LA-UR- 09-00060

Approved for public release; distribution is unlimited.

Title: The dog originated south of Yangtze River less than 16,000 years ago, from numerous wolves

Author(s): Jun-Feng Pang, State Key Laboratory of Genetic Resources and Evolution
Cornelya Kluetsch, KTH-Royal Institute of Technology
Thomas Leitner, Los Alamos National Laboratory
ET All

Intended for: Current Biology

Los Alamos
NATIONAL LABORATORY
EST. 1943

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Form 836 (7/06)

The dog originated south of Yangtze River less than 16,000 years ago, from numerous wolves

Jun-Feng Pang[1,2], Cornelya Kluetsch[3], Xiao-Ju Zou[2], Ai-bing Zhang[3], Li-Yang Luo[1,4], Helen Angleby[3], Arman Ardalan[3,5,6], Camilla Ekström[3], Anna Sköllermo[3], Joakim Lundeberg[3], Shuichi Matsumura[7,8], Thomas Leitner[9], Ya-Ping Zhang[1,2], Peter Savolainen[3]

[1]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China

[2]Laboratory for Conservation and Utilization of Bio-resource, Yunnan University, Kunming 650091, China

[3]KTH-Royal Institute of Technology, Gene Technology, Roslagstullsbacken 21, 10691 Stockholm, Sweden

[4]College of Wildlife Resource, Northeast Forestry University, No. 26 Hexing Road, Harbin 150040, China

[5]National Institute of Genetic Engineering and Biotechnology (NIGEB), 14965/161 Tehran, Iran

[6]School of Agronomy and Animal Science, University of Tehran, 4111 Karaj, Iran

[7]International Institute for Applied Systems Analysis, Schlossplatz 1, 2361 Laxenburg, Austria

[8]Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Muggelseedamm 310, 12587 Berlin, Germany

[9]Theoretical Biology and Biophysics, MS K710, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

We here present a detailed picture of the origins of the dog, giving strong and precise evidence for "where and when", and thereby also a first tentative picture of "how, why and by whom" the wolf was domesticated. Previous studies of mitochondrial DNA (mtDNA) have failed to definitely establish the time and place of origin because of lack in phylogenetic resolution for the so far studied 582 bp region, and inadequate sampling across the world. We therefore analysed 169 mtDNA genomes, selected from partial sequences (582 bp) from 1,576 dogs worldwide. This shows that dogs universally share a common gene pool, but the three earlier identified universally occurring phylogenetic clades of high age consist of ten much younger subclades, which originated 5,000-16,000 ya from at least 48 wolf founders. The full range of genetic diversity, all 10 subclades, is found only in south-eastern Asia south of Yangtze River, and the diversity decreases gradually across Eurasia down to only four subclades in Europe. This establishes that the dog has a single origin in time and space from a large number of wolves, less than 16,000 ya, probably in China south of Yangtze River. The place and time coincide with the origin of rice agriculture, suggesting an origin among sedentary hunter-gatherers or early rice farmers. The numerous founders indicate that wolf taming was an important cultural trait, and it is noticeable that in this region dogs are since ancient times used as food, offering a possible reason for the wolf domestication.

The dog was probably the first domesticated animal, and the only one accompanying humans to every continent in ancient times[1]. Despite this central position in human prehistory, there is not full agreement on where, when, and how dogs originated. Archaeological evidence indicates that domestic dogs existed by 12,000 ya[2], but has failed to tell where and at how many different places the dog originated (See the Supporting Information (SI) Text for a more detailed discussion). In an initial major study of mitochondrial DNA (mtDNA) data[3], it was suggested that the dog originated >100,000 ya, and at several different tines and/or places, based on the large age of the main phylogenetic group of mtDNA sequences (The basis for this theory is described in SI Text). However, a later study suggested a single much more recent origin, possibly ~15,000 ya, somewhere in Asia east of the Urals[4]. Since neither time nor exact geographical location has been definitely established, multiple and/or very ancient origins of dogs are still mostly maintained in literature[5-9]. Thus, there is no generally accepted picture of the time and place of origin of the dog, and the only geographical indication so far is for a vast territory covering the Asian continent East of the Urals and the Himalayas. Consequently, the human culture that performed the wolf taming remains unknown. Knowledge of the exact time and place of origin is necessary for identifying the related human culture, and the timing is crucial also for understanding the mechanisms of wolf domestication. An origin 15,000 years ago (ya)[2] requires numerous founders (wolves) to explain today's mtDNA diversity, which would indicate that the taming of wolves was an important custom of the related human culture, while an ancient origin (>100,000 ya)[3] could have involved a single female wolf at a singular chance event.

The earlier studies of mtDNA[3,4] have failed to determine the time of origin and number of founders for dogs because of lack in phylogenetic resolution for the studied region, 582 bp or less of the control region (CR). Since the mutation rate is at most 1 substitution per 40,000 year for this region, there has not been time for closely related founder haplotypes (from wolf) and their derived haplotypes to resolve into separate, identifiable, haplogroups, in the case the dog originated ! 12,000 ya as indicated by archaeological data[1,6] (A detailed discussion is given in SI Text). Therefore, analysis of a larger part of the mtDNA genome is necessary for establishing the time of origin and number of founders. Furthermore, incomplete sampling has hindered an exact determination of the region of origin of dogs. Here, we obtain an exhaustive survey of the global geographical diversity patterns by analysis of 582 bp of the CR for 1,576 domestic dogs, especially densely sampled across Asia (Table 1 and Supplementary Table 1), and based on this we generate the phylogenetic resolution necessary for determining the time of origin and number of founders, by analysing almost the entire mtDNA genome (16,195 bp) for a subsample of 169 dogs.

Results and Discussion

Phylogenetic analysis (Fig. 1a) grouped all CR sequences into the previously described six distinct groups; clades A-F[3] (see Materials and Methods for details). Except for the rare and geographically restricted clades D, E and F (probably derived from late-time wolf-dog hybridisation and therefore not further concerned: see SI Text), the gene pool of the Old World is very homogenous in that clades A, B and C were represented in 97.4% of the dogs and occurred universally and in similar proportions, and that the majority of individuals had a haplotype shared by virtually every population (Table 1, Fig. 1b and Fig. 2). Fourteen haplotypes (9 in clade A, 2 in clade B and 3 in clade C), which were represented in Europe, SW Asia and East Asia (China, SE Asia and Japan), we termed Universal Types (UTs). The UTs are universally very frequent (Fig. 1b), especially in the western part of the Old World (west of the Urals and Himalayas: "the West") where 81.5% of dogs in Europe and SW Asia have a UT (Table 1, Fig. 2) and consequently the same haplotype as a dog in East Asia. The frequency of UTs is considerably lower in East Asia (54.2%) with a minimum of 40.8% in the south-eastern part of Asia: China south of Yangtze River and SE Asia (Figure 2). We term this region "Asia south of Yangtze River" (ASY).

It is even more striking that, except for the UTs, almost all other haplotypes in the West are a single mutation from a UT. Thus, 98.7 % of dogs in Europe and 94.6 % in SW Asia have a haplotype which is either a UT or derives from a UT by a single mutation (collectively termed UT-derived: UTd), implying that close to 100% of dogs in the West have a haplotype which can be traced back to East Asia. In contrast, 69.2 % of dogs in East Asia and only 52.6% in ASY have an UTd. Accordingly, there is a gradient for almost every measure of diversity across the Old World, e.g. number of haplotypes and unique haplotypes, with a distinct maximum in ASY (Table 1 and Table S1). This is directly visible as a difference in coverage of the minimum-spanning networks, most distinctly for clade A (Fig. 1b). Western populations lack several parts of the network, largely the same in all populations, and eastern populations have most parts covered, but the only region with almost complete coverage is ASY.

To study this pattern in detail, we analysed the entire mtDNA genomes for 169 samples, chosen to cover the mtDNA diversity (Fig. 1c and Fig. S1). This showed that the seemingly dense clades

A, B and C contain distinct substructures (Fig. 3a). Clade A has six subclades, and B and C two each, giving totally 10 subclades with high bootstrap values (Fig. 3a) and separated by large genetic distances (Fig. 3b). The CR sequences group into the 10 subclades following a distinct geographical gradient (Fig. 2 and SI Text): only in ASY are all 10 subclades found, while 7 are represented in Central China and Japan, 5 in North China, India and SW Asia and only 4 in Europe. Only one of the six subclades of clade A is represented in Europe and SW Asia, the missing five corresponding to the empty parts of the networks in Fig. 1b. To conclude, the full extent of diversity for clades A, B and C is represented in the region comprising China south of Yangtze River and SE Asia, ASY. Outside this region only part of the total diversity is found, but it can be traced to a subset of the gene pool in ASY, basically the 14 UTs. This suggests that the domestic dog originated from ASY (exactly where in ASY is not possible to determine from this data set [Fig. 2, Table S1]), and probably also at a single time.

The proportion of individuals having clades A, B and C is very similar among geographical regions (Fig. 2, Table 1), and the most probable explanation for this is a single origin from a population having approximately these proportions. If originating in different regions, a majority of each clade in their respective region of origin would be anticipated, and if originating at different times the younger clades would have difficulty spreading to already populated regions. Only very thorough mixing across the entire Eurasian continent could have counteracted these patterns. A population genetic simulation analysis (SI Text) showed that multiple origins in time are virtually impossible, demanding extreme migration rates for sufficient mixing. Multiple origins in space, if occurring almost simultaneously, are also unlikely unless the migration rate was very high (e.g. 30% of dogs in each population migrating to the neighbouring population, in every generation since the time of origin, if an origin 20,000 ya is assumed). The only scenario not rejected at moderate migration rates is a single origin in time and space for the three clades.

The distance to the most recent common ancestors (MRCAs) for the ten subclades varies considerably (Fig. 3a, Table S2), giving times to the MRCAs of ~5,000-40,000 years for the three youngest, and 35,000-125,000 years for the oldest subclade (dating based on a tree-derived substitution rate [0.0296-0.0318 substitutions $site^{-1}$ $time^{-1}$], and time of wolf-coyote split. Since this split has no exact calibration point in the fossil record, but occurred 1.5-4.5 million ya[10], dates can only be given as a time range; see Materials and Methods). However, the older subclades contain a substructure of younger sub-subclades, several of which are represented by samples from across the Old World (Fig. 3b). Thus, several UTs (according to the CR sequence) in clades A, B and C (A11, A16, A17, A18, A19, A20, B1, B6, C1, C3), form sub-subclades each with a common MRCA for samples from East Asia as well as from Europe and/or SW Asia and Africa. For the European, SW Asian and African samples the mean distance to these universal MRCAs are very similar for clades A, B and C: 1.64 (s.e.m. = 0.068), 1.63 (0,034) and 1.71 (0,33) substitutions, respectively. This indicates a simultaneous introduction of clades A, B and C to the West, in agreement with the simulation analysis. The mean distance to the MRCAs for all three clades is 1.66 (s.e.m. = 0.035) substitutions, indicating an origin from the MRCAs (and thus spread of dogs to the West) 4,600-16,200 ya. This agrees with archaeological evidence for dogs in Europe and SW Asia by at least 9,000 ya and possibly up to 12,000 ya[1,6] (See SI Text for detailed discussion). Thus, the earlier inference[3] that mtDNA indicates (based on the assumption that dog clade A originated from a single wolf haplotype) a much earlier origin of dogs (>100,000 ya) is not valid (See SI Text for detailed discussion). Furthermore, the earlier favoured, archaeology based, theory that dogs originated from Europe or SW Asia[1,6] seems incompatible with the mtDNA data; the morphologically very diverse European dog population is obviously a peripheral population with only approximately 50% of the total genetic variation.

It is not possible to identify exactly which haplotypes were carried by the domesticated wolves, and thereby count the number of founders for the dog. However, we may count the minimum number of lineages existing before the spread of dogs from East Asia to the West. It is unlikely ($p<0.01$, Poisson distribution) that more than 6 substitutions have occurred in any lineage since the time of spread of dogs to the West (1.66 substitutions back in time). Counting the number of haplotypes separated by 14 substitutions or more (having MRCAs at least 7 substitutions back in time) we identify 48 lineages leading to today's dog haplotypes. Thus, assuming that the 10 subclades of clades A, B and C formed almost simultaneously shortly before spreading to the rest of the world, less than 16,000 ya, today's mtDNA gene pool must have originated from a minimum of 48 wolf haplotypes, and most probably an even larger actual number of female wolves. This agrees with studies of the dog MHC, identifying 41 founder alleles and therefore a minimum of 21, but according to simulations more probably several hundred, wolf founders[8].

This gives important clues about the mechanisms behind the origin of dogs, showing that the domestication of wolves was not a singular chance event, but rather an intentional act and important custom of the human culture involved. Wolf was present in China south of Yangtze into the 1950s[11], but is not recorded for SE Asia[12]. While not systematically surveyed in China[13], there is firm archaeological evidence of dogs by at least 7,100 ya in North China and 6,500 ya south of Yangtze River[13], but in SE Asia not until 4,000 ya with the spread of agriculture from

South China[11]. China had two centres of plant domestication and early agriculture, of millet by the Yellow river and rice in the Yangtze River area, both at least 8,500 ya[13,14]. Rice domestication probably started to develop first, at least 9,000 ya but possibly as early as 11,000 ya[13,14]. It is possible that the dog originated in this cultural context of increasingly sedentary hunter-gatherers or early rice farmers; the large number of domesticated wolves indicates a stable and widespread society. In contrast to most other parts of the world, dogs have been used as food on a large scale in southern East Asia, from ancient times until today[15,16]. This suggests that the wolf may have been domesticated for use as a source of food, rather than for hunting or guarding, as previously mostly suggested, perhaps from a European, non-dog eating, perspective.

Materials and Methods

Samples and DNA Sequence Analysis. Samples were assumed to represent geographical regions based on origin from either (i) a region with small influx of foreign dogs or (ii) a breed with known historical geographic origin. 907 samples (155 of which for 16,195 bp of the mtDNA genome, excluding repetitive and difficult-to-align regions) were analysed here: DNA was extracted from blood and hair[3] and buccal swabs[17], and analysed by PCR and DNA sequencing[18] as described, and remaining samples retrieved from GenBank (A list with sample information [haplotype, geographical origin, and GenBank accession number], alignments for the CR and mtDNA genome data, and PCR primers for the mtDNA genome, are given in editable format in Datasets S1 and S2).

Phylogenetic Analyses. The CR tree was inferred using a neighbor-joining method (BioNJ) with a HKY+I+G model (I=0.7799, shape=0.5921) and maximum likelihood (ML) evaluated ingroup midpoint rooting as described[3,19], and the whole mtDNA genome (WG) data tree was inferred by an optimized and parallelized ML code based on fastDNAml and DNArates[20,21], in iterative steps to refine the tree and substitution model parameters, as described[22]. The inference robustness of the clades for the CR tree was evaluated by non-parametric bootstrapping (1000 replicates) using BioNJ with a HKY+I+G model. As expected from the previous studies, poor support was found for clades A and B (Supplementary Fig. 2), because of relatively sparse mutation data in relation to the number of sequences (251 taxa; 601 characters with 113 variable sites, of which 81 were parsimony-informative), and the high rate of homoplasies (HI=0.64)[3,4,19]. However, the complete mtDNA genome data recreated all clades supported by high NJ bootstrap values (Fig. 3a and Supplementary Fig. 3), as well as Bayesian support values (Supplementary Fig. 4). Minimum spanning networks were inferred using Arlequin[23].

Molecular Clock Estimates. The rate of substitutions over time of the WG data was estimated using branch lengths in a ML tree without a clock constriction as well as Bayesian Markov Chain Monte Carlo (BMCMC) searches explicitly assuming various clock and population models. Based on the ML tree, with a least square optimization of the exact rooting point on the branch between the coyote and dog/wolf clade[24], the substitution rate was estimated at 0.031 substitutions site^{-1} time^{-1} (where *time* is the number of years since the MRCA of dog/wolves and coyotes [see further below], and the range was 0.0296-0.0318 substitutions site^{-1} time^{-1}, considering variation within the dog/wolf clade and the coyote clade). The divergence time estimates for all clades and sub-clades using all clocks agreed very well with that of the non-constricted tree estimates (R>0.92, Supplementary Table 2) without any significant differences (P>0.16), indicating that the rate was robust to assumptions on clock model and how the population had expanded. In comparison to a constant clock (cc) with rate 1, the Bayesian log normal relaxed clock (lc) had a mean rate of 1.011 (95% HPD 0.896-1.130) and the exponential clock (ec) a mean rate of 1.002 (95% HPD 0.697-1.324). A Bayes factor analysis, however, showed that the best fit was achieved with the exponential clock, followed by the log normal relaxed clock and the constant clock ($\log_{10}$ Bayes factor difference of the harmonic means of the tree likelihoods in each analysis were ec-lc=11.2, lc-cc=5.7, where values above 10 indicate "strong" support)[25-27]. Thus, the error in the clock rate may follow the relaxed exponential clock.

Calculation of Substitution Rates. The substitution rate for the mtDNA genome was calculated from the genetic distance between dog/wolf and coyote obtained above, and the time for their separation according to the fossil record. However, there is no exact calibration point in the fossil record for this split. Since the record is incomplete, the branching in the forming of the lines leading to wolves, coyotes and jackals is not totally clear. The lines leading to wolves and coyotes were well separated by ~1,5 million years ago, but it is not clear when the actual split occurred; a separation 1,8-2,5 million years ago seems most probable, but possibly occurred up to 4.5 million years ago[14]. We therefore used the time range 1.5-4.5 million years ago as calibration value. The substitution rate for the mtDNA genome sequence was therefore estimated as a range, based on the substitution rate estimated from the ML tree (0.0296-0.0318 substitutions site^{-1} time^{-1},) and the time of separation between wolf and coyote (1.5-4.5 million years ago), to $2.12\text{x}10^{-6}$ – $6.58\text{x}10^{-9}$ substitutions site^{-1} year^{-1}, or 1 substitution per 2,900-9,400 years for the 16,195 bp analysed.

1. Clutton-Brock J (1995) in *The Domestic Dog, Its Evolution, Behavior and Interactions With People*, ed Serpell J (Cambridge Univ. Press, Cambridge), pp 7-20.

2. Street M (2002) Ein Wiedersehen mit dem Hund von Bonn-Oberkassel. *Bonn Zool Beitr* 50(3):269-290.

3. Vilá C *et al.* (1997) Multiple and ancient origins of the domestic dog. *Science* 276:1687–1689.

4. Savolainen P, Zhang Y-P, Luo J, Lundeberg J, Leitner T (2002) Genetic evidence of an East Asian origin of dogs. *Science* 298:1610–1613.

5. Lindblad-Toh K *et al.* (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438:803-819.

6. Morey DF (2006) Burying key evidence: the social bond between dogs and people. *J Archaeol Sci* 33:158-175.

7. Ostrander EA, Wayne RK (2005) The Canine genome. *Genome Res* 15:1706-1716.

8. Vilá C, Seddon JM, Ellegren H (2005) Genes of domestic mammals augmented by backcrossing with wild ancestors. *Trends Genet* 21:214-218.

9. Zeder MA, Emshwiller E, Smith BD, Bradley DG (2006) Documenting domestication: the intersection of genetics and archaeology. *Trends Genet* 22:139-155.

10. Nowak RM (2003) in *Wolves: Behavior, Ecology, and Conservation*, eds Mech LD, Boitani L (Univ. of Chicago Press, Illinois), pp 239-258.

11. Gao Z-X (2006) Review of the Research on Wolf in China. *Chin J Zool* 41:134-136 (in Chinese).

12. Higham CFW (1996) A review of Archaeology in Mainland Southeast Asia. *J Archaeol Res* 4:3-49.

13. Underhill AP (1997) Current Issues in Chinese Neolithic Archaeology. *J World Prehistory* 11:103-160.

14. Bellwood P (2005) in *First Farmers: The origins of Agricultural Societies.* (Blackwell Publishing, UK), pp .

15. Simoons FJ (1991) in *Food in China, a Cultural and Historical Inquiry* (CRC Press, Boston), pp 200-252.

16. Higham CFW, Kijngam A, Manly, BFJ (1980) An analysis of prehistoric canid remains from Thailand. *J Arch Sci* 7:149165.

17. Natanaelsson C *et al.* (2006) Dog Y chromosomal DNA sequence: identification, sequencing and SNP discovery. *BMC Genet* 7:45.

18. Angleby H, Savolainen P (2005) A study of the forensic usefulness of the mitochondrial DNA variation among and within populations, breeds and types of domestic dogs. *Forensic Sci Int* 154:99-110.

19. Savolainen P, Leitner T, Wilton AN, Matisoo-Smith E, Lundeberg J (2004) A detailed picture of the origin of the Australian dingo, obtained from the study of mitochondrial DNA. *Proc Natl Acad Sci USA* 101:12387-12390.

20. Olsen G, Pracht S, Overbeek R (1998) Site-specific rates of nucleotide change: their justification and use. In: *5th Annual HIV Dynamics & Evolution Conference: Santa Fe, NM.*

21. Olsen GJ, Matsuda H, Hagstrom R, Overbeek R (1994) fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput Appl Biosci* 10:41-48.

22. Korber B. *et al.* (2000) Timing the ancestor of the HIV-1 pandemic strains. *Science* 288:1789-1796.

23. Excoffier L, Laval G, Schneider S. (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evol Bioinformatics Online* 1:7-50.

24. Athreya G. *et al.* (2008) A simple method for optimizing the root and evolutionary rate in phylogenetic trees with taxa collected at a minimum of two different time points. Submitted to *BMC Evol Biol*.

25. Suchard MA, Weiss RE, Sinsheimer JS (2001) Bayesian selection of continuous-time Markov chain evolutionary models. *Mol Biol Evol* 18:1001-1013.

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

27. Jeffreys H. (1998) *Theory of probability* (3 edn.) (USA: Oxford University Press).

Acknowledgements

This work was supported by grants from the National Basic Research Program of China (973 Program, 2007CB815700), Chinese Academy of Sciences (KSCX2-YW-N-018), Bureau of Science and Technology of Yunnan Province, National Natural Science Foundation of China (30621092), the Swedish Research Council, OE and Edla Johanssons Scientific Foundation, the Carl Trygger Foundation, the Wenner-Gren Foundations, and the Swedish Kennel Club.

Author Information J-F.P., C.K., X-J.Z., and A.Z. contributed equally to this study. DNA sequences have been deposited at GenBank under accession numbers XXXX-XXXX. Correspondence and requests for materials should be addressed to P.S. (savo@biotech.kth.se) or Y-P.Z. (zhangyp@mail.kiz.ac.cn, or zhangyp1@263.net.cn).

Table 1. Genetic diversity for CR data across the Old World

Region	N ABC(DEF)	nA(%)	nB(%)	nC(%)	HT	HTuniq	HTres	PropUT	PropUTd
East Asia	730 (5)	549(75.2)	124(17.0)	57(7.8)	121	84	28.5 ± 3.0	54.2	69.2
West	558 (27)	384(68.8)	123(22.0)	51(9.1)	68	41	23.0 ± 2.5	76.2	95.5
Europe	313 (23)	217(69.3)	68(21.7)	28(8.9)	39	19	19.5 ± 2.0	81.5	98.7
SW Asia	130 (3)	72(55.4)	45(34.6)	13(10.0)	30	8	19.5 ± 2.1	81.5	94.6
Africa	56 (1)	48(85.7)	6(10.7)	2(3.6)	22	6	22.0 ± 0.0	57.1	91.1
India	59 (0)	47(79.7)	4(6.8)	8(13.6)	22	6	21.5 ± 0.6	54.2	81.4
Siberia	60 (2)	39(65.0)	13(21.7)	8(13.3)	20	7	19.7 ± 0.5	46.7	75.0
Japan	118 (3)	76(64.4)	24(20.3)	18(15.3)	25	6	19.7 ± 1.7	58.5	84.7
Korea	90 (7)	80(88.9)	6(6.7)	4(4.4)	24	5	19.1 ± 1.6	63.3	77.8
N China	98 (0)	65(66.3)	25(25.5)	8(8.2)	24	5	18.8 ± 1.6	79.6	89.8
C China	141 (0)	109(77.3)	21(14.9)	11(7.8)	27	8	17.4 ± 1.9	70.9	85.8
S China	281 (0)	223(79.4)	44(15.7)	14(5.0)	71	40	27.7 ± 2.8	42.0	53.4
SE Asia	57 (2)	50(87.7)	3(5.3)	4(7.0)	30	11	29.7 ± 0.5	35.1	50.9
ASY	338 (2)	273(80.8)	47(13.9)	18(5.3)	87	53	30.1 ± 3.0	40.8	53.0

East Asia – China, 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. N/n – number of individuals, for the phylogenetic clades; HT – number of haplotypes; HTuniq – number of unique haplotypes; HTres – number of haplotypes obtained from resampling of size 56 (500 replications) to adjust for different sample size; PropUT/UTd – proportion of individuals carrying a UT and UTd.

Figure 1. Genetic relationships between the mtDNA CR sequences (582 bp).

a) Phylogenetic tree for dog (unlabeled) and wolf (filled square) haplotypes, rooted by coyote (Coy) sequences (distance reduced by 55%). Dog clades A-F are indicated.

b) Minimum spanning networks for dog clades A, B and C. Haplotypes (circles, UTs with bold lining) and empty nodes (solid dots) separated by one substitutional step. The presence (coloured circle) and frequency (proportional to circle size) of haplotypes is indicated for geographical regions; blue indicates haplotype shared with other regions and orange unique haplotype (excluding comparison to SE Asian islands). Circle size is reduced for the most frequent haplotypes for East Asia (eight most frequent reduced by 30%) and West (most frequent by 50%, next seven most frequent by 30%).

c) CR haplotypes chosen for analysis of the mtDNA genome, indicated by colour. The different colours refer to the subclades (six for clade A, two each for B and C) subsequently identified.

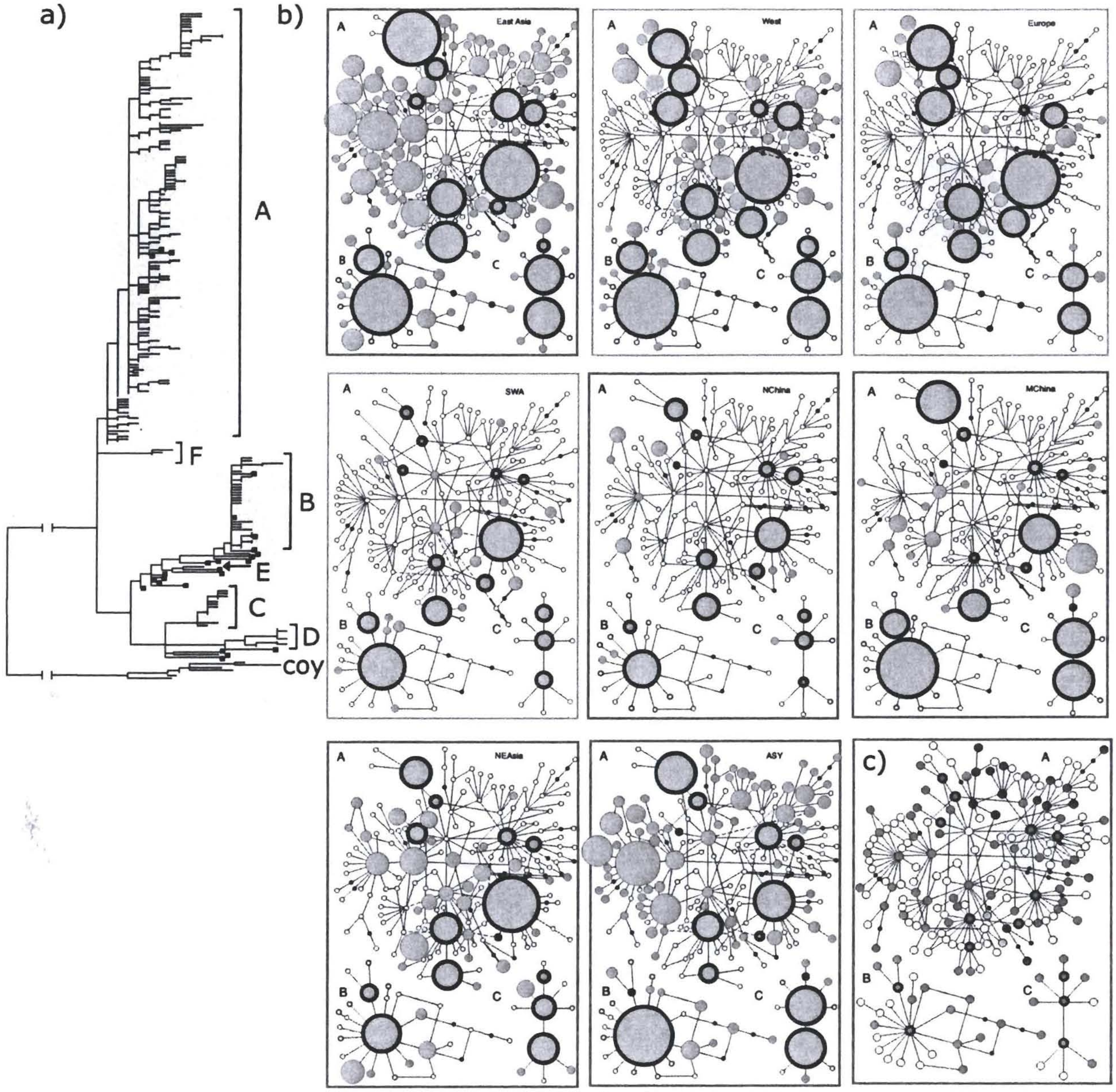
Figure 2. Genetic diversity across the Old World.

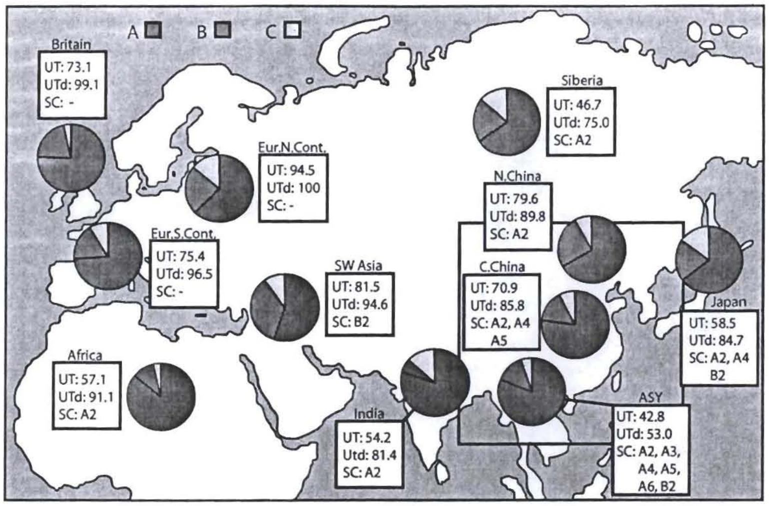
Pie diagrams: proportions of clades A, B and C; Boxes: proportion of individuals having a UT or UTd, and representation of the six non-universal (of totally ten) subclades (SC) of clades A, B and C. In lower panel, sampled regions are coloured. LHS: Liaoning, Hebei and Shanxi. Tibet includes samples from Quinghai and Nepal.

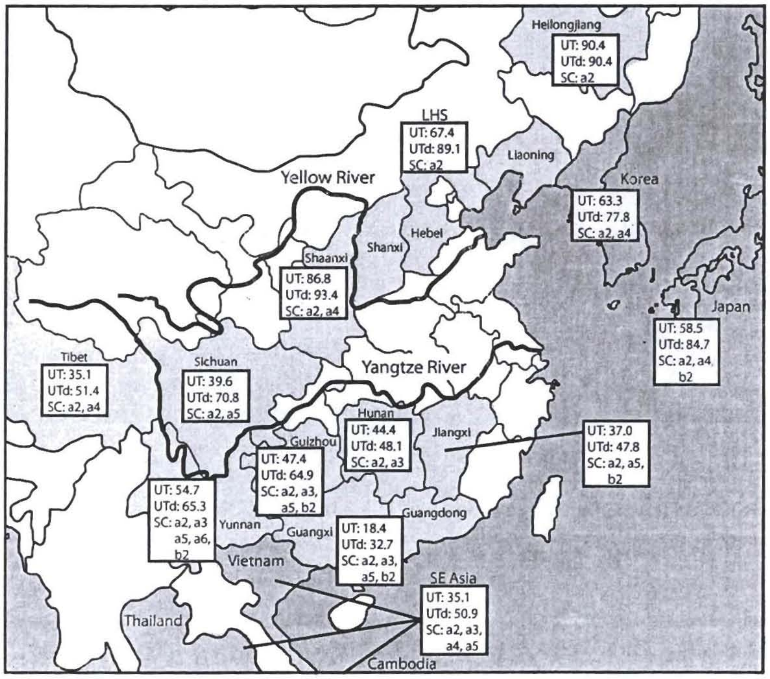
Figure 3. Genetic relationships between the mtDNA genome sequences (16,195 bp).

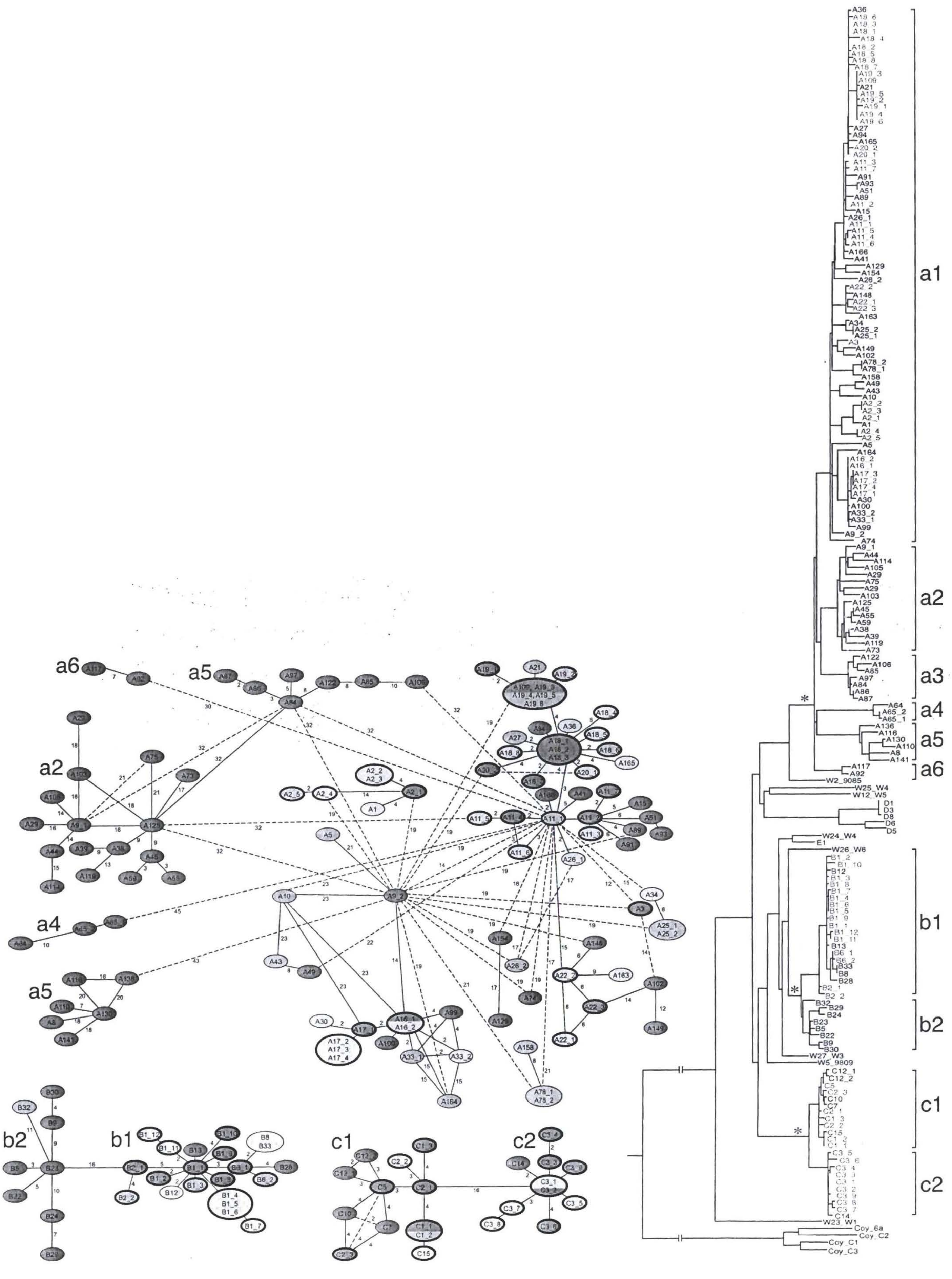
a) Phylogenetic tree for dog and wolf rooted by coyote sequences (distance reduced by 87%). Sample identity for dogs given as CR haplotype (red indicating UT) and individual number for samples identical for the CR. “W” indicates wolf and “Coy” coyote. Dog clades A, B and C (marked by *) and their subclades a1-a6, b1-b2 and c1-c2, were all supported by bootstrap values "87% and Bayesian values "99%.

b) Minimum spanning networks for dog clades A, B and C. Haplotypes (ellipses, UTs with bold lining) are separated by the number of substitutions indicated next to the connecting lines, no number indicating a single substitution. Sample identity the same as in panel b (identical samples given within the same ellipse). Colours indicate geographical origin: Red, East Asia; Blue, Europe; Green, SW Asia; Purple, Africa; Yellow, India.


a)
A
F
B
E
C
D
coy
b)
East Asia
West
Europe
SWA
NChina
MChina
NEAsia
ASY
c)
A
B
C


A
B
C
Britain
UT: 73.1
UTd: 99.1
SC: -
Eur.N.Cont.
UT: 94.5
UTd: 100
SC: -
Eur.S.Cont.
UT: 75.4
UTd: 96.5
SC: -
SW Asia
UT: 81.5
UTd: 94.6
SC: B2
Africa
UT: 57.1
UTd: 91.1
SC: A2
India
UT: 54.2
Utd: 81.4
SC: A2
Siberia
UT: 46.7
UTd: 75.0
SC: A2
N.China
UT: 79.6
UTd: 89.8
SC: A2
C.China
UT: 70.9
UTd: 85.8
SC: A2, A4
A5
Japan
UT: 58.5
UTd: 84.7
SC: A2, A4
B2
ASY
UT: 42.8
UTd: 53.0
SC: A2, A3,
A4, A5,
A6, B2


Heilongjiang
UT: 90.4
UTd: 90.4
SC: a2
LHS
UT: 67.4
UTd: 89.1
SC: a2
Liaoning
Yellow River
Korea
UT: 63.3
UTd: 77.8
SC: a2, a4
Hebei
Shanxi
Shaanxi
UT: 86.8
UTd: 93.4
SC: a2, a4
Japan
UT: 58.5
UTd: 84.7
SC: a2, a4,
b2
Tibet
UT: 35.1
UTd: 51.4
SC: a2, a4
Sichuan
UT: 39.6
UTd: 70.8
SC: a2, a5
Yangtze River
Hunan
UT: 44.4
UTd: 48.1
SC: a2, a3
Jiangxi
UT: 37.0
UTd: 47.8
SC: a2, a5,
b2
Guizhou
UT: 47.4
UTd: 64.9
SC: a2, a3,
a5, b2
UT: 54.7
UTd: 65.3
SC: a2, a3
a5, a6,
b2
Yunnan
Guangxi
UT: 18.4
UTd: 32.7
SC: a2, a3,
a5, b2
Guangdong
Vietnam
SE Asia
UT: 35.1
UTd: 50.9
SC: a2, a3,
a4, a5
Thailand
Cambodia


a6
a5
a2
a4
a5
b2
b1
c1
c2
a1
a2
a3
a4
a5
a6
b1
b2
c1
c2
*
*
*