Choi *et al. Mobile DNA* (2025) 16:1
https://doi.org/10.1186/s13100-024-00338-5

Mobile DNA

RESEARCH

Open Access

SINE retrotransposons import polyadenylation signals to 3'UTRs in dog (*Canis familiaris*)

Jessica D. Choi[1,2,3*], Lelani A. Del Pinto[1] and Nathan B. Sutter[1]

Abstract

Background Messenger RNA 3' untranslated regions (3'UTRs) control many aspects of gene expression and determine where the transcript will terminate. The polyadenylation signal (PAS) AAUAAA (AATAAA in DNA) is a key regulator of transcript termination and this hexamer, or a similar sequence, is very frequently found within 30 bp of 3'UTR ends. Short interspersed element (SINE) retrotransposons are found throughout genomes in high copy numbers. When inserted into genes they can disrupt expression, alter splicing, or cause nuclear retention of mRNAs. The genomes of the domestic dog and other carnivores carry hundreds of thousands of Can-SINEs, a tRNA-related SINE with transcription termination potential. Because of this we asked whether Can-SINEs may terminate transcript in some dog genes.

Results Each of the dog's nine Can-SINE consensus sequences carry an average of three AATAAA PASs on their sense strands but zero on their antisense strands. Consistent with the idea that Can-SINEs can terminate transcripts, we find that sense-oriented Can-SINEs are approximately ten times more frequent at 3' ends of 3'UTRs compared to further upstream within 3'UTRs. Furthermore, the count of AATAAA PASs on head-to-tail SINE sequences differs significantly between sense and antisense-oriented retrotransposons in transcripts. Can-SINEs near 3'UTR ends are likely to carry an AATAAA motif on the mRNA sense strand while those further upstream are not. We identified loci where Can-SINE insertion has truncated or altered a 3'UTR of the dog genome (dog 3'UTR) compared to the human ortholog. Dog 3'UTRs have peaks of AATAAA PAS frequency at 28, 32, and 36 bp from the end. The periodicity is partly explained by TAAA(n) repeats within Can-SINE AT-rich tails. We annotated all repeat-masked Can-SINE copies in the Boxer reference genome and found that the young SINEC_Cf type has a mode of 15 bp length for target site duplications (TSDs). All dog Can-SINE types favor integration at TSDs beginning with A(4).

Conclusion Dog Can-SINE retrotransposition has imported AATAAA PASs into gene transcripts and led to alteration of 3'UTRs. AATAAA sequences are selectively removed from Can-SINEs in introns and upstream 3'UTR regions but are retained at the far downstream end of 3'UTRs, which we infer reflects their role as termination sequences for these transcripts.

*Correspondence:
Jessica D. Choi
jaycee.choi@jax.org
[1] Department of Biology, La Sierra University, Riverside, CA, USA
[2] The Jackson Laboratory, Bar Harbor, ME, USA
[3] Graduate School of Biomedical Sciences, Tufts University, Boston, MA, USA

Background

A gene's 3' untranslated region (3'UTR) plays many roles in regulating gene expression [1]. Mutations in 3'UTRs can affect gene expression in part by creating multiple isoforms of the gene, which is common in vertebrates [2]. A 3'UTR also helps signal the end of transcription [3]. The sequence AAUAAA, or a closely related hexamer, is

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

the polyadenylation signal (PAS), a key regulator of termination and polyadenylation [4]. The AAUAAA motif is bound by a ternary complex of cleavage and polyadenylation specificity factor-160, WDR33, and CPSF-30 [5, 6].

Although the AAUAAA motif (5'-AATAAA-3' in the DNA sense strand) is the most commonly used PAS, Beaudoing et al. [7] identified 10 other variants that are also occasionally used. The most frequent PAS position is −15 / −16 bp upstream of the 3' end of the 3'UTR as measured to the near end of the PAS, or 22 bp to the farther end.

Gene expression can be impacted by retrotransposons, and some of them carry PASs. Short interspersed elements (SINEs) and long interspersed elements (LINEs) use a copy-paste method for insertion using an RNA intermediate (Pol III for SINEs, Pol II for LINEs) [8]. They have contributed to the expansion of the genome and provide a "fossil" record of genome change useful for cladistics. This is shown through mammalian-wide interspersed repeats (MIRs), which are old enough to be found in all mammals, while other SINEs, such as *Alu* in humans and Can-SINEs in carnivores are much younger.

The tRNA gene-like 5'region (head) of many SINEs contains internal RNA pol III promoter elements, the A and B boxes [9]. The insertion process of SINE RNA transcripts creates a direct repeat of sequence called the target site duplication (TSD), that flanks the SINE sequence and provides a useful means of identifying the insertion event [10, 11].

SINEs, especially those with TA-rich tail sequences, can provide PASs into transcripts. In eight ENCODE human cell lines, 9.4% of termination signals occur within transposable elements (TEs) [12], and termination signals are more often associated with TEs than are the canonical PASs [13]. Thus, it is implied that TE insertions into genes can provide alternate PASs. Furthermore, the human PASs not conserved in the mouse ortholog are associated with TEs; ~94% of human polyadenylation signals that are TE-associated are not conserved in the mouse [13]. TE insertion can therefore alter transcript termination.

Carnivore genomes carry Can-SINEs [14–16] present in hundreds of thousands of copies in carnivore genomes but absent outside the order [11, 17]. Similar to other T+ category SINEs, they contain sequence motifs with potential for signaling transcription termination [18, 19] such as AATAAA, and the TCT_{3-6} region, which is a transcription terminator sequence. This transcription terminator is also found, for example, in mouse B2 SINEs near the 3' end [20]. Despite being RNA pol III transcribed, some SINE transcripts are polyadenylated, which leads to a longer half-life [21]. It is currently unknown whether Can-SINEs are polyadenylated, but young insertions often have long homopolymer runs of A(n) at the 3' end of the SINE. Can-SINEs are present in both canoidea ("dog like") and feliform carnivores [22] and in every genome have distinct abundance and age profiles as measured by sequence divergence from a consensus [15, 23, 24].

Dogs have nine sub-types of Can-SINEs. SINEC_Cf is the youngest and its insertions often have less than 5% sequence divergence from the consensus. This type appears to still be active today, as many insertions are so recent they are polymorphic [11, 25, 26]. These polymorphic SINEC_Cf insertions have played an important role in trait diversification within dogs; they are either associated with or are the causal component of numerous dog traits/disorders: narcolepsy (intronic near *HCTR2* exon), merle coat pattern (intron-exon boundary in *SILV/PMEL*), black and tan coat (inverted SINEC_Cf repeat in an *ASIP* intron), retinal degeneration (within coding exon in *STK38L*), centronuclear myopathy (within coding exon in *PTPLA*), and retinal atrophy (near exon in *FAM161A*) [27–32].

Given the high copy number of SINE insertions in the dog genome, as well as the potential for transcription termination signaling, we wondered how often Can-SINEs have imported transcription termination signals into genes. Dog Can-SINEs, like other TEs, exhibit a strong orientation bias within transcripts, occurring at much higher frequency in antisense rather than sense orientation within introns [25, 33]. Selection against termination signals present in sense-oriented SINEs could be at least partly responsible. We hypothesize that dog Can-SINEs have imported AATAAA motifs to genes that use them as PASs in 3'UTRs. Here we analyze the orientation bias of Can-SINEs in transcripts and the changing counts of AATAAA motifs within SINEs when the SINE is inserted within 3'UTRs vs. introns or intergenic sequences.

Methods

Obtaining Dog CanFam3.1 Reference Genome Datasets. We downloaded from the UCSC Genome Browser Table Browser the genome-wide Ensembl gene models (version 99), gaps, repeat track SINEs and LINEs, and chromosome start and end bp values for CanFam3.1. Gene models without an open reading frame (e.g. RNA genes) were excluded from the dataset because they are annotated as "UTR" and if left in would confuse signals relevant for 3'UTRs of protein-coding genes. From NCBI we downloaded the assembled chromosome fastas and used them to slice out SINE and 3'UTR sub-sequences. We analyzed the dog autosomes and chrX.

Defining Gene Features. With custom scripts we defined segments within genes such that 3'UTRs could be redundant but all bases within introns could not be

part of any coding or untranslated exon segments. Intergenic sequence was defined as all sequences not present within gene model transcripts and not within the 500 bp upstream of transcription start sites.

Analysis of Gene Feature Intersections with Retrotransposons and Motif Finding. We wrote a Python script to count intersections of chromosome segment objects, find motifs such as AATAAA, and track strand orientations.

Annotation of Can-SINEs. We downloaded the positions for all SINEC repeat-masked SINEs from the table browser of the UCSC genome browser reference for dog, CanFam3.1. A custom Python script retrieved the sequence plus 100 bp flanks for each SINE and attempted to identify identical target site duplications (TSDs) ranging from 0.75 to 1.25 times the length of the consensus duplication for the given SINEC type (see Table 1 for SINE lengths). The longest identical TSD found was kept and we set a minimum length of 6 bp for keeping a SINE copy in our analysis. The sequence contained within the TSDs was considered to be the SINE. To locate A and B boxes, we searched sequences within 10 bp of the consensus position for the box. We permitted up to 5 mismatches from the consensus sequence and the lowest mismatch box (or boxes) was kept. Additionally, we identified the following components within each SINE sequence: the longest A(n) string, the longest string of A or T, and the longest string of {C or T}, allowing at most 5 inclusions of non-{C or T}.

Creation of sequence logos. We uploaded up to 10,000 TSD sequences for determining the sequence logos for Can-SINEs and up to 5000 A and 5000 B boxes sequences for their sequence logo determinations. All sequence logos were created using https://weblogo.berkeley.edu/logo.cgi.

Table 1 Consensus sequence length for SINE_C SINE types and MIRs

SINE Type	Consensus Sequence Length
SINEC_Cf	189
SINEC_Cf2	180
SINEC_Cf3	180
SINEC_a1	179
SINEC_a2	179
SINEC_b1	194
SINEC_b2	194
SINEC_c1	209
SINEC_c2	212
MIR	262
MIR3	207
MIRb	268
MIRc	268

Results

With an aim to better understand possible roles for Can-SINEs in 3'UTRs of the dog genome (dog 3'UTRs), we first confirmed the frequent presence of the AATAAA PAS near the 3' end of dog 3'UTRs. We calculated the proportion of AATAAA and ATTAAA motifs found in sense orientation at each position for all dog 3'UTRs in the reference genome (Fig. 1). Beaudoing et al. [7] show the most frequent location for PAS hexamers in human 3'UTRs. Our hexamer peak in dogs matches their finding at 22 bp, though we count distances from the 5' end of each hexamer, while they count from the 3' end. In addition to the overall peak with mode ~22–24 bp, we find the AATAAA motif is also frequent at 28, 32, and 36 bp from the 3' end of the 3'UTR. The other Beaudoing hexamers such as ATTAAA have at most only a small increase in frequency at ~22 bp in the dog (Table S1).

SINE and LINE retrotransposons exist throughout the dog genome at high copy numbers and in introns have an antisense orientation bias [25, 34]. We find the same bias in 3'UTRs when looking far (>225bp) from their 3' ends (Fig. 2). However, we found something very different at the 3' ends of 3'UTRs: sense-oriented Can-SINEs are much more frequent than antisense (Fig. 2a). We therefore hypothesized that Can-SINE PAS motifs can be incorporated into host gene 3'UTR ends.

Dog SINE consensus sequences carry PASs (Fig. 3 and Table S2). When counting on the sense strand from SINE head to tail, all Can-SINE consensus sequences contain at least three PASs clustered in their AT-rich tails. These tails typically contain some variant of AATA(n); SINEC_Cf, the youngest Can-SINE type, has four PASs within its tail (Fig. 3 and Table S2). Dog Can-SINEs have a polypyrimidine repeat as well as the PAS motifs. The SINEC_Cf2, a2, b1, b2, and c2 consensus sequences also have a TCTTT motif for RNA pol III termination that directly follows the PAS. Notably, among the Can-SINE subtypes only SINEC_c1 and SINEC_c2 contain any antisense-oriented PASs in their consensus sequences and in both cases the hexamer, which is located in the head of the SINE at ~41 bp, is not an AATAAA motif. Thus, the sense vs. antisense strands of dog Can-SINEs may have starkly different potential to signal termination of RNA polymerase.

To better understand variation within the hundreds of thousands of SINE copies in the dog genome, we collected all repeatMasked copies for the nine Can-SINE sub-types from the repeats track in the boxer reference genome CanFam3.1 (Fig. S1). Can-SINEs sub-types

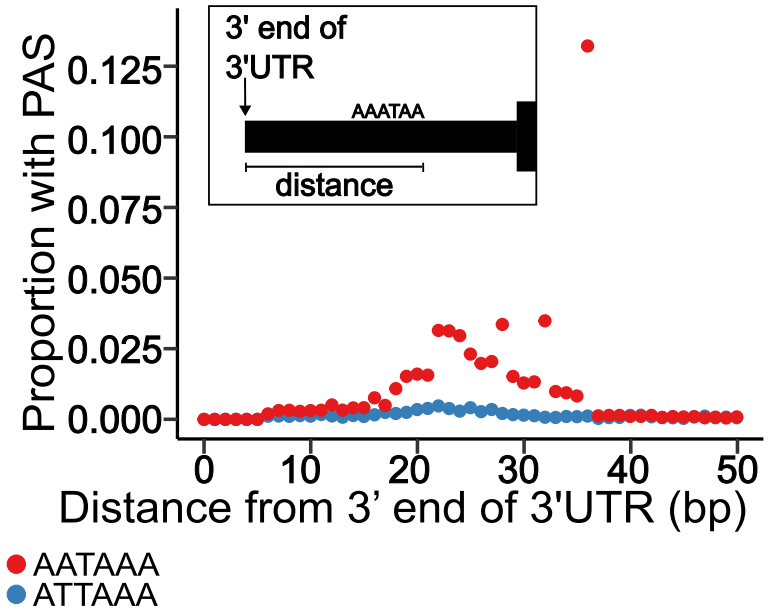

Fig. 1 PAS frequency in dog 3'UTRs. For each bp distance from the 3' end of the 3'UTR we calculate the proportion of PASs present, calculated from the 5'-most base in each PAS hexamer

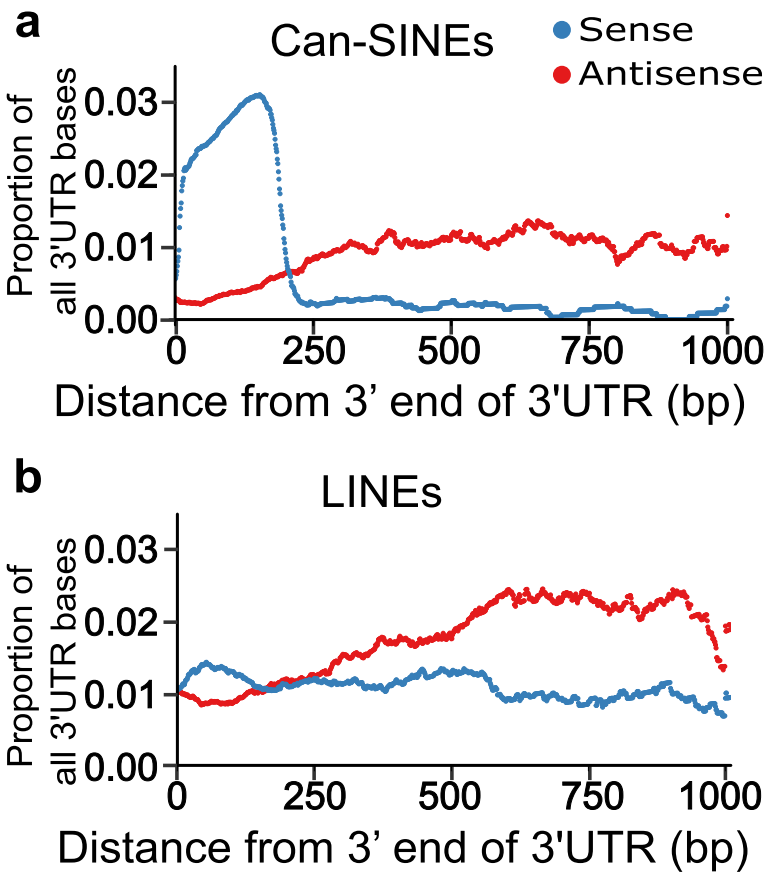

Fig. 2 Density at the 3' end of dog 3'UTRs for (**a**) Can-SINEs and (**b**) LINEs. At each base position, counting from the 3' end, the proportion of all 3'UTRs having a base with that retrotransposon orientation is indicated

SINE Type	SINE Consensus Sequence Tail	Sense	Antisense
SINEC_Cf	actatcataaataaataaaaattaaaaaaaaaaaa	4	0
SINEC_Cf2	tctcatgaataaataaataaaatctttaaaaaaaa	3	0
SINEC_Cf3	tctcatgaataaataaataaaatcttaaaaaaaaa	3	0
SINEC_a1	tctcatgaataaataaataaaatcttaaaaaaaaa	3	0
SINEC_a2	tctcatgaataaataaataaaatctttaaaaaaaa	3	0
SINEC_b1	tctctcaaataaataaataaaatctttaaaaaaaa	3	0
SINEC_b2	tctctcaaataaataaataaaatctttaaaaaaaa	3	0
SINEC_c1	ctctctctcaaaaaaaataaataaattaaaaaaaa	3	1
SINEC_c2	tctctcaaataaataaataaaatctttaaaaaaaa	3	1
MIR	acatagtaagcgctcaataaatgttggttattatt	3	0
MIR3	ccttccagctctgacattctatgattctatgattc	0	0
MIRb	acagtaagcgctcaataaatggtagctctattatt	2	0
MIRc	aaagtgctatacaaatgtaaggggttattattatt	0	0

Fig. 3 Can-SINE consensus sequences have multiple PASs in their sense (head to tail) strand. All 11 PAS sequences from Beaudoing et al, 2000, are included in the count. Counts cover the entire consensus sequence. Most Can-SINE PASs are "AATAAA" and are clustered in the A/T rich tail. Only the 3'-most 35 bp of each consensus are shown

all have a similar structure (Fig. 4a) and similar lengths in the range of approximately 160–220 bp (Fig. 4b, f). The AT-rich tails are longer in SINEC_Cf (Fig. 4c) than the older sub-type SINEC_a1 (Fig. 4g). We analyzed each element to identify Can-SINEs flanked by identical sequences that we designated as putative target site duplications (TSDs). Of the 171,386 total SINEC_Cf copies in the reference, 124,901 (73%) are flanked by identical sequences between 10–20 bp (inclusive) in length. Of these, 15 bp-long sequences are the mode, found in 26,055 SINEC_Cf copies (21%) (Fig. 4d).

In contrast to the relatively young SINEC_Cf type, the older dog Can-SINEs types have higher mean divergence from consensus sequences. This can be readily seen, for example, in the TSD distribution for SINEC_a1 (Fig. 4h) where only a shoulder on the curve is evident at 15 bp. For most SINEC_a1 copies, one or both of the duplicated target sites have evidently mutated away from the single, original sequence. In fact, the differing ages of the Can-SINE types are apparent by comparing the distributions of the TSDs in the relatively younger SINEC_Cf2 and SINEC_Cf3 (Fig. S1a, b), where identical TSD peaks are evident, vs. older types SINEC_c1 and SINEC_c2 that have very few 15 bp identical TSDs (Fig. S1).

The SINEC_Cf copies with identical flanking sequences 6 bp or greater (the "full set"; putative TSDs) have a mean length of 192 bp. This is the set used for Fig. 4a. Our data support the idea that the 15 bp TSD SINEC_Cfs are the youngest subset of Can-SINE: they have A and B boxes with Levenstein edit distances to the consensus of 0.63 and 0.57, respectively, while in the full set these distances are greater, at 1.01 and 1.02 respectively. Plus, while 90% of SINE copies in the full set have a detectable A box and 92% have a detectable B box (Fig. 4e), these figures rise to 98% and 99%, respectively, for the 15 bp TSD set. Finally, the longest string of A or T (the tail) is an average of 25.1 bp long in the full set (Fig. 4c) but slightly longer, 27.2 bp, in the 15-bp TSD subset.

For all Can-SINEs in the dog, the overriding sequence feature for the TSDs is an initial run of polyA (Fig. 4; Fig. S2). The A and B box sequences are very similar across the Can-SINE types, with just two positions (bases 5 and 9, G>T and C>T) varying in the A box between SINEC_Cf vs. SINEC_a1 (Fig. 4k, l, n, o).

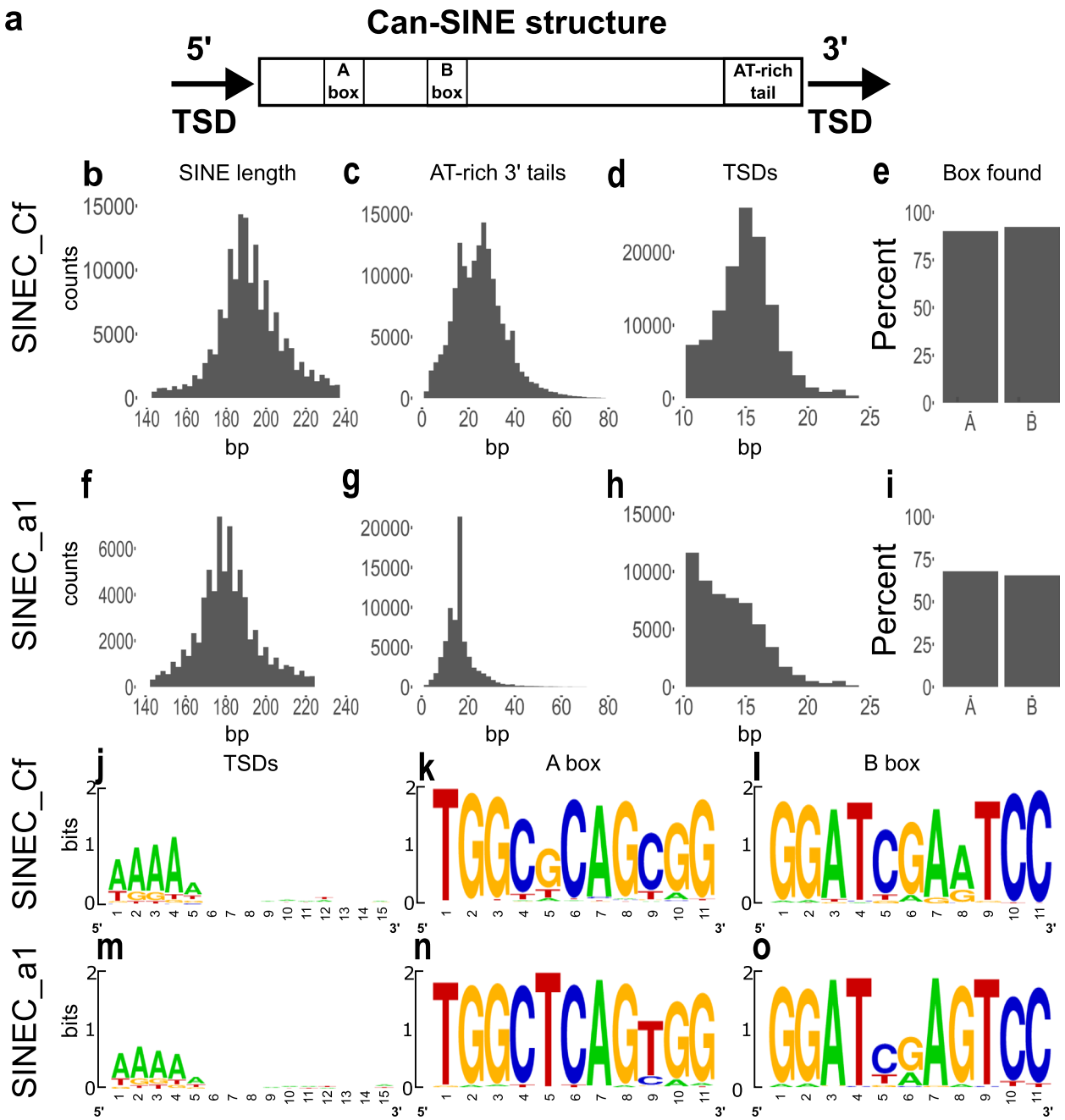

Fig. 4 Annotation of Can-SINE sub-types SINEC_Cf (rows 1 and 3: **b**-**e**, **j**-**l**) and SINEC_a1 (rows 2 and 4: **f**-**i**, **m**-**o**). Panel **a** is a schematic of the Can-SINE structure. In panels **b**-**d** and **f**-**h** the y-axis is the count of elements and the x-axis is bp. Length histograms of: (**b**, **f**) the reference genome's SINEs, (**c**, **g**) SINE A/T-rich 3' tails, (**d**, **h**) and target site duplications. **e**, **i** Barplots showing the percentage of SINE copies in which the A or B box internal promoter elements were identified. Compared to SINEC_a1 (row 2), SINEC_Cf (row 1) has a higher mean length, higher proportion of long putative target site duplications, and a higher proportion of A and B boxes identified. Sequence logos of: (**j**, **m**) target site duplications, (**k**, **n**) A box, and (**l**, **o**) B box

Given both the presence of PAS sequences in dog Can-SINE sense strands as well as the density patterns of Can-SINEs in 3'UTRs, we hypothesized that PASs in Can-SINEs are selected for or against depending on each Can-SINE's orientation and position within gene transcripts. To assess this, we categorized Can-SINE insertions into four genomic locations (ignoring SINEs falling into other locations): intergenic, intronic (any position

within an intron) or "near" or "far" from the 3' end of a 3'UTR (Fig. 5a). To be categorized as "near" a Can-SINE must occur entirely within 225 bp of the 3' end of the 3'UTR. "Far" Can-SINEs are wholly contained within a 3'UTR and none of the SINE's sequence is closer to the 3' end than 225 bp. Can-SINEs spanning across the 225 bp distance were excluded from this analysis. We categorized LINE and MIR insertions in the same way. The LINE subfamilies analyzed are LINE insertions of all types as designated LINE from the RepeatMasker run from the dog CanFam3.1 genome assembly. We then counted the number of AATAAA sequences on the head-to-tail strand of each retrotransposon and looked for differences in count distribution according to whether the retrotransposon was sense (RNA pol encounters the SINE head first) or antisense to the gene transcript.

We find that when Can-SINEs, LINEs, and MIRs are inserted into intergenic regions, there is no difference in the mean AATAAA motif counts between the two strands, as expected (Fig. 5b, f, j). However, the Can-SINEs, LINEs, and MIRs inserted in introns have significantly lower mean AATAAA motif counts on their head-to-tail strands when the retrotransposon is inserted in sense orientation (Fig. 5c, g, k), suggesting that AAT AAA motifs may be selectively removed from sense-oriented retrotransposons. Consistent with this, Can-SINEs inserted within 3'UTRs far upstream (>225bp) from the 3'UTRs' 3' end are five times more likely to be in antisense than sense orientation (Fig. 5d). Furthermore, the antisense-oriented Can-SINEs have over a seven times higher mean count of AATAAA sequences, with most sense-oriented Can-SINEs in the "far" category having zero AATAAA motifs on their head-to-tail strand (Fig. 5d). The situation is just the reverse for "near" retrotransposons, those wholly inserted within 225 bp of a 3' UTR's 3' end (Fig. 5e, i, m). For example, 92% of Can-SINEs inserted in this location are in sense orientation and have significantly more AATAAA motifs than their antisense counterparts (Fig. 5e). A small number of sense-oriented Can-SINEs in the "near" location apparently have zero AATAAA motifs but after manually checking all such cases we found that most are simply mis-annotated (Table S3).

LINEs on average have fewer AATAAA motifs in their sequence than Can-SINEs, but like Can-SINEs display a reversal in mean counts between the "far" vs. "near" categories (Fig. 5h, i). MIRs overall follow the same trend (Fig. 5j-m) albeit without a significant difference in mean in the "far" category after correction for multiple hypothesis testing (Fig. 5l). The pattern across all three types of retrotransposons is clear: the AATAAA motif count is high for sense-oriented retrotransposons near the end of a 3'UTR but low on average for elements inserted farther upstream in a 3'UTR or within an intron.

We found 38 RepeatMasker-annotated Can-SINEs within the "near" category that do not harbor any AAT AAA motifs. To determine whether the 3'UTRs in which they are inserted do have a PAS, and whether it is present within the SINE, we individually checked each of these 38 SINE insertions (Table S3). At 32 loci there is an AATAAA motif within 40 bp of the 3' end of the 3'UTR (which we define as an AATAAA motif close to the 3' end of the 3'UTR). In fact, in 26 of these 32 3'UTRs, the SINE insertion had been mis-annotated by being split into several separate repeat records. The record in each case is incorrectly marked as a simple sequence instead of a SINE and spans an expansion of the TAAA(n) STR from the Can-SINE AT-rich tail. Therefore, in 26/38 cases where we counted zero PASs, the SINE insertion actually does include the PAS, and often in several copies. The signal for counts of AATAAA motifs imported by Can-SINEs to 3' ends is therefore even stronger than the distribution shown in Fig. 5e, where we kept strictly to counts within reference genome-annotated repeats.

One particularly interesting SINEC_Cf3 insertion occurred in the dog *CTU1* (Fig. 6). Essentially, the entire 3'UTR modeled in the dog is the sense-oriented SINE. The small final coding exon and the SINE-become-3'UTR do not occur in the *CTU1* orthologs in two other carnivores: the ferret (*Mustela putorius furo*, reference musFur1) and cat (*Felis catus*, reference felCat9). In the reference genomes of both these species a trio of old repeat sequences, MIR, L2b and LTR106, are present in the middle of the intron, as in dogs, but the dog SINEC_Cf3 is not. Furthermore, the human (assembly hg38), ferret, and cat all show a completely

(See figure on next page.)

Fig. 5 AATAAA PAS count differs significantly between sense and antisense-oriented retrotransposons in transcripts. In all cases the number of AATAAA motifs on the head-to-tail strand of the retrotransposon was counted. Panel **a** shows the four different categories of retrotransposon insertions that were analyzed. Each panel's table summarizes sense ("S"; row 1) and antisense ("AS"; row 2)-oriented retrotransposons with the mean number of AATAAA motifs (column 2) and the total count of such elements in the dog reference genome CanFam3 (column 3). Row 3 reports the randomization test for difference in means between S and AS as raw p values uncorrected for the 12 hypothesis tests. The columns left to right are genomic location categories: intergenic (**b**, **f**, **j**), intronic (**c**, **g**, **k**), within a 3'UTR but nearest edge>225 bp from the 3' end ("far"; **d**, **h**, **l**), and within a 3'UTR but<225 bp from the 3' end ("near"; **e**, **i**, **m**). For the intergenic location the "S" and "AS" denote top and bottom strand. In all histograms the proportion at count=5 accounts for retrotransposons having 5 or more AATAAAs

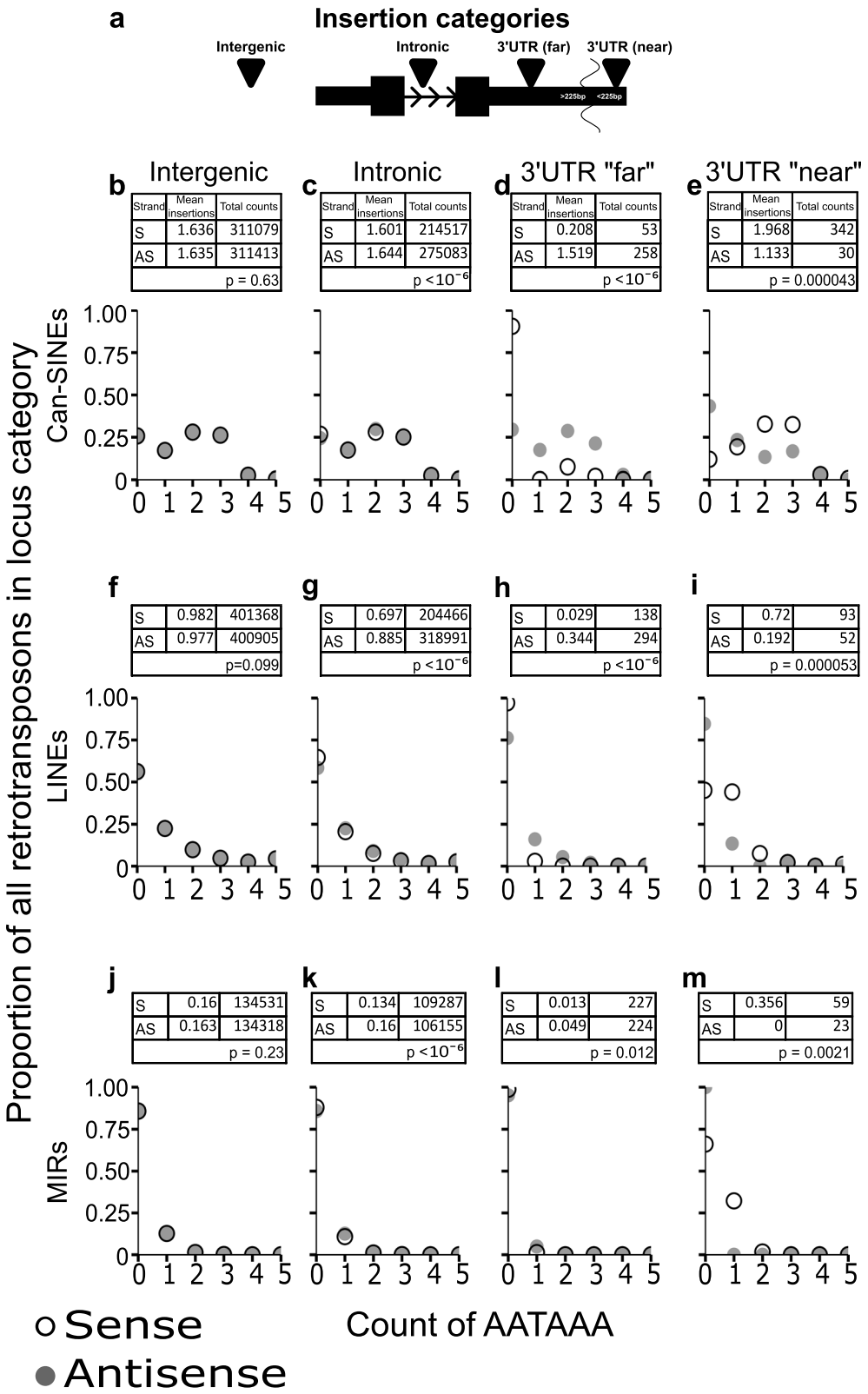

Fig. 5 (See legend on previous page.)

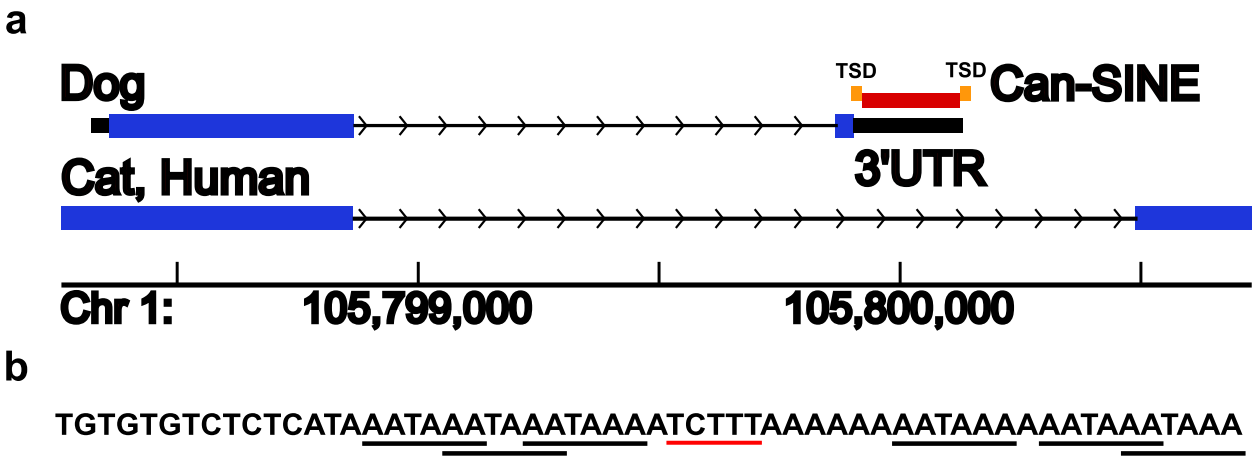

Fig. 6 A sense-oriented SINEC_Cf3 is the dog *CTU1* 3'UTR. **a** Taller filled boxes represent coding exons (blue), and shorter boxes represent untranslated regions (black). The SINE is mis-annotated as three separate repeat records that are in fact one SINE insertion (red). **b** Six AATAAA motifs occur within the SINE's AT-rich tail with a TCTTT RNA pol III terminator in the middle. A 15 bp target site duplication with two mismatches starts within the sixth AATAAA motif. The human and cat orthologs contain a different final coding exon and 3'UTR

different final coding exon and 3'UTR. Thus, a Can-SINE in one carnivore lineage but not others appears to have completely altered the downstream half of the *CTU1* gene.

A SINEC_Cf2 inserted into *SDHAF2* is another example of a sense-oriented Can-SINE near a 3'UTR 3' end (Fig. S3). Again, the entire 3'UTR in the dog gene model is the SINE, which is supported by 15 bp TSDs (with just a single base mismatch) flanking the Can-SINE. This SINEC_Cf2 element is relatively young at just 4.3% sequence divergence from the consensus sequence and is not present in either the ferret or cat orthologs of *SDHAF2*. There are several variant 3'UTRs for the human ortholog, several of which are much longer than the SINE-only 3'UTR in the dog.

The human *IL17RA* gene's 3'UTR is over 5800 bp long but in the dog ortholog a Can-SINE insertion has apparently truncated nearly all of the 3'UTR (Fig. S4). This SINEC_Cf2, at 12.8% sequence divergence from consensus, is likely older than many other insertions but is nevertheless not present in either the ferret or cat *IL17RA* orthologs.

The young sense-oriented SINEC_Cf copy at the 3' end of the dog *TTLL12* 3'UTR (Fig. S5) is just 4.4% diverged from the SINEC_Cf consensus sequence and is not present in either the ferret or cat orthologs. As in the examples above, the single SINE insertion has been annotated as multiple records. The two records should be annotated as a single SINE insertion, as shown by the putative 16 bp TSDs that surround the two segments. Strikingly, the AT-rich tail of this SINE has mutated into a pure TAAA(17) simple tandem repeat, providing 16 AATAAA sequences in the process.

The high copy number of Can-SINEs in the dog genome implies that short TAAA(n) simple tandem repeats are delivered to many thousands of loci by retrotransposon activity. STR mutation rates are often orders of magnitude higher than single base substitutions [35] and in some cases the TAAA(n) experiences a repeat expansion, as seen in the dog *TTLL12* locus (Fig. S5). Consistent with high Can-SINE activity, the dog has a high genome count of TAAA(n) STRs compared to at least some other vertebrates such as the rabbit, cow, sheep, horse, cat, chicken, golden eagle, lizard, garter snake, tibetan frog, (Table S4), with roughly twice as many of these STRs as are found in the human, chimpanzee, rat, or mouse genomes. The zebrafish contains a slightly elevated count of TAAA(n) STRs compared to the dog.

Aside from the expected 22 bp peak, the distribution of AATAAA sequences in dog 3'UTRs also spikes at distances of 28, 32, and 36 bp (Fig. 1). To validate our Python code's tallies we randomly sampled 50 3'UTRs from each of the top five peak bp positions for AATAAA motif frequency: 22, 24, 28, 32, and 36 bp from the 3'UTR 3' end; we then manually verified that an AATAAA motif does occur at that bp position. We found no errors in our script-based tallying (Table 2; Table S5). We count the distance to the 3'UTR 3' end from the promoter-proximal 5' end of all motifs. For each of these 250 3'UTRs we also recorded if there was an AATAAA motif found at any of the other four positions (Table S5). We found that Can-SINE TAAA(n) STRs contribute to the 4 bp interval peaks. In fact, Can-SINEs contributed the AATAAA motif at 60% of our 32 bp dataset and in every case the 3'UTR also has an AATAAA motif 4 bp upstream. We next tabulated the absolute counts of SINEs in sense and

Table 2 Random samples of 50 3'UTRs having an AATAAA PAS at one of five bp distances were manually checked to confirm the AATAAA and count if the PAS was within a repeat or not (other)

	AATAAA Identified at (bp)				
	22	**24**	**28**	**32**	**36**
Can-SINE	0 (0)	7 (4)	4 (3)	31 (31)	6 (6)
LINE	3 (0)	4 (0)	2 (2)	0 (0)	5 (5)
MIR	0 (0)	0 (0)	0 (0)	0 (0)	1 (1)
Other	47 (0)	39 (1)	44 (7)	19 (3)	38 (38)

In parentheses is the number of 3'UTRs that also have an AATAAA at 36 bp

antisense orientation that contribute AATAAA motifs to the 3' end of 3'UTRs (Table S6). Many times more AAT AAA motifs are imported to 3'UTR ends by sense-oriented rather than antisense-oriented Can-SINEs.

Discussion

Can-SINEs have been active throughout the evolution of the dog species. The youngest Can-SINE, SINEC_Cf, appears to still be active today as evidenced by the tens of thousands of polymorphic insertions in purebred dogs [25]. In contrast, other types, like SINEC_c1 and c2, are much older. We have annotated individual insertions of Can-SINE elements to better understand and predict their ages. Length of polyA tail has been used as a predictor for age of SINE insertions [36]. Insertions of non-A may "stabilize" the polyA sequence. We find a correlation between the length distribution of perfect TSDs and SINEC type (Fig. 4), where the younger SINEC_Cf, Cf2 and Cf3 types exist in many more copies with perfect 15 bp TSDs. SINEC_Cf copies with perfect 15 bp TSDs also have on average longer AT-rich tails. We found a number of specific SINE insertions in which the AT-rich tail underwent an expansion in the TAAA(n) STR, including in cases where the Can-SINE is providing the PAS AAT AAA (or set of AATAAA motifs) for a 3'UTR. It would be interesting to know the timing of this process. Are these STRs more mutable when the SINE is freshly inserted and less likely to have inclusions of sequence within the tandem repeat [37]? Perhaps an optimum time window exists in which a SINE is more likely to provide a PAS to a gene: after insertion but before it accumulates many mutations.

Why do we find Can-SINEs in 3'UTR ends that have an expanded TAAA(n) STR? Is this a chance occurrence resulting simply from the high mutability of the STR? Or are these cases of an adaptive strengthening of the PAS (if indeed additional AATAAA sequences do strengthen the termination signal)? Do added AATAAA motifs serve as a "work-around" for sub-optimal sequences surrounding the PAS? We have identified several hundred sense-oriented Can-SINEs at the 3' ends of 3'UTRs. Nearly all of them, from old SINEC_C1s to young SINEC_Cfs, contain AATAAA motifs. This suggests a long-lasting and ongoing importation/mutation of SINE-PASs at gene ends.

With hundreds of domestic dogs now being whole-genome sequenced, the data exist to compare orthologous Can-SINEs across canids and analyze the extent to which gene usage of Can-SINE PASs may be conserved. To what extent have Can-SINE insertions driven switching to new polyadenylation sites in canid evolution? Perhaps through comparative genomics, too, signatures of selection may reveal adaptive Can-SINE-altered gene ends.

We speculate that Can-SINE insertion-caused shortening of 3'UTRs (Fig. S4) could expose or remove miRNA binding sites and lead to changes in gene expression, as has been demonstrated in humans and other mammals [38]. Changes in 3'UTR length can bring miRNA binding sites closer to the ends of the 3'UTR where they are more effective [39]. In human *CDC6*, for example, 3'UTR shortening removes negative regulation of gene expression [40].

The high rate of recent SINEC_Cf insertions in the domestic dog genome provides a natural experiment for assessing the strength of transcription termination signals in Can-SINEs. We previously identified tens of thousands of putative polymorphic SINEC_Cf elements in a set of purebred dogs, including insertions within 3'UTRs. If these data were paired with gene expression datasets it would be possible to experimentally check use or non-use of Can-SINE-PASs in 3'UTRs and elsewhere in gene transcripts. Presumably, switching to a newly inserted Can-SINE PAS is a disruptive event for a gene and likely to be under negative selection. We expect such SINE insertions to be rare alleles and our evidence supports that [25].

Can-SINEs are a T+ category SINE with putative transcript termination potential. We previously found that SINEs in dog introns, as is true in other mammals' genomes, strongly favor the antisense orientation [25]. This is consistent with the idea that sense-orientated Can-SINEs are selected against, presumably due to a potential to disrupt gene expression. Here we've provided additional support for the idea that Can-SINEs have RNA pol II termination potential: Can-SINEs in transcripts have significant differences in AATAAA motif counts when in sense vs. antisense orientation (Fig. 5a-d). The magnitude of this difference is large in upstream portions of 3'UTRs but small (though significant) in introns. We find the same signal for insertions of LINEs and MIRs in introns, too. Within 3'UTR upstream regions the mean AATAAA motf count in Can-SINEs drops to 0.2, far

lower than the 1.5 average AATAAA motfs present on antisense Can-SINEs in the same gene region. Thus, AAT AAA motif removal via mutation appears to be favored for sense-oriented Can-SINEs in this region. Many Can-SINEs in sense orientation in 3'UTR upstream regions have zero AATAAA motifs. It would be worthwhile to investigate with expression datasets whether the Can-SINEs there that do have AATAAA motifs can in fact serve as polyadenylation signals in dog cells and if so, how frequently this occurs. Orthologs for at least some of these Can-SINEs likely exist in closely related canid species and a check could then be made for correlation between the presence or absence of AATAAA motifs in those species vs. Can-SINE PAS usage.

As more reference genomes are collected and annotated with gene models, it will be interesting to compare our results in canids with other T+ SINEs in other mammal clades. A dog Can-SINE newly inserting into an intron or 3'UTR very likely has more intrinsic transcription termination potential than human *Alu*, but perhaps not more than other T+ SINEs like rabbit CSINEs or mouse B2s [14]. We predict that other T+ SINE classes, when analyzed, will show a similar signal of AATAAA suppression for sense-oriented SINEs in transcripts.

The genomic era has seen an explosive increase in our knowledge of SINE and LINE retrotransposons and their roles in the evolution of genes and genomes. One theme that emerges is the specificity and particularity with which each lineage has been impacted by these "genome pests". Primate *Alu* has been shown to serve as a PAS [41] but first requires activating mutations that convert the sequence into a PAS. Sauria SINEs found in reptiles from geckos to snakes to *Anolis carolinensis* similarly appear to lack potential for transcription termination [42]. In contrast, the T+ SINEs in diverse mammal groups may be playing significant roles over evolutionary time in shaping the ends of gene transcripts. As we collect reference genomes from diverse lineages, we'll get to read each particular story of genomes being shaped by SINEs and other repeated sequences.

Conclusion

We find that dog Can-SINEs, like other T+ SINEs, carry multiple copies of the polyadenylation signal motif AAT AAA in their sense strand AT-rich 3' end tails. Can-SINEs inserted within introns and 3'UTRs have significantly different mean AATAAA motif counts depending on their strand and we conclude that the AATAAA motif is not a neutral sequence in either context. We find the same pattern for both LINEs and MIRs in introns and also for LINEs in 3'UTRs. Furthermore, we find that many more sense-oriented Can-SINEs are inserted in 3'UTR 3' ends than into positions farther upstream in 3'UTRs. We conclude that Can-SINE retrotransposition has imported AATAAA PASs into genes and in many loci these have then been used to signal transcription termination. For several loci we find evidence that a Can-SINE insertion has truncated a 3'UTR or created a new one. We speculate that other carnivores, also carrying Can-SINEs at high copy numbers, have experienced the same kinds of evolutionary changes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13100-024-00338-5.

Supplementary Material 1.
Supplementary Material 2.
Supplementary Material 3.
Supplementary Material 4.
Supplementary Material 5.
Supplementary Material 6.
Supplementary Material 7.

Acknowledgements
We thank Jonathan Specht for help with functions within our SINE annotation script.

Authors' contributions
J.D.C. analyzed most of the datasets and produced most figures. N.B.S. led the study, conducted the randomization tests, and wrote the Python programs for DNA segment intersection analysis, motif finding, and SINE annotation. L.A.D. analyzed all SINE annotations, produced Table S4, and helped produce some figures. J.D.C. and N.B.S. planned all analyses and wrote the paper with input from L.A.D.

Funding
Not applicable.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 4 November 2024 Accepted: 17 December 2024
Published online: 04 January 2025

References

1. Hesketh J. 3 -Untranslated regions are important in mRNA localization and translation: lessons from selenium and metallothionein. Biochem Soc Trans. 2004;32:4.
2. Mayr C. Evolution and Biological Roles of Alternative 3'UTRs. Trends Cell Biol. 2016;26(3):227–37. https://doi.org/10.1016/j.tcb.2015.10.012.

3. Mayr C. What Are 3′ UTRs Doing? Cold Spring Harb Perspect Biol. 2019;11(10):a034728. https://doi.org/10.1101/cshperspect.a034728.
4. Zarkower D, Stephenson P, Sheets M, Wickens M. The AAUAAA sequence is required both for cleavage and for polyadenylation of simian virus 40 pre-mRNA in vitro. Mol Cell Biol. 1986;6(7):2317–23.
5. Sun Y, Zhang Y, Hamilton K, Manley JL, Shi Y, Walz T, et al. Molecular basis for the recognition of the human AAUAAA polyadenylation signal. Proc Natl Acad Sci. 2018;115(7):E1419–28. Publisher: National Academy of Sciences Section: PNAS Plus. https://doi.org/10.1073/pnas.1718723115.
6. Hamilton K, Sun Y, Tong L. Biophysical characterizations of the recognition of the AAUAAA polyadenylation signal. RNA (New York, NY). 2019;25(12):1673–80. https://doi.org/10.1261/rna.070870.119.
7. Beaudoing E. Patterns of Variant Polyadenylation Signal Usage in Human Genes. Genome Res. 2000;10(7):1001–10. https://doi.org/10.1101/gr.10.7.1001.
8. Elbarbary RA, Lucas BA, Maquat LE. Retrotransposons as regulators of gene expression. Science. 2016;351(6274):aac7247. https://doi.org/10.1126/science.aac7247.
9. Daniels GR, Deininger PL. Repeat sequence families derived from mammalian tRNA genes. Nature. 1985;317(6040):819–22. https://doi.org/10.1038/317819a0.
10. Kajikawa M, Okada N. LINEs Mobilize SINEs in the Eel through a Shared 3′ Sequence. Cell. 2002;111(3):433–44. https://doi.org/10.1016/S0092-8674(02)01041-3.
11. Wang W, Kirkness EF. Short interspersed elements (SINEs) are a major source of canine genomic diversity. Genome Res. 2005;15(12):1798–808. https://doi.org/10.1101/gr.3765505.
12. Conley AB, Jordan IK. Cell type-specific termination of transcription by transposable element sequences. Mob DNA. 2012;3(1):15. https://doi.org/10.1186/1759-8753-3-15.
13. Lee JY, Ji Z, Tian B. Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3′-end of genes. Nucleic Acids Res. 2008;36(17):5581–90. https://doi.org/10.1093/nar/gkn540.
14. Borodulina OR, Kramerov DA. Short interspersed elements (SINEs) from insectivores. Two classes of mammalian SINEs distinguished by A-rich tail structure. Mamm Genome. 2001;12(10):779–86. https://doi.org/10.1007/s003350020029.
15. Vassetzky NS, Kramerov DA. CAN-a pan-carnivore SINE family. Mamm Genome. 2002;13(1):50–7. https://doi.org/10.1007/s00335-001-2111-1.
16. Coltman DW, Wright JM. Can SINEs: a family of tRNA-derived retroposons specific to the superfamily Canoidea. Nucleic Acids Res. 1994;22(14):2726–30. https://doi.org/10.1093/nar/22.14.2726.
17. Kirkness EF, Bafna V, Halpern AL, Levy S, Remington K, Rusch DB, et al. The Dog Genome: Survey Sequencing and Comparative Analysis. Science. 2003;301(5641):1898–903. https://doi.org/10.1126/science.1086432.
18. Borodulina OR, Golubchikova JS, Ustyantsev IG, Kramerov DA. Polyadenylation of RNA transcribed from mammalian SINEs by RNA polymerase III: Complex requirements for nucleotide sequences. Biochim Biophys Acta (BBA) Gene Regul Mech. 2016;1859(2):355–65. https://doi.org/10.1016/j.bbagrm.2015.12.003.
19. Kosushkin SA, Ustyantsev IG, Borodulina OR, Vassetzky NS, Kramerov DA. Tail Wags Dog's SINE: Retropositional Mechanisms of Can SINE Depend on Its A-Tail Structure. Biology. 2022;11(10):1403. https://doi.org/10.3390/biology11101403.
20. Borodulina OR, Kramerov DA. Transcripts synthesized by RNA polymerase III can be polyadenylated in an AAUAAA-dependent manner. RNA. 2008;14(9):1865–73. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. https://doi.org/10.1261/rna.1006608.
21. Ustyantsev IG, Tatosyan KA, Stasenko DV, Kochanova NY, Borodulina OR, Kramerov DA. Polyadenylation of Sine Transcripts Generated by RNA Polymerase III Dramatically Prolongs Their Lifetime in Cells. Mol Biol. 2020;54(1):67–74. https://doi.org/10.1134/S0026893319040150.
22. van der Vlugt HHJ, Lenstra JA. SINE elements of carnivores. Mamm Genome. 1995;6(1):49–51. https://doi.org/10.1007/BF00350894.
23. Peng C, Niu L, Deng J, Yu J, Zhang X, Zhou C, et al. Can-SINE dynamics in the giant panda and three other Caniformia genomes. Mob DNA. 2018;9(1):32. https://doi.org/10.1186/s13100-018-0137-0.
24. Walters-Conte KB, Johnson DLE, Allard MW, Pecon-Slattery J. Carnivore-Specific SINEs (Can-SINEs): Distribution, Evolution, and Genomic Impact. J Hered. 2011;102(Suppl 1):S2–10. https://doi.org/10.1093/jhered/esr051.
25. Kalla SE, Moghadam HK, Tomlinson M, Seebald A, Allen JJ, Whitney J, et al. Polymorphic SINEC_Cf Retrotransposons in the Genome of the Dog (Canis familiaris). Genomics. 2020. https://doi.org/10.1101/2020.10.27.358119.
26. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 2005;438(7069):803–19. https://doi.org/10.1038/nature04338.
27. Clark LA, Wahl JM, Rees CA, Murphy KE. Retrotransposon insertion in SILV is responsible for merle patterning of the domestic dog. Proc Natl Acad Sci. 2006;103(5):1376–81.
28. Lin L, Faraco J, Li R, Kadotani H, Rogers W, Lin X, et al. The Sleep Disorder Canine Narcolepsy Is Caused by a Mutation in the Hypocretin (Orexin) Receptor 2 Gene. Cell. 1999;98(3):365–76. https://doi.org/10.1016/S0092-8674(00)81965-0.
29. Dreger DL, Schmutz SM. A SINE Insertion Causes the Black-and-Tan and Saddle Tan Phenotypes in Domestic Dogs. J Hered. 2011;102(Suppl 1):S11–8. https://doi.org/10.1093/jhered/esr042.
30. Goldstein O, Kukekova AV, Aguirre GD, Acland GM. Exonic SINE insertion in STK38L causes canine early retinal degeneration (erd). Genomics. 2010;96(6):362–8. https://doi.org/10.1016/j.ygeno.2010.09.003.
31. Downs LM, Mellersh CS. An Intronic SINE Insertion in FAM161A that Causes Exon-Skipping Is Associated with Progressive Retinal Atrophy in Tibetan Spaniels and Tibetan Terriers. PLoS ONE. 2014;9(4):e93990. https://doi.org/10.1371/journal.pone.0093990.
32. Pelé M, Tiret L, Kessler JL, Blot S, Panthier JJ. SINE exonic insertion in the PTPLA gene leads to multiple splicing defects and segregates with the autosomal recessive centronuclear myopathy in dogs. Hum Mol Genet. 2005;14(11):1417–27. https://doi.org/10.1093/hmg/ddi151.
33. Smit AF. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev. 1999;9(6):657–63. https://doi.org/10.1016/S0959-437X(99)00031-3.
34. Zhang Y, Romanish MT, Mager DL. Distributions of Transposable Elements Reveal Hazardous Zones in Mammalian Introns. PLoS Comput Biol. 2011;7(5):e1002046. https://doi.org/10.1371/journal.pcbi.1002046.
35. Fan H, Chu JY. A brief review of short tandem repeat mutation. Genomics Proteomics Bioinforma. 2007;5(1):7–14.
36. Roy-Engel AM, El-Sawy M, Farooq L, Odom GL, Perepelitsa-Belancio V, Bruch H, et al. Human retroelements may introduce intragenic polyadenylation signals. Cytogenet Genome Res. 2005;110(1–4):365–71. https://doi.org/10.1159/000084968.
37. Eckert KA, Hile SE. Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. Mol Carcinog. 2009;48(4):379–88. https://doi.org/10.1002/mc.20499.
38. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating Cells Express mRNAs with Shortened 3′ Untranslated Regions and Fewer MicroRNA Target Sites. Science. 2008;320(5883):1643–7. Publisher: American Association for the Advancement of Science Section: Report. https://doi.org/10.1126/science.1155390.
39. Hoffman Y, Bublik DR, Ugalde AP, Elkon R, Biniashvili T, Agami R, et al. 3′UTR Shortening Potentiates MicroRNA-Based Repression of Pro-differentiation Genes in Proliferating Human Cells. PLOS Genet. 2016;12(2):e1005879. Publisher: Public Library of Science. https://doi.org/10.1371/journal.pgen.1005879.
40. Akman BH, Can T, Erson-Bensan AE. Estrogen-induced upregulation and 3′-UTR shortening of CDC6. Nucleic Acids Res. 2012;40(21):10679–88. Publisher: Oxford Academic. https://doi.org/10.1093/nar/gks855.
41. Chen C, Ara T, Gautheret D. Using Alu Elements as Polyadenylation Sites: A Case of Retroposon Exaptation. Mol Biol Evol. 2009;26(2):327–34. https://doi.org/10.1093/molbev/msn249.
42. Piskurek O, Austin CC, Okada N. Sauria SINEs: Novel Short Interspersed Retroposable Elements That Are Widespread in Reptile Genomes. J Mol Evol. 2006;62(5):630–44. https://doi.org/10.1007/s00239-005-0201-5.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.