Origin and recent expansion of an endogenous gammaretroviral lineage in canids

Julia V. Halo[1#], Amanda L. Pendleton[2], Abigail S. Jarosz[1], Robert J. Gifford[3], Malika L. Day[1], Jeffrey M. Kidd[2,4]

[1] Halo, J.V.[1#], Jarosz, A.S., Day, M.L., *Bowling Green State University, Department of Biological Sciences, Bowling Green, OH 43403.*

[2] Kidd, J.M., Pendleton, A.L., *University of Michigan Medical School, Department of Human Genetics, Ann Arbor, MI, 48109.*

[3] Gifford, R.J., *Centre for Virus Research, University of Glasgow, Glasgow, G12 8QQ Scotland.*

[4] Kidd, J.M., *University of Michigan Medical School, Department of Computational Medicine and Bioinformatics, 100 Washtenaw Ave., Ann Arbor, MI, 48109.*

[#] Correspondence should be sent:
Julia V. Halo, Ph.D.
Department of Biological Sciences
Bowling Green State University
521A Life Sciences Building
Bowling Green, OH 43403
Office phone (419) 372-4096
Contributing Authors' e-mail: JVH: juliahw@bgsu.edu; ALP: ampend@med.umich.edu; ASJ: ajarosz@bgsu.edu; MLD: mlday@bgsu.edu; JMK: jmkidd@med.umich.edu; RJG: robert.gifford@glasgow.ac.uk

Keywords (3 required)

Canine, retrovirus, endogenous retrovirus, insertional polymorphism, *Canidae*

Abstract

Mammalian genomes contain a fossilized record of ancient retroviral infections in the form of endogenous retroviruses (ERVs). We used whole genome sequence data to assess the origin and evolution of the recently active ERV-Fc gammaretroviral lineage based on the record of past infections retained in the genome of the domestic dog, *Canis lupus familiaris.* We identified 165 loci, including 58 insertions absent from the dog reference assembly, and characterized element polymorphism across 332 canids from nine species. Insertions were found throughout the dog genome including within and near gene models. Analysis of 19 proviral sequences identified shared disruptive mutations indicating defective proviruses were spread via complementation. The patterns of ERV polymorphism and sequence variation indicate multiple circulating viruses infected canid ancestors within the last 20 million to within 1.6 million years with a recent bust of germline invasion in the lineage leading to wolves and dogs.

Introduction

During a retroviral infection, the viral genome is reverse transcribed and the resulting DNA is then integrated into the host genome as a provirus. In principle, the provirus carries all requirements necessary for its replication, and typically consists of an internal region encoding the viral genes (*gag*, *pro*/*pol*, and *env*) flanked by two regulatory long terminal repeats (LTRs) that are identical at the time of integration. Outermost flanking the provirus are short, 4-6 bp target site duplications (TSDs) of host genomic sequence generated during integration. Infection of such a virus within a germ cell or germ tissue may lead to an integrant that is transmitted vertically to offspring as an endogenous retrovirus (ERV). Over time, the ERV may reach detectable frequencies within a population or even fixation within a species (Boeke and Stoye, 1997). Through repeated germline invasion and expansion over millions of years, ERVs have accumulated to such an extent that they account for considerable proportions of genetic sequence of many mammalian genomes.

ERVs have been commonly referred to as 'fossils' of their once-infectious counterparts, providing a record of exogenous retroviruses that previously infected a species (Boeke and Stoye, 1997). Because an ERV switches from a relatively rapid evolutionary state as an infectious virus to a relatively slow one while replicated as part of the host genome, recently formed ERVs tend to bear close resemblance to their exogenous equivalent and possess a greater potential to retain functional properties. Across species, the majority of ERVs are thought to provide no advantage to the host, and have, for the most part, been progressively degenerated over time due to accumulated mutations or from recombination between the proviral LTRs that replaces the full-length sequence with a solitary LTR, or 'solo LTR' (Boeke and Stoye, 1997). However, increasing evidence suggests evolutionary roles in host physiology via gene regulation, for example by providing alternative promoters, enhancers, splice sites, or termination signals (Chuong et al., 2016, Macfarlan et al., 2012, Rebollo et al., 2012). There are also instances in which ERV gene products have been co-opted for various host functions. Notable examples include syncytial trophoblast fusion in eutherian animals (Lavialle et al., 2013) and blocking of infection from exogenous viruses (Nethe et al., 2005, Stoye, 2012, Blanco-Melo et al., 2017, Weiss and Stoye, 2013).

In humans, ERVs (HERVs) make up over 8% of the genome, the majority being the degenerate remnants of ancient infections. However, a subset of these elements is relatively intact and displays signatures of relatively recent germline invasion. Specifically, the HERV-K (HML-2) group has many young, insertionally polymorphic integrants that display variability in prevalence among global populations, retain ORFs, and includes many copies with high LTR-LTR identity (Wildschutte et al., 2016). Other species are known to harbor similar 'young'

integrants resulting from relatively recent endogenization events that segregate as unfixed alleles within the species. Examples include the cervid gammaretrovirus in mule deer populations of North America (Elleder et al., 2012) and the insertionally polymorphic ERVs found in domestic and wild cats (Roca et al., 2004, Troyer et al., 2004). Other species are hosts to infection from exogenous viruses that have been shown to contribute to new germline infections, for example, the Koala retrovirus (KoRV) is in the midst of transitioning to an endogenous state in Australia (Ishida et al., 2015, Tarlinton et al., 2006, Lober et al., 2018). Recombination between distinct ERV RNAs that are co-packaged in the same virion may also contribute to new viruses that have altered pathogenic properties (Stocking and Kozak, 2008).

The endogenous retroviruses classified as ERV-Fc are distant relatives of extant gammaretroviruses (also referred to as gamma-like, or γ-like). As is typical of most ERV groups, ERV-Fc is named for its use of a primer binding site complementary to the tRNA used during reverse transcription ($tRNA^{phe}$). Previous analysis of the *pol* gene showed that ERV-Fc elements form a monophyletic clade with the human γ-like ERV groups HERV-H and HERV-W (Jern 2005). As is common to all γ-like representatives, members of the ERV-Fc group possess a 'simple' genome that encodes the canonical viral genes and lacks apparent accessory genes that are present among complex retroviruses. The ERV-Fc group was first characterized as a putatively extinct, low copy number lineage that first infected the ancestor of all simians and later contributed to independent germline invasions in primate lineages (Benit et al., 2003). It has since been shown that ERV-Fc related lineages were infecting mammalian ancestors as early as 30 million years ago and subsequently circulated and spread to a diverse range of hosts, including carnivores, rodents, and primates (Diehl et al., 2016). The spread of the ERV-Fc lineage included numerous instances of cross-species jumps and recombination events between different viral lineages, now preserved in the fossil record of their respective host genomes (Diehl et al., 2016).

In comparison to humans and other mammals, the dog (*Canis lupus familiaris*) displays a substantially lower ERV presence, with only 0.15% of the genome recognizably of retroviral origin (Lindblad-Toh et al., 2005, Martinez Barrio et al., 2011). No exogenous retrovirus has been confirmed in the dog or any other canid, though there have been reports of retrovirus-like particles and enzyme activities in affected tissues of lymphomic and leukemic dogs (Ghernati et al., 2000, Modiano et al., 2005, Modiano et al., 1995, Onions, 1980, Perk et al., 1992, Safran et al., 1992, Tomley et al., 1983). Nonetheless, the ERV fossil record in the canine genome demonstrates that retroviruses did infect canine ancestors. The vast majority of canine ERVs (or 'CfERVs') are of ancient origin, as inferred by sequence divergence and phylogenetic placement (Martinez Barrio et al., 2011), suggesting most CfERV lineages ceased replicating long ago. An exception comes

from a minor subset of ERV-Fc-derived proviruses within the reference genome that possess signatures of recent integration, including high LTR nucleotide identity and the presence of ORFs (Martinez Barrio et al., 2011). This ERV lineage has been recently detailed by Diehl, *et al.*, in which the authors described a distinct ERV-Fc lineage in the Caniformia suborder (Figure 1) classified therein as 'ERV-Fc1' (Diehl et al., 2016). The ERV-Fc1 lineage first spread to members of the Caniformia at least 20 million years ago (mya) as a recombinant virus of two otherwise distantly related γ-like lineages: the virus possessed ERV-Fc *gag*, *pol*, and LTR segments but had acquired an *env* gene most closely related to ERV-W (syncytin-like) (Diehl et al., 2016). A derived sublineage, CfERV-Fc1(a), later spread to and infected canid ancestors via a cross-species transmission from an unidentified source, after which the lineage remained active/mobile and endogenized canid members until at least the last 1-2 million years (Diehl et al., 2016). Phylogenetic analyses confirmed the few recently inserted loci belong to CfERV-Fc1(a) (Diehl et al., 2016).

The domestic dog belongs to the family *Canidae*, the oldest family of Carnivora, which arose in North America during the late Eocene (~46 mya) (Macdonald and Sillero-Zubiri, 2004, Kumar et al., 2017) (Figure 1). Following multiple crossings of the Bering Strait land bridge to Eurasia, canids underwent massive radiations, leading to the ancestors of most modern canids (Macdonald and Sillero-Zubiri, 2004). The now extinct progenitors of the wolf-like canids, belonging to the genus *Canis*, first appeared in North America ~6 mya and also entered Eurasia via the same route (Macdonald and Sillero-Zubiri, 2004). Slowly, canids colonized all continents excluding Antarctica, as the formation of the Isthmus of Panama permitted dispersal and radiations within South America starting around 3 mya (Macdonald and Sillero-Zubiri, 2004). Approximately 1.1 mya, *Canis lupus*, the direct ancestor of the dog, emerged in Eurasia (Koepfli et al., 2015). Along with many other canid species, the gray wolf migrated back to the New World during the Pleistocene when the land bridge formed once more (Macdonald and Sillero-Zubiri, 2004). Placed within the context of CfERV-Fc1(a) evolution, the initial insertions from this lineage would have occurred while early *Canidae* members were still in North America, and continued until the emergence of the gray wolf.

Utilizing genome data from canid species representing all four modern lineages of Canidae (Figure 1), we assessed the origin, evolution, and impact of the recently active γ-like CfERV-Fc1(a) lineage, yielding the most comprehensive assessment of ERV activity in carnivores to date. Aside from analyses utilizing the current dog reference assembly (CanFam3.1), relatively little is known in this regard. We used Illumina sequence data to characterize CfERV-Fc1(a) integrants in dogs and wild canids, resulting in the discoveries of numerous insertionally

polymorphic and novel copies and the further delineation of the presence of this ERV group by comparison of orthologous insertions across species to provide a rich evolutionary history of CfERV-Fc1(a) activity among the *Canidae*. Our analysis demonstrates that the spread of CfERV-Fc1(a) contributed to numerous germline invasions in the ancestors of modern canids, including proviruses with apparently intact ORFs and other signatures of recent integration. The data suggest mobilization of existing ERVs by complementation had a significant role in the proliferation of the CfERV-Fc1(a) lineage in canine ancestors.

Results

Discovery of CfERV-Fc1(a) insertions

Insertionally polymorphic CfERV-Fc1(a) loci in dogs and wild canids

We determined the presence of CfERV-Fc1(a) insertions using Illumina whole genome sequencing data from dogs and other *Canis* representatives in two ways (Figure 2). First, we searched for CfERV-Fc1(a) sequences in the dog reference genome that were polymorphic across a collection of resequenced canines. In total, our dataset contained 136 CfERV-Fc1(a) insertions, and was filtered to a curated set of 107 intact or near-intact loci, including two loci related by segmental duplication (see Methods). These insertions are referred to as 'reference' throughout the text due to their presence in the dog reference genome. Comparative BLAT searches demonstrated their absence from the draft genomes of other extant Caniformia species (*i.e.*, ferret and panda). We then intersected the reference loci with deletions predicted by Delly (Rausch et al., 2012) within a sample set of 101 resequenced *Canis* individuals, specifically including jackals, coyotes, gray wolves, and dogs (Table S1). Candidate deletions were classified as those that intersected with annotated 'CfERVF1'-related loci and were within the size range of the solo LTR or provirus (~457 and ~7,885 bp, respectively; Figure 2A). The analysis identified 11 unfixed reference insertions, including 10 solo LTRs and one full-length provirus.

Our second approach utilized aberrantly mapped read-pairs from the same set of 101 genomes to identify CfERV-Fc1(a) copies that are absent from the dog reference genome. We refer to such insertions as 'non-reference'. These sites were identified using a combined read mapping and *de novo* assembly approach previously used to characterize polymorphic retroelement insertions in humans (Wildschutte et al., 2015, Wildschutte et al., 2016) (Figure 2B; also see Methods). This process identified 58 unique non-reference insertions, all of which derived from 'CfERVF1'-related elements per RepeatMasker analysis. Twenty-six of the 58 assembled insertion loci were fully resolved as solo LTRs, 30 had non-resolved but linked 5' and 3' genome-LTR junctions, and two had one clear assembled 5' or 3' LTR junction. Due to the one-

sided nature of assembled reads, we note the latter two were excluded from the majority of subsequent analyses (also see Figure S1 and Table S2). The assembled flanking regions and TSDs of each insertion were unique, implying each was the result of an independent germline invasion. Together, our two approaches for discovery resulted in 69 candidate polymorphic CfERV-Fc1(a)-related elements.

Validation of allele presence and accuracy of read assembly

We initially surveyed a panel of genomic DNA samples from breed dogs to confirm the polymorphic status of a subset of insertions (Figure 3). We then confirmed the presence of as many of the identified non-reference insertions as possible (34/58 sites) in predicted carriers from the 101 samples, and performed additional screening of each site to discriminate solo LTR and full-length integrants (Table S2). We confirmed a non-reference insertion for each of the 34 sites for which DNA from a predicted carrier was available. A provirus was present at eight of these loci, both insertion alleles were detected at three loci, and a solo LTR was present for the remaining loci. The full nucleotide sequence was obtained for 33 of the 34 insertions, with preference for sequencing placed on the provirus allele when present. The provirus at the final site (chr5:78,331,579) could not be completely spanned due to the presence of highly repetitive sequence within the *gag* gene (~2,250 bp from the consensus start). We also confirmed the polymorphic nature of the 11 reference CfERV-Fc1(a) insertions predicted to be unfixed, however we did not detect variable insertion states for those sites.

We assessed the accuracy of read assembly by comparing the assembled alleles to Sanger reads obtained for the validated sites. Due to the inability of the Illumina reads to span a full-length provirus, we were limited to the evaluation of fully assembled solo LTRs. Base substitutions were observed for just two assembled non-reference loci. First, the assembled chr13:17,413,419 solo LTR had a predicted base change between its TSDs that was resolved in Sanger reads; all other validated TSDs were in agreement as 5 bp matches, as is typical of the lineage. Second, the chr16:6,873,790 solo LTR had a single change in the LTR relative to the assembled allele. All other validated loci were in complete agreement with predictions obtained by read assembly of those insertions.

Structural variants between assembled sequences and the reference genome were also observed. For example, the assembled contig at chr33:29,595,068 captured a deletion of a reference SINE insertion 84 bp downstream of the non-reference solo LTR (Figure 4A). Deletion of the reference SINE was also supported by Delly deletion calls using the same Illumina data. Sanger sequencing confirmed a 34 bp deletion in an assembled insertion situated within a $TA_{(n)}$

simple repeat near chr32:7,493,322 (Figure 4B). Finally, an assembled solo LTR that mapped to chr2:32,863,024 contained an apparent 8 bp extension from the canonical CfERVF1 Repbase LTR of its 3' junction (5' TTTTAACA 3'). We validated the presence of the additional sequence within matched TSDs flanking the LTR and confirmed its absence from the empty allele (Figure 4C). The extension is similar in sequence to the consensus CfERVF1 LTR (5' ACTTAACA 3') and maintains the canonical 3' CA sequence necessary for proviral integration. These properties support its presence as part of the LTR, possibly generated during reverse transcription or during post-integration sequence exchange.

The CfERV-Fc1(a) genomic landscape

In principle, upon integration a provirus contains the necessary regulatory sequences for its own transcription within its LTRs; solo LTR recombinants likewise retain the same regulatory ability. Indeed, ERVs have been shown to affect regulatory functions within the host and some have been exapted for functions in normal mammalian physiology (reviewed in (Jern and Coffin, 2008)). A previous analysis of the then-current CanFam2.0 reference build identified at least five γ-like ERVs within or near genes from proviruses that belonged to a distinct and older non-Fc1(a) sublineage (specifically the 'CfERV1z' ERV-P related group, per RepeatMasker) (Martinez Barrio et al., 2011). Given the discovery of numerous novel insertions in our study and the improved annotation of the CanFam3.1 reference assembly, we assessed CfERV-Fc1(a) presence in relation to dog gene models.

Genome-wide insertion patterns were assessed for 58 non-reference and all 107 reference CfERV-Fc1(a) insertions. Of the 165 insertions, 29 (17.6%) were present within the introns of Ensembl gene models while one exonic reference insertion was identified (Table S3). Nine of the genic insertions (30%) were in sense orientation in respect to the gene. Some insertions were also in the vicinity of genes. For example, thirteen additional Fc1 loci were within 5 kb of at least one dog gene model; four of seven insertions situated upstream of the nearest gene were in sense orientation. Another 15 Fc1 loci were within 10 kb of at least one gene, of which seven of ten upstream insertions were in sense orientation with respect to the nearest gene. ERV-related promoter and enhancer involvement has been reported for distances exceeding 50 kb both upstream and downstream of genes (for example, see (Maruggi et al., 2009)). We find that 96 (58.2%) of assessed CfERV-Fc1(a) elements are within 50 kb of a gene model. Compared with randomized placements, CfERV-Fc1(a) insertions are significantly depleted within genes ($p < 0.001$) and within 10 kb of genes ($p < 0.001$). However, no significant difference was observed at the 50 kb distance (Figure S2). Insertions were present on all chromosomes except chr35 and

the Y chromosome, which is incomplete and not part of the canonical CanFam3.1 assembly. Individual CfERV-Fc1(a) insertions have been annotated with gene identifiers, gene ontology terms, and distances to nearest gene(s) in Table S3.

Age and evolutionary relationship of CfERV-Fc1(a) insertions

Dating proviral integrants by LTR divergence

Nucleotide divergence between the 5' and 3' LTRs of a provirus has been commonly used to estimate the time since endogenization, assuming that ERV sequences evolve neutrally following integration (Johnson and Coffin, 1999, Hughes and Coffin, 2004). Using this dating method, we estimated broad formation times of CfERV-Fc1(a) proviruses that maintained both LTRs. This analysis excluded three truncated reference elements (chr1:48,699,324, chr8:73,924,489, and chrUnAAEX03024336:1) and one non-reference provirus with an internal 291 bp deletion of the 3' LTR (chr17:9,744,973). The 3' LTR of the chr33:22,146,581 non-reference insertion contained a 43 bp internal duplication, which we treated as a single change. We applied a host genome-wide dog neutral substitution rate of 1.33×10^{-9} changes per site per year (Botigue et al., 2017), yielding formation times of individual proviruses from 20.49 mya to within 1.64 mya.

These estimates are sensitive to the assumed mutation rate, in addition to the limited number of differences expected between LTRs for the youngest loci. Obtaining age estimates in this manner for the youngest proviruses (as assumed by high 5' and 3' LTR identity) is dependent on the time to accrue a single mutation between the LTRs (of ~457 bp in length). The youngest estimate (1.64 my) is driven by two proviruses whose LTRs differ by a single base change and five proviruses with identical 5' and 3' LTRs, although the inter-element LTR haplotype sequence differed between proviruses. Across these five proviruses, LTR identities ranged from 98.5% to 99.4% (average of 98.95%), with a total of five LTR pairs that shared private substitutions. The remaining provirus shared an average identity of 85.45% to the other four. We further identified solo LTRs with sequence identical to one of two respective proviral LTR haplotypes (chr3:82,194,219 and chr4:22,610,555; also see below), suggesting multiple germline invasions from related variants. These data are consistent with insertion of CfERV-Fc1(a) members from multiple exogenous forms in canine ancestors, during which related variants likely infected over a similar timeframe.

Prevalence of CfERV-Fc1(a) loci in canids

To more precisely delineate the expansion of the identified CfERV-Fc1(a) members and refine our dating estimates, we surveyed insertion prevalence within an expanded sample set that more

fully represent extant members of the *Canidae* family, including the genomes of the dhole (*Cuon alpinus*), dog-like Andean fox (*Lycalopex culpaeus*), red fox (*Vulpes vulpes*), as well as the furthest canid outgroups corresponding to the Island (*Urocyon littorali*) and gray foxes (*U. cinereoargenteus*) (Figure 1). Thus, the analysis provided a broad timeline to reconstruct the evolutionary history of this ERV lineage ranging from host divergences within the last tens of thousands of years (gray wolves) to several millions of years (true foxes).

In total, we *in silico* genotyped 145 insertions (89 reference and 56 non-reference loci) across 332 genomes of canines and wild canids (refer to Methods; Table S4). To more accurately facilitate the identification of putative population-specific CfERV-Fc1(a), and to distinguish possible dog-specific insertions that may have occurred since domestication, wolves with considerable dog ancestry were removed from subsequent analyses (see Methods). Alleles corresponding to reference (*i.e.*, CanFam3.1) and alternate loci were recreated based on the sequence flanking each insertion while accounting for TSD presence. We then inferred genotypes by re-mapping Illumina reads that spanned either recreated allele for each site per sample. Reference insertions were deemed suitable for genotyping only if matched TSDs were present with clear 5' and 3' LTR junctions. We excluded the two non-reference sites with only a single assembled LTR junction due to uncertainty of both breakpoints (above). To facilitate genotyping of the eight unresolved assemblies with linked 5' and 3' LTR junctions, we supplemented the Repbase CfERVF1_LTR consensus sequence over the missing region (lower case in Table S2). As has been discussed in earlier work (Wildschutte et al., 2016), this genotyping approach is limited by the inability of single reads to span the LTR; therefore, the data do not discriminate between the presence of a solo LTR from that of a provirus at a given locus.

Insertion allele frequencies ranged from 0.14% (inferred single insertion allele) to fixed across samples (Figure 5; all raw data is included in Table S5). The rarest insertions were found in gray wolves, the majority of which were also present in at least one village or breed dog (for example, see chr13:16,157,778 and chr15:32,084,977 in Figure 5). All non-reference insertions were variably present in *Canis* species, and only few had read support in outgroup species (*i.e.* foxes, dhole). Notably, there was no evidence for the presence of any loci specific to village or breed dogs. Of outgroup canids, ~33% (48 of 145) insertions were detected in the Andean fox, and ~50% (a total of 73) insertions were present in the dhole. The remaining foxes, representing the most distant splits of extant canids, had the lowest prevalence of occupied loci, with just five insertions found in the gray and Island foxes, respectively. However, this is not unexpected since insertions private to these lineages would not be ascertained in our discovery sample set.

311 The relative distribution of proviruses was in general agreement with dating via LTR
312 divergence, though some inconsistencies were observed. No proviruses were detected in the fox
313 outgroups (*Urocyon* and *Vulpes*) that have an estimated split time from other Canidae of >8 mya
314 (Kumar et al., 2017), but some were present in the Andean fox (chr2:65,300,388,
315 chr5:24,576,900) and dhole (chrX:50,661,637, chr11:12,752,994). LTR divergence calculations
316 using the inferred dog neutral substitution rate dated these insertions near 20.49, 14.80, 6.65,
317 and 4.94 mya, respectively, suggesting the dating based on LTR divergence may be
318 overestimated. The youngest proviruses were variably present in *Canis* representatives. Of the
319 most recent insertions, two (chr5:10,128,780, chr17:9,744,973) were present in both New and
320 Old World wolves, implying integration prior to the geographic split of this lineage (1.10 mya) (Fan
321 et al., 2016). The remaining proviruses were present in Old World wolves and dogs only. Among
322 these was the chr33:22,146,581 provirus that had an estimated date of formation of 6.58 mya by
323 LTR comparison, consistent with skewed dating of the site. Altogether, the data are consistent
324 with CfERV-Fc1(a) endogenization in the ancestors of all modern canids followed by numerous
325 invasions leading to a relatively recent burst of activity in the wolf and dog lineage of *Canis*.
326
327 ***Evolution of the CfERV-Fc1(a) lineage in Canidae***
328 LTR sequences are useful in a phylogenetic analysis for exploring the evolutionary patterns of
329 circulating variants prior to endogenization, as well as following integration within the host. To
330 infer the evolutionary history leading to CfERV-Fc1(a) presence in modern canids, we constructed
331 an LTR tree using as many loci as possible (from 19 proviral elements and 142 solo-LTRs) (Figure
332 6; Table S6).
333 In broadly comparing LTR placement to our inferred species presence (Figure 6), the
334 longer-branched clusters contained the few ancestral loci present in the outgroups (gray and red
335 foxes) and those that were mostly fixed among the other surveyed species. However, at least two
336 non-reference LTRs and other unfixed insertions were also in these clades, suggesting their more
337 recent formation from related variants therein. One provirus was present within the most basal
338 clade, and four (including the duplicated locus) were present within intermediate clades. We
339 observed a major lineage (upper portion of tree) that included the majority of recent integrants.
340 This lineage gave rise to the greatest number of polymorphic insertions, including a derived clade
341 of insertions that appears to be *Canis*-specific, with some sites restricted to one or two sub-
342 populations. This lineage also contains the majority of proviral LTRs (15 of 19 included in the
343 analysis), most possessing intact *pol* and/or *env* genes. The youngest proviral integrants, as
344 inferred from high LTR identities and prevalence among sampled genomes, tend to be on short

345 branches within derived clusters that contain the majority of unfixed loci, likely reflecting their
346 source from a relatively recent burst of activity in *Canis* ancestors.
347 Within the germline, the highest occurrence of recombination resulting in a solo LTR takes
348 place between identical LTRs (Belshaw et al., 2007, Stankiewicz and Lupski, 2002), implying the
349 LTR sequence itself is preserved in the solo form. Under this assumption, the presence of identical
350 solo LTR haplotypes should, in principle, indicate their origin from a common ancestral source.
351 We identified four such LTR haplotypes within the *Canis*-specific clades, including loci in co-
352 clusters with one of two proviruses (chr3:82,194,219 and chr4:22,610,555), therefore bounding
353 the inferred age of these insertions to within the last 1.64 mya (dashed lines in Figure 6). Between
354 the four identical clusters, the LTR haplotypes shared nucleotide identity ranging from 99.3%
355 (three substitutions from a consensus of the four clusters) to 99.7% (one substitution), suggesting
356 their origin from related variants over a common timeframe. We modified our dating method to
357 obtain an estimated time of formation across each cluster by considering the total concatenated
358 LTR length per cluster, as has been similarly employed elsewhere (Ishida et al., 2015). This
359 approach placed tentative formation times of the youngest insertions from a common variant
360 547,220 years ago (no change over 1,374 bp, or 3 LTRs) and 410,415 years ago (no change over
361 1,832 bp, or 4 LTRs). Comparison to the inferred prevalence of each cluster indicates the most
362 recent of these insertions arose in Old World wolves, consistent with this timeframe.
363 Since proviral LTRs begin as an identical pair, aberrant placement in a tree and/or the
364 presence of mismatched TSDs implies involvement of the locus in post-insertion conversion or
365 rearrangement (Hughes and Coffin, 2005). LTRs from the youngest proviruses tended to pair on
366 sister branches. An exception includes the LTRs of the chr33:22,146,581 provirus, whose
367 mispairing is consistent with conversion of at least one of its LTRs, possibly from the
368 chr1:48,699,324 provirus or a similar variant (see above). There were six instances of aberrant
369 LTR placement for the remaining eight CfERV-Fc1(a) proviruses that had both LTRs present
370 (labeled in Figure 6), suggesting putative post-insertion conversion and contributing to inflated
371 age estimates based on LTR divergence. The TSD repeats of individual proviruses had matched
372 5 bp repeats in all cases, suggesting none of the elements have seeded inter-element
373 chromosomal rearrangements. With exception of three instances of reference solo LTRs that
374 each had a base change between its flanking repeats, the TSDs for all other solo LTRs were also
375 intact.
376
377 **CfERV-Fc1(a) structure and biology**
378 ***Characterization of the inferred CfERV-Fc1(a) ancestor***

As an endogenous element, an ERV may retain close resemblance to its exogenous source over long periods of time with recent integrants assumed to possess a greater potential to retain the properties of the infectious progenitor. We combined the eight non-reference proviruses with the eleven reference insertions to generate an updated consensus (referred to here as CfERV-Fc1(a)$_{CON}$) as an inferred common ancestor of the CfERV-Fc1(a) sublineage. A detailed annotation of the updated consensus is provided in Figure S3 and summarized as follows.

Consistent with the analysis of Caniform ERV-Fc1 consensus proviruses (Diehl et al., 2016), CfERV-Fc1(a)$_{CON}$ shows an internal segment of uninterrupted ERV-Fc related ORFs for *gag* (~1.67 kb in length) and *pol* (~3.54 kb; in-frame with *gag,* beginning directly after the *gag* stop codon, as is typical of C-type gammaretroviral organization). The CfERV-Fc1(a)$_{CON}$ *gag* product was predicted to contain intact structural regions and functional motifs therein for matrix (including the PPPY late domain involved in particle release and the N-terminal glycine site of myristoylation that facilitates Gag-cell membrane association), capsid, and nucleocapsid domains (including the RNA binding zinc-binding finger CCHC-type domains). Likewise, the Fc1(a)$_{CON}$ *pol* ORF was predicted to encode a product with conserved motifs for protease, reverse transcriptase (the LPQG and YVDD motifs in the RT active center), Rnase H (the catalytic DEDD center of RNA hydrolysis), and integrase (the DDX$_{35}$E protease resistant core and N-terminal HHCC DNA binding motif). An *env* reading frame (absent from the Repbase CfERVF1 consensus) was also resolved in the updated consensus. The ERV-W like Fc1$_{CON}$ *env* ORF (~1.73 kb) was present within an alternate ORF overlapping the 3' end of *pol*. Its predicted product included the RRKR furin cleavage site of SU and TM, the CWIC (SU) and CX$_6$CC (TM) motifs involved in SU-TM interactions, and a putative RD114-and-D-type (RDR) receptor binding motif (Sinha and Johnson, 2017). A hydrophobicity plot generated for the translated sequence identified segments for a predicted fusion peptide, membrane-anchoring TM region, and immunosuppressive domain (ISD) (Cianciolo et al., 1985). Putative major splice donor (base 576 within the 5'UTR; 0.67 confidence) and acceptor sites (base 5,216 within *pol*; 0.85 confidence) were identified that would be predicted for the generation of *env* mRNA (see Figure S3 and the accompanying legend). The CfERV-Fc1(a)$_{CON}$ element possessed identical LTRs, a tRNA^{Phe} binding site for priming reverse transcription (GAA anticodon; bases 464 to 480), and the canonical 5'-TG…CA-3' terminal sequences required for integration (Boeke and Stoye, 1997).

Properties of individual CfERV-Fc1(a) proviruses

We assessed the properties of individual full-length elements for signatures of putative function (summarized in Figure 7). With the exception of the *gag* gene, we identified intact ORFs in several

413 reference copies and most of our non-reference sequenced proviruses. A reading frame for the
414 *pol* gene was present in six proviruses; of these, all contained apparent RT, RnaseH, and
415 integrase domains without any changes that would obviously be alter function. Likewise, an *env*
416 ORF was present among seven proviruses, of which all but one contained the above mentioned
417 functional domains (the SU-TM cleavage site is disrupted in the chr5:10,128,780 provirus: RRKA).
418 Comparison of the rate of nonsynonymous (d_N) to synonymous (d_S) nucleotide substitutions for
419 the seven intact *env* reading frames revealed an average d_N/d_S ratio of 0.525, indicating moderate
420 purifying selection (p = 0.02, Nei-Gojobori method). The hydrophobicity plot of each *env* ORF was
421 in agreement with that of the CfERVFc1$_{CON}$ provirus, with predicted segments for a fusion peptide,
422 TM region, and ISD. Comparison to the *pol* and *env* translated products that would be predicted
423 from the CfERVFc1$_{CON}$ inferred the individual proviruses shared 98.4% to 99.3% (Pol) and 98%
424 to 99.6% (Env) amino acid identity, respectively, and each was distinct from the inferred
425 consensus.

426 No complete *gag* reading frame was observed in any provirus. Particularly when
427 compared to *pol* and *env*, the *gag* gene had incurred a number of inactivating mutations, including
428 several shared frameshifts leading to premature stops. The longest *gag* reading frames
429 (chr3:82,194,219 and chr26:35,982,438) both possessed a premature stop within the first zinc
430 finger domain of the nucleocapsid; of note the terminal *gag* frameshift was the only obvious
431 inactivation of any gene in the latter chr26:35,982,438 provirus. This domain has roles in the
432 encapsidation of viral genomic RNAs via recognition of a particular packaging signal sequence
433 (Ali et al., 2016). Thus, absence of both zinc finger domains and the N-terminal myristoylation site
434 should interfere with canonical Gag functions, regardless of the presence of intact matrix and
435 capsid domains. Excluding the frameshift leading to the abortive stop in those proviruses, the
436 translated Gag would have respectively shared 97.8% and 98% amino acid identity to the
437 CfERVFc1$_{CON}$ Gag. Though none of the identified CfERV-Fc1(a) proviruses have retained
438 complete reading frames for all genes, this finding does not exclude the possibility that rare intact
439 proviruses remain to be identified, or that a putative infectious variant could be generated via
440 recombination of co-packaged RNAs.

441 The majority of the CfERV-Fc1(a) proviruses could be assigned to one of two proposed
442 subgroups based on the presence of a common deletion within the *env* gene (Figure 7). The
443 deletion spans a 1,073 bp region of *env* (we refer to the segment as *env*$_{\Delta 1073}$), removing the
444 internal majority portions of SU and TM (also refer to Figure S3; including the putative receptor
445 binding domain, motifs involved in SU-TM interactions, and transmembrane domain). Eight
446 proviruses possessed the *env*$_{\Delta 1073}$ deletion, including the duplicated locus. The frameshift for

three of those eight proviruses would result in a product of ~204 amino acids in size (chr2:65,300,388, chr4:22,610,555, and chr6:47,934,941), though the significance of such a product is unclear. The prevalence of the $env_{\Delta 1073}$ deletion was skewed toward proviruses that harbored multiple inactivating mutations, while only one possessed a retained ORF (chr11:12,752,994, *pol*), consistent with the older status of most of these loci. Additionally, the $env_{\Delta 1073}$ deletion was present in the oldest proviruses and inferred to have arisen at least prior to the split of the dog-like foxes (see chr2:65,300,387 in Figure 5), suggesting its formation early in CfERV-Fc1(a) evolution (at least 8.7 mya; Figure 1). However, three proviruses with the deletion could not be genotyped due to the absence of clear LTR-genome junctions or due to encompassing duplication, making it possible that the allele predates the Andean fox split, as would be consistent with their placement within the tree (for example, see chr8:73,924,489; Figure 6). The $env_{\Delta 1073}$ deletion was not monophyletic in gene or LTR-based phylogenies, as would be expected if proviruses carrying the allele arose from a 'master' source element (Clough et al., 1996, Nascimento and Rodrigo, 2016). Examination of the regions directly flanking the deletion did not reveal common base changes shared among members with the allele. Our data are also not consistent with its transfer to existing proviruses through gene conversion, which should display shared base changes between all elements with the deletion. Therefore, we propose the $env_{\Delta 1073}$ allele spread via template-switching of co-packaged $env_{\Delta 1073}$ RNAs. Any of the above scenarios would result in the spread of an otherwise defective *env* gene. In contrast, all but two of the most recently integrated proviruses contained an uninterrupted *env* reading frame. In addition to the $env_{\Delta 1073}$ deletion, unique *env* deletions were present in two other elements. Specifically, a 1,702 bp deletion which removed all but the first 450 bp of *env* and 291 bp of the chr17:9,744,973 3' LTR, as well as the 5' truncated provirus at chr1:148,699,324 with an 896 bp deletion situated within the common $env_{\Delta 1073}$ deletion.

CfERV-Fc1(a) proliferation in canine ancestors

Nucleotide signatures within ERVs may be used to infer the mode(s) of proliferation, of which several routes have been described. One such mechanism, *trans* complementation, involves the co-packaging and spread of transcribed viral RNA genomes by functional viral proteins, supplied by a virus within the same cell (either exogenous or endogenous), thereby prolonging the apparent activity of the ERV lineage. As a result, RNAs from otherwise defective proviruses may be spread in cases where the ERV retains intact structures for transcription by host cell machinery and RNA packaging (Boeke and Stoye, 1997). Molecular signatures of *trans* complementation

480 may be interpreted from the presence of inherited changes among multiple elements, particularly
481 ones that would render a provirus defective (Belshaw et al., 2004, Belshaw et al., 2005).
482 We observed evidence for the mobilization of CfERV-Fc1(a) copies via complementation.
483 For example, examination of the proviral gene regions revealed inherited frameshift-causing
484 indels and common premature stops that were variably present among the majority of elements
485 (a total of 12 of the 19 proviruses; also see Figure 7). At least three distinct frameshifts leading to
486 a stop within *gag* were shared over several elements (from the $Fc1_{CON}$ start, bp 882:
487 chr4:22,610,555, chr11:12,752,994, chr12:869,873; bp 1,911: chr17:9,744,973,
488 chr33:22,146,581; and bp 2,203: chr3:82,194,219, chr26:35,982,438, and the duplicated
489 chr3:219,396 and chrUn_JH373247:11,035 insertions). Proviruses also shared unique deletions
490 leading to abortive stops within *pol* (near $Fc1_{CON}$ bp 3,988: chr1:48,699,324, and
491 chr3:82,194,219). In addition to the common $env_{\Delta 1073}$ frameshift deletion, putative in-frame *pol*
492 deletions were also present ($Fc1_{CON}$ bp 5,263 Δ3 bp: chr3:82,194,219; chrUn_AAEX03024336:1;
493 bp 5,705 Δ27 bp: chr5:24,576,900, chrUn_AAEX03024336:1). Two proviruses contained a
494 shared stop within *env* ($Fc1_{CON}$ bp 6,240: chr3:82,194,219, chr6:47,934,941). The provirus on
495 chromosome 3 possessed a total of four of the above changes differentially shared with other
496 proviruses in *gag*, *pol*, and *env*; these were the only defective changes present within the element.
497 While successive conversion events of the provirus from existing loci cannot be ruled out, this
498 provirus appears to be a comparatively young element (only found in Old World wolves and dogs),
499 which more likely suggests formation of the element via multiple intermediate variants. No other
500 provirus contained multiple common indels.
501 We did not find evidence for expansion of the lineage having proliferated via
502 retrotransposition in *cis*, during which new insertions are generated in an intracellular process
503 akin to the retrotransposition of long interspersed elements (Ostertag and Kazazian, 2001). Such
504 post-insertion expansion is typically accompanied by a loss of the viral *env* gene, particularly
505 within recently mobilized insertions (as interpreted, for example, by the derived phylogenetic
506 placement), whereas *gag* and *pol* are retained. Our data suggest this scenario is unlikely given
507 the absence of a functional *gag* gene and presence of a conserved *env* ORF in several elements,
508 particularly young ones. In this regard, *cis* retrotransposition tends to facilitate rapid *env*-less copy
509 expansion and therefore tends to occur among derived copies of a given lineage (Magiorkinis et
510 al., 2012), and our data suggest the opposite regarding older (loss of *env*) and younger (*env*
511 present) CfERV-Fc1(a) proviruses.
512
513 **Discussion**

514 Mammalian genomes are littered with the remnants of retroviruses, the vast majority of which are
515 fixed among species and present as obviously defective copies. However, the genomes of several
516 species harbor some demonstrably ancient ERVs whose lineages contain relatively intact loci and
517 are sometimes polymorphic, despite millions of years since integration. Such ERVs have the
518 potential to exert expression (of either proviral-derived or host-derived products by donation of an
519 LTR) that may affect the host, especially for intact copies or those within new genomic contexts.
520 In particular, ERV expression from relatively recent integrants has been linked to disease
521 (reviewed in (Jern and Coffin, 2008, Mager and Stoye, 2015)). However, there is also growing
522 evidence that many fixed loci have been functionally co-opted by the host or play a role in host
523 gene regulation (reviewed in (Frank and Feschotte, 2017)). To begin to discriminate the
524 relationship of individual elements in the context of the host necessitates a population-level
525 investigation of the breadth of ERV abundance and prevalence within a species. Illustrating both
526 bursts of activity and putative extinction, our findings present a comprehensive assessment of the
527 evolutionary history of a single retroviral lineage through the genomic surveys of nine globally
528 distributed canid species, some represented by multiple subpopulations.
529 Relative to other animal models, ERV-host relationships within the dog have been
530 understudied. Until now, reports of canine ERVs have been from analysis of a single genome
531 assembly or limited screening of reference loci (Martinez Barrio et al., 2011, Jo et al., 2012,
532 Tarlinton et al., 2013). To further investigate a subset of apparent recent germline integrants
533 (Martinez Barrio et al., 2011) we surveyed the level of polymorphism and possible mechanisms
534 of spread of the γ-like ERV-Fc1(a) lineage across a diverse set of canid species. Our exhaustive
535 analysis of CfERV-Fc1(a) loci is the first population-level characterization of a recently active ERV
536 group in canids. From analysis of Illumina WGS from 101 representatives from the canid genus
537 *Canis*, we uncovered numerous insertionally polymorphic sites, corresponding to both reference
538 loci and insertions that are missing from the dog genome assembly, and performed a comparative
539 genotyping approach utilizing additional extant *Canidae* members to provide a broad evolutionary
540 history of this ERV lineage. We identified eight non-reference proviruses that contain ORFs,
541 display high LTR identities, and have derived placements within a representative phylogeny,
542 which are all characteristics of relatively young elements.
543 Insertions were located within dog gene models, although permutations indicated that
544 CfERV-Fc1(a) insertions are significantly depleted within and near genes (Figure S2). Given their
545 placement in previously unoccupied genomic locales, the presence of such insertions raises the
546 possibility of biological effects. For example, two intronic LTRs were fixed in all canids: one within
547 *AIG1*, a transmembrane hydrolase involved in lipid metabolism (Parsons et al., 2016); the other

548 in the diffuse panbronchiolitis region *DPCR1* of the dog major histocompatibility complex 1 (Yan
549 et al., 2018). Other intronic insertions were fixed in samples following the splits of the true and
550 dog-like foxes. These included genes with homologs involved in tumor suppression (*OPCML*),
551 cell growth regulation (*CDKL3*), DNA repair (*FANCL*), and innate immunity (*TMED7-TICAM2*). An
552 exonic *Canis*-specific solo LTR was located at chr1:107,628,579 within the 3' UTR of *BCAT2*, an
553 essential gene in metabolizing mitochondrial branched-chain amino acids. In humans, altered
554 expression of *BCAT2* is implicated in tumor growth and nucleotide biosynthesis in some forms of
555 pancreatic cancer (Mayers et al., 2016, Dey et al., 2017, Ananieva and Wilkinson, 2018). The
556 same LTR is situated ~550 bp upstream of *FUT2*, a fucosyltransferase involved ABH blood group
557 antigen biosynthesis in mucosal secretions (de Mattos, 2016, Ferrer-Admetlla et al., 2009). *FUT2*
558 variants affect secretion status and have been implicated in intestinal microbiota composition (Le
559 Pendu et al., 2006), viral resistance (Thorven et al., 2005), and slowed progression of HIV
560 (Kindberg et al., 2006). Other insertions were upstream and downstream of gene vicinities, again
561 raising the possibility of host effects (also refer to Table S3). Though connections between LTR
562 presence and physiology are yet to be determined, these findings will inform future investigations
563 into the potential effect of CfERVs on host biology.

564 CfERV-Fc1(a) integrants endogenized canid ancestors over a period of several millions
565 of years (Figure 8B-E). This activity included bouts of infectious activity/mobilization inferred from
566 the last 20.4 my to the most recent integrants formed within 1.6 mya, the latter of which are only
567 present in *Canis* sub-populations. The mutation rate we used to obtain these estimated
568 timeframes (1.33×10^{-9} changes per site per year (Botigue et al., 2017)) coincides with those from
569 two other ancient genome analyses, which utilized ancient DNA to calibrate wolf and dog mutation
570 rates (Skoglund et al., 2015, Frantz et al., 2016). However, our rate is substantially slower than
571 those used previously to date reference CfERV-Fc1(a) members including 2.2×10^{-9} (as an
572 "average" mammalian neutral substitution rate) (Martinez Barrio et al., 2011) and the faster rate
573 of 4.5×10^{-9} (as has been reported for the mouse) (Diehl et al., 2016). Applying those substitution
574 rates to our data would infer much younger integration times of 11.85 mya to <0.91 mya and 6.1
575 mya to <0.48 mya, respectively. We note the precision in ERV-Fc1(a) age estimations using this
576 method is subject to the accuracy of the inferred background mutation rate, but may be skewed
577 due to the presence of post-insertion sequence exchange between LTRs. As the latter cannot be
578 conclusively ruled out, we interpret our estimations as broad formation times only.

579 Due to their complete absence of LTR divergence, the youngest CfERV-Fc1(a) ages are
580 bounded to the estimate of 1.64 my, using the dog substitution rate. Therefore, we employed an
581 alternative approach that makes use of LTRs that shared haplotypes (Ishida et al., 2015) to narrow

582 the age estimations to ~547,220 and 410,415 years, again, as inferred from the time estimated
583 to accrue one mutation across multiple identical LTRs (respectively across three and four LTRs
584 per haplotype). For comparison, applying the average mammalian and mouse substitution rates
585 to the same data would place either event respectively at 303,251 and 161,734 years ago (no
586 change over three LTRs) and 227,438 and 121,300 years ago (no change over four LTRs). Both
587 estimates are consistent with CfERV-Fc1(a) circulation after the estimated emergence of the gray
588 wolf species 1.1 mya and pre-dating the split of the New and Old World gray wolves (Fan et al.,
589 2016) (Figure 8F). The branching patterns observed within our LTR phylogeny are consistent with
590 these findings, implying bursts of replication from closely related variants now recorded in clusters
591 of LTR haplotypes. In this regard, our findings suggest bouts of infection from multiple circulating
592 viruses over a relatively short evolutionary time period.
593 CfERV-Fc1(a) activity coincided with major speciation events in canine evolution (Figure
594 8B-E). Taking into consideration the above approaches for age estimations, we refined the dating
595 of endogenization events by integrating inferred ages with that of orthologous presence/absence
596 patterns across numerous canid lineages, many of which are recently diverged clades. The
597 analysis served two purposes. First, we made use of the tenet that ERV integration is permanent
598 and the likelihood of two independent integration events at the same locus is negligible. In this
599 way, the presence of an ERV insertion that is shared between individuals or species supports its
600 origin in a common ancestor. Therefore, integration prior to or following the split of two or more
601 species is supported by virtue of insertion presence/absence of occupied loci across those
602 species. Second, the analysis allowed us to infer insertion genotypes across highly diverse canid
603 representatives, thus providing the means to gauge the collective patterns of individual CfERV-
604 Fc1(a) loci among contemporary animals to infer putative sub-population or species-specific
605 integrants.
606 Comparisons of the approximated insertion dates discussed above in combination with
607 estimated species split times would place the earliest CfERV-Fc1(a) germline invasions prior to
608 or near the estimated divergence of the *Canidae* from now extinct ancestors (14.15 mya) (Kumar
609 et al., 2017), followed by invasions after the split of the true fox (12.9 mya) (Kumar et al., 2017)
610 and fox-like canid lineages (8.7 mya) (Koepfli et al., 2015). Subsequent insertions also occurred
611 prior to the split of the South American canid and wolf lineages (3.97 mya) (Koepfli et al., 2015).
612 According to this timeframe, and consistent with the detection of some young proviral insertions
613 private to gray wolves and dogs alone (Figure 5), the most recent invasions would have occurred
614 around the time of the branching event that gave rise to gray wolves (1.10 mya) (Koepfli et al.,
615 2015). Based on the lack of observed dog-specific loci, our data suggests that CfERV-Fc1(a)

616 replication ceased in wolf ancestors prior to domestication, which is estimated to have begun
617 around 40 kya (Botigue et al., 2017) (Figure 8G), but does not rule out continued activity. Analysis
618 of additional genomes, particularly from gray wolves, should clarify the presence of such variants
619 in future analysis.

620 ERV-Fc1(a) activity included the spread of defective recombinants. Our comparative
621 analysis of nucleotide differences shared among the proviruses supports a scenario in which
622 CfERV-Fc1(a) members proliferated in canine ancestors via complementation. Patterns of
623 discreet, shared changes among distinct elements in all viral genes were observed (*i.e.*,
624 premature stops and common base changes, indels, in addition to the $env_{\Delta 1073}$ segment; Figure
625 7), consistent with the spread of mutations present from existing Fc1(a) copies, probably via co-
626 packaging of the defective viral genomes. Of the 19 proviruses analyzed in full, the majority
627 displayed shared discreet stops or the $env_{\Delta 1073}$ deletion, in addition to in-frame indels. This pattern
628 is consistent with the hypothesis that degradation of ERV genomes, particularly involving the loss
629 of *env*, offers an evolutionary benefit to the host by preventing the potential horizontal spread of
630 infectious viruses between individuals, as has been suggested (Magiorkinis et al., 2012, Lober et
631 al., 2018). The presence of intact *env* genes, and sequence signatures of selective pressure
632 retained within those *env* reading frames, suggests involvement of Fc1(a) *env* leading to the
633 putative formation of recombinant proviruses, rather than having been intracellularly
634 retrotransposed (in *cis*) that would not require a functional envelope. Altogether such patterns of
635 reinfection may have predominantly occurred within a given individual, as none of these
636 mechanisms explicitly requires (but does not rule out) spread to other individuals within the
637 population; indeed concurrent reinfection of a single individual may also lead to unique proviruses
638 later transmitted to offspring (Young et al., 2012). Indeed, several retroviruses, including HIV,
639 have been shown to be capable of co-packaging RNA from other retroviruses, even ones with
640 low sequence homology (Ali et al., 2016). These findings suggest that complementation was a
641 predominant form of proliferation for the observed CfERV-Fc1(a) loci. In theory, a functional
642 provirus could arise in a spontaneous recombinant, raising the possibility of bursts of amplification
643 to come. Indeed, all viral genes in our consensus appear to be intact, illustrative that few changes
644 would be required to generate a putatively infectious virus.

645 Patterns of shared sequence changes, such as premature stops and in-frame shifts,
646 indicate that the oldest inherited change involved an in-frame shift in the *pol* gene (from the $Fc1_{CON}$
647 start, bp 5705 Δ27 bp). Aside from the $env_{\Delta 1073}$ deletion, all other common changes were present
648 in the lineage that led to the majority of young insertions (Figure 6). Among the earliest inferred
649 changes were premature stops in *gag* (CfERV-$Fc1_{CON}$ bp 882 and 2203, respectively) and *env*

(CfERV-Fc1$_{CON}$ bp 6240) that tended to have been present among elements within a *Canis*-specific subclade. Additional inherited changes were present in a separate *Canis*-specific subclade in the form of a third distinct stop in *gag* (CfERV-Fc1$_{CON}$ bp 1911). The shared *gag* stop was only observed within that cluster, suggesting its origin in a timeframe near variants contributing to the subclade followed by spread to new insertions therein. The stop is present in the chr17:9,744,973 and chr33:22,146,581 proviruses, therefore limiting LTR dating of the change; based on its restriction to assayed *Canis* members it likely originated within the last 2.74 my (Koepfli et al., 2015). Taken together, the data are consistent with independent origin and spread of multiple defective features that began prior to ancestors of the dog-like foxes and followed the Old and New World wolf split. The phylogenetic placement of defective proviruses suggests the co-occurrence of spread from multiple sources.

The apparent absence of any infectious retrovirus among canines is peculiar, particularly as individuals are likely to be challenged from viruses infecting prey species. While there have been reports of retroviral activities and particles displaying characteristic γ-like features in canine leukemias and lymphomas (Ghernati et al., 2000, Modiano et al., 2005, Modiano et al., 1995, Onions, 1980, Perk et al., 1992, Safran et al., 1992, Tomley et al., 1983), those findings have not been substantiated. It is well-known that canine cell lines are permissive for replication of retroviruses that infect other host species including human (Fadel and Poeschla, 2011), a property possibly reflecting the loss of the antiviral factor TRIM5α in canines (Sawyer et al., 2007). A recent report confirmed transcriptional activity from at least one γ-like CfERV group (non-Fc1(a)) in canine tissues and cell lines (Tarlinton et al., 2013). We have also preliminarily demonstrated expression of CfERV-Fc1(a) proviruses in canine tissues and tumor-derived cell lines (Jarosz and Halo, unpublished data).

Expression of ERV groups has been associated with both normal physiology and disease in several animal models, including humans, based on patterns of ERV-derived products observed within associated tissues (reviewed in (Jern and Coffin, 2008)). However, the consequences of this expression are not always clear. It is known from animal studies that ERVs with similarity to human ERVs, including those with extant forms with replicative activity, as well as proteins derived from related ERV members, are capable of driving aberrant cellular proliferation, tumorigenesis, and inciting immune responses (Jern and Coffin, 2008). Given our findings of the breadth and relative intactness of the CfERV-Fc1(a) lineage, we suggest that de-regulated expression from these loci is responsible for the γ-retroviral activities previously reported in canine tumors and cell lines, implying the potential for a pathogenic role of ERV-Fc1(a) loci and exogenous retroviruses in canines.

Materials and Methods

Whole genome sequence data

For ERV discovery, Illumina WGS data were obtained from a total of 101 samples corresponding to 37 breed dogs, 45 village dogs, and 19 wild canids (Auton et al., 2013, Botigue et al., 2017, Decker et al., 2015, Fan et al., 2016, Freedman et al., 2014, Marsden et al., 2016, Pendleton et al., 2018, Koepfli et al., 2015) (Table S1). Data were downloaded in fastq format and processed to Binary Alignment/Map BAM format using bwa version 7.15 and Picard v 2.9.0. SNV genotypes of sequenced samples were determined using Genome Analysis Toolkit (GATK) version 3.7 (McKenna et al., 2010). Information corresponding to all samples and sources of raw data is detailed in Table S1.

Identification of annotated CfERVF1 reference insertions

The dog ERV-Fc1(a) lineage is classified in Repbase as 'CfERVF1' derived (Repbase update 10.08) (Jurka et al., 2005). We therefore mined the CanFam3.1 RepeatMasker output for elements classified as 'CfERVF1_LTR' and 'CfERVF1-int' according to Repbase vouchers to identify dog ERV-Fc1(a) LTRs and proviral elements, respectively. We required the presence of at least one LTR and contiguous internal sequence for a provirus, and the absence of any proximal internal region for a solo LTR. A total of 136 insertions were identified, corresponding to 21 proviral elements and 115 solo LTRs. The integration breakpoint ±1kb of each locus was extracted and used in BLAT searches against the other available carnivoran reference assemblies corresponding to ferret (MusPutFur1.0) (Peng et al., 2014), panda (BGI_Shenzhen1.0) (Li et al., 2010), and cat (Felis_catus_8.0) (Pontius et al., 2007) to confirm specificity to the dog reference. Sequences for proviral loci were extracted from CanFam3.1 based on the start and end positions of the full-length insertions, and filtered to remove severely truncated elements, resulting in 11 CfERV-Fc1(a) full-length or near full-length elements (*i.e.*, containing at least one viral gene region and associated 5' or 3' LTR). This count is consistent with recent findings of this ERV group in the dog reference (Diehl et al., 2016). Solo LTR insertions were filtered similarly to remove truncated elements, resulting in 96 insertions for further analysis.

Deletion analysis of reference CfERV-Fc1(a) insertions

Reference insertions corresponding to deletion variants were inferred using the program Delly (v0.6.7) (Rausch et al., 2012), which processed BAM alignment files from samples indicated in Table S1 using a MAD score cutoff equal to 7, and a minimum map quality score threshold of at least 20. Resulting reference deletions with precise breakpoint predictions were next intersected

718 with 'CfERVF1' reference coordinates based on RepeatMasker annotations of CanFam3.1. Only
719 deletion calls corresponding to sizes of a solo LTR (400-500 bp) or a full-length provirus (7-9 kb)
720 were considered for further analysis.
721
722 **Identification of non-reference of CfERV-Fc1(a) insertions**
723 LTR-genome junctions corresponding to non-reference variants were assembled from supporting
724 Illumina reads (Wildschutte et al., 2015, Wildschutte et al., 2016), with modifications as follows.
725 The chromosomal positions of candidate non-reference ERVs were first identified using the
726 program RetroSeq (Keane et al., 2013). Individual BAM files were queried using RetroSeq
727 discovery to identify ERV-supporting discordant read pairs with one read aligned to the sequences
728 corresponding to 'CfERVF1' and 'CfERVF1_LTR' from RepBase (Jurka et al., 2005). Individual
729 BAM files were merged for subsequent steps using GATK as described (Wildschutte et al., 2016).
730 RetroSeq call was run on the merged BAM files requiring ≥2 supporting read pairs for a call and
731 output calls of levels 6, 7, and 8 further assessed, resulting in 2,381 candidate insertions. Output
732 calls within ±500 bp of an annotated CfERV from the above queried classes were excluded to
733 eliminate false calls of known loci. ERV-supporting read pairs and split reads within a 200 bp
734 window of the call breakpoint were subjected to *de novo* assembly using the program CAP3
735 (Huang and Madan, 1999). Output contigs were filtered to identify ERV-genome junctions
736 requiring ≥30 bp of assembled LTR-derived and genomic sequence in the form of (i) one LTR-
737 genome junction, (ii) linked assemblies of 5' and 3' LTR junctions, or (ii) a fully resolved LTR
738 (~457bp) with clear breakpoints that mapped to CanFam3.1. Contigs that contained putative
739 CfERV junctions were then aligned back to the reference to precisely map the insertion position
740 of each call. Assembly comparisons were visualized using the program Miropeats (Parsons,
741 1995).
742
743 **Validations and allele screening**
744 For validating non-reference calls, primers were designed to flank the predicted insertion within
745 ~200 bp based on the breakpoint position for a given site. Genomic DNA from a subset of samples
746 with predicted insertion variants was used for validations. DNA with limited material was subjected
747 to whole genome amplification (WGA) from ~10ng genomic DNA according to the manufacturer's
748 protocol (Repli-G, Qiagen). For each sample, WGA DNA was diluted 1:20 in nuclease free water
749 and 1 uL was utilized per PCR reaction. Two PCR reactions were run for each site in standard
750 conditions using Taq polymerase (Invitrogen): one reaction utilized primers flanking each
751 candidate call to detect the empty or solo LTR alleles; the second was to detect the presence of

752 a proviral junction, utilizing the appropriate flanking primer paired with a primer within the CfERV-
753 Fc1(a) proviral 5'UTR (near base ~506 from the start of the Repbase F1 consensus element).
754 Sanger sequencing was performed on at least one positive sample. When detected, provirus
755 insertions were amplified in overlapping fragments from a single sample in a Picomaxx reaction
756 per the manufacturer's instructions (Stratagene) and sequenced to ≥4x across the full element. A
757 consensus was then constructed for each insertion based on the Sanger reads obtained from
758 each site. All sequences corresponding to non-reference solo-LTR insertions and all sequenced
759 proviral elements have been made available in Table S2.
760
761 **Genomic distribution**
762 The positions of the reference and non-reference insertions were intersected with Ensembl dog
763 gene models (Release 81; ftp.ensembl.org/pub/release-81/gtf/canis_familiaris/). Intersections
764 were performed using bedtools (Quinlan, 2014) with window sizes of 0, 5, 10, 25, 50, and 100 kb.
765 To assess significant enrichment of insertions relative to genic regions, we performed one
766 thousand permutations of randomly shuffled insertion positions, intersected the new positions with
767 genes, and calculated the number of insertions intersecting genes within the varying window sizes
768 as above. P-values were calculated as the number of permuted insertion sets out of one thousand
769 that intersected with less than or equal to the number of genes observed in the true insertion set.
770
771 **Dating of individual proviruses**
772 A molecular clock analysis based on LTR divergence was used to estimate times of insertion
773 (Diehl et al., 2016, Johnson and Coffin, 1999, Wildschutte et al., 2016). For 7 non-reference and
774 8 reference proviruses that had 5' and 3' LTRs present, the nucleotide differences between those
775 LTRs was calculated, treating gaps >2bp as single changes. The total number of changes was
776 then divided by the LTR length (*e.g.* 457 bp), and the percent divergence normalized to the
777 inferred canine background mutation rate of $1.3x10^{-9}$ changes per site per year (Botigue et al.,
778 2017) to obtain age estimations in millions of years for individual insertions. The provirus at
779 chr17:97,449,73 was excluded from the analysis due to truncation of its 3' LTR. We extended
780 LTR dating to estimate times of formation for identical LTR groups that included solo LTRs using
781 a modification of the above approach as described elsewhere (Ishida et al., 2015). Briefly, the
782 total length in bp of the LTRs making up each cluster was collectively added and the age estimate
783 obtained by the percent divergence for a single base pair to have been introduced along the total
784 length utilizing the same mutation rate of $1.3x10^{-9}$ changes per site per year.
785

786 ***In silico* genotyping**

787 We genotyped 145 insertions (89 reference and 56 non-reference insertions) utilizing whole
788 genome Illumina reads and reconstructed alleles corresponding to the empty and occupied sites.
789 Genotyping was performed on 332 individuals including the 101 samples utilized for discoveries
790 of polymorphic variants (Kim et al., 2012, Vamathevan et al., 2013, Owczarek-Lipska et al., 2013,
791 Wang et al., 2013, Kim et al., 2013, Auton et al., 2013, Koepfli et al., 2015, Botigue et al., 2017,
792 Freedman et al., 2014, Li et al., 2014, Zhang et al., 2014, Decker et al., 2015, Wang et al., 2016,
793 Fan et al., 2016, Marsden et al., 2016, Robinson et al., 2016, Liu et al., 2014) (Table S4).
794 Reference insertions were deemed to be suitable for genotyping based on manual assessment
795 for the presence of paired TSDs and uninterrupted flanking sequence. Sites associated with
796 duplication events were identified by comparison of flanking regions and TSD presence, and
797 insertions within encompassing duplication (proviruses at chr3:219,396 and
798 chrUn_JH373247:11,035), or situated within duplicated pre-insertion segments
799 (chrUn_AAEX03025486:2,349) were excluded, as were sites with single assembled junctions
800 (chr13:20,887,612; chr27:44,066,943; Table S2). The sequences from validated and completely
801 assembled LTRs were utilized for allele reconstruction of non-reference sites. For example, the
802 validated sequences for the non-reference solo LTRs at chr2:32,863,024 (8 bp LTR extension)
803 and chr32:7,493,322 (associated with deletion of reference sequence) were included for
804 genotyping of alternate alleles. For sites with linked, but non-resolved, 5' and 3' assembled
805 junctions (*i.e.*, missing internal sequence), we substituted the internal portion of each element
806 from the Repbase CfERVF1 consensus (see Table S2), and used the inferred sequence for allele
807 reconstruction. Insertion and pre-insertion alleles were then recreated based on ±600bp flanking
808 each insertion point relative to the CanFam3.1 reference, accounting for each 5bp TSD pair. For
809 each sample, genotype likelihoods were then assessed at each site based on re-mapping of those
810 reads to either allele, with error probabilities based on read mapping quality (Li, 2011, Wildschutte
811 et al., 2015), excluding sites without re-mapped reads for a given sample. Read pairs for which
812 both reads mapped to the internal portion of the element were excluded to avoid false positive
813 calls potentially introduced by non-specific alignment. The pipeline for genotyping is available at
814 https://github.com/KiddLab/insertion-genotype. The genotyped samples were sorted by ancestral
815 population, and allele frequencies estimated for the total number of individuals per population
816 genotyped at each locus (Table S5).
817
818 **Admixture**

819 A sample set containing only dogs and wolves were previously genotyped at approximately 7.6
820 million SNPs determined to capture genetic diversity across canids (Botigue et al., 2017). Using
821 Plink (Purcell et al., 2007), sites were filtered to remove those with missing genotypes in at least
822 ten percent of samples, those in LD with another SNP within 50 bp (--indep-pairwise 50 10 0.1),
823 and randomly thinned to 500,000 SNPs. To reduce the bias of relatedness, the sample set was
824 further filtered to remove duplicates within a single modern breed, leaving 254 samples (Table
825 S7). Identification of wolf samples with high dog ancestry was made through five independent
826 ADMIXTURE (Alexander et al., 2009) analyses of the thinned SNP set with random seeds
827 (558905, 110684, 501738, 37781236, and 85140928) for K values 2 through 6. Since we aimed
828 to discern cfERV-Fc1(a) insertions that may be dog-specific (*i.e.* having occurred since
829 domestication), we removed any gray wolf that had high dog ancestry from further analysis. To
830 do this, we calculated average dog ancestry within gray wolves at K=3 across all runs, which was
831 the K value with the lowest cross validation error rate. Wolves with greater than 10% dog ancestry
832 (an Israeli (isw01) and Spanish (spw01) wolf) were excluded from subsequent species and sub-
833 population assessments.
834
835 **Phylogenetic analysis**
836 Nucleotide alignments were performed using MUSCLE (Edgar, 2004) followed by manual editing
837 in BioEdit (Hall, 1999) for intact CfERV-Fc1(a) LTRs from 19 proviral elements and 142 solo-
838 LTRs. Of non-reference elements, the solo LTR with a 388 bp internal deletion at
839 chr22:57,677,068 was excluded, as was the 141 bp truncated solo LTR at chr5:80,814,713. We
840 also excluded partially reconstructed insertions corresponding to 'one-sided' assemblies or sites
841 with linked 5' and 3' assembled junctions but that lacked internal resolution (Table S1). A
842 maximum likelihood (ML) phylogeny was reconstructed from the LTR alignment using FastTree
843 (Price et al., 2010) and the (GTR+CAT) model (generalized time reversible (GTR) model of
844 nucleotide substitution plus "CAT" rate approximation). To infer the robustness of inferred splits
845 in the phylogeny, local support values were calculated using the ML-based approach
846 implemented in FastTree, wherein the Shimodaira-Hasegawa test is applied to the three alternate
847 topologies (NNIs) around each node. The average d_N/d_S ratio for intact env genes was determined
848 using the codeml program in the PAML software package (version 4.8) (Xu and Yang, 2013)
849 based on a Neighbor-Joining tree. Statistical significance was determined using the Nei-Gojobori
850 method (Nei and Gojobori, 1986) implemented in MEGA7 (Kumar et al., 2016) with a null
851 hypothesis of strict neutrality ($d_N = d_S$).
852

Acknowledgments

We thank John Coffin, Michael Freeman, Welkin Johnson and Zachary Williams for meaningful discussion and comments, and all owners and donors involved in sample donations for genomic DNA sources. We thank Anna Kukekova for sharing red fox genome data and Adam Boyko, Tomàs Marquès-Bonet, Carles Vilà, and Robert Wayne for early access to genome sequence data. Images of canids were obtained for *Urocyon littoralis* ("Island Fox II" (CC BY 2.0) by Shanthanu Bhardwaj), *Vulpes vulpes* ("El pequeño amigo" (CC BY 2.0) by Minette Lang), *Lycalopex culpaeus* (by Christian Mehlführer; Wikimedia Commons), *Cuon alpinus* (Wikimedia Commons), and *Canis lupus* (www.usda.gov). This work was supported in part by a National Institutes of Health Academic Research Enhancement Award R15GM122028 to JVH, National Institutes of Health grant R01GM103961 to JMK, National Institutes of Health Training Fellowship T32HG00040 to ALP, and UK Medical Research Council MC_UU_12014/10 to RJG. DNA samples were provided by the Cornell Veterinary Biobank, a resource built with the support of NIH grant R24GM082910 and the Cornell University College of Veterinary Medicine.

Author Contributions

JVH, ALP, and JMK designed the study. JVH, ALP, and JMK were responsible for genome data processing. JVH, ASJ, MLD were responsible for sequence-based analysis. JVH, ALP, RJG and JMK were responsible for data analysis. JVH, ALP, and JMK wrote the paper. All authors have read and approved the final manuscript.

Competing Interests

The authors declare no competing interests exist.

877 **Figure Legends**

878

879 **Figure 1. Canidae evolution and representative extant species.** Relative to other Caniforms,
880 the evolutionary relationship of the four major canid lineages, along with estimated split times
881 (determined from (Kumar et al., 2017) and (Koepfli et al., 2015)) is shown. Species with asterisks
882 were included in CfERV-Fc1(a) discovery, and all canids here were used for *in silico* genotyping.
883 Images are provided for the underlined species. See acknowledgements for all image credits.

884

885 **Figure 2. Strategy for detecting insertionally polymorphic ERV variants. (**A) ERV allelic
886 presence. Upper: full-length provirus; Mid: solo LTR recombinant; Lower, unoccupied (pre-
887 integration) site. (B) Strategy for detection of reference ERV deletions. Illumina read pairs were
888 mapped to the CanFam3.1 reference, deletion-supporting read pairs and split reads identified
889 using the program Delly (Rausch et al., 2012), and candidate calls then intersected with
890 RepeatMasker outputs considering 'CFERVF1' repeats. Deletion calls within a size range
891 corresponding to a solo LTR or provirus were selected for further analysis. (C) Strategy for
892 detection of non-reference ERV insertions. ERV insertion-supporting anchored read pairs were
893 identified from merged Illumina data mapped to the CanFam3.1 reference using the RetroSeq
894 program (Keane et al., 2013). Insertion-supporting read pairs and intersecting split reads were
895 assembled, assemblies for which 'CfERVF1' sequence was present were identified by
896 RepeatMasker analysis, and the assembled contigs then re-mapped to the dog CanFam3.1
897 reference for precise breakpoint identification.

898

899 **Figure 3. Representative allele screening of polymorphic loci.** PCR screens of a subset of
900 non-reference CfERV-Fc1(a) integrants. Validation of insertionally polymorphic sites was
901 performed for seven candidate sites across genomic DNA from a panel of breed dogs. (A)
902 Strategy for primer design and allele detection. Primers were designed to target within 250 bp of
903 the insertion coordinates based on re-mapping of the assembled breakpoints to the CanFam3.1
904 reference. Two primers sets were used for each locus: one utilized an internal and flanking primer
905 to amplify the 5' LTR of a full-length element; another set was used for detection of the pre-
906 integration (unoccupied) or solo LTR alleles each locus. (B) Banding patterns supporting the
907 unoccupied, solo LTR, or full-length alleles. The chromosomal location of each integrant is
908 indicated at left; allele presence is indicated at right: (+) insertion presence and detected allele; (-
909) insertion absence. Samples: A, boxer; B, Labrador retriever; C, golden retriever; D, Springer
910 spaniel; E, standard poodle; F, German shepherd; G, shar-pei.

Figure 4. Assessment of assembled non-reference alleles. LTR insertions associated with structural variation as captured in assembled Illumina read data. Local three-way alignments were generated for each assembled locus using the program Miropeats (Parsons, 1995). Each consisted of the LTR allele obtained by read assembly, the validated LTR allele obtained by Sanger sequencing of the locus in one individual, and the empty locus as present within the CanFam3.1 reference. Alignments are shown for three representative LTR assemblies. The allele type is labeled at left in each alignment; lines are used to indicate the breakpoint position of the insertion and shared sequence between alleles. (A) An LTR assembly that includes captured deletion of a bimorphic SINE_Cf insertion present in the CanFam3.1 reference. (B) An assembled LTR associated with a short 34 bp deletion of sequence that is present in the reference. (C) A validated assembly of an LTR that included an 8 bp extension relative to the canonical CfERVF1 repeat.

Figure 5. Distribution of CfERV-Fc1(a) insertions in the genomes of modern canids. *In silico* genotyping was performed for 145 LTRs utilizing whole genome data across 347 sequenced canids, which were selected to represent extant members of all major *Canidae* lineages (Figure 1). Sample names are indicated above according to species or sub-population. Samples correspond to the Island and gray foxes (the furthest outgroup species; n=8), red fox (n=1), Andean fox (n=1), dhole (n=1), golden jackal (n=1), golden wolf (n=1), coyote (n=3), red wolf (n=2), and representatives of gray wolf sub-populations (n=33), village dogs (n=111), ancient breed dogs (n=38), and modern breed dogs (n=154). 'Insertion' and 'unoccupied' alleles were recreated utilizing the CanFam3.1 reference and genotypes were inferred by re-mapping Illumina reads that spanned either recreated allele for each sample. Samples lacking remapped reads across a given site were excluded from genotyping at that site alone (indicated with a '.'). Allele frequencies were calculated for each species or sub-population (see Methods) and plotted as a heat map (insertion frequency indicated by color bar at top). The locus identifier for each insertion (left) corresponds to the chromosome and the leftmost insertion breakpoint, irrespective of insertion orientation. Non-reference and reference insertions are indicated by an 'N' and 'R', respectively. Full-length proviruses are highlighted with a green diamond.

Figure 6. Evolutionary history of the CfERV-Fc1(a) lineage in canids. An approximately-maximum-likelihood phylogeny was reconstructed from an alignment of 157 ERV-Fc LTR sequences. The tree has been midpoint-rooted for display purposes. Asterisks below nodes

945 indicate local support values > 70%. Each insertion is denoted corresponding to chromosomal
946 position relative to CanFam3.1 coordinates. A color bar is shown at the right to denote element
947 presence as fixed among *Canis* (dark blue), insertionally polymorphic (light blue), or not
948 genotyped (gray). Elements identified as fixed among *Canis* insertionally polymorphic have been
949 further highlighted by blue shading. LTRs belonging to proviruses are indicated along with the
950 chromosomal position with a (5') or (3') as appropriate. Clusters of identical LTR haplotypes are
951 indicated with a vertical dashed line. Mis-matched pairs of proviral LTRs are indicated by a
952 diamond. LTRs from proviruses lacking cognate LTR pairs (*i.e.*, due to truncation of the element)
953 are indicated with a cross. The scale bar shown represents the evolutionary distance in
954 substitutions per site.
955
956 **Figure 7. Structural features of CfERV-Fc1(a) proviruses.** (A) Representation of the CfERV-
957 Fc1(a)$_{CON}$ provirus. Viral gene reading frames for ERV-Fc related *gag* and *pol* are shown in blue;
958 the ERV-W related *env* is shown in orange. Color usage is consistent with that of (Diehl et al.,
959 2016). LTRs are colored in gray: U3 is in medium tone; R is dark; U5 is light. The provirus and
960 open reading frames are shown to scale. (B) Structural features of non-reference and reference
961 proviruses. When present, open reading frames are indicated above the appropriate element.
962 Insertions and deletions >3 bases are depicted with blue and red flags, respectively. The *env*$_{\Delta 1073}$
963 deletion is labeled and indicated by a dashed line, as are other truncated or deleted element
964 features. Reference gaps present within are shown in light gray boxes to scale. Stop codons are
965 indicated with a black or red asterisk, where red is used to specify premature stops common to
966 two or more proviruses. Crosses at the left indicate proviruses that are unfixed among *Canis*
967 samples. The number of substitutions between LTRs is shown at right with the corresponding
968 calculated age as inferred based on the neutral substitution rate of 1.33×10^{-9} changes per site per
969 year (Botigue et al., 2017).
970
971 **Figure 8. History of CfERV-Fc1(a) germline invasion in the Canidae.** A timeline of major
972 events in canid or CfERV-Fc1(a) evolutionary history relative to estimated insertion events. At the
973 approximate time point, branching events of the major canid lineages are indicated by arrows
974 along the timeline with colors matching Figure 1. Indicated by proviruses to the right of the timeline
975 are estimated insertion times based on genotyping data from Figure 5. (A) Based on its presence
976 in all canids, the recombination event that formed the provirus (B), which infected canid ancestors
977 occurred sometime between the split of the major Caniform lineages (A) and the origins of canids
978 in North America (C). Following the migration to Eurasia (D), a major species radiation occurred

in the wolf-like canid lineage (E). Finally, the comparatively recent re-introduction of gray wolves in North America reflects the split between the Old and New World wolves (F), which likely partially coincided with the domestication of Old World Wolves (G). Estimated timings for events A-C are supported by (Kumar et al., 2017), D-E by (Wang and Tedford, 2008), F by (Koblmüller et al., 2016), and G by (Botigue et al., 2017).

Supplemental Figure Legends

Figure S1. Assembled CfERV breakpoints remapped to the CanFam3.1 reference. Three-way alignments for 58 non-reference insertions are shown. Alignments were used to depict CfERV-Fc1(a) LTR junctions obtained by assembled supporting reads (shown in red text) remapped to the CanFam3.1 reference sequence (shown in black text and underlined). The 5bp sequence corresponding to the target site duplication is underlined and bolded in the reference allele. The coordinates of the CanFam3.1 reference sequence shown is provided above each alignment; the first base of the LTR is labeled and indicated by an asterisk shown respective of orientation ('+' or '–'). Insertions for which a provirus was validated are labeled as appropriate. The single assembled junctions are provided for either of two insertions: chr13:20,998,612 (3' junction); chr27:44,066,943 (5' junction).

Figure S2. Depletion of CfERV-Fc1(a) insertions near dog gene models. Following one thousand permutations, the number of gene models that intersect with shuffled CfERV-Fc1(a) insertions are displayed in histograms. Permuted insertions that intersect with at least one Ensembl dog gene model precisely (green), within 10 kb (blue) or 50 kb (gray) are shown. Red lines indicate the observed number of insertions from the true set.

Figure S3. Annotated CfERV-Fc1(a) consensus provirus. A consensus provirus was deduced from 19 proviruses using BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html) based on the most commonly represented nucleotide at each site. The consensus nucleotide sequence is shown in black text. The 5' and 3' LTRs are labeled with black bars. The translated sequences for the viral genes are indicated below and with bars at the right, with the Gag sequence in blue, Pol in orange, and Env in green. Motifs pertaining to viral functions are labeled appropriately on their translated sequence and general annotated in the right sidebar. Translated start and stop sites are indicated for each of the three genes. Segments for a predicted fusion peptide,

membrane-anchoring TM region, and immunosuppressive domain (ISD) were determined using the program Phobius (http://phobius.sbc.su.se). Putative major splice donor and acceptor sites were determined using the program NetGene2 (http://www.cbs.dtu.dk/services/NetGene2/).

Supplemental Tables

Table S1. Canine sample information for discovery of CfERV-Fc1(a) insertions. Information for the resequencing dataset of 101 canines used for CfERV-Fc1(a) insertion discovery. The sample identifier, sex, breed/species/population information and canine group is given per sample. Also provided are the Short Read Archive (SRA) sequence identifiers (SRR) matching the files downloaded and processed in this study, along with the PubMed identifier for the accompanying published study (if available) for each sample.

Table S2. Information for non-reference sites considered in analyses. The coordinates relative to CanFam3.1 are provided for each identified non-reference insertion. For each site, information pertaining to the insertion orientation, target site duplication (relative to the CanFam3.1 reference), detected insertion alleles (provirus, solo LTR), and element sequence is provided. Primer sequences are provided for validated sites. (A) Information for sequenced loci and validated sequences. (B) Information for loci with complete assembled insertion alleles. (C) Information for loci with partially assembled insertion alleles.

Table S3. Gene region information and GO ontology analyses. The coordinates for each reference and non-reference insertion are provided along with Ensembl gene models from dog (release #81) that are within window distances of 0, 5, 10, 25, 50, and 100 kb of the insertion.

Table S4. Sample information for canid genotyping. Sample and data access information for the resequencing dataset of 332 canines genotyped at the discovered CfERV-Fc1(a) reference and non-reference insertions. Accompanying data descriptions provided for each sample match that of Table S1.

Table S5. Genotypes and inferred allele frequencies. Raw genotypes obtained across 332 resequenced samples for 56 non-reference and 89 reference insertions are provided in vcf format. Allele frequencies were calculated from raw genotypes per canid species or sub-population, as indicated above each column. Non-genotyped sites are noted with an “-”.

Table S6. LTR nucleotide alignment. LTR alignment for phylogenetic analysis using LTRs from a total of 19 proviruses and 142 solo LTRs, provided in fasta format.

Table S7. Samples included in admixture analysis. Sample information for the 254 samples included in admixture analysis. Accompanying data columns provided for each sample match that of Table S1.

References

ALEXANDER, D. H., NOVEMBRE, J. & LANGE, K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res,* 19**,** 1655-64.

ALI, L. M., RIZVI, T. A. & MUSTAFA, F. 2016. Cross- and Co-Packaging of Retroviral RNAs and Their Consequences. *Viruses,* 8.

ANANIEVA, E. A. & WILKINSON, A. C. 2018. Branched-chain amino acid metabolism in cancer. *Curr Opin Clin Nutr Metab Care,* 21**,** 64-70.

AUTON, A., RUI LI, Y., KIDD, J., OLIVEIRA, K., NADEL, J., HOLLOWAY, J. K., HAYWARD, J. J., COHEN, P. E., GREALLY, J. M., WANG, J., BUSTAMANTE, C. D. & BOYKO, A. R. 2013. Genetic recombination is targeted towards gene promoter regions in dogs. *PLoS Genet,* 9**,** e1003984.

BELSHAW, R., KATZOURAKIS, A., PACES, J., BURT, A. & TRISTEM, M. 2005. High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection. *Mol Biol Evol,* 22**,** 814-7.

BELSHAW, R., PEREIRA, V., KATZOURAKIS, A., TALBOT, G., PACES, J., BURT, A. & TRISTEM, M. 2004. Long-term reinfection of the human genome by endogenous retroviruses. *Proc Natl Acad Sci U S A,* 101**,** 4894-9.

BELSHAW, R., WATSON, J., KATZOURAKIS, A., HOWE, A., WOOLVEN-ALLEN, J., BURT, A. & TRISTEM, M. 2007. Rate of recombinational deletion among human endogenous retroviruses. *J Virol,* 81**,** 9437-42.

BENIT, L., CALTEAU, A. & HEIDMANN, T. 2003. Characterization of the low-copy HERV-Fc family: evidence for recent integrations in primates of elements with coding envelope genes. *Virology,* 312**,** 159-68.

BLANCO-MELO, D., GIFFORD, R. J. & BIENIASZ, P. D. 2017. Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors. *Elife,* 6.

BOEKE, J. D. & STOYE, J. P. 1997. Retrotransposons, endogenous retroviruses, and the evolution of retroelements. *In:* COFFIN, J., HUGHES, S. & VARMUS, H. (eds.) *Retroviruses.* New York, NY: CSHL Press.

BOTIGUE, L. R., SONG, S., SCHEU, A., GOPALAN, S., PENDLETON, A. L., OETJENS, M., TARAVELLA, A. M., SEREGELY, T., ZEEB-LANZ, A., ARBOGAST, R. M., BOBO, D., DALY, K., UNTERLANDER, M., BURGER, J., KIDD, J. M. & VEERAMAH, K. R. 2017. Ancient European dog genomes reveal continuity since the Early Neolithic. *Nat Commun,* 8**,** 16082.

CHUONG, E. B., ELDE, N. C. & FESCHOTTE, C. 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science,* 351**,** 1083-7.

CIANCIOLO, G. J., COPELAND, T. D., OROSZLAN, S. & SNYDERMAN, R. 1985. Inhibition of lymphocyte proliferation by a synthetic peptide homologous to retroviral envelope proteins. *Science,* 230**,** 453-5.

CLOUGH, J. E., FOSTER, J. A., BARNETT, M. & WICHMAN, H. A. 1996. Computer simulation of transposable element evolution: random template and strict master models. *J Mol Evol,* 42**,** 52-8.

DE MATTOS, L. C. 2016. Structural diversity and biological importance of ABO, H, Lewis and secretor histo-blood group carbohydrates. *Rev Bras Hematol Hemoter,* 38**,** 331-340.

DECKER, B., DAVIS, B. W., RIMBAULT, M., LONG, A. H., KARLINS, E., JAGANNATHAN, V., REIMAN, R., PARKER, H. G., DROGEMULLER, C., CORNEVEAUX, J. J., CHAPMAN, E. S., TRENT, J. M., LEEB, T., HUENTELMAN, M. J., WAYNE, R. K., KARYADI, D. M. & OSTRANDER, E. A. 2015. Comparison against 186 canid whole-genome sequences reveals survival strategies of an ancient clonally transmissible canine tumor. *Genome Res,* 25**,** 1646-55.

DEY, P., BADDOUR, J., MULLER, F., WU, C. C., WANG, H., LIAO, W. T., LAN, Z., CHEN, A., GUTSCHNER, T., KANG, Y., FLEMING, J., SATANI, N., ZHAO, D., ACHREJA, A., YANG, L., LEE, J., CHANG, E., GENOVESE, G., VIALE, A., YING, H., DRAETTA, G., MAITRA, A., WANG, Y. A., NAGRATH, D. & DEPINHO, R. A. 2017. Genomic deletion of malic enzyme 2 confers collateral lethality in pancreatic cancer. *Nature,* 542**,** 119-123.

DIEHL, W. E., PATEL, N., HALM, K. & JOHNSON, W. E. 2016. Tracking interspecies transmission and long-term evolution of an ancient retrovirus using the genomes of modern mammals. *Elife,* 5.

EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res,* 32**,** 1792-7.

ELLEDER, D., KIM, O., PADHI, A., BANKERT, J. G., SIMEONOV, I., SCHUSTER, S. C., WITTEKINDT, N. E., MOTAMENY, S. & POSS, M. 2012. Polymorphic integrations of an endogenous gammaretrovirus in the mule deer genome. *J Virol,* 86**,** 2787-96.

FADEL, H. J. & POESCHLA, E. M. 2011. Retroviral restriction and dependency factors in primates and carnivores. *Vet Immunol Immunopathol,* 143**,** 179-89.

FAN, Z., SILVA, P., GRONAU, I., WANG, S., ARMERO, A. S., SCHWEIZER, R. M., RAMIREZ, O., POLLINGER, J., GALAVERNI, M., ORTEGA DEL-VECCHYO, D., DU, L., ZHANG, W., ZHANG, Z., XING, J., VILA, C., MARQUES-BONET, T., GODINHO, R., YUE, B. & WAYNE, R. K. 2016. Worldwide patterns of genomic variation and admixture in gray wolves. *Genome Res,* 26**,** 163-73.

FERRER-ADMETLLA, A., SIKORA, M., LAAYOUNI, H., ESTEVE, A., ROUBINET, F., BLANCHER, A., CALAFELL, F., BERTRANPETIT, J. & CASALS, F. 2009. A natural history of FUT2 polymorphism in humans. *Mol Biol Evol,* 26**,** 1993-2003.

FRANK, J. A. & FESCHOTTE, C. 2017. Co-option of endogenous viral sequences for host cell function. *Curr Opin Virol,* 25**,** 81-89.

FRANTZ, L. A., MULLIN, V. E., PIONNIER-CAPITAN, M., LEBRASSEUR, O., OLLIVIER, M., PERRI, A., LINDERHOLM, A., MATTIANGELI, V., TEASDALE, M. D., DIMOPOULOS, E. A., TRESSET, A., DUFFRAISSE, M., MCCORMICK, F., BARTOSIEWICZ, L., GAL, E., NYERGES, E. A., SABLIN, M. V., BREHARD, S., MASHKOUR, M., BALASESCU, A., GILLET, B., HUGHES, S., CHASSAING, O., HITTE, C., VIGNE, J. D., DOBNEY, K., HANNI, C., BRADLEY, D. G. & LARSON, G. 2016. Genomic and archaeological evidence suggest a dual origin of domestic dogs. *Science,* 352**,** 1228-31.

FREEDMAN, A. H., GRONAU, I., SCHWEIZER, R. M., ORTEGA-DEL VECCHYO, D., HAN, E., SILVA, P. M., GALAVERNI, M., FAN, Z., MARX, P., LORENTE-GALDOS, B., BEALE, H., RAMIREZ, O., HORMOZDIARI, F., ALKAN, C., VILA, C., SQUIRE, K., GEFFEN, E., KUSAK, J., BOYKO, A. R., PARKER, H. G., LEE, C., TADIGOTLA, V., WILTON, A., SIEPEL, A., BUSTAMANTE, C. D., HARKINS, T. T., NELSON, S. F., OSTRANDER, E. A., MARQUES-BONET, T., WAYNE, R. K. & NOVEMBRE, J. 2014. Genome sequencing highlights the dynamic early history of dogs. *PLoS Genet,* 10**,** e1004016.

GHERNATI, I., CORBIN, A., CHABANNE, L., AUGER, C., MAGNOL, J. P., FOURNEL, C., MONIER, J. C., DARLIX, J. L. & RIGAL, D. 2000. Canine large granular lymphocyte leukemia and its derived cell line produce infectious retroviral particles. *Vet Pathol,* 37**,** 310-7.

HALL, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic acids symposium series,* 41**,** 95-98.

HUANG, X. & MADAN, A. 1999. CAP3: A DNA sequence assembly program. *Genome Res,* 9**,** 868-77.

HUGHES, J. F. & COFFIN, J. M. 2004. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. *Proc Natl Acad Sci U S A,* 101**,** 1668-72.

HUGHES, J. F. & COFFIN, J. M. 2005. Human endogenous retroviral elements as indicators of ectopic recombination events in the primate genome. *Genetics,* 171**,** 1183-94.

ISHIDA, Y., ZHAO, K., GREENWOOD, A. D. & ROCA, A. L. 2015. Proliferation of endogenous retroviruses in the early stages of a host germ line invasion. *Mol Biol Evol,* 32**,** 109-20.

JERN, P. & COFFIN, J. M. 2008. Effects of retroviruses on host genome function. *Annu Rev Genet,* 42**,** 709-32.

JO, H., CHOI, H., CHOI, M. K., SONG, N., KIM, J. H., OH, J. W., SEO, K., SEO, H. G., CHUN, T., KIM, T. H. & PARK, C. 2012. Identification and classification of endogenous retroviruses in the canine genome using degenerative PCR and in-silico data analysis. *Virology,* 422**,** 195-204.

JOHNSON, W. E. & COFFIN, J. M. 1999. Constructing primate phylogenies from ancient retrovirus sequences. *Proc Natl Acad Sci* 96**,** 10254-10260.

JURKA, J., KAPITONOV, V. V., PAVLICEK, A., KLONOWSKI, P., KOHANY, O. & WALICHIEWICZ, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res,* 110**,** 462-7.

KEANE, T. M., WONG, K. & ADAMS, D. J. 2013. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics,* 29**,** 389-90.

KIM, H. M., CHO, Y. S., KIM, H., JHO, S., SON, B., CHOI, J. Y., KIM, S., LEE, B. C., BHAK, J. & JANG, G. 2013. Whole genome comparison of donor and cloned dogs. *Sci Rep,* 3**,** 2998.

KIM, R. N., KIM, D. S., CHOI, S. H., YOON, B. H., KANG, A., NAM, S. H., KIM, D. W., KIM, J. J., HA, J. H., TOYODA, A., FUJIYAMA, A., KIM, A., KIM, M. Y., PARK, K. H., LEE, K. S. & PARK, H. S. 2012. Genome analysis of the domestic dog (Korean Jindo) by massively parallel sequencing. *DNA Res,* 19**,** 275-87.

KINDBERG, E., HEJDEMAN, B., BRATT, G., WAHREN, B., LINDBLOM, B., HINKULA, J. & SVENSSON, L. 2006. A nonsense mutation (428G-->A) in the fucosyltransferase FUT2 gene affects the progression of HIV-1 infection. *AIDS,* 20**,** 685-9.

KOBLMÜLLER, S., VILÀ, C., LORENTE-GALDOS, B., DABAD, M., RAMIREZ, O., MARQUES-BONET, T., WAYNE, R. K. & LEONARD, J. A. 2016. Whole mitochondrial genomes illuminate ancient intercontinental dispersals of grey wolves (Canis lupus). *Journal of Biogeography,* 43.

KOEPFLI, K. P., POLLINGER, J., GODINHO, R., ROBINSON, J., LEA, A., HENDRICKS, S., SCHWEIZER, R. M., THALMANN, O., SILVA, P., FAN, Z., YURCHENKO, A. A., DOBRYNIN, P., MAKUNIN, A., CAHILL, J. A., SHAPIRO, B., ALVARES, F., BRITO, J. C., GEFFEN, E., LEONARD, J. A., HELGEN, K. M., JOHNSON, W. E., O'BRIEN, S. J., VAN VALKENBURGH, B. & WAYNE, R. K. 2015. Genome-wide Evidence Reveals that African and Eurasian Golden Jackals Are Distinct Species. *Curr Biol,* 25**,** 2158-65.

KUMAR, S., STECHER, G., SULESKI, M. & HEDGES, S. B. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution,* 34**,** 834-845.

KUMAR, S., STECHER, G. & TAMURA, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol,* 33**,** 1870-4.

LAVIALLE, C., CORNELIS, G., DUPRESSOIR, A., ESNAULT, C., HEIDMANN, O., VERNOCHET, C. & HEIDMANN, T. 2013. Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. *Philos Trans R Soc Lond B Biol Sci,* 368**,** 20120507.

LE PENDU, J., RUVOEN-CLOUET, N., KINDBERG, E. & SVENSSON, L. 2006. Mendelian resistance to human norovirus infections. *Semin Immunol,* 18**,** 375-86.

LI, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics,* 27**,** 2987-93.

LI, R., FAN, W., TIAN, G., ZHU, H., HE, L., CAI, J., HUANG, Q., CAI, Q., LI, B., BAI, Y., ZHANG, Z., ZHANG, Y., WANG, W., LI, J., WEI, F., LI, H., JIAN, M., LI, J., ZHANG, Z., NIELSEN, R., LI, D., GU, W., YANG, Z., XUAN, Z., RYDER, O. A., LEUNG, F. C., ZHOU, Y., CAO, J., SUN, X., FU, Y., FANG, X., GUO, X., WANG, B., HOU, R., SHEN, F., MU, B., NI, P., LIN, R., QIAN, W., WANG, G., YU, C., NIE, W., WANG, J., WU, Z., LIANG, H., MIN, J., WU, Q., CHENG, S., RUAN, J., WANG, M., SHI, Z., WEN, M., LIU, B., REN, X., ZHENG, H., DONG, D., COOK, K., SHAN, G., ZHANG, H., KOSIOL, C., XIE, X., LU, Z., ZHENG, H., LI, Y., STEINER, C. C., LAM, T. T., LIN, S., ZHANG, Q., LI, G., TIAN, J., GONG, T., LIU, H., ZHANG, D., FANG, L., YE, C., ZHANG, J., HU, W., XU, A., REN, Y., ZHANG, G., BRUFORD, M. W., LI, Q., MA, L., GUO, Y., AN, N., HU, Y., ZHENG, Y., SHI, Y., LI, Z., LIU, Q., CHEN, Y., ZHAO, J., QU, N., ZHAO, S., TIAN, F., WANG, X., WANG, H., XU, L., LIU, X., VINAR, T., et al. 2010. The sequence and de novo assembly of the giant panda genome. *Nature,* 463**,** 311-7.

LI, Y., WU, D. D., BOYKO, A. R., WANG, G. D., WU, S. F., IRWIN, D. M. & ZHANG, Y. P. 2014. Population variation revealed high-altitude adaptation of Tibetan mastiffs. *Mol Biol Evol,* 31**,** 1200-5.

LINDBLAD-TOH, K., WADE, C. M., MIKKELSEN, T. S., KARLSSON, E. K., JAFFE, D. B., KAMAL, M., CLAMP, M., CHANG, J. L., KULBOKAS, E. J., 3RD, ZODY, M. C., MAUCELI, E., XIE, X., BREEN, M., WAYNE, R. K., OSTRANDER, E. A., PONTING, C. P., GALIBERT, F., SMITH, D. R., DEJONG, P. J., KIRKNESS, E., ALVAREZ, P., BIAGI, T., BROCKMAN, W., BUTLER, J., CHIN, C. W., COOK, A., CUFF, J., DALY, M. J., DECAPRIO, D., GNERRE, S., GRABHERR, M., KELLIS, M., KLEBER, M., BARDELEBEN, C., GOODSTADT, L., HEGER, A., HITTE, C., KIM, L., KOEPFLI, K. P., PARKER, H. G., POLLINGER, J. P., SEARLE, S. M., SUTTER, N. B., THOMAS, R., WEBBER, C., BALDWIN, J., ABEBE, A., ABOUELLEIL, A., AFTUCK, L., AIT-ZAHRA, M., ALDREDGE, T., ALLEN, N., AN, P., ANDERSON, S., ANTOINE, C., ARACHCHI, H., ASLAM, A., AYOTTE, L., BACHANTSANG, P., BARRY, A., BAYUL, T., BENAMARA, M., BERLIN, A., BESSETTE, D., BLITSHTEYN, B., BLOOM, T., BLYE, J., BOGUSLAVSKIY, L., BONNET, C., BOUKHGALTER, B., BROWN, A., CAHILL, P., CALIXTE, N., CAMARATA, J., CHESHATSANG, Y., CHU, J., CITROEN, M., COLLYMORE, A., COOKE, P., DAWOE, T., DAZA, R., DECKTOR, K., DEGRAY, S., DHARGAY, N., DOOLEY, K., DOOLEY, K., DORJE, P., DORJEE, K., DORRIS, L., DUFFEY, N., DUPES, A., EGBIREMOLEN, O., ELONG, R., FALK, J., FARINA, A., FARO, S., FERGUSON, D., FERREIRA, P., FISHER, S., FITZGERALD, M., et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature,* 438**,** 803-19.

LIU, D., XIONG, H., ELLIS, A. E., NORTHRUP, N. C., RODRIGUEZ, C. O., JR., O'REGAN, R. M., DALTON, S. & ZHAO, S. 2014. Molecular homology and difference between spontaneous canine mammary cancer and human breast cancer. *Cancer Res,* 74**,** 5045-56.

LOBER, U., HOBBS, M., DAYARAM, A., TSANGARAS, K., JONES, K., ALQUEZAR-PLANAS, D. E., ISHIDA, Y., MEERS, J., MAYER, J., QUEDENAU, C., CHEN, W., JOHNSON, R. N., TIMMS, P., YOUNG, P. R., ROCA, A. L. & GREENWOOD, A. D. 2018. Degradation and remobilization of endogenous retroviruses by recombination during the earliest stages of a germ-line invasion. *Proc Natl Acad Sci U S A*.

MACDONALD, D. W. & SILLERO-ZUBIRI, C. 2004. *The Biology and Conservation of Wild Canids,* New York, Oxford University Press Inc.

MACFARLAN, T. S., GIFFORD, W. D., DRISCOLL, S., LETTIERI, K., ROWE, H. M., BONANOMI, D., FIRTH, A., SINGER, O., TRONO, D. & PFAFF, S. L. 2012. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature,* 487**,** 57-63.

MAGER, D. L. & STOYE, J. P. 2015. Mammalian Endogenous Retroviruses. *Microbiol Spectr,* 3**,** MDNA3-0009-2014.

MAGIORKINIS, G., GIFFORD, R. J., KATZOURAKIS, A., DE RANTER, J. & BELSHAW, R. 2012. Env-less endogenous retroviruses are genomic superspreaders. *Proc Natl Acad Sci U S A,* 109**,** 7385-90.

MARSDEN, C. D., ORTEGA-DEL VECCHYO, D., O'BRIEN, D. P., TAYLOR, J. F., RAMIREZ, O., VILA, C., MARQUES-BONET, T., SCHNABEL, R. D., WAYNE, R. K. & LOHMUELLER, K. E. 2016. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proc Natl Acad Sci U S A,* 113**,** 152-7.

MARTINEZ BARRIO, A., EKERLJUNG, M., JERN, P., BENACHENHOU, F., SPERBER, G. O., BONGCAM-RUDLOFF, E., BLOMBERG, J. & ANDERSSON, G. 2011. The first sequenced carnivore genome shows complex host-endogenous retrovirus relationships. *PLoS One,* 6**,** e19832.

MARUGGI, G., PORCELLINI, S., FACCHINI, G., PERNA, S. K., CATTOGLIO, C., SARTORI, D., AMBROSI, A., SCHAMBACH, A., BAUM, C., BONINI, C., BOVOLENTA, C., MAVILIO, F. & RECCHIA, A. 2009. Transcriptional enhancers induce insertional gene deregulation independently from the vector type and design. *Mol Ther,* 17**,** 851-6.

MAYERS, J. R., TORRENCE, M. E., DANAI, L. V., PAPAGIANNAKOPOULOS, T., DAVIDSON, S. M., BAUER, M. R., LAU, A. N., JI, B. W., DIXIT, P. D., HOSIOS, A. M., MUIR, A., CHIN, C. R., FREINKMAN, E., JACKS, T., WOLPIN, B. M., VITKUP, D. & VANDER HEIDEN, M. G. 2016. Tissue of origin dictates branched-chain amino acid metabolism in mutant Kras-driven cancers. *Science,* 353**,** 1161-5.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. & DEPRISTO, M. A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res,* 20**,** 1297-303.

MODIANO, J. F., BREEN, M., BURNETT, R. C., PARKER, H. G., INUSAH, S., THOMAS, R., AVERY, P. R., LINDBLAD-TOH, K., OSTRANDER, E. A., CUTTER, G. C. & AVERY, A. C. 2005. Distinct B-cell and T-cell lymphoproliferative disease prevalence among dog breeds indicates heritable risk. *Cancer Res,* 65**,** 5654-61.

MODIANO, J. F., GETZY, D. M., AKOL, K. G., VAN WINKLE, T. J. & COCKERELL, G. L. 1995. Retrovirus-like activity in an immunosuppressed dog: pathological and immunological findings. *J Comp Pathol,* 112**,** 165-83.

NASCIMENTO, F. F. & RODRIGO, A. G. 2016. Computational Evaluation of the Strict Master and Random Template Models of Endogenous Retrovirus Evolution. *PLoS One,* 11**,** e0162454.

NEI, M. & GOJOBORI, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol,* 3**,** 418-26.

NETHE, M., BERKHOUT, B. & VAN DER KUYL, A. C. 2005. Retroviral superinfection resistance. *Retrovirology,* 2**,** 52.

ONIONS, D. 1980. RNA-dependent DNA polymerase activity in canine lymphosarcoma. *Eur J Cancer,* 16**,** 345-50.

OSTERTAG, E. M. & KAZAZIAN, H. H., JR. 2001. Biology of mammalian L1 retrotransposons. *Annu Rev Genet,* 35**,** 501-38.

OWCZAREK-LIPSKA, M., JAGANNATHAN, V., DROGEMULLER, C., LUTZ, S., GLANEMANN, B., LEEB, T. & KOOK, P. H. 2013. A frameshift mutation in the cubilin gene (CUBN) in Border Collies with Imerslund-Grasbeck syndrome (selective cobalamin malabsorption). *PLoS One,* 8**,** e61144.

PARSONS, J. D. 1995. Miropeats: graphical DNA sequence comparisons. *Comput Appl Biosci,* 11**,** 615-9.

PARSONS, W. H., KOLAR, M. J., KAMAT, S. S., COGNETTA, A. B., 3RD, HULCE, J. J., SAEZ, E., KAHN, B. B., SAGHATELIAN, A. & CRAVATT, B. F. 2016. AIG1 and ADTRP are atypical integral membrane hydrolases that degrade bioactive FAHFAs. *Nat Chem Biol,* 12**,** 367-372.

PENDLETON, A. L., SHEN, F., TARAVELLA, A. M., EMERY, S., VEERAMAH, K. R., BOYKO, A. R. & KIDD, J. M. 2018. Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. *BMC Biol,* 16**,** 64.

PENG, X., ALFOLDI, J., GORI, K., EISFELD, A. J., TYLER, S. R., TISONCIK-GO, J., BRAWAND, D., LAW, G. L., SKUNCA, N., HATTA, M., GASPER, D. J., KELLY, S. M., CHANG, J., THOMAS, M. J., JOHNSON, J., BERLIN, A. M., LARA, M., RUSSELL, P., SWOFFORD, R., TURNER-MAIER, J., YOUNG, S., HOURLIER, T., AKEN, B., SEARLE, S., SUN, X., YI, Y., SURESH, M., TUMPEY, T. M., SIEPEL, A., WISELY, S. M., DESSIMOZ, C., KAWAOKA, Y., BIRREN, B. W., LINDBLAD-TOH, K., DI PALMA, F., ENGELHARDT, J. F., PALERMO, R. E. & KATZE, M. G. 2014. The draft genome sequence of the ferret (Mustela putorius furo) facilitates study of human respiratory disease. *Nat Biotechnol,* 32**,** 1250-5.

PERK, K., SAFRAN, N. & DAHLBERG, J. E. 1992. Propagation and characterization of novel canine lentivirus isolated from a dog. *Leukemia,* 6 Suppl 3**,** 155S-157S.

PONTIUS, J. U., MULLIKIN, J. C., SMITH, D. R., AGENCOURT SEQUENCING, T., LINDBLAD-TOH, K., GNERRE, S., CLAMP, M., CHANG, J., STEPHENS, R., NEELAM, B., VOLFOVSKY, N., SCHAFFER, A. A., AGARWALA, R., NARFSTROM, K., MURPHY, W. J., GIGER, U., ROCA, A. L., ANTUNES, A., MENOTTI-RAYMOND, M., YUHKI, N., PECON-SLATTERY, J., JOHNSON, W. E., BOURQUE, G., TESLER, G., PROGRAM, N. C. S. & O'BRIEN, S. J. 2007. Initial sequence and comparative analysis of the cat genome. *Genome Res,* 17**,** 1675-89.

PRICE, M. N., DEHAL, P. S. & ARKIN, A. P. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One,* 5**,** e9490.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I., DALY, M. J. & SHAM, P. C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet,* 81**,** 559-75.

QUINLAN, A. R. 2014. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics,* 47**,** 11 12 1-34.

RAUSCH, T., ZICHNER, T., SCHLATTL, A., STUTZ, A. M., BENES, V. & KORBEL, J. O. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics,* 28**,** i333-i339.

REBOLLO, R., ROMANISH, M. T. & MAGER, D. L. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu Rev Genet,* 46**,** 21-42.

ROBINSON, J. A., ORTEGA-DEL VECCHYO, D., FAN, Z., KIM, B. Y., VONHOLDT, B. M., MARSDEN, C. D., LOHMUELLER, K. E. & WAYNE, R. K. 2016. Genomic Flatlining in the Endangered Island Fox. *Curr Biol,* 26**,** 1183-9.

ROCA, A. L., PECON-SLATTERY, J. & O'BRIEN, S. J. 2004. Genomically intact endogenous feline leukemia viruses of recent origin. *J Virol,* 78**,** 4370-5.

SAFRAN, N., PERK, K., EYAL, O. & DAHLBERG, J. E. 1992. Isolation and preliminary characterisation of a novel retrovirus isolated from a leukaemic dog. *Res Vet Sci,* 52**,** 250-5.

SAWYER, S. L., EMERMAN, M. & MALIK, H. S. 2007. Discordant evolution of the adjacent antiretroviral genes TRIM22 and TRIM5 in mammals. *PLoS Pathog,* 3**,** e197.

SINHA, A. & JOHNSON, W. E. 2017. Retroviruses of the RDR superinfection interference group: ancient origins and broad host distribution of a promiscuous Env gene. *Curr Opin Virol,* 25**,** 105-112.

SKOGLUND, P., ERSMARK, E., PALKOPOULOU, E. & DALEN, L. 2015. Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Curr Biol,* 25**,** 1515-9.

STANKIEWICZ, P. & LUPSKI, J. R. 2002. Molecular-evolutionary mechanisms for genomic disorders. *Curr Opin Genet Dev,* 12**,** 312-9.

STOCKING, C. & KOZAK, C. A. 2008. Murine endogenous retroviruses. *Cell Mol Life Sci,* 65**,** 3383-98.

STOYE, J. P. 2012. Studies of endogenous retroviruses reveal a continuing evolutionary saga. *Nat Rev Microbiol,* 10**,** 395-406.

TARLINTON, R. E., BARFOOT, H. K., ALLEN, C. E., BROWN, K., GIFFORD, R. J. & EMES, R. D. 2013. Characterisation of a group of endogenous gammaretroviruses in the canine genome. *Vet J,* 196**,** 28-33.

TARLINTON, R. E., MEERS, J. & YOUNG, P. R. 2006. Retroviral invasion of the koala genome. *Nature,* 442**,** 79-81.

THORVEN, M., GRAHN, A., HEDLUND, K. O., JOHANSSON, H., WAHLFRID, C., LARSON, G. & SVENSSON, L. 2005. A homozygous nonsense mutation (428G-->A) in the human secretor (FUT2) gene provides resistance to symptomatic norovirus (GGII) infections. *J Virol,* 79**,** 15351-5.

TOMLEY, F. M., ARMSTRONG, S. J., MAHY, B. W. & OWEN, L. N. 1983. Reverse transcriptase activity and particles of retroviral density in cultured canine lymphosarcoma supernatants. *Br J Cancer,* 47**,** 277-84.

TROYER, J. L., PECON-SLATTERY, J., ROELKE, M. E., BLACK, L., PACKER, C. & O'BRIEN, S. J. 2004. Patterns of feline immunodeficiency virus multiple infection and genome divergence in a free-ranging population of African lions. *J Virol,* 78**,** 3777-91.

VAMATHEVAN, J. J., HALL, M. D., HASAN, S., WOOLLARD, P. M., XU, M., YANG, Y., LI, X., WANG, X., KENNY, S., BROWN, J. R., HUXLEY-JONES, J., LYON, J., HASELDEN, J., MIN, J. & SANSEAU, P. 2013. Minipig and beagle animal model genomes aid species selection in pharmaceutical discovery and development. *Toxicol Appl Pharmacol,* 270**,** 149-57.

WANG, G. D., ZHAI, W., YANG, H. C., FAN, R. X., CAO, X., ZHONG, L., WANG, L., LIU, F., WU, H., CHENG, L. G., POYARKOV, A. D., POYARKOV, N. A., JR., TANG, S. S., ZHAO, W. M., GAO, Y., LV, X. M., IRWIN, D. M., SAVOLAINEN, P., WU, C. I. & ZHANG, Y. P. 2013. The genomics of selection in dogs and the parallel evolution between dogs and humans. *Nat Commun,* 4**,** 1860.

WANG, G. D., ZHAI, W., YANG, H. C., WANG, L., ZHONG, L., LIU, Y. H., FAN, R. X., YIN, T. T., ZHU, C. L., POYARKOV, A. D., IRWIN, D. M., HYTONEN, M. K., LOHI, H., WU, C. I., SAVOLAINEN, P. & ZHANG, Y. P. 2016. Out of southern East Asia: the natural history of domestic dogs across the world. *Cell Res,* 26**,** 21-33.

WANG, X. & TEDFORD, R. H. 2008. *DOGS: Their fossil relatives and evolutionary history,* New York Chichester, West Sussex, Columbia University Press.

WEISS, R. A. & STOYE, J. P. 2013. Virology. Our viral inheritance. *Science,* 340**,** 820-1.

WILDSCHUTTE, J. H., BARON, A., DIROFF, N. M. & KIDD, J. M. 2015. Discovery and characterization of Alu repeat sequences via precise local read assembly. *Nucleic Acids Res,* 43**,** 10292-307.

WILDSCHUTTE, J. H., WILLIAMS, Z. H., MONTESION, M., SUBRAMANIAN, R. P., KIDD, J. M. & COFFIN, J. M. 2016. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. *Proc Natl Acad Sci U S A,* 113**,** E2326-34.

XU, B. & YANG, Z. 2013. PAMLX: a graphical user interface for PAML. *Mol Biol Evol,* 30**,** 2723-4.

YAN, J., CHEN, G., ZHAO, X., CHEN, F., WANG, T. & MIAO, F. 2018. High expression of diffuse panbronchiolitis critical region 1 gene promotes cell proliferation, migration and

invasion in pancreatic ductal adenocarcinoma. *Biochem Biophys Res Commun,* 495**,** 1908-1914.

YOUNG, G. R., EKSMOND, U., SALCEDO, R., ALEXOPOULOU, L., STOYE, J. P. & KASSIOTIS, G. 2012. Resurrection of endogenous retroviruses in antibody-deficient mice. *Nature,* 491**,** 774-8.

ZHANG, W., FAN, Z., HAN, E., HOU, R., ZHANG, L., GALAVERNI, M., HUANG, J., LIU, H., SILVA, P., LI, P., POLLINGER, J. P., DU, L., ZHANG, X., YUE, B., WAYNE, R. K. & ZHANG, Z. 2014. Hypoxia adaptations in the grey wolf (Canis lupus chanco) from Qinghai-Tibet Plateau. *PLoS Genet,* 10**,** e1004466.

Figure 1

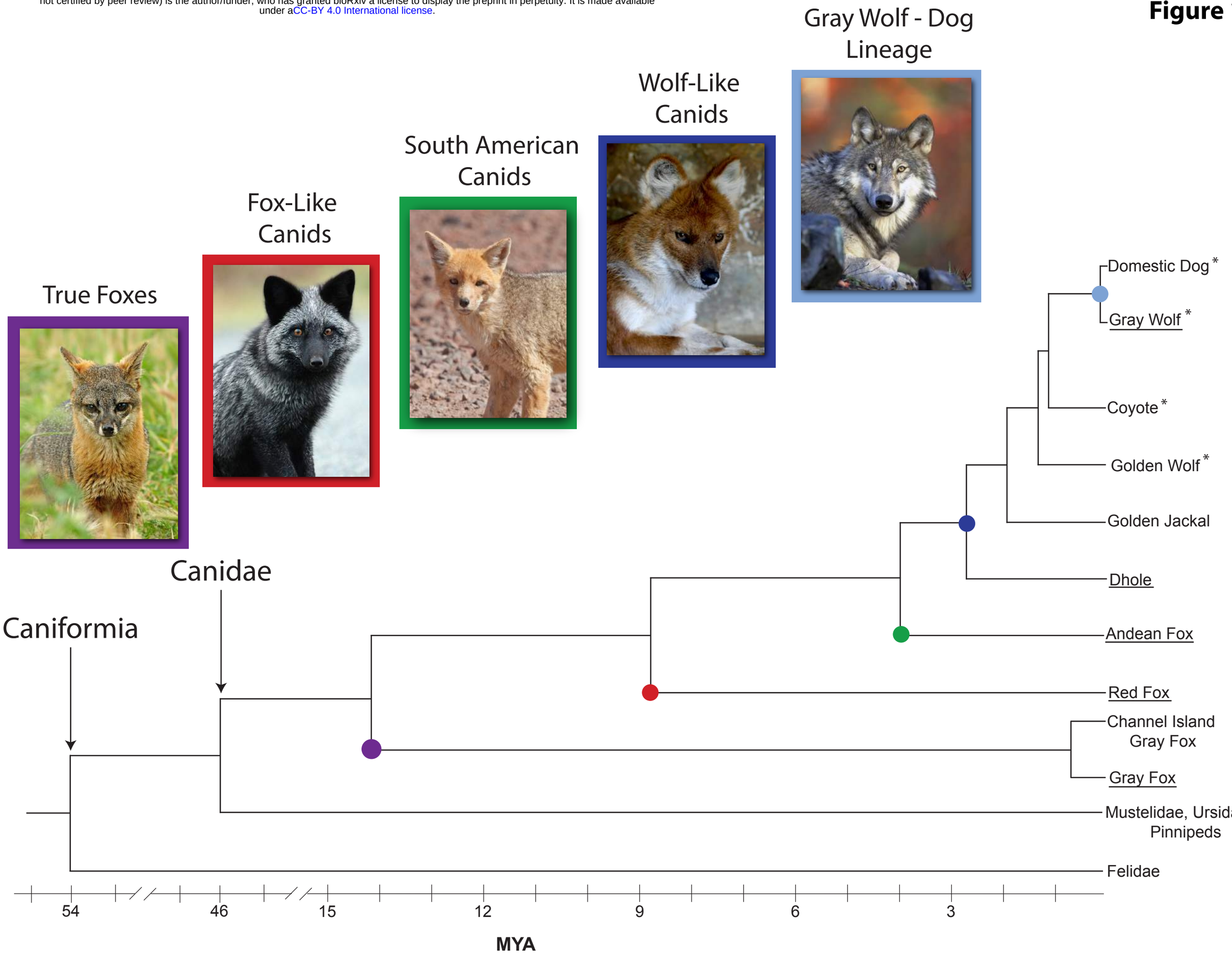


Figure 2

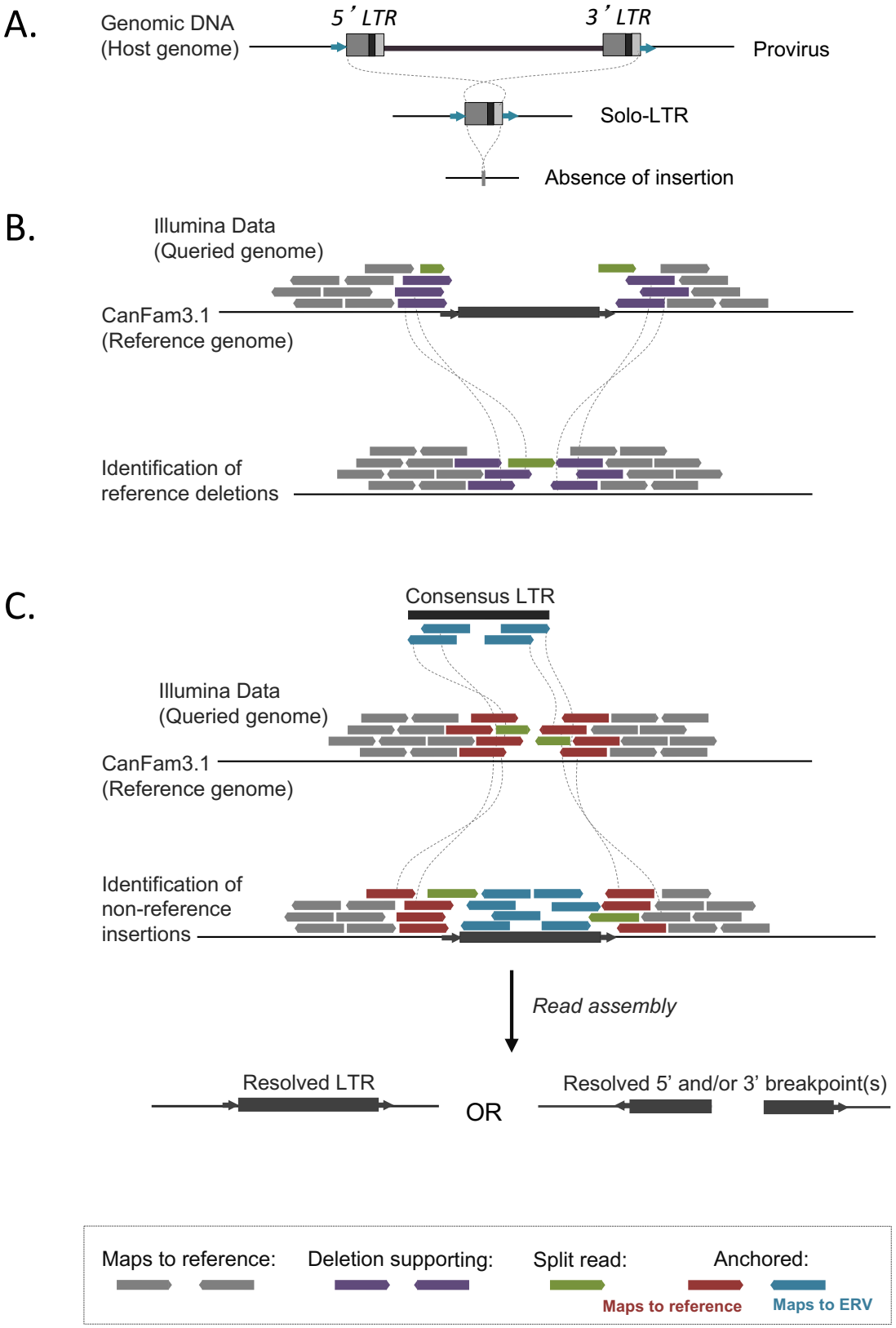
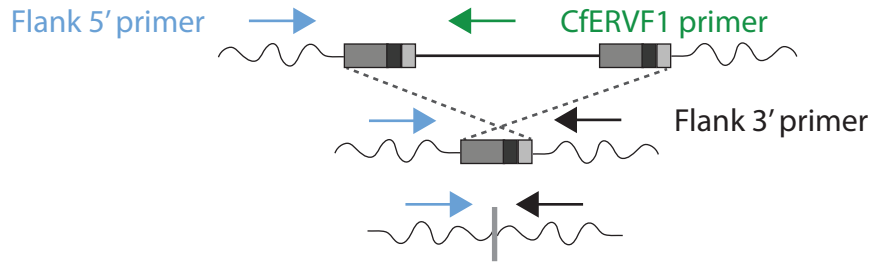


Figure 3

A.


Flank 5′ primer
CfERVF1 primer
Flank 3′ primer

B.

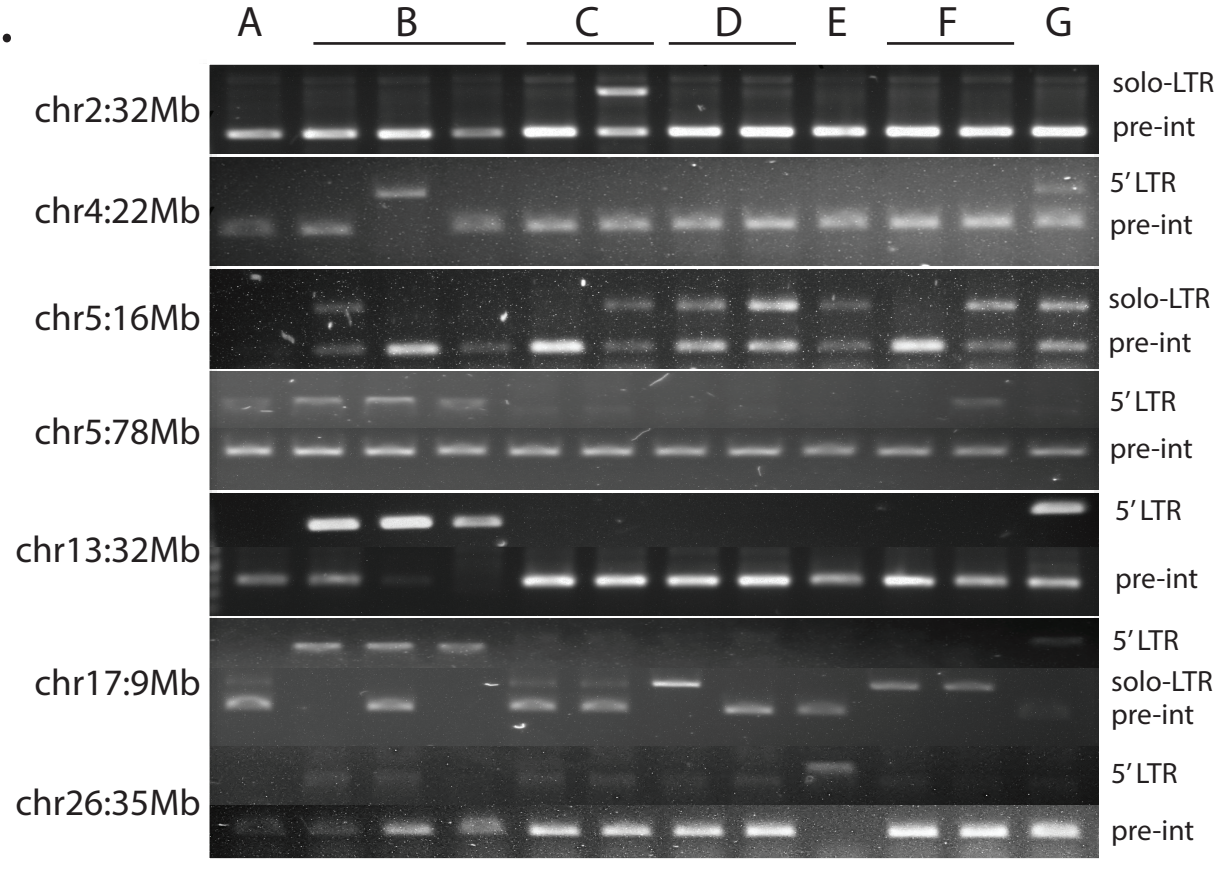

A
B
C
D
E
F
G
chr2:32Mb
chr4:22Mb
chr5:16Mb
chr5:78Mb
chr13:32Mb
chr17:9Mb
chr26:35Mb
solo-LTR
pre-int
5′ LTR
pre-int
solo-LTR
pre-int
5′ LTR
pre-int
5′ LTR
pre-int
5′ LTR
solo-LTR
pre-int
5′ LTR
pre-int

Figure 4

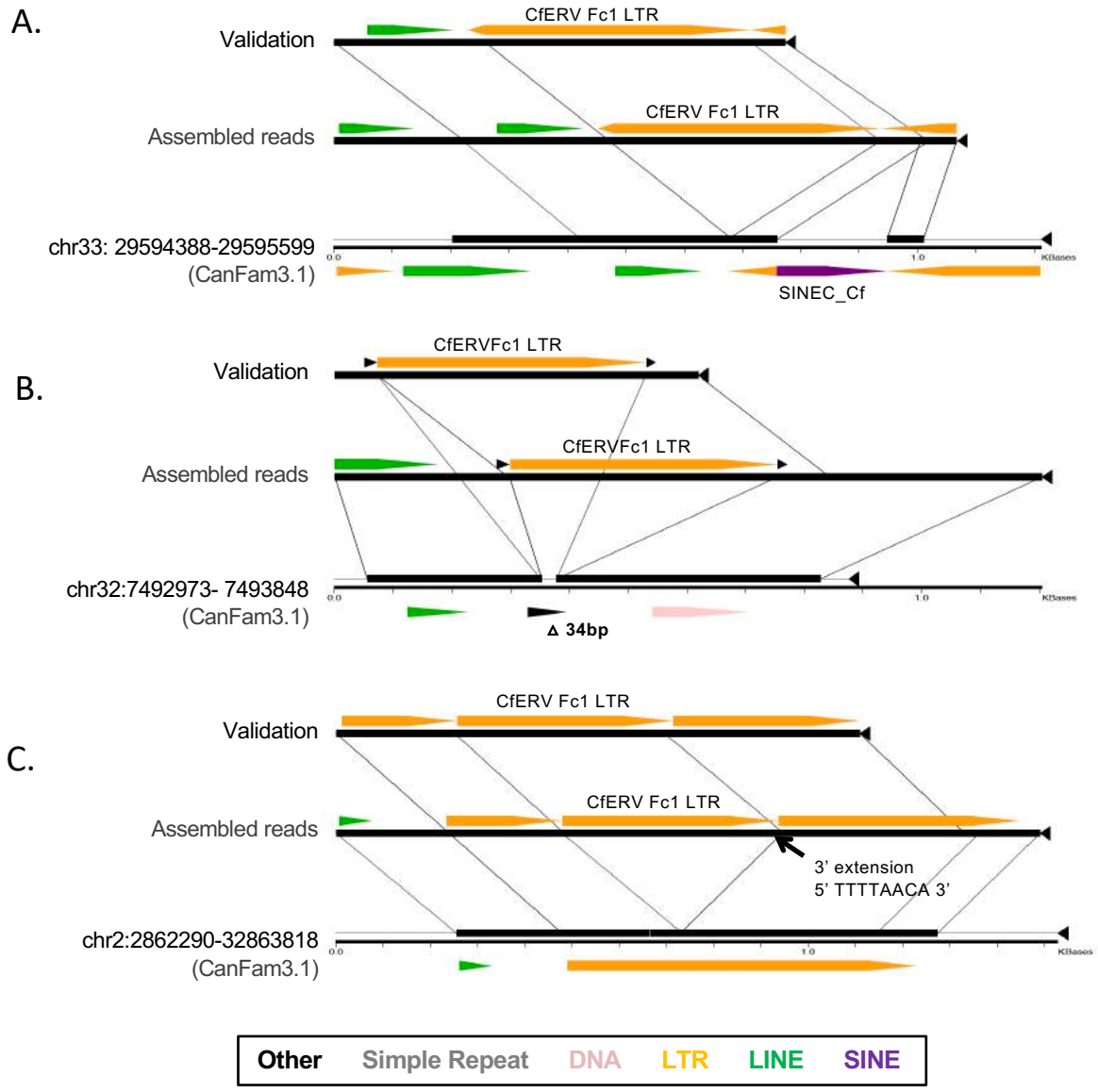


Figure 5

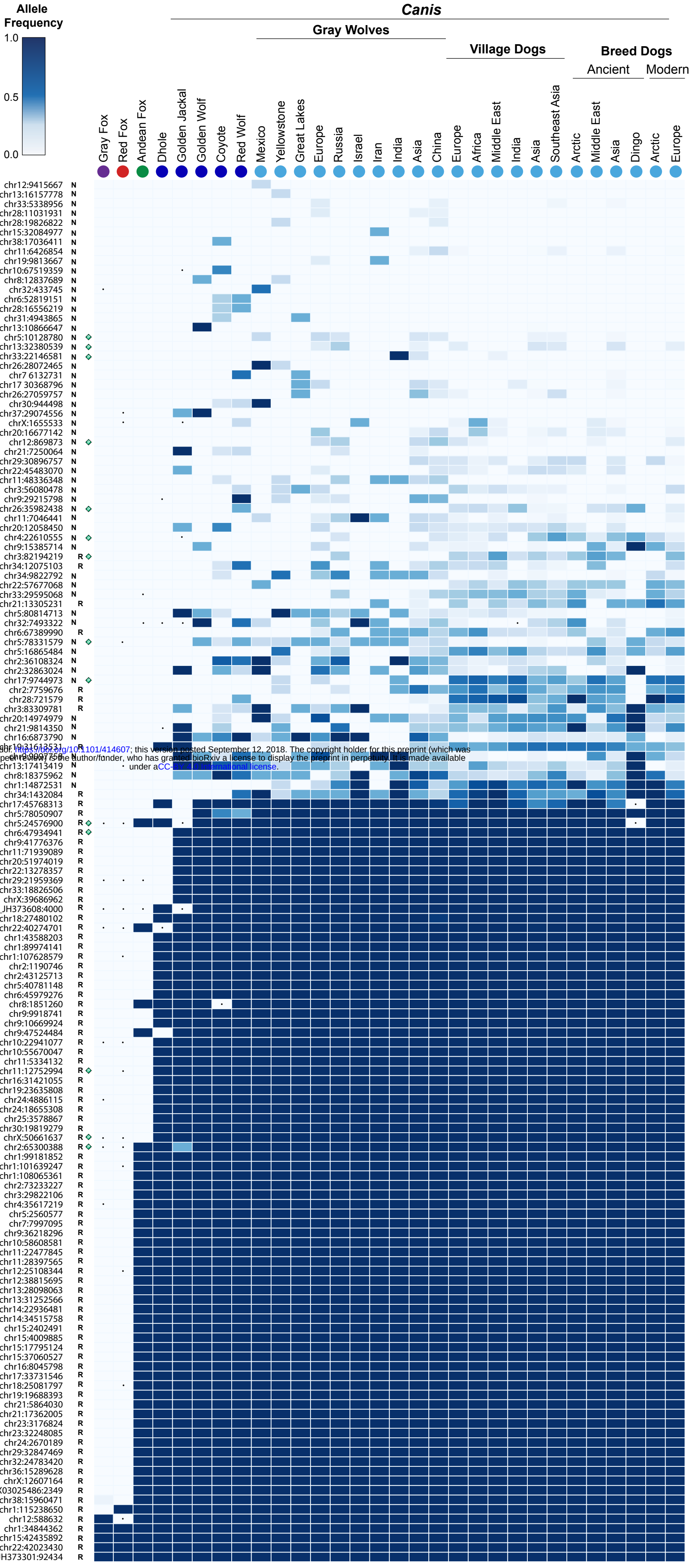


Figure 6

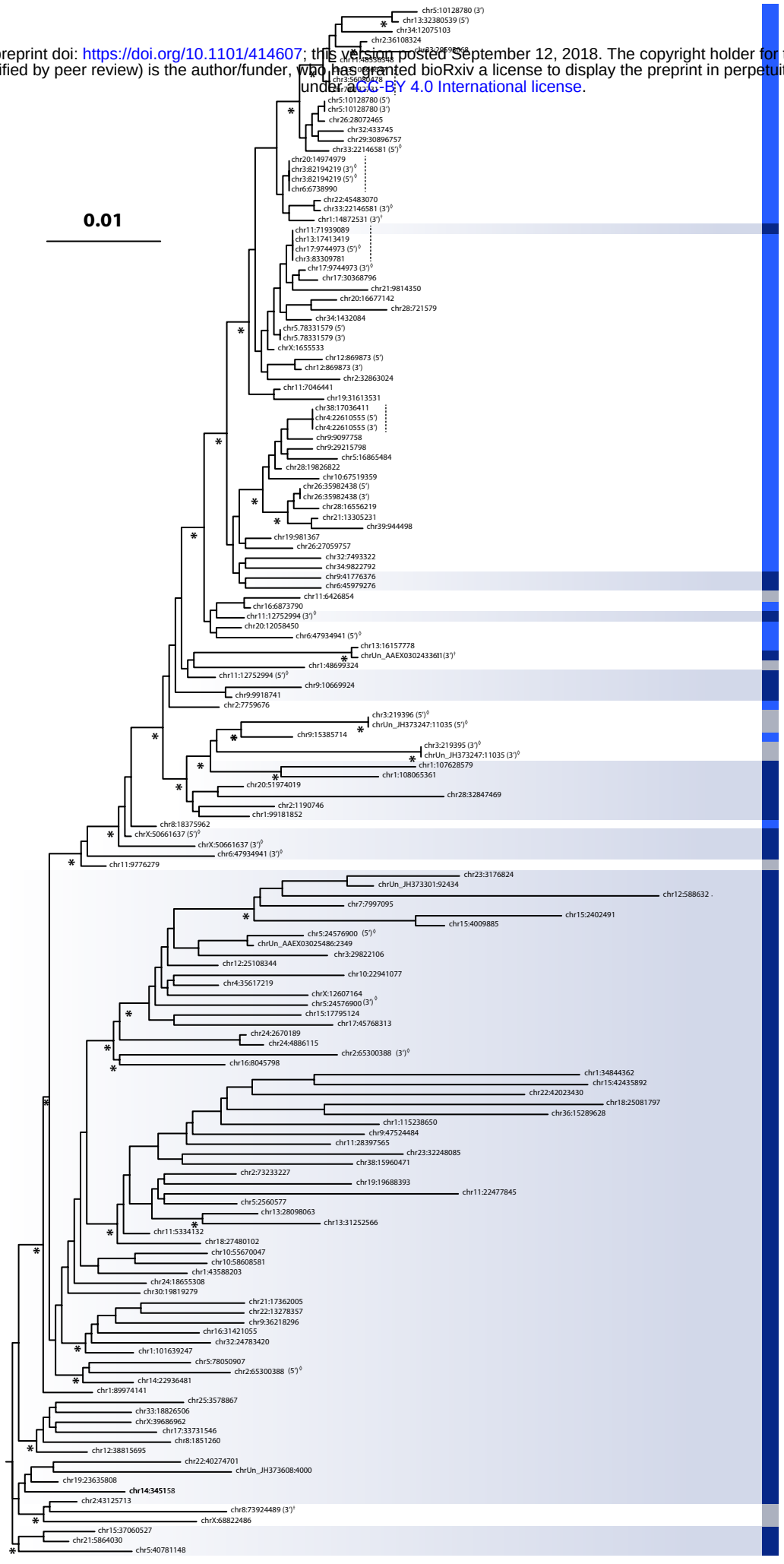
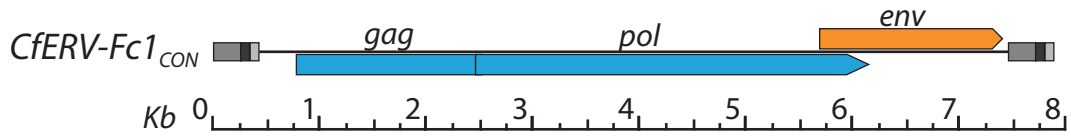


Figure 7

A.

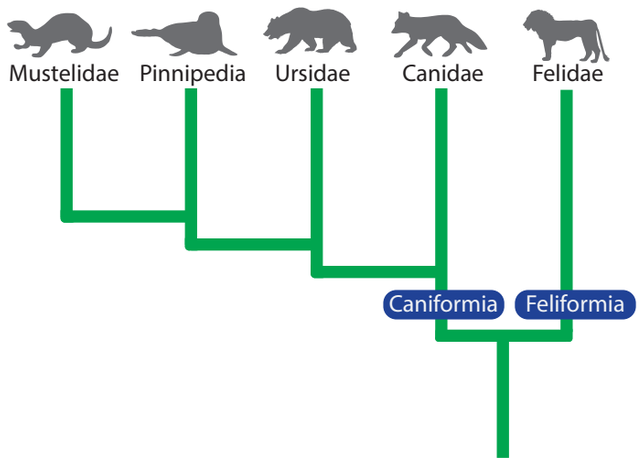


B.

	Unfixed in *Canis*†		5'LTR-3'LTR changes	Estimated T.O.F. (mya)
Non-reference *proviral Fc1(a) insertions*	†	*chr4:22610555*	*0*	*<1.64*
	†	*chr5:10128780*	*0*	*<1.64*
	†	*chr5:78331579*	*0*	*<1.64*
	†	*chr12:869873*	*1*	*~1.81*
	†	*chr13:32380539*	*1*	*~1.64*
	†	*chr17:9744973*	*n.a.*	*n.a.*
	†	*chr26:35982438*	*0*	*<1.64*
	†	*chr33:22146581*	*4*	*~6.58*
Reference *(CanFam3.1) proviral Fc1(a) insertions*		*chr1:48699324*	*n.a.*	*n.a.*
		chr2:65300388	*13*	*~20.49*
		chr3:219396	*12*	*~19.74*
	†	*chr3:82194219*	*0*	*<1.64*
		chr5:24576900	*9*	*~14.80*
		chr6:47934941	*10*	*~16.52*
		chr8:73924489	*n.a.*	*n.a.*
		chr11:12752994	*3*	*~4.94*
		chrX:50661637	*4*	*~6.62*
		chrUnAAEX03024336:1	*n.a.*	*n.a.*
		chrUnJH373247:11035	*12*	*~19.74*

Figure 8

A.



~54 million ya

The two major lineages of Carnivora, Feliformia and Caniformia, split approximately 54 million years ago.

55 million

B.

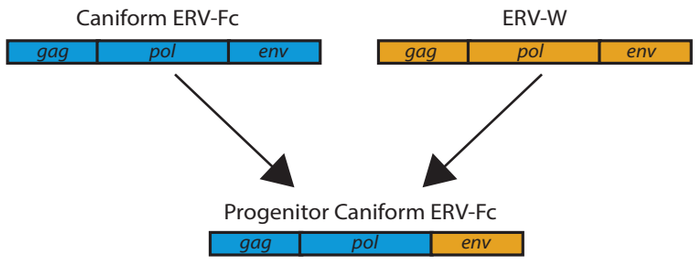
~46 million ya

Early *Canidae* species first arose in North America, likely in what is now the Southwestern United States.

Ancestral Germline Invasion

13 million

C.



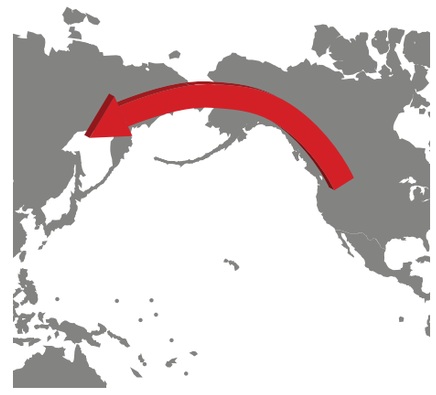
~46-12.9 million ya

Prior to the emergence of the true fox lineage, a recombination event formed a chimeric provirus that would eventually be transmitted to canid ancestors.

True Fox Lineage

Fox-Like Canids Lineage

D.



~6 million ya

Canis appeared approximately 6 mya in North America. Shortly thereafter, migration of *Canis* and other canids to Eurasia began via the Bering Strait land bridge.

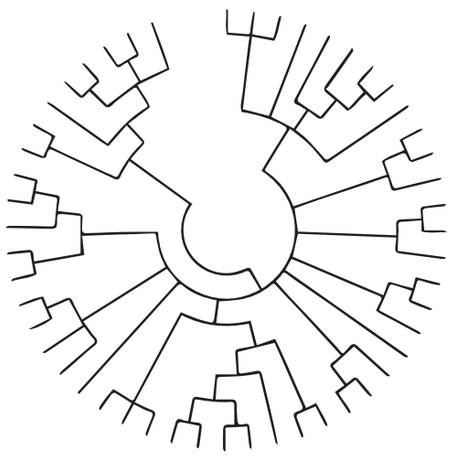
4 million

South American Canid Lineage

Canis "burst" of infections begins

Dhole Lineage

E.



~1.8 million ya

Known as the "wolf event", a species radiation of *Canis* members occurred in Eurasia, resulting in many modern and extinct wolf lineages, including the gray wolf.

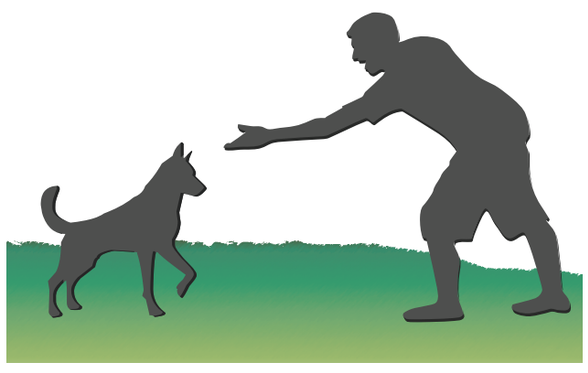
80 thousand

F.

~70,000 - 10,000 ya

The last Ice Age reformed the Bering Strait land bridge, permitting the re-entry of gray wolves and domesticated dogs into North America, up until the land bridge disappeared ~10,000 YBP.

G.



~40,000 ya - Present

Domestication of gray wolves occurred as far back as 40,000 years ago in Eurasia. Continued selection by humans over milennia has yielded hundreds of dog breeds.