bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

On the ancestry and evolution of the extinct dire wolf

AUTHORS

Gregory L. Gedman[1]*, Kathleen Morrill Pirovich[1]*, Jonas Oppenheimer[1,2], Chaz Hyseni[1], Molly Cassatt-Johnstone[3], Nicolas Alexandre[1], William Troy[4], Chris Chao[1], Olivier Fedrigo[1], Savannah J. Hoyt[5], Patrick G.S. Grady[1,4,5], Sam Sacco[2], William Seligmann[2], Ayusman Dash[1], Mithil Chokshi[1], Laura Knecht[1], James B. Papizan[1], Tyler Miyawaki[1], Sven Bocklandt[1], James Kelher[1], Sara Ord[1], Audrey T. Lin[6,7], Brandon R. Peecook[8], Angela Perri[9], Mikkel-Holger S. Sinding[10], Greger Larson[11], Julie Meachen[12], Love Dalén[13,14,15], Bridgett vonHoldt[16], M Thomas P Gilbert[17,18], Christopher E. Mason[19], Rachel J. O'Neill[5], Elinor K. Karlsson[20,21], Brandi L. Cantarel[4], George R. R. Martin[22], George Church[1,20,23], Ben Lamm[1], Beth Shapiro[1]

*Co-first authors

[1]Colossal Biosciences, Dallas, TX 75247, USA
[2]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA
[3]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA
[4]Form Bio, Dallas, TX 75247, USA
[5]Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA
[6]Richard Gilder Graduate School, American Museum of Natural History, New York, NY, USA
[7]Department of Anthropology, National Museum of Natural History, Smithsonian Institution, Washington, DC, USA
[8]Idaho Museum of Natural History & Department of Biological Sciences, Idaho State University
[9]Chronicle Heritage, Phoenix, Arizona
[10]Dept of Biology, University of Copenhagen, Denmark
[11]Palaeogenomics and Bio-Archaeology Research Network, School of Archaeology, University of Oxford, Oxford, UK
[12]Anatomy Department, Des Moines University
[13]Department of Zoology, Stockholm University, Sweden
[14]Centre for Palaeogenetics, Stockholm, Sweden
[15]Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden
[16]Department of Ecology & Evolutionary Biology, Princeton University
[17]Center for Evolutionary Hologenomics, The Globe Institute, University of Copenhagen, Denmark
[18]University Museum, NTNU, Trondheim, Norway
[19]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA
[20]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[21]Genomics and Computational Biology, University of Massachusetts Chan Medical School, Worcester, MA, USA
[22]Fevre River Packet Co.
[23]Harvard Medical School, Wyss Institute.

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SUMMARY

Dire wolves (*Aenocyon dirus*) are extinct predators of Pleistocene North America. Although phenotypically similar to living wolves (*Canis lupus*), dire wolves have yet to be placed confidently in the canid family tree. We generated 3.4× and 12.8× paleogenomes from two well-preserved dire wolves dating to > 13,000 and > 72,000 years ago, and estimated consensus species trees for these and 10 canid species. Our results revealed that ~2/3 of dire wolf ancestry is derived from a lineage sister to the clade comprising the gray wolf, coyote, and dhole, and the remaining ~1/3 from a lineage near the base of Canini diversity. We identified 80 genes evolving under diversifying selection in dire wolves. Our results underscore the power of paleogenomes to resolve long-standing taxonomic questions and contribute to growing evidence of the role of post-speciation gene flow as an evolutionary force.

KEYWORDS

evolution, wolf, canid, lineage, admixture, selection, ancient DNA, genomics, paleogenomics, pangenomics

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

INTRODUCTION

The dire wolf *(Aenocyon dirus*) first appeared in the North American fossil record during the late Middle Pleistocene ~250,000 years before present (ybp) [1] and was the most prominent canid species in continental North America until its extinction ~12,900 ybp (radiocarbon calibrated) [2]. Their large size and morphological distinctiveness, as well as their abundance in fossil deposits at the Rancho La Brea Tar Pits (Los Angeles, California) has established dire wolves as iconic predators of the American Ice Age.

Dire wolves were morphologically similar to gray wolves (*Canis lupus*) with some exceptions. Dire wolves had larger temporalis muscles than gray wolves, capable of generating a greater bite force [3]. They also had proportionately longer bacula, suggesting sexual competition [4,5]. Dire wolves were comparable in size to the largest gray wolves – *C. l. occidentalis* in the northwestern United States [6] – and exhibited similar variation by latitude compared to gray wolves. Based on differences in size and geography, dire wolves are ascribed to two subspecies: the smaller ~60 kg *A. d. guildayi* in the west, and the larger ~68 kg *A. d. diru*s in the east [7,8].

Dire wolves are believed to have originated in North America [9–11], but their evolutionary history and taxonomic relationships to other canid species remain uncertain. A recent analysis of ~0.01× to 0.23× coverage dire wolf genomes concluded that the dire wolf lineage first diverged from wolf-like canids, including gray wolves (*C. lupus*), coyotes (*C. latrans*), African golden wolves (*C. anthus*), golden jackals (*C. aureus*), Ethiopian wolves (*C. simensis*), dholes (*Cuon alpinus*), African wild dogs (*Lycaon pictus*), side-striped jackals (*Lupulella adustus*), and black-backed jackals (*L. mesomelas*), during the late Miocene, ~6 million years ago (Mya) [12]. This timing coincided with the first appearance in the Eurasian and African fossil records of the genus *Eucyon,* which is the lineage from which extant wolf-like canid species probably descended [13].

During the Pleistocene ice ages, shifting habitats and the ephemeral emergence of the Bering Land Bridge provided repeated opportunities for wolf-like lineages to disperse in and out of continental North America. Coyotes first appear in the North American fossil record 1-2 Mya [14,15], and gray wolves are present by the late Middle Pleistocene [9,16], although current North American gray wolf lineage entered North America from Eurasia around 25,000 years ago [17–19]. Despite the geographic and temporal overlap between the ranges of dire wolves, coyotes, and gray wolves, no evidence of admixture between dire wolves and either lineage has been identified [12]. Since admixture between closely related canid species is common [20–22], the absence of gene flow between dire wolves and sympatric canids suggests temporal, geographic, reproductive, or behavioral isolation.

Although Perri *et al.* [12] confirmed that dire wolves are a distinct evolutionary lineage, a high frequency of phylogenetic incongruence among estimated gene trees led to uncertainties about their early evolutionary history. About 50% of reported gene trees supported a topology in which dire wolves formed the earliest-diverging lineage of the wolf-like canid clade (Canina). Of the remaining gene trees, the side-striped and black-backed jackals were found to be an outgroup to a clade comprising dire wolves and other wolf-like canids [12]. Divergence time estimates suggest the dire wolf lineage emerged after the divergence of the South American Cerdocyonina subtribe 5-4 Mya [12,23] and before the divergence of the common ancestors of African jackals (*Lupulella* spp.) 7.6-3.5 My and wolf-like canids 8.5-4.0 Mya (Canina subtribe). Incomplete lineage sorting [24] resulting from

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

such rapid diversification could be one driver of this incongruence. In addition, while at least one admixture event between dire wolves and the common ancestor of dholes, coyotes, and wolves was detected, additional gene flow from unsampled populations, like the South American canids, could not be excluded.

Resolving the phylogenetic position of dire wolves and assessing evidence for gene flow is difficult given the millions of years of divergence between distinct canid lineages and technical challenges of analyzing ancient DNA. While low coverage paleogenomes can be useful for phylogenetic and demographic inference [23,24], ancient DNA is susceptible to DNA damage, deamination, and reference bias in mapping, in which divergent reads fail to map and similarity to the reference genome is artificially inflated [25]. This bias results in an overestimation of sequence similarity between the ancient genome and the mapped reference, which can impact phylogenetic reconstruction and reconstructions of population history [26,27]. This bias is particularly problematic when the reference genome is evolutionarily distant from the paleogenome [28,29], as is the case for the dire wolf.

Several factors contributed to the eventual extinction of dire wolves [30,2,31]. The extinction of megafaunal prey species may have led to nutritional stress among dire wolves. The fossil record shows evidence of shrinking skull sizes during the Last Glacial Maximum [32], which can result from changes in diet and lifestyle. The loss of dire wolf prey and habitat from human-induced fires during the Late Pleistocene has been one proposed driver of these changes leading to their decline [2]. Reduced population sizes may also have led to diversity loss and the accumulation of deleterious genetic variation due to increased rates of inbreeding. In support of this hypothesis, ossification disorders are observed in limb bones dating to several thousand years prior to dire wolf extinction [31].

To better understand the origin, evolution, and extinction of the dire wolf, we leveraged improved paleogenomics approaches and recovered additional ancient DNA from two well-preserved dire wolves [12]: DireSP, an isolated incisor (Cincinnati Museum Center VP1737) recovered from Sheriden Pit, Ohio, estimated to be ~13,000 years old based on calibrated radiocarbon dates from other bone material from the site [33]. and DireGB, a complete isolated skull (Idaho Museum of Natural History, IMNH 48001/52) from Gigantobison Bay, a beach around the American Falls Reservoir, Idaho, dated to at least 72±14 thousand years ago based on the age of volcanic material from a nearby stratum [34]. We generated paleogenomes for both specimens and performed phylogenetic, admixture, and selection analyses that provide new insights into the evolution and adaptation of the dire wolf to its unique niche.

RESULTS

We generated paleogenomes with 3.4× depth of coverage of DireSP and 12.8× coverage of DireGB. After filtering, trimming, and merging overlapping reads, we retained 2.8 billion unique sequence reads for DireSP and 33.3 billion unique sequence reads for DireGB (**Table S1**). We mapped the reads using BWA *aln* against the Greenland gray wolf chromosome-level (including Y) genome assembly (GenBank GCA_905319855.2 aka “canLor”) [35] (**Table S2**). After mapping, 6% of the genome was covered >5× for DireSP and 82% for DireGB (**Fig. 1A-B**, **Table S3**). Previously, ancient DNA isolated from these same individuals yielded 0.2× (DireSP) and 0.3× (DireGB) coverage [12] (**Table S3**). We found patterns of DNA damage expected for single-stranded libraries (5′ C>T: DireGB = 30.2%; DireSP=26.6%, **Fig. S1**). Mapping to the gray

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

wolf reference sex chromosomes versus autosomes yielded Rx ratios [36,37] indicating that both individuals were male (95% confidence intervals: DireSP=0.539-0.554, DireGB=0.538-0.554).

We next used an iterative process to reconstruct a progressively less divergent nuclear paleogenome reference ("aenDir") for mapping dire wolf ancient DNA sequencing libraries. After six iterations, we identified 14,569,091 changes (13,665,448 single nucleotide variants, 388,514 insertions, and 515,129 deletions) from the seed reference assembly (gray wolf). We then called a pseudo-haploid, variant-based consensus sequence that represented the highest depth alleles. In parallel, we used established methods [28,38–40] to reconstruct the mitochondrial paleogenome at high coverage (average depth = 158.4×). Merging both nuclear and mitochondrial reconstructed paleogenomes resulted in a final reference assembly with a hard-masked N content of 16.84% (**Table S4**).

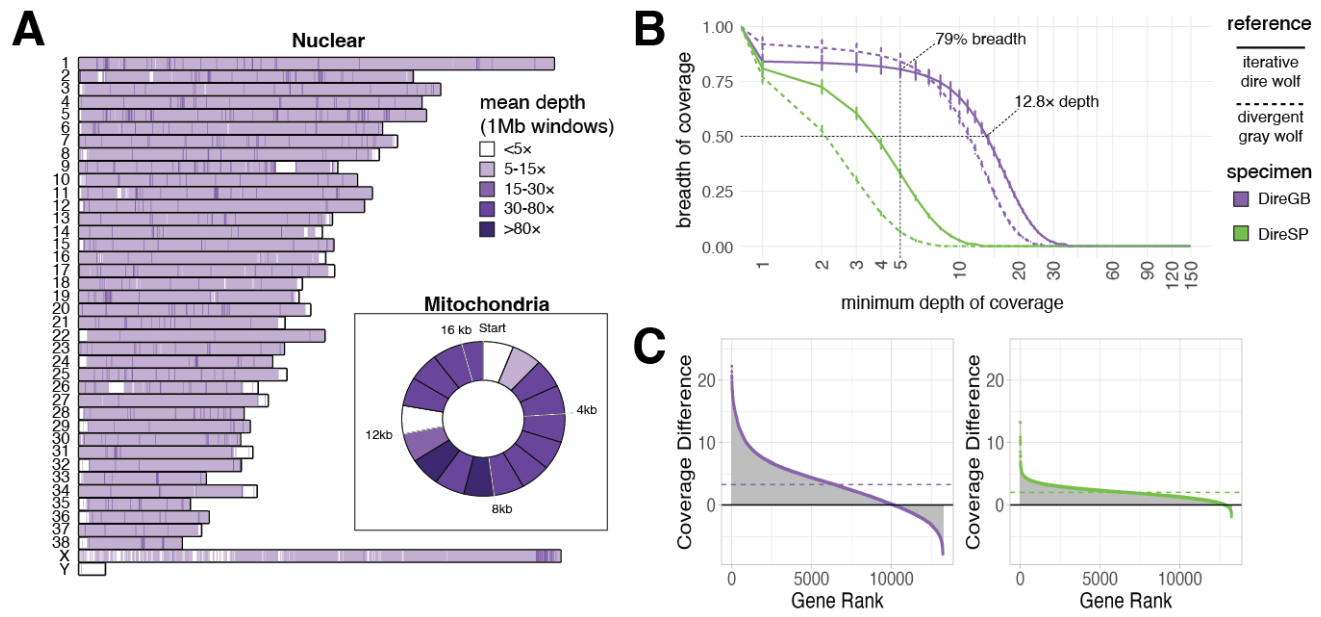


Figure 1. Iterative reconstruction of ancient dire wolf nuclear paleogenome improves whole genome and exome mappability. (**A**) Mean read depth of coverage by DireGB libraries (MQ>25) mapped over 1 Mb windows from the reference assembly of *A. dirus* reconstructed by iterative mapping and polishing. Inset displays mean read depth over the reconstructed full-length mitochondria sequence split by 1 Kb bins. (**B**) Cumulative coverage plot of DireGB (purple) and DireSP (green) libraries mapped to divergent gray wolf reference "canLor" (dotted line) and iterative dire wolf reference "aenDir" (solid line). Proportion of reference genome (i.e., breadth of coverage) at 5× and average depth for DireGB against the dire wolf reference are highlighted. (**C**) Average coverage difference over annotated orthologous genes (n=13,230) relative to dire wolf (positive) and gray wolf (negative). The majority of genes exhibited a boost in average coverage (dotted line) when using the reconstructed dire wolf paleogenome assembly in both DireGB (77% +3.3×) and DireSP (97% +2.1×).

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Mapping ancient DNA libraries to the reconstructed paleogenome reference (aenDir) increased coverage both at the whole genome level and at the exome level (**Fig. 1A-B**). Compared with the initial mapping to the Greenland gray wolf genome reference, mapping to aenDir resulted in an increase in average coverage of 2.8× for DireGB and 1.5× for DireSP libraries. Following annotation of the reconstructed paleogenome (aenDir), we found that the majority of protein-coding genes (71%) were retained compared to the gray wolf reference (canLor), with most missing genes due to masking of low coverage regions (**Figs. S2,3**). The depth of coverage in exons improved by ~2-fold for both DireSP and DireGB when mapped against the reconstructed dire wolf paleogenome (**Fig. 1C, Table S5**).

Mapping against the reconstructed paleogenome (aenDir) also reduced the impact of reference bias (**Fig. 2A**, **Table S6**). We found a significant decrease in allele sharing for the DireGB consensus called against the reconstructed dire wolf (aenDir) paleogenome compared to the gray wolf (canLor) genome, estimated using the *D*-statistic *D*(seed reference *Canis lupus*, non-seed reference *Canis lupus*; consensus *Aenocyon dirus*, outgroup *Vulpes vulpes*) (*D* for aenDir=-0.152, *Z*-score=-16.01; *D* for canLor=-0.521, *Z*-score=-76.51). In a complementary test, we assessed the identity-by-state (IBS) for DireSP reads mapped against the reconstructed aenDir paleogenome (**Fig. 2B**). We found that DireSP reads mapped to our reconstructed dire wolf genome exhibited higher IBS (i.e. greater sequence identity) compared to the gray wolf seed reference across all read lengths, especially for reads < 35 bp, suggesting the presence of residual reference bias, predominantly in shorter reads.

Dire wolf paleogenomes were unexpectedly more similar to extant Canina genomes than anticipated given a 5.8 Mya divergence time [12], a mutation rate calibrated for *Canis* species [41], and a wide range of generation times [42] for priors (**Fig. S4**). We computed sequence divergence on fixed autosomal transversions at sites covered at 5× relative to the total non-N sequence length of the reconstructed dire wolf paleogenome reference (2,310,286,013 bp). We found the dire wolf paleogenomes less diverged (on average, 0.09% divergent) than expected based on prior parameters (0.6% to 0.9% sequence divergence), and estimated sequence similarities of 99.91% to jackals and 99.94% to non-seed reference gray wolves at either end of wolf-like canid diversity (**Fig. S4A**). As residual reference bias in mapping could artificially inflate sequence similarity, we also assessed the distribution of unique 32-mers from the contamination-filtered sequence reads using *mash* for metagenome distance estimation [43]. We found that dire wolves exhibited equally low 32-mer diversity compared to gray wolf and jackal species (~0.01), with greater distance to red and gray foxes (0.02 to 0.03) (**Fig. S4B, Table S7**).

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

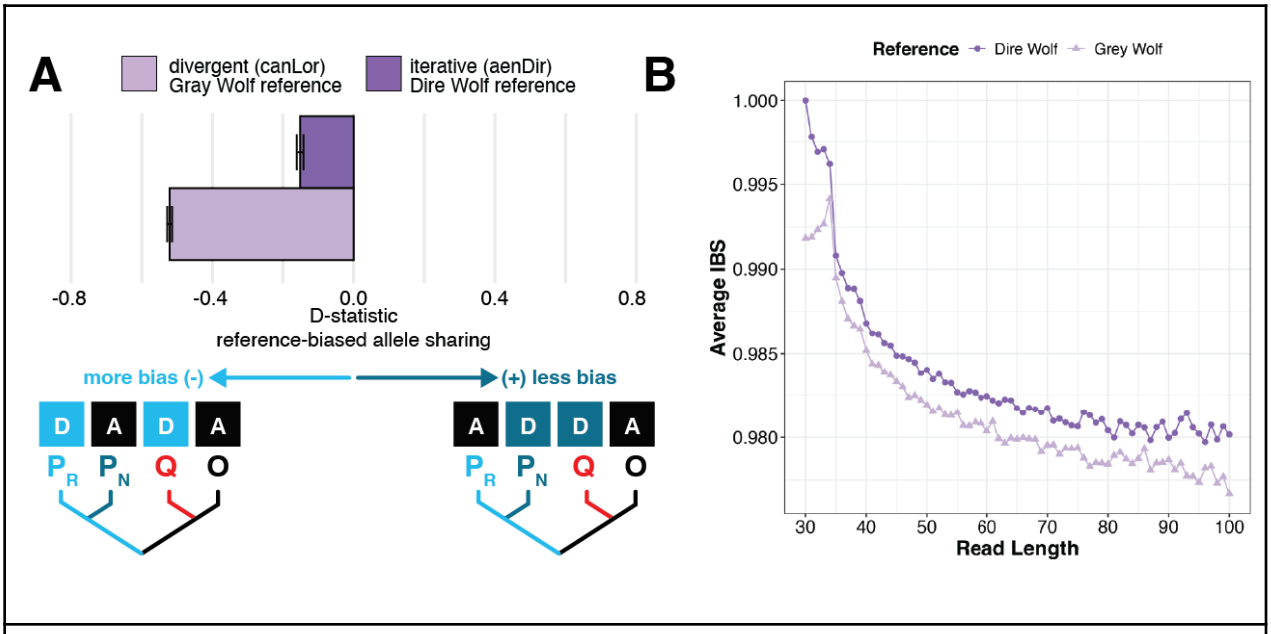


Figure 2. Iterative process to reconstruct the extinct dire wolf paleogenome sequence reduces reference bias. (**A**) Mapping to the reconstructed dire wolf paleogenome (aenDir) reduced reference bias 0.7 to 0.8 fold compared to mapping to the gray wolf reference genome (canLor). Residual reference bias was assessed using the ABBA/BABA test D statistic, which measures derived allele sharing of the DireGB consensus genome called against either the seed gray wolf reference genome (canLor) or reconstructed dire wolf paleogenome (aenDir) reference. Ancestral (A) and derived (D) alleles were localized by aligning either of the query (*A. dirus* in red, **Q**) consensus DireGB sequences (canLor- or aenDir-called), seed reference Greenland gray wolf (*C.lupus orion* in light blue, $\mathbf{P_R}$) and non-seed reference North American gray wolf (*C. lupus* in dark blue, $\mathbf{P_N}$) assemblies, and outgroup red fox assembly (*V. vulpes* in black, **O**) to a common reference genomic coordinates from a gray fox (*U.cinereoargenteus*) assembly. (**B**) Average identity-by-state (IBS) by read length for DireSP reads mapped to either the DireGB-based iteratively reconstructed paleogenome (dark purple) or divergent gray wolf (light purple) reference assemblies. IBS is defined as (matches - mismatch/aligned read length). Averages were calculated from a 500k read sample of the total alignment with an average of 5.7k (gray wolf) and 6.3k (dire wolf) reads per size bin.

Our estimated mitochondrial and nuclear phylogenies were similar to those published previously (**Figs. 3, S5-9, Tables S8-10**). The mitochondrial phylogeny showed the same mitonuclear discordance that is known to characterize the canid lineage [22]. Our nuclear phylogeny, which was based on 19.5Mb of non-coding dire wolf sequence at least 1Mb away from protein-coding genes and <25% missingness on average (**Figs. S6,7**), placed dire wolves into a monophyletic clade that diverges prior to the split between the black-backed jackal and wolf-like canids (**Figs. 3, S8**). However, this placement was only supported by 44% of input loci, while 39% of loci supported a closer relationship between dire wolves and non-Lupulellan wolf-like canids (**Fig. 3**). Divergence time estimates (**Fig. S9, Table S10**)

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

suggested that the dire wolf diverged from other canids ~4.5 million years ago (95% CI=5.84-2.89 Mya) and the jackals diverged ~4.08 million years ago (95% CI=5.34-2.58 Mya).

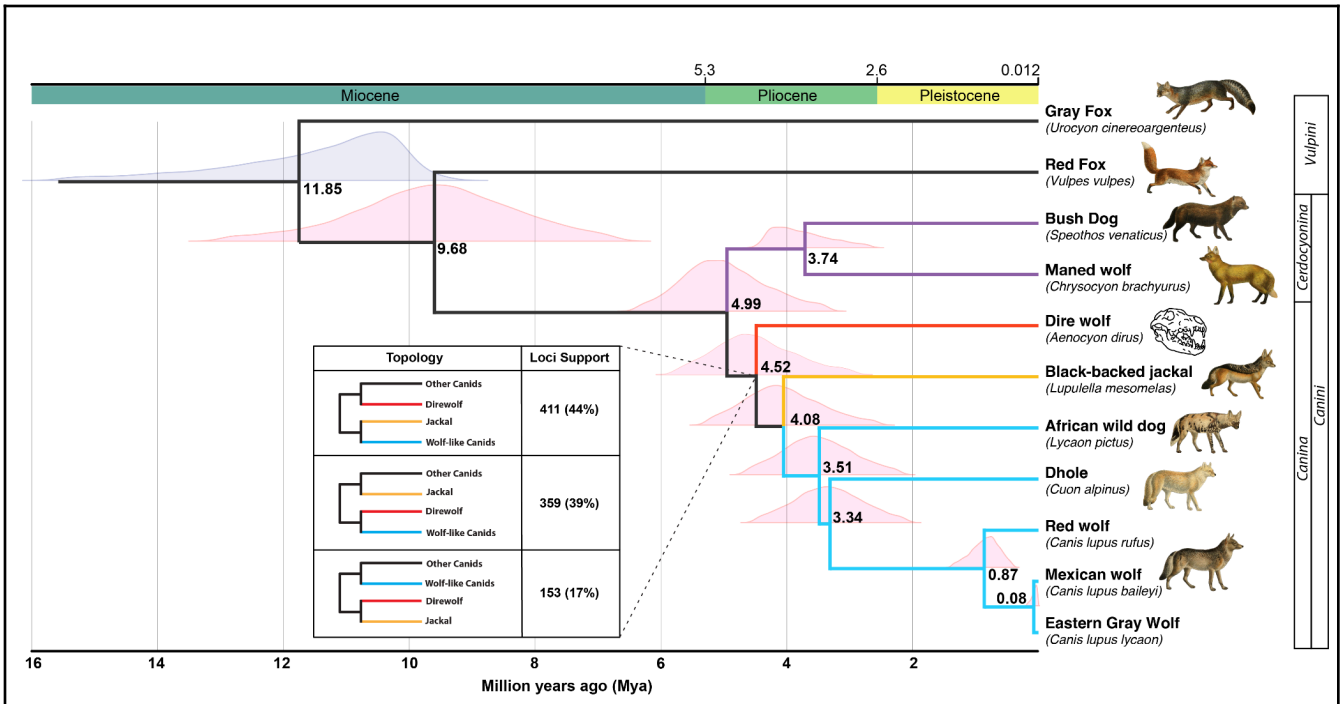


Figure 3. The dire wolf lineage is distinct from jackals and other wolf-like canids. Time-calibrated nuclear phylogenetic relationship of extinct dire wolves to extant canids using MCMCtree from 25 kb loci trees (n=924) using IQTree and species tree consensus using ASTRAL. Dire wolves (red) diverge prior to the split of black-backed jackals (yellow) and wolf-like canids (blue). Root ages are mean divergence time estimates in millions of years (Mya). Node distributions visualize confidence in placement. Node distributions with fossil calibrations are shown in purple. Inset details the alternative topologies for the dire wolf node, along with the proportion of loci supporting each alternate.

The discordance among estimated gene tree topologies appears to arise from admixed ancestry in dire wolves (**Fig. 4**). Dire wolves shared an excess of derived alleles with gray wolves relative to other canids, including dholes ($11.9 \leq Z \leq 15.3$), African wild dogs ($61.9 \leq Z \leq 65.4$), African jackals ($92.3 \leq Z \leq 99.6$), and South American canids ($173.4 \leq Z \leq 183.7$) (**Fig. 4A**). We did not detect similar affinity between dire wolves and gray wolves when comparing to wolf-like canids that are more closely related gray wolves, including coyotes and red wolves ($|Z| \leq 2.4$) (**Fig. 4A**). We also observed that, relative to dire wolves, African jackals ($66.0 \leq Z \leq 72.4$) and other wolf–like canids ($Z \geq 92.7$) share more derived alleles with gray wolves (**Fig. 4B**). Together, these patterns suggest that dire wolves had a mixture of ancestries from two distinct lineages, with one diverging prior to African jackals and another stemming before gray wolves and coyotes. Allele sharing between dire wolves and South American canids relative to extant wolf-like canids ($-7.2 \leq Z \leq -2.3$) is consistent with the deeper ancestry component falling within Cerdocyonina diversity (**Fig. 4B**). Both dire wolves

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

formed a clade relative to all individual coyotes and gray wolves (**Fig. S10**), implying that there was limited gene flow between dire wolves and specific *Canis* lineages following their divergence.

Admixture graph models of canid population history identified both wolf-like canid and deep ancestry components in the dire wolf lineage (**Fig. 4C**). We applied a heuristic algorithm [44] to find optimal admixture graphs modeling the relationships between seven canids (dire wolves, Eurasian gray wolves, coyotes, dholes, African wild dogs, maned wolves, using red foxes as an outgroup). No model permitting zero or one admixture event was supported by the data; however, the best-fitting model featuring two admixture events had statistical support (worst residual of Z=2.6) (**Fig. 4C**). This model found that dire wolves received a large portion of their ancestry (~39%) from a lineage that groups with South American canids, with the remaining ancestry (~61%) attributed to a sister lineage to dholes, coyotes, and gray wolves. This wolf-like ancestor may have been the source of the phylogenetically-discordant dire wolf mitochondrial genome, which groups within wolf-like canid diversity. When analyzed separately, both dire wolf individuals produced similar optimal graphs (**Fig. S13**), suggesting that admixture was a species-wide feature of dire wolf population history. Graphs that include both an additional outgroup (gray foxes) and additional South American canid species confirmed that the divergent portion of dire wolf ancestry falls near the base of Canini diversity (**Fig. S13**).

The phylogenetic placement of the wolf-like component in dire wolves is remarkably similar to a proposed unsampled canid lineage that contributed ancestry to the common ancestor of coyotes and wolves [45]. The best-fitting model similarly features gene flow from dire wolves into wolves and coyotes, which could be the source of this unsampled ancestry. However, scenarios in which dholes, rather than coyotes and wolves, are admixed also provide similar fits to the data (**Table S11**) and we observed that relationships among extant wolf-like canids can be modeled without requiring this unsampled admixture (**Fig. S12**). As suggested by evidence of excess derived allele sharing between dire wolves and South American canids, a similar model which instead places the minor ancestry component of dire wolves outside extant Canini diversity instead of on the South American canid branch provides a worse fit for the data (worst residual of Z=5.25). Models of canid population history that include additional extant canid lineages highlight other known and cryptic admixture events and identify the gene flow event leading to dire wolves (**Fig. S14**). We found mixed support for introgression from the dhole lineage into the African wild dog lineage, which could reflect cryptic admixture [45,46], and consistent support for other previously detected admixture events [20,45,47,48]. These findings reinforced that this approach is sensitive enough to detect hybridization across diverse canids (**Fig. S18**), and does not appear to be impacted by choice of reference genome (**Fig. S13**).

The ubiquity of gene flow among canids, especially among diverged wolf-like canid species [45–49], poses challenges for inferring the precise placement of admixture events among these taxa. Our goal was to infer the simplest models which described the data, both for the minimal (**Fig. 4C**) and expanded sets of canids (**Fig. S14**), which each provided evidence for the dual ancestry of dire wolves. However, given extensive hybridization among canids, the true population history may be more complex. More complicated models, such as those featuring three admixture events in the minimal canid species set, also largely suggest hybridization in dire wolves, though a much wider range of graphs fit the data, including many which have other lineages, rather than dire wolves, as admixed (**Table S11**). Challenges

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

arising from mapping sequence data diverged taxa also complicate inferences of canid population history. We observed several biases across alignment and genotype calling strategies (**Fig. S10, S19**). The Consensify-style pseudohaploid genotype calls using pangenome alignments presented here appeared to minimize these biases (**Fig. S13, S19)**, though do require greater agreement between reads, which has the potential to amplify biases arising in mapping.

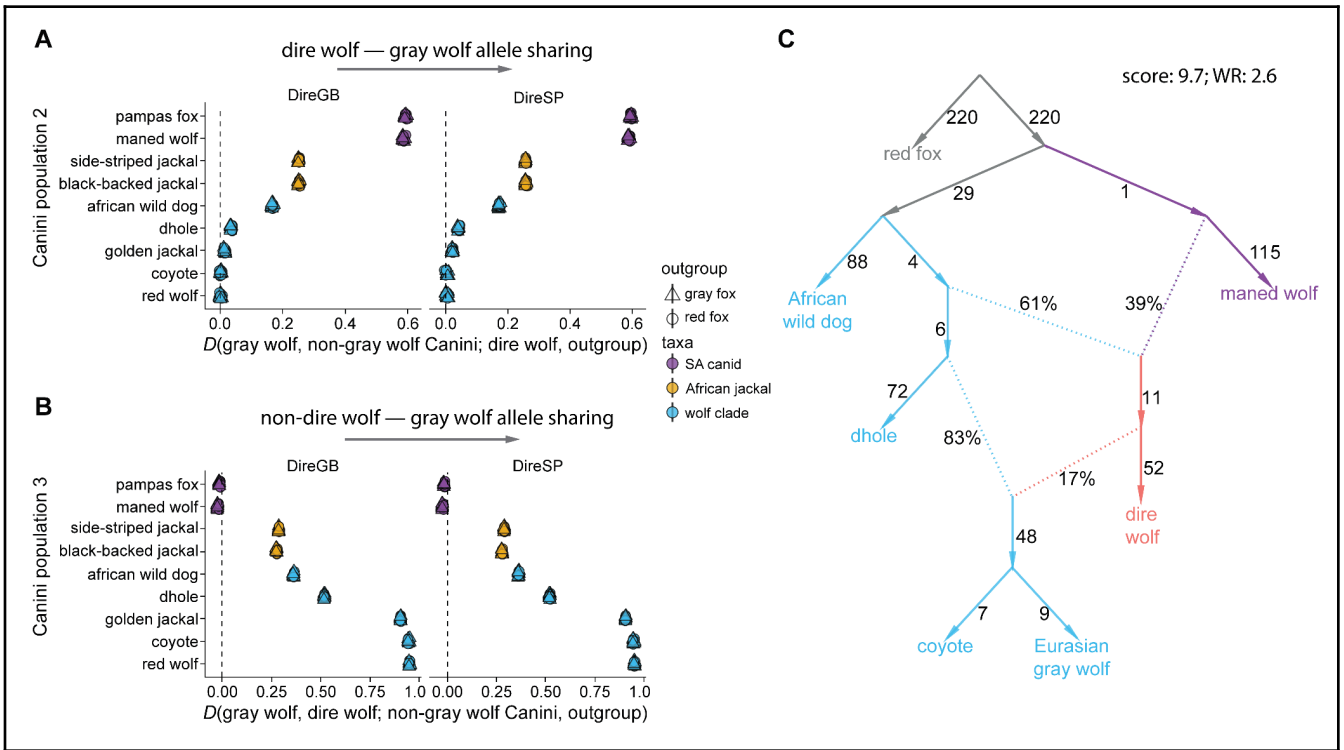


Figure 4. Investigating gene flow between dire wolves and other canid lineages. (**A**) D-statistics of the form *D*(gray wolf, other Canini group; dire wolf, red fox), investigating allele sharing between dire wolves and gray wolves, relative to other species and subspecies within Canini. (**B**) *D*-statistics of the form *D*(gray wolf, dire wolf; other Canini group, red fox), investigating allele sharing between dire wolves and other populations within Canini, relative to gray wolves. Bars depict 3 standard errors in (**A**) and (**B**). Each point is a comparison with a different individual gray wolf, with points colored by the Canini clade being of the non-gray wolf population in the comparison. (**C**) Best-fitting admixture graph with two admixture events modeling the relationship between dire wolves (n=2); wolf-like canids: Eurasian wolves (n=2); coyotes, (n=4), dholes (n=2), African wild dogs (n=3); and maned wolf (n=2); with red fox (n=2) as the outgroup. Additional graph fits are summarized in Table S11.

Comparative sequence analysis detected molecular signatures for diversifying and adaptive evolution in dire wolf protein-coding genes (**Fig. 5**). We used the HyPhy framework [50–52] to identify genes with signatures of positive selection among dire wolves and 10 other canid taxa, including episodic diversifying selection in dire wolves contrasted to other canids. We identified 49 genes under significant (BUSTED $p < 0.05$) episodic positive selection and another 31 genes under non-episodic positive selection in the dire wolf branch (free-ratio dN/dS > 2.0) (**Fig. 5, Table S12**). None of these genes exhibited pervasive positive selection over the canid tree (one-ratio dN/dS < 2.0), although dire wolves may share positive selection at a site level with other branches. Rarely, we found that the dire wolf codon matched a codon

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

found in another branch. For example, the 505th codon of *NCAPG* (MEME *p*= 3.31e-06) had two branches under positive selection: the root codon encodes serine (TCT) but both the dire wolf and *Canis* branches, and not other wolf-like canids, encode asparagine (AAT). Likewise, the 794th codon of *WNK4* (MEME *p*= 0.00513), encoded asparagine (AAC) in both dire wolf and *Lycaon* branches rather than the root alanine (GCC). We did not find extensive functional enrichment for gene sets across the 80 genes under diversifying positive selection in the dire wolf branch. Functional domain analysis [53] and protein language modeling [54] identified 42 dire wolf codons in 22 genes under episodic diversifying selection that intersect protein functional domains which could be impacted by amino acid substitution (**Table S13**).

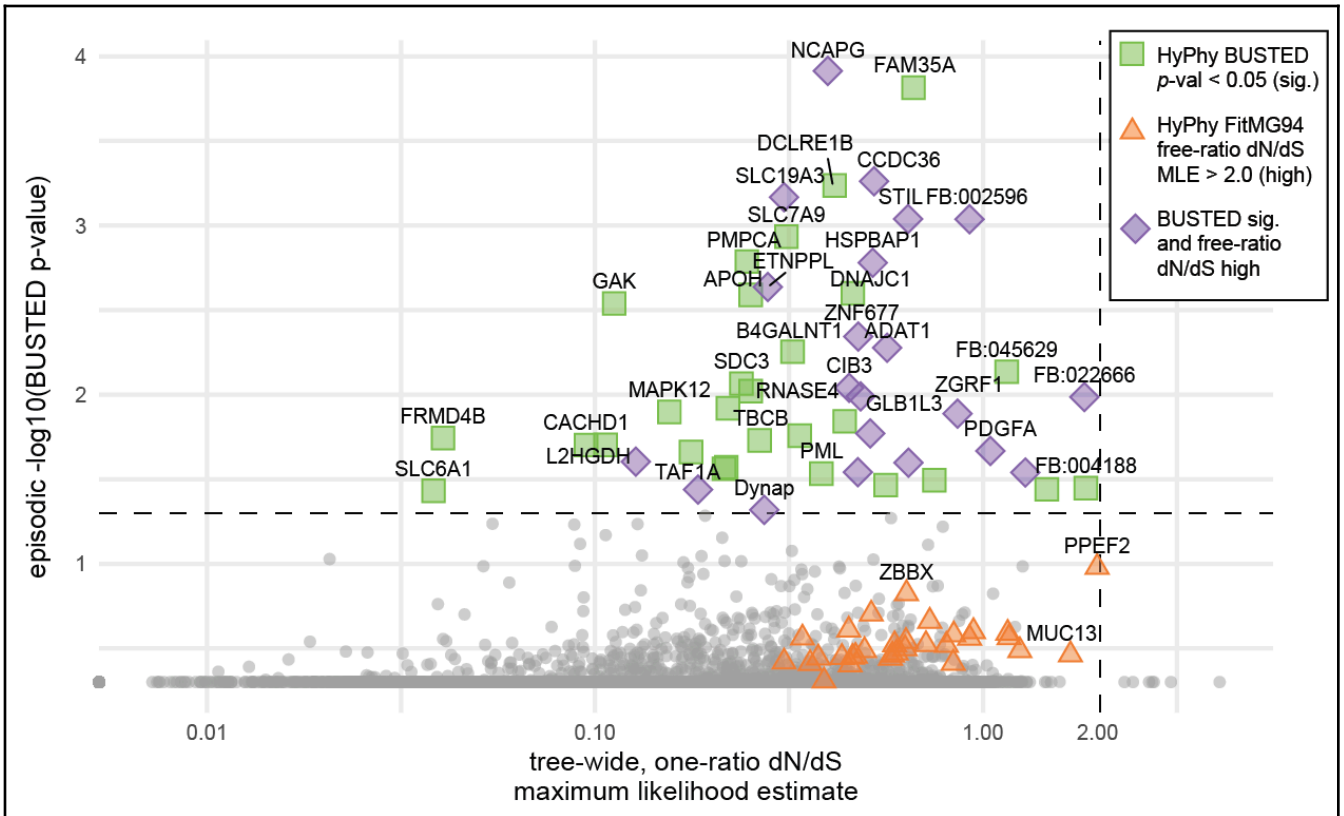


Figure 5. Comparative analysis of nuclear protein-coding gene sequences supported by dire wolf ancient DNA and 10 other canid genomes inform pathways under episodic or pervasive diversifying positive selection that may shape species adaptations and phenotypes. The -log10 HyPhy BUSTED *p*-values for episodic positive selection in the dire wolf branch plotted (y-axis) over the canid tree-wide, one-ratio dN/dS ratio maximum likelihood estimates (x-axis) for 5,984 genes where high one-ratio dN/dS implies pervasive positive selection across the canid lineage. We find 31 genes exhibiting branch-specific, free-ratio dN/dS > 2 (orange triangles), 28 genes significant for branch-specific, episodic selection in BUSTED test $p < 0.05$ (green squares), or 21 genes satisfying both conditions (purple diamonds). Annotated genes without gene symbols of orthologs identified are labeled with "FB:".

DISCUSSION

Dire wolf evolutionary history

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The Late Pleistocene dire wolf is descended from a canid lineage that emerged following admixture between two lineages: a lineage along the branch that gave rise to the South American canids (Cerdocyonina subtribe) after they diverged from African jackals and other wolf-like canids (Canina subtribe), and a more recently derived wolf-like canid lineage that diverged after the emergence of the African wild dog (*Lycaon*) lineage (**Figs. 3**, **4C**). This admixed ancestry explains the discordance in nuclear gene phylogenies observed both here and previously reported in Perri et al. (2021).

Given fossil evidence that dire wolves were only present in North America, the ancestral components of the dire wolf lineage might be attributed to other canids in the North American fossil record. The oldest wolf-like fossils in the North American fossil record are from the tribe Canini [10], including the extinct genus *Eucyon* (10.3-3.6 Mya), the proposed progenitor of *Canis* and *Lupulella*, and the earliest ancestors of the Cerdocyonina subtribe, which dispersed to South America [55–57]. It would be plausible, given an estimated divergence timing straddling Late Miocene and Pliocene, that the major ancestry component of the dire wolf is from an early North American *Eucyon* species, but our admixture graph modeling supports instead a major ancestor diverging after the *Lycaon* lineage (African wild dogs), which would be post *Eucyon* expansion outside of North America. *Eucyon* expanded into Eurasia and Africa during the Late Miocene, where it gave rise to *Lupulella*, the African jackals [57–61], as well as *Canis* and the related genera *Lycaon* and *Cuon* [46]. From the late Miocene through the Pleistocene, various *Canis* species expanded back into North America, supplanting or admixing with local populations [13]. *Canis lepophagus* expanded into the Americas during the late Miocene to early Pleistocene [10,62], for example, followed by *C. armbrusteri* and *C. edwardii* in the Pliocene and Pleistocene. The wolf-like lineage that contributes majority ancestry to dire wolves could be one of these earlier expanding lineages.

We note that the deep divergence (~10 Mya) among the canid lineages examined here makes it challenging to examine gene flow across these species while accounting for reference bias, particularly with highly fragmented ancient DNA. While our iterative mapping approach mitigated reference bias in our alignment of dire wolf sequences, it is possible that our more fragmented dire wolf reference reduced mapping performance and amplified other biases in the alignment of short reads from non-dire wolf canids. Thus, we complemented analyses of canid population history with a pangenome-mapping approach which mitigated these biases. Additionally, the prevalence of interspecific gene flow among canids makes it unlikely that any particular model will fully capture the underlying population history relating these species [47]. However, our conclusion regarding the dual ancestry of dire wolves is robustly supported across different models and approaches.

Our results underscore the role of post-divergence gene flow as an evolutionary force shaping species across the tree of life, and the potential for lineages to be forged through hybrid speciation – as uncovered in the Columbian mammoth (*Mammuthus columbi*), the same megafauna that roamed alongside dire wolves [63,64]. For example, excess basal ancestry in wolves and coyotes relative to other wolf-like canids was previously attributed to gene flow with an unknown wolf-like canid lineage that emerged following the divergence of African wild dogs [65], raising the possibility that the same unknown lineage interbred with both groups. Indeed, our results suggest that the wolf-like lineage found within dire wolves also contributed ancestry to wolves and coyotes, possibly mediated through gene flow with dire wolves. However, we infer that gene flow between diverged canid lineages was common, and that the relationships within the wolf-like canids can be modeled without invoking this

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

unknown lineage. Our admixture graph analyses do, however, uncover previously described episodes of gene flow between canids, resulting in admixed ancestries for species including eastern coyotes [48], red wolves [47], golden wolves [20,45], and dholes [46].

Evolutionary adaptations of dire wolves

We identified several genes in the dire wolf genome that may have contributed to one of their primary evolutionary features: their large builds. *NCAPG,* a gene associated with body size, was the most significant gene for episodic diversifying selection in the dire wolf branch. Yet, dire wolves and gray wolves exhibit unusually low rates of synonymous substitutions in *NCAPG* (both dS= 4e-10) – magnitudes lower than other canids (all dS > 0.001) – and together share a serine-to-asparagine substitution under site-specific positive selection (MEME *p*= 3.31e-06). This degree of sequence constraint could suggest strong purifying selection [66–69], perhaps upon non-coding regulatory elements influencing expression of adjacent gene *LCORL*, just 200 kilobases away. *LCORL* is an allosteric regulator of the polycomb repressor complex (PRC2) [70], and haplotypes spanning *NCAPG* and *LCORL* have been strongly associated with growth traits in domestic dogs[71], goats[72], sheep[73], horses[74], and cattle[75,76]. Expression levels of *NCAPG* and *LCORL* have inverse relationships on growth: positively correlated for *NCAPG* levels [77] and negatively correlated for *LCORL* levels [74]. We hypothesize that in both dire wolves and gray wolves, purifying selection has maintained ancestral regulatory elements at the *NCAPG* locus, while in other wolf-like canids of smaller body sizes, diversifying selection altered co-regulation of these genes.

We also found several genes under episodic diversifying selection in dire wolves annotated as affecting germ cell development, gamete function, and reproductive tissue structure, which may contribute to the reproductive biology of dire wolves distinct from wolf-like canids [4]. The genes *B4GALNT1*, *ROS1*, *TP73*, *AKAP3*, *NUTM1*, and *ABCG5* are all associated with “male infertility” in the Mammalian Phenotype Ontology (MP:0001925). While most of these fertility-related genes have not been implicated in reproductive structures, *ROS1* regulates the differentiation of male epididymis [78]. *PDGFA*, another gene under positive selection in dire wolves, has been shown as integral to epididymal morphology in mice [79]. None overlap the few genes associated with mammalian baculum size, however [80], meaning that genes shaping divergent reproductive biology in dire wolves are still a mystery.

An iterative approach to nuclear genome inference

We showed that iterative mapping combined with conservative consensus calling can help mitigate the impacts of reference bias and DNA damage in paleogenomics. To generate the dire wolf paleogenome, we mapped reads to the Greenland gray wolf as a seed reference, called consensus from aligned dire wolf reads, then re-mapped reads to this new consensus. We then repeated this process until no additional reads were added. Our intent was to estimate a conservative assembly, and thus we chose to hard-mask bases (i.e., Ns) not supported by >5× coverage, resulting in a final reference assembly comprising only the most abundant alleles from mapped dire wolf reads. Methods like EAGER use a similar conservative consensus calling approach when mapping to the nuclear genome, but employ only a single round of mapping and thus do not capture the maximum number of reads [81]. Iterative assembly tools are used in ancient DNA research, but mostly for mitochondrial genome assembly (e.g., MIA [82]) as they do not scale to whole nuclear genomes [83]. Iterative mapping was used to reconstruct the Denisovan genome, where Meyer et al. (2012) [38] performed two

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

rounds of iterative mapping and then filled in low coverage regions with human sequence, which is a source of reference bias. Given the improvements that we observed with multiple rounds of iterative mapping, we recommend a similar approach, in particular when no close reference genome is available.

One concern with an iterative mapping approach is that the process could actually exacerbate reference bias rather than reduce it. If the seed reference is too divergent, for example, only the most similar ancient reads will map, potentially masking informative sites and compounding bias. We approximated reference bias using *D*-statistics [38,84] by assessing derived allele sharing with distinct gray wolf subspecies, and found no evidence of compounding bias. In fact, we observed a reduction in bias towards the seed reference Greenland gray wolf genome. Since shorter reads have been shown to be more impacted by reference bias [40,85], we performed a secondary assessment of identity by state (IBS) by read length from an unrelated dire wolf individual (DireSP). While residual reference bias remains, particularly for shorter reads, we observed higher IBS on average for all DireSP reads regardless of length, suggesting we successfully generated a less divergent reference genome that facilitated better mapping of ancient reads.

While our approach estimated a paleogenome that was more similar to the dire wolf relative to the seed reference, steps could be taken to improve these genomes further. For example, while our approach can identify and incorporate SNPs and small insertions or deletions into the final reference, it cannot account for larger structural variants specific to the dire wolf. The reconstruction of large structural variants is unresolved for short read data sets that are typical of paleogenomes [86] and, given our average read length of 69 bp across all samples, is not achievable for the dire wolf. Nevertheless, by iteratively estimating a paleogenome reference, we could recover more reads that would have otherwise missed mapping to a divergent reference genome without those small sequence differences.

An alternative – or, perhaps, complementary – approach to mitigating reference bias is to construct a pangenome from multiple divergent species for initial mapping [87,88], which allows for capture of highly divergent reads and, potentially, larger structural variants. The benefits of a pangenome approach were demonstrated by Martiniano and colleagues (2020) [89], who saw a reduction of reference bias in mapping simulated ancient reads to the 1000 Genomes Project variational graph. By mapping real reads from ancient humans to this pangenome, they could capture sequence variation missed by initial mapping to a single reference without a loss of mapping sensitivity. However, divergent dire wolf haplotypes not represented by the available paths in the pangenome will still be affected by reference bias. Nevertheless, cost-efficient optimizations that enable a combination of iterative mapping and polishing approach will provide additional improvements to estimating increasingly accurate paleogenomes from short read data sets.

These data provide a comprehensive overview of dire wolf genomics, evolution, and adaptations, offering insights into this iconic species. The reconstructed reference paleogenome provides a resource for the broader research community and a roadmap for future studies aiming to leverage high-coverage ancient DNA. The genes identified under selection in dire wolves offer insights into possible molecular mechanisms for this species' role as a predator of extinct megafauna. This work furthers our understanding of the dire wolf and their place in the broader patterns of canid evolution.

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

STAR METHODS

Published ancient DNA sequencing data

We downloaded ancient DNA sequencing reads from dire wolf individuals previously sampled, including the Sheriden Pit individual (DireSP, BioSample accession SAMEA5559854) and the Gigantobison Bay individual (DireGB, BioSample accession SAMEA5559855) and three other individuals from American Falls Reservoir, Idaho (DireAFR, BioSample accession SAMEA5559852), Natural Trap Cave, Wyoming (DireNTC, BioSample accession SAMEA5559856), and Guy Wilson Cave, Tennessee (DireGWC, BioSample accession SAMEA5559853) publicly released under BioProject accession PRJEB31639 (**Table S1**).

Provenance and sampling of dire wolf individuals

We obtained permission to collect new samples from two dire wolf individuals that previously yielded authentic dire wolf mitochondrial and nuclear DNA [12]. The first dire wolf, referred to as "DireSP" (Cincinnati Museum Center; VP1737; ACAD1735; DireSP), is from Sheriden Pit site in Indian Trail Caverns, Ohio, and an incisor tooth previously sampled at the root (BioSample accession SAMEA5559854). We sampled an additional incisor from this individual for the present study. The second dire wolf, referred to as "DireGB" (Idaho Museum of Natural History; IMNH 48001/52; ACAD18742; DireGB) is from Gigantobison Bay site at American Falls Reservoir, Idaho, and presents a complete cranium with intact teeth. The left petrosal was previously sampled (BioSample accession SAMEA5559855) and the right petrous was sampled for this study. The Sheriden Pit individual could have been deposited before 13,000 years ago, based on calibrated radiocarbon dates from other bone material derived from the same layer [33], but has not been radiometrically dated, and the Gigantobison Bay individual is probably older than 72,000 years, based on the age of volcanic material from a nearby stratum [34].

Sample preparation and ancient DNA extraction

For the Sheriden Pit individual (BioSample accession SAMN46779696, we extracted ancient DNA from an incisor (Cincinnati Museum Center VP1737). We used a Dremel with drill attachment to obtain bone powder from the interior of the root. We rinsed the obtained bone powder in a 0.5% bleach solution for fifteen minutes followed by three water washes before proceeding with extraction. For the Gigantobison Bay individual (BioSample accession SAMN46779697), we extracted ancient DNA from the incus bone recovered from the right petrous of a complete skull (IMNH 48001/52). Ossicles are an optimal source of DNA preservation [90,91]. We rinsed the whole incus in a 0.5% bleach solution, followed by three water washes. We extracted DNA for both individuals using the Buffer D option, a protocol optimized for recovering short, degraded molecules [92]. For the Sheriden Pit sample, we extracted DNA from two 5 mg aliquots of bone powder following bleach pre-treatment. From these two DNA extracts we generated three single stranded libraries each [93]. Following the initial round of extractions we observed a non-dimineralized bone powder pellet in one of the

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

two lysate tubes and performed a re-digestion of this bone powder following Rohland et al. (2018) [92]. From the DNA isolated after this re-digestion we generated two more single-stranded libraries [93]. For the Gigantobison Bay individual, we extracted DNA from the whole incus, which was completely dissolved following overnight lysis. We then prepared single-stranded libraries from the extracts as per Kapp et al. (2021) [93].

We generated a single extract from the Gigantobison Bay ossicle, from which we made six libraries. All pre-PCR work, including sample preparation, DNA extraction, library construction, and quality control sequencing, was carried out in a dedicated ancient DNA lab at the University of California, Santa Cruz.

Massively parallel paired-end short DNA read sequencing

We prepared a total of nine Illumina libraries from the sample of the Sheriden Pit dire wolf (DireSP) and six Illumina libraries from the sample of the Gigantobison Bay dire wolf (DireGB). We sequenced all libraries on a Nova-seq 6000 using version 1.5 chemistry, loading 0.3nM final library concentration. For the DireSP sample, we obtained 1,320gb (R1+R2) of raw sequencing data or 13.2 billion total paired end reads (R1+R2). For the DireGB sample, we obtained 7,814gb (R1+R2) or 78 billion total paired end reads (R1+R2). We report the total number of raw sequencing reads generated per run, library, and individual in **Table S1**. For subsequent steps of parallelized processing, we split sequencing reads into batches of 50,000,000 reads.

Sequence read pre-processing and quality control

We trimmed raw sequencing reads using *fastp* [94] (*i*) to remove adapter sequence, low quality ends of reads, and reads less than 35 nucleotides, (*ii*) merge and correct overlapping reads with a minimum of 15 nucleotide overlap, and no more than three differences, and (*iii*) trimmed homopolymer G and X sequences at the 3' end (-q 25 -l 35 -m --overlap_len_require 15 --overlap_diff_limit 3 -c -g - x). We classified sequence contaminants using *Kraken 2* [95] and a confidence of 0.8 against the pre-compiled 8 GB database MiniKraken DB_8GB constructed from complete bacterial, archaeal, and viral genomes in RefSeq (2017-10-18). We used combine_kreports.py from *Kraken Tools* [96] to merge all metagenomic classification reports generated for split read batches. We report statistics on sequencing reads before and after processing for both published and novel ancient DNA sequencing libraries in **Table S1**.

Reference genome assemblies and annotations

We selected the Greenland gray wolf (*Canis lupus*) chromosome-level (including Y) genome assembly (GenBank GCA_905319855.2) [35] as our primary single-reference genome assembly of a modern canid species, and utilized additional canid genome assemblies [97–114] throughout this study as referenced in **Table S2**. We surmised that the Greenland gray wolf would serve as a suitable scaffold reference due to the geographic isolation of Greenland wolves relative to other gray wolves [115], the fully assembled chromosomes including Y and high contig N50 relative to other canid genome assemblies (**Table S2**), and to previously inferred phylogenies that suggest it is no more distantly related to dire wolves than any other extant members of the Canina subtribe.

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Read mapping to single reference genomes

We utilized two primary mapping strategies for the majority of the analyses in this study. Strategy 1 was a CPU-based accelerated Sentieon® implementation of the *BWA* `aln` algorithm [116] and seed lengths of `-l 16,500` nucleotides, maximum edit distances set by a fraction `-n 0.01` of missing alignments, and `-o 2` maximum gap opens [117]. Strategy 2 was a GPU-accelerated NVIDIA® Clara™ Parabricks `fq2bam` implementation of the *BWA* `mem` algorithm with settings for short reads (`-M -k 19 -r 2.5`).

Similarly, when mapping ancient reads, we used two strategies for removing duplicate reads and assessing DNA damage. First in the "standard" strategy, we used the *PALEOMIX* [118] `rmdup_collapsed` function and duplicate removal using *SAMtools* rmdup. We performed base recalibration with *mapDamage* [119] to rescale base quality scores in consideration of 12 nucleotides from both 5' and 3' ends of sequencing reads (`--rescale-length-5p 12` and `--rescale-length-3p 12`). We profiled Ancient DNA damage patterns using DamageProfiler [120]. The second strategy was "stringent," where we merged each sequencing run across libraries and marked duplicates again using `MarkDuplicates` from the Picard toolkit (broadinstitute.github.io/picard) to remove any remaining duplicate reads at the library level. After additional duplicate removal, we ran mapDamage again to assess overall DNA composition, read length distributions, fragmentation and misincorporation rates and C-to-T and G-to-A transition frequencies at relative positives across reads.

We assessed alignment statistics using *samtools* `idxstats` and `flagstat`, *mosdepth* [121], and *QualiMap* [122]`bamqc` after removal of marked duplicates: (*i*) *depth of coverage* as the # of unique reads mapping across all reference bases reported as a histogram and the mean # of unique reads mapping, (*ii*) *breadth of coverage* as % bases of reference genome covered by given #s of unique reads reported as a histogram and the % reference at 5× coverage by unique reads, and (*iii*) *physical coverage* or cumulative length of post- filtering and metagenomic classification of collapsed reads expressed as a multiple of the expected 2.8B nucleotides composing the total canid genome size reported from flow cytometry in coyotes, gray wolves, and domestic dogs [123,124]. As *QualiMap* does not exclude low mapping quality reads, unlike *mosdepth,* alignment statistics may be overestimated.

Sex inference of ancient specimens

We found support for both DireGB and DireSP individuals as male after mapping libraries to the Greenland gray wolf reference using *BWA* 'aln' with stringent duplication removal (see *Read Mapping 1B*). We calculated the Rx ratios [36,37] between chromosome X depth and autosomal mean depth of coverage and 95% confidence interval (CI) lower and upper bounds for both individuals, and designated a 95% CI upper bound under 0.6 as male and a 95% CI lower bound over 0.8 as female [36,37]. The 95% CIs for the ratios of reads aligning to chromosome X versus autosomes placed between 0.539 to 0.554 for DireSP and 0.538 to 0.554 for DireGB.

While the ratios of reads mapping to the reference gray wolf X chromosome (1.0× for SP and 5.7× for GB) and autosomes (2× for SP and 10× for GB) supports these individuals as male, we did not observe similar coverage over the reference gray wolf Y chromosome (0.1× for SP and 0.4× for GB), which may be unsurprising: Y chromosomes are notoriously difficult to assemble given highly repetitive regions and high homology to the X chromosome, and is roughly 1/20th the size of the X chromosome, so mapping sequencing reads from ancient

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DNA to the Y chromosome may be less stable than mapping to the X chromosome. Therefore, we took a closer look at reads mapping to Y-encoded genes as a secondary confirmation.

We examined depth of coverage by dire wolf reads at the Y-encoded gene *TSPY* because it exhibits relatively high copy number variation among modern gray wolves [125] and has high copy numbers in both humans and cattle [126,127]. Thus, for males we expect to see a relatively higher abundance of reads aligning to *TSPY* than other windows of the assembled Y chromosome sequence. We identified coordinates for the *TSPY* gene annotated on the Greenland gray wolf (*C. lupus*) assembly (GCA_905319855.2) at chrY:594247-596596 (ENSCAFG00865029519 in our gene annotation). We then calculated the depth of coverage by collapsed dire wolf reads aligned using *bedtools* [128] and *mosdepth* [121] across *TSPY* and 10,000 randomly sampled 2,349 nucleotide windows from assembled chromosome Y and 10,000 windows randomly sampled from any assembled autosome. We observed a high depth of coverage over *TSPY* (0.20× in DireSP and 0.72× in DireGB), relative to depths of coverage from equally-sized 2,349 nucleotide windows randomly sampled 10,000 times over the assembled gray wolf Y chromosome (0.103× ± SD: 0.397× in DireSP and 0.362× ± SD: 1.05× in DireGB), which, while lower than read depth over randomly sampled windows across autosomes (1.87× ± SD: 0.578× in DireSP and 10.6× ± SD: 2.86× in DireGB), supported the presence of male sex chromosomes.

Multi-canid pangenome construction and read mapping

Ancient DNA mapped to multi-species pangenomes is expected to reduce reference bias from mapping to a single, divergent reference [89]. Thus we generated a multi-species pangenome of canids using the graph-based *Minigraph-Cactus* (MC) pipeline [129], which starts from a reference assembly and iteratively introduces structural variation from aligned genome assemblies and, by excluding centromeric and highly repetitive sequences, can be more efficient for mapping sequencing read data [130]. We constructed the graph using the Greenland gray wolf (GenBank GCA_905319855.2) as our initial reference genome to be consistent with the coordinate system of our single reference alignments, and incorporated sequence variation from 6 chromosome-level assembled genomes that had the highest contig N50 and most complete set of chromosomes from two *Canis* spp. (domestic dog and dingo) and four other canid species and 3 scaffold-level assemblies from dhole (*Cuon alpinus*), bush dog (*Speothos venaticus*), and red fox (*Vulpes vulpes*) for broad phylogenetic representation of the canid lineage (**Table S2**). We implemented a custom script (`buildPangenome.sh`) launched from the *cactus docker image* (v2.7.1) to construct a pangenome bidirectional sequence variation graph. We built a graph index and mapped unclassified, trimmed reads using *Giraffe* [131] included in v*g tools* [132]. Only reads mapped to the mitochondria (MT) were extracted and utilized for dire wolf MT reconstruction. For downstream analysis, we surjected reads aligned to the multi-canid pangenome graph to the Greenland gray wolf (canLor) reference assembly. We observed that in pangenome alignments by *Giraffe*, short reads were occasionally aggressively clipped, resulting in a small fraction of bases within the overall read aligning. This phenomenon was most pronounced in the dire wolf individuals. We therefore filtered pangenome alignments to remove reads which had more than 20% of their bases soft-clipped, and reads which had a non-clipped length of less than 35 bases.

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Paleogenome reconstruction via iterative mapping and polishing

We reconstructed the dire wolf nuclear genome by iteratively mapping and polishing to a divergent seed reference, the Greenland gray wolf (GenBank GCA_905319855.2), using ancient DNA to progressively assemble a less divergent reference sequence. We performed initial polishing from the DireGB BAM file (see Read Depth) by employing the *Pilon* polisher [133] in diploid mode to correct sequences for insertions, deletions, and single nucleotide polymorphisms (SNPs) and perform local reassembly of each chromosome or contig (`--changes --fix all --diploid`).

For subsequent iterations, we allowed each library from DireGB to map to the polished sequences of autosomes, sex chromosomes, and unassembled contigs from the previous iteration, excluding the mitochondrial sequence which we assembled separately. We mapped unclassified and trimmed paired-end reads, merged and unmerged, using the *BWA* `mem` and standard deduplication with *PALEOMIX* (see Read Mapping 2A). Unmapped reads were discarded, libraries merged, and sequence polished. We carried out this process for 6 iterations in total and evaluated alignment statistics and quality metrics for each iteration (**Table S4**).

In order to reduce the computational load of assembling the mitochondrial genome and reduce reference bias in mapping, we first mapped the complete libraries of DireGB to the canid Minigraph-Cactus pangenome to identify mitochondrial reads in our data. We surjected all reads that mapped to mitochondrial scaffolds of the canid pangenome to the Greenland gray wolf mitochondrial (MT) sequence and reverse complemented the MT sequence to orient to strand with the D-loop control region. We then filtered reads that were unmerged, unmapped, and not in primary alignment using samtools view -F 261 and removed duplicates with Picard `MarkDuplicates` function. The filtered and duplicate-removed mitochondrial BAM was the final input from which we extracted reads using *samtools* `fastq` function (-n), resulting in 29,104 reads from the DireGB libraries.

We used *MIA* [82], an iterative mapping approach, to assemble the dire wolf mitochondrial genome with the gray wolf genome as the seed reference (NCBI Accession NC_008092.1), giving flags (-c -C -k 14 -U) and -s with an ancient DNA substitution matrix. We filtered the consensus assembly with “relaxed”, requiring 3× coverage and 66% consensus. All sites that did not pass these filters were set to ‘N’. For all downstream analyses, we used the relaxed filtered assembly (**Fig. S5**). We estimated a 34.5× coverage mitochondrial consensus genome for DireSP and a 90.171× coverage mitochondrial genome for DireGB, with 3.35-21.26% N content (**Table S3**).

We performed a final mapping of ancient DNA sequencing libraries to nuclear sequences resulting from the 6th iteration of mapping and polishing plus the iteratively assembled mitochondrial sequence. We aligned using *BWA* `aln` with standard *PALEOMIX* deduplication (Read Mapping 1A). We called pseudo-haploid consensus sequences by random read-sampling consensus and by variant-based consensus. In both cases, sequences below 5× unique read depth were rendered as hard-masked Ns, and we required a minimum base quality of BQ=30 and mapping quality of MQ=25. We performed random read-sampling using *ANGSD* [134] allele counting (`-doCounts 1 -minQ 30 -minMapQ 25 -dumpCounts 3`) followed by *Consensify* [135] to call consensus from reads mapped at 5× to 30× depth and 4/5 allelic support (`-min 5 -max 30 -n_matches 4 -n_random_reads 5`). We called a variant-based consensus to consider read depth and quality scores at variant and invariant sites. We called genotypes

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

using *GATK* [136] *HaplotypeCaller* emitting in gVCF mode, then converted to VCF using bcftools convert, filtered variant calls (minimum QUAL=20, BQ=30, MQ=25, DP=5). In the case of overlapping alleles at variant sites, we enforced each genotype to represent homozygous genotype of the allele with highest depth, which should minimize bias from DNA damage-induced cytosine deamination producing false variant calls, using the remove-overlaps plugin for *bcftools* (https://github.com/samtools/bcftools/blob/develop/plugins/remove-overlaps.c). For variable sites, overlapping variants with the highest QUAL values were retained, and for overlapping sites that contain invariant sites, we generated pseudo-QUAL values using the highest QUAL value of a known variant site scaled by the depth of the invariant site prior to filtering. Any sequences not covered in the resulting VCF were hard-masked as Ns in the final consensus.

Gene annotation and orthology analysis

We annotated genes using the fully automated *FLAG* tool (Troy et al. 2023) implemented in the `FLAG: Eukaryote Gene Annotation` workflow on the Form Bio platform (formbio.com). For the reconstructed dire wolf assembly, we included fully assembled transcript data from domestic dog (GCA_000002285.4), gray wolf (GCA_905319855.2), and dingo (GCA_003254725.2) as evidence for *Liftover*, *Augustus*, and *Helixer*. After employing a novel combine and filter algorithm, gene symbols were attributed to each model using *Entap*. We performed a BUSCO analysis using default settings and reported the total number of orthologs found in the mammalian database (n=9226). We assessed mean depth of coverage by all aligned short read libraries over all annotated genes (length-weighted average of mean depth per exon) in the reconstructed dire wolf genome sequence using mosdepth (**Table S5**).

Consensus calling from iterative reconstruction

We mapped short read genomes of modern canid species sourced from publicly available data sets (**Table S8**), the new sequencing libraries from DireGB and DireSP individuals, and the previously published libraries of dire wolf individuals [12] to the reconstructed dire wolf genome reference (aenDir) using *BWA* `mem` with standard deduplication with 'Mark Duplicates' (Read Mapping 2A).

We called consensus sequences using a random read-sampling approach. We used *ANGSD* (`-doCounts 1 -dumpCounts 3`) to quantify depth of sequence variants with a minimum mapping and sequence quality of 25. We used these counts to generate a consensus genome using Consensify for the modern (`-min 10 -max 500 -n_matches 8 -n_random_reads 10`) and ancient (`-min 5 -max 30 -n_matches 4 -n_random_reads 5`) read sets. Note in both cases, a variable site was incorporated as a change to the reference only if 80% of alternative reads samples were in agreement. Such random sampling of ancient data also decreases the likelihood of incorporating false positive variants from DNA damage signals [138].

Post hoc analysis of residual reference bias

We assessed residual reference bias as a disproportionate rate of derived allele sharing with the seed reference Greenland gray wolf (*C. lupus orion* aka "canLor") genome (GenBank

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

GCA_905319855.2) [35] relative to the genome of a distinct gray wolf subspecies (*C. lupus* aka "canLup"). We employed the ABBA/BABA *D*-statistic for introgression [38,84] to quantify derived allele sharing. We designated ancestral (A) alleles as those shared with outgroup species gray fox (*Urocyon cinereoargenteus* aka "uroCin") and/or red fox (*Vulpes vulpes* aka "vulVul"), where for site patterns of derived (D) alleles considered concordant under the expected species tree, `(((canLor,canLup), aenDir), outgroup/vulVul/uroCin)` would be `(((D,D),A),A)` or `(((D,D),D),A)` and discordant as `(((D,A),D),A)` and `(((A,D),D),A)`. We assessed these site patterns using haploid calls from long sequence alignment over syntenic regions to reference genomes of outgroup canid species varying in divergence.

First, we performed a long assembly-to-reference mapping with *minimap2* [139] using the preset for up to 10% sequence divergence (`asm10`), and applied the *minimap2* `paftools call` function to identify aligned regions (R lines) and sequence variants (V lines) relative to outgroup assemblies from most-to-least diverged from or within the wolf-like clade: the gray fox reference assembly (GenBank GCA_032313775.1) [112] and the red fox reference assembly (GenBank GCA_964106925.2) [111], the maned wolf (GenBank GCA_028533335.1, *Chrysocyon brachyurus* aka "chrBra") reference assembly [107], and the African wild dog (*Lycaon pictus* aka "lycPic") reference assembly (Yggdrasil DB12) [114]. To those reference genomes, we aligned assembled genome sequences of outgroups red fox and gray fox, when not used as reference; the seed-reference Greenland gray wolf (*C. lupus orion* aka canLor) and a non-reference genome assembled for a North American wolf (GenBank GCA_003160815.1, *C. lupus* aka canLup) [113]; and the pseudo-haploid consensus genome sequences of dire wolf individual DireGB called from random read-sampling ANGSD + Consensify from alignment to either the reconstructed ancient dire wolf (*Aenocyon dirus* aka "aenDir") genome reference ("iterative") or its seed-reference sequence of Greenland gray wolf ("divergent").

We restricted analyses to regions over each outgroup genome aligned by gray wolf and dire wolf genomes, comprising a total of 171,599,977 gray fox aligned sequence bases, 247,756,445 maned wolf aligned bases, and 244,509,933 African wild dog aligned bases, where these taxa should have long-range synteny for sequence alignment and variant calling. We then computed global *D*-statistic as (# ADDA sites - # DADA sites)/(# ADDA sites + # DADA sites) over canid syntenic regions of respective reference assemblies. We computed standard errors via a block jackknife procedure [38] to divide the outgroup genomes into 1 megabase equally-spaced contiguous blocks, and adjusted global *D*-statistics by this standard error to obtain Z-scores (**Table S6**).

We mapped DireSP sequence reads using BWA mem as described above to either the iteratively reconstructed dire wolf or gray wolf reference assemblies. Using the *Rsamtools* package (v2.22.0), we sampled 500,000 reads from each alignment file and parsed their cigar strings to calculate the identity-by-state (IBS) for each read. IBS was defined as the (number of aligned bases (M) - number of mismatches (NM))/aligned sequence length (M). Reads were binned by their aligned sequence lengths and average IBS was calculated for each bin.

Mitochondrial and nuclear phylogenetic analysis

In order to place our newly assembled mitochondrial sequence in phylogenetic context, we downloaded publicly available canid mitochondrial genomes (**Table S9**). We used *rotate* [140] to ensure all circularized sequences started at the same position and aligned the sequences with *MUSCLE* v3.8.1 [141]. We used *RAxML* v8.2.12 [142] to build a maximum likelihood phylogeny (-m

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

GTRGamma -f a -# 100 -p 12345 -x 12345) under the GTRGAMMA substitution model with 100 bootstrap replicates. The dire wolf mitochondrial genomes formed a clade outside of other wolf-like canids (**Fig. S5**) in a topology similar to that estimated previously [12,20,143].

To localize the placement of the dire wolf in the broader canid phylogeny, we set out to build a nuclear phylogeny using neutral windows of genomic sequence. We defined neutral windows using *bedtools* makewindows (-w 25000 -s 1025000 -i winnum), which defined uniquely named 25 Kb loci at least 1 Mb apart. Windows containing any annotated protein coding genes were removed using bedtools intersect (-v). This resulted in 1,332 possible loci to build trees from.

To assess canid sequence divergence over these windows, we aligned 14 representative canid data sets (**Table S8**), along with DireGB and DireSP, against the iteratively assembled paleogenome reference and obtained consensus sequencing using the *ANGSD* + *Consensify* approach outlined above. Sequences over the neutral windows were extracted and N content was measured using the "letterFrequency" function from the *BioStrings* package in R. Only windows that exhibited <= 50% missingness across all species (n=924) were retained for further analysis. Sequences that passed these thresholds were converted to multiple sequence alignments using *MAFFT* and trees were built from each loci using *IQTree2* with 100 bootstraps each with a GTR + I + G substitution model [144,145]. Each loci tree was merged and provided as input for *ASTRAL* II [146,147] to call a final consensus species tree.

Divergence time calibration

We used *MCMCtree* v4.10.7 from the *PAML* package [148] to perform divergence time estimation. We used an autocorrelated-rates relaxed clock model with approximate likelihood calculation and the HKY85 [149] substitution model with gamma-distributed (five categories) rate heterogeneity among sites. We used gamma priors for kappa (shape=2, scale=10) and alpha (1, 10) in the substitution model, and gamma-Dirichlet priors for the overall substitution rate (shape=2, scale=40, concentration=1) and rate drift parameter (2, 20, 1). These priors were selected to provide diffuse yet biologically realistic constraints on the nucleotide transition/transversion ratio, among-site rate variation, overall rate of molecular evolution, and rate drift (allowing for changes in evolutionary rate from parent to offspring branches) in the autocorrelated relaxed clock model, respectively.

We used two nodes as calibration points with priors based on previous studies: the root age calibrated for gray foxes (*Urocyon* spp.) (same root calibration as Perri et al. 2021) versus other canids [12] and the South American canid node represented by bush dog and maned wolf [107]. Lower and upper bounds for the root age were 10.3 Mya (*Metalopex macconnelli*) and 20 Mya (*Leptocyon* spp.), and 1.35-4.34 Mya for the South American canid node. For both nodes, we used log-normal distributions as priors, with 0.025 for location and scale parameters. This approximated a uniform distribution between upper and lower bounds, while softening the bounds with low-probability tails (see prior distributions of node ages in **Fig. S9**). The MCMC analysis was run for 1,000,000 iterations, with the first 10,000 discarded as burn-in, and samples collected every 100 iterations, resulting in 10,000 samples. We created visualizations of the resulting tree with divergence time estimates using the *MCMCtreeR* package [150] in R, showing both node age uncertainty and posterior distributions of node ages (**Fig. 3**, **Table S10**).

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sequence variant calling

We detected germline variants against the reconstructed dire wolf reference (aenDir) from BAMs of 35 modern canids (**Table S8**) and DireGB and DireSP using a joint-calling workflow implemented by *GATK4* [136]: *HaplotypeCaller* for local re-assembly of haplotypes, *GenomicsDBImport* for gVCF consolidation, and *GenotypeGVCFs* for joint variant calling. We also generated a variant call set comprising transversions only by filtering with *BCFtools* option, `-i '((REF="A" & ALT~"[TC]") | (REF="T" & ALT~"[AG]") | (REF="C" & ALT~"[AG]") | (REF="G" & ALT~"[TC]")) & (strlen(REF)=1 & strlen(ALT)=1)'`, which totaled 41,774,524 SNPs.

We also called variants from the multi-canid pangenome alignments surjected to the Greenland gray wolf reference (canLor) using the GPU-optimized NVIDIA® Clara™ Parabricks implementation of *DeepVariant* [151] for 23 modern canids (**Table S8**) and both dire wolves. *DeepVariant* gVCFs were combined using *GLNexus* [152] with the 'DeepVariant_unfiltered' configuration to avoid biasing allele frequencies of non-gray wolf species. We then set as missing any genotype with GQ<10 using the *bcftools* setGT plugin and filtered any site where fewer than 10 individuals were genotyped.

Sequence distance, diversity, and divergence

We estimated the sequence distance of dire wolves from extant canids using the formula: 2 x (mutations/generation) x (divergence time / generation time), assuming a mutation rate of 4 billion substitutions per generation calibrated for *Canis* species [41], a generation time between 5 and 8 years for black-backed jackals and wolf-like canids, respectively [42], and a divergence time of 5.8 Mya [12].

We assessed sequence similarity in two ways. First, we filtered the final species variant call set (described above) for homozygous transversions by species using *BCFTools* query and compared each using a custom script. Each species in a comparison had to have the same site vary in the same way to be considered a shared allele. We then subtracted the total differences from the non-N sequence length of the reconstructed dire wolf (2,310,286,013 nucleotides) and expressed a proportion of total available sequence. In parallel, we estimated sequence distance using *Mash* [43]. For each consensus sequence, we used *mash sketch* to hash the genome into 32-mers and assessed common 32-mers in each species in a pairwise manner.

Gene flow and admixture graph modeling

We converted autosomal variants to *EIGENSTRAT* format using *vcf2eigenstrat* from *SequenceTools* (https://github.com/stschiff/sequenceTools) for both the *GATK4* jointly-called variants against the dire wolf consensus genome and *DeepVariant* variants called against the gray wolf genome from the surjected pangenome alignments. To account for differential coverage across individuals, we then made pseudohaploid genotype calls for each individual for the full set of variants using a read sampling procedure modeled after *Consensify* [135] that controls for biases arising from differential sequencing coverage and mapping errors (**Fig. S10, S19**) by randomly selecting 3 reads at each site, or two reads at positions covered by exactly two reads, and requiring at least two bases among the selected reads match. Sites with fewer than two reads were set to missing. This procedure resulted in 33.91M aenDir-mapped and 35.06M pangenome-mapped autosomal, biallelic, mappability-filtered

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

transversions for 17 populations represented by 37 individuals. Ultimately, alignment to the reconstructed dire wolf reference increased reference bias (attraction toward DireGB) in a manner that pangenome alignment did not (**Fig. S12, S19**), particularly for more divergent outgroups. As admixture tests can be particularly sensitive to such biases [25], we focused pangenome-derived results for describing canid population history and gene flow, though note that patterns and ratios of admixture in the dire wolf lineage obtained from alignments to the reconstructed dire wolf genome were similar to those inferred from pangenome aligned reads (**Fig. S13**).

We used *D*-statistics to assess allele sharing between dire wolves and extant Canini lineages, using either red foxes or gray foxes as an outgroup *D*-statistics were calculated using *admixtools2* v2.0.6 [44]. We then used the findgraphs function from *admixtools2* to search for possible well-fitting admixture graph models across varying subsets of canid species. For each set of species, we performed 100 separate iterations of findgraphs across 0-8 admixture events, and extracted the graph with the lowest log-likelihood score as the best-fitting model across each iteration. In cases where less complex graphs provided adequate fits for a given set of species (worst residual of Z< 3), we did not model additional admixture events.

We began with a minimal set of species representing major canid lineages: dire wolves (n=2), Eurasian gray wolves (n=2), coyotes (n=4), dholes (n=2), African wild dogs (n=3), maned wolves (n=2), and red foxes (n=2) as an outgroup. We also explored similar models but with either gray foxes or both red and gray foxes as outgroups. We also explored graphs in which dire wolves were each modeled individually, rather than grouped into a population. Beyond this minimal species set, we also modeled a complex graph which featured African jackals (side-striped jackal and black-backed jackal), along with an additional South American canid lineage (Pampas fox), and a North American gray wolf population (Yellowstone). Finally, we assessed admixture graphs which featured only extant canid lineages and included populations with known history of admixture (golden jackals and red wolves), to assess the sensitivity of this method to detect known instances of admixture among canid lineages.

Diversifying positive selection analysis

We extracted all protein-coding nucleotide and translated amino acid sequences using *gffread* [153] on the pseudo-haploid consensus genome sequences of modern canid species and ancient dire wolf aligned to reconstructed dire wolf paleogenome and called from the random read-sampling *ANGSD* + *Consensify* method, which maintained identical genomic coordinates for gene annotation. We selected the longest coding transcript per distinct gene ID with <20% missing amino acids when translated using *seqkit* [154], yielding 13,830 protein-coding genes for positive selection tests. We ran HyPhy analysis programs BUSTED and FitMG94 in local and global modes on aligned CDS and the consensus species tree with dire wolf (DireGB) designated as foreground branch:

```
(uroCin_SAMN04495241,(vulVul_marble_Fox4,((chrBra_SAMN27257233,speVen_SAMN27257238)1:3.795527799956191,(DireGB,(lupMes_SAMEA5411047,(lycPic_SAMN09924608,(cuoAlp_SAMN10180424,(canRuf_SAMN02921317,(canLup_SAMN02921310,canBai_SAMN02921314)1:0.11623018789473029)1:4.18818635818363)1:0.39859685674669815)1:0.323225231778891)1:0.182689191931038 1)1:1.4874594855476162)1:6.424328629404271):0.0);
```

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We tested protein-coding genes using the BUSTED model for episodic positive selection to identify where at least one site experienced positive selection in the dire wolf branch. Additionally, we extracted from the FitMG94 output the branch-specific dN/dS rate estimates under a free-ratio model (`--type local`) and respective non-synonymous and synonymous branch lengths for dire wolf and 10 other canids and the likelihood-ratio test for overall dN/dS != 1 across the entire species tree using a one-ratio model (`--type global`). We considered those genes with tree-wide, one-ratio dN/dS estimate > 2 as indicating pervasive positive selection over the canid lineage.

We flagged genes as under episodic positive aka diversifying selection in the dire wolf branch that exhibited $p < 0.05$ in the BUSTED test and which flagged at least a single codon of evidence ratio > 2, and excluding genes for which the dire wolf consensus codon did not disagree with dire wolf reconstructed reference codon and neither the focal dire wolf codon nor root canid codon contained missing bases (Ns). The process of consensus calling at sites of low depth (<5×) or heterozygous alleles introduces missing bases that could result in misassignment of codons for the focal branch (e.g., dire wolf) and/or root branch (ancestral canid), which would in turn lead to false positives as HyPhy BUSTED needs only a single site to reject fitting the constrained model (no positive selection). Furthermore, BUSTED is insensitive to synonymous substitution rate variation (a constant dS=1 is held in all cases).

We also flagged additional genes as having strong evidence for positive aka diversifying selection in the dire wolf branch, if these had free-ratio maximum likelihood estimates of dN/dS > 2 from the standard MG94 codon evolution model [155] implemented in the HyPhy FitMG94 program. We filtered out such genes depleted of synonymous substitutions where dS > 0.001, to avoid extreme estimates, and no excess saturation of substitutions where dS or dN was greater than 4.0, and which exhibited overall dN/dS estimated under the one-ratio model meeting statistical significance (α=0.05) in the log-ratio test given degrees of freedom=(estimated parameters in the standard MG94 model - estimated parameters in the nucleotide GTR model) > 0 under the chi-squared distribution. We excluded those that exhibited extremely saturated rates of substitutions (dS or dN > 4.0) or depleted rates of synonymous substitutions (dS < 0.001), because dN/dS ratios could be inappropriately estimated or result from misalignment of reads [156].

We report all protein-coding genes tested and respective statistics from HyPhy BUSTED and FitMG94 programs in **Table S11**. As BUSTED evidence ratios only indicate whether a site better fits under the alternative, unconstrained model that permits positive selection, rather than providing direct evidence site-specific positive selection, we also ran HyPhy FEL (fixed effects indicating pervasive selection) and MEME (mixed effects aka partitioned selection) maximum likelihood models over all 80 genes under diversifying positive selection in dire wolves.

We performed gene set enrichment analysis using *g:Profiler* (`gprofiler2`) [157] and the domestic dog ortholog reference set (organism: “clfamiliaris”), testing the combined suite of 80 genes under episodic diversifying selection against the background set of 5,497 genes tested that pass all aforementioned filters. Corrections for multiple testing were performed using the Benjamini-Hochberg procedure.

We contextualized dire wolf specific substitutions using *InterProScan* to predict protein functional domains [53] and ESM1b model to predict substitution impact via *in silico* mutagenesis [54] over each dire wolf protein sequence. We replaced all invalid amino acid residues (`X` introduced from N content in coding sequences) with a neutral amino acid,

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

glycine (`G`), to enable modeling. For protein sequences longer than 1,022 amino acids that hit the sequence limit of the ESM1b model, we performed tiled predictions as in Brandes et al. 2023 [54] to obtain weighted average effect scores. We examined residues harboring dire wolf specific amino acids and compared to the canid-majority amino acid at each residue. As the "wild type" sequence modeled is sourced from the reconstructed dire wolf reference (aenDir), we considered the positive or negative pseudo log-likelihood scores between the consensus amino acid assigned to the dire wolf branch in BUSTED and the canid majority amino acid, depending which amino acid matched the root in BUSTED analysis. We report domain and effect predictions for flagged residues modeled in **Table S12**.

DATA AVAILABILITY

The sequencing data underlying this study are available in the NCBI Sequence Read Archive, and can be accessed with BioProject accession PRJNA1222369. For the HyPhy analyses, we included a tarball of the multiple sequence alignment FASTAs of the coding sequences for all 80 genes under diversifying selection in the dire wolf branch as well as the HyPhy results JSON files from BUSTED, FitMG94, FEL, and MEME programs. These data and all supplemental tables are available in a Zenodo repository associated with this study under the BioRxiv DOI.

ACKNOWLEDGEMENTS

We thank the Center for Genome Innovation and the Computational Biology Core in the Institute for Systems Genomics at the University of Connecticut for sequencing and data management support. We thank the Bureau of Reclamation for sampling permissions of the Idaho individual that produced DireGB (IMNH 48001/52) and staff at the Idaho Museum of Natural History for supporting resampling of specimens under their care. We thank the staff including Glenn Storrs and Cameron Schwalbach of the Cincinnati Museum Center for additional sampling of the DireSP specimen (VP1737). We also thank colleagues who lent their time and expertise to provide valuable comments on this work.

AUTHOR CONTRIBUTIONS

Conceptualization: G.G., K.M.P, J.O., M.CJ., B.L.C, B.S.
Data curation: G.G., K.M.P, J.O., B.L.C.
Formal analysis: G.G., K.M.P, J.O., C.H., M.CJ., W.T.
Funding acquisition: B.L.
Investigation: S.B., C.C., J.M., B.R.P, P.G.S.G., S.J.H., S.S., W.S., R.J.O, W.T.
Methodology: G.G., K.M.P, J.O., M.CJ., B.L.C, B.S.
Project administration: G.G., K.M.P
Resources: N.A, S.B., B.R.P., J.M., S.J.H., R.J.O, S.S., W.S., C.C., L.K., T.M., B.S.
Software: N.A., G.G., K.M.P, J.O., M.CJ.,
Supervision: B.L.C, B.S.

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Validation: G.G., K.M.P., J.O.
Visualization: G.G., K.M.P, J.O., M.CJ, C.H.
Writing – original draft: K.M.P., G.G.
Writing – review & editing: G.G., K.M.P., J.O. M.CJ., C.C., N.A, W.T, O.F, S.B., T.M., L.K., J.K., A.D., M.C., J.B.P, S.O., R.J.O, S.J.H., P.G.S.G., S.S., W.S., J.M., B.R.P, G.L., L.F, A.P., A.T.L, M.S.S., M.T.P.G, S.G., B.vH., E.K.K., C.E.M., L.D, B.L.C., G.R.R.M., G.C, B.S.

CONFLICTS OF INTEREST

Authors affiliated with Colossal Biosciences and/or Form Bio may hold stock and/or stock options in these companies.

E.K.K., C.E.M, B.vH., R.J.O., M.T.P.G, and L.D. are members of the scientific advisory board at Colossal Biosciences.

M.S.S. is a consultant of Colossal Biosciences.

G.R.R.M is an Investor and Cultural Advisor for Colossal Biosciences.

G.C. is co-founder of Colossal Biosciences and others available at arep.med.harvard.edu/t

SUPPLEMENTAL INFORMATION

Document S1: Figures S1-19
Table S1: Sequencing library statistics for all ancient samples.
Table S2: Reference assembly accessions used in study
Table S3: Alignment statistics for all ancient samples
Table S4: Iterative mapping and polishing statistics for paleogenome reconstruction
Table S5: Gene annotation information for paleogenome reconstruction
Table S6: Allele sharing statistics for residual reference bias assessment
Table S7: Genetic distance metrics
Table S8: Short read dataset accessions used in study
Table S9: Mitochondrial genome accessions used in study
Table S10: Age estimates for nodes in phylogenetic tree
Table S11: Summary of admixture methods and models
Table S12: Selection test results by gene
Table S13: Genes with dire wolf specific substitutions

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

1. Mead, J.I., Manganaro, C., Repenning, C.A., and Agenbroad, L.D. (1996). Early Rancholabrean mammals from Salamander Cave, Black Hills, South Dakota. In Palaeoecology and Palaeoenvironments of Late Cenozoic Mammals (University of Toronto Press), pp. 458–482. https://doi.org/10.3138/9781487574154-023.

2. O'Keefe, F.R., Dunn, R.E., Weitzel, E.M., Waters, M.R., Martinez, L.N., Binder, W.J., Southon, J.R., Cohen, J.E., Meachen, J.A., DeSantis, L.R.G., et al. (2023). Pre–Younger Dryas megafaunal extirpation at Rancho La Brea linked to fire-driven state shift. Science *381*, eabo3594. https://doi.org/10.1126/science.abo3594.

3. Anyonge, W., and Baker, A. (2006). Craniofacial morphology and feeding behavior in *Canis dirus*, the extinct Pleistocene dire wolf. J. Zool. *269*, 309–316. https://doi.org/10.1111/j.1469-7998.2006.00043.x.

4. De Latorre, D.V., and Marshall, C.R. (2024). Evolutionary allometry of the canid baculum (Carnivora: Mammalia). Biol. J. Linn. Soc., blae048. https://doi.org/10.1093/biolinnean/blae048.

5. Patterson, B.D., and Thaeler, C.S. (1982). The Mammalian Baculum: Hypotheses on the Nature of Bacular Variability. J. Mammal. *63*, 1–15. https://doi.org/10.2307/1380665.

6. O'Keefe, F.R., Meachen, J., Fet, E.V., and Brannick, A. (2013). Ecological determinants of clinal morphological variation in the cranium of the North American gray wolf. J. Mammal. *94*, 1223–1236. https://doi.org/10.1644/13-MAMM-A-069.

7. Dundas, R.G. (1999). Quaternary records of the dire wolf, *Canis dirus*, in North and South America. Boreas *28*, 375–385. https://doi.org/10.1111/j.1502-3885.1999.tb00227.x.

8. Anyonge, W. and Roman, Chris (2006). New Body Mass Estimates for Canis Dirus, the Extinct Pleistocene Dire Wolf. J. Vertebr. Paleontol. *26*, 209–212. https://doi.org/10.1671/0272-4634(2006)26[209:NBMEFC]2.0.CO;2.

9. Harris, A.H. (1981). Kurten, B., and E. Anderson. PLEISTOCENE MAMMALS OF NORTH AMERICA. Columbia Univ. Press, New York, 442 pp., 1980. Price $42.50. J. Mammal. *62*, 653–654. https://doi.org/10.2307/1380422.

10. Tedford, R.H., Wang, X., and Taylor, B.E. (2009). Phylogenetic Systematics of the North American Fossil Caninae (Carnivora: Canidae). Bull. Am. Mus. Nat. Hist. *325*, 1–218. https://doi.org/10.1206/574.1.

11. Nowak, R.M. (2010). Wolf Evolution and Taxonomy. In Wolves: Behavior, Ecology, and Conservation, L. D. Mech and L. Boitani, eds. (University of Chicago Press).

12. Perri, A.R., Mitchell, K.J., Mouton, A., Álvarez-Carretero, S., Hulme-Beaman, A., Haile, J., Jamieson, A., Meachen, J., Lin, A.T., Schubert, B.W., et al. (2021). Dire wolves were the last

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of an ancient New World canid lineage. Nature *591*, 87–91. https://doi.org/10.1038/s41586-020-03082-x.

13. Sotnikova, M., and Rook, L. (2010). Dispersal of the Canini (Mammalia, Canidae: Caninae) across Eurasia during the Late Miocene to Early Pleistocene. Quat. Int. *212*, 86–97. https://doi.org/10.1016/j.quaint.2009.06.008.

14. Meachen, J.A., and Samuels, J.X. (2012). Evolution in coyotes (*Canis latrans*) in response to the megafaunal extinctions. Proc. Natl. Acad. Sci. *109*, 4191–4196. https://doi.org/10.1073/pnas.1113788109.

15. Meachen, J.A., Janowicz, A.C., Avery, J.E., and Sadleir, R.W. (2014). Ecological Changes in Coyotes (Canis latrans) in Response to the Ice Age Megafaunal Extinctions. PLoS ONE *9*, e116041. https://doi.org/10.1371/journal.pone.0116041.

16. Nowak, R.M. (1979). North American Quaternary Canis (Museum of Natural History, University of Kansas) https://doi.org/10.5962/bhl.title.4072.

17. Heintzman, P.D., Froese, D., Ives, J.W., Soares, A.E.R., Zazula, G.D., Letts, B., Andrews, T.D., Driver, J.C., Hall, E., Hare, P.G., et al. (2016). Bison phylogeography constrains dispersal and viability of the Ice Free Corridor in western Canada. Proc. Natl. Acad. Sci. *113*, 8057–8063. https://doi.org/10.1073/pnas.1601077113.

18. Koblmüller, S., Vilà, C., Lorente-Galdos, B., Dabad, M., Ramirez, O., Marques-Bonet, T., Wayne, R.K., and Leonard, J.A. (2016). Whole mitochondrial genomes illuminate ancient intercontinental dispersals of grey wolves (*Canis lupus*). J. Biogeogr. *43*, 1728–1738. https://doi.org/10.1111/jbi.12765.

19. Loog, L., Thalmann, O., Sinding, M.S., Schuenemann, V.J., Perri, A., Germonpré, M., Bocherens, H., Witt, K.E., Samaniego Castruita, J.A., Velasco, M.S., et al. (2020). Ancient DNA suggests modern wolves trace their origin to a Late Pleistocene expansion from Beringia. Mol. Ecol. *29*, 1596–1610. https://doi.org/10.1111/mec.15329.

20. Koepfli, K.-P., Pollinger, J., Godinho, R., Robinson, J., Lea, A., Hendricks, S., Schweizer, R.M., Thalmann, O., Silva, P., Fan, Z., et al. (2015). Genome-wide Evidence Reveals that African and Eurasian Golden Jackals Are Distinct Species. Curr. Biol. *25*, 2158–2165. https://doi.org/10.1016/j.cub.2015.06.060.

21. vonHoldt, B.M., Cahill, J.A., Fan, Z., Gronau, I., Robinson, J., Pollinger, J.P., Shapiro, B., Wall, J., and Wayne, R.K. (2016). Whole-genome sequence analysis shows that two endemic species of North American wolf are admixtures of the coyote and gray wolf. Sci. Adv. *2*, e1501714. https://doi.org/10.1126/sciadv.1501714.

22. Gopalakrishnan, S., Sinding, M.-H.S., Ramos-Madrigal, J., Niemann, J., Samaniego Castruita, J.A., Vieira, F.G., Carøe, C., Montero, M.D.M., Kuderna, L., Serres, A., et al. (2018). Interspecific gene flow shaped the evolution of the genus Canis. Curr. Biol. *28*, 3441-3449.e5. https://doi.org/10.1016/j.cub.2018.08.041.

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

23. Westbury, M.V., Hartmann, S., Barlow, A., Preick, M., Ridush, B., Nagel, D., Rathgeber, T., Ziegler, R., Baryshnikov, G., Sheng, G., et al. (2020). Hyena paleogenomes reveal a complex evolutionary history of cross-continental gene flow between spotted and cave hyena. Sci. Adv. *6*, eaay0456. https://doi.org/10.1126/sciadv.aay0456.

24. Westbury, M.V., Barnett, R., Sandoval-Velasco, M., Gower, G., Vieira, F.G., de Manuel, M., Hansen, A.J., Yamaguchi, N., Werdelin, L., Marques-Bonet, T., et al. (2021). A genomic exploration of the early evolution of extant cats and their sabre-toothed relatives. Open Res. Eur. *1*, 25. https://doi.org/10.12688/openreseurope.13104.2.

25. Günther, T., and Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. PLoS Genet. *15*, e1008302. https://doi.org/10.1371/journal.pgen.1008302.

26. Prüfer, K., Stenzel, U., Hofreiter, M., Pääbo, S., Kelso, J., and Green, R.E. (2010). Computational challenges in the analysis of ancient DNA. Genome Biol. *11*, R47. https://doi.org/10.1186/gb-2010-11-5-r47.

27. Liu, Y., Wang, M., Chen, P., Wang, Z., Liu, J., Yao, L., Wang, F., Tang, R., Zou, X., and He, G. (2021). Combined Low-/High-Density Modern and Ancient Genome-Wide Data Document Genomic Admixture History of High-Altitude East Asians. Front. Genet. *12*, 582357. https://doi.org/10.3389/fgene.2021.582357.

28. Shapiro, B., and Hofreiter, M. (2014). A paleogenomic perspective on evolution and gene function: new insights from ancient dna. Science *343*, 1236573. https://doi.org/10.1126/science.1236573.

29. Akopyan, M., Genchev, M., Armstrong, E.E., and Mooney, J.A. (2024). Divergent reference genomes compromise the reconstruction of demographic histories, selection scans, and population genetic summary statistics. Preprint at bioRxiv, https://doi.org/10.1101/2024.11.26.625554 https://doi.org/10.1101/2024.11.26.625554.

30. O'Keefe, F., Binder, W., Frost, S., Sadleir, R., and Van Valkenburgh, B. (2014). Cranial morphometrics of the dire wolf, *Canis dirus*, at Rancho La Brea: temporal variability and its links to nutrient stress and climate. Palaeontol. Electron. https://doi.org/10.26879/437.

31. Schmökel, H., Farrell, A., and Balisi, M.F. (2023). Subchondral defects resembling osteochondrosis dissecans in joint surfaces of the extinct saber-toothed cat Smilodon fatalis and dire wolf Aenocyon dirus. PloS One *18*, e0287656. https://doi.org/10.1371/journal.pone.0287656.

32. O'Keefe, F., Binder, W., Frost, S., Sadleir, R., and Van Valkenburgh, B. (2014). Cranial morphometrics of the dire wolf, *Canis dirus*, at Rancho La Brea: temporal variability and its links to nutrient stress and climate. Palaeontol. Electron. https://doi.org/10.26879/437.

33. Tankersley, K.B. (1997). Sheriden: A Clovis cave site in eastern North America. Geoarchaeology *12*, 713–724.

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

https://doi.org/10.1002/(SICI)1520-6548(199709)12:6<713::AID-GEA9>3.0.CO;2-1.

34. Scott, W.E., Pierce, K.L., Bradbury, J.P., and Forester, R.M. (1982). Revised Quaternary stratigraphy and chronology in the American Falls area, southeastern Idaho. Cenozoic Geol. Ida. Ida. Bur. Mines Geol. Bull. *26*, 581–595.

35. Sinding, M.-H.S., Gopalakrishnan, S., Raundrup, K., Dalén, L., Threlfall, J., Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium, et al. (2021). The genome sequence of the grey wolf, *Canis lupus* Linnaeus 1758. Wellcome Open Res. *6*, 310. https://doi.org/10.12688/wellcomeopenres.17332.1.

36. Mittnik, A., Wang, C.-C., Svoboda, J., and Krause, J. (2016). A Molecular Approach to the Sexing of the Triple Burial at the Upper Paleolithic Site of Dolní Věstonice. PLOS ONE *11*, e0163019. https://doi.org/10.1371/journal.pone.0163019.

37. De Flamingh, A., Coutu, A., Roca, A.L., and Malhi, R.S. (2020). Accurate sex identification of ancient elephant and other animal remains using low-coverage DNA shotgun sequencing data. G3 GenesGenomesGenetics *10*, 1427–1432. https://doi.org/10.1534/g3.119.400833.

38. Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., De Filippo, C., et al. (2012). A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science *338*, 222–226. https://doi.org/10.1126/science.1224344.

39. Hahn, C., Bachmann, L., and Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. Nucleic Acids Res. *41*, e129. https://doi.org/10.1093/nar/gkt371.

40. Dolenz, S., Van Der Valk, T., Jin, C., Oppenheimer, J., Sharif, M.B., Orlando, L., Shapiro, B., Dalén, L., and Heintzman, P.D. (2024). Unravelling reference bias in ancient DNA datasets. Bioinformatics *40*, btae436. https://doi.org/10.1093/bioinformatics/btae436.

41. Skoglund, P., Ersmark, E., Palkopoulou, E., and Dalén, L. (2015). Ancient wolf genome reveals an early divergence of domestic dog ancestors and Aamixture into high-latitude Breeds. Curr. Biol. *25*, 1515–1519. https://doi.org/10.1016/j.cub.2015.04.019.

42. Pacifici, M., Santini, L., Di Marco, M., Baisero, D., Francucci, L., Grottolo Marasini, G., Visconti, P., and Rondinini, C. (2013). Generation length for mammals. Nat. Conserv. *5*, 89–94. https://doi.org/10.3897/natureconservation.5.5734.

43. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., and Phillippy, A.M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. *17*, 132. https://doi.org/10.1186/s13059-016-0997-x.

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

44. Maier, R., Flegontov, P., Flegontova, O., Işıldak, U., Changmai, P., and Reich, D. (2023). On the limits of fitting complex models of population history to f-statistics. eLife *12*, e85492. https://doi.org/10.7554/eLife.85492.

45. Gopalakrishnan, S., Sinding, M.-H.S., Ramos-Madrigal, J., Niemann, J., Samaniego Castruita, J.A., Vieira, F.G., Carøe, C., Montero, M. de M., Kuderna, L., Serres, A., et al. (2018). Interspecific Gene Flow Shaped the Evolution of the Genus *Canis*. Curr. Biol. *28*, 3441-3449.e5. https://doi.org/10.1016/j.cub.2018.08.041.

46. Chavez, D.E., Gronau, I., Hains, T., Kliver, S., Koepfli, K.-P., and Wayne, R.K. (2019). Comparative genomics provides new insights into the remarkable adaptations of the African wild dog (Lycaon pictus). Sci. Rep. *9*, 8329. https://doi.org/10.1038/s41598-019-44772-5.

47. Heppenheimer, E., Brzeski, K.E., Hinton, J.W., Chamberlain, M.J., Robinson, J., Wayne, R.K., and vonHoldt, B.M. (2020). A genome-wide perspective on the persistence of red wolf ancestry in southeastern canids. J. Hered. *111*, 277–286. https://doi.org/10.1093/jhered/esaa006.

48. Vilaça, S.T., Donaldson, M.E., Benazzo, A., Wheeldon, T.J., Vizzari, M.T., Bertorelle, G., Patterson, B.R., and Kyle, C.J. (2023). Tracing Eastern Wolf Origins From Whole-Genome Data in Context of Extensive Hybridization. Mol. Biol. Evol. *40*, msad055. https://doi.org/10.1093/molbev/msad055.

49. Koepfli, K.-P., Pollinger, J., Godinho, R., Robinson, J., Lea, A., Hendricks, S., Schweizer, R.M., Thalmann, O., Silva, P., Fan, Z., et al. (2015). Genome-wide Evidence Reveals that African and Eurasian Golden Jackals Are Distinct Species. Curr. Biol. *25*, 2158–2165. https://doi.org/10.1016/j.cub.2015.06.060.

50. Murrell, B., Weaver, S., Smith, M.D., Wertheim, J.O., Murrell, S., Aylward, A., Eren, K., Pollner, T., Martin, D.P., Smith, D.M., et al. (2015). Gene-wide identification of episodic selection. Mol. Biol. Evol. *32*, 1365–1371. https://doi.org/10.1093/molbev/msv035.

51. Kosakovsky Pond, S.L., and Frost, S.D.W. (2005). Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. Mol. Biol. Evol. *22*, 1208–1222. https://doi.org/10.1093/molbev/msi105.

52. Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K., and Kosakovsky Pond, S.L. (2012). Detecting individual sites subject to episodic diversifying selection. PLoS Genet. *8*, e1002764. https://doi.org/10.1371/journal.pgen.1002764.

53. Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., et al. (2021). The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. *49*, D344–D354. https://doi.org/10.1093/nar/gkaa977.

54. Brandes, N., Goldman, G., Wang, C.H., Ye, C.J., and Ntranos, V. (2023). Genome-wide

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

prediction of disease variant effects with a deep protein language model. Nat. Genet. *55*, 1512–1522. https://doi.org/10.1038/s41588-023-01465-0.

55. Perini, F.A., Russo, C. a. M., and Schrago, C.G. (2010). The evolution of South American endemic canids: a history of rapid diversification and morphological parallelism. J. Evol. Biol. *23*, 311–322. https://doi.org/10.1111/j.1420-9101.2009.01901.x.

56. Porto, L.M.V., Etienne, R.S., and Maestri, R. (2023). Evolutionary radiation in canids following continental colonizations. Evolution *77*, 971–979. https://doi.org/10.1093/evolut/qpad015.

57. Frosali, S., Bartolini-Lucenti, S., Madurell-Malapeira, J., Urciuoli, A., Costeur, L., and Rook, L. (2023). First digital study of the frontal sinus of stem-Canini (Canidae, Carnivora): evolutionary and ecological insights throughout advanced diagnostic in paleobiology. Front. Ecol. Evol. *11*, 1173341. https://doi.org/10.3389/fevo.2023.1173341.

58. Rook, L. (2009). The wide ranging genus *Eucyon* Tedford & Qiu, 1996 (Mammalia, Carnivora, Canidae, Canini) in the Mio-Pliocene of the Old World. Geodiversitas *31*, 723–741. https://doi.org/10.5252/g2009n4a723.

59. Werdelin, L., Lewis, M.E., and Haile-Selassie, Y. (2015). A critical review of A frican species of *E ucyon* (M ammalia; C arnivora; C anidae), with a new species from the P liocene of the W oranso- M ille A rea, A far R egion, E thiopia. Pap. Palaeontol. *1*, 33–40. https://doi.org/10.1002/spp2.1001.

60. Bartolini Lucenti, S., and Rook, L. (2021). “Canis” ferox Revisited: Diet Ecomorphology of Some Long Gone (Late Miocene and Pliocene) Fossil Dogs. J. Mamm. Evol. *28*, 285–306. https://doi.org/10.1007/s10914-020-09500-1.

61. Valenciano, A., Morales, J., and Govender, R. (2022). *Eucyon khoikhoi* sp. nov. (Carnivora: Canidae) from Langebaanweg ‘E’ Quarry (early Pliocene, South Africa): the most complete African canini from the Mio-Pliocene. Zool. J. Linn. Soc. *194*, 366–394. https://doi.org/10.1093/zoolinnean/zlab022.

62. Prassack, K.A., and Walkup, L.C. (2022). Maybe So, Maybe Not: Canis lepophagus at Hagerman Fossil Beds National Monument, Idaho, USA. J. Mamm. Evol. *29*, 313–333. https://doi.org/10.1007/s10914-021-09591-4.

63. van der Valk, T., Pečnerová, P., Díez-Del-Molino, D., Bergström, A., Oppenheimer, J., Hartmann, S., Xenikoudakis, G., Thomas, J.A., Dehasque, M., Sağlıcan, E., et al. (2021). Million-year-old DNA sheds light on the genomic history of mammoths. Nature *591*, 265–269. https://doi.org/10.1038/s41586-021-03224-9.

64. Mallet, J. (2007). Hybrid speciation. Nature *446*, 279–283. https://doi.org/10.1038/nature05706.

65. Gopalakrishnan, S., Sinding, M.-H.S., Ramos-Madrigal, J., Niemann, J., Samaniego

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Castruita, J.A., Vieira, F.G., Carøe, C., Montero, M.D.M., Kuderna, L., Serres, A., et al. (2018). Interspecific gene flow shaped the evolution of the genus Canis. Curr. Biol. *28*, 3441-3449.e5. https://doi.org/10.1016/j.cub.2018.08.041.

66. Parmley, J.L., Chamary, J.V., and Hurst, L.D. (2006). Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol. Biol. Evol. *23*, 301–309. https://doi.org/10.1093/molbev/msj035.

67. Lawrie, D.S., Messer, P.W., Hershberg, R., and Petrov, D.A. (2013). Strong purifying selection at synonymous sites in D. melanogaster. PLoS Genet. *9*, e1003527. https://doi.org/10.1371/journal.pgen.1003527.

68. Gudkov, M., Thibaut, L., and Giannoulatou, E. (2024). Quantifying negative selection on synonymous variants. Hum. Genet. Genomics Adv. *5*, 100262. https://doi.org/10.1016/j.xhgg.2024.100262.

69. Christmas, M.J., Dong, M., Meadows, J.R.S., Kozyrev, S.V., and Lindblad-Toh, K. (2024). Interpreting mammalian evolutionary constraint at synonymous sites in light of the unwanted transcript hypothesis. Preprint, https://doi.org/10.1101/2024.04.23.590689 https://doi.org/10.1101/2024.04.23.590689.

70. Conway, E., Jerman, E., Healy, E., Ito, S., Holoch, D., Oliviero, G., Deevy, O., Glancy, E., Fitzpatrick, D.J., Mucha, M., et al. (2018). A Family of Vertebrate-Specific Polycombs Encoded by the LCOR/LCORL Genes Balance PRC2 Subtype Activities. Mol. Cell *70*, 408-421.e8. https://doi.org/10.1016/j.molcel.2018.03.005.

71. Plassais, J., Kim, J., Davis, B.W., Karyadi, D.M., Hogan, A.N., Harris, A.C., Decker, B., Parker, H.G., and Ostrander, E.A. (2019). Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. Nat. Commun. *10*, 1489. https://doi.org/10.1038/s41467-019-09373-w.

72. Graber, J.K., Signer-Hasler, H., Burren, A., and Drögemüller, C. (2022). Evaluation of truncating variants in the *LCORL* gene in relation to body size of goats from Switzerland. Anim. Genet. *53*, 237–239. https://doi.org/10.1111/age.13177.

73. Yuan, Z., Ge, L., Su, P., Gu, Y., Chen, W., Cao, X., Wang, S., Lv, X., Getachew, T., Mwacharo, J.M., et al. (2023). NCAPG Regulates Myogenesis in Sheep, and SNPs Located in Its Putative Promoter Region Are Associated with Growth and Development Traits. Anim. Open Access J. MDPI *13*, 3173. https://doi.org/10.3390/ani13203173.

74. Metzger, J., Schrimpf, R., Philipp, U., and Distl, O. (2013). Expression levels of LCORL are associated with body size in horses. PloS One *8*, e56497. https://doi.org/10.1371/journal.pone.0056497.

75. Majeres, L.E., Dilger, A.C., Shike, D.W., McCann, J.C., and Beever, J.E. (2024). Defining a Haplotype Encompassing the LCORL-NCAPG Locus Associated with Increased Lean Growth in Beef Cattle. Genes *15*, 576. https://doi.org/10.3390/genes15050576.

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

76. Bai, F., Cai, Y., Qiu, M., Liang, C., Pan, L., Liu, Y., Feng, Y., Cao, X., Yang, Q., Ren, G., et al. (2025). *LCORL* and *STC2* variants increase body size and growth rate in cattle and other animals. Preprint, https://doi.org/10.1101/2025.01.01.630985 https://doi.org/10.1101/2025.01.01.630985.

77. Zhang, W., Li, J., Guo, Y., Zhang, L., Xu, L., Gao, X., Zhu, B., Gao, H., Ni, H., and Chen, Y. (2016). Multi-strategy genome-wide association studies identify the DCAF16-NCAPG region as a susceptibility locus for average daily gain in cattle. Sci. Rep. *6*, 38073. https://doi.org/10.1038/srep38073.

78. Jun, H.J., Roy, J., Smith, T.B., Wood, L.B., Lane, K., Woolfenden, S., Punko, D., Bronson, R.T., Haigis, K.M., Breton, S., et al. (2014). ROS1 signaling regulates epithelial differentiation in the epididymis. Endocrinology *155*, 3661–3673. https://doi.org/202007161214095980O.

79. Basciani, S., Mariani, S., Arizzi, M., Brama, M., Ricci, A., Betsholtz, C., Bondjers, C., Ricci, G., Catizone, A., Galdieri, M., et al. (2004). Expression of platelet-derived growth factor (PDGF) in the epididymis and analysis of the epididymal development in PDGF-A, PDGF-B, and PDGF receptor beta deficient mice. Biol. Reprod. *70*, 168–177. https://doi.org/10.1095/biolreprod.103.019232.

80. Schultz, N.G., Ingels, J., Hillhouse, A., Wardwell, K., Chang, P.L., Cheverud, J.M., Lutz, C., Lu, L., Williams, R.W., and Dean, M.D. (2016). The Genetic Basis of Baculum Size and Shape Variation in Mice. G3 GenesGenomesGenetics *6*, 1141–1151. https://doi.org/10.1534/g3.116.027888.

81. Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., and Nieselt, K. (2016). EAGER: efficient ancient genome reconstruction. Genome Biol. *17*, 60. https://doi.org/10.1186/s13059-016-0918-z.

82. Green, R.E., Malaspinas, A.-S., Krause, J., Briggs, A.W., Johnson, P.L.F., Uhler, C., Meyer, M., Good, J.M., Maricic, T., Stenzel, U., et al. (2008). A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing. Cell *134*, 416–426. https://doi.org/10.1016/j.cell.2008.06.021.

83. Westbury, M.V., and Lorenzen, E.D. (2022). Iteratively mapping ancient DNA to reconstruct highly divergent mitochondrial genomes: An evaluation of software, parameters and bait reference. Methods Ecol. Evol. *13*, 2419–2428. https://doi.org/10.1111/2041-210X.13990.

84. Durand, E.Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for Ancient Admixture between Closely Related Populations. Mol. Biol. Evol. *28*, 2239–2252. https://doi.org/10.1093/molbev/msr048.

85. Gopalakrishnan, S., Ebenesersdóttir, S.S., Lundstrøm, I.K.C., Turner-Walker, G., Moore, K.H.S., Luisi, P., Margaryan, A., Martin, M.D., Ellegaard, M.R., Magnússon, Ó.Þ., et al. (2022). The population genomic legacy of the second plague pandemic. Curr. Biol. CB *32*, 4743-4751.e6. https://doi.org/10.1016/j.cub.2022.09.023.

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

86. Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., and Pääbo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. PloS One *7*, e34131. https://doi.org/10.1371/journal.pone.0034131.

87. Zhang, X., Liu, T., Wang, J., Wang, P., Qiu, Y., Zhao, W., Pang, S., Li, X., Wang, H., Song, J., et al. (2021). Pan-genome of *Raphanus* highlights genetic variation and introgression among domesticated, wild, and weedy radishes. Mol. Plant *14*, 2032–2055. https://doi.org/10.1016/j.molp.2021.08.005.

88. Zhou, Y., Yang, L., Han, X., Han, J., Hu, Y., Li, F., Xia, H., Peng, L., Boschiero, C., Rosen, B.D., et al. (2022). Assembly of a pangenome for global cattle reveals missing sequences and novel structural variations, providing new insights into their diversity and evolutionary history. Genome Res. *32*, 1585–1601. https://doi.org/10.1101/gr.276550.122.

89. Martiniano, R., Garrison, E., Jones, E.R., Manica, A., and Durbin, R. (2020). Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. Genome Biol. *21*, 250. https://doi.org/10.1186/s13059-020-02160-7.

90. Sirak, K., Fernandes, D., Cheronet, O., Harney, E., Mah, M., Mallick, S., Rohland, N., Adamski, N., Broomandkhoshbacht, N., Callan, K., et al. (2020). Human auditory ossicles as an alternative optimal source of ancient DNA. Genome Res. *30*, 427–436. https://doi.org/10.1101/gr.260141.119.

91. Ibrahim, J., Brumfeld, V., Addadi, Y., Rubin, S., Weiner, S., and Boaretto, E. (2022). The petrous bone contains high concentrations of osteocytes: One possible reason why ancient DNA is better preserved in this bone. PLOS ONE *17*, e0269348. https://doi.org/10.1371/journal.pone.0269348.

92. Rohland, N., Glocke, I., Aximu-Petri, A., and Meyer, M. (2018). Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing. Nat. Protoc. *13*, 2447–2461. https://doi.org/10.1038/s41596-018-0050-5.

93. Kapp, J.D., Green, R.E., and Shapiro, B. (2021). A Fast and Efficient Single-stranded Genomic Library Preparation Method Optimized for Ancient DNA. J. Hered. *112*, 241–249. https://doi.org/10.1093/jhered/esab012.

94. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

95. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biol. *20*, 257. https://doi.org/10.1186/s13059-019-1891-0.

96. Lu, J., Rincon, N., Wood, D.E., Breitwieser, F.P., Pockrandt, C., Langmead, B., Salzberg, S.L., and Steinegger, M. (2022). Metagenome analysis using the Kraken software suite. Nat. Protoc. *17*, 2815–2839. https://doi.org/10.1038/s41596-022-00738-y.

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

97. Campana, M.G., Parker, L.D., Hawkins, M.T.R., Young, H.S., Helgen, K.M., Szykman Gunther, M., Woodroffe, R., Maldonado, J.E., and Fleischer, R.C. (2016). Genome sequence, population history, and pelage genetics of the endangered African wild dog (*Lycaon pictus*). BMC Genomics *17*, 1013. https://doi.org/10.1186/s12864-016-3368-9.

98. Kukekova, A.V., Johnson, J.L., Xiang, X., Feng, S., Liu, S., Rando, H.M., Kharlamova, A.V., Herbeck, Y., Serdyukova, N.A., Xiong, Z., et al. (2018). Red fox genome assembly identifies genomic regions associated with tame and aggressive behaviours. Nat. Ecol. Evol. *2*, 1479–1491. https://doi.org/10.1038/s41559-018-0611-6.

99. Armstrong, E.E., Taylor, R.W., Prost, S., Blinston, P., van der Meer, E., Madzikanda, H., Mufute, O., Mandisodza-Chikerema, R., Stuelpnagel, J., Sillero-Zubiri, C., et al. (2019). Cost-effective assembly of the African wild dog (*Lycaon pictus*) genome using linked reads. GigaScience *8*. https://doi.org/10.1093/gigascience/giy124.

100. Wang, G.-D., Shao, X.-J., Bai, B., Wang, J., Wang, X., Cao, X., Liu, Y.-H., Wang, X., Yin, T.-T., Zhang, S.-J., et al. (2019). Structural variation during dog domestication: insights from gray wolf and dhole genomes. Natl. Sci. Rev. *6*, 110–122. https://doi.org/10.1093/nsr/nwy076.

101. Allio, R., Tilak, M.-K., Scornavacca, C., Avenant, N.L., Kitchener, A.C., Corre, E., Nabholz, B., and Delsuc, F. (2021). High-quality carnivoran genomes from roadkill samples enable comparative species delineation in aardwolf and bat-eared fox. eLife *10*, e63167. https://doi.org/10.7554/eLife.63167.

102. Edwards, R.J., Field, M.A., Ferguson, J.M., Dudchenko, O., Keilwagen, J., Rosen, B.D., Johnson, G.S., Rice, E.S., Hillier, L.D., Hammond, J.M., et al. (2021). Chromosome-length genome assembly and structural variations of the primal Basenji dog (*Canis lupus familiaris*) genome. BMC Genomics *22*, 188. https://doi.org/10.1186/s12864-021-07493-6.

103. Halo, J.V., Pendleton, A.L., Shen, F., Doucet, A.J., Derrien, T., Hitte, C., Kirby, L.E., Myers, B., Sliwerska, E., Emery, S., et al. (2021). Long-read assembly of a Great Dane genome highlights the contribution of GC-rich sequence and mobile elements to canine genomes. Proc. Natl. Acad. Sci. *118*, e2016274118. https://doi.org/10.1073/pnas.2016274118.

104. Jagannathan, V., Hitte, C., Kidd, J.M., Masterson, P., Murphy, T.D., Emery, S., Davis, B., Buckley, R.M., Liu, Y.-H., Zhang, X.-Q., et al. (2021). Dog10K_Boxer_Tasha_1.0: a long-read assembly of the Dog reference genome. Genes *12*, 847. https://doi.org/10.3390/genes12060847.

105. Peng, Y., Li, H., Liu, Z., Zhang, C., Li, K., Gong, Y., Geng, L., Su, J., Guan, X., Liu, L., et al. (2021). Chromosome-level genome assembly of the Arctic fox (*Vulpes lagopus*) using PacBio sequencing and Hi-C technology. Mol. Ecol. Resour. *21*, 2093–2108. https://doi.org/10.1111/1755-0998.13397.

106. Wang, C., Wallerman, O., Arendt, M.-L., Sundström, E., Karlsson, Å., Nordin, J., Mäkeläinen, S., Pielberg, G.R., Hanson, J., Ohlsson, Å., et al. (2021). A novel canine reference genome

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

resolves genomic architecture and uncovers transcript complexity. Commun. Biol. *4*, 185. https://doi.org/10.1038/s42003-021-01698-x.

107. Chavez, D.E., Gronau, I., Hains, T., Dikow, R.B., Frandsen, P.B., Figueiró, H.V., Garcez, F.S., Tchaicka, L., De Paula, R.C., Rodrigues, F.H.G., et al. (2022). Comparative genomics uncovers the evolutionary history, demography, and molecular adaptations of South American canids. Proc. Natl. Acad. Sci. *119*, e2205986119. https://doi.org/10.1073/pnas.2205986119.

108. Field, M.A., Yadav, S., Dudchenko, O., Esvaran, M., Rosen, B.D., Skvortsova, K., Edwards, R.J., Keilwagen, J., Cochran, B.J., Manandhar, B., et al. (2022). The Australian dingo is an early offshoot of modern breed dogs. Sci. Adv. *8*, eabm5944. https://doi.org/10.1126/sciadv.abm5944.

109. Lan, T., Li, H., Yang, S., Shi, M., Han, L., Sahu, S.K., Lu, Y., Wang, J., Zhou, M., Liu, H., et al. (2022). The chromosome-scale genome of the raccoon dog: Insights into its evolutionary characteristics. iScience *25*, 105117. https://doi.org/10.1016/j.isci.2022.105117.

110. Lyu, T.-S., Wei, Q.-G., Wang, L.-D., Zhou, S.-Y., Shi, L.-P., Dong, Y.-H., Dou, H.-S., Sha, W.-L., Ga, T., Zhang, H.-H., et al. (2022). High-quality chromosome-level genome assembly of Tibetan fox (*Vulpes ferrilata*). Zool. Res. *43*, 362–366. https://doi.org/10.24272/j.issn.2095-8137.2021.399.

111. The Darwin Tree of Life Project Consortium, Blaxter, M., Mieszkowska, N., Di Palma, F., Holland, P., Durbin, R., Richards, T., Berriman, M., Kersey, P., Hollingsworth, P., et al. (2022). Sequence locally, think globally: The Darwin Tree of Life Project. Proc. Natl. Acad. Sci. *119*, e2115642118. https://doi.org/10.1073/pnas.2115642118.

112. Armstrong, E.E., Bissell, K.L., Heikkinen, M.A., Jessup, A., Junaid, M.O., Lee, D.H., Lieb, E.C., Liem, J.T., Martin, E.M., Moreno, M., et al. (2023). Chromosome-level assembly of the gray fox (*Urocyon cinereoargenteus*) confirms the basal loss of *PRDM9* in Canidae. Preprint at Genomics, https://doi.org/10.1101/2023.11.08.566296 https://doi.org/10.1101/2023.11.08.566296.

113. Bredemeyer, K.R., vonHoldt, B.M., Foley, N.M., Childers, I.R., Brzeski, K.E., and Murphy, W.J. (2024). The value of hybrid genomes: Building two highly contiguous reference genome assemblies to advance *Canis* genomic studies. J. Hered. *115*, 480–486. https://doi.org/10.1093/jhered/esae013.

114. Kliver, S., Kovacic, I., Mak, S., Sinding, M.-H.S., Stagegaard, J., Petersen, B., Nesme, J., and Gilbert, M.T.P. (2024). A chromosome phased diploid genome assembly of African hunting dog (*Lycaon pictus*). J. Hered., esae052. https://doi.org/10.1093/jhered/esae052.

115. Sinding, M.-H.S., Gopalakrishan, S., Vieira, F.G., Samaniego Castruita, J.A., Raundrup, K., Heide Jørgensen, M.P., Meldgaard, M., Petersen, B., Sicheritz-Ponten, T., Mikkelsen, J.B., et al. (2018). Population genomics of grey wolves and wolf-like canids in North America.

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

PLOS Genet. *14*, e1007745. https://doi.org/10.1371/journal.pgen.1007745.

116. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

117. Oliva, A., Tobler, R., Llamas, B., and Souilmi, Y. (2021). Additional evaluations show that specific *BWA-aln* settings still outperform *BWA-mem* for ancient DNA data alignment. Ecol. Evol. *11*, 18743–18748. https://doi.org/10.1002/ece3.8297.

118. Schubert, M., Ermini, L., Sarkissian, C.D., Jónsson, H., Ginolhac, A., Schaefer, R., Martin, M.D., Fernández, R., Kircher, M., McCue, M., et al. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat. Protoc. *9*, 1056–1082. https://doi.org/10.1038/nprot.2014.063.

119. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L.F., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics *29*, 1682–1684. https://doi.org/10.1093/bioinformatics/btt193.

120. Neukamm, J., Peltzer, A., and Nieselt, K. (2021). DamageProfiler: fast damage pattern calculation for ancient DNA. Bioinformatics *37*, 3652–3653. https://doi.org/10.1093/bioinformatics/btab190.

121. Pedersen, B.S., and Quinlan, A.R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics *34*, 867–868. https://doi.org/10.1093/bioinformatics/btx699.

122. Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics *32*, 292–294. https://doi.org/10.1093/bioinformatics/btv566.

123. Wurster-Hill, D.H., Ward, O.G., Davis, B.H., Park, J.P., Moyzis, R.K., and Meyne, J. (1988). Fragile sites, telomeric DNA sequences, B chromosomes, and DNA content in raccoon dogs, *Nyctereutes procyonoides,* with comparative notes on foxes, coyote, wolf, and raccoon. Cytogenet. Genome Res. *49*, 278–281. https://doi.org/10.1159/000132677.

124. Tiersch, T.R., Chandler, R.W., Wachtel, S.S., and Elias, S. (1989). Reference standards for flow cytometry and application in comparative studies of nuclear DNA content. Cytometry *10*, 706–710. https://doi.org/10.1002/cyto.990100606.

125. Smeds, L., Kojola, I., and Ellegren, H. (2019). The evolutionary history of grey wolf Y chromosomes. Mol. Ecol. *28*, 2173–2191. https://doi.org/10.1111/mec.15054.

126. Hamilton, C.K., Favetta, L.A., Di Meo, G.P., Floriot, S., Perucatti, A., Peippo, J., Kantanen, J., Eggen, A., Iannuzzi, L., and King, W.A. (2009). Copy Number Variation of Testis-Specific Protein, Y-Encoded *(TSPY)* in 14 Different Breeds of Cattle *(Bos taurus)*.

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sex. Dev. *3*, 205–213. https://doi.org/10.1159/000228721.

127. Xue, Y., and Tyler-Smith, C. (2011). An exceptional gene: evolution of the TSPY gene family in humans and other great apes. Genes *2*, 36–47. https://doi.org/10.3390/genes2010036.

128. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

129. Hickey, G., Monlong, J., Ebler, J., Novak, A.M., Eizenga, J.M., Gao, Y., Human Pangenome Reference Consortium, Abel, H.J., Antonacci-Fulton, L.L., Asri, M., et al. (2023). Pangenome graph construction from genome alignments with Minigraph-Cactus. Nat. Biotechnol. https://doi.org/10.1038/s41587-023-01793-w.

130. Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J., et al. (2023). A draft human pangenome reference. Nature *617*, 312–324. https://doi.org/10.1038/s41586-023-05896-x.

131. Sirén, J., Monlong, J., Chang, X., Novak, A.M., Eizenga, J.M., Markello, C., Sibbesen, J.A., Hickey, G., Chang, P.-C., Carroll, A., et al. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science *374*, abg8871. https://doi.org/10.1126/science.abg8871.

132. Hickey, G., Heller, D., Monlong, J., Sibbesen, J.A., Sirén, J., Eizenga, J., Dawson, E.T., Garrison, E., Novak, A.M., and Paten, B. (2020). Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biol. *21*, 35. https://doi.org/10.1186/s13059-020-1941-7.

133. Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS ONE *9*, e112963. https://doi.org/10.1371/journal.pone.0112963.

134. Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics *15*, 356. https://doi.org/10.1186/s12859-014-0356-4.

135. Barlow, A., Hartmann, S., Gonzalez, J., Hofreiter, M., and Paijmans, J.L.A. (2020). Consensify: A Method for Generating Pseudohaploid Genome Sequences from Palaeogenomic Datasets with Reduced Error Rates. Genes *11*, 50. https://doi.org/10.3390/genes11010050.

136. Auwera, G.A.V. de, and O'Connor, B.D. (2020). Genomics in the cloud: using Docker, GATK, and WDL in Terra First edition. (O'Reilly).

137. Troy, W., Damas, J., Titus, A.J., and Cantarel, B.L. (2023). FLAG: Find, Label, Annotate Genomes, a fully automated tool for genome gene structural and functional annotation

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of highly fragmented non-model species. Preprint, https://doi.org/10.1101/2023.07.14.548907 https://doi.org/10.1101/2023.07.14.548907.

138. Barlow, A., Hartmann, S., Gonzalez, J., Hofreiter, M., and Paijmans, J.L.A. (2020). Consensify: A Method for Generating Pseudohaploid Genome Sequences from Palaeogenomic Datasets with Reduced Error Rates. Genes *11*, 50. https://doi.org/10.3390/genes11010050.

139. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

140. Durbin, R., De Sanctis, B., and Blumer, M. (2023). Rotate: A command-line program to rotate circular DNA sequences to start at a given position or string. Wellcome Open Res. *8*, 401. https://doi.org/10.12688/wellcomeopenres.19568.1.

141. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340.

142. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033.

143. Atickem, A., Stenseth, N.Chr., Drouilly, M., Bock, S., Roos, C., and Zinner, D. (2018). Deep divergence among mitochondrial lineages in African jackals. Zool. Scr. *47*, 1–8. https://doi.org/10.1111/zsc.12257.

144. Hempel, E., Bibi, F., Faith, J.T., Brink, J.S., Kalthoff, D.C., Kamminga, P., Paijmans, J.L.A., Westbury, M.V., Hofreiter, M., and Zachos, F.E. (2021). Identifying the true number of specimens of the extinct blue antelope (Hippotragus leucophaeus). Sci. Rep. *11*, 2100. https://doi.org/10.1038/s41598-020-80142-2.

145. Hempel, E., Faith, J.T., Preick, M., de Jager, D., Barish, S., Hartmann, S., Grau, J.H., Moodley, Y., Gedman, G., Pirovich, K.M., et al. (2024). Colonial-driven extinction of the blue antelope despite genomic adaptation to low population size. Curr. Biol. *34*, 2020-2029.e6. https://doi.org/10.1016/j.cub.2024.03.051.

146. Mirarab, S., Reaz, R., Bayzid, Md.S., Zimmermann, T., Swenson, M.S., and Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics *30*, i541–i548. https://doi.org/10.1093/bioinformatics/btu462.

147. Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics *19*, 153. https://doi.org/10.1186/s12859-018-2129-y.

148. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591. https://doi.org/10.1093/molbev/msm088.

bioRxiv preprint doi: https://doi.org/10.1101/2025.04.09.647074; this version posted April 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

149. Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. *22*, 160–174. https://doi.org/10.1007/BF02101694.

150. Puttick, M.N. (2019). MCMCtreeR: functions to prepare MCMCtree analyses and visualize posterior ages on trees. Bioinformatics *35*, 5321–5322. https://doi.org/10.1093/bioinformatics/btz554.

151. Yun, T., Li, H., Chang, P.-C., Lin, M.F., Carroll, A., and McLean, C.Y. (2021). Accurate, scalable cohort variant calls using DeepVariant and GLnexus. Bioinformatics *36*, 5582–5589. https://doi.org/10.1093/bioinformatics/btaa1081.

152. Lin, M.F., Rodeh, O., Penn, J., Bai, X., Reid, J.G., Krasheninina, O., and Salerno, W.J. (2018). GLnexus: joint variant calling for large cohort sequencing. Preprint, https://doi.org/10.1101/343970 https://doi.org/10.1101/343970.

153. Pertea, G., and Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. F1000Research *9*, 304. https://doi.org/10.12688/f1000research.23297.1.

154. Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLOS ONE *11*, e0163962. https://doi.org/10.1371/journal.pone.0163962.

155. Muse, S., and Gaut, B. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. https://doi.org/10.1093/oxfordjournals.molbev.a040152.

156. Villanueva-Cañas, J.L., Laurie, S., and Albà, M.M. (2013). Improving Genome-Wide Scans of Positive Selection by Using Protein Isoforms of Similar Length. Genome Biol. Evol. *5*, 457–467. https://doi.org/10.1093/gbe/evt017.

157. Kolberg, L., Raudvere, U., Kuzmin, I., Adler, P., Vilo, J., and Peterson, H. (2023). g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic Acids Res. *51*, W207–W212. https://doi.org/10.1093/nar/gkad347.