# Recent, full-length gene retrocopies are common in canids

Kevin Batcher,[1] Scarlett Varney,[1] Daniel York,[2] Matthew Blacksmith,[3] Jeffrey M. Kidd,[3,4] Robert Rebhun,[2] Peter Dickinson,[2] and Danika Bannasch[1]

[1]*Department of Population Health and Reproduction,* [2]*Department of Surgical and Radiological Sciences, University of California, Davis, Davis, California 95616, USA;* [3]*Department of Human Genetics,* [4]*Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA*

Gene retrocopies arise from the reverse transcription and insertion into the genome of processed mRNA transcripts. Although many retrocopies have acquired mutations that render them functionally inactive, most mammals retain active LINE-1 sequences capable of producing new retrocopies. New retrocopies, referred to as retro copy number variants (retroCNVs), may not be identified by standard variant calling techniques in high-throughput sequencing data. Although multiple functional *FGF4* retroCNVs have been associated with skeletal dysplasias in dogs, the full landscape of canid retroCNVs has not been characterized. Here, retroCNV discovery was performed on a whole-genome sequencing data set of 293 canids from 76 breeds. We identified retroCNV parent genes via the presence of mRNA-specific 30-mers, and then identified retroCNV insertion sites through discordant read analysis. In total, we resolved insertion sites for 1911 retroCNVs from 1179 parent genes, 1236 of which appeared identical to their parent genes. Dogs had on average 54.1 total retroCNVs and 1.4 private retroCNVs. We found evidence of expression in testes for 12% (14/113) of the retroCNVs identified in six Golden Retrievers, including four chimeric transcripts, and 97 retroCNVs also had significantly elevated $F_{ST}$ across dog breeds, possibly indicating selection. We applied our approach to a subset of human genomes and detected an average of 4.2 retroCNVs per sample, highlighting a 13-fold relative increase of retroCNV frequency in dogs. Particularly in canids, retroCNVs are a largely unexplored source of genetic variation that can contribute to genome plasticity and that should be considered when investigating traits and diseases.

[Supplemental material is available for this article.]

Gene retrotransposition occurs when mRNA is reverse transcribed into DNA and inserted back into the genome, resulting in an intron-less copy of a gene referred to as a retrocopy or a processed pseudogene. This process is performed in mammals by long interspersed nuclear element 1 (LINE-1 or L1) proteins acting in *trans* on cellular mRNA (Jurka 1997; Esnault et al. 2000; Richardson et al. 2014). LINE-1-mediated retrotransposition results in the duplication of short (10- to 20-bp) segments of genomic DNA flanking the insertion, referred to as target site duplications (TSDs). Thousands of retrocopies have been identified in mammalian reference assemblies, although the exact number varies by species and by the annotation method applied (Rosikiewicz et al. 2017; Frankish et al. 2019). Most of these reference retrocopies are the consequence of ancestral retrotransposition events, evidenced by the accumulation of mutations that differentiate the retrocopy sequence from that of the parent gene (Rosikiewicz et al. 2017). Although these ancestral retrocopies tend to be fixed in a species (Ewing et al. 2013; Schrider et al. 2013; Kabza et al. 2015), most mammalian genomes contain active LINE-1s capable of producing novel retrocopy insertions (Penzkofer et al. 2016). These more recent retrocopy insertions may not be fixed in a species, resulting in gene copy number variation between individuals, referred to as retro copy number variants (retroCNVs) (Abyzov et al. 2013; Ewing et al. 2013; Schrider et al. 2013; Richardson et al. 2014; Casola and Betrán 2017). Although some of the retrocopies present in a reference genome assembly may be polymorphic in a species and thus may be reference retroCNVs, there are also nonreference retroCNVs that are not found in the assembly itself (Schrider et al. 2013).

RetroCNVs are a type of complex structural variant that requires specialized techniques for identification within whole-genome sequencing (WGS) data (Casola and Betrán 2017). When Illumina paired-end reads are aligned to a reference assembly, any reference retroCNV that is absent in an individual will appear as a deletion relative to the assembly. However, when an individual has a nonreference retroCNV, the reads coming from that retroCNV will align to the parent gene. Because the retrocopy lacks introns, discordant reads are observed aligning only to the exons of the parent gene. Discordant reads are also found at the 3′ and 5′ end of the gene mapping to the insertion site of the retroCNV. These two features can be used to identify nonreference retroCNVs from WGS data, with the gold standard in retroCNV discovery requiring identification of the parent gene and characterization of the insertion site (Richardson et al. 2014; Zhang et al. 2021). Estimates of gene retrotransposition rates have varied, although recent analyses using high-coverage WGS data have identified 1663 retroCNV parent genes in populations of mice (Zhang et al. 2021) and 503 in human populations (Zhang et al. 2017), indicating that gene retrotransposition is a common occurrence. There is

**Corresponding author: dlbannasch@ucdavis.edu**
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.276828.122.




32:1602–1611 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/22; www.genome.org

limited information about retroCNVs in dogs (Gao et al. 2019; Kim et al. 2019; Batcher et al. 2020); however, short interspersed nuclear element (SINE) insertions, which are also mobilized in *trans* by LINE-1 encoded proteins in a manner similar to gene retrocopies, have been shown to be highly dimorphic in dogs (Wang and Kirkness 2005; Halo et al. 2021).

Although retrocopies have historically been referred to as processed pseudogenes and presumed to be nonfunctional, evidence has accumulated for retrocopy expression and functionality (Cheetham et al. 2020; Troskie et al. 2021; Zhang et al. 2021). It has been argued that retroCNVs are likely to be deleterious based on negative selection in natural populations of mice (Zhang et al. 2021), and consistent with this hypothesis, retroCNVs in humans have been shown to be involved in cancer as well as neurodegenerative, mental, or cardiovascular disorders (Ciomborowska-Basheer et al. 2021). In dogs, two recently inserted and expressed *FGF4* retrocopies are associated with dominant skeletal dysplasias (Parker et al. 2009; Brown et al. 2017). Several additional *FGF4* retrocopies with no known phenotypic associations were also discovered, indicating that retroCNV formation may also be a common occurrence in dogs (Batcher et al. 2020). Artificial breed selection by humans could also increase the allele frequency of functional retroCNVs, as appears to have been the case with the *FGF4* retroCNVs, which are common in many breeds (Batcher et al. 2019). As such, analysis of the full landscape of retroCNVs in dogs could lead to interesting insights into retrocopy biology and may additionally help in identifying causative variants for phenotypic associations in dogs. The goal of this current study was to characterize the landscape of retroCNVs in dogs by performing retroCNV discovery on a diverse data set of canids and by further analyzing the retroCNV for evidence of function.

## Results

### retroCNV discovery

We used high-coverage WGS data from 293 canids (median coverage of 25.6×) aligned to the CanFam3.1 reference genome for our retroCNV discovery data set (Supplemental Table S1). Our approach to retroCNV discovery was to first identify mRNA-specific 30-mers, which are present in spliced gene sequences but absent from the CanFam3.1 reference assembly (Fig. 1A). These sequences are only found in genomic DNA when a nonreference retroCNV is present (Fig. 1B). No mRNA-specific 30-mers were identified from single-exon genes (N = 1574) or genes with recent retrocopies already present in the reference assembly (N = 75). We identified mRNA-specific 30-mers for 18,192 protein-coding genes and 10,807 long noncoding RNAs, which were then used for retroCNV parent gene discovery. In total, 1870 putative retroCNV parent genes were identified in the 293 canid data set based on the presence of these mRNA-specific 30-mers.

To resolve the insertion site of nonreference retroCNVs (Supplemental Fig. S1), we analyzed discordant paired-end reads. From the 1870 parent genes identified as having putative retroCNVs, insertion sites were resolved for 1911 total nonreference retroCNVs coming from 1179 parent genes (Fig. 2A). Many (808/1911, 42.3%) insertion sites were located within the introns of other protein-coding genes (Supplemental Table S2; Supplemental Fig. S2). Of the 1179 retroCNV parent genes, 1150 were protein coding, whereas 29 were lncRNAs. Four of the previously identified *FGF4* retroCNVs were successfully identified through our approach (Supplemental Table S2). The TSD was resolved for 1676 (87.7%) of the retroCNVs, and the median TSD length was 16 bp. No insertion sites could be identified for retroCNVs derived from 691 of the putative retroCNV parent genes; however, 125 of these parent genes had discordant reads mapping between exons, which may indicate the retroCNV inserted into repetitive or unresolved regions of the CanFam3.1 reference genome (Supplemental Table

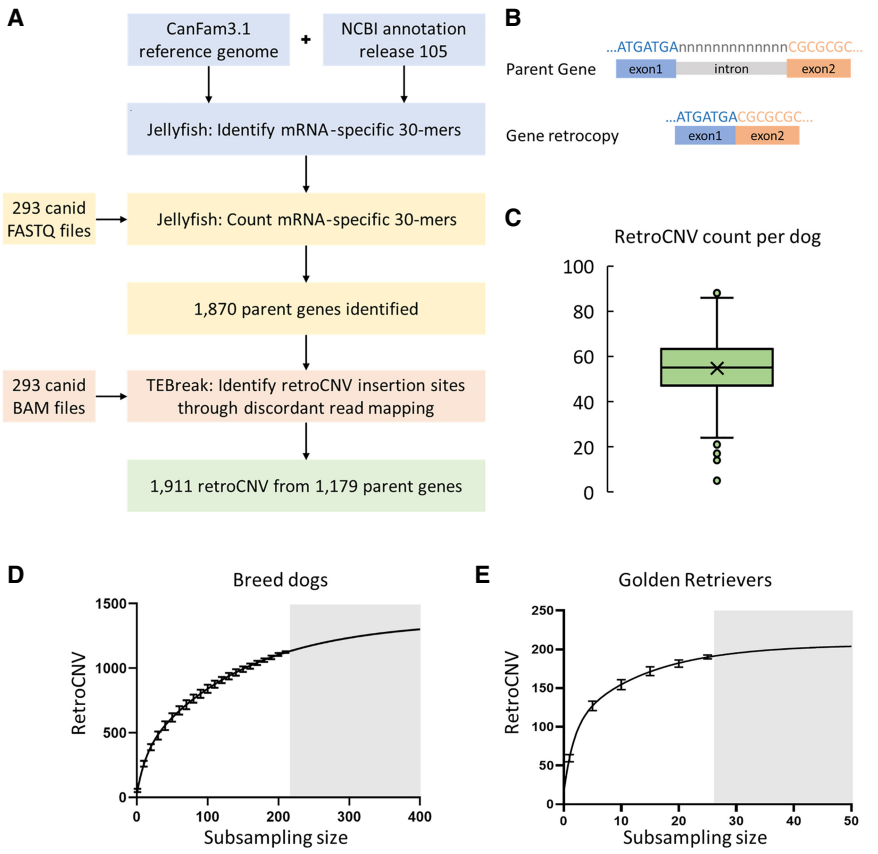


**Figure 1.** retroCNV discovery. (*A*) retroCNV discovery pipeline. (*B*) mRNA-specific sequences are formed owing to intron removal; these sequences are used to identify nonreference retroCNV parent genes in FASTQ reads from genomic DNA sequence. (*C*) Total nonreference retroCNV count seen in individual domestic dogs (average, 54.1). (*D*) Number of detected retroCNV insertions with increasing resampling sample sizes in breed dogs. Subsample sizes were selected from one to 210, with a step size of 10, and 100 replicates within each subsample. Bars represent standard deviation of the replicates at each subsample size, and the gray area shows the prediction of increasing the number of dogs beyond the number used in this study. (*E*) Number of detected retroCNV insertions with increasing resampling sample sizes in Golden Retrievers. Subsample sizes were selected from one to 25, with a step size of five, and 100 replicates within each subsample. Bars represent standard deviation of the replicates at each subsample size.

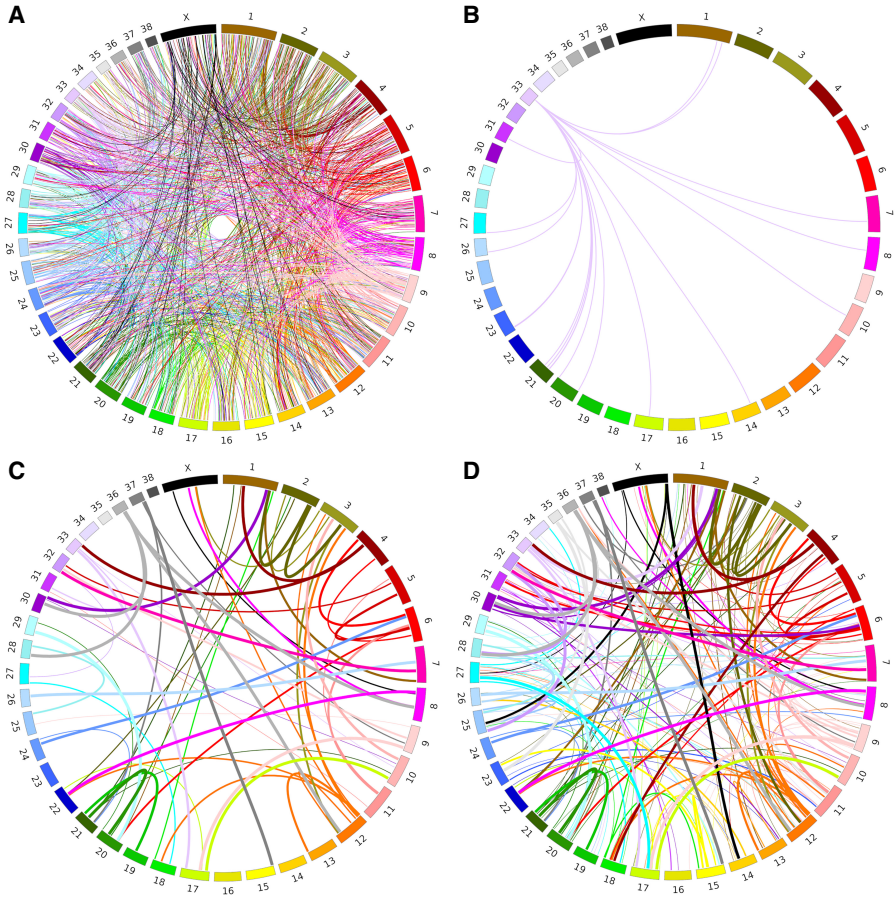

**Figure 2.** Circos plots highlighting the location of retroCNV parent genes and their insertion sites, with links colored based on the chromosome of the parent gene. (*A*) All recent retroCNVs identified in canids with no retroCNV-specific variants. (*B*) All retroCNVs of the *GAP43* parent gene. (*C*) All retroCNVs present in a single Golden Retriever (SRR7107792). (*D*) All retroCNVs present in all (N = 26) Golden Retrievers. (*C*, *D*) The thickness of the link represents how common the retroCNV is within the Golden Retriever data set.

S3). Further inspection revealed that 21 of the putative retroCNV parent genes had discordant reads mapping to satellite DNA, suggesting insertion of the retroCNV into a telomeric or centromeric region. Additionally, four parent genes (*MITF*, *NME7*, *TXNDC12*, and *PPP2CB*) that had discordant reads in all of the males and none of the females were identified, indicating that the retroCNVs were likely on the Y Chromosome, which is not represented in the CanFam3.1 assembly.

We also identified retrocopies that are present in the CanFam3.1 reference genome but missing in at least one of the 293 canids, which we refer to as reference retroCNVs. We identified 58 reference retroCNVs, which were confirmed through visual analysis of aligned WGS data (Supplemental Table S4). Several reference retroCNVs were highly prevalent in dogs (≥70%) and rare in wolves (≤10%), including the reference retroCNV *AKR1B1* (retro_cfam_63), which has previously been associated with domestication (Wang et al. 2019), as well as retroCNVs of *MGST3* (retro_cfam_35) and *RPL27A* (retro_cfam_145).

We focused all subsequent analysis on the 1911 nonreference retroCNVs with resolved insertion sites. A full matrix matching individual canids with retroCNVs is available in Supplemental Table S5. Although 880 of the retroCNV parent genes only had one retroCNV insertion site identified, 338 retroCNV parent genes had multiple insertion sites identified, including genes such as *GAPDH*, which had 21 retroCNV insertions (Supplemental Fig. S3; Supplemental Table S2). Most of the retroCNV parent genes (900/1179, 76.3%) had no retrocopies present in the CanFam3.1 reference assembly. Additionally, 231 retroCNV parent genes had no known retrocopies in any of the mammalian reference genome assemblies, including *GAP43*, which had 18 retroCNV insertion sites identified in canids (Fig. 2B; Supplemental Table S2).

Table 1 highlights some aspects of the nonreference retroCNVs by population. Dogs with an assigned breed, which we refer to as breed dogs, had 54.1 nonreference retroCNVs on average (95% CI 52.5–55.7) (Fig. 1C). Within the 227 breed dogs, there were 325 private (unique to an individual dog) retroCNVs, or 1.4 per breed dog on average (95% CI 1.2–1.7), whereas the 43 free-ranging dogs had 214 private retroCNVs and 5.0 each on average (95% CI 4.0–6.0). Most nonreference retroCNVs were identified in a small number of dogs and at a low allele frequency in the entire population (Supplemental Fig. S4). There were also 689 retroCNVs exclusive to breed dogs, 247 retroCNVs exclusive to the 18 wild canids, and 197 retroCNVs exclusive to the three African wild dogs (Table 1). We also observed that individuals that were sequenced at a higher depth tended to have more retroCNV insertion sites resolved (Supplemental Fig. S5). Although samples at greater than 30× coverage had an average of 57.9 retroCNVs (95% CI 56.2–59.6), samples between 10× and 30× coverage had 54.8 (95% CI 52.3–57.3) and samples less than 10× coverage had 40 (95% CI 35.2–44.8) on average.

## retroCNV-specific gene variants

To estimate how recently the retroCNVs inserted, we identified variants that occurred in the retroCNV after insertion through the analysis of variants at the parent gene locus. Most retroCNVs had not acquired any new variants after insertion, as retroCNV-specific variants were only identified in 153/1390 (11.0%) of the retroCNVs analyzed, and only eight of 1390 (0.6%) retroCNVs had high impact mutations (Supplemental Table S6). Highlighting their more recent origin, only 92/1212 (7.6%) of the dog exclusive retroCNV had any retroCNV-specific variants, whereas 61/178 (34.3%) of the retroCNVs that were shared across canids had retroCNV-specific variants.

## Resampling

We expected that more retroCNVs may be detected in even larger data sets. To test this, we performed random resampling of individual breed dogs. Although we identified 1140 retroCNVs in the 227 breed dogs, extrapolation from random resampling indicated that

**Table 1.** Population summary of nonreference retroCNVs

| Population | Total retroCNVs | Average retroCNVs | Private retroCNVs | Exclusive retroCNVs |
|---|---|---|---|---|
| Breed dogs (N = 227) | 1165 | 54.1 | 325 | 689 |
| Free-ranging dogs (N = 43) | 705 | 56.6 | 214 | 254 |
| Dingoes (N = 3) | 85 | 55.3 | 9 | 13 |
| Wolves (N = 10) | 354 | 71.5 | 144 | 185 |
| Coyotes (N = 5) | 104 | 33.8 | 31 | 49 |
| African wild dogs (N = 3) | 214 | 173.0 | 44 | 197 |

(Private) retroCNVs unique to a specific individual, (exclusive) retroCNVs unique to a specific population.

we would expect to identify over 1300 retroCNVs in a data set of 400 breed dogs (Fig. 1D). We similarly performed resampling within Golden Retrievers and observed that smaller sample sizes (N = 26) can sufficiently capture the majority of retroCNVs within a single breed (Figs. 1E, 2C,D). Resampling within the free-ranging dogs indicated that a large number of retroCNVs remains to be discovered, highlighting the heterogeneous nature of the free-ranging dogs (Supplemental Fig. S6).

## retroCNV validation

We expected that some of the nonreference retroCNVs, although absent from the CanFam3.1 reference genome assembly, would be present in the alternative canid genome assemblies. This would confirm them as true retroCNVs and validate our discovery method. To test this, we first applied our retroCNV discovery pipeline to Illumina WGS data generated from four canids that have been used to create alternate canid reference assemblies, and identified 248 nonreference retroCNVs (Table 2). We then directly confirmed the presence of 173 of the 248 (69.8%) retroCNVs within their respective assemblies (Supplemental Table S7). We also identified the insertion site for six retroCNVs that had not been resolved through discordant read mapping (Supplemental Table S7). Most of the retroCNVs were full length with respect to the parent genes, with 153/179 (85.5%) containing the entire parental gene coding sequence.

Genome assembly involves the collapsing of heterozygous haplotypes, a process that could have resulted in heterozygous retroCNVs being excluded from the assembly sequence (Feng and Li 2021). Therefore, we also analyzed Pacific Biosciences (PacBio) long-read data from each sample aligned to the CanFam3.1 assembly for evidence of retrocopy insertions (Supplemental Data S1). We found evidence for 236/248 (95.1%) of the retroCNVs within the long-read data (Supplemental Table S8). A poly(A) tail ≥10 bp was observed in 211/236 (89.4%) of the retroCNVs, with median length of 29 bp. A TSD was identified for 218/236 (92.4%) of the retroCNVs, with median length of 14 bp. Overall, 244/248 (98.4%) of the retroCNVs were validated either within their respective genome assembly or in the long-read data (Table 2).

We also developed PCR assays for nine of the predicted retroCNVs and validated them through Sanger sequencing (Fig. 3A). In individuals positive for the retroCNV, we observed a poly(A) tail at one end of the insertion site and the 5′ gene sequence of the retroCNV parent gene on the other end of the insertion site (Fig. 3C), which matched the expectation from the discordant reads and confirmed the presence of a retrocopy. For each retroCNV sequenced, the TSD was identified as the duplicated genomic sequence present at both ends of the insertion (Fig. 3C; Supplemental Table S9). The TSD identified through Sanger sequencing for these nine retroCNVs matched the predicted TSD from the discordant reads. We additionally identified dogs lacking the retroCNVs (Fig. 3B), confirming the retrocopies as polymorphic insertions. The list of retroCNVs validated through PCR and Sanger sequencing and the associated breeds is available in Supplemental Table S9.

## retroCNV selection and expression

We calculated a fixation index ($F_{ST}$) for the retroCNVs across breed clades (Parker et al. 2017) and identified 97 retroCNVs that had significantly elevated $F_{ST}$ (see Methods). This included the two previously identified *FGF4* retroCNVs (Table 3; for full list and clade distribution, see Supplemental Tables S10, S11).

To determine if the retroCNVs showed evidence of expression, we first performed WGS and retroCNV discovery in six Golden Retrievers and then RNA-seq using testes from the same six Golden Retrievers. There were 113 total nonreference retroCNVs identified in the six individuals. RetroCNV-specific variants were present in 24 of the 113 retroCNVs, allowing for the distinction between parent and retroCNV-derived transcripts, and confirmed the expression for two retroCNVs. Among the retroCNVs that had inserted within the introns of another gene, chimeric reads between a retroCNV parent gene and a gene at the insertion site were observed in four of 42. This included the *COILL2* retroCNV, which is inserted within the 5′ UTR of *LOC100686934*, producing a novel chimeric transcript in two of the six Golden Retrievers (Supplemental Fig. S7). However, five of the 42 insertion site genes were not sufficiently expressed in testes (fewer than one transcript per million in all samples) to allow for the evaluation of chimera formation with the retroCNVs. Additionally, discordant reads mapping to the parent gene loci were observed at 11 of the insertion sites. Overall, at least one form of evidence for expression was observed for 12.4% (14/113) of the retroCNVs in six Golden Retriever testes (Supplemental Table S12). The expressed retroCNV *FARSBL1* was present in 60% of all dogs and only a single wild canid.

**Table 2.** Analysis of retroCNVs in alternative canid genome assemblies

| Assembly | Predicted retroCNVs | RetroCNVs in assembly | RetroCNVs in assembly or PacBio |
|---|---|---|---|
| UMICH_Zoey_3.1 | 49 | 33 | 47 |
| UU_Cfam_GSD_1.0 | 59 | 41 | 59 |
| Canfam_GSD | 60 | 41 | 59 |
| CanLup_DDS | 80 | 58 | 79 |
| Total | 248 | 173 | 244 |

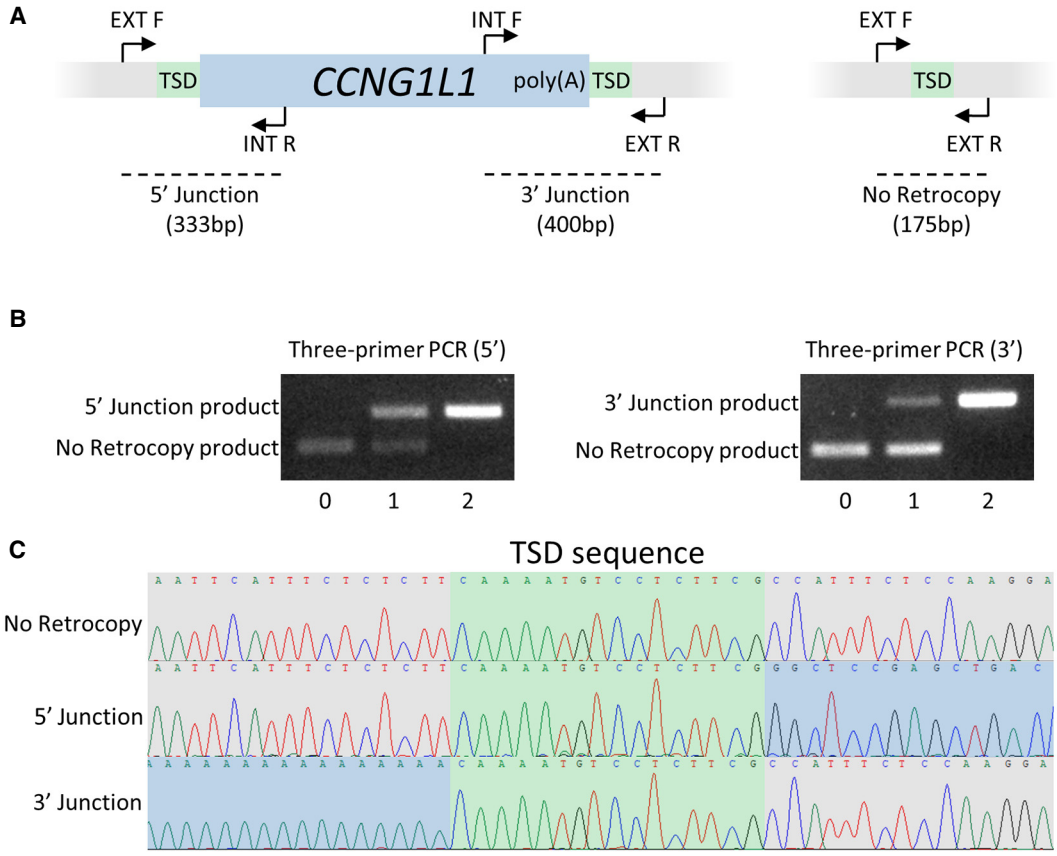


**Figure 3.** retroCNV validation. (*A*) Primer design for retroCNV genotyping, with *CCNG1L1* as an example. When *CCNG1L1* is present, the EXT_F and INT_R primers produce a 333-bp product at the 5′ junction, and the INT_F and EXT_R primers produce a 400-bp product at the 3′ junction. When the retrocopy is absent, the EXT_F and EXT_R primers produce a 175-bp product. (*B*) Three-primer PCR results for *CCNG1L1* at the 5′ and 3′ junctions for individuals with zero, one, and two copies of *CCNG1L1*. The two external primers EXT_F and EXT_R are included in both reactions, as well as one of the internal primers, INT_F (5′ Junction) or INT_R (3′ Junction). (*C*) Sanger sequencing results for the *CCNG1L1* retroCNV. The TSD is identified as the genomic sequence from the insertion site, which is present at both the 5′ and 3′ ends of the retroCNV.

### retroCNV discovery in humans

To determine the rate of retroCNV occurrence in another species for comparison, we performed retroCNV discovery in 78 individuals from 26 populations using The 1000 Genomes Project Consortium phase 3 high-coverage data set (The 1000 Genomes Project Consortium 2015; Byrska-Bishop et al. 2021). We resolved insertion sites for 46 nonreference retroCNVs from 44 parent genes in the 78 samples (Supplemental Table S13). Of the 44 retroCNV parent genes, 40 have been previously identified in human data sets, whereas 34 of the 46 retroCNV insertion sites had been previously identified. A full analysis of which human retroCNVs have been identified in previous studies is available in Supplemental Table S14. Individuals in this data set had 4.2 retroCNVs on average (95% CI 3.9–4.6).

### Rate of retroCNV formation in canids

We observed a total of 214 nonreference retroCNVs in the three African wild dogs (*Lycaon pictus*), which had 173 retroCNVs each on average. Most of these retroCNVs were private to the African wild dogs, and only 17 retroCNVs were shared between the African wild dogs and any other canid, indicating that most of the retroCNVs identified in either species inserted after the species had diverged. Similarly, we identify 194 retroCNVs that are exclusive to gray wolves and 1010 retroCNVs exclusive to breed dogs. Genetic analyses have indicated that domestication in dogs occurred ~25,000 yr ago, whereas breed formation largely occurred within the last 200 yr (Ostrander et al. 2019). If the 1010 breed dog–specific retroCNVs inserted after domestication, we can estimate the rate of retroCNV accumulation at approximately four per 100 yr. Alternatively, within our data set of Golden Retrievers (N = 26), we identify 10 retroCNVs that are exclusive to the breed and not found in any other breed dogs or canids. Because the Golden Retriever breed was formed ~200 yr ago, 10 Golden Retriever–exclusive retroCNVs would indicate a similar rate of retroCNV accumulation at five per 100 yr.

## Discussion

Previous analyses of canid retrocopies have focused on those retrocopies present in the CanFam3.1 reference assembly, which was produced from a single dog (Hoeppner et al. 2014; Rosikiewicz et al. 2017; Gao et al. 2019). In this study, we characterized a rich landscape of retroCNVs in canids, consistent with an active LINE-1. By applying a novel approach to retroCNV discovery on a diverse data set of 293 canids, we identified 1911 retroCNVs,

**Table 3.** RetroCNVs with the highest $F_{ST}$ between breed clades

| retroCNV | $F_{ST}$ | Insertion site | Total dogs | Insertion site gene |
|---|---|---|---|---|
| *NDUFAF4L1* | 0.484 | Chr 2: 6,242,423–6,242,444 | 5 | *LOC111091106* |
| *FGF4L2* | 0.439 | Chr 12: 33,710,166–33,710,178 | 16 | — |
| *RHEBL1* | 0.430 | Chr 9: 27,506,566–27,506,581 | 4 | *CA10* |
| *RPS2* | 0.430 | Chr 8: 5,138,945–5,139,948 | 155 | — |
| *S100PL4* | 0.413 | Chr 34: 34,324,611–34,324,631 | 13 | — |
| *PREPL1* | 0.403 | Chr 8: 7,565,459–7,565,475 | 5 | — |
| *ARHGAP5L1* | 0.399 | Chr 5: 18,750,413–18,750,430 | 9 | — |
| *RPS16L1* | 0.399 | Chr 3: 26,720,828–26,720,840 | 9 | — |
| *ARPC1BL1* | 0.398 | Chr 14: 34,916,675–34,916,690 | 5 | — |
| *NAA20L1* | 0.395 | Chr 18: 34,191,916–34,191,933 | 9 | *KIAA1549L* |
| *RESTL1* | 0.388 | Chr 12: 36,658,275–36,658,284 | 13 | *COL12A1* |
| *NAP1L1L2* | 0.365 | Chr 17: 29,923,085–29,923,100 | 7 | *LOC102154187* |
| *NAP1L1L3* | 0.358 | Chr 18: 49,079,285–49,079,285 | 11 | *IGHMBP2* |
| *C16orf87* | 0.350 | Chr 22: 35,188,887–35,190,448 | 113 | *C22H16orf87* |
| *FAM133BL4* | 0.348 | Chr 29: 31,980,926–31,981,012 | 8 | *CA2* |
| *RPL10L1* | 0.348 | Chr 3: 32,009,305–32,009,323 | 8 | *NIPA1* |
| *ST13* | 0.342 | Chr 1: 55,086,310–55,087,949 | 107 | *UNC93A* |
| *LSM2L1* | 0.341 | Chr 22: 35,508,014–35,508,030 | 13 | — |
| *FGF4L1* | 0.337 | Chr 18: 20,443,708–20,443,726 | 15 | — |
| *EIF4BL3* | 0.336 | Chr 18: 16,670,414–16,670,429 | 14 | *RELN* |

most of which had inserted recently. Domestic dogs have 54.1 nonreference retroCNVs each on average, and as many of the retroCNVs are private to a single individual, we also expect to find additional retroCNVs in larger discovery data sets. We observed many retroCNVs that appear under selection and that, even within a single tissue type, 12% of the retroCNVs were expressed or forming novel chimeric transcripts with nearby genes, indicating that some of the retroCNVs may have functional and phenotypic consequences in canids.

Our approach to retroCNV discovery successfully resolved a large number of retroCNVs in canids. In humans, previous studies using low-coverage WGS data sets have underestimated the rate of retrocopy insertion (Richardson et al. 2014). When long-read assemblies were analyzed, the estimated rate of retroCNV formation in humans was increased from 39 events per 939 individuals to 40 events per 22 individuals, or 4.1 per individual on average (Feng and Li 2021). Similar to the analysis by Feng and Li, we found that humans had on average 4.2 retroCNV insertions using our discovery method, which identified 46 total retroCNV insertions in 78 human genomes. Although most of the retroCNV parent genes have been reported in previous analyses of human genomes, eight of the insertion sites have not been previously identified (Zhang et al. 2017). Previous analysis of canine SINEs, which are also mobilized via LINE-1 proteins acting in *trans*, highlighted a rate of SINE insertions 10- to 100-fold higher than that observed in humans (Wang and Kirkness 2005). Although phenotypic associations with retroCNVs are rare, SINE and LINE-1 insertions are a significant source of phenotypic variability in dogs, being responsible for 10% of the phenotype-associated variants identified to date (Online Mendelian Inheritance in Animals [OMIA] Sydney School of Veterinary Science, March 29, 2022; https://omia.org/), whereas in humans, transposable elements are responsible for only 0.27% of all disease mutations (Callinan and Batzer 2006). In a recent analysis of the CanFam3.1 and UMICH_Zoey_3.1 reference assemblies, 16,221 dimorphic SINEs and 1121 dimorphic LINE-1s were identified, which represented a 17-fold increase in SINE differences and an eightfold increase in LINE-1 differences compared with the number found in humans (Halo et al. 2021). Domestic dogs, at 54.1 retroCNVs on average, also have a 13-fold increase in retroCNVs relative to humans, data that are consistent with either a highly active or a highly promiscuous LINE-1 in dogs. We also identified 231 retroCNV parent genes that have no known retrocopies in any other mammalian species, which might indicate that canine LINE-1 proteins are less selective. Among these retroCNV parent genes was *GAP43*, which had 18 retroCNV insertions and which may have implications in cognitive function in dogs (Routtenberg et al. 2000).

In this study, we only considered retroCNVs with identifiable insertion sites as valid, although we provide evidence for the presence of 125 additional retroCNVs with unresolved insertion sites. Our method of retroCNV parent gene discovery also cannot identify single-exon genes or genes with recent retrocopies present in the CanFam3.1 reference assembly. Still, we estimated the rate of retroCNV formation in domestic dogs at around four retroCNVs per 100 yr, which is still an underestimate as our data set of 228 breed dogs does not capture every retroCNV in the entire population. A recent analysis of retroCNVs in mice estimated their rate of retroCNV formation at two per 100 yr (Zhang et al. 2021). This would indicate that domestic dogs have a rate of retroCNV formation even 2× greater than mice, which are known to have a large number of active LINE-1s (Penzkofer et al. 2016) and have been shown to have fivefold as many retroCNVs as humans (Ewing et al. 2013). In natural populations of mice, however, retroCNVs were also shown to be under negative selection owing to deleterious effects, in which retroCNVs are quickly removed from the population (Zhang et al. 2021). Also, in human populations, retroCNV insertions are not found in evolutionarily conserved regions, which indicates that highly deleterious retroCNVs are under negative selection (Zhang et al. 2017). We observed that many retroCNVs appear to be under positive selection in dogs, with significantly elevated $F_{ST}$ values. Like the *FGF4* retrogenes, other retroCNVs may be under positive selection by breeders owing to their phenotypic effects (Batcher et al. 2019), although they may also be neutral variants whose frequencies differ owing to the dynamics of breed formation and low genetic diversity within breeds (Bannasch et al. 2021).

Although gene retrocopies were historically considered nonfunctional pseudogenes, more recently, it has been recognized

that retrocopies, which are a form of structural variant, are often functional through a variety of mechanisms that have been explored in recent reviews (Cheetham et al. 2020; Ciomborowska-Basheer et al. 2021). We found evidence for expression in 14 out of 113 retroCNVs in testes tissue from six Golden Retrievers, including four novel chimeric transcripts. However, this is likely an underestimate owing to our use of a single tissue type for transcriptional assessment. Additionally, many of the retroCNVs were indistinguishable from their parent genes and thus cannot be effectively queried for evidence of expression through RNA-seq analysis alone. Comparison of overall gene expression in larger RNA-seq data sets including individuals with and without specific retroCNVs may be required to determine expression of retroCNVs that are identical to their parent genes. When such analyses were performed in mice, differences in retroCNV parent gene expression were found between individuals with and without the respective retroCNVs, including many cases in which overall expression was significantly reduced in the individuals with the retroCNVs (Zhang et al. 2021). Our data confirms that the *AKR1B1* reference retroCNV, which has previously been shown to be expressed and associated with dog domestication, is common in dogs and rare in wolves (Wang et al. 2019). Similarly, the *MGST3* and *RPL27A* reference retroCNVs and the expressed *FARSBL1* retroCNV are both common across dogs and rare in wild canids and may play roles in domestication. *AKR1B1*, *MGST3*, and *FARSB* are also involved in metabolism (Jakobsson et al. 1997; Crosas et al. 2003), whereas the *RPL27A* retrogene is inserted within *NCOA3*, a coactivator of transcription involved in thyroid function (Nolan et al. 2021).

The four Y Chromosome retroCNVs were present in all male canids and thus were not true retroCNVs, and their sequence deviation from their parent genes of origin indicates that they are likely ancestral. Two of the Y Chromosome retroCNVs had been previously identified through the identification of autosomal variants that were actually sex-linked and owing to the retroCNV insertion (Tsai et al. 2019). In this study, we also identified variants at the retroCNV parent gene loci, which are likely attributable to nonreference retroCNVs. These variants would normally be attributed to variation in the parent genes, potentially hindering any analyses looking for causative mutations; it is noteworthy that some variants identified in the coding sequences of genes may in fact be attributable to nonreference retroCNVs elsewhere in the genome.

Domesticated dog breeds have undergone artificial selection, which has led to extreme phenotypic diversity between breeds as well as breed predispositions to many heritable disorders (Wayne and Ostrander 2007; Asher et al. 2009; Summers et al. 2010; Gough et al. 2018; Bannasch et al. 2021). In particular, many dog breeds are at a substantially higher risk for specific cancers compared with human populations (Schiffman and Breen 2015). We have shown that dogs have many retroCNVs, consistent with a highly active LINE-1, and we also observed that some retroCNVs are under selection or are also capable of expression or insertional mutagenesis through the formation of novel chimeric transcripts with nearby genes. These functional retroCNVs are likely a contributing factor to the phenotypic diversity seen in canids and are also strong candidates for disease associations in dogs, including susceptibility to cancers. We hope this list of retroCNV insertion sites will be a useful resource for the canine research community and that further assessment of the retroCNVs for evidence of function will provide insight into the genetics of phenotypic traits under selection in dogs.

## Methods

### Data selection

Illumina sequencing data aligned to the CanFam3.1 reference were downloaded from the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) (Hoeppner et al. 2014). The data set included 227 dogs from 76 different breeds (referred to as "breed dogs"), 43 free-ranging dogs (marked as either "village" or "indigenous" by the data provider), three dingoes, 10 gray wolves, two red wolves, five coyotes, and three African wild dogs. A full list of samples used in this study and their accession numbers is available (Supplemental Table S1). Ten samples were removed from the analysis owing to a low number of retroCNV insertion sites being resolved, possibly owing to the quality of the sequencing data.

### Parent gene discovery using mRNA-specific 30-mers

The nucleotide sequence of a gene retrocopy resembles the processed mRNA transcript of its parent gene, with unique nucleotide sequences formed at the exon–exon junctions. These unique nucleotide sequences are only observed in genomic DNA when a retrocopy insertion is present and thus can be used to identify the parent genes of retroCNVs from WGS data. We used Gffread (Pertea and Pertea 2020) to obtain the spliced gene sequences for each gene transcript found in the NCBI CanFam3.1 annotation release 105 (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Canis_lupus_familiaris/105/). For each spliced gene sequence, we created a set of mRNA-specific 30-mers that are absent from the CanFam3.1 reference assembly using Jellyfish count (Marçais and Kingsford 2011). This identified all unique 30-mer sequences within the transcriptome that are absent from CanFam3.1 owing to intron removal, giving a maximum of 29 mRNA-specific 30-mers per exon–exon junction. We attributed mRNA-specific 30-mers from alternatively spliced gene transcripts to their respective parent gene. To reduce false positives owing to sequencing errors, we filtered out 30-mers with an edit distance of two substitutions from the reference genome using mrsFAST (Hach et al. 2014). We removed any gene that had fewer than five mRNA-specific 30-mers from further analysis, which included 1574 single-exon genes and 75 genes with recent retrocopies in the CanFam3.1 assembly. In total, 5,884,280 30-mers from 30,792 genes (median per gene, 106) were retained for retrocopy parent gene discovery. WGS data were then queried for the presence of the mRNA-specific 30-mers using Jellyfish. Genes that had at least five mRNA-specific 30-mers and at least 10% of the total 30-mers for that gene identified were considered as putative retroCNV parent genes for further analysis.

### retroCNV insertion site discovery through discordant read analysis

Gene retrocopies are derived from processed mRNA transcripts and thus lack introns. When Illumina paired-end reads containing retroCNV sequences are aligned to a reference genome, they align to the parent gene locus, resulting in discordant read pairs that align only to the exons of the parent gene. Discordant read pairs can also be observed at the 5′ and 3′ ends of the parent gene, wherein one read aligns at the parent gene loci, and the other read aligns at the insertion site elsewhere in the genome. Discordant reads can thus be used to verify the presence of a retroCNV as well as identify the insertion site. We performed discordant read analysis on aligned WGS data using TEBreak (Carreira et al. 2016). The "--disc_only" option was used to obtain a list of discordant read clusters of at least four reads ("--min_disc_reads 4") mapping from a putative parent gene to elsewhere in the

genome via "--disco_target." Putative insertion sites for retroCNVs were visually confirmed in the Integrative Genomics Viewer (IGV) (Supplemental Fig. S8; Thorvaldsdóttir et al. 2013). A retrocopy insertion site was considered valid if discordant reads were observed mapping to the same genomic locus from both the 3′ and 5′ end of the parent gene or if discordant reads were found mapping from either the 3′ or 5′ as well as exon–exon discordant reads at the parent gene. Any discordant reads mapping from parent genes to known CanFam3.1 reference retrocopy insertion sites were ignored, as reference retroCNVs were analyzed separately. The TSD sequence was identified as the overlap between forward and reverse discordant reads at the insertion site. The 5′ and 3′ junction sequences for the retroCNV insertions were resolved using TEBreak and are available in Supplemental Table S15. Visual representations of retroCNV insertion sites were produced using Circos (Krzywinski et al. 2009). As has been proposed by Cheetham et al. (2020), we chose, rather than "pseudogene," a term that does not make functional inferences for the retrocopies; for example, retrocopies of the *FGF4* gene were labeled *FGF4L1*, *FGF4L2*, etc.

### CanFam3.1 reference assembly retroCNVs

We also examined retrocopies present in the CanFam3.1 assembly for evidence of being retroCNVs; an individual lacking a CanFam3.1 reference retrocopy would appear to have a deletion at those loci when aligned to CanFam3.1. A list of reference retrocopy locations was downloaded from RetrogeneDB (Rosikiewicz et al. 2017). The data set of 293 canids previously aligned to CanFam3.1 was analyzed using DELLY with default settings to identify structural variants within 1 kb of reference retrocopies (Rausch et al. 2012). All deletions were confirmed visually in IGV. Additionally, aligned sequence data from a gray wolf and a coyote were visually analyzed in IGV at all recent canine RetrogeneDB retrocopy loci (>95% identity with parent gene) to identify any retroCNVs that may have gone undetected by DELLY.

### retroCNV-specific variant identification

Sequence variants between a retroCNV and its parent gene sequence are owing either to germline variants present within the parent gene or to new polymorphisms unique to the retroCNV that occurred after insertion, which we refer to as retroCNV-specific variants. As all retroCNV-derived reads align to the parent gene loci, we analyzed variants at the parent gene loci in order to identify retroCNV-specific variants. We first identified variants at the retroCNV parent gene loci using BCFtools mpileup (Danecek et al. 2021). We then compared variant allele frequencies between individuals positive or negative for each retroCNV, and variants that only appeared in individuals with the retroCNV were considered unique to the retroCNV. RetroCNVs that were unique to wild canids were excluded from this analysis as the wild canids contained many unique variants that could not be easily differentiated between variation within the parent genes or the retroCNV sequences. retroCNVs that had multiple insertions from the same parent gene were also excluded. Variant effect prediction was performed using the UCSC Genome Browser Variant Annotation Integrator tool (Hinrichs et al. 2016).

### retroCNV validation

We performed retroCNV discovery using Illumina data aligned to CanFam3.1 on four individuals that were previously used to generate additional dog genome assemblies: UMICH_Zoey_3.1 (Halo et al. 2021), UU_Cfam_GSD_1.0 (Wang et al. 2021), Canfam_GSD (Field et al. 2020), and CanLup_DDS (Field et al. 2022). We then assessed the presence of the retroCNVs within their respective assemblies using BLAST (Madden 2013).

PacBio data were also examined for evidence of the retroCNV insertions. Individual long-read FASTQ files were aligned to the CanFam3.1 reference with minimap2 version 2.17 (Li 2018). Alignment files were sorted, merged, and indexed with SAMtools version 1.5 (Danecek et al. 2021). The predicted retroCNV insertion sites ±100 bp were analyzed using a modified version of a pipeline designed to detect LINE-1 insertions in long-read sequenced genomes. The pipeline extracts the raw long reads, which align to a locus of interest, and uses a combination of Canu and wtdbg2 to assemble the reads into contigs (Koren et al. 2017; Ruan and Li 2020). The contigs are then polished using Racon (Vaser et al. 2017), aligned to the reference using minimap2 version 2.20, and put in orientation with the reference. Precise breakpoints were identified using AGE (Abyzov and Gerstein 2011).

We developed three primer PCR assays for retroCNVs using Primer3 software (Untergasser et al. 2012), with forward and reverse primers flanking the insertion site and internal primers at the 5′ or 3′ ends of the parent gene. A panel of 10 dogs from a breed identified as carrying each retroCNV were selected at random from the Bannasch laboratory DNA repository for testing (Batcher et al. 2019). A list of the primers used in this study and their expected product sizes is available in Supplemental Table S16. Sanger sequencing was performed on an Applied Biosystems 3500 genetic analyzer using a BigDye terminator sequencing kit (Applied Biosystems).

### Population analysis

For $F_{ST}$ calculations, dog breeds were placed into the multibreed clades identified by Parker et al. (2017). Only clades containing individuals from at least three breeds were included in this analysis. Additionally, only three Golden Retrievers were selected at random to include in the retriever clade. $F_{ST}$ between clades was calculated as described by Zhang et al. (2017), including the calculation of a null distribution from 1000 fake population sets generated through shuffling individual labels for significance estimates. retroCNVs for which 1000 fake population sets never produced an equal or higher $F_{ST}$ than the real population were considered significant.

### WGS and RNA-seq

Adult Golden Retriever testes were obtained from routine castration procedures. Tissue samples were flash-frozen in liquid nitrogen and stored at −80°C. Genomic DNA was extracted using a Gentra Puregene DNA extraction kit (Qiagen), and RNA was extracted using an RNeasy fibrous tissue mini kit (Qiagen). Library preparation and NovaSeq S4 Illumina paired-end sequencing were performed at the UC Davis Genome Center. Reads were aligned to the CanFam3.1 reference assembly using minimap2 (Li 2018). PCR duplicate reads were removed, and the aligned files were sorted and indexed using SAMtools (Danecek et al. 2021). Evidence for chimeric transcripts in the RNA-seq data set was found through visual analysis of the retroCNV insertion sites and nearby genes in IGV. Because of the small sample size and heterogeneity of retroCNVs between individuals, evidence of expression was determined visually through examination of the insert site, the 5′ UTR of the parent gene, and insertion site genes for chimeric transcripts. For retroCNVs that contained any retroCNV-specific variants, the parent gene loci were examined for evidence of the retroCNV-specific variant, which would indicate expression of the retroCNV. A minimum of two discordant reads was used to consider a retroCNV expressed.

### Human comparative analysis

We performed retroCNV discovery in a subset of individuals from The 1000 Genomes Project Consortium high-coverage phase 3 data set (Sudmant et al. 2015; The 1000 Genomes Project Consortium 2015). We selected three individuals from each of 26 human populations at random for this analysis (Supplemental Table S8) and used the GRCh38 annotation release 109.20210514 for 30-mer construction. We then performed retroCNV discovery in the same manner as was performed on the canid data set, and compared the retroCNVs identified by our approach to those identified in a previous study that used the same individuals (Zhang et al. 2017). We compared the retroCNVs identified in this study to those retroCNVs identified in the same 78 individuals by Zhang et al. as well as retroCNVs identified in four other studies that used different data sets (Abyzov et al. 2013; Ewing et al. 2013; Schrider et al. 2013; Feng and Li 2021).

## Data access

The WGS and RNA-seq data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA776905. The source code is available as Supplemental Code. The retroCNV insertion sites in bigBed format are available in the Supplemental Material, and at a track hub for the UCSC Genome Browser available at GitHub (https://github.com/klbatcher/retroCNV_insertions).

## Competing interest statement

The authors declare no competing interests.

## Acknowledgments

We thank Laura Kirby for helpful discussions related to the use of long-read data to detect retrogene insertions. We thank Guide Dogs for the Blind, Inc., San Rafael, California, for the samples. This work was supported in part by the Lodric Maddox Graduate Fellowship, Maxine Adler Endowed Chair funds, and the Center for Companion Animal Health. Additional financial support was provided by the Students Training in Advanced Research (STAR) program through a UC Davis School of Veterinary Medicine Endowment Fund. M.B. and J.M.K. were supported by the National Institute of General Medical Sciences of the National Institutes of Health under award numbers R01GM140135 and T32GM007544.

## References

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526:** 68–74. doi:10.1038/nature15393

Abyzov A, Gerstein M. 2011. AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics* **27:** 595–603. doi:10.1093/bioinformatics/btq713

Abyzov A, Iskow R, Gokcumen O, Radke DW, Balasubramanian S, Pei B, Habegger L, Lee C, The 1000 Genomes Project Consortium, Gerstein M. 2013. Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. *Genome Res* **23:** 2042–2052. doi:10.1101/gr.154625.113

Asher L, Diesel G, Summers JF, McGreevy PD, Collins LM. 2009. Inherited defects in pedigree dogs. Part 1: disorders related to breed standards. *Vet J* **182:** 402–411. doi:10.1016/j.tvjl.2009.08.033

Bannasch D, Famula T, Donner J, Anderson H, Honkanen L, Batcher K, Safra N, Thomasy S, Rebhun R. 2021. The effect of inbreeding, body size and morphology on health in dog breeds. *Canine Med Genet* **8:** 12. doi:10.1186/s40575-021-00111-4

Batcher K, Dickinson P, Giuffrida M, Sturges B, Vernau K, Knipe M, Rasouliha SH, Drögemüller C, Leeb T, Maciejczyk K, et al. 2019. Phenotypic effects of *FGF4* retrogenes on intervertebral disc disease in dogs. *Genes (Basel)* **10:** 435. doi:10.3390/genes10060435

Batcher K, Dickinson P, Maciejczyk K, Brzeski K, Rasouliha SH, Letko A, Drögemüller C, Leeb T, Bannasch D. 2020. Multiple *FGF4* retrocopies recently derived within canids. *Genes (Basel)* **11:** 839. doi:10.3390/genes11080839

Brown EA, Dickinson PJ, Mansour T, Sturges BK, Aguilar M, Young AE, Korff C, Lind J, Ettinger CL, Varon S, et al. 2017. *FGF4* retrogene on CFA12 is responsible for chondrodystrophy and intervertebral disc disease in dogs. *Proc Natl Acad Sci* **114:** 11476–11481. doi:10.1073/pnas.1709082114

Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. 2021. High coverage whole genome sequencing of the expanded 1000 genomes project cohort including 602 trios. bioRxiv doi:10.1101/2021.02.06.430068

Callinan P, Batzer M. 2006. Retrotransposable elements and human disease. *Genome Dyn* **1:** 104–115. doi:10.1159/000092503

Carreira PE, Ewing AD, Li G, Schauer SN, Upton KR, Fagg AC, Morell S, Kindlova M, Gerdes P, Richardson SR, et al. 2016. Evidence for L1-associated DNA rearrangements and negligible L1 retrotransposition in glioblastoma multiforme. *Mob DNA* **7:** 21. doi:10.1186/s13100-016-0076-6

Casola C, Betrán E. 2017. The genomic impact of gene retrocopies: What have we learned from comparative genomics, population genomics, and transcriptomic analyses? *Genome Biol Evol* **9:** 1351–1373. doi:10.1093/gbe/evx081

Cheetham SW, Faulkner GJ, Dinger ME. 2020. Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nat Rev Genet* **21:** 191–201. doi:10.1038/s41576-019-0196-1

Ciomborowska-Basheer J, Staszak K, Kubiak MR, Makałowska I. 2021. Not so dead genes: retrocopies as regulators of their disease-related progenitors and hosts. *Cells* **10:** 912. doi:10.3390/cells10040912

Crosas B, Hyndman DJ, Gallego O, Martras S, Parés X, Flynn TG, Farrés J. 2003. Human aldose reductase and human small intestine aldose reductase are efficient retinal reductases: consequences for retinoid metabolism. *Biochem J* **373:** 973–979. doi:10.1042/bj20021818

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* **10:** giab008. doi:10.1093/gigascience/giab008

Esnault C, Maestre J, Heidmann T. 2000. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet* **24:** 363–367. doi:10.1038/74184

Ewing AD, Ballinger TJ, Earl D, Harris CC, Ding L, Wilson RK, Haussler D. 2013. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biol* **14:** R22. doi:10.1186/gb-2013-14-3-r22

Feng X, Li H. 2021. Higher rates of processed pseudogene acquisition in humans and three great apes revealed by long-read assemblies. *Mol Biol Evol* **38:** 2958–2966. doi:10.1093/molbev/msab062

Field MA, Rosen BD, Dudchenko O, Chan EK, Minoche AE, Edwards RJ, Barton K, Lyons RJ, Tuipulotu DE, Hayes VM, et al. 2020. Canfam_GSD: *de novo* chromosome-length genome assembly of the German shepherd dog (*Canis lupus familiaris*) using a combination of long reads, optical mapping, and Hi-C. *Gigascience* **9:** giaa027. doi:10.1093/gigascience/giaa027

Field MA, Yadav S, Dudchenko O, Esvaran M, Rosen BD, Skvortsova K, Edwards RJ, Keilwagen J, Cochran BJ, Manandhar B, et al. 2022. The Australian dingo is an early offshoot of modern breed dogs. *Sci Adv* **8:** eabm5944. doi:10.1126/sciadv.abm5944

Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. 2019. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47:** D766–D773. doi:10.1093/nar/gky955

Gao X, Li Y, Adetula AA, Wu Y, Chen H. 2019. Analysis of new retrogenes provides insight into dog adaptive evolution. *Ecol Evol* **9:** 11185–11197. doi:10.1002/ece3.5620

Gough A, Thomas A, O'Neill D. 2018. *Breed predispositions to disease in dogs and cats*. Wiley, Hoboken, NJ.

Hach F, Sarrafi I, Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. 2014. mrsFAST-Ultra: a compact, SNP-aware mapper for high performance sequencing applications. *Nucleic Acids Res* **42:** W494–W500. doi:10.1093/nar/gku370

Halo JV, Pendleton AL, Shen F, Doucet AJ, Derrien T, Hitte C, Kirby LE, Myers B, Sliwerska E, Emery S, et al. 2021. Long-read assembly of a Great Dane genome highlights the contribution of GC-rich sequence and mobile elements to canine genomes. *Proc Natl Acad Sci* **118:** e2016274118. doi:10.1073/pnas.2016274118

Hinrichs AS, Raney BJ, Speir ML, Rhead B, Casper J, Karolchik D, Kuhn RM, Rosenbloom KR, Zweig AS, Haussler D, et al. 2016. UCSC Data Integrator

and Variant Annotation Integrator. *Bioinformatics* **32:** 1430–1432. doi:10.1093/bioinformatics/btv766
Hoeppner MP, Lundquist A, Pirun M, Meadows JR, Zamani N, Johnson J, Sundström G, Cook A, FitzGerald MG, Swofford R, et al. 2014. An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. *PLoS One* **9:** e91172. doi:10.1371/journal.pone.0091172
Jakobsson P-J, Mancini JA, Riendeau D, Ford-Hutchinson AW. 1997. Identification and characterization of a novel microsomal enzyme with glutathione-dependent transferase and peroxidase activities. *J Biol Chem* **272:** 22934–22939. doi:10.1074/jbc.272.36.22934
Jurka J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci* **94:** 1872–1877. doi:10.1073/pnas.94.5.1872
Kabza M, Kubiak MR, Danek A, Rosikiewicz W, Deorowicz S, Polański A, Makałowska I. 2015. Inter-population differences in retrogene loss and expression in humans. *PLoS Genet* **11:** e1005579. doi:10.1371/journal.pgen.1005579
Kim S, Mun S, Kim T, Lee K-H, Kang K, Cho J-Y, Han K. 2019. Transposable element-mediated structural variation analysis in dog breeds using whole-genome sequencing. *Mamm Genome* **30:** 289–300. doi:10.1007/s00335-019-09812-5
Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res* **27:** 722–736. doi:10.1101/gr.215087.116
Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19:** 1639–1645. doi:10.1101/gr.092759.109
Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34:** 3094–3100. doi:10.1093/bioinformatics/bty191
Madden T. 2013. The BLAST sequence analysis tool. In *The NCBI Handbook*, 2nd ed. (ed. McEntyre J, Ostell J), pp. 425–436. National Center for Biotechnology Information, Bethesda, MD. https://www.ncbi.nlm.nih.gov/books/NBK148670/.
Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27:** 764–770. doi:10.1093/bioinformatics/btr011
Nolan J, Campbell PJ, Brown SJ, Zhu G, Gordon S, Lim EM, Joseph J, Cross SM, Panicker V, Medland SE, et al. 2021. Genome-wide analysis of thyroid function in Australian adolescents highlights SERPINA7 and NCOA3. *Eur J Endocrinol* **185:** 743–753. doi:10.1530/EJE-21-0614
Ostrander EA, Wang G-D, Larson G, Vonholdt BM, Davis BW, Jagannathan V, Hitte C, Wayne RK, Zhang Y-P, et al. 2019. Dog10K: an international sequencing effort to advance studies of canine domestication, phenotypes and health. *Natl Sci Rev* **6:** 810–824. doi:10.1093/nsr/nwz049
Parker HG, VonHoldt BM, Quignon P, Margulies EH, Shao S, Mosher DS, Spady TC, Elkahloun A, Cargill M, Jones PG, et al. 2009. An expressed *fgf4* retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science* **325:** 995–998. doi:10.1126/science.1173275
Parker HG, Dreger DL, Rimbault M, Davis BW, Mullen AB, Carpintero-Ramirez G, Ostrander EA. 2017. Genomic analyses reveal the influence of geographic origin, migration, and hybridization on modern dog breed development. *Cell Rep* **19:** 697–708. doi:10.1016/j.celrep.2017.03.079
Penzkofer T, Jäger M, Figlerowicz M, Badge R, Mundlos S, Robinson PN, Zemojtel T. 2016. L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. *Nucleic Acids Res* **45:** D68–D73. doi:10.1093/nar/gkw925
Pertea G, Pertea M. 2020. GFF utilities: GffRead and GffCompare. *F1000Res* **9:** 304. doi:10.12688/f1000research.23297.1
Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28:** i333–i339. doi:10.1093/bioinformatics/bts378
Richardson SR, Salvador-Palomeque C, Faulkner GJ. 2014. Diversity through duplication: whole-genome sequencing reveals novel gene retrocopies in the human population. *Bioessays* **36:** 475–481. doi:10.1002/bies.201300181
Rosikiewicz W, Kabza M, Kosiński JG, Ciomborowska-Basheer J, Kubiak MR, Makałowska I. 2017. RetrogeneDB—a database of plant and animal retrocopies. *Database* **2017:** bax038. doi:10.1093/database/bax038
Routtenberg A, Cantallops I, Zaffuto S, Serrano P, Namgung U. 2000. Enhanced learning after genetic overexpression of a brain growth protein. *Proc Natl Acad Sci* **97:** 7657–7662. doi:10.1073/pnas.97.13.7657
Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* **17:** 155–158. doi:10.1038/s41592-019-0669-3
Schiffman JD, Breen M. 2015. Comparative oncology: what dogs and other species can teach us about humans with cancer. *Philos Trans R Soc Lond B Biol Sci* **370:** 20140231. doi:10.1098/rstb.2014.0231
Schrider DR, Navarro FC, Galante PA, Parmigiani RB, Camargo AA, Hahn MW, de Souza SJ. 2013. Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS Genet* **9:** e1003242. doi:10.1371/journal.pgen.1003242
Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH-Y, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526:** 75–81. doi:10.1038/nature15394
Summers JF, Diesel G, Asher L, McGreevy PD, Collins LM. 2010. Inherited defects in pedigree dogs, part 2: disorders that are not related to breed standards. *Vet J* **183:** 39–45. doi:10.1016/j.tvjl.2009.11.002
Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinformatics* **14:** 178–192. doi:10.1093/bib/bbs017
Troskie R-L, Jafrani Y, Mercer TR, Ewing AD, Faulkner GJ, Cheetham SW. 2021. Long-read cDNA sequencing identifies functional pseudogenes in the human transcriptome. *Genome Biol* **22:** 146. doi:10.1186/s13059-020-02207-9
Tsai KL, Evans JM, Noorai RE, Starr-Moss AN, Clark LA. 2019. Novel Y chromosome retrocopies in canids revealed through a genome-wide association study for sex. *Genes (Basel)* **10:** 320. doi:10.3390/genes10040320
Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3—new capabilities and interfaces. *Nucleic Acids Res* **40:** e115. doi:10.1093/nar/gks596
Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27:** 737–746. doi:10.1101/gr.214270.116
Wang W, Kirkness EF. 2005. Short interspersed elements (SINEs) are a major source of canine genomic diversity. *Genome Res* **15:** 1798–1808. doi:10.1101/gr.3765505
Wang G-D, Shao X-J, Bai B, Wang J, Wang X, Cao X, Liu Y-H, Wang X, Yin T-T, Zhang S-J, et al. 2019. Structural variation during dog domestication: insights from gray wolf and dhole genomes. *Natl Sci Rev* **6:** 110–122. doi:10.1093/nsr/nwy076
Wang C, Wallerman O, Arendt M-L, Sundström E, Karlsson Å, Nordin J, Mäkeläinen S, Pielberg GR, Hanson J, Ohlsson Å, et al. 2021. A novel canine reference genome resolves genomic architecture and uncovers transcript complexity. *Commun Biol* **4:** 185. doi:10.1038/s42003-020-01566-0
Wayne RK, Ostrander EA. 2007. Lessons learned from the dog genome. *Trends Genet* **23:** 557–567. doi:10.1016/j.tig.2007.08.013
Zhang Y, Li S, Abyzov A, Gerstein MB. 2017. Landscape and variation of novel retroduplications in 26 human populations. *PLoS Comput Biol* **13:** e1005567. doi:10.1371/journal.pcbi.1005567
Zhang W, Xie C, Ullrich K, Zhang YE, Tautz D. 2021. The mutational load in natural populations is significantly affected by high primary rates of retroposition. *Proc Natl Acad Sci* **118:** e2013043118. doi:10.1073/pnas.2013043118

*Received April 8, 2022; accepted in revised form July 19, 2022.*