Dog10K_Boxer_Tasha_1.0: A Long-Read Assembly of the Dog Reference Genome

Vidhya Jagannathan, Christophe Hitte, Jeffrey M Kidd, Patrick Masterson, Terence D Murphy, Sarah Emery, Brian Davis, Reuben M Buckley, Yan-Hu Liu, Xiang-Quan Zhang, et al.

genes

MDPI

Article

Dog10K_Boxer_Tasha_1.0: A Long-Read Assembly of the Dog Reference Genome

Vidhya Jagannathan [1], Christophe Hitte [2], Jeffrey M. Kidd [3,4], Patrick Masterson [5], Terence D. Murphy [5], Sarah Emery [3], Brian Davis [6], Reuben M. Buckley [7], Yan-Hu Liu [8,9], Xiang-Quan Zhang [8,9], Tosso Leeb [1], Ya-Ping Zhang [8,9], Elaine A. Ostrander [7] and Guo-Dong Wang [8,9,*]

1 Vetsuisse Faculty, Institute of Genetics, University of Bern, 3001 Bern, Switzerland; vidhya.jagannathan@vetsuisse.unibe.ch (V.J.); tosso.leeb@vetsuisse.unibe.ch (T.L.)
2 Institute Genetics Development Rennes, University of Rennes, CNRS—UMR 6290, F-35000 Rennes, France; hitte@univ-rennes1.fr
3 Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA; jmkidd@umich.edu (J.M.K.); sbherman@umich.edu (S.E.)
4 Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI 48109, USA
5 National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MA 20894, USA; patrick.masterson@nih.gov (P.M.); murphyte@nih.gov (T.D.M.)
6 Department of Integrative Biological Sciences, Texas A and M University, College Station, TX 77840, USA; bdavis@cvm.tamu.edu
7 National Human Genome Research Institute, National Institutes of Health, Bethesda, MA 20894, USA; reuben.buckley@nih.gov (R.M.B.); eostrand@mail.nih.gov (E.A.O.)
8 State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650201, China; liuyanhu@mail.kiz.ac.cn (Y.-H.L.); zhangxqkiz@163.com (X.-Q.Z.); zhangyp@mail.kiz.ac.cn (Y.-P.Z.)
9 Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650201, China
* Correspondence: wanggd@mail.kiz.ac.cn

Citation: Jagannathan, V.; Hitte, C.; Kidd, J.M.; Masterson, P.; Murphy, T.D.; Emery, S.; Davis, B.; Buckley, R.M.; Liu, Y.-H.; Zhang, X.-Q.; et al. Dog10K_Boxer_Tasha_1.0: A Long-Read Assembly of the Dog Reference Genome. *Genes* **2021**, *12*, 847. https://doi.org/10.3390/genes12060847

Academic Editor: Benjamin N. Sacks

Received: 5 May 2021
Accepted: 27 May 2021
Published: 30 May 2021

Abstract: The domestic dog has evolved to be an important biomedical model for studies regarding the genetic basis of disease, morphology and behavior. Genetic studies in the dog have relied on a draft reference genome of a purebred female boxer dog named "Tasha" initially published in 2005. Derived from a Sanger whole genome shotgun sequencing approach coupled with limited clone-based sequencing, the initial assembly and subsequent updates have served as the predominant resource for canine genetics for 15 years. While the initial assembly produced a good-quality draft, as with all assemblies produced at the time, it contained gaps, assembly errors and missing sequences, particularly in GC-rich regions, which are found at many promoters and in the first exons of protein-coding genes. Here, we present Dog10K_Boxer_Tasha_1.0, an improved chromosome-level highly contiguous genome assembly of Tasha created with long-read technologies that increases sequence contiguity >100-fold, closes >23,000 gaps of the CanFam3.1 reference assembly and improves gene annotation by identifying >1200 new protein-coding transcripts. The assembly and annotation are available at NCBI under the accession GCF_000002285.5.

Keywords: *Canis lupus familiaris*; high quality; contiguity; Pacific Biosciences; annotation; resource

1. Introduction

High-quality reference genomes are fundamental assets for the study of genetic variation in any species. The ability to link genotype to phenotype and the subsequent identification of functional variants rely on high fidelity assessment of variants throughout the genome. This reliance is well illustrated by the domestic dog, which offers specific challenges for any genetic study. Featuring over 350 pure breeding populations, each breed is a mosaic of ancient and modern variants, and each reflects a complex history linking it to

other related breeds. As a result of bottlenecks associated with domestication (15,000–30,000 years before present) and more recent individual breed formation (50–250 years before present), dog genomes contain long and frequent stretches of linkage disequilibrium (LD). While helpful for identifying loci of interest, long LD makes the necessary fine mapping for moving from marker to gene to variant both labor-intensive and error-prone.

In 2005, the first high-quality draft (7.5×) sequence of a Boxer dog, named Tasha, was made publicly available [1]. The reference sequence has proven useful in discoveries of canine-associated molecular variants [2,3], including single-nucleotide variants (SNVs) and small indels, regulatory sequences [4,5], large rearrangements and copy number variants [6,7] associated with both inter and intra gene variation. The resulting SNV arrays, designed based on variation relative to the Tasha-derived assembly, have led to the success of hundreds of genome-wide association studies (GWAS), advancing the dog as a system for studies of disease susceptibility and molecular pathomechanisms, evolution, and behavior. However, for the dog system to advance further, long-read high-quality assemblies from different individuals are needed. This will greatly improve the sensitivity of variant detection, especially for large structural variation. Furthermore, high-quality assemblies are an essential prerequisite for accurate annotation, which is required to assay the potential functional effects of detected variants. Recently, several high-quality genomes from different dogs became available [8–10].

However, using the same dog as used for the initial assembly offers specific advantages, including the ability to integrate new findings with previous observations. A high-quality genome assembly from the boxer Tasha will mean that the value of existing resources, such as existing bacterial artificial chromosome (BAC) libraries, and the wealth of experience and knowledge gained using previous versions of this dog's genome, will be preserved for future research efforts. The dog genome assembly reported here was built using a combination of Pacific Biosciences (PacBio) continuous long-read (CLR) sequencing technology, 10x Chromium-linked reads, BAC pair-end sequences and the draft reference genome sequence CanFam3.1.

2. Materials and Methods

2.1. Whole Genome Sequencing

A single blood draw from which genomic DNA was isolated from blood leukocytes of a female Boxer, Tasha, and which was also used to generate the previous CanFam 1, CanFam 2 and CanFam 3 genome assemblies, was utilized here. Continuous long-read (CLR) sequencing was carried out at Novogene Bioinformatics Technology Co., Ltd. (Beijing, China) with a PacBio Sequel sequencer (Pacific Biosciences, Menlo Park, CA, USA). Approximately 100 μg of genomic DNA were used for sequencing. SMRTbell libraries were prepared using a DNA Template Prep Kit 1.0 (PacBio), and 56 20-kb SMRTbell libraries were constructed. A total of 252 Gb of sequence data were collected. High molecular weight DNA from Tasha was also sequenced with Chromium libraries (10x Genomics, Pleasanton, CA, USA) on Illumina (San Diego, CA, USA) HiSeq X (2 × 150 bp), generating 589,824,390 read pairs or 176 Gb of data.

2.2. Genome Assembly Workflow

We assembled the genome using the Canu (v1.6) [11] and wtdbg2 [12] assembly algorithms. Briefly, the pipeline was composed of assembly, scaffolding and a final polishing step. PacBio reads had a mean read length of 8.5 kb and were used for the de novo assembly. The reads were corrected using the Canu error correction module, which generates a consensus sequence for each read using its best set of long read overlaps. The corrected consensus reads were then assembled using the wtdbg2 algorithm [12], which is designed for assembly of long reads produced by the PacBio or Nanopore technologies. The assembled contigs were polished with raw PacBio reads using the WTPOA-CNS tool of the WTDBG2 package. This was followed by misassembly detection and correction

with TIGMINT [13]. End sequences from BAC clones were extracted from the TraceDB of NCBI and used for scaffolding corrected contigs using the BESST algorithm (v2.2.8) [14]. Gap filling was performed using the PacBio subreads with PBjelly (from PBSuite v15.8.24) [15] and one additional round of genome polishing was carried out using Pilon v1.23. [16] with the 10x Chromium reads. Finally, RaGOO (v1.1) [17] was used for reference-guided scaffolding, using CanFam3.1 as the reference. The draft scaffolds were subjected to additional gap closure using PBJelly.

2.3. Assembly Quality Control

The scaffold order and orientation of the assembly was assessed by aligning it to an existing radiation hybrid map (RH-map) comprising 10,000 markers [18]. A chromosome-wide review of scaffold discrepancies was determined visually, by aligning the sequences of RH map markers against the assembled scaffolds. The generated dot plots were examined for contigs that were incorrectly ordered in scaffolds and these were manually inspected and eventually reordered. The assembly was also assessed for completeness using BUSCO [19], which provides a summary of genome completeness using a database of expected gene content based on near-universal single-copy orthologs from mammalian species with genomic sequence data. This includes 4104 single copy genes that are evolutionarily conserved between mammals.

2.3.1. Fosmid End Sequence Alignment

End sequences from previously constructed fosmid libraries from Tasha were aligned to the assembly as previously described [1]. Concordant clones were considered to be those with an inward read orientation and a size between 35,328 and 43,453 bp. Using bedtools [20], the physical coverage of concordant clones in 5 kb windows along the genome was determined. Segments of the primary chromosome assemblies that were not supported by any concordant fosmids were also identified. Analysis was limited to the primary chromosome assemblies (chr1-chr38, chrX) and any interval that intersected with chromosome ends was discarded. This resulted in a total of 1004 regions, of which 282 intersected with a segmental duplication interval (considering the union of assembly and read-depth-based annotations). To assess the significance of the intersection with segmental duplications, we performed 1000 random permutations of the intervals using bedtools and found that 49 to 103 of the intervals intersected with a duplication, with a mean intersection rate of 75.

2.3.2. Alignment of Finished BAC Clone Sequences

A list of assembled BAC clones from the CH-82 library was obtained from ftp://ftp.ncbi.nih.gov/repository/clone/reports/Canis_familiaris/CH82.clone_acstate_9615.out (accessed on 26th June 2019). The sequence of 395 finished clones was aligned to the long-read Tasha assembly using minimap2 (v2.17) [21]. One clone (AC190394.3) did not have a minimap2 alignment, 124 clones returned multiple alignment positions, 124 clones aligned to a single position annotated as duplicated in the Dog_10K_Boxer_Tasha_1.0 assembly, and four clones returned alignments that did not include the entire BAC sequence. We therefore focused on a set of 142 clones that had alignment to a single locus based on minimap2 with a query alignment that encompassed the entire clone length and that did not overlap with regions annotated as segmental duplications in the Dog_10K_Boxer_Tasha_1.0 assembly. An optimal global sequence alignment between the BAC sequence and the assembly was then determined using a stretcher [22] with default parameters.

2.4. Detection of Common Repeats and Segmental Duplications

Common repeats were identified with RepeatMasker (v4.0.7) using the rmblastn (v2.2.27+) search engine and a combined repeat database consisting of the Dfam_Consensus-20170127 [23] and RepBase-20170127 [24] releases.

Segmental duplications in the assembly were detected using two approaches. First, duplicated regions were identified based on assembly self-alignment using the program SEDEF [25]. Duplications with at least 90% sequence identity and length of 1 kb were retained. Second, duplications were defined based on an analysis of the depth of coverage of Illumina sequencing data using the fastCN [26] program. Copy number was estimated in non-overlapping windows each containing 3 kbp of unmasked sequence. Control regions for normalization were converted to Dog10K_Boxer_Tasha_1.0 coordinates using the liftOver tool [27,28]. Segmental duplications were defined as segments of four or more consecutive windows with an estimated copy number of at least 2.5. Comparable annotations for the CanFam3.1 assembly were obtained from [8].

2.5. Gene Annotation

The assembly was annotated using the previously described NCBI pipeline [29,30]. The pipeline uses a WindowMasker-masked genome for building gene models substantiated with RNA-seq data and protein alignments. RNA-sequencing data from various dog tissues were used for gene prediction (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Canis_lupus_familiaris/106/) (accessed on 1st Feb 2021).

2.6. Genome Assembly Alignment

The Dog10K_Boxer_Tasha_1.0 assembly was aligned to the CanFam3.1 assembly using minimap2 (v2.17) [21] with the 'asm5' option. Insertions and deletions were identified using the paftools.js program distributed with minimap2 with default options. Analysis was restricted to the primary chromosome sequences (chr1-38 and chrX). Regions that overlapped with assembly gaps, segmental duplications detected based on assembly self-alignment, or segmental duplications identified by read depth were removed.

2.7. Structural Variant Detection

Raw PacBio reads were aligned to the CanFam3.1 and Dog10K_Boxer_Tasha_1.0 assemblies using minimap2 (v2.17) [21]. Structural variants were identified using sniffles (v1.0.12) [31]. Only calls with precise breakpoints on the primary chromosome sequences (chr1-38 and chrX) were considered. Calls were filtered to remove insertions and deletions that intersect with assembly gaps.

2.8. BAC Assembly

Bacterial artificial chromosome (BAC) clones that mapped to the amylase locus were received from the BACPAC resources center (Emeryville, CA, USA). BACs were streaked to obtain single clones on LB agar with 100 ug/ul chloramphenicol and singe clones were cultured 20–24 h at 37 °C in 100 mL LB broth with 100 ug/uL chloramphenicol. BAC DNA was isolated using NucleoBond Xtra Midi kit for transfection-grade plasmid DNA without NucleoBond® Finalizer (Machery-Nagel, Bethlehem, PA, USA) and, after precipitation and drying, resuspended in 500 uL H_2O by incubating 72–96 h at 4 °C. Within 48 h of resuspension, BAC DNA was sequenced on a Minion with the Flongle adapter (Oxford Nanopore Technologies, Oxford, UK). Libraries were made using the Rapid Barcoding Sequencing kit (Oxford Nanopore Technologies, SQK-RBK004) according to the manufacturer's protocol, except for fragmentation where 0.25 uL of Fragmentation Mix was mixed with 200 ng of DNA in 4.75 uL of water, incubated 30 °C for 1 min then 80 °C for 1 min, and cooled on ice. Following fragmentation, BAC libraries were pooled by adding 1.67 uL of each library prep; 0.5 uL Rapid Primer (RAP) was added, and the mix was incubated

for 5 min at room temperature. Flow cells were primed and loaded according to the manufacturer's protocol.

Nanopore reads from the BAC were assembled using the pipeline described in https://github.com/KiddLab/run_canu_bac (accessed on 1st Feb 2021). Briefly, raw reads were filtered for hits of *Escherichia coli* and assembled using canu (v2.1) [11]. The unique portion of the resulting circular contig was then extracted and polished using racon (v1.4.10) [32]. Finally, the vector backbone sequence was removed and the contig was rotated to begin at the appropriate position. The final CH82-451P03 sequence was compared to the Dog10K_Boxer_Tasha_1.0 assembly using MUMmer (v3.23) [33].

2.9. Mapping SNV Array Probes

Chromosomal sequences from CanFam3.1 and Dog10K_Boxer_Tasha_1.0 were aligned to each other using blat [34]. The aligned fragments were processed using UCSC tools to create the necessary chain file for use with the liftOver tool. The liftOver was performed using the default settings with the "-multiple" option included. Genomic positions from both the Affymetrix (Santa Clara, CA, USA) Axiom Canine HD Array and Illumina (San Diego, CA, USA) CanineHD BeadChip were converted from CanFam3.1 to Dog10K_Boxer_Tasha_1.0. Genomic positions were obtained for the Axiom Canine HD Array from: https://sec-assets.thermofisher.com/TFS-Assets/LSG/Support-Files/Axiom_K9_HD.na35.r5.a7.annot.csv.zip (accessed on 21st Nov 2020); and for the CanineHD BeadChip: ftp://webdata2:webdata2@ussd-ftp.illumina.com/downloads/ProductFiles/CanineHD/CanineHD_B.csv (accessed on 21st Nov 2020). The bed files resulting from the lift over were converted to Plink map files. All markers were included in each map file and markers were ordered sequentially according to the order they were downloaded from their corresponding URLs. Markers for which no position was obtained were placed on chromosome "0" at position "0".

3. Results

DNA isolated and stored at −80 °C at NHGRI from the same female Boxer, Tasha, used for the CanFam 3.1 draft genome sequence was utilized to generate a new assembly. Frozen DNA from the same aliquot was thawed and used to prepare high molecular weight DNA libraries, which were sequenced using PacBio single-molecule real-time (SMRT) and 10x Genomics Linked-Reads sequencing technologies. Approximately 100-fold coverage (252 Gb) and 74-fold coverage (176 Gb) of the genome were generated using PacBio and 10x Genomics reads, respectively.

3.1. Dog10K_Boxer_Tasha_1.0 Assembly

PacBio SMRT cells produced 27,878,642 reads with a mean length of 8514 bp and N50 read length—a length at which 50% of the bases are in reads longer or equal to—was 13,189 bp. All PacBio reads were used for the assembly. The assembly pipeline (Figure 1) underwent initial read correction with Canu. After correction, 558,6195 reads were used for assembling with wtdbg2, obtaining a corrected read cut-off of 14 kb that provided 43-fold (104,569,563,638 bases) genome coverage for input. The initial ungapped assembly of WTDBG2 contained 1562 contigs with an N50 of 23.8 Mb. Tigmint (v.0.4) was used to correct initial assembly errors by incorporating the linked reads generated by 10x Genomics Chromium long-read technology. Tigmint split 75 misassembled contigs, which resulted in an assembly featuring 1786 contigs, of which 1724 were >500 bp. The assembly contig N50—the contig length in the assembly where equal or longer contigs contain half the bases of the genome—was 23.72 Mb.

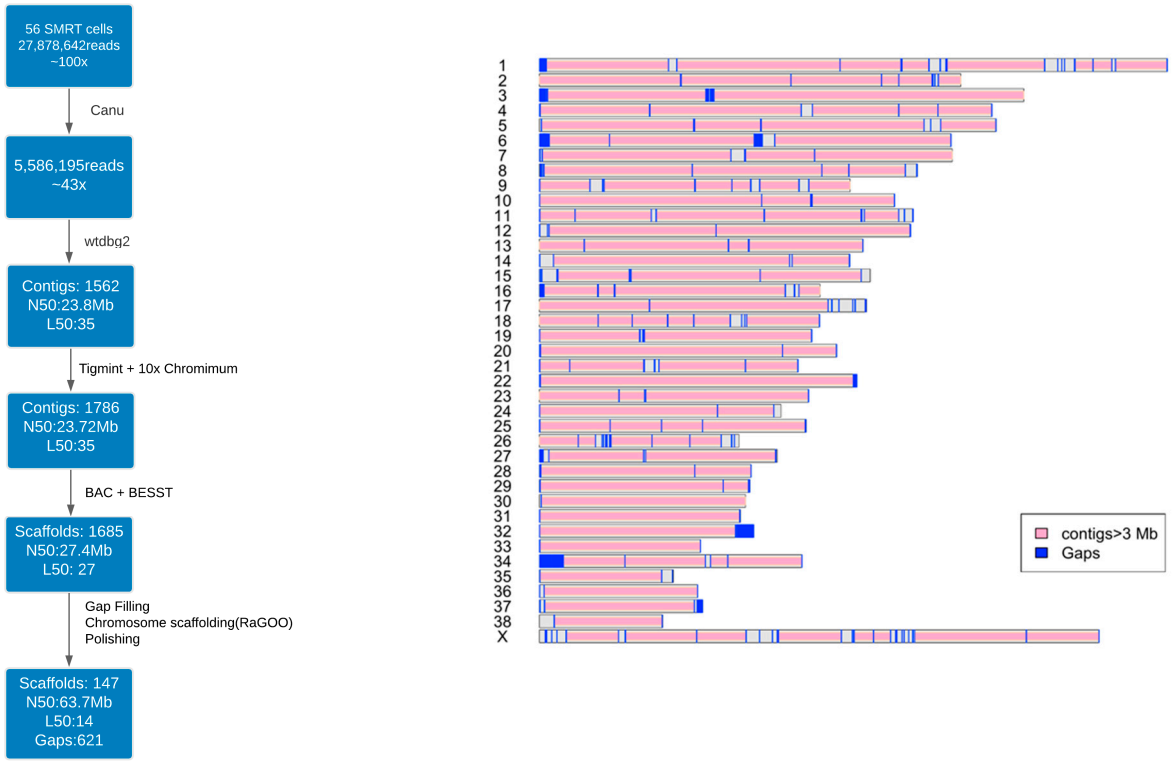
The Tigmint-corrected assembly was then scaffolded with BAC end sequences. The resultant scaffolding, constructed with the BESST algorithm (v 2.2.8), resulted in an assembly of 1685 scaffolds, which increased the N50 to 27.4 Mb. RaGOO was then used to

scaffold the data into 39 chromosomes based on CanFam3.1. The chromosome level scaffolds had a minimum of four contigs as noted on chromosomes 28, 30 and 36 and a maximum of 82 contigs on the X chromosome. The N50 of the scaffolded assembly was 63,738,581 bp (Table 1). The assembly contained 621 spanned gaps closing >23,000 of the CanFam3.1 assembly (18.25 Mb) (Figure 1). The number of unplaced scaffolds was 107 with an average length of 19.8 kb and consisting of 2,127,951 bases.

Table 1. Summary statistics for the Dog10K_Boxer_Tasha_1.0 genome assembly and comparison with current dog reference genome CanFam3.1.

Statistic	CanFam3.1	Dog10K_Boxer_Tasha_1.0
Total sequence length	2,410,976,875	2,312,802,206
Total ungapped length	2,392,715,236	2,312,743,367
No. of scaffolds	3310	147
No. of unplaced scaffolds	3228	107
Scaffold N50	45,876,610	63,738,581
Scaffold L50	20	14
No. of unspanned gaps	80	399
No. of spanned gaps	23,796	621
No. of contigs	27,106	1162
Contig N50	267,478	27,487,084
Contig L50	2436	31
No. of chromosomes	39	39

The quality of the Dog10K_Boxer_Tasha_1.0 assembly was assessed by comparison with an existing RH-map of 10,000 markers. The comparison strongly supported the overall accuracy of the assembly. There were two major discordances between the RH map and the draft assembly order of the contigs, one on chromosome 6 and the other on chromosome 11. The order was corrected, and gaps were again closed using PBJelly and PacBio SMRT raw reads. Discrepancies involving blocks of ~1 Mb on chromosome 9 and 0.2 Mb on chromosome 16 could not be resolved and will require further investigation.



(**a**) (**b**)

Figure 1. Dog_10k_Boxer_Tasha_1.0 assembly. (**a**) Assembly workflow pipeline. The different algorithms used in the pipeline have been indicated. N50 is the contig/scaffold length in the assembly where equal or longer contigs contain 50% of the genome. L50 count is the number of contigs whose length sum makes N50. (**b**) Ideogram showing chromosomes, contigs, and gaps. The grey regions indicate contigs of size less than 3Mb.

3.2. Assembly Quality Assessment

We used fosmid clone end sequences to identify regions that may be misassembled in Dog_10k_Boxer_Tasha_1.0. We identified 895,746 clones with a concordant mapping based on the orientation of the end-sequences and the apparent size of the cloned fragment (Figure S1), yielding a median genomic physical coverage of 17 concordant clones (Figure S2). Using this map of fosmid coverage, we identified 1004 intervals (32.5 Mb) on the primary chromosomes that do not intersect with a concordantly mapping fosmid (Table S1). We found that 282 of these intervals intersected with regions of segmental duplication in Dog_10K_Boxer_Tasha_1.0, a value greater than that observed in any of 1000 random permutations. This indicates that duplicated regions are enriched for potential misassembly.

We also assessed the per-bp sequence accuracy of the Dog_10k_Boxer_Tasha_1.0 assembly using 142 finished BAC clones from the CH-82 library that have a unique alignment to Dog_10k_Boxer_Tasha_1.0. Discarding alignment gaps, mismatches were observed at 14,255 of 26,778,153 aligned nucleotides (Figure S3). Assuming that all mismatches represent errors in the Dog_10k_Boxer_Tasha_1.0 sequence—a conservative assumption since heterozygous sites as well as errors in the BAC sequence are expected—the observed mismatch rate corresponds to an estimated per-base sequence quality [35] of Q33. We note, however, that the apparent number of alignment gaps is higher than the apparent single base substitution rate, suggesting that indels remain the primary error mode in long-read assemblies (Table S2).

3.3. Assembly Completeness

The completeness of the assembly was assessed using BUSCO, which uses a set of universal single-copy orthologs. This analysis showed an improvement of BUSCO completeness from 92.2% in CanFam3.1 to 95.3% in Dog10K_Boxer_Tasha_1.0 (Table 2).

Table 2. Comparison of BUSCO analysis of genomes.

Statistic	Dog10k_Boxer_Tasha_1.0	CanFam3.1
Complete BUSCOs	95.3%	92.2%
Complete and single copy BUSCOs	94.1%	91.1%
Complete and duplicated BUSCOs	1.2%	1.1%
Fragmented BUSCOs	2.1%	4.0%
Missing BUSCOs	2.6%	3.8%

We further compared the structural accuracy of the RaGOO arranged chromosome-level scaffolds to that of the CanFam3.1 chromosomes. We identified several regions known to be misassembled in CanFam3.1 and were now corrected in the Dog10k_Boxer_Tasha_1.0 assembly. These regions were supported by corresponding BAC end sequences (Figure S4).

Additionally, the orientation of chromosomes 27 and 32 was reversed compared to CanFam3.1. The two chromosomal re-orientations were backed by evidence in [36] and [37], based on recombination rates in dog chromosomes and fluorescence in situ hybridization experiments by Matthew Breen (personal communication).

3.4. Gene Annotation

Annotation of the Dog10K_Boxer_Tasha_1.0 assembly was carried out using the NCBI annotation pipeline and released via the NCBI ftp site [38]. The annotation pipeline used RNA-seq data from more than 25 tissues, along with known RefSeq, Genbank transcripts and canine expressed sequence tags. Statistics from the annotation release 106 are listed in Table 3. The annotation includes 20,100 protein-coding genes, which is comparable to annotations of other carnivores (average 20,105, stdev 1078, from 27 species). A total of 1299 protein-coding transcripts from 737 genes were identified as novel as they do not align to CanFam3.1 assembly. We found 78 out of 2473 known RefSeq transcripts did not map to the Dog10k_Boxer_Tasha_1.0 assembly [38]. Significantly, we observed a 7.0% increase (17,721 vs. 16,554) in the number of annotated protein-coding genes with very high coverage (≥ 90%) alignments compared to their best hits in SwissProt, with 88% of all protein-coding genes having at least one isoform exceeding 90% coverage. In addition, the new Tasha assembly has only 4.5% (891) of protein-coding genes represented with corrected models that compensate for suspected frameshifts or premature stop codons in the genome, compared to 5.5% for the prior NCBI annotation of CanFam3.1, or 5.6–11.3% for NCBI annotations of several other canine assemblies. These improvements can be largely attributed to fewer assembly gaps and the fact that gaps comprising exons of several genes have now been closed (Figure S5). For example, 5770 genes in CanFam3.1 have gaps within and flanking them. Only 12 of these genes still have gaps overlapping their exons and introns in Dog_10k_Boxer_Tasha_1.0.

Table 3. Annotation statistics for NCBI annotation release 106. * are non-coding RNA genes that cannot be classified.

Feature	Dog10k_Boxer_Tasha_1.0/Annotation Release 106
Protein-coding genes	20,100
Non-coding genes	15,306
Small non-coding genes	2083
Long non-coding genes	12,667
Miscellaneous * non-coding genes	10
Pseudogenes	4887

3.5. SNV Array Probes Mapped to Dog10k_Boxer_Tasha_1.0

Marker positions from the Axiom Canine HD Array and CanineHD BeadChip were mapped from CanFam3.1 to Dog10K_Boxer_Tasha_1.0. For the Axiom Canine HD Array and CanineHD BeadChip, 98.12% and 97.91% of markers, respectively, were successfully mapped to the new assembly. The data are available as supplementary files S1 and S2. The majority of markers on both arrays mapped to the same chromosome on both assemblies, with marker order remaining mostly intact. The largest contiguous off-diagonal collection of markers was found on chromosome 16 in CanFam3.1 and on chromosome 34 in Dog10K_Boxer_Tasha_1.0.

3.6. Analysis of Duplications

We identified segmental duplications in the Dog10K_Boxer_Tasha_1.0 assembly using two approaches. First, based on assembly self-alignment, we defined segmental duplications as segments at least 1 kb in length with a sequence identity of 90% or greater. This identified 5730 intervals encompassing 28.7 Mb of sequence on the primary chromosome assemblies (Table S3). Second, we identified 321 intervals encompassing 38.3 Mb of sequence based on excess depth of coverage from Illumina sequencing reads. These measures of duplication content are both less than that found in the Great Dane Zoey or CanFam3.1 assemblies [8], indicating that these duplicated sequences are not correctly resolved in the Dog10K_Boxer_Tasha_1.0 genome assembly.

3.7. Analysis of Repetitive Sequences

We identified common repeats in the Dog10K_Boxer_Tasha_1.0 assembly using RepeatMasker. A total of 41.1% of the assembly is comprised of repeats, with most falling into one of three categories: LINEs (469 Mb), SINEs (241 Mb) and LTRs (110 Mb). A complete summary of the repeat element composition is available in Table 4. We compared the results with an equivalent annotation of CanFam3.1. As before, we limited analyses to the primary chromosome sequences. At a high level, the repeat content of the Dog10K_Boxer_Tasha_1.0 and CanFam3.1 assemblies is similar (Table 4). However, the primary chromosome sequences in the Dog10K_Boxer_Tasha_1.0 assembly includes substantially more sequence classified as 'satellite', reflecting the ability of long-read sequencing to extend into subtelomeric and pericentromeric chromosomal regions. Although RepeatMasker analysis indicates that the CanFam3.1 contains more sequence classified as both LINE and SINE than the Dog10K_Boxer_Tasha_1.0 assembly, closer analysis revealed unexpected differences in repeat content (Table 5). LINE and SINE retrotransposons move via a copy-and-paste mechanism and new insertions accumulate mutations over evolutionary time scales [39]. Focusing on the youngest sequences shows that CanFam3.1 contains over 9000 more copies of a family of carnivore SINEs (SINECs) that show less than 10% sequence divergence, while the Dog10K_Boxer_Tasha_1.0 assembly contains 576 more LINEs that have less than 10% sequence divergence and are longer than

4 kb. We aligned the Dog10K_Boxer_Tasha_1.0 and CanFam3.1 assemblies to further explore this difference in the content of SINEC and LINE elements that have a low sequence divergence and identified 55,329 insertion–deletion differences between the assemblies longer than 10 bp. The variant size distribution has clear peaks corresponding to the expected sizes of dimorphic LINEs and SINEs (Figure 2).

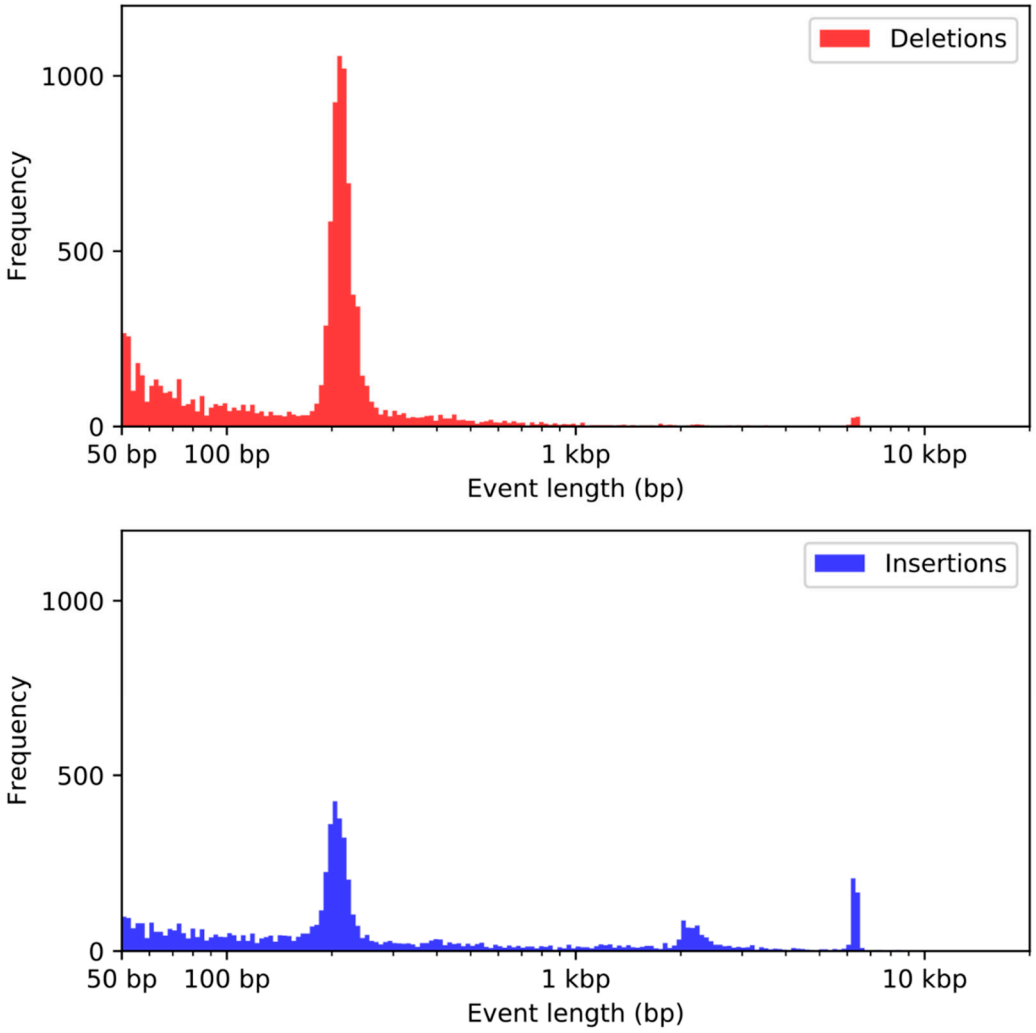


Figure 2. Size distribution of insertion–deletion differences identified between the Dog10K_Boxer_Tasha_1.0 and CanFam3.1 assemblies. The sizes of 22,330 sequences present in CanFam3.1 but absent in Dog10K_Boxer_Tasha_1.0 (red, deletions) and of 32,999 sequences present in Dog10K_Boxer_Tasha_1.0 but absent in CanFam3.1 (blue, insertions) are shown. The bins of each histogram are of equal size on a logarithmic scale.

Since LINE and SINE insertions are highly polymorphic among canines [8,40], we reasoned that the representation in the Dog10K_Boxer_Tasha_1.0 and CanFam3.1 assemblies may reflect the differential inclusion of heterozygous insertions. To assess this possibility, we identified structural variants relative to each assembly using the Tasha PacBio reads. Given the challenges associated with accurately discovering large insertions, we focused our analysis on deletion variants. We identified 35,187 deletions based on alignment to CanFam3.1 and 26,667 deletions based on alignment to Dog10K_Boxer_Tasha_1.0 (Supporting Files S3 and S4). Analysis of the variant size distribution is consistent with differential representation of heterozygous SINEs and LINEs in the two assemblies; there is an excess of ~200 bp deletions when mapping to CanFam3.1, while there is an excess of ~6 kb deletions when mapping to Dog10K_Boxer_Tasha_1.0 (Figure 3).

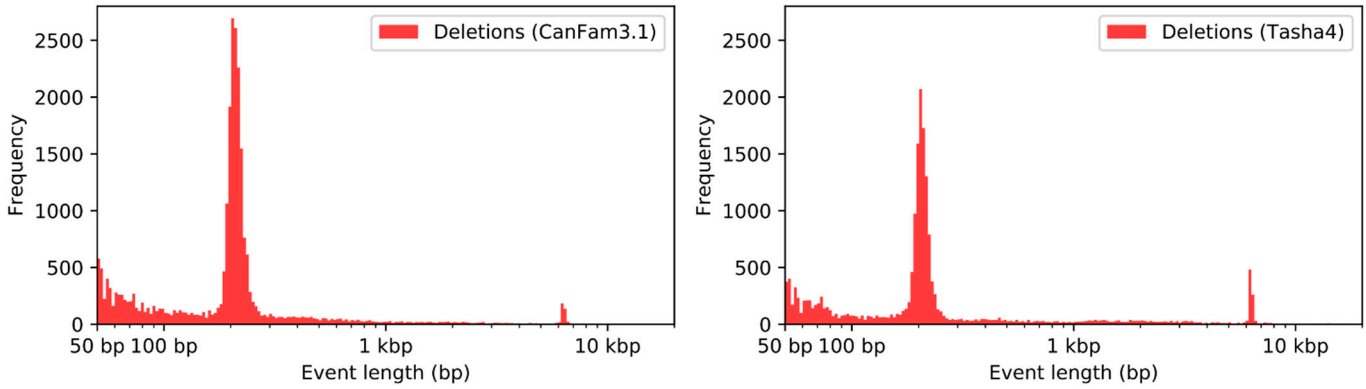


Figure 3. Discovery of deletion variants using PacBio reads. Deletions were identified based on alignment of PacBio reads to the CanFam3.1 (**left**) or Dog10K_Boxer_Tasha_1.0 (**right**) assemblies. The bins of each histogram are of equal size on a logarithmic scale.

Table 4. Repeat content of the Dog10K_Boxer_Tasha_1.0 and CanFam3.1 assemblies. Results are shown for the primary chromosome sequences.

	Dog10K_Boxer_Tasha_1.0		CanFam3.1	
Repeat Class	**Elements**	**bp**	**Elements**	**bp**
DNA	341,866	65,043,282	347,025	65,997,048
LINE	1,286,663	467,394,285	1,307,498	470,518,469
LTR	378,505	111,520,139	384,551	113,151,392
Low_complexity	123,075	6,525,287	120,803	6,009,804
RC	1636	345,889	1649	347,342
RNA	489	103,097	504	105,770
SINE	1,579,792	240,791,186	1,605,511	244,461,861
Satellite	5730	11,298,647	635	624,881
Simple_repeat	891,331	40,450,974	895,091	38,358,719
Unknown	3449	559,562	3487	565,722
rRNA	953	129,078	958	115,711
scRNA	70	4996	71	5156
snRNA	4492	278,022	4617	285,578
srpRNA	45	8900	47	9496
tRNA	35,501	2,608,084	35,906	2,636,278

Table 5. Repeat content for the lowly diverged SINE and LINE sequences.

	Dog10K_Boxer_Tasha_1.0		CanFam3.1	
Repeat Class	**Elements**	**bp**	**Elements**	**bp**
SINEC	1,125,416	177,104,238	1,146,663	180,147,553
SINEC < 10% divergence	454,869	71,490,885	464,113	72,819,234
LINE/L1	853,212	379,452,954	869,259	381,738,114
LINE/L1 < 10% divergence and≥ 4kb	4805	26,935,018	4229	23,359,516

3.8. Duplications at the Pancreatic Amylase Locus

Pancreatic amylase (AMY2B) catalyzes starch to maltose sugar in the small intestine. Changes in amylase gene copy number and expression have been correlated with dietary preferences across mammals [41]. Carnivores such as wolf, coyote and golden jackal have

two copies of the gene [42,43]. Increased copy number of the gene AMY2B, has been associated with adaptation to a starch-rich diet in modern dogs [42,44,45]. AMY2B copy number is variable both within and among modern dog breeds [46], suggesting a dynamic copy number state, perhaps reflecting recurrent expansion and contraction of a tandemly duplicated array. Long-read assembly data from a Basenji, named China [9], and a German Shepherd, named Nala [47], support the presence of a tandemly duplicated architecture at the AMY2B locus. In addition to tandem duplications, large segmental duplications encompassing AMY2B have also been described [26,48].

In the Dog10K_Boxer_Tasha_1.0 assembly, *AMY2B* is represented as a single copy on chromosome 6. Using Illumina read data, we estimate that the diploid *AMY2* copy number in Tasha is 12 (Figure 4). We found that Tasha is also heterozygous for a large duplication encompassing this locus. Examination of aligned fosmid end-sequence pairs revealed two clusters of clones that have an everted orientation consistent with a tandem duplication structure [49]. We identified the boundaries of these tandem duplications using the raw PacBio reads, defining the boundaries of tandem duplication units that are 1.9 Mb and 14.9 kb in length. Due to the presence of the larger duplication, the 12 *AMY2B* copies found in the Dog10K_Boxer_Tasha_1.0 genome are distributed among three structural alleles. Using Nanopore sequencing, we assembled a BAC mapping to this locus (CH82-451P03), and found that it contains a single copy of the *AMY2B* gene (Figure S6). Thus, at least one of the three structural *AMY2B* alleles in Tasha contains a single copy of this gene.

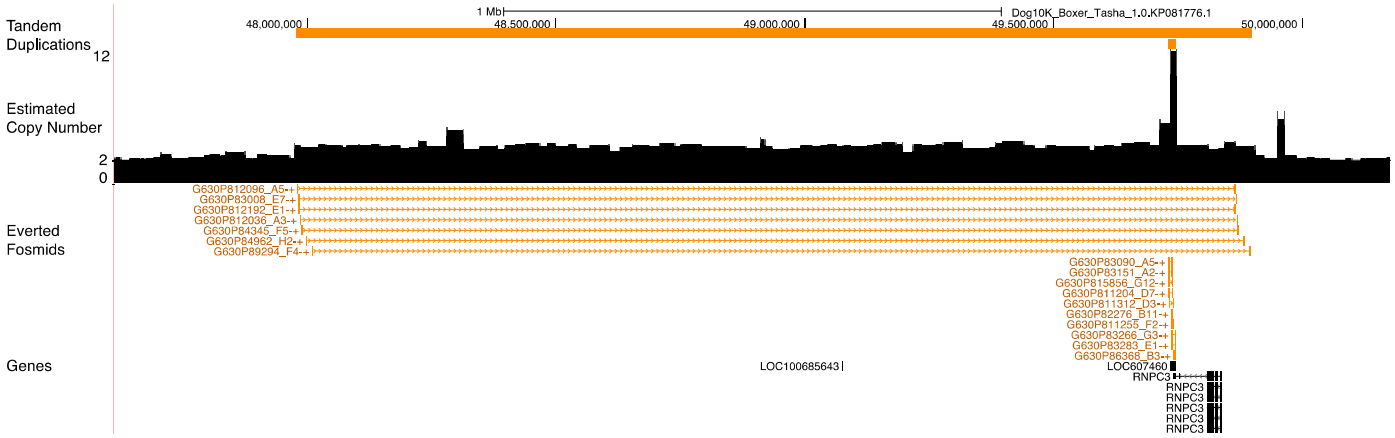


Figure 4. Structural variation at the amylase locus. A genome browser view illustrating structural variation at the amylase locus in Tasha is shown. The orange bars at the top indicate the locations of tandem duplications identified using the raw PacBio long-read data. This includes a large, 1.9 Mbp duplication (chr6:47977592-49898283) as well as a 14.8 kbp duplication (chr6:49729008-49743863). A read depth profile showing copy number estimated from Illumina sequencing data is depicted as a bar plot across the interval. An elevated copy number of 3, corresponding to the 1.9 Mb duplication, is observed, as well as a spike in copy number overlapping with the *AMY2B* gene. Mappings of discordant fosmid end sequences are shown in orange below the copy number profile. Each depicted clone has end sequences that align in an everted orientation consistent with the presence of a tandem duplication. The position of gene models derived from the NCBI gene annotation, release 106, are shown at the bottom of the figure. The *LOC607460* gene model corresponds to pancreatic α-amylase (*AMY2B*).

4. Discussion

Canis lupus familiaris, the domestic dog, is now well-established as a genetic system for studies of disease susceptibility, physiology and morphology, all of which inform our understanding of human health. Major advances in human disease genetics have resulted directly from observations made in the dog. Some prominent examples include the identification of *PNPLA1* variants in human patients with autosomal recessive congenital ichthyosis 10 that was enabled by results obtained in Golden Retrievers [50] or the elucidation of the role of the *PRCD* gene in dogs with progressive cone-rod dystrophy and human patients with retinitis pigmentosa [51]. In addition, because of the availability of a canine

genome assembly, canine disease models are now well established for several diseases including Duchenne muscular dystrophy [52], hypohidrotic ectodermal dysplasia [53], and Leber congenital amaurosis [53]. Similarly, canid evolution has revealed new insights into shifts in canine behaviors that are both surprising and informative, and evinced human dependence on dogs for early survival. While early canine studies relied on segregation studies in families, and later, GWAS studies in case–control cohorts, the most informative studies now rely on large numbers of SNVs and small indels retrieved from publicly available sequences aligned to the reference genome. As such, the reference genome is of critical importance, as current sequence-based GWAS studies highlight not just gene regions, but genic or regulatory variants of interest.

Using PacBio and 10x Chromium long-reads, Dog10k_Boxer_Tasha_1.0 was generated as a new dog genome resource, with a dramatically increased continuity. CanFam3.1 had a contig size of only 267 kb, while the Dog10k_Boxer_Tasha_1.0 assembly has an N50 contig size of 27.3 Mb featuring a >100-fold increase in sequence continuity. The improvements in the Dog10k_Boxer_Tasha_1.0 genome sequence relative to the CanFam3.1 assembly included not only greater continuity and fewer gaps, but also led to the correction of misassembled gene regions such as *OCA2* (Figure S4), which were supported by concordant alignments of BAC end sequences to the Dog10k_Boxer_Tasha_1.0 assembly.

The improvements in continuity and quality yielded a stronger template for annotation, resulting in better gene models. There is a 7.0% increase in protein-coding genes with high-coverage (≥90%) alignments in SwissProt, likely resulting from the increased contiguity, and the percentage of protein-coding genes annotated with corrections for suspected frameshifts or premature stop codons is the lowest of any current canine assembly (4.5%, vs. 5.6–11.3%), which may reflect the use of CLR reads and an additional polishing step. There are 78 of 2743 known RefSeq transcripts (2.8%) that do not map to the Dog10k_Boxer_Tasha_1.0 assembly, which is higher than observed for other assemblies and for which we cannot rule out transcript sequence characteristics or undetected chromosomal deletions, which requires further investigation. In particular, whole genome alignments between Dog10k_Boxer_Tasha_1.0 and previous Tasha assemblies highlight two major deletions on the X chromosome in the new assembly: an 8 Mb deletion (NC_051843.1: 14.2M..22.2M) and a 4.5 Mb deletion (NC_051843.1: 72M..76.5M). Additional sequencing of the X chromosome is required to resolve these regions.

There is a systematic underrepresentation of GC-rich sequences in CanFam3.1, as the necessary cloning and sequencing steps did not amplify GC-rich DNA particularly well. Long-read sequencing for the new assembly did not use any cloning steps or PCR and, as a result, GC-rich sequences are better represented and many gaps that were present in CanFam3.1 could be closed. This is critical as GC-rich sequences are often found in the first exons and promoter regions of genes, and play important roles in regulation, such as through differential methylation of CpG islands. As a result, the Dog10k_Boxer_Tasha_1.0 assembly will allow for more accurate identification of genetic variation in GC-rich regulatory regions and methylome studies.

Since the assembly approach we employed results in a haploid assembly representation, heterozygous loci are not uniformly represented. Essentially, only a single allele at a heterozygous site is included in the assembly. The effect of the haploid representation is most pronounced at heterozygous sites of structural variation where the two alleles may differ by hundreds or thousands of nucleotides. Intriguingly, the CanFam3.1 and Dog10k_Boxer_Tasha_1.0 assemblies have a systematic difference in the inclusion of alleles for dimorphic SINEC and LINE-1 sequences. Thus, although long-read sequencing approaches can resolve the full sequence of large insertions, genome assemblies that represent a diploid sample as a single haplotype may yield an incomplete representation of the true extent of mobile element diversity in canines.

To date, five long-read-based *de novo* dog genome assemblies [8–10] have been made available at the NCBI genome repository with comparable parameters such as number of

genes annotated and number of gaps between the new assemblies. The NCBI has annotated all five genomes and made them available on their genome browser https://www.ncbi.nlm.nih.gov/genome/gdv/?org=canis-lupus-familiaris (accessed on 1st Feb 2021). The comparative results indicate a strong likelihood that more protein-coding transcripts, pseudogenes, and non-coding genes remain to be discovered and annotated. However, the highly continuous genome sequence reported here provides a greatly improved framework which will enhance the characterization of functional sequences, genetic variation, and improve the utility of the thousands of canid sequences already generated, setting the stage for genetic studies of high accuracy and resolution.

The availability of *de novo* assemblies from different breeds will help to characterize structural variants (SVs), including copy-number variations (CNV), mobile element diversity, chromosomal rearrangements, missing sequences and non-redundant sequences. In all species and, especially in dogs, a single reference genome from one individual is unable to represent the full spectrum of divergent sequences in populations worldwide. Dog genomes vary in gene content, including tandem duplicated genes, CNVs distributed throughout the genome and in repetitive parts of the genome such as transposable elements. By characterizing genetic and structural variation within the canine species, *de novo* assemblies will better reveal the extensive variation in genome content among canine subpopulations defined by breeds, clades, and geography. The extensive analysis of the genetic variability of the canine genome will constitute the next paradigm shift for canine genomics.

5. Conclusions

We provide the Dog10k_Boxer_Tasha_1.0 genome assembly derived from the female boxer Tasha, the same dog that was used for the previous genome assemblies CanFam1, 2 and 3. Our assembly represents a substantial improvement in continuity and completeness and, together with the associated annotation, will be a valuable resource for canine and comparative genetics research.

Supplementary Materials: The following are available online at www.mdpi.com/2073-4425/12/6/847/s1. Figure S1. Apparent fosmid library insert size; Figure S2. Coverage of concordant fosmid clones; Figure S3 Apparent sequence errors based on finished BAC clones; Figure S4. Corrected misassembled regions; Figure S5 Filled Gaps in exons; Figure S6 Alignment of CH82-451P03 to the Dog_10K_Boxer_Tasha_1.0 assembly; TableS1 Regions with no concordant fosmid coverage; Table S2 BACalignedStats.xlsx; TableS3 SegDup-align-table; File S1 Axiom Canine HD Array map file for 10k_Boxer_Tasha; File S2 CanineHD BeadChip map file for 10k_Boxer_Tasha.

Author Contributions: Conceptualization, E.A.O., Y.-P.Z. and G.-D.W.; methodology, V.J., J.M.K., S.E., C.H., B.D.; sequencing: G.-D.W., S.E., J.M.K.; software, V.J. and J.M.K.; Whole genome gene annotation, P.M.; gene annotation analysis, T.D.M.; formal analysis, V.J., J.M.K., C.H., Y.-H.L., R.M.B. and X.-Q.Z.; resources, E.A.O., Y.-P.Z. and G.-D.W.; writing—original draft preparation, V.J., T.L., J.M.K. and E.A.O.; writing—review and editing, V.J., T.L., J.M.K., E.A.O., G.-D.W.; funding acquisition, E.A.O. and Y.-P.Z. All authors have read and agreed to the published version of the manuscript.

Funding: E.A.O. was funded by the Intramural Program of the National Human Genome Research Institute. J.M.K and S.E. were supported by grant R01GM140135 from the National Institutes of Health, The National Key R&D Program of China (2019YFA0707101), Key Research Program of Frontier Sciences of the CAS (ZDBS-LY-SM011), and Innovative Research Team (in Science and Technology) of Yunnan Province (202005AE160012). G.D.W. is supported by the Youth Innovation Promotion Association of CAS. Funding for the NCBI annotation and analysis, to P.M. and T.D.M., was provided by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The genome assembly is deposited at NCBI under accession number GCF_000002285.5. The associated BioProject accession number is PRJNA13179. BAC clone sequence has been deposited under accession MW972226. UCSC Track hub for Dog_10K_Boxer_Tasha_1.0 assembly https://github.com/KiddLab/tasha_genome_hub (accessed on 29 May 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lindblad-Toh, K.; Wade, C.M.; Mikkelsen, T.S.; Karlsson, E.K.; Jaffe, D.B.; Kamal, M.; Clamp, M.; Chang, J.L.; Kulbokas, E.J.; Zody, M.C.; et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **2005**, *438*, 803–819, doi:10.1038/nature04338.
2. Jagannathan, V.; Drögemüller, C.; Leeb, T.; Aguirre, G.; André, C.; Bannasch, D.; Becker, D.; Davis, B.; Ekenstedt, K.; Faller, K.; et al. A comprehensive biomedical variant catalogue based on whole genome sequences of 582 dogs and eight wolves. *Anim. Genet.* **2019**, *50*, 695–704, doi:10.1111/age.12834.
3. Plassais, J.; Kim, J.; Davis, B.W.; Karyadi, D.M.; Hogan, A.N.; Harris, A.C.; Decker, B.; Parker, H.G.; Ostrander, E.A. Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nat. Commun.* **2019**, *10*, doi:10.1038/s41467-019-09373-w.
4. Xie, X.; Lu, J.; Kulbokas, E.J.; Golub, T.R.; Mootha, V.; Lindblad-Toh, K.; Lander, E.S.; Kellis, M. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **2005**, *434*, 338–345, doi:10.1038/nature03441.
5. Dermitzakis, E.T.; Kirkness, E.; Schwarz, S.; Birney, E.; Reymond, A.; Antonarakis, S.E. Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res.* **2004**, *14*, 852–859, doi:10.1101/gr.1934904.
6. Ramirez, O.; Olalde, I.; Berglund, J.; Lorente-Galdos, B.; Hernandez-Rodriguez, J.; Quilez, J.; Webster, M.T.; Wayne, R.K.; Lalueza-Fox, C.; Vilà, C.; et al. Analysis of structural diversity in wolf-like canids reveals post-domestication variants. *BMC Genom.* **2014**, *15*, 465, doi:10.1186/1471-2164-15-465.
7. Serres-Armero, A.; Povolotskaya, I.S.; Quilez, J.; Ramirez, O.; Santpere, G.; Kuderna, L.F.K.; Hernandez-Rodriguez, J.; Fernandez-Callejo, M.; Gomez-Sanchez, D.; Freedman, A.H.; et al. Similar genomic proportions of copy number variation within gray wolves and modern dog breeds inferred from whole genome sequencing. *BMC Genom.* **2017**, *18*, 977, doi:10.1186/s12864-017-4318-x.
8. Halo, J.V.; Pendleton, A.L.; Shen, F.; Doucet, A.J.; Derrien, T.; Hitte, C.; Kirby, L.E.; Myers, B.; Sliwerska, E.; Emery, S.; et al. Long-read assembly of a great dane genome highlights the contribution of GC-rich sequence and mobile elements to canine genomes. *Proc Natl Acad Sci USA* **2021**, *118*, doi:10.1073/pnas.2016274118.
9. Edwards, R.J.; Field, M.A.; Ferguson, J.M.; Dudchenko, O.; Keilwagen, J.; Rosen, B.D.; Johnson, G.S.; Rice, E.S.; Hillier, L.D.; Hammond, J.M.; et al. Chromosome-length genome assembly and structural variations of the primal basenji dog (canis lupus familiaris) genome. *BMC Genom.* **2021**, *22*, 188, doi:10.1186/s12864-021-07493-6.
10. Wang, C.; Wallerman, O.; Arendt, M.-L.; Sundström, E.; Karlsson, Å.; Nordin, J.; Mäkeläinen, S.; Pielberg, G.R.; Hanson, J.; Ohlsson, Å.; et al. A novel canine reference genome resolves genomic architecture and uncovers transcript complexity. *Commun. Biol.* **2021**, *4*, 1–11, doi:10.1038/s42003-021-01698-x.
11. Koren, S.; Walenz, B.P.; Berlin, K.; Miller, J.R.; Bergman, N.H.; Phillippy, A.M. Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Res.* **2017**, *27*, 722–736, doi:10.1101/gr.215087.116.
12. Ruan, J.; Li, H. Fast and accurate long-read assembly with Wtdbg2. *Nat. Methods* **2020**, *17*, 155–158, doi:10.1038/s41592-019-0669-3.
13. Jackman, S.D.; Coombe, L.; Chu, J.; Warren, R.L.; Vandervalk, B.P.; Yeo, S.; Xue, Z.; Mohamadi, H.; Bohlmann, J.; Jones, S.J.M.; et al. Tigmint: Correcting assembly errors using linked reads from large molecules. *BMC Bioinform.* **2018**, *19*, doi:10.1186/s12859-018-2425-6.
14. Sahlin, K.; Chikhi, R.; Arvestad, L. Assembly scaffolding with PE-contaminated mate-pair libraries. *Bioinformatics* **2016**, *32*, 1925–1932, doi:10.1093/bioinformatics/btw064.
15. English, A.C.; Richards, S.; Han, Y.; Wang, M.; Vee, V.; Qu, J.; Qin, X.; Muzny, D.M.; Reid, J.G.; Worley, K.C.; et al. Mind the gap: Upgrading Genomes with pacific biosciences RS long-read sequencing technology. *PLoS ONE* **2012**, *7*, doi:10.1371/journal.pone.0047768.
16. Walker, B.J.; Abeel, T.; Shea, T.; Priest, M.; Abouelliel, A.; Sakthikumar, S.; Cuomo, C.A.; Zeng, Q.; Wortman, J.; Young, S.K.; et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **2014**, *9*, doi:10.1371/journal.pone.0112963.
17. Alonge, M.; Soyk, S.; Ramakrishnan, S.; Wang, X.; Goodwin, S.; Sedlazeck, F.J.; Lippman, Z.B.; Schatz, M.C. RaGOO: Fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* **2019**, *20*, doi:10.1186/s13059-019-1829-6.
18. Hitte, C.; Madeoy, J.; Kirkness, E.F.; Priat, C.; Lorentzen, T.D.; Senger, F.; Thomas, D.; Derrien, T.; Ramirez, C.; Scott, C.; et al. Facilitating genome navigation: Survey sequencing and dense radiation-hybrid gene mapping. *Nat. Rev. Genet.* **2005**, *6*, 643–648, doi:10.1038/nrg1658.

19. Seppey, M.; Manni, M.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness. In *Methods in Molecular Biology*; Humana Press Inc., 2019; Vol. 1962, pp. 227–245
20. Quinlan, A.R.; Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842, doi:10.1093/bioinformatics/btq033.
21. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **2018**, *34*, 3094–3100, doi:10.1093/bioinformatics/bty191.
22. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European molecular biology open software suite. *Trends Genet.* **2000**, *16*, 276–277, doi:10.1016/s0168-9525(00)02024-2.
23. Hubley, R.; Finn, R.D.; Clements, J.; Eddy, S.R.; Jones, T.A.; Bao, W.; Smit, A.F.A.; Wheeler, T.J. The Dfam database of repetitive DNA families. *Nucleic. Acids Res.* **2016**, *44*, D81–D89, doi:10.1093/nar/gkv1272.
24. Bao, W.; Kojima, K.K.; Kohany, O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **2015**, *6*, 11, doi:10.1186/s13100-015-0041-9.
25. Numanagic, I.; Gökkaya, A.S.; Zhang, L.; Berger, B.; Alkan, C.; Hach, F. Fast characterization of segmental duplications in genome assemblies. *Bioinformatics* **2018**, *34*, i706–i714, doi:10.1093/bioinformatics/bty586.
26. Pendleton, A.L.; Shen, F.; Taravella, A.M.; Emery, S.; Veeramah, K.R.; Boyko, A.R.; Kidd, J.M. Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. *BMC Biol.* **2018**, *16*, 64, doi:10.1186/s12915-018-0535-2.
27. Kent, W.J.; Baertsch, R.; Hinrichs, A.; Miller, W.; Haussler, D. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci USA* **2003**, *100*, 11484–11489, doi:10.1073/pnas.1932072100.
28. Kuhn, R.M.; Haussler, D.; Kent, W.J. The UCSC genome browser and associated tools. *Brief. Bioinform.* **2013**, *14*, 144–161, doi:10.1093/bib/bbs038.
29. Pruitt, K.D.; Tatusova, T.; Brown, G.R.; Maglott, D.R. NCBI reference sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic. Acids Res.* **2012**, *40*, D130–D135, doi:10.1093/nar/gkr1079.
30. Françoise Thibaud-Nissen; Souvorov, A.; Terence, M.; DiCuccio, M.; Kitts, P. *The NCBI Handbook [Internet]*, 2nd ed.; 2013. Available online: https://www.ncbi.nlm.nih.gov/books/NBK143764/ accessed 1 February 2021).
31. Sedlazeck, F.J.; Rescheneder, P.; Smolka, M.; Fang, H.; Nattestad, M.; von Haeseler, A.; Schatz, M.C. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **2018**, *15*, 461–468, doi:10.1038/s41592-018-0001-7.
32. Vaser, R.; Sović, I.; Nagarajan, N.; Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **2017**, *27*, 737–746, doi:10.1101/gr.214270.116.
33. Kurtz, S.; Phillippy, A.; Delcher, A.L.; Smoot, M.; Shumway, M.; Antonescu, C.; Salzberg, S.L. Versatile and open software for comparing large genomes. *Genome Biol.* **2004**, *5*, R12, doi:10.1186/gb-2004-5-2-r12.
34. Kent, W.J. BLAT—the BLAST-like Alignment Tool. *Genome Res.* **2002**, *12*, 656–664, doi:10.1101/gr.229202.
35. Ewing, B.; Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **1998**, *8*, 186–194.
36. Campbell, C.L.; Bhérer, C.; Morrow, B.E.; Boyko, A.R.; Auton, A. A Pedigree-based map of recombination in the domestic dog genome. *G3 Genes Genomes Genet.* **2016**, *6*, 3517–3524, doi:10.1534/g3.116.034678.
37. Wong, A.K.; Ruhe, A.L.; Dumont, B.L.; Robertson, K.R.; Guerrero, G.; Shull, S.M.; Ziegle, J.S.; Millon, L.V.; Broman, K.W.; Payseur, B.A.; et al. A comprehensive linkage map of the dog genome. *Genetics* **2010**, *184*, 595–605, doi:10.1534/genetics.109.106831.
38. Canis Lupus Familiaris Annotation Report. Available online: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Canis_lupus_familiaris/106/#TranscriptAlignmentStats (accessed on 15 April 2021).
39. Richardson, S.R.; Doucet, A.J.; Kopera, H.C.; Moldovan, J.B.; Garcia-Perez, J.L.; Moran, J.V. The influence of LINE-1 and SINE retrotransposons on mammalian genomes. *Microbiol. Spectr.* **2015**, *3*. doi:10.1128/microbiolspec.MDNA3-0061-2014.
40. Wang, W.; Kirkness, E.F. Short Interspersed elements (SINEs) are a major source of canine genomic diversity. *Genome Res.* **2005**, *15*, 1798–1808, doi:10.1101/gr.3765505.
41. Pajic, P.; Pavlidis, P.; Dean, K.; Neznanova, L.; Romano, R.-A.; Garneau, D.; Daugherity, E.; Globig, A.; Ruhl, S.; Gokcumen, O. Independent amylase gene copy number bursts correlate with dietary preferences in mammals. *Elife* **2019**, *8*, doi:10.7554/eLife.44628.
42. Axelsson, E.; Ratnakumar, A.; Arendt, M.-L.; Maqbool, K.; Webster, M.T.; Perloski, M.; Liberg, O.; Arnemo, J.M.; Hedhammar, A.; Lindblad-Toh, K. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* **2013**, *495*, 360–364, doi:10.1038/nature11837.
43. Freedman, A.H.; Gronau, I.; Schweizer, R.M.; Vecchyo, D.O.-D.; Han, E.; Silva, P.M.; Galaverni, M.; Fan, Z.; Marx, P.; Lorente-Galdos, B.; et al. Genome sequencing highlights the dynamic early history of dogs. *PLOS Genetics* **2014**, *10*, e1004016, doi:10.1371/journal.pgen.1004016.
44. Ollivier, M.; Tresset, A.; Bastian, F.; Lagoutte, L.; Axelsson, E.; Arendt, M.-L.; Bălăşescu, A.; Marshour, M.; Sablin, M.V.; Salanova, L.; et al. Amy2B copy number variation reveals starch diet adaptations in Ancient European dogs. *R Soc. Open Sci.* **2016**, *3*, 160449, doi:10.1098/rsos.160449.
45. 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, doi:10.1111/age.12179.

46. Reiter, T.; Jagoda, E.; Capellini, T.D. Dietary variation and evolution of gene copy number among dog breeds. *PLoS ONE* **2016**, *11*, e0148899, doi:10.1371/journal.pone.0148899.
47. Field, M.A.; Rosen, B.D.; Dudchenko, O.; Chan, E.K.F.; Minoche, A.E.; Edwards, R.J.; Barton, K.; Lyons, R.J.; Tuipulotu, D.E.; Hayes, V.M.; et al. Canfam_GSD: De Novo chromosome-length genome assembly of the German Shepherd dog (Canis Lupus Familiaris) using a combination of long reads, optical mapping, and Hi-C. *Gigascience* **2020**, *9*, doi:10.1093/gigascience/giaa027.
48. Botigué, L.R.; Song, S.; Scheu, A.; Gopalan, S.; Pendleton, A.L.; Oetjens, M.; Taravella, A.M.; Seregély, T.; Zeeb-Lanz, A.; Arbogast, R.-M.; et al. Ancient european dog genomes reveal continuity since the Early Neolithic. *Nat. Commun.* **2017**, *8*, 16082, doi:10.1038/ncomms16082.
49. Cooper, G.M.; Zerr, T.; Kidd, J.M.; Eichler, E.E.; Nickerson, D.A. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat. Genet.* **2008**, *40*, 1199–1203, doi:10.1038/ng.236.
50. Grall, A.; Guaguère, E.; Planchais, S.; Grond, S.; Bourrat, E.; Hausser, I.; Hitte, C.; Le Gallo, M.; Derbois, C.; Kim, G.-J.; et al. PNPLA1 mutations cause autosomal recessive congenital ichthyosis in Golden Retriever dogs and humans. *Nat. Genet.* **2012**, *44*, 140–147, doi:10.1038/ng.1056.
51. Zangerl, B.; Goldstein, O.; Philp, A.R.; Lindauer, S.J.P.; Pearce-Kelling, S.E.; Mullins, R.F.; Graphodatsky, A.S.; Ripoll, D.; Felix, J.S.; Stone, E.M.; et al. Identical mutation in a novel retinal gene causes progressive rod-cone degeneration in dogs and retinitis pigmentosa in humans. *Genomics* **2006**, *88*, 551–563, doi:10.1016/j.ygeno.2006.07.007.
52. Kornegay, J.N. The Golden Retriever model of duchenne muscular dystrophy. *Skelet. Muscle* **2017**, *7*, 9, doi:10.1186/s13395-017-0124-z.
53. Margolis, C.A.; Schneider, P.; Huttner, K.; Kirby, N.; Houser, T.P.; Wildman, L.; Grove, G.L.; Schneider, H.; Casal, M.L. Prenatal Treatment of X-linked hypohidrotic ectodermal dysplasia using recombinant ectodysplasin in a canine model. *J. Pharmacol. Exp. Ther.* **2019**, *370*, 806–813, doi:10.1124/jpet.118.256040.