Title

Comparative genomics of Balto, a famous historic dog, captures lost diversity of 1920s sled dogs

Journal

Science, 380(6643)

ISSN

0036-8075

Authors

Moon, Katherine L
Huson, Heather J
Morrill, Kathleen
et al.

Publication Date

2023-04-28

DOI

10.1126/science.abn5887

Peer reviewed

Published in final edited form as:

Science. 2023 April 28; 380(6643): eabn5887. doi:10.1126/science.abn5887.

*Corresponding authors. katielouisemoon@gmail.com (KMM).
§Consortium authors and affiliations listed at end of manuscript.

Author contributions:
Conceptualization: HJH, GS, EKK, BS
Data Acquisition: KLM, HJH, BS, GS
Analysis: KLM, HJH, KM, MSW, XL, KS, EKK
Writing: KLM, HJH, KM, XL, EKK, BS

Competing interests: Authors declare no competing interests.

Consortium list:
Gregory Andrews[1], Joel C. Armstrong[2], Matteo Bianchi[3], Bruce W. Birren[4], Kevin R. Bredemeyer[5], Ana M. Breit[6], Matthew J. Christmas[3], Hiram Clawson[2], Joana Damas[7], Federica Di Palma[8,9], Mark Diekhans[2], Michael X. Dong[3], Eduardo Eizirik[10], Kaili Fan[1], Cornelia Fanter[11], Nicole M. Foley[5], Karin Forsberg-Nilsson[12,13], Carlos J. Garcia[14], John Gatesy[15], Steven Gazal[16], Diane P. Genereux[4], Linda Goodman[17], Jenna Grimshaw[14], Michaela K. Halsey[14], Andrew J. Harris[5], Glenn Hickey[18], Michael Hiller[19,20,21], Allyson G. Hindle[11], Robert M. Hubley[22], Graham M. Hughes[23], Jeremy Johnson[4], David Juan[24], Irene M. Kaplow[25,26], Elinor K. Karlsson[1,4,27], Kathleen C. Keough[17,28,29], Bogdan Kirilenko[19,20,21], Klaus-Peter Koepfli[30,31,32], Jennifer M. Korstian[14], Amanda Kowalczyk[25,26], Sergey V. Kozyrev[3], Alyssa J. Lawler[4,26,33], Colleen Lawless[23], Thomas Lehmann[34], Danielle L. Levesque[6], Harris A. Lewin[7,35,36], Xue Li[1,4,37], Abigail Lind[28,29], Kerstin Lindblad-Toh[3,4], Aya Mackay-Smith[38], Voichita D. Marinescu[3], Tomas Marques-Bonet[39,40,41,42], Victor C. Mason[43], Jennifer R. S. Meadows[3], Wynn K. Meyer[44], Jill E. Moore[1], Lucas R. Moreira[1,4], Diana D. Moreno-Santillan[14], Kathleen M. Morrill[1,4,37], Gerard Muntané[24], William J. Murphy[5], Arcadi Navarro[39,41,45,46], Martin Nweeia[47,48,49,50], Sylvia Ortmann[51], Austin Osmanski[14], Benedict Paten[2], Nicole S. Paulat[14], Andreas R. Pfenning[25,26], BaDoi N. Phan[25,26,52], Katherine S. Pollard[28,29,53], Henry E. Pratt[1], David A. Ray[14], Steven K. Reilly[38], Jeb R. Rosen[22], Irina Ruf[54], Louise Ryan[23], Oliver A. Ryder[55,56], Pardis C. Sabeti[4,57,58], Daniel E. Schäffer[25], Aitor Serres[24], Beth Shapiro[59,60], Arian F. A. Smit[22], Mark Springer[61], Chaitanya Srinivasan[25], Cynthia Steiner[55], Jessica M. Storer[22], Kevin A. M. Sullivan[14], Patrick F. Sullivan[62,63], Elisabeth Sundström[3], Megan A. Supple[59], Ross Swofford[4], Joy-El Talbot[64], Emma Teeling[23], Jason Turner-Maier[4], Alejandro Valenzuela[24], Franziska Wagner[65], Ola Wallerman[3], Chao Wang[3], Juehan Wang[16], Zhiping Weng[1], Aryn P. Wilder[55], Morgan E. Wirthlin[25,26,66], James R. Xue4,57, Xiaomeng Zhang[4,25,26]
Affiliations:
[1]Program in Bioinformatics and Integrative Biology, UMass Chan Medical School; Worcester, MA 01605, USA.
[2]Genomics Institute, University of California Santa Cruz; Santa Cruz, CA 95064, USA.
[3]Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University; Uppsala, 751 32, Sweden.
[4]Broad Institute of MIT and Harvard; Cambridge, MA 02139, USA.
[5]Veterinary Integrative Biosciences, Texas A&M University; College Station, TX 77843, USA.
[6]School of Biology and Ecology, University of Maine; Orono, ME 04469, USA.
[7]The Genome Center, University of California Davis; Davis, CA 95616, USA.
[8]Genome British Columbia; Vancouver, BC, Canada.
[9]School of Biological Sciences, University of East Anglia; Norwich, UK.
[10]School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul; Porto Alegre, 90619–900, Brazil.
[11]School of Life Sciences, University of Nevada Las Vegas; Las Vegas, NV 89154, USA.
[12]Biodiscovery Institute, University of Nottingham; Nottingham, UK.
[13]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University; Uppsala, 751 85, Sweden.
[14]Department of Biological Sciences, Texas Tech University; Lubbock, TX 79409, USA.
[15]Division of Vertebrate Zoology, American Museum of Natural History; New York, NY 10024, USA.
[16]Keck School of Medicine, University of Southern California; Los Angeles, CA 90033, USA.
[17]Fauna Bio Incorporated; Emeryville, CA 94608, USA.
[18]Baskin School of Engineering, University of California Santa Cruz; Santa Cruz, CA 95064, USA.
[19]Faculty of Biosciences, Goethe-University; 60438 Frankfurt, Germany.
[20]LOEWE Centre for Translational Biodiversity Genomics; 60325 Frankfurt, Germany.
[21]Senckenberg Research Institute; 60325 Frankfurt, Germany.
[22] Institute for Systems Biology; Seattle, WA 98109, USA.
[23]School of Biology and Environmental Science, University College Dublin; Belfield, Dublin 4, Ireland.
[24]Department of Experimental and Health Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra; Barcelona, 08003, Spain.
[25]Department of Computational Biology, School of Computer Science, Carnegie Mellon University; Pittsburgh, PA 15213, USA.
[26]Neuroscience Institute, Carnegie Mellon University; Pittsburgh, PA 15213, USA.
[27]Program in Molecular Medicine, UMass Chan Medical School; Worcester, MA 01605, USA.
[28]Department of Epidemiology & Biostatistics, University of California San Francisco; San Francisco, CA 94158, USA.
[29]Gladstone Institutes; San Francisco, CA 94158, USA.
[30]Center for Species Survival, Smithsonian's National Zoo and Conservation Biology Institute; Washington, DC 20008, USA.

Comparative genomics of Balto, a famous historic dog, captures lost diversity of 1920s sled dogs

Katherine L. Moon[1,2,*], **Heather J. Huson**[3], **Kathleen Morrill**[4,5,6], **Ming-Shan Wang**[1,2], **Xue Li**[4,5,6], **Krishnamoorthy Srikanth**[3],
Zoonomia Consortium[§],
Kerstin Lindblad-Toh[6,7], **Gavin J. Svenson**[8], **Elinor K. Karlsson**[4,5], **Beth Shapiro**[1,2]

[1] Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA, USA

[2] Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA, USA

[3] Department of Animal Sciences, Cornell University College of Agriculture and Life Sciences, Ithaca, NY, 14853, USA

[31]Computer Technologies Laboratory, ITMO University; St. Petersburg 197101, Russia.
[32]Smithsonian-Mason School of Conservation, George Mason University; Front Royal, VA 22630, USA.
[33]Department of Biological Sciences, Mellon College of Science, Carnegie Mellon University; Pittsburgh, PA 15213, USA.
[34]Senckenberg Research Institute and Natural History Museum Frankfurt; 60325 Frankfurt am Main, Germany.
[35]Department of Evolution and Ecology, University of California Davis; Davis, CA 95616, USA.
[36]John Muir Institute for the Environment, University of California Davis; Davis, CA 95616, USA.
[37]Morningside Graduate School of Biomedical Sciences, UMass Chan Medical School; Worcester, MA 01605, USA.
[38]Department of Genetics, Yale School of Medicine; New Haven, CT 06510, USA.
[39]Catalan Institution of Research and Advanced Studies (ICREA); Barcelona, 08010, Spain.
[40]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST); Barcelona, 08036, Spain.
[41]Department of Medicine and LIfe Sciences, Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra; Barcelona, 08003, Spain.
[42] Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona; 08193, Cerdanyola del Vallès, Barcelona, Spain.
[43] Institute of Cell Biology, University of Bern; 3012, Bern, Switzerland.
[44]Department of Biological Sciences, Lehigh University; Bethlehem, PA 18015, USA.
[45]BarcelonaBeta Brain Research Center, Pasqual Maragall Foundation; Barcelona, 08005, Spain.
[46]CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST); Barcelona, 08003, Spain.
[47]Department of Comprehensive Care, School of Dental Medicine, Case Western Reserve University; Cleveland, OH 44106, USA.
[48]Department of Vertebrate Zoology, Canadian Museum of Nature; Ottawa, Ontario K2P 2R1, Canada.
[49]Department of Vertebrate Zoology, Smithsonian Institution; Washington, DC 20002, USA.
[50]Narwhal Genome Initiative, Department of Restorative Dentistry and Biomaterials Sciences, Harvard School of Dental Medicine; Boston, MA 02115, USA.
[51]Department of Evolutionary Ecology, Leibniz Institute for Zoo and Wildlife Research; 10315 Berlin, Germany.
[52]Medical Scientist Training Program, University of Pittsburgh School of Medicine; Pittsburgh, PA 15261, USA.
[53]Chan Zuckerberg Biohub; San Francisco, CA 94158, USA.
[54]Division of Messel Research and Mammalogy, Senckenberg Research Institute and Natural History Museum Frankfurt; 60325 Frankfurt am Main, Germany.
[55]Conservation Genetics, San Diego Zoo Wildlife Alliance; Escondido, CA 92027, USA.
[56]Department of Evolution, Behavior and Ecology, School of Biological Sciences, University of California San Diego; La Jolla, CA 92039, USA.
[57]Department of Organismic and Evolutionary Biology, Harvard University; Cambridge, MA 02138, USA.
[58]Howard Hughes Medical Institute; Chevy Chase, MD, USA.
[59]Department of Ecology and Evolutionary Biology, University of California Santa Cruz; Santa Cruz, CA 95064, USA.
[60]Howard Hughes Medical Institute, University of California Santa Cruz; Santa Cruz, CA 95064, USA.
[61]Department of Evolution, Ecology and Organismal Biology, University of California Riverside; Riverside, CA 92521, USA.
[62]Department of Genetics, University of North Carolina Medical School; Chapel Hill, NC 27599, USA.
[63]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet; Stockholm, Sweden.
[64] Iris Data Solutions, LLC; Orono, ME 04473, USA.
[65]Museum of Zoology, Senckenberg Natural History Collections Dresden; 01109 Dresden, Germany.
[66]Allen Institute for Brain Science; Seattle, WA 98109, USA

[4] Bioinformatics and Integrative Biology, UMass Chan Medical School, Worcester, MA 01655, USA

[5] Morningside Graduate School of Biomedical Sciences, UMass Chan Medical School, Worcester, MA 01655, USA

[6] Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

[7] Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University; Uppsala, 751 32, Sweden.

[8] Cleveland Museum of Natural History, Cleveland, OH 44106, USA

Abstract

We reconstruct the phenotype of Balto, the heroic sled dog renowned for transporting diphtheria antitoxin to Nome, Alaska in 1925, using evolutionary constraint estimates from the Zoonomia alignment of 240 mammals and 682 genomes from dogs and wolves of the 21st century. Balto shares just part of his diverse ancestry with the eponymous Siberian husky breed. Balto's genotype predicts a combination of coat features atypical for modern sled dog breeds, and a slightly smaller stature. He had enhanced starch digestion compared with Greenland sled dogs and a compendium of derived homozygous coding variants at constrained positions in genes connected to bone and skin development. We propose that Balto's population of origin, which was less inbred and genetically healthier than modern breeds, was adapted to the extreme environment of 1920s Alaska.

One-Sentence Summary:

Comparative genomics uncovers genotype-phenotype links between Balto, famed sled dog of the 1925 Serum Run, and modern dogs.

Technological advances in the recovery of ancient DNA make it possible to generate high-coverage nuclear genomes from historic and fossil specimens, but interpreting genetic data from past individuals is difficult without data from their contemporaries. Comparative genomic analysis offers a solution: by combining population-level genomic data and catalogs of trait associations in modern populations, we can infer the genetic and phenotypic features of long-dead individuals and the populations from which they were born. Zoonomia is a new comparative resource that addresses limitations of previous datasets (1) to support interpretation of paleogenomics data. With 240 placental mammal species, Zoonomia has sufficient power to distinguish individual bases under evolutionary constraint - a useful predictor of functional importance (2) - in coding and regulatory elements (3). Zoonomia's reference-free genome alignment (4, 5) allows evolutionary constraint to be scored in any of its 240 species, including dogs.

Here, we generate a genome for Balto, the famous sled dog who delivered diphtheria serum to the children of Nome, Alaska, during a 1925 outbreak. Following his death, Balto was taxidermied and his remains are held by the Cleveland Museum of Natural History. We generated a 40.4-fold coverage nuclear genome from Balto's underbelly skin using protocols for degraded samples. His DNA was well preserved, with an average endogenous content of

87.7% in sequencing libraries, low (<1%) damage rates (fig. S1) and short (68bp) average fragment sizes, consistent with the age of the sample.

Balto was born in the kennel of sled dog breeder Leonard Seppala in 1919. Although Seppala's small fast dogs were known as Siberian huskies (6), they were a working population that differed from the dog breed recognized by the American Kennel Club (AKC) today. Modern dog breeds are genetically closed populations that conform to a tightly delineated physical standard (7). Balto's relationship to AKC-recognized sled dog breeds like the Siberian husky (established in 1930) and Alaskan malamute (1935) (8) is unclear. Balto himself was neutered at six months of age and had no offspring.

Working populations of sled dogs survive. Alaskan sled dogs are bred solely for physical performance, including outcrossing with various breeds (9). Greenland sled dogs are an indigenous land-race breed that have been used for hunting and sledging by Inuit in Greenland for 850 years, where they have been isolated from contact with other dogs (10). Here, we use the term "breed" exclusively to refer to modern breeds recognized by the AKC or other kennel clubs (e.g. sled dog breeds), as distinct from the less rigidly defined populations of Greenland sled dogs and Alaskan sled dogs (working sled dogs). This is a genetic distinction; AKC-registered dogs can be successful working sled dogs.

We compared Balto to working sled dogs, sled dog breeds, other breeds, village dogs (free-breeding dogs without known breed ancestry), and other canids. Our whole genome dataset comprised 688 dogs (table S1) representing 135 breeds/populations, including three Alaskan sled dogs and five Greenland sled dogs (10). We identified evolutionarily constrained bases using phyloP evolutionary constraint scores from the dog-referenced version of the 240 species Zoonomia alignment (3).

Ancestry analysis places Balto in a clade of sled dog breeds and working sled dogs and closest to the Alaskan sled dogs (Fig. 1A,B). Most of his ancestry is assigned to clades of Arctic-origin dogs (68%) and, to a lesser extent, Asian-origin dogs (24%) in an unsupervised admixture analysis with 2166 dogs and 116 clusters (Fig. 1C, table S2, S3). He carried no discernible wolf ancestry. The more recently established Alaskan sled dog population (9) did not fall into a distinct ancestry cluster in the unsupervised analysis, but comprised 34% of Balto's ancestry in a supervised analysis defining them as a cluster (fig. S2).

Balto was more genetically diverse than breed dogs today and similar to working sled dogs (Fig. 1D). Balto had shorter runs of homozygosity than any breed dog, and fewer runs of homozygosity than all but one Tibetan mastiff (table S4). When inbreeding is calculated from runs of homozygosity, Balto and the two working sled dog populations are lower than almost any breed dog (fig. S3). When inbreeding is calculated using an allele frequency approach (method-of-moment), Greenland sled dogs have high inbreeding coefficients, reflecting their long genetic isolation in Greenland (fig. S3).

To evaluate the genetic health of Balto's population of origin, we developed an analytical approach that leveraged the Zoonomia 240 species constraint scores and required only a single dog from each population (necessary since Balto is the only available representative of his population). Briefly, we selected one individual at random from each breed or

population (57 dogs total) and scored variant positions as either evolutionarily constrained (and more likely to be damaging (2)) or not using the Zoonomia phyloP scores (3). We also identified variants likely to be “rare” (low frequency) in each dog’s breed or population. Because we couldn’t directly measure population allele frequencies with only a single representative dog, we defined “rare” variants as heterozygous or homozygous variants unique to that dog among all 57 representative dogs. This metric effectively identifies variants occurring at unusually low frequencies (fig. S4).

Balto and modern working sled dogs had a lower burden of rare, potentially damaging variation, indicating they represent genetically healthier populations (11) than breed dogs. Balto and the working sled dogs had significantly fewer potentially damaging variants (missense or constrained) than any breed dog, including the sled dog breeds (Fig. 1E). The pattern persists even in the less genetically diverse Greenland sled dog. Selection for fitness in working sled dog populations appears more effective in removing damaging genetic variation than selection to meet a breed standard.

Balto’s physical appearance predicted from his genome sequence (Fig. 2A, table S5) matches historical photos (Fig. 2B) and his taxidermied remains, indicating that the same variants shaping modern breed phenotypes also explained natural variation in his pre-breed working population. We predict that he stood 55cm tall at his shoulders (12)(Fig. 2C), within the acceptable range for today’s Siberian husky breed (53–60cm (8)), and had a double layered coat (13) that was mostly black with only a little bit of white (14). He was homozygous for an allele conferring tan points (15) and one for blue eyes (16), but both were masked by his melanistic facial mask (17), and his predicted light-tan pigmentation (18) may have been indistinguishable from white. He carried neither the “wolf agouti” nor “Northern domino” patterns that are common in the Siberian husky and other sled dog breeds today (19).

Both Balto and Alaskan sled dogs had unexpected evidence of adaptation to starch-rich diets. They carry the dog version of *MGAM*, a gene involved in starch processing that is differentiated between dogs and wolves (20) and one of fourteen regions analyzed for evidence of selective pressure in Balto’s lineage using a gene tree analysis (table S6). In earlier work, the high frequency of the wolf version of *MGAM* in Greenland sled dogs prompted speculation that reduced starch digestion might be a working sled dog trait (10). Our findings suggest this phenomenon is specific to Greenland sled dogs. Gene tree analysis places one of Balto’s chromosomes in the ancestral wolf cluster, and one to the derived dog cluster (fig. S5). Most Alaskan sled dogs carry the dog version (frequency=0.83). However, read coverage of the gene *AMY2B* suggests Balto had fewer copies of this gene than many modern dogs, and thus comparatively lower production of the starch-digesting enzyme amylase (21, 22). Taken together, we suggest Balto’s ability to digest starch was enhanced compared to wolves and Greenland sled dogs, but reduced compared to modern breeds.

Of the other 14 regions tested, most (10/14) lacked sufficient diversity in dogs to resolve phylogenetic relationships. Bootstrap support was weak for two other genes selected in Greenland sled dogs (*CACNA1A* and *MAGI2*). As expected, Balto did not carry versions of *EPAS1* associated with high altitude adaptation (23).

We found an enrichment for unusual function variation in Balto's population consistent with adaptation to the extreme environments in which early 20th century sled dogs worked. We identified variants in Balto's genome that were new (not seen in wolves) and likely to be common in his population (homozygous in Balto; fig. S4). We further filtered for variants that were both protein-altering (missense) and evolutionarily constrained (FDR<0.01), and thus likely to be functional. Balto was no more likely to carry such variants than dogs from 54 other populations (fig. S6), but in Balto these variants tended to disrupt tissue development genes (GO:0009888; 24 genes; 3.02-fold enrichment; p_{FDR}=0.013)(table S7). This enrichment was unique to Balto (Fig. 2D, fig. S7), and most of the variants were rare or missing in other dog populations (fig. S8). Even when all GO biological process gene sets are tested in all 57 dogs, Balto's enrichment in tissue development genes is highly unusual. It ranks 4th out of 888,573 dog/set pairs tested (fig. S7, table S8). Phenotype associations from human disease studies suggest that these variants could have influenced skeletal and epithelial development including joint formation, body weight, coordination, and skin thickness (table S9)(24). Modern sled dog breeds and working sled dogs are only slightly more similar to Balto than other dogs at these variants (fig. S9).

Balto was part of a famed population of small, fast, and fit sled dogs imported from Siberia. Following his famous run, the Siberian husky breed was recognized by the AKC. By sequencing his genome from his taxidermied remains and analyzing it in the context of large comparative and canine datasets, we show that Balto shared only part of his ancestry with today's Siberian huskies. Balto's working sled dog contemporaries were healthier and more genetically diverse than modern breeds, and may have carried variants that helped them survive the harsh conditions of 1920s Alaska (6). Further work is still needed to assess the impact of the evolutionarily constrained missense variants that Balto carried. While the era of Balto and his fellow huskies has passed, comparative genomics, supported by a growing collection of modern and past genomes, can provide a snapshot of individuals and populations from the past, as well as insights into the selective pressures that shaped them.

Materials and Methods:

Assembly of comparative canid genetic variants

We collated a reference set of comparative canid genetic variants starting from the curated Broad-UMass Canid Variant set (https://data.broadinstitute.org/DogData/) and comprising whole genome sequencing data for 531 dogs of known breed ancestry distributed among 132 breeds, 28 dogs of mixed breed ancestry, 12 dogs of unknown ancestry, 69 worldwide indigenous or village dogs, 33 wolves, and 1 coyote (see stable S1).

Ancient DNA extraction, library preparation, and genome assembly

We extracted DNA from a ~5mm × 5mm piece of Balto's underbelly skin tissue, in two replicates (HM246 and HM247) with an extraction negative, using the ancient DNA specific protocol in Dabney et al. 2013 (28). We prepared 32 ~1pmol input Illumina libraries from these extracts following the Santa Cruz library preparation method (29), including positive and negative controls. All 32 libraries passed quality control (QC), and so we sequenced

them to a depth of ~2.3 billion on a NovaSeq 6000 platform 150bp paired end (see table S11 for the number of reads produced per library).

We used SeqPrep v.1.1 (30) to trim adapters, remove reads shorter than 28bp, and merge remaining paired-end reads with a minimum overlap of 15 bp. We then used the Burrows-Wheeler Aligner (BWA) v.0.7.12 (31) with a minimum quality cut off of 20 to align reads to the *Canis lupus familiaris* (dog) reference genome (CanFam3.1) (NCBI: GCA_000002285.2). All 32 bam files (one for each library) were merged into one with PCR duplicates removed. We used both Qualimap (v2.2.1) and samtools (v1.7) to calculate metrics and assess the quality of the alignment (see table S12).

Variant calling

We used GATK HaplotypeCaller to call variants in Balto as well as 10 previously published Greenland sled dogs (10) and 3 Alaskan sled dogs sequenced for this study (see Supplementary Methods for details on sampling, DNA extraction, and sequencing) against the UMass-Broad Canid Variant set using parameter *--genotyping-mode* GENOTYPE_GIVEN_ALLELES *--alleles* (known alleles). Then, we merged variant call records from these 14 dogs with records from the UMass-Broad Candid Variants set, for variant calls in a full set of 688 individuals: Balto (this study), 3 modern Alaskan sled dogs (this study), 10 modern Greenland sled dogs (10), 531 dogs from modern breeds, 40 dogs of unknown or admixed ancestry, 69 village or indigenous dogs, 33 wolves, and 1 coyote.

Phylogenetic analysis and neighbor-joining trees

Using a dataset of 100 representative canids (see table S1 for samples selected in the `Phylogenetic Analysis`) we confirmed Balto's phylogenetic position by generating a neighbor-joining (NJ) phylogenetic tree and conducting a principal component analysis (PCA). We converted the variant calls into a FASTA file and used MEGA-CC(33) with 1000 bootstraps to assess tree topology. We also ran a PCA on this set using *PLINK* (v1.9), and then visualized the first two principal components in R (v. 3.6.3) using the `ggplot2` package.

Global ancestry inference

We inferred Balto's ancestral similarity to modern dog breeds, sled dog type breeds, and working sled dogs using a custom built reference panel of modern dogs and canids of the 21st century (table S3). In *PLINK* (v2.00a3LM) (35), we identified 4,267,732 biallelic single nucleotide polymorphisms with <10% missing genotypes, and calculated Wright's F-statistics using Hudson method (36, 37) for (1) each dog breed and sled dog population versus all other dogs; (2) all village dogs versus all other dogs; (3) each regional village dog population; (4) all wolves versus all other dogs; (5) all coyotes versus all other canids; and (6) North American wolves versus Eurasian wolves. We selected 1,858,634 SNPs with F_{ST}>0.5 across all comparisons, and performed LD-based pruning in 250kb windows for r^2>0.2 to extract 136,779 markers for global ancestry inference. We merged Balto's genotypes for these SNPs with genotypes from the reference samples. For reference samples also represented in the whole genome dataset, population labels used in the admixture analysis are given in the `Representative in Global Ancestry Inferencè column of table S1.

We performed global ancestry inference using *ADMIXTURE (38)* in both supervised mode (random seed: 43) with 20 bootstrap replicates to estimate parameter standard errors, and in unsupervised mode for the same number of populations (K=116), which showed low levels of error (0.3) in ten-fold cross-validation analysis of chromosome 1 for K clusters between 50 and 150 (table S13).

Homozygosity and inbreeding metrics

We removed samples with any missing data from the dataset of 100 representative individuals used in the phylogenetic analyses, leaving 86 individuals (see table S1 for samples selected in the 'Homozygosity Analysis'). Using this pruned dataset, we detected runs of homozygosity (RoH) using a window-based approach implemented in *PLINK* (v1.9) (35). We calculated two measures of inbreeding: the method-of-moments coefficient in *PLINK* (F_{MoM}) and the metric based on runs-of-homozygosity (F_{RoH}), as recommended by Zhao et al. 2020 (40) (table S4). Using the *R* (v. 3.6.3) function `cor.test`, we confirmed that F_{RoH} and F_{MoM} are significantly correlated (R_{Pearson}= 0.6752819, p= 9.958e-13, t= 8.3913, df= 84).

Population representative sampling

As Balto is the sole representative of his population, we randomly selected one representative sample from each of 57 populations for the discovery of individually-represented, population-relevant genetic variants (see table S1 for samples selected in the `Population Variants Analysis`) among 67,085,518 biallelic single nucleotide polymorphisms. These populations included Balto, 1 Alaskan sled dog, 1 Greenland sled dog, and 54 modern purebred dogs, including 1 Siberian husky and 1 Alaskan malamute. Likewise, we selected, where available, another 5 to 11 random samples from 10 modern breeds, and all remaining Greenland sled dog samples, to assess the population-wide allele frequency of these variants (see table S1 `Population Frequency Analysis`).

Dog-referenced mammalian evolutionary constraint

We selected biallelic SNPs under evolutionary constraint by examining sites overlapping phyloP evolutionary constraint scores from the dog-referenced version of the 240 species Cactus alignment (3). We calculated the constraint score cutoffs at various false discovery rates (FDR).

Unique, rare, and potentially deleterious variants

We first identified all "population-unique" variants, defined as those observed in the representative dog from a population (either once or twice) and not observed in representatives from any of the other populations. With this method, we identified 206,164 population-unique variants for Balto, 120,279 for the Alaskan sled dog, 119,482 variants for the Greenland sled dog, 120,780 unique to the Alaskan malamute, and 133,200 unique to the Siberian husky. We confirmed that population-unique variants tend to be uncommon by calculating the allele frequencies in its population. We used Zoonomia PhyloP scores and SnpEff(42) annotations to identify which population-unique variants were either "evolutionarily constrained" (phyloP score above the FDR 0.05 cutoff of 2.56) or a

missense mutation and thus more likely to have functional consequences (table S15). We grouped the dogs into working dog groups including Balto, Alaskan sled dog, and Greenland sled dog, and modern breeds including all the other 54 dogs. We then applied Student's t-test on the percentage of "evolutionarily constrained" or missense mutation for the two groups.

Derived, common, and potentially beneficial variants

We identified "homozygous derived" variants, defined as those observed twice in the representative dog from a population and not observed in wolves, for each of the populations. With this method, we identified 176,135 homozygous derived variants for Balto, 148,036 variants for Alaskan sled dog, 260,457 variants for Greenland sled dog, 225,270 variants for Alaskan Malamute, and 189,188 variants for Siberian husky. We confirmed that homozygous variants in each representative dog tend to be "common" in their population by calculating the allele frequency of the homozygous derived variants in its own breed. We also used a Wilcox test against randomly selected SNPs to show that population-unique SNPs are rare, whereas homozygous derived SNPs are rather common, among their population.

We further defined variants likely to be functional as those that were both "highly evolutionarily constrained" (defined by phyloP score above the FDR>0.01 cutoff of 3.52) and a missense mutation. We annotated the variant by genes, and performed gene set enrichment against all Gene Ontology Biological Process gene sets (http://geneontology.org/) using the R package rbioapi v. 0.7.4 (43, 44) (table S7, S8). We also tested for overlap between Balto's variant genes and genes implicated in particular phenotypes in human studies using the Human Phenotype Ontology (24) and the "Investigate gene sets" feature provided by GSEA (http://www.gsea-msigdb.org/) (table S9).

Prediction of Balto's aesthetic phenotypes

We extracted Balto's genotypes for a panel of 27 genetic variants associated with physical appearance in domestic dogs (table S5) to infer his coat coloration, patterning, and type. We also phased haplotypes from Balto's genotypes using *EAGLE* (v.2.4.1) (51) with reference haplotypes from the phased UMass-Broad Canid Variants and constructed the haplotype consensus sequences of the *MITF*-M promoter length polymorphism locus (chr20: 21,839,331 – 21,839,366) and upstream SINE insertion locus (chr20: 21,836,232 – 21,836,429) using *BCFtools* in order to investigate the *MITF* variants that putatively affect white spotting. We also ran a body size prediction for Balto using a random forest model (*R* packages `caret` and `randomForest`) built on the relative heights (defined as where a dog's shoulders fall relative to an "average person", and surveyed on a Likert scale from ankle-high and shorter, or survey option 0, to hip-high and taller, or survey option 4) of 1,730 modern pet dogs surveyed and 2,797 size-associated SNPs genotyped by the Darwin's Ark project described previously (12) (see supporting files for model and scripts used to run prediction).

Balto's physiological adaptations

We examined the genotypes underlying 14 regions (table S6), which included 1 region under selection in high altitude individuals (53) (Endothelial PAS domain-containing protein

1-*EPAS1*), 2 regions previously identified as under selection in sled dogs (10) (Calcium Voltage-Gated Channel Subunit Alpha1 A - *CACNA1A* and Maltase-Glucoamylase - *MGAM*), 8 regions identified by population branch statistics as potentially under selection in sled dog breeds (12), and 3 regions responsible for aesthetic phenotypes described previously in domestic dogs (Melanocortin 1 Receptor - *MC1R (45)*, Agouti Signaling Protein - *ASIP (52)*, and a chr28 cis-regulatory region associated with single-layered coats (13)). Following the method outlined in Bergström et al. 2020 (21), we also investigated the number of Amylase Alpha 2B (*AMY2B*) copies Balto had by quantifying the ratio of reads (reads/total length of region) mapping to the *AMY2B* regions in CanFam3.1 (ratio: 0.20) to the number of reads mapping to 75 randomly chosen 1kb windows of the genome (ratio: 0.59), given that higher copy numbers are suggested for dog adaptation to starch-rich diets (22).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments:

We thank the Cleveland Museum of Natural History for their contributions to Balto's preservation and history, and the owners of the three working Alaskan sled dogs sequenced for this work (IACUC #2014–0121).

Funding:

NIH grant R01 HG008742 (EKK)

NIH grant U19 AG057377 (EKK)

The Siberian Husky Club of America

Data and materials availability:

Raw sequencing reads for Balto and Alaskan sled dogs have been deposited to the NCBI Sequence Read Archive under BioProject accession PRJNA786530.

References

1. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, Ward LD, Lowe CB, Holloway AK, Clamp M, Gnerre S, Alföldi J, Beal K, Chang J, Clawson H, Cuff J, Di Palma F, Fitzgerald S, Flicek P, Guttman M, Hubisz MJ, Jaffe DB, Jungreis I, Kent WJ, Kostka D, Lara M, Martins AL, Massingham T, Moltke I, Raney BJ, Rasmussen MD, Robinson J, Stark A, Vilella AJ, Wen J, Xie X, Zody MC, Broad Institute Sequencing Platform and Whole Genome Assembly Team, Baldwin J, Bloom T, Chin CW, Heiman D, Nicol R, Nusbaum C, Young S, Wilkinson J, Worley KC, Kovar CL, Muzny DM, Gibbs RA, Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Cree A, Dihn HH, Fowler G, Jhangiani S, Joshi V, Lee S, Lewis LR, Nazareth LV, Okwuonu G, Santibanez J, Warren WC, Mardis ER, Weinstock GM, Wilson RK, Genome Institute at Washington University, Delehaunty K, Dooling D, Fronik C, Fulton L, Fulton B, Graves T, Minx P, Sodergren E, Birney E, Margulies EH, Herrero J, Green ED, Haussler D, Siepel A, Goldman N, Pollard KS, Pedersen JS, Lander ES, Kellis M, A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 478, 476–482 (2011). [PubMed: 21993624]

2. Meadows J, Gazal S, Sullivan P, Zoonomia Consortium, Karlsson EK, Lindblad-Toh K, Leveraging Base Pair Mammalian Constraint to Understand Genetic Variation and Human Disease. Science.

3. Christmas MJ, Kaplow IM, Genereux DP, Dong MX, Hughes GM, Li X, Sullivan PF, Hindle AG, Andrews G, Armstrong JC, Bianchi M, Breit AM, Diekhans M, Fanter C, Foley NM, Goodman L, Keough KC, Kirilenko B, Kowalczyk A, Lawless C, Lind A, Meadows JRS, Moreira L, Ryan L, Swofford R, Valenzuela A, Wagner F, Wallerman O, Damas J, Fan K, Grimshaw J, Johnson J, Kozyrev SV, Lawler AJ, Marinescu VD, Osmanski A, Paulat NS, Phan BN, Reilly SK, Schäffer DE, Steiner C, Supple MA, Wilder AP, Wirthlin ME, Xue JR, Birren BW, Gazal S, Hubley RM, Koepfli K-P, Marques-Bonet T, Meyer W, Nweeia M, Shapiro B, Smit AFA, Springer M, Teeling E, Weng Z, Hiller M, Levesque DL, Lewin H, Murphy WJ, Navarro A, Paten B, Pollard KS, Ray DA, Ruf I, Ryder OA, Pfenning AR, Lindblad-Toh K, Karlsson EK, Evolutionary constraint and innovation across hundreds of placental mammals. Science.

4. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng S, Stiller J, Genereux D, Johnson J, Marinescu VD, Alföldi J, Harris RS, Lindblad-Toh K, Haussler D, Karlsson E, Jarvis ED, Zhang G, Paten B, Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature. 587, 246–251 (2020). [PubMed: 33177663]

5. Zoonomia Consortium A comparative genomics multitool for scientific discovery and conservation. Nature. 587, 240–245 (2020). [PubMed: 33177664]

6. Salisbury G, Salisbury L, The Cruelest Miles: The Heroic Story of Dogs and Men in a Race Against an Epidemic (W. W. Norton & Company, 2003).

7. Sutter NB, Mosher DS, Gray MM, Ostrander EA, Morphometrics within dog breeds are highly reproducible and dispute Rensch's rule. Mamm. Genome. 19, 713–723 (2008). [PubMed: 19020935]

8. American Kennel Club, The Complete Dog Book: 20th Edition (Random House Publishing Group, 2007).

9. Huson HJ, Parker HG, Runstadler J, Ostrander EA, A genetic dissection of breed composition and performance enhancement in the Alaskan sled dog. BMC Genet. 11, 71 (2010). [PubMed: 20649949]

10. Sinding M-HS, Gopalakrishnan S, Ramos-Madrigal J, de Manuel M, Pitulko VV, Kuderna L, Feuerborn TR, Frantz LAF, Vieira FG, Niemann J, Samaniego Castruita JA, Carøe C, Andersen-Ranberg EU, Jordan PD, Pavlova EY, Nikolskiy PA, Kasparov AK, Ivanova VV, Willerslev E, Skoglund P, Fredholm M, Wennerberg SE, Heide-Jørgensen MP, Dietz R, Sonne C, Meldgaard M, Dalén L, Larson G, Petersen B, Sicheritz-Pontén T, Bachmann L, Wiig Ø, Marques-Bonet T, Hansen AJ, Gilbert MTP, Arctic-adapted dogs emerged at the Pleistocene–Holocene transition. Science. 368, 1495–1499 (2020). [PubMed: 32587022]

11. Shindyapina AV, Zenin AA, Tarkhov AE, Santesmasses D, Fedichev PO, Gladyshev VN, Germline burden of rare damaging variants negatively affects human healthspan and lifespan. Elife. 9 (2020), doi:10.7554/eLife.53449.

12. Morrill K, Hekman J, Li X, McClure J, Logan B, Goodman L, Gao M, Dong Y, Alonso M, Carmichael E, Snyder-Mackler N, Alonso J, Noh HJ, Johnson J, Koltookian M, Lieu C, Megquier K, Swofford R, Turner-Maier J, White ME, Weng Z, Colubri A, Genereux DP, Lord KA, Karlsson EK, Ancestry-inclusive dog genomics challenges popular breed stereotypes. Science (2022).

13. Whitaker DT, Ostrander EA, Hair of the Dog: Identification of a Cis-Regulatory Module Predicted to Influence Canine Coat Composition. Genes. 10 (2019), doi:10.3390/genes10050323.

14. Karlsson EK, Baranowska I, Wade CM, Salmon Hillbertz NHC, Zody MC, Anderson N, Biagi TM, Patterson N, Pielberg GR, Kulbokas EJ 3rd, Comstock KE, Keller ET, Mesirov JP, von Euler H, Kämpe O, Hedhammar A, Lander ES, Andersson G, Andersson L, Lindblad-Toh K, Efficient mapping of mendelian traits in dogs through genome-wide association. Nat. Genet. 39, 1321–1328 (2007). [PubMed: 17906626]

15. Dreger DL, Parker HG, Ostrander EA, Schmutz SM, Identification of a mutation that is associated with the saddle tan and black-and-tan phenotypes in Basset Hounds and Pembroke Welsh Corgis. J. Hered. 104, 399–406 (2013). [PubMed: 23519866]

16. Deane-Coe PE, Chu ET, Slavney A, Boyko AR, Sams AJ, Direct-to-consumer DNA testing of 6,000 dogs reveals 98.6-kb duplication associated with blue eyes and heterochromia in Siberian Huskies. PLoS Genet. 14, e1007648 (2018). [PubMed: 30286082]

17. Schmutz SM, Berryere TG, Ellinwood NM, Kerns JA, Barsh GS, MC1R studies in dogs with melanistic mask or brindle patterns. J. Hered. 94, 69–73 (2003). [PubMed: 12692165]

18. Slavney AJ, Kawakami T, Jensen MK, Nelson TC, Sams AJ, Boyko AR, Five genetic variants explain over 70% of hair coat pheomelanin intensity variation in purebred and mixed breed domestic dogs. PLoS One. 16, e0250579 (2021). [PubMed: 34043658]
19. Anderson H, Honkanen L, Ruotanen P, Mathlin J, Donner J, Comprehensive genetic testing combined with citizen science reveals a recently characterized ancient MC1R mutation associated with partial recessive red phenotypes in dog. Canine Med Genet. 7, 16 (2020). [PubMed: 33292722]
20. Axelsson E, Ratnakumar A, Arendt M-L, Maqbool K, Webster MT, Perloski M, Liberg O, Arnemo JM, Hedhammar A, Lindblad-Toh K, The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature. 495, 360–364 (2013). [PubMed: 23354050]
21. Bergström A, Frantz L, Schmidt R, Ersmark E, Lebrasseur O, Girdland-Flink L, Lin AT, Storå J, Sjögren K-G, Anthony D, Antipina E, Amiri S, Bar-Oz G, Bazaliiskii VI, Bulatović J, Brown D, Carmagnini A, Davy T, Fedorov S, Fiore I, Fulton D, Germonpré M, Haile J, Irving-Pease EK, Jamieson A, Janssens L, Kirillova I, Horwitz LK, Kuzmanovic-Cvetković J, Kuzmin Y, Losey RJ, Dizdar DL, Mashkour M, Novak M, Onar V, Orton D, Pasarić M, Radivojević M, Rajković D, Roberts B, Ryan H, Sablin M, Shidlovskiy F, Stojanović I, Tagliacozzo A, Trantalidou K, Ullén I, Villaluenga A, Wapnish P, Dobney K, Götherström A, Linderholm A, Dalén L, Pinhasi R, Larson G, Skoglund P, Origins and genetic legacy of prehistoric dogs. Science. 370, 557–564 (2020). [PubMed: 33122379]
22. Arendt M, Fall T, Lindblad-Toh K, Axelsson E, Amylase activity is associated with *AMY 2B* copy numbers in dog: implications for dog domestication, diet and diabetes. Animal Genetics. 45 (2014), pp. 716–722. [PubMed: 24975239]
23. Gou X, Wang Z, Li N, Qiu F, Xu Z, Yan D, Yang S, Jia J, Kong X, Wei Z, Lu S, Lian L, Wu C, Wang X, Li G, Ma T, Jiang Q, Zhao X, Yang J, Liu B, Wei D, Li H, Yang J, Yan Y, Zhao G, Dong X, Li M, Deng W, Leng J, Wei C, Wang C, Mao H, Zhang H, Ding G, Li Y, Whole-genome sequencing of six dog breeds from continuous altitudes reveals adaptation to high-altitude hypoxia. Genome Res. 24, 1308–1315 (2014). [PubMed: 24721644]
24. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, Danis D, Balagura G, Baynam G, Brower AM, Callahan TJ, Chute CG, Est JL, Galer PD, Ganesan S, Griese M, Haimel M, Pazmandi J, Hanauer M, Harris NL, Hartnett MJ, Hastreiter M, Hauck F, He Y, Jeske T, Kearney H, Kindle G, Klein C, Knoflach K, Krause R, Lagorce D, McMurry JA, Miller JA, Munoz-Torres MC, Peters RL, Rapp CK, Rath AM, Rind SA, Rosenberg AZ, Segal MM, Seidel MG, Smedley D, Talmy T, Thomas Y, Wiafe SA, Xian J, Yüksel Z, Helbig I, Mungall CJ, Haendel MA, Robinson PN, The Human Phenotype Ontology in 2021. Nucleic Acids Res. 49, D1207–D1217 (2021). [PubMed: 33264411]
25. Thomas B, Thomas P, Leonhard Seppala: The Siberian Dog and the Golden Age of Sleddog Racing 1908–1941 (Pictorial Histories Publishing Company, Incorporated, 2015).
26. The Cleveland Museum of Natural History, Balto FAQs, (available at https://www.cmnh.org/science-news/blog/march-2020/balto-faq).
27. Sled dog central: The Inuit sled dog by sue Hamilton, (available at http://www.sleddogcentral.com/inuit.htm).
28. Dabney J, Knapp M, Glocke I, Gansauge M-T, Weihmann A, Nickel B, Valdiosera C, García N, Pääbo S, Arsuaga J-L, Meyer M, Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc. Natl. Acad. Sci. U. S. A. 110, 15758–15763 (2013). [PubMed: 24019490]
29. Kapp JD, Green RE, Shapiro B, A Fast and Efficient Single-stranded Genomic Library Preparation Method Optimized for Ancient DNA. J. Hered. 112, 241–249 (2021). [PubMed: 33768239]
30. John JS, SeqPrep: tool for stripping adaptors and/or merging paired reads with overlap into single reads. URL: https://githubcom/jstjohn/SeqPrep (2011).
31. Li H, Durbin R, Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 26, 589–595 (2010). [PubMed: 20080505]
32. Plassais J, Kim J, Davis BW, Karyadi DM, Hogan AN, Harris AC, Decker B, Parker HG, Ostrander EA, Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. Nat. Commun. 10, 1489 (2019). [PubMed: 30940804]

33. Kumar S, Stecher G, Peterson D, Tamura K, MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. Bioinformatics. 28, 2685–2686 (2012). [PubMed: 22923298]
34. Hayward JJ, Castelhano MG, Oliveira KC, Corey E, Balkman C, Baxter TL, Casal ML, Center SA, Fang M, Garrison SJ, Kalla SE, Korniliev P, Kotlikoff MI, Moise NS, Shannon LM, Simpson KW, Sutter NB, Todhunter RJ, Boyko AR, Complex disease and phenotype mapping in the domestic dog. Nat. Commun. 7, 10460 (2016). [PubMed: 26795439]
35. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC, PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007). [PubMed: 17701901]
36. Weir BS, Cockerham CC, ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE. Evolution. 38, 1358–1370 (1984). [PubMed: 28563791]
37. Bhatia G, Patterson N, Sankararaman S, Price AL, Estimating and interpreting FST: the impact of rare variants. Genome Res. 23, 1514–1521 (2013). [PubMed: 23861382]
38. Alexander DH, Lange K, Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics. 12, 246 (2011). [PubMed: 21682921]
39. Foote AD, Hooper R, Alexander A, Baird RW, Baker CS, Ballance L, Barlow J, Brownlow A, Collins T, Constantine R, Dalla Rosa L, Davison NJ, Durban JW, Esteban R, Excoffier L, Martin SLF, Forney KA, Gerrodette T, Gilbert MTP, Guinet C, Hanson MB, Li S, Martin MD, Robertson KM, Samarra FIP, de Stephanis R, Tavares SB, Tixier P, Totterdell JA, Wade P, Wolf JBW, Fan G, Zhang Y, Morin PA, Runs of homozygosity in killer whale genomes provide a global record of demographic histories. Mol. Ecol. (2021), doi:10.1111/mec.16137.
40. Zhao G, Zhang T, Liu Y, Wang Z, Xu L, Zhu B, Gao X, Zhang L, Gao H, Liu GE, Li J, Xu L, Genome-Wide Assessment of Runs of Homozygosity in Chinese Wagyu Beef Cattle. Animals (Basel). 10 (2020), doi:10.3390/ani10081425.
41. McQuillan R, Leutenegger A-L, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, Smolej-Narancic N, Janicijevic B, Polasek O, Tenesa A, Macleod AK, Farrington SM, Rudan P, Hayward C, Vitart V, Rudan I, Wild SH, Dunlop MG, Wright AF, Campbell H, Wilson JF, Runs of homozygosity in European populations. Am. J. Hum. Genet. 83, 359–372 (2008). [PubMed: 18760389]
42. Cingolani P, snpEff: Variant effect prediction (2012).
43. Rezwani M, Pourfathollah AA, Noorbakhsh F, rbioapi: User-Friendly R Interface to Biologic Web Services' API. Bioinformatics (2022), doi:10.1093/bioinformatics/btac172.
44. Mi H, Ebert D, Muruganujan A, Mills C, Albou L-P, Mushayamaha T, Thomas PD, PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res. 49, D394–D403 (2021). [PubMed: 33290554]
45. Schmutz SM, Berryere TG, Goldfinch AD, TYRP1 and MC1R genotypes and their effects on coat color in dogs. Mamm. Genome. 13, 380–387 (2002). [PubMed: 12140685]
46. Cargill EJ, Famula TR, Schnabel RD, Strain GM, Murphy KE, The color of a Dalmatian's spots: linkage evidence to support the TYRP1 gene. BMC Vet. Res. 1, 1 (2005). [PubMed: 16045797]
47. Kiener S, Kehl A, Loechel R, Langbein-Detsch I, Müller E, Bannasch D, Jagannathan V, Leeb T, Novel Brown Coat Color (Cocoa) in French Bulldogs Results from a Nonsense Variant in HPS3. Genes. 11 (2020), doi:10.3390/genes11060636.
48. Kerns JA, Newton J, Berryere TG, Rubin EM, Cheng J-F, Schmutz SM, Barsh GS, Characterization of the dog Agouti gene and a nonagouti mutation in German Shepherd Dogs. Mamm. Genome. 15, 798–808 (2004). [PubMed: 15520882]
49. Kerns JA, Cargill EJ, Clark LA, Candille SI, Berryere TG, Olivier M, Lust G, Todhunter RJ, Schmutz SM, Murphy KE, Barsh GS, Linkage and segregation analysis of black and brindle coat color in domestic dogs. Genetics. 176, 1679–1689 (2007). [PubMed: 17483404]
50. Monteagudo LV, Tejedor MT, The b(c) allele of TYRP1 is causative for the recessive brown (liver) colour in German Shepherd dogs. Anim. Genet. 46, 588–589 (2015). [PubMed: 26370740]
51. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, Schoenherr S, Forer L, McCarthy S, Abecasis GR, Durbin R, L Price A, Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. 48, 1443–1448 (2016). [PubMed: 27694958]

52. Berryere TG, Kerns JA, Barsh GS, Schmutz SM, Association of an Agouti allele with fawn or sable coat color in domestic dogs. Mamm. Genome. 16, 262–272 (2005). [PubMed: 15965787]

53. vonHoldt B, Fan Z, Ortega-Del Vecchyo D, Wayne RK, EPAS1 variants in high altitude Tibetan wolves were selectively introgressed into highland dogs. PeerJ. 5, e3522 (2017). [PubMed: 28717592]

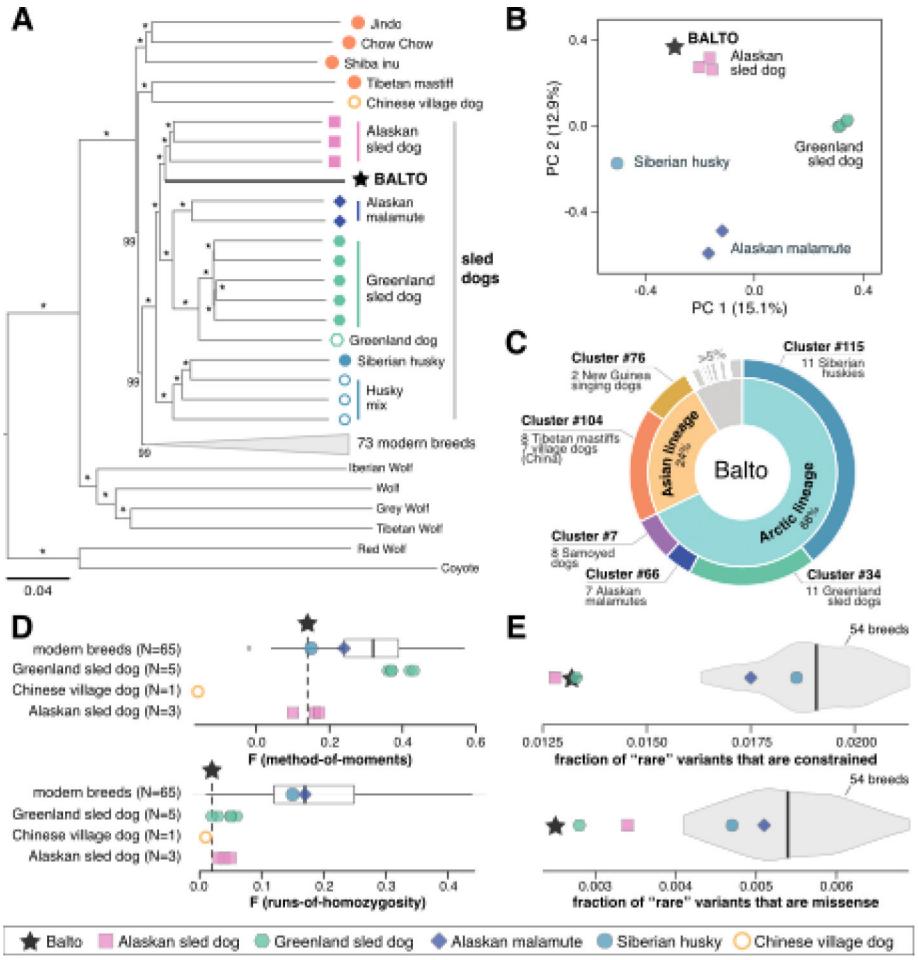

Figure 1. Balto clusters most closely with Alaskan sled dogs, but had high genetic diversity and a lower burden of potentially damaging variants.

(**A**) Neighbor-joining tree clusters Balto (★) most closely with the outbred, working population of Alaskan sled dogs, and a part of a clade of sled dog populations. (**B**) Similarly, principal component analysis puts Balto near, but not in, a cluster of Alaskan sled dogs. (**C**) Unsupervised admixture analysis of Balto alongside the Alaskan sled dogs and other dogs and canids (K= 116 putative populations and N= 2166 individuals) infers substantial ancestral similarity to Siberian huskies, Greenland sled dogs, and outbred dogs from Asia (table S2). The remainder of his ancestry (8%) matches poorly (<5%) to any other clusters. Balto and working sled dogs (**D**) had lower levels of inbreeding, and (**E**) carried fewer constrained (p_{wilcox}=0.0019) and missense (p_{wilcox}= 0.0023) rare variants than modern dog breeds (table S10).

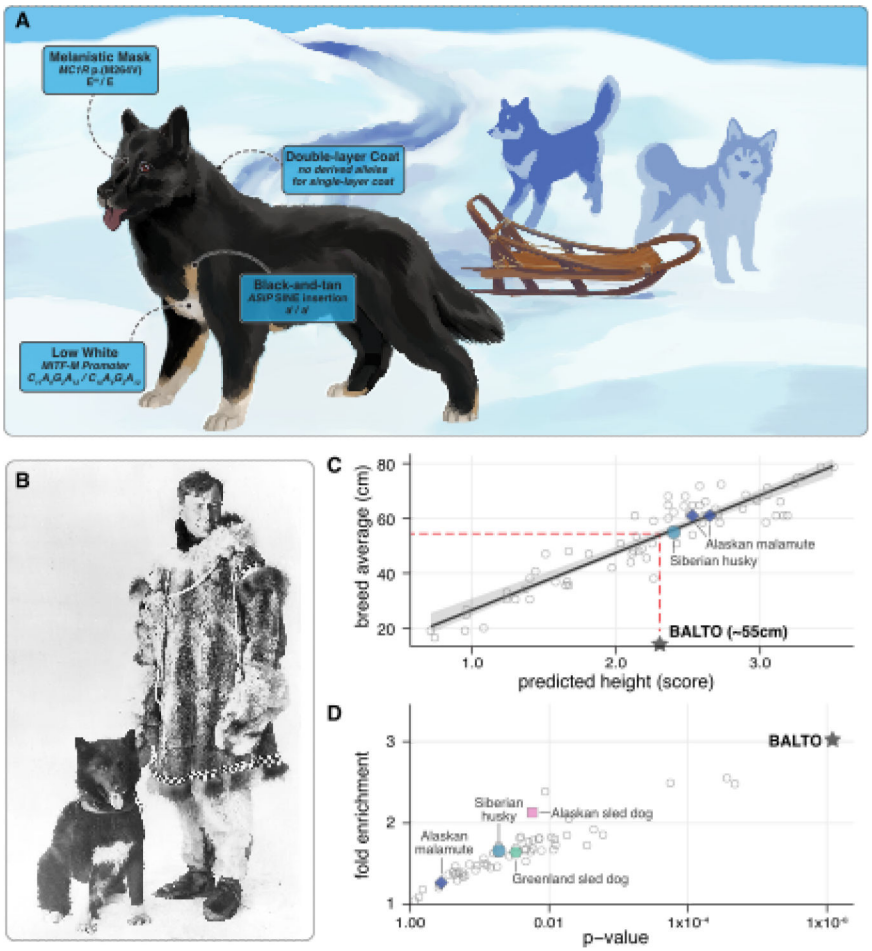

Figure 2. Genomic recreation of Balto's physical appearance.

(**A**) Prediction of Balto's coat features based on his genome sequence with details on each trait and genotype in blue boxes. (**B**) A photo of Balto with musher Gunnar Kaasen. From the photo and his taxidermied remains, Balto was a black dog with dark eyes and some white patches on his chest and legs. He had a double-layered coat, and stood just under knee-high relative to Kaasen. Photo credit: Cleveland Museum of Natural History. (**C**) Using a random forest model based on 1,730 dogs and 2,797 height-associated genetic variants (12), we predicted that Balto would stand around 55 cm tall (value: 2.3) at his withers, close to the average height for the Siberian husky breed. Circles show dogs from other breeds. (**D**) Gene set enrichment testing of genes with common and constrained missense variants in 57 different dog populations shows a significant enrichment (p_{FDR}=0.013) in the GO Tissue Development pathway only for Balto's population.