Copy number variation underlies complex phenotypes in domestic dog breeds and other canids

Aitor Serres-Armero[1], Brian W. Davis[2,3], Inna S. Povolotskaya[4], Carlos Morcillo-Suarez[1], Jocelyn Plassais[2], David Juan[1§], Elaine A. Ostrander[2§] and Tomas Marques-Bonet[1,5,6,7§]

[1] IBE, Institut de Biologia Evolutiva (Universitat Pompeu Fabra/CSIC), Ciencies Experimentals i de la Salut, Barcelona, 08003, Spain

[2]Cancer Genetics and Comparative Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, 20892, USA

[3] Department of Veterinary Integrative Biosciences, College of Veterinary Medicine, Texas A&M University, College Station, TX 77843

[4] Veltischev Research and Clinical Institute for Pediatrics of the Pirogov Russian National Research Medical University, Moscow, Russia

[5] CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain

[6] Institucio Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia 08010, Spain

[7] Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Catalonia, Spain

[§] Corresponding authors

Running title: CNV underlies complex phenotypes in dogs

Keywords: copy number variation, dog genomics, dog breeds, morphometrics, whole genome sequencing, genome wide association study.

Email addresses:

ASA: aitor.serres@upf.edu

BWD: bdavis@cvm.tamu.edu

ISP: ipovolotskaya@gmail.com

CMS: carlos.morcillo@upf.edu

JP: jocelyn.plassais@gmail.com

DJ: david.juan@upf.edu

EAO: eostrand@mail.nih.gov

TMB: tomas.marques@upf.edu

Abstract

Extreme phenotypic diversity, a history of artificial selection, and socioeconomic value make domestic dog breeds a compelling subject for genomic research. Copy number variation (CNV) is known to account for a significant part of inter-individual genomic diversity in other systems. However, a comprehensive genome-wide study of structural variation as it relates to breed-specific phenotypes is lacking. We have generated whole genome CNV maps for more than 300 canids. Our dataset extends the canine structural variation landscape to more than 100 dog breeds, including novel variants that cannot be assessed using microarray technologies. We have taken advantage of this dataset to perform the first CNV-based GWAS in canids. We identify 96 loci that display copy number differences across breeds, which are statistically associated with a previously compiled set of breed-specific morphometrics and disease susceptibilities. Among these, we highlight the discovery of a long-range interaction involving a CNV near *MED13L* and *TBX3*, which could influence breed standard height. Integration of the CNVs with chromatin interactions, long non-coding RNA expression, and single nucleotide variation highlights a subset of specific loci and genes with potential functional relevance and the prospect to explain trait variation between dog breeds.

Introduction

Dogs have been the subject of intense study over many decades (Vilà et al. 1999; Ostrander and Wayne 2005; Freedman et al. 2014; Ostrander et al. 2019), providing valuable insight into human history, disease and evolution (Ní Leathlobhair et al. 2018; Wang et al. 2018; Coelho et al. 2018). Much has been learned about canines through traditional approaches, including genotype studies with microsatellites (Irion 2003), single nucleotide polymorphisms (SNP) (Gundry et al. 2007; Boyko et al. 2010; Vaysse et al. 2011) and, finally, whole genome sequencing (WGS) (Lindblad-Toh et al. 2005;

Freedman et al. 2014; Plassais et al. 2019).

As a result of the extensive history of genetic studies in dogs, remarkable advances have been made towards the resolution of the canine phylogeny (vonHoldt et al. 2010; Parker et al. 2017) and the temporal, geographic and demographic history of dog domestication (Freedman et al. 2014; Skoglund et al. 2015; Shannon et al. 2015). Studies suggest that dogs were initially domesticated from grey wolves 15,000 to 40,000 years ago (Freedman et al. 2014; Skoglund et al. 2015; Freedman and Wayne 2017; Ostrander et al. 2019) with a rapid diversification of breeds occurring within the past few hundred years. Currently, about 400 dog breeds exist worldwide, 193 recognized by the American Kennel Club and 360 by the Fédération Cynologique Internationale. Breed classification schemes have been proposed based on occupation, morphology, and geographic origin (American Kennel Club 2007; Wucher et al. 2017). The most recent genetic analysis encompassing nearly 200 breeds and populations suggests a monophyletic origin for most modern breeds and provides data regarding their origins and timing (Parker et al. 2017). Clusters of genetically similar breeds were identified and assigned to clades, which often reflected occupational and geographical origins.

Targeted and genome wide genotyping approaches have led to the discovery of nearly 400 variants associated with more than 270 traits, over 220 of which correspond to possible models for human diseases (Online Mendelian Inheritance in Animals, OMIA. Sydney School of Veterinary Science, {2019-10-21}. World Wide Web URL: https://omia.org/) (Sydney School of Veterinary Science 2019). Particularly, genome wide association studies (GWAS) involving modest size cohorts of dogs have led to the identification of variants controlling a variety of morphological, behavioral and disease traits (Rimbault et al. 2013; MacLean et al. 2019; Vaysse et al. 2011; Hayward et al. 2016; Plassais et al. 2019; Akey et al. 2010).

The recent and intense artificial selective pressure exerted on dogs has induced pronounced inter-breed phenotypic differences while preserving intra-breed homogeneity. This process makes dogs of the same breed more likely to share not only morphometric traits but also disease susceptibilities (Chase et al. 2009; Boyko et al. 2010; Karlsson and Lindblad-Toh 2008; Ostrander et

al. 2019; Marchant et al. 2017; Mansour et al. 2018; Akey et al. 2010). The level of anatomic similarity among dogs of any one breed is sufficiently strong that genetic studies have been successfully executed using breed standards as phenotypes, thus unraveling the genetic bases of some complex traits such as body size or behavior (MacLean et al. 2019; Plassais et al. 2019; Hayward et al. 2016; Vaysse et al. 2011; Boyko et al. 2010; Akey et al. 2010), which remain elusive, even in humans.

However, all these analyses have been performed using a subset of indicative SNPs and, more recently, SNPs from WGS (Jagannathan et al. 2019; Plassais et al. 2019), but other forms of genomic variation have rarely been studied systematically. In fact, there is still a lack of fine-scale, genome-wide analyses of any variants other than SNPs across dog breeds, a notable exception when compared to humans and other model organisms (Yalcin et al. 2011; Brown et al. 2012; Sudmant et al. 2015). Copy number variation has been previously studied in canines to elucidate specific phenotypes (Arendt et al. 2014; Waldo and Diaz 2015; Deane-Coe et al. 2018; Karyadi et al. 2013). However, most studies have focused on the comparison of dogs and wolves using array-based technologies, rather than undertaking a comprehensive and unbiased examination of all CNVs across the genome of distinct breeds (Berglund et al. 2012; Schoenebeck et al. 2012). Most CNV related studies published to date only aimed to identify segmentally duplicated regions and did not aim to produce quantitative copy-number (CN) genotypes (Quilez et al. 2012; Molin et al. 2014). Knowing the exact number of copies at a locus is crucial for an accurate comparison of closely related organisms, such as distinct dog breeds and wild canids.

Here we present a fine-scale CNV map of over 300 canid samples using WGS to produce the most extensive, high-resolution CNV panel in dogs to date. We examine more than 145 individual breeds, as well as non-breed dogs, including village dogs, dingoes, captive New Guinea singing dogs and wild canids such as wolves. We employ this dataset to determine the ability of CNVs to recreate a current dog phylogeny. Moreover, we test for breed-phenotype associations using an extensive dataset of breed standards as individual phenotypes in the first CNV-based GWAS performed in dogs to date.

Results

We created a fine-scale CNV map using a panel of 263 purebred dog genomes, 59 village dogs from diverse locations, and 17 grey and Tibetan wolves (Supplemental Fig. S1). All the samples were previously sequenced at moderate coverage (Methods; SRA BioProject numbers: PRJNA232497, PRJNA448733, PRJNA186960, PRJNA176193, PRJNA192935, PRJNA233638, PRJNA247491, PRJNA263947, PRJNA261736, PRJEB6079, PRJEB6076, PRJEB2162, PRJNA188158, PRJNA208087, PRJEB5500; Supplemental Data S1,2; (Kim et al. 2012; Zhang et al. 2014)). We find over 95% concordance between the structural variants generated in this study and those that we previously reported (Serres-Armero et al. 2017). The breed frequencies of the main CNVs presented here have additionally been validated using array comparative genomic hybridization data (Ramirez et al. 2014; Nicholas et al. 2011; Berglund et al. 2012) (Supplemental Fig. S2). All phenotypic analyses are based on breed standards from the American Kennel Club and the Fédération Cynologique Internationale, which were published previously (Plassais et al. 2019) (Methods, Supplemental Data S3).

Copy Number Statistics of modern dogs, village dogs, and wolves

We report a total of 348.26 Mb of CNVs larger than one kilobase pair across all samples, amounting to approximately 14.69% of the entire canine genome. Of note, the sex chromosomes and unassembled chromosomes have not been considered in this work. A total of 6,765 events (169.96 Mb) with an average size of 25.53 kb are gains, and 66,254 events (126.94 Mb) with an average size of 2.36 kb are losses relative to the CanFam3.1 dog genome reference build (Table 1). Gains, defined as any region with CN above two, are more often shared across samples than losses, and therefore most rare variants (MAF<0.05) tend to be deletions (Table 1). We note that the gains cluster together across most autosomes while losses do not (Fisher aggregated p-value$_{gains}$=0.006 and

p-value$_{losses}$=0.906 and Supplemental Fig. S3), a phenomenon previously described in other species (Li et al. 2009; Upadhyay et al. 2017). In terms of possible function alterations, 4,989 gains and 10,366 losses overlap with at least 5% of a gene annotation, although we note that this overlap most often involves introns. In contrast, 1,362 gains overlap entire genes as opposed to only 686 losses, a remarkable difference considering the larger raw number of deletion events reported here (Table 1). This suggests that the overlap between CN losses and entire genes could be constrained and therefore possibly deleterious in dogs.

	Sample average		Union CNVs	
	Gains (CN ≥ 2)	Losses (CN ≤ 2)	Gains (CN ≥ 2)	Losses (CN ≤ 2)
5% Gene overlap	3,540.92 ± 14.91 (104.95 ± 3.60)	1,805.86 ± 1,035 (3.56 ± 2.32)	4,989 (60.43)	10,366 (20.83)
100% Gene overlap	1,081.77 ± 13.44 (65.36 ± 3.25)	82.41 ± 60.39 (0.28 ± 0.21)	1,362 (7.97)	686 (0.80)
Frequency < 0.05	12.69 ± 7.55 (0.08 ± 0.06)	558.05 ± 495.74 (0.83 ± 0.78)	1,847 (17.17)	21,094 (36.23)
Totals	5,071.35 ± 36.31 (80.61 ± 0.71)	10,108 ± 4,597.22 (17.32 ± 9.37)	6,765 (163.66)	66,254 (126.94)

Table 1. Global and average CNV summary statistics. Statistics for the number of CNVs overlapping 5 and 100 percent of a gene are shown, as well as the frequency of rare events (events shared by less than 5% of all samples). In each cell, the number of events is reported together with the number of cumulative megabase pairs they span in parentheses. Per-sample averages are separated from their standard deviation values by a plus-minus sign.

We assessed how much of the current breed phylogeny, constructed using 150,000 SNPs (Parker et al. 2017), can be recapitulated using the duplications and deletions reported here (Supplemental Fig. S4,5). Altogether, we are able to separate breeds resulting from the first domestication bottleneck (i.e., Arctic and Asian spitz, Tibetan Mastiffs, and ancient sighthounds) (Freedman et al. 2014). However, we do not achieve a fully monophyletic separation of breeds derived in the eighteenth century and after, even when accounting for possible described admixture and inbreeding effects (Parker et al. 2017). Some possible explanations include homoplasies (Gazave et al. 2011; Bickhart et al. 2016), non-neutrality of the CNVs (Chen et al. 2009), and a poor genotype quality. Therefore, the correlation of CNV with geography and genealogy is not as clear as previously

observed using single nucleotide variation (SNV) (Parker et al. 2017). This observation suggests that CNV variability is not strongly driven by population structure, as are SNVs. Consequently, population stratification is not expected to be a prevalent confounding factor in CN genotype-phenotype correlations within domestic breeds.

Comparative of modern dogs, village dogs, and wolves

Domestic dogs, wolves and, to a smaller extent, village dogs can be discriminated via principal component (PC) analysis or by pairwise Euclidean distance (Fig. 1A). Most domestic dogs cluster together across the first four PCs, with a few exceptions overlapping village dogs. Wolves exhibit a greater dispersion but still constitute a distinct group (Fig. 1A). The first principal component (PC1) recapitulates the variation cline expected to result from dog domestication, where non-domestic dogs appear between domestic dogs and wolves. In contrast, the third principal component (PC3) hints at the opposite pattern, which endorses our previous observation that dogs may preserve CNV similarities with wolves (Serres-Armero et al. 2017). We did not observe a significant reduction in the number of CNV sites in purebred dogs when compared to wolves (Supplemental Fig. S6). This is in stark contrast to the SNV decline reported using whole genome sequencing data in numerous domesticated organisms (Freedman et al. 2014; Makino et al. 2018). In fact, village dogs show a slightly reduced number of CNV sites compared to dogs and wolves.

We applied the pairwise V_{ST} statistic (Redon et al. 2006) to scan for regions overlapping genes with different CN genotypes between dogs and wolves. Of note, V_{ST} is a fixation index generally ranging from 0 to 1, which compares the variance in absolute CN within and between two groups of samples, a small intra-group, and large inter-group CN variance will yield V_{ST} values close to 1. We identified 11 Mb of high V_{ST} CNVs overlapping with 61 genes. Some of these CNVs have previously been reported, such as those affecting the *AMY2B* (Chr6:46,948,529-46,957,042) or *MAGI2* (Chr18:18,447,653-18,449,673) genes (Chen et al. 2009; Arendt et al. 2014). However, we also discovered some novel gene-overlapping CNVs (Supplemental Table S1). We observed an

unexpectedly large proportion of CNVs overlapping genes involved in fatty acid biosynthesis (GO p-value < 0.001), some of which have been previously reported (Axelsson et al. 2013). We also report differences in a CNV within a hemoglobin chain gene cluster (Chr21:28,187,060-28,188,467) (ENSCAFT00000009978, ENSCAFT00000022860, ENSCAFT00000009984) (Fig. 1C) (CN below two in many domestic and village dogs) and *SLIT2* (Chr3:88,965,914-88,966,914) (Fig. 1D) (CN below two in many wolves) genes which, in humans, have been associated with adaptation to high altitude and neural development respectively (Hu 1999; Bigham 2016). We note that the CN distribution for *SLIT2* is unexpected since *SLIT1/2* mice knockouts suffer optic chiasm and kidney development problems (Plump et al. 2002; Grieshammer et al. 2004), and a deletion segregating within the population at such frequencies would likely appear in homozygosity. A more cautious hypothesis would be to consider that this event corresponds to a sequence rearrangement or omission in the dog genome to which wolf reads map poorly.

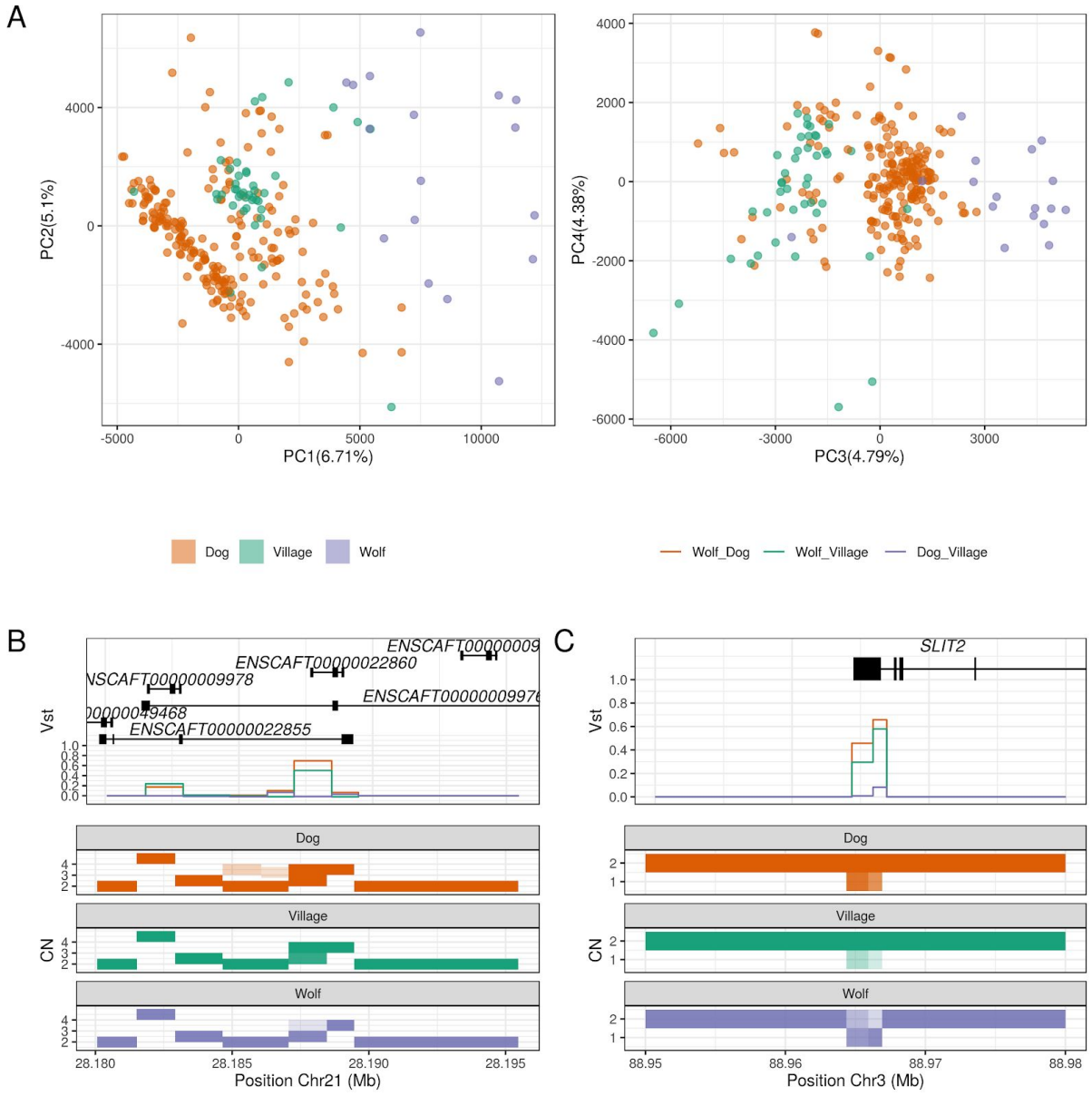

Figure 1. **(A)** Copy number-based principal component analysis of breed dogs (green), village dogs (red), and wolves (blue). **(B-C)** Depictions of the copy number values for two highly differentiated loci: the *HBB* chain gene cluster (*ENSCAFT00000009978*, *ENSCAFT00000022860*, *ENSCAFT00000009984*) and *SLIT2*. Discrete copy number values for all samples are depicted in rectangle plots. Blue: wolves; Green: village dogs; Orange: modern dogs. Top panel: V_{ST} values for the same genomic windows; Bottom panel: copy number window values.

CNV-GWAS

Given the lack of global maps for genome-wide CNV analyses, absolute copy number has never been globally assessed for trait associations in dogs. Here, we used 58 non-redundant phenotypes based on breed standards in a search for associated CNVs. In order to assess different association trends, we implemented and compared discrete and continuous generalizations of widely used association tests (see Methods and Supplemental Table S3), controlling for population stratification only when an excess of significant p-values was observed after accounting for inflation (Tsepilov et al. 2013).

CNV associations were found for 31 of the 58 non-redundant phenotypes assayed. An additional 17 phenotypes showed associated CNVs with p-values above the secondary threshold (Methods), one order of magnitude below the Bonferroni correction. The most frequently represented phenotypes were body height, hair length, and tail-to-body ratio. In contrast, our analyses were not able to assess some commonly studied phenotypes such as eye pigmentation or fur color and density, for which phenotype-driving CNVs are known, for instance, those affecting the *ASIP*, *RALY*, *RSPO2, FOXI3,* and *ALX4* genes (Dreger and Schmutz 2011; Cadieu et al. 2009; Drögemüller et al. 2008; Deane-Coe et al. 2018). These traits are variable within breeds and thus require individual phenotypes in order to be resolved, or they are particular to a unique breed for which there were few representatives in our panel. Additionally, our median sample sequence coverage of 12.66X is unsuitable for accurately detecting CNVs below 500 bp. For these phenotypes, our panel and phenotype imputation approach using breed standards lacked the power to detect the associated CNVs.

We report a duplication associated with body height (Chr26:12,739,546-12,754,676) (Fig. 2A,C, Supplemental Fig. S7A), previously reported as a SNP result (Hayward et al. 2016; Plassais et al. 2019), which harbors a CpG island 20 kb upstream of the *MED13L* gene. We detect a well-supported Hi-C interaction (Vietri Rudan et al. 2015) between this duplication and the Hi-C window containing the *TBX3* gene, located almost one Mb downstream (Supplemental Table S5).

Haploinsufficiency of the syntenic region containing *TBX3* and *MED13L* (Jin et al. 2013) has been reported to cause short stature, developmental delay and intellectual disability, among other conditions in humans (Adegbola et al. 2015) and is established as a major contributor to height in horses (Kader et al. 2015). This interaction contains a CCCTC-binding factor (CTCF) motif, which is conserved in placental mammals (Pollard et al. 2010), and contributes to the formation of a 3D chromatin regulatory domain that isolates *TBX3* from *MED13L* and *TBX5* in mouse (van Weerd et al. 2014).

About 50 small-sized dogs in our panel, defined as breeds with an average adult male and female height of 22 cm or less, carried either a homozygous or heterozygous ~14 kb deletion (Chr7:43,787,168-43,801,320) located approximately 10 kb downstream from the *SMAD2* gene. This deletion encompasses a CpG island on the 3' end of the gene and has been previously hypothesized to have regulatory functions (Rimbault et al. 2013) (Fig. 2A,B, Supplemental Fig. S7B). This region contains a conserved cluster of transcription binding sites located in a mammalian-level synteny block, and its corresponding orthologous region (mm10_Chr18:76324482-76336682) physically interacts with the promoter of the *HOXA10* gene in a different chromosome in mouse ESCs (Denholtz et al. 2013). *HOXA10* is a developmental gene essential for osteoblastogenesis and skeletal development (Favier et al. 1996; Hassan et al. 2007), as well as in sexual differentiation in mammals (Wilhelm and Koopman 2006; Kobayashi and Behringer 2003).

In this study, we sought to determine if genetic associations could be found for established breed propensities for cardiac, thyroid, orthopedic, and eye diseases (available from the Orthopedic Foundation for Animals (OFA)) using a similar approach as that described for morphological traits. A minimum threshold of one affected individual in 2000 (0.05%) reported in the OFA database was required for a breed to be included in the analysis of any disease. Also, we selected as “cases” breeds with the highest risk of developing the disorder while breeds with the lowest risk served as “controls”. We report a minimum of 21 “case” breeds for any of the eight diseases analyzed here.

Consistent with the lack of clinical phenotypes in our study and the assumption that most

dogs in the database were healthy, the GWAS performed here, especially those involving cardiac and thyroid conditions, suffered from a p-value deflation. However, our approach found a few noteworthy CNV candidates overlapping provocative candidate genes (Supplemental Table S2). We detected an association for generalized-progressive retinal atrophy (gPRA) risk in a CNVs covering more than ten exons of the *DMBT1* multi-copy gene (Chr28:32,220,591-32,260,415) (Fig. 2D,E). Of note, the genetic basis of canine progressive retinal atrophy has for long been a field of intense study in humans and dogs (Hitti et al. 2019; Bunel et al. 2019; Lippmann et al. 2007; Acland et al. 1994; Downs et al. 2011). The lowest gPRA risk group (0.17% average prevalence), which included 34 breeds, had more copies of this complex duplication than the higher risk group (2.38% average prevalence), which included 26 breeds. This could suggest a potential protective role for an increased CN. CNVs in this gene have previously been hypothesized to be associated with macular degeneration in humans, a distinct condition but one which nevertheless affects the retina (Polley et al. 2016).

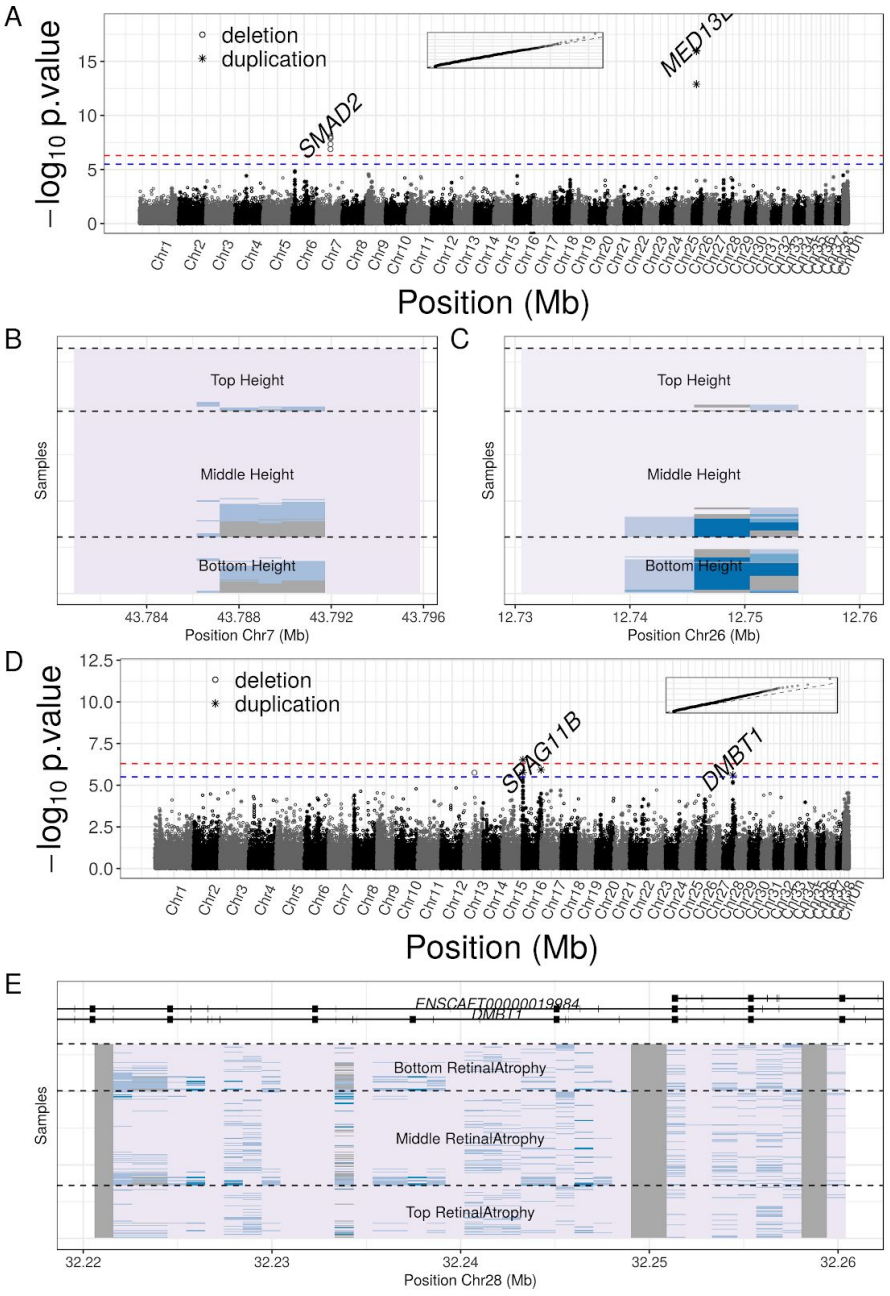

Figure 2. (A, D) Manhattan plots of the copy number GWAS for breed standard height, retinal atrophy susceptibility, respectively (Jones et al. 2008). Red line: Bonferroni correction ($-\log_{10}$ p-value=6.417). Blue line: one order of magnitude below Bonferroni correction ($-\log_{10}$ p-value=5.417). P-values were calculated using different tests (Supplemental Table S3 and Methods). **(B-C, E)** Close-up of the relevant regions for each trait, respectively*:* *SMAD2* (Chr7:43,787,168-43,801,320) and *MED13L* (Chr26:12,739,546-12,754,676) loci for height, and *DMBT1* (Chr28:32,220,591-32,260,415) for retinal atrophy. Each sample corresponds to a line along the y-axis and is ordered according to the trait in question. The x-axis shows the genomic position of each window. CN windows for each sample are colored according to their normalized distance to the median CN in the window,the darker the shade of blue, the more the CN of a sample differs from the window median. Gray CN windows correspond to uncertain genotypes.

CNV-GWAS annotation

To investigate the relationship between CNV and SNP driven associations, we gathered data on genomic variants from different studies to test whether our secondary-threshold GWAS associations followed any discernible patterns. In particular, we focused on signals that were close to significance, as concordance between multiple non-coding or intergenic regions and their previous, independent annotations could both serve as a validation and potentially point to polygenic effects.

We cross-referenced the CNV-GWAS signals with a preceding WGS-GWAS study for the same traits (Plassais et al. 2019). For each reported SNV association, we assessed whether the closest CNV signals had higher p-values than expected (Methods). Even if we were able to identify this trend in a few cases, most prominently deletions, the majority of CNV associations were independent of SNP associations (Supplemental Fig. S8). Of interest, one of our most significant GWAS results, the *SMAD2* locus, segregates together with a previously reported SNP at frequencies of 0.6 ± 0.29 depending on the breed (Chase et al. 2009; Rimbault et al. 2013; Plassais et al. 2019).

In order to assess whether intergenic and intronic CNV associations could point to unannotated regulatory regions, we studied the enrichment in conserved motifs. For this, we intersected the 75 way GERP score Ensembl annotation (Hunt et al. 2018) with our significant calls (Methods). We found no significant increase in the number of conserved and associated CNVs compared to the global background of all non-genic structural variation (Supplemental Table S4). This means that associated CNVs behave the same as the rest of CN events in terms of sequence conservation. Indeed, there seems to be an overall depletion in highly conserved motifs in the canine structural variation space, in part due to the poor alignment within complex regions when constructing conservation scores. This depletion is consistent with previous findings involving ultraconserved elements in mammals (Derti et al. 2006), suggesting that dosage alteration of these elements via CNVs could have deleterious effects.

A substantial part of the CNV-GWAS associations reported here either overlapped or were close to long noncoding RNA genes. Therefore, we assessed the concordance between the lncRNA

tissue of expression and the CNV-GWAS trait as a possible indicator of a non-spurious distribution of these association results. We used the dog lncRNA database (Le Béguec et al. 2018) to annotate the GWAS significant signals within a 10 kb range of a lncRNA based on the tissue where the lncRNA is most abundant. We compared the empirical GWAS-lncRNA contingency table against an independent distribution of both features (Methods). We found that some trait associations were enriched in concordant lncRNA tissues (e.g., brain lncRNA expression for intelligence) while retaining the expected counts in all other tissues. Particular examples are adrenal gland expression for temperament associations and muscle, blood, and heart expression for racing, i.e., whether a breed is commonly used for dog racing (Supplemental Fig. S9). While these results should be interpreted cautiously due to the low robustness of sparse and low contingency table counts, the consistency between traits with no major protein-coding associations and lncRNA expression encourages further exploration.

Finally, we also assessed whether any associated region, aside from *TBX3* mentioned above, displayed any distal Hi-C contacts (Fig. 3A). We verified all contacts reported here using ChIP-seq data for dog CTCF motifs (Schmidt et al. 2012), assessing whether each end of the contact contains at least one CTCF in inward opposing directions (Methods). Any long-range interactions involving lncRNAs were also accounted for in the corresponding analysis. We found seven significant, well-supported interactions in our dataset. Most prominently, an association signal for hair length in a largely unannotated genomic region (Chr9:16,780,483-16,782,227) interacts with the *MAP2K6* gene located almost one Mb away (Fig. 3B). The role of the gene in hypertrichosis in both humans and foxes is a topic of debate (Clark et al. 2016). Of note, the *KCNJ2* and *KCNJ16* genes, together with four lncRNAs are also within the range of this interaction.

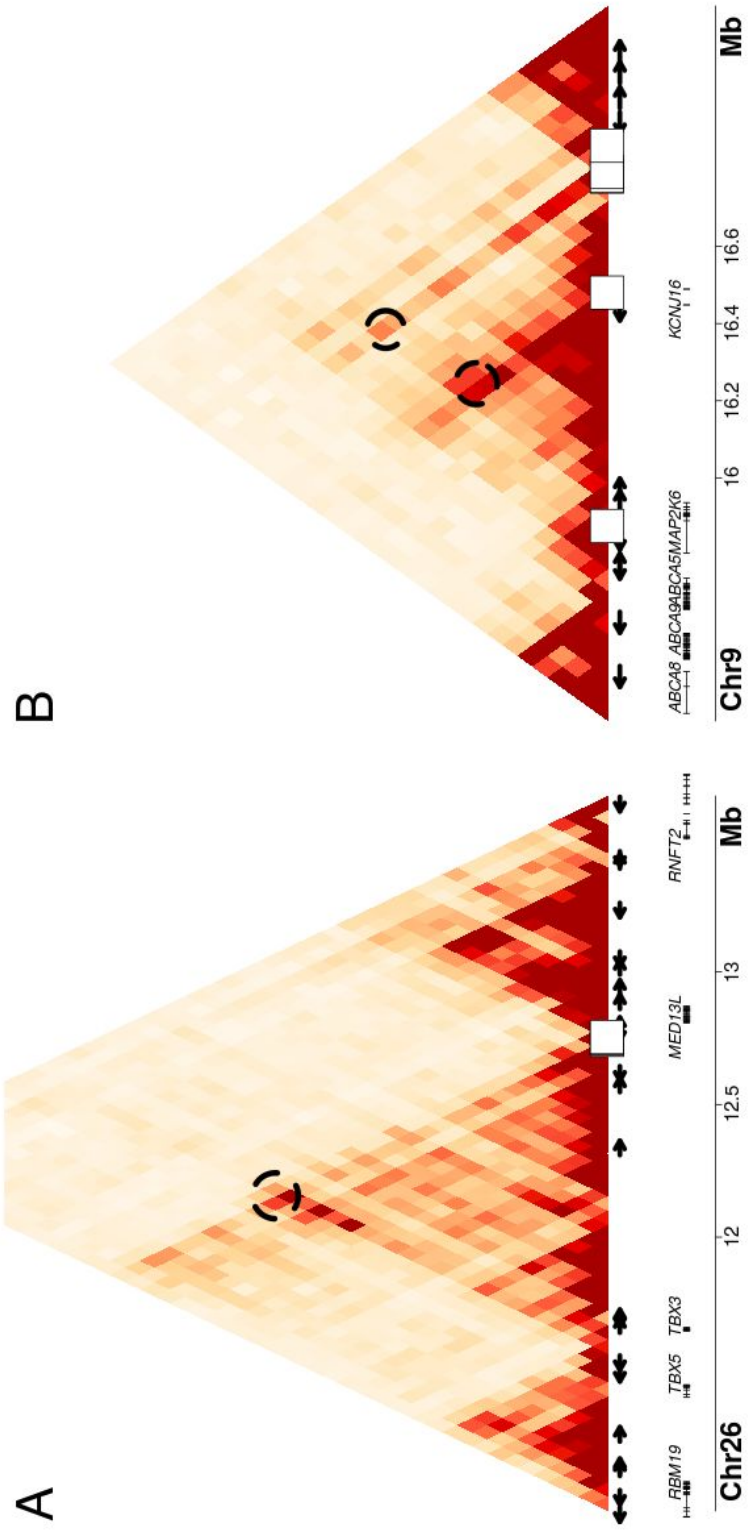

Figure 3. (A-C) Genomic interaction plots for our significant interaction hits on Chromosomes 26, 4, and 9. The white squares on the interaction plot mark the position the GWAS hits. The arrows mark the position and directionality of the most significant CTCF motifs of the area. The dashed circles within the Hi-C plot mark the interactions involving the relevant genes and the associated CNVs.

Breed V_{ST}

We next analyzed possible differences in copy number arising from breed differentiation without consideration of any specific phenotype. We applied the pairwise V_{ST} statistic (Redon et al. 2006) (Methods, Supplemental Data S4) to all pairs of breed clades consisting of more than six individuals (Methods). Overall, we found some highly differentiated loci in a subset of established clades (primarily Tibetan Mastiffs, Arctic Spitz, Shepherds, Ancient Sighthounds, and Scent hounds), largely corresponding to gene-poor regions. However, a few of these differentiated CNVs contained one or more members of extensive protein families (such as olfactory receptors, solute carriers, and late cornified envelope proteins).

We detected a high V_{ST} signal in a CNV locus (Chr1:101,853,329-101,921,437) involving two adjoining genes involved in innate immune response *NLRP13* and *NLRP8* (Fig. 4A). This association results from German Shepherds and Rottweilers having a different CN distribution than other breeds at this locus, probably due to their common ancestry. We also observed a deletion of ~35 kb (Chr4:17,945,894-17,986,191) in the third intron of *CTNNA3* (Fig. 4B), a gene involved in cell adhesion. Terriers and retrievers possess, on average, fewer copies of this CNV than most other breeds. Both *CTNNA3* and *NLRP8/13* are secondary threshold GWAS associations for cataract propensity and herding (i.e. whether a breed is used for herding) respectively (Supplemental Table S2). Finally, we also found a homozygous deletion in many German Shepherds (Chr27:25,735,696-25,844,995) encompassing two *SLC7A* orthologs (ENSCAFT00000025818 and ENSCAFT00000025829) (Fig. 4C).

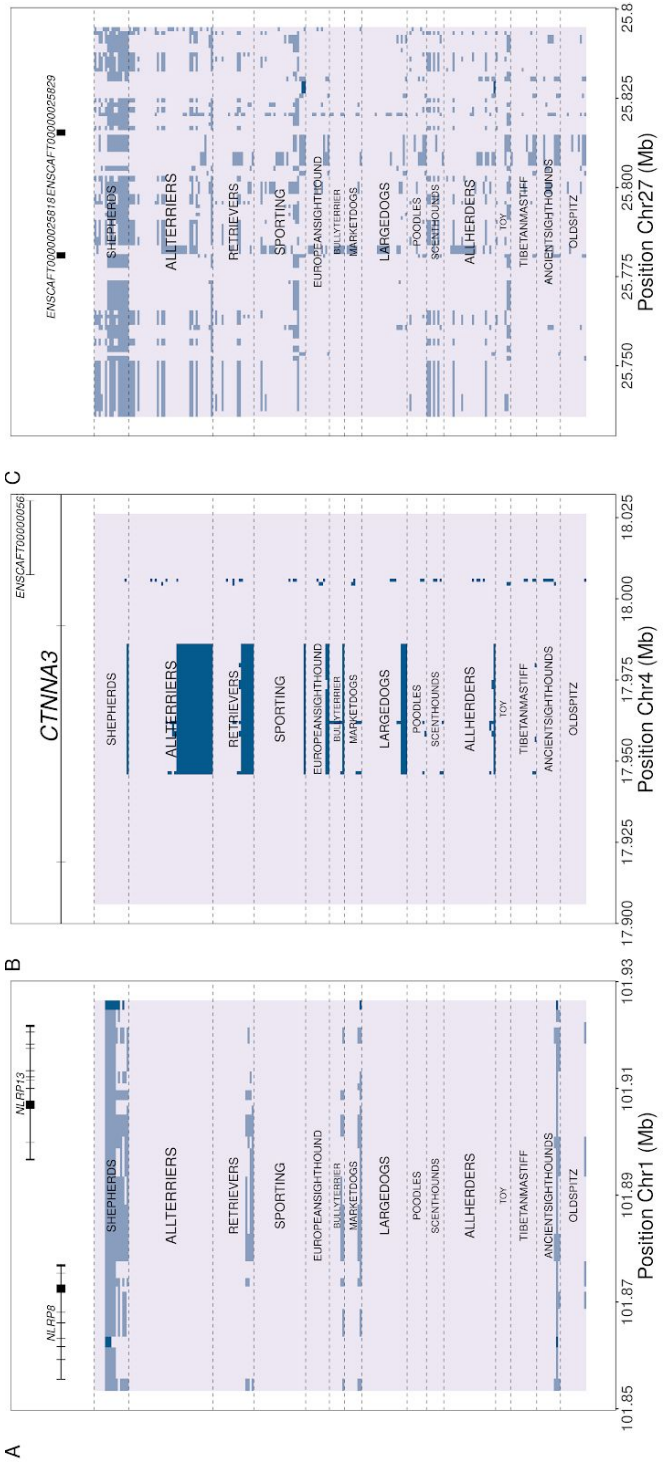

Figure 4. (A-C) Representation of high inter-breed V_{ST} regions for the *NLRP8* and *NLRP13* genes, *CTNNA3*, and *SLC7A* (ENSCAFT00000025818 and ENSCAFT00000025829), respectively. Normalized copy number is represented by color where the modal CN is the lightest color, and deviations from the mode (either deletions or duplications) are colored darker. Each row represents a sample ordered by clade, as described in (Parker et al. 2017).

Discussion

This study represents the first comprehensive and cohesive whole genome analysis specific to CNVs in the dog. We herein take advantage of our newly created map of structural variation in domestic dogs and other canids to explore the association between CNVs and phenotypes, showing clear and reproducible differences between established breed clades. GWAS with absolute CN as the testable variable are not common in the literature, especially outside of the field of human genetics (Wellcome Trust Case Control Consortium et al. 2010), and even less so in domesticated animals. In part, this is due to the technical difficulties of working with CNVs (Zhao et al. 2013) compared to SNPs, and the more complex genetic scenario associated with their evolution (Sudmant et al. 2015; Xu et al. 2016). These hindrances ultimately result in a lack of specific tools, validation methods, and difficulties in the workflow. However, CN events have been correlated with distinct traits in a vast array of organisms (Hegele 2007; Karyadi et al. 2013; Chain et al. 2014; Upadhyay et al. 2017). Many structural variants and their phenotypic effects have been explored in dogs (Alvarez and Akey 2012), even shedding light on the molecular mechanisms of complex diseases such as skin cancer (Karyadi et al. 2013).

Using a curated panel of breed standards as a proxy for individual phenotypes, we discover new associations between phenotypes of interest and CNVs within or adjacent to excellent candidate genes. The replication of previous findings, especially those for achondroplasia, height, and body mass (Supplemental Fig. S7), provide proof of concept for this global approach. Our exploration of previously published chromatin contact maps for dogs and other mammals has provided additional insight into a distal association involving the candidate gene *MED13L* and has revealed remarkable candidate genes for hair length associations. Additionally, the CNV-GWAS for breed disease risks highlights some compelling associations, providing testable hypotheses in future studies.

Overall we detect more than 110 clear instances (Supplemental Data S4) of CNVs that are at an increased frequency in a subset of modern dog clades. These CNVs are candidates for recent

selection or could have been swept along in the process of clade origination. The much lower incidence of overall CNV events compared to other kinds of genomic variation, such as SNVs, suggests that CNVs may drive stronger phenotypic effects and their evolution through domestication is a relevant topic in several organisms (Zoe N. Lye 2019; Strillacci et al. 2019; Solé et al.). The inability to completely purge these potentially deleterious variants could result from a combination of artificial selective pressures and high inbreeding (Gaut et al. 2018), two phenomena that have been extensively reported in dogs (Akey et al. 2010; Calboli et al. 2008). Nevertheless, functional studies are needed to further validate such CNV candidates. Because of limitations in the number of dogs sequenced per breed, expanded data sets should be used to reexamine marginally significant results. In addition, realignment of the 341 sequences used here with long-read *de novo* assemblies (Wang et al.) will further refine these results.

Finally, we highlight the importance of including multi-omics data for conducting complete genetic analyses in any system, particularly those involving complex traits. The use of orthogonal genomic variants, such as tissue-specific lncRNA expression profiles, can help to contextualize the more abstruse CNV associations and explore their molecular bases. Whole genome copy number analyses provide a powerful approach for identifying the role of genomic variation in diversity within many understudied organisms.

Materials and methods

Samples

We analyzed a panel of 431 canid samples containing purebred dogs, free ranging (village) dogs, and wolves. Four wolves (Wolf34, WOLF6116, Wolf23, Wolf18) were used to train the Hidden Markov Model transition matrix (explained below) and discarded from the final panel. After quality control (described below), a total of 263 dog genomes, 59 village dogs, and 17 wolves were kept. The purebred dog samples classify into more than 130 breeds, which altogether can be divided into more than 30 breed clades (Supplemental Data S1). The breed status of each sample was used to infer its phenotype.

Phenotypes

A database of anatomical, behavioral, and disease susceptibility records was composed for each dog breed for the association studies. For most morphometrics, we used a phenotype database containing information from the FCI (http://www.fci.be/en/Nomenclature/) and the AKC (https://www.akc.org/dog-breeds/) published by Plassais et al. (Plassais et al. 2019). Behavioral data were retrieved from the CBAR-Q survey (https://datarepository.wolframcloud.com/resources/C-BARQ-Survey). Temperament and intelligence data were available from the ATTS (http://atts.org/breed-statistics/) database and "the Intelligence of dogs" book (Coren 1994), respectively. Disease susceptibility data were exclusively extracted from the OFA database (https://www.ofa.org/diseases/breed-statistics#detail). Data for purebred dog litter size were obtained from a comprehensive study on the litters from a variety of dog breeds (Borge et al. 2011). Additional morphometrics data from a previous publication (Jones et al. 2008) were included. Matching morphometrics data from multiple data sources were found to be consistent in most of the cases.

Copy number genotyping

Sample pre-processing:

The initial collection of sample sequencing formats was coerced into FASTQ format using the appropriate tools (biobambam, qseq2fastq, fastq dump), and all sequencing qualities were standardized to Phred 33 encoding. Adapters were trimmed with TrimGalore (Martin 2011), using paired-end data when possible and restricting the output length to a minimum of 36 base pairs. The trimmed sequencing reads were then further split into 78mers to facilitate the mapping process.

Reference assembly preparation:

In order to use an exhaustive mapper and further perform the necessary read depth calculations, the CanFam3.1 assembly was prepared as indicated below:

1. Standard repeat masking: masking of the corresponding genome wide tandem repeat finder annotations (Haeussler et al. 2019).
2. Assembly k-mer masking: in order to identify potentially hidden repeats, the assembly was split into 36mers with a 5 bp overlap and re-mapped against itself using GEM (Marco-Sola et al. 2012) at 6% divergence with a 10% edit distance. K-mers mapping to more than 20 positions were additionally masked. This version of the assembly was indexed (BWA, GEM, SAMTools) and used for all subsequent sample mappings.
3. Padding and assembly windowing: all the masked locations described in steps one and two were extended for 78 bps on each side. This aims to correct for the general effect of read depth deflation around masked loci. Next, the assembly was partitioned into 1000 bp windows of non-masked sequence as described in (Alkan et al. 2009). The resulting one kb genomic window coordinates were used for copy number estimation and are theoretically comparable across samples due to the common reference.

Mapping and read depth post-processing:

The pre-processed samples were aligned against the masked CanFam3.1 reference using the GEM exhaustive mapper at 6% divergence and 10% edit distance. The resulting files were processed with mrCanavar (Alkan et al. 2009), a tool for absolute copy number prediction based on read depth normalization, which performs GC correction and discriminates between CN 2 (aka control or diploid regions) and potentially duplicated windows.

Quality control:

Three primary parameters were assessed to decide which samples to include in these analyses:

- CR deviations from a gaussian distribution were measured using the Kolmogorov distance. Extreme deviations from a bell shape or distribution mean shifts could be a product of faulty normalization.

- A hard threshold (0.45) was imposed on the CR standard deviation to avoid excessive scatter of the HMM emissions.

- Local (i.e., neighboring window) control region copy number correlations were assessed using Pearson's coefficient. An excess of non-independent and non-homoscedastic CR windows was detected in a few samples, which were discarded.

Copy number genotyping and smoothing

We sought to discretize the copy number estimations to enhance comparability and produce a more biologically consistent CN measure. In order to do so, a similar setup to the one described in (Serres-Armero et al. 2017) was used.

We implemented a Hidden Markov Model in which the observed read depth (emissions) was linked to a certain integer CN value (hidden state) via a Gaussian distribution. Briefly, a set of hidden states ranging from 0 to 20 (plus Gaussian mixtures of states with CN above 20) with variance proportional to the empirical diploid dispersion and the hidden CN was declared. The transition

matrix was trained using the Baum-Welch algorithm coded in the Python pomegranate (https://github.com/jmschrei/pomegranate) package. Then, the forward-backward probability of each state for each one kb window was predicted in every sample.

Additionally, the CN genotypes were updated using the predicted probabilities of all samples together. A sliding window range of five windows with four window overlap was defined, and the expected probability of each state within it was computed. The expected local probabilities of each state were then used as priors to apply Bayes' rule on the third one kb window within the range for each sample. Finally, the range of CN states whose cumulative posterior distribution summed up to 0.95 was output.

$$p(CN=N \mid cn \in [x+dx]) = \frac{p(cn \in [x+dx] \mid CN=N) * p(CN=N)}{p(cn \in [x+dx])} = \frac{PDF(cn, N, \sqrt{0.5N}\sigma_{CR})}{\sum_{CN} PDF(cn, N, \sqrt{0.5N}\sigma_{CR})} * p(N);$$

We defined as duplications any windows where at least one individual had a CN range above (and not overlapping) CN=2. Similarly, all windows with a CN range below (and not overlapping) CN=2 were considered deletions. Most analyses were restricted to the duplication/deletion space defined here.

Copy number classification and deletion re-calling

Working with ranged CN genotypes can make it difficult to find natural sample clusters or perform genotype classification. Therefore, for each duplicated one kb window, the set of the most distant, non-overlapping CN interval(s) compared to the modal CN was defined. The rest of the CN ranges were then assigned to any of the defined intervals based on the overlap, with the option to define intermediate, non-overlapping intervals. The process was repeated until no range was re-classified.

Additionally, we aimed to emit definite, un-ranged genotypes for the set of deletions (defined

via HMM) by refitting the empirical observations with a Gaussian Mixture Model. The R (R Core Team 2018) mixtools package was used (Benaglia 2009) to fit the mixture weights of a model with fixed means 0, 1, 2, and variances sd(CN=2)/2, sd(CN=2)/2, sd(CN=2), where sd(CN=2) is the standard deviation of the control region read depth. The expected probabilities of each CN averaged over all samples were used to update the individual probabilities on each site using Bayes' rule and only the most likely genotype was output.

Segregation of structural variants by breeds

V_{ST} analyses

An in-house implementation of the pairwise V_{ST} statistic (Redon et al. 2006) was applied to each non-diploid one kb window in all breed clades containing six or more individuals. Much like F_{ST}, V_{ST} compares the statistical variance of copy number values within each breed to that of both breeds taken together.

As we had previously detected that small sample sizes could bias the genomic V_{ST} distribution, all breed groups were subsampled to six individuals 1000 times, and the median value was kept for each window and comparison.

$$V_{ST}(B_1, B_2) = 1 - \frac{len(B_1)Var(B_1) + len(B_2)Var(B_2)}{len([B_1, B_2])Var([B_1, B_2])}$$

PCA

Principal component analyses were performed using the prcomp function from the R stats package. In order to prevent sample size biases, a common PCA basis was created using a random balanced subset of all breed clades. All other samples were projected into this common basis by applying the centering, scaling, and rotation matrices output by prcomp.

CNV-based phylogeny

Tree construction:

All Euclidean distance matrices were calculated directly from the CN values using the R stats package. The distance matrices were then used to construct phylogenetic trees with the ape (Paradis et al. 2004) R package.

Tree comparisons:

When trees containing different samples, breeds, and metrics had to be compared, we extracted the common tree topologies by projecting the different distance matrices against the column space of their respective indicator matrices (where each ordered column signals which samples belong to a common breed). The column values of the resulting matrices were collapsed by breed and propagated across the diagonal to create a symmetric, synthetic distance matrix which retains the topological properties of the original matrix. The resulting distance matrices were thus ordered, filtered, and comparable under common scaling conditions. In our case, we applied simple correlation and 2-norm comparisons.

$$B\left(B^T B\right)^{-1} B^T D$$

$$\forall\, sample \in [1,2,\ldots I],\ \forall\, breed \in [1,2,\ldots J] \quad b_{\vec{ij}} := \begin{bmatrix} 1 & if\ sample \in breed \\ 0 & if\ sample \notin breed \end{bmatrix}$$

$$d_{\vec{ij}}^2 = \left\| \boldsymbol{CN}_{samp\ i} - \boldsymbol{CN}_{samp\ j} \right\|_2^2$$

Haplotype sharing tree

In order to avoid the effect of possible excessive haplotype sharing across seemingly unrelated breeds on the tree topologies, these potentially confounding loci were omitted. For this, the positions of the pairwise shared haplotype locations in (Parker et al. 2017) were removed from the

deletion space, and the sample distances were re-calculated based on the remaining deletions, correcting for the amount of subtracted positions. A similar setup was designed that removes haplotypes sharing breeds in any pairwise comparisons. The resulting topologies were compared as described above.

GWAS

Generalizations of three widely used statistical tests were used to accommodate population stratification into p-value calculations. The tests were chosen to match the requirements of each phenotype distribution, favoring linear regression for continuous, potentially additive traits, and chi-square independence tests for tabulated, categorical data. Stratification on categorical data was only applied if inflation was detected in the p-values (Supplemental Table S3).

A "primary" Bonferroni multiple-testing correction threshold (with value $-\log_{10}$. p-value=6.417) was established by dividing the significance value (α=0.05) by the total number of CNV windows. However, real CNV events are generally composed of many successive windows, and therefore, their p-values will be statistically non-independent. As the total number of independent tests should be notably lower than the total number of windows, a "secondary" and more permissive threshold (with value $-\log_{10}$. p-value=5.417) was also defined to be one order of magnitude below the primary. We considered p-values above either threshold to be of interest, prioritizing those above the primary threshold.

Categorical phenotypes:

We applied an in-house implementation of the generalized Cochran–Mantel–Haenszel (CMH) test by Richard Landis (Landis et al. 1979), as explained in Alan Agresti's 2002 statistical handbook (Agresti 2002). This generalization allows for stratification of data into subpopulations and with the ordinal nature of phenotypes and copy number.

The phenotype data was split into the top 70 and bottom 30 percentiles (two groups). Copy

number was also classified into categories, as previously described. When necessary, population stratification was accounted for by dividing the data into two similarly sized substrata based on the breed tree proposed by Parker et al. 2017 (Parker et al. 2017).

Continuous phenotypes:

Assessment of specific copy number and phenotype trends was carried out using linear regression (R base software). The four first principal components of the scaled copy number data matrix (see below) were used as covariates to correct for population stratification. Regression analyses and any further GWAS recalculations were restricted to the duplications/deletions space.

GWAS comparisons and validations

All genome arithmetics were performed using the BEDTools suite (Quinlan and Hall 2010), enforcing the necessary parameters. In broad terms, window-based association signals were mapped to their respective structural variants and then intersected with the corresponding annotation files.

lncRNA:

We proposed the independent joint distribution of all copy number lncRNA tissues and the proportion of association signals across traits as the null hypothesis to test for deviations in the associated copy number variant (ptissue $\otimes$ passociation). Multinomial distributions over the association table were assumed, and excessive cell counts were reported in terms of standard deviations. lncRNA data was downloaded from (Le Béguec et al. 2018).

Conservation scores:

Highly conserved regions were defined by binning GERP scores according to their 95th quantile value (~3). All non-exonic structural variants were used as a background to test whether

non-exonic associated variants were enriched in highly conserved elements. We tested the null hypothesis of variable independence using Fisher's test (variables: association & conservation). GERP scores were downloaded from (Hunt et al. 2018).

Hi-C:

The ratio of main contact read support (region against itself) and every other region involving that same contact was computed in CNV regions. A threshold was set at the 95th quantile of the distribution to call significant contacts (Supplemental Fig. S10). Hi-C data was downloaded from (Vietri Rudan et al. 2015).

ChIP-seq:

All putative significant contacts were verified by assessing both that they contained at least one CTCF motif on each side (Vietri Rudan et al. 2015) and that the CTCF motifs were correctly oriented i.e. facing each other. The ChIP-seq data were downloaded from (Schmidt et al. 2012) and lifted over from the CanFam2 genome build to CanFam3.1 (Haeussler et al. 2019). We re-annotated the CTCF orientation for the relevant loci using the dog-specific CTCF position weight matrix (https://www.ebi.ac.uk/research/flicek/publications/FOG03) and the software PWMTools (Ambrosini et al. 2018).

Leading SNP:

We gathered all structural variation GWAS p-values within one Mb surrounding the leading SNP GWAS signals proposed by (Plassais et al. 2019). Next, for each leading SNP, the structural variation data were binned into equally sized blocks to assess if the block containing the leading SNP contained more significant p-values than the rest.

Data access

The sequencing data used in this study is deposited in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession numbers PRJNA232497, PRJNA448733, PRJNA186960, PRJNA176193, PRJNA192935, PRJNA233638, PRJNA247491, PRJNA263947, PRJNA261736, PRJEB6079, PRJEB6076, PRJEB2162, PRJNA188158, PRJNA208087 and PRJEB5500. The aCGH data used for validation is deposited in the NCBI GEO (https://www.ncbi.nlm.nih.gov/gds/) dataset under accession numbers GSE26170, GSE40210 and GSE58195. The processed CNV files, the phenotype tables, the structural variants file, the differentiated CNV file and the sample accessions file are available as Supplemental Data S1-5.

Acknowledgments

J.P. and E.A.O. were funded by the Intramural Program of the National Human Genome Research Institute of the National Institutes of Health. T.M.B was funded by ERC-CON-2019-864203, BFU2017-86471-P (MINECO/FEDER, UE), Howard Hughes International Early Career, Obra Social "La Caixa" and Secretaria d'Universitats i Recerca and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 880).

We thank EAO, Adam Boyko, Robert Wayne, and many other contributors to dog sequencing data for making the samples publically available. We thank A. Harris and A. Hogan for their help with experimental validations. We acknowledge the role of the Orthopedic Foundation for Animals, the American Kennel Club, and the Fédération Cynologique Internationale for the public availability of their data regarding dog breeds.

Author contributions: TMB, DJ, BWD, ISP and EAO designed the study and analyses; EAO and BWD provided a compendium of all the data; ASA, DJ and ISP performed most of the analyses;

CM and JP contributed to the analyses; TMB, ASA and DJ wrote the manuscript; all authors read and approved the manuscript.

Competing Interests

The authors declare no conflict of interest.

References:

Acland GM, Blanton SH, Hershfield B, Aguirre GD. 1994. XLPRA: a canine retinal degeneration inherited as an X-linked trait. *Am J Med Genet* **52**: 27–33.

Adegbola A, Musante L, Callewaert B, Maciel P, Hu H, Isidor B, Picker-Minh S, Le Caignec C, Delle Chiaie B, Vanakker O, et al. 2015. Redefining the MED13L syndrome. *Eur J Hum Genet* **23**: 1308–1317.

Agresti A. 2002. Categorical Data Analysis. *Wiley Series in Probability and Statistics*. http://dx.doi.org/10.1002/0471249688.

Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, Madeoy J, Nicholas TJ, Neff MW. 2010. Tracking footprints of artificial selection in the dog genome. *Proceedings of the National Academy of Sciences* **107**: 1160–1165. http://dx.doi.org/10.1073/pnas.0909918107.

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, et al. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**: 1061–1067.

Alvarez CE, Akey JM. 2012. Copy number variation in the domestic dog. *Mamm Genome* **23**: 144–163.

Ambrosini G, Groux R, Bucher P. 2018. PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. *Bioinformatics* **34**: 2483–2484.

American Kennel Club. 2007. *The Complete Dog Book: 20th Edition*. Ballantine Books.

Arendt M, Fall T, Lindblad-Toh K, Axelsson E. 2014. Amylase activity is associated with AMY2B copy numbers in dog: implications for dog domestication, diet and diabetes. *Anim Genet* **45**: 716–722.

Axelsson E, Ratnakumar A, Arendt M-L, Maqbool K, Webster MT, Perloski M, Liberg O, Arnemo JM, Hedhammar A, Lindblad-Toh K. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* **495**: 360–364.

Benaglia T. 2009. *Mixtools: An R Package for Analyzing Finite Mixture Models*.

Berglund J, Nevalainen EM, Molin A-M, Perloski M, André C, Zody MC, Sharpe T, Hitte C, Lindblad-Toh K, Lohi H, et al. 2012. Novel origins of copy number variation in the dog genome. *Genome Biol* **13**: R73.

Bickhart DM, Xu L, Hutchison JL, Cole JB, Null DJ, Schroeder SG, Song J, Garcia JF, Sonstegard TS, Van Tassell CP, et al. 2016. Diversity and population-genetic properties of copy number variations and multicopy genes in cattle. *DNA Res* **23**: 253–262.

Bigham AW. 2016. Genetics of human origin and evolution: high-altitude adaptations. *Curr Opin Genet Dev* **41**: 8–13.

Borge KS, Tønnessen R, Nødtvedt A, Indrebø A. 2011. Litter size at birth in purebred dogs—A retrospective study of 224 breeds. *Theriogenology* **75**: 911–919. http://dx.doi.org/10.1016/j.theriogenology.2010.10.034.

Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, Lohmueller KE, Zhao K, Brisbin A, Parker HG, vonHoldt BM, et al. 2010. A simple genetic architecture underlies morphological variation in dogs. *PLoS Biol* **8**: e1000451.

Brown KH, Dobrinski KP, Lee AS, Gokcumen O, Mills RE, Shi X, Chong WWS, Chen JYH, Yoo P, David S, et al. 2012. Extensive genetic diversity and substructuring among zebrafish strains revealed through copy number variant analysis. *Proc Natl Acad Sci U S A* **109**: 529–534.

Bunel M, Chaudieu G, Hamel C, Lagoutte L, Manes G, Botherel N, Brabet P, Pilorge P, André C, Quignon P. 2019. Natural models for retinitis pigmentosa: progressive retinal atrophy in dog breeds. *Hum Genet* **138**: 441–453.

Cadieu E, Neff MW, Quignon P, Walsh K, Chase K, Parker HG, Vonholdt BM, Rhue A, Boyko A, Byers A, et al. 2009. Coat variation in the domestic dog is governed by variants in three genes. *Science* **326**: 150–153.

Calboli FCF, Sampson J, Fretwell N, Balding DJ. 2008. Population structure and inbreeding from pedigree analysis of purebred dogs. *Genetics* **179**: 593–601.

Chain FJJ, Feulner PGD, Panchal M, Eizaguirre C, Samonte IE, Kalbe M, Lenz TL, Stoll M, Bornberg-Bauer E, Milinski M, et al. 2014. Extensive copy-number variation of young genes across stickleback populations. *PLoS Genet* **10**: e1004830.

Chase K, Jones P, Martin A, Ostrander EA, Lark KG. 2009. Genetic mapping of fixed phenotypes: disease frequency as a breed characteristic. *J Hered* **100 Suppl 1**: S37–41.

Chen W-K, Swartz JD, Rush LJ, Alvarez CE. 2009. Mapping DNA structural variation in dogs. *Genome Res* **19**: 500–509.

Clark J-ABJ, Whalen D, Marshall HD. 2016. Genomic analysis of gum disease and hypertrichosis in foxes. *Genet Mol Res* **15**. http://dx.doi.org/10.4238/gmr.15025363.

Coelho LP, Kultima JR, Costea PI, Fournier C, Pan Y, Czarnecki-Maulden G, Hayward MR, Forslund SK, Schmidt TSB, Descombes P, et al. 2018. Similarity of the dog and human gut microbiomes in gene content and response to diet. *Microbiome* **6**: 72.

Coren S. 1994. *The intelligence of dogs: canine consciousness and capabilities*.

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

Denholtz M, Bonora G, Chronis C, Splinter E, de Laat W, Ernst J, Pellegrini M, Plath K. 2013. Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and polycomb proteins in genome organization. *Cell Stem Cell* **13**: 602–616.

Derti A, Roth FP, Church GM, Wu C-T. 2006. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat Genet* **38**: 1216–1220.

Downs LM, Wallin-Håkansson B, Boursnell M, Marklund S, Hedhammar Å, Truvé K, Hübinette L, Lindblad-Toh K, Bergström T, Mellersh CS. 2011. A frameshift mutation in golden retriever dogs with progressive retinal atrophy endorses SLC4A3 as a candidate gene for human retinal degenerations. *PLoS One* **6**: e21452.

Dreger DL, Schmutz SM. 2011. A SINE insertion causes the black-and-tan and saddle tan phenotypes in domestic dogs. *J Hered* **102 Suppl 1**: S11–8.

Drögemüller C, Karlsson EK, Hytönen MK, Perloski M, Dolf G, Sainio K, Lohi H, Lindblad-Toh K, Leeb T. 2008. A mutation in hairless dogs implicates FOXI3 in ectodermal development. *Science* **321**: 1462.

Favier B, Rijli FM, Fromental-Ramain C, Fraulob V, Chambon P, Dollé P. 1996. Functional cooperation between the non-paralogous genes Hoxa-10 and Hoxd-11 in the developing forelimb and axial skeleton. *Development* **122**: 449–460.

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. 2014. Genome sequencing highlights the dynamic early history of dogs. *PLoS Genet* **10**: e1004016.

Freedman AH, Wayne RK. 2017. Deciphering the Origin of Dogs: From Fossils to Genomes. *Annu Rev Anim Biosci* **5**: 281–307.

Gaut BS, Seymour DK, Liu Q, Zhou Y. 2018. Demography and its effects on genomic variation in crop domestication. *Nat Plants* **4**: 512–520.

Gazave E, Darré F, Morcillo-Suarez C, Petit-Marty N, Carreño A, Marigorta UM, Ryder OA, Blancher A, Rocchi M, Bosch E, et al. 2011. Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. *Genome Res* **21**: 1626–1639.

Grieshammer U, Le Ma, Plump AS, Wang F, Tessier-Lavigne M, Martin GR. 2004. SLIT2-mediated ROBO2 signaling restricts kidney induction to a single site. *Dev Cell* **6**: 709–717.

Gundry RL, Allard MW, Moretti TR, Honeycutt RL, Wilson MR, Monson KL, Foran DR. 2007. Mitochondrial DNA analysis of the domestic dog: control region variation within and among breeds. *J Forensic Sci* **52**: 562–572.

Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Hinrichs AS, Gonzalez JN, et al. 2019. The UCSC Genome Browser database: 2019

update. *Nucleic Acids Res* **47**: D853–D858.

Hassan MQ, Tare R, Lee SH, Mandeville M, Weiner B, Montecino M, van Wijnen AJ, Stein JL, Stein GS, Lian JB. 2007. HOXA10 controls osteoblastogenesis by directly activating bone regulatory and phenotypic genes. *Mol Cell Biol* **27**: 3337–3352.

Hayward JJ, Castelhano MG, Oliveira KC, Corey E, Balkman C, Baxter TL, Casal ML, Center SA, Fang M, Garrison SJ, et al. 2016. Complex disease and phenotype mapping in the domestic dog. *Nat Commun* **7**: 10460.

Hegele RA. 2007. Copy-number variations and human disease. *Am J Hum Genet* **81**: 414–5; author reply 415.

Hitti RJ, Oliver JAC, Schofield EC, Bauer A, Kaukonen M, Forman OP, Leeb T, Lohi H, Burmeister LM, Sargan D, et al. 2019. Whole Genome Sequencing of Giant Schnauzer Dogs with Progressive Retinal Atrophy Establishes NECAP1 as a Novel Candidate Gene for Retinal Degeneration. *Genes* **10**: 385. http://dx.doi.org/10.3390/genes10050385.

Hu H. 1999. Chemorepulsion of neuronal migration by Slit2 in the developing mammalian forebrain. *Neuron* **23**: 703–711.

Hunt SE, McLaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, Parton A, Armean IM, Trevanion SJ, Flicek P, et al. 2018. Ensembl variation resources. *Database* **2018**. http://dx.doi.org/10.1093/database/bay119.

Irion DN. 2003. Analysis of Genetic Variation in 28 Dog Breed Populations With 100 Microsatellite Markers. *J Hered* **94**: 81–87.

Jagannathan V, Drögemüller C, Leeb T, Dog Biomedical Variant Database Consortium (DBVDC). 2019. A comprehensive biomedical variant catalogue based on whole genome sequences of 582 dogs and eight wolves. *Anim Genet* **50**: 695–704.

Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen C-A, Schmitt AD, Espinoza CA, Ren B. 2013. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**: 290–294.

Jones P, Chase K, Martin A, Davern P, Ostrander EA, Lark KG. 2008. Single-nucleotide-polymorphism-based association mapping of dog stereotypes. *Genetics* **179**: 1033–1044.

Kader A, Li Y, Dong K, Irwin DM, Zhao Q, He X, Liu J, Pu Y, Gorkhali NA, Liu X, et al. 2015. Population Variation Reveals Independent Selection toward Small Body Size in Chinese Debao Pony. *Genome Biol Evol* **8**: 42–50.

Karlsson EK, Lindblad-Toh K. 2008. Leader of the pack: gene mapping in dogs and other model organisms. *Nat Rev Genet* **9**: 713–725.

Karyadi DM, Karlins E, Decker B, vonHoldt BM, Carpintero-Ramirez G, Parker HG, Wayne RK, Ostrander EA. 2013. A copy number variant at the KITLG locus likely confers risk for canine squamous cell carcinoma of the digit. *PLoS Genet* **9**: e1003409.

Kim RN, Kim D-S, Choi S-H, Yoon B-H, Kang A, Nam S-H, Kim D-W, Kim J-J, Ha J-H, Toyoda A, et al. 2012. Genome analysis of the domestic dog (Korean Jindo) by

massively parallel sequencing. *DNA Res* **19**: 275–287.

Kobayashi A, Behringer RR. 2003. Developmental genetics of the female reproductive tract in mammals. *Nature Reviews Genetics* **4**: 969–980. http://dx.doi.org/10.1038/nrg1225.

Landis JR, Cooper MM, Kennedy T, Koch GG. 1979. A computer program for testing average partial association in three-way contingency tables (PARCAT). *Comput Programs Biomed* **9**: 223–246.

Le Béguec C, Wucher V, Lagoutte L, Cadieu E, Botherel N, Hédan B, De Brito C, Guillory A-S, André C, Derrien T, et al. 2018. Characterisation and functional predictions of canine long non-coding RNAs. *Sci Rep* **8**: 13444.

Li J, Yang T, Wang L, Yan H, Zhang Y, Guo Y, Pan F, Zhang Z, Peng Y, Zhou Q, et al. 2009. Whole Genome Distribution and Ethnic Differentiation of Copy Number Variation in Caucasian and Asian Populations. *PLoS One* **4**: e7958.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ 3rd, Zody MC, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**: 803–819.

Lippmann T, Jonkisz A, Dobosz T, Petrasch-Parwez E, Epplen JT, Dekomien G. 2007. Haplotype-defined linkage region for gPRA in Schapendoes dogs. *Mol Vis* **13**: 174–180.

MacLean EL, Snyder-Mackler N, vonHoldt BM, Serpell JA. 2019. Highly heritable and functionally relevant breed differences in dog behaviour. *Proc Biol Sci* **286**: 20190716.

Makino T, Rubin C-J, Carneiro M, Axelsson E, Andersson L, Webster MT. 2018. Elevated Proportions of Deleterious Genetic Variation in Domestic Animals and Plants. *Genome Biol Evol* **10**: 276–290.

Mansour TA, Lucot K, Konopelski SE, Dickinson PJ, Sturges BK, Vernau KL, Choi S, Stern JA, Thomasy SM, Döring S, et al. 2018. Whole genome variant association across 100 dogs identifies a frame shift mutation in DISHEVELLED 2 which contributes to Robinow-like syndrome in Bulldogs and related screw tail dog breeds. *PLoS Genet* **14**: e1007850.

Marchant TW, Johnson EJ, McTeir L, Johnson CI, Gow A, Liuti T, Kuehn D, Svenson K, Bermingham ML, Drögemüller M, et al. 2017. Canine Brachycephaly Is Associated with a Retrotransposon-Mediated Missplicing of SMOC2. *Curr Biol* **27**: 1573–1584.e6.

Marco-Sola S, Sammeth M, Guigó R, Ribeca P. 2012. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods* **9**: 1185–1188.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10. http://dx.doi.org/10.14806/ej.17.1.200.

Molin A-M, Berglund J, Webster MT, Lindblad-Toh K. 2014. Genome-wide copy number variant discovery in dogs using the CanineHD genotyping array. *BMC Genomics* **15**: 210.

Nicholas TJ, Baker C, Eichler EE, Akey JM. 2011. A high-resolution integrated map of copy number polymorphisms within and between breeds of the modern domesticated dog. *BMC Genomics* **12**: 414.

Ní Leathlobhair M, Perri AR, Irving-Pease EK, Witt KE, Linderholm A, Haile J, Lebrasseur O, Ameen C, Blick J, Boyko AR, et al. 2018. The evolutionary history of dogs in the Americas. *Science* **361**: 81–85.

Ostrander EA, Wang G-D, Larson G, vonHoldt BM, Davis BW, Jagannathan V, Hitte C, Wayne RK, Zhang Y-P, Dog10K Consortium. 2019. Dog10K: an international sequencing effort to advance studies of canine domestication, phenotypes and health. *Natl Sci Rev* **6**: 810–824.

Ostrander EA, Wayne RK. 2005. The canine genome. *Genome Res* **15**: 1706–1716.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**: 289–290. http://dx.doi.org/10.1093/bioinformatics/btg412.

Parker HG, Dreger DL, Rimbault M, Davis BW, Mullen AB, Carpintero-Ramirez G, Ostrander EA. 2017. Genomic Analyses Reveal the Influence of Geographic Origin, Migration, and Hybridization on Modern Dog Breed Development. *Cell Rep* **19**: 697–708.

Plassais J, Kim J, Davis BW, Karyadi DM, Hogan AN, Harris AC, Decker B, Parker HG, Ostrander EA. 2019. Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nat Commun* **10**: 1489.

Plump AS, Erskine L, Sabatier C, Brose K, Epstein CJ, Goodman CS, Mason CA, Tessier-Lavigne M. 2002. Slit1 and Slit2 cooperate to prevent premature midline crossing of retinal axons in the mouse visual system. *Neuron* **33**: 219–232.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**: 110–121.

Polley S, Cipriani V, Khan JC, Shahid H, Moore AT, Yates JRW, Hollox EJ. 2016. Analysis of copy number variation at DMBT1 and age-related macular degeneration. *BMC Med Genet* **17**: 44.

Quilez J, Martínez V, Woolliams JA, Sanchez A, Pong-Wong R, Kennedy LJ, Quinnell RJ, Ollier WER, Roura X, Ferrer L, et al. 2012. Genetic Control of Canine Leishmaniasis: Genome-Wide Association Study and Genomic Selection Analysis. *PLoS One* **7**: e35349.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Ramirez O, Olalde I, Berglund J, Lorente-Galdos B, Hernandez-Rodriguez J, Quilez J, Webster MT, Wayne RK, Lalueza-Fox C, Vilà C, et al. 2014. Analysis of structural diversity in wolf-like canids reveals post-domestication variants. *BMC Genomics* **15**: 465.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. *Nature* **444**: 444–454.

Rimbault M, Beale HC, Schoenebeck JJ, Hoopes BC, Allen JJ, Kilroy-Glynn P, Wayne RK,

Sutter NB, Ostrander EA. 2013. Derived variants at six genes explain nearly half of size reduction in dog breeds. *Genome Res* **23**: 1985–1995.

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**: 335–348.

Schoenebeck JJ, Hutchinson SA, Byers A, Beale HC, Carrington B, Faden DL, Rimbault M, Decker B, Kidd JM, Sood R, et al. 2012. Variation of BMP3 Contributes to Dog Breed Skull Diversity. *PLoS Genet* **8**: e1002849.

Serres-Armero A, Povolotskaya IS, Quilez J, Ramirez O, Santpere G, Kuderna LFK, Hernandez-Rodriguez J, Fernandez-Callejo M, Gomez-Sanchez D, Freedman AH, et al. 2017. Similar genomic proportions of copy number variation within gray wolves and modern dog breeds inferred from whole genome sequencing. *BMC Genomics* **18**: 977.

Shannon LM, Boyko RH, Castelhano M, Corey E, Hayward JJ, McLean C, White ME, Abi Said M, Anita BA, Bondjengo NI, et al. 2015. Genetic structure in village dogs reveals a Central Asian domestication origin. *Proc Natl Acad Sci U S A* **112**: 13639–13644.

Skoglund P, Ersmark E, Palkopoulou E, Dalén L. 2015. Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Curr Biol* **25**: 1515–1519.

Solé M, Ablondi M, Binzer-Panchal A, Velie BD, Hollfelder N, Buys N, Ducro BJ, François L, Janssens S, Schurink A, et al. Inter- and intra-breed genome-wide copy number diversity in a large cohort of European equine breeds. http://dx.doi.org/10.21203/rs.2.10580/v1.

Strillacci MG, Gorla E, Ríos-Utrera A, Vega-Murillo VE, Montaño-Bermudez M, Garcia-Ruiz A, Cerolini S, Román-Ponce SI, Bagnato A. 2019. Copy Number Variation Mapping and Genomic Variation of Autochthonous and Commercial Turkey Populations. *Front Genet* **10**. https://www.frontiersin.org/articles/10.3389/fgene.2019.00982/pdf (Accessed November 11, 2020).

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, et al. 2015. Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**: aab3761.

Sydney School of Veterinary Science. 2019. OMIA. *OMIA database*. https://omia.org/ (Accessed October 21, 2019).

Tsepilov YA, Ried JS, Strauch K, Grallert H, van Duijn CM, Axenovich TI, Aulchenko YS. 2013. Development and Application of Genomic Control Methods for Genome-Wide Association Studies Using Non-Additive Models. *PLoS One* **8**: e81431.

Upadhyay M, da Silva VH, Megens H-J, Visker MHPW, Ajmone-Marsan P, Bâlteanu VA, Dunner S, Garcia JF, Ginja C, Kantanen J, et al. 2017. Distribution and Functionality of Copy Number Variation across European Cattle Populations. *Front Genet* **8**: 108.

van Weerd JH, Badi I, van den Boogaard M, Stefanovic S, van de Werken HJG, Gomez-Velazquez M, Badia-Careaga C, Manzanares M, de Laat W, Barnett P, et al. 2014. A large permissive regulatory domain exclusively controls Tbx3 expression in the

cardiac conduction system. *Circ Res* **115**: 432–441.

Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, Sigurdsson S, Fall T, Seppälä EH, Hansen MST, Lawley CT, et al. 2011. Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS Genet* **7**: e1002316.

Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S. 2015. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep* **10**: 1297–1309.

Vilà C, Maldonado JE, Wayne RK. 1999. Phylogenetic relationships, evolution, and genetic diversity of the domestic dog. *J Hered* **90**: 71–77.

vonHoldt BM, Pollinger JP, Lohmueller KE, Han E, Parker HG, Quignon P, Degenhardt JD, Boyko AR, Earl DA, Auton A, et al. 2010. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* **464**: 898–902. http://dx.doi.org/10.1038/nature08837.

Waldo JT, Diaz KS. 2015. Development and validation of a diagnostic test for Ridge allele copy number in Rhodesian Ridgeback dogs. *Canine Genet Epidemiol* **2**: 2.

Wang C, Wallerman O, Arendt M-L, Sundström E, Karlsson Å, Nordin J, Mäkeläinen S, Pielberg GR, Hanson J, Ohlsson Å, et al. A new long-read dog assembly uncovers thousands of exons and functional elements missing in the previous reference. http://dx.doi.org/10.1101/2020.07.02.185108.

Wang X, Zhou B-W, Yin T-T, Chen F-L, Esmailizadeh A, Turner MM, Poyarkov AD, Savolainen P, Wang G-D, Fu Q, et al. 2018. Canine transmissible venereal tumor genome reveals ancient introgression from coyotes to arctic sled dogs. *Evolutionary Biology*.

Wellcome Trust Case Control Consortium, Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, et al. 2010. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**: 713–720.

Wilhelm D, Koopman P. 2006. The makings of maleness: towards an integrated view of male sexual development. *Nat Rev Genet* **7**: 620–631.

Wucher V, Legeai F, Hédan B, Rizk G, Lagoutte L, Leeb T, Jagannathan V, Cadieu E, David A, Lohi H, et al. 2017. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res* **45**: e57.

Xu L, Hou Y, Bickhart DM, Zhou Y, Hay EHA, Song J, Sonstegard TS, Van Tassell CP, Liu GE. 2016. Population-genetic properties of differentiated copy number variations in cattle. *Sci Rep* **6**: 23161.

Yalcin B, Wong K, Agam A, Goodson M, Keane TM, Gan X, Nellåker C, Goodstadt L, Nicod J, Bhomra A, et al. 2011. Sequence-based characterization of structural variation in the mouse genome. *Nature* **477**: 326–329.

Zhang W, Fan Z, Han E, Hou R, Zhang L, Galaverni M, Huang J, Liu H, Silva P, Li P, et al. 2014. Hypoxia adaptations in the grey wolf (Canis lupus chanco) from Qinghai-Tibet

Plateau. *PLoS Genet* **10**: e1004466.

Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. 2013. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14 Suppl 11**: S1.

Zoe N. Lye MDP. 2019. Copy Number Variation in Domestication. *Trends Plant Sci* **24**: 352–365.