Exclusion of SINE Inverted Pairs from the Genome of the Dog (*Canis familiaris*)

Cassandra Ward[1], Sara E. Kalla[2], Allison Seebald[2], and Nathan B. Sutter[1,2]

(1) La Sierra University, Riverside, CA and (2) work completely at College of Veterinary Medicine, Cornell University, Ithaca, NY

Abstract

Retrotransposons make up about one third of mammal genomes. The genome of the domestic dog (*Canis familiaris)* is no exception: there are 171,386 copies of the dog short interspersed element SINEC_Cf in the reference genome. This SINE is so young that many insertions have not yet gone to fixation making the dog a prime model for research on genome patterns, retrotransposon insertions, and disruption of gene expression. To discover polymorphic SINEC_Cfs, we collected a total of 279M next-gen sequence reads from 62 libraries enriched for flanks of the head end of SINEC_Cf. The libraries represent 59 distinct pure breeds. While most reads map to reference SINEC_Cfs that are presumably fixed in all dog chromosomes, approximately 8% of reads map to insertion loci not present in the reference genome, which we define as polymorphic. We found 81,747 such putatively polymorphic SINEs. We used these polymorphic SINEs and reference SINEs to analyze pairs. We define a SINE pair as two SINE insertions that are within a certain distance of each other with no intervening SINEs. There are four orientations: head-to-head, tail-to-tail, head-to-tail, and tail-to-head (Figure 1); we also track the orientation of pairs relative to a gene transcript, if present. The head-to-tail and tail-to-head orientations are inverted SINE pairs. In other mammals, including humans (with Alu pairs), such inverted SINE pairs are observed at a much lower density than direct repeats and when present in transcripts have been shown to disrupt gene expression. The inference is that inverted pairs are under negative selection. We looked for a similar loss of SINE pairs in the dog genome. Inverted pairs less than 100 base pairs apart are much less frequent than pairs in the same orientation. This relation holds for pairs within introns as well as in nongenic sequences. We also found an orientation bias for inverted SINEC_Cfs with high pairwise alignment scores. We also looked at pairs in which one is present in the reference genome and the other is a polymorphic SINE detected in our libraries (the majority of which are SINEC_Cf type). At short spacer distances we find low proportions of pairs in which the polymorphic SINEs are paired with SINEC_Cf and SINEC_Cf2. No such ratio change occurs for pairs with LINE1 or MIR-type SINEs.

Questions

1. Are inverted pairs selected against?
2. Does sequence similarity between inverted pairs affect their frequency of insertion in the dog genome?
3. Does the spacer distance between the SINE pair impact the ratio of inverted pairs?

Methods

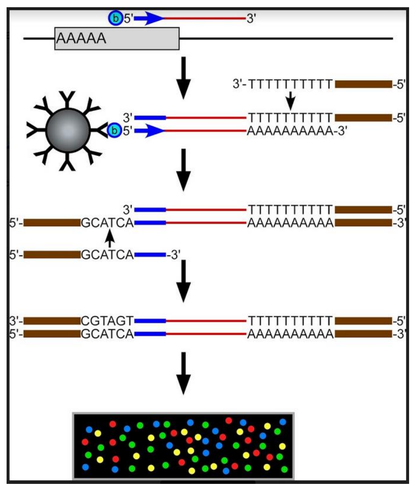

Figure 1 We created 62 libraries for Illumina Hi-Seq sequencing. Specifically, DNA synthesis is primed from dog genomic DNA with a biotinylated oligonucleotide that hybridizes to conserved sequence in SINEC_Cf. The primer's 3' end hybridizes to base 18 within the SINE. DNA extension products are captured to streptavidin-coated magnetic beads, polyadenylated, and used for 2nd-end strand synthesis. After recovery from the beads, one of 16 barcoded primers tailed with Illumina HiSeq forward adaptor sequence is used for several rounds of PCR. After quality assessment by gel sizing, 16 libraries are pooled and subjected to one lane of 100 bp single-end sequencing.

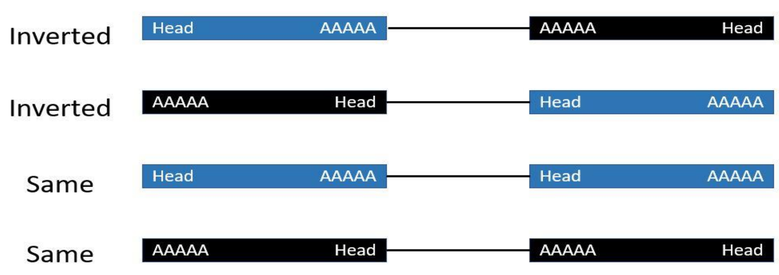

Figure 2 SINE pair orientations and how we called pairs "inverted" versus "same".

Results

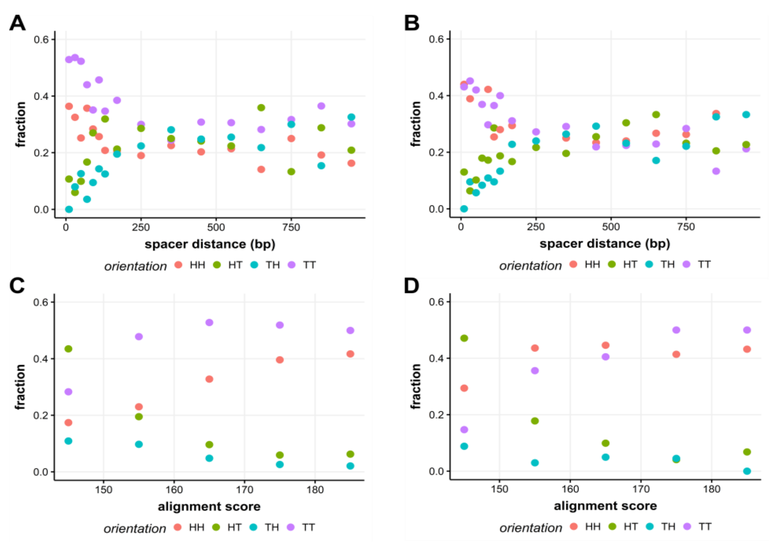

Figure 3 Inverted pairs of reference SINEC_Cf copies are rare when the SINEs are close and similar in sequence. (A) The four possible orientations of pairs in introns were tracked relative to the gene. HH = both SINEs are in sense orientation (RNA pol encounters the heads first); TT = both antisense; HT = RNA pol hits the head of the first SINE then the tail of the second; TH = RNA pol hits the tail of the first SINE then the head of the second. {HH,TT} are direct pairs while {HT, TH} are inverted pairs. (B) Nongenic pairs, H = top strand, T = bottom strand. (C) All intronic pairs with spacers up to 100 bp were included. (D) All nongenic pairs with spacers up to 100 bp were included

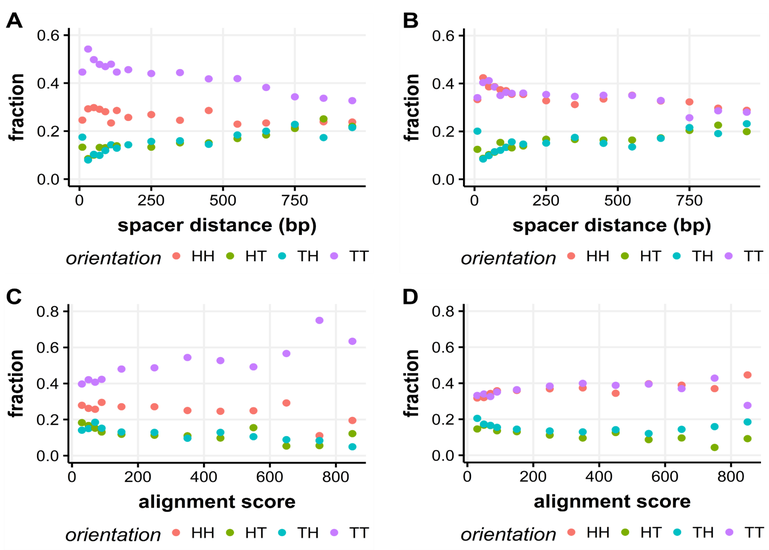

Figure 4 Inverted pairs of reference LINE copies are rare when the LINEs are close and similar in sequence. (A) The four possible orientations of pairs in introns were tracked relative to the gene. HH = both LINEs are in sense orientation (RNA pol encounters the heads first); TT = both antisense; HT = RNA pol hits the head of the first LINE then the tail of the second; TH = RNA pol hits the tail of the first LINE then the head of the second. {HH,TT} are direct pairs while {HT, TH} are inverted pairs. (B) Nongenic pairs, H = top strand, T = bottom strand. (C) All intronic pairs with spacers up to 500 bp were included. (D) All nongenic pairs with spacers up to 500 bp were included.

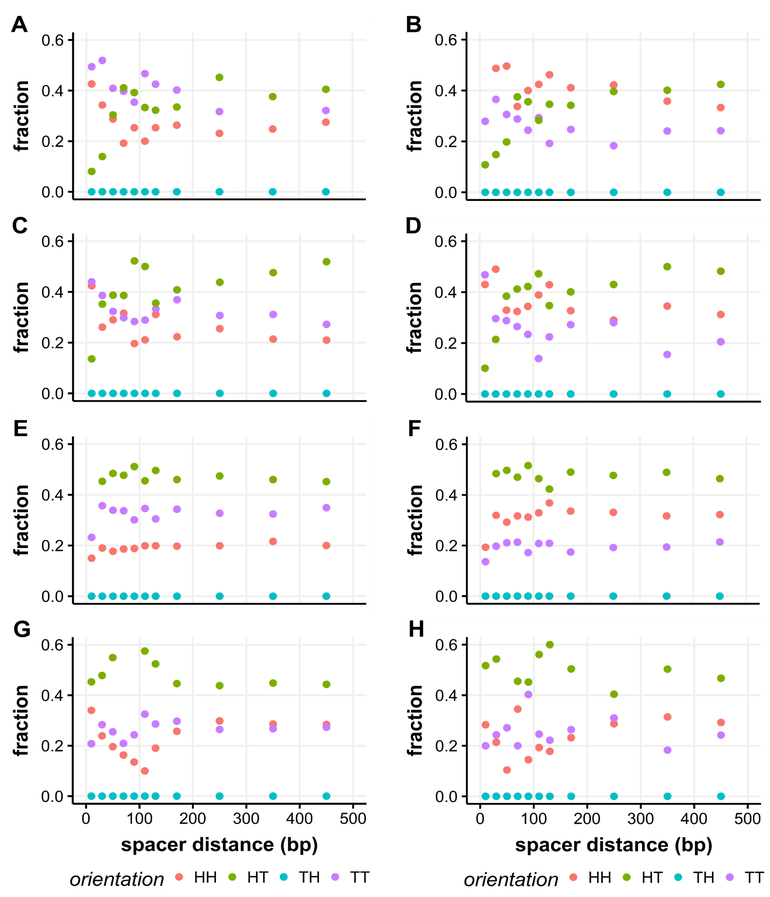

Figure 5 Polymorphic SINEs near SINEC_Cf and SINEC_Cf2 are rarely in inverted orientation. Because Head-proximal flanks containing repeat sequences were filtered out of the dataset during our SINE discovery we had to exclude from this analysis all pairs in which the polymorphic SINE's head faces the reference SINE. This means we can never find the "TH" orientation in which the heads point toward each other. The three remaining orientations of pairs were tracked in relation to the gene's direction of transcription (left column) or top strand (right column). HH = both in pair are in sense orientation; TT = both antisense; HT = heads face out (inverted orientation). Polymorphic SINEs are paired with: (A) reference SINEC_Cf in introns and (B) nongenic sequence, (C) SINEC_Cf2 in introns and (D) nongenic, (E) LINEs in introns and (F) nongenic sequence, and (G) MIRs in introns and (H) nongenic sequence.

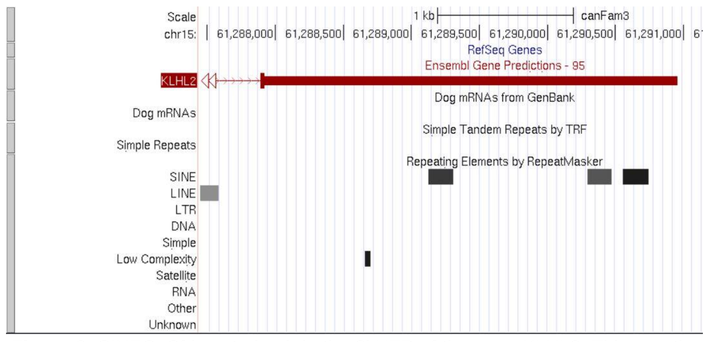

Figure 6 SINEC_Cf inverted pair in the 3' UTR of the gene KLHL2. This was the only reference SINEC_Cf inverted pair in a 3' UTR in the dog genome. The alignments of the two SINEs were compared and found to be at 79.2% and had the spacer distance of 83 base pairs.

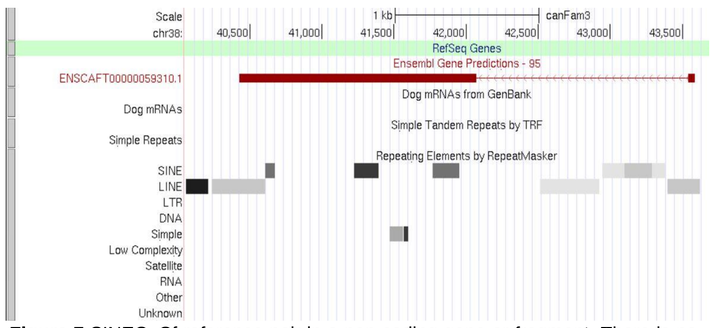

Figure 7 SINEC_Cf reference pair in a non-coding gene or fragment. There is no no syntenic similarity with human, so we suspect this is a pseudogene.

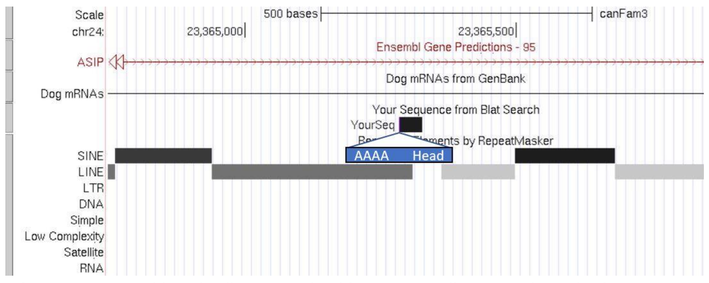

Figure 8 Inverted head-tail SINE pair with a spacer distance of 214 bp in ASIP gene. This polymorphic SINE paired with a reference SINE is reported to cause black and tan saddle coat patterns in dogs (Dreger, 2011).

Results and Discussion

While a null model would predict all four SINE pair orientations at equal frequency, we observed inverted pairs at a much lower rate in some circumstances. Inverted pairs of reference SINEC_Cfs <50 bp apart are only 20% not 50% of the total. This bias stands for both intronic as well as nongenic pairs. LINE inverted pairs are also rare in both introns and nongenic sequence and the effect is apparent at a 10 times greater distance (up to approximately 500 bp) than SINEC_Cfs. While looking at the alignment scores in SINEC_Cf reference pairs we found very few inverted pairs with the highest scores. LINEs are longer than SINEs and the pairwise alignment scores are comparatively higher, which may explain why we observe less of a bias against inverted pairs. With the evidence that reference (mostly fixed) SINEC_Cfs are rarely inverted when close together we hypothesized that this same trend would persist when comparing our polymorphic SINEs to SINEC_Cf. At very close distances (0-50bp) we see the same loss of inverted pairs in both intronic and nongenic sequence for polymorphic SINEs paired with both SINEC_Cf and SINEC_Cf2. Finally, we found just a single reference SINEC_Cf inverted pair in a 3' UTR. The rarity of this event supports the idea that retrotransposon insertions in a 3' UTR can disrupt gene expression.

Acknowledgements

The authors thank David Kupiec, Florencia Ardon and Lalit Ponnala for technical assistance, and Peter Schweitzer at the Cornell University Life Sciences Core Lab for help trouble shooting barcoding issues with our custom Illumina HiSeq libraries. We also thank the dog owners and breeders who contributed samples from their animals. Research reported here was supported by the National Human Genome Research Institute of the National Institutes of Health under award number R21HG006051. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research was also supported by internal funds from Cornell University and La Sierra University.

Bibliography

Cook, G. W., M. K. Konkel, J. D. Major, J. A. Walker, K. Han et al., (2011) *Alu pair exclusions in the human genome*. Mobile DNA 2: 10.

Terry Fitzpatrick & Sui Huang (2012) *3'-UTR-located inverted Alu repeats facilitate mRNA translational repression and stress granule accumulation*. Nucleus, 3:4, 359-369, DOI: 10.4161/nucl20827

Tajaddod, M., Tanzer, A., Licht, K. et al. (2016) *Transcriptome-wide effects of inverted SINEs on gene expression and their impact on RNA polymerase II activity*. Genome Biol 17, 220.

Dayna L. Dreger, Sheila M. Schmutz (2011) *A SINE Insertion Causes the Black-and-Tan and Saddle Tan Phenotypes in Domestic Dogs*. Journal of Heredity, Volume 102, Issue Suppl_1, September-October 2011, Pages S11–S18.