Chromosome-length genome assembly and structural variations of the primal Basenji dog (*Canis lupus familiaris*) genome

Richard J. Edwards[1†] , Matt A. Field[2,3†], James M. Ferguson[4†], Olga Dudchenko[5,6,7†], Jens Keilwagen[8], Benjamin D. Rosen[9], Gary S. Johnson[10], Edward Rice[10], LaDeanna Hillier[11], Jillian M. Hammond[4], Samuel G. Towarnicki[1], Arina Omer[5,6], Ksenia Skvortsova[12,13], Ozren Bogdanovic[1,12], Robert A. Zammit1[14], Erez Lieberman Aiden[5,6,7,15,16*], Wesley C. Warren[10*], J. William O. Ballard[14*]

[†] Richard Edwards, Matt Field, James Ferguson and Olga Dudchenko contributed equally to this work.

* Correspondence: jwoballard@gmail.com, warrenwc@missouri.edu erez@erez.com

1 School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney NSW 2052, Australia. Richard.edwards@unsw.edu.au, o.bogdanovic@garvan.org.au, s.towanicki@unsw.edu.au

2 Centre for Tropical Bioinformatics and Molecular Biology, Australian Institute of Tropical Health and Medicine, James Cook University, Cairns, QLD 4878, Australia. matt.field@jcu.edu.au

3 John Curtin School of Medical Research, Australian National University, Canberra, ACT 2600, Australia. matt.field@jcu.edu.au

4 Kinghorn Center for Clinical Genomics, Garvan Institute of Medical Research, Victoria Street, Darlinghurst, NSW 2010, Australia. j.ferguson@garvan.org.au, j.hammond@garvan.org.au

5 The Center for Genome Architecture, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. Olga.Dudchenko@bcm.edu, Arina.Omer@bcm.edu, Zane.Colaric@bcm.edu, erez@erez.com

6 Department of Computer Science, Rice University, Houston, TX, USA. Olga.Dudchenko@bcm.edu, Arina.Omer@bcm.edu, Zane.Colaric@bcm.edu, erez@erez.com

7 Center for Theoretical and Biological Physics, Rice University, Houston, TX, USA. Olga.Dudchenko@bcm.edu, erez@erez.com

8 Julius Kühn-Institut, Erwin-Baur-Str. 27 06484 Quedlinburg, Germany. jens.keilwagen@julius-kuehn.de

9 Animal Genomics and Improvement Laboratory, Agricultural Research Service USDA, Beltsville, MD 20705. ben.rosen@usda.gov

10 College of Agriculture, Food & Natural Resources, University of Missouri, MO 65211 USA johnsongs@missouri.edu, Edward.rice@missouri.edu, warrenwc@missouri.edu

11 Genome Sciences, University of Washington, WA 98195 USA. lhillier@uw.edu

12 Genomics and Epigenetics Division, Garvan Institute of Medical Research, Victoria Street, Darlinghurst, NSW 2010, Australia. k.skvortsova@garvan.org.au, o.bogdanovic@garvan.org.au

13 St Vincent's Clinical School, Faculty of Medicine, University of New South Wales, Sydney, NSW, 2010, Australia, k.skvortsova@garvan.org.au

14 Vineyard Veterinary Hospital, 703 Windsor Rd, Vineyard, NSW, 2765. Robert@vineyardvet.com.au

15 Broad Institute of MIT and Harvard, Cambridge, MA, USA. erez@erez.com

16 Shanghai Institute for Advanced Immunochemical Studies, ShanghaiTech University, Shanghai, China. erez@erez.com

Abstract

Background: Basenjis are considered an ancient dog breed of central African origins that still live and hunt with tribesmen in the African Congo. Nicknamed the barkless dog, Basenjis possess unique phylogeny, geographical origins and traits make understanding their genome structure relative to more modern dog breeds of great interest. Here, we report the *de novo* assemblies of two Basenji: a female, China, and a male, Wags. We conduct pairwise comparisons and report structural variations between assembled genomes of three dog breeds: Basenji (CanFam_Bas), Boxer (CanFam3.1) and German Shepherd Dog (GSD) (CanFam_GSD). We then align representative whole genome sequences from 58 dog breeds and show the importance of genome reference when assessing variation among dog breeds.

Results: Here we present two high quality Basenji genome assemblies, CanFam_Bas (China) and Wags. CanFam_Bas is superior to CanFam v3,1 is terms of genome contiguity and comparable overall to the high quality CanFam_GSD assembly. The increasing number of available canid reference genomes allows us to examine the impact the choice of reference genome makes with regard to reference genome quality and breed relatedness. By aligning short read data from 58 representative dog breeds to three reference genomes, we demonstrate how the choice of reference genome significantly impacts both read mapping and variant detection. Further, we generate a conservative list of structural variant calls using a consensus of both Pacific Bioscience and Oxford Nanopore long reads to identify large structural breed differences. Collectively this work highlights the importance the choice of reference genome makes in canid variation studies.

Conclusions: The growing number of high-quality canid reference genomes means the choice of reference genome is an increasingly critical decision in subsequent canid variant analyses. The basal position of the Basenji makes it suitable for variant analysis for targeted applications of specific dog breeds. However, as is increasingly being employed in other

model organisms, we believe more comprehensive analyses across the entire family of canids is more suited to a pangenome approach.

Keywords: canine genome**,** domestication, comparative genomics, artificial selection

80

Background

Dogs were the first animals to be domesticated by humans some 30,000 years ago [1] and exhibit exceptional levels of breed variation as a result of extensive artificial trait selection [2]. It is not clear whether dogs were domesticated once or several times, though the weight of accumulating evidence suggests multiple events [3-9]. By establishing genome resources for more ancient breeds of dog, we can explore genetic adaptations perhaps unique to the modern dog breeds. Basenjis are an ancient breed that sits at the base of the currently accepted dog phylogeny [10]. Basenji-like dogs are depicted in drawings and models dating back to the Twelfth Dynasty of Egypt [11] and they share many unique traits with pariah dog types. Like dingoes and New Guinea Singing dogs (NGSD), Basenji's come oestrus annually—as compared to most other dog breeds, which have two or more breeding seasons every year. Basenji's, dingoes and NGSDs are prone to howls, yodels, and other vocalizations over the characteristic bark of modern dog breeds. One explanation for the unusual vocalisation of the Basenji is that the larynx is flattened [12]. The shape of the dingo and NGSD larynx is not reported.

Basenjis were originally indigenous to central Africa, wherever there was tropical forest. Primarily, what is now the DRC Congo, Southern Sudan, Central African Republic and the small countries on the central Atlantic coast. Today their territory has shrunk to the more remote parts of central Africa. The Basenji probably made its debut in the western world in around 1843. In a painting of three dogs belonging to Queen Victoria and Prince Albert entitled "Esquimaux, Niger and Neptune". Niger is clearly a Basenji. In total, 71 Basenji's have been exported from Africa and, to date, ~ 56 have been incorporated into the registered Basenji breeding population.

The first dog genome to be sequenced was of Tasha the Boxer [13], which was a tremendous advance and continues to be the resource guiding canine genomics research

today. The Boxer is a highly derived brachycephalic breed that has been subject to
110 generations of artificial selection. Further, due to its discontiguous sequence representation it has been difficult to accurately detect structural variations (SV's) in other domestic dog breeds. Now, a new generation of breed-specific chromosome-level genome reference assemblies are becoming available (5 breeds in October 2020 according to the NCBI assembly archive). For example, we previously published a chromosome-level German
115 Shepherd dog (GSD) genome assembly (CanFam_GSD) that is comprised of only 410 scaffolds and 716 contigs [14].

Here, we first report the sequence and *de novo* assembly of two Basenji genomes, female and male. We then compare these assemblies with the Boxer (CanFam3.1) and GSD (CanFam_GSD) [13, 14]. We conduct pairwise comparisons and report single-nucleotide
120 variants (SNVs) and structural variants (SV) between Basenji, Boxer [13] and GSD [14]. We distinguish an SNV as a variation in a single nucleotide without any limitations on its frequency. SV comprises a major portion of genetic diversity and its biological impact is highly variable. Chen et al. [15] used high-resolution array comparative genome hybridization to create the first map of DNA copy number variation (CNV) in dogs. Many
125 canine CNVs were shown to effect protein coding genes, including disease and candidate disease genes, and are thus likely to be functional. In this study, we find all types of genetic variation are impacted by the choice of reference genome. The basal position of the Basenji makes it useful as a general reference for variant analysis, but the generation of clade-specific genomes is likely to be important for canine nutrition and disease studies. We recommend a
130 pan-genome approach for comprehensive analyses of canid variation.

Results

Basenji female assembly, CanFam_Bas

The female Basenji, China (Fig. 1A), was initially assembled from 84.5 Gb Oxford Nanopore Technologies (ONT) PromethION reads (approx. 35X depth based on a 2.41 Gb genome size) using Flye (v2.6) [16, 17] and subjected to long read polishing with Racon v1.3.3 [18] and Medaka 1.03 [19] (Supplementary Fig 1A). Additional short read Pilon [20] error-correction was performed with 115.1 Gb (approx. 47.7X) BGIseq data. Hi-C proximity ligation was used with the DNA zoo pipeline [21-23] to scaffold 1,657 contigs into 1,456 scaffolds, increasing the N50 from 26.3 Mb to 63.1 Mb and decreasing the L50 from 33 to 14 (Figs 2 and 3, Supplementary Table 1). Scaffolds were gap-filled by PBJelly (pbsuite v.15.8.24) [24] using the ONT data, reducing the number of gaps from 348 to 148 and the number of scaffolds to 1,407. Following a final round of Pilon [20] BGIseq-polishing, scaffolds were mapped onto the CanFam3.1 [13] using PAFScaff v0.40 [25]. Diploidocus vector filtering removed one 5.7kb contig and masked a 3.3kb region of Chromosome X as lambda phage (J02459.1) contamination. Seven rounds of iterative Diploidocus tidying of the remaining sequences removed 277 (832 kb) as low coverage/quality and 481 (1.58 Mb) as probable haplotigs, retaining 483 core scaffolds and 165 probable repeat-heavy sequences [14] as China v1.0 (Supplementary Fig 2, Supplementary Table 2).

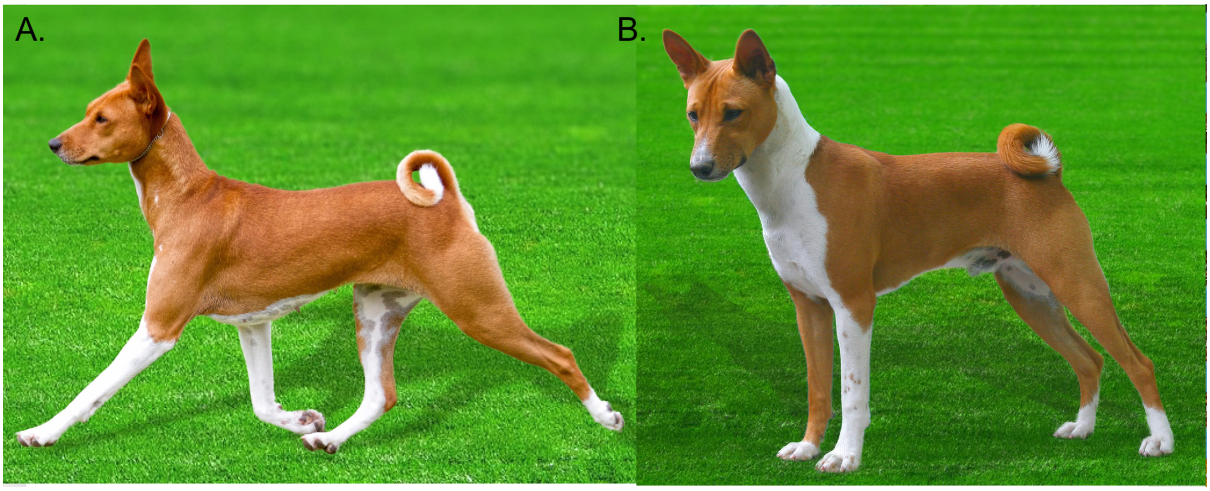

Figure 1. The Basenji dogs included in the study. A. China. Is registered as Australian Kennel Club Supreme Champion Zanzipow Bowies China Girl. Her registration is #2100018945. She was born in 2016 and she is free of all known genetic diseases. Her sire and dam are Australian bred and her most recent ancestor from Africa was 18 generations ago. **B**. Wags. is registered as American Kennel Club Champion Kibushi the Oracle, born in in 2008. His registration number is HP345321/01. His sire is an American bred dog while his dam was imported from the Haut-Ule district of the DRC congo, 3°24'04.0"N 27°19'04.6"E, in 2006 .

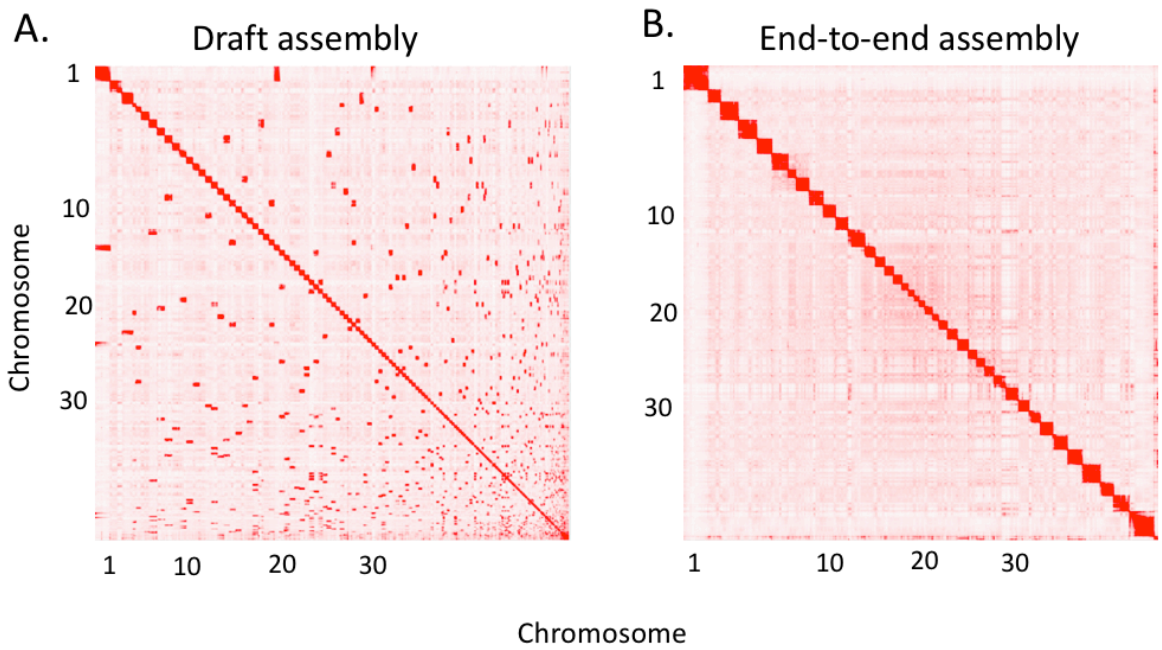

Figure 2. Contact matrices generated by aligning the CanFam_Bas (China) Hi-C data set to the genome assembly A. before the Hi-C upgrade (draft assembly). **B.** After Hi-C scaffolding (End-to-end assembly).

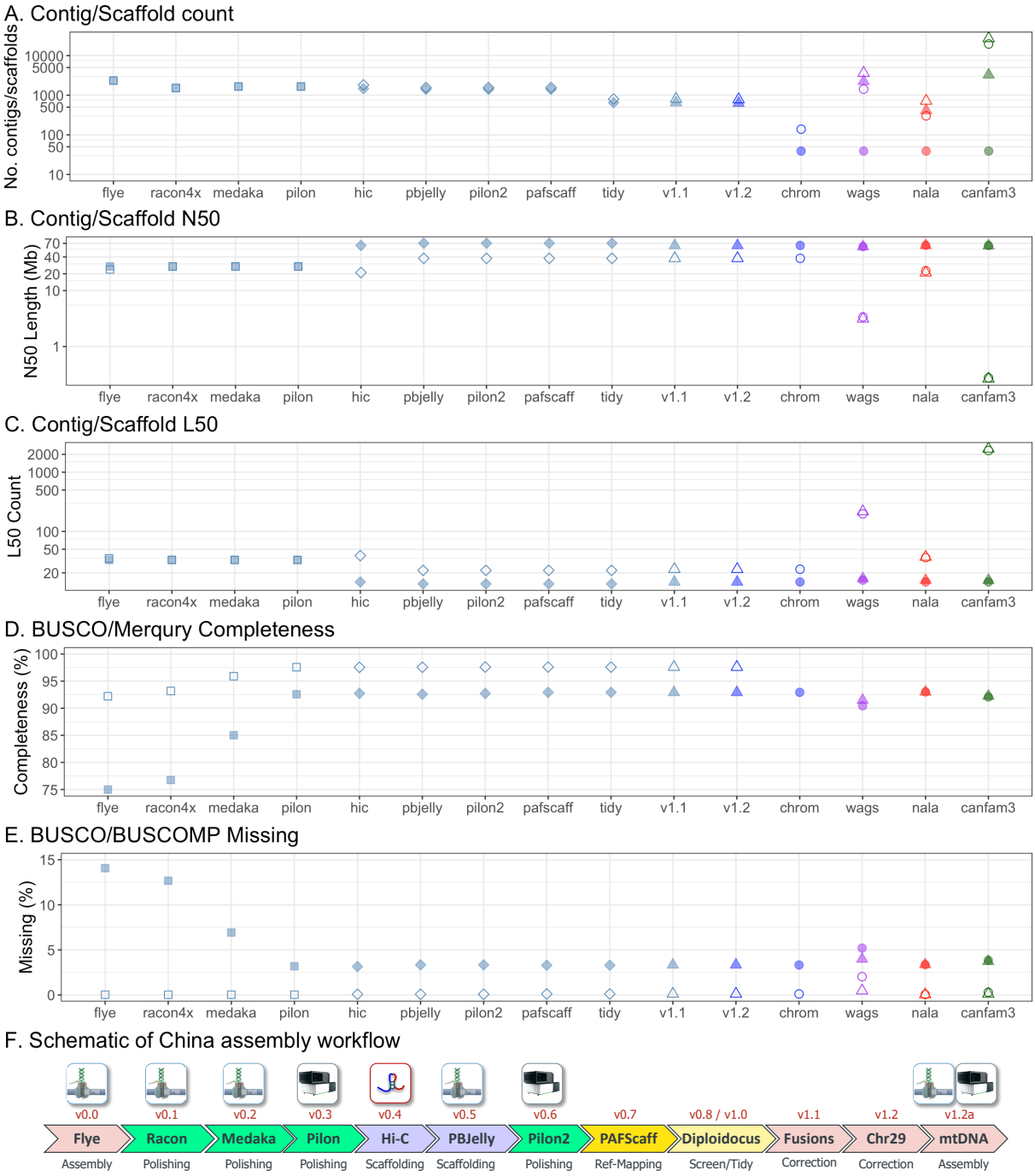

Figure 3. Key contiguity, quality and completeness metrics for different assembly stages and comparison dog genomes. Square, pre-scaffolded China; Diamond, scaffolded China; Triangle, complete assembly; Circle, main chromosome scaffolds only; Blue, China; Purple, Wags; Red, CanFam_GSD; Green, CanFam3.1. **A.** Contig (open) and scaffold (filled) numbers. **B.** Contig (open) and scaffold (filled) N50. **C.** Contig (open) and scaffold (filled) L50. **D.** Genome completeness estimated by BUSCO v3 (filled) and Merqury (open). **E.** The percentage of missing BUSCO genes (filled) and BUSCOMP genes (those found to be Complete in any assembly) (open). **F.** Schematic of China assembly workflow. CanFam_Bas is China v1.2.

Genome assembly correction. Two pairs of fused chromosomes in China v1.0 were incorrectly joined by PBJelly. Pre-gap-filled HiC scaffolds were mapped onto the assembly using Minimap2 v2.17 [26] and parsed with GABLAM v2.30.5 [27] to identify the gap corresponding to the fusion regions. These were manually separated into four individual chromosomes, gap lengths standardised, and scaffolds re-mapped onto CanFam v3.1 using PAFScaff v0.4.0. D-GENIES [28]. Analysis against CanFam_GSD chromosomes confirmed that PBJelly had incorrectly fused two pairs of chromosomes: chromosomes 8 with 13, and chromosome 18 with 30. These were manually separated and the assembly re-mapped onto CanFam 3.1 as China v1.1. PAFScaff assigned 112 scaffolds to chromosomes, including 39 nuclear chromosome-length scaffolds.

It was observed that the mitochondrial chromosome was missing and China v1.1 Chromosome 29 contained a 33.2 kb region consisting of almost two complete copies of the mitochondrial genome that were not found in other dog genome assemblies. The 26 ONT reads that mapped onto both flanking regions were reassembled with Flye v.2.7.1 [16, 17] into a 77.2 kb chromosome fragment, which was polished with Racon v1.3.3 [18] and Medaka 1.03 [19]. This was mapped back on to the Chromosome 29 scaffold with GABLAM v2.30.5 [27] (blast+ v2.9.0 megablast [29]) and the mitochondrial artefact replaced with the corrected nuclear mitochondrial DNA (NUMT) sequence. Finally, scaffolds under 1kb were removed to produce the China v1.2 nuclear genome.

Mitochondrial genome assembly. In total, 4,740 ONT reads (52.1 Mb) mapping on to mtDNA were extracted. To avoid NUMT contaminants, a subset of 80 reads (1.32 Mb) with 99%+ mtDNA assignment and 99%+ mtDNA coverage, ranging in size from 16,197 kb to 17,567 kb, were assembled with Flye 2.7b-b1526 [16, 17] (genome size 16.7 kb) into a 33,349 bp circular contig consisting of two mtDNA copies. This contig was polished with Racon [18] and Medaka [19], before being re-circularised to a single-copy matching the

CanFam3.1 mtDNA start position. After final Pilon [20] correction of SNPs and indels, the 16,761 bp mitochondrial genome was added to the China v1.2 nuclear assembly.

205 ***CanFam_Bas (China) reference genome.*** The resulting chromosome-length CanFam_Bas reference genome is 2,345,002,994 bp on 632 scaffolds with 149 gaps (76,431 bp gaps) (Table 1). The 39 nuclear plus mitochondrial chromosome scaffolds account for 99.3% of the assembly and show a high level of synteny with CanFam3.1 and CanFam_GSD (Fig 4). CanFam_Bas represents the most contiguous dog chromosomes to date, with a contig N50 of

210 37.8 Mb and contig L50 of 23, which is slight improvement over CanFam_GSD and considerably more contiguous than the standard dog reference genome, CanFam3.1 (Fig 3, Table 1). The completeness and accuracy of the genome as measured by BUSCO v3 [30] (laurasiatherian, *n*=6253) is also superior to CanFam3.1 and approaches that of CanFam_GSD (92.9% Complete, 3.75% Fragmented, 3.34% Missing).

215 **Methylomic identification of putative regulatory elements**

Additionally, we profiled whole genome methylation of Basenji's blood DNA using MethylC-seq [31]. The number of unmethylated (un) highly methylated (high) CpG sites in Basenji's genome were similar to that of GSD (Supplementary Fig. 3A). Importantly, high resolution DNA methylation data can be utilised to identify the putative regulatory elements

220 in a given tissue type. That is, CpG-rich unmethylated regions (UMRs) mostly correspond to gene promoters, while CpG-poor low-methylated regions (LMRs) correspond to distal regulatory elements, both characterised by DNAse I hypersensitivity [32] Using MethylSeekR algorithm [33], we performed the segmentation of Basenji DNA methylome and discovered 20,660 UMRs and 54,807 LMRs (Supplementary Fig. 3B,C), in line with

225 previously reported numbers in mammalian genomes [14, 33, 34].

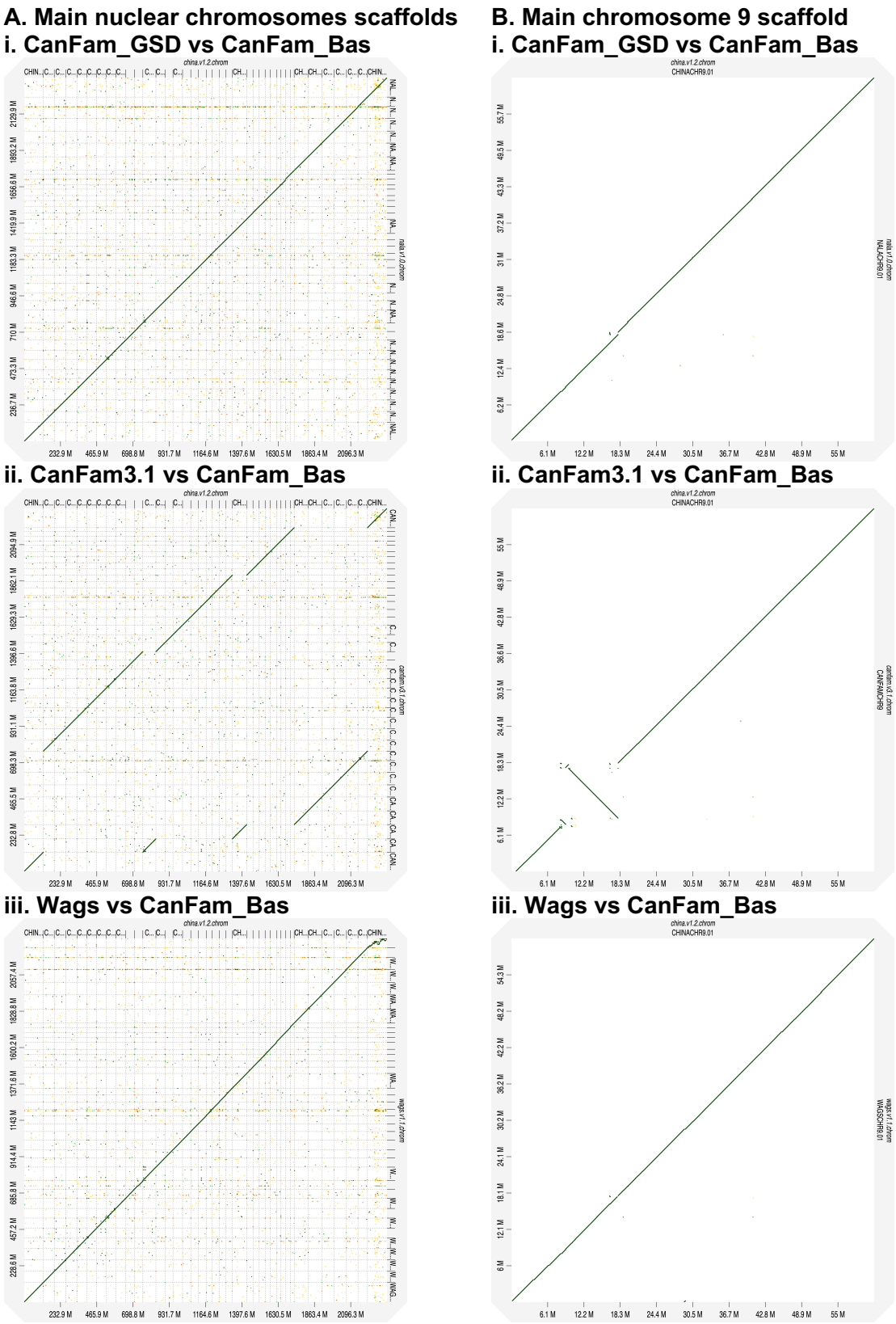

Figure 4. D-GENIES synteny plots of main chromosome scaffolds for three dog genome assemblies against CanFam_Bas. In each case, CanFam_Bas (China v1.2) is on the x-axis and the comparison assembly on the y-axis. Gridlines demarcate scaffolds. Thick black lines indicate regions of genomic alignment. **A**. All-by-all main chromosome scaffold alignments with **(i)** CanFam_GSD, **(ii)** CanFam_3.1, and **(iii)** Wags. **B**. Main chromosome 9 scaffold alignment with **(i)** CanFam_GSD, **(ii)** CanFam_3.1, and **(iii)** Wags.

Table 1. Genome assembly and annotation statistics for Basenji assemblies vs CanFam3.1 and CanFam_GSD.

	CanFam_Bas (China)	Wags	CanFam3.1	CanFam_GSD
Total sequence length	2,345,002,994	2,410,429,933	2,410,976,875	2,407,308,279
Total ungapped length	2,344,926,563	2,410,291,233	2,392,715,236	2,401,163,822
Number of scaffolds	632	2,243	3,310	430
Scaffold N50	64,291,023	61,087,166	45,876,610	64,346,267
Scaffold L50	14	16	20	15
Number of contigs	780	3,630	27,106	736
Contig N50	37,759,230	3,131,423	267,478	20,914,347
Contig L50	23	217	2,436	37
No. chromosomes	40	40	40	40
Percentage genome in main chromosomes	99.3%	94.8%	98.3%	96.5%
BUSCO complete (genome)	92.9% (1.14% Duplicated)	91.5% (1.31% Duplicated)	92.2% (1.17% Duplicated)	93.0% (1.38% Duplicated)
BUSCO fragmented (genome)	3.74%	4.53%	4.03%	3.73%
BUSCO missing (genome)	3.34%	3.98%	3.73%	3.37%
BUSCO complete (proteome)	98.5% (1.9% Duplicated)	97.8% (2.4% Duplicated)	95.1% (1.0% Duplicated)	98.9% (2.4% Duplicated)
BUSCO fragmented (proteome)	1.2%	1.5%	1.9%	1.0%
BUSCO missing (proteome)	0.3%	0.7%	3.0%	0.1%

235 **Male Basenji assembly, Wags**

For the male Basenji, Wags, (Fig 1B) we generated Pacific Bioscience Single Molecule Real Time (SMRT) sequences to approximately 45X genome coverage and assembled the genome to ungapped size of 2.41 Gb (Supplementary Fig 1B). Assembly contiguity metrics of 3,630 total contigs show N50 contig and scaffold lengths of 3.1 and 61 Mb length, respectively

240 (Table 1). Wags alignment to China revealed a high level of synteny. However, the Wags assembly of the X chromosome is smaller in size (59 Mb vs 125 Mb) and shows multiple rearrangements as a result of lower sequence coverage on the sex chromosomes (~21x).

Therefore, all comparative analyses reported herein were done with China, the female Basenji, CanFam_Bas. In addition, the Wags assembly includes 3.6 Mb of the Basenji dog Y
245 for future comparative studies of this unique chromosome.

Genome annotation

The CanFam_Bas and Wags assemblies were annotated using the homology-based gene prediction program GeMoMa v1.6.2beta [35] and nine reference species [14]. In total, CanFam_Bas (China) and Wags had similar numbers of predicted protein-coding genes at
250 27,129 (68,251 transcripts) and 27,783 (65,511) transcripts, respectively (Supplementary Table 3). Analysing the longest protein isoform per gene with BUSCO v3 [30] (laurasiatherian, *n*=6253, proteins mode), CanFam_Bas was measured to be 98.5 % complete (1.9% duplicated) and Wags was measured as 97.8% complete (2.4% duplicated). Both proteomes compare favourably with CanFam3.1 in terms of completeness (Table 1). To
255 correct for differences introduced by the annotation method, CanFam3.1 was annotated with the same GeMoMa pipeline. Approximately 90% of the Quest For Orthologues (QFO) reference dog proteome [36] is covered by each GeMoMa proteome, confirming comparable levels of completeness (Supplementary Table 3).

When the CanFam_Bas GeMoMa proteome was compared to Wags, CanFam3.1
260 and CanFam_GSD, over 91% genes had reciprocal best hits for at least one protein isoform (Supplementary Table 3). To investigate this further, the Wags, CanFam3.1 and CanFam_GSD genomes were mapped onto CanFam_Bas and the coverage for each gene calculated with Diploidocus v0.10.0. Of the 27,129 predicted genes, 26,354 (97.1%) are found at least 50% covered in all four dogs, whilst only 30 (0.11%) are completely unique to
265 CanFam_Bas. In total, Wags is missing 302 predicted genes, CanFam_GSD is completely missing 95 predicted genes, and CanFam3.1 is missing 211 predicted genes (Table 2). A considerably greater proportion of the missing genes in Wags (64.2% versus 11.4% in

CanFam3.1 and 15.8% in CanFam_GSD) were on the X chromosome. To test for artefacts due to assembly errors we mapped the long read data for Wags and CanFam_GSD onto
270 CanFam_Bas. Only 7 of the 302 missing Wags genes (2.3%) had no long read coverage, whilst 21/95 (22.1%) of genes missing in CanFam_GSD were confirmed by an absence of mapped long reads.

Table 2. Predicted copy numbers for CanFam_Bas GeMoMa genes based on A.
275 **assembly mapping, and B. long read mapping.**

A. Dog	Missing	Partial (<50%)	Single (1n)	Duplicate (2n)	3n+
CanFam_Bas	0 (0)	0 (0)	27129 (25788)	0 (0)	0 (0)
CanFam_Wags	302 (108)	120 (58)	26103 (25035)	486 (472)	118 (115)
CanFam3.1	211 (187)	167 (161)	26404 (25125)	251 (223)	96 (92)
CanFam_GSD	95 (80)	48 (42)	26586 (25304)	299 (266)	101 (96)

B. Data	0n	0.5n	1n	1.5n	2n	2.5n+
CanFam_Bas (ONT)	2 (2)	1257 (1201)	24116 (22954)	1508 (1403)	80 (68)	166 (160)
CanFam_Wags (PacBio)	7 (2)	4717 (3476)	21140 (21049)	1028 (1024)	79 (79)	158 (158)
CanFam_GSD (ONT)	21 (18)	1893 (1795)	22412 (21350)	2503 (2349)	109 (98)	191 (178)

Figures in brackets exclude predicted genes on X chromosome.

Amylase copy number

Two copies of the *Amy2B* gene were identified in a tandem repeat on Chromosome 6 of the
280 CanFam_Bas assembly. The single-copy read depth for CanFam_Bas, calculated as the modal read depth across single copy complete genes identified by BUSCO v3 [30], was estimated to be 34X. This was verified by BUSCO complete genes, which gave mean predicted copy numbers of 1.008±0.005 (95% C.I.) (Supplementary Fig 4). The two complete *Amy2B* coding sequence copies had a mean depth of 97.5X, equating to 2.87N, or a total

285 copy number estimate of 5.78N (2 x 97.5 / 34). The full CanFam_GSD *Amy2B* repeat region was also found in two complete copies with a mean depth of 98.1X, estimating 5.77 copies (2 x 98.1 / 34). Similar results were obtained restricting analysis to reads at least 5kb (6.01 gene copies; 5.98 region copies) or 10kb (6.18 gene copies; 6.04 region copies) to minimise repeat-based fluctuations in read depth. In contrast, droplet digital PCR (ddPCR) estimated
290 that China had 4.5 copies per individual chromosome. This slight difference suggests that the raw sequence data slightly overestimated copies or the ddPCR primers did not capture all the genetic variation.

Wags assembly has a single copy of the *Amy2B* region, which includes 90% of the *Amy2B* coding sequence (data not shown). Single-copy depth analysis of Wags estimates
295 4.97 (90% at 253.8X) and 5.51 (100% at 253.5X) copies of the *AMY2B* coding sequence and tandem repeat unit, respectively.

Nuclear mitochondrial DNA fragments

During the assembly of China, the mitochondrial genome was erroneously assembled into a nuclear mitochondrial DNA (NUMT) fragment on chromosome 29. A blastn search (e < 10-
300 4) identified 291 putative NUMTs, ranging in size from 34 bp to 6,580 bp (Table 3). These total 190.5 kb (approx. 11.4 mtDNA copies) and span the entire mtDNA with reasonably even coverage and no obvious bias (Supplementary Fig. 5A). All 291 NUMTs are well-supported with at least 3 reads that span the entire NUMT plus at least 5 kb each side (Supplementary Table 4). Only 1 NUMT was not found to be full-length in CanFam_GSD
305 (Supplementary Fig. 5B). An additional 26 NUMTs are partially covered in CanFam3.1 and 9 are entirely absent. Whilst this could represent a breed difference, 19 of the 35 additional incomplete NUMTs in CanFam3.1 (65.5%) are also incomplete in Wags, whilst Wags has a further ten incomplete NUMTs that are present in CanFam3.1 (Supplementary Table 4). This

is consistent with these regions being generally harder to assemble and/or align. Further
310 analyses of these regions may provide insight into domestic dog genealogies.

Whole genome assembly comparisons

To discover unique large-scale structural differences in assembled genomes of the three breeds – Basenji, German Shepard and Boxer – we performed pairwise alignments of CanFam_Bas, CanFam3.1 and CanFam_GSD. Overall, genome synteny was maintained and
315 there were limited large scale genomic rearrangements observed (Fig 4A). There was, however a large inversion in CanFam3.1 that was not present in CanFam_Bas or CanFam_GSD (Fig 4B). To investigate this further, we aligned CanFam_Bas against Wags (Fig 4A iii). As expected there was no inversion on Chromosome 9 (Fig 4B iii), suggesting the inversion or perhaps an assembly error, occurs in CanFam3.1.

320 **Long read structural variant detection**

To generate a conservative set of high-quality structural variants, the consensus of ONT and SMRT long read sequences data from both Basenji (China ONT and Wags SMRT) and GSD (Nala ONT and SMRT) was mapped onto Basenji (CanFam_BAS), Boxer (CanFam3.1) and GSD (CanFam_GSD) reference genomes (Supplementary Table 5). One difference in the
325 data sets is that both ONT and SMRT reads from GSD were sourced from the same individual dog whereas for Basenji the ONT and SMRT reads are from different individual Basenji dogs. These high-quality SVs were additionally annotated for their overlap to both genes and exons. Of these high-quality ONT/SMRT consensus SVs, the Basenji long reads overlap 814 CanFam3.1 exons and 568 CanFam_GSD exons while the GSD long reads
330 overlap 825 CanFam3.1 exons and 495 CanFam_Bas exons. In total, these SVs represent 92.19 Mb, 97.21 Mb, and 78.99 Mb of deleted sequence and 4.11 Mb, 4.09 Mb, and 7.69 Mb of inserted sequence for the CanFam_Bas, CanFam3.1, and CanFam_GSD assemblies respectively.

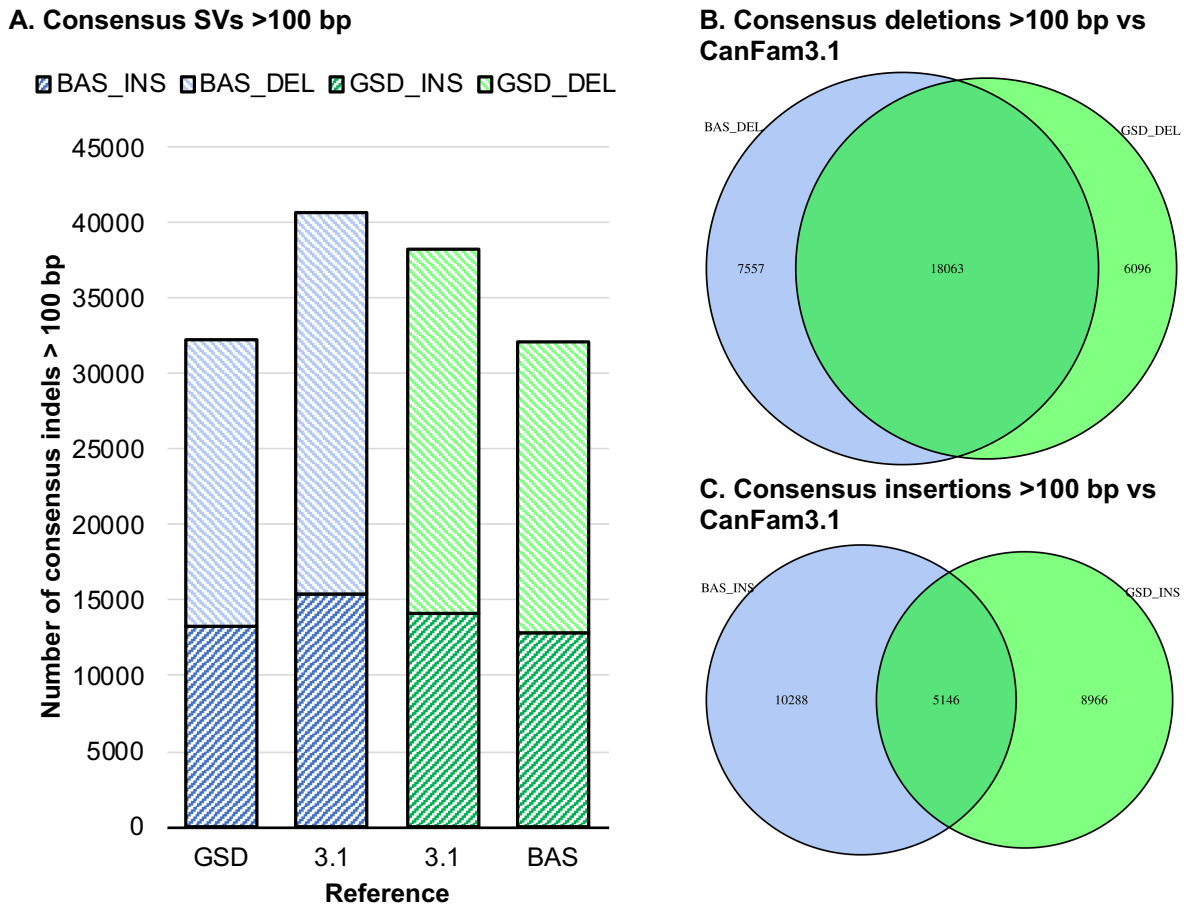

Figure 5. Consensus structural variant calls for combined ONT and PacBio data. High-quality set of consensus structural variant (SV) calls generated from the intersection of the ONT and PacBio SV calls for each breed versus reference comparison, limited to SVs >100 bp long. **A.** Total numbers of SVs called from Basenji CanFam_Bas reads (red) versus CanFam3.1 and CanFam_GSD, and German Shepherd long reads (green) versus CanFam3.1 and CanFam_Bas. **B.** Numbers and overlap of consensus deletion calls for Basenji reads (blue) and GSD reads (green) versus CanFam3.1. **B.** Numbers and overlap of consensus insertion calls for Basenji reads (blue) and GSD reads (green) versus CanFam3.1.

To reduce the impact of small indels arising from local mis-assembly errors, the high-quality consensus SVs were further reduced to those over 100 bp in length (Fig 5). Overall, we observe similar number of total SV calls from Basenji long reads relative to CanFam_GSD and GSD long reads relative to CanFam_BAS. Both breeds had a substantially larger number of consensus SV calls against the CanFam3.1 reference overall, with Basenji long reads generating more SVs calls than GSD long reads. We next overlapped the CanFam SV calls relative to CanFam v3.1 and found 18,063 long read deletion calls overlapped between Basenji and GSD. For Basenji this represented 70.00% of the total 25,260 Basenji

deletions while for GSD this represented 73.25% of the total 24,138 GSD deletions (Fig. 5B). Insertions were fewer in number and degree of overlap, however we still found 5,146 overlapping insertions between Basenji and GSD long reads representing 33.33% of the total 15,434 Basenji insertions and 36.46% of the total 14,111 GSD insertions (Fig. 5C). The high degree of overlap in SVs from GSD and Basenji relative to CanFam v3.1 represent Boxer-specific SVs or potential issues with the current canid reference assembly.

Short read mapping, SNV / small indel detection

Mapping multiple individuals to a reference genome for variant calling is standard practice in population genomics and is known to be prone to biases in the reference genome. To investigate whether differences identified in variant analyses were due to the Basenji being a basal breed or due to assembly quality difference, short read data from 58 different dog breeds belonging to sixteen different well-supported clades from Parker et al. [10] (Supplementary Table 6) were mapped on to three reference genomes Basenji (CanFam_BAS), Boxer (CanFam3.1) and GSD (CanFam_GSD). Large-scale structural differences between breeds would be expected to significantly affect read mapping efficiencies for closely-related versus distantly-related breeds, whilst missing assembly data would be expected to result in a systematic reduction in mapping across all breeds. In our analysis, we observe such systematic and breed specific changes in both the number of mapped reads and variants detected.

For the systematic changes, overall trends are exhibited in the total percentage of reads mapped across the three references, with the highest percent of reads mapping to CanFam_GSD, followed closely by CanFam_BAS and the lowest percent of reads mapping to CanFam3.1 (Fig 6A). ANOVA shows these differences are significant ($F_{2, 171}$=819.53, $P<0.0001$). To investigate this result further and test for interactions we focused upon breeds within each of the monophyletic clades close to or associated with the three reference

genomes. Specifically, we included the short-read sequences from the four breeds in the Asian Spitz clade as this close to the basal Basenji lineage. We also included the six breeds European Mastif clade containing the Boxer and three breeds within the New World clade containing the GSD. Overall, the CanFam_Bas and CanFam_GSD performed equally well while the relative mapping was lowest for CanFam3.1 (Fig 6B). Once again ANOVA shows this result is significant ($F_{8, 30}$=32.01, P<0.0001). Next, we considered the capacity of each reference genome to detect SNV's and indels. In this case CanFam_BAS detected higher number of changes than did either CanFam3.1 or CanFam_GSD (Fig 6C and D; ANOVA $F_{2, 171}$=30.71, and $F_{2, 171}$= 12.08, respectively with P<0.0001 for each). In combination these data attest to the quality of the CanFam_Bas assembly. There is a noticeable depletion of variant calls for the reference breed (Supplementary Fig 6), but there were no significant interactions between clades and reference genome (data not shown).

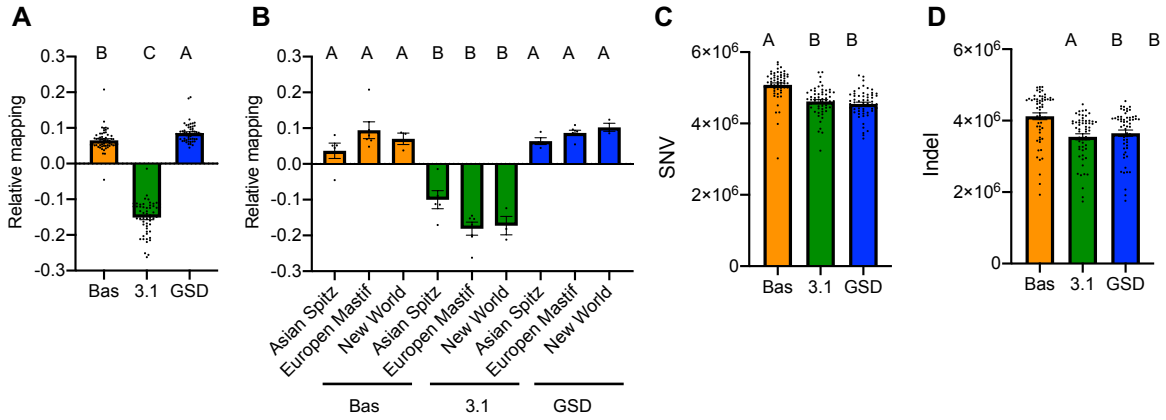

Figure 6. Comparative short read mapping and single nucleotide variant calling for 58 dog breeds versus three reference genomes: CanFam_Bas (Bas), CanFam3.1 (3.1) and CanFam_GSD (GSD). A. Relative percentage of reads mapped to each reference calculated by subtracting the mean percentage of reads for the three reference genomes from the number of reads for each reference. **B.** Relative percentage of reads mapped to each reference for three closely related clades determined from Parker et al. [13]. The Basenji is closely related to the Asian Spitz clade. The Boxer is a member of the European Mastiff clade and the GSD is a member of the New world clade. **C.** The number of SNVs. **D.** The number of small indels.

Discussion

In this manuscript we present a reference-quality assembly of a female Basenji (China), which we designate CanFam_Bas. We also present a second Basenji genome, Wags. The Wags build is of high-quality but not as good as CanFam_Bas, partly because Wags was a male. In total, 64.2% of the missing genes are on the X chromosome, compared to under 16% in the other two individuals. Analysis of long read data mapped onto the CanFam_Bas genome annotation further supports assembly issues as a contributor, with only two autosomal genes lacking any coverage in Wags SMRT data. Equivalently, two other autosomal genes also lacked any coverage in China ONT data, suggesting that even high quality assemblies can have missing genes in an assembly that may not be biologically lost.

CanFam_Bas is considerably more contiguous and complete than CanFam3.1 (Boxer breed) and slightly more contiguous (Contig N50 37.8 kb vs 20.9 kb) but has slightly lower completeness (BUSCO completeness 92.9% vs 93.0%) than CanFam_GSD [14] (Table 1, Supplementary Table 1). The ungapped nuclear chromosomes lengths of CanFam_Bas, CanFam3.1, and CanFam_GSD and 2.33 Gb, 2.32 Gb, 2.36 Gb and, respectively. Likely, both assembly quality and breed differences contribute to the 211 missing genes in CanFam3.1. CanFam_GSD had fewer predicted genes missing (95) but a higher percentage (22.0%) also lacking long read coverage. This uncertainty notwithstanding, our analysis highlights the need to consider multiple reference genomes in a pan-genomic approach when a comprehensive analysis is required. These results also highlight the need for considering missing genes carefully on a case-by-case basis as, even with high quality genome assemblies such as those compared in this study, an apparent absence may reflect missing data rather than missing biology.

We noted that the mitochondrial chromosome was missing from the China 1.0 Basenji assembly. Exhaustive search of the genome detected a 33.2kb region consisting of almost two

complete copies of the mitochondrial genome on Chromosome 29 that was not present in other dog assemblies. The real NUMT fragment was 4,302 bp in size and a total of 291 (190.5 kb) were detected across the genome. NUMTs are troublesome insertions that are universal in nuclear DNA, but they are not ubiquitous in nuclear genome assemblies [37]. Very often NUMTs are edited out during the genome annotation process [38], but it is becoming increasingly clear that that such organelle insertions must be properly accounted for in nuclear genome assemblies as they have the potential to provide unique insight into population histories and animal well-being [39]. In domestic cats, the nuclear cat genome includes 7.9 kb fragment that is duplicated in tandem 38–76 times [40]. In humans NUMTs are polymorphic at 141 loci. NUMT polymorphisms have been used to estimate the age of of insertions in human and other primate lineages [39] while five insertions have been implicated in human disease [37, 41].

Sequence analyses detected six copies of *Amy2B* in Chia and five in Wags. Previous studies have shown that Basenji's may have 4-18 copies [6]. This high variation in *Amy2B* copy number suggests at least three evolutionary histories of the gene in Basenji's. First, *Amy2B* copies may have differentially accumulated in specific lineages since the divergence from other dog breeds. Second, the ancestral founding population of the modern Basenji may have been polymorphic for *Amy2B.* Third, it remains possible that the *Amy2B* region from other breeds have differentially introgressed into the modern breed dog. Basenji-like dogs are depicted in In Egyptian drawings and models dating back to 1550-1900 BC [11]. Possibly, the Basenji is derived from the Abuwtiyuw, which was a lightly built hunting dog with erect ears and a curly tail that died before 2280BC [42]. Likely, sequencing the region around *Amy2B* in Basenji's with a higher copy numbers will aid resolving the alternate hypotheses. This large genetic diversity in Basenji *Amy2B* copy has important dietary and veterinary implications. Veterinarians frequently recommend a rice based diet following complex

intestinal surgeries and following diarrhoea because it is thought to be bland and easily digestible. However, dogs with low *Amy2B* copy numbers may not be able to digest rice
450 readily as their serum amylase levels are lower [43].

CanFam_Bas and CanFam_GSD represent two high-quality reference genomes from different breeds. The availability of corresponding Basenji and GSD long read data, provided an excellent opportunity to further investigate the contributions of breed and assembly differences, as real SVs will be represented in the raw read data even if assembly
455 errors have introduced false positives or negatives. Possessing ONT and SMRT data for both GSD and Basenji allows us to overlap the disparate SV call sets to generate a conservative list of SV calls relative to CanFam3.1. This analysis identified over 70,000 SVs in CanFam_Bas relative to CanFam3.1 and over 64,000 SVs in GSD relative to CanFam3.1 (Supplementary Table 5). There is a high degree of overlap in SVs from GSD and Basenji
460 relative to CanFam v3.1 (68% of total basenji SV calls and 77% of GSD SV calls), which highlights potential issues with the current canid reference assembly. Further, each consensus set contains several hundred SVs overlapping annotated exons, highlighting the importance of the selection of appropriate reference genome for analysis of specific genomic regions.

465 Next, we examined the overlap of the consensus calls for SVs over 100bp of GSD and Basenji relative to CanFam v3.1. We find a high degree of agreement for deletions, with 70.1% of GSD calls and 74.75% of Basenji calls overlapping (Fig 5B), compared to insertions where 33.3% of basenji calls and 36.4% of calls overlap (Fig 5C). Basenji and GSD long reads have 25.9% and 19.1% more SVs called when CanFam3.1 is used as the
470 reference, respectively, and the high degree of overlap is consistent with Boxer-specific SVs and/or assembly issues. GSD-specific and Basenji-specific SV calls are also evident.

To get additional insight into breed differences and the influence of reference genome selection, we mapped short read data from 58 breeds onto CanFam_Bas, CanFam3.1, and CanFam_GSD. If assembly differences are dominated by real differences between breeds then we might expect different breeds to map with different efficiencies onto the three genomes. Differences to due quality, on the other hand, should be reflected across all breeds. With some minor exceptions, short read data from the different breeds consistently mapped better onto CanFam_GSD than CanFam_Bas, which was in turn better than CanFam3.1 (Fig 6A). This is consistent with assembly quality being a dominating factor.

After adjusting for read mapping difference, different reference genomes also produce different SNV and small indel densities for the 58 mapped breeds (Supplementary Fig 6). Using a Basenji reference genome consistently identifies more variants than either GSD or Boxer (Fig 6, Supplementary Fig 6). This probably reflects the basal position of Basenji in the breed phylogeny [10]. The basal position of the Basenji makes it useful as a reference for variant analysis as there are clear biases affecting related breeds seen for both the GSD and Boxer reference genomes. On the other hand, CanFam_GSD has slightly higher read mapping levels across breeds, which may provide better total coverage.

Together, these data suggest that a single high-quality reference should be sufficient for most general analyses, but the generation of clade specific genomes is likely to be important for canine nutrition and disease studies. Fanconi Syndrome has been reported in several dog breeds, but the vast majority (75%) have occurred in Basenjis. Approximately, 10-30% of Basenji's in North America are affected. The most severe and common ailment in Basenjis is Fancioni Syndrome, in which the renal tubes fail to reabsorb electrolytes and nutrients [44]. In 2011, it was shown that Basenji Fanconi Syndrome is caused by a 370bp deletion on canine chromosome 3 [45]. To date no other breeds have been recoded with this same deletion, although other mutations can cause the disease. Likely complex diseases in

dogs belonging to different clades may have different underlying causes. For example, it remains unclear whether the same suite of mutations causes hip dysplasia in the GSD (New World clade) and the Labrador (Retriever/ Spaniel clade).

500 ## Conclusions

Here, we present two high quality *de novo* Basenji genome assemblies: CanFam_Bas (China, female) and Wags (male). CanFam_Bas offers improved genome contiguity relative to CanFam v3.1 and can serve as a representative basal breed in future canid studies. We generate core genomic information for the Basenji that has the potential to inform future
505 studies of population history and aid disease management. We generate high-quality variants (SV's, SNVs, small indels, and SVs) relative to CanFam_Bas, CanFam v3.1, and CanFam_GSD. We demonstrate the impact that the reference genome makes on both read mapping and variant detection, illustrating the importance of either selecting the appropriate reference genome or employing a pan-genome approach in future canid studies.

510 ## Methods

Sequencing and genome assembly of female Basenji, China

Sampling. China or Zanzipow Bowies China Girl is an Australian Supreme Champion Kennel Club show dog. Her Australian National Kennel Council registration number is 2100018945. She is bred primarily from Australian lines, with her most recent common
515 ancestor coming from Africa 18 generations previously. She was born on 14 Jan 2016 and is free from all known currently reported diseases.

Sequencing. High molecular weight DNA was extracted from 100 µl of blood using the DNeasy Blood and Tissue kit (Qiagen). For long read (Oxford Nanopore) sequencing, 1 µg of DNA was prepared via the Genomic DNA by Ligation kit (SQK-LSK109) as per the
520 manufacturer's protocol. Prepared DNA (180 ng) was loaded onto a PromethION (FLO-

PRO002) flowcell and sequenced with standard parameters. After 48 hours, a nuclease flush was performed and an additional 115 ng was loaded onto the run. Base-calling was performed after sequencing with the GPU-enabled guppy basecaller (v3.0.3) using the PromethION high accuracy flip-flop model with config 'dna_r9.4.1_450bps_hac.cfg'.

525 For short read sequencing whole blood was shipped to BGI, Hong Kong. High molecular weight DNA was extracted, a paired-end library prepared, and the sample run on the BGISEQ-500 sequencing system to generate high quality PE100 data. A total of 767,111,208 clean reads (115.1Gb) were produced with a lower base call accuracy (Q20) of 95.92%.

530 ***Assembly.*** An overview of the China assembly workflow is given in Supplementary Fig 1A. The ONT reads were assembled with the Flye (v2.6-release) assembler [16, 17]. The resulting contigs were polished with ONT reads using four rounds of Racon (v1.4.3) [18] followed by Medaka (v0.10.0) [19] to minimise error propagation. BGI-seq reads were aligned to the polished assembly with BWA-mem (v 0.7.17) [46] and Pilon (v1.23) (diploid)

535 [20] was used for further error correction. A second assembly was performed using Canu assembler v1.8.0 [47] and correcting the sequencing errors using 2 rounds of Arrow polishing [48]. The Flye assembly was considered more contiguous and therefore selected as the core assembly.

Scaffolding. The ONT assembly was scaffolded to chromosome-length by the DNA Zoo

540 following standard DNA zoo methodology (www.dnazoo.org/methods). Briefly, an *in situ* Hi-C library was prepared [23] from a blood sample from the same individual and sequenced to ~30x coverage (assuming 2.4 Gb genome size). The Hi-C data was processed using Juicer [49], and used as input into the 3D-DNA pipeline [50] to produce a candidate chromosome-length genome assembly. We performed additional finishing on the scaffolds using Juicebox

545 Assembly Tools [21]. The matrices are visualized in Juicebox.js, a cloud-based visualization

system for Hi-C data [22] and are available for browsing at multiple resolutions. After scaffolding, all ONT reads were aligned to the assembly with Minimap2 (v2.16) (-ax map-ont) [26] and used by PBJelly (pbsuite v.15.8.24) [24] to fill gaps. BGI reads were re-mapped with BWA-mem (v 0.7.17) and another round of polishing was performed with Pilon (v1.23) (diploid, SNP and indel correction) [20] .

Final clean-up. The Pilon-polished genome underwent a final scaffold clean-up to generate a high-quality core assembly following Field and colleagues' processing of Canfam_GSD [14]. Scaffolds were reoriented and assigned to CanFam3.1 chromosomes using PAFScaff v0.3.0 [14] (Minimap2 v2.16 mapping) [26]. Diploidocus v0.9.0 [51] was used to screen contamination, remove low-coverage artefacts and haplotig sequences, and annotate remaining scaffolds with potential issues as described in [14]. ONT reads were mapped onto the assembly using Minimap2 v2.17 (-ax map-ont --secondary=no) [26] and read depth summaries calculated with BBMap v38.51 pileup.sh [52]. Any scaffolds with median coverage less than three (e.g., less than 50% of the scaffold covered by at least three reads) were filtered out as low-coverage scaffolds. Single-copy read depth was estimated using the modal read depth of 35X across the 5,5578 single copy complete genes identified by BUSCO v3.0.2b [30]. This was used to set low-, mid- and high-depth thresholds for PurgeHaplotigs v20190612 [53] (implementing Perl v5.28.0, BEDTools v2.27.1 [54, 55], R v3.5.3, and SAMTools v1.9 [56] at 8X, 25X and 68X. PurgeHaplotig coverage was adjusted to exclude gap regions and any scaffolds with 80%+ bases in the low/haploid coverage bins and 95%+ of their length mapped by PurgeHaplotigs onto another scaffold were filtered as haplotigs or assembly artefacts. Any other scaffolds with 80%+ low coverage bases were filtered as Low Coverage. Remaining scaffolds were further classified based on read depth profiles. Scaffolds with <20% diploid coverage and 50%+ high coverage were marked as probable collapsed repeats. Scaffolds with "Diploid" depth as the dominant PurgeHaplotigs

coverage bin and >50% match to another scaffold were marked as a possible repeat sequences [14].

Genome assembly correction. Main chromosome scaffold integrity was checked using D-GENIES [28] comparisons of different assembly stages with CanFam_GSD chromosomes [14]. Two pairs of fused chromosomes were identified following incorrect joins made by PBJelly (pbsuite v.15.8.24) [24]. Pre-gap-filled HiC scaffolds were mapped onto the assembly using Minimap2 v2.17 [26] and parsed with GABLAM v2.30.5 [27] to identify the gap corresponding to the fusion regions. These were manually separated into four individual chromosomes, gap lengths standardised, and scaffolds re-mapped onto CanFam v3.1 using PAFScaff v0.4.0 [25]. Finally, scaffolds under 1kb were removed following correction of Chromosome 29 (below).

Correction of mitochondrial insertion into Chromosome 29. NUMT analysis identified a 33.2 kb region consisting of almost two complete copies of the mitochondrial genome, not present in other dog genome assemblies. GABLAM v2.30.5 [27] was used to confirm that this region was also absent from the Canu assembly of *China*. ONT reads that mapped onto both flanking regions of the 33.2 kb putative NUMT were extracted and reassembled with Flye v2.7.1 [16, 17]. The new NUMT was approx. 2.8 kb long. To avoid repeated problems with mitochondrial reads mis-polishing the sequence, a subset of 4.82M ONT reads (72.7 Gb, ~30X) was extracted and mapped onto the assembled region with Minimap 2.17 [26]. Reads mapping to at least 5 kb of the assembled region including some immediate flanking sequence were extracted (66 reads, 1.50 Mb) and polished with one round of Racon v1.4.5 [18] (-m 8 -x -6 -g -8 -w 500) and Medaka v0.7.1 [19] (model r941_prom_high). The polished NUMT region was mapped on to the Chromosome 29 scaffold with GABLAM v2.30.5 [27] (blast+ v2.9.0 [57] megablast) and stretches of 100% sequence identity

595 identified each side of the NUMT. The mtDNA sequence between these regions of identity was replaced with the re-assembled NUMT sequence.

Mitochondrial genome assembly. To assemble the mitochondrion, ONT reads were mapped onto a construct of three tandem copies of the CanFam3.1 mtDNA with minimap2 v2.17 [26] (-ax map-ont --secondary=no). Reads with hits were extracted using SAMTools v1.9 [56]
600 fasta and mapped onto a double-copy CanFam3.1 mtDNA with GABLAM v2.30.5 [27] (blast+ v2.9.0 [57] megablast). "Pure complete" mtDNA ONT reads were identified as those with 99%+ of their length mapping to mtDNA and 99%+ coverage of the mtDNA. These reads were assembled with Flye 2.7b-b1526 [16, 17] (genome size 16.7 kb) and polished with Racon v1.4.5 [18] (-m 8 -x -6 -g -8 -w 500) followed by Medaka v0.7.1 [19] (model
605 r941_prom_high). The polished mtDNA assembly was mapped onto CanFam3.1 mtDNA with GABLAM v2.30.5 [27] (blast+ v2.9.0 [57] megablast) and circularised by extracting a region from the centre of the assembly corresponding to a single full-length copy with the same start and end positions. Final correction of SNPs and indels was performed by adding the mtDNA to the nuclear assembly, mapping BGI reads with BWA v0.7.17 and polishing
610 the mtDNA with Pilon v1.23 [20]. The polished mtDNA was then added back to the nuclear genome for the final *China* assembly.

Genome assembly quality assessment. At each stage of the assembly, summary statistics were calculated with SLiMSuite SeqList v1.45.0 [58], quality was assessed with Merqury v20200318 (Meryl v20200313, bedtools v2.27.1 [54, 55], SAMTools v1.9 [56], java v8u45,
615 igv v2.8.0) and completeness assessed with BUSCO v3.0.2b [30] (BLAST+ v2.2.31 [57], HMMer v3.2.1 [59], Augustus v3.3.2, EMBOSS v6.6.0, laurasiatherian lineage (n=6253)). To account for fluctuations in BUSCO ratings, presence of complete BUSCO genes across assembly stages was also assessed with BUSCOMP v0.9.4 [60]. Final assembly scaffold statistics and quality assessment was performed with Diploidocus v0.10.2 (KAT v2.4.2, perl

620 v5.28.0, BEDtools v2.27.1 [54, 55], SAMTools v1.9 [56], purge_haplotigs v20190612, java v8u231-jre, bbmap v38.51, minimap2 v2.17 [26], BLAST+ v2.9.0 [57]).

DNA methylation calling. China's blood DNA methylome libraries were sequenced on the Illumina HiSeq X platform (150 bp, PE), generating 336 million read pairs and yielding 14x sequencing coverage. Sequenced reads were trimmed using Trimmomatic [61] and mapped

625 to the China v1.0 genome reference using WALT [62] with the following parameters: -m 10 -t 24 -N 10000000 -L 2000. The mappability of the MethylC-seq library was 86%. Duplicate reads were removed using Picard Tools v2.3.0. Genotype and methylation bias correction were performed using MethylDackel with additional parameters: minOppositeDepth 5 --maxVariantFrac 0.5 --OT 10,140,10,140 --OB 10,140,10,140. The numbers of methylated

630 and unmethylated calls at each CpG site were determined using MethylDackel (https://github.com/dpryan79/MethylDackel). Bisulphite conversion efficiency was 99.71%, estimated using unmethylated lambda phage spike-in control.

UMR and LMR calling. Segmentation of basenji's blood DNA methylome into CpG-rich unmethylated regions (UMRs) and CpG-poor low-methylated regions (LMRs) was

635 performed using MethylSeekR [33] (segmentUMRsLMRs(m=meth, meth.cutoff=0.5, nCpG.cutoff=5, PMDs = NA, num.cores=num.cores, myGenomeSeq=build, seqLengths=seqlengths(build), nCpG.smoothing = 3, minCover = 5).

Sequencing and genome assembly of male Basenji, Wags

Wags DNA was derived from blood of a single male. He is registered as American Kennel

640 Club Champion Kibushi the Oracle, born on December 3, 2008. His registration number is HP345321/01. Sire is AM Ch C-Quests Soul Driver, HM827502/02, and his dam is Avongara Luka, HP345312/01, a native bitch imported from the Haut-Ule district of the DRC Congo, 3°24'04.0"N 27°19'04.6"E, in 2006. SMRT sequences for Wags (Fig. 1B) were generated on the Pacific Biosciences Sequel instrument (V2 chemistry) to approximately 45x

645 genome coverage based on a genome size estimate of 2.5 Gb. An overview of the China assembly workflow is given in Supplementary Fig 1B. All SMRT sequences were assembled with the HGAP4 algorithm, a Falcon based assembly pipeline available through the SMRT Link interphase (SMRT Link v5.0.1.9585) [63]. The assembly was then error corrected with the original SMRT sequences using the Arrow error-correction module [63]. Additional
650 polishing of the assembly for residual indels was done by aligning 32x coverage of Illumina data and the Pilon algorithm [20]. Chromosomal level scaffolds were generated with the same DNA source using the Proximo^{TM} Hi-C genome scaffolding software (Phase Genomics Inc) and finalized by alignment to the CanFam3.1 reference.

Locus copy number estimation

655 Copy numbers for specific assembly regions were calculated using Diploidocus v0.10.0 (runmode=regcnv) [51]. For each animal, long reads were mapped onto the assembly with Minimap2 v2.17 (no secondary mapping) [26] and the modal read depth across single-copy Complete genes identified by BUSCO v3 [30] (laurasiatherian_db) calculated using SAMTools v1.9 [56] mpileup. This set the expected single-copy read depth, X_{SC}. Copy
660 number for a region, N_{reg} was estimated by dividing the mean read depth across that region, X_{reg}, by X_{SC}. The variance and standard deviation of the estimate was calculated using X_{reg} for all single copy BUSCO genes. For Wags (a male), genes on the X chromosome were excluded from this analysis.

Amylase copy number. The copy number of the beta amylase gene *Amy2B* was calculated
665 using Diploidocus v0.10.0 (runmode=regcnv) [51] using a modification of the single locus copy number estimation (above) to account for multiple copies of the gene in the assembly. First, the AMY2B protein sequence from CanFam3.1 (UniprotKB: J9PAL7) was used as a query and searched against the genome with Exonerate v2.2.0 [64] to identify assembled copies of the *Amy2B* gene. A full-length (14.8 kb) *Amy2B* gene repeat was also extracted

670 from the CanFam_GSD assembly and mapped onto the assembly with Minimap2 v2.17 (-x asm5) [26]. Estimated N_{reg} values were converted into a number of copies by multiplying by the proportion of the query found covered by that region. The total genome *Amy2B* copy number was then calculated as the summed copy number over each hit. To further investigate the robustness of the method and improve the *Amy2B* copy number estimate in China,
675 analysis was repeated with ONT reads at least 5 kb in length and at least 10 kb in length. These reads should be less susceptible to poor mapping at repeat sequences, but at a cost of reduced coverage.

We also used droplet digital PCR (ddPCR) to directly quantify the *Amy2B* copy number of China [65]. ddPCR was performed using a QX100 ddPCR system (Bio-rad). Each
680 reaction was performed in a 20µl reaction volume containing 10 µl of 2x ddPCR Supermix (Bio-rad), 1µl of each 20x primer/probe, 1µl of DraI restriction enzyme (New England BioLabs #R0129S), 5µl of DNA template (4ng/µl) and 2µl ddH_2O. Primer sequence for Amy2B: forward 5′-CCAAACCTGGACGGACATCT-3′ and reverse 5′-TATCGTTCGCATTCAAGAGCAA-3′ with FAM probe: 6FAM–
685 TTTGAGTGGCGCTGGG-MGBNFQ. Primer sequence for C7orf28b-3: 5′-GGGAAACTCCACAAGCAATCA-3′ and reverse 5′-GAGCCCATGGAGGAAATCATC-3′ with HEX probe HEX-CACCTGCTAAACAGC-MGBNFQ.

Nuclear mitochondrial DNA (NUMT) fragment analysis

To make sure that any nuclear mitochondrial DNA fragments (NUMTs) were identified as a
690 single region, a double-copy CanFam3.1 mtDNA sequence was constructed and then searched against the CanFam_Bas nuclear genome and compressed to unique hits using GABLAM v2.30.5 [27] (blast+ v2.9.0 [57] blastn, localunique) with a blastn e-value cut-off of 1e-4 [29]. Predicted copy number was calculated for each NUMT in *China* using the method described above. Diploidocus v0.10.0 was also used to calculate number of reads

695 spanning each entire NUMT plus flanking regions of 0 bp, 100 bp, 1 kb and 5 kb. In addition, assembly coverage for each NUMT was calculated for Wags, CanFam3.1 and CanFam_GSD. Each genome was split into 1 Mb tiled fragments and mapped onto CanFam_Bas with Minimap2 v2.17 [26]. Each BAM file was used for Diploidocus v0.10.0 regcnv [51] analysis with a single-copy read depth of 1X. Mitochondrial genome coverage was analysed by
700 extracting all 291 NUMT regions with SeqSuite v1.23.3 [66] and mapping them onto the CanFam3.1 mtDNA chromosome using GABLAM v2.30.5 [27] (blast+ v2.10. [57] tblastn).

Genome annotation

Following Field et al. [14] each genome was annotated using the homology-based gene prediction program GeMoMa [35] (v1.6.2beta) and nine reference organisms. For each
705 reference organism, coding exons of full-length transcript were extracted and translated to peptides using the GeMoMa module Extractor. These peptides were searched in each Basenji genome using mmseqs2 [67] (v5877873cbcd50a6d954607fc2df1210f8c2c3a4b). Based on the results of mmseqs2 and Extractor, transcripts were predicted for each Basenji genome from each reference organism independently. These nine gene annotation sets were then
710 combined into a final gene annotation using the GeMoMa module GAF. To make a fair comparison of the influence of genome quality and completeness on annotation, CanFam3.1 was annotated with the same pipeline. Annotation of CanFam_GSD using the same pipeline was obtained from Field et al. [14].

Annotation summary and quality assessment. Annotation summary statistics and the longest
715 protein isoform per gene were generated with SAAGA v0.4.0 [68]. Annotation completeness was estimated using BUSCO v3 [30] (laurasiatherian, *n*=6253, proteins mode), run on a reduced annotation consisting of the longest protein per gene. To check for truncated or fragmented protein predictions, predicted proteins were mapped onto the Quest For Orthologues reference dog proteome [36] with mmseqs2 v [67]. The best protein hit for each

720 gene was used to calculate a protein length ratio (length of predicted protein / length of reference protein). Percentage coverage of query and hit proteins was calculated with mmseqs2 v [67]. A reciprocal search identified the closest predicted protein for each reference protein. Any reciprocal best hits were marked as predicted orthologues.

Annotation copy number and coverage analysis. Predicted copy number was calculated for

725 every protein-coding gene in CanFam_Bas using the method described above. In addition, assembly coverage for each CanFam_Bas gene was calculated for Wags, CanFam3.1 and CanFam_GSD. Each genome was split into 1 Mb tiled fragments and mapped onto CanFam_Bas with Minimap2 v2.17 (-ax asm5 -L) [26]. Each BAM file was used for Diploidocus v0.10.0 regcnv analysis with a single-copy read depth of 1X. In addition, SMRT

730 reads from Wags and ONT reads from the GSD were mapped onto CanFam_Bas with Minimap2 v2.17 (--secondary=no -L -ax map-pb or -ax map-ont) [26] and the standard predicted copy number calculation applied. Genes with zero coverage were marked 0n. Other genes were binned according to coverage: for mapped assemblies, coverage was rounded to the nearest integer; for long read mapping, coverage was rounded to the nearest 0.5n. Genes

735 with greater than zero but less than 50% coverage were assigned to 0.5n. Any genes exceeding a rounded coverage of 2n were grouped as “3+”.

Ribosomal RNA prediction. For each genome, genes for rRNA were predicted with Barrnap v0.9 [69] in the eukaryotic mode, implementing Perl v5.28.0, HMMer v3.2.1 [59] and BEDTools v2.27.1 [54, 55].

740 **Whole genome assembly comparisons**

Whole genome synteny analysis were performed for the main chromosome scaffolds using the D-GENIES [28] web portal.

Long read structural variant detection

Structural variant calls were generated using a combination of minimap2 v2.17-r943-dirty

745 [26], SAMTools v1.9 [56], and sniffles v1.0.11 [70]. In total, four sets of long reads from three samples were analyzed consisting of China the Basenji (Oxford Nanopore), Wags the Basenji (SMRT) and Nala the German Shepherd (Oxford Nanopore and SMRT). Variants for the Basenji and Nala were annotated with gene model predictions generated using GeMoMa [35] (v1.6.2beta) while CanFam3.1 variants were annotated with Ensembl gene annotations
750 v100 for CanFam v3.1 [71]

Short read mapping, SNV / small indel detection

Representative Illumina data was identified from [72]. All SRA IDs associated with the DBVDC bioproject on NCBI (SRP144493) were downloaded along with their metadata and reduced to 126 samples representing the biggest sequencing run (no. bases) per annotated
755 breed. These 126 samples were used for initial read mapping and variant calling (Supplementary Table 7). Following removal of unknown/mixed/village dog samples, canids other than domestic dogs, and duplicates the remaining breeds were mapped onto those used by Parker et al. [10]. In total, 58 breeds in the significantly monophyletic clades (greater than 70% bootstrap) designated by Parker et al. [10] were considered for analyses (Supplementary
760 Table 6).

SNVs and small indels were called from the Illumina reads of the 58 representative breeds against three reference genomes (Basenji China v1.0, CanFam3.1, and CanFam_GSD). All Illumina reads were downloaded from the short read archive using the SRA toolkit v2.10.4 [73]. All samples were analysed using a modified version of an existing
765 variant detection pipeline [74]. Briefly, the pipeline employs the Burrows-Wheeler Aligner v0.7.17-r1188 (BWA) for read alignment [75], SAMTools v1.9 [56] and picard v2.4.1 (http://broadinstitute.github.io/picard) for binary alignment map (BAM) file preprocessing, and Genome Analysis ToolKit v3.6 (GATK) for calling SNVs and small indels [76]. The workflow follows GATK best-practices using default parameters throughout. Additionally,

770 SRA reads from each reference genome were aligned using BWA and SAMTools and a consensus variant list generated using an approach described previously [77]. The consensus variant lists for each reference genome were utilized in GATK's BaseRecalibrator step as the '-knownSites' argument to serve as the required variant truthset. Variant alignment statistics were generated using SAMTools flagstat. Joint variant calls were generated for the larger
775 dataset of 126 samples relative to the three reference genomes and the total number of SNVs and small indels for individual samples tabulated. Variants for the Basenji and CanFam_GSD were annotated with gene model predictions generated using GeMoMa (see above) while CanFam variants were annotated with ENSEMBL gene annotations v100 for CanFam v3.1. Read mapping statistics for each sample were calculated using SAMTools to remove
780 secondary mapping and then summarized with v38.51 pileup.sh [52]. Numbers were then converted into relative values for each reference by averaging the score for each breed over the three reference genomes and then calculating the difference from the mean. Breeds were considered individually, but mean values for each clade were also calculated. Three clades are of particular importance as they are closely related, or include, the three reference
785 genome assemblies. The Asian Spitz clade is considered closely related to the Basenji. This clade contained the Alaskan Malamute, Shar-Pei, Shiba Inu and Tibetan Mastiff. The European Mastiff Clade contained the Boxer, Bull Terrier, Cane Corseo, Great Dane, Mastiff and Rhodesian Ridgeback. The New World clade contained the Berger Picard, Chinook and German Shepherd. One-way ANOVA's were employed to detect significant differences
790 between groups. As the same short read samples were examined relative to the three reference genomes statistical significance was set to be $P<0.01$.

Acknowledgements

We would like to thank Jennifer Power for introducing us to China and Barbara Reisinger, Jon Curby and Jared Reisinger for access to Wags. For China, Brooke Morgan-Burke at the

Vineyard Veterinary Hospital provided constant encouragement. ONT data were collected at the Garvan Institute and the nuclease flush by Jillian Hammond, the BGI -seq data was provided by BGI and the Hi-C data collected at Baylor College of Medicine. For Wags, DNA sequencing was completed at the The Hudson Alpha Institute and assembly at the The McDonnell Genome Institute, Washington University School of Medicine.

Funding

This work was supported by the the University of New South Wales/ School of Biotechnology and Biomolecular Sciences Genomics Initiative and the Basenji Health Endowment Inc, Poynette, WI and. The DNA Zoo initiative of E.A.L. funded the Hi-C data collection and analyses. M.F. is funded by NHMRC APP5121190. E.L.A. was supported by an NSF Physics Frontiers Center Award (PHY1427654), the Welch Foundation (Q-1866), a USDA Agriculture and Food Research Initiative Grant (2017-05741), an NIH 4D Nucleome Grant (U01HL130010), and an NIH Encyclopaedia of DNA Elements Mapping Center Award (UM1HG009375).

Author contributions

J.W.O.B. coordinated the project. R.J.E., M.A.F., and J.W.O.B. designed the study. J.W.O.B and W.W funded the project. For China, R.A.Z. provided the sample. J.F. and B.D.R. performed the initial assembly and polishing, O.D., A.O. and Z.C. performed the Hi-C experiment, and O.D. and E.L.A. conducted the Hi-C analyses. R.J.E performed the final assembly polishing, curation, correction, analyses of genome completeness, mitochondrial genome assembly, NUMT analysis, and read-depth gene copy number analyses. S.G.T. completed the digital droplet PCR. K.S. and O.B. conducted the DNA methylation analyses. M.A.F. performed the structural comparisons, dog breed short read mapping and variant calling. M.A.F., R.J.E. and J.W.O.B. performed the dog breed variant analysis. E.R., L.H., G.S.J., and W.W. performed the Wags assembly. J.K. performed the genome annotation.

820 R.J.E. M.F., M.A.F, B.D.R., O.D., O.B. and J.W.O.B. wrote the manuscript. All authors edited and approved the final manuscript.

Ethics declarations

Ethics approval and consent to participate

For China, all experimentation was performed under the approval of the University of New
825 South Wales Ethics Committee (ACEC ID: 18/18B).

Competing interests

JF has received travel and accommodation expenses to speak at Oxford Nanopore Technologies conferences. Otherwise, the authors declare that the submitted work was carried out in the absence of any professional or financial relationships that could potentially
830 be construed as a conflict of interest.

Supplementary information

Supplementary information accompanies this paper at:
http://www.slimsuite.unsw.edu.au/research/basenji/

Availability of data and materials

835 Chromosome-length genome assemblies for both Basenji are available at NCBI (CanFam_Bas (*China*), GCA_013276365.1; *Wags*, GCA_004886185.2). The CanFam_Bas mitochondrial genome is available at NCBI Genbank: MW051511. To review GEO accession GSE159396: Go to
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE159396
840 Enter token ylclqswyvhitlcr into the box

References

1. 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.
2. Drake AG, Klingenberg CP. Large-scale diversification of skull shape in domestic dogs: disparity and modularity. Am Nat. 2010;175:289-301.
3. Thalmann O, Shapiro B, Cui P, Schuenemann VJ, Sawyer SK, Greenfield DL, Germonpre MB, Sablin MV, Lopez-Giraldez F, Domingo-Roura X, et al. Complete mitochondrial genomes of ancient canids suggest a European origin of domestic dogs. Science. 2013;342:871-874.
4. 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. Similar genomic proportions of copy number variation within gray wolves and modern dog breeds inferred from whole genome sequencing. BMC Genomics. 2017;18:977.
5. Ostrander EA, Wayne RK, Freedman AH, Davis BW. Demographic history, selection and functional diversity of the canine genome. Nat Rev Genet. 2017;18:705-720.
6. 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.
7. Zhang S-j, Wang G-D, Ma P, Zhang L-l, Yin T-T, Liu Y-h, Otecko NO, Wang M, Ma. Y-p, Wang L, et al. Genomic regions under selection in the feralization of the dingoes Nat Comm. 2020;11:671.
8. Wang GD, Zhai W, Yang HC, Wang L, Zhong L, Liu YH, Fan RX, Yin TT, Zhu CL, Poyarkov AD, et al. Out of southern East Asia: the natural history of domestic dogs across the world. Cell Res. 2016;26:21-33.
9. Gopalakrishnan S, Sinding MS, Ramos-Madrigal J, Niemann J, Samaniego Castruita JA, Vieira FG, Caroe C, Montero MM, Kuderna L, Serres A, et al. Interspecific gene flow shaped the evolution of the genus *Canis*. Curr Biol. 2018;28:3441-3449 e3445.
10. 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.
11. Dollman G. The Basenji dog. J R Africa Soc. 1937;36:148-149.
12. Ashdown RR, Lea T. The larynx of the Basenji dog. J Small Anim Pract. 1979;20:675-679.
13. 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;438:803-819.
14. Field MA, Rosen BD, Dudchenko O, Chan EKF, Minoche AE, Barton K, Lyons RJ, Tuipulotu DE, Edwards RJ, 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;accepted 14/02/20.
15. Chen WK, Swartz JD, Rush LJ, Alvarez CE. Mapping DNA structural variation in dogs. Genome Res. 2009;19:500-509.
16. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37:540-546.

17. Lin Y, Yuan J, Kolmogorov M, Shen MW, Chaisson M, Pevzner PA. Assembly of long error-prone reads using de Bruijn graphs. Proc Natl Acad Sci U S A. 2016;113:E8396-E8405.
18. Vaser R, Sovic I, Nagarajan N, Sikic M. Fast and accurate *de novo* genome assembly from long uncorrected reads. Genome Res. 2017;27:737-746.
19. ONT. Sequence correction provided byONT research [Online].https://github.com/nanoporetech/medaka**.** 2018.
20. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9:e112963.
21. Dudchenko O, M.S. S, Batra SS, Durand NC, Musial NT, Mostofa R, Pham M, St Hilaire BG, Yao W, Stamenova E, et al. The Juicebox Assembly Tools module facilitates *de novo* assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv. 2018:1-26.
22. Robinson JT, Turner D, Durand NC, Thorvaldsdottir H, Mesirov JP, Aiden EL. Juicebox.js provides a cloud-based visualization system for Hi-C data. Cell Syst. 2018;6:256-258 e251.
23. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665-1680.
24. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, Gibbs RA. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One. 2012;7:e47768.
25. Edwards RJ. PAFScaff. https://github.com/slimsuite/pafscaff.; Accessed 17 January 2019.
26. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094-3100.
27. Davey NE, Shields DC, Edwards RJ. SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent. Nucleic Acids Res. 2006;34:3546-3554.
28. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ. 2018;6:e4958.
29. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schaffer AA. Database indexing for production MegaBLAST searches. Bioinformatics. 2008;24:1757-1764.
30. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210-3212.
31. Urich MA, Nery JR, Lister R, Schmitz RJ, Ecker JR. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. Nat Protoc. 2015;10:475-483.
32. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011;480:490-495.
33. Burger L, Gaidatzis D, Schubeler D, Stadler MB. Identification of active regulatory regions from DNA methylation data. Nucleic Acids Res. 2013;41:e155.
34. Mo A, Mukamel EA, Davis FP, Luo C, Henry GL, Picard S, Urich MA, Nery JR, Sejnowski TJ, Lister R, et al. Epigenomic signatures of neuronal diversity in the mammalian brain. Neuron. 2015;86:1369-1384.

35. Keilwagen J, Hartung F, Grau J. GeMoMa: Homology-based gene prediction utilizing intron position conservation and RNA-seq data. Methods Mol Biol. 2019;1962:161-177.
36. Altenhoff AM, Garrayo-Ventas J, Cosentino S, Emms D, Glover NM, Hernandez-Plaza A, Nevers Y, Sundesha V, Szklarczyk D, Fernandez JM, et al. The Quest for Orthologs benchmark service and consensus calls in 2020. Nucleic Acids Res. 2020;48:W538-W545.
37. Hazkani-Covo E, Martin WF. Quantifying the Number of Independent Organelle DNA Insertions in Genome Evolution and Human Health. Genome Biol Evol. 2017;9:1190-1203.
38. Pesole G, Allen JF, Lane N, Martin W, Rand DM, Schatz G, Saccone C. The neglected genome. EMBO Rep. 2012;13:473-474.
39. Dayama G, Emery SB, Kidd JM, Mills RE. The genomic landscape of polymorphic human nuclear mitochondrial insertions. Nucleic Acids Res. 2014;42:12640-12649.
40. Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J Mol Evol. 1994;39:174-190.
41. Chen JM, Chuzhanova N, Stenson PD, Ferec C, Cooper DN. Meta-analysis of gross insertions causing human genetic disease: novel mutational mechanisms and the role of replication slippage. Hum Mutat. 2005;25:207-221.
42. Martin EC. Dr. Johnson's Apple Orchard: The Story of America's First Pet Cemetery. Hartsdale Canine Cemetery; 1997.
43. 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;45:716-722.
44. Mainka SA. Fanconi syndrome in a Basenji. Can Vet J. 1985;26:303-305.
45. Johnson GS, Farias F. Progress in a 15 year search for the genetic causes for Fanconi syndrome in Basenjs. In; *Direct Fanconi Syndrome DNA test FAQ*. Basenji Club of Ameica Website; 2011
46. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. . 1303. . ArXiv. 2013;1303:1-3.
47. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:722-736.
48. Biosciences P. https://github.com/PacificBiosciences/GenomicConsensus**.** Accessed 1 November 2019.
49. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 2016;3:99-101.
50. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, Aiden EL. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356:92-95.
51. Edwards RJ. Diploidocus: Diploid genome assembly analysis tools https://github.com/slimsuite/diploidocus**.** 2020.
52. BBMap. https://sourceforge.net/projects/bbmap/**.** Accessed 1 November 2019.
53. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19:460.
54. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 2014;47:11 12 11-34.

55. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841-842.
56. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078-2079.
57. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403-410.
58. Edwards, R.J. SeqList. Edwards RJ. PAFScaff. https://github.com/slimsuite/Seqlist. Accessed 17 January 2020.
59. Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs. Bioinformatics. 2013;29:2487-2489.
60. Edwards RJ. BUSCOMP: BUSCO compilation and comparison – Assessing completeness in multiple genome assemblies https://github.com/slimsuite/buscomp. 2019.
61. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114-2120.
62. Chen H, Smith AD, Chen T. WALT: fast and accurate read mapping for bisulfite sequencing. Bioinformatics. 2016;32:3507-3509.
63. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10:563-569.
64. Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6:31.
65. Ollivier M, Tresset A, Bastian F, Lagoutte L, Axelsson E, Arendt ML, Balasescu A, Marshour M, Sablin MV, Salanova L, et al. Amy2B copy number variation reveals starch diet adaptations in ancient European dogs. Royal Society Open Science. 2016;3.
66. Edwards RJ. Miscellaneous biological sequence analysis tools suite. 2014.
67. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35:1026-1028.
68. Edwards RJ. SAAGA: Summarise, Annotate & Assess Genome Annotations: https://github.com/slimsuite/saaga. 020.
69. Seeman T. Barrnap. https://github.com/tseemann/barr. Accessed 1 November 2019.
70. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15:461-468.
71. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, et al. Ensembl 2020. Nucleic Acids Res. 2020;48:D682-D688.
72. 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.
73. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C. The sequence read archive. Nucleic Acids Res. 2011;39:D19-21.
74. Field MA, Cho V, Andrews TD, Goodnow CC. Reliably Detecting Clinically Important Variants Requires Both Combined Variant Calls and Optimized Filtering Strategies. PLoS One. 2015;10:e0143199.
75. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754-1760.

76. 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;20:1297-1303.
77. Waardenberg AJ, Field MA. consensusDE: an R package for assessing consensus of multiple RNA-seq algorithms with RUV correction. PeerJ. 2019;7:e8206.