



**HAL**  
open science

## Genome sequencing of 2000 canids by the Dog10K consortium advances the understanding of demography, genome function and architecture

Jennifer R S Meadows, Jeffrey M Kidd, Guo-Dong Wang, Heidi G Parker, Peter Z Schall, Matteo Bianchi, Matthew J Christmas, Katia Bougiouri, Reuben M Buckley, Christophe Hitte, et al.

### ► To cite this version:

Jennifer R S Meadows, Jeffrey M Kidd, Guo-Dong Wang, Heidi G Parker, Peter Z Schall, et al.. Genome sequencing of 2000 canids by the Dog10K consortium advances the understanding of demography, genome function and architecture. *Genome Biology*, 2023, 24 (1), pp.187. 10.1186/s13059-023-03023-7. hal-04197751v1

**HAL Id: hal-04197751**

**<https://univ-rennes.hal.science/hal-04197751v1>**

Submitted on 6 Sep 2023 (v1), last revised 14 Nov 2023 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

RESEARCH

Open Access



# Genome sequencing of 2000 canids by the Dog10K consortium advances the understanding of demography, genome function and architecture

Jennifer R. S. Meadows<sup>1\*†</sup>, Jeffrey M. Kidd<sup>2\*†</sup>, Guo-Dong Wang<sup>3</sup>, Heidi G. Parker<sup>4</sup>, Peter Z. Schall<sup>2</sup>, Matteo Bianchi<sup>1</sup>, Matthew J. Christmas<sup>1</sup>, Katia Bougiouri<sup>5</sup>, Reuben M. Buckley<sup>4</sup>, Christophe Hitte<sup>6</sup>, Anthony K. Nguyen<sup>2</sup>, Chao Wang<sup>1</sup>, Vidhya Jagannathan<sup>7</sup>, Julia E. Niskanen<sup>8</sup>, Laurent A. F. Frantz<sup>9</sup>, Meharji Arumilli<sup>8</sup>, Sruthi Hundi<sup>8</sup>, Kerstin Lindblad-Toh<sup>1,10</sup>, Catarina Ginja<sup>11</sup>, Kadek Karang Agustina<sup>12</sup>, Catherine André<sup>6</sup>, Adam R. Boyko<sup>13</sup>, Brian W. Davis<sup>14</sup>, Michaela Drögemüller<sup>7</sup>, Xin-Yao Feng<sup>3</sup>, Konstantinos Gkagkavouzis<sup>15</sup>, Giorgos Iliopoulos<sup>16</sup>, Alexander C. Harris<sup>4</sup>, Marjo K. Hytönen<sup>8</sup>, Daniela C. Kalthoff<sup>16</sup>, Yan-Hu Liu<sup>3</sup>, Petros Lymberakis<sup>17,18,19</sup>, Nikolaos Poulakakis<sup>17,18,19</sup>, Ana Elisabete Pires<sup>11</sup>, Fernando Racimo<sup>5</sup>, Fabian Ramos-Almodovar<sup>2</sup>, Peter Savolainen<sup>20</sup>, Semina Venetsani<sup>21</sup>, Imke Tammen<sup>22</sup>, Alexandros Triantafyllidis<sup>15</sup>, Bridgett vonHoldt<sup>23</sup>, Robert K. Wayne<sup>24</sup>, Greger Larson<sup>25</sup>, Frank W. Nicholas<sup>22</sup>, Hannes Lohi<sup>8</sup>, Tosso Leeb<sup>7</sup>, Ya-Ping Zhang<sup>3†</sup> and Elaine A. Ostrander<sup>4\*†</sup>

<sup>†</sup>Jennifer R. S. Meadows and Jeffrey M. Kidd contributed equally to this work.

<sup>†</sup>Ya-Ping Zhang and Elaine A. Ostrander co-senior authors.

\*Correspondence: jennifer.meadows@imb.uu.se; jmkidd@umich.edu; eostrand@mail.nih.gov

<sup>1</sup> Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, 75132 Uppsala, Sweden

<sup>2</sup> Department of Human Genetics, University of Michigan, Ann Arbor, MI 48107, USA

<sup>4</sup> National Human Genome Research Institute, National Institutes of Health, 50 South Drive, Building 50 Room 5351, Bethesda, MD 20892, USA

Full list of author information is available at the end of the article

## Abstract

**Background:** The international Dog10K project aims to sequence and analyze several thousand canine genomes. Incorporating 20× data from 1987 individuals, including 1611 dogs (321 breeds), 309 village dogs, 63 wolves, and four coyotes, we identify genomic variation across the canid family, setting the stage for detailed studies of domestication, behavior, morphology, disease susceptibility, and genome architecture and function.

**Results:** We report the analysis of > 48 M single-nucleotide, indel, and structural variants spanning the autosomes, X chromosome, and mitochondria. We discover more than 75% of variation for 239 sampled breeds. Allele sharing analysis indicates that 94.9% of breeds form monophyletic clusters and 25 major clades. German Shepherd Dogs and related breeds show the highest allele sharing with independent breeds from multiple clades. On average, each breed dog differs from the UU\_Cfam\_GSD\_1.0 reference at 26,960 deletions and 14,034 insertions greater than 50 bp, with wolves having 14% more variants. Discovered variants include retrogene insertions from 926 parent genes. To aid functional prioritization, single-nucleotide variants were annotated with SnpEff and Zoonomia phyloP constraint scores. Constrained



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

positions were negatively correlated with allele frequency. Finally, the utility of the Dog10K data as an imputation reference panel is assessed, generating high-confidence calls across varied genotyping platform densities including for breeds not included in the Dog10K collection.

**Conclusions:** We have developed a dense dataset of 1987 sequenced canids that reveals patterns of allele sharing, identifies likely functional variants, informs breed structure, and enables accurate imputation. Dog10K data are publicly available.

**Keywords:** Canine, Dog, Genomics, Variation, Demographic history, Mitochondrial DNA, Genetic diversity

## Background

Recent advances in comparative genomics have enhanced the utility of the domestic dog and other canines for studies of mammalian biology, disease, and domestication. The initial dog reference genome, derived from a single boxer, was released in 2004 [1] and has since been augmented with reference patches [2, 3], variation catalogs (e.g., [4–6]), and functional annotations (e.g., [2, 3, 7]). The resulting data has been important for identifying genes and variants controlling simple Mendelian traits (e.g., [4, 8, 9]), tracing migration of human populations [10–12], building a vocabulary for mammalian behavior [13, 14], and enabling studies of both aging [15, 16] and disease susceptibility [17].

Many complex genetic questions remain, and answering them has been limited by the reliance of both reference and test datasets comprised of dogs of largely western European descent, incomplete catalogs of copy number variants [5], and the exclusion of village and feral dogs and other canid species from most datasets [18]. Exome-based sequencing approaches have made useful contributions, but have been limited by dataset size [19]. Also, while studies of ancient canids have revealed key events in canine history (e.g., [11, 20–24]), this research relies on high-quality reference genomes supported by sequence variation from large numbers of wild and domestic canids. At present, these resources are insufficient. In response to this demand, a group of canine geneticists and biologists joined forces in 2016 to initiate Dog10K, a worldwide consortium with a goal of producing and analyzing DNA sequences from 10,000 canids [25].

Since 2004, several hundred canine genomes have been partially or fully sequenced by individual groups or laboratories, most with the aim of amassing markers for genome-wide association studies (GWAS) and subsequent fine mapping and functional studies, or for inferring canine history. As the diversity, density, and quality of available sequences have improved, so too has the resolution for identifying putative functional variants, although these studies have not kept pace with the larger field of mammalian biology [4, 26, 27]. The publication of new, high-quality, long-read assemblies of the Basenji [28], Great Dane [29], German Shepherd Dog [30, 31], Labrador Retriever [32], a revised version of the original Boxer [33], dingo [34], and gray wolf [35] have aided the community's effort to address historical topics of interest and permit the analysis of previously inaccessible genomic features such as gene promoters, regulatory elements, repeated sequences, and mobile elements [29, 30].

Although phase-resolved canine assemblies are not currently available, the continued development of long-read assemblies will enable future analyses of variation using a pangenome approach [36]. In this study, we discover and characterize canine variation through alignment of Illumina sequencing reads to the recently published assembly of Mischka, a German Shepherd Dog (UU\_Cfam\_GSD\_1.0) [30].

The Dog10K dataset includes samples from 321 dog breeds, with 261 breeds represented by three or more individuals, containing a worldwide distribution of rare and common breeds, collectively spanning variation in morphology, disease susceptibility, and behavior. Our dataset uniquely possesses a worldwide sampling of village dogs and niche populations, both of which fall outside the umbrella of pure or mixed breed dogs. The inclusion of 1929 individuals makes the Dog10K reference panel the largest to date, allowing for the imputation of canine genotypes across diverse breeds and genotyping platforms, including low-pass sequencing data. Finally, the inclusion of wild canid populations, including wolves and coyotes, completes the most comprehensive and inclusive dataset of canines assembled, allowing us to perform detailed analyses of genome architecture.

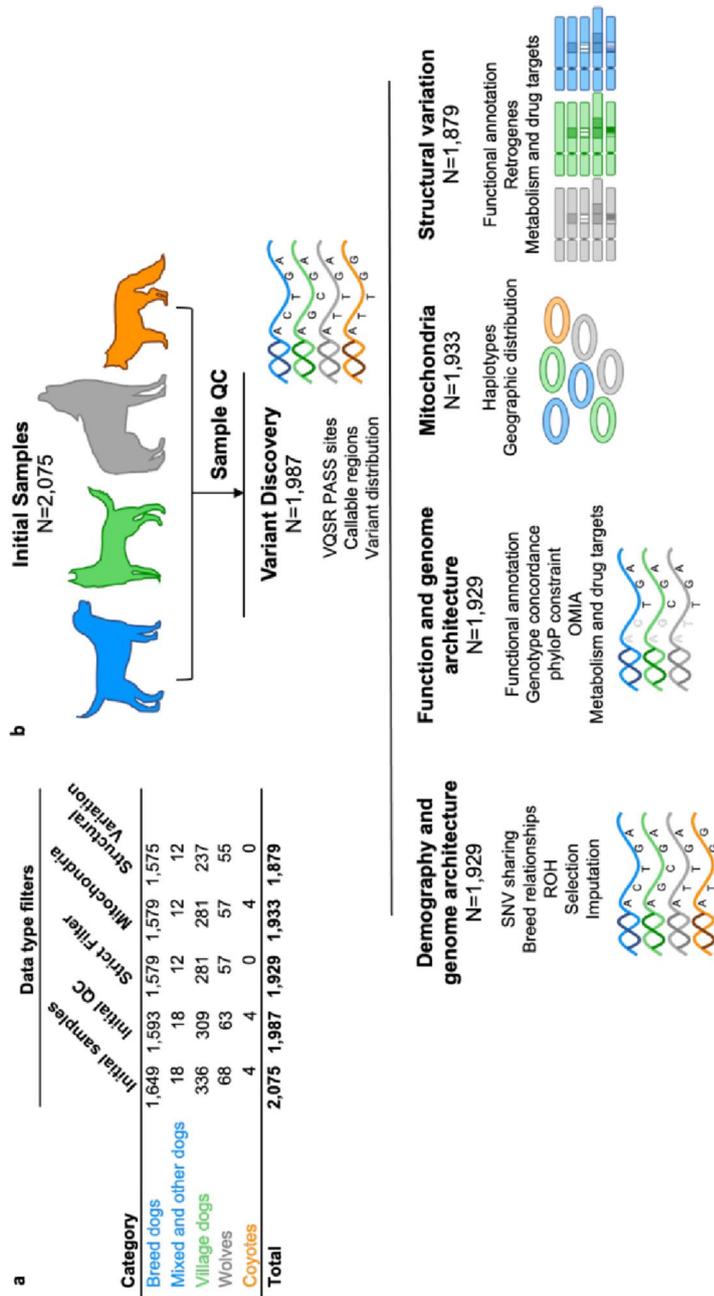
In the analysis herein, we present Illumina sequencing data from 1987 canids, with joint calling across the mitochondrial and nuclear genomes revealing over 144,000 structural variants (deletions, insertions, duplications, and inversions  $\geq 50$  bp in size), 14.4 million indels, and 34 million single-nucleotide variants (SNVs), the most extensive variant catalog produced in canines to date. Clade analysis with the nuclear SNV dataset reveals both expected and new relationships among breed dogs sampled. The sequencing of > 330 village dogs and wolves demonstrates a wealth of variation previously undiscovered in breed dogs, with almost one third of all observed variation exclusive to these two groups. Analysis of mitochondrial data reveals surprisingly few haplotypes in dogs, with greater observed variation in wild canids.

## Results

### Sample selection and data harmonization

The 2075 samples collected for Dog10K were selected to represent a wide variety of breeds of differing morphology, history, and behavior (1649 samples); dogs that represent local niche populations or breeds that are not nationally registered (18 samples); village dogs from multiple locations (336 samples); and wild canids (68 wolves, 4 coyotes) (Fig. 1, Additional file 1: Table S1). At the time of collection, breed samples were free from known disease, and efforts were made to balance sex across all populations (52.6% female).

A reference genome consisting of the German Shepherd Dog genome assembly [30] (UU\_Cfam\_GSD\_1.0, GCF\_011100685.1), supplemented by three Y chromosome contigs from a Labrador Retriever (ROS\_Cfam\_1.0, GCF\_014441545.1), was used as the foundation for all analyses. A pipeline based on bwa-mem2 and GATK best practices was used for the uniform sequence alignment and processing across four centers (Additional file 2: Sect. 1) [37–39]. Variant calling (mitochondrial genome: SNVs and indels; nuclear genome: SNVs, indels, and SVs) and quality filtering were performed across the entire sample set. Sample and variant filters were used to generate different datasets for addressing specific questions (Fig. 1).



**Fig. 1** Overview of the Dog10K collection. **a** Sample collection and sample filtering for **(b)** the varied demographic, genome function, and architecture examined in the program. QC, quality control; ROH, runs of homozygosity; OMIA, Online Mendelian Inheritance in Animals

In brief, primary SNV and small indel variant discovery was performed on 1987 samples fulfilling initial quality thresholds (Additional file 2: Sect. 2). Of these, 1929 samples passed additional quality control thresholds and were used in most SNV analyses. In these steps, variants were selected using either VQSR PASS criteria or additional strict variant filters (Additional file 2: Sect. 9). For SV analyses, 1824 samples were available after additional quality controls (Additional file 2: Sect. 7).

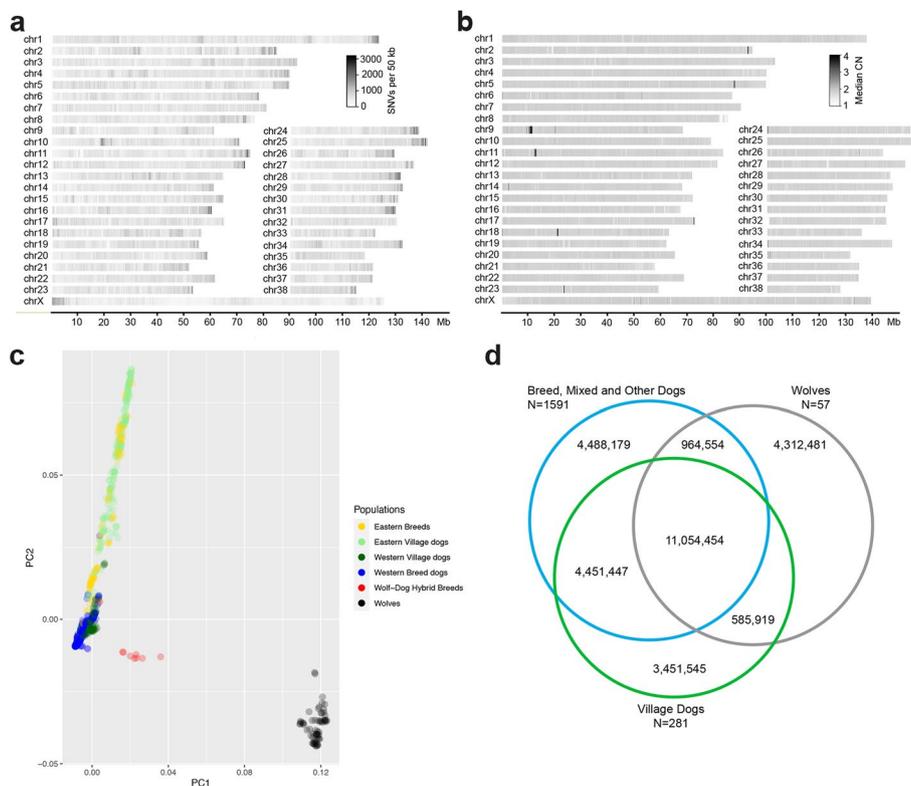
### Genome-wide pattern of sequence variation in canines

Our initial variant callset, derived from 1987 dogs, wolves, and coyotes, contains 33,374,690 SNVs across the autosomes and pseudoautosomal region of the X chromosome (X-PAR), and 1,191,860 SNVs on the non-homologous portion of the X chromosome. Using hard filters, we identify a total of 14,414,501 indel and mixed variants. Subsequent analyses are focused on SNVs due to the paucity of validated canine indels available to train refined filters. Based on read depth and mapping quality profiles, we developed a “callable” region annotation and estimated that 96% of the assembly is amenable to short-read variant calling. SNVs show an uneven distribution across chromosomes with a 65% increase in SNV density observed near chromosome ends ( $p < 1 \times 10^{-30}$ , Welch’s unequal variances *t*-test) and a moderate correlation with GC content as measured in 50-kb windows (Pearson’s  $r = 0.37$ ) (Fig. 2a).

We subsequently performed a deep analysis of variation in 1929 dog and wolf samples that passed more stringent quality filters (Fig. 1, Additional file 2: Sects. 3 and 9). These samples include 321 breeds, with 261 breeds represented by three or more individuals, 281 village dog samples from 26 different countries, and 57 wolf samples from across Eurasia. Principal component (PC) analysis of SNV genotypes reflects the ancestry of dog and wolf samples (Fig. 2c). The first component accounts for 4.1% of total variation and separates wolves and dogs. PC2 (1.7% of variation) stratifies village dogs and breed dogs based on their origin, with Eastern Eurasian breeds and village dogs at one end of the continuum, and Western Eurasian samples at the other. Samples from the Saarloos Wolfdog and Czechoslovakian Wolfdog breeds, both of which have recent wolf ancestry, show an intermediate placement along PC1. The sole Shiloh Shepherd in the dataset, a breed which may have partial Czechoslovakian Wolfdog ancestry [40], is placed among other Western Eurasian breeds.

### Variation among and within sample groupings

Direct ascertainment of variation from whole genome sequencing permits an assessment of shared genetic variation among breed dogs (including mixed and other breeds), village dogs, and wolves. We first assessed the level of allele sharing among the 1929 analyzed samples (Fig. 2d). The alternative allele was detected in all three sample categories at 37.7% of the 29,308,579 biallelic autosomal SNV sites. Despite making up only 3% of the analyzed samples (57/1929), 14.7% of the variants were present only in sampled wolves, while 15.2% of total sites are absent from wolves. This may be a reflection of the small number of wolves in the study. A total of 11.8% of the variants are found only in village dogs, which represent 14.6% of the samples (281/1929). The combined breed and mixed/other samples represent 82.4% of the total samples (1591/1929), yet only 15.3% of variants are private to this group.



**Fig. 2** Variant distribution across the genome. **a** SNV density in 50-kb windows, drawn from 1987 samples. Increased SNV density is observed at the X-PAR region and the ends of most autosomes. **b** Median copy number (CN) for 1824 dogs reveals a large, common duplication on chr9 relative to the reference genome. **c** Principal component (PC) analysis separates dog and wolf samples along the first axis while axis two separates dogs from Eastern and Western Eurasia. **d** SNV sharing between the three categories of samples

Since rare variants may be informative for inferring recent genetic relationships, we examined variants that were found in only two individuals, i.e.,  $F_2$  sites [41, 42]. We identified 2,384,354 autosomal SNVs which are found in exactly two of the 1929 samples. Most  $F_2$  sharing was found within groups: 87% of wolf  $F_2$  sites are shared with another wolf, while 69% of  $F_2$  sites in breed dogs or village dogs were shared with another breed or village dog, respectively. Reflective of recent shared ancestry, we identified 10 breed dogs who share  $\geq 20\%$  of their  $F_2$  sites with at least one wolf (Additional file 1: Table S2). As expected, this includes the Saarloos Wolfdog and Czechoslovakian Wolfdog breeds [43, 44]. The sole Shiloh Shepherd shares 78% of  $F_2$  sites with wolves; however, we note that D-statistic analyses do not detect significant allele sharing between the Shiloh Shepherd and wolves relative to that observed in German Shepherd Dogs (Additional file 2: Sect. 3).

To guide future sequencing studies, we estimated the fraction of total variation found in the diverse breeds sampled. Using the observed distribution of non-reference allele counts observed in each breed, we estimated the total number of SNVs expected in a hypothetical set of 100 individuals [45], and compared this value to the total already found in the existing Dog10K call set. Not surprisingly, the predicted fraction of discovered variation varies widely among the 261 breeds represented by at least three

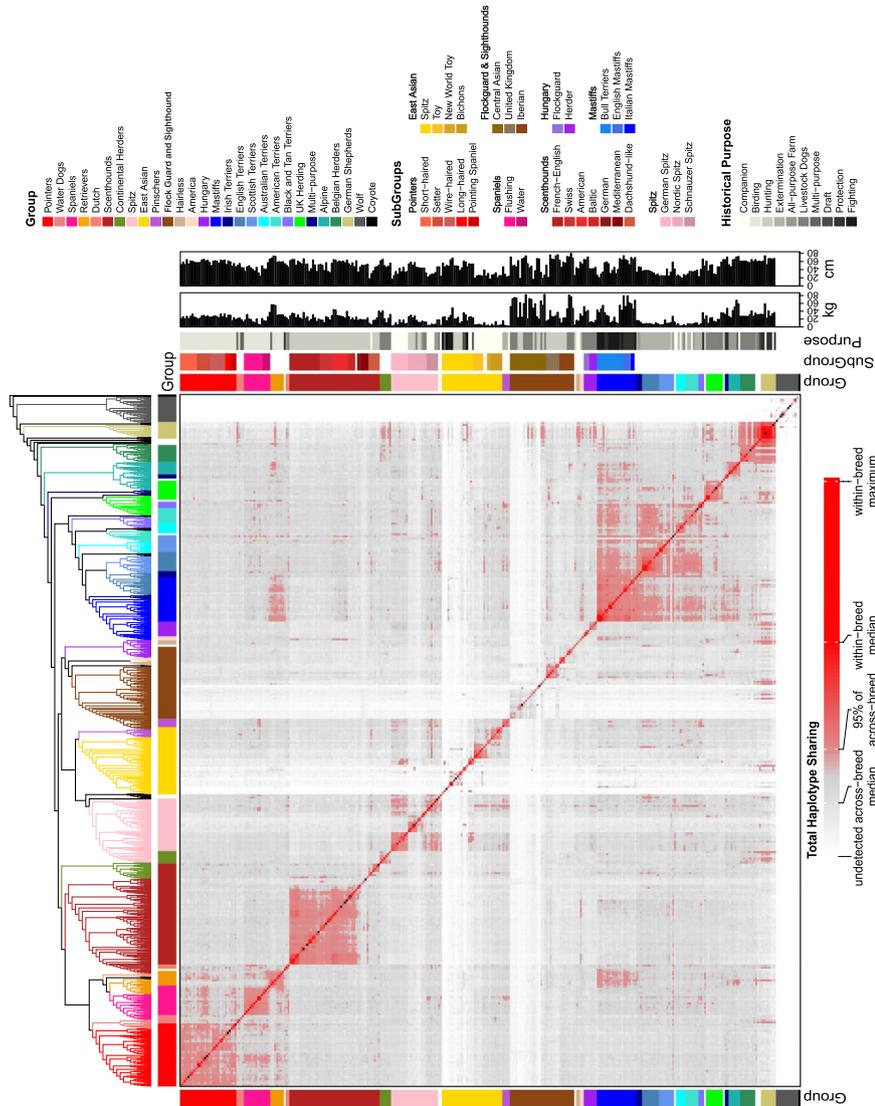
individuals (mean = 82.8%, range 47.1–98.4%), and is weakly correlated with the number of individuals sampled per breed (Pearson's  $r=0.29$ ). For 22 breeds, we determine that >90% of the total predicted variants have been identified, while for 20 breeds,  $\leq 75\%$  of the total variation has been discovered (Additional file 1: Table S3). For instance, we estimate that the five Norwegian Lundehunds sequenced here capture 98.4% of variation that would likely be discovered in 100 individuals from the same breed. This reflects their well-established closed breeding population structure that was derived from 5–6 individuals [46–48]. In contrast, four Czechoslovakian wolfdog samples capture only 47% of the variation that would be captured in a sample of 100 such dogs. These estimates assume that the sampled individuals are representative of the breed as a whole, and so may be biased if there is within-breed population structure. As these calculations do not account for variation shared between breeds, these predictions represent the lower bounds of the total fraction of variation for each breed already captured in the Dog10K collection.

### Breed relationships and haplotype sharing

We used the Dog10K SNVs to assess the relationships among the sampled breeds. We combined breed subtypes and varieties, resulting in a dataset of 292 breeds represented by more than one dog (Additional file 1: Table S4, Additional file 2: Sect. 3). The output cladogram is based on genomic distance and assessed through 100 resampled data sets. We defined clades as clusters of two or more breeds that share the same branch in >65% of samplings. We found that 277 of 292 breeds (94.9%) formed monophyletic clusters with 100% confidence, and two additional breeds formed monophyletic clusters with >90% confidence (Fig. 3). Seven of the 13 breeds that did not comprise a single branch were within the Scenthound clade, where breeds are frequently defined by single morphological features such as color or height.

Overall, the analyzed breeds form 25 major clades comprising two to 49 breeds that cluster in >65% of permutations (average 90%). Only two breeds, the Norwegian Lundehund and the Löwchen, cluster consistently within clades with which they do not appear to share obvious clade-defining traits. However, the Norwegian Lundehund does not display significant haplotype sharing with any breed, including the Terriers with which it clusters, or with Nordic Spitz types. This is likely related to the drift associated with high levels of homozygosity and random IBS with Terriers. By comparison, the Löwchen shares haplotypes with other small fluffy-type dogs, but clusters with the small hounds and has recent haplotype sharing with the dachshund, suggesting a common origin for the size and coat variation found in these breeds.

The major clades are made up of breeds sharing occupation, morphological traits, and/or geographic origin. Within the larger clades, additional structure can be found with subclades (97% average cluster confidence) displaying a second layer of similarity. In some cases, clade structure reflects the relationships among breed varieties. For example, the German Spitz are split by size, with the Klein and Mittel varieties clustering with the Pomeranian and Volpino Italiano breeds, while the German Giant Spitz and the samples labeled simply as German Spitz clustered with the Keeshond breed. The next most closely related group contains the American Eskimo Dog and Japanese Spitz, two breeds that were created from the German Spitz. Within the scenthound clade, a clade



**Fig. 3** Genetic relationships between 1563 samples spanning 292 breeds. The plot is annotated for major and minor breed groups, breed purpose, and key morphological features. The fraction of haplotype sharing is indicated by the heatmap

described by occupation, subclades correspond to the geographical origin of the breeds. Alternatively, in the Hungary clade, a clade defined by geographical origin, subgroups can be found that indicate the occupation of member breeds.

We next assessed levels of haplotype sharing among breed dogs. Consistent with previous studies [49], the average haplotype sharing of dogs within a breed is >40 times greater than the average among dogs from different breeds (average across breeds = 23.5 Mb). Dogs representing breeds within the same clade, as identified on the consensus neighbor joining cladogram, share haplotypes at 3.6 times the average observed in breeds from different clades, and sharing is seven-fold higher for breeds within subclades compared to breeds in distinct clades [(Mann–Whitney test for all above comparisons is  $p < 2.2 \times 10^{-16}$ ) (Fig. 3)]. The Asian clade, as well as the Flockguards and Sighthounds clade, have the lowest amount of sharing with other clades.

We observe excess haplotype sharing among the terrier clades as well as between the terriers and the Mastiff clades. This is reflective of breed development via admixture or recent ancestry involving multiple clades, where the extent of haplotype sharing correlates with the method of breed development. For instance, there is a long-standing history of terrier and mastiff-type breeds being crossed in the mid-1800s to form multiple bull terrier- and terrier-like breeds such as the Staffordshire Bull Terrier and the Boston Terrier (see the Bull Terrier subclade). There is also excessive sharing between the Mastiff clade and the Retriever clade that has not been observed in previous phylogenies, but suggests recent admixture between these breeds or their ancestors. German Shepherd Dogs and related breeds show the largest number of admixture events with independent breeds from multiple clades (Fig. 3). German Shepherd Dogs, specifically, have sharing values greater than 95% of background levels with 29 breeds from 13 clades and three of the non-clade breeds. Breeds within the German Shepherd clade are the only ones showing significant levels of haplotype sharing with wolves. Since a similar analysis with SNV genotyping arrays and the CanFam 3.1 Boxer reference genome revealed the same result [49], using a German Shepherd Dog reference genome is unlikely to contribute significantly to this observation.

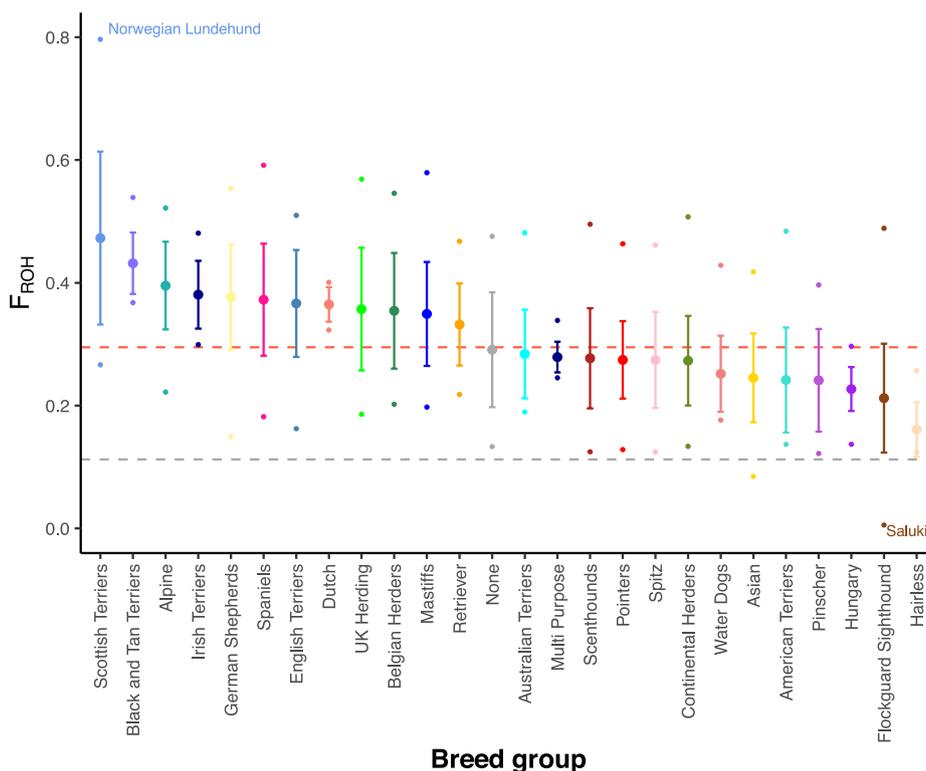
### Runs of homozygosity within sample categories

Runs of homozygosity (ROH) in an individual's genome result from the inheritance of two copies of an ancestral haplotype in that individual, and so ROHs are autozygous (homozygous by descent). The estimated proportion of a genome(s) that is in ROH gives a measure of individual or population level inbreeding. For all dog breeds, selection has involved some level of inbreeding and this has resulted in a wide range in ROH across breeds [50–52]. For each genome in the Dog10K collection, we estimated the proportion in ROH ( $F_{ROH}$ ) (Additional file 1: Table S5). This provides high-resolution estimates of historical levels of inbreeding within breeds and breed groups, as well as the genomic coordinates of regions where ROH are never found. Regions lacking ROH may indicate locations where heterozygosity is maintained for correct function. As expected, wild canids show the lowest genome proportions in ROH, with coyotes possessing the smallest total average ROH length (45.2 Mb), and breed dogs having the largest (665 Mb, Table 1). However, there is large variation in these averages, with some individuals and breeds showing particularly elevated, and others particularly low, ROH (Fig. 4,

**Table 1** Runs of homozygosity (ROH) by sample category

Average ROH		Total ROH		
Category	Count	Length (Mb)	Largest (Mb; genome %)	Smallest (Mb; genome %)
Breed dogs	1267	0.525	1842; 79.6% (Norwegian Lundehund)	12.8; 0.6% (Saluki)
Village dogs	670	0.373	872; 37.7% (Nepal)	8.3; 0.4% (China)
Wolves	570	0.438	946; 40.9% (Sweden)	20.2; 0.9% (Tajikistan)
Coyotes	152	0.298	61.2; 2.6%	38.8; 1.7%

The sample or breed population with the largest or smallest total amount of ROH is indicated



**Fig. 4** Proportion of the genome covered by ROH ( $F_{ROH}$ ). Mean and standard deviation are plotted for breed groups and are colored as per Fig. 3. Red dashed line shows mean  $F_{ROH}$  for breed dogs; gray dashed line shows mean  $F_{ROH}$  for wolves. Breeds containing individuals with the highest and lowest  $F_{ROH}$  are labeled

Table 1). For example, a Norwegian Lundehund had the largest number of ROH bases (total ROH = 1842 Mb;  $F_{ROH}$  = 78.8%), while a Saluki (sighthound) had the fewest (total ROH = 12.8 Mb;  $F_{ROH}$  = 0.56%).

We identified 389 genomic regions that were devoid of ROH in any sample. The ROH-free regions have a mean length of 64.5 kb and range in size from nine bp to a 1.3-Mb region at the start of chr35. The telomeres of all 38 autosomes lack any ROH. We compared all 389 ROH-absent regions to regions of the genome with low depth and mapping quality (“uncallable” regions). We found that 369 of the 389 overlapped for at least 80% of their length with uncallable regions, and all but two overlapped with uncallable regions along at least 20% of their length. The two regions with ROH outside of uncallable regions are gene free, short stretches (1.4 and 0.9 kb) at the ends

of chrs22 and 30, respectively. The Dog10K dataset therefore provides sufficient resolution to show the presence of ROH across almost the entirety of the dog genome. A previous study [52] identified a set of 27 genes where at least one exon did not overlap ROH in any of 4342 dogs analyzed by SNV genotyping arrays. Exons of all 27 of these genes were found to be present in at least one ROH in our dataset, suggesting that the lack of ROH in the previous analysis was likely due to the lower-density data derived from genotyping arrays rather than the presence of recessive lethal variants.

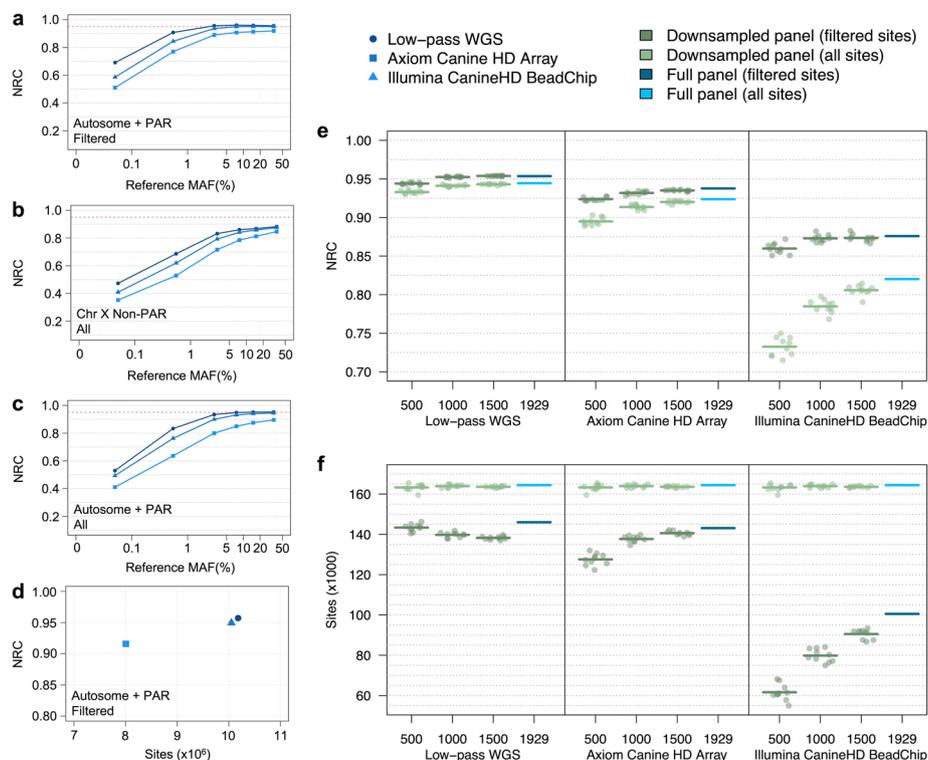
### Imputation

The size and breed diversity within the Dog10K dataset provide an excellent opportunity for genotype imputation. The Dog10K imputation reference panel includes all 1929 samples phased for biallelic SNVs with a missing genotype rate < 5%. To test imputation utility, we analyzed 10 publicly available WGS samples representing 10 breeds; five of these breeds were included in the Dog10k collection (Additional file 1: Table S6). Data from each WGS sample were downsampled to represent three separate genotyping platforms; (i) low-pass WGS, (ii) Axiom Canine HD Array, and (iii) Illumina CanineHD BeadChip. Imputation accuracy was positively correlated with platform variant density. For example, imputation based on autosomal and X-PAR sites from low-pass WGS data achieved non-reference concordance (NRC) rates of 0.95 using a reference MAF > 1%. Accuracy rates were maintained for genotypes imputed from the Axiom Canine HD Array sites, but only at a higher reference MAF (> 5%) (Fig. 5a). In contrast, the X chromosome non-PAR had lower imputation accuracy for all three platforms (NRC rates < 0.90, Fig. 5b). Requiring an INFO score > 0.9 improved NRC rates across all platforms, with the largest gain noted for rare alleles (reference MAF < 1%) (Fig. 5a,c). Accuracy rates were similar between the majority of individuals, regardless of whether the imputed individual's breed was represented in the Dog10K reference panel or not (Additional file 2: Fig. S1).

We next tested the impact of the Dog10K imputation reference panel size on imputation quality and genotype ascertainment. Here, chr38 genotypes from the publicly accessed samples were assessed. From the full Dog10K panel, ten reference panels were created for each of 500, 1000, or 1500 randomly selected individuals. Independent of the modeled genotype platform, the larger reference panels show increased imputation accuracy, although the gains in NRC rates were reduced using panel sizes > 1000 (Fig. 5e). Specifically, NRC rates differed by only 0.001 between the 1000 and 1929 sample panels. Compared to the low-pass WGS platform, NRC rates for the Axiom Canine HD Array and Illumina CanineHD BeadChip array differed by 0.006 and 0.003, respectively. Despite small gains in NRC rates, the larger reference panels revealed increased counts of imputed variants with high-quality scores. For example, the transition from 1000 to 1929 samples resulted in the ascertainment of 6268 chr38 variants for the low-pass WGS platform, 5394 for the Axiom Canine HD Array platform and 20,707 for the Illumina CanineHD BeadChip platform (Fig. 5f).

### Mitochondrial sequence analysis

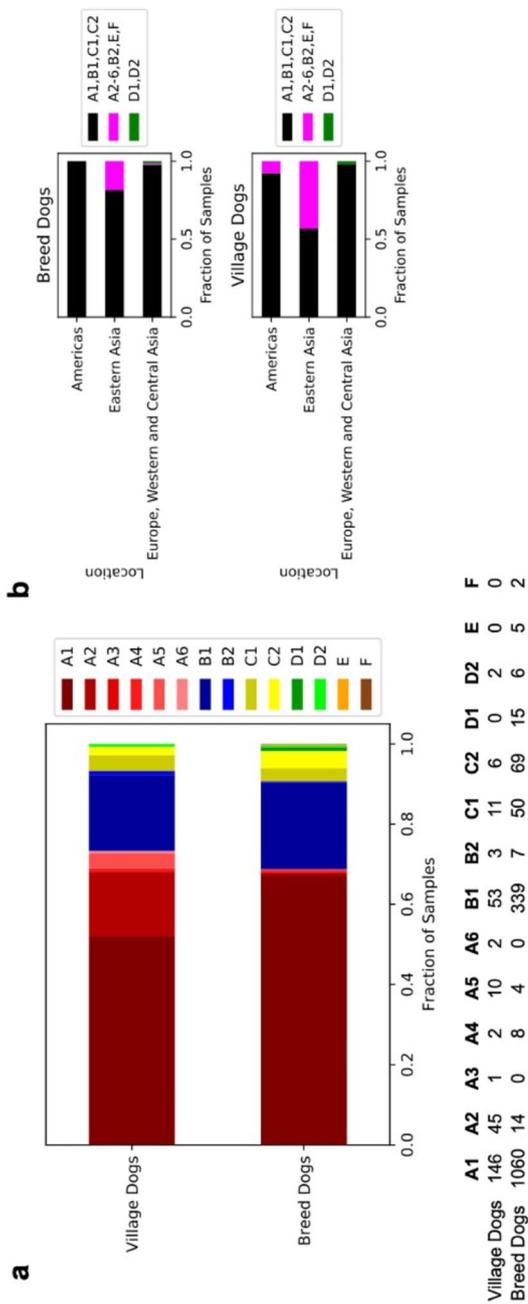
The mitochondrial genome is often overlooked in large nuclear genome sequencing projects, despite the importance of mitochondrial variation for forensics [53–56] and studies of ancient and modern canine diversity [10, 57–60]. Here, we reconstruct the



**Fig. 5** Genotype imputation accuracy of the Dog10K reference panel. **a** NRC rates of imputed genotypes across autosomes and the PAR segment of chromosome X. Variant sites are filtered according to GLIMPSE and IMPUTE5 imputation quality scores (INFO > 0.9). **b** NRC rates of imputed genotypes across the non-PAR segment of chromosome X. Variants are not filtered by imputation quality score, as imputation software does not provide scores for haploid genotypes. **c** NRC rates of imputed genotypes across autosomes and the PAR segment of chromosome X prior to filtering on imputation quality. **d** NRC rates and total number of imputed sites for each platform. Sites were filtered according to imputation quality score > 0.9 and reference MAF > 1%. **e** NRC rates for downsampled and full chromosome 38 reference panels for sites with reference MAF > 1%. Results show both quality and non-quality filtered sites. Data points show NRC rates for a single downsampled reference panel. Horizontal bars indicate mean NRC rates for each reference panel population size. **f** Number of imputed variants for downsampled and full chromosome 38 reference panels for sites with reference MAF > 1%. Results show both quality and non-quality filtered sites. Data points show the number of imputed variants for a single downsampled reference panel. Horizontal bars indicate the mean number of variants for each reference panel population size

mitochondrial genome of 1933 samples, including 1929 dogs and wolves, and four coyotes (Additional file 2: Sect. 6). Consistent with previous expectations [61], most dog mitochondrial genomes (85.8% of dogs) belong to the A1 or B1 haplogroup. Other subclades of A and B as well as clades C, D, E, and F are represented at lower frequencies (Fig. 6a, Additional file 1: Table S1). The most common haplogroups (A1, B1, C1, and C2) have a broad geographic distribution. In contrast, rarer haplogroups such as A2, A3, A4, A5, A6, B2, E, and F are found primarily in Eastern Asia (Fig. 6b).

Across the 1933 individuals, only 887 unique mitochondrial sequences (haplotypes) were observed. The most common was present in 52 individuals (2.69% frequency), and the 12 most common haplotypes were observed in 20% of samples (393/1933 individuals). We calculated the average number of pairwise differences for each group containing at least three samples (six wolf and 18 village dog populations, based on country of origin, and 261 breeds). On average, village dogs contain the highest level of mitochondrial diversity, but a range of variability is seen across groupings. Remarkably, 23 of the



**Fig. 6** Assignment and distribution of haplogroups in breed and village dogs. **a** Each breed and village dog was assigned to an existing mitochondrial haplotype. **b** Broad geographic categories based on sample origin

261 breeds with at least three individuals contain only a single mitochondrial haplotype. Linking the mitochondrial and nuclear genomes, we observe a weak correlation between within-breed mitochondrial and autosomal sequence diversity (Spearman correlation of 0.29,  $p = 1.7 \times 10^{-6}$ ). This correlation is reduced when breeds with no mitochondrial diversity are omitted (Spearman correlation of 0.19,  $p = 0.004$ ).

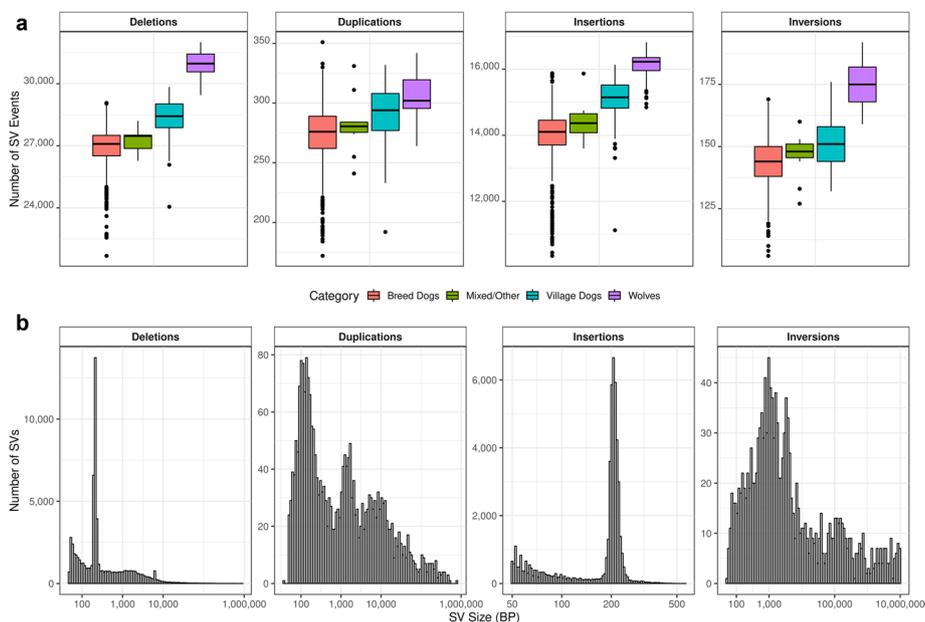
### Structural variation

Structural variation is an important source of genome variation, and it plays a variety of roles in genome evolution, adaptation, and gene expression [62, 63]. We assessed structural variation (> 50 bp) in a reduced set of 1879 samples with uniform read depth profiles. We constructed paralog-specific copy-number maps based on uniquely mapping 30-mers, and observed notable regions of increased copy number (Fig. 2b), including previously described duplications on chrs9 (*SOX9*) and 18 (*MAGI2*) (Additional file 2: Sect. 7). The UU\_Cfam\_GSD\_1.0 reference, similar to other dogs, lacks the *MAGI2* duplication inserted within the *SOX9* locus. However, both these regions are polymorphic across dogs, contributing to the increased copy number patterns and genome assembly errors [30]. The multi-locus copy number pattern was also evident in wolves (Additional file 2: Fig. S2).

We also noted a 32-kb locus with an extremely large copy-number range located at chr26:31,435,296–31,467,885. The region is duplicated in the UU\_Cfam\_GSD\_1.0 assembly and is highly polymorphic in the Dog10K collection, with Quick-mer2 [64] copy number estimates of 60–70 copies in dogs, and up to 120 copies for wolves. The region lacks annotated genes and is not found in the human reference genome (hg38). We intersect Quick-mer2 copy number estimates with coordinates of 18,162 protein-coding genes and observe only 22 genes with a median copy number > 3, including the expected *AMY2B* locus [65] (Additional file 1: Table S7). In total, 1745 protein-coding genes have a copy number range > 2.5 across the Dog10K collection; of these, 546 genes have a single sample that has an outlier estimated copy number.

Using Manta [66], which utilizes read-pair and split-read signatures to identify variation, and GraphTyper2 [67], which genotypes structural variants using pangenome graphs, we quantified 147,113 deletion, tandem duplication, insertion, and inversion structural variants (Fig. 7). We assessed linkage disequilibrium (LD) between genotyped structural variants and SNVs and found that 64.7% of deletions, 58.6% of insertions, and 43.8% of duplications are in strong LD ( $r^2 > 0.8$ ) with a flanking SNV. The lower levels of LD found with duplications likely reflect both a higher mutational recurrence rate and lower genotype accuracy for this SV type. On average, we find 26,960 deletions (affecting a total of 69,950,356 bp) and 14,034 insertions (affecting a total of 2,566,573 bp) in each purebred dog. We detect an average of 14% more structural variants in wolves than breed dogs, including 30,943 deletions (affecting a total of 66,291,676 bp) and 16,071 insertions (affecting a total of 2,761,848 bp) per sample.

Insertion and deletion variants were further queried for intersection with genes. Due to the length range of the deletions, some structural variants impacted multiple genomic feature types (e.g., intron and exon, splice region and untranslated region. Additional file 2: Sect. 7). A total of 31,950 deletions were identified that intersected 12,522 genes, including 5372 genes with an exon deleted. This includes deletion variants identified



**Fig. 7** Structural variation detected across 1879 samples. **a** Boxplots of the number of deletion, duplication, insertion, and inversion variants are shown broken down by sample category. **b** Histograms of the size distribution of each class of detected structural variants are shown. An increase in variant count at the ~200 bp size bin is apparent for deletions and insertions. This corresponds to the size of SINEC elements

by Manta that perfectly correspond to the coordinates of introns; additional examination revealed that many Manta deletion calls correspond to the presence of a retrogene. This includes the *FGF4* locus, where both the retrogene variation associated with the chondrodysplasia and chondrodystrophy phenotypes are present in multiple dog breeds [68–70], as well as a 133-kb multi-gene duplication responsible for the dorsal hair ridge in Rhodesian and Thai Ridgebacks [71]. The 133 kb duplication is present in all Rhodesian and Thai Ridgebacks in the Dog10K collection, as well as three African village dogs (Congo, VILLCG000006; Kenya, VILLKE000001; Liberia, VILLLR000017).

We searched the Manta deletion calls for the intron-deletion signature indicative of retrogenes and identified 926 parent genes that have candidate retrogenes (Additional file 1: Table S8). Strikingly, 464 candidate retrogenes were not identified in a recently completed survey of retrogenes in 293 canids [72]. Additional retrogene examples include *G3BP1*, found in 50.4% of breed dogs and only 1.8% of wolves, and *MCMBP*, found in 62.6% of breed dogs and 1.8% of wolves. Both genes were previously identified as having retrogenes in dogs [72, 73].

Retrogene formation utilizes the reverse transcription activity encoded by LINE-1 transposable elements [74, 75]. A LINE-1 encoded protein is also required to mobilize SINEs [76], including the carnivore-specific SINEC elements that make a large contribution to genome differences among canines [29, 77]. The contribution of SINEC elements to canine genomic diversity is apparent as a visible spike in insertion and deletion variant counts ~200 bp in size (Fig. 7). RepeatMasker analysis indicates that SINEC sequence represents 31.7% of the deletion and 52.7% of the insertion variants identified (Additional file 2: Fig. S3). Of the 51,950 insertions, 701 intersect with an annotated exon. This includes a 223-bp insertion at chr15:18,164,073, in the second of two exons in

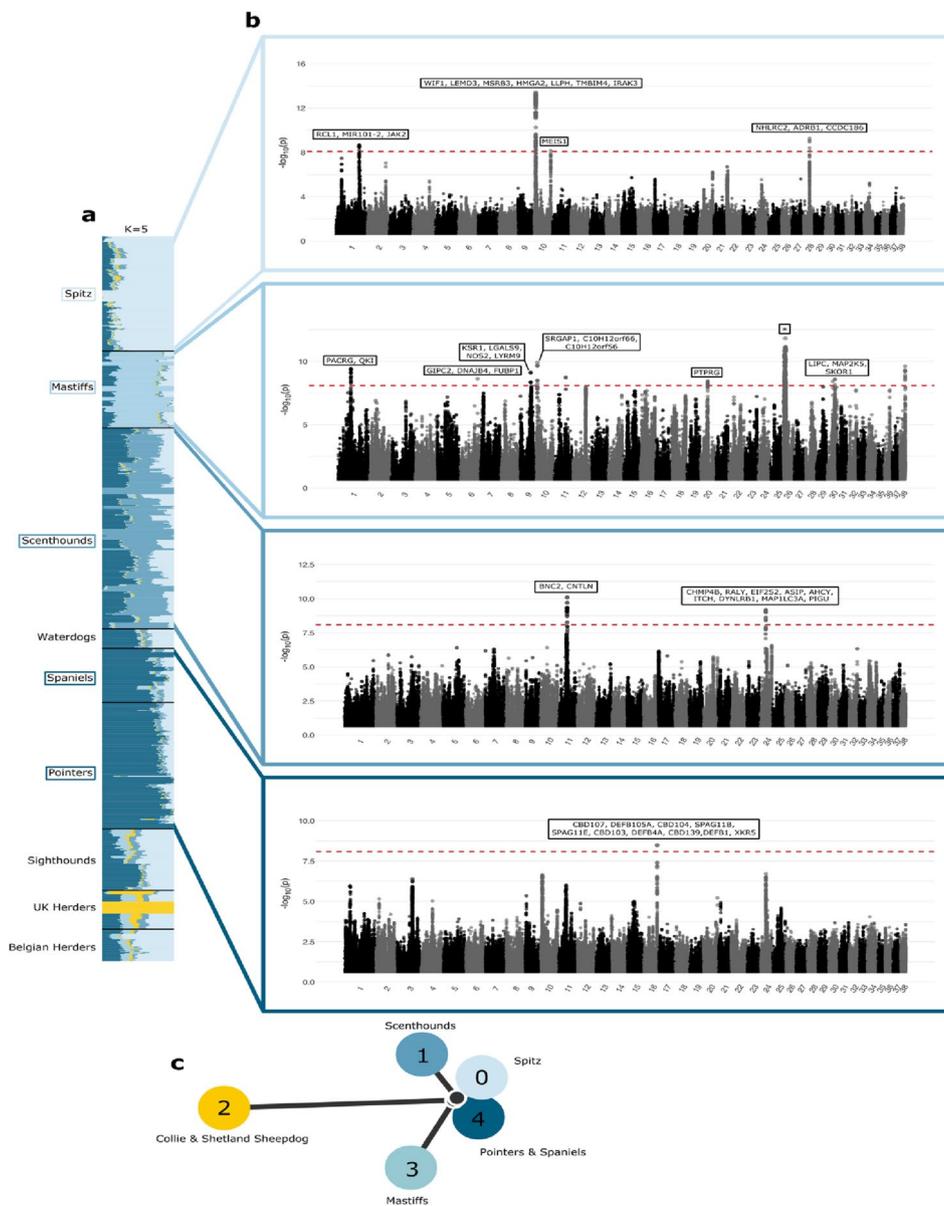
*RNASE1* (NM\_001313784.1), which is present in 47% of wolves and 0.06% of breed dogs, as well as a 214-bp insertion at chr1:108,879,297 in the third of eight exons of *ELSPBP1* (NM\_001002931.1) (found in 47% of breed dogs and 11% of wolves). RepeatMasker analysis indicates that both insertion sequences are SINEC\_Cf elements.

### Signatures of selection across breed ancestries

To test for signatures of selection among major breed clades, we assigned 790 breed dogs into nine groups (Spitz, Sighthounds, Waterdogs, Scenthounds, Pointers, Belgian Herders, UK Herders, Spaniels, and Mastiffs) based on genetic similarity and morphological features (Fig. 3, Additional file 2: Sect. 8). To balance the risks of overfitting with the interpretability of results, we focus on analysis of  $K=5$  ancestral components. These five components are distributed across the analyzed breed dogs and are maximized in the Spitz, Mastiffs, Scenthounds, Pointers, and Spaniels, and a subset of the UK Herders (Collies and Shetland Sheepdogs) (Fig. 8). Using Ohana [78], we then searched for signals of selection in each ancestral component by identifying variants with population differentiation that is not consistent with the genome-wide estimated allele frequency covariance matrix. We set significance levels based on the number of tests performed and considered genes either overlapping or within 100 kb of the significant sites as potential candidates for selection, resulting in 15 loci (Fig. 8, Additional file 1: Table S9).

Several of the candidate regions contain genes associated with variation in size, morphology, and coloration. A region on chr24 that shows selection in the ancestral component maximized in the Scenthounds contains *ASIP*, a major contributor to coat color variation [79–81]. A region on chr16 that shows selection in the ancestral component maximized for Pointers and Spaniels contains multiple beta-defensin and sperm-associated antigen genes (e.g., nine genes across *CBD*, *DEFB*, *SPAG* families). Beta-defensins have been previously linked to coat color; notably, *CBD103* (the K locus) which is located in this region, contains variants associated with black or brindle coat color [82, 83], and may be under selection in wolf populations for resistance to canine distemper virus [84]. The region on chr26 that shows selection in the ancestral component maximized in the Mastiffs is under selection in boxers (five boxers are included in the Mastiff group) [85], contains genes involved in skeletal and muscular development and function and tissue morphology, and is associated with canine body size and height [4].

We found four regions with signals of selection in the ancestral component that is maximized in the Spitz (Fig. 8). The region on chr1 includes *RCL1*, which has been associated with snout ratio and tail curl [86] as well as *JAK2*, which contributes to human [87] and canine primary polycythemia [88]. The strong peak on chr10 includes genes previously linked to ear morphology (*WIF1*, *LEMD3*, *MSRB3*, *LLPH*, and *IRAK3*) [4, 13, 86, 89], body size in humans and dogs [13, 89–91] and beak size in Darwin's finches (*HMG2*) [92]. We applied iSAFE [93], a method which ranks candidate favored mutations during a selective sweep based on haplotype and allele frequency patterns, to disentangle the signature in the chr1 locus. Setting Spitz dogs as cases and the remaining samples as controls, we found that the sites with the highest iSAFE scores, including several sites identified by Ohana, cluster in *HMG2*. Application of iSAFE to other loci revealed broad patterns of high-scoring variants that do not pinpoint a single gene (Additional file 1: Table S10, Additional file 2: Sect. 8).



**Fig. 8** Signatures of selection inferred with ancestry components. **a** Population structure inferred from Ohana for the nine selected dog groups using  $K=5$ . **b** Manhattan plots for four of the five ancestral components. Top to bottom: Spitz, Mastiffs, Scenthounds, Pointers, and Spaniels. The red dotted line represents the Bonferroni cutoff and genes either overlapping or within 100 kb from a significant site are indicated at each peak. The asterisk within the Manhattan plot for the Mastiff component contains 88 candidate genes listed in Table S7. **c** Population tree connecting the ancestral components with colors corresponding to the ancestries are shown in the admixture plot. Each ancestral component is labeled based on the dog group(s) for which it is maximized

### Function inference from Dog10K variation

Since variation predicted to alter gene function is expected to be enriched for false positives [94], we applied additional depth and genotype quality filters to the biallelic SNVs identified in 1929 dogs and wolves. These filters removed 0.7% of total available VQSR PASS sites, resulting in 27,878,361 autosomal and 847,128 chrX SNVs utilized

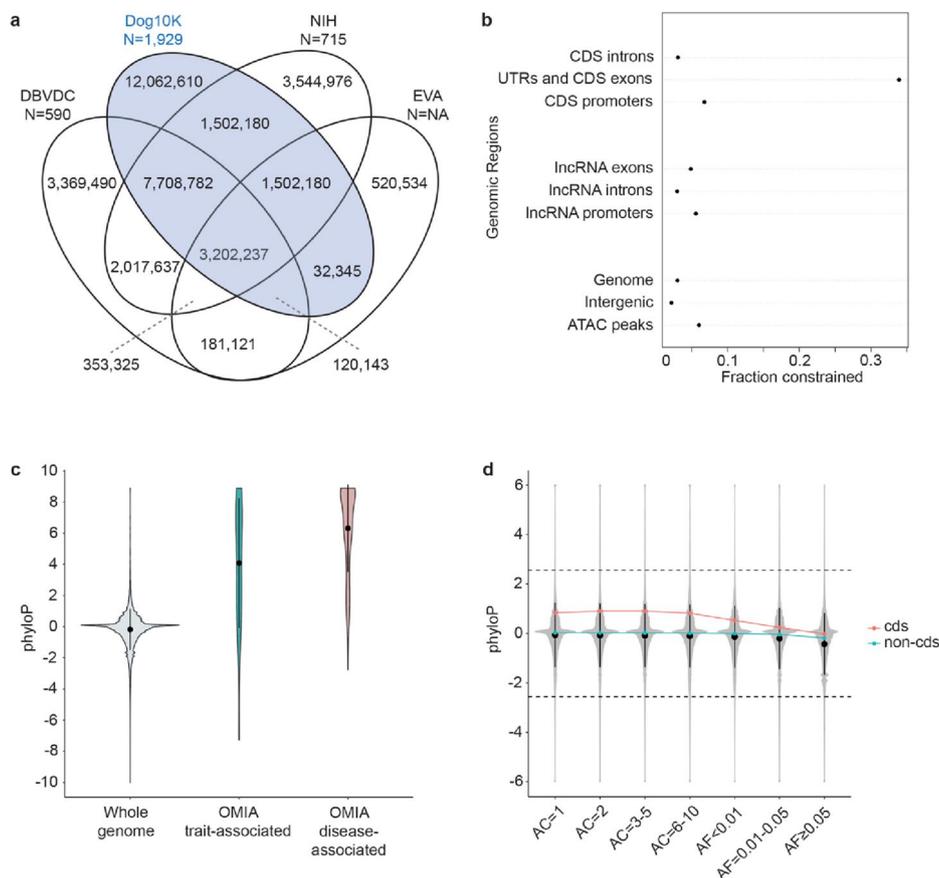
for functional analysis (Additional file 2: Sect. 9). On autosomes, 78.9% of filtered sites had an observed MAF > 1%, but the allelic profile for removed sites on chrX differed, with 58.5% of sites observed with a MAF > 1% (Additional file 2: Sect. 9). Both the VQSR PASS and strict-filtered biallelic SNV sets had concordance rates  $\geq 99.6\%$ , based on the genotypes of 168 individuals also typed on the Illumina Canine HD Array (Additional file 2: Sect. 10). The strict-filtered SNV set used for functional analyses captures 1.27% of the theoretically possible chrX and autosomal variation (Additional file 1: Table S11), with one SNV every 80 bp when all 1929 individuals were considered.

Panels of normal variation are key to prioritizing SNVs for downstream functional analyses. We compared the composition and biallelic SNV sites contributed from only dogs and wolves (when known) for three such panels, DBVDC (590 samples, 20,443,472 SNVs) [6], NIH panel (715 samples, 18,468,060 SNVs) [4], and the European Variation Archive (EVA) RS Release 3. Additional file 1: Table S12 summarizes the sample acquisition and distinct alignment and site filtering strategies for each panel. These factors, including minimum coverage depth (ranging from  $2 \times$  to  $10 \times$ ), impact the number of samples and variants available for downstream analyses (Additional file 2: Sect. 11). We find that 43% of SNVs are unique to the Dog10K collection (Fig. 9) and that 98% of these unique variants are rare (AF < 1%) and are not due to differences between the CanFam3.1 [2] and UU\_Cfam\_GSD\_1.0 [30] assemblies. This variation is in part a reflection of the diversity and uniqueness of the dog breeds included in Dog10K (60% of breeds are only found in the Dog10K collection, Additional file 1: Table S13), as well as the limited sample sharing between this and the other sets (only 10/1929 samples were shared; Additional file 2: Sect. 11). As expected, given gene density, recombination rates and other demographic pressures, genetic variation within the Dog10K dataset was not evenly spread across the genome, with example outlier peaks observed on chrs12 and 18 which harbor the dog leukocyte antigen (DLA) and olfactory receptor genes, respectively (Additional file 2: Fig. S4).

### Base-pair constraint for functional prioritization

One of the goals of the Dog10K consortium is to provide the community with a set of SNVs that can be used to aid in the identification of phenotypic associations. Within the coding region of the genome, we identify 7607 high and 129,766 moderately deleterious SNVs within the 1591 breed dog dataset (67% with AF < 1%; Additional file 1: Tables S14 and S15). Across the entire genome, we used estimates of evolutionary constraint to infer function, with Zoonomia single base-pair phyloP scores calculated from an alignment of 240 mammalian species [95]. Here, CanFam3.1 referenced phyloP scores were converted to UU\_Cfam\_GSD\_1.0 coordinates, revealing that 3.5% of the genome is under constraint (purifying selection; 5% FDR, phyloP  $\geq 2.56$ ). A large fraction of constraint bases is observed in protein-coding genes (CDS and UTRs), but an appreciable 2.2% of intergenic space is also constrained (Fig. 9b).

To benchmark the utility of the Zoonomia phyloP scores, we examined the distribution of positions classified as disease-associated or other trait-associated in the curated Online Mendelian Inheritance in Animals (OMIA) database [96]. The median phyloP score for both sets was greater than the 5% FDR for constraint, indicating that the associated bases are enriched for regions of the genome under selection (Fig. 9c). Within



**Fig. 9** Comparison among variant data sets. **a** Number of variable positions shared between major databases. **b** Fraction of genome spaces under constraint (5% FDR, phyloP > 2.56). **c** Enrichment of constrained bases in OMIA trait (blue), and disease (red) sets compared to the genome as a whole (gray). **d** Relationship between allele count (AC), allele frequency (AF), and phyloP score for the coding (CDS, red) and non-coding (non-CDS, green) bases in the whole genome (gray)

breed dogs, a negative correlation was observed between allele frequency and phyloP score (Fig. 9d), although as noted from studies in other species, common variants in constrained positions may be involved in local adaptation. In breed dogs under selection, variation at these positions may also result in favorable trait outcomes, such as the *HPS3* g.44487038G > A variant (phyloP = 7.03), associated with the “cocoa” brown color segregating in French Bulldogs [97].

### OMIA variants found in the Dog10K collection

Breed dogs were submitted to Dog10K with the expectation that they were free from known disease. However, they cannot be evaluated for phenotypes that develop late in adulthood, nor can health be ascertained for village dogs or wolf samples. We therefore examine the frequency distribution of trait-associated (morphology or other) and disease-associated variants accessed from OMIA [96]. For this analysis, 337 SNVs and small indels, each variant included in the OMIA database, and each spanning less than 20 consecutive bases, were selected for interrogation. Of these, 76 variants were detected in at least one individual in the Dog10K

collection (Additional file 1: Table S16). As expected, the alternative allele frequency of the morphology-associated variants spanned all frequency classes (14 variants, Alt AF = 0.2–72%), and for each, all three genotype classes were observed (Additional file 1: Table S17). Overall, 58 OMIA disease-associated variant positions were detected in Dog10K, with 62 homozygous occurrences detected across 15 disease traits (Table 2, Additional file 1: Table S16). This was not unexpected for some diseases where affected individuals have variants associated with a mild phenotype, sex-limited inheritance, late-onset, or incomplete penetrance.

### HWE deviation to identify candidate disease variants

Since individuals within the Dog10K collection were assumed to be healthy at the time of sampling, we hypothesize that disease variants would show depleted homozygous frequencies [98]. Using the VQSR PASS SNV VCF as an input, we find 42 missense variants that pass the initial filtering criteria, with genotype and variant site quality statistics further narrowing the candidate list to seven SNVs. After visual inspection of alignment files, it was evident that in each case, other factors such as the existence of pseudogenes, assembly artifacts, and structural variation provided more parsimonious explanations for observed departures from HWE. We note that 13/42 HWE deviation candidates are retained within the strict-filtered VCF (Additional file 1: Table S18).

**Table 2** Dog10K samples with likely causal homozygous genotypes for autosomal recessive diseases, risk factors, or traits

Trait	OMIA ID	Gene	Homozygous samples (N)	Dog10K breeds or village dogs carriers
Lens luxation	000588–9615	<i>ADAMTS17</i>	1	American Toy Terrier
Persistent Mullerian duct syndrome	000791–9615	<i>AMHR2</i>	1	Miniature Schnauzer
Laryngeal paralysis and polyneuropathy	002301–9615	<i>CNTNAP1</i>	1	Pyrenean Shepherd
Exercise-induced collapse	001466–9615	<i>DNM1</i>	2	Curly Coated Retriever
Dwarfism, growth-hormone deficiency	001473–9615	<i>GH1</i>	6	Bolonka, Brussel Griffon, Petit Brabancon Griffon
Lundehund syndrome	002031–9615	<i>P3H2</i>	4	Norwegian Lundehund
Ichthyosis, PNPLA1-related	001588–9615	<i>PNPLA1</i>	2	Golden Retriever
Progressive rod-cone degeneration	001298–9615	<i>PRCD</i>	4	Australian Cattle Dog, Entlebucher Mountain Dog, Portuguese Podengo, Swedish White Elkhound
Hypotrichosis, recessive	001279–9615	<i>SGK3</i>	4	American Hairless Terrier
Urolithiasis	001033–9615	<i>SLC2A9</i>	6	Dalmatian, Majorca Mastiff
Oculocutaneous albinism, type IV	001821–9615	<i>SLC45A2</i>	1	Bullmastiff
Degenerative myelopathy (risk factor)	000263–9615	<i>SOD1</i>	22	many (incl. village dogs)
Thrombocytopenia, TUBB1-related	002434–9615	<i>TUBB1</i>	2	Norfolk Terrier
Von Willebrand disease I	001057–9615	<i>VWF</i>	2	Kromfohländer
Von Willebrand disease II	001339–9615	<i>VWF</i>	4	Boykin Spaniel, German Spitz

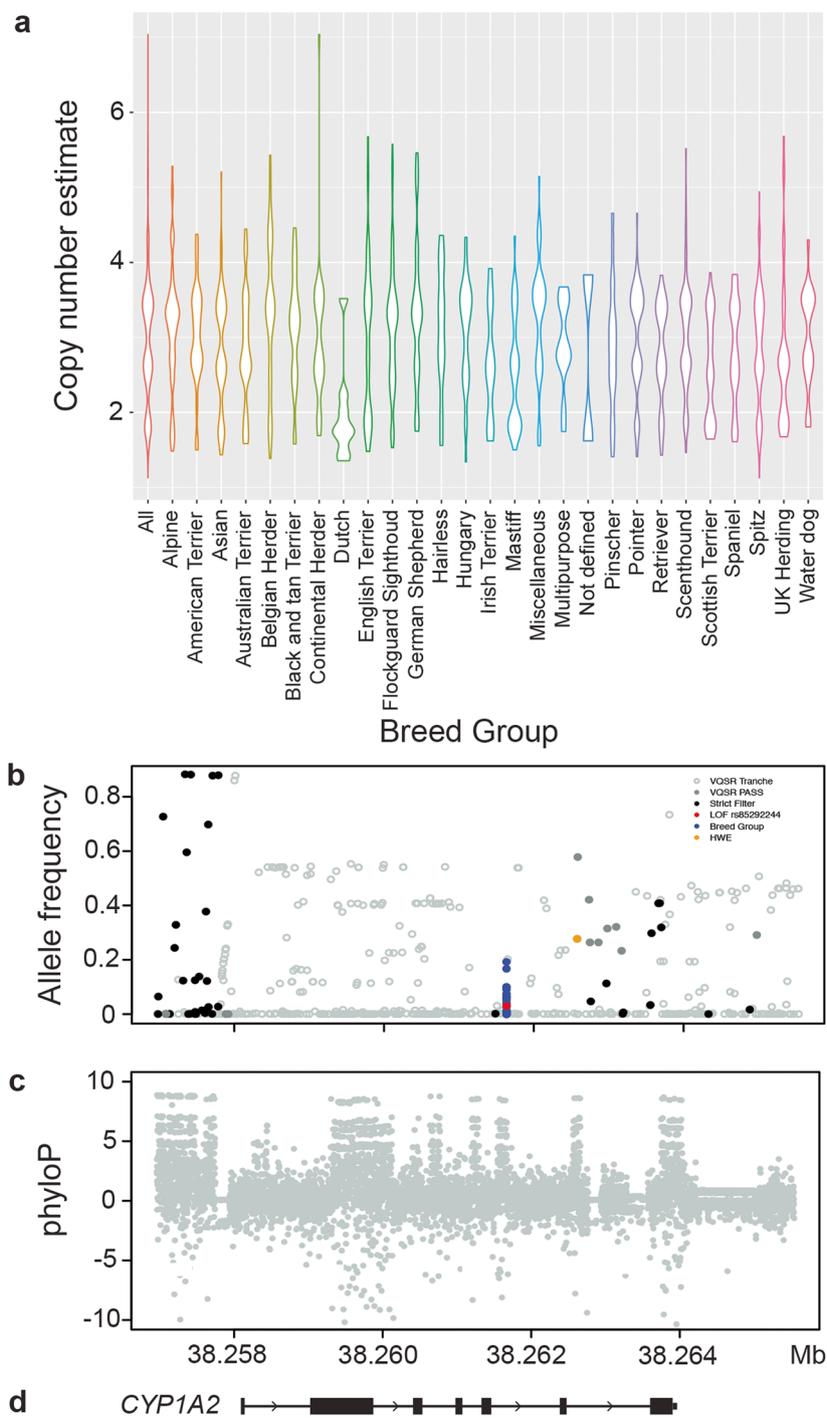
### Variants affecting metabolism and drug targets

To illustrate the utility of the Dog10K collection, we searched for genetic variation affecting druggable targets. From the 1427 genes in the human Tier 1 druggable gene set [99], we identified 79 genes with their full coding sequence impacted by SVs, and 249 genes with high-impact snpEff SNV annotations (375 SNVs, median phyloP = 3.16). At the known SV variable selective phosphodiesterase type 4 inhibitor, *CYP1A2* (28), 49.7% of samples (934/1,879) are estimated to have a copy number  $\geq 3$ . This variability is noted in all major breed clades (Fig. 10a). This locus provides the opportunity to visualize the impact of SNV filtration and functional consequence (Fig. 10b). The allele frequency of SNVs failing the VQSR PASS tranche (open gray circles), passing this tranche (filled gray circles), and available after strict filters (black circles) are plotted. Highlighted in red is the loss of function SNV, rs852922442, for which the C > T causes a premature stop codon and decreased *CYP1A2* expression [100, 101]. While rare across dog breeds overall (AF = 0.03), the rs852922442-T allele is common in individuals from the German Shepherd and Scenthound clades (AF = 0.19 and 0.07, 26 and 202 samples respectively), while notably absent from the Mastiff, American Terriers, Australian Terriers and Belgian Herder clades (96, 31, 27, and 26 samples, respectively; Fig. 10b, blue dots). This wide AF distribution likely explains the observed interindividual variability associated with the pharmacokinetics of *CYP1A2*-substrate drugs in dogs. This locus also includes an HWE deviation candidate variant (orange circle), discounted due to the presence of the locus spanning SV.

Another interesting druggable target is a previously unknown SV, which spans the entire coding sequence of *SLC28A3* (Fig. 11a). Across breeds, we observe clearly defined profiles corresponding to gene copy numbers of two, three, and four (Fig. 11b). All Grand Basset Griffon Vendéen dogs (GBGV) have a copy number of  $\geq 4$  at this locus, with one individual (GBGV000003) inferred to have a copy number of six (Fig. 11a). *SLC28A3* (previously *CNT3*) is a concentrative nucleoside transporter with many functions. While no high-impact coding variants are present in the Dog10K strict variant catalog, coding variation in the human *SLC28A3* ortholog are known to influence the metabolism of gemcitabine [102, 103], a drug used to treat solid tumors in human [104] and canine patients [105, 106]. Outside of the clinic, one of the most interesting observations has been the *SLC28A3* association with advanced maternal age in studies of the methylome [107]. At least one study, which profiled the DNA methylomes of paired parental peripheral blood and cord bloods from nuclear families, revealed that methylome-associated expression changes in many genes, including *SLC28A3*, are related to adverse outcomes in advanced maternal age pregnancy, a serious consideration in canines.

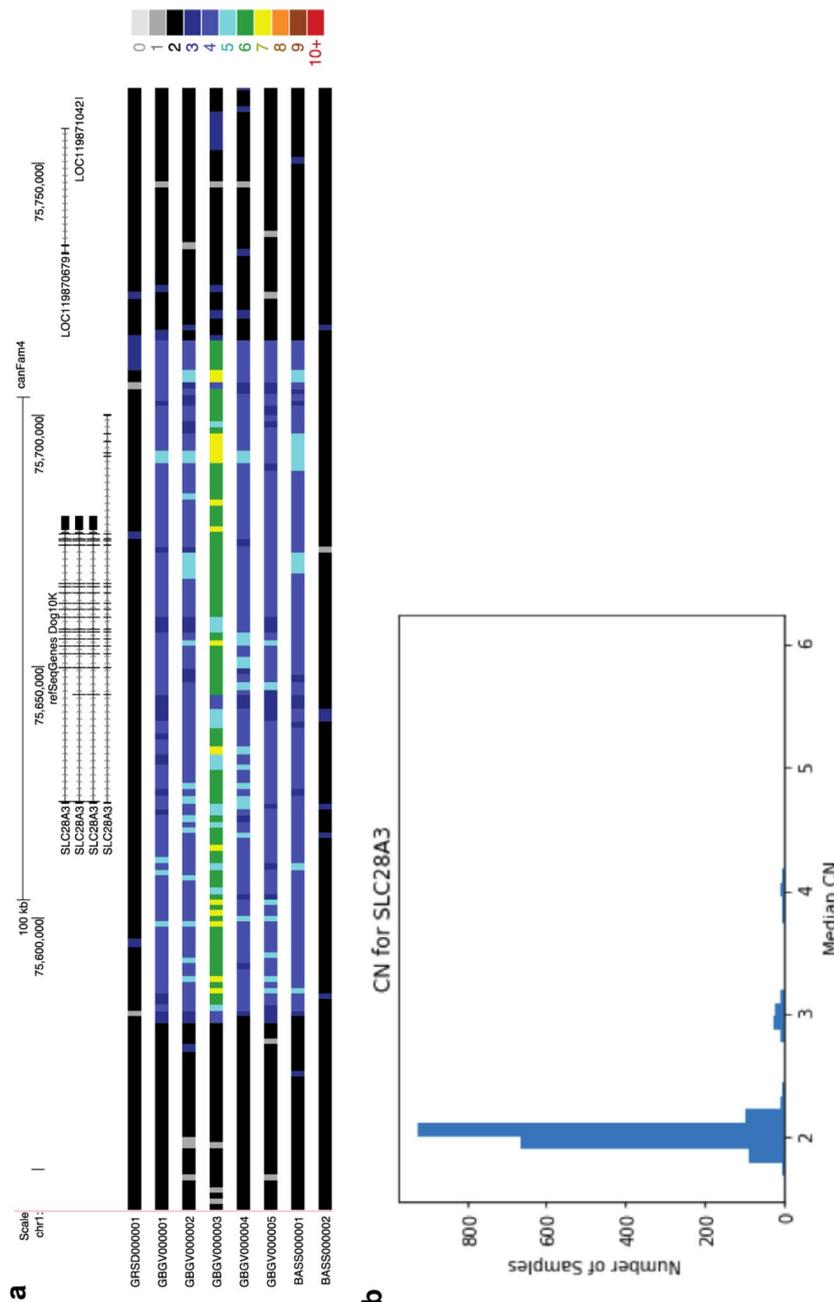
### Discussion

The Dog10K consortium sequenced and analyzed a large and diverse genomic sampling of canids. Our results encompass a harmonized resource of more than 48 million SNVs, indels, structural variants, and mitochondrial sequences as well as a set of pipelines and best practices for expansion to larger data sets. The identified variants are a valuable dataset that will enable future studies into the demographic and selective history of canids and serve as a panel of variation useful for the exploration of diseases and other phenotypes.



**Fig. 10** Structural and point variants impacting *CYP1A2*. **a** Distribution of copy number variation within all breed dogs, and the major clades. **b** Location of SNVs across *CYP1A2* inclusive of filtering or impact status. **c** Mammalian phyloP scores (bounded by 10, -10). **d** Illustration of the region from the reference genome perspective

Our refined dataset includes 1579 samples from 321 breeds or defined populations, of which 261 are represented by three or more dogs (Additional file 2: Sect. 3). In addition, we sequenced 293 mixed breed and village dogs, and 57 wolves sampled from multiple



**Fig. 11** *SLC28A3* locus copy number expansion. **a** Location and span of the copy number element in polymorphic Grand Basset Griffon Vendéen (GBGV) and basset hound (BASS) individuals contrasted with copy number two samples from the German Shepherd Dog (GRSD) and BASS breeds. **b** Distribution of the median copy number count across the 1879 samples assayed

geographic regions which, in aggregate, allows us to capture not only considerable levels of phenotypic diversity but permits the ascertainment of substantial levels of genetic variation. Our comprehensive ROH analysis is likely to prove key to understanding the historical relationships among modern breeds, the history of breed development, and the relationships between modern, historical, and ancient canids. With one variant every 80 bp, the Dog10K collection has captured most of the genetic variation present in 22 common breeds. This allows future efforts to focus on other rare breeds or geographically isolated populations to reveal the role of undiscovered variation in canine biology and evolution.

Our variant filtering pipeline leveraged sites routinely genotyped in commercial arrays as a training set to identify 34.5 million high-quality SNVs. Since a robust truth set is not available for indel variants, we applied hard filters based on criteria recommended by the GATK best practices to identify 14.4 million indels. The indel total includes sites with a mixture of SNV and indel alleles. Our indel to SNV ratio of 2.4 is similar to that reported by two other recent surveys of dog and wolf variation [6, 30]. However, another study of canines, which included additional outgroup samples from the *Canis*, *Cuon*, and *Lycalopex* genera, reports an SNV to indel ratio of 4.2 [4], while studies of equines [108], bovines [109], and humans [110] report SNV to indel ratios greater than 10. It is not clear to what degree the apparent excess of indel variation in canines reflects true biological differences or technical artifacts in calling. Given this uncertainty, our analysis is primarily focused on SNVs.

Our analyses demonstrate the utility of the Dog10K variant dataset as a reference panel for use in genotype imputation [111], an approach which has been shown to be effective in making use of low-pass or poor-quality sequence data [112, 113]. Canine studies have successfully incorporated this approach [114–117], particularly for disease GWAS, leading to identification of a risk haplotype for congenital laryngeal paralysis in Alaska sled dogs [118], and a locus for canine idiopathic pulmonary fibrosis in West Highland white terriers [114], among others.

The largest previous study, based on a panel of 676 dogs from 91 breeds with 97 high-coverage WGS dog samples downsampled to approximately  $1 \times$  coverage per sample, demonstrates that both quality filtering and MAF were critical to accuracy [117]. Both affect power to conduct successful GWAS, with a previous study demonstrating that as the MAF difference between cases and controls is reduced, the number of samples required for imputation of low-pass WGS to reach the same power in a GWAS as high-coverage WGS grows exponentially [117]. While this study suggested discarding sites with a  $MAF < 0.05$ , our data argues for selecting variants with imputation quality  $> 0.90$  and reference MAFs  $> 1\%$ . This reflects both the large number of dogs and breeds as well as the variation captured in village dogs in our dataset, both of which are critical for the development of any reference panel, in dogs [119, 120] or otherwise [121]. For the Illumina CanineHD BeadChip platform, the criteria we propose will provide imputed genotypes with NRC rates  $> 0.85$  for over 8 M sites, whereas for the low-pass WGS and Axiom Canine HD Array platforms, these criteria provide NRC rates of approximately 0.95 for over 10 M sites (Fig. 5d). It is important to note, however, that any imputation analysis is only as accurate as the samples in the reference panel, and expansion to a panel even larger than we present here, is an important long-term goal.

By modeling allele frequency changes and admixture, we identified regions that show frequency differentiation across five ancestral components found throughout the analyzed breeds. Analysis reveals that many of these signals likely reflect selection for body size and coat color during the establishment of breeds, several of which were identified previously, thus validating the completeness of the Dog10K dataset. The precise identification of the genes targeted by selection during breed formation is hindered by the extended range of linkage disequilibrium in breed dogs [122]. As a result, identified loci often contain multiple genes previously associated with disparate phenotypes. For example, a large region on chr26 shows signatures of selection in the ancestral component that is maximized in the Mastiff group. This 7.5-Mb region shows an extended reduction in nucleotide diversity relative to other clades (Additional file 2: Fig. S5), and includes genes associated with canine body size and height [4] as well as glioma risk [123] and other cancer phenotypes [124–126].

To refine the selection candidates, we applied iSAFE [93], a method for fine mapping mutations favored during selective sweeps, to the regions we identified. Our analysis nominates *HMG2*, a known regulator of canine body size, as the likely target in the chr10 locus that was selected in the ancestral component that is maximized in Spitz dogs. However, for the remaining loci, we observe a broad pattern of high-scoring candidate variants distributed throughout the candidate region. Further dissection of such loci will require combinations of selection scans, association studies with well measured phenotypes, preferably including samples from multiple breeds, and functional follow-up.

By combining tools for the discovery and genotyping of structural variants, we present a genome-wide catalog of insertion, deletion, duplication, and inversion variants. Association of these variants can now be assessed in ongoing genome-wide studies. Examination of the size spectrum of the detected variants highlights the disproportionate contribution of LINE-1 encoded proteins to canine genome diversity. Although the variant discovery approach we used is unable to resolve large insertions associated with LINE-1 sequences, we found that 31.7% of deletions and 52.7% of discovered insertions are SINEC sequences that are variably present among the samples in the Dog10K collection. Unexpectedly, many of the deletion variants we identified reflect the presence of retrogenes. These insertions are missing from the UU\_Cfam\_GSD\_1.0 reference and are derived from 926 parent genes. Since our retrogene discovery was limited to deletions corresponding to the full length of introns, a targeted discovery approach is likely to identify additional retrogenes present in the Dog10K collection, reinforcing retrogenes as an important class of canine genetic variation [72].

Rigorous quality controls and filters are essential to identify rare variants that have a functional impact. The Dog10K collection utilizes a common sequencing source, as well as harmonized alignment and variant calling pipelines that aim to reduce the impact of batch effects on variant quality. We note however, from comparisons with OMIA and analysis of Hardy–Weinberg Equilibrium, that even our most strictly filtered callset is not free of false positives. These errors are multicausal and illustrate the challenges encountered in analysis of large-scale sequencing studies or diverse breeds. For example, the genotype of a female village dog from Azerbaijan was heterozygous for the chrX variant, NSDHL:c.700G>A (p.Gly234Arg). The same genotype has been reported in a Chihuahua with verrucous epidermal keratinocytic nevi [127], a disease with X-chromosomal

semi-dominant inheritance and presumed embryonic lethality in hemizygous males (OMIA002117-9615). Inspection of the short-read alignments revealed that the village dog heterozygous variant call was a false-positive, a technical artifact caused by the insertion of an *NSDHL* retrogene with a c.[700G>A] allele on chr14.

This example further highlights the challenges that retrogene insertions pose for canine clinical genetics [17, 128]. Additionally, our variant set includes a site in *CYP1A2* that is out of Hardy–Weinberg equilibrium (Fig. 10). This position is targeted on existing SNV genotyping arrays and was included in the VQSR truth training set, and thus survived the resulting variant filters. In both the *NSDHL* and *CYP1A2* examples, access to alignment files and additional SNV quality metrics allowed the errors to be identified. Despite advances in variant filtration methodologies, manual curation of rare, functionally important variants remains essential.

The scale of the Dog10K variant collection makes this a valuable resource for functional prioritization. Again using the OMIA analyses, we find 179 samples homozygous for a *BTBD17* variant associated with a 46,XX disorder of sex development and embryonic lethality [129]. This non-coding single-base insertion (XM\_038546704.1: c.85 + 206\_85 + 207insG) was observed at a frequency of 22%, which is higher than expected for a variant that causes a severe disease (Additional file 1: Table S17). While the original finding of homozygous lethality was reported only in German Short-haired Pointers, this is a timely reminder that putative disease-associated variants should be carefully investigated prior to use in non-discovery breeds or populations, where the association between variant and pathogenic effect has yet to be confirmed.

We also examined the potential of the Dog10K collection to aid in the translation of pharmacogenetics. Here, the collection could point towards breed groups fixed for LOF variants, or highlight groups that need additional care in veterinary prognostic treatment. The *CYP1A2* locus has known clinical significance, as it plays a rate limiting step in the metabolism of multiple veterinary drugs including theophylline, clozapine, and tacrine [130]. We find that the *CYP1A2* locus is copy number variable across all defined breed groups (Fig. 10a), suggesting that this expansion predates breed formation. We also see breed group variability at the *CYP1A2* loss of function allele, rs852922442-T (Fig. 10b). Here, both sampled Keeshonds were homozygous LOF, providing a spontaneous canine model to study the compensatory effects of this gene knockout [131, 132]. While the role of *CYP1A2* is well-established, the roles of other potential drug targets examined in this study remain to be elucidated and require cautionary comment. Genes such as *SLC28A3* play roles in many biological processes spanning nutrient metabolism to COVID-19 therapy pharmacogenomics [103, 133], and we cannot assume a phenotypic outcome from gain or LOF variants. It is important to establish an a priori hypothesis, collect large numbers of samples, and phenotype each sample meticulously before applying genome level observations to clinical decisions.

## Conclusions

Variants identified in the Dog10K collection represent a global view of canine genome diversity that informs functional interpretation and enables future studies. The mapping and processing pipelines of Dog10K are open to the community, allowing for the expansion of additional samples to capture and exploit the full extent of canine diversity.

## Methods

### Sample selection, sequencing and read alignment

DNA was isolated from 2075 canids, comprising 1649 breed dogs, 336 village dogs, 18 dogs of mixed origin or that are not recognized by any international registering body (labeled as “mixed/other”), 68 wolves, and four coyotes supplied by investigators from eight sites (Additional file 1: Table S1). Samples were collected as per each institution’s animal care, collection, and use protocols (Additional file 1: Table S19). Whole genome sequencing was carried out using the Illumina HiSeq X Ten platform by Novogene (Inc.) in Tianjin, China. Approximately 0.2 µg of DNA from each sample was sheared into ~350 bp with the Covaris system, and sample index libraries generated using the NEB Next® Ultra™ DNA Library Prep Kit for Illumina (NEB, USA) following the manufacturer’s recommendations.

To analyze the nuclear genome, raw sequencing reads were aligned to a modified version of the Wang et al. German Shepherd Dog genome assembly [30] (UU\_Cfam\_GSD\_1.0, GCF\_011100685.1), supplemented by three Y chromosome sequences from a Labrador Retriever (ROS\_Cfam\_1.0, GCF\_014441545.1) assembly. Read data was processed across multiple centers using a shared GATK-based pipeline prior to centralized genotyping and filtration of candidate variants on the autosomes and chrX. Following alignment, samples were removed due to low coverage ( $< 10\times$ ), the presence of sample duplicates, mislabeled or unknown breed identity, or potential contamination indicated by reference read fraction at heterozygous positions. To identify candidate SNVs, we applied the Variant Quality Score Recalibration (VQSR) procedure with cutoffs that retain 99% of variants present on the Illumina CanineHD BeadChip and Axiom K9 HD genotyping arrays. The alignment pipeline is described in more detail in Additional file 2: Sect. 1.

Different analyses require different levels of variant stringency and sample membership. These are summarized in Fig. 1 and in Additional file 2: Sects. 2 and 9. For genome-wide assessment of SNV density, the VQSR tranche 99 primary PASS SNV (VQSR PASS) call set was used. This is derived from 1987 samples (1593 breed dogs, 309 village dogs, 18 mixed/other, 63 wolves, and 4 coyotes). Most analyses utilize a set of 1929 samples (1579 breed dogs, 281 village dogs, 12 mixed/other, 57 wolves) that pass more stringent quality controls. For *SNV functional analyses*, additional filters were applied to the VQSR PASS SNV set. These included filters based on depth ( $-\text{minDP } 5$ ), genotype quality ( $-\text{minGQ } 20$ ), and an in-house allelic balance ( $0.70 \geq \text{AB} \leq 0.30$ ) filter based on the vcf4.2 allele depth (AD) INFO field. Parameters were adjusted to suit autosomes or chrX. Filtering was followed by iterative steps of variant and sample missingness. Mitochondrial analysis utilized the set of 1929 samples supplemented by the additional inclusion of four coyotes (Additional file 2: Sect. 6). For *structural variant studies*, 1879 samples with uniform depth profiles were utilized. SVs detected using Manta v1.6.0 [66] and genotyped using GraphTyper2 v2.7.2 [67]. Genome-wide copy-number estimates were created using QuicK-mer2 [64]. Detailed methods are included in Additional file 2: Sect. 7.

### Reference annotation

The reference genome was annotated with the reference appropriate NCBI gene annotation files ([https://ftp.ncbi.nlm.nih.gov/genomes/all/annotation\\_releases/9615/106/](https://ftp.ncbi.nlm.nih.gov/genomes/all/annotation_releases/9615/106/)). All gene, transcript, exon, and CDS annotation attribute fields were updated following annotation conventions used by Ensembl. When duplicate copies of the gene ID

annotation were observed, version numbers were modified in accordance with annotation copy number within the UU\_Cfam\_GSD\_1.0 assembly, rather than annotation copy number across all assemblies annotated under NCBI release 106. This process identified 118 duplicate gene IDs, including 33 with protein-coding biotype (Additional file 1: Table S20, Additional file 2: Sect. 12). Using liftover, the reference genome was further annotated with Zoonomia CanFam3.1 phyloP scores (accessed March 2022) and regions of open chromatin (BarkBase ATAC annotation [3]).

#### Identification of regions amenable to SNV calling with short sequencing reads

A genome callability mask was created to facilitate downstream analyses. Positions marked (i) “N” in the genome reference, (ii) where  $\geq 10\%$  of aligned reads have a mapping quality (MQ) of 0, or (iii) where the total coverage was more than 50% away from the median coverage were identified as “unmappable regions”. Separate cutoffs were determined for the autosomes and X-PAR region, and the non-PAR segment of the X chromosome.

#### SNV variation within and between groups

To interrogate recent shared ancestry, we measure the statistic  $F_2$  in 1929 samples.  $F_2$  notes variants found in only two samples regardless of their zygosity and is similar to the count of  $f_2$  variants (or doubletons), i.e., those present exactly twice in a sample [41, 42]. We utilized  $F_2$  rather than  $f_2$  due to the wide range of inbreeding found across individuals. We utilized 2,384,354 autosomal  $F_2$  sites that have no missing genotypes. We predicted the total number of variants present in each breed that has at least three individuals based on the distribution of non-reference allele counts (the non-reference site frequency spectrum). Based on this distribution, we predicted the number of non-reference variants that would be discovered in a sample of 100 individuals of the same breed by applying a linear program method to the observed site frequency spectrum [45].

#### Breed genetic distance and haplotype sharing

Autosomal variation from the VQSR PASS set (26,486,238 SNVs) sourced from 1579 breed dogs, 57 wolves, and four coyotes were used as inputs. Genomic distance (1-IBS) was calculated in PLINK (v1.9) [134]. The distance matrix was transformed into a cladogram using *neighbor* in the PHYLIP suite of programs [135] and visualized with FigTree (v1.44, <http://tree.bio.ed.ac.uk/software/figtree>). To determine the significance of branch placement in the cladogram, the dataset was resampled 100 times by pulling a random 10% of the SNVs to make 100 distance matrices. The cladograms created from each of the random variant-set matrices were combined using *consense* in the PHYLIP suite of programs. Clades were defined as clusters of two or more breeds that share the same branch in  $> 65\%$  of samplings. We additionally made a comparison dataset by first randomly identifying one dog from each breed then removing SNVs with a linkage disequilibrium value of  $r^2 > 0.5$  within a 500-kb window leaving 3,106,329 SNVs. Bootstrapped distance matrices were created by randomly drawing 3.1 million SNVs from this dataset with replacement 100 times. Cladograms were created with *neighbor* and combined using *consense* (part of PHYLIP). The placement of samples relative to each other was assessed and individuals were removed from the breed analysis if they (i) did not cluster with the multi-breed clade that contained all other members of the same

breed; (ii) they were listed as an ambiguous breed; or (iii) they were part of a population that included first-generation hybrids. Shared haplotypes of at least 250 kb were estimated using BEAGLE v4.1 [136] and haplotypes with a  $\text{LOD} > 3.0$  were predicted to be identical-by-descent. The length of all shared segments was totaled for every pair of dogs and these totals were averaged within each breed, within each clade, and across clades. D-stats were calculated for German Shepherd-like breeds to assess wolf admixture. The R package *admixr* [137] was used to run Admixtools v7.0.2 [138] on the tree structure (W, X)(Y, Z) where W = German Shepherd Dog, Z = Coyote, X = list of German Shepherd-related breeds, and Y = list of wolf populations. Significance was set at  $|Z| \geq 3$ . Additional detail is provided in the Additional file 2: Sect. 3.

### Runs of homozygosity

Runs of homozygosity (ROH) for all samples were defined using the sliding-window approach implemented in PLINK v1.90b4.9 [134] with the “--homozyg” function. Settings were based on those previously recommended for high-density SNP datasets [139], with minimum average SNP density (--homozyg-density 50), maximum gap between adjacent SNPs (--homozyg-gap 1000), the size of the sliding window (--homozyg-kb 200), and the minimum number of variants needed to detect ROH (--homozyg-window-snp 100) set to reflect the average SNP density of the dataset. The “--homozyg-window-het” and “--homozyg-window-missing” flags were set to 3 and 2 respectively to account for potential sequencing errors and missing data. The number of heterozygous sites to allow within a window (--homozyg-window-het 3) was set based on the average number of heterozygous sites called in male dogs outside of the pseudoautosomal regions on chrX (i.e., where all males are haploid and therefore any heterozygous calls are errors). The coefficient of inbreeding was calculated from our ROH estimates ( $F_{\text{ROH}}$ ) by dividing the total length of all ROH within a sample by the genome size (i.e.,  $F_{\text{ROH}}$  is the proportion of the genome within ROH). The BEDTools v2.29.2 [140] *subtract* function was used to identify all regions in the genome that are absent of ROH across all samples. These were intersected with genome callability mask (BEDTools *intersect*).

### Imputation

The Dog10K reference panel was created using all 1929 samples and the VQSR PASS SNV VCF. Multiallelic and sites with missingness  $> 5\%$  were removed. SHAPEIT5 was used for phasing [141]. In total, 29,234,830 autosomal and 965,534 chrX SNVs are included. Public WGS were used to assess imputation outputs. The 10 samples are drawn from 5 breeds in, and 5 not in, the Dog10K collection (Additional file 1: Table S6). Each WGS was processed as described above. To represent low-pass WGS, alignment files were downsampled to  $1 \times$  coverage. To represent array genotypes, Axiom CanineHD Array sites (530,104 sites) and Illumina CanineHD BeadChip sites (134,037 sites) were first lifted to UU\_Cfam\_GSD\_1.0 (liftover [142]) and subsequently extracted from each WGS. Different methods were required to impute the three downsampled genotype methods. For low-pass WGS data, genotype likelihoods were calculated using bcftools v1.17 mpileup and call commands, followed by GLIMPSE v1.1.1 imputation of genotypes from genotype likelihoods [143]. For Axiom and Illumina array data, genotypes were phased using SHAPEIT5, rare allele MAF cutoff set to 0, and the Dog10K reference

panel for reference haplotypes followed by IMPUTE5 imputation of the phased array data [144]. Both genotype imputation tools were run with default parameters. For chrX non-PAR, males and females were imputed separately, with males run using the haploid settings of both imputation tools. Genotype imputation accuracy was measured as the non-reference genotype concordance (NRC) between imputed genotypes and WGS genotypes. NRC rates were assessed for site imputation quality (imputation info metric > 0.9), MAF within the reference panel, and genotype chromosomal context (autosome/PAR or X chromosome). Ten additional chr38 smaller reference panels were created for each of 500, 1000, and 1500 individuals by selecting samples at random from the full sample list of 1929 individuals. Chromosome 38 genotypes from the selected dogs were then extracted from the VQSR PASS VCF, processed, and phased in accordance with the methodology used to create the full Dog10K reference panel.

### Mitochondrial analyses

A modified GATK Mutect2 pipeline [145] was used to call mitochondrial variation. Read-pairs from the nuclear genome alignment process were extracted if (i) at least one read aligned to the UU\_Cfam\_GSD\_1.0 chrM sequence, or (ii) to a nuclear mitochondrial segment that is at least 300 bp long with at least 95% identity to the reference mitochondrial genome sequence. Extracted read-pairs were aligned to two linear references based on the NC\_002008.4 mitochondrial genome reference sequence. The first reference was identical to NC\_002008.4, the second is rotated to start at position 8000. Using two genome sequences compensates for the bias in the lower rate of alignment for reads derived from the ends of the linear sequence (bwa-MEM [146] v0.7.15). Read depth was calculated (GATK v4.2.5.0 *CollectHsMetrics*) and downsampled to a depth of 5000 (GATK v4.2.5.0 *DownsampleSam*). Mutect2 was used to identify candidate variants from each alignment with options `--mitochondria-mode`, `--max-reads-per-alignment-start 75`, `--max-mnp-distance 0`, and `--annotation StrandBiasBySample`. The resulting VCF was filtered with `GATK FilterMutectCalls --mitochondria-mode`. VCF files from both references are then merged, with variants in the first and last 4 kb taken from the alignment to the rotated reference. Sites where the most frequent alternative allele fails the `strand_bias` filter or represents a heteroplasmy (an allele fraction less than 0.5) were removed. Regions with a coverage less than 100 and regions that overlap positions 15,512–15,535 or 15,990 were masked to “N”. The accuracy of the mitochondrial variation discovery pipeline was assessed by comparing the mitochondrial sequence constructed from Illumina data to that reported in five published long-read canine genome assemblies. Assignment to mitochondrial haplogroups was performed based on similarity to previously defined samples [60]. More detail is provided in the Additional file 2: Sect. 6.

### Structural variation

CNVs were detected with the QuicK-mer2 [64] search command with default parameters ( $k=30$ , edit distance = 2, depth-threshold 100). Control regions for copy number and GC normalization were defined by excluding non-autosomal chromosomal sequence, regions that are duplicated in the genome assembly based on assembly self-alignment [147], reported CNVs [30], and regions with an elevated copy number identified in a preliminary analysis using fastCN [148]. Samples with a median absolute copy number estimate

deviation greater than 0.25 were excluded from the analysis. The paralog-specific copy number for each gene was estimated based on the median QuicK-mer2 estimate of intersecting windows for each sample. This analysis was limited to the 18,162 protein-coding genes that were fully encompassed by at least one k-mer window. Structural variants were identified with Manta v1.6.0 and default parameters [66]. Inversions were converted to event representation using the Manta convertInversion.py utility. Raw calls were merged using svimmer and genotyped across all samples using GraphTyper2 v2.7.2 with default parameters [67]. For break-end (BND), insertion (INS), deletion (DEL), and duplication (DUP) calls, the “AGGREGATED” genotyping model was used. For inversion (INV) candidates, the breakpoint model was used as reported by GraphTyper2. SVs were filtered for quality, depth, and allelic balance, with a maximum size of 10 Mb. Candidate retrogenes were identified by deletions that have a 99% reciprocal overlap with annotated introns (Additional file 2: Fig. S6). See Additional file 2: Sect. 7 for full details.

### Signatures of selection

Ohana [78] was used to detect signals of selection shared across nine dog breed groups we defined based on allele sharing patterns (See “Breed genetic distance and haplotype sharing”). The 790 individuals within the broader Spitz, Sighthounds, Waterdogs, Scenthounds, Pointers, Belgian Herder, UK Herding, Spaniel, and Mastiffs groups possess similar morphological traits (Additional file 1: Table S21, Additional file 2: Sect. 8). Only biallelic PASS SNVs with a minor allele frequency (MAF) of 5% and no missing data were considered (6,181,086 autosomal sites). Ohana ran with the number of ancestral components ranging from  $K=2$  up to  $K=11$ , with  $K=5$  selected as a compromise between low risk of overfitting and interpretability of component identity. The five inferred ancestral components were maximized for the following dog groups: Mastiffs, Scenthounds, Spitz, Pointers and Spaniels, and the Collie and Shetland Sheepdog. The log-likelihood ratio test statistic of Ohana’s *selscan* module was used to evaluate the likelihood of selection for each variant. Genomic control was carried out, and  $p$ -values were calibrated using a mixed chi-squared distribution with the *emdbook* R package (version 1.3.12) [149]. A 5% Bonferroni threshold for the number of sites analyzed was used as a significance threshold ( $-\log_{10}p=8.09$ ). The *intersect* function of BEDTools v2.30.0 [140] was used to identify genes overlapping or within 100 kb of the significant sites.

To refine the targets of selection, we applied iSAFE [93], a method to fine-map variants targeted by selection that does not rely on additional demographic information of the study populations or functional annotation of the mutations under focus. We applied iSAFE to each region, including flanking regions, setting the cases as the clade in which each ancestral component was maximized and using the remaining clades as controls. The loci on chr26 and chr38 were not analyzed due to their large size. Additional details are provided in the Additional file 2: Sect. 8.

### Variant concordance

The quality of our variant collection was assessed by comparing genotypes for 168 sequenced samples that were previously genotyped on the Illumina CanineHD array. Concordance was calculated for both sites retained in the VQSR PASS and strict filter sets (151,197 and 145,271 polymorphic sites respectively, Additional file 2: Sect. 10).

### SNV functional annotation

The annotated VCF catalog (See “Reference annotation”) was further filtered by sample category (Breed Dog And Other  $N=1591$ , Village Dog  $N=281$ , Wolf  $N=57$ ) and functional annotation with *snpSift* from snpEFF 4.3t [150]. SNV density was calculated in 100-kb bins for various sample categories, allele frequencies (vcftools 0.1.16 [151]: rare,  $AF \leq 1\%$ ; Intermediate,  $1\% < AF < 5\%$ ; common,  $AF \geq 5\%$ ) in both the coding and non-coding fractions of the genome (Additional file 2: Sect. 12).

### Comparison of public variation catalogs

The strict-filtered Dog10K dataset was compared to three other publically available datasets in multiple ways, (i) methods used to call variants within each set, (ii) sharing of individuals between sets, and (iii) sharing of breed types. The sets were strict-filtered Dog10K VCF (1929 samples, 28,725,482 SNVs), DBVDC (590 samples, 20,443,472 SNVs) [6], NIH (715 samples, 18,468,060 SNVs) [4], and EVA v3 (4,548,628 SNVs) ([http://ftp.ebi.ac.uk/pub/databases/eva/rs\\_releases/release\\_3/by\\_species/canis\\_lupus\\_familiaris/](http://ftp.ebi.ac.uk/pub/databases/eva/rs_releases/release_3/by_species/canis_lupus_familiaris/)). CanFam3.1 referenced datasets were lifted to UU\_Cfam\_GSD\_1.0 coordinates, with variants on unplaced scaffolds excluded from further analysis. The full NIH panel contains multiple canid outgroups (Additional file 1: Table S12). These were removed, allowing for the comparison of positions variable in dogs and wolves. For (i) the methods and filters used to call variants was tabulated, and due to this variability, for (ii) individuals were considered shared between datasets if their proportion of IBD was in excess of that observed for the closest pair in the Dog10K dataset (i.e., PLINK (v1.9) [134] PiHAT  $> 0.9451$  based on 145,845 random SNVs). For (iii) breed types, breed names and descriptors were harmonized and compared across sets.

### Fraction of theoretical variation discovered

The fraction of possible variants captured by the strict filter set was calculated by first summing the number of each base contained in the callable region of UU\_Cfam\_GSD\_1.0. In this analysis, complementary bases were combined, i.e., C and G, and T and A. Observed base changes were extracted using the strict-filtered VCF with BCFtools *stats* function [152]. Calculations were performed separately for the autosomes and chrX.

### Intersection with OMIA database

Variant information for 463 published likely causative variants for canine inherited traits and diseases were downloaded as a CSV file from OMIA [96] (omia.org; March 2022). The analysis was restricted to SNVs, and small indels spanning less than 20 consecutive nucleotides, leaving 352 “small” variants. Positions were lifted from CanFam3.1 to UU\_Cfam\_GSD\_1.0 using the chain file downloaded from UCSC (<https://hgdownload.soe.ucsc.edu/goldenPath/canFam4/liftOver/>). The lifted positions were extracted from the functional Dog10K dataset using BCFtools *isec* (samtools version 1.10) [152]. The resulting file was manually curated, genotype distributions were tabulated, and the OMIA traits for the identified variants were annotated.

### Variants affecting metabolism and drug targets

The Tier 1 of 1427 human druggable target genes [99] was downloaded and where matching with a UU\_Cfam\_GSD\_1.0 referenced NCBI annotation (“Reference annotation”) taken forward for analysis. Tier 1 includes gene targets of approved small molecules or biotherapeutic drugs, as well as clinical-phase drug candidates from the time of publication. The Tier 1 gene space was intersected with high effect coding variants from the “SNV functional annotation” and CNVs from the Quick-mer2 [64] “Structural variation” analysis. Only genes completely covered by a Quick-mer2 window were considered. See Additional file 2: Sect. 12.

### SNV deviations from HWE

Using the VQSR PASS VCF as an input, deviations from HWE were determined for each biallelic, missense, or loss of function variant. Calculations were based on the chi-square distribution at a Bonferroni adjusted  $p$ -value  $< 0.05$  (R v4.2.0). For each of 42 HWE candidate positions, variant site statistics (BaseQRankSum, FS, MQ, MQRankSum, QD, ReadPosRankSum, SOR, and VQSLOD) and genotype statistics (depth, allele depth, and phred-scaled genotype likelihoods) were extracted using the *vcfR* package [153]. Each genotype statistic was analyzed according to its assigned genotype of either reference or alternate. Based on quality scores and potential biological interest, the relevant alignment files of seven variants were selected for additional visual analysis (IGV v2.10.0 [154]). In addition, reads containing the variant of interest were mapped to the ROS\_Cfam\_1.0, UMICH\_Zoey\_3.1, UNSW\_CanFamBas\_1.0, UU\_Cfam\_GSD\_1.0, and Dog10K\_Boxer\_Tasha long-read assemblies using the NCBI blast tool to identify potential alternative causes of HWE deviation.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-023-03023-7>.

**Additional file 1: Table S1.** Sample metadata and analyses where used. **Table S2.** Samples with high F2 allele sharing with wolves. **Table S3.** Variation yet to be discovered based on 100 samples per breed. **Table S4.** Breed group placement of each sample. **Table S5.** Runs of homozygosity (ROH) statistics per sample. **Table S6.** Publicly available test samples used to measure imputation accuracy. **Table S7.** Copy number variable genes. **Table S8.** Potential retrogenes. **Table S9.** Candidate genes either overlapping or within a 100kb distance of a significant site for each targeted ancestry. **Table S10.** iSAFE top 10 sites per selection signature and ancestry component. **Table S11.** Observed fraction of the theoretically possible SNVs. **Table S12.** Alignment and filtering strategies for three panels of normal variation. **Table S13.** Breed categories included in three panels of normal variation. **Table S14.** Summary of variant counts in Dog10K sample sets. **Table S15.** Distribution of SNVs across functional classes and Dog10K sample sets. **Table S16.** Genotypes observed for 76 OMIA variants. **Table S17.** Allele frequency distributions detected in Dog10K for OMIA categories. **Table S18.** Variant positions deviating from HWE. **Table S19.** Animal protocols, approving board, and date of approval. **Table S20.** Summary of duplicate genes within NCBI release 106. **Table S21.** Samples used in Ohana selection analysis.

**Additional file 2:** Supplementary Methods [165–182]. **Fig. S1.** Imputation accuracy of individual samples for sites with MAF > 1%. **Fig. S2.** Median copy-number across the genome for wolves. **Fig. S3.** Repeatmasker classification of SINE variation. **Fig. S4.** Distribution of variation across the genome for breed and other dogs ( $n=1,591$ ). **Fig. S5.** Nucleotide diversity along chr26. **Fig. S6.** Signature of a retrogene detected at the *TEX2* locus.

**Additional file 3.** Review history.

### Acknowledgements

We thank Jeffrey J. Schoenebeck (University of Edinburgh), for access to chromosome Y contigs and Michael Dong (Uppsala University) for access to CanFam3.1 referenced Zoonomia phyloP scores. Computation and data handling were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX projects, SNIC 2021/5-296, SNIC 2021/6-208), partially funded by the Swedish Research Council through grant agreement no. 2018-05973, CSC—IT center for science Ltd., Espoo, Finland, and the Interfaculty Bioinformatics Unit of the University of Bern. We thank the University of Michigan Advanced

Research Computing Group for assistance with data transfer and processing. We thank several people for DNA isolation and sample handling: Jessica Hale and Andrew Hogan from NHGRI; Richard Allen from Oxford University; Nadine Botherel, Edouard Cadieu, Benoit Hédan, Laetitia Lagoutte from the Cani-DNA BRC from the University of Rennes; Sini Karjalainen and Ileana Quintero from the University of Helsinki; Paul Maza from Cornell University School of Veterinary Medicine; Bruce Elfstrom, Doug Lally, Dietrich Gotzke, Tom Asmus from Cornell University. We thank Henrique Carvalho from the Portuguese Nature and Forests Conservation Institute for providing access to Iberian wolf samples from the tissue bank Sistema de Monitorização de Lobos Mortos (SMLM). We thank Elinor Karlsson from the University of Massachusetts Chan Medical School for help discussions. We also thank our many colleagues who provided additional samples for this project as well as guidance and suggestions, particularly Cathryn Mellersh from the Kennel Club Genetics Centre of the University of Cambridge. Finally, we thank the thousands of canine owners who provided DNA samples for this project.

#### Review history

The review history is available as Additional File 3.

#### Peer review information

Anahita Bishop was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

#### Authors' contributions

RKW, EAO, GDW, and YPZ conceived the study, and EAO, JRSM, JMK, GDW, and GL supervised the project. JRSM, MA, VJ, AN, CW, CH, BWD, AH, RMB, JEN, JMK, and MB processed and analyzed the raw sequencing data, including annotation and variant filters; HGP analyzed the haplotype sharing and breed group membership; RMB created the imputation reference panel and analyzed imputation accuracy; MJC analyzed the runs of homozygosity; KB, LAF, and FR analyzed the signatures of selection; GL, FR-A, JRSM, PS, G-D W, and JMK analyzed the mitochondrial data; AKN, PZS, and JMK processed and analyzed the structural variant data; JEN, VJ, CH, RMB, Y-HL, MJC, CW, MB, and JRSM analyzed the functional consequence of the SNV data; VJ, IT, FWN, and TL analyzed the OMIA dataset; AKN, PZS, JMK, EAO, and JRSM analyzed the metabolism and drug targets. BVH, CA, ARB, MD, HL, IG, KG, PL, PN, VS, CG, KKA, AEP, and AT acquired and curated samples. KL-T, JMK, JRSM, HL, TL, and EAO provided compute resources and SH, X-YF, and MKH assisted with analysis. JRSM, JMK, and EAO wrote the initial manuscript draft. HL, GL, MC, RKW, MB, KL-T, CH, AEP, KB, and TL provided edits. All authors read and approved the final manuscript.

#### Funding

Funding is acknowledged by Y-PZ from National Science and Technology Innovation 2030 Major Project of China (2021ZD0203900) and the National Key R&D Program of China (2019YFA0707101), JMK from NIH Grant R01GM140135; EAO, HGP, RB, and ACH from the Intramural program of the National Human Genome Research Institute of the National Institutes of Health, USA; HL, JES, MA, and MKH from the Jane and Aatos Erkko Foundation, Helsinki, Finland; LAFF and GL from the European Research Council grants (ERC-2013-StG-337574-UNDEAD and ERC-2019-StG-853272-PALAEOFARM); IT and FWN from the Ronald Bruce Anstee Bequest to the Sydney School of Veterinary Science, Sydney, Australia; AB from NIH grant R24 GM082910 and the Cornell University College of Veterinary Medicine; CH and CA by the CRB-Anim project, PIA1, INBS Infrastructure in Biology Health ANR-11-INBS-0003 (2012–2022) and from the CNRS and the University of Rennes; KL-T from the Swedish Research Council; AEP and CG from PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF); KKA from the Bali Animal Welfare Association; Y.-H.L. and G.-D.W. are supported by the Youth Innovation Promotion Association, Chinese Academy of Sciences. This work was supported by the Animal Branch of the Germplasm Bank of Wild Species, Chinese Academy of Sciences (the Large Research Infrastructure Funding).

#### Availability of data and materials

Raw sequence data is available from the SRA under accessions PRJNA648123 [155] and PRJNA188158 [156] and are listed in Additional file 1: Table S1. SNV and SV variants have been deposited to the European Variation Archive (PRJEB62420) [157] and mitochondrial genomes are available in GenBank (accessions OQ339232-OQ341164). Variant files and associated annotations are available at <https://kiddlabshare.med.umich.edu/dog10K/> and at <https://zenodo.org/record/8084059> [158]. Code to perform genome alignment, variant calling, and mitochondrial processing is available under the MIT Open Access License at <https://github.com/jmkidd/dogmap> [159], <https://github.com/jmkidd/doggenotype> [160], and <https://github.com/jmkidd/callmito> [161]. Archival versions are available under the MIT Open Access License at <https://zenodo.org/record/8087879> [162], <https://zenodo.org/record/8087891> [163] and <https://zenodo.org/record/8087897> [164].

#### Declarations

##### Ethics approval and consent to participate

All samples were collected and processed with Institutional approval from the participating organization (Additional file 1: Table S19).

##### Competing interests

BWD is Director of Veterinary Research at Volition Veterinary LLC; ARB is a Co-Founder and CSO of Embark Veterinary Inc. and serves on the Board of EMBARK and the Morris Animal Foundation. HL has consulted with Wisdom Panel Kinship, is an owner and chairman of the board of Petbiomics Ltd., and Petmeta Labs Ltd, and is an owner and advisor of DeepScan Diagnostics Ltd. IT is the curator of Online Mendelian Inheritance which is supported by the University of Sydney.

**Author details**

<sup>1</sup>Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, 75132 Uppsala, Sweden. <sup>2</sup>Department of Human Genetics, University of Michigan, Ann Arbor, MI 48107, USA. <sup>3</sup>State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China. <sup>4</sup>National Human Genome Research Institute, National Institutes of Health, 50 South Drive, Building 50 Room 5351, Bethesda, MD 20892, USA. <sup>5</sup>Section for Molecular Ecology and Evolution, Globe Institute, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen, Denmark. <sup>6</sup>University of Rennes, CNRS, Institute Genetics and Development Rennes - UMR6290, 35000 Rennes, France. <sup>7</sup>Institute of Genetics, Vetsuisse Faculty, University of Bern, 3001 Bern, Switzerland. <sup>8</sup>Department of Medical and Clinical Genetics, Department of Veterinary Biosciences, University of Helsinki and Folkhälsan Research Center, 02900 Helsinki, Finland. <sup>9</sup>School of Biological and Behavioural Sciences, Queen Mary University of London, London E14NS, UK and Palaeogenomics Group, Department of Veterinary Sciences, Ludwig Maximilian University, D-80539 Munich, Germany. <sup>10</sup>Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. <sup>11</sup>BIOPOLIS-CIBIO-InBIO-Centro de Investigação Em Biodiversidade E Recursos Genéticos - ArchGen Group, Universidade Do Porto, 4485-661 Vairão, Portugal. <sup>12</sup>Department of Public Health, Udayana University, Bali 80361, Indonesia. <sup>13</sup>Department of Biomedical Sciences, Cornell University, 930 Campus Road, Ithaca, NY 14853, USA. <sup>14</sup>Department of Veterinary Integrative Biosciences, School of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX 77843, USA. <sup>15</sup>Department of Genetics, School of Biology, Aristotle University of Thessaloniki, Thessaloniki, Macedonia 54124, Greece and Genomics and Epigenomics Translational Research (GENeTres), Center for Interdisciplinary Research and Innovation (CIRI-AUTH, Balkan Center, Thessaloniki, Greece. <sup>16</sup>NGO "Callisto", Wildlife and Nature Conservation Society, 54621 Thessaloniki, Greece. <sup>17</sup>Natural History Museum of Crete & Department of Biology, University of Crete, 71202 Irakleio, Greece. <sup>18</sup>Biology Department, School of Sciences and Engineering, University of Crete, Heraklion, Greece. <sup>19</sup>Palaeogenomics and Evolutionary Genetics Lab, Institute of Molecular Biology and Biotechnology (IMBB), Foundation for Research and Technology - Hellas (FORTH), Heraklion, Greece. <sup>20</sup>Department of Gene Technology, Science for Life Laboratory, KTH - Royal Institute of Technology, 17121 Solna, Sweden. <sup>21</sup>Department of Genetics, School of Biology, Aristotle University of Thessaloniki, 54124 Thessaloniki, Macedonia, Greece. <sup>22</sup>Sydney School of Veterinary Science, The University of Sydney, Sydney, NSW 2570, Australia. <sup>23</sup>Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA. <sup>24</sup>Department of Ecology and Evolutionary Biology, Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095-7246, USA. <sup>25</sup>Palaeogenomics and Bio-Archaeology Research Network, School of Archaeology, University of Oxford, Oxford OX1 3TG, UK.

Received: 23 November 2022 Accepted: 25 July 2023

Published online: 15 August 2023

**References**

- Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ 3rd, Zody MC, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*. 2005;438:803–19.
- Hoepfner MP, Lundquist A, Pirun M, Meadows JR, Zamani N, Johnson J, Sundstrom G, Cook A, FitzGerald MG, Swofford R, et al. An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. *PLoS ONE*. 2014;9:e91172.
- Megquier K, Genereux DP, Hekman J, Swofford R, Turner-Maier J, Johnson J, Alonso J, Li X, Morrill K, Anguish LJ, et al. BarkBase: epigenomic annotation of canine genomes. *Genes (Basel)*. 2019;10(6):433.
- 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*. 2019;10:1489.
- Serres-Armero A, Davis BW, Povolotskaya IS, Morcillo-Suarez C, Plassais J, Juan D, Ostrander EA, Marques-Bonet T. Copy number variation underlies complex phenotypes in domestic dog breeds and other canids. *Genome Res*. 2021;31:762–74.
- Jagannathan V, Drogemuller C, Leeb T. Dog Biomedical Variant Database C: a comprehensive biomedical variant catalogue based on whole genome sequences of 582 dogs and eight wolves. *Anim Genet*. 2019;50:695–704.
- Wucher V, Legeai F, Hedan B, Rizk G, Lagoutte L, Leeb T, Jagannathan V, Cadieu E, David A, Lohi H, et al. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res*. 2017;45:e57.
- Moura E, Tasqueti UI, Mangrich-Rocha RMV, Engracia Filho JR, Farias MR, Pimpao CT. Inborn errors of metabolism in dogs: historical, metabolic, genetic, and clinical aspects. *Top Companion Anim Med*. 2022;51:100731.
- Karlsson EK, Baranowska I, Wade CM, Salmon Hillbertz NH, Zody MC, Anderson N, Biagi TM, Patterson N, Pielberg GR, Kulbokas EJ 3rd, et al. Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat Genet*. 2007;39:1321–8.
- Witt KE, Judd K, Kitchen A, Grier C, Kohler TA, Ortman SG, Kemp BM, Malhi RS. DNA analysis of ancient dogs of the Americas: identifying possible founding haplotypes and reconstructing population histories. *J Hum Evol*. 2015;79:105–18.
- Bergstrom A, Frantz L, Schmidt R, Ersmark E, Lebrasseur O, Girdland-Flink L, Lin AT, Stora J, Sjogren KG, Anthony D, et al. Origins and genetic legacy of prehistoric dogs. *Science*. 2020;370:557–64.
- Perri AR, Feuerborn TR, Frantz LAF, Larson G, Malhi RS, Meltzer DJ, Witt KE. Dog domestication and the dual dispersal of people and dogs into the Americas. *Proc Natl Acad Sci U S A*. 2021;118(6):e2010083118.
- Morrill K, Hekman J, Li X, McClure J, Logan B, Goodman L, Gao M, Dong Y, Alonso M, Carmichael E, et al. Ancestry-inclusive dog genomics challenges popular breed stereotypes. *Science*. 2022;376:eabk0639.

14. Zapata I, Eyre AW, Alvarez CE, Serpell JA. Latent class analysis of behavior across dog breeds reveal underlying temperament profiles. *Sci Rep.* 2022;12:15627.
15. Horvath S, Lu AT, Haghani A, Zoller JA, Li CZ, Lim AR, Brooke RT, Raj K, Serres-Armero A, Dreger DL, et al. DNA methylation clocks for dogs and humans. *Proc Natl Acad Sci U S A.* 2022;119:e2120887119.
16. Yarborough S, Fitzpatrick A, Schwartz SM. Dog Aging Project C: Evaluation of cognitive function in the Dog Aging Project: associations with baseline canine characteristics. *Sci Rep.* 2022;12:13316.
17. Leeb T, Bannasch D, Schoenebeck JJ. Identification of genetic risk factors for monogenic and complex canine diseases. *Annu Rev Anim Biosci.* 2023;11:183–205.
18. Boyko AR, Boyko RH, Boyko CM, Parker HG, Castelhanos M, Corey L, Degenhardt JD, Auton A, Hedimbi M, Kityo R, et al. Complex population structure in African village dogs and its implications for inferring dog domestication history. *Proc Natl Acad Sci U S A.* 2009;106:13903–8.
19. Broeckx BJG, Derrien T, Mottier S, Wucher V, Cadieu E, Hedan B, Le Beguec C, Bothereil N, Lindblad-Toh K, Saunders JH, et al. An exome sequencing based approach for genome-wide association studies in the dog. *Sci Rep.* 2017;7:15680.
20. Chavez DE, Gronau I, Hains T, Dikow RB, Frandsen PB, Figueiro HV, Garcez FS, Tchaicka L, de Paula RC, Rodrigues FHG, et al. Comparative genomics uncovers the evolutionary history, demography, and molecular adaptations of South American canids. *Proc Natl Acad Sci U S A.* 2022;119:e2205986119.
21. Bergstrom A, Stanton DWG, Taron UH, Frantz L, Sinding MS, Ersmark E, Pfrengle S, Cassatt-Johnstone M, Lebrasseur O, Girdland-Flink L, et al. Grey wolf genomic history reveals a dual ancestry of dogs. *Nature.* 2022;607:313–20.
22. Skoglund P, Ersmark E, Palkopoulou E, Dalen L. Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Curr Biol.* 2015;25:1515–9.
23. Frantz LA, Mullin VE, Pionnier-Capitan M, Lebrasseur O, Ollivier M, Perri A, Linderholm A, Mattiangeli V, Teasdale MD, Dimopoulos EA, et al. Genomic and archaeological evidence suggest a dual origin of domestic dogs. *Science.* 2016;352:1228–31.
24. Botigue LR, Song S, Scheu A, Gopalan S, Pendleton AL, Oetjens M, Taravella AM, Seregely T, Zeeb-Lanz A, Arbogast RM, et al. Ancient European dog genomes reveal continuity since the Early Neolithic. *Nat Commun.* 2017;8:16082.
25. Ostrander EA, Wang GD, Larson G, vonHoldt BM, Davis BW, Jagannathan V, Hitte C, Wayne RK, Zhang YP, Dog KC. Dog10K: an international sequencing effort to advance studies of canine domestication, phenotypes and health. *Natl Sci Rev.* 2019;6:810–24.
26. Geiger M, Schoenebeck JJ, Schneider RA, Schmidt MJ, Fischer MS, Sanchez-Villagra MR. Exceptional changes in skeletal anatomy under domestication: the case of brachycephaly. *Integr Org Biol.* 2021;3:obab023.
27. Bannasch DL, Baes CF, Leeb T. Genetic variants affecting skeletal morphology in domestic dogs. *Trends Genet.* 2020;36:598–609.
28. Edwards RJ, Field MA, Ferguson JM, Dudchenko O, Keilwagen J, Rosen BD, Johnson GS, Rice ES, Hillier D, Hammond JM, et al. Chromosome-length genome assembly and structural variations of the primal Basenji dog (*Canis lupus familiaris*) genome. *BMC Genomics.* 2021;22:188.
29. Halo JV, Pendleton AL, Shen F, Doucet AJ, Derrien T, Hitte C, Kirby LE, Myers B, Sliwerska E, Emery S, et al. 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 U S A.* 2021;118(11):e2016274118.
30. Wang C, Wallerman O, Arendt ML, Sundstrom E, Karlsson A, Nordin J, Makelainen S, Pielberg GR, Hanson J, Ohlsson A, et al. A novel canine reference genome resolves genomic architecture and uncovers transcript complexity. *Commun Biol.* 2021;4:185.
31. Field MA, Rosen BD, Dudchenko O, Chan EKF, Minoche AE, Edwards RJ, Barton K, Lyons RJ, Tuipulotu DE, Hayes VM, et al. Canfam\_GSD: De novo chromosome-length genome assembly of the German Shepherd Dog (*Canis lupus familiaris*) using a combination of long reads, optical mapping, and Hi-C. *Gigascience.* 2020;9(4):giaa027.
32. Player RA, Forsyth ER, Verratti KJ, Mohr DW, Scott AF, Bradburne CE. A novel *canis lupus familiaris* reference genome improves variant resolution for use in breed-specific GWAS. *Life Sci Alliance.* 2021;4(4):e202000902.
33. Jagannathan V, Hitte C, Kidd JM, Masterson P, Murphy TD, Emery S, Davis B, Buckley RM, Liu YH, Zhang XQ, et al. Dog10K\_Boxer\_Tasha\_1.0: a long-read assembly of the dog reference genome. *Genes (Basel).* 2021;12(6):847.
34. Field MA, Yadav S, Dudchenko O, Esvaran M, Rosen BD, Skvortsova K, Edwards RJ, Keilwagen J, Cochran BJ, Manandhar B, et al. The Australian dingo is an early offshoot of modern breed dogs. *Sci Adv.* 2022;8:eabm5944.
35. Sinding MS, Gopalakrishnan S, Raundrup K, Dalen L, Threlfall J. Darwin Tree of Life Barcoding c, Wellcome Sanger Institute Tree of Life p, Wellcome Sanger Institute Scientific Operations DNAPc, Tree of Life Core Informatics c, Darwin Tree of Life C, Gilbert T: The genome sequence of the grey wolf, *Canis lupus Linnaeus 1758*. *Wellcome Open Res.* 2021;6:310.
36. Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, Chang X, Seaman JD, Rounthwaite R, Ebler J, et al. Pangenome graphs. *Annu Rev Genomics Hum Genet.* 2020;21:139–62.
37. Regier AA, Farjoun Y, Larson DE, Krasheninina O, Kang HM, Howrigan DP, Chen BJ, Kher M, Banks E, Ames DC, et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat Commun.* 2018;9:4038.
38. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 2018.
39. Vasimuddin M, Misra S, Li H, Aluru S: Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS); 20–24 May 2019. 2019: 314–324.
40. The ISSA Outcrossing Plan [<https://shiloh-shepherd.com/pages/outcross2.html>]
41. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. *Nature.* 2012;491:56–65.
42. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. *Nature.* 2015;526:68–74.

43. Caniglia R, Fabbri E, Hulva P, Bolfikova BC, Jindrichova M, Stronen AV, Dykyy I, Camatta A, Carnier P, Randi E, Galaverni M. Wolf outside, dog inside? The genomic make-up of the Czechoslovakian Wolfdog. *BMC Genomics*. 2018;19:533.
44. Moravcikova N, Kasarda R, Zidek R, Vostry L, Vostra-Vydrova H, Vasek J, Cilova D. Czechoslovakian Wolfdog Genomic Divergence from Its Ancestors *Canis lupus*, German Shepherd Dog, and Different Sheepdogs of European Origin. *Genes (Basel)*. 2021;12(6):832.
45. Gravel S, National Heart L. Blood Institute GOESP: predicting discovery rates of genomic features. *Genetics*. 2014;197:601–10.
46. Pfahler S, Distl O. Effective population size, extended linkage disequilibrium and signatures of selection in the rare dog breed lundehund. *PLoS ONE*. 2015;10:e0122680.
47. Kettunen A, Daverdin M, Helfjord T, Berg P. Cross-breeding is inevitable to conserve the highly inbred population of puffin hunter: the Norwegian lundehund. *PLoS ONE*. 2017;12:e0170039.
48. Melis C, Borg AA, Espelien IS, Jensen H. Low neutral genetic variability in a specialist puffin hunter: the Norwegian Lundehund. *Anim Genet*. 2013;44:348–51.
49. Parker HG, Dreger DL, Rimbault M, Davis BW, Mullen AB, Carpintero-Ramirez G, Ostrander EA. Genomic analyses reveal the influence of geographic origin, migration, and hybridization on modern dog breed development. *Cell Rep*. 2017;19:697–708.
50. Marsden CD, Ortega-Del Vecchyo D, O'Brien DP, Taylor JF, Ramirez O, Vila C, Marques-Bonet T, Schnabel RD, Wayne RK, Lohmueller KE. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proc Natl Acad Sci U S A*. 2016;113:152–7.
51. Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva PM, Galaverni M, Fan Z, Marx P, Lorente-Galdos B, et al. Genome sequencing highlights the dynamic early history of dogs. *PLoS Genet*. 2014;10:e1004016.
52. Mooney JA, Yohannes A, Lohmueller KE. The impact of identity by descent on fitness and disease in dogs. *Proc Natl Acad Sci U S A*. 2021;118(16):e2019116118.
53. Savolainen P, Rosen B, Holmberg A, Leitner T, Uhlen M, Lundeberg J. Sequence analysis of domestic dog mitochondrial DNA for forensic use. *J Forensic Sci*. 1997;42:593–600.
54. Arenas M, Pereira F, Oliveira M, Pinto N, Lopes AM, Gomes V, Carracedo A, Amorim A. Forensic genetics and genomics: much more than just a human affair. *PLoS Genet*. 2017;13:e1006960.
55. Barrientos LS, Crespi JA, Fameli A, Posik DM, Morales H, Peral Garcia P, Giovambattista G. DNA profile of dog feces as evidence to solve a homicide. *Leg Med (Tokyo)*. 2016;22:54–7.
56. Clarke M, Vandenberg N. Dog attack: the application of canine DNA profiling in forensic casework. *Forensic Sci Med Pathol*. 2010;6:151–7.
57. Thalmann O, Shapiro B, Cui P, Schuenemann VJ, Sawyer SK, Greenfield DL, Germonpre MB, Sablin MV, Lopez-Giraldes F, Domingo-Roura X, et al. Complete mitochondrial genomes of ancient canids suggest a European origin of domestic dogs. *Science*. 2013;342:871–4.
58. Verginelli F, Capelli C, Coia V, Musiani M, Falchetti M, Ottini L, Palmirotta R, Tagliacozzo A, De Grossi MI, Mariani-Costantini R. Mitochondrial DNA from prehistoric canids highlights relationships between dogs and South-East European wolves. *Mol Biol Evol*. 2005;22:2541–51.
59. Savolainen P, Zhang YP, Luo J, Lundeberg J, Leitner T. Genetic evidence for an East Asian origin of domestic dogs. *Science*. 2002;298:1610–3.
60. Pang JF, Kluetsch C, Zou XJ, Zhang AB, Luo LY, Angleby H, Ardalan A, Ekstrom C, Skolleremo A, Lundeberg J, et al. mtDNA data indicate a single origin for dogs south of Yangtze River, less than 16,300 years ago, from numerous wolves. *Mol Biol Evol*. 2009;26:2849–64.
61. Duleba A, Skonieczna K, Bogdanowicz W, Malyarchuk B, Grzybowski T. Complete mitochondrial genome database and standardized classification system for *Canis lupus familiaris*. *Forensic Sci Int Genet*. 2015;19:123–9.
62. Hujuel MLA, Sherman MA, Barton AR, Mukamel RE, Sankaran VG, Terao C, Loh PR. Influences of rare copy-number variation on human complex traits. *Cell*. 2022;185:4233–4248 e4227.
63. Zhang F, Gu W, Hurler ME, Lupski JR. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet*. 2009;10:451–81.
64. Shen F, Kidd JM. Rapid, Paralog-Sensitive CNV Analysis of 2457 Human Genomes Using QuickK-mer2. *Genes (Basel)*. 2020;11(2):141.
65. Axelsson E, Ratnakumar A, Arendt ML, 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*. 2013;495:360–4.
66. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Kallberg M, Cox AJ, Kruglyak S, Saunders CT. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;32:1220–2.
67. Eggertsson HP, Kristmundsdottir S, Beyter D, Jonsson H, Skuladottir A, Hardarson MT, Gudbjartsson DF, Stefansson K, Halldorsson BV, Melsted P. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat Commun*. 2019;10:5402.
68. Parker HG, VonHoldt BM, Quignon P, Margulies EH, Shao S, Mosher DS, Spady TC, Elkahoulou A, Cargill M, Jones PG, et al. An expressed *fgf4* retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science*. 2009;325:995–8.
69. Brown EA, Dickinson PJ, Mansour T, Sturges BK, Aguilar M, Young AE, Korff C, Lind J, Ettinger CL, Varon S, et al. FGF4 retrogene on CFA12 is responsible for chondrodystrophy and intervertebral disc disease in dogs. *Proc Natl Acad Sci U S A*. 2017;114:11476–81.
70. Bannasch D, Batchler K, Leuthard F, Bannasch M, Hug P, Marcellin-Little DJ, Dickinson PJ, Drogemuller M, Drogemuller C, Leeb T. The Effects of FGF4 Retrogenes on Canine Morphology. *Genes (Basel)*. 2022;13(2):325.
71. Salmon Hillbertz NH, Isaksson M, Karlsson EK, Hellmen E, Pielberg GR, Savolainen P, Wade CM, von Euler H, Gustafson U, Hedhammar A, et al. Duplication of FGF3, FGF4, FGF19 and ORAOV1 causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. *Nat Genet*. 2007;39:1318–20.

72. Batchner K, Varney S, York D, Blacksmith M, Kidd JM, Rebhun R, Dickinson P, Bannasch D. Recent, full-length gene retrocopies are common in canids. *Genome Res.* 2022;32:1602–11.
73. Gao X, Li Y, Adetula AA, Wu Y, Chen H. Analysis of new retrogenes provides insight into dog adaptive evolution. *Ecol Evol.* 2019;9:11185–97.
74. Jurka J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci U S A.* 1997;94:1872–7.
75. Esnault C, Maestre J, Heidmann T. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet.* 2000;24:363–7.
76. Dewannieux M, Esnault C, Heidmann T. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet.* 2003;35:41–8.
77. Wang W, Kirkness EF. Short interspersed elements (SINEs) are a major source of canine genomic diversity. *Genome Res.* 2005;15:1798–808.
78. Cheng JY, Stern AJ, Racimo F, Nielsen R. Detecting selection in multiple populations by modeling ancestral admixture components. *Mol Biol Evol.* 2022;39(1):msab294.
79. Kerns JA, Newton J, Berryere TG, Rubin EM, Cheng JF, Schmutz SM, Barsh GS. Characterization of the dog Agouti gene and a nonagouti mutation in German Shepherd Dogs. *Mamm Genome.* 2004;15:798–808.
80. Berryere TG, Kerns JA, Barsh GS, Schmutz SM. Association of an Agouti allele with fawn or sable coat color in domestic dogs. *Mamm Genome.* 2005;16:262–72.
81. Bannasch DL, Kaelin CB, Letko A, Loechel R, Hug P, Jagannathan V, Henkel J, Roosje P, Hytonen MK, Lohi H, et al. Dog colour patterns explained by modular promoters of ancient canid origin. *Nat Ecol Evol.* 2021;5:1415–23.
82. Candille SI, Kaelin CB, Cattanaach BM, Yu B, Thompson DA, Nix MA, Kerns JA, Schmutz SM, Millhauser GL, Barsh GS. A -defensin mutation causes black coat color in domestic dogs. *Science.* 2007;318:1418–23.
83. 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.* 2007;176:1679–89.
84. Cubaynes S, Brandell EE, Stahler DR, Smith DW, Almberg ES, Schindler S, Wayne RK, Dobson AP, vonHoldt BM, MacNulty DR, et al. Disease outbreaks select for mate choice and coat color in wolves. *Science.* 2022;378:300–3.
85. Quilez J, Short AD, Martinez V, Kennedy LJ, Ollier W, Sanchez A, Altet L, Francino O. A selective sweep of >8 Mb on chromosome 26 in the Boxer genome. *BMC Genomics.* 2011;12:339.
86. Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, Sigurdsson S, Fall T, Seppala EH, Hansen MS, Lawley CT, et al. Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS Genet.* 2011;7:e1002316.
87. Chen E, Mullally A. How does JAK2V617F contribute to the pathogenesis of myeloproliferative neoplasms? *Hematology Am Soc Hematol Educ Program.* 2014;2014:268–76.
88. Beurlet S, Krief P, Sansonetti A, Briend-Marchal A, Kiladjian JJ, Padua RA, Chomienne C, Cassinat B. Identification of JAK2 mutations in canine primary polycythemia. *Exp Hematol.* 2011;39:542–5.
89. Webster MT, Kamgari N, Perloski M, Hoepfner MP, Axelsson E, Hedhammar A, Pielberg G, Lindblad-Toh K. Linked genetic variants on chromosome 10 control ear morphology and body mass among dog breeds. *BMC Genomics.* 2015;16:474.
90. Rimbault M, Beale HC, Schoenebeck JJ, Hoopes BC, Allen JJ, Kilroy-Glynn P, Wayne RK, Sutter NB, Ostrander EA. Derived variants at six genes explain nearly half of size reduction in dog breeds. *Genome Res.* 2013;23:1985–95.
91. Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson BV, Zusmanovich P, Sulem P, Thorlacius S, Gylfason A, Steinberg S, et al. Many sequence variants affecting diversity of adult human height. *Nat Genet.* 2008;40:609–15.
92. Lamichhane S, Han F, Berglund J, Wang C, Almen MS, Webster MT, Grant BR, Grant PR, Andersson L. A beak size locus in Darwin's finches facilitated character displacement during a drought. *Science.* 2016;352:470–4.
93. Akbari A, Vitti JJ, Iranmehr A, Bakhtiari M, Sabeti PC, Mirarab S, Bafna V. Identifying the favored mutation in a positive selective sweep. *Nat Methods.* 2018;15:279–82.
94. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science.* 2012;335:823–8.
95. Zoonomia C. A comparative genomics multitool for scientific discovery and conservation. *Nature.* 2020;587:240–5.
96. Nicholas FW. Online Mendelian Inheritance in Animals (OMIA): a record of advances in animal genetics, freely available on the Internet for 25 years. *Anim Genet.* 2021;52:3–9.
97. Kiener S, Kehl A, Loechel R, Langbein-Detsch I, Muller E, Bannasch D, Jagannathan V, Leeb T. Novel brown coat color (Cocoa) in French bulldogs results from a nonsense variant in HPS3. *Genes (Basel).* 2020;11(6):636.
98. Abramovs N, Brass A, Tassabehji M. Hardy-Weinberg equilibrium in the large scale genomic sequencing era. *Front Genet.* 2020;11:210.
99. Finan C, Gaulton A, Kruger FA, Lumbers RT, Shah T, Engmann J, Galver L, Kelley R, Karlsson A, Santos R, et al. The druggable genome and support for target identification and validation in drug development. *Sci Transl Med.* 2017;9(383):eaag1166.
100. Tenmizu D, Endo Y, Noguchi K, Kamimura H. Identification of the novel canine CYP1A2 1117 C > T SNP causing protein deletion. *Xenobiotica.* 2004;34:835–46.
101. Mise M, Yadera S, Matsuda M, Hashizume T, Matsumoto S, Terauchi Y, Fujii T. Polymorphic expression of CYP1A2 leading to interindividual variability in metabolism of a novel benzodiazepine receptor partial inverse agonist in dogs. *Drug Metab Dispos.* 2004;32:240–5.
102. Khatri A, Williams BW, Fisher J, Brundage RC, Gurvich VJ, Lis LG, Skubitiz KM, Dudek AZ, Greeno EW, Kratzke RA, et al. SLC28A3 genotype and gemcitabine rate of infusion affect dFdCTP metabolite disposition in patients with solid tumours. *Br J Cancer.* 2014;110:304–12.

103. Lee SY, Im SA, Park YH, Woo SY, Kim S, Choi MK, Chang W, Ahn JS, Im YH. Genetic polymorphisms of SLC28A3, SLC29A1 and RRM1 predict clinical outcome in patients with metastatic breast cancer receiving gemcitabine plus paclitaxel chemotherapy. *Eur J Cancer*. 2014;50:698–705.
104. Rizzo S, Scala I, Robayo AR, Cefali M, De Dosso S, Cappio S, Xhepa G, Del Grande F. Body composition as a predictor of chemotherapy-related toxicity in pancreatic cancer patients: a systematic review. *Front Oncol*. 2022;12:974116.
105. Marconato L, Finotello R, Bonfanti U, Dacasto M, Beatrice L, Pizzoni S, Leone VF, Balestra G, Furlanello T, Rohrer Bley C, Aresu L. An open-label phase 1 dose-escalation clinical trial of a single intravenous administration of gemcitabine in dogs with advanced solid tumors. *J Vet Intern Med*. 2015;29:620–5.
106. Elbadawy M, Usui T, Mori T, Tsunedomi R, Hazama S, Nabeta R, Uchide T, Fukushima R, Yoshida T, Shibutani M, et al. Establishment of a novel experimental model for muscle-invasive bladder cancer using a dog bladder cancer organoid culture. *Cancer Sci*. 2019;110:2806–21.
107. Hua L, Chen W, Meng Y, Qin M, Yan Z, Yang R, Liu Q, Wei Y, Zhao Y, Yan L, Qiao J. The combination of DNA methylation and transcriptome revealed the intergenerational inheritance on the influence of advanced maternal age. *Clin Transl Med*. 2022;12:e990.
108. Jagannathan V, Gerber V, Rieder S, Tetens J, Thaller G, Drogemuller C, Leeb T. Comprehensive characterization of horse genome variation by whole-genome sequencing of 88 horses. *Anim Genet*. 2019;50:74–7.
109. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brondum RF, Liao X, Djari A, Rodriguez SC, Grohs C, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet*. 2014;46:858–65.
110. Eggertsson HP, Jonsson H, Kristmundsdottir S, Hjartarson E, Kehr B, Masson G, Zink F, Hjorleifsson KE, Jonasdottir A, Jonasdottir A, et al. GraphTyper enables population-scale genotyping using pangenome graphs. *Nat Genet*. 2017;49:1654–60.
111. Browning BL, Browning SR. Genotype imputation with millions of reference samples. *Am J Hum Genet*. 2016;98:116–26.
112. Martin AR, Atkinson EG, Chapman SB, Stevenson A, Stroud RE, Abebe T, Akena D, Alemayehu M, Ashaba FK, Atwoli L, et al. Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *Am J Hum Genet*. 2021;108:656–68.
113. Li JH, Mazur CA, Berisa T, Pickrell JK. Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays. *Genome Res*. 2021;31:529–37.
114. Piras IS, Bleul C, Siniard A, Wolfe AJ, De Both MD, Hernandez AG, Huentelman MJ. Association of common genetic variants in the CPSF7 and SDHAF2 genes with canine idiopathic pulmonary fibrosis in the west highland white terrier. *Genes (Basel)*. 2020;11(6):609.
115. Hayward JJ, White ME, Boyle M, Shannon LM, Casal ML, Castelhana MG, Center SA, Meyers-Wallen VN, Simpson KW, Sutter NB, et al. Imputation of canine genotype array data using 365 whole-genome sequences improves power of genome-wide association studies. *PLoS Genet*. 2019;15:e1008003.
116. Friedrich J, Antolin R, Edwards SM, Sanchez-Molano E, Haskell MJ, Hickey JM, Wiener P. Accuracy of genotype imputation in Labrador Retrievers. *Anim Genet*. 2018;49:303–11.
117. Buckley RM, Harris AC, Wang GD, Whitaker DT, Zhang YP, Ostrander EA. Best practices for analyzing imputed genotypes from low-pass sequencing in dogs. *Mamm Genome*. 2022;33:213–29.
118. Srikanth K, von Pfeil DJF, Stanley BJ, Griffiths C, Huson HJ. Genome wide association study with imputed whole genome sequence data identifies a 431 kb risk haplotype on CFA18 for congenital laryngeal paralysis in Alaskan sled dogs. *Genes (Basel)*. 2022;13(10):1808.
119. Jenkins CA, Schofield EC, Mellersh CS, De Risio L, Ricketts SL, Dog Biomedical Variant Database C. Improving the resolution of canine genome-wide association studies using genotype imputation: a study of two breeds. *Anim Genet*. 2021;52:703–13.
120. Friedenberg SG, Meurs KM. Genotype imputation in the domestic dog. *Mamm Genome*. 2016;27:485–94.
121. Das S, Abecasis GR, Browning BL. Genotype imputation from large reference panels. *Annu Rev Genomics Hum Genet*. 2018;19:73–96.
122. Schlamp F, van der Made J, Stambler R, Chesebrough L, Boyko AR, Messer PW. Evaluating the performance of selection scans to detect selective sweeps in domestic dogs. *Mol Ecol*. 2016;25:342–56.
123. Truve K, Dickinson P, Xiong A, York D, Jayashankar K, Pielberg G, Koltookian M, Muren E, Fuxelius HH, Weishaupt H, et al. Utilizing the dog genome in the search for novel candidate genes involved in glioma development-genome wide association mapping followed by targeted massive parallel sequencing identifies a strongly associated locus. *PLoS Genet*. 2016;12:e1006000.
124. Shyrokovaya EY, Prassolov VS, Spirin PV. The role of the MCTS1 and DENR proteins in regulating the mechanisms associated with malignant cell transformation. *Acta Naturae*. 2021;13:98–105.
125. Tomaszewski WH, Waibl-Polania J, Chakraborty M, Perera J, Ratiu J, Miggelbrink A, McDonnell DP, Khasraw M, Ashley DM, Fecci PE, et al. Neuronal CaMKK2 promotes immunosuppression and checkpoint blockade resistance in glioblastoma. *Nat Commun*. 2022;13:6483.
126. Baroja-Mazo A, Penin-Franch A, Lucas-Ruiz F, de Torre-Minguela C, Alarcon-Vila C, Hernandez-Caselles T, Pelegrin P. P2X7 receptor activation impairs antitumour activity of natural killer cells. *Br J Pharmacol*. 2023;180:111–28.
127. Leuthard F, Lehner G, Jagannathan V, Leeb T, Welle M. A missense variant in the NSDHL gene in a Chihuahua with a congenital cornification disorder resembling inflammatory linear verrucous epidermal nevi. *Anim Genet*. 2019;50:768–71.
128. Dickinson PJ, Bannasch DL. Current understanding of the genetics of intervertebral disc degeneration. *Front Vet Sci*. 2020;7:431.
129. Meyers-Wallen VN, Boyko AR, Danko CG, Grenier JK, Mezey JG, Hayward JJ, Shannon LM, Gao C, Shafquat A, Rice EJ, et al. XX Disorder of Sex Development is associated with an insertion on chromosome 9 and downregulation of RSP01 in dogs (*Canis lupus familiaris*). *PLoS ONE*. 2017;12:e0186331.
130. Faber MS, Jetter A, Fuhr U. Assessment of CYP1A2 activity in clinical practice: why, how, and when? *Basic Clin Pharmacol Toxicol*. 2005;97:125–34.

131. Sun D, Lu J, Zhang Y, Liu J, Liu Z, Yao B, Guo Y, Wang X. Characterization of a Novel CYP1A2 Knockout Rat Model Constructed by CRISPR/Cas9. *Drug Metab Dispos.* 2021;49:638–47.
132. Kapelyukh Y, Henderson CJ, Scheer N, Rode A, Wolf CR. Defining the Contribution of CYP1A1 and CYP1A2 to Drug Metabolism Using Humanized CYP1A1/1A2 and Cyp1a1/Cyp1a2 Knockout Mice. *Drug Metab Dispos.* 2019;47:907–18.
133. Takahashi T, Luzum JA, Nicol MR, Jacobson PA. Pharmacogenomics of COVID-19 therapies. *NPJ Genom Med.* 2020;5:35.
134. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81:559–75.
135. Felsenstein J. PHYLIP-Phylogeny Inference Package (Ver. 3.2). *Cladistics.* 1989;5:164–6.
136. Browning BL, Browning SR. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics.* 2013;194:459–71.
137. Petr M, Vernot B, Kelso J. admix-R package for reproducible analyses using ADMIXTOOLS. *Bioinformatics.* 2019;35:3194–5.
138. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. Ancient admixture in human history. *Genetics.* 2012;192:1065–93.
139. Duntsch L, Whibley A, Brekke P, Ewen JG, Santure AW. Genomic data of different resolutions reveal consistent inbreeding estimates but contrasting homozygosity landscapes for the threatened Aotearoa New Zealand hihi. *Mol Ecol.* 2021;30:6006–20.
140. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26:841–2.
141. Hofmeister RJ, Ribeiro DM, Rubinacci S, Delaneau O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat Genet.* 2023;55(7):1243–1249.
142. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al. The UCSC genome browser database: update 2006. *Nucleic Acids Res.* 2006;34:D590–598.
143. Rubinacci S, Ribeiro DM, Hofmeister RJ, Delaneau O. Publisher Correction: Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat Genet.* 2021;53:412.
144. Rubinacci S, Delaneau O, Marchini J. Genotype imputation using the positional burrows wheeler Transform. *PLoS Genet.* 2020;16:e1009049.
145. Laricchia KM, Lake NJ, Watts NA, Shand M, Haessly A, Gauthier L, Benjamin D, Banks E, Soto J, Garimella K, et al. Mitochondrial DNA variation across 56,434 individuals in gnomAD. *Genome Res.* 2022;32:569–82.
146. Li H: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. In *ArXiv e-prints*, vol. 1303; 2013.
147. Numanagic I, Gokkaya AS, Zhang L, Berger B, Alkan C, Hach F. Fast characterization of segmental duplications in genome assemblies. *Bioinformatics.* 2018;34:i706–14.
148. Pendleton AL, Shen F, Taravella AM, Emery S, Veeramah KR, Boyko AR, Kidd JM. Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. *BMC Biol.* 2018;16:64.
149. Bolker BM. *Ecological models and data* in R. Princeton, NJ: Princeton University Press; 2008.
150. Cingolani P, Platts A, Le Wang L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin).* 2012;6:80–92.
151. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. The variant call format and VCFtools. *Bioinformatics.* 2011;27:2156–8.
152. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. *Gigascience.* 2021;10(2):giab008.
153. Knaus BJ, Grunwald NJ. vcf: a package to manipulate and visualize variant call format data in R. *Mol Ecol Resour.* 2017;17:44–53.
154. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* 2013;14:178–92.
155. 10,000 Dog Genome Consortium. Phase One Resequencing for 10,000 Dog Genome Consortium. Datasets. Sequence Read Archive. <https://www.ncbi.nlm.nih.gov/sra/PRJNA648123> (2022)
156. University of Bern. *Canis lupus familiaris* 1000 genomes project. Datasets. Sequence Read Archive. <https://www.ncbi.nlm.nih.gov/sra/PRJNA188158> (2022)
157. European Nucleotide Archive. <https://identifiers.org/ena.embl:PRJEB62420> (2023)
158. Genome sequencing of 2,000 canids advances the understanding of demography, genome function and architecture. 2023. Zenodo. <https://doi.org/10.5281/zenodo.8084059>.
159. Kidd, Jeffrey M. Dogmap. Github. <https://github.com/jmkidd/dogmap> (2022)
160. Kidd, Jeffrey M. Doggenotype. Github. <https://github.com/jmkidd/doggenotype> (2022)
161. Kidd, Jeffrey M. Callmito. Github. <https://github.com/jmkidd/callmito> (2022)
162. Kidd, Jeffrey M. Dogmap. Zenodo. <https://doi.org/10.5281/zenodo.8087879> (2023)
163. Kidd, Jeffrey M. Doggenotype. Zenodo. <https://doi.org/10.5281/zenodo.8087891> (2023)
164. Kidd, Jeffrey M. Callmito. Zenodo. <https://doi.org/10.5281/zenodo.8087897> (2023)
165. Ellis N, Goodfellow PN. The mammalian pseudoautosomal region. *Trends Genet.* 1989;5:406–10.
166. Young AC, Kirkness EF, Breen M. Tackling the characterization of canine chromosomal breakpoints with an integrated in-situ/in-silico approach: the canine PAR and PAB. *Chromosome Res.* 2008;16:1193–202.
167. Li G, Davis BW, Raudsepp T, Pearks Wilkerson AJ, Mason VC, Ferguson-Smith M, O'Brien PC, Waters PD, Murphy WJ. Comparative analysis of mammalian Y chromosomes illuminates ancestral structure and lineage-specific evolution. *Genome Res.* 2013;23:1486–95.
168. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics.* 2018;34:867–8.

169. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20:1297–303.
170. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nat Biotechnol.* 2017;35:316–9.
171. Wang, C. Dog\_10k\_mapping. Github. [https://github.com/Chao912/dog\\_10k](https://github.com/Chao912/dog_10k) (2022)
172. Wang C. Dog\_10k\_mapping. 2023. Zenodo. <https://doi.org/10.5281/zenodo.8087147>.
173. vonHoldt BM, Cahill JA, Fan Z, Gronau I, Robinson J, Pollinger JP, Shapiro B, Wall J, Wayne RK. Whole-genome sequence analysis shows that two endemic species of North American wolf are admixtures of the coyote and gray wolf. *Sci Adv.* 2016;2:e1501714.
174. Monzon J, Kays R, Dykhuizen DE. Assessment of coyote-wolf-dog admixture using ancestry-informative diagnostic SNPs. *Mol Ecol.* 2014;23:182–97.
175. vonHoldt BM, Kays R, Pollinger JP, Wayne RK. Admixture mapping identifies introgressed genomic regions in North American canids. *Mol Ecol.* 2016;25:2443–53.
176. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7.
177. Kim KS, Lee SE, Jeong HW, Ha JH. The complete nucleotide sequence of the domestic dog (*Canis familiaris*) mitochondrial genome. *Mol Phylogenet Evol.* 1998;10:210–20.
178. Fregel R, Suarez NM, Betancor E, Gonzalez AM, Cabrera VM, Pestano J. Mitochondrial DNA haplogroup phylogeny of the dog: Proposal for a cladistic nomenclature. *Mitochondrion.* 2015;22:75–84.
179. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol.* 2011;7:539.
180. Arendt M, Fall T, Lindblad-Toh K, Axelsson E. Amylase activity is associated with *AMY2B* copy numbers in dog: implications for dog domestication, diet and diabetes. *Anim Genet.* 2014;45:716–22.
181. Reiter T, Jagoda E, Capellini TD. Dietary variation and evolution of gene copy number among dog breeds. *PLoS ONE.* 2016;11:e0148899.
182. Larson DE, Abel HJ, Chiang C, Badve A, Das I, Eldred JM, Layer RM, Hall IM. svtools: population-scale analysis of structural variation. *Bioinformatics.* 2019;35:4782–7.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

