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% of unplaced contigs. Highlighting the continued difficulty in properly assembling duplications, less than half of read-depth and assembly alignment duplications overlap, but the mCanLor1.2 Greenland wolf assembly shows greater concordance. Further study shows the presence of multiple segments that have alignments to four or more duplicate copies. These high-recurrence duplications correspond to gene retrocopies. We identified 3,892 candidate retrocopies from 1,316 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 between the wolf and other canine assemblies found that approximately 92% of retrocopy insertions are shared between assemblies. By calculating the number of generations since genome divergence, we estimate that new retrocopy insertions appear, on average, in 1 out of 3,514 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, genome

Significance

Duplications are highly influential on evolution, but are commonly misassembled, especially in lagging genomic groups like canines. We assessed nine canine assemblies for duplication presence, and found enrichment for acrocentric regions, misattribution of duplications to unplaced contigs, and the presence of short, high-recurrence duplications. Investigating further, we find high numbers of retrocopies retaining hallmarks present in the canine assemblies, and determine a rate of novel retrocopy insertion at 1 in 3,514 births.

Introduction

Gene duplication is a major driver of evolution, but correctly resolving duplications in genome assemblies remains challenging. Gene duplication is a form of structural variation, which refers to chromosomal alterations that are typically larger than 1000 bases (1 Kb) in size, and includes other changes such as insertions, deletions, inversions, and other copy number changes [1]. Copy number gains 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 retrocopies 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 in 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 [17]. 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 [18]. False duplications can also arise due to misassembly. This can occur when contigs created from divergent alleles are mistakenly represented as being duplicated [19-21]. 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 [20]. Longer sequencing reads coupled with new methodologies for haplotype-resolved assembly [22, 23] 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 [24]. The first canine reference genome was released in 2005, derived from Tasha the Boxer [25]. This was further refined into canFam3.1, released in 2014, also based on Tasha [26]. 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 [27]. Array-CGH data showed that many duplications were copy-number polymorphic [28]. Additional studies of canine CNVs have linked these variations to conditions such as ACL rupture [29] and breed-specific morphological structure [30]. 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 [31, 32].

Since the release of canFam3.1, multiple assemblies of dogs and wolves have been generated using long-read technology [33-40]. 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. However, limited analyses of the duplication content of these assemblies have been performed, with most description focused on specific loci, such as the amylase region, or other regions of suspected misassembly [41]. To address the lack of analysis of duplications in these long-read genomes, 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 [31, 41]. We further perform a search for retrocopy 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 canine assemblies, and identify retrocopies as a major driver of gene duplications in canine genomes.

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, 33, 42]. 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 pairwise 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.

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 **(Supplementary Figure 1)** [43].

158 We defined a high-recurrence segment as any area that contained at least four or more
159 paralogous duplications smaller than 2.5 Kb, and identified 1,492 high-recurrence duplications,
160 totaling approximately 3.0234 Mb. Overall, this makes up 3.095% of duplicated base pairs
161 located by self-alignment. The segment with the largest number of copies is found at
162 chr1:100216700-100219200, with 159 intersecting duplications. These high-recurrence
163 duplications appear to be distributed throughout the Mischka genome.
164 Because the same locus can be included in multiple duplications at varying levels of
165 sequence similarity, we merged segments together, resulting in 6,040 duplicated intervals in the
166 assembled chromosomes and 1,940 duplicated intervals in the unplaced contigs (**Supplementary**
167 **Figure 2, Supplementary Table 1 and 5**). Duplications occupy 96.89 Mb, or about 4.12%, of
168 the assembled chromosomes. In sharp contrast, about 90.15% (115.8 Mb) of the sequence on
169 unplaced contigs is duplicated. In total, including unplaced contigs, 8.58% of the Mischka
170 assembly is duplicated according to genome assembly self-alignment.
171 Duplications are not distributed uniformly along canine autosomes. Of the 5,525
172 duplicated segments located on the autosomes, 193 (3.49%) overlap the first megabase of the
173 beginning of the chromosome, while 172 (3.11%) are located within the last megabase of the
174 autosomes. Relative to 1000 random permutations, this is a 2.5-fold increase for the first
175 megabase (mean: 1.7%; max: 2.2%) and a 1.75-fold increase for the last megabase (mean: 1.6%;
176 max: 2.2%).

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

178 To complement duplications identified via genome assembly self-alignments, we
179 identified duplications based on read-depth. Illumina sequencing reads derived from Mischka
180 were aligned to the Mischka assembly and processed using fastCN [31]. Repetitive sequences are
181 masked in the assembly prior to alignment and the read-depth profiles are used to estimate
182 absolute copy number in windows consisting of 1,000 unmasked positions. We define read-depth
183 duplications as segments consisting of four or more consecutive windows with an estimated copy
184 number exceeding 2.5 and with a minimum size of 10 Kb, resulting in 1,573 segmental
185 duplications defined by read-depth **(Supplementary Table 2 and 6)**. We find that approximately
186 60% (30.6 Mb) of unplaced contigs that contain at least four windows are duplicated. In
187 comparison, 56.21 Mb (1033 segments) is duplicated on the assembled chromosomes, making up
188 2.39% of the 38 autosomes and X chromosome.
189 The duplications identified by read-depth range in size from 10 Kb to 1.6 Mb and have a
190 mean size of 48 Kb and a median of 21 Kb. We find that 57 (6.1%) duplications are present in
191 the first megabase of the autosomes, while 36 (3.8%) are located in the last 1 Mb of each
192 autosome. Relative to random permutations, we find a 3.19-fold increase in duplication content
193 in the first megabase of autosomes (mean: 1.9%; max: 3.5%), and a 2.1-fold increase in the last
194 megabase (mean: 1.8%; max: 3.2%). The median copy number of all the duplications was 3.62,
195 and the mean was 4.85. The maximum copy number of a given duplication was 78.3, located at
196 chr27:33386819-33407975.

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

202 *Read-depth analysis and self-alignment show low levels of agreement*

203 We compared the position of duplications identified by genome self-alignment and read-
204 depth to search for concordance. To account for detection differences between the methods, we
205 only considered duplications that were located on assembled chromosomes, greater than 15 Kb
206 in size, have at least 95% identity, and were not located in the first megabase of an autosome. A
207 total of 1,122 duplications identified by genome assembly self-alignment and 918 duplications
208 identified by read-depth met these criteria **(Supplementary Table 1, 2, 5, 6)**.
209 We explored how many duplications in each set were found to be unique. By intersecting
210 the self-alignment set and the read-depth set, we find that 528/1,122 self-alignment duplications
211 (47.1%) and 504/918 read-depth duplications (54.9%) share any overlap. The union of these two
212 sets contains approximately 77.78 Mb in total, 26.6 (34.2%) of which is duplicated in both sets.
213 Self-alignment uniquely detects 29.88 Mb (38.41%) and read-depth uniquely detects 21.02 Mb
214 (27.02%) **(Figure 1B)**. Despite only 35% agreement between the two methods, both analyses
215 show similar trends at a chromosomal level, including an increase in duplication count on chr26
216 **(Supplementary Figure 3).**

217 *Duplication content varies across canine assemblies*

218 We expanded our self-alignment and read-depth analysis to eight additional recently
219 published canine assemblies **(Table 1, Figure 1A)** [33-40]. Most assemblies used some method
220 of PacBio sequencing, predominantly Sequel, but the China and Yella assemblies used Oxford
221 Nanopore (ONT) sequencing. We note that China and Wags are different Basenjis sequenced
222 using different technologies by the same study. While the authors of that paper state that China is
223 the better assembly, we exclude China from the read-depth analysis due to a lack of available
224 Illumina data [40]. Notably, among the genomes is mCanLor1.2, a Grey wolf genome that serves
225 as an outgroup to the other breed dog genomes [34].
226 Canine assemblies have between 1.8% (Tasha) and 5.6% (mCanLor1.2) of their genome
227 identified as duplicated by genome assembly self-alignment **(Table 1)**. Assemblies with a low
228 duplication content (below 2.5%) include Sandy, Tasha, and Zoey, whereas assemblies with a
229 high percentage of duplications (above 4%) include Mischka, mCanLor1.2, Nala, Wags, and
230 Yella. A similar divide in duplication content among assemblies remains after filtering for
231 duplications above 15 Kb and increasing the sequence identity minimum to 95%. By analyzing
232 read-depth, we see a slight change, where mCanLor1.2, Nala, and Yella have the highest fraction
233 of their genome annotated as duplicated. The number and duplication status of unplaced contigs
234 differs widely among the assemblies **(Supplementary Table 1)**. The highest duplicated fraction

235 is found in Mischka (90%) and Zoey (70%), with the others having less than 50% of their
236 unplaced contigs duplicated.
237 Additionally, we explored 13,817 protein-coding genes for their duplication status as
238 determined by read-depth. Illumina data from each sample was aligned to the Mischka assembly;
239 China and Nala are removed from this analysis, due to lack of Illumina data and strong bias
240 based on GC composition respectively **(Supplementary Figure 4)**. The copy number of each
241 gene was then estimated using read-depth. We identify a total of 1,068 genes as duplicated in at
242 least one canine; 222 of those genes are considered duplicated in all seven canines. Further
243 analysis indicated that the most highly duplicated genes have no annotated function and are not
244 associated with a named gene **(Supplementary Table 7).** We performed a Gene Ontology (GO)
245 enrichment analysis of the 423 genes identified as duplicated based on read-depth from Mischka.
246 Enriched functional categories include 'intermediate filament organization' (GO:0045109),
247 'signal transduction' (GO:0007165), 'positive regulation of cellular process' (GO:0048522),
248 'multicellular organismal process' (GO:0032501), and 'negative regulation of cellular process'
249 (GO:0048523) **(Supplementary Table 9)** [44-46].
250 To explore the impact of PacBio HiFi reads on representing duplications, we compared
251 duplications identified by assembly self-alignment and read-depth in the mCanLor1.2 assembly.
252 Using the same filtering criteria as used in Mischka, we find that a total of 115.3 Mb is
253 duplicated in the union of self-alignment and read-depth, a 1.5-fold increase over Mischka. We
254 also find an increase in congruency between the two methods, as 60.07 Mb is duplicated in both
255 the self-alignment set and the read-depth set, making up 52.11% of the total base pairs, a notable
256 increase from the 34.2% overlap found in Mischka **(Figure 1C)**.

257 *Retrocopies proliferate throughout canine genome assemblies and show hallmarks of retrotransposition*

258 While exploring self-alignment duplications, we found that high-recurrence duplications
259 in the Mischka assembly often corresponded to the exons of transcripts found in other species,
260 suggesting the presence of gene retrocopies **(Figure 2A)**. To confirm this, we identified
261 candidate retrogenes in the Mischka assembly using BLAT. From an initial set of 19,972 multi-
262 exon protein-coding genes, we identified 3,892 retrocopies from 1,316 parental genes, which
263 includes 66 retrocopies that retained the full open reading frame of the parental gene **(Table 2,**
264 **Supplementary Table 4)**. 3,658 (94%) of these retrocopies are located on assembled
265 chromosomes. We found an enrichment of retrocopies on chrX, with a 1.18-fold increase over
266 expectation from 1,000 random permutations **(Supplementary Figure 5).** Additionally,
267 retrocopies that retained the full open reading frame of their parental gene had much greater
268 sequence identity with the parent gene (median: 99.92%) vs. those that did not (median: 96.14%)
269 (p-value: $3.81*10^{-48}$). (**Supplementary Figure 6).**
270 By intersecting high-recurrence duplications with annotated retrocopies, we determine
271 that 1,080/1,492 (72.39%) high-recurrence duplications segments overlap with a retrocopy; in
272 turn, 1,078/3,658 (29.47%) retrocopies on assembled chromosomes intersect with high-
273 recurrence duplications. On a base-pair level, retrocopies located in highly duplicated segments
274 make up a total of 4.35 Mb. Thus, multiple pair-wise alignments among retrocopies derived from

275 the same parent gene account for the highly duplicated regions identified by genome self-
276 alignment.
277 We additionally investigated how many retrocopies retained the signature hallmarks of
278 retrotransposition, which are the presence of target site duplications (TSDs) and poly(A) tails
279 [47] **(Figure 2B)**. Of the 3,496 retrocopies on assembled chromosomes in the Mischka genome
280 that passed filtering, 3,442 (98.46%) show at least one of the two hallmarks **(Figure 2C)**. In
281 total, 1,387 (39.67%) only have TSDs of at least 5 bp, 94 (2.69%) only have poly(A) tails, and
282 1,961 (56.09%) have both signatures. The mean length of poly(A) tails is 11.91 bases, and the
283 mean length of TSDs is 11.19 bases, with 1082 having TSDs 10 bp or longer **(Figures 2D and**
284 **2E)**. All 66 retrocopies that retain the parental ORF have TSDs, while 20 also have a poly(A)
285 tail.
286 Another known hallmark of retrotransposition is the presence of a consensus LINE-1
287 endonuclease recognition sequence, commonly read as TTTT/AA, where "/" marks the cleavage
288 site. We searched for the presence of the expected consensus sequence across retrocopies and
289 recovered the expected endonuclease recognition sequence **(Figure 2F)** [48-51]. We find that
290 approximately 40% of retrocopies annotated with a TSD 5bp or longer contain the expected cut
291 site sequence.
292 When analyzing only the retrocopy loci that are dimorphic among assemblies that passed
293 filters (n=212), thus permitting a more precise definition of event boundaries based on
294 comparison with the pre-integration empty site, we find a confirmed TSD of at least 5 bp or
295 greater, with an average TSD length of 15.1 bp. Additionally, loci that possessed a confirmed
296 TSD also possessed an endonuclease cleavage site **(Supplementary Figure 7).**
297 We expanded the retrocopy analysis to all other canine assemblies, using protein-coding
298 genes converted from Mischka coordinates as queries. Retrocopy counts are broadly similar
299 across assemblies (**Table 2, Supplementary Table 8**). Mischka's slightly greater number of
300 retrocopies appears to be accounted for by the presence of unplaced contigs, many of which
301 correspond to duplicated sequence **(Supplementary Table 1).** Similar trends of identified TSDs
302 and poly(A) tails are observed across the other assemblies, with over 98% of all identifies
303 retrocopies showing at least one hallmark **(Supplementary Table 3)**. We also specifically
304 analyzed retrocopies that retain the parental ORF for presence of hallmarks **(Supplementary**
305 **Table 4)**. Over 96% of full-ORF retrocopies show at least one hallmark, usually presence of a
306 target-site duplication.
307 The top ten retrocopy-producing parent genes were determined for each assembly
308 **(Figure 4)**. There were only 11 parent genes present across the top ten lists of all assemblies,
309 highlighting a high level of shared parental genes across the assemblies. Many of the top
310 annotated genes are previously known retrocopy producers in human and/or canine (*GAPDH,*
311 *HNRNPA1, EEF1A1, HMGN2* and the *RPL* genes) [52-56]. One gene, *BTF3*, was unique to the
312 Labrador Retriever genome, Yella, but retrocopies from this gene are present at slightly
313 decreased frequency in the other canine assemblies. Additionally, there were several instances
314 where we could not determine a unique progenitor among related parental genes. Most notable

315 among this group are LOC100686340 and LOC100687749, which are present in the top ten
316 retrocopy producer lists of all canine assemblies except Yella and Wags, the latter of which only
317 has LOC100686340 annotated.
318 To explore the function of common retrocopy progenitors, we performed a GO
319 enrichment analysis of the most common parent genes in the Mischka assembly. There is one
320 significant enriched GO category among the 50 parent genes in the Mischka assembly that had at
321 least 10 retrocopies, 'cytoplasmic translation' (GO:0002181), which likely derives from the high
322 number of ribosomal genes (RPL and RPS). There are 5 RPL genes in the top 10 parent genes,
323 and 15 RPL or RPS genes in the top 50. We expanded this analysis to the 117 genes that had at
324 least 5 retrocopies, which identified several enriched categories, including 'ribosomal small
325 subunit assembly' (GO:0000028) and 'cytoplasmic translation' (GO:0002181) **(Supplementary**
326 **Table 9)** [44-46].
327 Finally, we filtered retrocopies for any overlap with segmental duplications identified by
328 read-depth. We find that the Mischka assembly has a high prevalence of retrocopies present in
329 segmental duplications, with 466 compared to the range of 62-263 (avg: 126) found in other
330 assemblies. We note that 162/466 (34.7%) are present on unplaced contigs, further highlighting
331 Mischka's duplication bias towards unplaced contigs. No other canine assembly has any
332 retrocopies mapped to unplaced contigs, indicating that Mischka retrocopies are frequently
333 present in segmental duplications that get assembled onto unplaced contigs, in line with the
334 increased duplication rate present in the self-alignment data.

335 *Estimating the rate of retrocopy insertion*

336 To identify the rate of retrocopy insertion, we compared five canine assemblies from
337 diverse dogs (Mischka, Nala, China, Sandy, and Zoey) against mCanLor1.2. Only retrocopies
338 located on the autosomal chromosomes that did not overlap a read-depth segmental duplication
339 were considered. First, we assessed which retrocopies were shared across these six assemblies.
340 After applying strict filtering, we found a total of 3,377 unique retrocopies across the six
341 assemblies, 3,111 of which were located in all six canines (92.12%). 46 were unique to the
342 mCanLor1.2 assembly, and the largest categories tended to be singletons **(Supplementary**
343 **Figure 8, Supplementary Table 10)**. Mischka and Nala, both German Shepherds, have the
344 fewest singletons, reflecting their similarity relative to the other assemblies.
345 We also searched for the presence of retrocopies in other Canines. We searched for 3,658
346 retrocopies from Mischka in the fox (VulVul2.2) and dhole (GWHAAAC00000000) genome
347 assemblies [57, 58]. In total, 1923 retrocopies (52.2%) were found in the fox and/or the dhole
348 genomes, 468 retrocopies (12.79%) were present in both the fox and the dhole, 549 were unique
349 to fox (15%), and 906 were unique to the dhole (24.76%) **(Supplementary Table 11)**.
350 Consistent with their more recent origin, none of the 66 Mischka retrocopies with intact ORFs
351 were found in either the fox or dhole assemblies. We also find that sequence similarity of
352 retrocopies unique to Mischka (median: 96.75%) vs. those found in either the fox or the dhole
353 (median: 95.72%) tended to be slightly higher (p-value: $6.02*10^{-44}$) **(Supplementary Figure 6).**

354 To estimate the rate of retrocopy insertions, we first focus on the Mischka genome. We
355 estimated the number of generations since genomic divergence from the mCanLor1.2 wolf
356 genome. Excluding duplicated segments and short tandem repeats, there are 4,219,105 autosomal
357 single nucleotide differences between the Mischka and mCanLor1.2 genome assemblies. Using
358 the wolf mutation rate of $4.5 * 10^{-9}$ per generation [59], and with an analyzed genome size of
359 2,022,154,146 bp, we calculated that 231,827 generations have elapsed since genome divergence
360 between Mischka and mCanLor1.2. By comparing the retrocopies between Mischka and
361 mCanLor1.2, we identified 48 retrocopies that are unique to Mischka, and 79 that are unique to
362 mCanlor1.2. Retrocopies that were present uniquely to Mischka had a much greater sequence
363 similarity to their parent gene (mean: 99.92%) vs. those shared in the mCanLor1.2 assembly
364 (mean: 96.14%) (p-value: $9.396*10^{-29}$) **(Supplementary Figure 6)**. By dividing the average
365 amount of unique retrocopies over the number of generations, we estimate that the rate of
366 retrocopy insertion as $2.74 * 10^{-4}$ per generation, or approximately 1 in 3,650 births, a higher rate
367 than expected. Insertion rates estimated by comparing multiple canine assemblies, including
368 Mischka, to the mCanLor1.2 wolf genome range from $2.74 * 10^{-4}$ per generation to $3.1 * 10^{-4}$ per
369 generation, or from 1 in 3226 to 1 in 3650 (avg: 1/3514 births) **(Table 3)**. When considering a
370 calibration based on the lower-bound and higher-bound wolf SNV mutation rate, which was
371 defined as a range of 2.6×10^{-9} to 7.1×10^{-9} [59], we estimate the range of novel insertion in
372 Mischka is $1.58 * 10^{-4}$ per generation to $4.32 * 10^{-4}$ per generation (1/2,315-1/6,329 births).

373 ## Discussion

374 We analyzed the duplication content of nine recent canine assemblies using two methods:
375 genome assembly self-alignment and read-depth analysis, identifying 1.74-7.05% of each
376 genome as duplicated. We found that duplications were enriched in the first and last megabases
377 of canine autosomes. Since canine autosomes are acrocentric [54], this suggests an enrichment of
378 duplications in pericentromeric and subtelomeric regions, as found in other species [60].
379 The assembly self-alignment and read-depth approaches identified a range of duplication
380 content in each assembly, ranging from 17.27 to 109.78 Mb by self-alignment and 41.17-166.47
381 Mb by read-depth. Overlap analysis shows a generally poor level of agreement, with most
382 duplicated base pairs only identified by one of the methods. Thus, each of the assemblies offers
383 an incomplete representation of the true duplication content in each sample. The assemblies
384 differ in how they represent partially resolved duplications. For example, in the Mischka
385 assembly, unresolved duplications were often retained as separate unplaced contigs that were not
386 assigned to a chromosomal location. Other assemblies report few or no duplications on unplaced
387 contigs. This difference may have implications in downstream analyses that utilize these
388 assemblies.
389 Our initial analysis of duplications in the Mischka assembly revealed the presence of
390 many short, high-recurrence duplications. Further investigation showed that these segments
391 corresponded to multiple related retrogene insertions. To explore this further, we annotated
392 1,263-1,316 retrocopies across all assemblies, and found that a large majority displayed the
393 hallmarks of retrotransposition such as poly(A) tails and target site duplications. This includes

394 49-66 retrogenes per assembly that appear to maintain intact parental ORFs. Whether these intact
395 retrogenes are expressed or have any functional impact requires further investigation. We also
396 note that our definition of retrogene boundaries is based on the existing annotation of genes in
397 canines. This annotation of untranslated regions (UTRs) in these gene models is often
398 incomplete, resulting in an underestimation of the size of retrogene insertions (**Figure 2B**), and
399 underestimating the count of retrocopies with target-site duplications and/or poly(A) tails. We
400 also find that 53% of retrocopies from Mischka are present in the dhole or fox assemblies,
401 implying an origin prior to wolf speciation. However, it should be noted that the dhole and red
402 fox assemblies are highly fragmented (n=29,680 and 82,424 contigs, respectively), limiting the
403 ability to interpret sharing among the two genomes [57, 58].
404 Additionally, in the Mischka assembly, we find that 466 retrocopies are located within
405 larger segmental duplications with over half of those retrocopies being present on unplaced
406 contigs. This suggests a pattern where a retrocopy inserts into a locus, which is then
407 subsequently duplicated, perhaps repeatedly, throughout the genome. A similar pattern of
408 insertion-duplication has been described for retrogenes that may have functional effects in
409 elephants and horses [61, 62].
410 We take advantage of retrocopy comparisons between dog and wolf assemblies to
411 estimate an average retrogene insertion rate. Methods for estimating mutation rates from patterns
412 of sequence differences can be divided into three broad categories [63]. First, rates can be
413 estimated by sequencing trios and counting the number of new mutations observed in children
414 that are absent in their parents. This estimate is the least affected by natural selection, as only
415 lethal mutations should be unobserved in the offspring. However, a large number of individuals
416 may need to be analyzed, particularly for mutation types that arise at a slow rate. This method is
417 also extremely sensitive to false positive calls (in the children) and false negative calls (in the
418 parents). Second, rates can be estimated from the distribution of variants identified in multiple
419 individuals using population genetic models. The expected distribution of differences among
420 individuals is a function of multiple population genetic parameters. If some parameters are
421 known, an estimate of the mutation rate can be obtained. For example, if the effective population
422 size is known, or can be estimated using other data, then the mutation rate can be estimated
423 based on the distribution of segregating sites in the sample. Similarly, mutation rates can be
424 estimated from differences between chromosome segments that are inherited identical-by-
425 descent (IBD) if the number of generations separating the IBD segments can be determined.
426 Related methods have been applied over both recent and ancient time scales [64-67]. However,
427 these methods may require models of how population sizes change over time, are sensitive to
428 missing rare variants, and can be affected by natural selection. The third class of estimates takes
429 a phylogenetic approach and estimates the mutation rate based on the number of differences that
430 have occurred along two sampled lineages. The divergence time between sampled lineages can
431 be estimated from species separation times inferred from the fossil record. However, for closely
432 related groups, much of the lineage divergence time occurs as coalescence in the ancestral
433 population, making estimates of the ancestral population size critical when a fossil calibration is

434 used. As in population-based estimates, an alternative is to calibrate phylogenetic estimates
435 based on a sequence divergence time estimated using a previously known rate for a different type
436 of mutation.
437 In this study, we utilize phylogenetic comparisons between dog and wolf genomes to
438 estimate the rate of retrocopy formation. We calibrate these estimates using a single nucleotide
439 mutation rate inferred from sequencing a wolf pedigree. This mutation rate, $4.5*10^{-9}$/bp/gen,
440 implies a divergence between assembled dog genomes and the mCanLor1.2 Greenland wolf of
441 231,827 generations, or 695,481 years, assuming 3 years per generation [68]. This deep
442 divergence may seem inconsistent with the domestication of dogs in the past 15,000-30,000
443 thousand years [68-70]. Dogs are thought to have been domesticated from a now-extinct
444 progenitor population of wolves somewhere in Asia and the ancestral wolf population may have
445 had a large effective size. When correcting for differences in the assumed mutation rate, a
446 previous Bayesian coalescent analysis of dog and wolf divergence estimated that the ancestral
447 population size of new-world and old-world wolves was quite large, 143,000 individuals [71].
448 This implies a mean coalescence of two lineages in this ancestral population of 286,000
449 generations, which, along with the population separation time, is broadly consistent with our
450 phylogenetically inferred estimate. We note that these estimates of population size may be biased
451 by the presence of admixture from coyotes in some of new-world wolf samples used to create the
452 estimate [71, 72] or by unmodeled population structure among ancestral wolves.
453 We note that the our initial annotation of retrocopies in each assembly (Table 2) applies
454 conservative filters of alignment quality to reduce false positive annotations. However, this
455 induces false negative calls across assemblies. For the rate estimate, we corrected for this by
456 assessing the presence of retrocopies called in any assembly among all other assemblies through
457 an alignment of the flanking sequences (Supplementary Figure 8). Although our estimate of
458 1/3,514 births is affected by several assumptions including the absence of selection and of a
459 uniform rate, the estimate of the SNP mutation rate is likely the largest single source of
460 uncertainty in our estimate. Utilizing the reported confidence interval of the SNP mutation rate
461 ($2.6*10^{-9}$ – $7.1*10^{-9}$) implies a retrogene insertion rate of between 1/2,226 and 1/6,075 births.
462 The inferred retrocopy insertion rate in canines is substantially higher than that found in
463 humans; between humans and chimps, one new gene retrocopy insertion was estimated for every
464 6000 births, based on the 1000 Genomes Project [73]. However, this estimate is calibrated based
465 on an assumed human effective population size of 10,000 which is largely consistent with an
466 assumed human mutation rate of ~$2.5*10^{-8}$ bp/gen [74]. Adjusting this for the human mutation
467 rate obtained from pedigree sequencing, ~$1.3\ 10^{-8}$ bp/gen [75], decreases the estimated rate of
468 retrocopy insertion in humans to ~1/11,500 births, or about 3-times lower than the canine
469 estimate. However, it is noted that long-read data enables a higher retrocopy insertion detection
470 rate [76] so this may be an underestimate of the true human insertion rate. Other observations are
471 consistent with a high rate of retrogene formation in dogs. A recent survey suggests that
472 polymorphic retrogene insertions are 13-fold more common in dogs than in humans [13]. The
473 most well-studied retrogenes in dogs are derived from the *FGF4* gene which cause the short-leg

474 phenotype found in some breeds such as corgis and beagles [12]. *FGF4* retrogenes have occurred
475 multiple times independently, with 7 independent copies identified to date [13, 77]. Additionally,
476 dogs have a high rate of SINE polymorphism [37, 78]. Both SINE and retrogene insertions
477 require the activity of LINE-1 elements. In contrast to SINEs, which do not encode any proteins,
478 LINE-1s are autonomous retroelements that encode proteins required for their own
479 retrotransposition [79]. LINE-1 encoded proteins predominantly mobilize the RNA that encodes
480 them through a process known as *cis-preference* [80]. However, one or more LINE-1 encoded
481 proteins can rarely act to mobilize other RNAs such as SINEs or cellular mRNAs [81]. We
482 speculate that the high rate of SINE and retrogene insertion found in dogs may result from a
483 shared mechanism that involves a reduced level of *cis-preference* for the proteins encoded by the
484 canine LINE-1 element.

485 Materials and Methods

486 *Assemblies considered*

487 Our analysis involves 9 published long-read canine assemblies [33-40]. For simplicity,
488 we refer to them by name. Assembly accessions and other details are provided in **Table 1**.

489 *Genome assembly self-alignment*

490 Genome assembly self-alignment was performed using BISER (Brisk Inference of
491 Segmental duplication Evolutionary stRucture) (v1.2.3) which was developed to detect
492 segmental duplications at low identity across multiple genome assemblies [16]. BISER defines a
493 segmental duplication with the following criteria: below a certain error threshold based on shared
494 sequence identity ε(approximately 15% for 75% shared sequence identity), longer than 1,000 bp,
495 and paralog sequences may overlap up to a certain amount of base pairs, also defined by the error
496 threshold. Before alignment, repetitive regions in the genome are soft-masked, prohibiting the
497 initialization of alignments in low-complexity and highly-repetitive sequences, such as LINEs,
498 SINEs, and simple repeats. Duplications involving the mitochondrial chromosome were removed
499 from analysis since nuclear mitochondrial insertions (nuMTS) represent a distinct class of
500 variation [82], and two categories of duplications were defined: one where the paralogous
501 sequence between duplication copies is at least 90% identical, and a subset where the identity
502 rate is at least 95%. These sets are further divided based on their placement on assembled
503 chromosomes (the 38 canine autosomes and the X-chromosome) vs. unplaced contigs in each
504 assembly. Because BISER identifies duplicated segments as pairs of sequences and duplication
505 pairs often overlap each other, we used BEDTools merge (v2.30) to create a non-redundant set of
506 duplicated intervals [83].

507 *Distribution of duplications along the genome*

508 We analyzed the distribution of duplications located on the 38 canine autosomes and the
509 X chromosome by permutation. The distribution of duplications was compared against 1000
510 random permutations using BEDTools shuffle (v2.30), limiting placement to the assembled

511 chromosome sequence [83]. This was performed three times, on the self-alignment duplications,
512 the read-depth duplications, and on the retrogene set to look for chromosomal enrichment.

513 *Locating highly duplicated segments*

514 Short segments with a high recurrent duplication level were identified using the
515 duplication pairs found by self-alignment located on the assembled chromosomes. All
516 duplications smaller than 2.5 Kb were located, and their coordinates rounded to the nearest
517 hundred (i.e., a coordinate of 2,349 would become 2,300, and a coordinate of 3,175 would
518 become 3,200), to account for minor difference in alignment endpoints. Duplications were
519 merged into segments, and segments that contained at least four duplications were labeled as
520 highly recurrent.

521 *Read-depth analysis*

522 Read-depth analysis was applied to Illumina data from the assembled individuals using
523 fastCN [31]. First, a version of the assembly was created in which elements identified by
524 RepeatMasker, tandem repeat finder [84], WindowMasker [85], and over-represented 50-mers
525 were masked. Next, all placements within an edit distance of two for 36-bp segments of each
526 Illumina read determined using mrsFAST (v3.4.1) [86]. The resulting depth profiles were
527 tabulated, corrected for local GC content per fastCN, and converted to copy number estimates
528 based on a set of control regions. For copy number normalization we excluded: reported
529 deletions, duplications, and insertions [33], duplicated intervals from a preliminary self-
530 alignment of the Mischka genome, duplicated segments identified by a preliminary fastCN
531 analysis, and all chromosomes other than chr1-38. Copy number profiles were calculated based
532 on the mean normalized depth tabulated in windows containing 1,000 unmasked positions.
533 Duplicated regions are defined as at least four consecutive windows whose copy numbers all
534 exceed 2.5 and have a total combined size equal to or greater than 10 Kb. An overall copy
535 number was determined by taking the median copy number of all windows within a duplicated
536 region. Masking of the additional genomes was performed as described above. Normalization
537 controls were converted from Mischka coordinates using minimap2 (v2.26) [87]. China was
538 omitted from these analyses because, while it does have BGI short read sequencing, we opted to
539 only utilize Illumina-based data. Nala was omitted due to extreme GC bias observed in the
540 associated Illumina data **(Supplementary Figure 4)**.

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

542 To compare the results found by self-alignment and the results found by read-depth, both
543 sets of duplications (BISER and read-depth) were filtered to retain only a) duplications on the
544 assembled chromosomes; b) duplications greater than 15 Kb; c) duplications with at least 95%
545 identity; d) duplications that are not located in the first megabase of any autosome. Analysis was
546 limited to assembled chromosomes due to differences in the treatment of unplaced contigs
547 between assemblies. The read-depth analysis has a mean window size of 3,822 bp, with four
548 windows of an average size equaling approximately 15 kb. We utilized this metric to filter our

549 read-depth and our self-alignment sets to match this new criterion. The first megabase of each
550 autosome is excluded because, as a result of masking, there are no informative read-depth
551 windows in these regions. Self-alignment duplications were intersected with 1 Kb unique
552 sequence read-depth windows, and a self-alignment duplication was confirmed supported if the
553 windows had a median copy number of at least 2.5. Read-depth duplications were counted as
554 supported if they overlapped with any self-alignment duplication. The methods described above
555 for the Mischka assembly were expanded to the eight other canine genome assemblies with
556 control region coordinates using liftoff (v1.6.3) [87, 88]. In the gene analysis, read-depth
557 windows are intersected with the protein-coding genes annotated in Mischka.
558 For the gene set of choice, we used a subset of 13,817 genes identified in Mischka, with
559 the restriction that genes must be protein-coding, must be located on the 38 autosomes or the X
560 chromosome, and that a gene must fully encompass at least 3 windows, where each window
561 represents 1 Kb of non-repetitive DNA sequence used by fastCN. A table was built to observe
562 copy number across the genome, where a row represents a canine sample and a column
563 represented a pre-determined window in the Mischka coordinates, with entries being the copy
564 number. Copy number for genes were determined as the median of all copy numbers in windows
565 encompassed by a gene. A canine assembly is considered to have a duplicated gene if the copy
566 number exceeds 2.5.

567 *Retrogene analysis*

568 We searched for candidate retrocopies in the Mischka assembly based on the annotation
569 of 19,972 protein coding genes used by the Dog10K consortium [14]. A single transcript was
570 selected for each gene, with priority given to the longest isoforms and fewest number of exons.
571 We removed any genes with an intron less than 50 bp, to account for many uninvestigated genes
572 with single base pair introns, and any gene that only had a single exon.
573 We constructed cDNA sequences using SAMtools faidx (v1.13) to extract the sequence
574 of each exon [89]. Using BLAT, cDNA sequences were mapped against the genome with
575 repetitive 11mers filtered (--*ooc* option) (v.35) [90]. The resulting alignments were examined for
576 hits that could be retrocopies. Alignments were removed if they were: less than 100 bp,
577 overlapped the genomic coordinates of the parent gene, had less than 90% sequence identity with
578 the parent gene, or do not cross an exon-exon boundary of the parent gene. The same analysis
579 procedure was applied to the other eight genome assemblies after converting the genes annotated
580 in Mischka to the coordinates of each assembly using liftoff based on minimap2 alignments
581 (v2.26) [88, 91]. 1000 random permutations were conducted in the same manner described in the
582 section *Distribution of duplications along the genome.* Due to paralogous genes, parent genes
583 could not be resolved for a few retrocopies in each canine assembly.

584 *Conducting Gene Ontology (GO) Analysis*

585 At several instances, large lists of genes were submitted to the Gene Ontology (GO)
586 database in order to find functional links [44-46]. These were submitted to the GO Enrichment
587 Analysis (Powered by Panther) under the 'Biological Process' tab to *Canis lupus familiaris*. A

588 list of the genes submitted and their results can be found in **Supplementary Table 9**. We utilized
589 the 2024-01-17 release (10.5281/zenodo.10536401).

590 *Examination of retrogene hallmarks*

591 Retrogene containing loci from all nine canines were explored for presence of retrogene
592 hallmarks: target-site duplications (TSDs), poly(A) tails, and LINE-1 endonuclease cleavage
593 sites. Before detection proceeded, loci within one hundred base pairs of reference gaps were
594 removed with BEDTools (v2.26.0) window with parameters *-w 100 and -u* [83]. Additionally,
595 any loci within the first or last 100 base pairs of each chromosome as well as loci which were
596 comprised of multiple segments did not undergo hallmark detection. For each locus, a
597 symmetrical number of base pairs were extracted upstream and downstream of the retrogene
598 with SAMtools faidx (v1.13) [89].
599 To detect TSDs, the upstream and downstream fasta files were aligned to each other
600 using the water function within the EMBOSS module (v6.6.0) [92]. Water is a modified version
601 of a Smith-Waterman local alignment and reports the highest scoring alignment only [93]. The
602 scoring matrix used was a custom matrix which used +2 for all matches and -6 for all
603 mismatches (-1000 for any alignment containing "N"). Gap opening and gap extension penalties
604 were both set to 10. Alignment output containing "N" as a base pair were discarded, as were
605 final alignments shorter than 5 base pairs in length. If alignments started within 5bp of the start
606 of the upstream query or 5bp of the end of the downstream query, the flanking distance was
607 increased by 5 and the alignment was repeated.
608 Poly(A) detection followed TSD detection. The original flanking distance worth of base
609 pairs were again extracted beyond the 3' end of the retrogene with SAMtools faidx (v1.13).
610 From there, strings of A or T, depending on retrogene orientation, were detected within the
611 region. Poly(A)s were not extended into previously detected TSDs. If multiple putative poly(A)
612 tails were detected, the poly(A) closest to the TSD was reported if one was present. If no TSD
613 was detected, the Poly(A) closest to the retrogene was reported.
614 To assess if insertions possess a LINE-1 endonuclease cleavage site, loci which possess a
615 detected TSD as determined above had the first five base pairs of TSD sequence extracted along
616 with the two base pairs before the TSD using SAMtools faidx (v1.13). If the retrocopy is in the
617 same orientation as the chromosome, the sequence is reverse complimented so that all cleavage
618 site sequences are reported on the minus strand. Cleavage sites were then aggregated on a
619 sample-by-sample basis and plotted using matplotlib [94] and logomaker (v0.8) [95] .
620 Various extract distances were trialed ranging from 10 base pairs to 2000 base pairs with
621 60 base pairs being chosen to maximize concordance between TSDs and poly(A) (concordance
622 defined as the number of loci in which the TSD and poly(A) are within 5 bases of each other).
623 Seven of the nine samples had a maximum concordance with flanks of 50bp, 60bp, 70bp, 80bp,
624 or 90bp, while the other two had a maximum concordance with flanks of 200bp. Though this
625 method is limited in its ability to detect TSDs/poly(A) combinations that are more than 60bp
626 combined, as well as hallmarks present in retrocopies containing unannotated UTR sequence,

627 these limitations are thought to be minor as the vast majority of retrogenes contained at least one
628 hallmark.
629 To assess fullORF loci, the list of retrocopies with full open reading frames was
630 intersected with the loci which had their hallmarks identified (with 60bp of flanking distance) in
631 the previous step using bedtools intersect (v2.26).
632 Polymorphic loci were investigated for an alternate pattern of TSDs or endonuclease
633 cleavage sites. For each locus, an additional 5kb of flanking sequence of a retrocopy (or lack
634 thereof, referred to as the "empty site") was extracted using samtools faidx (v1.13). At each
635 locus, canines with starting coordinates longer than 20,001bp were excluded from the analyses,
636 as well as loci with multiple parent gene candidates. The extracted sequences were aligned to one
637 another using age -both (v0.4) [97]. If an insertion was detected, the inserted sequence was used
638 as a query to blat against the cDNA of the retrocopy of interest (v35). Blat commands followed
639 as such: "blat -minIdentity=85 extracted_insertion_sequence cDNA_file blat_output_file". The
640 corresponding longest blat hit must be greater than 75% of the length of the insertion. Individual
641 canine comparisons underwent separate processing if the blat alignment was of insufficient
642 length, no insertion longer than 20bp was detected, or there was an insertion in both the query
643 and the reference. These hits were not investigated in downstream steps. For remaining
644 comparisons, AGE also detects locations of TSDs and is more stringent than our previous
645 method, with detected TSDs being exactly at the boundary of the detected insertion and requires
646 perfect sequence identity between upstream and downstream TSDs. For loci that contained a
647 TSD, an endonuclease cleavage site was determined by extracting the last two bases before the
648 TSD and the first 5 bp of the TSD using samtools faidx (v1.13). Cleavage sites were reverse
649 complimented if the orientation of the retrocopy was forward. Logo plots and histograms of this
650 dataset were generated using logomaker(v0.8) and matplotlib [95]. At each locus, only one
651 confirmed filled site was processed. Priority was given to canines with at least 5bp of TSD
652 sequence in the following order: mCanLor, then Mischka, Sandy, Nala, China, Zoey.

653 *Estimating the insertion rate of retrocopies*

654 We selected contiguous genome assemblies from diverse dog samples to search for the
655 insertion rate of retrocopies, including Mischka, Nala, China, Sandy, and Zoey relative to the
656 mCanLor1.2 assembly. We discounted all retrocopies that overlap a read-depth segmental
657 duplication. For each canine retrocopy set, pairs of 500 bp flanks were extracted from each
658 retrocopy and aligned to mCanLor1.2 using minimap2 (v2.26) with the -c and -x asm5 options
659 [87]. Retrocopies whose flanks did not map were discounted entirely from the analysis, due to an
660 inability to determine presence of sharing between samples. For all retrocopies, if a retrocopy's
661 presence across all six assemblies had not yet been determined in the initial mapping, both the
662 flanks and the intervening sequence were derived from an assembly that did have the sequence
663 (in order, priority was given to mCanLor, then Mischka, Sandy, Nala, China, and Zoey in that
664 order) and mapped to all dogs to determine whether or not the retrocopy was truly present. If the
665 flanks were present and the intervening sequence was also present, the retrocopy was marked as
666 shared. If only the flanks were present, the retrocopy was marked as absent in that dog, and if

flanks did not map, the entire retrocopy was discounted, once again due to inability to determine sharing. The remaining retrocopies were totaled for presence or absence in each assembly, and converted into an upset plot using UpSet (v.0.9.0) [96].

Then, a retrocopy insertion rate was estimated based on a comparison between each assembly and mCanLor1.2. For each independent comparison, a retrocopy was considered as long as it could be determined whether or not it was present in mCanLor1.2 and the compared assembly, regardless of whether it failed to map in another assembly. This allowed for determination of retrocopies unique to mCanLor1.2 vs. a comparison assembly.

To identify sequence identity of retrocopies to the parent gene, retrocopy sequences were compared against the parent gene cDNA using BLAT (v.35) [90]. Sequence identity was determined as number of matches over number of matches plus mismatches. Retrocopies were then divided into categories for comparison: one group with fullORF retrocopies vs. those that were not; one group with retrocopies that were also found in either the dhole or the fox vs. those that were not; and one group with retrocopies unique to the Mischka assembly vs. those that were also found in the mCanLor1.2. assembly.

To further estimate rates of differences between genome assemblies, we identified single nucleotide differences among assemblies. Autosomal assembly size was determined by called regions between an assembly and mCanLor, as determined by minimap2 (v2.20) [87], with the additional flags of *-c --cs*, outputting differences in *PAF* format using the paftools.js call command. Regions are filtered to remove duplicated segments identified by read-depth, duplications identified by self-alignment, and simple tandem repeats identified by Tandem Repeat Finder [84]. SNPs called by minimap2 were then intersected with filtered regions. 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 [59] and the size of the analyzed genome after filtering 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 also calculated a range for the generations based on the range of the wolf mutation rate, which was 2.6×10^{-9} and 7.1×10^{-9}. 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, and additionally calculated a range of rate of insertion based on the generation range [25, 97, 98].

Locating ancestral retrocopies

We selected 3,658 retrocopies from the Mischka assembly to search for in the red fox (*Vulpes vulpes,* VulVul2.2) and dhole (*Cuon alpinus,* GWHAAAC00000000) genome assemblies. It should be noted that both genome assemblies are composed of contigs, as opposed to chromosomes. We extracted 1 Kb flanks from all the retrocopies with samtools faidx (v1.13)

705 and used minimap2 (v2.26) with the -c and -x asm20 options to align these flanks to the fox or
706 dhole genome. We filtered out any retrocopies whose flank pairs either had at least one member
707 missing, or whose flank pairs mapped to different contigs. We then took the intervening
708 sequence between the flanks and removed any that were not at least 75% of the size of the
709 original retrocopy in Mischka. With the remaining set, we extracted the intervening sequence
710 with samtools faidx from their respective genomes and used minimap2 with the same options to
711 map the putative retrocopies back to the Mischka assembly. If a candidate retrocopy from the fox
712 or dhole did not overlap the original location in Mischka, it was discounted from the analysis.

713 Acknowledgements

714 We thank the numerous authors of the genome assemblies used in this analysis for publicly
715 providing their data for the investigation. In addition, we thank the editor and anonymous
716 reviewers for their helpful comments.

717 Author Contributions

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

722 Data Availability

723 Final counts and coordinates of each set of items (self-alignment duplications, read-depth
724 duplications, genes in read-depth duplications, and retrogenes with hallmarks) are located in the
725 provided supplementary files. Relevant and custom code is available on GitHub at
726 https://github.com/antnguye/dups_and_retrogenes_in_canine_assemblies. Hallmark analysis is
727 available on GitHub at
728 https://github.com/mblacksmith1996/Hallmark_Detection_Python_Scripts. A UCSC genome
729 browser TrackHub depicting duplications identified in the Mischka and mCanLor1.2 assemblies
730 is available at https://github.com/KiddLab/dog-segdup-tracks.

731 # 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

732

733 ***Table 1. Canine Assemblies Investigated.***

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

738

Sample Name	Protein-Coding Genes	All Retrocopies (Parental Genes)	Assembled Retrocopies (Parental Genes)	Full-length ORF Retrocopies (Parental Genes)
Mischka (canFam4)	19972	3892 (1316)	3658 (1305)	66 (55)
mCanLor1.2	19449	3649 (1307)	3634 (1304)	59 (52)
China	19393	3601 (1305)	3583 (1298)	49 (45)
Nala	19490	3695 (1315)	3693 (1315)	64 (50)
Sandy	19382	3649 (1311)	3646 (1310)	62 (53)
Tasha	19020	3474 (1263)	3472 (1263)	50 (42)
Wags	19394	3626 (1281)	3401 (1234)	51 (44)
Yella	19192	3568 (1281)	3568 (1281)	55 (48)
Zoey	19380	3632 (1297)	3572 (1286)	49 (41)

739

740 *Table 2. Retrocopies identified in each canine assembly.*

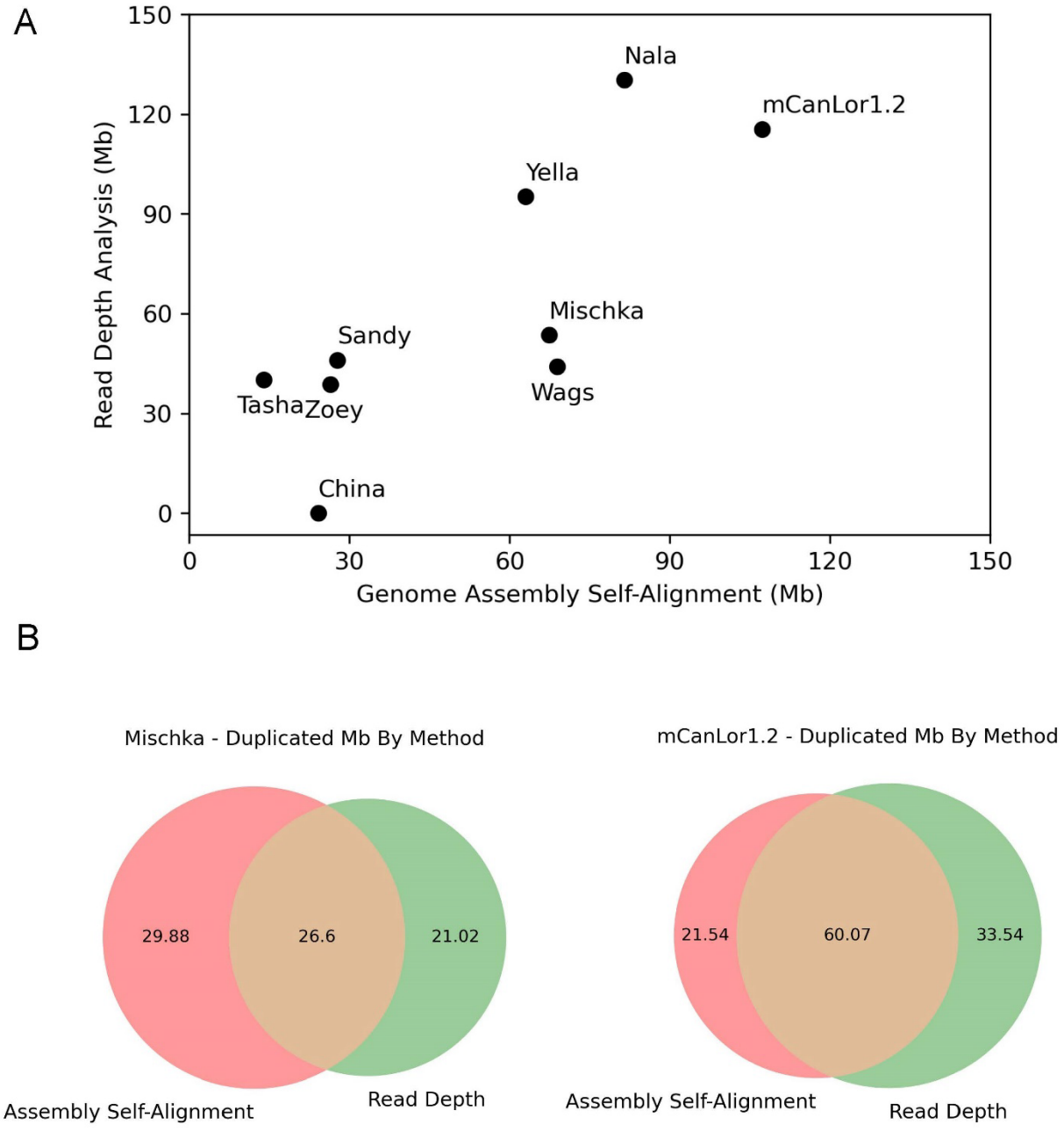
741 Column two gives the number of protein-coding genes considered as potential gene parents in
742 each assembly. Retrocopies in column 3 include all chromosomes, including mitochondrial and
743 unplaced contigs, and column 4 contains only the assembled chromosomes, chr1-38 + X. The
744 total number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

Assembly Name	Total Retrocopies Between Assemblies	Retrocopies Unique to Assembly	Retrocopies Unique to mCanLor1.2	Estimated Generations Since Divergence	Insertion Rate * 10^{-4} (1 in N births)	Range of Generations	Range of Insertion Rate * 10^{-4} (1 in N births)
Mischka	3,489	48	79	231,827	2.74 (3,650)	146,932-401,239	1.58-4.32 (2,315-6,329)
China	3,513	48	92	226,008	3.1 (3,226)	143,245-391,168	1.79-4.89 (2,045-5,587)
Nala	3,503	52	79	235,551	2.78 (3,597)	149,293-407,685	1.61-4.39 (2,278-6,211)
Sandy	3,510	54	79	234,571	2.83 (3,534)	148,672-405,989	1.64-4.47 (2,237-6,098)
Zoey	3,487	47	78	224,470	2.78 (3,597)	142,270-388,507	1.61-4.39 (2,278-6,211)

Table 3. Rate of Retrocopy Formation in Canine Assemblies

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. Columns 3 and 4 show how many retrocopies are unique to the assembly and to mCanLor1.2 respectively. 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^{-4}, and presented as 1 in n births. Column 7 shows the range of generations elapsed, based on the upper and lower bounds of the wolf mutation rate. Column 8 shows the range of insertion rate, based on the generation range in Column 7.

758 **Figures**

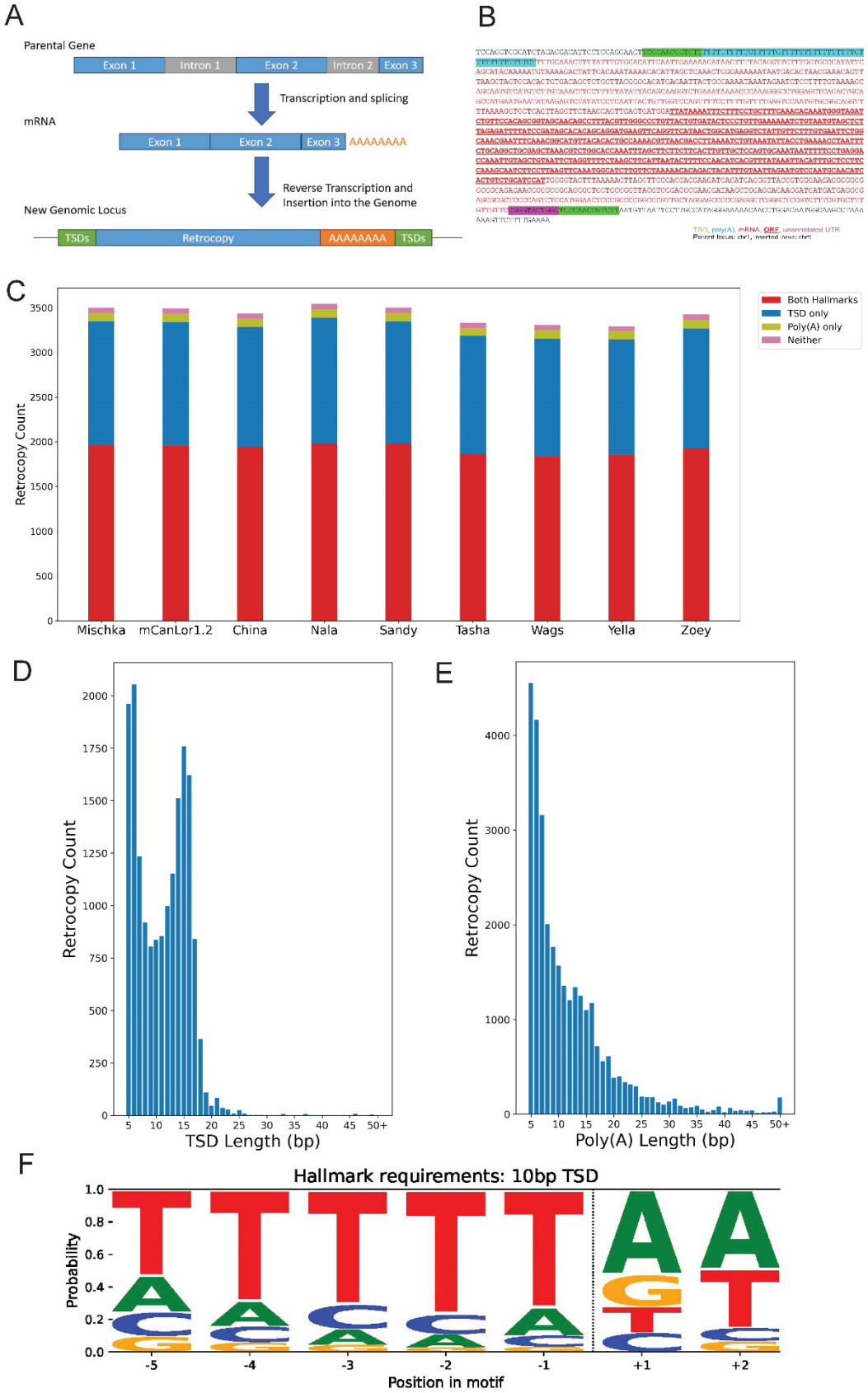


759

760 *Figure 1. Estimated duplication content of canine assemblies.*

761 Duplication content was determined based on read-depth and genome assembly self-alignment.
762 (A) The amount of each genome assembly detected as duplicated using each method is shown.
763 Since the read-depth approach is best able to detect larger, more similar duplications, the regions
764 detected by genome self-alignment were filtered to segments with 95% sequence identity at least
765 15 Kb in length. China lacks Illumina data and thus has no read-depth analysis. (B) A Venn

766 Diagram showing the intersection of the two methods for the Mischka assembly and for the
767 mCanLor1.2 assembly. Values are plotted in Mb.


A
Parental Gene
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
Transcription and splicing
mRNA
Exon 1
Exon 2
Exon 3
AAAAAAAA
Reverse Transcription and
Insertion into the Genome
New Genomic Locus
TSDs
Retrocopy
AAAAAAAA
TSDs
B
C
Retrocopy Count
Both Hallmarks
TSD only
Poly(A) only
Neither
Mischka
mCanLor1.2
China
Nala
Sandy
Tasha
Wags
Yella
Zoey
D
Retrocopy Count
TSD Length (bp)
E
Retrocopy Count
Poly(A) Length (bp)
F
Hallmark requirements: 10bp TSD
Probability
Position in motif

769 *Figure 2. Hallmarks of retrotransposition in detected retrocopies.*

770 Analysis of retrocopy presence across canine assemblies is shown. A) A diagram depicting the
771 key steps of retrocopy formation, including the presence of target site duplications and poly(A)
772 tails at the insertion, is shown. B) The annotated sequence of a retrocopy from the TBPL1 parent
773 gene, which is inserted in chr9 in the Mischka assembly, is shown. Highlighted sequence depicts
774 the aforementioned hallmarks as indicated. This retrocopy retains the parental ORF, shown in
775 bold. The short purple highlighted sequence corresponds to UTR sequence included in the
776 retrocopy that is not part of the existing TBPL1 annotation. C) The distribution of how many
777 retrocopies possess TSDs, poly(A) tails, both, or neither in each of the nine canine assemblies.
778 D) A histogram of the TSD lengths found for retrocopies in the Mischka assembly. TSD lengths
779 of 50 or greater have been grouped into one bin. E) A histogram of the lengths of poly(A) tails
780 found for retrocopies in the Mischka assembly. Lengths of 50 or greater have been grouped into
781 one bin. F) An examination of target site cleavage sequence for retrocopies in Mischka for
782 retrocopies with at least a 10bp TSD (N=1615). It shows the sequence composition spanning 5
783 bases before and 2 bases after the inferred cut site.

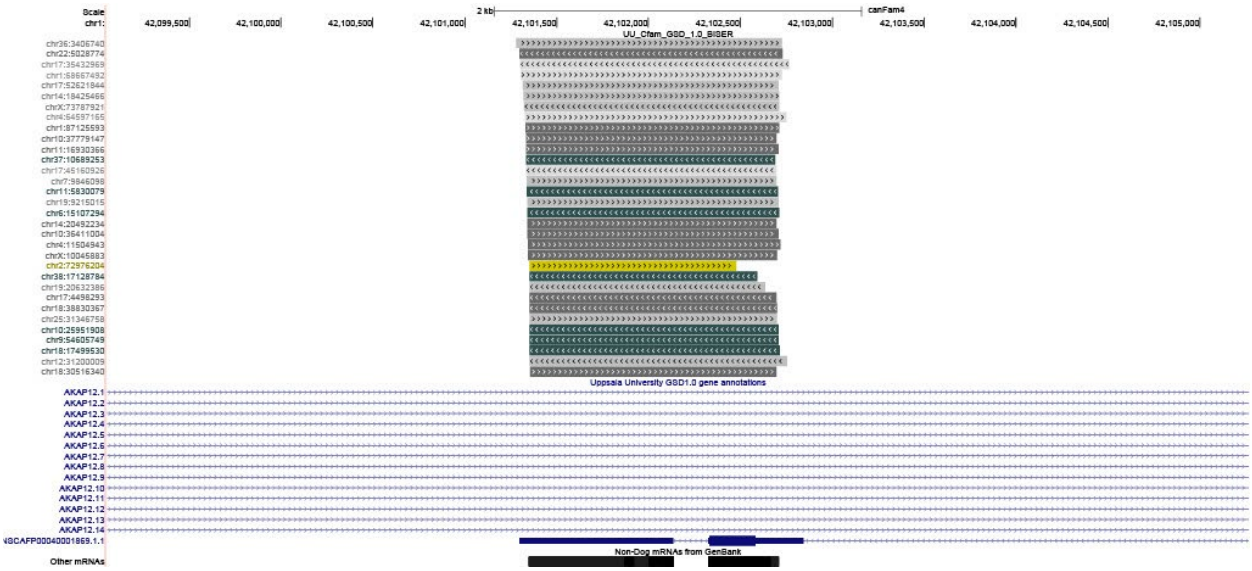
mCanLor1.2
EEF1A1 (169)
HNRNPA1 (128)
GAPDH (114)
RPL17 (59)
RPL31 (59)
RPL21 (53)
LOC119863874 (34)
RPL32_2 (33)
RPL7 (31)
LOC100687749/LOC100686340 (26)

Mischka	China	Nala	Sandy	Tasha	Wags	Yella	Zoey
EEF1A1 (169)	EEF1A1 (167)	EEF1A1 (173)	EEF1A1 (176)	EEF1A1 (167)	EEF1A1 (165)	EEF1A1 (168)	EEF1A1 (169)
HNRNPA1 (123)	HNRNPA1 (129)	HNRNPA1 (125)	HNRNPA1 (130)	HNRNPA1 (122)	HNRNPA1 (117)	HNRNPA1 (126)	HNRNPA1 (123)
GAPDH (113)	GAPDH (111)	GAPDH (112)	GAPDH (113)	GAPDH (109)	GAPDH (106)	GAPDH (114)	GAPDH (110)
RPL17 (64)	RPL17 (66)	RPL17 (67)	RPL17 (68)	RPL17 (63)	RPL17 (57)	RPL17 (67)	RPL17 (67)
RPL31 (59)	RPL31 (57)	RPL31 (58)	RPL31 (59)	RPL31 (56)	RPL31 (52)	RPL31 (61)	RPL31 (58)
RPL21 (53)	RPL21 (55)	RPL21 (54)	RPL21 (54)	RPL21 (52)	RPL21 (51)	RPL21 (53)	RPL21 (53)
LOC119863874 (35)	LOC119863874 (34)	RPL7 (35)	LOC119863874 (34)	RPL32_2 (34)	LOC119863874 (33)	RPL32_2 (34)	RPL32_2 (33)
RPL32_2 (33)	RPL32_2 (33)	LOC119863874 (34)	RPL32_2 (32)	LOC119863874 (33)	RPL32_2 (32)	LOC119863874 (34)	LOC119863874 (33)
RPL7 (31)	RPL7 (29)	RPL32_2 (33)	RPL7 (30)	RPL7 (30)	RPL7 (32)	RPL7 (30)	RPL7 (32)
LOC100687749/LOC100686340 (29)	LOC100687749/LOC100686340 (27)	LOC100687749/LOC100686340 (28)	LOC100687749/LOC100686340 (29)	LOC100687749/LOC100686340 (27)	LOC100686340 (27)	BTF3 (28)	LOC100687749/LOC100686340 (27)

Figure 3. Most frequent retrocopy parental genes.

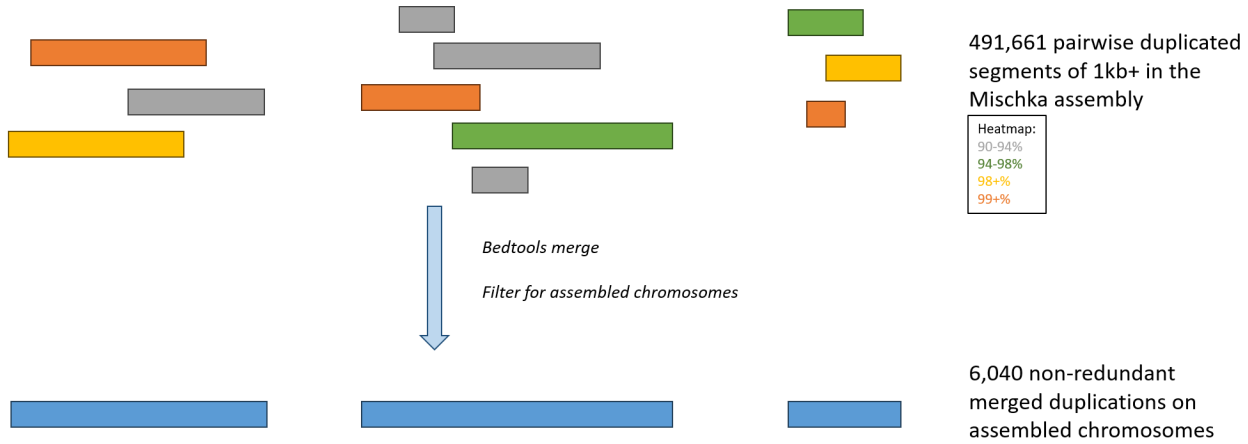
Each column lists the 10 genes that give rise to the most identified retrocopies in each assembly. Retrocopies presented here are limited to the assembled chromosomes. The number of retrocopies from each parent are given in parentheses. The same genes are colored identically to one another to quickly determine presence of reoccurring genes. Only 11 parent genes are represented across all nine assemblies, with retrocopies from BTF3 found in the other assemblies at slightly decreased frequencies that removed it from the top ten.

Supplementary Figures



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

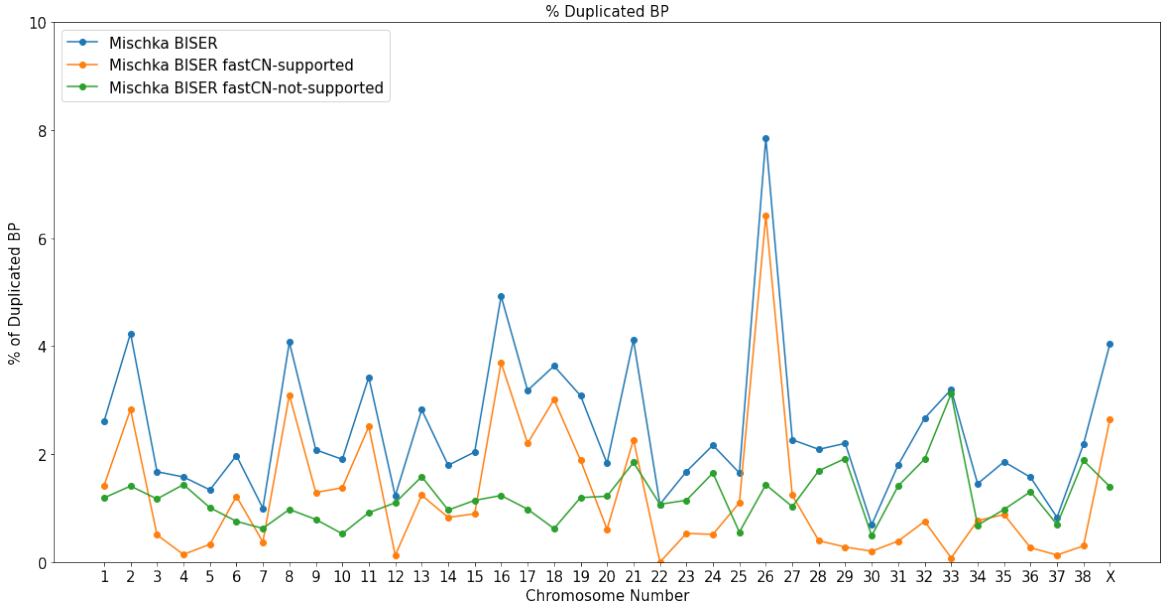
This screenshot of UCSC Genome Browser depicts the region chr1:42,099,050-42,105,269 in Mischka, a segment of the genome located in an intron of the gene *AKAP12*. The green and gray bars depict duplicated intervals detected by genome assembly self-alignment (BISER, top), whereas the blue lines with 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.



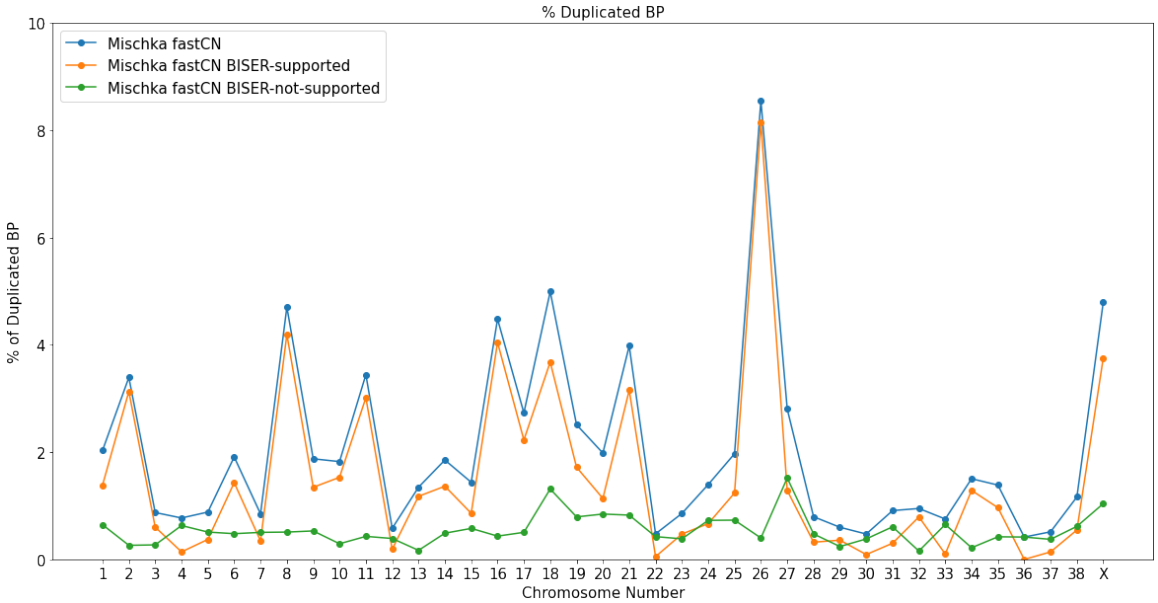
Supplementary Figure 2. Merging Pairwise Duplications Identified by Genome Self Alignment.

In the genome assembly self-alignment analysis, we identify pairwise duplicated segments 1kb in size with greater than 90% sequence similarity (represented in color). These duplicated segments appear across all chromosomes present in canine assemblies, including on unplaced contigs. Because these duplications are pairwise, a single locus can be duplicated multiple times and be present in multiple pairwise alignments with varying levels of sequence similarity (depicted in color). When merging, bedtools merge (v2.30) was used to combine overlapping segments together, regardless of sequence similarity We also removed duplications present on non-assembled chromosomes at this stage (chrM and unplaced contigs), since the treatment of unplaced contigs varies among the assemblies we studied.

A



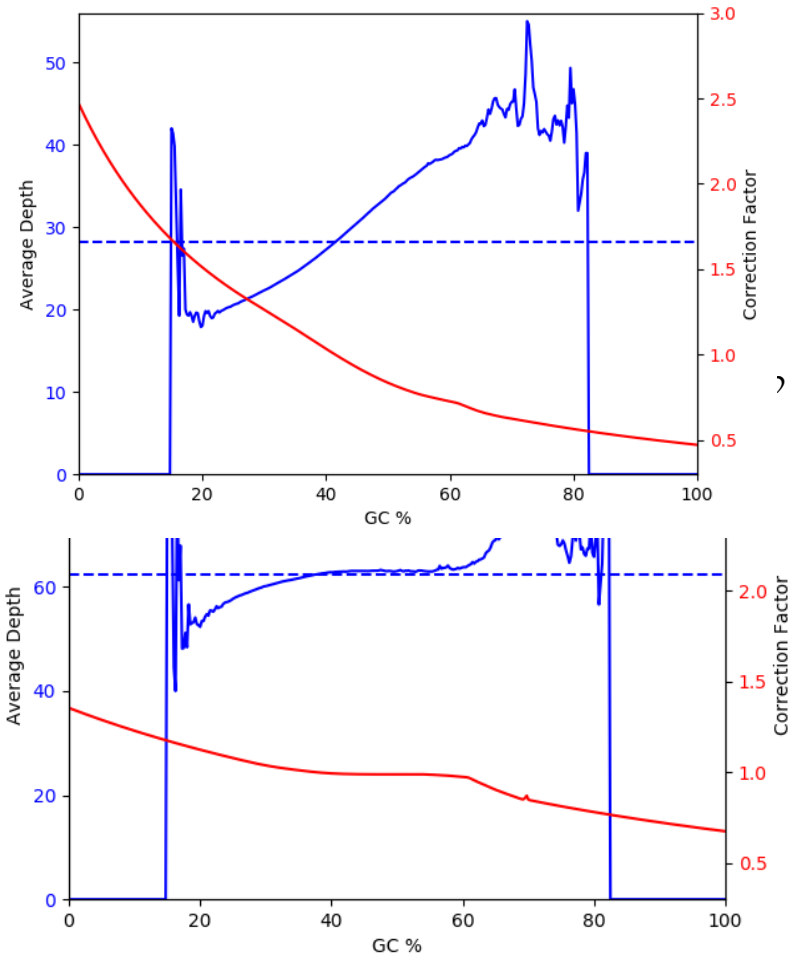
B



Supplementary Figure 3. 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 percentage. (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.

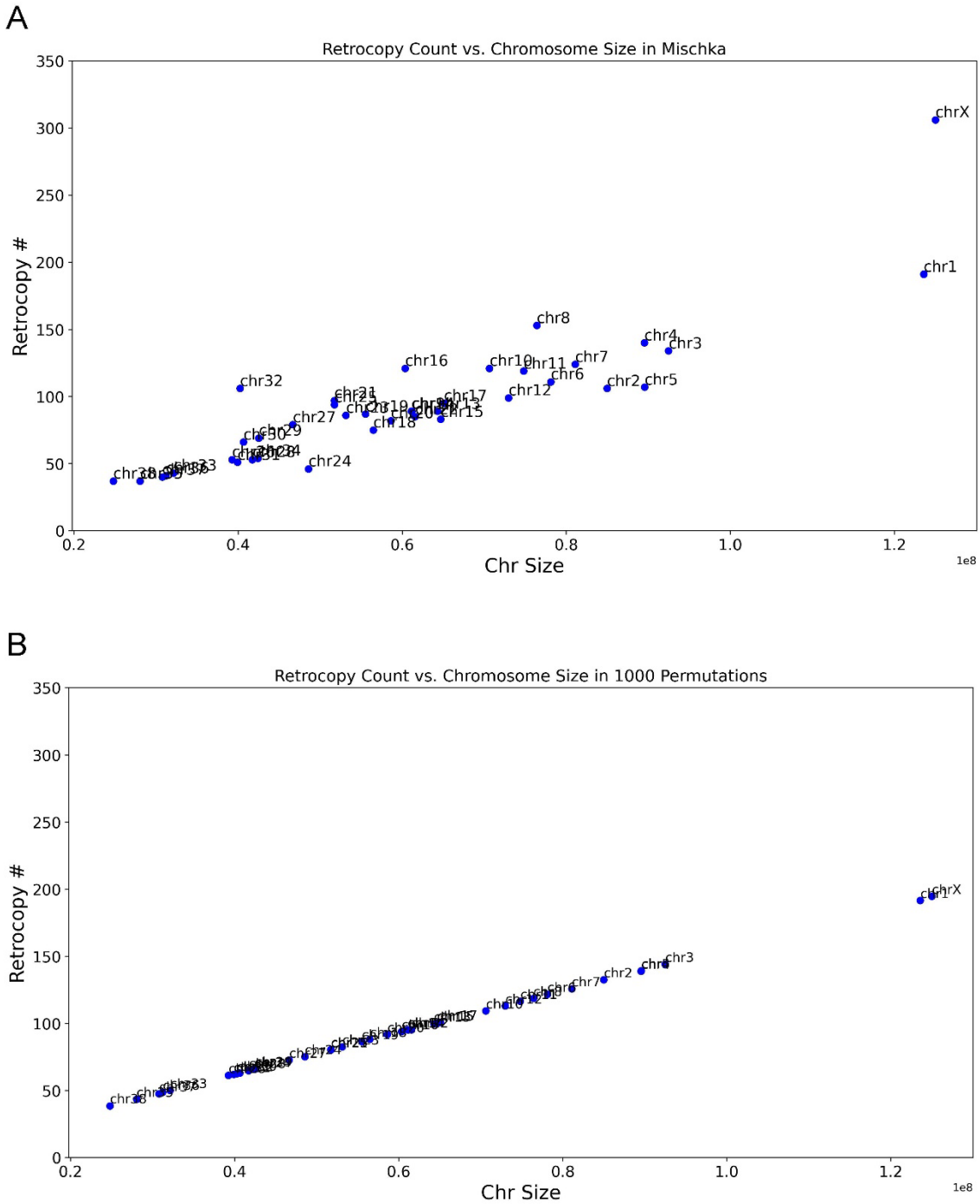
A)



B)

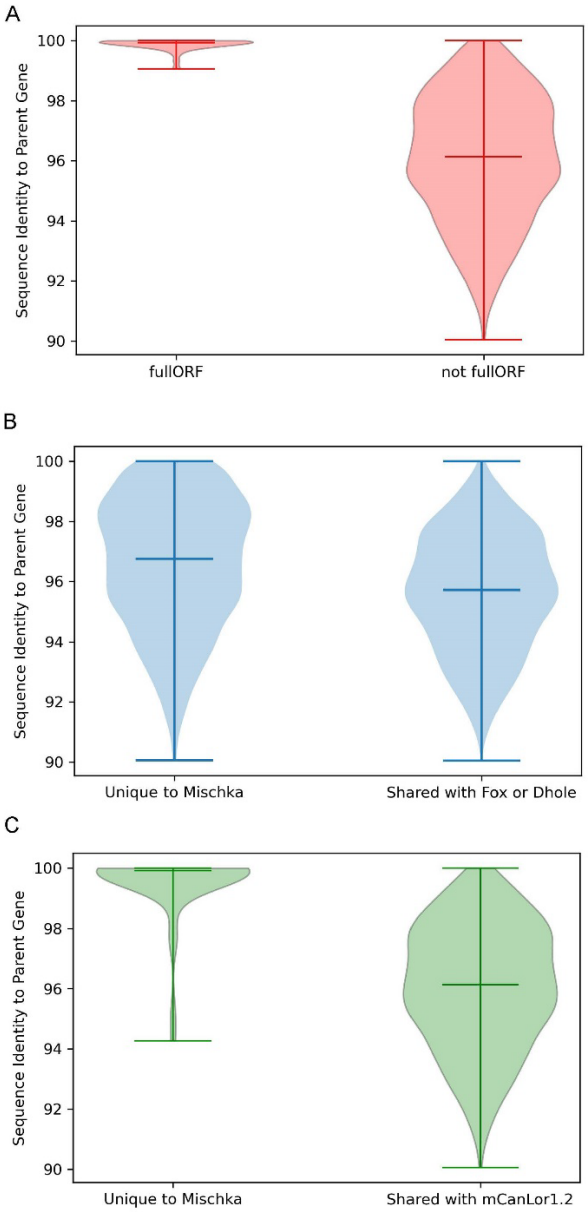
Supplementary Figure 4. GC Coverage Profiles.

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.



Supplementary Figure 5. Retrocopy Count vs. Chromosome Size.

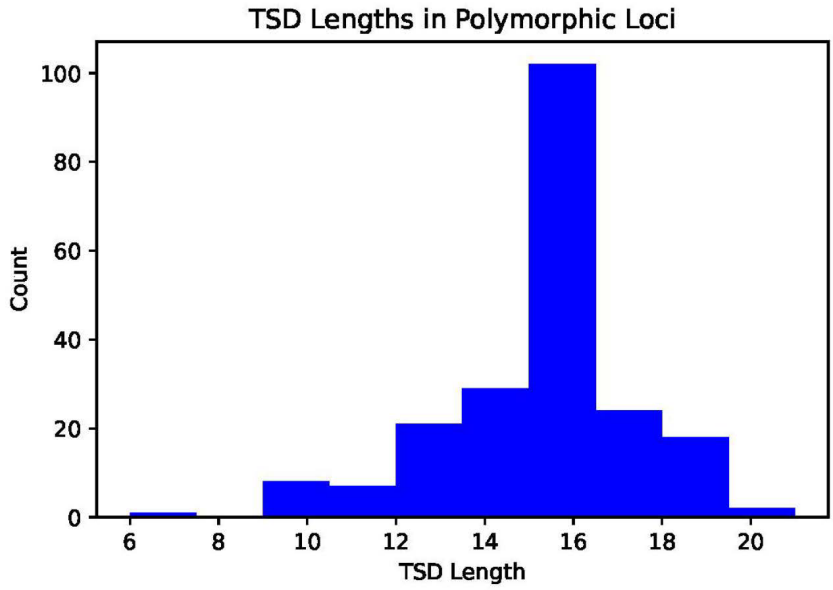
A) Depiction of the number of retrocopies on each of the assembled chromosomes (chr1-38 + X) found in the Mischka assembly vs. the length of each chromosome. ChrX has 228 total retrocopies. B) Depiction of the number of retrocopies on each of the assembled chromosomes found in 1000 random permutations vs. the length of each chromosome. ChrX has a mean of 146 retrocopies and a max of 188 retrocopies. Mischka contains a 1.18-fold increase of detected retrogenes on chrX.



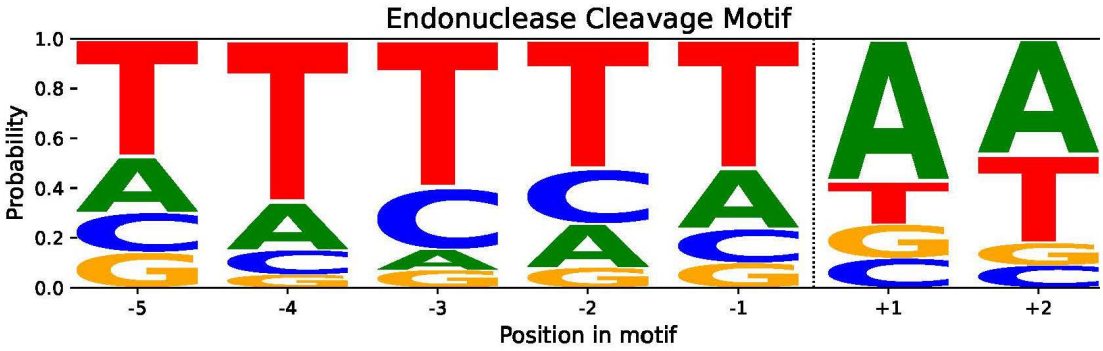
Supplementary Figure 6. Sequence Similarity Between Mischka Retrocopies and Parent Gene

This violin plot depicts three categories of comparison of sequence similarity between retrocopies found in the Mischka assembly and the parent gene. A) Sequence similarity of retrocopies between those that retain the parental open-reading frame (Mean: 99.86%; Median: 99.93%) vs. those that do not (Mean: 96.04%; Median: 96.14%). P-value: $3.81*10^{-48}$. B) Sequence similarity of retrocopies unique to Mischka (Mean: 96.57%; Median: 96.75%) vs. those found in either the dhole or the fox assemblies (Mean: 95.62%; Median: 95.72%) P-value: $6.01*10^{-44}$. C) Sequence similarity of retrocopies unique to Mischka (Mean: 99.54%; Median: 99.92%) vs. those found in the mCanLor1.2 assembly (Mean: 96.04%; Median: 96.14%). P-value: $9.396*10^{-29}$.

A

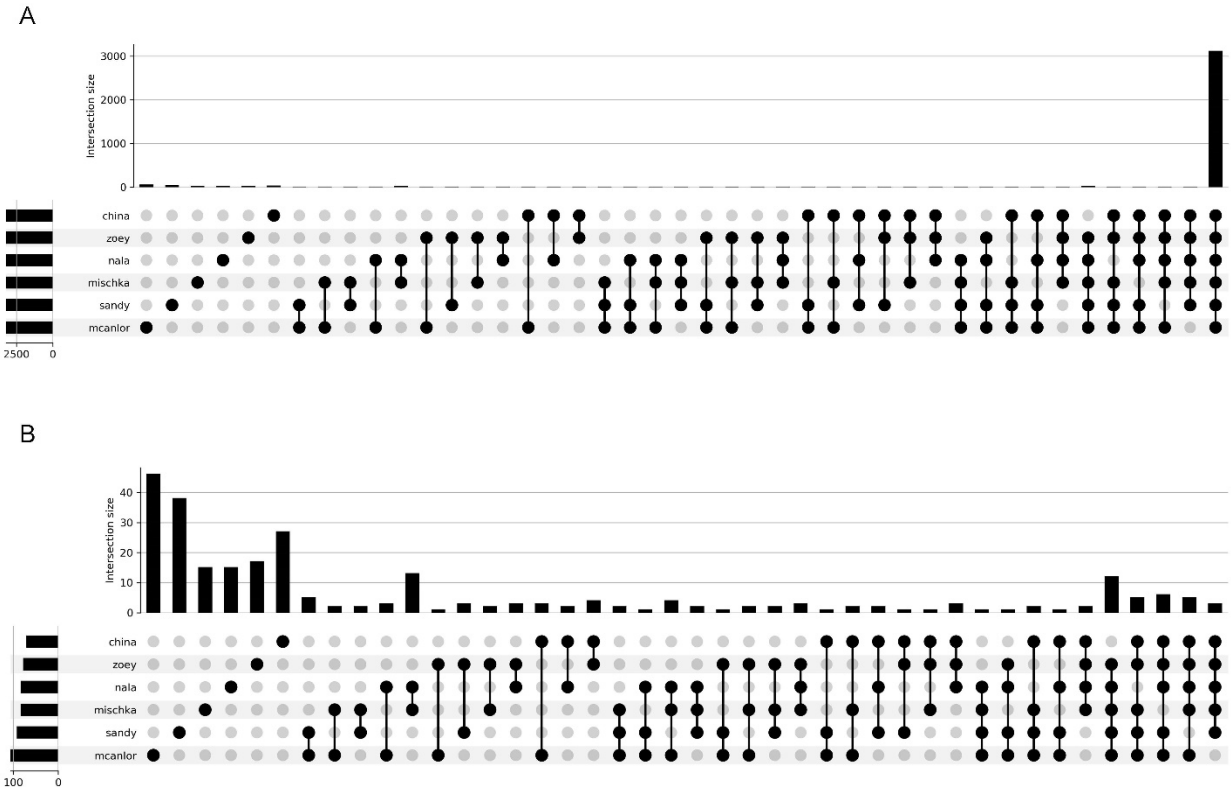


B



Supplementary Figure 7. Hallmark Analysis of Polymorphic Retrocopies.

An analysis of the polymorphic retrocopies (n=226 resolved loci). A) Detected TSD lengths for all polymorphic retrocopy containing loci with TSDs 5bp or longer (N=212). Only one canine is represented at each locus. Priority is given as to the first canine with an identified TSD as follows mCanLor, then Mischka, Sandy, Nala, China, and Zoey. The average TSD length across the dataset is 15.1 bp. B) The endonuclease cleavage motif for polymorphic loci containing 5bp TSDs, which follows the expected sequence of TTTTT/AA.



Supplementary Figure 8. Retrocopies Shared Across Canine Assemblies.

This upset plot depicts 3,377 retrocopies that were found to be present in at least one assembly between Mischka, mCanLor1.2, Sandy, China, Nala, and Zoey. A) depicts all retrocopies grouped by presence in canine assemblies, and B) removes the category for shared between all six assemblies. Of the retrocopies depicted, 3,111 are shared between all six assemblies. The second largest category is mCanLor1.2-specific retrocopies (n=56). The remaining categories have between 1 and 38 retrocopies.

907 **Supplementary Tables**

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%)

908
909 ***Supplementary Table 1. All Duplications found by Genome Assembly Self-Alignment.***

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

917

Name	Read-depth Duplications	Duplicated Mbp (% of genome)	Read-depth Duplications on Assembled Chromosomes	Mbp on Assembled Chromosomes (% of assembled genome)
Mischka	1573	87.1873 (3.5%)	1033	56.2099 (2.39%)
mCanLor1.2	2093	130.4306 (5.33%)	2069	123.14 (5.11%)
Nala	4709	152.5044 (6.34%)	4706	152.3764 (6.39%)
Sandy	1702	57.157 (2.43%)	1681	55.1293 (2.36%)
Tasha	990	42.5327 (1.84%)	990	42.5327 (1.84%)
Wags	1007	63.9808 (2.65%)	688	45.5681 (1.99%)
Yella	1060	98.3779 (4.11%)	1059	98.3612 (4.11%)
Zoey	929	48.9654 (2.09%)	695	40.4794 (1.74%)

918

919 ***Supplementary Table 2. All Duplications found by Read-depth.***

920 Duplications were defined as at least four consecutive windows found by fastCN to have a copy
921 number of at least 2.5, with a size cutoff of 10 Kb. Column 1 contains canine name, column 2
922 contains the total number of duplications detected by read-depth, column 3 contains the total
923 number of duplicated Mbp, and columns 4 and 5 mimics 2 and 3, but restricted to chr1-38 + X,
924 referred to as the “assembled chromosomes”.

Sample Name	Total Retrocopies on Assembled Chromosomes	Filtered Retrocopies	Any Hallmark	Both Hallmarks	TSD Only	Poly(A) Only	Neither
Mischka (canFam4)	3658	3496	3446	1086	2326	34	50
mCanLor1.2	3634	3493	3448	1006	2407	35	45
China	3583	3434	3381	982	2367	32	53
Nala	3693	3543	3480	1018	2430	32	63
Sandy	3646	3502	3451	1024	2393	34	51
Tasha	3472	3328	3278	1035	2216	27	50
Wags	3401	3307	3253	960	2260	33	54
Yella	3568	3291	3241	949	2264	28	50
Zoey	3572	3423	3371	963	2370	38	52

Supplementary Table 3. 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.

Sample Name	Total FullORF Retrocopies	Any Hallmark	Both Hallmarks	TSD Only	Poly(A) Only	Neither
Mischka (canFam4)	66	66	46	20	0	0
mCanLor1.2	59	57	42	14	1	2
China	49	49	40	8	1	0
Nala	64	63	42	21	0	1
Sandy	62	62	50	11	1	0
Tasha	50	50	35	13	2	0
Wags	51	51	40	10	1	0
Yella	51	50	34	14	2	1
Zoey	49	48	32	14	2	1

Supplementary Table 4. FullORF Retrocopy Hallmarks.

This table mimics the same pattern as seen in Table 2, 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. 65-81% of fullORF retrocopies have evidence of both hallmarks.

942 *Supplementary Files:*

943 Table S5A. All Pairwise Self-Alignment Duplications
944 Table S5B. All Self-Alignment Duplicated Intervals
945 Table S6. All Read-depth Segmental Duplications
946 Table S7. Duplicated Genes found by Read-depth
947 Table S8. All Retrocopies across All Assemblies
948 Table S9. GO Analysis of Duplicated and Parent Genes
949 Table S10. Shared Retrocopies Across Selected Canine Assemblies
950 Table S11. Mischka Retrocopies Converted to Fox and Dhole
951

952 # References

953 1. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW,
954 Tyler-Smith C, Hurles ME, et al: **Copy number variation: new insights in genome diversity.**
955 *Genome Res* 2006, **16:**949-961.
956 2. Conrad B, Antonarakis SE: **Gene Duplication: A Drive for Phenotypic Diversity and Cause of**
957 **Human Disease.** *Annual Review of Genomics and Human Genetics* 2007, **8:**17-35.
958 3. Dehal P, Boore JL: **Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate.**
959 *PLOS Biology* 2005, **3:**e314.
960 4. Abdullaev ET, Umarova IR, Arndt PF: **Modelling segmental duplications in the human genome.**
961 *BMC Genomics* 2021, **22:**496.
962 5. Rody HVS, Baute GJ, Rieseberg LH, Oliveira LO: **Both mechanism and age of duplications**
963 **contribute to biased gene retention patterns in plants.** *BMC Genomics* 2017, **18:**46.
964 6. Payen C, Koszul R, Dujon B, Fischer G: **Segmental Duplications Arise from Pol32-Dependent**
965 **Repair of Broken Forks through Two Alternative Replication-Based Mechanisms.** *PLOS Genetics*
966 2008, **4:**e1000175.
967 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**
968 **copy number variations in a segmental duplication-mediated rearrangement hotspot using**
969 **multiplex competitive amplification.** *Journal of Human Genetics* 2012, **57:**545-551.
970 8. Esnault C, Maestre J, Heidmann T: **Human LINE retrotransposons generate processed**
971 **pseudogenes.** *Nat Genet* 2000, **24:**363-367.
972 9. Kaessmann H, Vinckenbosch N, Long MY: **RNA-based gene duplication: mechanistic and**
973 **evolutionary insights.** *Nature Reviews Genetics* 2009, **10:**19-31.
974 10. Casola C, Betran E: **The Genomic Impact of Gene Retrocopies: What Have We Learned from**
975 **Comparative Genomics, Population Genomics, and Transcriptomic Analyses?** *Genome Biol Evol*
976 2017, **9:**1351-1373.
977 11. Tutar Y: **Pseudogenes.** *Comparative and Functional Genomics* 2012.
978 12. Brown EA, Dickinson PJ, Mansour T, Sturges BK, Aguilar M, Young AE, Korff C, Lind J, Ettinger CL,
979 Varon S, et al: **FGF4 retrogene on CFA12 is responsible for chondrodystrophy and**
980 **intervertebral disc disease in dogs.** *Proceedings of the National Academy of Sciences* 2017,
981 **114:**11476-11481.
982 13. Batcher K, Varney S, York D, Blacksmith M, Kidd JM, Rebhun R, Dickinson P, Bannasch D: **Recent,**
983 **full-length gene retrocopies are common in canids.** *Genome Res* 2022, **32:**1602-1611.
984 14. Meadows JRS, Kidd JM, Wang GD, Parker HG, Schall PZ, Bianchi M, Christmas MJ, Bougiouri K,
985 Buckley RM, Hitte C, et al: **Genome sequencing of 2000 canids by the Dog10K consortium**
986 **advances the understanding of demography, genome function and architecture.** *Genome Biol*
987 2023, **24:**187.
988 15. Bianchi CA, Marcellin-Little DJ, Dickinson PJ, Garcia TC, Li CF, Batcher K, Bannasch DL: **FGF4L2**
989 **retrogene copy number is associated with intervertebral disc calcification and vertebral**
990 **geometry in Nova Scotia Duck Tolling Retrievers.** *Am J Vet Res* 2023, **84**.
991 16. Išerić H, Alkan C, Hach F, Numanagić I: **Fast characterization of segmental duplication structure**
992 **in multiple genome assemblies.** *Algorithms for Molecular Biology* 2022, **17:**4.
993 17. Salzberg SL, Yorke JA: **Beware of mis-assembled genomes.** *Bioinformatics* 2005, **21:**4320-4321.
994 18. Pan W, Lonardi S: **Accurate detection of chimeric contigs via Bionano optical maps.**
995 *Bioinformatics* 2019, **35:**1760-1762.
996 19. Hartasanchez DA, Braso-Vives M, Heredia-Genestar JM, Pybus M, Navarro A: **Effect of Collapsed**
997 **Duplications on Diversity Estimates: What to Expect.** *Genome Biol Evol* 2018, **10:**2899-2905.

20. Alkan C, Coe BP, Eichler EE: **APPLICATIONS OF NEXT-GENERATION SEQUENCING Genome structural variation discovery and genotyping.** *Nature Reviews Genetics* 2011, **12:**363-375.
21. Ko BJ, Lee C, Kim J, Rhie A, Yoo DA, Howe K, Wood J, Cho S, Brown S, Formenti G, et al: **Widespread false gene gains caused by duplication errors in genome assemblies.** *Genome Biol* 2022, **23:**205.
22. 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.
23. 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-+.
24. 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.
25. 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.
26. 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.
27. 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.
28. Chen WK, Swartz JD, Rush LJ, Alvarez CE: **Mapping DNA structural variation in dogs.** *Genome Res* 2009, **19:**500-509.
29. 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.
30. 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.
31. 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.
32. 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.
33. 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.
34. 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.
35. 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**.

1046 36. Jagannathan V, Hitte C, Kidd JM, Masterson P, Murphy TD, Emery S, Davis B, Buckley RM, Liu Y-
1047 H, Zhang X-Q, et al: **Dog10K_Boxer_Tasha_1.0: A Long-Read Assembly of the Dog Reference**
1048 **Genome.** *Genes* 2021, **12:**847.
1049 37. Halo JV, Pendleton AL, Shen F, Doucet AJ, Derrien T, Hitte C, Kirby LE, Myers B, Sliwerska E,
1050 Emery S, et al: **Long-read assembly of a Great Dane genome highlights the contribution of GC-**
1051 **rich sequence and mobile elements to canine genomes.** *Proceedings of the National Academy*
1052 *of Sciences* 2021, **118**.
1053 38. Field MA, Rosen BD, Dudchenko O, Chan EKF, Minoche AE, Edwards RJ, Barton K, Lyons RJ,
1054 Tuipulotu DE, Hayes VM, et al: **Canfam_GSD: De novo chromosome-length genome assembly**
1055 **of the German Shepherd Dog (Canis lupus familiaris) using a combination of long reads, optical**
1056 **mapping, and Hi-C.** *Gigascience* 2020, **9**.
1057 39. Field MA, Yadav S, Dudchenko O, Esvaran M, Rosen BD, Skvortsova K, Edwards RJ, Keilwagen J,
1058 Cochran BJ, Manandhar B, et al: **The Australian dingo is an early offshoot of modern breed**
1059 **dogs.** *Sci Adv* 2022, **8:**eabm5944.
1060 40. Edwards RJ, Field MA, Ferguson JM, Dudchenko O, Keilwagen J, Rosen BD, Johnson GS, Rice ES,
1061 Hillier LD, Hammond JM, et al: **Chromosome-length genome assembly and structural variations**
1062 **of the primal Basenji dog (Canis lupus familiaris) genome.** *BMC Genomics* 2021, **22:**188.
1063 41. Shen F, Kidd JM: **Rapid, Paralog-Sensitive CNV Analysis of 2457 Human Genomes Using QuicK-**
1064 **mer2.** *Genes* 2020, **11:**141.
1065 42. Wang G-D, Larson G, Kidd JM, vonHoldt BM, Ostrander EA, Zhang Y-P: **Dog10K: the**
1066 **International Consortium of Canine Genome Sequencing.** *National Science Review* 2019, **6:**611-
1067 613.
1068 43. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human**
1069 **genome browser at UCSC.** *Genome Res* 2002, **12:**996-1006.
1070 44. Gene Ontology C, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, Feuermann
1071 M, Gaudet P, Harris NL, et al: **The Gene Ontology knowledgebase in 2023.** *Genetics* 2023, **224**.
1072 45. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS,
1073 Eppig JT, et al: **Gene ontology: tool for the unification of biology. The Gene Ontology**
1074 **Consortium.** *Nat Genet* 2000, **25:**25-29.
1075 46. Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou LP, Mi H: **PANTHER: Making**
1076 **genome-scale phylogenetics accessible to all.** *Protein Sci* 2022, **31:**8-22.
1077 47. Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD: **Molecular archeology**
1078 **of L1 insertions in the human genome.** *Genome Biol* 2002, **3:**research0052.
1079 48. Feng Q, Moran JV, Kazazian HH, Jr., Boeke JD: **Human L1 retrotransposon encodes a conserved**
1080 **endonuclease required for retrotransposition.** *Cell* 1996, **87:**905-916.
1081 49. Flasch DA, Macia A, Sanchez L, Ljungman M, Heras SR, Garcia-Perez JL, Wilson TE, Moran JV:
1082 **Genome-wide de novo L1 Retrotransposition Connects Endonuclease Activity with**
1083 **Replication.** *Cell* 2019, **177:**837-851 e828.
1084 50. Jurka J: **Sequence patterns indicate an enzymatic involvement in integration of mammalian**
1085 **retroposons.** *Proc Natl Acad Sci U S A* 1997, **94:**1872-1877.
1086 51. Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, Batzer MA, Moran JV: **DNA**
1087 **repair mediated by endonuclease-independent LINE-1 retrotransposition.** *Nat Genet* 2002,
1088 **31:**159-165.
1089 52. Gao X, Li Y, Adetula AA, Wu Y, Chen H: **Analysis of new retrogenes provides insight into dog**
1090 **adaptive evolution.** *Ecology and Evolution* 2019, **9:**11185-11197.
1091 53. Liu YJ, Zheng D, Balasubramanian S, Carriero N, Khurana E, Robilotto R, Gerstein MB:
1092 **Comprehensive analysis of the pseudogenes of glycolytic enzymes in vertebrates: the**

1093 **anomalously high number of GAPDH pseudogenes highlights a recent burst of retrotrans-**
1094 **positional activity.** *BMC Genomics* 2009, **10:**480.
1095 54. Breen M: **Canine cytogenetics--from band to basepair.** *Cytogenet Genome Res* 2008, **120:**50-60.
1096 55. Marquez LM, Miller DJ, MacKenzie JB, Van Oppen MJ: **Pseudogenes contribute to the extreme**
1097 **diversity of nuclear ribosomal DNA in the hard coral Acropora.** *Mol Biol Evol* 2003, **20:**1077-
1098 1086.
1099 56. Ciomborowska-Basheer J, Staszak K, Kubiak MR, Makalowska I: **Not So Dead Genes-Retrocopies**
1100 **as Regulators of Their Disease-Related Progenitors and Hosts.** *Cells* 2021, **10**.
1101 57. Kukekova AV, Johnson JL, Xiang X, Feng S, Liu S, Rando HM, Kharlamova AV, Herbeck Y,
1102 Serdyukova NA, Xiong Z, et al: **Red fox genome assembly identifies genomic regions associated**
1103 **with tame and aggressive behaviours.** *Nat Ecol Evol* 2018, **2:**1479-1491.
1104 58. Wang GD, Shao XJ, Bai B, Wang J, Wang X, Cao X, Liu YH, Wang X, Yin TT, Zhang SJ, et al:
1105 **Structural variation during dog domestication: insights from gray wolf and dhole genomes.**
1106 *Natl Sci Rev* 2019, **6:**110-122.
1107 59. Koch E, Schweizer RM, Schweizer TM, Stahler DR, Smith DW, Wayne RK, Novembre J: **De novo**
1108 **mutation rate estimation in wolves of known pedigree.** *Mol Biol Evol* 2019, **36:**2536-2547.
1109 60. Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ: **Human subtelomeres**
1110 **are hot spots of interchromosomal recombination and segmental duplication.** *Nature* 2005,
1111 **437:**94-100.
1112 61. Sulak M, Fong L, Mika K, Chigurupati S, Yon L, Mongan NP, Emes RD, Lynch VJ: **TP53 copy**
1113 **number expansion is associated with the evolution of increased body size and an enhanced**
1114 **DNA damage response in elephants.** *Elife* 2016, **5**.
1115 62. Batcher K, Varney S, Raudsepp T, Jevit M, Dickinson P, Jagannathan V, Leeb T, Bannasch D:
1116 **Ancient segmentally duplicated LCORL retrocopies in equids.** *PLoS One* 2023, **18:**e0286861.
1117 63. Segurel L, Wyman MJ, Przeworski M: **Determinants of mutation rate variation in the human**
1118 **germline.** *Annu Rev Genomics Hum Genet* 2014, **15:**47-70.
1119 64. Lipson M, Loh PR, Sankararaman S, Patterson N, Berger B, Reich D: **Calibrating the Human**
1120 **Mutation Rate via Ancestral Recombination Density in Diploid Genomes.** *PLoS Genet* 2015,
1121 **11:**e1005550.
1122 65. Narasimhan VM, Rahbari R, Scally A, Wuster A, Mason D, Xue Y, Wright J, Trembath RC, Maher
1123 ER, van Heel DA, et al: **Estimating the human mutation rate from autozygous segments reveals**
1124 **population differences in human mutational processes.** *Nat Commun* 2017, **8:**303.
1125 66. Palamara PF, Francioli LC, Wilton PR, Genovese G, Gusev A, Finucane HK, Sankararaman S,
1126 Genome of the Netherlands C, Sunyaev SR, de Bakker PI, et al: **Leveraging Distant Relatedness**
1127 **to Quantify Human Mutation and Gene-Conversion Rates.** *Am J Hum Genet* 2015, **97:**775-789.
1128 67. Tian X, Browning BL, Browning SR: **Estimating the Genome-wide Mutation Rate with Three-**
1129 **Way Identity by Descent.** *Am J Hum Genet* 2019, **105:**883-893.
1130 68. Skoglund P, Ersmark E, Palkopoulou E, Dalen L: **Ancient wolf genome reveals an early**
1131 **divergence of domestic dog ancestors and admixture into high-latitude breeds.** *Curr Biol* 2015,
1132 **25:**1515-1519.
1133 69. Bergstrom A, Frantz L, Schmidt R, Ersmark E, Lebrasseur O, Girdland-Flink L, Lin AT, Stora J,
1134 Sjogren KG, Anthony D, et al: **Origins and genetic legacy of prehistoric dogs.** *Science* 2020,
1135 **370:**557-564.
1136 70. Botigue LR, Song S, Scheu A, Gopalan S, Pendleton AL, Oetjens M, Taravella AM, Seregely T,
1137 Zeeb-Lanz A, Arbogast RM, et al: **Ancient European dog genomes reveal continuity since the**
1138 **Early Neolithic.** *Nat Commun* 2017, **8:**16082.

1139 71. Fan Z, Silva P, Gronau I, Wang S, Armero AS, Schweizer RM, Ramirez O, Pollinger J, Galaverni M,
1140 Ortega Del-Vecchyo D, et al: **Worldwide patterns of genomic variation and admixture in gray**
1141 **wolves.** *Genome Res* 2016, **26:**163-173.
1142 72. Sinding MS, Gopalakrishan S, Vieira FG, Samaniego Castruita JA, Raundrup K, Heide Jorgensen
1143 MP, Meldgaard M, Petersen B, Sicheritz-Ponten T, Mikkelsen JB, et al: **Population genomics of**
1144 **grey wolves and wolf-like canids in North America.** *PLoS Genet* 2018, **14:**e1007745.
1145 73. Ewing AD, Ballinger TJ, Earl D, Broad Institute Genome S, Analysis P, Platform, Harris CC, Ding L,
1146 Wilson RK, Haussler D: **Retrotransposition of gene transcripts leads to structural variation in**
1147 **mammalian genomes.** *Genome Biol* 2013, **14:**R22.
1148 74. Nachman MW, Crowell SL: **Estimate of the mutation rate per nucleotide in humans.** *Genetics*
1149 2000, **156:**297-304.
1150 75. Jonsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT, Hjorleifsson
1151 KE, Eggertsson HP, Gudjonsson SA, et al: **Parental influence on human germline de novo**
1152 **mutations in 1,548 trios from Iceland.** *Nature* 2017, **549:**519-522.
1153 76. Feng X, Li H: **Higher Rates of Processed Pseudogene Acquisition in Humans and Three Great**
1154 **Apes Revealed by Long-Read Assemblies.** *Mol Biol Evol* 2021, **38:**2958-2966.
1155 77. Batcher K, Dickinson P, Maciejczyk K, Brzeski K, Rasouliha SH, Letko A, Drogemuller C, Leeb T,
1156 Bannasch D: **Multiple FGF4 Retrocopies Recently Derived within Canids.** *Genes (Basel)* 2020,
1157 **11**.
1158 78. Wang W, Kirkness EF: **Short interspersed elements (SINEs) are a major source of canine**
1159 **genomic diversity.** *Genome Res* 2005, **15:**1798-1808.
1160 79. Beck CR, Garcia-Perez JL, Badge RM, Moran JV: **LINE-1 elements in structural variation and**
1161 **disease.** *Annu Rev Genomics Hum Genet* 2011, **12:**187-215.
1162 80. Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV: **Human L1**
1163 **retrotransposition: cis preference versus trans complementation.** *Mol Cell Biol* 2001, **21:**1429-
1164 1439.
1165 81. Dewannieux M, Esnault C, Heidmann T: **LINE-mediated retrotransposition of marked Alu**
1166 **sequences.** *Nat Genet* 2003, **35:**41-48.
1167 82. Verscheure S, Backeljau T, Desmyter S: **In silico discovery of a nearly complete mitochondrial**
1168 **genome Numt in the dog (Canis lupus familiaris) nuclear genome.** *Genetica* 2015, **143:**453-458.
1169 83. Quinlan AR: **BEDTools: The Swiss-Army Tool for Genome Feature Analysis.** *Curr Protoc*
1170 *Bioinformatics* 2014, **47:**11 12 11-34.
1171 84. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res*
1172 1999, **27:**573-580.
1173 85. Morgulis A, Gertz EM, Schaffer AA, Agarwala R: **WindowMasker: window-based masker for**
1174 **sequenced genomes.** *Bioinformatics* 2006, **22:**134-141.
1175 86. Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, Sahinalp SC: **mrsFAST: a cache-**
1176 **oblivious algorithm for short-read mapping.** *Nat Methods* 2010, **7:**576-577.
1177 87. Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics* 2018, **34:**3094-
1178 3100.
1179 88. Shumate A, Salzberg SL: **Liftoff: accurate mapping of gene annotations.** *Bioinformatics* 2021,
1180 **37:**1639-1643.
1181 89. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T,
1182 McCarthy SA, Davies RM, Li H: **Twelve years of SAMtools and BCFtools.** *Gigascience* 2021, **10**.
1183 90. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12:**656-664.
1184 91. Li H: **New strategies to improve minimap2 alignment accuracy.** *Bioinformatics* 2021, **37:**4572-
1185 4574.

1186 92. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.**
1187 *Trends Genet* 2000, **16:**276-277.
1188 93. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981,
1189 **147:**195-197.
1190 94. Hunter JD: **Matplotlib: A 2D Graphics Environment.** *Computing in Science & Engineering* 2007,
1191 **9:**90-95.
1192 95. Tareen A, Kinney JB: **Logomaker: beautiful sequence logos in Python.** *Bioinformatics* 2020,
1193 **36:**2272-2274.
1194 96. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H: **UpSet: Visualization of Intersecting**
1195 **Sets.** *IEEE Trans Vis Comput Graph* 2014, **20:**1983-1992.
1196 97. Freedman AH: **Genome Sequencing Highlights the Dynamic Early History of Dogs (vol 10,**
1197 **e1004016, 2014).** *Plos Genetics* 2014, **10**.
1198 98. Larson G, Bradley DG: **How Much Is That in Dog Years? The Advent of Canine Population**
1199 **Genomics.** *Plos Genetics* 2014, **10**.

1200