# Science

## AAAS

*This copy is for your personal, non-commercial use only.*

# Transmissible Dog Cancer Genome Reveals the Origin and History of an Ancient Cell Lineage

Elizabeth P. Murchison,[1,2]*† David C. Wedge,[1]* Ludmil B. Alexandrov,[1] Beiyuan Fu,[1] Inigo Martincorena,[1] Zemin Ning,[1] Jose M. C. Tubio,[1] Emma I. Werner,[1] Jan Allen,[3] Andrigo Barboza De Nardi,[4] Edward M. Donelan,[3] Gabriele Marino,[5] Ariberto Fassati,[6] Peter J. Campbell,[1] Fengtang Yang,[1] Austin Burt,[7] Robin A. Weiss,[6] Michael R. Stratton[1]†

Canine transmissible venereal tumor (CTVT) is the oldest known somatic cell lineage. It is a transmissible cancer that propagates naturally in dogs. We sequenced the genomes of two CTVT tumors and found that CTVT has acquired 1.9 million somatic substitution mutations and bears evidence of exposure to ultraviolet light. CTVT is remarkably stable and lacks subclonal heterogeneity despite thousands of rearrangements, copy-number changes, and retrotransposon insertions. More than 10,000 genes carry nonsynonymous variants, and 646 genes have been lost. CTVT first arose in a dog with low genomic heterozygosity that may have lived about 11,000 years ago. The cancer spawned by this individual dispersed across continents about 500 years ago. Our results provide a genetic identikit of an ancient dog and demonstrate the robustness of mammalian somatic cells to survive for millennia despite a massive mutation burden.

Canine transmissible venereal tumor (CTVT) is a naturally occurring transmissible cancer. It is a clonal cell lineage that spreads within the domestic dog population by the allogeneic transfer of living cancer cells, usually during coitus. The disease manifests itself with the appearance of tumors most often associated with the external genitalia of male and female dogs. The first known report of CTVT was made in 1810, when it was described by a London veterinary practitioner as "an ulcerous state, accompanied with a fungous excrescence" that arises in "organs concerned in generation" (*1*). It has subsequently been reported in dog populations worldwide (*2, 3*) and is, to our knowledge, the oldest and most widely disseminated cancer in the natural world.

We sequenced the genomes of two CTVT tumors, collected in Maningrida, Australia (24T), and Franca, Brazil (79T), as well as the genomes of their respective hosts 24H, an Aboriginal camp dog, and 79H, an American cocker spaniel (Fig. 1A). We also prepared metaphases for cytogenetic analysis from two CTVTs collected in Cape Verde and Italy. Metaphase fluorescence in situ hybridization (FISH) using red fox chromosomes as probes revealed massive karyotypic rearrangement in the CTVT genome, which was highly consistent between the two tumors analyzed (Fig. 1B). Despite the aneuploidy in CTVT detected by using cytogenetics, a copy-number analysis revealed that the genome is largely diploid [including a large proportion of the genome that is diploid with loss of heterozygosity (LOH)] and that there has been minimal change in copy-number status in 24T and 79T lineages since their divergence (Fig. 1C and tables S1 and S2). In contrast to human tumors, most of which contain several detectable subclones (*4*), presumably because of positive selection for newly acquired mutations conferring selective advantage (*5*), we found no evidence for subclonality in CTVT metaphases or copy-number plots. This suggests that CTVT is not undergoing positive selection at high frequency, possibly indicating that it is well adapted to its niche.

CTVT's massive burden of karyotypic abnormalities despite its largely diploid genomic copy number indicates that the genome has undergone large-scale copy neutral structural rearrangement. We found 2118 candidate somatic structural variants that were shared between 24T and 79T and 216 and 72 candidate somatic structural variants in 24T and 79T, respectively, that were unique to one tumor. We also searched for evidence of transposon mobilization in CTVT. We found 348 and 352 transposon insertions (involving both long interspersed nuclear elements and short interspersed nuclear elements) that were unique to 24T and 79T, respectively, and are thus likely to represent somatic retrotransposition events that occurred after the divergence of 24T and 79T.

We identified 3.04 million substitution variants in 24T and 2.77 million in 79T after removing all single-nucleotide polymorphisms (SNPs) known to segregate in the dog or the wolf germ line (including those identified in 24H or 79H). These variants will include somatic mutations as well as SNPs that have not been captured in previous canine sequencing efforts. We estimated the true number of somatic substitution mutations in CTVT by calculating the ratio of homozygous to heterozygous known SNPs in diploid regions that retain both parental chromosomes. Assuming that all homozygous variants in these regions are SNPs, we were able to estimate the number of unannotated heterozygous SNPs in each diploid segment by using the homozygous-to-heterozygous ratio among annotated SNPs. This analysis indicated that at least 65% of unannotated variants in CTVT are likely to be somatic, corresponding to a total of ~1.9 million somatic mutations in CTVT. A total of 103,667 and 109,119 variants were unique to 24T and 79T, respectively, the majority of which probably arose as somatic mutations after the two tumors' divergence because only 2056 and 5647, respectively, of these occur in regions that have been lost in the other tumor. Although a range of total mutation counts is observed in human cancers, the majority have between 1000 and 5000 somatic single-base-substitution mutations (*6*). Thus, CTVT has acquired several hundred times more somatic mutations than most human cancers.

To ascertain the processes responsible for the mutations in CTVT, we characterized CTVT's mutational spectrum and searched for known mutational signatures (*6*) in the CTVT genomes. The CTVT mutation spectrum was dominated by C>T (or G>A) mutations and CC>TT (or GG>AA) dinucleotide mutations (Fig. 2, A and B). Four mutational signatures were identified in CTVT (labeled A to D, Fig. 2C), which were sufficient to explain 98% of the mutations in CTVT (0.96 Pearson correlation between the mutation set observed in CTVT and the mutation set reconstructed by using four mutational signatures). The chemical events associated with some of these mutational signatures have been

[1]Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK. [2]Department of Veterinary Medicine, University of Cambridge, Madingley Road, Cambridge CB3 0ES, UK. [3]Animal Management in Rural and Remote Indigenous Communities (AMRRIC), Post Office Box 1464, Nightcliff, NT 0814, Australia. [4]Department of Clinical and Veterinary Surgery, São Paulo State University (UNESP), Via de Acesso Prof. Paulo Donato Castellane, s/n. CEP: 14884-900, Jaboticabal, São Paulo, Brazil. [5]Department of Veterinary Sciences, University of Messina, 98168 Messina, Italy. [6]Wohl Virion Centre and Medical Research Council Centre for Medical and Molecular Virology, Division of Infection and Immunity, University College London, Gower Street, London WC1E 6BT, UK. [7]Department of Life Sciences, Imperial College London, Silwood Park, Ascot, Berks. SL5 7PY, UK.

*These authors contributed equally to this work.
†Corresponding author. E-mail: epm27@cam.ac.uk (E.P.M.); mrs@sanger.ac.uk (M.R.S.)

characterized. Signature A is associated with germline SNPs, and its contribution to CTVT variant sets probably reflects incomplete removal of germline-inherited variants. Signature B defines the mutational signature characterized by C>T at CpG dinucleotides that is widely found in human cancers and known to be correlated with patient age (6). Signature C [known as signature 5 in (6)] is also frequently present in a spectrum of human cancers; however, its etiology is unknown (6). Signature D, which is characterized by C>T and CC>TT mutations and in humans is predominantly observed in cancers of the skin and known to be associated with exposure to ultraviolet light (6), explains 42% of the mutations in CTVT. These observations suggest that CTVT has been exposed at a low level to ultraviolet light during its evolution. Although CTVT tumors usually occur inside the genital orifice, they may be exposed to sunlight when they protrude from the vulva, ulcerate through preputial skin, or occur on external surfaces such as the skin or conjunctiva (for example, see 24T and 79T, Fig. 1A). Furthermore, it is the very cells that are exposed
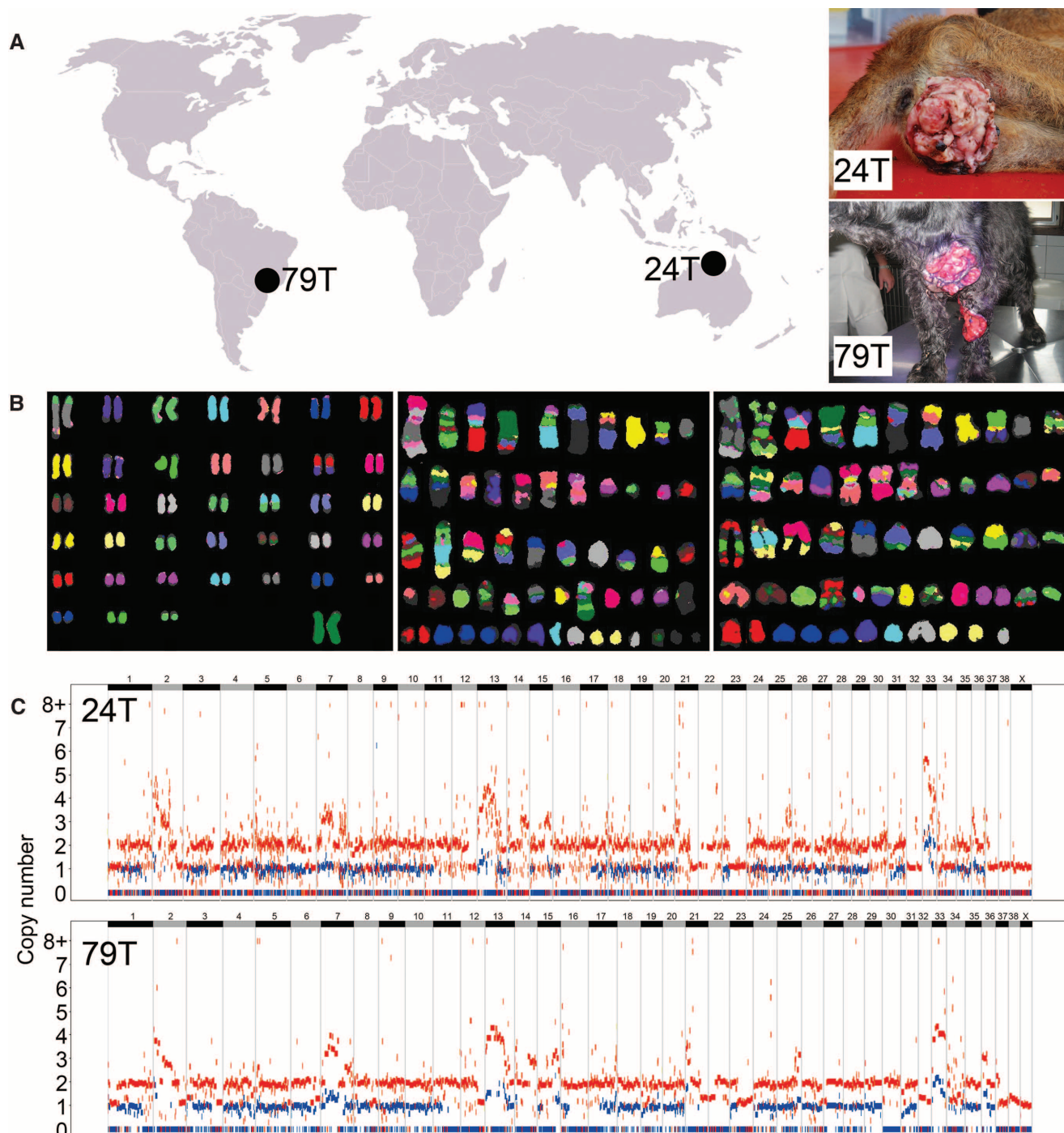


**Fig. 1. CTVT tumors, karyotypes, and copy number.** (**A**) Samples sequenced in this study. Both tumor (24T, 79T) and host (24H, 79H) DNA was sequenced from the two individuals shown. (**B**) Multiplex FISH using red fox probes to investigate karyotypes of a normal female dog (left) and CTVTs collected in Cape Verde (center) and Italy (right). (**C**) CTVT genomic copy number for 24T (top) and 79T (bottom). Red and blue points represent total copy number and minor copy number (i.e., copy number of the allele present in fewer copies), respectively, calculated by using normalized read counts at each of 2,544,508 SNP loci. Chromosomes are represented by horizontal alternating black and gray bars.

on the surface of a tumor that are most likely to propagate the CTVT lineage by passage to new hosts.

More than 10,000 genes (10,955 in 24T and 10,546 in 79T) in CTVT carry at least one non-synonymous substitution variant that is not a known germline SNP. Table S3 lists those genes that we consider to be the highest-confidence driver mutations in CTVT. These include a known rearrangement involving *MYC* (*7*), a homozygous deletion of *CDKN2A*, a hemizygous nonsense mutation in *SETD2*, and a rearrangement involving *ERG* that creates a potential in-frame *NEK1-ERG* fusion gene. A census of genes that have been lost in CTVT by homozygous deletion or hemizygous nonsense mutation indicated that at least 646 genes, 2.8% of the 22,874 protein-coding genes annotated in the dog genome, are collectively dispensable for survival and proliferation of a somatic cell (table S4).

We next sought to reconstruct the phenotype of the CTVT founder animal and to estimate the age of the cancer that it spawned by using the variants found in the CTVT genome. We compared the genotypes of 24T and 79T as well as 24H and 79H at 23,782 polymorphic SNP loci with those of 1106 previously genotyped dogs, wolves, and coyotes (*8*, *9*). The result, displayed by using principal components analysis (Fig. 3A), indicates that the CTVT founder animal was likely to have been a dog belonging to one of the ancient breeds [previous analyses were unable to distinguish between a wolf or an ancient-breed dog origin (*10*)]. Analysis of a pairwise distance tree indicated that, of the 86 breeds included in the analysis (*9*), the CTVT founder animal clusters most closely with Alaskan malamutes and

huskies (11 Alaskan malamutes have >95% probability after resampling genotypes of having one of the 16 genotypes closest to CTVT) (Fig. 3B and table S5). As expected, 79H clustered most closely with cocker spaniels within the modern breeds (>95% probability after resampling that each of the six closest genotypes are English cocker spaniels) (Fig. 3, A and B, and table S5). Host 24H, an Aboriginal camp dog, appears to have genetic contributions from both ancient and modern breeds (Fig. 3, A and B, and table S5).

Recent studies have mapped several loci conferring canine phenotypic features, such as coat color, morphology, and behavior (*11–21*). We examined the sequence of CTVT at a number of these loci to determine the likely phenotype of the founder animal (table S6). Our analysis indicated that the founder animal was likely to have been of medium or large size with an agouti or solid black coat. It carried a mixture of wolflike and doglike alleles at loci that have been linked to dog domestication (*21*). Tumors 24T and 79T each carried a single X chromosome and had no evidence of a Y chromosome, as found in a previous analysis (*22*). This is consistent with either a male (after somatic Y chromosome loss) or a female (after somatic X chromosome loss) founder animal (Fig. 3C). Analysis of genome-wide heterozygosity indicated that the founder animal was relatively inbred (Fig. 3D).

Previous studies have estimated that CTVT is between 200 and 70,000 years old (*10*, *23*). We sought to clarify the age of CTVT by using the mutations associated with mutational signature B (Fig. 2C), which is correlated with patient age at diagnosis in many human cancer types (*6*, *24*). We estimated that 492,533 mutations in CTVT are

likely to have been caused by this mutational process. By using the mutation rate of signature B in human medulloblastoma as a molecular clock [43.3 mutations of this signature genome-wide per year; we chose medulloblastoma because it is the human cancer with the closest correlation between number of mutations of this signature and patient age (*6*)], we estimate that CTVT may have first arisen about 11,368 years ago (lower and upper confidence intervals are 10,179 and 12,873 years, respectively). There is uncertainty in this estimate introduced by the possibilities that the accumulation of mutations of this signature is not clocklike in CTVT and that there are tissue- or species-specific differences in the rate of mutation accumulation between CTVT and human medulloblastoma. Applying this molecular clock to the mutations that occurred after divergence of the two tumors, we suggest that the most recent common ancestor of 24T and 79T may have existed about 460 years ago (458.2 years for 24T, 459.8 years for 79T) (Fig. 3E). We note that the estimated timing of this divergence coincides with the era of rapid human global exploration.

The founder animal whose somatic cells first gave rise to CTVT was an ancient-breed dog that may have lived about 11,000 years ago. The date of CTVT emergence, together with the structure of its phylogenetic tree (*23*) and evidence for both wolflike and doglike alleles at loci associated with domestication, is consistent with the possibility that CTVT may have first arisen within a genetically isolated population of early dogs whose limited genetic diversity facilitated the cancer's escape from its hosts' immune systems. Similarly, the Tasmanian devil facial tumor disease, the only other known naturally occurring

**Fig. 2. CTVT mutations.** Analyses were performed on a set of 395,306 CTVT variants that were annotated as somatic because of their heterozygous status within genomic regions that have undergone both LOH and duplication. (**A**) Simple mutation spectrum in CTVT. Mutations are labeled in pyrimidine context. (**B**) Dinucleotide mutation spectrum in CTVT. The first base was defined as the mutation with the lower chromosome coordinate, and the second base is immediately adjacent to the first base on the same strand. The strand is displayed relative to the pyrimidine context of the first base. A total of 3518 dinucleotide mutations were included in the analysis. (**C**) The proportion of mutations in CTVT explained by mutational signatures A to D.
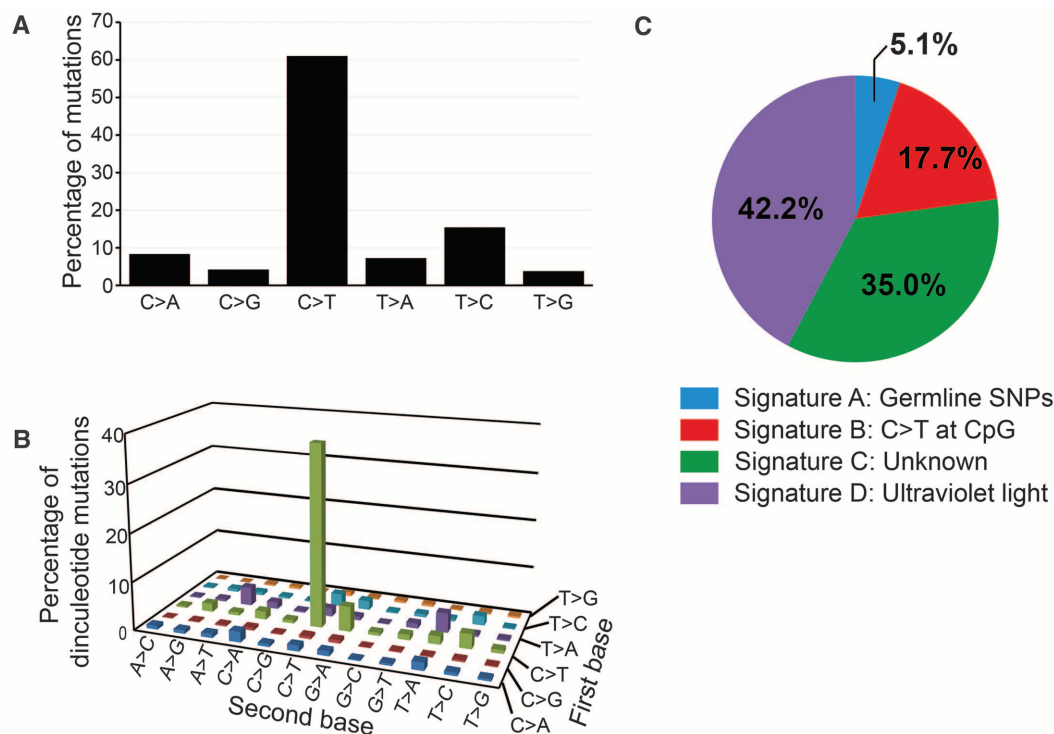
**Fig. 3. Tracing the CTVT founder animal.** (**A**) Principal components (PCs) analysis of 1106 wolves, dogs, and coyotes using genotypes at 23,782 polymorphic SNP loci (*8, 9*). Each individual is represented by a single colored dot, and positions of CTVT (inferred from the genotypes of 24T and 79T), 24H, and 79H are indicated. Breeds were classified as modern or ancient according to (*27*). (**B**) Positions of CTVT (left), 24H (center), and 79H (right) on pairwise distance tree comparing genotypes at 23,782 SNP loci with 1106 other dogs, wolves, and coyotes (*8, 9*). Only the closest breeds to CTVT, 24H, and 79H are shown. Breeds containing members that most strongly clustered with CTVT and 79H after genotype resampling are marked with red text and asterisks (see table S5; 24H did not cluster strongly with individuals from any single breed). NGSD, New Guinea singing dog. (**C**) Sex chromosome copy number of 24H (a female dog), 79H (a male dog), 24T, and 79T determined by counting the number of reads aligning to X and Y chromosomes and normalized to 79H. Y chromosome reads in 79T are likely to be derived from contaminating host DNA. (**D**) Proportion of annotated SNP loci in germline diploid regions that are heterozygous in 24H, 79H, 24T, and 79T. (**E**) Timeline for CTVT origin and divergence.
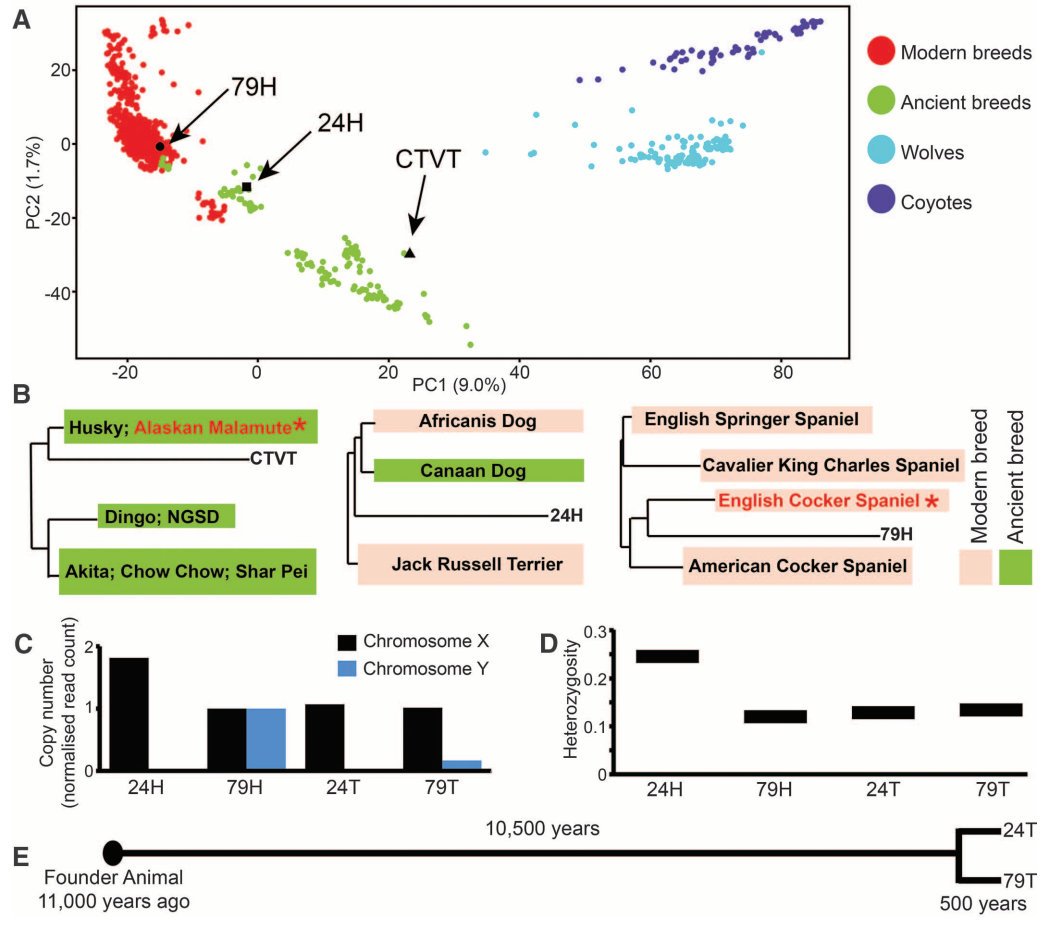


clonally transmissible cancer, arose in an island population with low genetic diversity (*25, 26*). Populations with limited genetic diversity may be particularly susceptible to the emergence and spread of transmissible cancers.

The CTVT genome has illuminated the origins, history, and evolution of the world's oldest known cancer. It is remarkable that a somatic genome whose DNA would normally have survived for no more than 15 years during the life of one dog has continued to exist for several millennia as a parasitic life form. CTVT's survival and global dominance are a testament to the ability of the mammalian somatic cell genome to adapt to and persist in a new ecological niche.

### References and Notes

1. D. P. Blaine, *A Domestic Treatise on the Diseases of Horses and Dogs* (T. Boosey, London, 1810), pp. 161–162.
2. E. P. Murchison, *Oncogene* **27** (suppl. 2), S19–S30 (2008).
3. B. Ganguly, U. Das, A. K. Das, *Vet. Comp. Oncol.*, published online 25 August 2013 (10.1111/vco.12060).
4. S. Nik-Zainal *et al.*, *Cell* **149**, 994–1007 (2012).
5. M. Gerlinger *et al.*, *N. Engl. J. Med.* **366**, 883–892 (2012).
6. L. B. Alexandrov *et al.*, *Nature* **500**, 415–421 (2013).
7. N. Katzir *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 1054–1058 (1985).
8. B. M. vonHoldt *et al.*, *Genome Res.* **21**, 1294–1305 (2011).
9. B. M. vonHoldt *et al.*, *Nature* **464**, 898–902 (2010).
10. C. Murgia, J. K. Pritchard, S. Y. Kim, A. Fassati, R. A. Weiss, *Cell* **126**, 477–487 (2006).
11. E. Cadieu *et al.*, *Science* **326**, 150–153 (2009); 10.1126/science.1177808.
12. S. I. Candille *et al.*, *Science* **318**, 1418–1423 (2007); 10.1126/science.1147880.
13. L. A. Clark, J. M. Wahl, C. A. Rees, K. E. Murphy, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 1376–1381 (2006).
14. N. H. Dodman *et al.*, *Mol. Psychiatry* **15**, 8–10 (2010).
15. R. E. Everts, J. Rothuizen, B. A. van Oost, *Anim. Genet.* **31**, 194–199 (2000).
16. B. C. Hoopes, M. Rimbault, D. Liebers, E. A. Ostrander, N. B. Sutter, *Mamm. Genome* **23**, 780–790 (2012).
17. S. M. Schmutz, T. G. Berryere, *Anim. Genet.* **38**, 539–549 (2007).
18. N. B. Sutter *et al.*, *Science* **316**, 112–115 (2007).
19. Y. Takeuchi *et al.*, *Anim. Genet.* **40**, 616–622 (2009).
20. A. Vaysse *et al.*, *PLOS Genet.* **7**, e1002316 (2011).
21. E. Axelsson *et al.*, *Nature* **495**, 360–364 (2013).
22. R. Thomas, C. Rebbeck, A. M. Leroi, A. Burt, M. Breen, *Chromosome Res.* **17**, 927–934 (2009).
23. C. A. Rebbeck, R. Thomas, A. M. Leroi, A. Burt, *Evolution* **63**, 2340–2349 (2009).
24. P. J. Stephens *et al.*, *Nature* **486**, 400–404 (2012).
25. E. P. Murchison *et al.*, *Cell* **148**, 780–791 (2012).
26. W. Miller *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 12348–12353 (2011).
27. G. Larson *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 8878–8883 (2012).