Characterization of Polymorphic SINE Insertions and Genes in Dog Retrotransposon Free Regions

Yun Seok Lee, Sara E. Kalla, Allison Seebald, Jessica D. Choi, Jeremy J. Allen, Nathan B. Sutter

La Sierra UNIVERSITY

Abstract

Retrotransposons are mobile genetic elements that have played a major role in mammalian genome evolution. For example, retrotransposon insertions in the dog genome have introduced novel open reading frames and splice acceptor sites, and caused phenotypes ranging from narcolepsy and other diseases to the merle coat pattern selected within some breeds. One dog retrotransposon in particular, SINEC_Cf, is so young that thousands of insertions have not yet gone to fixation. Despite the presence in the dog reference genome of 1,351,940 LINEs and 1,134,572 SINEs (of which 171,386 are SINEC_Cf), we have identified 1375 "free regions" that are at least 10,000 bp long and contain no SINEs, LINEs, or assembly gaps. There are 16,901 free regions at least 5000 bp long, many of which span over gene upstream or downstream ends. We have analyzed the genes found in these dog SINE+LINE free regions because transposon free regions in the human and mouse genomes were previously shown to be rich in genes crucial for early development and transcriptional regulation. We have also analyzed patterns of polymorphic SINE insertion into our free regions to check whether SINEs in these loci have lower than average insertion frequencies or tend to insert at free region edges. To make this possible we Illumina sequenced 434 libraries created by extending into flanking non-repeat sequence from a primer hybridizing to conserved SINEC_Cf sequence. The libraries represent 356 dogs from 125 breeds.

Questions

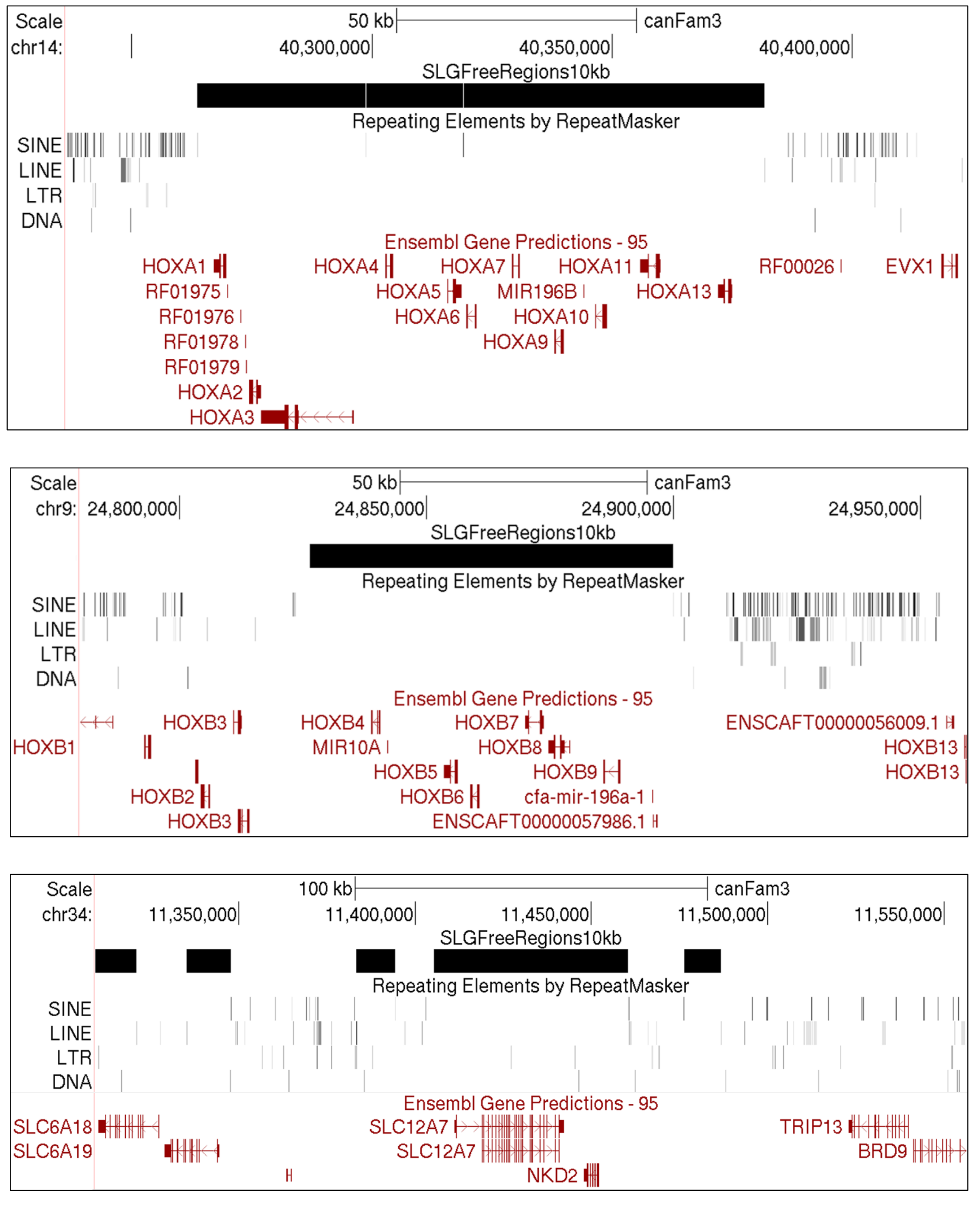
1. Which genes are present or overrepresented in the retrotransposon free regions?
2. In what manner do polymorphic SINEs insert themselves into free regions?

Methods

We created custom libraries enriched for SINEC_Cf flanking sequences in two ways: 1) extension from a SINE hybridizing biotinylated primer and 16-plex HiSeq sequencing, and 2) restriction digest of gDNA and circularization with ligase followed by inverse PCR with two SINE hybridizing primers and HiSeq of 95-plex pools. SINE flanks with >20% repeated sequence were filtered out of the dataset to improve genome alignment quality and remaining seq. read alignments were used to discover SINE insertion loci. Genes, SINEs and free regions were intersected using a custom python script.

Results

We found 1376 free regions in the dog genome. Many contain protein-coding genes, including all of the longest free regions (Figure 1, below). Of 167,000 total polymorphic SINEs discovered, 1302 are found in reference SINE-free regions.



Results

To determine which types of sequences are prevalent in free regions we compared their base pair composition to the genome as a whole. Free regions are 6 fold richer in coding exons. In contrast, the composition of nongenic sequences in the genome overall was higher than free regions by 52% (Table 1).

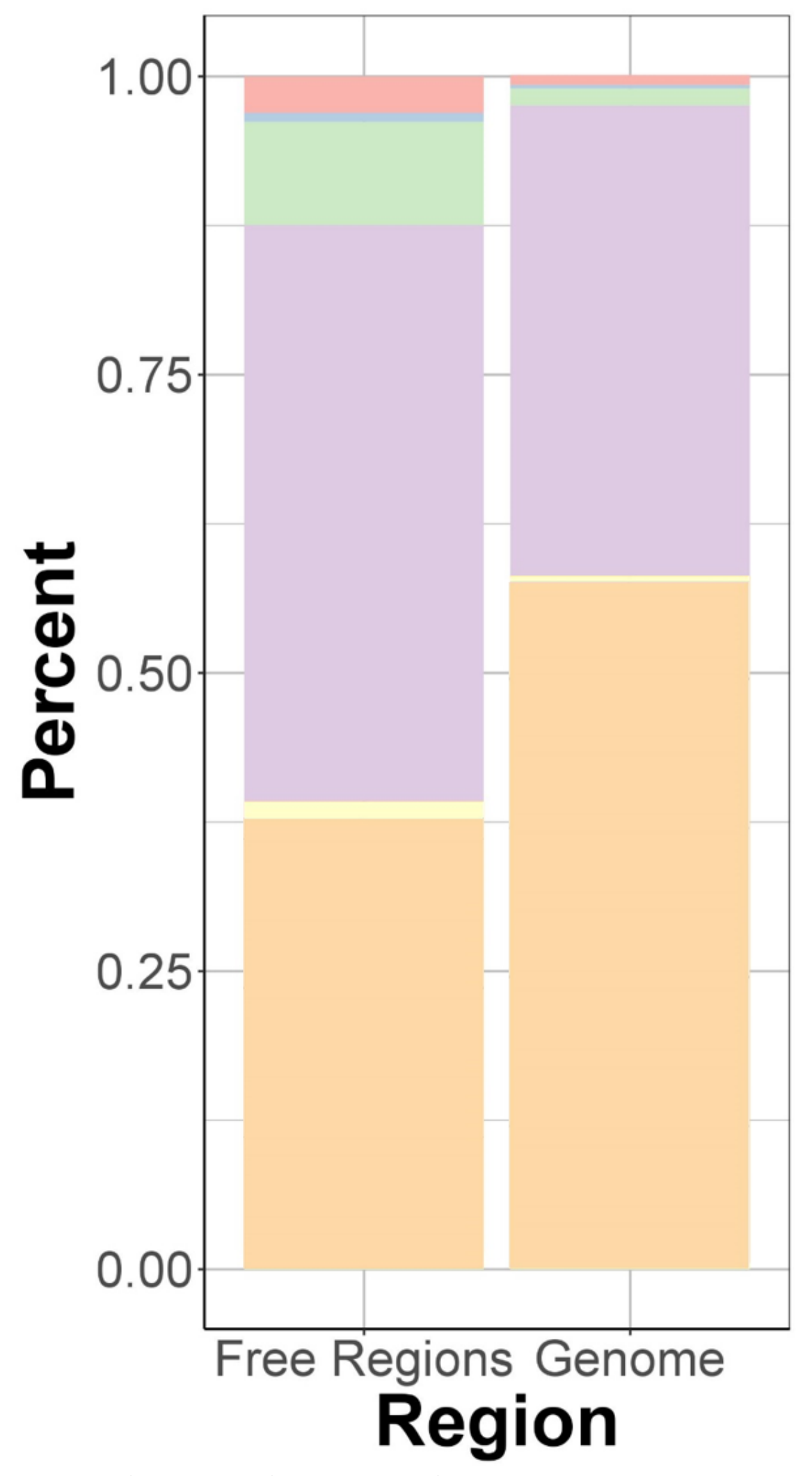


Figure 2 (left). Stacked bar plots of the base pair composition of free regions and the genome. Free regions contain more 3'UTR, 5'UTR, exon, intron, and promoter sequences.

Table 1 (bottom). Percent composition of free regions and the genome. Free regions are 32% richer in genic sequences than the genome.

	% in Free Regions	% in the Genome
3'UTR	3.1%	0.8%
5'UTR	0.8%	0.3%
Exon	8.7%	1.4%
Intron	48.3%	39.5%
Promoter	1.4%	0.5%
Nongenic	37.8%	57.6%

GO analysis shows that gene products in free regions that serve as transcription factors regulating early morphogenesis are overrepresented (Table 2), with their fold enrichment scores ranging from 1.88 to 2.39.

GO Biological Process	Fold Enrichment	P-value
multicellular organism development	1.96	3.06E-22
anatomical structure development	1.88	2.22E-21
system development	2	2.48E-21
developmental process	1.8	1.71E-19
regulation of gene expression	1.87	2.54E-19
anatomical structure morphogenesis	2.39	3.37E-19
regulation of RNA metabolic process	1.94	6.89E-19
animal organ development	2.11	1.70E-18
regulation of transcription, DNA-templated	1.95	1.15E-17
regulation of RNA biosynthetic process	1.92	2.54E-17

GO Cellular Component	Fold Enrichment	P-value
nucleus	1.57	1.69E-13
intracellular membrane-bounded organelle	1.38	3.68E-10
nucleoplasm	1.75	1.21E-09
membrane-bounded organelle	1.32	1.66E-08
synapse	2.14	6.29E-08
nuclear lumen	1.57	6.34E-08
organelle	1.26	1.30E-07
intracellular organelle	1.27	1.42E-07
organelle lumen	1.52	1.71E-07
intracellular organelle lumen	1.52	1.71E-07

GO Molecular Function	Fold Enrichment	P-value
sequence-specific DNA binding	3.24	4.81E-23
DNA binding	2.45	1.47E-21
binding	1.36	7.01E-21
nucleic acid binding	1.87	3.42E-18
DNA-binding transcription factor activity	2.77	1.94E-15
transcription regulator activity	2.46	1.94E-14
sequence-specific double-stranded DNA binding	2.99	5.30E-13
transcription regulatory region sequence-specific DNA binding	2.99	1.37E-12
RNA polymerase II regulatory region DNA binding	3.03	2.92E-12
RNA polymerase II regulatory region sequence-specific DNA binding	3.03	2.92E-12

Table 2. The ten most statistically significant GO analysis results for the three domains: biological process (top), cellular component (middle), and molecular function (bottom). Transcription factors that regulate morphogenesis and early development are overrepresented.

Figure 1 (left). The three longest free regions in the dog genome. The two longest regions (top and middle) harbor HOX A and B clusters which contain genes for pattern formation and embryonic development. The third longest region (bottom) encompasses SLC12A7, which codes for a solute carrier protein.

Insertions of polymorphic SINEs within free regions were surveyed in order to analyze their position preferences within the regions. There was no significant bias shown (Figure 4).

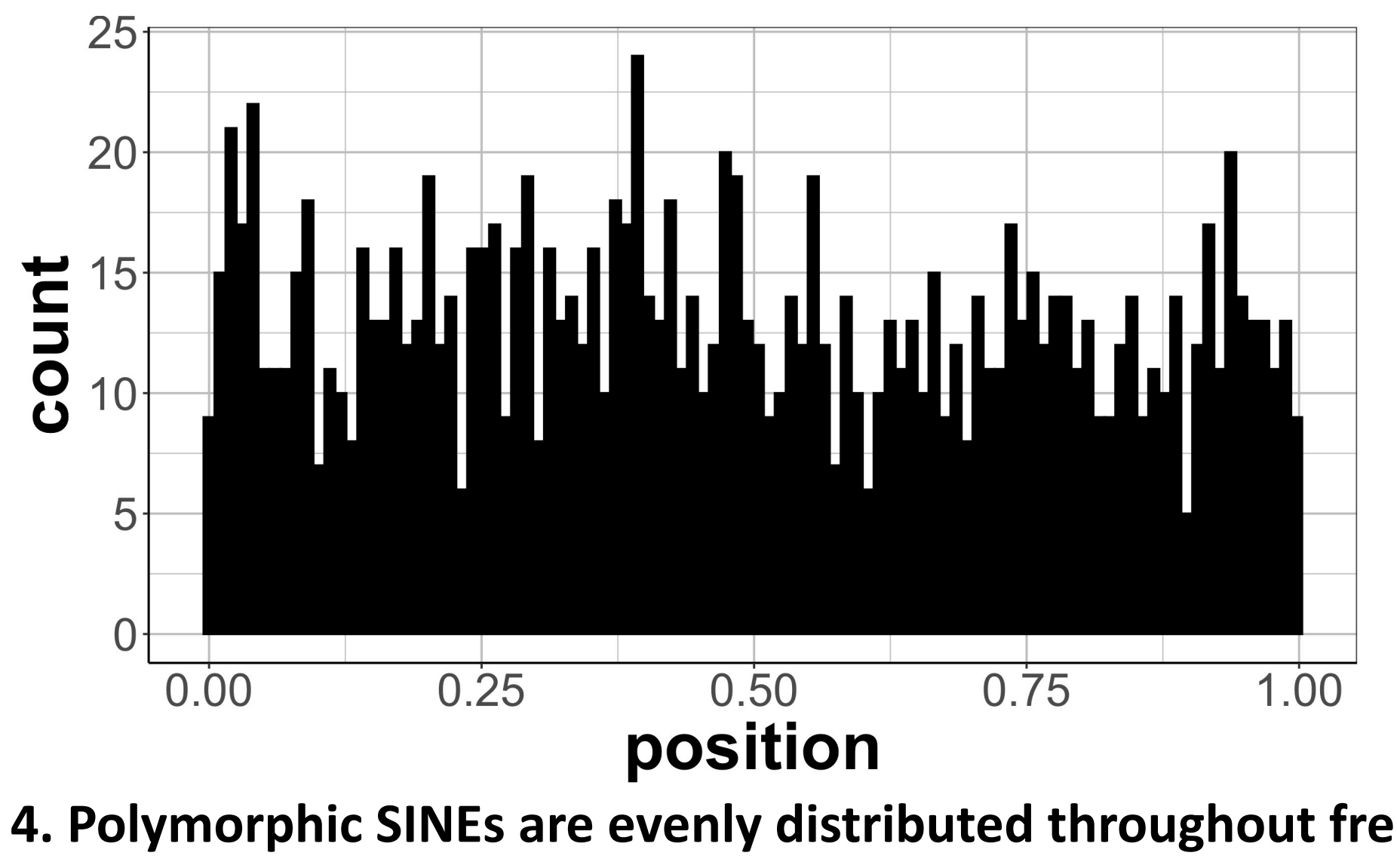


Figure 4. Polymorphic SINEs are evenly distributed throughout free regions. Specifically, SINEs don't cluster at the edges of the regions.

SINE ID	Chr	bp	Gene	Location	Samples	SINE ID	Chr	bp	Gene	Location	Samples
5349691	chr1	75510486	HNRNPK	3'UTR	1	5463840	chr14	53837050	FOXP2	exon	1
5376421	chr1	107062791	SLC17A7	exon	1	5262590	chr15	6887959	SFPQ	exon	30
5412079	chr1	120454209	TSHZ3	exon	2	5437052	chr15	45032222	POU4F2	exon	1
5366261	chr2	17455379	BAMBI	exon	1	5413488	chr16	6702820	TRPV6	exon	1
5463420	chr2	55246392	MAP1B	exon	1	5375896	chr16	6702898	TRPV6	exon	1
5450184	chr2	85213370	CASZ1	3'UTR	1	5473350	chr16	6703867	TRPV6	exon	1
5365725	chr3	27230034	CMYA5	exon	2	5453202	chr17	1000317	MYT1L	exon	1
5281283	chr3	30965079	CERT1	3'UTR	13	5390803	chr17	1009431	MYT1L	exon	1
5295580	chr3	61356683	ADD1	3'UTR	14	5383672	chr17	15885284	APOB	exon	1
5379123	chr3	61362798	ADD1	exon	1	5364331	chr17	51780695	HIPK1	3'UTR	2
5410823	chr3	61362804	ADD1	intron	1	5386524	chr17	60355867	ZNF687	exon	1
5389195	chr4	6690144	PCNX2	exon	2	5397979	chr17	60355916	ZNF687	exon	1
5360944	chr5	30817112	PITPNM3	exon	1	5403506	chr17	60356179	ZNF687	exon	1
5426339	chr5	30821239	SKIDA1	3'UTR	1	5362799	chr17	60360349	PI4KB	exon	2
5395423	chr5	42238126	MPRIP	intron	1	5365608	chr18	1737207	IKZF1	exon	1
5427526	chr5	56438136	C1QTNF12	5'UTR	1	5395930	chr18	22516766	PCLO	exon	1
5431192	chr5	56487033	INTS11	exon	1	5468533	chr18	25360743	BET1L	exon	1
5427514	chr5	56510877	DVL1	exon	1	5402946	chr18	45523815	MUC5B	exon	1
5418247	chr5	57370342	PANK4	3'UTR	1	5438206	chr18	46005052	IFITM10	exon	1
5473045	chr5	58015159	PRDM16	3'UTR	1	5398028	chr18	46078936	TNNT3	intron	1
5378020	chr5	58060768	MEGF6	exon	1	5420180	chr18	54106559	AHNAK	intron	1
5450070	chr5	62936002	CLSTN1	3'UTR	1	5411518	chr20	37249686	NT5DC2	exon	1
5426088	chr5	64566887	CDT1	exon	2	5422628	chr20	37257202	STAB1	exon	1
5426084	chr5	65440389	KLHDC4	exon	1	5371678	chr20	40052452	LAMB2	exon	1
5379888	chr5	81605230	EDC4	exon	1	5395998	chr20	40054482	LAMB2	exon	1
5472172	chr5	81787584	PARD6A	exon	1	5405310	chr20	40067491	USP19	exon	1
5478741	chr5	81800361	CARMIL2	exon	1	5415752	chr20	40150270	WDR6	exon	1
5389275	chr6	9025155	GIGYF1	exon	1	5408430	chr20	40465935	CELSR3	exon	1
5430434	chr6	9032956	GNB2	exon	1	5376986	chr20	40470134	CELSR3	exon	1
5442441	chr6	9033102	GNB2	exon	1	5473367	chr20	40526798	COL7A1	exon	1
5439976	chr6	9115461	SAP25	exon	1	5374861	chr20	44757687	IQCN	exon	1
5421081	chr6	11112285	TECPR1	3'UTR	1	5453769	chr20	44758258	IQCN	exon	1
5466183	chr6	11553933	USP42	3'UTR	1	5399860	chr20	52344210	PNPLA6	exon	1
5428662	chr6	11554006	USP42	3'UTR	2	5467395	chr20	57559977	STK11	intron	1
5338637	chr6	11555074	USP42	3'UTR	2	5474612	chr21	40305049	KCNC21	exon	1
5427548	chr6	11557712	CYTH3	intron	1	5468803	chr22	39386473	SLITRK5	exon	1
5431203	chr6	12872230	FOXK1	exon	1	5383139	chr22	60410409	ATP11A	exon	1
5414933	chr6	14872760	SNX8	3'UTR	1	5470111	chr22	60978562	GAS6	intron	1
5378896	chr6	37533969	CREBBP	exon	2	5405728	chr23	47503578	RAP2B	exon	1
5379147	chr6	38853505	PKD1	exon	1	5362642	chr24	27146054	SLC32A1	exon	1
5455494	chr6	39135856	MAPK8IP3	exon	1	5416604	chr24	29210254	PLCG1	exon	1
5442443	chr6	39513067	CACNA1H	exon	1	5358169	chr24	47438292	SAMD10	exon	1
5408024	chr6	39920714	WDR90	exon	2	5392963	chr24	47512885	RGS19	3'UTR	1
5389265	chr6	43042311	PRPF38B	3'UTR	1	5398249	chr25	44092678	DIS3L2	exon	1
5299826	chr6	43043398	PRPF38B	3'UTR	13	5392997	chr26	1722213	RIMBP2	exon	1
5415971	chr6	68965112	NEXN	exon	1	5428768	chr26	24295277	PATZ1	exon	1
5364519	chr7	26818584	DNM3`	intron	2	5398258	chr26	24295589	PATZ1	exon	1
5463645	chr7	35128830	ZBTB18	exon	1	5408981	chr30	8868383	PLA2G4B	exon	1
5473950	chr7	78593053	CXXC1	exon	1	5406052	chr30	18425483	ONECUT1	exon	1
5449956	chr9	509483	HGS	exon	1	5444185	chr31	37264463	U2AF	exon	1
5397681	chr9	1030128	RPTOR	exon	1	5404096	chr31	37709796	PDXK	3'UTR	4
5389371	chr9	1731079	TBCD16	intron	1	5422955	chr31	38166824	PFKL	exon	1
5399591	chr9	1732473	TBCD16	3'UTR	1	5428780	chr31	38178325	PFKL	exon	1
5442460	chr9	1851749	CBX2	exon	1	5459443	chr32	21653351	SLC25A12	intron	1
5394409	chr9	4755865	UNC13D	intron	1	5412238	chr34	8073797	ICE1	exon	2
5361744	chr9	45049427	TLCD3A	intron	1	5423834	chr34	11300066	TERT	exon	1
5428686	chr9	48432395	TUBB4B	exon	1	5362673	chr34	14853378	SOX2	exon	1
5421856	chr9	48486446	ANAPC2	exon	1	5456507	chr34	33786162	MECOM	exon	1
5376727	chr9	49546528	KCNT1	exon	2	5313835	chr36	16480048	METAP1D	3'UTR	18
5421857	chr9	49689668	LOC491263	exon	1	5396372	chr36	22174018	TTN	exon	1
5408853	chr9	49689712	LOC491263	exon	1	5456537	chr36	22188584	TTN	exon	1
5376115	chr10	17293142	CRELD2	exon	1	5448217	chr36	22342221	TTN	exon	1
5392566	chr10	17315660	ZBED4	exon	1	5427184	chr37	11056915	FZD7	exon	1
5405579	chr10	17317128	ZBED4	exon	1	5264676	chr37	11060880	FZD7	3'UTR	4
5424720	chr10	65767564	MEIS1	exon	1	5393171	chr37	14891106	DGBF	exon	1
5361984	chr10	65767610	MEIS1	exon	1	5378503	chr37	30726961	DLGAP2	exon	2
5401725	chr12	1575317	AGER	exon	1	5383378	chr37	30727011	DLGAP2	exon	2
5470284	chr12	1580787	PBX2	exon	1	5461333	chr38	21994642	DCAF8	exon	1
5475734	chr12	2663493	RXRB	exon	1	5363201	chrX	25098715	IL1RAPL1	exon	1
5430031	chr13	37327230	FAM83H	exon	1	5431063	chrX	101589449	LOC611589	exon	1
5473451	chr13	37818713	CPSF1	exon	1	5439279	chrX	101589762	LOC611589	exon	1
5302159	chr14	39640741	SNX10	exon	1	5421992	chrX	122175828	PLXNA3	exon	1

Table 3. List of 144 polymorphic SINEs within the genes in free regions. We hypothesize these insertions are likely to be negatively selected.

Discussion

The 1376 regions in the dog genome that are free of retrotransposons and are over 10kb long were found to be rich in genic sequences. A subset of free regions harbored overrepresented genes coding for the transcription factors that regulate developmental processes. Despite the fact that these regions contain genes that are crucial to organismal development, polymorphic SINEs do insert themselves without positional bias in these regions, with a similar frequency as the entire genome.

Acknowledgements

We thank the current and former members of the Sutter lab for their contribution: Bryn Batin, Kristen Simental, Cassandra Ward, Jonathan Specht, Jose Salome Correa, Jordan Orosco, Nancy Kalaj, Hooman Moghadam, Max Tomlinson, and Jordan Whitney. We also thank dog owners and breeders who provided the DNA samples. This work would not be possible without them. Research reported in this poster 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

Clark LA, Wahl JM, Rees CA, Murphy KE. 2006. From The Cover: Retrotransposon insertion in SILV is responsible for merle patterning of the domestic dog. *Proceedings of the National Academy of Sciences*. 103(5):1376–81

Lin L, Faraco J, Li R, Kadotani H, Rogers W, et al. 1999. The Sleep Disorder Canine Narcolepsy Is Caused by a Mutation in the Hypocretin (Orexin) Receptor 2 Gene. *Cell*. 98(3):365–76

Simons C. 2005. Transposon-free regions in mammalian genomes. *Genome Research*. 16(2):164–72

Wang W. 2005. Short interspersed elements (SINEs) are a major source of canine genomic diversity. *Genome Research*. 15(12):1798–1808