



HAL
open science

Long-read assembly of a Great Dane genome highlights the contribution of GC-rich sequence and mobile elements to canine genomes

Julia Halo, Amanda Pendleton, Feichen Shen, Aurélien Doucet, Thomas Derrien, Christophe Hitte, Laura Kirby, Bridget Myers, Elzbieta Sliwerska, Sarah Emery, et al.

► To cite this version:

Julia Halo, Amanda Pendleton, Feichen Shen, Aurélien Doucet, Thomas Derrien, et al.. Long-read assembly of a Great Dane genome highlights the contribution of GC-rich sequence and mobile elements to canine genomes. 2020. hal-03007309

HAL Id: hal-03007309

<https://hal.science/hal-03007309>

Preprint submitted on 5 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

1 Long-read assembly of a Great Dane genome highlights the contribution of GC-rich sequence
2 and mobile elements to canine genomes

3
4 Julia V. Halo^{1,2*}, Amanda L. Pendleton^{2*}, Feichen Shen^{2*}, Aurélien J. Doucet², Thomas Derrien³,
5 Christophe Hitte³, Laura E. Kirby², Bridget Myers², Elzbieta Sliwerska², Sarah Emery², John V.
6 Moran^{2,4}, Adam R. Boyko⁵, Jeffrey M. Kidd^{2,6,7}

7
8 ¹Department of Biological Sciences, Bowling Green State University, Bowling Green, OH, USA

9 ²Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

10 ³Univ. Rennes 1, CNRS, IGDR – UMR 6290, F-35000 Rennes, France

11 ⁴Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA

12 ⁵Department of Biomedical Sciences, Cornell University, Ithaca, New York, USA.

13 ⁶Department Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor,
14 MI USA

15
16 *These authors contributed equally

17
18 ⁷Correspondence should be addressed to Jeffrey M. Kidd at jmkidd@umich.edu

19

20

21 **Key words**

22 *Canis familiaris*, long-read assembly, mobile elements, structural variation

23 **Abstract**

24 Technological advances have allowed improvements in genome reference sequence assemblies.
25 Here, we combined long- and short-read sequence resources to assemble the genome of a female
26 Great Dane dog. This assembly has improved continuity compared to the existing Boxer-derived
27 (CanFam3.1) reference genome. Annotation of the Great Dane assembly identified 22,182
28 protein-coding gene models and 7,049 long non-coding RNAs, including 49 protein-coding
29 genes not present in the CanFam3.1 reference. The Great Dane assembly spans the majority of
30 sequence gaps in the CanFam3.1 reference and illustrates that 2,151 gaps overlap the
31 transcription start site of a predicted protein-coding gene. Moreover, a subset of the resolved
32 gaps, which have an 80.95% median GC content, localize to transcription start sites and
33 recombination hotspots more often than expected by chance, suggesting the stable canine
34 recombinational landscape has shaped genome architecture. Alignment of the Great Dane and
35 CanFam3.1 assemblies identified 16,834 deletions and 15,621 insertions, as well as 2,665
36 deletions and 3,493 insertions located on secondary contigs. These structural variants are
37 dominated by retrotransposon insertion/deletion polymorphisms and include 16,221 dimorphic
38 canine short interspersed elements (SINECs) and 1,121 dimorphic long interspersed element-1
39 sequences (LINE-1_Cfs). Analysis of sequences flanking the 3' end of LINE-1_Cfs (*i.e.*, LINE-
40 1_Cf 3'-transductions) suggests multiple retrotransposition-competent LINE-1_Cfs segregate
41 among dog populations. Consistent with this conclusion, we demonstrate that a canine LINE-
42 1_Cf element with intact open reading frames can retrotranspose its own RNA and that of a
43 SINEC_Cf consensus sequence in cultured human cells, implicating ongoing retrotransposon
44 activity as a driver of canine genetic variation.

45 **Significance**

46 Advancements in long-read DNA sequencing technologies provide more comprehensive views
47 of genomes. We used long-read sequences to assemble a Great Dane dog genome that provides
48 several improvements over the existing reference derived from a Boxer dog. Assembly
49 comparisons revealed that gaps in the Boxer assembly often occur at the beginning of protein-
50 coding genes and have a high-GC content, which likely reflects limitations of previous
51 technologies in resolving GC-rich sequences. Dimorphic LINE-1 and SINEC retrotransposon
52 sequences represent the predominant differences between the Great Dane and Boxer assemblies.
53 Proof-of-principle experiments demonstrated that expression of a canine LINE-1 could promote
54 the retrotransposition of itself and a SINEC_Cf consensus sequence in cultured human cells.
55 Thus, ongoing retrotransposon activity may contribute to canine genetic diversity.

56 **Introduction**

57 The domestic dog (*Canis lupus familiaris*) is an established model system for studying the
58 genetic basis of phenotype diversity, assessing the impact of natural and artificial selection on
59 genome architecture, and identifying genes relevant to human disease. The unique genetic
60 structure of dogs, formed as a result of trait selection and breed formation, has particularly aided
61 genetic mapping of dog traits (1, 2).

62
63 Canine genetics research has taken advantage of a growing collection of genomics tools
64 including high density single nucleotide polymorphism (SNP) arrays, thousands of genome
65 sequences acquired with short-read technologies, the existence of rich phenotype information,
66 and the availability of DNA obtained from ancient samples (3). This research has relied on the
67 reference genome, CanFam, derived from a Boxer breed dog named Tasha and originally
68 released in 2005 (4). The CanFam assembly was constructed at the end of the first phase of
69 mammalian genome sequencing projects and used a whole-genome shotgun approach that
70 included the end-sequencing of large genomic DNA inserts contained within bacterial artificial
71 chromosome (BAC) and fosmid libraries in conjunction with a limited amount of finished BAC
72 clone sequence (4). Subsequent analyses of CanFam and other genomes sequenced in this
73 manner have suggested that there is an incomplete representation of duplicated and repetitive
74 sequences in the resultant assemblies. Although multiple updates have improved the CanFam
75 assembly, yielding the current CanFam3.1 reference assembly (5), numerous assembly errors,
76 sequence gaps, and incomplete gene models remain. Thus, a more complete and comprehensive
77 dog genome will aid the identification of mutations that cause phenotypic differences among
78 dogs and enable continued advances in comparative genomics (6).

79
80 Genome analyses have revealed that canine genomes contain an elevated number of high GC-
81 content segments relative to other mammalian species (7, 8). Genetic recombination may
82 contribute to the evolution of these segments. Studies across a number of mammalian species
83 have indicated that genetic recombination events cluster in specific regions known as hotspots
84 (9). In many species, the PRDM9 zing-finger protein binds to specific nucleotide sequences to
85 promote the initiation of recombination (10-12). In addition to cross-overs, the molecular
86 resolution of recombination events involves gene conversion (13). Gene conversion shows a bias

87 in favor of copying G/C sequences and makes an important contribution to the evolution of
88 genome content (14, 15). Due to gene conversion events and changes in the DNA binding
89 domains of PRDM9, the locations of recombination hotspots in many species are not stable over
90 evolutionary time (16, 17). However, dogs lack a functional PRDM9 protein (18), and canine
91 recombination maps indicate that recombination events are concentrated in GC-rich segments
92 that reside near gene promoters (19, 20). Thus, the observed distribution of GC-rich sequence
93 segments in the canine genome may be a consequence of the stable recombination landscape in
94 canines.

95
96 Analysis of the CanFam3.1 reference has demonstrated that a large fraction of the dog genome
97 has resulted from the expansion of transposable elements belonging to the short and long
98 interspersed element (SINE and LINE) families. Fine mapping has implicated mobile element
99 insertions, and associated events such as retrogene insertions, as the causal mutation underlying
100 morphological differences, canine diseases, and selectively bred phenotypes (21-31). A
101 comparison between the Boxer-derived CanFam reference and a low coverage (~1.5x) draft
102 genome from a Poodle identified several thousand dimorphic copies of a recently active lysine
103 transfer RNA (tRNA)-derived canine SINE element, SINEC_Cf, implying a variation rate 10-
104 100 fold higher than observed for still active SINE lineages in humans (32). Similarly, insertions
105 derived from a young canine LINE-1 lineage, L1_Cf, were found to be >3-fold more prevalent
106 than L1Hs, the active LINE-1 lineage found in humans (32, 33). However, the assembly of long
107 repetitive sequences with a high nucleotide identity is technically challenging, leaving many
108 LINE sequences incorrectly represented in existing reference genomes. Consequently, the
109 biological impact of these elements has remained largely unexplored and the discovery of
110 dimorphic canine LINE-1 sequences is limited to a few reports (32, 34, 35).

111
112 Following the era of capillary sequencing, genome reference construction shifted toward high
113 coverage assemblies that utilized comparatively short-sequencing reads. These approaches
114 offered a great reduction in cost and an increase in per-base accuracy, but still were largely
115 unable to resolve duplicated and repetitive sequences, often yielding assemblies that contained
116 tens of thousands of contigs (36). Methods based on linked-read or chromosome conformation
117 sequencing are capable of linking the resulting contigs into larger scaffolds, including entire

118 chromosome arms, but these scaffolds are typically littered with sequence gaps reflecting the
119 poor representation of repetitive sequences (37-39). Here, we analyze the genome of a female
120 Great Dane named Zoey that we sequenced using PacBio long-read technology. We integrated
121 this long-read data with additional sequencing resources, including standard high coverage short-
122 read sequence data, as well as sequence data derived from mate-pair and pooled fosmid libraries,
123 to generate a high-quality assembly. Using this new assembly, we annotate novel gene structures
124 and GC-rich sequences that are absent from CanFam3.1 and under-represented in existing
125 Illumina canine short-read sequence datasets. We demonstrate that gaps in the CanFam3.1
126 assembly are enriched with sequences that have an extremely high GC content and that overlap
127 with transcription start sites and recombination hotspots. We identify thousands of mobile
128 element insertions, including intact LINE-1 copies, and make use of our fosmid library to
129 subclone an intact L1_Cf element. We demonstrate that a cloned canine L1_Cf is capable of high
130 levels of retrotransposition of its own mRNA (*in cis*) and can drive the retrotransposition of a
131 consensus SINEC_Cf RNA (*in trans*) in cultured human cells. Our analysis provides a more
132 complete view of the canine genome and demonstrates that the distribution of extremely GC-rich
133 sequences and the activity of mobile elements are major factors affecting the content of canine
134 genomes.

135

136 **Results**

137 **Long-read assembly of a Great Dane genome**

138 We performed a genome assembly of a female Great Dane, Zoey, using multiple genome
139 sequencing resources that included a standard Illumina short-read sequencing library, a 3 kb
140 Illumina mate-pair sequencing library, sequences from a pooled fosmid library, and ~50X raw
141 long-read coverage generated using the PacBio RSII system. PacBio long reads were assembled
142 using the Falcon assembler (40), yielding 2,620 primary contigs longer than 1 kbp that
143 encompassed 2.3 Gbp of sequence. In addition, 6,857 secondary contigs, with a total length of
144 178.5 Mbp, that represent the sequence of heterozygous alleles were assembled (see
145 *Supplementary Information*, Section 1).

146

147 The assembly process is based on detecting overlaps among sequencing reads. As a result, reads
148 that end in long stretches of sequence which map to multiple genomic locations and that have

149 high sequence identity, can give rise to chimeric contigs that falsely conjoin discontinuous
150 genomic segments. Using Illumina mate-pair and fosmid pool data from Zoey, clone end
151 sequences from the Boxer Tasha, and alignments to the existing CanFam3.1 assembly, we
152 identified 20 contigs that appeared to be chimeric. We split these contigs at the chimeric
153 junctions, yielding a total of 2,640 contigs with an N_{50} length of 4.3 Mbp and a maximum contig
154 length of 28.8 Mbp. As expected, alignment against the CanFam3.1 assembly indicated
155 comprehensive chromosome coverage (Figure 1). Consistent with the problems in assembly
156 caused by segmental duplications, we found that long contigs (> 3 Mbp) ended in duplicated
157 sequence greater than 10 kbp more often than expected by chance ($p < 0.001$ by permutation, see
158 *Supplementary Information*, Section 1).

159
160 Alignment of the 2,640 contigs and the raw PacBio reads against the CanFam3.1 assembly
161 revealed apparent gaps between contigs, many of which were spanned by PacBio reads.
162 Reasoning that these reads may have been excluded from the assembly due to length cutoff
163 parameters used in the Falcon pipeline, we performed a locus specific assembly utilizing the
164 Canu assembler (v1.3) (41.). This process yielded 373 additional contigs with a total length of
165 10.5 Mbp and a N_{50} length of 30 kbp. Based on the mapping of the Zoey derived mate-pair
166 sequences and end sequences from the Tasha-derived fosmid and BAC libraries, we linked the
167 2,640 primary contigs and 373 gap-filling contigs into scaffolds (42). Gap-filling contigs that
168 were not linked using paired reads were excluded from further analysis, resulting in a total of
169 1,759 scaffolds with a N_{50} of 21 Mbp. Scaffolds were assigned to chromosomes and ordered
170 based on alignment to CanFam3.1. Sequences that appeared to represent allelic variants based on
171 sequence identity and read depth were removed, yielding a chromosomal representation that
172 included 754 unlocalized sequences (see *Supplementary Information*, Section 1).

173

174 **Annotation of genome features**

175 We identified segmental duplications in the Zoey and CanFam3.1 assemblies based on assembly
176 self-alignment (43) and read-depth (44) approaches (see *Supplementary Information*, Section 3).
177 Although the number of duplications is similar in each genome, the Zoey assembly contains a
178 smaller total amount of sequence classified as “duplicated”, which likely reflects the continued
179 challenges in properly resolving duplications longer than 10 kbp (Table 1) (45). To compare the

180 large scale organization of the assemblies, we constructed reciprocal liftOver tracks that identify
181 corresponding segments between the CanFam3.1 and Zoey assemblies (46). Based on this
182 comparison, we identified 44 candidate inversions >5kb. Of these candidate large inversions,
183 68% (30 of 44) were associated with duplicated sequences. The X chromosome, which
184 contributes 5.3% of the genome length, contains 41% (18/44) of the predicted inversions.

185
186 We created a new gene annotation based on previously published RNA sequencing data using
187 both genome-guided and genome free approaches (47-49) (see *Supplementary Information*,
188 Section 2). Following filtration, this process resulted in a final set of 22,182 protein coding gene
189 models; forty-nine of these gene models are absent from the CanFam3.1 assembly. Full-length
190 matches were found for only 84.9% (18,834) of all protein-coding gene models, while near-full
191 length alignments were found for 93% (20,670) of the models. We additionally annotated 7,049
192 long non-coding RNAs (50), including 84 with no or only partial alignment to CanFam3.1.
193 Using existing RNA-Seq data (5), we estimated expression values for each protein-coding gene
194 across eleven tissues and report the results as tracks on a custom UCSC Genome Browser
195 assembly hub (51) (Figure 2). The assembly hub illustrates correspondence between the
196 CanFam3.1 and Zoey assemblies and displays the annotation of additional features including
197 structural variants, segmental duplications, common repeats, and BAC clone end-sequences (see
198 *Supplementary Information*, Section 7).

199
200 **Resolved assembly gaps include GC-rich segments underrepresented in Illumina libraries**

201 Alignment indicates that 12,806 of the autosomal gaps in CanFam3.1 are confidently localized to
202 a unique location in the Zoey genome assembly. In total, 16.8% (2,151) of the gap segments
203 overlap with a transcription start site of a protein-coding gene, which makes it possible to better
204 understand the importance of these previously missing sequences in canine biology (5, 52).
205 Surprisingly, analysis of unique k-mer sequences that map to the CanFam3.1 gap sequences
206 suggested that these DNA segments often are absent from existing Illumina short-read data sets,
207 even though analysis of DNA from the same samples using a custom array comparative genomic
208 hybridization platform indicates their presence. Interrogation of read-pair signatures also suggest
209 that these sequences are systematically depleted in Illumina libraries, which is due to their
210 extreme GC-rich sequence composition (see *Supplementary Information*, Section 4).

211
212 The sequences corresponding to gaps in CanFam3.1 have an extremely high GC content, with a
213 median GC content of 67.3%, a value substantially higher than the genome-wide expectation of
214 39.6% (Figure 3). Given the relationship between GC content and recombination in dogs (19,
215 20), we examined the distance between CanFam3.1 gap sequences and recombination hotspots.
216 We found that 11.8% of gap segments (1,457 of 12,304 on the autosomes) are located within 1
217 kbp of a hotspot, compared to only 2.9% of intervals expected by chance. These patterns are
218 driven by a subset of segments that have the most extreme GC content. We identified 5,553
219 segments with a GC content greater than that obtained from 1,000 random permutations. These
220 extreme-GC segments span a total of 4.03 Mbp in the Zoey assembly, have a median length of
221 531 bp, a median GC content of 80.95%, and are located much closer to transcription start sites
222 (median distance of 290 bp) and recombination hotspots (median distance of 68.7 kbp) than
223 expected by chance.

224 225 **Mobile elements account for the majority of structural differences between canine genomes**

226 We compared the CanFam3.1 and Zoey assemblies to identify insertion-deletion differences at
227 least 50bp in length. After filtering variants that intersect with assembly gaps or segmental
228 duplications, we identified 16,834 deletions (median size: 207 bp) and 15,621 insertions (median
229 size of 204 bp) in the Zoey assembly relative to CanFam3.1 (see *Supplementary Information*,
230 Section 5). In total, these structural variants represent 13.2 Mbp of sequence difference between
231 the two assemblies. The length distribution of the detected variants shows a striking bimodal
232 pattern with clear peaks at ~200 bp and ~6 kbp, consistent with the size of canine SINEC and
233 LINE-1 sequences (Figure 4). We inspected the sequence of the events in the 150-250 bp size
234 range and found that 7,298 deletions and 6,071 insertions were dimorphic SINEC sequences.
235 Additionally, LINE-1 sequences accounted for 339 deletions and 581 insertions longer than 1
236 kbp.

237
238 Our assembly also contains 6,857 secondary contigs, which represent alternative sequences at
239 loci where Zoey is heterozygous for a structural variant. Alignment of these secondary contigs
240 against the CanFam3.1 assembly yielded an additional 2,665 deletion and 3,493 insertion events,
241 encompassing a total of 2.67 Mbp of sequence. We further inspected the sequence of these

242 variants and found 1,259 deletions and 1,593 insertions consistent with dimorphic SINEC
243 elements, and 75 deletions and 126 insertions consistent with dimorphic LINE-1 elements.
244 Together, comparison of the Zoey and CanFam3.1 genomes identified at least 16,221 dimorphic
245 SINEC and 1,121 dimorphic LINE-1 sequences (see *Supplementary Information*, Section 5).

246
247 LINE-1 transcription often bypasses the polyadenylation signal encoded within the element,
248 resulting in the inclusion of flanking genomic sequence in the LINE-1 RNA (53-55). Thus, after
249 retrotransposition, the resulting 3'-transductions can be used as sequence signatures to identify
250 the progenitor source elements of individual LINE-1 insertions (56, 57). We identified 18
251 transduced sequences among the dimorphic LINE-1 sequences in our data set. Of these
252 transduced sequences, 17 aligned elsewhere in the genome at a location that is not adjacent to an
253 annotated LINE-1. This includes a pair of LINE-1 copies on chr25 and chrX which share the
254 same transduced sequence, as well as a locus on chr19 that has the same transduction as a
255 duplicated sequence present on chr2 and chr3. Such "parentless" 3'-transductions suggest the
256 presence of additional dimorphic LINE-1 sequences that are capable of retrotransposition (see
257 *Supplementary Information*, Section 5).

258

259 **Canine genomes contain LINE-1s and SINEs capable of retrotransposition**

260 The high degree of dimorphic LINE-1 and SINEC sequences found between the two assemblies
261 suggests that mobile element activity represents a mutational process that is ongoing in canines.
262 The canine LINE-1 (L1_Cf) consensus sequence contains segments of GC-rich sequence and
263 homopolymer runs, including a stretch of 7 'C' nucleotides in the ORF1p coding sequence that
264 likely are prone to errors incurred during DNA replication, PCR, and sequencing. Thus, a
265 bioinformatic search for L1_Cf sequences with intact open reading frames is biased by
266 uncorrected sequencing errors. We therefore searched the Zoey assembly for sequences that have
267 long matches with low sequence divergence from the L1_Cf consensus. The Zoey assembly
268 encodes 837 L1_Cf sequences that have less than 2% divergence and are greater than 99.4% of
269 full-length; an additional 169 elements are present on the secondary contigs. This set includes
270 187 full length LINE-1s, of which, 31 were found in secondary contigs. For comparison, these
271 values represent a 65% increase over the 113 elements present in CanFam3.1 that meet the same
272 criteria (see *Supplementary Information*, Section 5).

273
274 To more thoroughly characterize canine LINE-1 copies that may remain active, we isolated and
275 sequenced individual fosmid clones from Zoey predicted to contain full-length LINE-1s. We identified
276 one sequence, from fosmid clone 104_5 on chr1 (L1_Cf-104_5), possessing intact open reading
277 frames, which encode the ORF1p and ORF2p predicted proteins, that lack mutations expected to
278 disrupt protein function (see *Supplementary Information*, Section 6). We subcloned this element
279 for functional analysis in a cultured cell assay that uses an indicator cassette that is only
280 expressed following a successful round of retrotransposition (58-60), yielding G418-resistance
281 foci. We found that the L1_Cf-104_5 element is capable of retrotransposition of its own mRNA
282 *in cis* in human HeLa cells (Figure 5 and *Supplementary Information*, Section 6).

283
284 SINE sequences are non-autonomous elements that utilize the function of LINE-1 ORF2p to
285 mediate their retrotransposition *in trans* (60, 61). To test the capability of L1_Cf-104_5 to
286 mobilize SINE RNA *in trans*, we constructed a second reporter vector containing the SINEC_Cf
287 consensus sequence marked by an appropriate indicator cassette (61). We found that expression
288 of L1_Cf-104_5 was capable of mobilizing both canine SINEC and human Alu RNAs *in trans*
289 (Figure 5 and *Supplementary Information*, Section 6).

290

291 **Discussion**

292 Due to their unique breed structure, history of selection for disparate traits, and extensive
293 phenotypic data, dogs are an essential model for dissecting the genetic basis of complex traits
294 and understanding the impact of evolutionary forces on genome diversity. The era of long-read
295 sequencing is revolutionizing genomics by enabling a more complete view of genomic variation
296 (62). Here, we describe the assembly and annotation of the genome of Great Dane dog and
297 compare it with the Boxer-derived CanFam3.1 reference assembly. Comparison of our Great
298 Dane genome to the CanFam3.1 reference revealed several key findings important to canine
299 genome biology. Several other long-read assemblies of canines are planned or have been recently
300 released (63, 64). The availability of these resources, along with other long-read canine
301 assemblies that are planned or have been recently released (63, 64), will provide significant
302 benefits to the canine genomics community.

303

304 Our Great Dane assembly has improved sequence continuity, resolves novel gene structures, and
305 identifies several features important to canine genome biology. For example, we created a new
306 gene annotation that includes 49 predicted protein coding genes that are absent from the
307 CanFam3.1 reference genome. Our analysis also identified 2,151 protein-coding gene models
308 whose transcription start position corresponds to a gap in the CanFam3.1 assembly. This finding
309 largely resolves prior observations that many dog genes appear to have incomplete first exons
310 and promoters (5, 6). Analysis of the Great Dane assembly further revealed that gaps in the
311 CanFam3.1 assembly are enriched for sequence that has extremely high-GC content, providing a
312 probable explanation of their absence from the CanFam3.1 assembly (65).

313

314 The presence of extremely GC-rich segments likely reflects a key aspect of canine genome
315 biology. In contrast to humans and many other mammals, genetic recombination in canines is
316 targeted towards gene promoter regions due to the absence of a functional *PRDM9* gene (20). In
317 other species, the PRDM9 protein binds to specific nucleotide sequences and targets the
318 initiation of recombination to distinct loci in the genome. It has been hypothesized that
319 recombination in dogs is instead localized by general chromatin marks, which are associated
320 with promoters, resulting in a fine-scale genetic map that is more stable over evolutionary time
321 (19, 20). In addition to crossing-over, recombination events result in gene conversion, a process
322 with a bias in favor of G/C alleles. Biased gene conversion can be modeled as positive selection
323 in favor of G/C alleles at a locus (14, 15, 66) and has been previously proposed as an explanation
324 for the unusual GC content of the dog genome (19, 20). Our analysis indicates that the GC rich
325 segments associated with recombination hotspots are larger than expected previously. These
326 expanded segments have an unknown effect on the expression of their associated gene, have
327 been largely absent from previous genome assemblies, and are depleted from Illumina
328 sequencing data. A more extensive examination of the long-term consequence of stable
329 recombination hotspots on genome sequence structure will require assessment of genomes of
330 other species which lack PRDM9 using long-read technologies.

331

332 Long-read sequencing offers a less biased view of structural variation between genomes,
333 particularly for insertions (67). The profile of genomic structural variation between the Zoey and
334 CanFam3.1 assemblies is dominated by dimorphic SINEC and LINE-1 sequences, with 16,221

335 dimorphic SINEC and 1,121 dimorphic LINE-1 sequences. Although analogies between humans
336 and dogs can be problematic (68), a comparison with humans illustrates the magnitude of the
337 mobile element diversity found between the Great Dane and Boxer genome assemblies. In terms
338 of human mobile element diversity, the 1000 Genomes Project estimates that humans differ from
339 the reference genome by an average of 915 Alu insertions and 128 LINE-1 insertions (69). A
340 recent study collated these findings, along with other published data sets, and identified a total of
341 13,572 dimorphic Alu elements in humans (70), though we note that these estimates are based on
342 Illumina sequencing data, which has limitations in mapping to repetitive regions and in fully
343 capturing insertion alleles (67). Finally, an approach specifically designed to identify dimorphic
344 human LINE-1 insertions utilizing long-read sequencing data identified 203 non-reference
345 insertions in the benchmark sample NA12878, of which 123 which were greater than 1 kbp in
346 length (71).

347
348 Illumina sequencing data indicate that Zoey differs from the CanFam3.1 reference at 3.57
349 million single nucleotide variants (SNVs). This number is lower than the number of differences
350 typically found in a globally diverse collection of human genomes (4.1-5.0 million SNVs) (69),
351 and is comparable to the number found in the NHLBI TOPMed data set (72) (median of 3.3
352 million SNVs among 53,831 humans sequenced as part of the National Heart, Lung, and Blood
353 Institute's Trans-Omics for Precision Medicine program). Relative to the number of SNVs, the
354 level of LINE-1 and SINEC dimorphism we found between two dog genomes is
355 disproportionately large. This total represents an approximately 17-fold increase in SINE
356 differences (16,221/915) and an eight-fold increase in LINE differences (1,121/128) compared to
357 the numbers found among humans. Remarkably, more dimorphic SINEs were found between
358 these two breed dogs than have been found in studies of thousands of humans (70, 73). Our data
359 will aid systematic studies of the potential contribution of these elements to canine phenotypes,
360 including cancers (74).

361
362 Further study is required to determine the relative contribution of (i) new insertions in breeds or
363 populations versus (ii) the assortment of segregating variants that were present in the progenitor
364 populations. However, our study suggests that retrotransposition is an ongoing process that
365 continues to affect the canine genome. We provide proof-of-principle evidence that dog genomes

366 contain LINE-1 and SINEC elements that are capable of retrotransposition in a cultured cell
367 assay. We also identified two LINE-1 lineages with the same 3'-transduced sequence associated
368 with multiple elements, suggesting the presence of multiple canine LINE-1s that are capable of
369 spawning new insertions. Additionally, analysis of 3'-transduction patterns suggests the presence
370 of additional active LINE-1s in canines that have yet to be characterized. Thus, a full
371 understanding of canine evolution and phenotypic differences requires consideration of these
372 important drivers of genome diversity.

373

374 **Methods**

375 Genome assembly and analysis utilized long-read and short data from a female Great Dane
376 named Zoey, a pooled fosmid library (75) constructed from Zoey, sequence data generated from
377 a female Boxer, named Tasha, as part of the CanFam genome assembly (4), and results from a
378 custom comparative genomics hybridization array (array-CGH). Data accessions and detailed
379 methods are available in the Supplementary Information.

380

381 Genome assembly of ~50-fold whole-genome, single-molecule, real-time sequencing (SMRT)
382 data from Zoey was performed on DNAnexus using the FALCON 1.7.7 pipeline (40) and the
383 Damasker suite (76). Chimeric contigs were identified based on mapped reads from the Zoey
384 mate-pair jumping library, the Zoey fosmid pools, the Tasha BAC end sequences, and Tasha
385 fosmid end sequences. Regions that showed a lack of concordant paired end coverage were
386 identified as potential chimeric junction sites and split apart prior to scaffolding. Primary contigs
387 were supplemented with contigs obtained from a local assembly of reads aligning to gaps
388 between contigs on CanFam3.1 using Canu v1.3 (41). Contigs were linked into scaffolds using
389 mapping of the Zoey mate -data, Tasha BAC end sequence data and Tasha fosmid data using the
390 BESST scaffolding algorithm (version 2.2.7)(42) and assigned to chromosomes based on
391 alignment to CanFam3.1. Chain files for use with the UCSC liftOver tool were constructed based
392 on blat (46) alignments. A UCSC TrackHub hosting the Zoey assembly, as well as relevant
393 annotations of both the Zoey assembly and CanFam3.1 is available at

394 https://github.com/KiddLab/zoey_genome_hub

395

396 Common repeats in both the CanFam3.1 and Zoey assemblies were identified using
397 RepeatMasker version 4.0.7 with option ‘-species dog’, using the rmblastn (version 2.2.27+)
398 search engine and a combined repeat database consisting of the Dfam_Consensus-20170127 and
399 RepBase-20170127 releases. Self-alignment analysis of each assembly was performed using
400 SEDEF (43) with default parameters. Results were filtered for alignments at least 1kb in length
401 and at least 90% sequence identity. Read-depth analysis was performed using fastCN as
402 described previously (44). Copy-number estimates were constructed in non-overlapping
403 windows each containing 3 kbp of unmasked sequence. Segmental duplications were identified
404 as runs of four windows in a row with an estimated copy-number ≥ 2.5 . To provide an unbiased
405 assessment of duplication content, read-depth analysis was performed based on Illumina data from
406 Penelope, an Iberian Wolf, in addition to sequences from Zoey and Tasha.

407
408 Forty-two canine RNA-Seq runs representing eleven tissue types were used to annotate genes in
409 the Zoey genome (5). *De novo* gene models were created based on alignment of RNA-Seq reads
410 using Cufflinks (v2.2.1) (47, 48) and in a non-reference guided fashion using Trinity (v2.3.2)
411 (49). Gene models were merged and annotated using PASA-Lite (77) and the transdecoder
412 pipeline (version 5.0.1) (78). Gene names and functional annotations were determined using
413 BLAST2GO (79). Expression levels for each of the 22,182 protein-coding gene models were
414 estimated using Kallisto (version 0.46.0) (80). Long non-coding RNAs in the Zoey genome were
415 identified using the FEELnc program (50).

416
417 To identify large insertion and deletion variants, the Zoey assembly and 6,857 secondary contigs,
418 were aligned to CanFam3.1 using minimap2 (version 2.9-r720) with the -asm5 option (81). The
419 output from the alignment was parsed using the paftools.js program released as part of minimap2
420 to identify candidate variants. Breakpoint coordinates were refined by performing targeted
421 alignment of the flanking and variant sequence for each candidate using AGE (82).

422
423 Individual fosmids containing potentially full-length L1_Cf elements were isolated from pools
424 using a lifting procedure coupled with hybridization of a probe containing digoxigenin (DIG)
425 labeled dUTP. Isolated fosmids were sequenced in small pools via RS II PacBio sequencing and
426 assembled using the HGAP2 software (83). An intact L1_Cf was subcloned from fosmid 104_5,

427 equipped with an *mneoI* retrotransposition indicator cassette, and tested for retrotransposition in
428 HeLa-HA cells (58, 59). The construction of the L1_Cf expression vector and the conditions
429 used to assay for retrotransposition are detailed in *Supplemental Information*, Section 6.

430
431 To monitor SINEC_Cf mobilization, we modified the *Alu neo^{tet}* vector, which contains an active
432 human AluYa5 element equipped with a reporter cassette engineered to monitor the
433 retrotransposition of RNA polymerase III (pol III) transcripts (61). Briefly, the *Alu neo^{tet}* vector
434 consists of a 7SL RNA Pol III enhancer sequence upstream of AluYa5 that is equipped with a
435 ‘backward’ *neo^R* gene under the control of an SV40 promoter. The *neo^R* gene is disrupted by a
436 tetrahymena self-splicing group I intron that is in the same transcriptional orientation as the Alu
437 element. This arrangement only allows the expression of the *neo^R* gene (yielding G418-resistant
438 foci) upon a successful round of retrotransposition in HeLa-HA cells, yielding G418-resistant
439 foci (61). We replaced the AluYa5 sequence with the SINEC_Cf consensus sequence, obtained
440 from Repbase (84). The resultant construct was used to assay SINEC_Cf mobilization, *in trans*,
441 in the presence of either an active human LINE-1 or the newly cloned L1_Cf-104_5 expression
442 plasmid that lacks the retrotransposition indicator cassette (JM101/L1.3Δneo or ADL1Cf-
443 104_5Δneo, respectively). The construction of the SINEC_Cf expression vector and the
444 conditions used to assay for retrotransposition are detailed in *Supplemental Information*, Section
445 6.

446

447 **Competing interests**

448 J.V.M. is an inventor on patent US6150160, is a paid consultant for Gilead Sciences, serves on
449 the scientific advisory board of Tessera Therapeutics Inc. (where he is paid as a consultant and
450 has equity options), and currently serves on the American Society of Human Genetics Board of
451 Directors. A.R.B. is the co-founder and Chief Science Officer of Embark Veterinary.

452

453 **Acknowledgments**

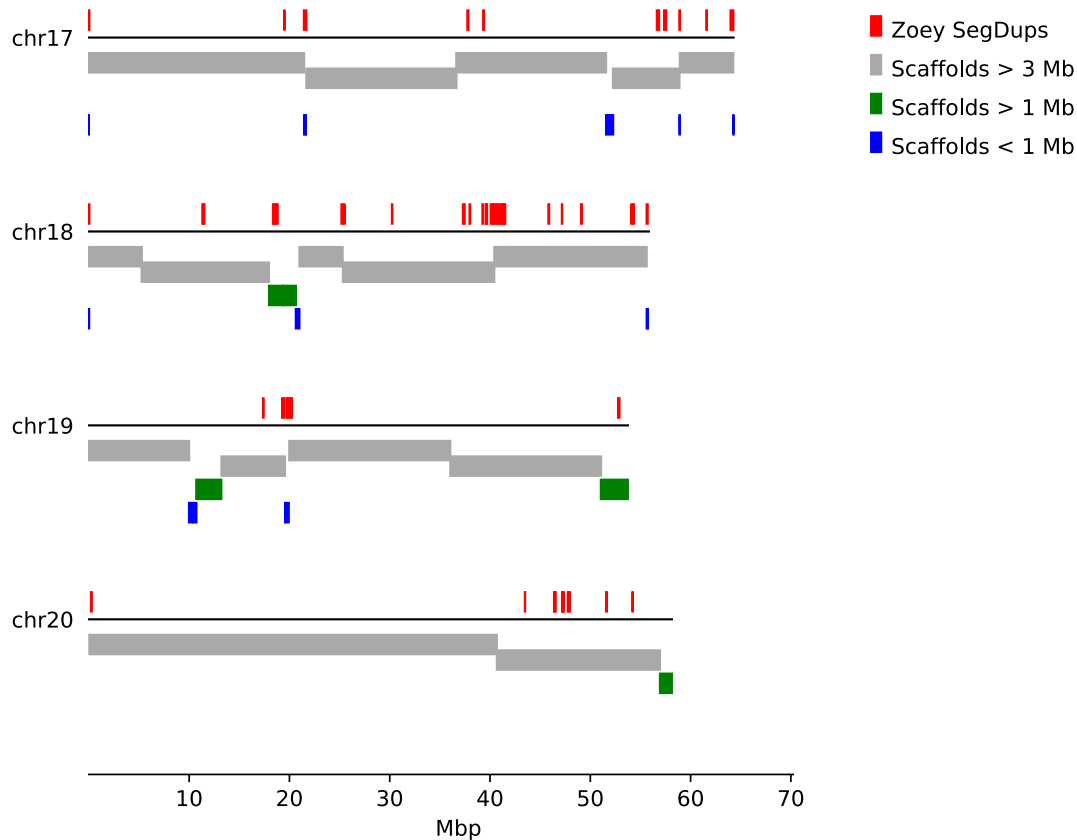
454 This work was supported in part by National Institutes of Health (NIH) grant R01GM103961 to
455 J.M.K. and A.R.B., NIH Academic Research Enhancement Award R15GM122028 to J.V.H., and
456 NIH Training Fellowship T32HG00040 to A.L.P. DNA samples were provided by the Cornell
457 Veterinary Biobank, a resource built with the support of NIH Grant R24GM082910, and the

458 Cornell University College of Veterinary Medicine. Additional DNA samples were kindly
459 provided by Brian Davis, Elaine Ostrander, and Linda Gates. We thank Dorina Twigg, Chai
460 Fungtammasan, Brett Hannigan, Mark Mooney, Dylan Pollard, and DNAnexus for assistance
461 with sequence data processing and the University of Michigan Advanced Genomics Core for
462 assistance with data production. We especially thank Linda Gates for her continued devotion to
463 all Great Danes and her assistance with this project.

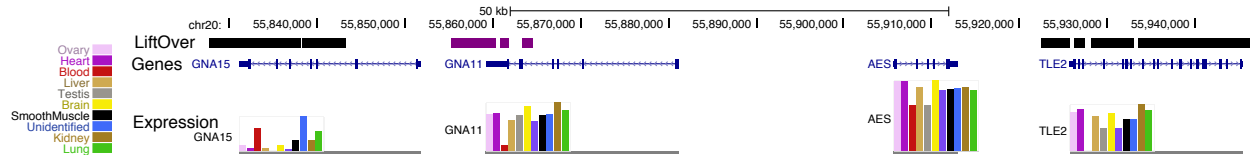
	CanFam3.1 autosomes + X	Zoey autosomes + X
Total length	2,327,633,984	2,326,329,672
<i>non-N</i>	2,317,593,971	2,320,292,846
Number of gaps	19,553	997
Longest contiguous segment	2,428,071	28,813,894
Mean contiguous segment length	118,523	2,239,665
Median contiguous segment length	54,641	1,107,836
N ₅₀ segment length	277,468	4,765,928
Segmental Duplications Genome Alignment, >1 kbp, >90% ID	6,250	6,371
<i>bp</i>	49,339,683	45,425,166
Segmental Duplications Penelope Read Depth	459	468
<i>bp</i>	47,757,534	40,836,807

464 **Table 1: Comparison of the Boxer and Great Dane assemblies.** Presented are general
465 assembly statistics for the primary autosomal and X chromosome sequence of the CanFam3.1
466 and Zoey assemblies. Contiguous segment refers to the length of sequence uninterrupted by an
467 ‘N’ nucleotide. Segmental duplications were identified in each assembly based on an assembly
468 self-alignment and by the depth of coverage of Illumina sequencing reads from Penelope, an
469 Iberian Wolf. See *Supplementary Information*, Section 3 for additional details.

470

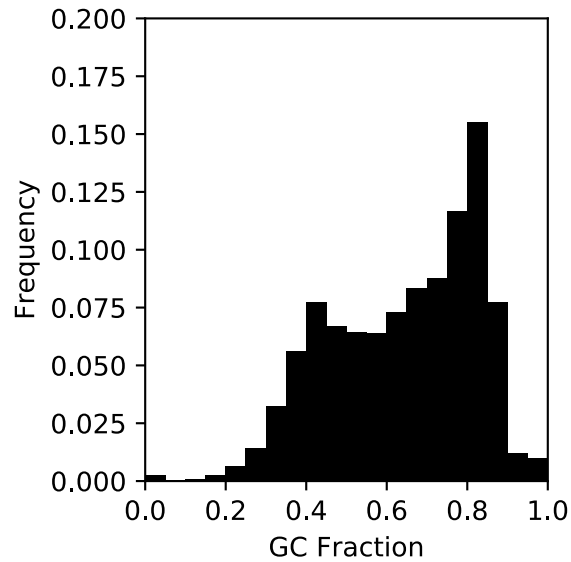


471
472 **Figure 1: Alignment of assembled contigs to the CanFam3.1 genome.** Each of the 2,640
473 primary contigs were aligned to the CanFam3.1 reference genome. Results are shown for four
474 chromosomes. The colored bars below each line indicate the corresponding position of each
475 contig, colored based on their indicated length. Above each line, regions of segmental
476 duplications based on read depth in the Zoey Illumina data are indicated by red boxes.
477 Permutation tests indicate that long contigs end at regions of segmental duplication more often
478 than expected by chance. See *Supplementary Information*, Section 1 for additional details.
479



480
481
482
483
484
485
486
487
488
489
490
491
492

Figure 2: Annotation of genes missing from the CanFam3.1 assembly. Shown is a genome browser view of chr20 on the Zoey assembly is shown. The top track summarizes a comparison between the Zoey and CanFam3.1 assemblies using the UCSC liftOver tool. Black segments show alignment to the corresponding chromosome on the CanFam3.1 assembly. Purple segments match to an unlocalized contig (chrUn_JH374124) in the CanFam3.1 assembly. The large region in the middle between the purple and black segments is absent from the CanFam3.1 assembly. The track below shows the position of four genes in this region annotated using RNA-Seq data: *GNA15*, *GNA11*, *AES*, and *TLE2*. The colored bars below each gene model show the expression levels across different tissues, as indicated by the color key in the left of the figure. See *Supplementary Information*, Section 2 for additional details.



493

494 **Figure 3: CanFam3.1 assembly gaps are enriched for sequence with extreme GC content.**

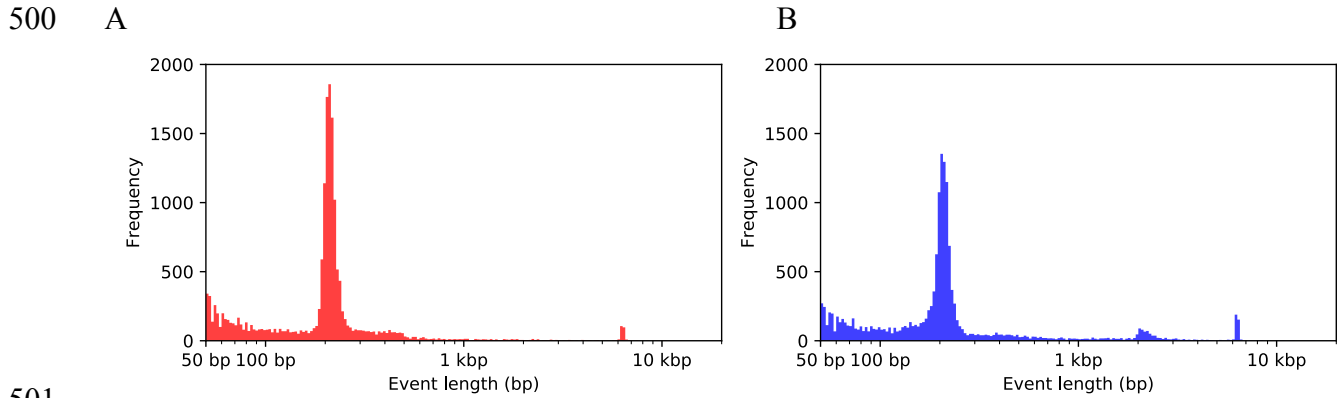
495 Depicted is the distribution of GC content for 12,806 resolved assembly gaps. A subset

496 consisting of 5,553 of the 12,806 segments have a GC content greater than that found in 99% of

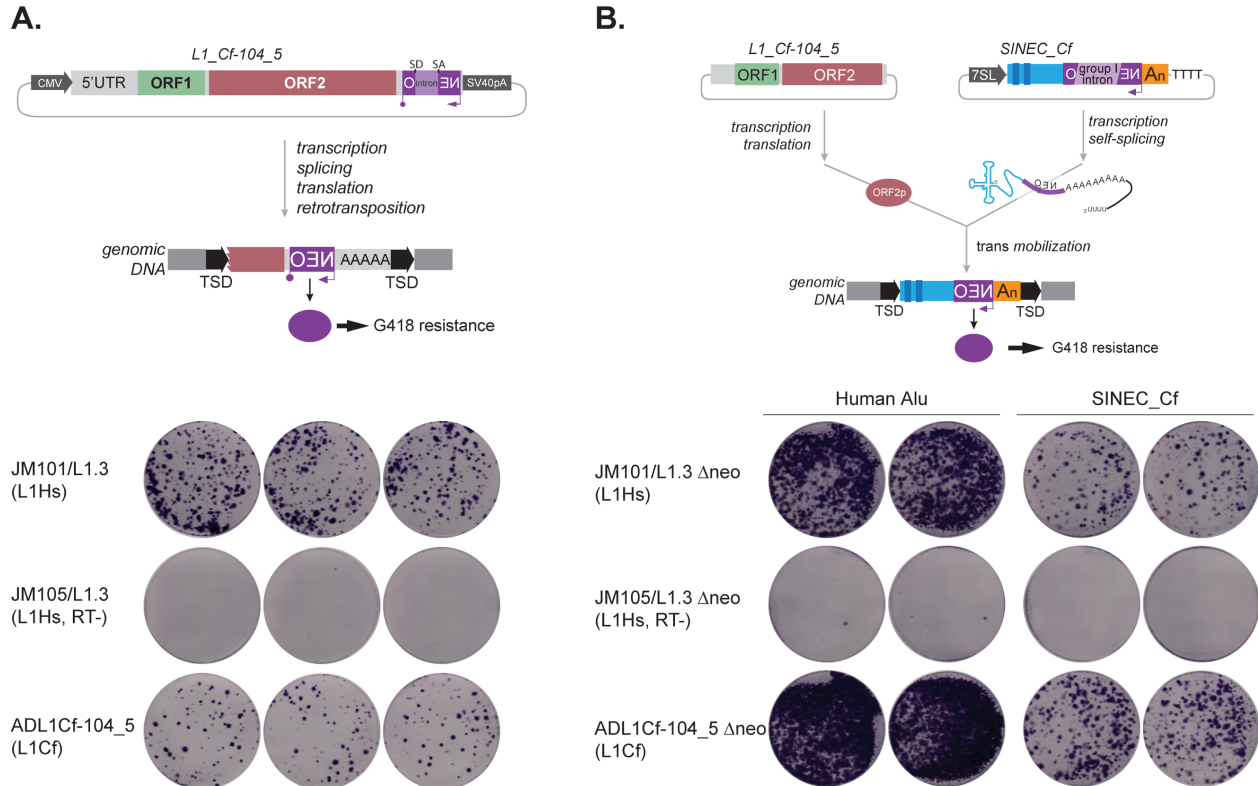
497 randomly selected segments. See *Supplementary Information*, Section 4 for additional details.

498

499



501 **Figure 4: Size of structural variants identified between CanFam3.1 and Zoey assemblies.**
502 Shown are histograms depicting the size distribution of 16,834 deletions (panel A) and 15,621
503 insertions (panel B) between the Zoey and CanFam3.1 assemblies. Variant size is plotted on a
504 logarithmic scale such that the bins in the histogram are of equal size in the log scale. Large
505 increases at ~200bp and ~6kbp indicate the disproportionate contribution of dimorphic LINE1
506 and SINEC sequences to the genetic differences between the two assemblies. See *Supplementary*
507 *Information*, Section 5 for additional details.
508



509

510

Figure 5: Identification of canine LINE-1 and SINEC elements capable of

511 **retrotransposition.** (Panel A, top): A full length L1_Cf equipped with a retrotransposition

512 indicator cassette (*mneoI*) was assayed for retrotransposition in human HeLa-HA cells. TSD,

513 indicates a target site duplication generated upon retrotransposition. (Panel A, bottom): Results

514 of the retrotransposition assay. JM101/L1.3 (positive control) contains an active human LINE-1.

515 JM105/L1.3 (negative control) contains a human LINE-1 that harbors an inactivating missense

516 mutation in the reverse transcriptase domain of ORF2p (85). ADL1Cf-104_5 contains the full-

517 length canine LINE-1 identified in this study. (Panel B, top): A consensus SINEC_Cf element

518 equipped with an indicator cassette to monitor the retrotransposition of RNA polymerase III

519 transcripts (*neo^{tet}*) (61) was assayed for retrotransposition in human HeLa-HA cells in the

520 presence of either an active human LINE-1 or the newly cloned L1_Cf-104_5 sequence that

521 lacks a retrotransposition indicator cassette (JM101/L1.3Δneo or ADL1Cf-104_5Δneo,

522 respectively). (Panel B, bottom): Results of the retrotransposition assay. JM101/L1.3Δneo

523 (positive control) contains an active human LINE-1. JM105/L1.3 Δneo (negative control)

524 contains a human LINE-1 that harbors an inactivating missense mutation in the reverse

525 transcriptase domain of ORF2p (85). ADL1Cf-104_5Δneo contains an active canine LINE-1 (see

526 panel A). The expression of either JM101/L1.3Δneo or ADL1-Cf-105Δneo could drive human

527 Alu and SINEC_Cf retrotransposition. In both assays, the blue stained foci represent G418-
528 resistant foci containing a presumptive retrotransposition event. See *Supplementary Information*,
529 Section 6 for additional details.

530 References

- 531 1. Karlsson EK & Lindblad-Toh K (2008) Leader of the pack: gene mapping in dogs and
532 other model organisms. *Nat Rev Genet* 9(9):713-725.
- 533 2. Boyko AR (2011) The domestic dog: man's best friend in the genomic era. *Genome Biol*
534 12(2):216.
- 535 3. Ostrander EA, Wayne RK, Freedman AH, & Davis BW (2017) Demographic history,
536 selection and functional diversity of the canine genome. *Nat Rev Genet* 18(12):705-720.
- 537 4. Lindblad-Toh K, *et al.* (2005) Genome sequence, comparative analysis and haplotype
538 structure of the domestic dog. *Nature* 438(7069):803-819.
- 539 5. Hoepfner MP, *et al.* (2014) An improved canine genome and a comprehensive catalogue
540 of coding genes and non-coding transcripts. *PLoS One* 9(3):e91172.
- 541 6. Ricketts SL & Marchant TW (2018) Meeting report from the Companion Animal Genetic
542 Health conference 2018 (CAGH 2018): a healthy companionship: the genetics of health
543 in dogs. *Canine Genet Epidemiol* 5(Suppl 1):6.
- 544 7. Han L, Su B, Li WH, & Zhao Z (2008) CpG island density and its correlations with
545 genomic features in mammalian genomes. *Genome Biol* 9(5):R79.
- 546 8. Han L & Zhao Z (2009) Contrast features of CpG islands in the promoter and other
547 regions in the dog genome. *Genomics*. 94(2):117-124.
- 548 9. Paigen K & Petkov P (2010) Mammalian recombination hot spots: properties, control and
549 evolution. *Nat Rev Genet* 11(3):221-233.
- 550 10. Baudat F, *et al.* (2010) PRDM9 is a major determinant of meiotic recombination hotspots
551 in humans and mice. *Science* 327(5967):836-840.
- 552 11. Myers S, *et al.* (2010) Drive against hotspot motifs in primates implicates the PRDM9
553 gene in meiotic recombination. *Science* 327(5967):876-879.
- 554 12. Parvanov ED, Petkov PM, & Paigen K (2010) Prdm9 controls activation of mammalian
555 recombination hotspots. *Science* 327(5967):835.
- 556 13. Kauppi L, Jeffreys AJ, & Keeney S (2004) Where the crossovers are: recombination
557 distributions in mammals. *Nat Rev Genet* 5(6):413-424.
- 558 14. Duret L & Galtier N (2009) Biased gene conversion and the evolution of mammalian
559 genomic landscapes. *Annu Rev Genomics Hum Genet* 10:285-311.
- 560 15. Meunier J & Duret L (2004) Recombination drives the evolution of GC-content in the
561 human genome. *Mol Biol Evol* 21(6):984-990.
- 562 16. Paigen K & Petkov PM (2018) PRDM9 and Its Role in Genetic Recombination. *Trends*
563 *Genet* 34(4):291-300.
- 564 17. Baker Z, *et al.* (2017) Repeated losses of PRDM9-directed recombination despite the
565 conservation of PRDM9 across vertebrates. *Elife* 6.
- 566 18. Oliver PL, *et al.* (2009) Accelerated evolution of the Prdm9 speciation gene across
567 diverse metazoan taxa. *PLoS Genet* 5(12):e1000753.
- 568 19. Axelsson E, Webster MT, Ratnakumar A, Ponting CP, & Lindblad-Toh K (2012) Death
569 of PRDM9 coincides with stabilization of the recombination landscape in the dog
570 genome. *Genome Res* 22(1):51-63.
- 571 20. Auton A, *et al.* (2013) Genetic recombination is targeted towards gene promoter regions
572 in dogs. *PLoS Genet* 9(12):e1003984.
- 573 21. Credille KM, *et al.* (2009) Transglutaminase 1-deficient recessive lamellar ichthyosis
574 associated with a LINE-1 insertion in Jack Russell terrier dogs. *Br J Dermatol*
575 161(2):265-272.

- 576 22. Wolf ZT, *et al.* (2014) A LINE-1 insertion in DLX6 is responsible for cleft palate and
577 mandibular abnormalities in a canine model of Pierre Robin sequence. *PLoS Genet*
578 10(4):e1004257.
- 579 23. Brooks MB, Gu W, Barnas JL, Ray J, & Ray K (2003) A Line 1 insertion in the Factor
580 IX gene segregates with mild hemophilia B in dogs. *Mamm Genome* 14(11):788-795.
- 581 24. Lin L, *et al.* (1999) The sleep disorder canine narcolepsy is caused by a mutation in the
582 hypocretin (orexin) receptor 2 gene. *Cell* 98(3):365-376.
- 583 25. Pele M, Tired L, Kessler JL, Blot S, & Panthier JJ (2005) SINE exonic insertion in the
584 PTPLA gene leads to multiple splicing defects and segregates with the autosomal
585 recessive centronuclear myopathy in dogs. *Hum Mol Genet* 14(11):1417-1427.
- 586 26. Clark LA, Wahl JM, Rees CA, & Murphy KE (2006) Retrotransposon insertion in SILV
587 is responsible for merle patterning of the domestic dog. *Proceedings Of The National*
588 *Academy Of Sciences Of The United States Of America*. 103(5):1376-1381.
- 589 27. Sutter NB, *et al.* (2007) A single IGF1 allele is a major determinant of small size in dogs.
590 *Science* 316(5821):112-115.
- 591 28. Gray MM, Sutter NB, Ostrander EA, & Wayne RK (2010) The IGF1 small dog
592 haplotype is derived from Middle Eastern grey wolves. *BMC Biol* 8:16.
- 593 29. Parker HG, *et al.* (2009) An expressed fgf4 retrogene is associated with breed-defining
594 chondrodysplasia in domestic dogs. *Science* 325(5943):995-998.
- 595 30. Brown EA, *et al.* (2017) FGF4 Retrogene On CFA12 Is Responsible For
596 Chondrodystrophy And Intervertebral Disc Disease In Dogs. *bioRxiv*.
- 597 31. Downs LM & Mellersh CS (2014) An Intronic SINE insertion in FAM161A that causes
598 exon-skipping is associated with progressive retinal atrophy in Tibetan Spaniels and
599 Tibetan Terriers. *PLoS One* 9(4):e93990.
- 600 32. Wang W & Kirkness EF (2005) Short interspersed elements (SINEs) are a major source
601 of canine genomic diversity. *Genome Res* 15(12):1798-1808.
- 602 33. Boissinot S, Entezam A, Young L, Munson PJ, & Furano AV (2004) The insertional
603 history of an active family of L1 retrotransposons in humans. *Genome Res* 14(7):1221-
604 1231.
- 605 34. Everson R, *et al.* (2017) An intronic LINE-1 insertion in MERTK is strongly associated
606 with retinopathy in Swedish Vallhund dogs. *PLoS One* 12(8):e0183021.
- 607 35. Katzir N, Arman E, Cohen D, Givol D, & Rechavi G (1987) Common origin of
608 transmissible venereal tumors (TVT) in dogs. *Oncogene* 1(4):445-448.
- 609 36. Alkan C, Sajjadian S, & Eichler EE (2011) Limitations of next-generation genome
610 sequence assembly. *Nat Methods* 8(1):61-65.
- 611 37. Weisenfeld NI, Kumar V, Shah P, Church DM, & Jaffe DB (2017) Direct determination
612 of diploid genome sequences. *Genome Res* 27(5):757-767.
- 613 38. Burton JN, *et al.* (2013) Chromosome-scale scaffolding of de novo genome assemblies
614 based on chromatin interactions. *Nat Biotechnol* 31(12):1119-1125.
- 615 39. Kaplan N & Dekker J (2013) High-throughput genome scaffolding from in vivo DNA
616 interaction frequency. *Nat Biotechnol* 31(12):1143-1147.
- 617 40. Chin CS, *et al.* (2016) Phased diploid genome assembly with single-molecule real-time
618 sequencing. *Nat Methods* 13(12):1050-1054.
- 619 41. Koren S, *et al.* (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer
620 weighting and repeat separation. *Genome Res* 27(5):722-736.

- 621 42. Sahlin K, Vezzi F, Nystedt B, Lundeberg J, & Arvestad L (2014) BESST--efficient
622 scaffolding of large fragmented assemblies. *BMC Bioinformatics* 15:281.
- 623 43. Numanagic I, *et al.* (2018) Fast characterization of segmental duplications in genome
624 assemblies. *Bioinformatics* 34(17):i706-i714.
- 625 44. Pendleton AL, *et al.* (2018) Comparison of village dog and wolf genomes highlights the
626 role of the neural crest in dog domestication. *BMC Biol* 16(1):64.
- 627 45. Vollger MR, *et al.* (2019) Long-read sequence and assembly of segmental duplications.
628 *Nat Methods* 16(1):88-94.
- 629 46. Kent WJ (2002) BLAT--the BLAST-like alignment tool. *Genome Res* 12(4):656-664.
- 630 47. Trapnell C, *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq
631 experiments with TopHat and Cufflinks. *Nature protocols* 7(3):562-578.
- 632 48. Trapnell C, *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals
633 unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*
634 28(5):511-515.
- 635 49. Grabherr MG, *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data
636 without a reference genome. *Nat Biotechnol* 29(7):644-652.
- 637 50. Wucher V, *et al.* (2017) FEELnc: a tool for long non-coding RNA annotation and its
638 application to the dog transcriptome. *Nucleic Acids Res* 45(8):e57.
- 639 51. Raney BJ, *et al.* (2014) Track data hubs enable visualization of user-defined genome-
640 wide annotations on the UCSC Genome Browser. *Bioinformatics* 30(7):1003-1005.
- 641 52. Holden LA, *et al.* (2018) Assembly and Analysis of Unmapped Genome Sequence Reads
642 Reveal Novel Sequence and Variation in Dogs. *Sci Rep* 8(1):10862.
- 643 53. Moran JV, DeBerardinis RJ, & Kazazian HH, Jr. (1999) Exon shuffling by L1
644 retrotransposition. *Science* 283(5407):1530-1534.
- 645 54. Goodier JL, Ostertag EM, & Kazazian HH, Jr. (2000) Transduction of 3'-flanking
646 sequences is common in L1 retrotransposition. *Hum Mol Genet* 9(4):653-657.
- 647 55. Pickeral OK, Makalowski W, Boguski MS, & Boeke JD (2000) Frequent human genomic
648 DNA transduction driven by LINE-1 retrotransposition. *Genome Res* 10(4):411-415.
- 649 56. Szak ST, Pickeral OK, Landsman D, & Boeke JD (2003) Identifying related L1
650 retrotransposons by analyzing 3' transduced sequences. *Genome Biol* 4(5):R30.
- 651 57. Macfarlane CM, *et al.* (2013) Transduction-specific ATLAS reveals a cohort of highly
652 active L1 retrotransposons in human populations. *Hum Mutat* 34(7):974-985.
- 653 58. Kopera HC, *et al.* (2016) LINE-1 Cultured Cell Retrotransposition Assay. *Methods Mol*
654 *Biol* 1400:139-156.
- 655 59. Moran JV, *et al.* (1996) High frequency retrotransposition in cultured mammalian cells.
656 *Cell* 87(5):917-927.
- 657 60. Doucet AJ, Wilusz JE, Miyoshi T, Liu Y, & Moran JV (2015) A 3' Poly(A) Tract Is
658 Required for LINE-1 Retrotransposition. *Molecular cell* 60(5):728-741.
- 659 61. Dewannieux M, Esnault C, & Heidmann T (2003) LINE-mediated retrotransposition of
660 marked Alu sequences. *Nat Genet* 35(1):41-48.
- 661 62. Pollard MO, Gurdasani D, Mentzer AJ, Porter T, & Sandhu MS (2018) Long reads: their
662 purpose and place. *Hum Mol Genet* 27(R2):R234-R241.
- 663 63. Field MA, *et al.* (2020) Canfam_GSD: De novo chromosome-length genome assembly of
664 the German Shepherd Dog (*Canis lupus familiaris*) using a combination of long reads,
665 optical mapping, and Hi-C. *Gigascience* 9(4).

- 666 64. Wang C, *et al.* (2020) A new long-read dog assembly uncovers thousands of exons and
667 functional elements missing in the previous reference. *bioRxiv*.
- 668 65. Kieleczawa J (2006) Fundamentals of sequencing of difficult templates--an overview. *J*
669 *Biomol Tech* 17(3):207-217.
- 670 66. Marais G (2003) Biased gene conversion: implications for genome and sex evolution.
671 *Trends Genet* 19(6):330-338.
- 672 67. Chaisson MJP, *et al.* (2019) Multi-platform discovery of haplotype-resolved structural
673 variation in human genomes. *Nat Commun* 10(1):1784.
- 674 68. Norton HL, Quillen EE, Bigham AW, Pearson LN, & Dunsworth H (2019) Human races
675 are not like dog breeds: refuting a racist analogy. *Evolution: Education and Outreach*
676 12(1):17.
- 677 69. Genomes Project C, *et al.* (2015) A global reference for human genetic variation. *Nature*
678 526(7571):68-74.
- 679 70. Payer LM, *et al.* (2017) Structural variants caused by Alu insertions are associated with
680 risks for many human diseases. *Proceedings Of The National Academy Of Sciences Of*
681 *The United States Of America*. 114(20):E3984-E3992.
- 682 71. Zhou W, *et al.* (2019) Identification and characterization of occult human-specific LINE-
683 1 insertions using long-read sequencing technology. *Nucleic Acids Res*.
- 684 72. Taliun D, *et al.* (2019) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed
685 Program. *bioRxiv*.
- 686 73. Sudmant PH, *et al.* (2015) An integrated map of structural variation in 2,504 human
687 genomes. *Nature* 526(7571):75-81.
- 688 74. Burns KH (2017) Transposable elements in cancer. *Nat Rev Cancer* 17(7):415-424.
- 689 75. Song S, Sliwerska E, Emery S, & Kidd JM (2017) Modeling Human Population
690 Separation History Using Physically Phased Genomes. *Genetics* 205(1):385-395.
- 691 76. Anonymous (2014) *Algorithms in bioinformatics : 14th International Workshop, WABI*
692 *2014, Wroclaw, Poland, September 8-10, 2014. Proceedings* (Springer, New York) 1st
693 edition. Ed p pages cm.
- 694 77. Haas BJ, *et al.* (2003) Improving the Arabidopsis genome annotation using maximal
695 transcript alignment assemblies. *Nucleic Acids Res* 31(19):5654-5666.
- 696 78. Haas BJ, *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using
697 the Trinity platform for reference generation and analysis. *Nature protocols* 8(8):1494-
698 1512.
- 699 79. Gotz S, *et al.* (2008) High-throughput functional annotation and data mining with the
700 Blast2GO suite. *Nucleic Acids Res* 36(10):3420-3435.
- 701 80. Bray NL, Pimentel H, Melsted P, & Pachter L (2016) Near-optimal probabilistic RNA-
702 seq quantification. *Nat Biotechnol* 34(5):525-527.
- 703 81. Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*
704 34(18):3094-3100.
- 705 82. Abyzov A & Gerstein M (2011) AGE: defining breakpoints of genomic structural
706 variants at single-nucleotide resolution, through optimal alignments with gap excision.
707 *Bioinformatics* 27(5):595-603.
- 708 83. Chin CS, *et al.* (2013) Nonhybrid, finished microbial genome assemblies from long-read
709 SMRT sequencing data. *Nat Methods* 10(6):563-569.
- 710 84. Bao W, Kojima KK, & Kohany O (2015) Repbase Update, a database of repetitive
711 elements in eukaryotic genomes. *Mob DNA* 6:11.

- 712 85. Wei W, *et al.* (2001) Human L1 retrotransposition: cis preference versus trans
713 complementation. *Mol Cell Biol* 21(4):1429-1439.
714