Duplications and retrogenes are numerous and widespread in modern canine genomic assemblies

Anthony K. Nguyen[1,2], Matthew S. Blacksmith[1,2], and Jeffrey M. Kidd[1,2,*]

[1]Department of Human Genetics, University of Michigan, Ann Arbor, Michigan

[2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan

*Correspondence to J.M.K at jmkidd@umich.edu

Abstract

Recent years have seen a dramatic increase in the number of canine genome assemblies available. Duplications are an important source of evolutionary novelty and are also prone to misassembly. We explored the duplication content of nine canine genome assemblies using both genome self-alignment and read-depth approaches. We find that 8.58% of the genome is duplicated in the canFam4 assembly, derived from the German Shepherd Dog Mischka, including 90.15% percent of unplaced contigs. Highlighting the continued difficulty in properly assembling duplications, less than half of read-depth and assembly alignment duplications overlap. Further study shows the presence of multiple segments that have alignments to four or more duplicate copies. These highly duplicated segments correspond to gene retrocopies. We identified 2749 candidate retrocopies from 968 parental genes in the canFam4 assembly and find that approximately 8.82% of duplicated base pairs involve a retrocopy, confirming this mechanism as a major driver of gene duplication in canines. Similar patterns are found across eight other recent canine genome assemblies, with multiple metrics supporting the high-quality of the mCanLor1.2 wolf assembly constructed using PacBio HiFi reads. Comparison of the wolf and German Shepherd assemblies found that approximately 54% of retrocopy insertions are shared between assemblies. By calculating the number of generations since the divergence of these genomes, we estimate that new retrocopy insertions appear in 1 out of 345 births. Together, our analyses illustrate the impact of retrogene formation on canine genomes and highlight the variable representation of duplicated sequences among recently completed canine assemblies.

Keywords: read-depth, alignment, segmental duplication, retrocopy

Introduction

Structural variation refers to chromosomal alterations, typically larger than 1000 bases (1 kB) in size, and includes changes such as insertions, deletions, inversions, and other copy number changes [1]. Copy number gains, or duplications, that affect genes are a key mechanism in the evolution of novel gene functions and the emergence of gene families. Duplicated genes are associated with a number of core biological processes, including development, transcriptional regulation, and environmental response [2].

Gene duplications arise through two means: the reverse transcription and integration of RNA transcripts, or the duplication of DNA segments. Duplications on the DNA-level can arise via whole genome duplications, in which the entire genome is replicated. Whole genome duplications have occurred multiple times in evolutionary history and have contributed to increased genome complexity in vertebrates [3]. Alternatively, DNA-level duplications may be smaller and affect only a segment of the genome. These segmental duplications may encompass genes, giving rise to additional copies of genes and the expansion of related gene families. Segmental duplications are typically defined as non-repetitive sequences longer than 1 kB that are found in more than one copy in a haploid genome [4]. This definition differentiates duplications from common repeats, such as LINEs (Long INterspersed Element) and SINEs (Short INterspersed Element), which proliferate throughout genomes through distinct mechanisms.

Segmental duplications are further divided into three types based on the relative locations of duplicates: tandem duplications, in which duplications appear adjacent to one another; intra-chromosomal dispersed duplications, which is when duplications occur on the same chromosome but with an intervening sequence; and inter-chromosomal dispersed, when duplication copies occur on separate chromosomes. The intervening sequence between related duplicates can include genes and is susceptible to additional duplication or deletion [5]. DNA duplications likely first arise as a result of replication-based errors including break-induced replication and template switching mechanisms [6]. Once formed, additional copies can be gained or lost due to nonallelic homologous recombination [7]. No matter their type, when these copies differ in count between individuals or populations they are referred to as copy number variants (CNVs).

Gene duplications may also be caused by a retrotransposition-associated mechanism, in which mRNAs are reverse transcribed and inserted into the genome at new locations [8]. The resulting duplications, known as retrocopies, lack introns and show the hallmarks of retrotransposition, including a poly(A) tail and flanking target site duplications [9]. While retrocopies refer to any derivation of a gene transcript from retrotransposition, whereas the term retrogene implies functionality, both terms will be used interchangeably within this study [10]. The retrogenes may be full-length or truncated at their 5' end. Over time, mutations in retrocopies may accumulate that disrupt gene function creating pseudogenes [11]. However, retrogenes may be functional; several have been documented to have associations with important biological processes, including immune function, metabolism, environmental response, physical appearance, and disease, in both humans and canines [10, 12]. Gene conversion and changes in pseudogene sequence can also give rise to actual and novel function, altering

the name from pseudogenes into functional retrogenes. Retrocopies commonly occur in dogs [13, 14] and confound the interpretation of genetic variation [15].

Novel developments in sequencing technology have allowed for the rapid and thorough analysis of thousands of samples. Despite this, systematically identifying and analyzing duplications remains challenging. To investigate CNVs samples, the most common method of analysis is to look for an increase in read-depth, typically obtained from whole genome sequencing, relative to a reference genome assembly. Duplications can also be found by aligning a genome reference to itself to create a genome self-alignment that identifies sequences present multiple times [16]. When constructing such genome self-alignments, it is critical to first filter out common repeats since the proliferation of repeats overwhelms the alignment process.

There are several challenges associated with properly assembling duplicated regions. Assembly errors may lead to collapsed duplications in which duplicated sequences are only represented once in an assembly. This commonly occurs at the end of assembled contigs due to a failure to correctly assemble sequences spanning the duplication. Such assembly errors can also give rise to chimeric contigs that have been incorrectly assembled as a result of mis-joining non-adjacent segments [17]. False duplications can also arise due to misassembly. This can occur when contigs created from divergent alleles are mistakenly represented as being duplicated [18, 19]. Correctly assembling duplications is further complicated by the diploid nature of most analyzed genomes. As a result of challenges in correctly assembling duplicated sequences, the duplication content of a genome assembly may not be an accurate representation of the true duplication content of the analyzed genome. Assemblies derived from short sequencing reads are plagued by incorrect representation of duplicates [19]. Longer sequencing reads coupled with new methodologies for haplotype-resolved assembly [20, 21] promise to alleviate these concerns resulting in a more accurate representation of genome content and structure.

The canine remains a powerful system for genetics and genomics. The domestic dog is in the unique position of having co-evolved alongside humans, while simultaneously experiencing human-driven extreme artificial selection for over ten thousand years [22]. The first canine reference genome was released in 2005, derived from Tasha the Boxer [23]. This was further developed into canFam3.1, released in 2014, also based on Tasha [24]. The canine reference genome was assessed for copy number variation across multiple studies. Initial studies estimated that approximately 4.21% of the canine reference genome is composed of recent segmental duplication, including 841 genes in predicted segmental duplications [25]. Additional studies of canine CNVs have linked these variations to conditions such as ACL rupture [26] and breed-specific morphological structure [27]. Other studies have concluded that CNVs likely did not have major effects on canid domestication, with suggestions that dogs and wolves had similar proportions of CNV loci [28, 29].

Since the release of canFam3.1, multiple assemblies of dogs and wolves have been generated using long-read technology [30-37]. These new assemblies have helped elucidate novel genes, regulatory elements, and variants, while revealing new aspects of canine genome biology. These assemblies have been created using a variety of sequencing technologies (Oxford Nanopore and PacBio) and algorithms. In this study, we aim to characterize the duplication content of different recently published long-read canine genomes and assess patterns of gene duplication found throughout canines.

We apply two computational analyses to nine recently-released canine genome assemblies: genome assembly self-alignment, which involves the mapping of an assembly to itself in order to find matching duplicated segments [16] and read-depth analysis, which searches for regions of unusually high coverage in Illumina sequencing data [28, 38]. We further perform a search for retrogene insertions in each genome assembly. We find notable and large regions of duplicated sequence that is not properly represented in existing dog assemblies, describe polymorphic duplications that differ between the representative canine assemblies, and identify retrogenes as a major driver of gene duplications in canine genomes.

Materials and Methods

Assemblies considered

Our analysis involves 9 published long-read canine assemblies [30-37]. For simplicity, we refer to them by name. Assembly accessions and other details are provided in **Table 1**.

Genome assembly self-alignment

Genome assembly self-alignment was performed using BISER (Brisk Inference of Segmental duplication Evolutionary stRucture) (v1.2.3) which was developed to detect segmental duplications at low identity across multiple genome assemblies [16]. BISER defines a segmental duplication with the following criteria: below a certain error threshold based on shared sequence identity ε (approximately 15% for 75% shared sequence identity), longer than 1,000 bp, and paralog sequences may overlap up to a certain amount of base pairs, also defined by the error threshold. Before alignment, repetitive regions in the genome are soft-masked, prohibiting the initialization of alignments in low-complexity and highly-repetitive sequences, such as LINEs, SINEs, and simple repeats. Duplications involving the mitochondrial chromosome were removed from analysis since nuclear mitochondrial insertions (nuMTS) represent a distinct class of variation [39], and two categories of duplications were defined: one where the paralogous sequence between duplication copies is at least 90% identical, and a subset where the identity rate is at least 95%. These sets are further divided based on their placement on assembled chromosomes (the 38 canine autosomes and the X-chromosome) vs. unplaced contigs in each assembly. Because BISER identifies duplicated segments as pairs of sequences and duplication pairs often overlap each other, we used BEDTools merge (v2.30) to create a non-redundant set of duplicated intervals [40].

Distribution of duplications along the genome

We analyzed the distribution of duplications located on the 38 canine autosomes and the X chromosome by permutation. The distribution of duplications was compared against 1000 random permutations using BEDTools shuffle (v2.30), limiting placement to the assembled chromosome sequence [40]. This was performed three times, on the self-alignment duplications, the read-depth duplications, and on the retrogene set to look for chromosomal enrichment.

Locating highly duplicated segments

Short segments with a high recurrent duplication level were identified using the duplication pairs found by self-alignment located on the assembled chromosomes. All duplications smaller than 2.5 kB were located, and their coordinates rounded to the nearest hundred (i.e., a coordinate of 2,349

would become 2,300, and a coordinate of 3,175 would become 3,200), to account for minor difference in alignment endpoints. Duplications were merged into segments, and segments that contained at least four duplications were labeled as highly recurrent.

Read-depth analysis

Read-depth analysis was applied to Illumina data from the assembled individuals using fastCN [28]. First, a version of the assembly was created in which elements identified by RepeatMasker, tandem repeat finder [41], and over-represented 50-mers were masked. Next, all placements within an edit distance of two for 36-bp segments of each Illumina read determined using mrsFAST (v3.4.1) [42]. The resulting depth profiles were tabulated, corrected for local GC content per fastCN, and converted to copy number estimates based on a set of control regions. For copy number normalization we excluded: reported deletions, duplications, and insertions [30], duplicated intervals from a preliminary self-alignment of the Mischka genome, duplicated segments identified by a preliminary fastCN analysis, and all chromosomes other than chr1-38. Copy number profiles were calculated based on the mean normalized depth tabulated in windows containing 1,000 unmasked positions. Duplicated regions are defined as at least four consecutive windows whose copy numbers all exceed 2.5 and have a total combined size is equal or greater than 10 kB. An overall copy number was determined by taking the median copy number of all windows within a duplicated region. Masking of the additional genomes was performed as described above. Normalization controls were converted from Mischka coordinates using minimap2 (v2.26) [43]. China was omitted from these analyses because, while it does have BGI short read sequencing, we opted to only utilize Illumina-based data. Nala was omitted due to extreme GC bias observed in the associated Illumina data **(Supplementary Figure 1)**.

Comparison of duplications identified using self-alignment and read-depth approaches

To compare the results found by self-alignment and the results found by read-depth, both sets of duplications (BISER and read-depth) were filtered to retain only a) duplications on the assembled chromosomes; b) duplications greater than 10 kB; c) duplications with at least 95% identity; d) duplications that are not located in the first megabase of any autosome. The first megabase of each autosome is excluded because, as a result of masking, there are no informative read-depth windows in these regions. Self-alignment duplications were intersected with 1 kB unique sequence read-depth windows, and a self-alignment duplication was confirmed supported if the windows had a median copy number of at least 2.5. Read-depth duplications were counted as supported if they overlapped with any self-alignment duplication. The methods described above for the Mischka assembly were expanded to the eight other canine genome assemblies with control region coordinates using liftoff (v1.6.3) [43, 44]. In the gene analysis, read-depth windows are intersected with the protein-coding genes annotated in Mischka.

For the gene set of choice, we used a subset of 13,817 genes identified in Mischka, with the restriction that genes must be protein-coding, must be located on the 38 autosomes or the X chromosome, and that a gene must fully encompass at least 3 windows, where each window represents 1 kB of non-repetitive DNA sequence used by fastCN. A table was built to observe copy number across the genome, where a row represents a canine sample and a column represented a pre-determined window in the Mischka coordinates, with entries being the copy number. Copy number for genes were

determined as the median of all copy numbers in windows encompassed by a gene. A canine assembly is considered to have a duplicated gene if the copy number exceeds 2.5.

Retrogene analysis

We searched for candidate retrocopies in the Mischka assembly based on the annotation of 19,972 protein coding genes used by the Dog10K consortium [14]. A single transcript was selected for each gene, with priority given to the longest isoforms and fewest number of exons. We removed any genes with an intron less than 50 bp, to account for many uninvestigated genes with single base pair introns, and any gene that only had a single exon.

We constructed cDNA sequences using SAMtools faidx (v1.13) to extract the sequence of each exon [45]. Using BLAT, cDNA sequences were mapped against the genome with repetitive 11mers filtered (--*ooc* option) (v.35) [46]. The resulting alignments were examined for hits that could be retrocopies. Alignments were removed if they were: less than 100 bp, overlapped the genomic coordinates of the parent gene, had less than 90% sequence identity with the parent gene, or do not cross an exon-exon boundary of the parent gene. The same analysis procedure was applied to the other eight genome assemblies after converting the genes annotated in Mischka to the coordinates of each assembly using liftoff based on minimap2 alignments (v2.26) [44, 47]. 1000 random permutations were conducted in the same manner described in the section *Distribution of duplications along the genome.* Due to paralogous genes, parent genes could not be resolved for a few retrocopies in each canine assembly.

Examination of retrogene hallmarks

Retrogene containing loci from all nine canines were explored for presence of retrogene hallmarks: target-site duplications (TSDs) and poly(A) tails. Before detection proceeded, loci within one hundred base pairs of reference gaps were removed with BEDTools (v2.26.0) window with parameters -*w 100 and -u* [40]. Additionally, any loci within the first or last 100 base pairs of each chromosome did not undergo hallmark detection. For each locus, a symmetrical number of base pairs were extracted upstream and downstream of the retrogene with SAMtools faidx (v1.13) [45]. To detect TSDs, the upstream and downstream fasta files were aligned to each other using the water function within the EMBOSS module (v6.6.0) [48]. Water is a modified version of a Smith-Waterman local alignment and reports the highest scoring alignment only [49]. The scoring matrix used was a custom matrix which used +2 for all matches and -6 for all mismatches (-1000 for any alignment containing “N”). Gap opening and gap extension penalties were both set to 10. Alignment output containing “N” as a base pair were discarded, as were final alignments shorter than 5 base pairs in length. If alignments started within 5bp of the start of the upstream query or 5bp of the end of the downstream query, the flanking distance was increased by 5 and the alignment was repeated. Poly(A) detection followed TSD detection. The original flanking distance worth of base pairs were again extracted beyond the 3’ end of the retrogene with SAMtools faidx (v1.13). From there, strings of A or T, depending on retrogene orientation, were detected within the region. Poly(A)s were not extended into previously detected TSDs. If multiple putative Poly(A) tails were detected, the Poly(A) closest to the TSD was reported if one was present. If no TSD was detected, the Poly(A) closest to the retrogene was reported. Various extract distances were trialed ranging from 10 base pairs to 2000 base pairs with 70 base pairs being chosen to maximize

concordance between TSDs and poly(A) (concordance defined as the number of loci in which the TSD and poly(A) are within 20 bases of each other). Six of the nine samples had a maximum concordance with flanks of 50bp, 60bp, 70bp, 80bp, or 90bp, while the other three had a maximum concordance with flanks of 200bp. Though this method is limited in its ability to detect TSDs /Poly(A) combinations that are more than 70bp combined, as well as hallmarks present in retrocopies containing unannotated UTR sequence, these limitations are thought to be minor as the vast majority of retrogenes contained at least one hallmark.

Estimating the insertion rate of retrocopies

We selected contiguous genome assemblies from diverse dog samples insertion rate of retrocopies, including Mischka, Nala, China, Sandy, and Zoey relative to the mCanLor1.2 assembly. We discount all retrocopies that overlap a read-depth segmental duplication. For each canine retrocopy set, pairs of 1kB flanks were extracted from each retrocopy and using minimap2, mapped to mCanLor1.2 [43]. To be considered a counterpart, a retrocopy's flanks in one assembly must fall within 1.5 kB of a retrocopy in the comparison assembly. If one flank matches and the other does not, the pair as a whole is discounted. This analysis was performed by both mapping all retrocopy flanks to mCanLor1.2, and mCanLor1.2 retrocopy flanks to the selected canine assemblies. Resulting retrocopies were then compared to determine how many retrocopies were shared between mCanLor1.2 and a given canine assembly, how many were unique to mCanLor1.2, and how many were unique to the given canine assembly.

To further estimate rates of differences between genome assemblies, we identified single nucleotide differences among assemblies. All involved numbers are filtered to remove duplicated segments identified by read-depth. The autosomes from each assembly were aligned to mCanLor1.2 using minimap2 (v2.20) [43], with the additional flags of *-c --cs*, outputting differences in *PAF* format. Resultant *PAF* data was sorted by chromosome and position, and variants called using paftools from minimap2, with the flags *-I1000 -L1000*, outputting data in *VCF* format [43]. The resulting SNPs were used to estimate the number of generations that had passed since divergence between mCanLor1.2 and the other genome assemblies. We utilize the wolf SNP mutation rate estimated by Koch et al. of 4.5 * 10^{-9} per generation [50] and the size of an assembly's autosomal genome, minus duplicated segments, to determine generations passed as:

$$\frac{\frac{1}{2} * SNPs}{Size\ of\ the\ canine\ autosomal\ genome} * \frac{1}{Wolf\ mutation\ Rate}$$

This formula calculates the approximate number of generations since divergence. We divide the average number of unique retrocopies detected between a given canine assembly and mCanLor1.2 and divide by the estimated generational since divergence to get an estimate for the rate of retrocopy insertion [23, 51, 52].

Results

Duplications identified by self-alignment of the Mischka genome include segments with multiple duplicated copies

The UU_Cfam_GSD_1.0 assembly (accession GCA_011100685.1), derived from a female German Shepherd Dog named Mischka, has extensive annotation, and serves as the reference genome used by the Dog10K consortium [14, 30, 53]. This genome assembly was constructed using PacBio Sequel sequencing and consists of sequence assigned to 39 assembled chromosomes (chr1-38 and chrX), the mitochondria (chrM), and 2,158 unplaced contigs not localized to a chromosome. Due to the importance of this reference to ongoing research, we first performed a detailed analysis of duplications in this assembly. Using genome assembly self-alignment, we identified 491,661 unique duplicated segments of 1,000 bp or greater with at least 90% sequence identity in the Mischka assembly. Of these, 122,521 segments are located in the 38 canine autosomes or chromosome X and 369,135 duplications are located on unplaced contig sequences. Of the duplicated segments on the assembled chromosomes, 50,629 have a paralog on an unplaced contig (41.03%). The largest single duplicated segment is nearly 1.8 mB in length, located at chr8:74446738-76245570.

Because the same position can be included in multiple duplications, we merged segments together, resulting in 6,040 duplicated intervals in the assembled chromosomes and 1,940 duplicated intervals in the unplaced contigs (**Supplementary Table 3**). Duplications occupy 96.89 mB, or about 4.12%, of the assembled chromosomes. In sharp contrast, about 90.15% (115.8 mB) of the sequence on unplaced contigs is duplicated. In total, including unplaced contigs, 8.58% of the Mischka assembly is duplicated according to genome assembly self-alignment.

Duplications are not distributed uniformly along canine autosomes. Of the 5,525 duplicated segments located on the autosomes, 193 (3.49%) overlap the first megabase of the beginning of the chromosome, while 172 (3.11%) are located within the last megabase of the autosome. Relative to 1000 random permutations, this is a 2.5-fold increase for the first megabase (mean: 1.7%; max: 2.2%) and a 1.75-fold increase for the last megabase (mean: 1.6%; max: 2.2%).

Many of the duplications detected on the assembled chromosomes are small; 46.62% (57,120/122,521) of duplicated segments are shorter than 2 kB. Examination of the size distribution of duplicated segments smaller than 2 kB revealed three notable peaks at approximately 1,250, 1,375, and 1,775 bp. Inspection of the duplication landscape on UCSC Genome Browser identified numerous high-recurrence duplications that contained many short alignments linking a locus with multiple paralogous segments **(Figure 1)** [54]. We defined a high-recurrence segment as any area that contained at least four or more paralogous duplications smaller than 2.5 kB, and identify 1,492 high-recurrence duplications, totaling approximately 3.0234 mB. Overall, this makes up 3.095% of duplicated base pairs located by self-alignment. The segment with the largest number of copies is found at chr1:100216700-100219200, with 159 intersecting duplications. These high-recurrence duplications appear to be distributed throughout the Mischka genome.

Duplications identified in Mischka using read-depth are not uniformly distributed

To complement duplications identified via genome assembly self-alignments, we identified duplications based on read-depth. Illumina sequencing reads derived from Mischka were aligned to the Mischka assembly and processed using fastCN [28]. Repetitive sequences are masked in the assembly prior to alignment and the read-depth profiles are used to estimate absolute copy number in windows consisting of 1,000 unmasked positions. We define read-depth duplications as segments consisting of

four or more consecutive windows with an estimated copy number exceeding 2.5, resulting in 1,719 segmental duplications defined by read-depth **(Supplementary Table 4)**. We find that approximately 60% (30.6 mB) of unplaced contigs that contain at least four windows are duplicated. In comparison, 56.21 mB (1180 segments) is duplicated on the assembled chromosomes, making up 2.26% of the 38 autosomes and X chromosome **(Table 3)**.

The duplications identified by read-depth range in size from 10 kB to 1.6 mB and have a mean size of 48 kB and a median of 21 kB. We find that 57 (6.1%) duplications are present in the first megabase of the autosomes, while 36 (3.8%) are located in the last 1 mB of each autosome. Relative to random permutations, we find a 3.19-fold increase in duplication content in the first megabase of autosomes (mean: 1.9%; max: 3.5%), and a 2.1-fold increase in the last megabase (mean: 1.8%; max: 3.2%). The median copy number of all the duplications was 3.62, and the mean was 4.85. The maximum copy number of a given duplication was 78.3, located at chr27:33386819-33407975.

Most chromosomes have between 1 and 5% of their sequence annotated as a duplication by read-depth; however, 8.85% of chromosome 26 is duplicated. Examination revealed that a 2.4 mB region encompassing 6 segmental duplications, located at chr26:25,387,954-27,841,554, accounts for this high duplication percentage. This region has a median estimated copy number of 9.5.

Read-depth analysis identifies segments that are not duplicated by assembly self-alignment

To assess the concordance of the methods, we compared the position of duplications identified by genome self-alignment and read-depth. To account for detection differences between the methods, we only considered duplications that were located on assembled chromosomes, greater than 10 kB in size, have at least 95% identity, and were not located in the first megabase of an autosome. A total of 1,275 duplications identified by genome assembly self-alignment and 974 duplications identified by read-depth met these criteria **(Supplementary Table 3, 4)**.

By intersecting the 1,275 self-alignment duplications with read-depth copy number windows, we can determine a relative copy number for each self-alignment duplication. Only 1,211 duplications fully overlap at least one read-depth window. Of those 1,211, 575 (47.5%) have an estimated copy number greater than 2.5. Despite low agreement, both methods show similar trends at a chromosomal level, including an increase in duplication count on chr26 **(Supplementary Figure 2)**.

We next explored how many duplications in each set were found to be unique. By intersecting the self-alignment set and the read-depth set, we find that 543/1,275 self-alignment duplications (42.7%) and 440/974 read-depth duplications (45.2%) share any sort of overlap. The union of these two sets contains approximately 69.16 mB in total, 28.26 (40.9%) of which is duplicated in both sets. Self-alignment uniquely detects 27.87 mB (40.3%) and read-depth uniquely detects 13.02 mB (18.8%).

Similar duplication content is found in other canine assemblies

We expanded our self-alignment and read-depth analysis to eight additional recently published canine assemblies **(Table 1)** [30-37]. Most assemblies used some method of PacBio sequencing, predominantly Sequel, but the China and Yella assemblies used Oxford Nanopore (ONT) sequencing. We note that China and Wags are different Basenjis sequenced using different technologies by the same

study. While the authors of that paper state that China is the better assembly, we exclude China from the read-depth analysis due to a lack of available Illumina data [37].

Canine assemblies have between 1.8% (Tasha) and 5.6% (mCanLor1.2) of their genome identified as duplicated by genome assembly self-alignment **(Table 1)**. Assemblies with a low duplication content (below 2.5%) include Sandy, Tasha, and Zoey, whereas assemblies with a high percentage of duplications (above 4%) include Mischka, mCanLor1.2, Nala, Wags, and Yella. A similar divide in duplication content among assemblies remains after filtering for duplications above 10kB and increasing the sequence identity minimum to 95%. By analyzing read-depth, we see a slight change, where mCanLor1.2, Nala, and Yella have the highest fraction of their genome annotated as duplicated. The number and duplication status of unplaced contigs differs widely among the assemblies **(Table 3)**. The highest duplicated fraction is found in Mischka (90%) and Zoey (70%), with the others having less than 50% of their unplaced contigs duplicated.

Additionally, we explored 13,817 protein-coding genes for their duplication status as determined by read-depth. Illumina data from each sample was aligned to the Mischka assembly; China and Nala are removed from this analysis, due to lack of Illumina data and strong bias based on GC composition respectively **(Supplementary Figure 1)**. The copy number of each gene was then estimated using fastCN. We identify a total of 1,068 genes as duplicated in at least one canine; 222 of those genes are considered duplicated iny all seven canines. Further analysis indicated that the most highly duplicated genes have no annotated function and are not associated with a named gene **(Supplementary Table 5).**

To explore the impact of PacBio HiFi reads on representing duplications, we compared duplications identified by assembly self-alignment and read-depth in the mCanLor1.2 assembly. Using the same filtering criteria as used in Mischka, we find that a total of 121.67 mB is duplicated in the union of self-alignment and read-depth, a 1.7-fold increase over Mischka. We also find an increase in congruency between the two methods, as 60.9 mB is duplicated in both the self-alignment set and the read-depth set, making up 50.05% of the total base pairs, rising from Mischka's 40.9%.

Retrocopies found throughout canine genome assemblies are known frequent pseudogene producers

While exploring self-alignment duplications, we found that high-recurrence duplications in the Mischka assembly often corresponded to the exons of transcripts found in other species, suggesting the presence of gene retrocopies **(Figure 1)**. To confirm this, we identified candidate retrogenes in the Mischka assembly using BLAT. From an initial set of 19,972 multi-exon protein-coding genes, we identified 2,911 retrocopies from 988 parental genes, which includes 47 retrocopies that retained the full open reading frame of the parental gene **(Table 4, Supplementary Table 7)**. 2743 (94.2%) of these retrocopies are located on assembled chromosomes. We also detected an enrichment of retrocopies on chrX, with a 1.56-fold increase over 1,000 random permutations **(Supplementary Figure 3).**

By intersecting high-recurrence duplications with annotated retrocopies, we determine that 1,068/1,493 (71.58%) highly duplicated segments overlap with a retrocopy; in turn, 1,498/2,743 retrocopies on assembled chromosomes intersect with high-recurrence duplications (54.61%). On a base-pair level, retrocopies located in highly duplicated segments make up a total of 8.61 mB. Thus,

multiple pair-wise alignments among retrogenes derived from the same parent gene account for the highly duplicated regions identified by genome self-alignment.

We additionally investigated how many retrogenes retained the signature hallmarks of retrotransposition: the presence of target site duplication (TSD) and poly(A) tails [55]. Of the 2,738 retrocopies on assembled chromosomes in the Mischka genome that passed filtering, 2,702 (96.69%) show at least one of the two hallmarks **(Table 5)**. In total, 1,739 (63.51%) have only TSDs of at least 5 bp, 29 (1.06%) have only poly-A tails, and 934 (34.11%) have both signatures. The mean length of poly-A tails is 11.15 bases, and the mean length of TSDs is 11.29 bases, with 515 (55.14%) having TSDs 10 bp or longer. Of the 47 retrogenes that retain the parental ORF, all 47 have TSDs, while 23 also have a poly(A) tail. An example of an annotated parental retrocopy can be found in **Figure 2**.

We expanded the retrocopy analysis to all other canine assemblies, using protein-coding genes converted from Mischka coordinates as queries. Retrocopy counts are broadly similar across assemblies (**Table 4, Supplementary Table 7**). Mischka's slightly greater number of retrocopies appears to be accounted for by the presence of unplaced contigs, many of which correspond to duplicated sequence **(Table 2).**

The top ten retrocopy-producing parent genes were determined for each assembly **(Figure 3)**. In total, only eleven genes share the top ten spots across all assemblies. Most of the top annotated genes identifies known retrocopy producers in human and/or canine (*GAPDH, HNRNPA1, EEF1A1, HMGN2* and the *RPL* genes) [56-60].

Across all nine assemblies, an average of 859.7 retrocopies retain both hallmarks (32.4%). An average of 1702.7 retrocopies have only TSDs (64.85%), 27.8 retrocopies have only poly(A) tails (1.06%), and 35.4 retrocopies have neither hallmarks (1.35%) **(Table 5)**. The mean length of poly(A) tails is 11.21 bases, and the mean length of TSDs is 11.14 bases.

We also specifically analyze retrocopies that retain the parental ORF for presence of hallmarks **(Supplementary Table 1)**. In all cases, over 94% of full-ORF retrocopies show at least one hallmark, usually presence of a target-site duplication.

Finally, we filtered retrocopies for any overlap with segmental duplications identified by read-depth. We find that the Mischka assembly has a high prevalence of retrocopies present in segmental duplications, with 339 compared to the range of 42-193 found in other assemblies. We note that 125 of these retrocopies are present on unplaced contigs. No other canine assembly has any retrocopies mapped to unplaced contigs, indicating that Mischka retrocopies are frequently present in segmental duplications that get assembled onto unplaced contigs, in line with the increased duplication rate present in the self-alignment data.

Estimating the rate of retrocopy insertion

To identify the rate of retrocopy insertion, we compared five canine assemblies from diverse dogs (Mischka, Nala, China, Sandy, and Zoey) against mCanLor1.2. Only retrocopies located on the autosomal chromosomes that did not overlap a read-depth segmental duplication were considered (**Supplementary Table 2)**. We determined whether a retrocopy in a given canine assembly had a

counterpart in mCanLor1.2, and vice-versa. When comparing Mischka to mCanLor1.2, a total of 3099 unique retrocopies were present across both assemblies: 1591 (51.34%) are shared, 722 (23.3%) are unique to Mischka, and 786 (25.36%) are unique to mCanLor1.2. Thus, an average of 754 retrocopies have occurred since divergence between Mischka and mCanLor1.2. Across the other assembly comparisons, mCanLor1.2 has between 693-838 unique retrocopies, and the other assemblies have between 677-722 unique retrocopies.

We next estimated the number of generations since from mCanLor1.2. Excluding duplicated segments, there are 5,047,880 autosomal single nucleotide differences between the Mischka and mCanLor1.2 genome assemblies. Using the wolf mutation rate of $4.5 * 10^{-9}$ per generation [50], and with the Mischka non-duplicated autosomal genome size of 2,154,974,313 bp, we calculated that 260,270 generations have elapsed since divergence. By dividing the average amount of unique retrocopies over the number of generations, we estimate that the rate of retrocopy insertion as $2.898 * 10^{-3}$ per generation, or approximately 1 in 345 births, a surprisingly high rate. Insertion rates estimated using other canine assemblies range from $2.679 * 10^{-3}$ per generation to $2.959 * 10^{-3}$ per generation, or from 1 in 338 to 1 in 374 (average: 1/349.8) **(Supplementary Table 2)**.

Discussion

We analyzed the duplication content of nine recent canine assemblies using two methods: genome assembly self-alignment and read-depth analysis, identifying 1.74-7.05% of each genome as duplicated. We found that duplications were enriched in the first and last megabases of canine autosomes. Since canine autosomes are acrocentric [58], this suggests an enrichment of duplications in pericentromeric and subtelomeric regions, as found in other species [61].

The assembly self-alignment and read-depth approaches identified a range of duplication content in each assembly, ranging from 17.27 to 109.78 mB by self-alignment and 41.17-166.47 mB by read-depth. Overlap analysis shows a generally poor level of agreement, with most duplicated base pairs only identified by one of the methods. Thus, each of the assemblies offers an incomplete representation of the true duplication content in each sample. The assemblies differ in how they represent partially resolved duplications. For example, in the Mischka assembly, unresolved duplications were often retained as separate unplaced contigs that were not assigned to a chromosomal location. Other assemblies report few or no duplications on unplaced contigs. This difference may have implications in downstream analyses that utilize these assemblies.

Our initial analysis of duplications in the Mischka assembly revealed the presence of many short, high-recurrence duplications. Further investigation showed that these segments corresponded to multiple related retrogene insertions. To explore this further, we annotated 2,596-2,911 retrogenes across all assemblies, and found that a large majority displayed the hallmarks of retrotransposition such as poly(A) tails and target site duplications. This includes 27-47 retrogenes per assembly that appear to maintain intact parental ORFs. Whether these intact retrogenes are expressed or have any functional impact requires further investigation. We also note that our definition of retrogene boundaries is based on the existing annotation of genes in canines. This annotation of untranslated regions (UTRs) in these gene models is often incomplete, resulting in an underestimation of the size of retrogene insertions

(**Figure 2**), and underestimating the count of retrocopies with target-site duplications and/or poly(A) tails.

Additionally, in the Mischka assembly, we find that 339 retrocopies are located within larger segmental duplications with over half of those retrocopies being present on unplaced contigs. This suggests a pattern where a retrogene inserts into a novel locus, which is then subsequently duplicated, perhaps repeatedly, throughout the genome. A similar pattern of insertion-duplication has been described for retrogenes that may have functional effects in elephants and horses [62, 63].

Finally, we take advantage of retrocopy comparisons between dog and wolf assemblies to estimate an average retrogene insertion rate of $2.898 * 10^{-3}$ per generation, or about 1 in 345 births. This rate is based on SNP differences between assemblies, and utilizes several simplifying assumptions, including a uniform mutation rate and an average genome-wide divergence time [50]. This rate is substantially higher than that found in humans; between humans and chimps, one new gene retrocopy insertion was estimated for every 6000 births, based on the 1000 Genomes Project [64]. However, it is noted that long-read data enables a higher retrocopy insertion detection rate [65] so this is may be an underestimate of the true human insertion rate.

Other observations are consistent with a high rate of retrogene formation in dogs. A recent survey suggests that polymorphic retrogene insertions are 13-fold more common in dogs than in humans [13]. The most well-studied retrogenes in dogs are derived from the *FGF4* gene which cause the short-leg phenotype found in some breeds such as corgis and beagles [12]. *FGF4* retrogenes have occurred multiple times independently, with 7 independent copies identified to date [13, 66]. Additionally, dogs have a high rate of SINE polymorphism [34, 67]. Both SINE and retrogene insertions require the activity of LINE-1 elements. In contrast to SINEs, which do not encode any proteins, LINE-1s are autonomous retroelements that encode proteins required for their own retrotransposition [68]. LINE-1 encoded proteins predominantly mobilize the RNA that encodes them through a process known as *cis-preference* [69]. However, one or more LINE-1 encoded proteins can rarely act in *trans* to mobilize other RNAs such as SINEs or cellular mRNAs. We speculate that the high rate of SINE and retrogene insertion found in dogs results from a shared mechanism that involves a reduced level of *cis-preference* for the proteins encoded by the canine LINE-1 element.

Acknowledgements

We thank the numerous authors of the genome assemblies used in this analysis for publicly providing their data for the investigation. In addition, we thank Peter Schall and Emily Koch for helpful comments on this study.

Author Contributions

AKN and JMK conceived the study. AKN carried out the data-gathering and primary analyses, and MSB analyzed retrogenes for transposition hallmarks. MSB and JMK advised on the methodology and interpretation of results. AKN drafted the manuscript, and all authors reviewed and edited the final manuscript.

Data Availability

Final counts of each set of items (self-alignment duplications, read-depth duplications, genes in read-depth duplications, and total retrogenes) are located in the provided supplementary files. Individualized files for each canine assembly for each method are available on request. Relevant and custom code is available on GitHub at https://github.com/antnguye/dups_and_retrogenes_in_canine_assemblies. Hallmark analysis is available on GitHub at https://github.com/mblacksmith1996/Hallmark_Detection_Python_Scripts.

Tables

Sample Name	Sample Type	N_{50} (Mbp)	Primary Sequencing Technology	Accession Number
Mischka (canFam4)	German Shepherd	14.96	PacBio Sequel	GCA_011100685.1
Sandy	Dingo	40.7	PacBio Sequel/ONT PromethION	GCA_003254725.2
China	Basenji	37.76	ONT PromethION	GCA_013276365.2
mCanLor1.2	Greenland Wolf	34	PacBio HiFi	GCA_905319855.2
Tasha	Boxer	27.45	PacBio Sequel	GCA_000002285.4
Nala	German Shepherd	20.9	PacBio Sequel	GCA_008641055.3
Zoey	Great Dane	4.72	PacBio RSII	GCA_005444595.1
Wags	Basenji	3.13	PacBio Sequel	GCA_004886185.2
Yella	Labrador Retriever	0.091	ONT GridION	GCA_012045015.1

Table 1. Canine Assemblies Investigated.

Listed in each row is the name of the sample used for each assembly, the sample type, the N_{50} in megabases as reported by each publication, and the main sequencing technology used to construct the assembly. Aside from Mischka/canFam4 at the top, the rest of the assemblies are ordered by their contig N_{50}.

Name	Self-alignment Duplication Count	mBP occupied by duplications (% of the genome)	Self-Alignment Duplications filtered to 10kB and 95% Homology	mBP occupied by Self Alignment duplications filtered to 10kB and 95% (% of the genome)	mBP occupied by BISER duplications on unplaced contigs (% of the genome)
Mischka	7980	97.68 (4.15%)	3293	82.22 (3.49%)	115.79 (90.15%)
mCanLor1.2	5256	134.54 (5.58%)	781	109.78 (4.55%)	15.78 (53.05%)
China	5535	52.99 (2.27%)	841	27.35 (1.17%)	5.46 (34.73%)
Nala	5887	111.84 (4.69%)	1384	84.53 (3.55%)	5.41 (23.58%)
Sandy	5490	56.65 (2.42%)	848	30.79 (1.32%)	4.22 (33.31%)
Tasha	5571	42.26 (1.83%)	566	17.27 (0.75%)	0.43 (20.14%)
Wags	6993	94.4 (4.13%)	2398	72.39 (3.17%)	44.67 (36.98%)
Yella	6751	103.47 (4.32%)	1139	68.66 (2.87%)	N/A
Zoey	6571	55.9 (2.4%)	1494	31.36 (1.35%)	11.73 (69.54%)

Table 2. All Duplications found by Genome Assembly Self-Alignment.

Column 1 indicates the canine assembly's name. Column 2 indicates the number of merged duplications detected by genome assembly self-alignment. Column 3 indicates the total amount of base pairs, in mBP, occupied by column 2, along with the percentage of the genome represented. Columns 4 and 5 are the same as 2 and 3, but filtered to only include duplicated intervals greater than 10 kB in size, and higher than 95% sequence identity. Column 6 is the mBP of BISER duplications located on unplaced contigs, and the percentage of base pairs within contigs that are considered duplicated per assembly.

Name	Read-depth Duplications	Read-depth Duplications on Assembled Chromosomes	mBP on Assembled Chromosomes
Mischka	1719	1180	57.37 (2.43%)
mCanLor1.2	2420	2411	132.13 (5.6%)
Nala	6473	6470	166.47 (7.05%)
Sandy	2445	2424	61.05 (2.59%)
Tasha	1076	1076	43.26 (1.83%)
Wags	1094	804	49.05 (2.08%)
Yella	1239	1238	99.85 (4.23%)
Zoey	1036	781	41.17 (1.74%)

Table 3. All Duplications found by Read-depth.

Duplications were defined as at least four consecutive windows found by fastCN to have a copy number of at least 2.5. No further size limitation was implemented at this step. Column 1 contains canine name, column 2 contains the total number of duplications detected by read-depth, column 3 contains all duplications located on Chr1-38 + ChrX, and column 4 contains the total number of mBP made up by column 3.

Sample Name	Protein-Coding Genes	All Retrocopies (Parental Genes)	Assembled Retrocopies (Parental Genes)	Full-length ORF Retrocopies (Parental Genes)
Mischka (canFam4)	19972	2911 (988)	2743 (968)	47 (42)
mCanLor1.2	19449	2711 (952)	2703 (948)	39 (37)
China	19393	2655 (944)	2641 (937)	27 (25)
Nala	19490	2723 (970)	2721 (970)	35 (29)
Sandy	19382	2692 (959)	2690 (959)	35 (30)
Tasha	19020	2602 (914)	2602 (913)	31 (28)
Wags	19394	2596 (948)	2427 (907)	36 (31)
Yella	19192	2631 (922)	2631 (922)	32 (31)
Zoey	19380	2664 (940)	2617 (931)	30 (28)

Table 4. Retrocopies identified in each canine assembly.

Column two gives the number of protein-coding genes considered as potential gene parents in each assembly. Retrocopies in column 3 include all chromosomes, including mitochondrial and unplaced contigs, and column 4 contains only the assembled chromosomes, chr1-38 + X. The total number of retrocopies and the number that retain the open reading frame of the parental gene are shown.

Sample Name	Total Retrocopies on Assembled Chromosomes	Filtered Retrocopies	Any Hallmark	Both Hallmarks	TSD Only	Poly(A) Only	Neither
Mischka (canFam4)	2743	2738	2702	934	1739	29	36
mCanLor1.2	2703	2699	2666	867	1770	29	33
China	2641	2640	2606	853	1724	29	34
Nala	2721	2717	2678	881	1767	30	39
Sandy	2690	2690	2653	887	1736	30	37
Tasha	2602	2594	2558	885	1652	21	36
Wags	2427	2419	2380	804	1549	27	39
Yella	2631	2520	2488	796	1670	22	32
Zoey	2617	2613	2580	830	1717	33	33

Table 5. Hallmarks found in Retrocopies on Assembled Chromosomes.

Total retrocopies found in Table 4 were filtered prior to hallmark analysis. Retrocopies were only considered for hallmark analysis if they were at least 100 bp away from either end of a chromosome, did not overlap any genome assembly gap, and were on chr1-38 + chrX. Hallmarks were only searched for within 60 bp outside of a retrocopy, with total number found in Column 3: Filtered Retrocopies. Target site duplications (TSDs) were required to be a minimum of 5 bp. Numbers correspond to the number of retrocopies that have the presence of either hallmark, both hallmarks, only target site duplications, only poly(A) tails, or no evidence of either hallmark.

Figures

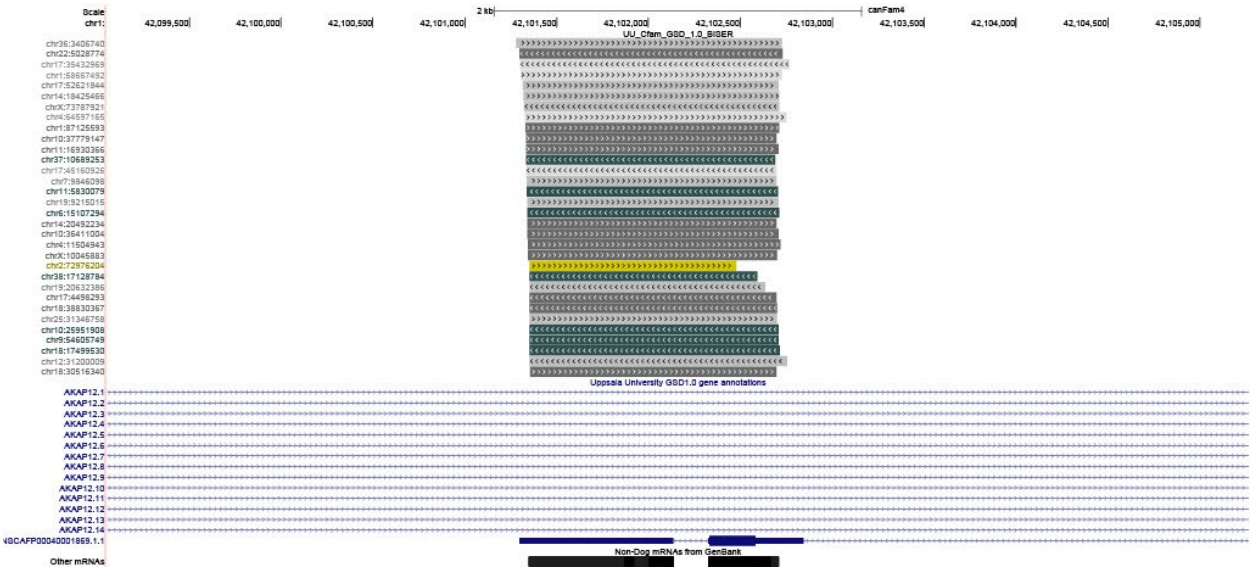


Figure 1. A High-Recurrence Duplicated Segment Detected by Genome Assembly Self-Alignment.

This screenshot of UCSC Genome Browser depicts the region chr1:42,099,050-42,105,269 in Mischka, a segment of the genome located in an intron of the gene *AKAP12*. The green and gray bars depict duplicated intervals detected by genome assembly self-alignment (BISER, top), whereas the blue lines with arrows indicate genes annotated in the Mischka assembly. The final track represents non-canine mRNAs from GenBank. Each gray/green bar has a location where the duplicated sequence is found elsewhere as indicated at locations given on the left at homology levels between 90 and 98%, with the singular yellow bar indicating an even greater homology between 98% and 99%. The parental retrogene for this duplicated sequence is *HMGN2*, a common parent gene.

TCCAGGTGGGATCTAGAGGACATTCCTCCAGGAAGTTCCCAACCGTCTTTTTTTTTTTGTTTTTGTTTTTTTTTTTTTTTTGT
TTTTTTTTTTTCTTTTGCAAACTTTTATTTGTGCACATTCAATTGAAAAAGATAACTTTTACAGGTTCTTTGGTGCCCATATTC
AGCATACAAAAATGTAAAAGACTATTCACAAATAAAACACATTAGCTCAAACTGGGAAAAAATAATGACACTAACGAAACAGTT
TAAGGTAGTGCACACTGTGACAGCTCTGCTTACGGGGACATGAGAATTACTGCGAAAATAAATAGAATGTCCTTTTGTAAAAGC
AGCAATGTCATGTCTTGTAAACTTCCTTTTTATATTACAGCAAGGTCTGAAATAAAACCCAAAGGGCCTGGAGCTCACACTGCA
GCCATGAATGAATATAAGAGTCTATATCCTCAATTACTGTTGGTCCAGTTTTCCTTTTGTTTTGAGTCCAATGTGCGGCAGGTT
TTAAAAGGTGCTCAGTTAGGTTCTAACCAGTTGAGTGATGGA**TTATAAAATTTCTTTCCTGCTTTCAAACACAAATGGGTAGAT**
CTGTTCCACAGCGGTAGCAACAGCCTTTACGTTGGGCCCTGTTACTGTGATACTCCCTGTTGAAAAAATCTGTAATGTAGCTCT
TAGAGATTTTATCCGATAGCACACAGCAGGATGAAGTTCAGGTTCATAACTGGCATGAGGTCTATTGTTCTTTGTGAATTCTGG
CAAACGAATTTCAAACGGCATGTTACACACTGCCAAAACGTTAACGACCTTAAAATCTGTAAATATTACCTGAAAACCTAATTT
CTGCAGGCTGCGAGCTAAACGTCTGGCACCAAATTTAGCTTCTTCTTCACTTGTTGCTCCAGTGCAAATAATTTTTCCTGAGGA
CCAAATTGTAGCTGTAATTCTAGGTTTTCTAAGCTTCATTAATACTTTTCCAACATCACGTTTATAAATTACATTTGCTCCTTC
CAAAGCAATCTTCCTTAAGTTCAAATGGCATCTTGTTCTAAAAACACAGACTACATTTGTAATTAGAATGTCCAATGCAACATC
ACTGTCTGCATCCATTGGGGTAGTTTAAAAACTTAGCTTCCCACCACGAAGATCACATCAGGGTTACCGTGGGAAGAGGGGGCG
GCGGCAGAGAACGCGGCGGCAGCGCTGCTGCCGGTTACGTCGGACGCGAAGGATAAGCTGGAGGACAAGGATGATGATGAGGCG
AGCGCGCTCCCCAGTCTCCATGGAACTCCCGCCCCGGCCCGTTGCTAGGAGCCCCCGAGGCTCGGGCTCCGTCTTTCGTGCTTT
GTTGTTCCGGGTACTGGCTCCCAACCGTCTTAATGTTAATTCCTTGCCATAGGGAAAAACAACCTGGACAATGGCAAGCCTAAA
AAAGTTCTTTTGAAAA

Figure 2. Annotated Retrocopy in Mischka.

Depicted is a retrocopy for the gene *TBPL1*. The original gene is located at chr1:27,131,596-27,168,730, and the retrocopy is found at chr9:29,216,113-29,217,556. This retrocopy is in the reverse orientation relative to the cDNA. Green highlights indicate target-site duplications, and blue highlights indicate the poly(A) tail, both signs of retrocopy formation. Pink denotes the unannotated UTR sequence, red denotes the original retrocopy hit, and the bolded and underlined sequence is the original ORF of *TBPL1.*

Mischka	mCanLor1.2	China	Nala	Tasha	Wags	Yella	Zoey	Sandy
GAPDH (103)	HNRNPA1 (108)	HNRNPA1 (98)	HNRNPA1 (93)	GAPDH (97)	HNRNPA1 (90)	HNRNPA1 (94)	HNRNPA1 (90)	HNRNPA1 (97)
HNRNPA1 (97)	EEF1A1 (93)	EEF1A1 (86)	EEF1A1 (89)	HNRNPA1 (94)	EEF1A1 (88)	EEF1A1 (87)	EEF1A1 (89)	EEF1A1 (92)
EEF1A1 (92)	GAPDH (66)	GAPDH (63)	GAPDH (61)	EEF1A1 (90)	GAPDH (58)	GAPDH (63)	GAPDH (60)	GAPDH (62)
RPL17 (52)	RPL21 (46)	RPL17 (55)	RPL17 (56)	RPL17 (51)	LOC119863874 (30)	RPL17 (55)	RPL17 (55)	RPL17 (57)
LOC100686340 (34)	LOC100686340 (31)	LOC119863874 (31)	LOC100686340 (32)	LOC119863874 (31)	LOC100686340 (27)	LOC119863874 (33)	RPL31 (30)	LOC100686340 (35)
LOC119863874 (31)	LOC119863874 (29)	LOC100686340 (30)	LOC119863874 (31)	LOC100686340 (31)	HMGN2 (27)	LOC100686340 (32)	LOC119863874 (30)	LOC119863874 (31)
HMGN2 (28)	RPL17 (29)	HMGN2 (29)	RPL31 (28)	RPL32_2 (29)	RPL32_2 (26)	RPL31 (30)	LOC100686340 (30)	RPL21 (29)
RPL31 (28)	RPL31 (27)	RPL31 (28)	HMGN2 (28)	HMGN2 (28)	RPL21 (25)	RPL32_2 (29)	HMGN2 (28)	RPL31 (28)
RPL32_2 (28)	LOC100687749 (27)	RPL32_2 (28)	RPL32_2 (28)	RPL31 (27)	RPL17 (25)	HMGN2 (28)	RPL32_2 (28)	HMGN2 (28)
LOC100687749 (26)	HMGN2 (24)	RPL21 (28)	RPL21 (27)	RPL21 (26)	RPL31 (24)	RPL21 (28)	RPL21 (27)	RPL32_2 (28)

Figure 3. Most Frequent Retrocopy Parental Genes.

Each column lists the 10 genes that give rise to the most identified retrocopies in each assembly. Retrocopies presented here are limited to the assembled chromosomes. The number of retrocopies from each parent are given in parentheses. The same genes are colored identically to one another to quickly determine presence of reoccurring genes. Only 11 parent genes are represented across all nine assemblies.

Supplementary Tables

Sample Name	Total FullORF Retrocopies	Any Hallmark	Both Hallmarks	TSD Only	Poly(A) Only	Neither
Mischka (canFam4)	47	47	23	24	0	0
mCanLor1.2	46	45	21	24	1	0
China	35	35	14	21	0	0
Nala	43	42	16	26	0	1
Sandy	42	42	18	23	1	0
Tasha	36	36	16	20	0	0
Wags	38	37	16	21	0	1
Yella	40	40	16	23	1	0
Zoey	35	34	13	21	0	1

Supplementary Table 1. FullORF Retrocopy Hallmarks.

This table mimics the same pattern as seen in Table 4, except the only retrocopies present in this table retain the full open reading frame as their parent gene. A majority of fullORF retrocopies have target-site duplications, with extremely rare occurrences without. 37-48% of fullORF retrocopies have evidence of both hallmarks.

Assembly Name	Retrocopies after Filtering for Duplications	Total Retrocopies Between Assemblies	Retrocopies Unique to Assembly (%)	Retrocopies Unique to mCanLor1.2 (%)	Estimated Generations Since Divergence	Insertion Rate * 10^{-3} (1 in n births)
Mischka	2335	3099	722 (23.3)	786 (25.4)	260,270	2.898 (345)
China	2341	3080	682 (22.1)	750 (24.4)	267,321	2.679 (374)
Nala	2326	3144	713 (22.7)	838 (26.7)	262,277	2.959 (338)
Sandy	2387	3104	718 (23.1)	743 (23.9)	261,321	2.956 (338)
Zoey	2329	3073	677 (22.0)	693 (22.6)	240,094	2.826 (354)

Supplementary Table 2. Retrocopies Unique to Each Assembly Compared to mCanLor1.2

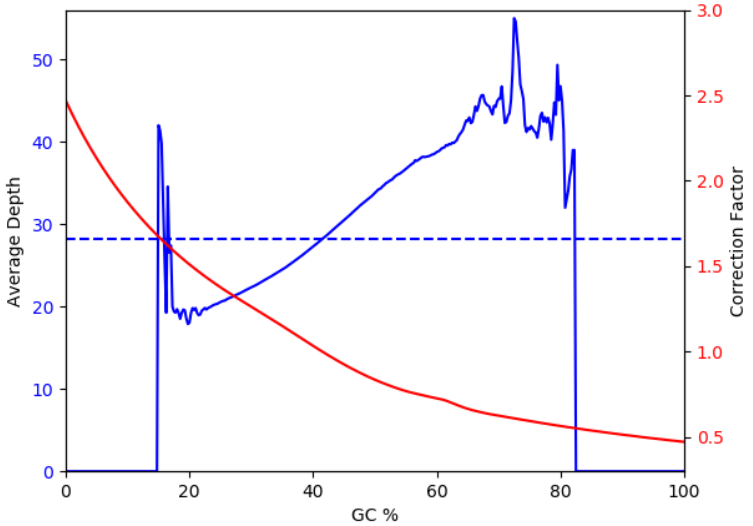
This table describes the five assemblies retained for retrocopy comparison against mCanLor1.2, the Greenland Wolf Assembly. Column 1: Assembly Name. Column 2 contains the total sum of retrocopies between the assembly and mCanLor1.2, only counting shared retrocopies once. Columns 3 and 4 show how many retrocopies are unique to the assembly and to mCanLor1.2 respectively, with the total percentage that that makes up of the sum in Column 2. Column 5 shows the number of generations elapsed since divergence between the assembly in mCanLor1.2. Column 6 shows the final determined insertion rate, calculated by dividing the average between Columns 3 and 4 over the generations in Column 5, shown as an insertion rate of X * 10^{-3}, and presented as 1 in n births.

Supplementary Files:

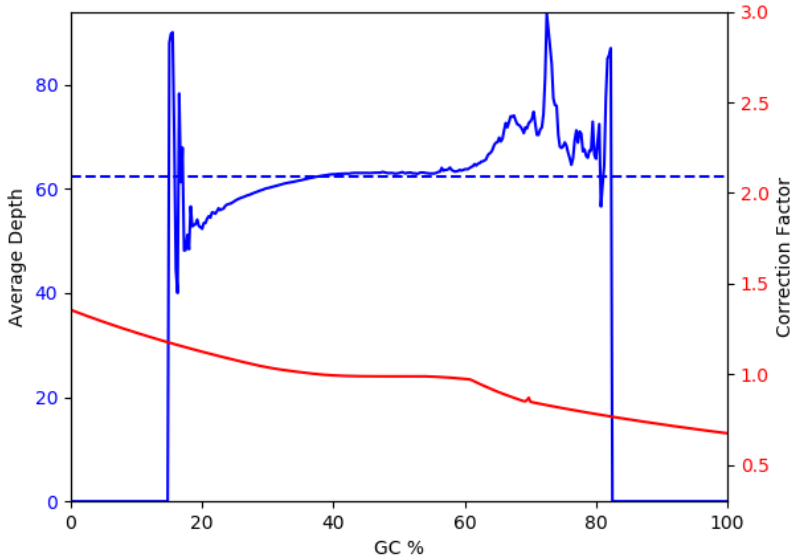
Table S3. All Self-Alignment Duplicated Intervals
Table S4. All Read-depth Segmental Duplications
Table S5. Duplicated Genes found by Read-depth
Table S6. All Retrocopies across All Assemblies

Supplementary Figures

A)



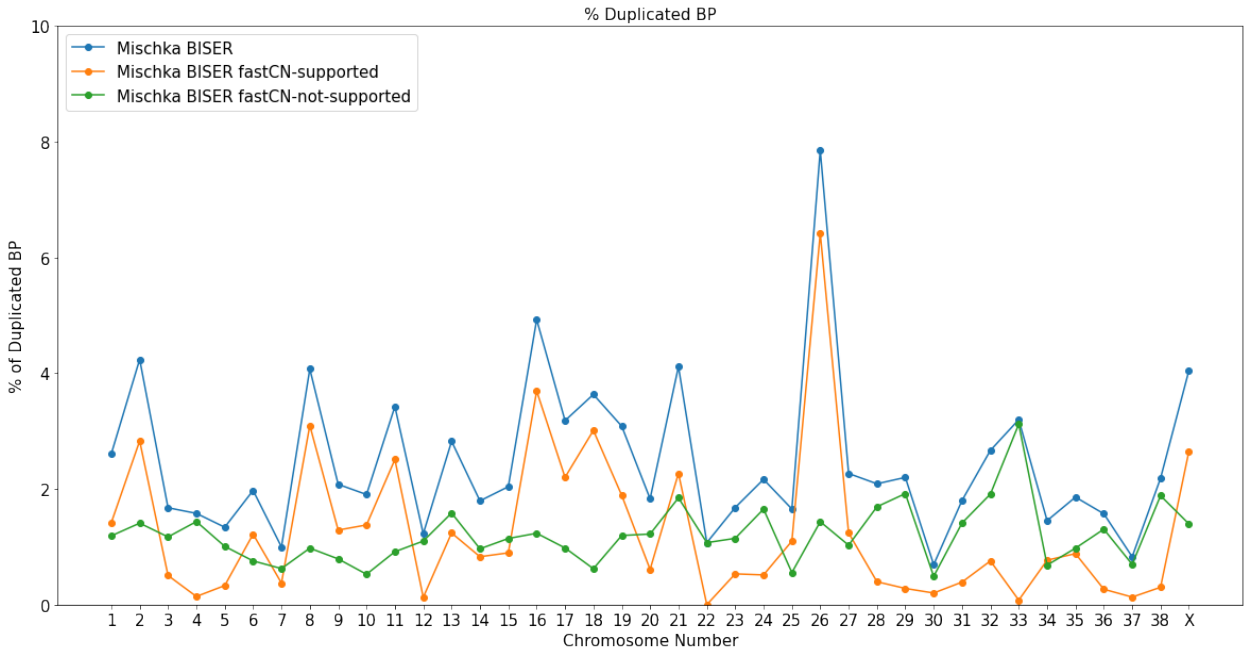
B)



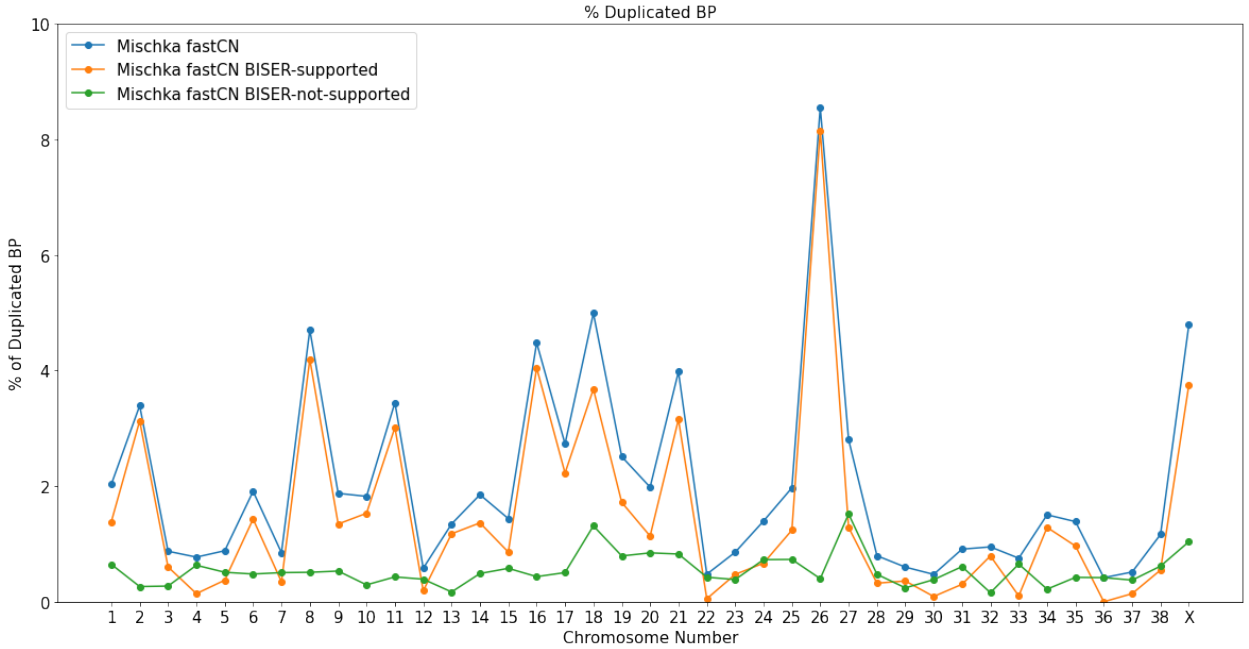
Supplementary Figure 1. GC Coverage Profiles.

Plots of coverage by local GC content are produced by fastCN. The dashed blue line shows mean read-depth. The solid blue line shows the coverage as a function of local GC content. The red line shows the correction factor calculated to normalize coverage based on GC content. (A) Depicted is the curve for Nala, which shows a skewed coverage across different GC levels, suggesting bias in the Illumina data. (B) Depicted is the curve for Mischka, showcasing a more uniform coverage profile across GC levels.

A



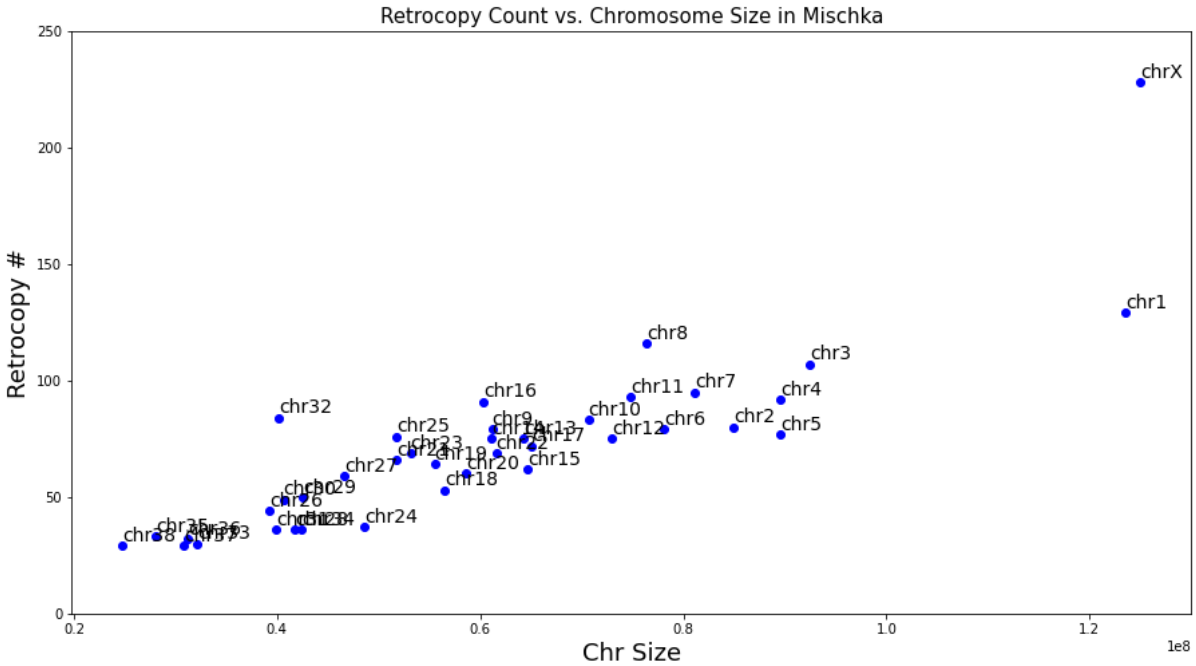
B



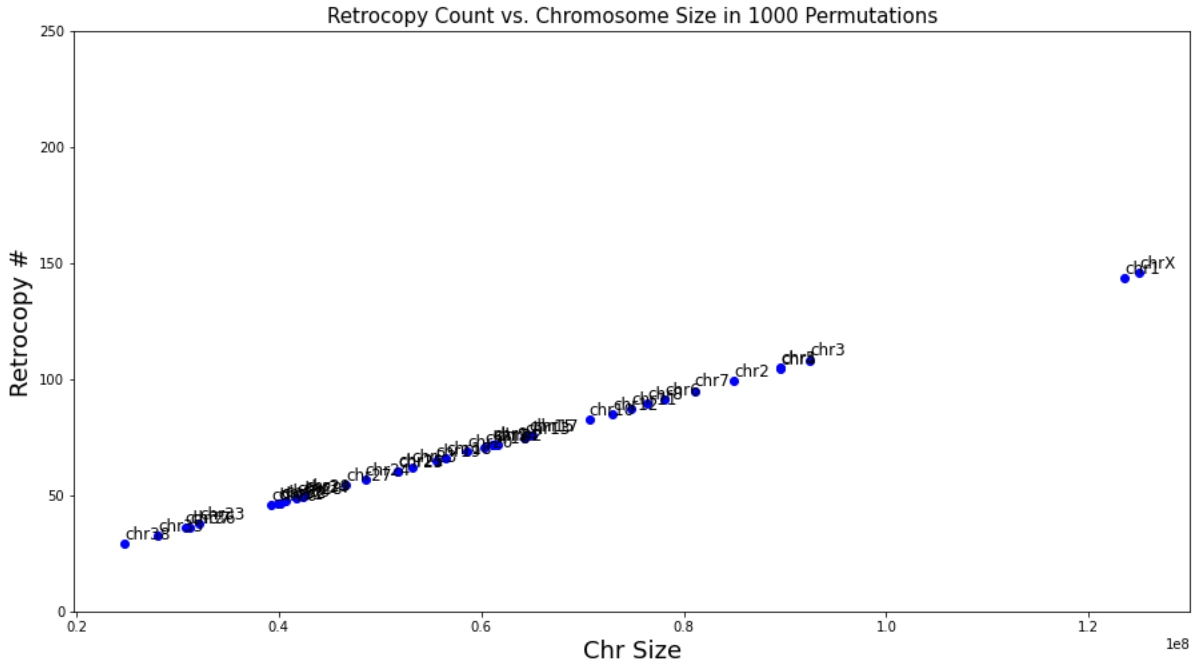
Supplementary Figure 2. Duplication Content of Each Chromosome.

The X-axis represents each assembled chromosome, and the Y-axis represents what fraction of that chromosome is duplicated as a % (A) Blue represents duplications found by genome assembly self-alignment (BISER). Orange is BISER duplications supported by read-depth, and green is unsupported. (B) Blue represents duplications detected by read-depth. Orange shows support from BISER, and green is unsupported.

A



B



Supplementary Figure 3. Retrocopy Count vs. Chromosome Size.

A) Depiction of the number of retrocopies on each of the assembled chromosomes (chr1-38 + X) found in the Mischka assembly vs. the length of each chromosome. ChrX has 228 total retrocopies. B) Depiction of the number of retrocopies on each of the assembled chromosomes found in 1000 random permutations vs. the length of each chromosome. ChrX has a mean of 146 retrocopies and a max of 188 retrocopies. Mischka contains a 1.56-fold increase of detected retrogenes on chrX.

References

1. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, et al: **Copy number variation: new insights in genome diversity.** *Genome Res* 2006, **16:**949-961.
2. Conrad B, Antonarakis SE: **Gene Duplication: A Drive for Phenotypic Diversity and Cause of Human Disease.** *Annual Review of Genomics and Human Genetics* 2007, **8:**17-35.
3. Dehal P, Boore JL: **Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate.** *PLOS Biology* 2005, **3:**e314.
4. Abdullaev ET, Umarova IR, Arndt PF: **Modelling segmental duplications in the human genome.** *BMC Genomics* 2021, **22:**496.
5. Rody HVS, Baute GJ, Rieseberg LH, Oliveira LO: **Both mechanism and age of duplications contribute to biased gene retention patterns in plants.** *BMC Genomics* 2017, **18:**46.
6. Payen C, Koszul R, Dujon B, Fischer G: **Segmental Duplications Arise from Pol32-Dependent Repair of Broken Forks through Two Alternative Replication-Based Mechanisms.** *PLOS Genetics* 2008, **4:**e1000175.
7. Du RQ, Lu CC, Jiang ZW, Li SL, Ma RX, An HJ, Xu MF, An Y, Xia YK, Jin L, et al: **Efficient typing of copy number variations in a segmental duplication-mediated rearrangement hotspot using multiplex competitive amplification.** *Journal of Human Genetics* 2012, **57:**545-551.
8. Esnault C, Maestre J, Heidmann T: **Human LINE retrotransposons generate processed pseudogenes.** *Nat Genet* 2000, **24:**363-367.
9. Kaessmann H, Vinckenbosch N, Long MY: **RNA-based gene duplication: mechanistic and evolutionary insights.** *Nature Reviews Genetics* 2009, **10:**19-31.
10. Casola C, Betran E: **The Genomic Impact of Gene Retrocopies: What Have We Learned from Comparative Genomics, Population Genomics, and Transcriptomic Analyses?** *Genome Biol Evol* 2017, **9:**1351-1373.
11. Tutar Y: **Pseudogenes.** *Comparative and Functional Genomics* 2012.
12. Brown EA, Dickinson PJ, Mansour T, Sturges BK, Aguilar M, Young AE, Korff C, Lind J, Ettinger CL, Varon S, et al: **FGF4 retrogene on CFA12 is responsible for chondrodystrophy and intervertebral disc disease in dogs.** *Proceedings of the National Academy of Sciences* 2017, **114:**11476-11481.
13. Batcher K, Varney S, York D, Blacksmith M, Kidd JM, Rebhun R, Dickinson P, Bannasch D: **Recent, full-length gene retrocopies are common in canids.** *Genome Res* 2022, **32:**1602-1611.
14. Meadows JRS, Kidd JM, Wang GD, Parker HG, Schall PZ, Bianchi M, Christmas MJ, Bougiouri K, Buckley RM, Hitte C, et al: **Genome sequencing of 2000 canids by the Dog10K consortium advances the understanding of demography, genome function and architecture.** *Genome Biol* 2023, **24:**187.
15. Bianchi CA, Marcellin-Little DJ, Dickinson PJ, Garcia TC, Li CF, Batcher K, Bannasch DL: **FGF4L2 retrogene copy number is associated with intervertebral disc calcification and vertebral geometry in Nova Scotia Duck Tolling Retrievers.** *Am J Vet Res* 2023, **84**.
16. Išerić H, Alkan C, Hach F, Numanagić I: **Fast characterization of segmental duplication structure in multiple genome assemblies.** *Algorithms for Molecular Biology* 2022, **17:**4.
17. Pan W, Lonardi S: **Accurate detection of chimeric contigs via Bionano optical maps.** *Bioinformatics* 2019, **35:**1760-1762.
18. Hartasanchez DA, Braso-Vives M, Heredia-Genestar JM, Pybus M, Navarro A: **Effect of Collapsed Duplications on Diversity Estimates: What to Expect.** *Genome Biol Evol* 2018, **10:**2899-2905.
19. Alkan C, Coe BP, Eichler EE: **APPLICATIONS OF NEXT-GENERATION SEQUENCING Genome structural variation discovery and genotyping.** *Nature Reviews Genetics* 2011, **12:**363-375.

20. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al: **The complete sequence of a human genome.** *Science* 2022, **376:**44-53.
21. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM: **De novo assembly of haplotype-resolved genomes with trio binning.** *Nature Biotechnology* 2018, **36:**1174-+.
22. Frantz LAF, Mullin VE, Pionnier-Capitan M, Lebrasseur O, Ollivier M, Perri A, Linderholm A, Mattiangeli V, Teasdale MD, Dimopoulos EA, et al: **Genomic and archaeological evidence suggests a dual origin of domestic dogs.** *Science* 2016, **352:**1228-1231.
23. 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.
24. Hoeppner MP, Lundquist A, Pirun M, Meadows JRS, Zamani N, Johnson J, Sundström G, Cook A, FitzGerald MG, Swofford R, et al: **An Improved Canine Genome and a Comprehensive Catalogue of Coding Genes and Non-Coding Transcripts.** *PLOS ONE* 2014, **9:**e91172.
25. Nicholas TJ, Cheng Z, Ventura M, Mealey K, Eichler EE, Akey JM: **The genomic architecture of segmental duplications and associated copy number variants in dogs.** *Genome Research* 2009, **19:**491-499.
26. Binversie EE, Baker LA, Engelman CD, Hao Z, Moran JJ, Piazza AM, Sample SJ, Muir P: **Analysis of copy number variation in dogs implicates genomic structural variation in the development of anterior cruciate ligament rupture.** *PLoS One* 2020, **15:**e0244075.
27. Serres-Armero A, Davis BW, Povolotskaya IS, Morcillo-Suarez C, Plassais J, Juan D, Ostrander EA, Marques-Bonet T: **Copy number variation underlies complex phenotypes in domestic dog breeds and other canids.** *Genome Res* 2021, **31:**762-774.
28. Pendleton AL, Shen F, Taravella AM, Emery S, Veeramah KR, Boyko AR, Kidd JM: **Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication.** *BMC Biology* 2018, **16:**64.
29. 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.
30. Wang C, Wallerman O, Arendt ML, Sundstrom E, Karlsson A, Nordin J, Makelainen S, Pielberg GR, Hanson J, Ohlsson A, et al: **A novel canine reference genome resolves genomic architecture and uncovers transcript complexity.** *Commun Biol* 2021, **4:**185.
31. Sinding MS, Gopalakrishnan S, Raundrup K, Dalen L, Threlfall J, Darwin Tree of Life Barcoding c, Wellcome Sanger Institute Tree of Life p, Wellcome Sanger Institute Scientific Operations DNAPc, Tree of Life Core Informatics c, Darwin Tree of Life C, Gilbert T: **The genome sequence of the grey wolf, Canis lupus Linnaeus 1758.** *Wellcome Open Res* 2021, **6:**310.
32. Player RA, Forsyth ER, Verratti KJ, Mohr DW, Scott AF, Bradburne CE: **A novel canis lupus familiaris reference genome improves variant resolution for use in breed-specific GWAS.** *Life Science Alliance* 2021, **4**.
33. Jagannathan V, Hitte C, Kidd JM, Masterson P, Murphy TD, Emery S, Davis B, Buckley RM, 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.
34. Halo JV, Pendleton AL, Shen F, Doucet AJ, Derrien T, Hitte C, Kirby LE, Myers B, Sliwerska E, Emery S, et al: **Long-read assembly of a Great Dane genome highlights the contribution of GC-rich sequence and mobile elements to canine genomes.** *Proceedings of the National Academy of Sciences* 2021, **118**.

35. Field MA, Rosen BD, Dudchenko O, Chan EKF, Minoche AE, Edwards RJ, Barton K, Lyons RJ, Tuipulotu DE, Hayes VM, et al: **Canfam_GSD: De novo chromosome-length genome assembly of the German Shepherd Dog (Canis lupus familiaris) using a combination of long reads, optical mapping, and Hi-C.** *Gigascience* 2020, **9**.
36. Field MA, Yadav S, Dudchenko O, Esvaran M, Rosen BD, Skvortsova K, Edwards RJ, Keilwagen J, Cochran BJ, Manandhar B, et al: **The Australian dingo is an early offshoot of modern breed dogs.** *Sci Adv* 2022, **8:**eabm5944.
37. Edwards RJ, Field MA, Ferguson JM, Dudchenko O, Keilwagen J, Rosen BD, Johnson GS, Rice ES, Hillier LD, Hammond JM, et al: **Chromosome-length genome assembly and structural variations of the primal Basenji dog (Canis lupus familiaris) genome.** *BMC Genomics* 2021, **22:**188.
38. Shen F, Kidd JM: **Rapid, Paralog-Sensitive CNV Analysis of 2457 Human Genomes Using QuicK-mer2.** *Genes* 2020, **11:**141.
39. Verscheure S, Backeljau T, Desmyter S: **In silico discovery of a nearly complete mitochondrial genome Numt in the dog (Canis lupus familiaris) nuclear genome.** *Genetica* 2015, **143:**453-458.
40. Quinlan AR: **BEDTools: The Swiss-Army Tool for Genome Feature Analysis.** *Curr Protoc Bioinformatics* 2014, **47:**11 12 11-34.
41. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573-580.
42. Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, Sahinalp SC: **mrsFAST: a cache-oblivious algorithm for short-read mapping.** *Nat Methods* 2010, **7:**576-577.
43. Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics* 2018, **34:**3094-3100.
44. Shumate A, Salzberg SL: **Liftoff: accurate mapping of gene annotations.** *Bioinformatics* 2021, **37:**1639-1643.
45. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H: **Twelve years of SAMtools and BCFtools.** *Gigascience* 2021, **10**.
46. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12:**656-664.
47. Li H: **New strategies to improve minimap2 alignment accuracy.** *Bioinformatics* 2021, **37:**4572-4574.
48. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16:**276-277.
49. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147:**195-197.
50. Koch E, Schweizer RM, Schweizer TM, Stahler DR, Smith DW, Wayne RK, Novembre J: **De novo mutation rate estimation in wolves of known pedigree.** *Mol Biol Evol* 2019, **36:**2536-2547.
51. Freedman AH: **Genome Sequencing Highlights the Dynamic Early History of Dogs (vol 10, e1004016, 2014).** *Plos Genetics* 2014, **10**.
52. Larson G, Bradley DG: **How Much Is That in Dog Years? The Advent of Canine Population Genomics.** *Plos Genetics* 2014, **10**.
53. Wang G-D, Larson G, Kidd JM, vonHoldt BM, Ostrander EA, Zhang Y-P: **Dog10K: the International Consortium of Canine Genome Sequencing.** *National Science Review* 2019, **6:**611-613.
54. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12:**996-1006.
55. Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD: **Molecular archeology of L1 insertions in the human genome.** *Genome Biol* 2002, **3:**research0052.
56. Gao X, Li Y, Adetula AA, Wu Y, Chen H: **Analysis of new retrogenes provides insight into dog adaptive evolution.** *Ecology and Evolution* 2019, **9:**11185-11197.

57. Liu YJ, Zheng D, Balasubramanian S, Carriero N, Khurana E, Robilotto R, Gerstein MB: **Comprehensive analysis of the pseudogenes of glycolytic enzymes in vertebrates: the anomalously high number of GAPDH pseudogenes highlights a recent burst of retrotrans-positional activity.** *BMC Genomics* 2009, **10:**480.
58. Breen M: **Canine cytogenetics--from band to basepair.** *Cytogenet Genome Res* 2008, **120:**50-60.
59. Marquez LM, Miller DJ, MacKenzie JB, Van Oppen MJ: **Pseudogenes contribute to the extreme diversity of nuclear ribosomal DNA in the hard coral Acropora.** *Mol Biol Evol* 2003, **20:**1077-1086.
60. Ciomborowska-Basheer J, Staszak K, Kubiak MR, Makalowska I: **Not So Dead Genes-Retrocopies as Regulators of Their Disease-Related Progenitors and Hosts.** *Cells* 2021, **10**.
61. Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ: **Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication.** *Nature* 2005, **437:**94-100.
62. Sulak M, Fong L, Mika K, Chigurupati S, Yon L, Mongan NP, Emes RD, Lynch VJ: **TP53 copy number expansion is associated with the evolution of increased body size and an enhanced DNA damage response in elephants.** *Elife* 2016, **5**.
63. Batcher K, Varney S, Raudsepp T, Jevit M, Dickinson P, Jagannathan V, Leeb T, Bannasch D: **Ancient segmentally duplicated LCORL retrocopies in equids.** *PLoS One* 2023, **18:**e0286861.
64. Ewing AD, Ballinger TJ, Earl D, Broad Institute Genome S, Analysis P, Platform, Harris CC, Ding L, Wilson RK, Haussler D: **Retrotransposition of gene transcripts leads to structural variation in mammalian genomes.** *Genome Biol* 2013, **14:**R22.
65. Feng X, Li H: **Higher Rates of Processed Pseudogene Acquisition in Humans and Three Great Apes Revealed by Long-Read Assemblies.** *Mol Biol Evol* 2021, **38:**2958-2966.
66. Batcher K, Dickinson P, Maciejczyk K, Brzeski K, Rasouliha SH, Letko A, Drogemuller C, Leeb T, Bannasch D: **Multiple FGF4 Retrocopies Recently Derived within Canids.** *Genes (Basel)* 2020, **11**.
67. Wang W, Kirkness EF: **Short interspersed elements (SINEs) are a major source of canine genomic diversity.** *Genome Res* 2005, **15:**1798-1808.
68. Beck CR, Garcia-Perez JL, Badge RM, Moran JV: **LINE-1 elements in structural variation and disease.** *Annu Rev Genomics Hum Genet* 2011, **12:**187-215.
69. Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV: **Human L1 retrotransposition: cis preference versus trans complementation.** *Mol Cell Biol* 2001, **21:**1429-1439.