Supplementary Materials for

The Evolutionary History of Dogs in the Americas

Máire Ní Leathlobhair1*, Angela R. Perri2,3*, Evan K. Irving-Pease4*, Kelsey E. Witt5*, Anna Linderholm4,6*, James Haile4,7, Ophelie Lebrasseur4, Carly Ameen8, Jeffrey Blick9,†, Adam R. Boyko10, Selina Brace11, Yahaira Nunes Cortes12, Susan J. Crockford13, Alison Devault14, Evangelos A. Dimopoulos4, Morley Eldridge15, Jacob Enk14, Shyam Gopalakrishnan7, Kevin Gori1, Vaughan Grimes16, Eric Guiry17 , Anders J. Hansen7,18, Ardern Hulme-Beaman4,8, John Johnson19, Andrew Kitchen20, Aleksei K. Kasparov21, Young-Mi Kwon1, Pavel A. Nikolskiy21,22, Carlos Peraza Lope23, Aurélie Manin24,25, Terrance Martin26, Michael Meyer27, Kelsey Noack Myers28, Mark Omura29, Jean-Marie Rouillard14,30, Elena Y. Pavlova21,31, Paul Sciulli32, Mikkel-Holger S. Sinding7,18,33, Andrea Strakova1, Varvara V. Ivanova34 , Christopher Widga35, Eske Willerslev7, Vladimir V. Pitulko21, Ian Barnes11, M. Thomas P. Gilbert7,36, Keith M. Dobney8, Ripan S. Malhi37,38, Elizabeth P. Murchison1,§,a, Greger Larson4,§,a and Laurent A. F. Frantz4,39,§,a

* These authors contributed equally to this work a These authors co-supervised this work § corresponding authors: epm27@cam.ac.uk (E.P.M); greger.larson@arch.ox.ac.uk (G.L.); laurent.frantz@qmul.ac.uk (L.A.F.F)

This PDF file includes:

Material and Methods Figs. S1 to S28 Tables S3 to S8, S12 to S16 Captions for Tables S1, S2, S9 to S11

Other Supporting Online Material for this manuscript includes the following:

Tables S1, S2, S9 to S11 (Excel)

Materials and Methods

Archaeological Site Descriptions Here we describe all sites associated with pre-contact samples that were used in the analysis (see Table S1 and Fig 1a for more information). We include the sample number(s) assigned by the home depository along with our study sample number(s) (e.g., depository; sample). Where only one set of numbers is given the depository number and sample number are the same.

Aachim Lighthouse (AM1, AM2; CGG10, CGG11) The Aachim Lighthouse site is located on the Aachim Peninsula on the Eastern Siberian Sea coast. This site has the remains of sea mammals and stone tools, as well as the mandibles of two dogs that have been radiocarbon dated to 1,700 BP (26).

Angel Mounds (AM310A, AM310B, AM310C, AM474; FS3305, FS3283.33, FS3283.22, FS2187)

Angel Mounds is located along the Ohio River just east of present day Evansville, Indiana. The site consists of eleven mounds, several plazas, and numerous residential neighborhoods surrounded by a large semi-circular palisade (27). More recent research suggests that site construction began just before AD 1050 and continued through AD 1450 with the primary occupation occurring during the Late Mississippian period between AD 1350-1450 (28). Extensive excavations were conducted at Angel Mounds, from the WPA era in the late 1930's through the 1960's, and several dog burials and dog skulls are recorded from excavations. Of the four dog burials listed in the Angel Mounds catalog, only one contains an articulated dog skeleton that can be considered an intentional and undisturbed dog burial. This dog, FS 3305 (AM310A), was photographed during WPA excavations and was recovered from the block known as X11B. There are also the remains of two dogs (which are mixed in the same context) recorded as "dog burial", FS 3283.33 (AM310B) and FS 3283.22 (AM310C), which comes from the same excavation block as the articulated dog; and FS 2187 (AM474), which comes from block W11A. These second and third "burials" consist of several disarticulated bone fragments representing incomplete remains, and do not have any hand-drawn or photographed in situ representations associated with them. FS 3283 includes a smaller dog's cranial and mandibular fragments (partially burned) as well as a large dog's limb bones and postcranial elements. Materials associated with FS 2187 include the cranium, without mandibles or postcranial elements, of what is an average-sized Mississippian dog along with the shaft portion of a single right humerus from a sandhill crane (Grus canadensis) from context W11A.

Anker (ISM21C; AL2705) The Anker site (IAS CK 21) is a Mississippian village site on the north side of the Little Calumet River, just south of Chicago (29). Part of the site was destroyed from previous construction efforts, but a dome-shaped house, fire pits, storage pits, and more than 30 human burials were identified. The individuals interred at this site were likely wealthy, as they were buried with numerous grave goods, including animal skulls and copper pieces. The

pottery that has been recovered is fairly homogenous, all jars in the

Oneonta style, which suggests that this site dates to AD 1400–1500. A small number of Canis bones were recovered from the site, including some dog teeth that were used as grave goods. The Anker sample included in this study was morphologically identified as Canis lupus.

Apple Creek (ISM070; AL2707) The Apple Creek site is a Woodland village on the north bank of Apple Creek, four miles north of Eldred, IL (30). The site was excavated in 1962 and 1963, and has cultural components from the Hopewell and Woodland periods. The majority of faunal bones recovered from the site were white-tailed deer (Odocoileus virginianus), suggesting that they were an important food source for the area (31). Seven dog burials have been recovered from the site, as well as some additional isolated elements, which show evidence of consumption. The dog burials are likely from the Middle or Late Woodland periods (2,000-1,000 BP). The dogs are all terrier-sized, which is smaller on average than most other Woodland dogs. The dog used in this analysis, Burial 1, was a small mature male dog with more tooth wear on the right side of the maxilla and mandible than the left.

Baum, Ohio (OSU_13320; AL2748) The Baum site is a prehistoric village settlement in Paint Creek River Valley, Ross County, Ohio. This is the type site for the Baum Phase of the Fort Ancient culture, dating between AD 950-1200 (32, 33). The village extends over 10 acres of ground, and consists of dozens of circular post structures 3-4 meters in diameter, surrounding a large central mound (34, 35). The mound has been known since the mid-19th century, and is interpreted as a large ceremonial pyramid structure with two levels, containing 17 human burials (32, 34, 36), and is perhaps the best-preserved mound in the region. In three seasons of excavations covering 2 acres, Mills uncovered 47 tepee structures, 127 burials, and 234 refuse pits, concluding that the site was occupied year-round by a minimum of several hundred occupants (34).

Abundant, well preserved faunal remains were found at the site within the refuse pits, and remains of domestic dogs were found ubiquitously throughout the village. These remains were described, in the early 20th century, by A.F. Lucas as belonging to Indian dogs of a size and proportion similar to the bull terrier (34). A total of 50 dog remains, including 7 crania, were collected. Some of the bones had evidence of cut marks, while others were broken consistent with other prey species, and some were made into ornaments. All crania collected were broken, which Mills suggested was to remove the brain (34). Comparisons with other prehistoric dogs suggest these were of the general Indian dog type found at contemporary sites in the region, as well as at prehistoric sites in Texas, and the old Pueblos (34, 37).

While individual dog burials and associated ritual activity are known from contemporary sites at the time (e.g., (38, 39) the evidence from Baum suggests a more domestic role for these dogs. The breakage of dog bones similar to other food animals indicates that at least on some occasions dogs were used as food. Gnaw marks from dogs on other animal bones found throughout the middens suggests that dogs had access to food waste. Unlike earlier Archaic

period sites where dogs were regularly afforded

special treatment indicated by their careful burial, these later prehistoric dogs from Baum appear to have been a utilitarian presence at the site, likely similar to modern feral or village dogs, with deposition predominantly occurring in refuse pits.

Cox (1Jo176-66-6; Cox6) The Cox mound is located in Jackson County, Alabama, and had a number of flint tools and pot sherds associated with the mound (40). The site dates to the Middle Woodland period (1-500 AD). To the east of the mound, there were thirty human burials, many of which included grave goods such as shell beads and tempered bowls. Multiple dogs have also been recovered from the site, one of which is included in this analysis.

Channel Islands (CA-SNI-4, CA-SNI-21 133.21-A, CA-SNI-21 133.21-7, CA-SBA-27 F.492, CA-SRI-41, CAW2, CAO1; CISNI4, CINHA, CINH7, CIAS, CICVD, CAW2, CAO1)

The Channel Islands are 20 to 98 km from the coast of Southern California, and have been occupied for 13,000 years (41). The peoples living on the island had a primarily marine diet and complex trade networks both between islands and with the mainland. Dogs have lived on the island for at least 6,000 years, and were almost certainly brought over by humans (42). They were likely not used as hunters, but may have been an occasional food source, and were found on all of the Channel Islands. Seven individuals from three islands, as well as one mainland archaeological site, have been used in this study. Two dogs were recovered from Santa Cruz Island, one from the Orizaba Cove area (SCRI-434) that dates to the Late Holocene, and the other from Willows Canyon. Three dogs are from San Nicolas Island, two of which come from the North Head site (SNI-21), which was occupied during the Terminal Early Period (5,000 BP) and the Middle Period (2,000 BP). The third dog is from site SNI-4, although its age is uncertain. One dog derives from Santa Rosa Island, at the Canada Verde site (CA-SRI- 41), which dates to the terminal Early Period (4,000 BP). The dog from the mainland was recovered from the coastal Chumash town of Syuxtun, which is now a part of modern- day Santa Barbara, and was important both for trade and for interactions between mainland populations and the people from Santa Cruz Island.

Flint River (1Mo48 – 40-11; FR11) The Flint River site was a village with a large circular shell mound (2 m tall, 15 m in diameter) located on the Flint River near Courtland, Alabama (40). The mound is made of raw clay, with a large shell deposit. Over 200 human burials have been recovered from the site, as well as 19 dog burials (43). Dogs were interred throughout the period of occupation, and did not share graves with humans. Large numbers of artifacts were also recovered from the mound, including pottery and sandstone bowls, pipes, awls, and other tools. The majority of the components are from the Late Archaic period, however intrusive Mississippian human burials have also been recovered from the mound (44). The dog analyzed in this study may be Late Archaic or Mississippian.

Grass Mesa (5MT23-16; 5MT316) Grass Mesa Village was a large village located east of the Dolores River in Colorado (45). It was occupied from AD 700 to the early 900s AD, and has one

of the highest

artifact densities of any archaeological site in the Mesa Verde region, with a large ceramic assemblage spanning the whole length of human occupation. The earliest period of occupation consisted of dispersed homesteads clustered around pit structures, while later periods of occupation (starting around AD 760) saw the construction of a great kiva, as well as room blocks and more closely-spaced residential structures. Grass Mesa Village had its height of occupation (roughly 150 households) around AD 850, although areas of the village began to be abandoned not long after. The majority of faunal remains from the site were of multiple species of deer and rabbits, although there were also small numbers of dog and unidentifiable Canis spp. bones.

Janey B. Goode, Illinois (11S1232 98-1, 11S1232 34-2, 11S1232 1671-1, 11S1232 845-1, 11S1232 939-1, 11S1232 2601-1, 11S1232 1724-1, 11S1232 4109-2, 11S1232 2793-2, 11S1232 5267-1, 11S1232 5819-1, 11S1232 6287-1, 11S1232 7023-1, 11S1232 7892-1, 11S1232 F9 PP4407-1, 11S1232 I15 PP4747-1, 11S1232 L17 P 2-1, 11S1232 3993-2, 11S1232 Q25 PP4811-1; JBG1M, JBG5, JBG11, JBG12, JBG13, JBG17, JBG19, JBG21, JBG24, JBG26, JBG32, JBG35, JBG37, JBG41, JBG42, JBG43, JBG45, JBG48, JBG50)

The Janey B. Goode site (11S1232) is a settlement near Brooklyn, Illinois, that was occupied from AD 650 to 1400 and was excavated from 1998 to 2004 (46). This site includes a small Late Woodland component (AD 650-900), and is primarily composed of Terminal Late Woodland period features (AD 900-1050), which include storage pits and structures that were likely houses. A small Mississippian component (AD 1050-1400) has also been identified, and this site is only 8 km from Cahokia, the largest Mississippian mound site as well as the center of the Mississippian world. Interestingly, Janey B. Goode also has one of the largest numbers of dog burials of any archaeological site – dogs have been recovered from 102 different features (47). The majority of these dogs have been recovered from the Terminal Late Woodland component. Many of these dogs were recovered as complete skeletons, and were found in storage pits or buried under the settlement structures. A small number of individuals may have been ritually sacrificed – six males were buried without skulls, suggesting they were beheaded prior to their burial. Cut marks on a Mississippian dog indicate that dogs may have been consumed during the Mississippian period, but dogs from the Late Woodland and Terminal Late Woodland periods show no evidence of consumption, but instead have vertebral fractures, which may be suggestive of their use as pack animals prior to the Mississippian period.

Koster, Illinois (ISM_F2256A; AL2135) The Koster site is a complex, highly-stratified site located in a tributary valley of the lower Illinois River in west-central Illinois, excavated between 1969 and 1978 by the Foundation for Illinois Archaeology, Northwestern University and the Center for American Archeology (48–51). Cultural deposits dated from the early Archaic through Mississippian periods are buried to depths of over 10m, providing a continuous record of Holocene human occupation in the region (52, 53). Given this record, the Koster site has become a valuable resource for investigating changes in the environment, settlement patterns, technology, and human adaptations to climatic and social variation (53–55); see (56) for

overview).

The early Archaic layers at Koster (Horizons 11 and 12), dating from c. 9,900-9,000 calibrated BP (57), are some of the most well-studied, revealing dense layers of habitation, artifacts, middens, and hearths. Horizon 11 also included the burial of nine humans, four infants and five adults ((53), outside of the living area. Excavators also discovered the burial of four dogs, found as complete, articulated skeletons in shallow pits (58, 59). At the time of excavation the dogs were dated based on associated material to around 8,500 years ago, making them the oldest examples of intentional dog burials in the world and some of the oldest domesticated dogs ever identified; certainly the earliest in the New World. Our recent direct dating of two of the Koster dog burials (ISM_F2256A and ISM_F2357) dated them to between 10,130-9,680 calibrated BP respectively (2), making them among the earliest dated material from Koster and confirming their position as the earliest identified domesticated dogs in the Americas and the earliest dog burials in the world.

One of the specimen included in this study (ISM_F2256A; AL2135) was found in a shallow, basin-shaped pit with a metate and mano placed near its cranium, though it is not clear whether this is associated with the burial. The skeleton is complete and the animal was buried lying on its side, with no evidence of intentional cut marks or other trauma. Given the presence of preserved bacula in the other dogs, the missing baculum here suggests the dog was a female, and it was an adult (58). Morey and Wiant (58):225- 226) previously discussed the confirmed status of the Koster canids as domesticated dogs based on comparisons to modern and ancient wolves and coyotes, due to their small size. Particularly, they noted the wide palate and cranial vault of this specimen in comparison to coyotes, a finding that seem to be corroborated by our genetic analyses (see below).

Little Bear (CT8 D-2 2-90; LB2) The Little Bear Creek site is an Archaic shell mound located at the mouth of the Little Bear Creek, in Colbert County, Alabama (60). The mound contained nearly 200 burials, some of which were cremated or interred with grave goods including shell beads. Other artifacts recovered from the mound include shell-tempered pottery and flint knives and projectile points. The dog remains from this site are all from the Late Archaic period, and were deliberately buried in the mound, similar to the human interments, but dogs and humans were not found buried together at this site (61).

Mayapan (M2: Q152 bag 1, M3: Q152 bag 1, M4: Q152 bag 1, M10: Q162 bag 4; May2, May3, May4, May10)

Mayapan was the largest Mayan city during the Postclassic period, and was occupied from its founding at 900 BP until its abandonment at 450 BP ((62). It is located in Yucatan, Mexico, and consisted of a monumental center containing temples and ritual buildings, surrounded by residential areas which contained both palaces for the elite and smaller houselots ((63, 64)). Dog remains were highly concentrated in the Templo Redondo Group, located in the main plaza of the monumental center (Masson and Lope 2013). Over 2,000 dog bone fragments have been identified from the site (65). Young dogs were consumed regularly, and older dogs were often

used in ritual contexts.

McPhee Pueblo (5MT4475-20; 5MT520)

6

McPhee Pueblo was part of McPhee Village, which is located 6 km northwest of Dolores, Colorado (66). It was occupied for 200 years during the Pueblo I and II periods (roughly AD 800-980). The pueblo consisted of housing clusters with both living and storage areas, clustered in groups with pit structures. McPhee Pueblo was the largest pueblo in the village at its height of occupation (AD 870-900), and was also the only part of the village to be occupied for the full range of occupation. Over 200 bones have been recovered from the site that are identified as Canis spp., including coyotes, wolves, and dogs.

Modoc (ISML50, ISM090; AL2810, AL2706) The Modoc Rock Shelter is at the edge of the Mississippi floodplain, located 2 miles SE of Prairie du Rocher, IL, and is one of the earliest sites of human occupation in Illinois (67). The lowest strata has been dated to 11,000-9,000 BP, and the site was occupied for 6,000 years. In the earliest strata, tools and projectile points are primarily all that has been recovered, but slightly later in time fire pits have been identified as well. Trash piles of vertebrate remains suggests that the individuals living in the rock shelter were hunters, and also ate large numbers of gastropods. Post molds have been identified starting at 8,000 BP, along with six human burials, but there is no trace of human occupation after 4,000 BP. Two complete dog burials were recovered from the same strata as the human burials (68).

Perry (1Lu25-Dog 35, 1Lu25-Dog 39, 1Lu25-Dog59; P35, P39, P59) The Perry site is a large shell mound located in northern Alabama, on the Tennessee River (69). The shell mound contains hundreds of human burials that have been dated to the middle Archaic to Mississippian periods, based on the artifacts present with the burials, as well as a large assemblage of stone tools (70). Over 100 dogs have also been buried in the shell mound, some of them with humans. The age of these dog burials are largely unknown, as the majority of the dogs were buried without associated grave goods.

Port au Choix, Newfoundland (MU_NP50A_1; AL3194) The Port au Choix site is located on Newfoundland's northwest coast on the Port au Choix peninsula, projecting into the confluence of the Gulf of St. Lawrence and the Strait of Belle Isle. The area encompasses a number of well-preserved localities, including a Maritime Archaic burial ground (Port au Choix-3) with over 100 preserved burials (Port au Choix-3, Locus II), excavated from 1967-1969 by Memorial University of Newfoundland (71–73). The Maritime Archaic are defined as native groups in the Atlantic Provinces, dating from approximately 9,000-3,500 years ago, who were well- adapted to a coastal environment (71, 74). The burial ground at Port au Choix is thought to date to approximately 4,400-3,300 years ago (75).

The remains of four dogs were recovered from the Port au Choix-3 burial ground (for review see (76) – two complete and articulated associated burials (c.f. (77) accompanying human Burial 50 (Locus II), fragments of one dog from Locus I, and another from burial fill at Locus II (73). All the dogs appear to be of the Large or Common Indian dog size (c.f. (78):459), though there is some variation between them which is within the range of sexual dimorphism (73). The

dog analyzed here

7

(MU_MP50A_1, AL3194) is an older male, likely weighing between 45-55 pounds, killed by a blow to the head, and included, along with another male dog, in a multi- human burial (73): 77-78). Direct radiocarbon dating of the dog resulted in a date of 4,402-3,912 calibrated BP (UCIAMS159456), indicating it comes from the earliest use of the burial ground within the Maritime Archaic period. Tuck ((73):78) proposed these dogs were likely used as companions, hunting aids, and occasionally as travois dogs due to their well-developed muscle attachments. Comparisons to Eskimo sled dogs were unfavorable, but the Port au Choix specimens were similar to other large breed native dogs seen at the prehistoric site of Frontenac Island, New York (79). Early descriptions of large breed native dogs described them as slender and wolf-like, with large erect ears and a long, pointed snout (78):461).

Prince Rupert Harbour (D1-184, D9-985, D10-973, W5-74, W8-969; PRD1, PRD9, PRD10, PRW5, PRW89)

Prince Rupert Harbour is on the northern Northwest Coast of North America, just south of the Alaskan boundary. Two sites, GbTo-13 and GbTo-54, had large-scale excavations in 2012-2013 (80). The work used a novel method of electronic data collection where virtually all data was spatially linked to real-world coordinates (81). The larger excavation, at GbTo-54, had nearly 500 m3 excavated. The oldest component here dates from 900 BC (calibrated radiocarbon) to about 250 BC; the main occupation was between 250 BC and AD 800; and there was a slightly less intensive occupation from about AD 900 to 1300. Some late pre-contact activity also occurred, but there were no European trade goods, suggesting no admixture with European dogs likely occurred in the population of dogs at this site (also suggested by our mtDNA analysis; see below).

GbTo-13 had smaller an occupation limited to AD 1000–1300.

Both sites were remarkable for the large numbers of exotic and high-prestige artifacts and animal bones found. These latter include an unprecedented amount of mountain goat (82) as well as grizzly and black bear, sea lion, northern fur seal, wolf, and other species otherwise rare in Northwest Coast assemblages. The mammal NISP was 5,810 for GbTo-54 and 499 for GbTo-13 (82). Rank-linked artifact and faunal distribution indicated that one house was a chief's residence, yet even the poorest houses appear to have had much more wealth than at almost all other sites in Prince Rupert Harbour, suggesting regional and intra-village ranking. One purposeful dog burial was found just outside the chief's house at GbTo-54. Some 300 dog bones and 21 wolf bones were recovered from GbTo-54, and 25 dog from GbTo-13. The dogs are smaller than the wolves in the area, and so the two species can be distinguished. Three dogs and two wolves were used for this analysis.

Reinhardt, Ohio (OSU_F5607; AL2772) The Reinhardt site (33PI880) is a middle period (AD 1200-1400) Fort Ancient village site in the Scioto Valley, Ohio. The site was first systematically excavated in 1988, and was identified as an approximately 91-meter diameter midden ring

visible on the surface immediately adjacent to a terrace edge located 150 meters from the Scioto River (83, 84). More recent survey and excavations confirm that the site was a circular, or arc shaped, village, with a small central plaza (83). Associated AMS dates from wood

charcoal and a single human burial support the dating of the site to within the Middle Fort Ancient period (83), though bifaces from at least four earlier time periods were recovered during excavations.

Excavations in 2008 revealed a small number of canid remains, including two individual dog burials (39, 84). The specimen in this study is from a fully articulated, adult, male dog with heavily worn teeth, that was found buried under a layer of sand with a turkey bone awl (84). Significant pathologies to the dog were recorded, including vertebral pathologies representing either severe arthritis, or a healed infection, and an abscess in the right M2 (39). In their analysis of the two dog burials, Nolan and Sciulli (39) conclude that there was significant variation between dogs at the Reinhardt site, though their sample size was prohibitively small. Comparisons with other dogs from the region indicate little change in morphology between the Archaic and later prehistoric periods, despite this being a time of intense cultural transmission and changes in subsistence patterns and lifeways (39, 85).

The role of dogs in the prehistoric Eastern Woodlands is intimately connected with the prevalence of dog burials throughout the Archaic period, a tradition which continued to a lesser degree into the later prehistoric periods (38). Dogs in this region were known to serve a variety of purposes, such as for transportation, hunting, companionship and as a food source. Dogs also participated in ceremonial and ritual activity, and dog sacrifices were not uncommon (38, 86, 87). The individual burial of dogs at Reinhardt suggests a level of intimacy remained between the Fort Ancient people and at least some of their dogs, though isolated dog remains found throughout the site indicate that this intimacy was not afforded to all dogs equally.

Scioto Caverns (OSU_4816_1_1, OSU_4816_2_2, OSU_4816_2_4, OSU_4816_2_6, OSU_4816_2_8, OSU_4816_3_4, OSU_4816_3_8; OSU611, OSU622, OSU624, OSU626, OSU628, OSU634, OSU638)

The Scioto Caverns site in Ohio are a series of three limestone caverns located near the Scioto River and Wright-Holder earthworks complex (88). The bones of 25 dogs have also been recovered from this deposit, including 5 nearly-complete skulls and 11 mandibles. The dogs are likely from the Hopewell period (200 BC - 500 AD), based on the nature of their burial, in which they were covered with limestone slabs. Based on their size, the dogs seem similar to Archaic dogs identified from Kentucky and Alabama, and are smaller than Woodland period dogs from the Midwest.

Simonsen Bison Kill (ISM_13CK61_7_2; AL2699/ISM172) The Simonsen site is located in northwest Iowa, near Quimby, and its usage dates to the Paleoindian and Archaic periods (89). It has been suggested to be a bison kill site, given that the majority of remains at the site are of bison, and there are large numbers of points associated with the bison (90). Three zones at the site contained cultural material (89). The deepest has a small number of artifacts and the most bison bones, likely from Bison occidentalis. The middle horizon contains charcoal as well as a

single artifact, while the highest zone contains charred wood and ash suggestive of a firepit, as well as the ramus of a large dog, which was used in this analysis.

Tizayuca (MT04_1661, MT02_707; AL2546, AL2552) Tizayuca (Hidalgo, Mexico) is located in the Basin of Mexico, at only 20 km of Teotihuacan, the largest urban center of Mesoamerica from 1,800-1,400 BP. Operations of rescue archaeology, between 2004 and 2009, allow the identification of at least three successive occupations during the Classic (c. 1,800-1,500 BP), Early Postclassic (c. 1,100-800 BP) and Middle/Late Postclassic (c. 800-500 BP) under the respective influence of Teotihuacan, Toltec and Aztec cultures (91). Over 800 dog bones have been identified, from neonatal to adults, representing at least 41 individuals. Cut marks and burning patterns suggest the consumption of dogs, but several individuals have also been buried in residential and ceremonial areas (92).

Sample AL2552 comes from a partial articulated skeleton buried in a Teotihuacan residential compound. The presence of a baculum indicates it was a male. Sample AL2546 comes from a disarticulated skeleton found in the basement of a domestic household related to Aztec occupation.

Uyak, Alaska (HMCZ_38342; AL3198) The Uyak Site (KOD-145), also known as “Our Point”, is a substantial prehistoric midden at least three meters deep, covering hundreds of acres and located on the western side of Kodiak Island, Alaska. The site was excavated by physical anthropologist Aleš Hrdlička from 1933-1936 for the United States National Museum (now the National Museum of Natural History), and consisted of house structures and hearths, stone and organic artefacts, human remains, and a large faunal assemblage (93). The stratigraphy of the midden deposits was divided into three layers by Hrdlička’s team based on artifact type and deposit descriptions, and occupation dates from 2,000 BP to Russian contact in the mid-18th century were suggested (93–95). More recent AMS dating of a selection of dog and fox bones confirms that the site was occupied during this period (96). Due to the long occupation history of the site, and without direct dating, the cultural context of the study specimen can only be summarized to span both the earlier Kachemak phase (4,000- 900 cal BP) and the later Koniag phase (900-200 cal BP), as it is clear dogs were present throughout the occupation (93, 97). Both cultures were maritime hunter/gatherers, though debate still surrounds whether the nature of the relationship between the two cultures was a direct ancestral relationship or a replacement event (93, 98).

Primarily concerned with the recovery of human remains, Hrdlička and Allen prioritized the collection of only well-preserved and intact canid remains (97), and noted that dogs were ubiquitous throughout the site. Hundreds of domestic dog remains were collected, and early metric analysis identified two types of dogs, classified as a “small” and a “large” type (97). While Hrdlička (93) originally suggested that the small type was restricted to the earliest deposits, later replaced by the large type, other studies have noted size variability among contemporary prehistoric canids throughout the Kodiak archipelago (99), and more recent work suggests the variation could be a result of sexual dimorphism (96).

Excavation of the specimens was poorly recorded, so it is unknown what context the dog remains were found in, and whether they represent intentional burials, midden/refuse deposits or some combination of deposition types. This makes any interpretation of the role of dogs at the site difficult, though cut marks suggesting butchery were recorded on a portion of both mandibular and cranial elements (100). Dogs are a known food source, both preferentially or as a "fall-back food" during times of hardship, and additionally serve a variety of other functions in the Arctic, including assisting in prey acquisition, transportation, sanitation through food waste disposal, and even warmth (101–103). As the region's sole domesticate, it is likely the dogs at Uyak fulfilled several of these roles simultaneously.

Weyanoke Old Town, Virginia (DB49-LC1, DBU2-LM1; AL3226, AL3223) The Weyanoke Old Town site, also known as the Hatch site, is located on a tributary of the James River on the coastal plain southeast of Richmond, Virginia. The site was excavated between 1975 and 1989 by the Virginia Foundation for Archaeological Research, Inc. (104, 105) and covers over 4,329 sq. meters (106). Based on associated artifacts, occupation at the site spans from at least the early Archaic through the early English Colonial period, when the Virginia Algonquians occupied the region ((104). The site was home to a Weyanoke (Weanoc tribe) village dating from the prehistoric through the early Colonial period (107, 108)).

Over 112 domestic dogs have been recovered from the site, one of the largest discoveries of domestic dogs in the Americas (86, 106, 109). The dog was the only domesticated animal of the Virginia Algonquians ((110)) and what little is known about the use of native dogs in the region is pieced together from early accounts in contact- period historic documents. There is no evidence that dogs at Weyanoke were used for food, as all dogs were recovered in isolated and associated burials (c.f. (77)), articulated with no evidence of cut marks or burning (106). Dogs of the Virginia Algonquians are documented partaking in small and large game hunting and as protection from predators like wolves and bears (111, 112). Like other regions of the Midcontinent and Eastern Woodlands (113–115), dogs, both puppies and adults, played an important role in burial with humans at Weyanoke. It is unclear whether dogs at Weyanoke were viewed as companions, ritual offerings, or both, but dog sacrifices were not uncommon among contact-period tribes (e.g., (116, 117).

Descriptions of the Virginian native dog often mention a wolf-like ((117–120) or fox-like ((121) appearance with an inability to bark and a propensity for howling. In an analysis of the Weyanoke Old Town skeletal material, Blick ((109)) noted the dogs there are clearly domesticated and not wolf-like in their cranial and postcranial skeletal morphology. He later proposed that admixture between local wolves and aboriginal dogs may explain the reported "wolf-like" behavior and appearance of Algonquian dogs ((106):6). They stood an average of 42 cm high and weighed approximately 10kg, similar to a medium-sized Algonquian native village dog depicted in a 1585 painting by John White (see (122):62, Plate 32). This painting, the earliest depiction of a Native American dog by a European, shows a knee-height, medium-sized

yellow-brown dog with pricked ears and a long tail, similar to many modern village dogs or dingoes. In contrast to this, in

1602 John Brereton described dogs in coastal Massachusetts as “fox-like, black, and sharp-nosed” ((121):13). The Weyanoke canids are associated with Late Woodland period artifacts, confirming their native ancestry and lack of inbreeding with later introduced European breeds. This is supported by direct radiocarbon dating of one of the dog used for DNA analysis (AL3223) to 985-935 cal BP (OxA-35,516).

Yellow Jacket Pueblo (5MT501) Yellow Jacket Pueblo was a large village in the Central Mesa Verde region that was occupied during the Late Pueblo II through the Pueblo III periods (roughly AD 1050- 1260) (123). The village had a number of public and ceremonial structures, including room blocks, plazas, kivas, small towers, and what may be the a Chacoan Great House. A small number of canid remains have been identified (NISP = 27), including one bone that was specifically identified as belonging to a dog. Some of the canid bones show evidence of burning or cut marks, suggesting that these dogs were used as food.

Zhokhov (Zhokh2004-18, Zh-90-2, Zhokh2004-19, Zh-90-1, Zhokh2004-113, Zh-03- 97, Zh-90-3, Zh-04-154, Zh-05-29; CGG1, CGG2, CGG3, CGG4, CGG5, CGG6, CGG7, CGG8, CGG9)

Zhokhov Island is located off of the northeastern coast of Siberia (124, 125). Human occupation of Zhokhov began in the Late Pleistocene period, with the earliest evidence for humans at the site dating to 9,000 years ago ((125) . House pits have been identified from this time period, as well as bone and antler fragments and abundant amounts of wood. Small blades and bone and ivory points have also been recovered from the site, as well as a fairly sophisticated sledge runner for the time period. Two fragmented dog mandibles, as well as a small number of postcranial bones, have been recovered from the site, and date to the earliest period of human occupation. They are smaller than wolf mandibles, and are similar in size to other ancient Arctic dog remains that have been recovered. Previous mitochondrial DNA sequencing has demonstrated that the dogs belong to Haplogroup A, and are genetically indistinguishable from modern domestic dogs in the hypervariable region of the mitochondrial genome (26).

Ancient DNA - University of Oxford DNA extraction DNA was extracted from teeth or bone samples (see Table S1) in a dedicated ancient DNA laboratory using the appropriate sterile techniques and equipment. Extraction was carried out following the Dabney extraction protocol (126) but with the addition of a 30 minutes pre-digest stage (127).

DNA Sequencing Illumina libraries were built following (128), with the addition of a six base-pair barcode added to the IS1_adapter.P5 adapter. The libraries were then amplified on an Applied Biosystems StepOnePlus Real-Time PCR system to check that library building was successful, and to determine the optimum number of cycles to use during the indexing amplification PCR reaction. A six base-pair barcode was used during the indexing amplification reaction resulting in each library being double-barcoded with an “internal adapter” directly adjacent to the ancient DNA strand and which would form the

first bases sequenced, and an external barcode that would be sequenced during Illumina barcode sequencing. Libraries were sequenced on an Illumina HiSeq 2500 (Single End 80bp) sequencer at the Danish National High-Throughput Sequencing Centre and on a Illumina NexSeq 500 (Single End 80bp) at the Natural History Museum (London).

Ancient DNA - University of Illinois Most of the DNA extractions were performed at the University of Illinois at Urbana Champaign, at the Carl R. Woese Institute for Genomic Biology, with methods described in Witt et al. (129), but a subset of extractions was performed at the Centre for GeoGenetics at the University of Copenhagen, with methods described in Allentoft et al. (130).

At the University of Illinois, genomic libraries were built for the extracts using the NEBNext DNA Library Prep Kit for Illumina (New England Biolabs). They were amplified twice, first using the NEBNext DNA polymerase and the associated index primers, to allow the samples to be pooled and sequenced together. The first amplification followed manufacturer instructions, and was repeated for twelve cycles. For the second amplification, Phusion High-Fidelity PCR Master Mix with HF buffer (New England Biolabs) was used, and four PCR reactions were made for each sample. Five uL of PCR product from the first amplification was added to each reaction, and the DNA was amplified according to manufacturer's instructions for 12 cycles. The four reactions for each sample were pooled and cleaned using Ampure XP (Beckman Coulter) and MagSi-DNA NGSPREP beads (MagnaMedics), using an 80% ethanol: sample ratio. The library was visually examined on an agarose gel, and quantitated using a Qubit 1.0 fluorometer (Thermo Fisher Scientific). Only libraries with concentrations of at least 20 ng/uL were used for capture.

To enrich for mitochondrial DNA, we developed a custom set of RNA baits as part of a myBaits kit (Arbor Biosciences) that covered the complete dog mitogenome with 4x tiling density (see Table S1 for details on the samples that were captured). Captures were performed using the myBaits manual version 3.01 at the University of Illinois, with a 60° C incubation for 28 hours. The heat elution step was skipped, and the capture was amplified using KAPA Hi-Fi polymerase, following manufacturer's instructions for 16 cycles. The PCR reaction was cleaned using MagSi-DNA NGSPREP beads (MagnaMedics), and the capture was visually examined on an agarose gel and quantitated using a Qubit 1.0 Fluorimeter, following the manufacturer's instructions. If the DNA concentration of the amplified capture was lower than 20 ng/uL, the capture was reamplified for 8 cycles using the KAPA polymerase prior to sequencing.

At the University of Copenhagen, genomic libraries were built from the extracts using the NEBNext DNA Library Master Mix Set 2 (New England Biolabs), with modifications. The End Repair mix was incubated for 20 minutes at 12° C and 15 minutes at 37° C. The Quick Ligation mix was incubated for 20 minutes at 20° C. The Fill-In mix was incubated for 20 minutes at 65° C and 20 minutes at 80° C. Each step was purified using a Qiagen MinElute PCR Purification Kit. The protocol was followed as directed except that a differing amount of EB Buffer was used

for each mix (30 uL for

13

End Repair, and 42 uL for Quick Ligation) and the column was incubated for 15 minutes at 37° C prior to elution. The finished libraries were amplified using Taq Gold in a mix that included 10 uL Taq Gold Buffer, 2.5 mM $MgCl_2$, 0.8 mM uL BSA, 0.08 mM dNTPs, 0.2 μM of each of Illumina's Multiplexing PCR primer and a custom-designed index primer with a six-nucleotide index and 2 uL Taq Gold, in a total volume of 100 uL. qPCR was performed on the libraries to assess the quantity of DNA. The PCR conditions were followed according to manufacturer's directions, amplifying for 10-14 cycles, depending on the qPCR results. The PCR reaction was purified using the QIAQuick PCR Purification Kit, with elution in 30 uL EB Buffer and an incubation at 37 C for 10 minutes prior to the elution step. DNA concentration was assayed using a Qubit 2.0 Fluorimeter, following manufacturer's instructions. If the DNA concentration was less than 20 ng/uL, a second PCR amplification was performed using Phusion. The mix included 20 uL template DNA, 2 uL each of primers IS5 and IS6, 50 uL Phusion Master Mix, and 26 uL H2O. The PCR program followed manufacturer's instructions but for 6- 10 cycles, and was purified with a QIAQuick PCR Purification Kit as described above. The capture procedure was performed following manual version 2.3.1 at the University of Copenhagen, with a 65 C incubation for 18 hours. The heat elution step was skipped, and the capture was amplified using KAPA Hi-Fi polymerase, following manufacturer's instructions for 16 cycles. The PCR reaction was cleaned using a QiaQuick PCR Purification Kit, eluting 30 uL of EB Buffer after a 15 minute incubation at 37 C. The capture was visually examined and quantitated using an Agilent 3300 Bioanalyzer. If the DNA concentration was lower than 20 ng/uL, the capture was re amplified using the KAPA polymerase.

Samples were pooled 8-10 individuals to a sequencing lane, and were sequenced on an Illumina HiSeq 2500 (80 or 100 bp). The samples captured at Copenhagen were sequenced at the Danish National DNA Sequencing Center, and the samples captured at the University of Illinois were sequenced at the Roy J. Carver Biotechnology Center at the University of Illinois.

Data processing - ancient DNA Raw reads were filtered allowing one mismatch to the indices used in library preparation. Adapter sequences were removed using AdapterRemoval (131). Reads were aligned using Burrows-Wheeler Aligner (BWA) version 0.7.5ar405 (132) to canFam3.1, with default parameters apart from disabling the seed option ("l 1024") (133). FilterUniqueSAMCons (134) was then used to remove duplicates. BAM files from different sequencing lanes were merged using the MergeSamFiles tool from Picard v1.129 (http://broadinstitute.github.io/picard/). To accommodate the low coverage of the nuclear genome of our newly sequenced North American dogs, genotypes were called by randomly sampling a single read of 20 base pair minimum and with a mapping quality (MAQ) and base quality (BQ) of at least 30 at each covered position in the genome, excluding bases within 5bp of the start and end of a read (135–137). The Newgrange dog was genotyped similarly as modern

data (except for 5bp at start and end of a read; see below)(138).

For the mtDNA we generated majority consensus (using reads with BQ>=20 and MAPQ>=30) sequence for all samples that had at least 3x average coverage (71 samples; Table S1) excluding bases within 5bp of the start and end of a read.

Molecular damage was assessed using MapDamage2.0 using default parameters (139) (Figure S1; Figure S2). Most samples display clear signs of deamination (Figure S1; Figure S2). Samples AL3231 and AL2696 display limited deamination patterns consistent with these being from a relatively recent time period (1000-1400 AD; Table S1).

Publically available data Raw reads/bam files for 47 canid genomes (13–16) were downloaded from NCBI or DoGSD (140)(Table S2). Samples downloaded from NCBI were aligned to the CanFam3.1 reference genome using BWA mem (132). We computed depth of coverage (DoC) for each sample using bedtools (141). These genomes were chosen due to their high coverage and geographic spread covering North and South American, East Asian, and Western European (African, Indian and European) dogs, as well as Eurasian and American gray wolves and Coyotes and an outgroup (Lycalopex culpaeus; Andean fox). We also obtained data from two additional publically available CTVT genomes (142), genome data from an ancient wolf from the Taimyr peninsula (143) and from an ancient Irish dog (138)

Lastly we obtained data from 5,406 modern dogs that were genotyped on the semi- custom CanineHD SNP array (~185K SNPs) developed by (9).

Genotyping Dog samples We used samtools ‘mpileup’ (0.1.19) (144) to call genotypes with default settings. Pileup files were further filtered, for each sample, using the following criteria:

Minimum DoC >= 6 Excluded all sites in region of high DoC (top 5%) Excluded all sites within 3bp of an indel Only bases with quality >=30 within reads with mapping quality >= 30 were used. Minimum fraction of reads supporting heterozygous (variant allele frequency [VAF] >= 0.3) - all sites that did not pass this criteria (0 < VAF < 0.3) were coded as missing (N).

For high coverage ancient sample (Newgrange dog) we also discarded the first and last 5bp of each read for genotype calling, to avoid incorporating errors from deaminated sites (see above).

The Taimyr wolf was processed using the same random read approach used for the other ancient data (see above).

CTVT

Ancestry analyses were performed using data from two CTVT genomes that have been previously described, 24T and 79T (142). The goal was to determine the phylogenetic placement of CTVT within a cohort of modern and ancient dogs. CTVT genomes carry two types of genetic variation: germline variation inherited by the CTVT founder dog, and somatic variation acquired during the somatic evolution of the CTVT clone. The goal of this part of the analysis was to capture CTVT germline variation, and to use this to include CTVT in a phylogenetic analysis.

We generated a list of callable sites in CTVT using the criteria outlined in 'Genotyping - Dog samples'. Only regions that retained germline diploidy, as previously described (142), were considered. Sites were further filtered to retain only those sites in which the variant allele fraction (VAF) for a non-reference allele was ≥ 0.1 in at least one CTVT tumor, and for which no more than two nucleotides were detected at VAF ≥ 0.1 (i.e. multi-allelic sites were rejected). These sites were defined as single nucleotide variant (SNV) candidates

CTVT tumor biopsies contain both CTVT cells and host cells. The latter derive from stromal, immune and blood vessel components. Thus DNA derived from CTVT tumors is an amalgam of CTVT and matched host DNA. In order to identify and exclude SNVs derived exclusively from the matched host, as well as to correctly genotype alleles shared between CTVT and matched hosts, we took the following approach. Sites identified as SNV candidates (see above) were genotyped in genomes 24H and 79H, the matched hosts for tumors 24T and 79T(142). The genotypes of 24H and 79H were inferred using the following VAF thresholds:

- homozygous reference: VAF < 0.2
- heterozygous: VAF = [0.2-0.8]
- homozygous alternative: VAF > 0.8.

Using the known host contamination fractions for 24T and 79T (142), we used the following VAF thresholds to genotype SNV candidates in CTVT cells. SNVs that were homozygous reference in the matched host were genotyped in CTVT using the following VAF thresholds:

Homozygous reference: VAF<0.2 Heterozygous: VAF=[0.2-0.6] Homozygous alternative:VAF>0.6 SNVs that were heterozygous in the matched host were genotyped in CTVT using the following VAF thresholds:

Homozygous reference: VAF<0.3 Heterozygous: VAF=[0.3-0.7] Homozygous alternative:VAF>0.7 SNVs that were homozygous alternative in the matched host were genotyped in CTVT using the following VAF thresholds: Homozygous reference: VAF<0.4 Heterozygous: VAF=[0.4-0.8] Homozygous alternative:VAF>0.8

CTVT SNVs were further processed as described below (Ascertainment panel).

Ascertainment panel All genotypes from all genome-wide samples were then merged using bedtools. Ascertainment was done without outgroups (but including Coyote). We selected all bi-allelic markers excluding sites that 1) were heterozygous in only one sample (Table S2) (required a minimum of two chromosomes in our set of samples to carry the derived allele; in the case of sites only variable in CTVT we required the two genomes (24T and 79T) to be homozygous to limit the inclusion of somatic mutations into the list of SNPs 2) sites that were not covered in our outgroup (Andean fox) 3) sites with more than 20% missing data across samples. All low coverage ancient samples were excluded from this step. As the two CTVT matched hosts, 24H and 79H, were not included in our ascertainment panel (Table S2), the genotypes of these two individuals were not taken into account when determining which CTVT SNPs were represented in other dog genomes. This resulted in ~6.21M high quality SNPs. We then excluded all sites that were outside of the germline diploid region in CTVT (142). This resulted in ~2.03M SNPs, including ~600K transversions.

mtDNA analysis RAxML We used all samples with at least 3x average coverage and consensus sequences with at least 80% coverage over the entire mtDNA genome were considered for further analysis. We further obtained ancient and modern mtDNA genomes from (145). This data set contains representative samples of all four major haplogroups (A, B, C, D) including 3 ancient American dogs. We aligned the data using mafft v7.2 (146, 147). We built a maximum likelihood tree, with 100 bootstrap replicates using GTR+G model as implemented in RAxML (148).

All but one ancient American mtDNAs formed a monophyletic clade within haplogroup A, (bootstrap value=87; Figure S3). North American dogs further cluster with ancient sled dogs from the the island of Zhokhov in Eastern Siberia (125). Unsurprisingly, samples CGG10-11 (Aachim dogs) fall outside of the Zhokhov / pre- contact clade as these are recent sled dogs from Siberia (~1.5kya; Table S1). Lastly, one sample from British Columbia (Prince Rupert Harbour site; PRW89; ~1.5Kya) clusters with North American wolves. Wolves and dogs are poorly distinguished at this site (see above) so the sequence of this sample might be from a wolf - although interbreeding between wolves and dogs is also a possibility (see below).

We then assessed whether previous studies that used control region of the mtDNA were able to identify the pre-contact monophyletic clade we have identified here (Figure S3). To do so, we extracted the control regions from all samples, that overlapped with the fragments analysed in (5) and (149) (605 bp in total), filtering out samples with more than 10% missing data and a ML tree with RAxML. The result of this analysis are presented in Figure S4. We found that while the control region has the power to distinguish between the major dog haplogroup (A, B, C, D) it did not possess the power to distinguish between pre-contact dog and other dogs within haplogroup A.

We expanded our mitogenome sample size to assess whether the mtDNA haplogroup that we had identified in pre-contact dogs exists in modern America dogs. To do so we used 942 additional mitogenomes from a worldwide sample of dogs, including CTVT and hosts genomes as well as 169 village and breed dogs that were sampled in North and South America (150–155). Description and accession number of all additional samples can be found in (150) (in Supplementary file 1 and 8 of (150)). We combined this data with the mtDNA genome analysed above and built a maximum likelihood tree, with 100 bootstrap replicates using a GTR+G model as implemented in RAxML (148). Out of 667 modern domestic dog genomes analysed here we found only five modern samples with a pre-contact mtDNA haplogroup (Figure S5): 1) Terrier cross from San Juan del Sur, Nicaragua (Accession: KU291094), 2) Chihuahua (Accession: EU408262) 3) Japanese Spitz (Accession: EU789755) 4) non-breed dog from Shanxixian, China (Accession: EU789669) 5) non-breed dog from Laem Ngop, Thailand (Accession: EU789664).

Interestingly, two out of five of these sequences are from American dogs (Chihuahua and Nicaragua dog). Three, however, are from East Asia. This is surprising and suggests a very low frequency of the PCD haplogroup in East Asia (~2.5%). All five modern samples cluster together with ancient Mexican dogs from Mayapan (Figure S5). Multiple scenarios could explain the finding of East Asian dogs within the PCD clade: 1) the clade to which these East Asian samples belong diverged from PCD dogs prior to their introduction into the Americas 2) there was some back and forth migration of dogs between America and Asia after the flooding of the land bridge between Western and Eastern Beringia ~11,000 years ago (156) 3) these sequences were mislabelled.

BEAST We used BEAST v1.8.4 (157) to calibrate the evolutionary rate of our canid data set. We restricted this analysis to sequences with at least 10x average coverage (Table S1). The mtDNA was partitioned into four categories (tRNA, rRNA, control region and coding sequence). We fitted a separate substitution model to each partition: tRNA (HKY+I), rRNA (TN93+G), control region (HKY+G+I) and coding sequence (SDR06) as selected by Akaike information criteria (AIC) using partitionfinder (158). The same tree was used for all four partitions. Age of archeological samples was used as prior (uniform distribution of tip age). We used a Bayesian Skyline prior (159)(group size parameter = 10) and a strict molecular clock as in (145) (uncorrelated clock was also tested and did not result in noticeable changes). We ran 50 million Markov chain Monte Carlo (MCMC) chains and sampled tree parameters every 5,000 iterations. Convergence was evaluated with Tracer v 1.6.0 (ESS for each parameter >=100). Trees were summarized using Maximum Clade Credibility as implemented in TreeAnnotator v1.8.4 (10% burn-in).

BEAST retrieved the same topology as RAxML, with all pre-contact dogs forming a highly supported monophyletic clade (with a posterior probability of 0.99; Figure S6). We estimate that TMRCA of all pre-contact dogs is ~14666 years (95% HPD: 12965-

16484) and that the time to most recent common ancestor (TMRCA) of all sampled

~15606 years ago (95% HPD:13739-17646).

We also inferred the age of the MRCA between the Mayapan dogs and the five modern dogs identified as monophyletic with PCD (Figure S5). We found that these five modern dogs diverged from the Mayapan dogs between 9,865 and 6,289 years ago (95% HPD) suggesting that their divergence postdates the flooding of the land bridge between Western and Eastern Beringia. Given this results, it is unlikely that the mtDNA haplotype of these dogs originated in Eurasia. These results instead suggest that dogs carrying PCD ancestry have been transported from Americas into East Asia. This most likely took place during recent times and could be linked to the creation of hairless dog breeds in Asia (160). However, more research is needed to further test these possibilities, especially given the possibility of mislabelling in databases such as GenBank.

Nuclear ancestry analyses PCA Using smartpca (161) we performed Principal Components Analysis (PCA) using various projections and data sets on our 2.03M SNPs:

All canids (including wolves and coyotes) - PCD samples projected (Figure S7)

Only dogs (excluding wolves and coyotes) - PCD samples projected (Figure S8)

Only dogs (excluding wolves and coyotes) - PCD and CTVT samples projected (Figure S9)

For PCD we used all 7 samples for which we could call at least 10,000 sites (minimum number of sites suggested for ancient DNA analysis (130)). We used all available sites (sites covered in at least 1 ancient sample; ~1.5M SNPs) to compute the eigenvectors and then projected PCD onto that space. We also projected CTVT to ensure that their placement was not an artefact of somatic mutations.

Figure S7 shows that PCD are more closely related to dogs (except for one sample, AL2135 from Koster; see below) than wolves or coyotes. It also shows that dogs are less variable than wolves or coyotes. Figure S8 shows a distinction between Arctic, East Asian, and European dogs. Lastly, Figure S9 recapitulates the same results demonstrating that this result is not induced by somatic mutations in CTVT and also shows how PCD and CTVT are more closely related to each other, and fall in between Arctic dogs and all other dogs. A PCD sample (AL2135; Koster, Illinois) was projected in between dogs and wild canids (Figure S7). This suggests that this sample is admixed with wild canids (see D-statistics analyses below). Its mtDNA haplotype, however, clusters with other PCD dogs (Figure S3).

Neighbour joining tree We used plink v1.9 (162) to compute an Identity By State (IBS) matrix using all 2.03M SNPs. This matrix was used to build a neighbour joining tree (NJ) using the R

package “ape” (163); (Figure S10). The tree recapitulates the deep split between East Asian and Western Eurasian dogs (138) and confirms that CTVT is more closely related to the PCD dogs than to any other dog population (bootstrap=100). It also shows that CTVT/PCD form a monophyletic group with Arctic breeds that fall outside of the rest of the dogs. The tree also confirms that PCD form a monophyletic clade with high support (bootstrap = 100).

Admixture analyses (see D-statistics below) show that all East Asian dogs, excluding Vietnamese Village dogs, are significantly admixed with European dog populations. Such disproportionate admixture could affect the topology of the tree. To test this we built a tree excluding all East Asian dogs except Vietnamese. This tree shows a different topology with Vietnamese still grouping with Dingoes, however, PCD, Arctic dogs and CTVT are now more closely related to Western dogs than to Asian Dogs (Figure S11).

Bayesian Tree We built a phylogeny using nuclear genotypes with MrBayes 3.2 (164). To do so we used PGDSpider 2.0.9.2 (165) to build a Nexus file with discrete SNP format (0=reference, 1=heterozygous, 2=homozygous alternative). We used the Mkv model (166) implemented in MrBayes (Ordered character), which provides a likelihood framework for data sets that contain only variable characters. We also imposed a minimum distance of 10Kb between SNPs to limit the influence of linkage disequilibrium (LD) and lastly included only PCD samples with higher coverage (AL3194 and AL3223; Table S1; ~30K SNP total).

We ran two independent runs of four MCMC chains with two million samples. Trees were summarized discarding 25% as burnin. To limit biases from missing data we limited this analysis to transversions that were covered in 90% of our samples. Convergence was assessed by ensuring that average standard deviation of split frequencies was below 0.01 and that the potential scale reduction factor was close to 1 for all parameters. This analysis confirms that CTVT and PCD are monophyletic with high support (Posterior probability [PP] = 1; Figure S12) and the basal placement of the CTVT/PCD clade. However, this analysis suggests that modern Arctic dogs are more closely related to Eurasian dogs than to PCD. This is most likely due to the complex ancestry of Arctic dogs such as admixture from European dogs (see below; Table S4).

f3 statistics We computed outgroup f3-statistics as f3(pre-contact dogs [PCD], X; outgroup) using ADMIXTOOLS (167) where X is any other dog population (see Table S2), to quantify the amount of genetic drift shared between pre-contact dogs and other dogs using only transversions (Figure S13). For this analysis, we used only two PCD samples (AL3194 and AL3223; Table S1), with ~1.9x and ~0.5x coverage, respectively. Our results support our NJ tree (Figure S10) demonstrating that PCD is more closely related to CTVT and Arctic dogs than any other dog population. These results also support the observation that PCD/CTVT and Arctic breeds are equally related to all other dogs,

except for Basenji and one Indian dog, which could be due to admixture from wolves into these two samples (e.g. Indian wolf or golden wolf).

D-statistics We only used two PCD samples with ~1.9x and ~0.5x coverage (AL3194 and AL3223; Table S1) for these analyses, except when explicitly mentioned (e.g. Koster dog AL2135; see below). We used all 2.03M SNPs.

PCD is more closely related to CTVT and Arctic breeds We computed D(Outgroup, PCD, Pop3, Pop4) using ADMIXTOOLS (167) where Pop3 was fixed as either European dogs, Asian dogs, Arctic dogs, or CTVT and Pop4 represented any possible other sample. We plotted, as box plots, the results of these combinations (Figure S14; Figure S15). Positive values imply that PCD shares more derived alleles with the population on the y axis, while negative values imply that pre- contact dogs are closer to the other dog populations. The results indicate that PCD do not share any more derived alleles with European dogs than they do with Asian dogs, suggesting that they are equally related to both. This result supports our f3-statistics and NJ tree finding that PCD and CTVT are equally related to European and Asian dogs (Figure S10; Figure S13). Arctic breeds also appear more closely related to PCD/CTVT than any other dog population (Figure S14; Figure S15; Figure S13).

Admixture between Coyote / North American wolves and PCD We tested for admixture from wild North American canids into higher coverage PCD genomes (AL3194; AL3223). To do so we computed D(Outgroup, Coyote/North American Wolf, Pop3, Pop4) where Pop3/4 can be any dog genome. We found that in both cases Z values were mostly above 3 in most cases (Figure S16; Figure S17) for both AL3194 and AL3223 indicating admixture from Coyotes / North American Wolves in PCD samples. We also tested for extra admixture from wild canids into our higher coverage PCD genomes (AL3194; AL3223). To do so we computed D(Outgroup, Coyote, AL3194, AL3223) and D(Outgroup, American Wolf, AL3194, AL3223). We found no evidence of extra admixture from wild canids into these samples (Table S3).

We computed D(Outgroup, Coyote, CTVT, B) and D(Outgroup, American Wolf, CTVT, B) where B represented every possible pair of populations to determine whether there was any detectable admixture from Coyote and American wolf populations into the CTVT founder dog (Figure S16; Figure S17). We also found evidence that the CTVT founder dog shared more derived alleles with Coyotes than other non-PCD population suggestive of admixture. This is consistent with TreeMix and Qpgraph analyses (see below). We note, however, that this pattern could be consistent with admixture in both direction (dogs to wolves / wolves to dogs; e.g. see (168)).

Estimating Eurasian ancestry in Arctic dogs Using D-statistics based on whole genome data we found evidence that all Arctic breeds are a mixture of the basal lineage (that leads to CTVT and PCD) and of the Eurasian dog lineage (Table S4).

Taimyr admixture into Arctic dogs, PCD and CTVT We used whole genome data to assess Taimyr wolf admixture into PCD, Arctic dogs and CTVT (143). We found few values with $|Z| > 3$ (AL3194 and Alaskan malamute; Table S5). Lowering the threshold to $|Z|>2.5$, we found admixture from the Taimyr wolf into PCD (both AL3194 and AL3223) as well as in all Arctic dogs (husky, Greenland sledge dog, and Alaskan malamute) and CTVT. We find no evidence for additional admixture from the Taimyr wolf into either CTVT, PCD or Arctic breed (Table S6). This suggests that Taimyr admixture into Arctic dogs suggested in (143) may have taken place after the PCD, Arctic dog and CTVT lineage diverged from Eurasian dogs but before the divergence of the Arctic dog and PCD/CTVT lineages.

Admixture from European dogs into East Asian dogs We tested for admixture from European dogs into east Asian dogs. Following (9, 138) we used Vietnamese village dogs as the reference East Asian population to test for by computing D(Outgroup, Portugal, Vietnam, X). We found evidence of admixture in all East Asian populations test in this study (Table S7).

Potential evidence for Coyote admixture in Koster dog (AL2135) Our PCA analysis suggests that AL2135 is admixed with wild canids. To test this hypothesis we computed D(Outgroup, North American canid / Taimyr wolf, AL2135, AL3194). We restricted this analysis to AL3194 as it is the highest coverage PCD dog available in this study. We found borderline significant results ($|Z|>2$; Table S8) suggestive of admixture from Coyote into AL2135. This sample, however, is very low coverage (only ~17K SNPs were called). Its placement on the PCA and this positive admixture signal might therefore be due to this low coverage.

Estimating pre-contact ancestry in modern North American and Arctic dogs We used the SNP array data obtained from (9) to assess the degree to which modern dog populations found in North America retained ancestry from pre-contact dogs. This SNP panel contained 28 genotyped populations from North America, such as Peruvian village dogs, Alaskan village dogs or Carolina dogs (see Table S9 for the full list). We computed f4 ratios using ADMIXTOOLS (167, 169) to estimate admixture proportion (α) from pre-contact dogs into these populations by computing:

α =f4(A,O;X, C) ÷f4(A, O; B, C)

Where A is CTVT, O is the Andean fox (outgroup), B is PCD (AL3194 and AL3223), C is any European or East Asian population (see Table S9) and X is any American dog (see Table S9 and Figure S18). We computed α for all combinations of European/Asian and modern North American populations (jackknifing was performed with a block sizes of 1 cM).

Besides the Alaskan Village dogs, we found no significant signal of pre-contact ancestry in modern North American populations (α always <4% and Z always < 3; Table S10). Alaskan Village dogs, on the other hand, have ~17% (11-20% and Z always > 4.5; Table S10) ancestry derived from pre-contact dogs. We also used outgroup f3 statistics to

assess the degree of shared drift between various populations available on the SNP array and PCD (Figure 2b).

To further assess this result we used ADMIXTURE (170) on a subset of the SNP array samples including all modern North American populations as well as Arctic dogs, "basal" breeds (171) and selected European and Asian populations (e.g. Boxer and Chow-Chow). K=4 was selected as the best K value based on 10 fold cross validation (Figure S19). This analysis support previous f4 ratio analysis showing that most modern North American dog populations have little pre-contact ancestry (<4%; Figure S20). ADMIXTURE, however, detects some evidence of limited PCD/Arctic ancestry in Carolina dogs ranging from 0-33% (Figure S20; population CD). Such signal might not have been detected by our F4 analysis as a result of the variable amount of ancestry in this population. This analysis also reveals an affinity between Chinook and PCD/Arctic breeds (12-15%; Figure S20). This is not surprising given that Chinook dogs are considered as Sledge dogs. With K=4, however, we cannot distinguish between PCD and Arctic dogs ancestry. This PCD/Arctic component in Carolina dogs and Chinook might therefore be the result of admixture with Arctic dogs rather than PCD. To test this we tried to separate PCD/Arctic ancestry with higher K values. Both K=10 and K=15, however, failed to differentiate PCD and Arctic dog ancestry (Figure S20) but instead differentiated New World Arctic dogs (Alaskan malamute and Greenland sledge dogs) from Old World Arctic dogs/PCD (Figure S20).

Alaskan Village dogs were the population of north American village dog with the most PCD admixture. This is not surprising as these are closely related to Arctic breeds (6, 9). The f4 ratio conducted above is thus not appropriate for this population (as it assumes close relatedness to Eurasian dogs; see Figure S18). Here we wanted to test whether these dogs have any pre-contact ancestry (interbred with pre-contact dogs). To do so we computed every possible combination of the same f4 ratio as above but using only Arctic breeds. We found that both Alaskan malamute and Greenland sledge dog have a significant amount of ancestry from PCD (~4-14%; Table S11). This however, might be the result of substructure among Arctic dogs (see below).

We assessed whether these results could be affected by the ascertainment of the SNP array by repeating the analysis above (PCD admixture fraction into Alaskan malamute and Greenland sledge dogs) using whole genome data. We used only transversions for this analysis (~600K SNPs). We found very little difference in admixture fraction (~7- 14%; Table S11) indicating that the ascertainment of the array did not introduce much bias.

We also used D-statistics on genome-wide data to test for admixture from PCD into Arctic breeds since their MRCA. As for the f4 ratio we found that both Alaskan malamute and Greenland sledge dogs have a significant amount of ancestry from PCD (Table S12). We tested whether this signal could be due to admixture from Eurasian dogs into Siberian husky dogs (making derived alleles in Alaskan malamute and Greenland sledge dogs (GSD) match PCD

more often; Table S13). We found evidence that the Siberian husky and Alaskan malamute genomes that we analysed here received gene-

flow from European dogs (Table S13). However, we found no evidence that GSD received gene-flow since the MRCA of Arctic breeds. This suggests that Eurasian admixture did not affect our result. As stated above, this could also be the result of ancient substructure within Arctic dogs.

We found, however, no signal that either Alaskan malamute or Greenland sledge dogs shared an excess of derived alleles with PCD compared with each other (D(Outgroup, AL3194, Alaskan malamute, Greenland sledge dog)=-0.0001, sd=-0.010). This shows that these dog lineages did not receive additional gene-flow from PCD since their divergence from each other. This result suggests that the signal detected above (excess shared derived alleles between American Arctic dogs and PCD) is due to ancient substructure within Arctic dogs (172). More precisely, we hypothesise that the Eurasian Arctic dogs that were recently brought into the Americas, all the way to Greenland, originated from a population that was more closely related to PCD dogs than other Arctic dogs. The high degree of mtDNA divergence within ancient Eurasian Arctic dogs from Zhokhov (~9,000 BP; Figure 1b) suggests that ancient substructure with Arctic dogs is a plausible scenario.

TreeMix In order to test the topology suggested by our phylogenetics and f3 statistics analyses we used Treemix (18) to build a tree with admixture edges. We only used 3 representatives from each major dog group:

Western Eurasian dogs - Portuguese village dogs (DEU), German Shepherd (DGS) East Asian dogs - Vietnamese village dogs (DVN) because they lack admixture from European dogs (see above) and Tibetan village dogs (DTI)

Pre-contact dogs (PCD), including both Port au Choix (AL3194) and Weyanoke Old Town (AL3223; Table S1)

Arctic dogs - Malamute (DMA) and Greenland dogs (DGL) because they seem least admixed with Western dogs (see above)

CTVT - (79T and 24T)

Eurasian wolves (WEU) from Spain and Portugal North American wolves (WAM) from Yellowstone

Coyotes (COY) as an outgroup

We only used transversions in order to limit the effect of DNA damage on the analysis and only used sites that were covered in all samples (~60,000 SNPs).

The results of these analyses are presented in Figure S21, Figure S22, Figure S23 and Figure S24. The placement of the East Asian dog (DVN) population was affected by adding admixture edges (Figure S22, Figure S23). With two admixture edges DVN

outgroup PCD/CTVT and Arctic dogs but has strong admixture into DTI. This support results from our D-statistics analysis (Table S7) and NJ analysis (Figure S10; Figure S11) that suggests that DTI is mixed with European ancestry. We also found evidence for European ancestry in Arctic dogs, supporting our D-statistics analyses (Table S4). Lastly we also found admixture from COY into PCD/CTVT, consistent with D-statistics (Figure S16).

qpGraph We used qpGraph (167) to fit admixture graphs to nine populations representing PCD, CTVT, and each of the three major dog groups, plus wolves and coyotes.

Western Eurasian dogs - Portuguese village dogs (DEU)

East Asian dogs - Vietnamese village dogs (DVN)

Pre-contact dogs (PCD), including both Port au Choix (AL3194) and Weyanoke Old Town (AL3223; Table S1)

Canine transmissible venereal tumor (CTVT), including 24T, 79T

Arctic dogs - Alaskan malamute (DMA)

Eurasian wolves (WEU) from Spain and Portugal

North American wolves (WAM) from Yellowstone

Coyotes (COY) from California

Andean Fox (OUT) as the outgroup

We only used transversions in order to limit the effect of DNA damage on the analysis. This resulted in 600,991 high quality SNPs.

To explore the space of all possible admixture graphs we implemented a heuristic search algorithm. Given an outgroup with which to root the graph, a stepwise addition order algorithm was used for adding leaf nodes to the graph. At each step, insertion of a new node was tested at all branches of the graph, except the outgroup branch. Where a node could not be inserted without producing f4 outliers (i.e. $|Z| >= 3$) then all possible admixture combinations were also attempted. If a node could not be inserted via either approach, that sub-graph was discarded. If the node was successfully inserted, the remaining nodes were recursively inserted into that graph. All possible starting node orders were attempted to ensure full coverage of the graph space.

As the number of possible graphs grows super-exponentially with each additional leaf node, we initially excluded CTVT from the search space and looked for models with fit the remaining eight populations. We fitted 480,166 unique admixture graphs for these

8 populations and recorded the 892 graphs that left no f4 outliers (i.e. |Z| < 3). We then fitted a further 309,525 unique models, testing all possible insertions of CTVT into the 892 eight-population graphs, and recorded the 1,655 graphs that left no f4 outliers.

Treemix analysis was also performed using the same nine populations, with six admixture edges (the maximum number seen in the qpGraph analyses). We chose the most plausible qpGraph model (Figure S25) by comparing all fitted models to the Treemix tree with the same sampling (Figure S26), Neighbour joining tree (Figure S10), Bayesian tree (Figure S12) and D-statistics analyses (see above).

Phenotypic information Considering the evidence of introgression between wild North American canids into the pre-contact domestic dog population (Table S3), we assessed the presence of a marker associated with melanism, which has introgressed from dogs into North American gray wolves and coyotes (168), in the higher coverage PCD genomes (Port du Choix sample: AL3194, ~2x; Weyanoke old town sample: AL3223, ~0.5x). We found no evidence for the CBD103ΔG23 / KB mutation in either of these samples.

CTVT Mutation rate analysis Overall rationale Our goal was to estimate a lower bound for the CTVT somatic mutation rate and to use this to estimate an upper range for the time at which CTVT originated. To do this, we collected biopsies from a pair of CTVT tumors involved in a naturally occurring direct transmission event, and identified mutations that had arisen during the known transmission time interval to define a somatic mutation rate. We then estimated the number of somatic mutations in the entire CTVT lineage and applied our somatic mutation rate to estimate the time of CTVT origin.

Previous estimates of CTVT time of origin have relied on microsatellite mutation rates in mammalian germlines (173, 174), or mutation rates in human cancer (142). These have led to estimates of 250 to 2,500 years since the most recent common ancestor of a group of globally dispersed tumors (173), or 6,500 to 65,000 years (174) and 10,179- 12,873 years (142) since the origin of CTVT.

Case histories Dog 609 was a mixed-breed free-ranging dog from the Gambia with an approximately 31cm3 vaginal CTVT tumor. Her ten-month-old male puppy, Dog 608, had several CTVT tumors on the ventral skin. This unusual CTVT presentation in Dog 608 suggested that CTVT cells may have transmitted from mother to puppy during parturition.

Samples This project was approved by the Department of Veterinary Medicine, University of Cambridge, Ethics and Welfare Committee (reference CR174). A 1-2 mm3 biopsy was

sampled from Dog 609's vaginal tumor (609T). A 1-2 mm3 biopsy was sampled from one of Dog 608's skin tumors on the same day (608T). Biopsies were also collected from host tissues (ovary (609H) or testis (608H)). Genomic DNA was extracted using the Qiagen DNeasy Blood and Tissue extraction kit (Qiagen, Hilden, Germany). CTVT diagnosis was confirmed as previously described (150). Whole genome sequencing libraries were prepared with insert size 450 bp and sequenced with 150 bp paired end reads using the Illumina HiSeq X Ten platform (Illumina, San Diego, CA). Reads were aligned to CanFam3.1 using BWA-MEM (132). Average sequencing depth is reported in Table S2.

Variant Calling

Variant extraction and filtering We used Somatypus (https://github.com/baezortega/somatypus), a Platypus (175) based variant calling and genotyping pipeline, to identify SNVs and small insertions and deletions (indels). In order to make an initial call, SNVs were required to have ≥3 supporting reads in at least one of the four sequenced samples (608T, 608H, 609T, 609H). Indels were inputted to GATK Realigner Target Creator (176) for local realignment and SNVs were re-called from realigned genomes.

The following in-built Platypus flags were used to exclude SNVs at two stages, before and after genotyping: badReads, MQ, QD, strandBias, SC.

The following post-processing filters were also implemented:

Strand bias filter. For each SNV, the total coverage, as well as forward and reverse strand read support were extracted. For low total coverage positions (≤10 reads across all four samples), we discarded calls with less than two supporting reads in either forward or reverse direction. For high total coverage positions (>10 reads across all four samples), we discarded calls with less than 20% support on either the forward or reverse sequencing strands.

Simple repeat filter. SNVs within simple repeats, as defined by the UCSC table browser (CanFam3.1), were excluded.

Extreme depth filter. SNVs within regions of high read depth were also excluded. To detect high read depth (HRD) regions we first generated BigWig coverage files from matched normal whole genome sequence data files (608H, 609H). We then identified areas with coverage 12 standard deviations higher than the mean read coverage, on a chromosome by chromosome basis. Common intervals between normal samples were identified using bedtools multiinter (141) and were merged using bedtools merge. The maximum allowed distance between regions to be merged was 250 bp. HRD regions spanning less than 500 bp were excluded. Any HRD region that overlapped with gene regions as defined by the UCSC table browser (CanFam3.1, Genes and Gene Predictions, Ensembl Genes) was excluded.

Low VAF filter. SNVs with VAF>0 and VAF<0.2 in both 608H and 609H were discarded if (i) they were not detected in 608T or 609T or (ii) they were found with VAF>0 and VAF<0.1 in either or both of 608T and 609T.

Regions filter. SNVs occurring in the mitochondrial genome or on unassigned scaffolds were excluded. In addition, to avoid problems caused by variable coverage in hosts, SNV analysis was restricted to autosomes.

Germline and consensus filtering SNVs identified in 608T and 609T will belong to one or more of the following categories:

(i) Contaminating germline SNVs from matched host (ii) Germline SNVs inherited by the CTVT founder dog (iii) Somatic mutation SNVs

In order to enrich for somatic mutations, we filtered our candidate SNVs against a panel of 28,812,954 canid germline SNVs. Specifically, we excluded any genomic site that was reported in any of the following variant catalogues:

- 608H and 609H (sites with ≥5 reads coverage and ≥2 reads supporting a non-reference allele were considered SNVs)
- The Variant and Systematic Error Catalogue (VSEC) (177)
- The CanineHD 170K SNP array (178)
- The ascertainment panel generated in this study prior to incorporating CTVT samples (Genotyping - Ascertainment panel)
- A complete genome from a Greenland sledge dog (14)(Table S2) included in the ascertainment panel was additionally genotyped. This provided additional SNVs beyond those in the ascertainment panel, as the ascertainment panel excluded SNVs that were found on only one chromosome; thus SNVs that were exclusively found in the Greenland sledge dog individual and CTVT would not have been included in the ascertainment panel that we filtered against (see previous bullet point), but were excluded in this step (Genotyping - Dog samples)(14)

Next, we further filtered the remaining SNVs using the following criteria:

- We retained only those SNVs that had that had ≥2 reads supporting the variant, all with minimum base quality of 20 and minimum mapping quality of 35, in at least one of the two tumors using the alleleCount package (http://cancerit.github.io/alleleCount/).
- We retained only those SNVs that were identified by GATK Haplotype Caller. The GATK engine was restricted processing candidate loci.

● We required that the matched host must have coverage of at least 20 reads total at the candidate SNV position. Candidate SNVs that did not reach this threshold in one or both matched hosts were discarded.

● We discarded SNVs where one or both matched hosts had ≥10 reads total (regardless of whether they supported the variant) with base quality <20 and mapping quality <35 in matched hosts

1,934,103 and 1,934,125 "tumor-only" SNVs remained in 608T and 609T respectively after these steps; 1,933,897 of these were shared by 608T and 609T. Of the SNVs in this set that mapped to genomic regions retaining both parental copies, almost all SNVs were heterozygous. Thus, the majority of these SNVs are likely to be somatic; however, some germline variation that was present in the CTVT founder dog, but that is not represented in the germline panel used here likely still remains. It is also likely that some somatic mutations, which occurred in the same sites as germline SNVs represented in our panel, have been removed.

Tumor-unique SNVs

We next filtered tumor-only SNVs, as defined above, for those unique to either 608T or 609T. In order to be considered unique to a single tumor, a variant was required to be present with ≥2 supporting reads in only one tumor, with minimum base quality of 20 and minimum mapping quality of 35 for those reads supporting the variant.

This method yielded 206 tumor-unique SNVs in 608T and 228 tumor-unique SNVs in 609T.

Mutational spectrum Each tumor-only SNV was classified as one of six possible mutation types in the pyrimidine context (C>A, C>G, C>T, T>A, T>C, T>G). The immediate 5' and 3' sequence contexts for each mutation was extracted from the CanFam3.1 dog reference genome (179) yielding 96 mutation types. The mutational spectrum for the 1,933897, CTVT tumor-only SNVs (shared between 608T and 609T) is shown in Figure S27A.

Mutational signature fitting We performed mutational signature fitting in order to estimate the number of mutations contributed by different exposures to the CTVT mutational spectrum.

Validated mutational signatures were obtained from the Catalogue of Somatic Mutations in Cancer (COSMIC; http://cancer.sanger.ac.uk/cosmic/signatures) database and renormalized to the CanFam3.1 dog reference genome (179). In addition, we generated a "Dog Germline" signature, from the germline mutational spectrum of a Greenland sledge dog (Variant calling - Germline and consensus filtering; Table S2).

Reference (142) previously showed that COSMIC mutational signature 1 (5-methyl-cytosine deamination), signature 5 (unknown etiology), and signature 7 (ultraviolet light exposure) are operative in CTVT, and that these three signatures are sufficient to describe

the pattern of somatic substitutions observed in CTVT. We therefore fitted these three signatures, together with the Dog Germline signature (see above), to the CTVT tumor- only mutation spectrum (Table S14; Figure S27B and C). Results are similar to previous findings (142). Signatures were fitted using sigfit (https://github.com/kgori/sigfit). Simulations were run using 100 chains with 10,000 iterations each. Importantly, the Dog Germline signature accounted for only 5.5% of the tumor-only SNVs, suggesting that the majority of the SNVs in this set are indeed somatic.

N[C>T]G CTVT tumor-only SNVs Mutational signature 1 is largely composed of 5'-N[C>T]G-3' mutations (where N is any base) (http://cancer.sanger.ac.uk/cosmic/signatures). Of the 1,933,897 tumor-only SNVs shared by 608T and 609T, 222,072 are N[C>T]G.

Copy-number analysis Average mappability and GC-content were generated for the dog reference genome (179) with the generateMap, mapCounter, and gcCounter tools in the HMMcopy package (180). GC content and genomic mappability biases for read counts in non-overlapping 1kb windows were corrected using HMMcopy. Copy number estimation was then performed on GC- and mappability-corrected read counts using a bespoke copy number calling pipeline (https://github.com/ymk1/cnv_pipeline.git)

Tumor purity in 608T and 609T was evaluated based on the VAF distribution of tumor-only SNVs. Tumor purity was estimated as follows:

Purity=2*VAFmed Where VAF

med

is the median VAF value of tumor-only SNVs. Using this method, 608T was estimated to be 49.3% CTVT cells and 609T was estimated to be 65% CTVT cells.

Identifying clonal mutations in tumor-unique variant sets We categorised tumor-unique SNVs in 608T and 609T as either clonal or subclonal, that is, present in all or a fraction of tumor cells within a sample, respectively. To do this, we first examined the VAF distributions of germline SNPs in 608T and 609T for each copy number (CN) state (CN1, CN2, CN3, CN4, CN6). We used a Gaussian mixture model (k = 2), implemented using the R package MCLUST (181), to define VAF clusters for heterozygous and homozygous SNPs. Next, we fitted this model to VAF distributions of tumor-unique SNVs. SNVs that fell below the 5% lower bound were defined as subclonal; all other SNVs were considered clonal. The results of this analysis are shown in Table S15.

N[C>T]G CTVT tumor-unique clonal SNVs Of the 183 and 174 clonal mutations identified uniquely in 608T and 609T respectively, 27 and 23 were N[C>T]G in 608T and 609T respectively. 26/27 and 21/23 (609T) were validated using read alignment visualisation.

CTVT mutation rate

We have determined that 608T and 609T acquired 183 and 174 clonal mutations, and 27 and 23 clonal N[C>T]G mutations respectively since they diverged from their most recent common ancestor (MRCA).

In order to estimate the CTVT mutation rate, we need to know the time intervals during which the clonal tumor-unique mutations arose in 608T and 609T. These time intervals (i

608T

and i

609T

) correspond to:

$$i_{608T} = t_{MRCA\text{-}608T} - t_{MRCA\text{-}608T/609T}$$

and

$$i_{609T} = t_{MRCA\text{-}609T} - t_{MRCA\text{-}608T/609T}$$

where $t_{MRCA\text{-}608T}$ and $t_{MRCA\text{-}609}$ are time-points defining the MRCA cells of 608T and 609T respectively, and $t_{MRCA\text{-}608T/609T}$ is the time-point defining the MRCA cell of both 608T and 609T.

We assumed that $t_{MRCA\text{-}608T/609T}$ occurred during 609T tumor development; i.e. the clones

that spawned the sampled 608T and 609T biopsies diverged in the period after infection of Dog 609 (the mother) but before transmission to Dog 608 (the son). Thus, the earliest time-point for $t_{MRCA\text{-}608T/609T}$ would coincide with the time at which Dog 609 (the mother) was infected with CTVT, i.e. month 0.

We assumed that Dog 609 was infected during the heat cycle in which she conceived the puppy, Dog 608. Although we cannot be certain that this assumption is valid, we observed that Dog 609's tumor appeared to be of a similar size to Dog 608's tumor. Unless CTVT tumors have large variation in growth rate, we believe that it is unlikely that Dog 609 was infected with CTVT in the heat cycle previous to that in which she conceived Dog 608. Given that the gestation period in domestic dogs can range from 57-72 days (182–184), we estimated that Dog 609 was infected with CTVT approximately 2 months prior to when her son, Dog 608, was born and infected. This implies that the latest time-point for $t_{MRCA\text{-}608T/609T}$ is 2 months after Dog 609 was infected with CTVT. Thus $t_{MRCA\text{-}608T/609T}$ = 0 to 2 months, defining month 0 as the month at which Dog 609 (the mother) was infected with CTVT.

Assuming no polyclonal seeding, $t_{MRCA\text{-}608T}$ either occurred within Dog 608, or was the cell that transmitted from Dog 609 to Dog 608, and thus $t_{MRCA\text{-}608T}$ = 2 to 12 months, where month 0 is the month at which Dog 609 (the mother) was infected with CTVT, and 12 months corresponds to the time of sampling.

$t_{MRCA\text{-}609T}$ could have occurred at any time during 609T tumor development. Thus $t_{MRCA\text{-}609T}$ = 0 to 12 months months, where month 0 is the month at which Dog 609 (the mother) was infected with CTVT, and 12 months corresponds to the time of sampling.

Thus, we estimate both i_{608}

and i

609

to be 0 to 12 months.

We determined that 608T and 609T had acquired 183 and 174 mutations since their divergence from their MRCA (MRCA

608T-609T

) and before the MRCA of the clone biopsied in 608T (MRCA

608T

) and the MRCA and the clone biopsied in 609T

). (MRCA

609T

These mutations will likely have arisen as part of clock-like ageing- associated mutational signatures 1 and 5 and possibly as part of mutational signature 7 (exposure to ultraviolet light). Signature 1 mutation rate is believed to be highly dependent on cell division (185). Due to the small number of tumor-unique mutations, signature fitting cannot give us an accurate estimate of the respective contributions of these signatures to tumor-unique SNV sets. As mutational signature 1 is largely composed of N[C>T]G mutations (where N is any base), we used N[C>T]G as a proxy for signature 1. Table S16 shows the number of N[C>T]G mutations unique to 608T and 609T, as well as the number found in the somatic lineage from the CTVT founder dog until MRCA

608T-609T

.

As 608T harbours more clonal N[C>T]G mutations than 609T (Table S16), we infer that MRCA

608T

existed more recently than MRCA

609T

. Assuming that i

608T

is up to 12 months (see above), then the slowest rate at which N[C>T]G mutations accumulate is 27 N[C>T]G mutations / year or 12.56 N[C>T]G mutations / Gigabase (Gb) / year, usingthe callable dog genome size (excluding simple repeats and regions of high read depth) of 2.15 Gb (179). Applying this rate to the whole lineage (222,072 N[C>T]G mutations across the callable genome ~ 2.15 Gigabase pairs, see above), we obtain an upper bound of 8,225 years for the origin of CTVT. Our method cannot directly infer an upper bound for the CTVT mutation rate, and hence a lower bound for the age of CTVT. However, assuming that the disease described by Blaine in 1810 (186) was indeed CTVT, then CTVT must have arisen at least 200 years ago.

Comparison with mutation rates in human cancer Signature 1 mutation rate varies between human cancer tissue types(185). The mutation rate lower bound that we have derived for CTVT N[C>T]G mutations (>12.56 mutations / Gb / year) is comparable to the N[C>T]G mutation rates found in human cancers(185). Cervical cancer was reported to have the highest estimated rate of accumulation of N[C>T]G somatic mutations of 36 human cancer types (16.61 N[C>T]G somatic mutations / Gb / year)(185). Applying the cervical cancer N[C>T]G mutation rate to the CTVT lineage would provide an estimate of 6,195 years since CTVT origin.

Limitations of the approach The approach to deriving the CTVT mutation rate that we have presented here is based on a number of assumptions. These are outlined below:

● No polyclonal seeding. We have discounted the possibility of polyclonal seeding of Dog 609's tumor. If polyclonal seeding occurred, then t

MRCA-608T/609T

may have existed in Dog 609 tumor's donor, rather than in Dog 609's tumor itself. If this is the case, then the mutation rate would be slower, leading to older estimates for CTVT time- of-origin.

● Age of Dog 608. Dog 608 was estimated to be ten months old at the time of sampling. Given that Dog 609 was approximately 14 days pregnant at the time of sampling, and given that dogs have heat cycles every six months, we believe that Dog

608's age estimate is likely accurate. However, an inaccurate age estimate could affect our mutation rate estimates.

● Time of infection of Dog 609. We assumed that Dog 609 (the mother) was infected with CTVT at the time of the heat cycle during which she conceived Dog 608. However, if Dog 609 was in fact infected with CTVT during a previous heat cycle, then t

MRCA-608T/609T

may have existed at an earlier time point: this would lead to estimation of a slower mutation rate and older estimates for CTVT time-of-origin.

● Estimation of total mutation burden. We estimated the total mutation burden by filtering against a large panel of variation in normal dogs (see Germline and consensus filtering). If our set of tumor-only SNVs (see section tumor-unique SNVs above) is substantially over-filtered (i.e. somatic mutations were removed as they occur at the same site as germline SNVs), then CTVT could have arisen earlier than our estimates suggest. If, on the other hand, substantial numbers of germline SNVs remain in the set of total mutations, then we may have over-estimated the time of CTVT origin.

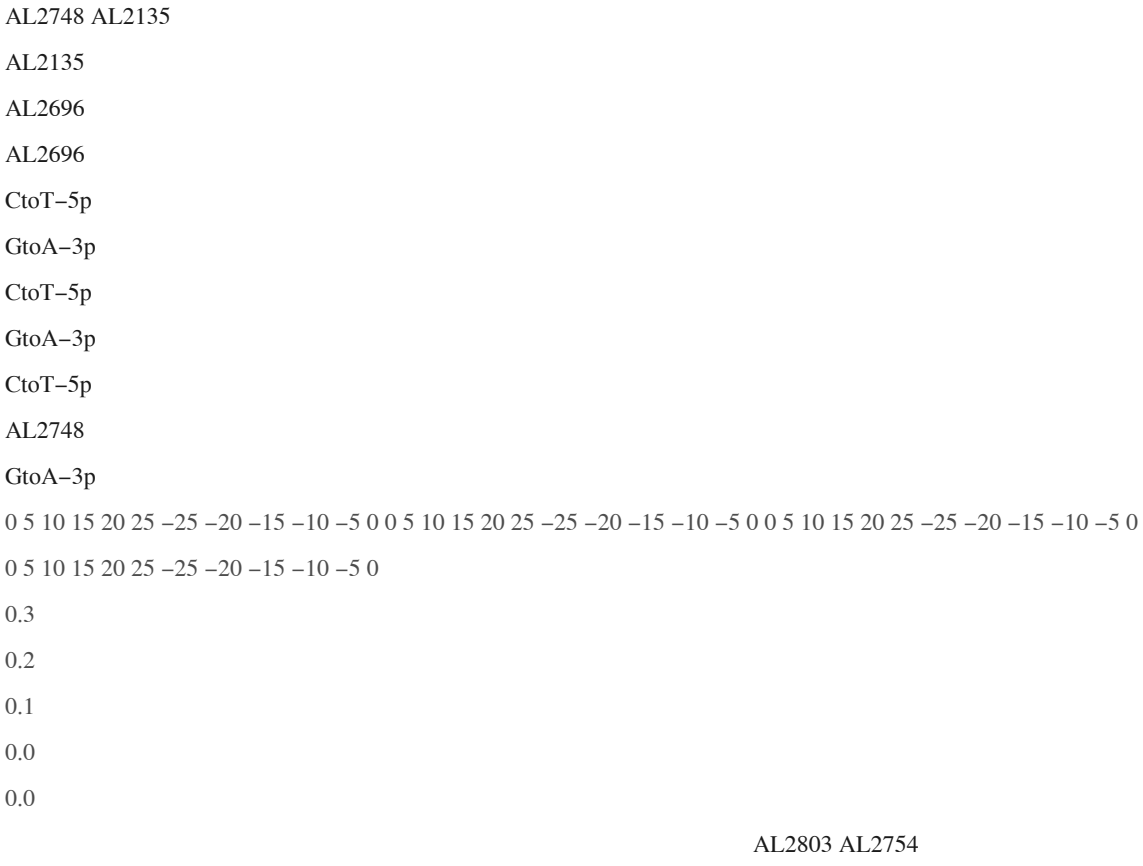
● Back-mutation. We discounted back-mutation as a significant factor in our estimates.

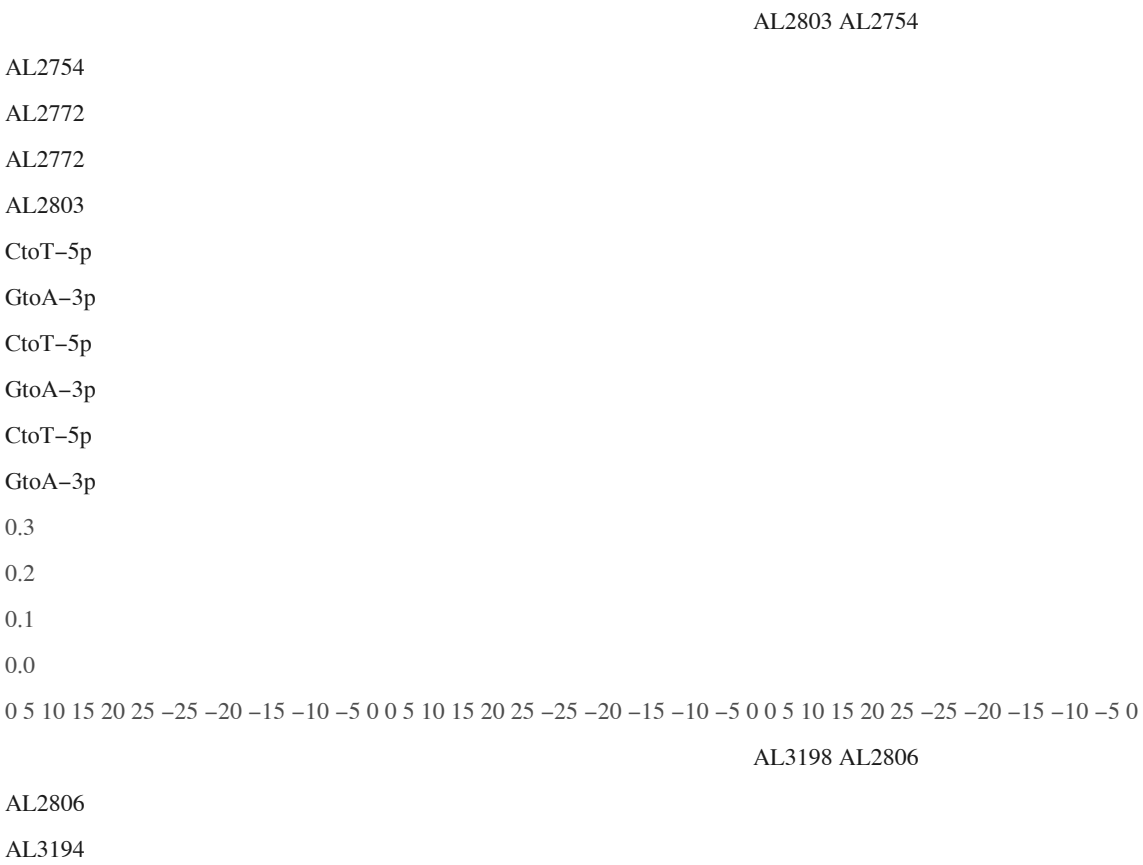
● Mutation false discovery rate. The expected proportion of false positive SNVs should be the same in the tumor-unique and tumor-only variant sets as we have used the same filters in both cases. Thus, this is unlikely to have substantially affected estimates.

● Mutation opportunity and variable mutation rate. We have discounted the effects of variable mutation opportunity from our estimates. Mutation opportunity may change over time due to (i) altered DNA methylation or chromatin states; (ii) decrease in number of available NCG sites over time; (iii) copy number alterations. Furthermore, although N[C>T]G is usually considered constant, a recent study has detected germline mutations in the MBD4 gene which alter the rate of signature 1 mutation accumulation (187). As we detected a number of mutations in MBD4, we cannot exclude the possibility that somatic alteration of this or other loci in CTVT may have caused variation in the rate of N[C>T]G mutation accumulation.

● Sampling error. We based our estimates of the CTVT mutation rate on one observation and did not account for sampling variation. Future studies can address this by measuring mutations in additional CTVT time intervals.

Despite these limitations, our analysis provides a plausible estimate of CTVT somatic mutation rate, and is comparable with clock-like mutation rates observed in some human cancers (185).


AL2748 AL2135
AL2135
AL2696
AL2696
CtoT−5p
GtoA−3p
CtoT−5p
GtoA−3p
CtoT−5p
AL2748
GtoA−3p
0 5 10 15 20 25 −25 −20 −15 −10 −5 0 0 5 10 15 20 25 −25 −20 −15 −10 −5 0 0 5 10 15 20 25 −25 −20 −15 −10 −5 0
0 5 10 15 20 25 −25 −20 −15 −10 −5 0
0.3
0.2
0.1
0.0
0.0


AL2803 AL2754
AL2754
AL2772
AL2772
AL2803
CtoT−5p
GtoA−3p
CtoT−5p
GtoA−3p
CtoT−5p
GtoA−3p
0.3
0.2
0.1
0.0
0 5 10 15 20 25 −25 −20 −15 −10 −5 0 0 5 10 15 20 25 −25 −20 −15 −10 −5 0 0 5 10 15 20 25 −25 −20 −15 −10 −5 0

AL3198 AL2806
AL2806
AL3194

AL3194

AL3198

y c n

0.3

CtoT−5p

GtoA−3p

CtoT−5p

GtoA−3p

CtoT−5p

GtoA−3p

e u q e

0.2

0.1

CtoT−5p

r F

GtoA−3p

0 5 10 15 20 25 −25 −20 −15 −10 −5 0 0 5 10 15 20 25 −25 −20 −15 −10 −5 0 0 5 10 15 20 25 −25 −20 −15 −10 −5 0

AL3202

AL3202

AL3223

AL3223

AL3226

AL3226

CtoT−5p

GtoA−3p

CtoT−5p

GtoA−3p

CtoT−5p

GtoA−3p

0.3

0.2

0.1

0.0

0 5 10 15 20 25 −25 −20 −15 −10 −5 0 0 5 10 15 20 25 −25 −20 −15 −10 −5 0 0 5 10 15 20 25 −25 −20 −15 −10 −5 0

AL3231

AL3231

CtoT−5p

Fig. S1. Per library C to T (red) and G to A (blue) frequency of mis-incorporation at 3’ and 5’ end of read for samples used in nuclear genome analyses.

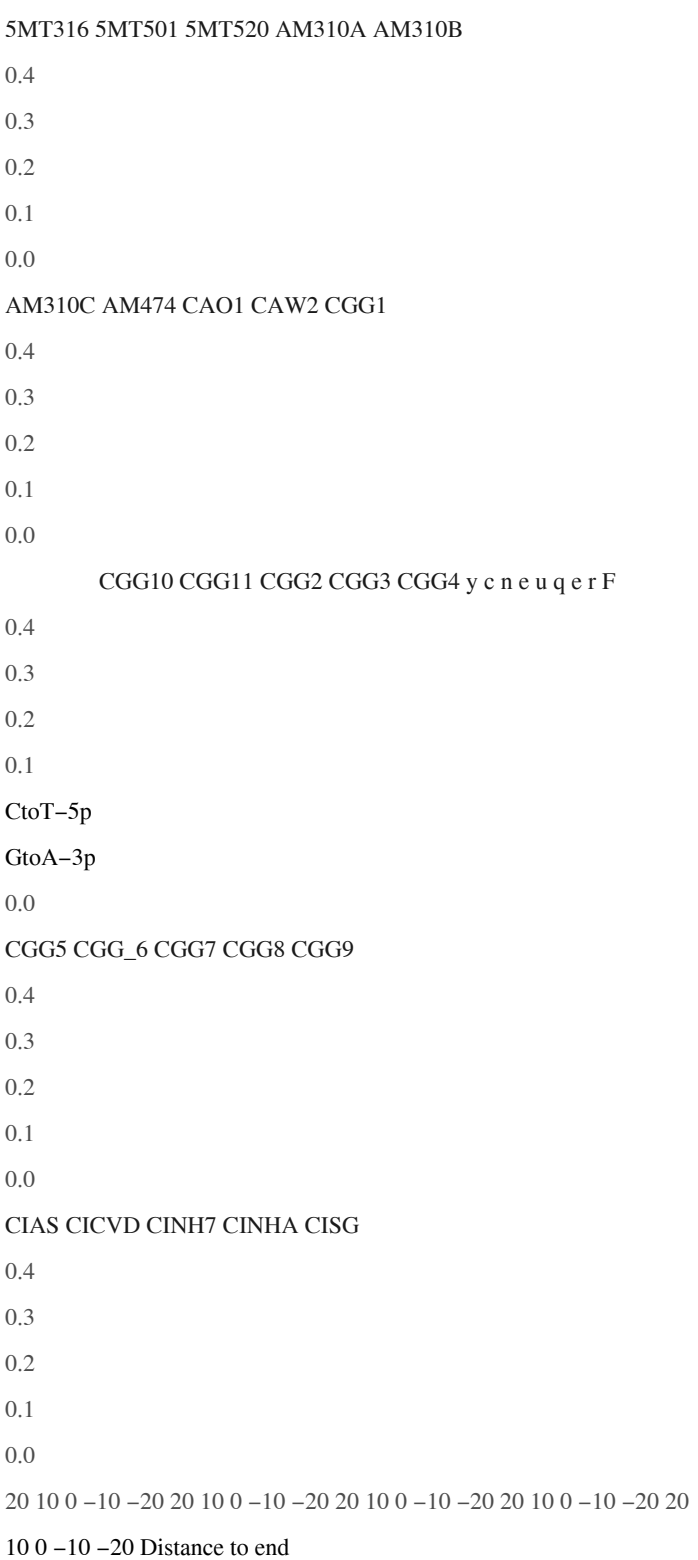


Fig. S2 Per library C to T (red) and G to A (blue) frequency of mis-incorporation at 3’ and 5’ end of read for samples used in mtDNA analyses. Lack of 5’ damage in some libraries is due to

library preparation protocol (see Ancient DNA - Illinois section).

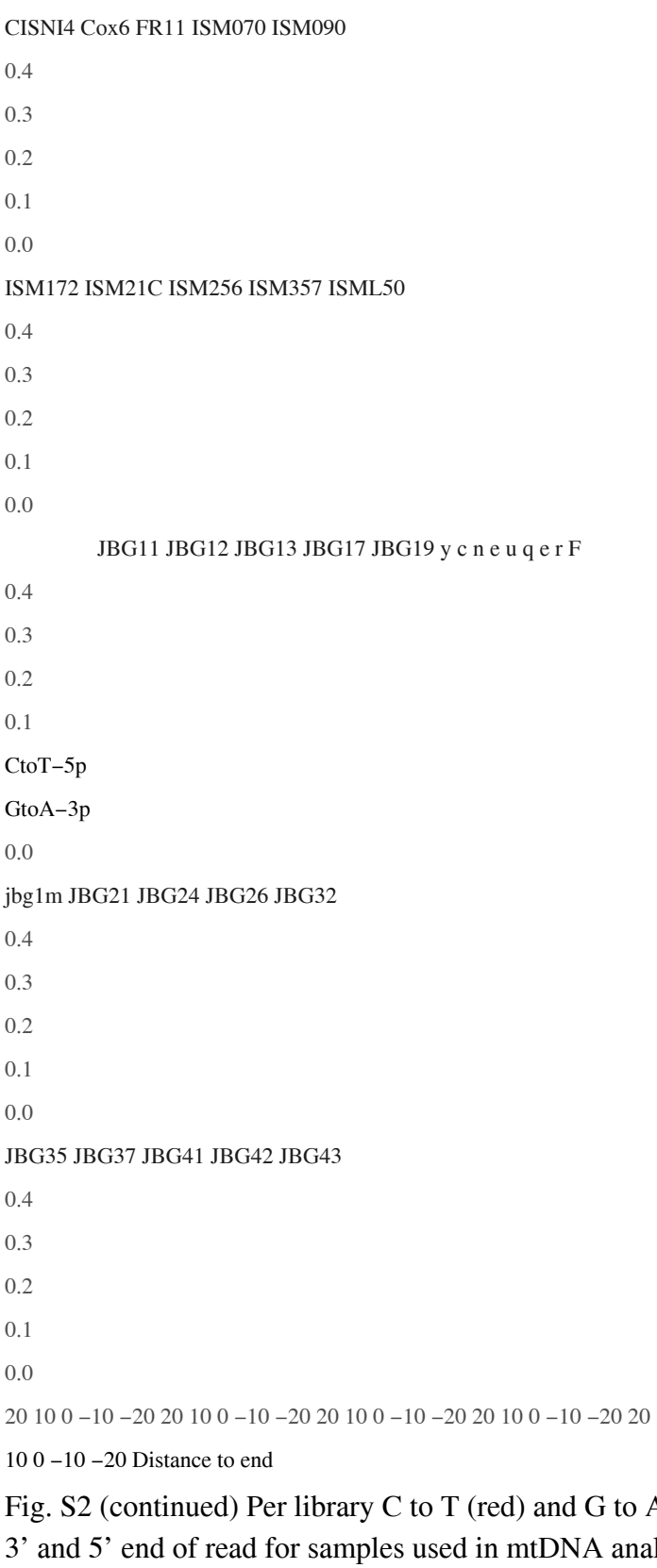


Fig. S2 (continued) Per library C to T (red) and G to A (blue) frequency of mis-incorporation at 3’ and 5’ end of read for samples used in mtDNA analyses. Lack of 5’ damage in some libraries

is due to library preparation protocol (see Ancient DNA - Illinois section).

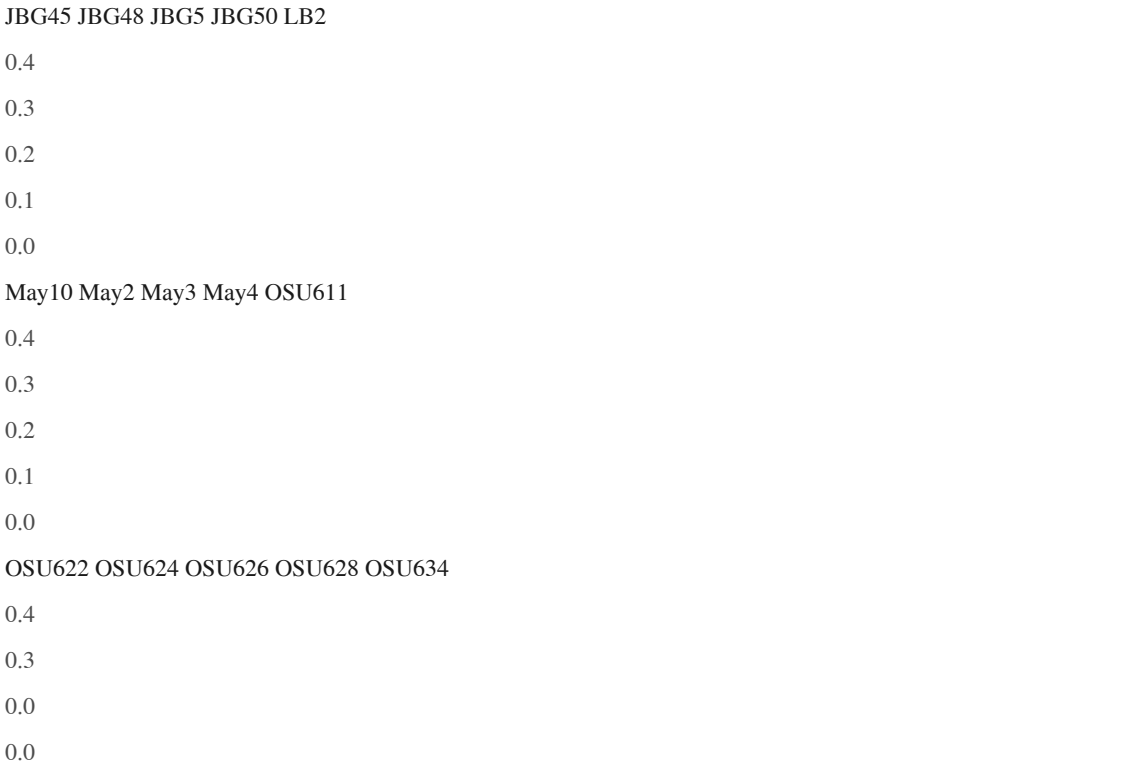
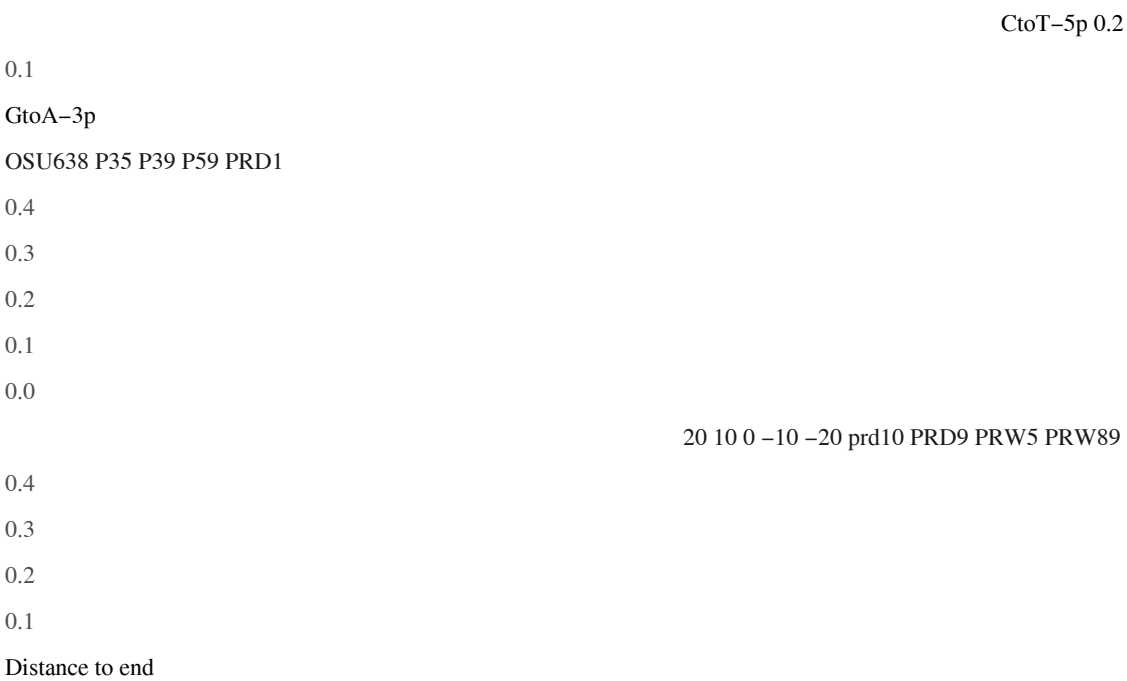


Fig. S2 (continued) Per library C to T (red) and G to A (blue) frequency of mis-incorporation at 3' and 5' end of read for samples used in mtDNA analyses. Lack of 5' damage in some libraries is due to library preparation protocol (see Ancient DNA - Illinois section).



0.0050

Belgium_30_000 Alaska_8_000 Switzerland2_4_500 Haplogroup A Dingo_A CGG10 CGG11 D01_A D61_A CGG5 CGG3 CGG1 CGG8 CGG7 CGG4 CGG2 CGG9

AM474

FR11

Cox6

ISM090

prd10

Canada6

PRW89 Haplogroup B

USA1

USA2

Alaska4 Canada2

Alaska_1_000 Russia1

Oman

Poland1 Haplogroup D

Switzerland3_4_500

Russia_18_000 Belgium_36_000

China2

100 77

38

51

100

99

100

33

91 1

78

100

99

100

99

98

100

67

26 63 39

66

45

33

61 15

42

92

61

26

90

58

66

76

88

88

99 59

17

100

89

64

26

98 84

53

87

11

54

99

12

69

100

100

96

100

78

94

100

100

75

73

17

99

98

82

51

92

100 100

100

CGG_6 jbg1m AL3226 USA_1_000_A ISM070 AL3194 AL3223 5MT501 AM310A ISM357 USA_8_500_A ISM256 5MT316

CAW2

LB2

JBG32

JBG5

JBG50

CINHA

Haplogroup C

Canada5

USA3

China3

Switzerland1_4_500

Japan

Belgium_26_000

98

26

56

97

100

100

Argentina_1_000_A May2 May3 May4 May10 JBG13

Canada1 Canada3

Alaska2

Alaska6

Sweden3

coyote1

93

JBG17

5MT520 CAO1 CIAS

Canada7

Ukraine

Mexico1

53

100

61

71

69

17

100

CISG P59

P35

Israel2

Iran

45

87

61

AL2772

OSU611 OSU634 OSU626 OSU624

JBG48

JBG26 JBG11

PRD9

Spain

Poland2

India

China4

Alaska_0_800

24

52

46

64

0

67

87

100

99

OSU628

OSU638

AM310B AM310C

JBG12

92

JBG35

100 2

93

JBG45

JBG37

69

99

77

11

78

JBG42

92

88

62

Canada8

100

100

85

78

86

JBG43

0

JBG21

39

68

Saudi_Arabia1

27

JBG19

27

10

JBG24 JBG41

OSU622 AL3198 100

99

PRD1

92

CICVD CINH7

Canada10

22

87

Canada11 USA4 Alaska1 Alaska3

58

Canada9

100

Sweden1

100

47

21

10

Alaska5 90 47

32

Canada4

96

100

30

17

Russia2

44

Sweden2 Russia3 Croatia

70

100

87

Israel1

Mongolia

Mexico2

100

100

Finland Russia_15_000 Russia_22_000

82

Saudi_Arabia2

Russia_3_500 Italy

100

100

China1

38

Fig. S3 Maximum likelihood tree based on mtDNA data. The four major dog haplogroups are indicated: A (red; includes all but one pre-contact dogs), B (purple), C (yellow), D (green). Blue tip label represent newly sequenced samples (this study). Dark blue highlighted clade represents American dogs (monophyletic, bootstrap support value=87). Light blue highlighted clades (CGG1-10) represent Zhokhov Island samples (~9Kya sled dogs from Eastern Siberia; see Table S1). CGG10-11 (outside of the Zhokhov / pre- contact clade) are more recent sled dogs from Siberia (~1.5kya; Table S1). Node labels indicate bootstrap replicates.

36 32

74

40

11 19 47

0

2

80 52 56

18

4 33

0

1

16

46

1

0

0

0

0

0

6

2

44

2

0

3

20 0

1

7

36

1

0

18

65

21

0 0 0

0

0 5 0 1

16

4

42 5 6 0 4 7

21

15

4 20

0

8

21

1 14 0 0 64

97

8

0

0

0

12

0

2

1 0

3

21

0

4 2

0

0

13

12

30

0 2

35

36

27

0

5

37

45

61

82

42

41

22

2

84

13 4 46 35 63 0

0

0

4

0 0

0 5

16

6

52

29

62

0

1

23

0

0 0

0 0

13 4 49

37

0

8

28

0

25 40 9

1

3

0

0 0 0 0

0

0

8

57

2

1

0

0

29

0

0

27

20

0

1 0 2

54

0

7 24

28 0

7

100

12

0

29

30

1

0

0 0

21

0

0

12

6

0

36

36

0

0

0.02

1

Mongolia Israel1 India China4 China3 Spain Israel2 Alaska2 USA3 USA1 Alaska5 USA2 Sweden3 Sweden2 Poland2 Russia2 Alaska4

99

Canada4 Canada2 Alaska6 China2 China1 Russia3 HQ452435 D44_B

5

D26_B D48_B D33_B D42_B D40_B

16

D35_B D41_B D49_B Ukraine Sweden1 D98_B D51_B D31_B D99_B

37

AY163889.1 D47_B D32_B HQ452436 D46_B D50_B PRW89 Canada3 USA4 Canada5 Canada7 Canada1 Canada6 Canada11 Canada10 Canada9 Alaska3

98

Alaska1 Mexico1 Mexico2 Finland Russia1

D52_C

ChineseDog1_C

HQ452437

D84_C

Saudi_Arabia2

Switzerland2_4_500

AM310B JBG45

Cox6 0

OSU628

jbg1m FR11

AY163891.1

CINH7

AM474 JBG17 CISG CAO1

CIAS LB2

CGG9

5MT316

5MT501 AL3223 AM310A D22_A

0

AY163883.1 D78_A D15_A D66_A

D04_A

May4 May2

ChineseDog3_A

D73_A

D71_A

D25_A

D16_A AL3194 AY163896.1

D85_A D83_A D27_A

PRD9 64

prd10 D75_A

0

JBG12

JBG35 JBG43

JBG48

AY163885.1 AY163880.1 AY163884.1

AY163881.1

D90_A

Russia_18_000 76

Belgium_30_000

7

100

Alaska_0_800 11

Germany_2_500_C 32

D09_C

D63_C D60_C

73

D11_C D12_C D55_C HQ452438 48

Russia_22_000

Japan

JBG24

JBG19 OSU624 OSU634 JBG26

JBG41 JBG21

7

OSU626

CGG10 Dingo_A

ISM357

AY163892.2

HQ452428

5

D30_A

D67_A

0

17 0

JBG11 1

JBG32 AL2772 1

OSU622 CGG11 D01_A

10

CAW2

CGG1

D59_A

AY163894.1

AY163886.1 HQ452425

24

AY163887.1

2

0

D07_A

24

D69_A

D94_A

0

JBG42

JBG50

JBG5

AM310C Basenji_A

P59

Argentina_1_000_A

D93_A Alaska_8_000

16

0

CICVD

AY163888.1 JBG13

0

52

84

CGG2 CGG_6

0

USA_8_500_A ISM256 CGG3 CGG7

CGG5

CGG4

CGG8 USA_1_000_A

HQ452424

16

49

D21_A D61_A D79_A D81_A

26

D103_A D28_A

D68_A

1

D65_A

D82_A D56_A 0

D89_A AY163890.1

5MT520

0

92

May3

6

May10

HQ452430 AY163895.1 HQ452429

0

ChineseDog2_A D72_A

ISM070

D88_A

HQ452427

26

D18_A CINHA

PRD1

D95_A D87_A D80_A

D86_A D97_A D96_A

D91_A ISM090

5

OSU638 JBG37 OSU611

HQ452426

P35 AY163882.1 AY163879.1

AY163878.1

AY163893.1

D05_A

Belgium_26_000 Belgium_36_000

Switzerland3_4_500 Italy D03_D D34_D Poland1 Russia_3_500 Switzerland1_4_500 Saudi_Arabia1 Oman Iran Canada8 Croatia coyote1

Fig. S4 Maximum likelihood tree based on 605bp of the control region. All samples starting with prefix HQ were obtained from (5) while all samples starting with prefix AY were obtained from (8). The four major dog haplogroups are indicated with different branch colours: A (red; includes all but one pre-contact dogs), B (purple), C (yellow), D (green). All pre-contact dogs from this study are highlighted in light blue. Node labels indicate bootstrap replicates.

0.0050

CGG1 94

59

52

CGG4

46

64

CGG9

CGG2

CGG8

CGG7 OSU622

P59

53

74

OSU611

JBG12

AM310C

P35 LB2

JBG21

JBG43

11

OSU624

JBG50

5MT316

61

AM474

41

71

May3

87

AM310A

32

97

CINH7

CICVD

PRD1

PRD9

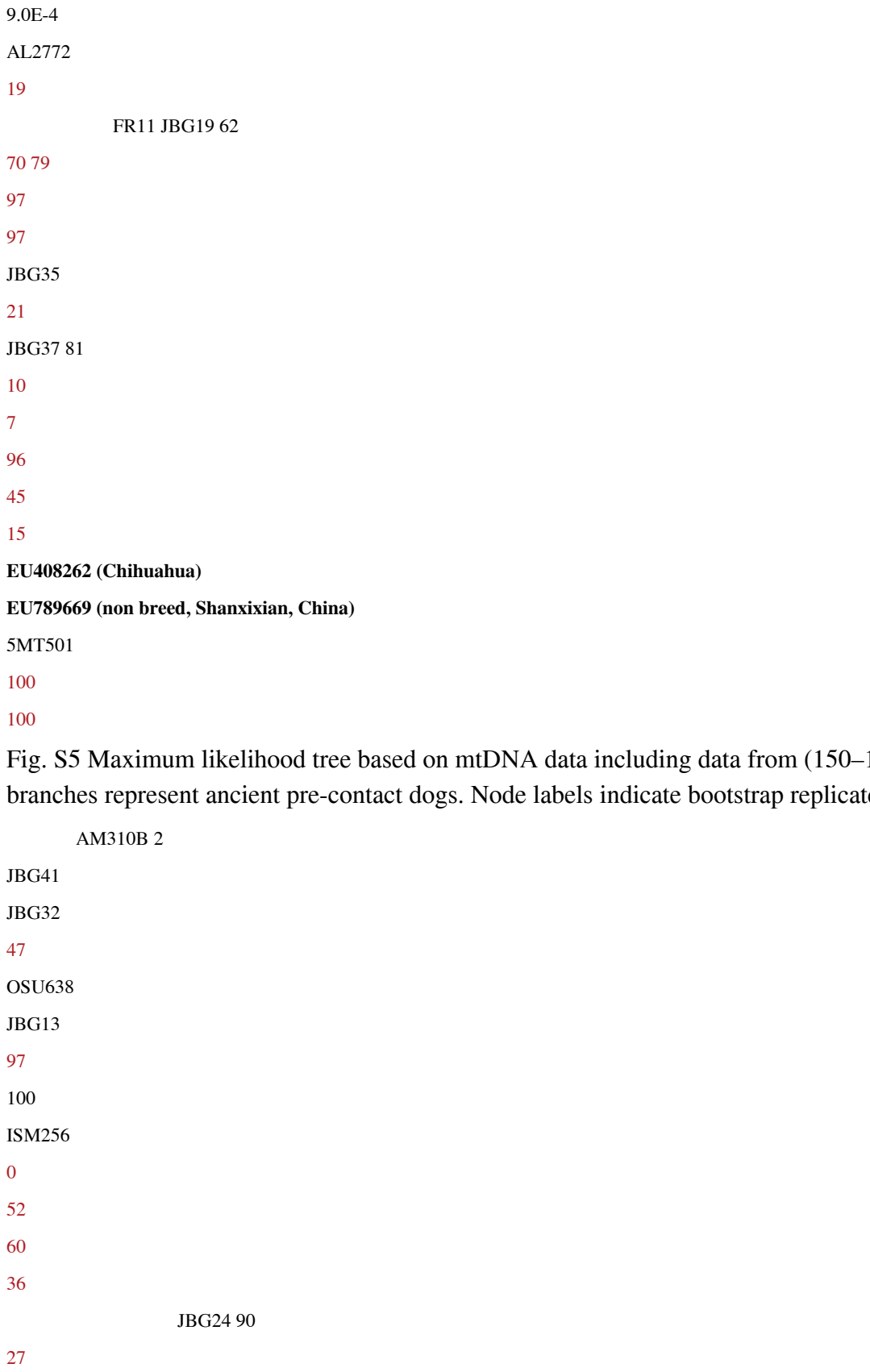


Fig. S5 Maximum likelihood tree based on mtDNA data including data from (150–155). Red branches represent ancient pre-contact dogs. Node labels indicate bootstrap replicates.

0

21 71

JBG26

0

OSU634

22

7

21

72

ISM172

64

0

72

50

JBG11

75

JBG42

63

67

100

48

68

0

JBG48 OSU628 67

OSU626

3

34

JBG45

JBG5

80

ISM090

Cox6

19

jbg1m

CISG

JBG17

51

Argentina-1-000-A

98

96

May10

69

KU291094 (Terrier cross, San Juan del Sur, Nicaragua)

EU789664 (non breed, Laem Ngop, Thailand)

USA-8-500-A 12

86

AL3223

46

ISM070 AL3194

AL3198 CINHA

prd10

AL3226

CAW2

99

85

48

31

CIAS

5MT520 CAO1

May2

EU789755 (Japanese Spitz)

USA-1-000-A

May4

99

ISM357

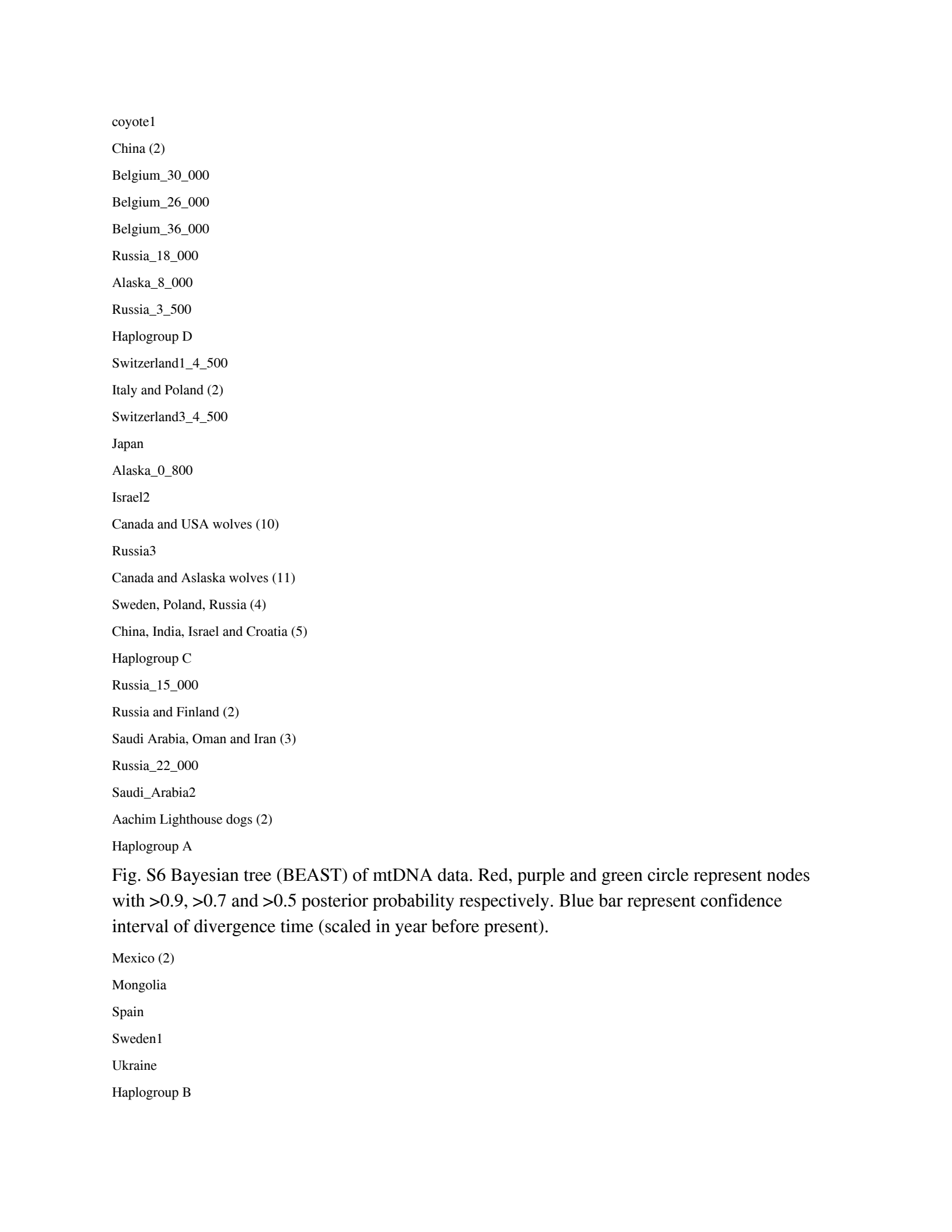


Fig. S6 Bayesian tree (BEAST) of mtDNA data. Red, purple and green circle represent nodes with >0.9, >0.7 and >0.5 posterior probability respectively. Blue bar represent confidence interval of divergence time (scaled in year before present).

Alaska_1_000

Switzerland2_4_500

Zhokhov dogs (2)

Zhokhov dogs (5)

Pre-contact dogs (75)

400000 300000 200000 100000 0

Fig. S7 Principal Components Analysis (PC1 versus PC2) of 57 canid samples (including wolves and coyotes) based on 2,063,129 SNPs ascertained using the genome-wide data-set. All pre-contact dog samples were projected.

Fig. S8 Principal Components Analysis (PC1 versus PC2) of 44 dog samples (excluding wolves and coyotes) based on 2,063,129 SNPs ascertained using the genome-wide data-set. All pre-contact dog samples were projected.

Fig. S9 Principal Components Analysis (PC1 versus PC2) of 44 dog samples (excluding wolves and coyotes) based on 2,063,129 SNPs ascertained using the genome-wide data-set. All pre-contact dog and CTVT samples were projected.

AndeanFox

100

C_MidW

C_Cal

100

100

TAI

100

100 100

100 100 100

100

D_Mal68

100

W_Yellow1

W_Yellow2

W_Mex1

100

W_Mongo

W_Altai

W_Spa W_India W_Iran

W_Port

D_Green

92

100

100

D_Husky89

D_Husky

100

D_AHusky91 100

100 C_79T

C_24T

89

33

100

2

26

33

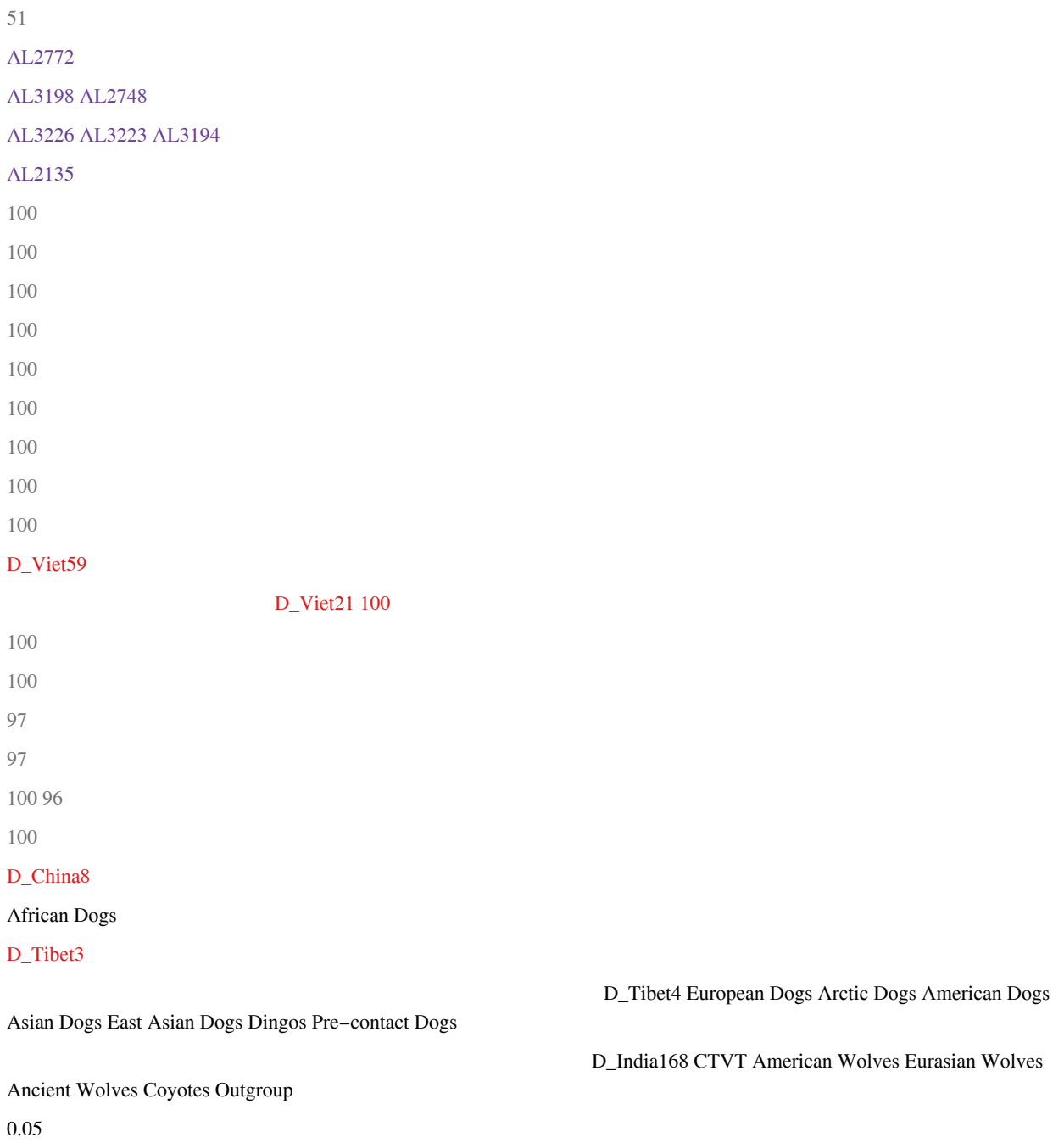


Fig. S10 Neighbour Joining (NJ) tree based on Identity By State (IBS). This figure is the same as in Figure 1c. Confirms that CTVT is more closely related to pre-contact dogs than any other dog population. Confirms that PCD form a monophyletic clade.

D_TMastif5

D_TMastif4

D_China9

D_Dingo

D_India60

100

100

100

100

100

100 58 65

D_Qatar5

D_SLaika

D_Leb85 D_Mex D_Na8

D_Port61

D_Leb79 D_Portt71 D_Peru D_Qatar27

D_Na89

D_NGDG

D_GerShep3 D_GerShep6

D_Basenji

47


AndeanFox
100
C_Cal C_MidW
100
100
TAI
100
D_Viet59
100
D_Viet21
100
100
W_Mex1 W_Yellow1
W_Yellow2
100
100
100
100
100 100
100
W_Altai W_Mongo
W_Port W_Spa W_India W_Iran
100
D_Dingo
100
AL3194
African Dogs European Dogs
100
100
C_24T
D_Mal68
D_India168
AL3223
Arctic Dogs American Dogs Asian Dogs East Asian Dogs Dingos Pre−contact Dogs CTVT American Wolves
100
Eurasian Wolves Ancient Wolves Coyotes
100 Outgroup

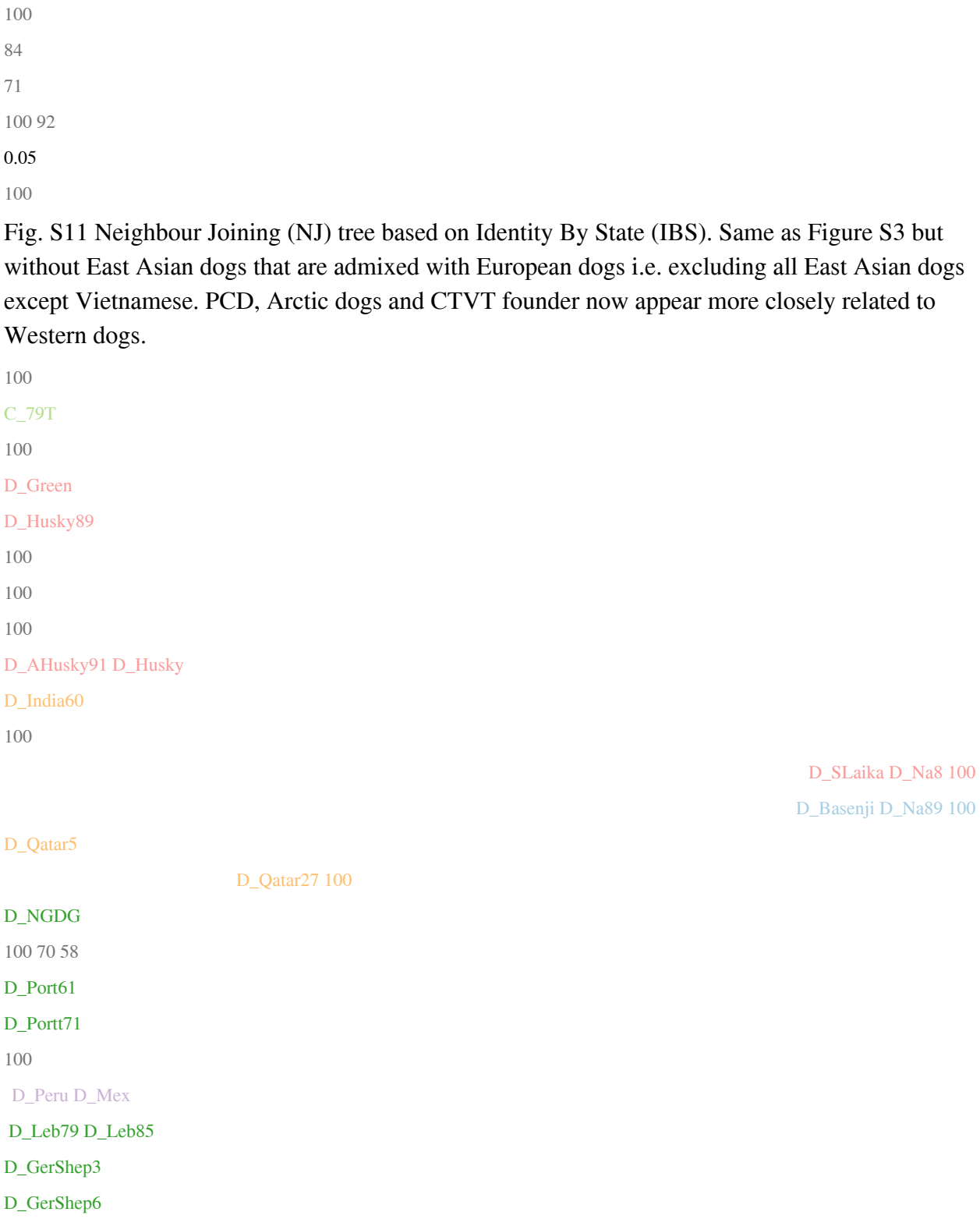


Fig. S11 Neighbour Joining (NJ) tree based on Identity By State (IBS). Same as Figure S3 but without East Asian dogs that are admixed with European dogs i.e. excluding all East Asian dogs except Vietnamese. PCD, Arctic dogs and CTVT founder now appear more closely related to Western dogs.

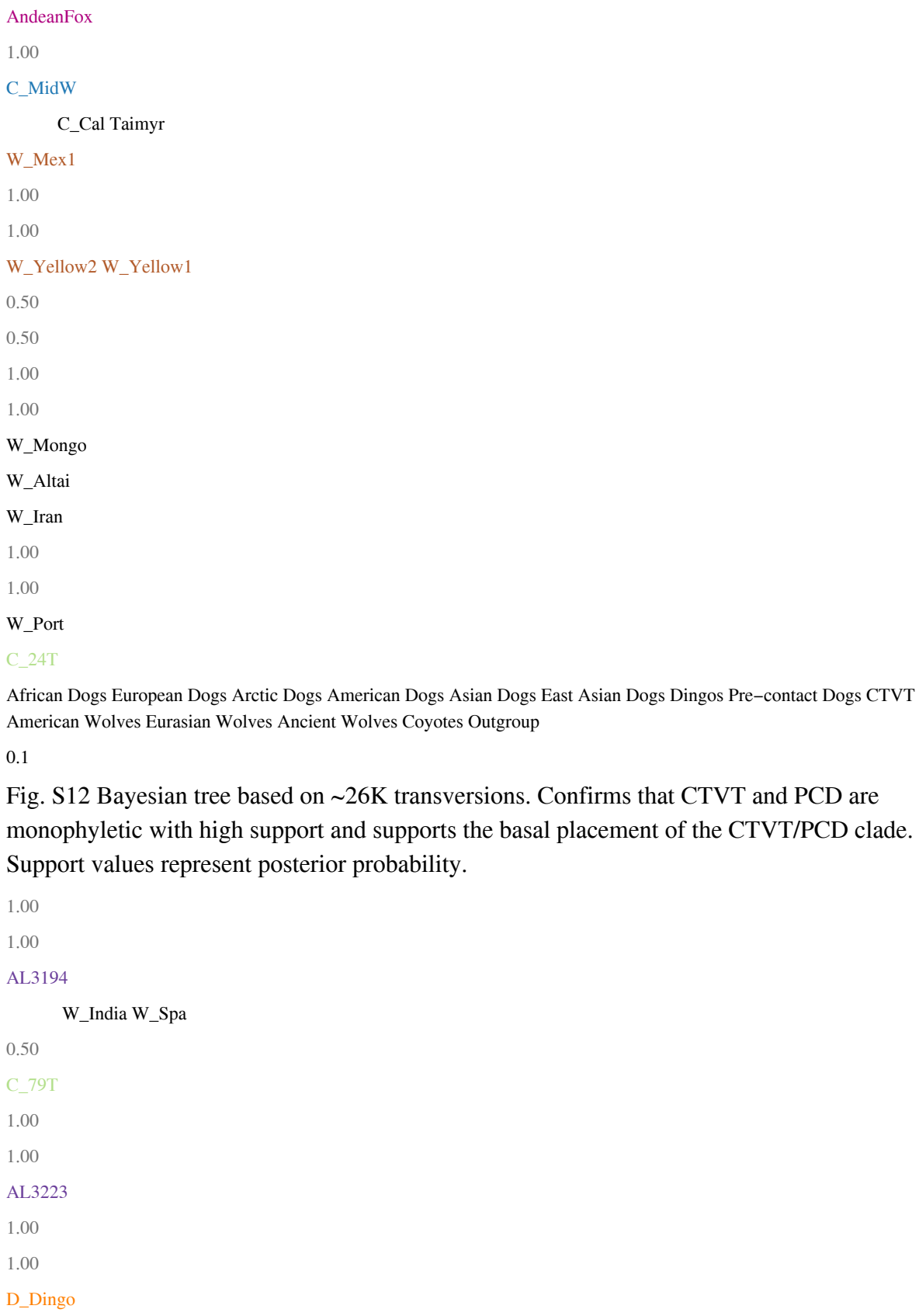


Fig. S12 Bayesian tree based on ~26K transversions. Confirms that CTVT and PCD are monophyletic with high support and supports the basal placement of the CTVT/PCD clade. Support values represent posterior probability.

D_Mal68

1.00 0.80

1.00

1.00

D_Green D_Husky89

D_Husky

D_AHusky91

1.00

1.00

1.00

1.00

D_Viet59 D_Viet21

D_China9 1.00

D_India168

D_China8

1.00

1.00

1.00

1.00

1.00

D_TMastif5 D_TMastif4 D_Tibet4

D_Tibet3 D_India60

1.00

1.00

1.00

D_Basenji D_Na89

D_Na8 0.72

1.00

D_Qatar5

D_Qatar27

1.00

0.50

D_NGDG

D_SLaika

0.98

0.98

0.98

0.50

D_Leb79

D_Leb85

1.00

D_GerShep6 D_GerShep3

0.52 0.50

0.50

D_Peru D_Portt71

D_Mex D_Port61

D_Green AL3194 AL3223

D_Mal68 AL3223 C_24T

AL3194 C_24T

D_Mal68

D_Green

C_79T

C_79T

D_Husky89

D_Husky89 D_Mal68

D_Mal68

D_AHusky91

D_AHusky91

D_AHusky91 D_Green

D_Green

D_Husky

D_Husky

D_Husky

D_AHusky91

AL3223

AL3223

D_Husky89

D_Husky

AL3194

AL3194

D_NGDG D_Viet21

D_Husky89

C_24T

C_24T

D_SLaika

D_Viet21

C_79T

C_79T

D_Mex

D_NGDG

D_NGDG

D_NGDG

D_Viet59

D_SLaika

D_SLaika

D_SLaika D_GerShep3

D_Mex

D_Mex D_GerShep6

D_Mex

D_Port61 D_Tibet3

D_Leb85

D_Leb85

D_China8

W_Iran

D_Basenji 0.175 0.200 0.225 0.19 0.20 0.21 0.22 0.23 0.24

0.20 0.21 0.22 0.23 0.24 0.20 0.21 0.22 0.23 0.24 f3

f3 D_Viet59 D_GerShep3 D_GerShep6

D_GerShep3

D_GerShep3 D_GerShep6

D_GerShep6 D_Portt71

D_Port61

D_Port61

Pre-contact Dogs

D_China9 D_Peru D_Leb79

D_Tibet3 D_Port61

D_Portt71

D_Portt71

African Dogs

D_Leb85 D_Tibet4

D_China8 D_Portt71

D_Peru

D_Peru

American Dogs

American Wolf Ancient Wolf Arctic Dogs Asian Dogs

Coyotes CTVT Dingo

East Asian Dogs Eurasian Wolf European Dogs Spitz Dogs

Fig. S13 Shared genetic drift measured by f

3

D_Qatar5

D_Qatar5 D_Leb79

D_Leb79 D_India168 D_Dingo

D_China9

D_Viet59

D_Viet59

D_TMastif5 D_TMastif4

D_Peru

D_Viet21

D_Viet21 D_Qatar27

D_Qatar27 D_Qatar5

D_Leb79 D_Leb85

D_Tibet3

D_Tibet3 D_Qatar27 D_Na8

D_Tibet4 D_India168

D_China8

D_China8

D_Na89

D_Dingo

D_China9

D_China9 D_India60

D_Tibet4

D_Tibet4 D_Basenji W_Spa

D_TMastif5 D_TMastif4

D_India168

D_India168

W_Port

D_Qatar5

D_TMastif4

D_TMastif4

W_Yellow1

D_Qatar27

D_TMastif5

D_TMastif5 W_Mex1 W_India

D_Na8

D_Na8

W_Yellow2

D_Na8 D_Na89

D_Na89

D_Na89

W_Mongo

D_India60

D_Dingo

D_Dingo

W_Altai

D_Basenji

D_India60

D_India60 D_Basenji

(Outgroup; Y, X) where Y is either Port au Choix dog (AL3194), Weyanoke Old town dog (AL3223), Alaskan Malamute (D_Mal68), Greenland sledge dog (D_Green) and X represents modern dog populations. Error bars represent 1 SE.

a

AL3223

b

D_Basenji

D_India60 D_India60

D_Basenji D_Na89

D_Na8 D_Na8

D_Na89 D_Qatar27

D_Qatar27 D_India168

D_India168 D_Tibet4

D_Tibet4 D_Qatar5

D_Portt71 D_Leb79

D_Peru D_Peru

D_Leb79 D_Portt71

D_Mex D_Mex

D_TMastif4

African Dogs

D_Port61

D_Port61

American Dogs

D_TMastif4

D_Qatar5

Arctic Dogs

n o i t a l u p o P

D_GerShep3 D_Leb85 D_China9 D_Tibet3 D_GerShep6

n o i t a l u p o P

D_China9 D_GerShep3

Asian Dogs

D_Leb85

CTVT D_Tibet3 D_GerShep6

Dingo

D_China8

D_China8

East Asian Dogs

D_TMastif5

D_TMastif5

European Dogs D_SLaika

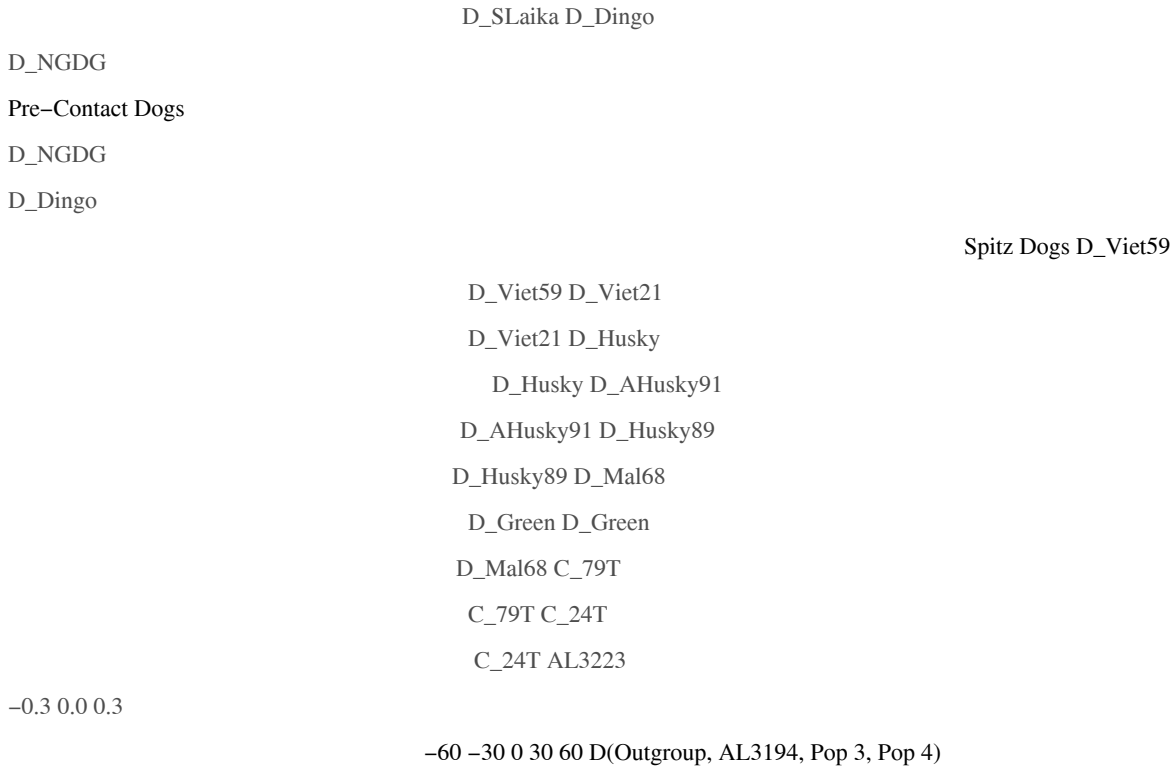


Fig. S14 Box plot representing a. D-statistics and b. significance of D-statistics (Z) for every combination of D(Outgroup, AL3194[Port au Choix], Pop3, Pop4), where Pop3 is fixed and Pop4 represents any other genome. Positive values support a close relationship between Pop3 and PCD while negative values imply PCD are closer to other dog populations. If Pop3 is not admixed with PCD, we expect -4<Z<4 (x-axis).

a

AL3194

b

D_India60

D_India60 D_Basenji

D_Basenji D_Na89

D_Na8 D_Na8

D_Na89 D_India168

D_India168 D_Qatar27

D_Qatar27 D_Peru

D_TMastif4 D_Portt71

D_Peru D_Tibet4

D_Portt71 D_Port61

D_Tibet4 D_TMastif4

D_China9 D_Qatar5

D_Port61

African Dogs

D_China9

D_Mex

American Dogs

D_Mex

D_Leb79

Arctic Dogs

n o i t a l u p o P

D_Leb79 D_China8 D_GerShep3

n o i t a

D_China8 D_GerShep3

Asian Dogs

D_Leb85 D_GerShep6 D_Tibet3

l u p o P

D_GerShep6 D_TMastif5 D_Qatar5 D_Tibet3

CTVT

Dingo

East Asian Dogs

D_TMastif5

D_Leb85

European Dogs D_Dingo

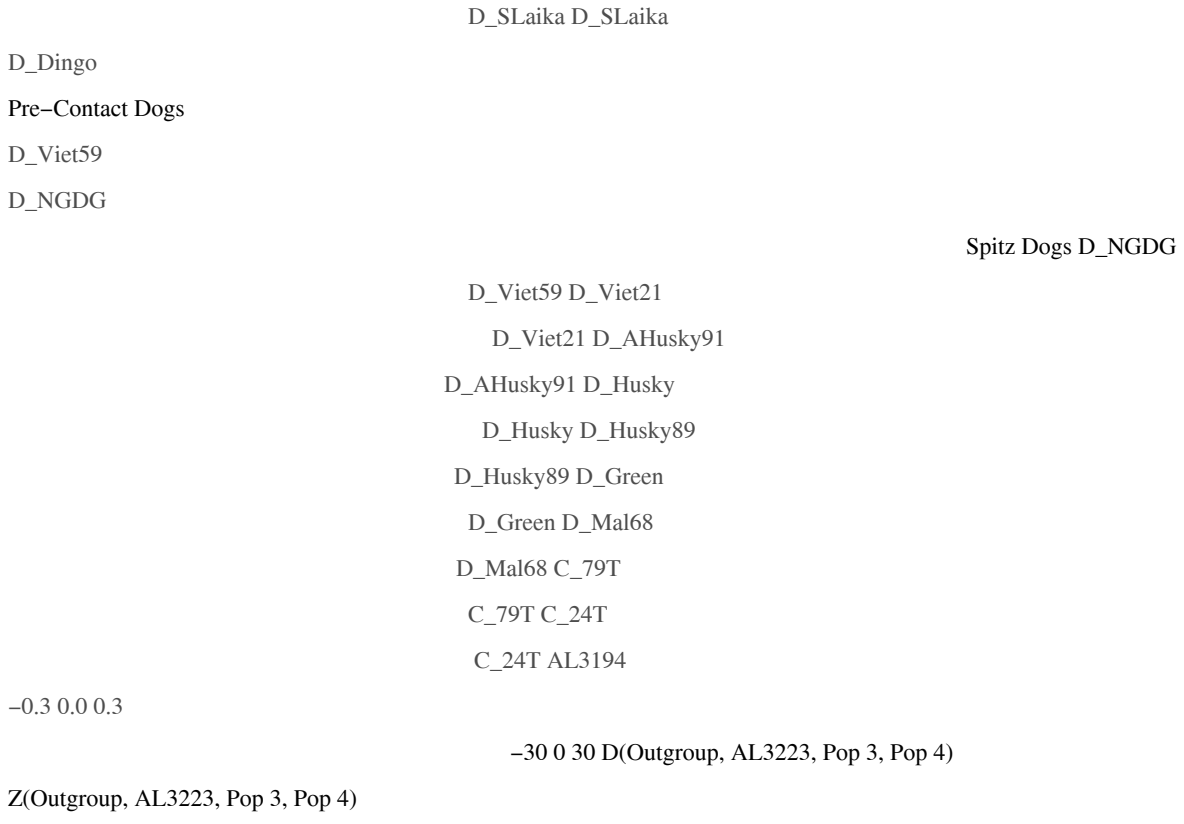


Fig. S15 Box plot representing a. D-statistics and b. significance of D-statistics (Z) for every combination of D(Outgroup, AL3223[Weyanoke Old town], Pop3, Pop4), where Pop3 is fixed and Pop4 represents any other genome. Positive values support a close relationship between Pop3 and PCD while negative values imply PCD are closer to other dog populations. If Pop3 is not admixed with PCD, we expect -4<Z<4 (x-axis).

a

AL3194

b

D_GerShep3

D_GerShep3 D_Leb79

D_GerShep6 D_GerShep6

D_Leb79 D_Na89

D_Portt71 D_Portt71

D_Na89 D_Qatar27

D_Qatar27 D_Husky

D_Husky D_Port61

D_Port61 D_Peru

D_Peru D_India60

D_India60 D_Leb85

D_Leb85 D_Tibet4

D_Tibet4

African Dogs

D_India168

D_SLaika

American Dogs D_Viet21

D_India168 D_SLaika

D_Mex

Arctic Dogs

n o i t a l u p o P

D_Mex D_Viet59

n o i

D_Viet21

D_Qatar5

Asian Dogs

D_Na8

t a l u p o P

D_Viet59 D_Na8

CTVT

D_Qatar5

Dingo D_TMastif5

D_TMastif5 D_TMastif4

D_TMastif4

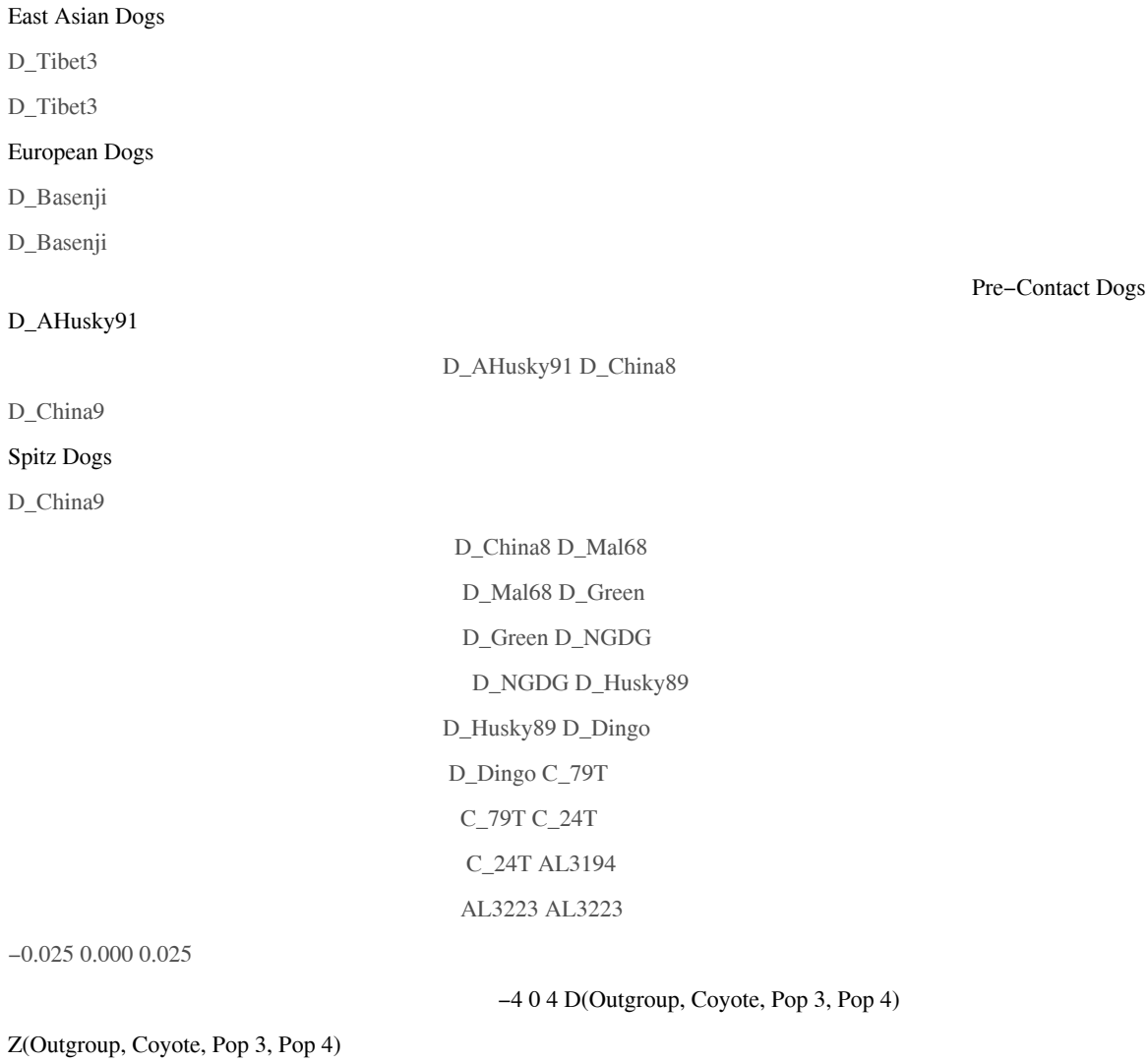


Fig. S16 Box plot representing a. D-statistics and b. significance of D-statistics (Z) for every combination of D(Outgroup, Coyote, Pop3, Pop4), where Pop3 is fixed (x-axis) and Pop4 represents any other genome. Positive values support a close relationship between Pop3 and coyotes while negative values imply coyotes are closer to other dog populations. If Pop3 is not admixed with coyotes, we expect -4<Z<4 (x-axis).

Fig. S17 Box plot representing a. D-statistics b. significance of D-statistics (Z) for every combination of D(Outgroup, north American wolf, Pop3, Pop4), where Pop3 is fixed and Pop4

represents any other genome. Positive values support a close relationship between Pop3 and north American wolves while negative values imply NA wolves are closer to other dog populations. If Pop3 is not admixed with north American wolves, we expect - 4<Z<4 (x-axis).

1- α α

CTVT Pre-Contact

Dog

Fig. S18 Schematic representation of the assumed phylogeny for the f4 ratio test used to estimate pre-contact ancestry into modern North American dogs. Alpha represents the degree of ancestry from pre-contact dogs.

American

Eurasian Dog

Dog

Andean Fox

#

0 7 . 0

#

2 4 6 8 10 12 14

K

#

8 6 . 0

#

#

#

V C

6 6 . 0

#

#

#

4 6 . 0

#

#

2 6 . 0

# #

#

#

Fig. S19 Cross validation (CV) values for ADMIXTURE analysis of SNP array data.

Fig. S20 ADMIXTURE results based on SNP array data for K=4, 10 and 15. Population code: AED=American Eskimo Dog, AM=Alaskan Malamute, APBT=American Pit Bull Terrier, AST=American Staffordshire Terrier, BEA=Beagle, BOX=Boxer, CBR=Chesapeake Bay Retriever, CC=Chow Chow, CD=Carolina Dog, CHI=Chihuahua, CLD=Catahoula Leopard Dog, COO=Chinook, CSP=Chinese Shar-pei, CTVT=CTVT, DAL=Alaskan Husky, DCH=Chinese Village Dog, DEU=European Village Dog, DGL=Greenland Sledge Dog, DHU=Husky, DLB=Lebanese Village Dog, DMA=Malamute, DME=Mexican Hairless Dog, DPU=Peruvian Hairless Dog, DSL=Siberian Laika, EUR=Eurasier, FS=Finnish Spitz, GSD=Greenland Sledge Dog, NEW=Newfoundland, NSDTR=Nova Scotia Duck Tolling Retriever, PCD=Pre- Colombian Dogs, PIO=Peruvian Inca Orchid, SAM=Samoyed, SH=Siberian Husky, VDB=Village Dog Belize, VDB2=Village Dog Brazil, VDC=Village Dog Colombia, VDCR=Village Dog Costa Rica, VDDR=Village Dog Dominican Republic, VDH=Village Dog Honduras, VDP=Village Dog Panama, VDPA=Village Dog Peru- Arequipa, VDPC=Village Dog Peru-Cusco, VDPI=Village Dog Peru-Ica, VDPL=Village Dog Peru-Loreto, VDPP=Village Dog Peru-Puno, VDPR=Village Dog Puerto Rico, VDUA=Village Dog US-Alaska, XOL=Xoloitzcuintli (see Table S9 for more information).

COY

DVN

DGL

DMA

DTI WAM

CTVT

D

5.5 SE

Fig. S21 Admixture graph without migration edge and matrix of residuals, expressed as the number of standard errors, inferred using TreeMix (based on transversions). Western Eurasian dogs - Portuguese village dogs (DEU), German Shepherd (DGS), East Asian dogs - Vietnamese village dogs (DVN) and Tibetan village dogs (DTI), Pre-contact dogs (PCD), including both Port au Choix (AL3194) and Weyanoke Old Town (AL3223), Arctic dogs - Malamute (DMA) and Greenland dogs (DGL), CTVT - (79T and 24T), Eurasian wolves (WEU) from Spain and Portugal, North American wolves (WAM) from Yellowstone, Coyotes (COY) as an outgroup.

DMA

WEU

DGS

Migration weight

0.5

DGL

DTI

CTVT

DGS

PCD

DEU

PCD

D

D

D

C

P

−5.5 SE

0

COY

DVN

10 s.e.


DEU
WAM
WEU
I T
A M
S G
L G
T V T
D C
Y O C
N V D
U E D 0.00 0.05 0.10 0.15 0.20
M A W
U E W
Drift parameter

COY

DVN

DTI

DTI WEU

DGL

PCD

D

5.4 SE

Fig. S22 Admixture graph with a single migration edge and matrix of residuals, expressed as the number of standard errors, inferred using TreeMix (based on transversions). We see evidence for admixture from coyotes (COY) into the pre-contact dog lineage (PCD/CTVT), consistent with Figure S16. Western Eurasian dogs - Portuguese village dogs (DEU), German Shepherd (DGS), East Asian dogs - Vietnamese village dogs (DVN) and Tibetan village dogs (DTI), Pre-contact dogs (PCD), including both Port au Choix (AL3194) and Weyanoke Old Town (AL3223), Arctic dogs - Malamute (DMA) and Greenland dogs (DGL), CTVT - (79T and 24T), Eurasian wolves (WEU) from Spain and Portugal, North American wolves (WAM) from Yellowstone, Coyotes (COY) as an outgroup.

DMA

DEU

DGS

Migration weight

0.5

DGS

DGL

CTVT

PCD

CTVT

WAM

D

D

D

C

P

−5.4 SE

0

COY

DMA


DVN
10 s.e.
DEU
WAM
WEU
I T
A M
S G
L G
T V T
D C
Y O C
N V D
U E D 0.00 0.05 0.10 0.15 0.20 0.25
M A W
U E W
Drift parameter

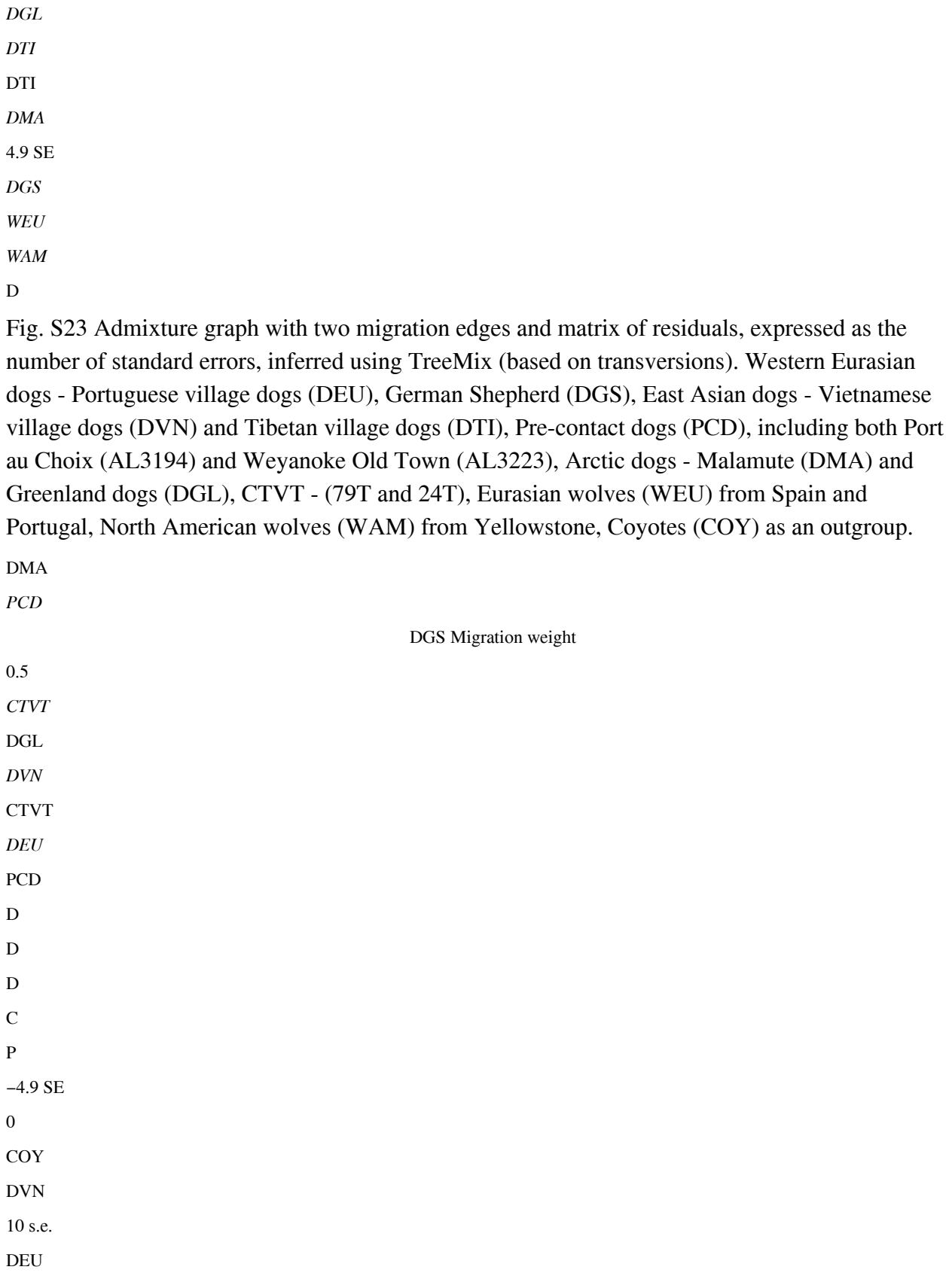
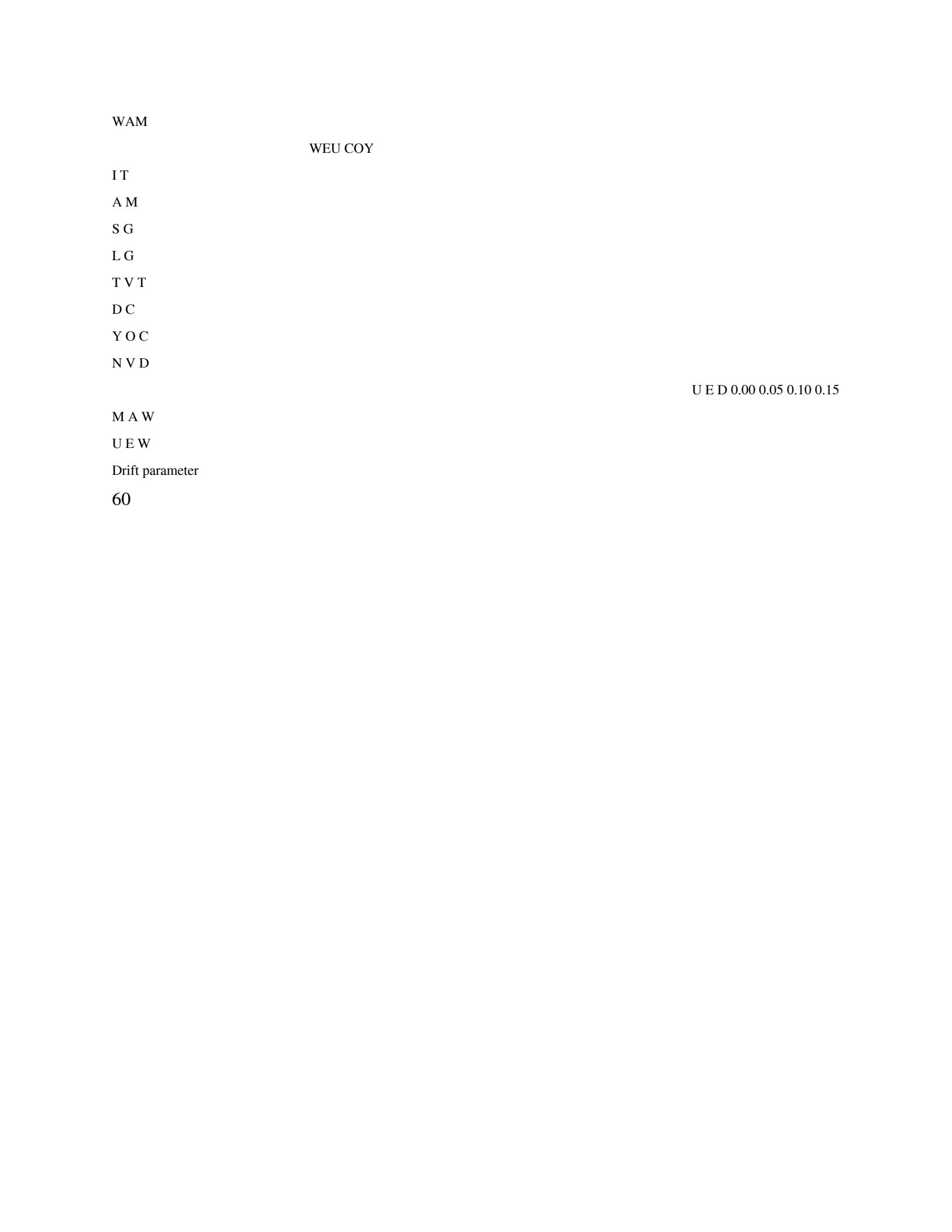


Fig. S23 Admixture graph with two migration edges and matrix of residuals, expressed as the number of standard errors, inferred using TreeMix (based on transversions). Western Eurasian dogs - Portuguese village dogs (DEU), German Shepherd (DGS), East Asian dogs - Vietnamese village dogs (DVN) and Tibetan village dogs (DTI), Pre-contact dogs (PCD), including both Port au Choix (AL3194) and Weyanoke Old Town (AL3223), Arctic dogs - Malamute (DMA) and Greenland dogs (DGL), CTVT - (79T and 24T), Eurasian wolves (WEU) from Spain and Portugal, North American wolves (WAM) from Yellowstone, Coyotes (COY) as an outgroup.


WAM
WEU COY
I T
A M
S G
L G
T V T
D C
Y O C
N V D
U E D 0.00 0.05 0.10 0.15
M A W
U E W
Drift parameter
60

WEU

DEU

DTI DVN

DGL

WAM

COY

D

2.7 SE

Fig. S24 Admixture graph with three migration edges and matrix of residuals, expressed as the number of standard errors, inferred using TreeMix (based on transversions). Western Eurasian dogs - Portuguese village dogs (DEU), German Shepherd (DGS), East Asian dogs - Vietnamese village dogs (DVN) and Tibetan village dogs (DTI), Pre-contact dogs (PCD), including both Port au Choix (AL3194) and Weyanoke Old Town (AL3223), Arctic dogs - Malamute (DMA) and Greenland dogs (DGL), CTVT - (79T and 24T), Eurasian wolves (WEU) from Spain and Portugal, North American wolves (WAM) from Yellowstone, Coyotes (COY) as an outgroup.

DMA PCD

DGS

Migration

CTVT weight

0.5

DGL

CTVT DMA

PCD

D

D

D

C

P

−2.7 SE

0

DTI

COY

DGS

DVN

10 s.e.

DEU

WAM


WEU
I T
A M
S G
L G
T V T
D C
Y O C
N V D
U E D 0.00 0.05 0.10 0.15 0.20
M A W
U E W
Drift parameter

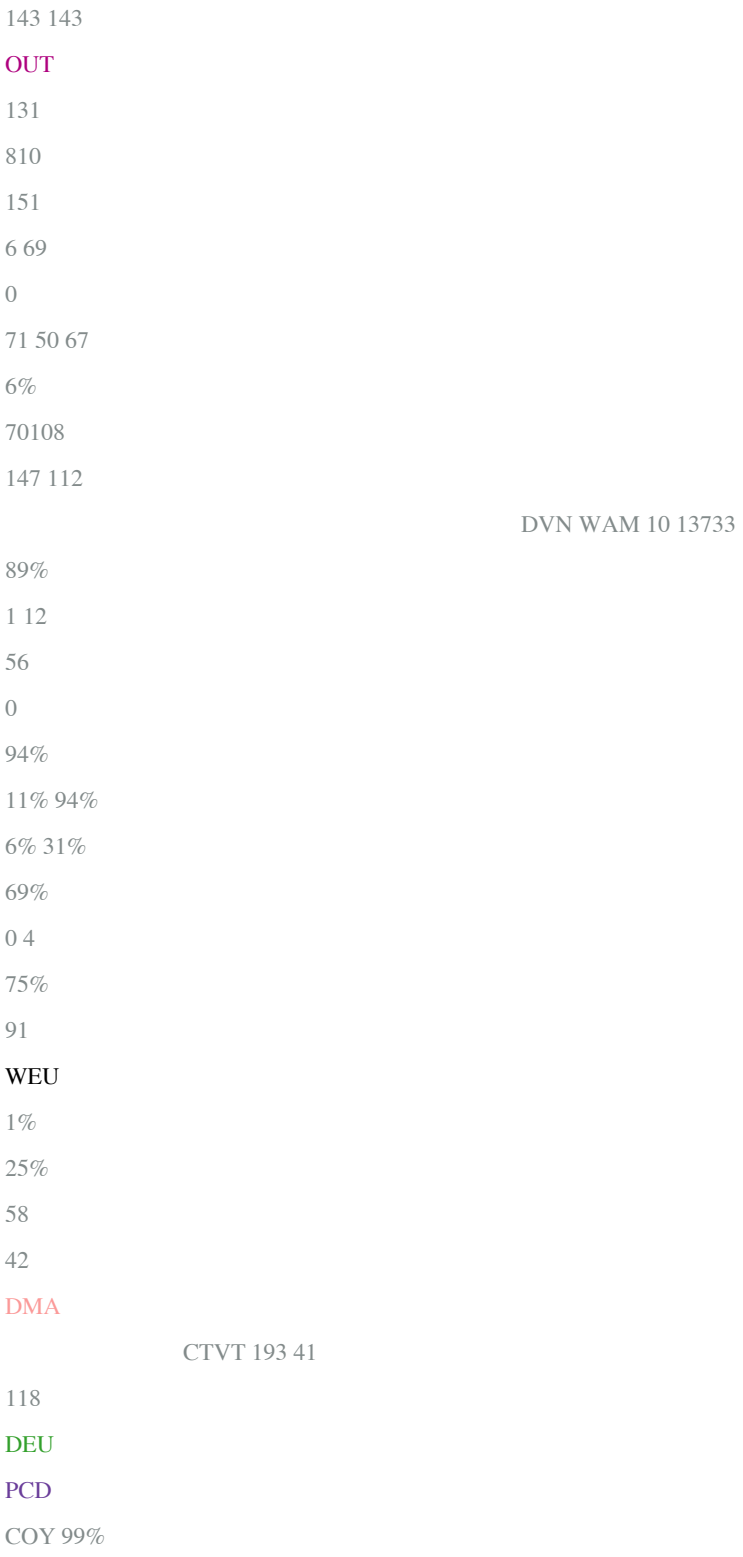


Fig. S25 Qpgraph model with admixture fractions. Western Eurasian dogs - Portuguese village dogs (DEU), East Asian dogs - Vietnamese village dogs (DVN), Pre-contact dogs (PCD), including both Port au Choix (AL3194) and Weyanoke Old Town (AL3223), Arctic dogs -

Malamute (DMA), CTVT - (79T and 24T), Eurasian wolves (WEU) from Spain and Portugal, North American wolves (WAM) from Yellowstone, Coyotes (COY) as an outgroup.

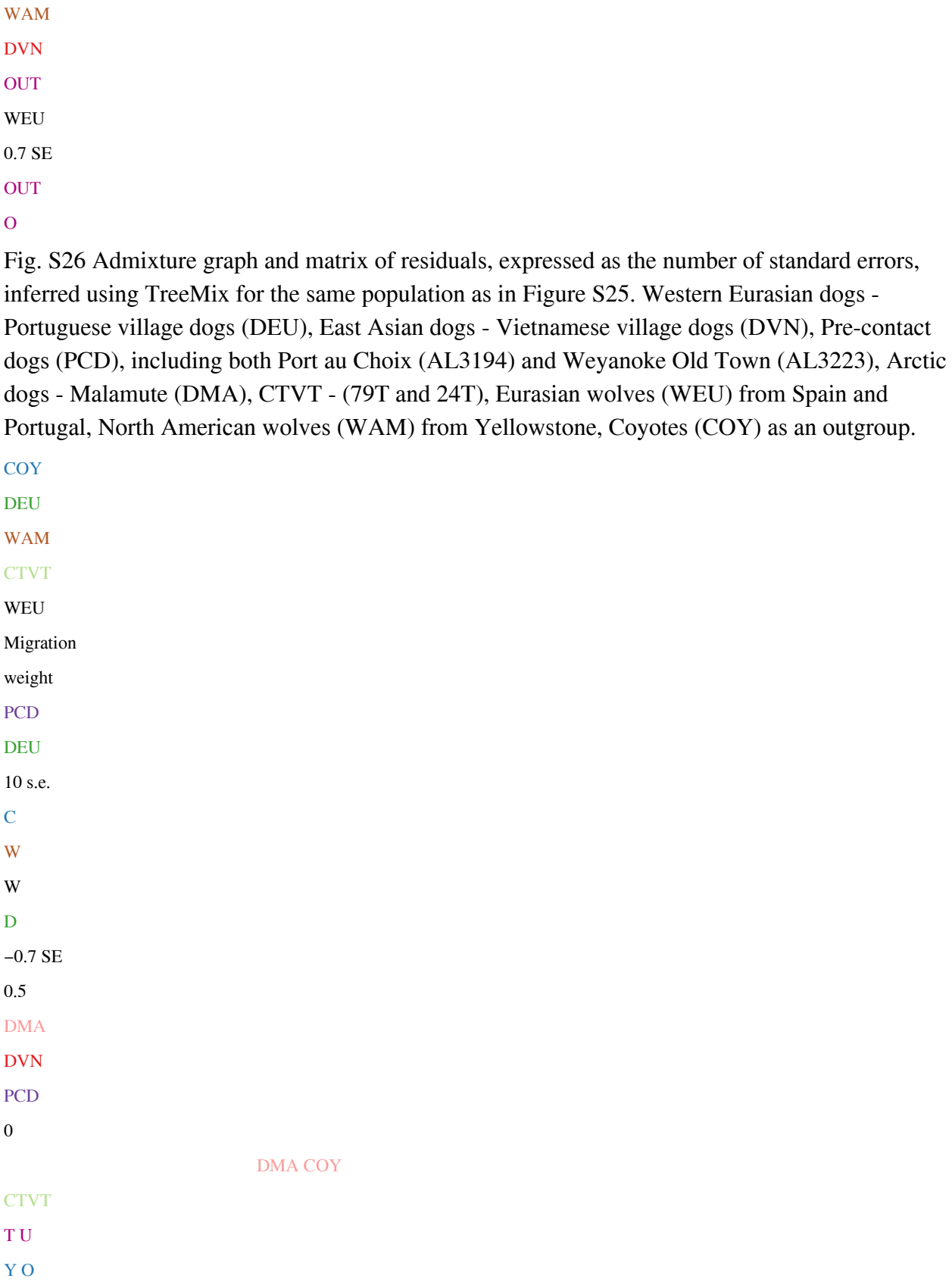


Fig. S26 Admixture graph and matrix of residuals, expressed as the number of standard errors, inferred using TreeMix for the same population as in Figure S25. Western Eurasian dogs - Portuguese village dogs (DEU), East Asian dogs - Vietnamese village dogs (DVN), Pre-contact dogs (PCD), including both Port au Choix (AL3194) and Weyanoke Old Town (AL3223), Arctic dogs - Malamute (DMA), CTVT - (79T and 24T), Eurasian wolves (WEU) from Spain and Portugal, North American wolves (WAM) from Yellowstone, Coyotes (COY) as an outgroup.


M A
U E
U E
N V D
D C P 0.00 0.05 0.10 0.15 0.20 0.25
A M D
T V T C
Drift parameter

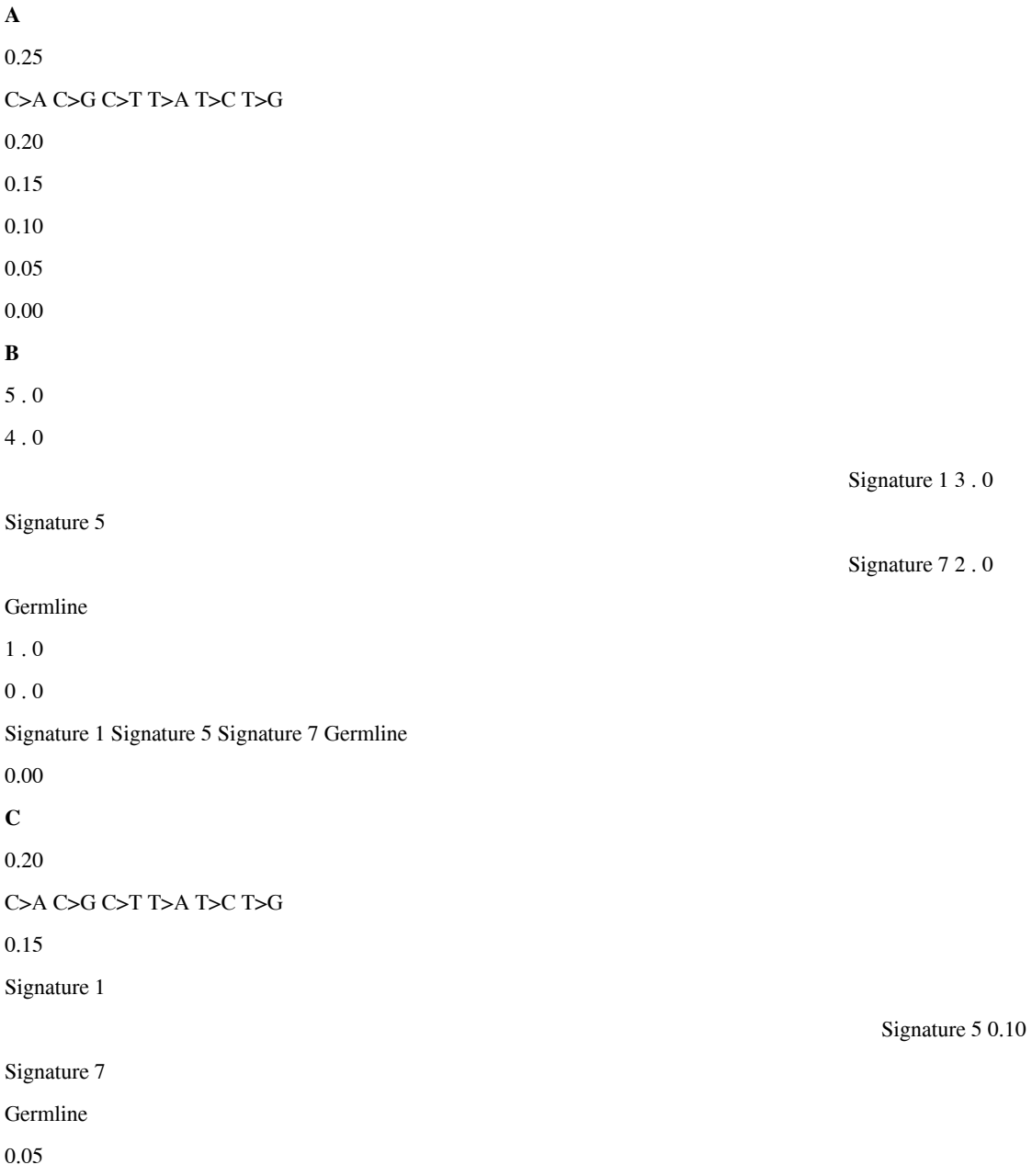


Fig. S27 A. CTVT mutation spectrum. 1,933,897 tumor-only mutations in CTVT are displayed by mutation type (in pyrimidine context) with immediate 5' and 3' context. Each of the 96 mutation classes is displayed on the horizontal axis. Mutation proportions are displayed relative to CanFam3.1 B. Fraction of CTVT tumor-only mutations attributable to COSMIC Signatures 1, 5, 7, and the dog germline signature, as estimated using sigfit. C.Reconstruction of CTVT tumor-only spectrum using COSMIC signatures 1,5 and 7 and the dog germline signature.

Table S1. Information about samples ancient samples sequenced in this study including provenance, age (radiocarbon or stratigraphic information), and sequencing statistics (endogeneous content etc.). N_SNP_call corresponds to the number of sites from our ~6M sites that were call in each sample. Analysed Nuclear and mtDNA Analysed columns corresponds to sample that were (1) and were not (0) analysed for nuclear and mtDNA analyses respectively. mtDNA capture column indicates whether mtDNA was enriched using target capture (0, no; 1, yes).

Table S2. Table containing information (coverage, accession etc.) of modern whole genomes used in this study.

Table S3. Table containing information (coverage, accession etc.) of modern whole genomes used in this study.

Pop 2 Pop 3 Pop 4 D Z BABA ABBA n

C_Cal AL3194 AL3223 0.0152 1.460 5,681 5,511 285,638

C_Cal AL3223 AL3194 -0.0152 -1.460 5,511 5,681 285,638

C_MidW AL3194 AL3223 0.0035 0.277 3,443 3,419 181,671

C_MidW AL3223 AL3194 -0.0035 -0.277 3,419 3,443 181,671

W_Mex1 AL3194 AL3223 -0.0130 -1.086 7,394 7,588 282,977

W_Mex1 AL3223 AL3194 0.0130 1.086 7,588 7,394 282,977

W_Yellow1 AL3194 AL3223 -0.0083 -0.785 7,501 7,626 281,684

W_Yellow1 AL3223 AL3194 0.0083 0.785 7,626 7,501 281,684

W_Yellow2 AL3194 AL3223 -0.0067 -0.622 7,513 7,615 282,189

W_Yellow2 AL3223 AL3194 0.0067 0.622 7,615 7,513 282,189

TAI AL3194 AL3223 -0.0001 -0.005 3,026 3,026 116,212

TAI AL3223 AL3194 0.0001 0.005 3,026 3,026 116,212

Table S4. D-statistics for D(Outgroup, Asian or European dogs, PCD or CTVT, Arctic dogs). The Andean Fox was used as an outgroup for these analyses. n corresponds to the number of SNPs where all populations have data. Standard error for these statistics was obtained by performing a weighted block jackknife over 1Mb blocks. Statistics with Z>3 and Z<-3 are shown in bold.

Pop 2	Pop 3	Pop 4	D	Z	BABA	ABBA	n
D_China8	AL3194	D_AHusky91	0.0752	7.65	47,384	40,759	879,166
D_China8	AL3194	D_Green	0.0606	6.01	47,406	41,994	920,449
D_China8	AL3194	D_Husky	0.0693	7.253	51,412	44,755	936,527
D_China8	AL3194	D_Husky89	0.0825	8.407	41,900	35,513	768,218
D_China8	AL3194	D_Mal68	0.0655	6.538	46,698	40,958	907,759
D_China8	C_24T	D_AHusky91	0.0907	9.461	61,440	51,230	1,115,093
D_China8	C_24T	D_Green	0.0753	7.523	61,012	52,472	1,167,919
D_China8	C_24T	D_Husky	0.0819	8.781	66,255	56,230	1,189,650
D_China8	C_24T	D_Husky89	0.0987	9.798	54,693	44,866	985,295
D_China8	C_24T	D_Mal68	0.0796	8.128	60,682	51,739	1,151,059
D_China9	AL3194	D_AHusky91	0.064	6.644	45,860	40,348	849971
D_China9	AL3194	D_Green	0.0812	8.035	47,827	40,641	893422
D_China9	AL3194	D_Husky	0.0684	7.323	50,481	44,022	905549
D_China9	AL3194	D_Husky89	0.087	8.783	41,939	35,228	760651
D_China9	AL3194	D_Mal68	0.065	6.504	45,784	40,195	877630
D_China9	C_24T	D_AHusky91	0.0776	7.936	59,705	51,110	1,078553
D_China9	C_24T	D_Green	0.0926	9.696	61,461	51,046	1,133912

D_China9 C_24T D_Husky 0.0799 8.709 65,471 55,786 1,150989

D_China9 C_24T D_Husky89 0.1001 9.815 54,894 44,904 974829

D_China9 C_24T D_Mal68 0.0743 7.814 59,571 51,330 1,113510

D_Port61 AL3194 D_AHusky91 0.0981 9.344 49,646 40,780 897394

D_Port61 AL3194 D_Green 0.0733 7.059 49,098 42,391 939,704

D_Port61 AL3194 D_Husky 0.0994 10.286 54,206 44,402 956,025

D_Port61 AL3194 D_Husky89 0.1006 9.791 43,556 35,593 784,550

D_Port61 AL3194 D_Mal68 0.1044 9.88 49,860 40,438 926,900

D_Port61 C_24T D_AHusky91 0.1115 10.706 64,216 51,337 1,137,495

D_Port61 C_24T D_Green 0.088 8.191 63,226 53,005 1,191,621

D_Port61 C_24T D_Husky 0.1125 11.651 70,044 55,878 1,213,643

D_Port61 C_24T D_Husky89 0.1181 11.113 57,125 45,064 1,005,524

D_Port61 C_24T D_Mal68 0.1167 11.029 64,747 51,223 1,174,636

D_Portt71 AL3194 D_AHusky91 0.0949 9.146 48,273 39,904 877,094

D_Portt71 AL3194 D_Green 0.0814 8.348 48,157 40,909 918,351

D_Portt71 AL3194 D_Husky 0.0967 10.277 52,649 43,367 934,632

D_Portt71 AL3194 D_Husky89 0.1035 10.427 42,550 34,572 765,611

D_Portt71 AL3194 D_Mal68 0.1047 10.09 48,497 39,308 906,275

D_Portt71 C_24T D_AHusky91 0.1122 10.895 62,687 50,038 1,112,724

D_Portt71 C_24T D_Green 0.0994 9.888 62,205 50,964 1,165,580

D_Portt71 C_24T D_Husky 0.1139 12.371 68,354 54,374 1,187,444

D_Portt71 C_24T D_Husky89 0.124 11.829 55,911 43,573 982,083

D_Portt71 C_24T D_Mal68 0.1214 11.729 63,196 49,518 1,149,527

Table S5. D-statistics for D(Outgroup, Taimyr, PCD/Arctic dogs, European, Asian, or Arctic dogs). The Andean Fox was used as an outgroup for these analyses. n corresponds to the number of SNPs where all populations have data. Standard error for these statistics was obtained by performing a weighted block jackknife over 1Mb blocks. Statistics with Z>3 and Z<-3 are shown in bold.

Pop 2	Pop 3	Pop 4	D	Z	BABA	ABBA	n
TAI	**AL3194**	**D_India168**	**-0.0274**	**-3.377**	**13,074**	**13,812**	**297,919**
TAI	**AL3194**	**D_Peru**	**-0.0267**	**-3.293**	**17,139**	**18,079**	**384,525**
TAI	**AL3194**	**D_India60**	**-0.0247**	**-3.254**	**17,383**	**18,264**	**372,784**
TAI	**AL3194**	**D_Na8**	**-0.0256**	**-3.199**	**16,614**	**17,489**	**366,726**
TAI	**D_Mal68**	**D_Na8**	**-0.0206**	**-3.127**	**19,530**	**20,350**	**459,311**
TAI	**AL3194**	**D_Port61**	**-0.0236**	**-3.006**	**17,398**	**18,239**	**387,777**
TAI	D_Mal68	D_Peru	-0.0195	-2.856	20,182	20,983	482,639
TAI	D_Mal68	D_India60	-0.0189	-2.789	20,930	21,735	465,930
TAI	D_Mal68	D_Port61	-0.0204	-2.768	20,358	21,206	485,553
TAI	AL3194	D_Viet59	-0.0227	-2.753	13,212	13,824	308,811
TAI	D_Green	D_Na8	-0.0202	-2.674	20,383	21,224	465,676
TAI	D_Husky89	D_India168	-0.0205	-2.664	13,042	13,588	314,547
TAI	D_Green	D_India60	-0.0198	-2.607	21,505	22,372	472,462
TAI	D_Husky89	D_Port61	-0.0215	-2.571	17,284	18,043	413,104
TAI	D_Husky89	D_Peru	-0.0201	-2.525	17,191	17,896	410,031
TAI	D_Green	D_Port61	-0.0197	-2.506	21,204	22,058	492,659
TAI	AL3194	D_Husky	-0.0199	-2.498	16,444	17,111	394,720
TAI	AL3194	D_SLaika	-0.0194	-2.485	16,821	17,487	379,149

TAI D_Green D_India168 -0.0178 -2.481 15,728 16,296 377,110

TAI D_Mal68 D_Na89 -0.0195 -2.481 12,195 12,680 289,545

TAI D_Green D_Peru -0.0194 -2.455 20,782 21,605 488,730

TAI D_Husky89 D_India60 -0.0186 -2.44 17,456 18,117 394,520

TAI AL3194 D_Mex -0.0193 -2.426 17,007 17,675 380,509

TAI AL3194 D_Viet21 -0.0204 -2.419 13,088 13,634 303,723

TAI AL3194 D_Tibet3 -0.0184 -2.41 16,776 17,403 375,015

TAI AL3194 D_Portt71 -0.0185 -2.366 16,997 17,637 379,175

Table S6. D-statistics for D(Outgroup, Taimyr, PCD or Arctic dogs, CTVT/Arctic dogs or PCD). The Andean Fox was used as an outgroup for these analyses. n corresponds to the number of SNPs where all populations have data. Standard error for these statistics was obtained by performing a weighted block jackknife over 1Mb blocks. Statistics indicate that there has been no extra admixture from the Taimyr wolf into other populations.

Pop 2 Pop 3 Pop 4 D Z BABA ABBA n

TAI AL3194 C_24T -0.0239 -2.258 10,176 10,674 380,938

TAI AL3194 C_79T -0.0267 -2.364 8,948 9,438 346,188

TAI AL3194 D_AHusky91 -0.0126 -1.58 15,165 15,554 370,227

TAI AL3194 D_Green -0.0059 -0.681 15,468 15,650 387,439

TAI AL3194 D_Husky -0.0199 -2.498 16,444 17,111 394,720

TAI AL3194 D_Husky89 -0.0076 -0.915 13,258 13,462 321,758

TAI AL3194 D_Mal68 -0.0025 -0.312 15,225 15,301 382,193

TAI C_24T AL3194 0.0239 2.258 10,674 10,176 380,938

TAI C_24T C_79T -0.0014 -0.654 5,721 5,737 441,374

TAI C_24T D_AHusky91 0.0045 0.587 19,419 19,247 461,923

TAI C_24T D_Green 0.009 1.179 19,495 19,147 483,404

TAI C_24T D_Husky -0.0028 -0.381 20,900 21,018 493,025

TAI C_24T D_Husky89 0.0071 0.929 17,107 16,865 405,779

TAI C_24T D_Mal68 0.0102 1.321 19,467 19,075 476,690

TAI C_79T AL3194 0.0267 2.364 9,438 8,948 346,188

TAI C_79T C_24T 0.0014 0.654 5,737 5,721 441,374

TAI C_79T D_AHusky91 0.0041 0.539 17,112 16,972 416,326

TAI C_79T D_Green 0.0062 0.786 17,041 16,832 435,794

TAI C_79T D_Husky -0.0021 -0.279 18,489 18,566 446,651

TAI C_79T D_Husky89 0.0081 1.039 14,690 14,454 358,654

TAI C_79T D_Mal68 0.0113 1.419 17,079 16,698 429,798

TAI D_AHusky91 AL3194 0.0126 1.58 15,554 15,165 370,227

TAI D_AHusky91 C_24T -0.0045 -0.587 19,247 19,419 461,923

TAI D_AHusky91 C_79T -0.0041 -0.539 16,972 17,112 416,326

TAI D_AHusky91 D_Green 0.0083 1.034 17,900 17,607 470,340

TAI D_AHusky91 D_Husky -0.0081 -1.105 15,523 15,775 478,592

TAI D_AHusky91 D_Husky89 0.0009 0.115 12,481 12,458 394,759

TAI D_AHusky91 D_Mal68 0.0072 1.009 18,163 17,904 463,584

TAI D_Green AL3194 0.0059 0.681 15,650 15,468 387,439

TAI D_Green C_24T -0.009 -1.179 19,147 19,495 483,404

TAI D_Green C_79T -0.0062 -0.786 16,832 17,041 435,794

TAI D_Green D_AHusky91 -0.0083 -1.034 17,607 17,900 470,340

TAI D_Green D_Husky -0.0137 -1.839 19,216 19,750 501,352

TAI D_Green D_Husky89 -0.0025 -0.296 15,501 15,578 416,390

TAI D_Green D_Mal68 -0.0002 -0.029 16,258 16,266 485,928

TAI D_Husky AL3194 0.0199 2.498 17,111 16,444 394,720

TAI D_Husky C_24T 0.0028 0.381 21,018 20,900 493,025

TAI D_Husky C_79T 0.0021 0.279 18,566 18,489 446,651

TAI D_Husky D_AHusky91 0.0081 1.105 15,775 15,523 478,592

TAI D_Husky D_Green 0.0137 1.839 19,750 19,216 501,352

TAI D_Husky D_Husky89 0.0123 1.655 14,207 13,860 419,790

TAI D_Husky D_Mal68 0.013 1.913 19,767 19,258 494,030

TAI D_Husky89 AL3194 0.0076 0.915 13,462 13,258 321,758

TAI D_Husky89 C_24T -0.0071 -0.929 16,865 17,107 405,779

TAI D_Husky89 C_79T -0.0081 -1.039 14,454 14,690 358,654

TAI D_Husky89 D_AHusky91 -0.0009 -0.115 12,458 12,481 394,759

TAI D_Husky89 D_Green 0.0025 0.296 15,578 15,501 416,390

TAI D_Husky89 D_Husky -0.0123 -1.655 13,860 14,207 419,790

TAI D_Husky89 D_Mal68 0.0043 0.587 15,803 15,667 408,271

TAI D_Mal68 AL3194 0.0025 0.312 15,301 15,225 382,193

TAI D_Mal68 C_24T -0.0102 -1.321 19,075 19,467 476,690

TAI D_Mal68 C_79T -0.0113 -1.419 16,698 17,079 429,798

TAI D_Mal68 D_AHusky91 -0.0072 -1.009 17,904 18,163 463,584

TAI D_Mal68 D_Green 0.0002 0.029 16,266 16,258 485,928

TAI D_Mal68 D_Husky -0.013 -1.913 19,258 19,767 494,030

TAI D_Mal68 D_Husky89 -0.0043 -0.587 15,667 15,803 408,271

Table S7. D-statistics for D(Outgroup, Portuguese Village Dogs, Vietnamese Village Dogs, Asian Dogs). The Andean Fox was used as an outgroup for these analyses. n corresponds to the number of SNPs where all populations have data. Standard error for these statistics was obtained by performing a weighted block jackknife over 1Mb blocks.

Pop 2	**Pop 3**	**Pop 4**	**D**	**Z**	**BABA**	**ABBA**	**n**
D_Port61	D_Viet21	D_China8	0.1029	11.106	52,952	43,074	934,713
D_Port61	D_Viet21	D_China9	0.0557	6.079	48,963	43,796	924,557
D_Port61	D_Viet59	D_China8	0.0965	10.469	52,395	43,172	941,437
D_Port61	D_Viet59	D_China9	0.0481	5.589	47,976	43,576	920,022
D_Port61	D_Viet21	D_Tibet3	0.0663	6.823	49,504	43,352	922,535
D_Port61	D_Viet21	D_Tibet4	0.0599	6.610	49,756	44,134	914,979
D_Port61	D_Viet59	D_Tibet3	0.0575	6.277	49,026	43,693	929,474
D_Port61	D_Viet59	D_Tibet4	0.0513	5.871	49,127	44,333	921,845
D_Port61	D_Viet21	D_Mastif4	0.0714	7.278	49,690	43,069	894,260
D_Port61	D_Viet21	D_Mastif5	0.0643	7.229	49,757	43,747	913,367
D_Port61	D_Viet59	D_Mastif4	0.0639	6.791	48,007	42,240	898,585
D_Port61	D_Viet59	D_Mastif5	0.0587	6.929	48,421	43,054	917,851

Table S8. D-statistics for D(Outgroup, North American canids (wolf or coyote) and Taimyr wolf, PCD (AL2135 or AL3194), PCD (AL2135 or AL3194)). The Andean Fox was used as an outgroup for these analyses. n corresponds to the number of SNPs where all populations have data. Standard error for these statistics was obtained by performing a weighted block jackknife over 1Mb blocks. Results suggest that the most likely source of admixture into the Koster dog (AL2135) is a Mid Western coyote.

Pop 2	Pop 3	Pop 4	D	Z	BABA	ABBA	n
W_Mex1	AL2135	AL3194	0.0690	1.024	112	98	2,762
W_Mex1	AL3194	AL2135	-0.0690	-1.024	98	112	2,762
W_Yellow1	AL2135	AL3194	0.0808	1.304	124	106	2,758
W_Yellow1	AL3194	AL2135	0.0808	-1.304	106	124	2,758
W_Yellow2	AL2135	AL3194	0.0792	1.248	120	102	2,764
W_Yellow2	AL3194	AL2135	-0.0792	-1.248	102	120	2,764
C_Cal	AL2135	AL3194	-0.0004	-0.006	87	87	2,794
C_Cal	AL3194	AL2135	0.0004	0.006	87	87	2,794
C_MidW	**AL2135**	**AL3194**	**-0.1814**	**-2.034**	**43**	**62**	**1,863**
C_MidW	**AL3194**	**AL2135**	**0.1814**	**2.034**	**62**	**43**	**1,863**
TAI	AL2135	AL3194	0.1284	1.217	53	41	1,098
TAI	AL3194	AL2135	-0.1284	-1.217	41	53	1,098

Table S9. Sample code used as population name for SNP array analysis.

Table S10. Results of f4 ratio analysis depicted in Figure S18 (see Table S9 for population code).

Table S11. Results of f4 ratio analysis used to estimate proportion of PCD ancestry in Arctic breeds. This table include results for both SNP array and Genome-wide analyses.

Table S12. D-statistics for D(Outgroup, PCD (AL3194), Arctic dogs, Arctic dogs). The Andean Fox was used as an outgroup for these analyses. n corresponds to the number of SNPs where all populations have data. Standard error for these statistics was obtained by performing a weighted block jackknife over 1Mb blocks. Statistics with Z>3 and Z<-3 are shown in bold.

Pop 3 Pop 4 D Z BABA ABBA n

D_Mal68 D_Green 0.0001 0.010 38,564 38,555 926,919

D_Husky89 D_AHusky91 0.0059 0.416 29,524 29,174 749,696

D_Husky D_Husky89 0.0166 1.254 33,365 32,275 797,067

D_Husky D_AHusky91 0.0176 1.324 37,461 36,164 913,221

D_Husky89 D_Green 0.0437 3.261 38,311 35,101 790,543

D_Husky89 D_Mal68 0.0455 3.384 39,192 35,781 775,433

D_AHusky91 D_Green 0.0470 3.555 44,116 40,158 897,276

D_AHusky91 D_Mal68 0.0450 3.596 44,737 40,883 884,940

D_Husky D_Green 0.0619 5.021 48,948 43,246 955,940

D_Husky D_Mal68 0.0609 5.050 49,177 43,528 942,665

Table S13. D-statistics for D(Outgroup, Asian or European Dogs, Arctic dogs, Arctic dogs). The Andean Fox was used as an outgroup for these analyses. n corresponds to the number of SNPs where all populations have data. Standard error for these statistics was obtained by performing a weighted block jackknife over 1Mb blocks. Statistics with Z>3 are shown in bold.

Pop 2	**Pop 3**	**Pop 4**	**D**	**Z**	**BABA**	**ABBA**	**n**
D_Portt71	D_AHusky91	D_Husky89	0.0000	0.001	37,394	37,394	954,618
D_Portt71	D_Mal68	D_Husky89	0.0005	0.040	47,184	47,141	988,232
D_Portt71	D_Husky	D_Mal68	0.0009	0.086	58,360	58,259	1,190,049
D_Port61	D_AHusky91	D_Mal68	0.0016	0.149	55,075	54,905	1,143,740
D_Port61	D_Mal68	D_Husky	0.0018	0.186	59,906	59,692	1,217,824
D_Portt71	D_Husky89	D_Husky	0.0026	0.250	42,139	41,923	1,014,889
D_Port61	D_Husky89	D_Mal68	0.0029	0.271	48,766	48,481	1,013,144
D_Portt71	D_AHusky91	D_Mal68	0.0046	0.412	53,778	53,282	1,117,362
D_Port61	D_AHusky91	D_Husky	0.0060	0.586	48,323	47,748	1,179,281
D_Portt71	D_AHusky91	D_Husky	0.0060	0.596	46,884	46,321	1,151,944
D_Port61	D_Husky89	D_AHusky91	0.0077	0.675	38,794	38,202	979,116
D_Port61	D_Husky89	D_Husky	0.0098	0.948	43,569	42,725	1,040,592
D_Portt71	D_Green	D_AHusky91	0.0197	1.780	53,880	51,794	1,133,450
D_Portt71	D_Green	D_Husky	0.0242	2.401	59,660	56,844	1,207,389

D_Portt71 D_Green D_Mal68 0.0281 2.433 49,870 47,151 1,171,857

D_Portt71 D_Green D_Husky89 0.0267 2.466 47,462 44,995 1,007,512

D_Port61 D_Green D_Husky89 0.0283 2.625 48,938 46,241 1,033,241

D_Port61 D_Green D_AHusky91 0.0301 2.628 55,922 52,657 1,160,778

D_Port61 D_Green D_Husky 0.0341 3.348 61,488 57,440 1,236,161

D_Port61 D_Green D_Mal68 0.0375 3.355 51,677 47,947 1,199,625

D_China8 D_Husky89 D_AHusky91 0.0017 0.158 37,346 37,218 959,372

D_China9 D_Mal68 D_AHusky91 0.0040 0.424 52,229 51,810 1,085,883

D_China9 D_Green D_Husky89 0.0063 0.586 46,453 45,871 1,003,859

D_China8 D_Green D_Mal68 0.0066 0.628 48,381 47,747 1,175,269

D_China8 D_Husky D_AHusky91 0.0068 0.744 46,638 46,006 1,155,853

D_China8 D_Mal68 D_Husky 0.0071 0.780 58,261 57,441 1,193,177

D_China8 D_Husky D_Husky89 0.0078 0.820 42,094 41,441 1,019,519

D_China9 D_Husky D_Green 0.0081 0.846 57,408 56,484 1,178,134

D_China9 D_AHusky91 D_Husky 0.0085 0.957 45,878 45,108 1,120,029

D_China9 D_Mal68 D_Husky 0.0096 1.121 57,254 56,163 1,156,324

D_China8 D_Green D_Husky 0.0124 1.296 58,444 57,014 1,211,379

D_China9 D_Husky D_Husky89 0.0141 1.428 42,263 41,090 1,009,765

D_China9 D_AHusky91 D_Green 0.0152 1.493 52,310 50,746 1,106,207

D_China8 D_Mal68 D_AHusky91 0.0146 1.553 53,877 52,324 1,120,670

D_China8 D_Mal68 D_Husky89 0.0160 1.601 47,812 46,307 992,341

D_China8 D_Green D_AHusky91 0.0193 1.872 53,594 51,565 1,137,595

D_China9 D_Husky89 D_AHusky91 0.0208 1.901 37,871 36,329 950,029

D_China9 D_Mal68 D_Green 0.0205 2.167 48,324 46,381 1,142,965

D_China9 D_Mal68 D_Husky89 0.0224 2.288 47,913 45,810 982,870

D_China8 D_Green D_Husky89 0.0250 2.488 47,327 45,021 1,012,257

Table S14 Fitting of COSMIC mutational signatures 1, 5, 7 and Dog Germline signature to 1,933,897 CTVT tumor-only SNVs shared between 608T and 609T.

Mutational signature Fit Number of SNVs

(n=1,933,897)

1 Mean (15.4%) 297,820.1

Lower bound (15.2%) 293,952.3

Upper bound (15.5%) 299,754

5 Mean (40.1%) 775,492.7

Lower bound (40.0%) 773,558.8

Upper bound (40.3%) 779,360.5

7 Mean (38.95%) 753,252.9

Lower bound (38.9%) 752,285.9

Upper bound (39%) 754,219.8

Dog Germline Mean (5.5%) 106,364.3

Lower bound (5.3%) 102,496.5

Upper bound (5.7%) 110,232.1

Table S15 Number of clonal tumor-unique SNVs in 608T and 609T. CN, copy number. “All” indicates the complete set of tumor-unique SNVs, “Clonal”, only those that are clonal.

Sample Copy number state All Clonal

608T CN1 40 33

CN2 140 125

CN3 22 21

CN4 3 3

CN6 1 1

206 183

609T CN1 43 30

CN2 159 125

CN3 21 15

CN4 5 4

228 174

82

Table S16 Number of clonal N[C>T]G mutations unique to 608T and 609T, as well as the number found in the somatic lineage from the CTVT founder dog until MRCA

608T-609T

.

Variant set N[C>T]G mutations

Clonal 608T-unique 27

Clonal 609T-unique 23

CTVT (origin to MRCA

608T-609T

) 222,072

83