two-allele case:

Differentiation Index (DI) = [P(world) – P(region)] / P(world)

P(world): probability that two alleles drawn at random from the entire population is different from each other.

P(region): probability that that two alleles drawn from a region is different from each other.

The index quantifies the degree of genetic differentiation between regions. We calculated a single DI value for the world from each simulated data and compared it with the one calculated from the observed mtDNA data. DI varies between 0 (no differentiation, i.e. all the regions show the identical "clade" composition) and 1 (complete differentiation).

A discrete-generation coalescent method (Laval and Excoffier 2004) was used to follow the change in the frequency of a particular mtDNA type during the history since domestication. It allowed multiple coalescent and/or migration events per generation. We started with 1062 samples from the five regions (Britain 108, Europe 151, SW Asia 130, China 555, and Japan 118) and simulated genealogies backwards in time. When we reached the assumed time of domestication T, we allocated an mtDNA type to each ancestor randomly according to the assumed mtDNA proportions (e.g. A=0%, B=50%, C=50%) in the original population. By counting the number of descendants in the modern samples for all the ancestors, we obtained the resulting proportion of the three types in the five regions, and thus DI. For each hypothesis, we iterated simulation runs for 1000 times and calculated the number of runs where (1) every regional sample contains all the three types, and (2) simulated DI $\leq$ observed DI (0.02395). We did not have to consider mutation in the model because we examined the proportion of clades.

Population dynamics (table S4) in each deme basically follow a discrete-type logistic growth

$$N_{t+1} = N_t \exp\left[r(1 - N_t/K)\right]$$

where N_t is the population size at generation t, r is an intrinsic growth rate, and K is a carrying capacity. We assumed that the migration rate m depends on the population size of the destination deme, because the number of migrants to a vacant land is expected to be much larger than that to a fully-populated place. Therefore, the migration rate m to the deme i is

$$m_i = m_0 + m_1\left(1 - N_i/K_i\right)$$

where m_0 is the basic migration rate, m_0+m_1 is the maximum rate when the neighbouring deme is vacant, and N_i and K_i is the population size and carrying capacity of the deme i, respectively.

Hypotheses about the origin of domestic dogs, in terms of time and space, can be categorised into four basic types: Multiple-Sequential (i.e. multiple origins in space and time), Single-Sequential (i.e. multiple origins in time), Multiple-Simultaneous (i.e. multiple origins in space), and Single-Simultaneous (i.e. single origin in space and time) (table S5). For simplicity, when we examine the "multiple origins in space" scenarios, we assumed that clade A originated from one region, while the remaining two clades B and C originated from a different region.

As a basic scenario, we assumed that clade A originated from the central deme in the region China, while B and C originated from the central deme in the region Southwest Asia. First, we tested sequential origin scenarios (Multiple-Sequential): clade A spread all over the world first, and then B and C appeared in Southwest Asia. The results of simulation suggest that this scenario is highly unlikely even when the age of domestication is old or the migration rate is extreme. The same conclusion was obtained for the Single-Sequential scenario. This dismisses the hypothesis that clades B and C appeared in the dog gene pool through dog-wolf hybridization.

Next, we investigated simultaneous multiple origin scenarios where clades B and C appeared in Southwest Asia independently at the same time when clade A appeared in China (Multiple-Simultaneous). Our simulation showed that the observed level of homogenous distribution of the major three clades are hardly seen unless the migration rate m_0 is extremely high (m_0=0.3) (table S5, fig. S6). The migration rate required for sufficient mixing becomes lower if the age of domestication becomes older. However, the default rate (m_0=0.1 or Nm_0=50), which was chosen as a relatively high value for a long-term scenario (see e.g. (Slatkin 1985), the highest Nm estimate was 42 for a species of molluscs and the highest among mammals was 2.2, for mice), is still insufficient even if the domestication date is 10,000 generations ago.

Finally, for the Single-Simultaneous (i.e. single origin in space and time) we see the observed level of homogeneous distribution when we assume that all the clades originated from China simultaneously (table S5). This is the case also if the migration rate m_0 is low (m_0=0.01) although regional differentiation can happen if the age of domestication is large (T=10,000).

We carried out sensitivity analyses with a range of carrying capacities, smaller migration rates to a vacant place (m_0+m_1), or increased numbers of demes. The two Sequential origin hypotheses do not hold under any conditions because it is impossible for the latecomers to spread over the world. The Multiple-Simultaneous hypothesis always requires a very high migration rate AND a very old age of domestication. When we tested a smaller carrying capacity K=2000 in the Single-Simultaneous scenario, we found that regional differentiation can happen more easily due to genetic drift (table S6). Similarly, a series of founder effects caused a considerable regional differentiation when we used smaller migration rates to a vacant place (results are not shown). This may suggest a rapid spread throughout the world from the place where the first domesticated dog population has appeared. The conclusions were unchanged when we doubled the number of demes.

In summary, if clade A, and clades B and C originated from different places, the distribution of the three types is unlikely to be so uniform as observed in the real dog mtDNA sample. If the second domestication happened in the same region as the first one, the haplotypes from the second one would have difficulty in spreading over the world. This implies that all three clades have appeared simultaneously at the same place at the same time. In other words, they originated through one major domestication event.

Some alternative scenarios for the dog origins

Alternative explanations for the dog origins, other than the here suggested single origin from ASY, cannot be excluded but demand much more complicated scenarios. We want to exemplify with a few of the innumerable possibilities. (i) A scenario with two major origin events, for example one in Europe (so that the mtDNA haplotypes for clades A, B and C in Europe, basically the UTs, would have a mainly European origin) and the other in ASY, seems improbable. This would demand that clades A and B both originated from wolf in two different regions (subclades a1and b1 in Europe and subclades a2-a6 and b2 in ASY), drawing very similar haplotypes from two separate wolf populations. Furthermore, the full repertoire of subclades a1, b1 and clade C would have had to "migrate" from Europe to East Asia with very small losses (since almost 100% of CR sequences in Europe are a UTd and therefore closely related to haplotypes in East Asia) and this must have happened in ancient times (since mtDNA genomes are mostly not identical in Europe and East Asia, respectively). (ii) Alternatively, an origin of clade C in Europe and clades A and B elsewhere was suggested in a recent paper (Deguilloux 2009) observing a high frequency of clade C among ancient dog samples from Europe (however, most of the samples giving clade C were from a single locale and therefore possibly from related individuals). In this scenario it is, among other things, hard to explain why clade C would have decreased in Europe from a frequency of 100% to only 9%, while spreading, in ancient times, to all parts of the Old World, everywhere keeping a frequency of around 10% and ending up at 15% in Japan at the far end of the Eurasian continent. (iii) In a third scenario, clade B might have originated in SW Asia (explaining that the highest frequency of clade B, and also the earliest archaeological evidence for dog, is found in this region) while clades A and C originated in East Asia. If so, a high amount of migration during thousands of years must have occurred to adjust the frequencies of clade B from 100% down to 35% in SW Asia, and from 0% to 17% in East Asia. It is hard to conceive why the B haplotypes in SW Asia and the surrounding regions have not been retained in a majority of the dogs; instead the frequency of clade B is unusually low in nearby Africa (11%) and India (7%). Furthermore, subclade b2 probably originated in ASY rather than SW Asia, since almost all haplotypes for b2 is

found only in ASY and surrounding regions. Therefore, in this scenario, the closely related haplotypes in subclades b1 and b2 must have originated independently from different wolf populations.

Altogether, the strongest argument for a single origin in time and space is the universal sharing of haplotypes (with close to 100% of dogs in Europe and SW Asia having CR haplotypes identical or almost identical to those in East Asia), the similar proportions of clades A, B and C among regions, and the gradient of diversity. If the dog would have originated from several independent events of domestication, large proportions of regionally unique and distinct haplotypes in each centre of origin, and a non-even distribution of diversity and of proportions of phylogenetic clades, would be expected.

Additional references

Deguilloux MF, Moquel J, Pemonge MH, Colombeau G. 2009. Ancient DNA supports lineage replacement in European dog gene pool: insight into Neolithic southeast France. J Arch Sci. 36:513-519.

Laval G, Excoffier L. 2004. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics. 20:2485-2487.

Nei M. 1977. F-statistics and analysis of gene diversity in subdivided populations. Ann Hum Genet. 41:225-233.

Nobis G. 1979. Der älteste Haushund lebte vor 14,000 Jahren. Umschau. 79:610.

Sablin MV, Khlopachev GA. 2002. The earliest ice age dogs: evidence from Eliseevichi 1. Current Anthropology. 43:795-798.

Slatkin M. 1985. Rare alleles as indicators of gene flow. Evolution. 39:53-65.

Street M. 2002. Ein Wiedersehen mit dem Hund von Bonn-Oberkassel. Bonn Zool Beitr. 50:269-290. (In German)

Table S1. Genetic diversity for CR data across the World (continued below)

Population	N ABC(DEF)	nA(%)	nB(%)	nC(%)	HT (ABC)	HT (A)	HT (B)	HT (C)	HTres (37)	HTres (56)
East Asia	730(5)	549(75.2)	124(16.9)	57(7.8)	121	97	16	8	22.0 ± 2.5	28.5 ± 3.0
West	558(27)	384(68.8)	123(22.0)	51(9.1)	68	53	11	4	18.1 ± 2.3	23.0 ± 2.5
Europe	313(23)	217(69.3)	68(21.7)	28(8.9)	39	29	6	4	16.2 ± 2.0	19.5 ± 2.0
SW Asia	130(3)	72(55.4)	45(34.6)	13(10.0)	30	20	7	3	15.3 ± 2.1	19.5 ± 2.1
Africa	56(1)	48(85.7)	6(10.7)	2(3.6)	22	19	2	1	17.8 ± 1.4	22.0 ± 0.0
India	59(0)	47(79.7)	4(6.8)	8(13.6)	22	18	1	3	17.6 ± 1.5	21.5 ± 0.6
Siberia	60(2)	39(65.0)	13(21.7)	8(13.3)	20	15	3	2	17.3 ± 1.2	19.7 ± 0.5
Japan	118(3)	76(64.4)	24(20.3)	18(15.3)	25	17	5	3	16.7 ± 1.7	19.7 ± 1.7
Korea	90(7)	80(88.9)	6(6.7)	4(4.4)	24	17	3	4	15.1 ± 1.7	19.1 ± 1.6
N. China	98(0)	65(66.3)	25(25.5)	8(8.2)	24	16	4	4	15.3 ± 1.7	18.8 ± 1.6
C. China	141(0)	109(77.3)	21(14.9)	11(7.8)	27	22	2	3	14.4 ± 1.8	17.4 ± 1.9
S. China	281(0)	223(79.4)	44(15.7)	14(4.9)	71	59	8	4	21.2 ± 2.3	27.7 ± 2.8
SE Asia	57(2)	50(87.7)	3(5.3)	4(7.0)	30	25	2	3	22.6 ± 1.7	29.7 ± 0.5
ASY	338(2)	273(80.8)	47(13.9)	18(5.3)	87	73	9	5	22.6 ± 2.4	30.1 ± 3.0
Britain	108(0)	82(75.9)	22(20.4)	4(3.7)	24	19	3	2	15.3 ± 1.7	18.7 ± 1.6
Scandinavia	38(20)	22(57.9)	12(31.6)	4(10.5)	9	7	1	1	8.9 ± 0.2	NA
Eur. N Cont	91(0)	57(62.6)	21(23.1)	13(14.3)	14	8	3	3	11.7 ± 1.0	12.9 ± 0.8
Eur. S Cont	57(2)	42(73.7)	10(17.5)	5(8.8)	18	12	3	3	14.2 ± 1.5	17.9 ± 0.3
Eur. Cont	151(2)	101(66.9)	31(20.5)	19(12.6)	25	16	5	4	14.9 ± 1.7	17.4 ± 1.8
East	930(14)	710(76.3)	148(15.9)	72(7.7)	140	112	18	10	22.1 ± 2.5	29.2 ± 3.1
Arctic Amer.	33(0)	33(100.0)	0(0.0)	0(0.0)	8	8	0	0	NA	NA
China	555(0)	423(76.2)	97(17.5)	35(6.3)	97	78	13	6	20.7 ± 2.4	27.0 ± 2.7
China + SEA	612(2)	473(77.3)	100(16.3)	39(6.4)	110	89	14	7	21.7 ± 2.6	28.4 ± 2.9
Helongjiang	52(0)	34(65.4)	14(26.9)	4(7.7)	13	9	2	2	11.4 ± 1.0	NA
Liaoning	6(0)	3(50.0)	3(50.0)	0(0.0)	5	3	2	0	NA	NA
Hebei	17(0)	15(88.2)	2(11.8)	0(0.0)	8	6	2	0	NA	NA
Shanxi	23(0)	13(56.5)	6(26.1)	4(17.4)	12	7	1	4	NA	NA
LHS	46(0)	31(67.4)	11(23.9)	4(8.7)	19	12	3	4	16.7 ± 1.2	NA
Shaanxi	91(0)	60(65.9)	21(23.1)	10(10.9)	20	15	2	3	12.5 ± 1.6	15.3 ± 1.5
Shichuan	48(0)	47(97.9)	0(0.0)	1(2.1)	10	9	0	1	9.2 ± 0.8	NA
Tibet, Q, N	37(0)	28(75.7)	7(18.9)	2(5.4)	15	12	2	1	15.0 ± 0.0	NA
Guangdong	14(0)	11(78.6)	1(7.14)	2(14.3)	9	7	1	1	NA	NA
Guangxi	35(0)	32(91.4)	3(8.6)	0(0.0)	18	16	2	0	NA	NA
GdGx	49(0)	43(87.8)	4(8.2)	2(4.1)	24	20	3	1	20.0 ± 1.4	NA
Hunan	54(0)	36(66.7)	13(24.1)	5(9.3)	16	12	1	3	13.1 ± 1.3	NA
Guizhou	57(0)	44(77.2)	10(17.5)	3(5.3)	28	24	2	2	21.6 ± 1.6	27.7 ± 0.5
Jiangxi	46(0)	32(69.6)	13(28.3)	1(2.2)	23	17	5	1	19.9 ± 1.2	NA
Yunnan	75(0)	68(90.7)	4(5.3)	3(4.0)	29	23	3	3	19.1 ± 1.9	24.8 ± 1.5
Hainansanya	31(0)	27(87.1)	2(6.5)	2(6.5)	13	11	1	1	NA	NA
ISEA	12(0)	11(91.7)	1(8.3)	0(0)	7	6	1	0	NA	NA

East Asia – China (N/C/S China, Tibet, Quinghai), Southeast (SE) Asia, Japan; West - Europe, SW Asia, Africa, India; N/C/S China – China north of Yellow River/between Yellow and Yangtze/south of Yangtze River; ASY – S China, SE Asia; East – East Asia, Siberia, Korea, Mongolia; LHS – Liaoning, Hebei, and Shanxi; Tibet, Q, N – Tibet, Qinghai, and Nepal; GdGx – Guangdong and Guangxi; ISEA – Islands of South East Asia. HT – the number of haplotypes; HTres – the number of haplotypes obtained from resampling of size 37 and 56 (500 replications) to adjust for different sample size; PropUT/UTd – the proportion of individuals carrying a UT and UTd.

Table S1. (continued)

Population	PropUT (A)	PropUTdUT (A)	PropUT (B)	PropUTdUT (B)	PropUT (C)	PropUTdUT (C)	PropUT (totalABC)	PropUTdUT (totalABC)	Gene Diversity
East Asia	46.1	60.3	75.8	94.4	86.0	100.0	54.2	69.2	0.9528
West	69.5	94.0	88.6	98.4	96.1	100.0	76.2	95.5	0.9216
Europe	77.9	98.6	88.2	98.5	92.9	100.0	81.5	98.7	0.9174
SW Asia	75.0	91.7	86.7	97.8	100.0	100.0	81.5	94.6	0.8707
Africa	50.0	89.6	100.0	100.0	100.0	100.0	57.1	91.1	0.9292
India	42.6	76.6	100.0	100.0	100.0	100.0	54.2	81.4	0.9184
Siberia	33.3	61.5	92.3	100.0	37.5	100.0	46.7	75.0	0.9350
Japan	55.3	76.3	41.7	100.0	94.4	100.0	58.5	84.7	0.9342
Korea	62.5	75.0	66.7	100.0	75.0	100.0	63.3	77.8	0.8214
N. China	73.8	84.6	92.0	100.0	87.5	100.0	79.6	89.8	0.8992
C. China	65.1	81.7	100.0	100.0	72.7	100.0	70.9	85.8	0.9003
S. China	32.7	43.9	75.0	86.4	85.7	100.0	42.0	53.4	0.9486
SE Asia	30.0	44.0	66.7	100.0	75.0	100.0	35.1	50.9	0.9526
ASY	32.2	44.0	74.5	87.2	83.3	100.0	40.8	53.0	0.9572
Britain	68.3	98.8	86.4	100.0	100.0	100.0	73.1	99.1	0.9125
Scandinavia	81.8	100.0	100.0	100.0	100.0	100.0	89.5	100.0	0.8253
Eur. N Cont	98.2	100.0	90.5	100.0	84.6	100.0	94.5	100.0	0.9061
Eur. S Cont	71.4	97.6	80.0	90.0	100.0	100.0	75.4	96.5	0.8694
Eur. Cont	87.1	99.0	87.1	96.8	89.5	100.0	87.4	98.7	0.9167
East	46.3	61.0	77	95.3	80.6	100.0	53.9	69.5	0.9525
Arctic Amer.	18.2	21.2	NA	NA	NA	NA	18.2	21.2	0.6751
China	46.3	59.3	84.5	92.8	82.9	100.0	55.3	67.7	0.9452
China + SEA	44.6	57.7	84.0	93.0	82.1	100.0	53.4	66.2	0.9502
Helongjiang	85.3	85.3	100.0	100.0	100.0	100.0	90.4	90.4	0.8648
Liaoning	66.7	100.0	66.7	100.0	NA	NA	66.7	100.0	0.8485
Hebei	60.0	93.3	50.0	100.0	NA	NA	58.8	94.1	0.8342
Shanxi	61.5	69.2	100.0	100.0	75.0	100.0	73.9	82.6	0.8696
LHS	61.3	83.9	81.8	100.0	75.0	100.0	67.4	89.1	0.9126
Shaanxi	85.0	90.0	100.0	100.0	70.0	100.0	86.8	93.4	0.8667
Shichuan	38.3	70.2	NA	NA	100.0	100.0	39.6	70.8	0.8088
Tibet, Q, N	17.9	35.7	85.7	100.0	100.0	100.0	35.1	51.4	0.9197
Guangdong	18.2	54.5	0.0	100.0	100.0	100.0	28.6	64.3	0.9101
Guangxi	9.4	12.5	66.7	100.0	NA	NA	14.3	20.0	0.8646
GdGx	11.6	23.3	50.0	100.0	100.0	100.0	18.4	32.7	0.9072
Hunan	19.4	22.2	100.0	100.0	80.0	100.0	44.4	48.1	0.8411
Guizhou	34.1	54.5	90.0	100.0	100.0	100.0	47.4	64.9	0.9446
Jiangxi	25.0	34.4	61.5	76.9	100.0	100.0	37.0	47.8	0.9240
Yunnan	55.9	66.2	25.0	25.0	66.7	100.0	54.7	65.3	0.8805
Hainansanya	33.3	48.2	100.0	100.0	100.0	100.0	41.9	54.8	0.9180
ISEA	18.2	27.3	100.0	100.0	NA	NA	25.0	33.3	0.8551

Table S2. Geographical representation of subclades of clades A, B and C

Region (No. of individuals)	Number of subclades	a1	a2	a3	a4	a5	a6	b1	b2	c1	c2	Haplotypes not assigned to subclade
Britain (108)	**4**	(82)	-	-	-	-	-	(22)	-	(2)	(2)	
Europe Cont. (151)	**4**	(99)	-	-	-	-	-	(31)	-	(12)	(7)	A40 (2)
Eur. N. Cont. (91)	**4**	(57)	-	-	-	-	-	(21)	-	(8)	(5)	
Eur. S. Cont. (57)	**4**	(40)	-	-	-	-	-	(10)	-	(3)	(2)	A40 (2)
SW Asia (130)	**5**	(71)	-	-	-	-	-	(41)	(4)	(9)	(4)	A9 (1)
Africa (56)	**4**	(46)	(1)	-	-	-	-	(6)	-	-	(2)	A153 (1)
India (59)	**5**	(45)	(1)	-	-	-	-	(4)	-	(2)	(6)	A153 (1)
Siberia (60)	**4**	(36)	(3)	-	-	-	-	(13)	-	(8)	-	
Scandinavia (38)	**3**	(22)	-	-	-	-	-	(12)	-	-	(4)	
Japan (118)	**7**	(61)	(13)	-	(2)	-	-	(20)	(4)	(4)	(14)	
Korea (90)	**7**	(62)	(11)	-	(1)	-	-	(5)	(1)	(2)	(2)	A153 (2) A161 (4) A162 (1)
N. China (98)	**5**	(55)	(10)	-	-	-	-	(25)	-	(7)	(1)	
C. China (141)	**7**	(89)	(18)	-	(1)	(1)	-	(21)	-	(10)	(1)	
S. China (281)	**9**	(86)	(99)	(14)	-	(18)	(3)	(35)	(9)	(8)	(6)	A9 (3)
SE Asia (57)	**9**	(20)	(13)	(3)	(1)	(7)	-	(2)	(1)	(2)	(2)	A42 (1)/ A128 (1)/ A132 (3)/ A9 (1)
ASY (338)	**10**	(106)	(112)	(17)	(1)	(25)	(3)	(37)	(10)	(10)	(8)	A42 (1)/ A128 (1)/ A132 (3)/ A9 (4)
Subregions N. China												
Heilanongjiang (52)	**4**	(29)	(5)	-	-	-	-	(14)	-	(4)	-	
Hebei (17)	**3**	(14)	(1)	-	-	-	-	(2)	-	-	-	
Liaoning (6)	**2**	(3)	-	-	-	-	-	(3)	-	-	-	
Shanxi (23)	**5**	(9)	(4)	-	-	-	-	(6)	-	(3)	(1)	
Subregions C. China												
Anhui (2)	**1**	(2)	-	-	-	-	-	-	-	-	-	
Shaanxi (91)	**6**	(54)	(5)	-	(1)	-	-	(21)	-	(9)	(1)	
Sichuan (48)	**4**	(33)	(13)	-	-	(1)	-	-	-	(1)	-	
Subregions S. China												
Guangdong (14)	**6**	(4)	(4)	(1)	-	(2)	-	-	(1)	-	(2)	
Guangxi (35)	**5**	(3)	(22)	(3)	-	(4)	-	(3)	-	-	-	
Guizhou (57)	**8**	(21)	(16)	(4)	-	(2)	-	(9)	(1)	(1)	(2)	A9 (1)
Hunan (54)	**6**	(10)	(25)	(1)	-	-	-	(13)	-	(4)	(1)	
Jiangxi (46)	**6**	(10)	(19)	-	-	(3)	-	(9)	(4)	(1)	-	
Yunnan (75)	**9**	(38)	(13)	(5)	-	(7)	(3)	(1)	(3)	(2)	(1)	A9 (2)
West China												
Tibet (26)	**4**	(10)	(7)	-	(2)	-	-	(5)	-	-	-	A52 (2)
Qinghai (8)	**3**	-	(5)	-	-	-	-	(1)	-	-	(2)	
Nepal (3)	**3**	(1)	(1)	-	-	-	-	(1)	-	-	-	

Note: number of individuals given within parentheses for each subclade, and for non-assigned haplotypes.

Table S3. Divergence estimates for clades and subclades in the mitochondrial genome tree (Figure 2a), for different clock and population models

clade	tree dist	cc-cs	cc-es	cc-sky	lc-sky	ec-sky
A	1.41E-03	1.62E-03	1.63E-03	1.56E-03	1.59E-03	1.75E-03
a1	7.95E-04	9.72E-04	9.67E-04	9.24E-04	9.45E-04	1.14E-03
a2	7.49E-04	7.95E-04	7.95E-04	7.97E-04	8.20E-04	7.91E-04
a3	3.71E-04	4.03E-04	4.06E-04	3.87E-04	3.95E-04	5.27E-04
a4	2.95E-04	2.65E-04	2.67E-04	2.40E-04	2.75E-04	3.63E-04
a5	8.07E-04	6.42E-04	6.30E-04	6.55E-04	6.94E-04	6.88E-04
a6	2.22E-04	2.51E-04	2.51E-04	2.24E-04	2.16E-04	2.79E-04
B	8.50E-04	6.98E-04	7.00E-04	7.00E-04	7.03E-04	7.64E-04
b1	3.74E-04	3.10E-04	3.08E-04	3.11E-04	3.21E-04	4.74E-04
b2	5.38E-04	5.63E-04	5.63E-04	5.87E-04	5.87E-04	5.70E-04
C	6.28E-04	7.11E-04	7.14E-04	6.92E-04	7.14E-04	1.05E-03
c1	2.11E-04	2.68E-04	2.70E-04	2.48E-04	2.56E-04	4.76E-04
c2	1.68E-04	1.90E-04	1.92E-04	1.94E-04	1.96E-04	3.41E-04
D		1.16E-03	1.17E-03	1.15E-03	1.20E-03	9.64E-04
coy		1.88E-03	1.88E-03	1.84E-03	1.84E-03	1.84E-03
R		0.96	0.96	0.97	0.97	0.92
p		0.39	0.38	0.43	0.38	0.16

Note: *tree dist* is an estimate based on an unconstrained clock ML tree; *cc-cs* is a BMCMC estimate with constant clock rate and constant population size; *cc-es* constant clock rate and exponential population size; *cc-sky* constant clock rate and a population size estimated from a skyline plot; *lc-sky* lognormal relaxed clock rate and a population size estimated from a skyline plot; and *ec-sky* relaxed exponential clock rate and a population size estimated from a skyline plot. *R* is the correlation coefficient of the results from one BMCMC estimate to the *tree dist* estimate, and *p* is estimated using Welch's t test. Note that, as mentioned in the main text, *ec-sky* had a low estimated sample size (ESS) for the tree height; we show the mean values for completeness and to demonstrate the similarity among all clock estimates.

Table S4. List of parameters for the population dynamics

parameter	default value	Remark
r	0.405	$\exp(r) = 1.5$
K	5000	carrying capacity of each deme
m_0	0.1	migration rate
m_0+m_1	0.5	maximum rate when neighbouring demes are empty
T	5000	age of domestication (in generations)

Table S5. Simulation results for typical hypotheses of the dog domestication.

	Age [a]	Migration rate (m_0)						
	T	0.01	0.05	0.1	0.15	0.2	0.25	0.3
Multiple-Sequential [b]	5000	0.000	0.000	0.001	0.002	0.006	0.009	0.005
	6000	0.000	0.000	0.000	0.002	0.002	0.002	0.003
	7000	0.000	0.000	0.000	0.002	0.002	0.011	0.003
	8000	0.000	0.000	0.001	0.001	0.005	0.003	0.005
	9000	0.000	0.000	0.002	0.003	0.009	0.008	0.008
	10000	0.000	0.001	0.004	0.005	0.005	0.003	0.008
Single-Sequential [c]	5000	0.000	0.000	0.000	0.000	0.000	0.001	0.002
	10000	0.000	0.000	0.001	0.002	0.008	0.007	0.009
Multiple-Simultaneous [d]	5000	0.000	0.000	0.000	0.000	0.001	0.013	**0.097**
	6000	0.000	0.000	0.000	0.000	0.027	**0.118**	**0.270**
	7000	0.000	0.000	0.000	0.008	**0.077**	**0.223**	**0.494**
	8000	0.000	0.000	0.001	0.022	**0.164**	**0.375**	**0.628**
	9000	0.000	0.000	0.005	**0.064**	**0.249**	**0.500**	**0.711**
	10000	0.000	0.000	0.011	**0.118**	**0.371**	**0.610**	**0.764**
Single-Simultaneous [e]	5000	**0.068**	**0.327**	**0.490**	**0.663**	**0.727**	**0.798**	**0.849**
	10000	0.044	**0.250**	**0.467**	**0.607**	**0.764**	**0.790**	**0.861**
Single-Simultaneous [f]	5000	**0.052**	**0.294**	**0.460**	**0.573**	**0.693**	**0.742**	**0.823**
	10000	0.019	**0.181**	**0.405**	**0.563**	**0.650**	**0.697**	**0.777**

Note: Probabilities that the distribution of the three major types was as homogeneous as observed in the real dog mtDNA CR sample are shown. The default parameter values were used unless mentioned. For details, see text.

[a] The age (generations) of the domestication, the second domestication in the sequential cases.

[b] Type A spread over the world first, after which 99% of the central deme of Southwest Asia was replaced by a new population (B: 50%, C: 50%).

[c] Type A spread over the world first, after which 99% of the central deme of China was replaced by a new population (B: 50%, C: 50%).

[d] One population (A: 100%) appeared at the central deme of China, while the other population (B: 50%, C: 50%) appeared at the central deme of Southwest Asia.

[e] One population (A: 33.3%, B: 33.3%, C: 33.3%) appeared at the central deme of China.

[f] One population (A: 70%, B: 20%, C: 10%) appeared at the central deme of China.

Table S6. Sensitivity analysis in relation to the deme size.

Note: Probabilities that the distribution of the three major clades was as

	Age T	Carrying capacity (K)	Migration rate (m_0)						
			0.01	0.05	0.1	0.15	0.2	0.25	0.3
Multiple-Simultaneous	5000	2000	0.000	0.000	0.000	0.001	0.009	0.038	**0.103**
	5000	5000	0.000	0.000	0.000	0.000	0.001	0.013	**0.097**
	5000	10000	0.000	0.000	0.000	0.000	0.000	0.008	**0.062**
	10000	2000	0.000	0.000	0.018	**0.081**	**0.215**	**0.361**	**0.500**
	10000	5000	0.000	0.000	0.011	**0.118**	**0.371**	**0.610**	**0.764**
	10000	10000	0.000	0.000	0.001	**0.115**	**0.445**	**0.754**	**0.912**
Single-Simultaneous[a]	5000	2000	0.002	**0.061**	**0.174**	**0.254**	**0.348**	**0.456**	**0.535**
	5000	5000	**0.068**	**0.327**	**0.490**	**0.663**	**0.727**	**0.798**	**0.849**
	5000	10000	**0.280**	**0.655**	**0.803**	**0.873**	**0.932**	**0.941**	**0.972**
	10000	2000	0.001	0.046	**0.120**	**0.253**	**0.339**	**0.459**	**0.510**
	10000	5000	0.044	**0.250**	**0.467**	**0.607**	**0.764**	**0.790**	**0.861**
	10000	10000	**0.159**	**0.556**	**0.744**	**0.856**	**0.914**	**0.941**	**0.962**

homogeneous as observed in the real dog mtDNA sample are shown. See Table S5 for further explanations.

[a] One population (A: 33.3%, B: 33.3%, C: 33.3%) appeared at the central deme of China.

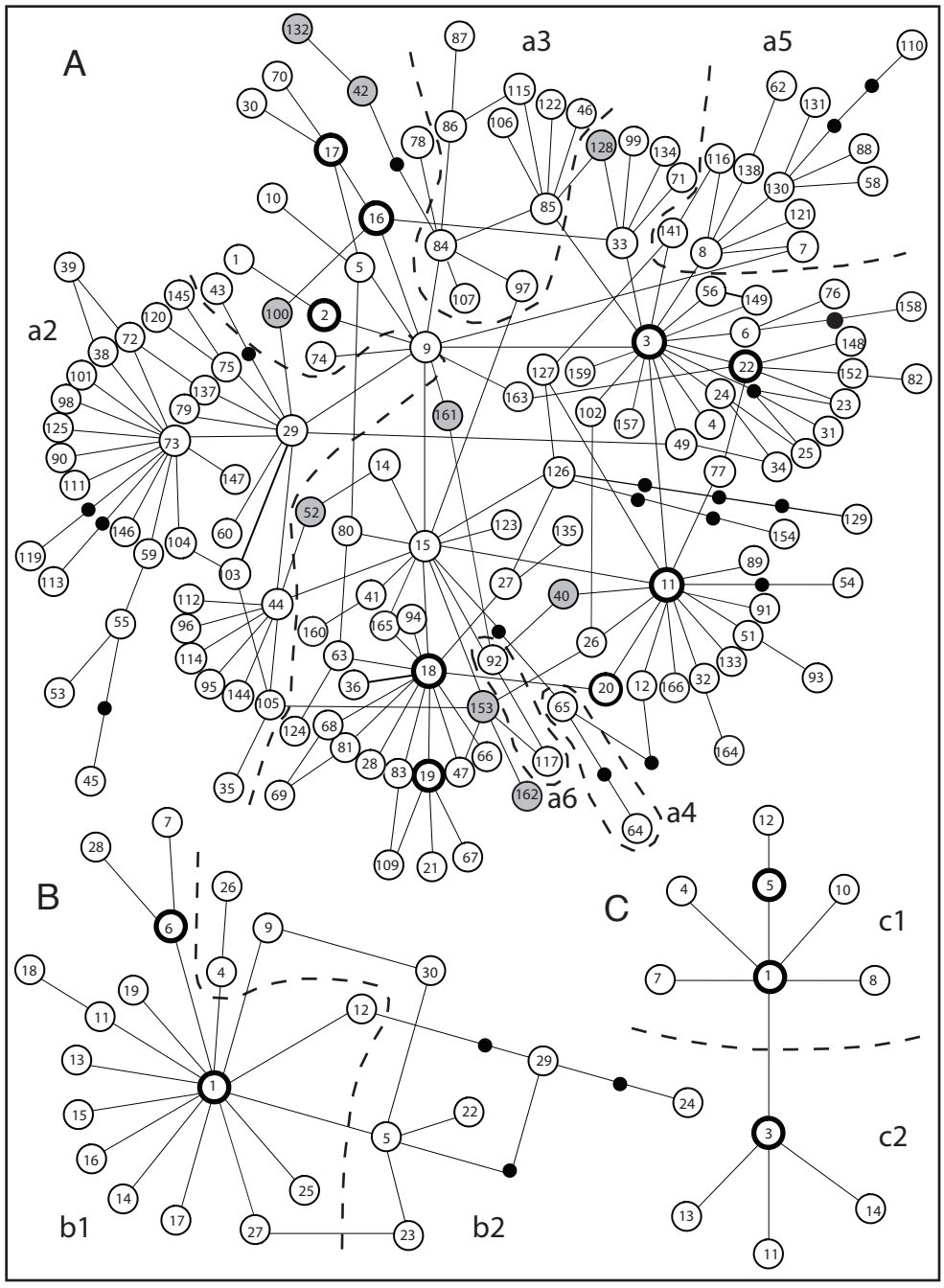

Fig. S1. Minimum spanning networks for dog clades A, B and C, giving the names of the haplotypes. The haplotypes (circles, UTs with bold lining) and empty nodes (solid dots) are separated by one substitutional step (indels are ignored). Subclades a2-a6, b1, b2, c1 and c2 are indicated within dashed lines. Haplotypes not assigned to any subclade are shaded.

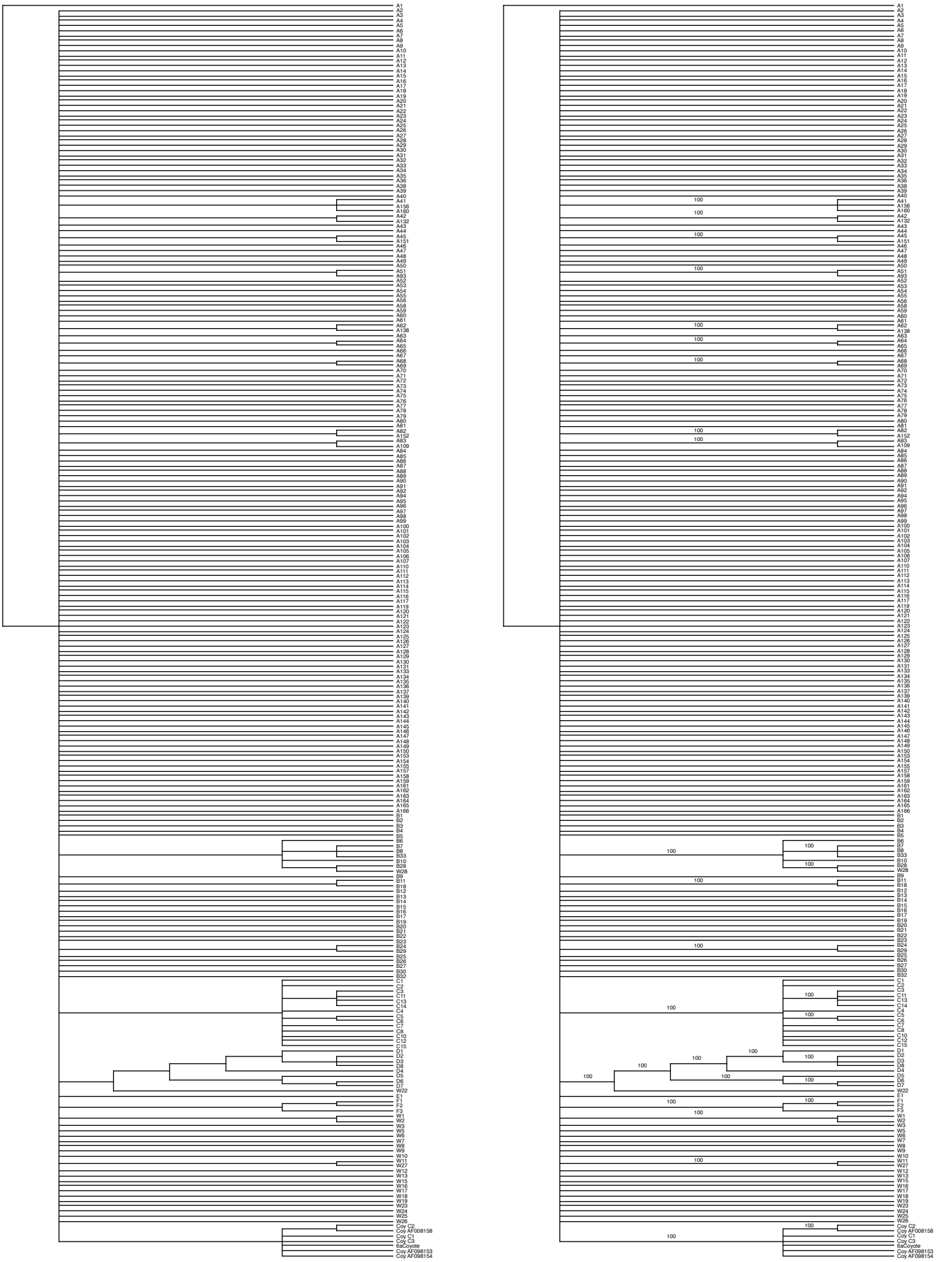

Fig. S2. NJ bootstrap results (in percent) for the CR data, based on 1000 replicates

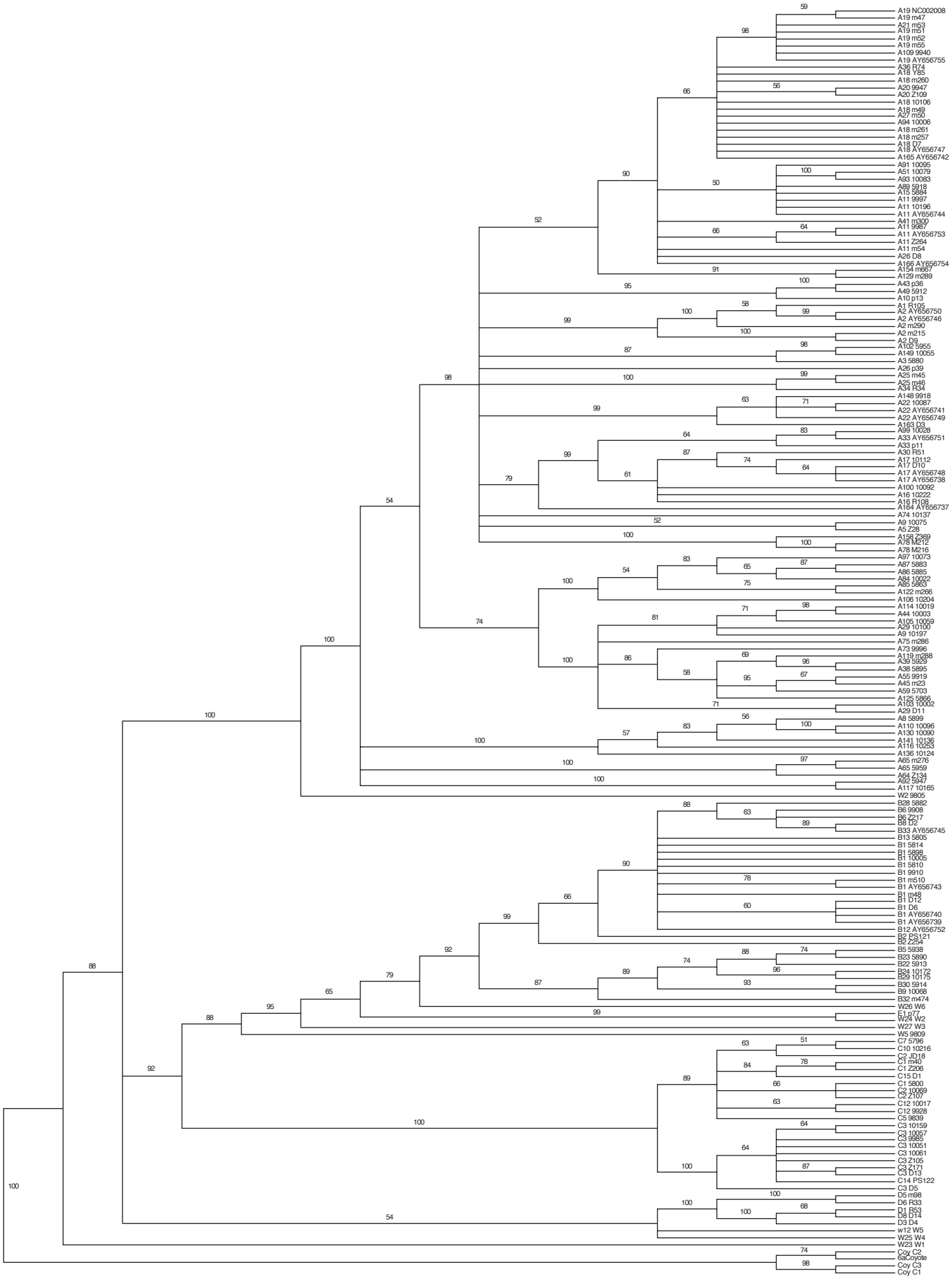

Fig. S3. NJ bootstrap results (in percent) for the whole genome data, based on 1000 replicates.

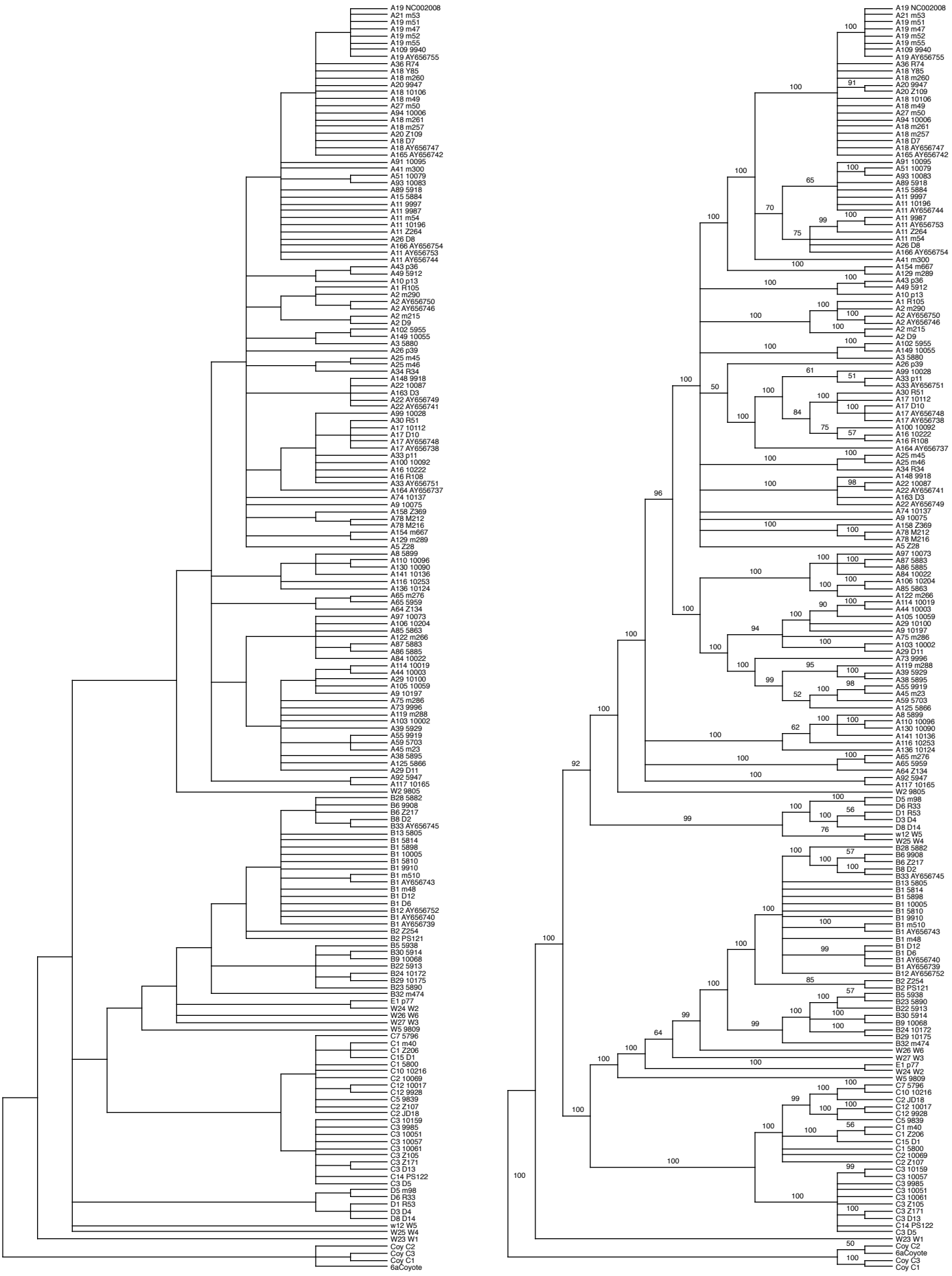

Fig. S4. Bayesian support value results (in percent) for the whole genome data, based on 10000 trees and a constant clock model.

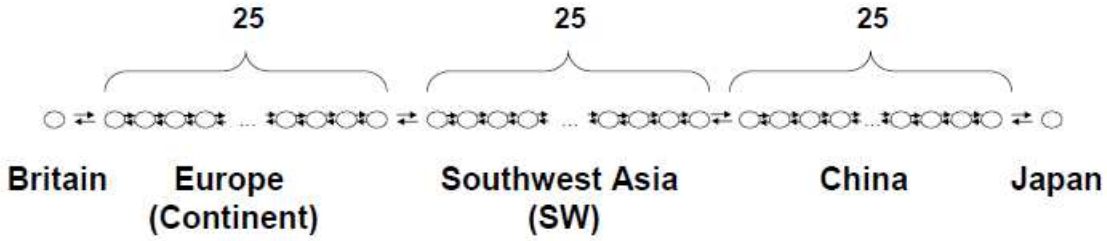

Fig. S5. Stepping-stone model used in the simulation. The three major regions contain 25 demes, while the two minor regions consist of just one deme. Each deme exchanges *N*m* migrants with the neighbouring demes for each generation. The edge demes (i.e. Britain and Japan) exchange only 1/2**N***m* migrants with their neighbours.

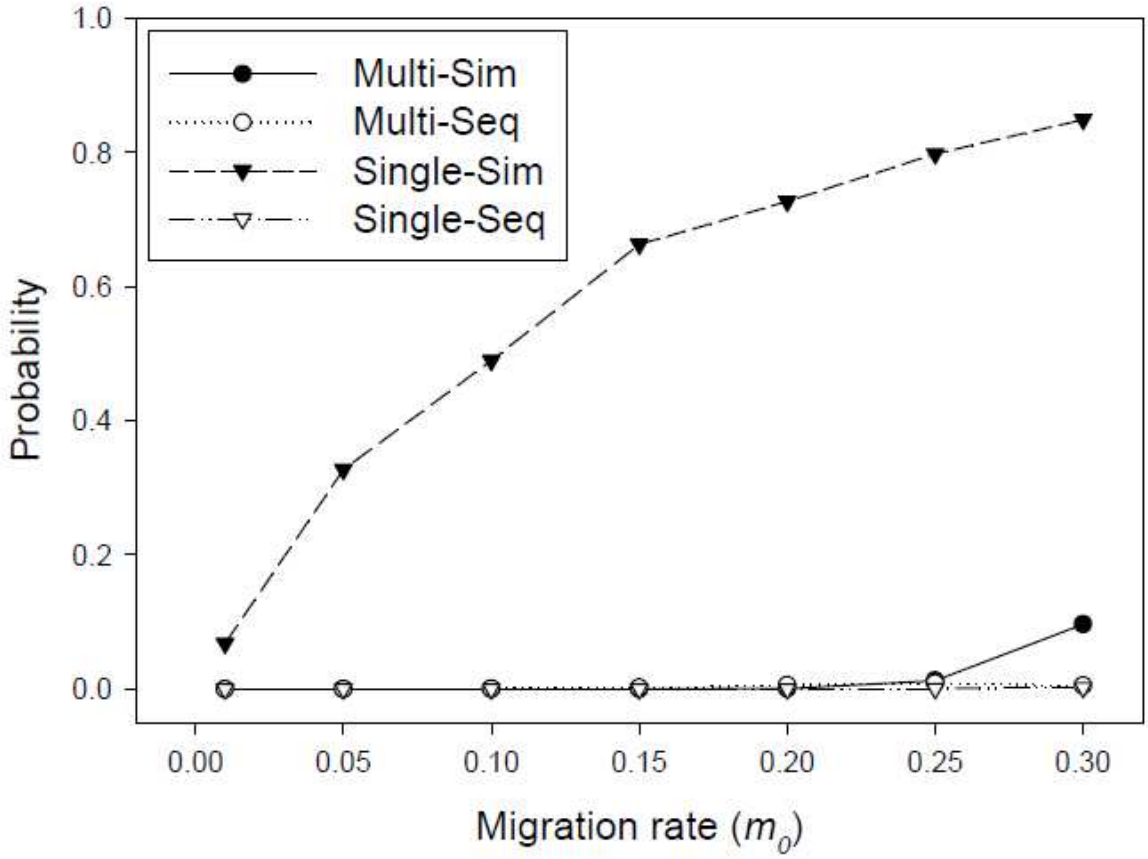

Fig. S6. Probability that the distribution of the three clades was as homogeneous as observed in the real dog mtDNA sample. *K*=5000, *T*=5000. Multi-Sim: Multiple-Simultaneous, Multi-Seq: Multiple-Sequential, Single-Sim: Single-Simultaneous, Single-Seq: Single-Sequential. For details, see text.