


RESEARCH

Open Access



Origin and recent expansion of an endogenous gammaretroviral lineage in domestic and wild canids

Julia V. Halo^{1*}, Amanda L. Pendleton², Abigail S. Jarosz¹, Robert J. Gifford³, Malika L. Day¹ and Jeffrey M. Kidd^{2,4} 

Abstract

Background: Vertebrate genomes contain a record of retroviruses that invaded the germlines of ancestral hosts and are passed to offspring as endogenous retroviruses (ERVs). ERVs can impact host function since they contain the necessary sequences for expression within the host. Dogs are an important system for the study of disease and evolution, yet no substantiated reports of infectious retroviruses in dogs exist. Here, we utilized Illumina whole genome sequence data to assess the origin and evolution of a recently active gammaretroviral lineage in domestic and wild canids.

Results: We identified numerous recently integrated loci of a canid-specific ERV-Fc sublineage within *Canis*, including 58 insertions that were absent from the reference assembly. Insertions were found throughout the dog genome including within and near gene models. By comparison of orthologous occupied sites, we characterized element prevalence across 332 genomes including all nine extant canid species, revealing evolutionary patterns of ERV-Fc segregation among species as well as subpopulations.

Conclusions: Sequence analysis revealed common disruptive mutations, suggesting a predominant form of ERV-Fc spread by *trans* complementation of defective proviruses. ERV-Fc activity included multiple circulating variants that infected canid ancestors from the last 20 million to within 1.6 million years, with recent bursts of germline invasion in the sublineage leading to wolves and dogs.

Keywords: Canine, Retrovirus, Endogenous retrovirus, Insertional polymorphism, *Canidae*

Background

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 integration that is transmitted vertically to offspring as an endogenous retrovirus (ERV). Over time, the ERV may reach high frequency within a population and eventual fixation within a species [1]. Through repeated germline invasion and expansion over millions of years, ERVs have accumulated to considerable proportions in the genomes of many vertebrates.

ERVs have been referred to as ‘genomic fossils’ of their once-infectious counterparts, providing a limited record of exogenous retroviruses that previously infected a species, became endogenized, and spread throughout a species [1]. Among vertebrate species, the majority of ERVs are thought to provide no advantage to the host and have progressively degenerated over time due to accumulated mutations or from recombination between the proviral

*Correspondence: juliahw@bgsu.edu

¹ Department of Biological Sciences, Bowling Green State University, Bowling Green, OH 43403, USA

Full list of author information is available at the end of the article



LTRs resulting in a solo LTR [1]. An ERV is replicated as part of the host genome and evolves with a slower rate than an infectious virus, with recently formed ERVs tending to bear close resemblance to their exogenous equivalent and possessing a greater potential to retain functional properties. Indeed, several species' genomes are known to harbor ERVs bearing signatures of relatively recent germline invasion [2–12]. These properties include the presence of some or all viral reading frames, transcriptional activation, high LTR–LTR nucleotide identity, and integrants segregating as unfixed alleles among species or within populations. Other evidence suggests evolutionary roles in host physiology, for example by altering genomic structure or gene regulation by providing alternative promoters, enhancers, splice sites, or termination signals [13–15]. There are also instances in which ERV gene products have been co-opted for host functions. Notable examples include syncytial trophoblast fusion in eutherian animals [16] and blocking of infection from exogenous viruses [17–21].

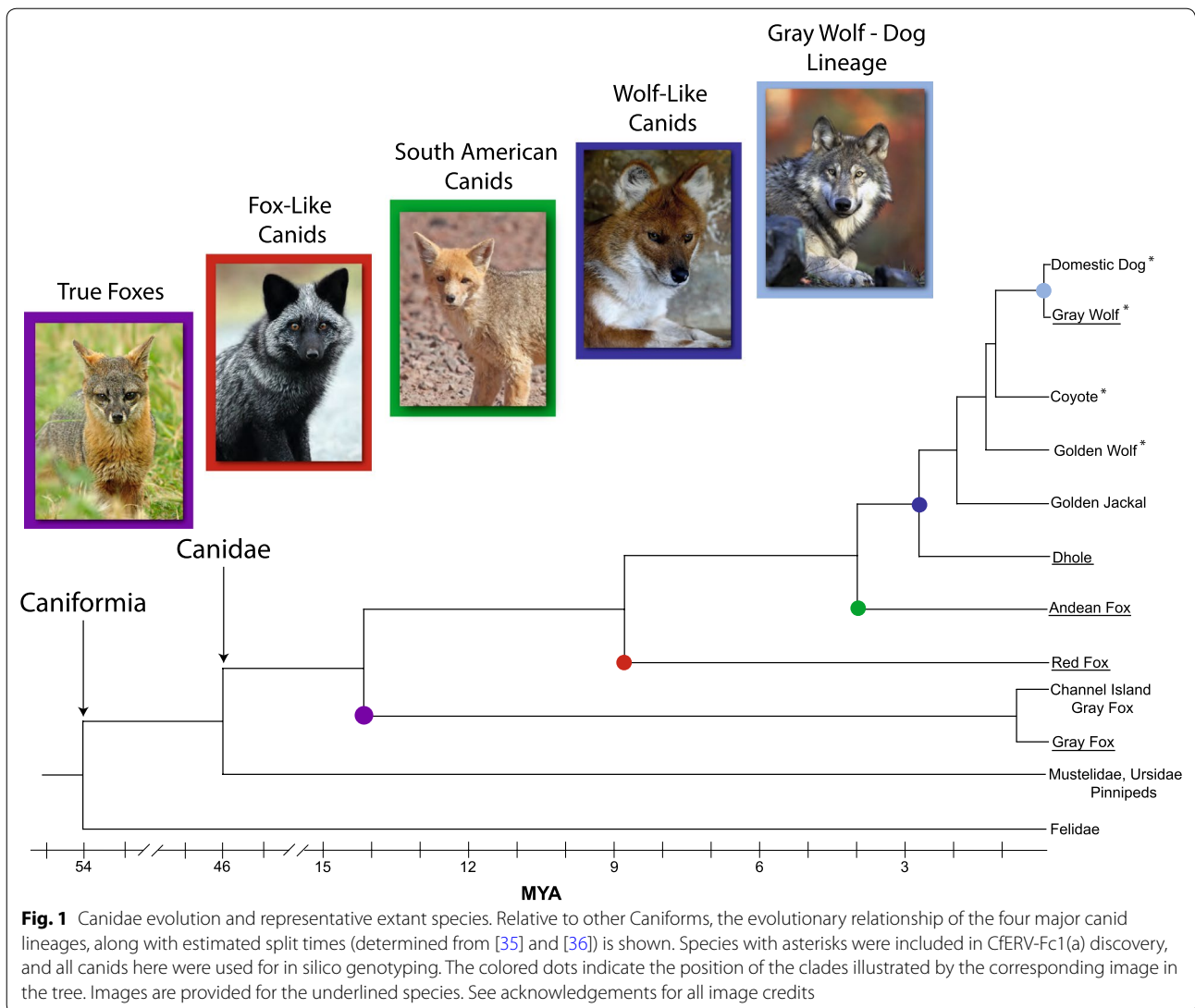
The endogenous retroviruses classified as ERV-Fc are distant relatives of extant gammaretroviruses (also referred to as gamma-like, or γ -like) [11, 22]. As is typical of most ERV groups, ERV-Fc was originally named for its use of a primer binding site complementary to the tRNA used during reverse transcription (tRNA^{phe}) [23]. 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 [24]. 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. ERV-Fc was first characterized as a putatively extinct, low copy number lineage that infected the ancestor of all simians and later contributed to independent germline invasions in primate lineages [22]. 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 [10]. 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 [10].

In comparison to humans and other mammals, the domestic dog (*Canis lupus familiaris*) displays a substantially lower ERV presence, with only 0.15% of the genome recognizably of retroviral origin [11, 25]. To date, 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

[26–32]. Nonetheless, the ERV fossil record in the dog 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 [11], suggesting most CfERV lineages ceased replicating long ago. An exception comes from a minor subset of ERV-Fc-related proviruses that possess high LTR nucleotide identity and ORFs [11]. This ERV lineage was recently detailed by Diehl, et al., in which the authors described a distinct ERV-Fc lineage in the Caniformia suborder, to which dogs and other canids belong, classified therein as ERV-Fc1 [10]. 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) [10]. This recombination event most likely arose from reverse transcription of co-packaged but distinct ERV RNAs in the same virion, and may have contributed to altered pathogenic properties of the chimeric virus, as has been shown [33]. A derived sublineage of the recombinant, CfERV-Fc1(a), later spread to and infected canid ancestors via a cross-species transmission from an unidentified source, after which the lineage endogenized canids until at least the last 1–2 million years [10]. It is this lineage that accounts for the few recent CfERV integrants in the dog reference assembly [10].

The domestic dog belongs to the family Canidae which arose in North America during the late Eocene (~46 mya) and is the oldest family of Carnivora [34, 35]. Following multiple crossings of the Bering Strait land bridge to Eurasia, canids underwent massive radiations, leading to the ancestors of most modern canids [34]. 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 [34]. 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 [34]. Approximately 1.1 mya, *Canis lupus*, the direct ancestor of the dog, emerged in Eurasia [36]. 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 [34]. 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 (Fig. 1), we assessed the origin, evolution, and impact of the recently active γ -like CfERV-Fc1(a) lineage, yielding



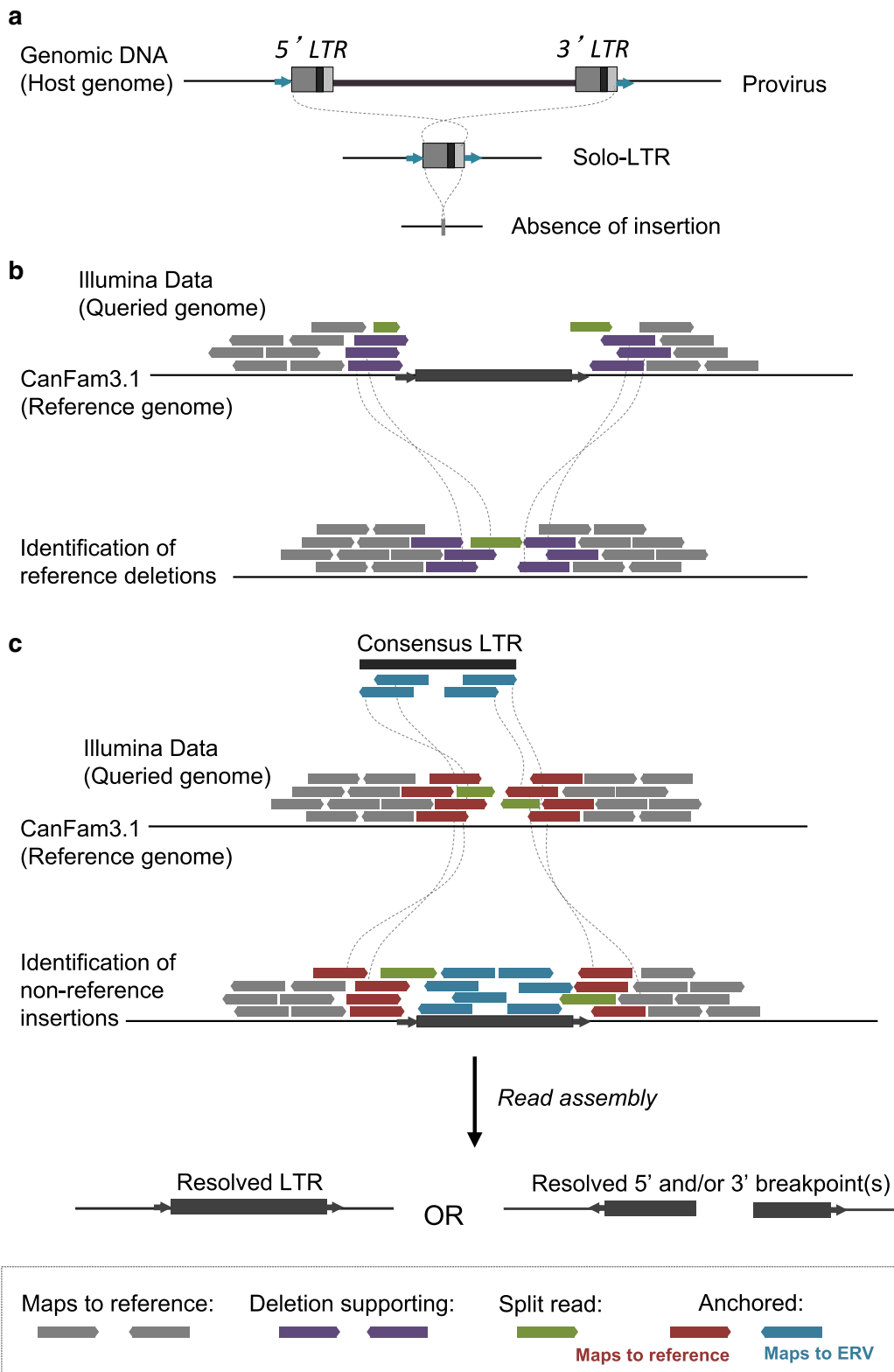
the most comprehensive assessment of ERV activity in carnivores to date. We used Illumina sequence data to characterize CfERV-Fc1(a) integrants in dogs and wild canids, resulting in the discoveries of numerous polymorphic and novel insertions. We further delineate the presence of this ERV group through comparisons of orthologous insertions across species in order to provide a rich evolutionary history of CfERV-Fc1(a) activity. 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 (Fig. 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, which are absent from the draft genomes of other extant Caniformia species. These insertions are referred to as ‘reference’ throughout the text due to their presence in the dog



(See figure on previous page.)

Fig. 2 Strategy for detecting insertionally polymorphic ERV variants. **a** ERV allelic presence. Upper: full-length provirus; Mid: solo LTR recombinant; Lower, unoccupied (pre-integration) site. **b** Strategy for detection of reference ERV deletions. Illumina read pairs were mapped to the CanFam3.1 reference, deletion-supporting read pairs and split reads identified using the program Delly [37], and candidate calls then intersected with RepeatMasker outputs considering 'CFERV1' repeats. Deletion calls within a size range corresponding to a solo LTR or provirus were selected for further analysis. **c** Strategy for detection of non-reference ERV insertions. ERV insertion-supporting anchored read pairs were identified from merged Illumina data mapped to the CanFam3.1 reference using the RetroSeq program [90]. Insertion-supporting read pairs and intersecting split reads were assembled, assemblies for which 'CFERV1' sequence was present were identified by RepeatMasker analysis, and the assembled contigs then re-mapped to the dog CanFam3.1 reference for precise breakpoint identification

reference genome. We then intersected the reference loci with deletions predicted by Delly [37] within a sample set of 101 resequenced *Canis* individuals, specifically including jackals, coyotes, gray wolves, and dogs (Additional file 1: Table S1). Candidate deletions were classified as those that intersected with annotated 'CFERV1'-related loci and were within the size range of the solo LTR or provirus (~457 and ~7885 bp, respectively; Fig. 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 [9, 38] (Fig. 2b). This process identified 58 unique non-reference insertions, all of which derived from 'CFERV1'-related elements per RepeatMasker analysis, as well as one insertion located in a gap in the existing CanFam3.1 reference assembly. 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 Additional file 2: Figure S1 and Additional file 3: 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 (Fig. 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 for which genomic DNA was available, and performed additional screening of each site to

discriminate solo LTR and full-length integrants (Additional file 3: 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. Locus-specific sequencing was used to obtain the full nucleotide sequence for 33 of the 34 insertions, with preference for sequencing placed on the provirus allele when present (8 proviruses). The provirus at the final site (chr5:78,331,579) was obtained using PCR-free PacBio sequencing and contained a segment of A-rich, low complexity sequence as part of an insertion of non-ERV sequence within the *gag* gene (~2250 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 (Fig. 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 (Fig. 4b). Finally, an assembled

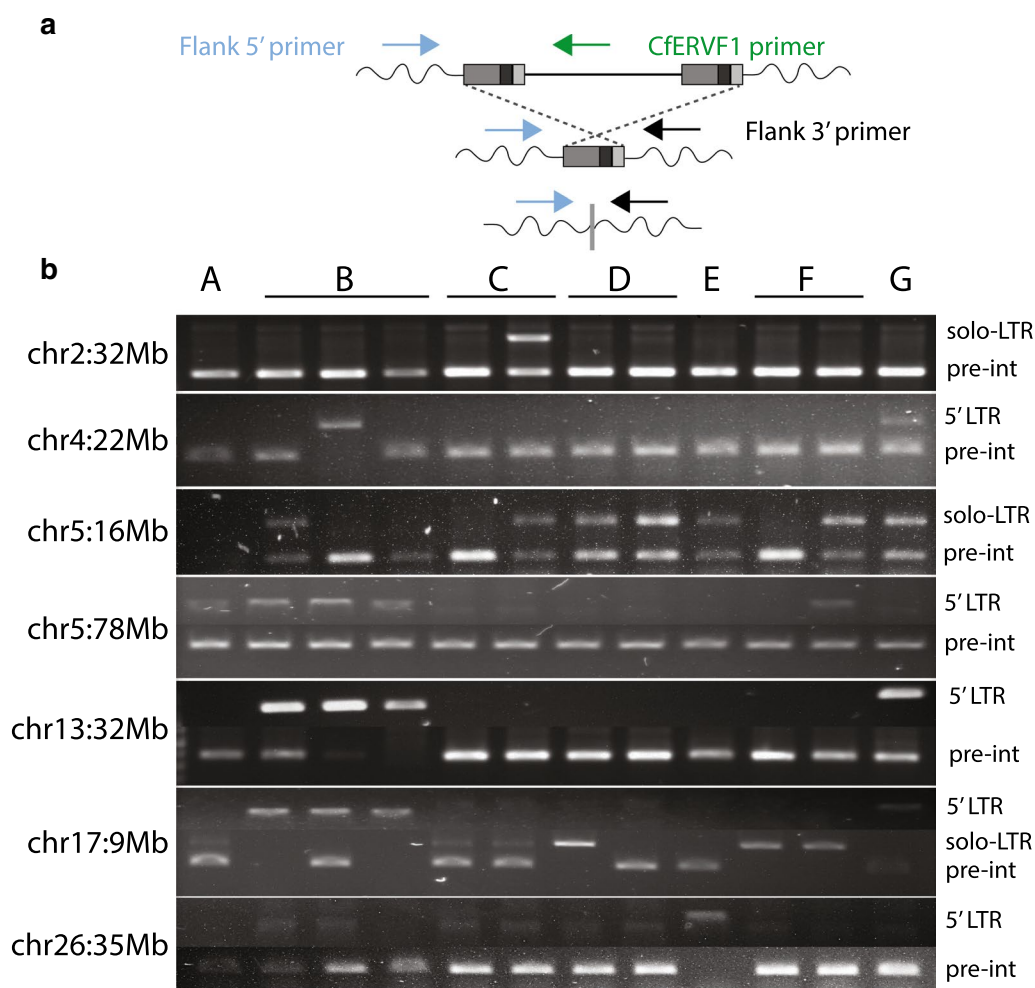


Fig. 3 Representative allele screening of polymorphic loci. PCR screens of a subset of non-reference CfERV-Fc1(a) integrants. Validation of insertionally polymorphic sites was performed for seven candidate sites across genomic DNA from a panel of breed dogs. **a** Strategy for primer design and allele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

solo LTR that mapped to chr2:32,863,024 contained an apparent 8 bp extension from the canonical CfERV1 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 (Fig. 4c). The extension is similar in sequence to the consensus CfERV1 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 [39, 40]). 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

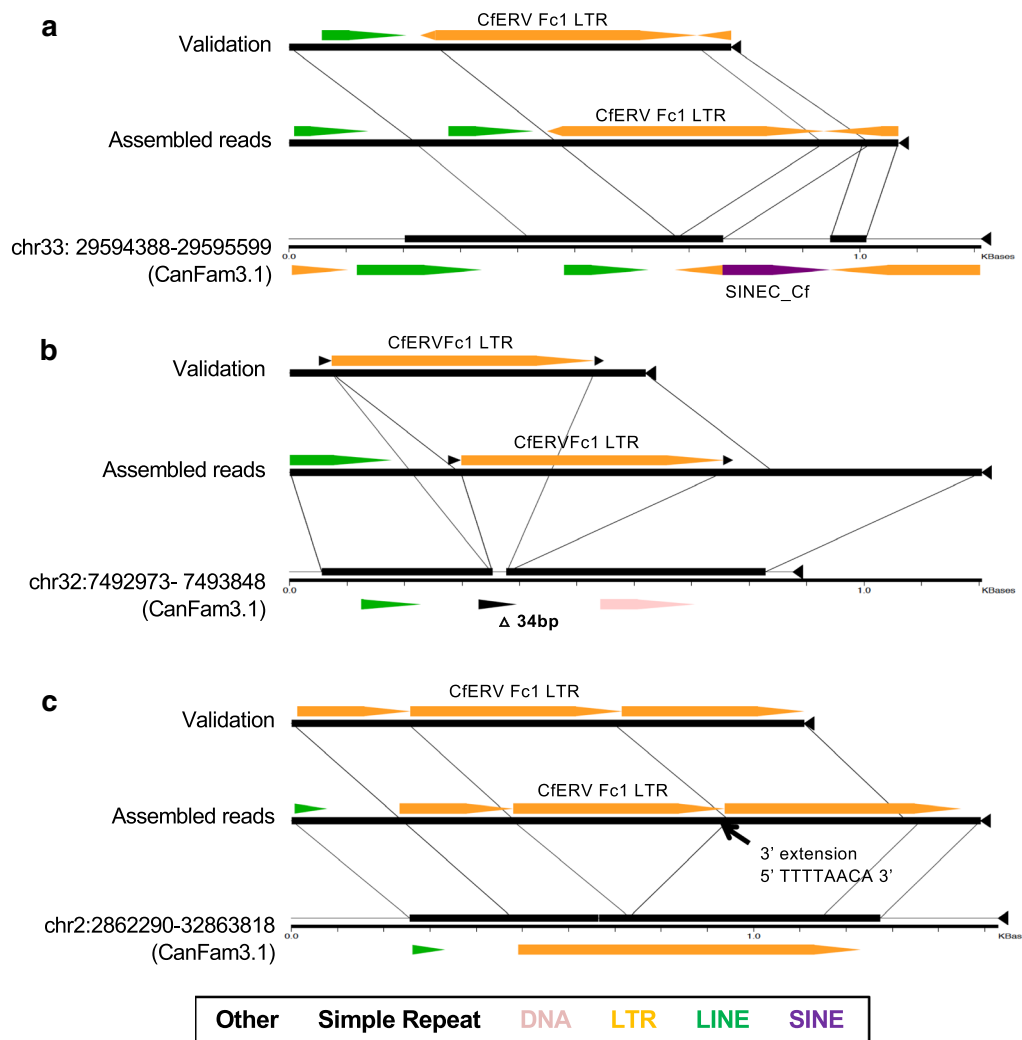


Fig. 4 Assessment of assembled non-reference alleles. LTR insertions associated with structural variation as captured in assembled Illumina read data. Local three-way alignments were generated for each assembled locus using the program Miropeats [92]. Each consisted of the LTR allele obtained by read 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 CfERV1 repeat

related group, per RepeatMasker) [11]. 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 (Additional file 4: 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 [41]). 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 (Additional file 5: 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.

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 [42, 43]. 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 [44], yielding formation times of individual proviruses from 20.49 mya to within 1.64 mya.

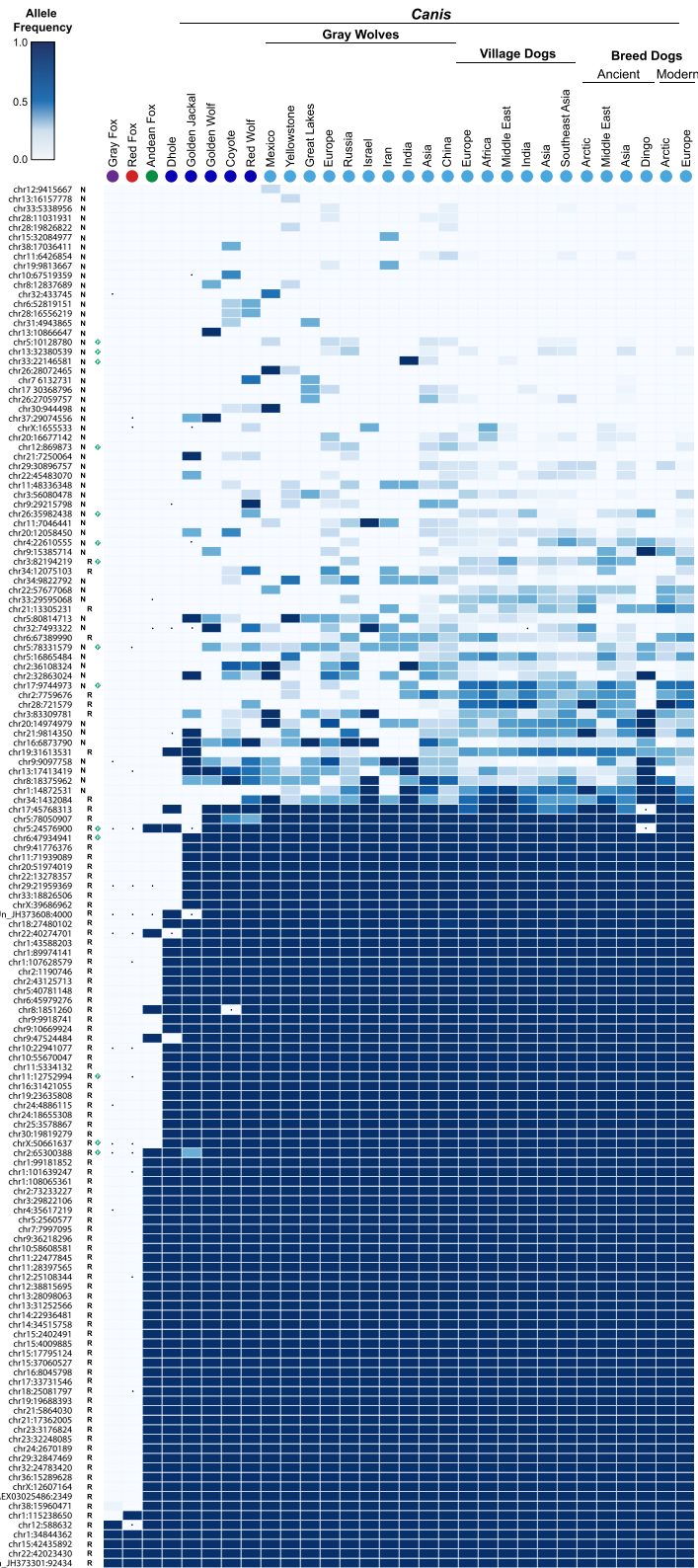
These estimates are sensitive to the assumed mutation rate, in addition to the limited number of differences expected between LTRs for the youngest loci. The youngest estimate (1.64 my) is driven by two proviruses whose LTRs differ by a single base change and five proviruses with identical 5' and 3' 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. A potential confounding factor is the presence of proviral loci within duplicated sequences, which are incorrectly represented as unique in the CanFam3.1 reference. Comparison with genomic copy number profiles from a diverse collection of 43 village dogs and 10 wolves shows that three proviral loci (chr3:219,396, chr5:7,8331,579, chr8:7,3924,489) are found in regions that have an expanded copy number [45]. Despite this

reference sequence duplication, TSDs and internal sequence of each provirus were unique. Overall, 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 littoralis*) and gray foxes (*U. cinereoargenteus*) (Fig. 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 (Additional file 6: 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. 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. To facilitate genotyping of the eight unresolved assemblies with linked 5' and 3' LTR junctions, we supplemented the Repbase CfERV1_LTR consensus sequence over the missing region (lower case in Additional file 3: Table S2). As has been discussed in earlier work [9], 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. Read-based genotypes show 87.5%(42/48) agreement with genotypes determined by PCR, with each of the six disagreements being cases where a heterozygous genotype which was incorrectly classified as homozygous reference, likely due to low read support.



(See figure on previous page.)

Fig. 5 Distribution of CfERV-Fc1(a) insertions in the genomes of modern canids. *In silico* genotyping was performed for 145 LTRs using Illumina read pairs across 347 sequenced canids representing extant members of all major *Canidae* lineages (Fig. 1). Sample names are indicated above by species or sub-population. Samples correspond to the Island and gray foxes ($n=8$), 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. 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. A green diamond is used to indicate loci with full-length alleles

Insertion allele frequencies ranged from 0.14% (inferred single insertion allele) to fixed across samples (Fig. 5; all raw data is included in Additional file 7: 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 Fig. 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. For 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 Island and gray foxes, representing the most distant splits of extant canids, had the lowest prevalence of occupied loci, with just five insertions each. However, this is not unexpected since insertions private to these lineages would not be ascertained in our discovery sample set.

The relative distribution of proviruses was in general agreement with dating via LTR divergence, though some inconsistencies were observed. No proviruses were detected in the fox outgroups (*Urocyon* and *Vulpes*) that have an estimated split time from other *Canidae* of >8 mya [35], but some were present in the Andean fox (chr2:65,300,388, chr5:24,576,900) and dhole (chrX:50,661,637, chr11:12,752,994). LTR divergence calculations using the inferred dog neutral substitution rate dated these insertions near 20.49, 14.80, 6.65, and 4.94 mya, respectively, suggesting the dating based on LTR divergence may be overestimated, as has been observed for other ERV groups [46, 47]. The youngest proviruses were variably present in *Canis* representatives. Of the most recent insertions, two (chr5:10,128,780, chr17:9,744,973) were present in both New and Old World wolves, implying integration prior to the geographic split of this lineage (1.10 mya) [48]. The remaining proviruses were present in Old World wolves and dogs only. Among these was the chr33:22,146,581 provirus that had an estimated date of formation of 6.58 mya by LTR comparison, consistent with skewed dating of the site. Altogether, the data are consistent

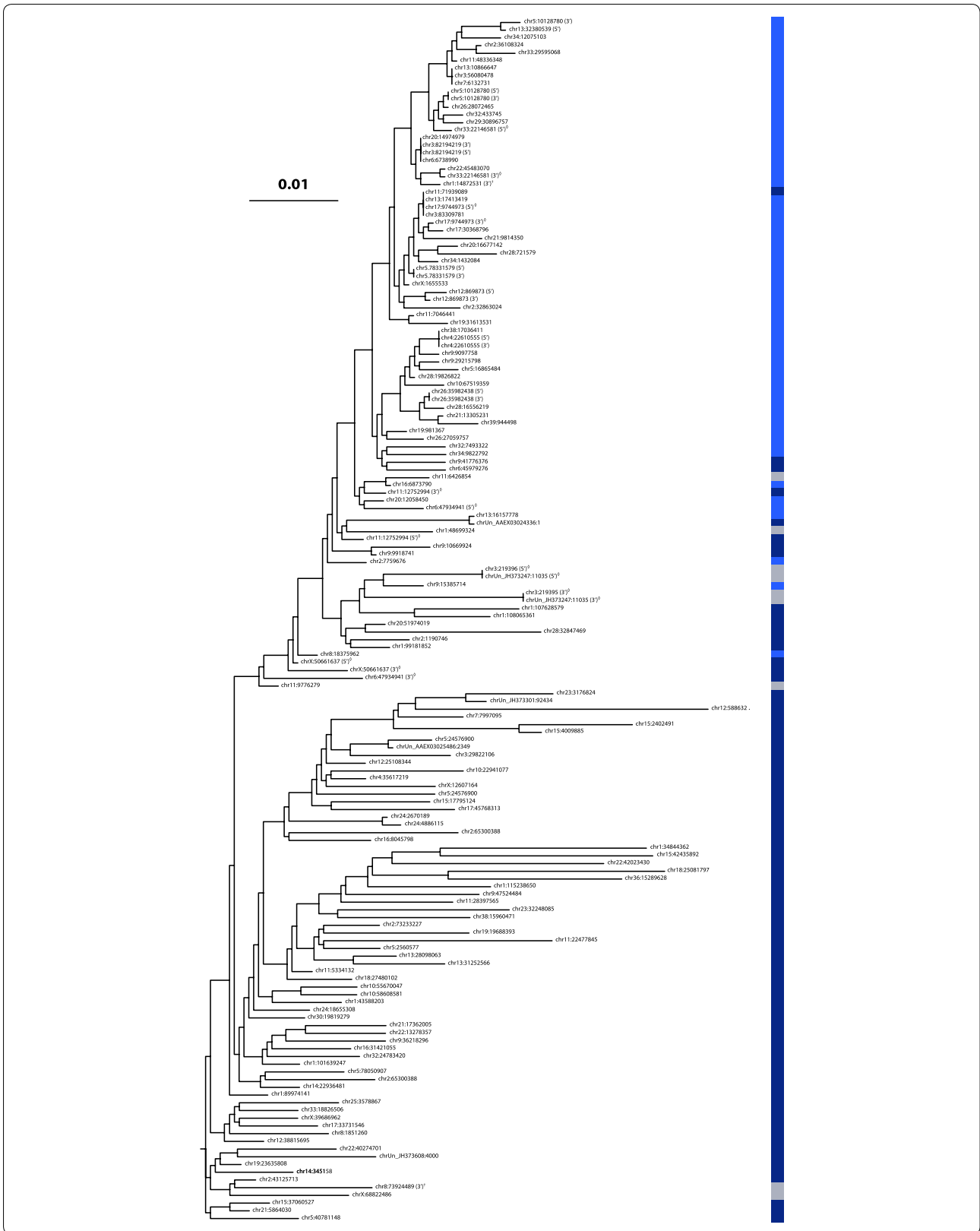
with CfERV-Fc1(a) endogenization in the ancestors of all modern canids followed by numerous invasions leading to a relatively recent burst of activity in the wolf and dog lineage of *Canis*.

Evolution of the CfERV-Fc1(a) lineage in *Canidae*

LTR sequences are useful in a phylogenetic analysis for exploring the evolutionary patterns of circulating variants prior to endogenization, as well as following integration within the host. To infer the evolutionary history leading to CfERV-Fc1(a) presence in modern canids, we constructed an LTR tree using as many loci as possible (from 19 proviral elements and 142 solo-LTRs) (Fig. 6; Additional file 8: Table S6).

In broadly comparing LTR placement to our inferred species presence (Fig. 6), the longer-branched clusters contained the few ancestral loci present in the outgroups (gray and red foxes) and those that were mostly fixed among the other surveyed species. However, at least two non-reference LTRs and other unfixed insertions were also in these clades, suggesting their more recent formation from related variants therein. One provirus was present within the most basal clade, and four (including the duplicated locus) were present within intermediate clades. We observed a major lineage (upper portion of tree) that included the majority of recent integrants. This lineage gave rise to the greatest number of polymorphic insertions, including a derived clade of insertions that appears to be *Canis*-specific, with some sites restricted to one or two sub-populations. This lineage also contains the majority of proviral LTRs (15 of 19 included in the analysis), most possessing intact *pol* and/or *env* genes. The youngest proviral integrants, as inferred from high LTR identities and prevalence among sampled genomes, tend to be on short branches within derived clusters that contain the majority of unfixed loci, likely reflecting their source from a relatively recent burst of activity in *Canis* ancestors.

Within the germline, the highest occurrence of recombination resulting in a solo LTR takes place between identical LTRs [49, 50], implying the LTR sequence itself is preserved in the solo form. Under this



(See figure on previous page.)

Fig. 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 indicate local support values > 70%. Chromosomal positions are relative to CanFam3.1 coordinates. A color bar is shown at the right to denote element presence as fixed among *Canis* (dark blue), insertionally polymorphic (light blue), or not genotyped (gray). LTRs belonging to proviruses are indicated along with the chromosomal position with a (5') or (3') as appropriate. Clusters of identical LTR haplotypes are indicated with a vertical dashed line. Mismatched proviral LTRs are indicated by a diamond. LTRs from proviruses lacking cognate LTR pairs (*i.e.*, due to truncation of the element) are indicated with a cross. The scale bar shown represents the evolutionary distance in substitutions per site

assumption, the presence of identical solo LTR haplotypes should imply a common ancestral source. We identified four such LTR haplotypes within the *Canis*-specific clades, including loci in co-clusters with one of two proviruses (chr3:82,194,219 and chr4:22,610,555), therefore bounding the inferred age of these insertions to within the last 1.64 mya (dashed lines in Fig. 6). Between the four identical clusters, the LTR haplotypes shared nucleotide identity ranging from 99.3% (three substitutions from a consensus of the four clusters) to 99.7% (one substitution), suggesting their origin from related variants over a common timeframe. We modified our dating method to obtain an estimated time of formation across each cluster by considering the total concatenated LTR length per cluster, as has been similarly employed elsewhere [5]. This approach placed tentative formation times of the youngest insertions from a common variant 547,220 years ago (no change over 1374 bp, or 3 LTRs) and 410,415 years ago (no change over 1832 bp, or 4 LTRs). Comparison to the inferred prevalence of each cluster indicates the most recent of these insertions arose in Old World wolves, consistent with this timeframe.

Since proviral LTRs begin as an identical pair, aberrant placement in a tree and/or the presence of mismatched TSDs implies post-insertion conversion or rearrangement at the locus [51]. LTRs from the youngest proviruses tended to pair on sister branches. An exception includes the LTRs of the chr33:22,146,581 provirus, whose mispairing is consistent with conversion of at least one of its LTRs, possibly from the chr1:48,699,324 provirus or a similar variant (see above). There were six instances of aberrant LTR placement for the remaining eight CfERV-Fc1(a) proviruses that had both LTRs present (labeled in Fig. 6), suggesting putative post-insertion conversion and contributing to inflated age estimates based on LTR divergence. The TSD repeats of individual proviruses had matched 5 bp repeats in all cases, suggesting none of the elements have seeded inter-element chromosomal rearrangements. With exception of three instances of reference solo LTRs that each had a base change between its flanking repeats, the TSDs for all other solo LTRs were also intact.

CfERV-Fc1(a) structure and biology

Characterization of the inferred CfERV-Fc1(a) ancestor

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 Additional file 9: Figure S3 and summarized as follows.

Consistent with the analysis of Caniform ERV-Fc1 consensus proviruses [10], 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 DDX₃₅E protease resistant core and N-terminal HHCC DNA binding motif). An *env* reading frame (absent from the Repbase CfERV-Fc1 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 CX₆CC (TM) motifs involved in SU-TM interactions, and a putative RD114-and-D-type (RDR) receptor binding motif [52]. A hydrophobicity plot generated for the translated sequence identified segments for a predicted fusion peptide, membrane-anchoring TM region, and immunosuppressive domain (ISD) [53]. Putative major splice donor (base 576 within the 5'UTR; 0.67 confidence) and acceptor sites (base 5216 within *pol*; 0.85 confidence) were identified that would be predicted for the generation of *env* mRNA (see Additional file 9: Figure S3). The CfERV-Fc1(a)_{CON}

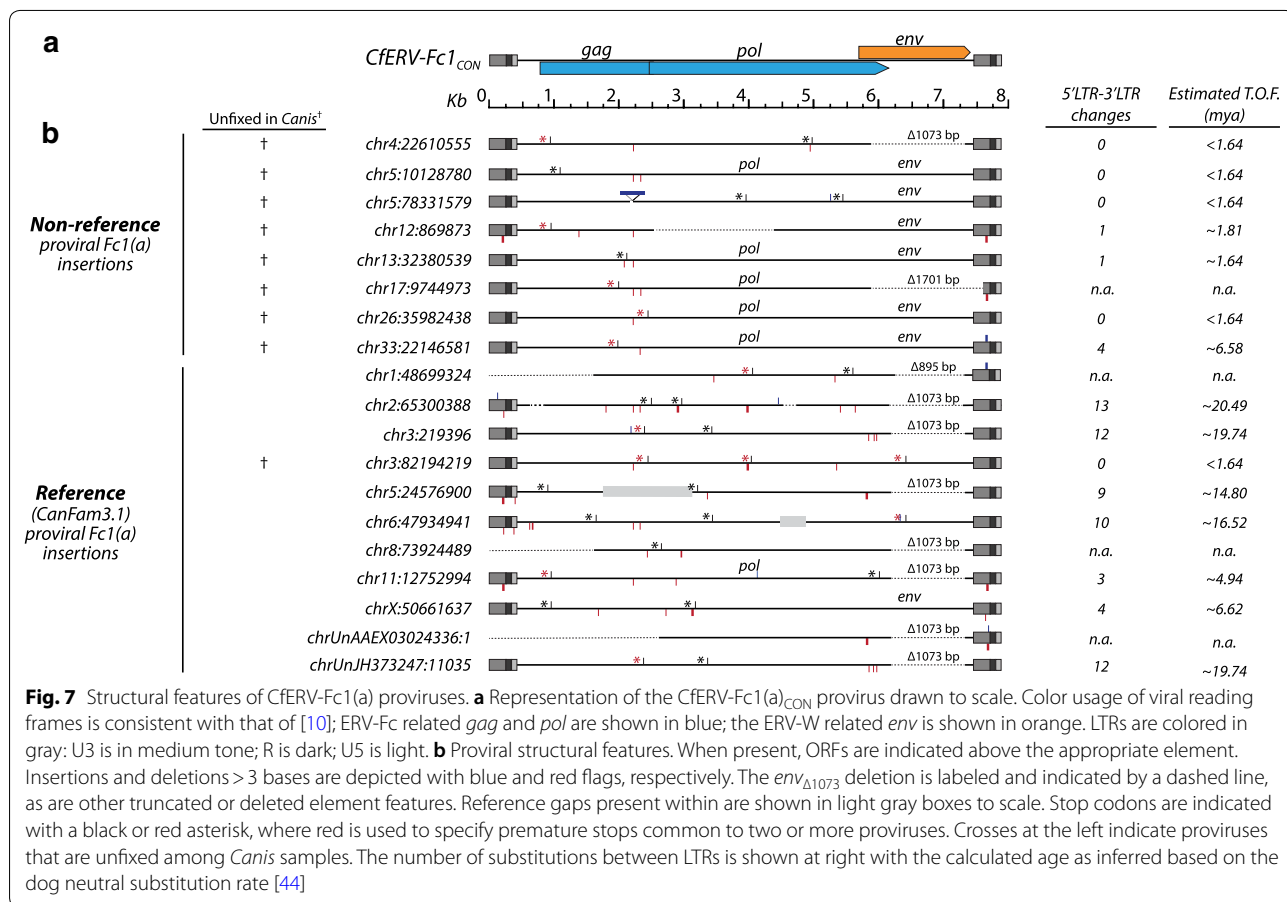
element possessed identical LTRs, a tRNA^{Phe} binding site for priming reverse transcription (GAA anticodon; bases 464 to 480), and the canonical 5'-TG...CA-3' terminal sequences required for integration [1].

Properties of individual CfERV-Fc1(a) proviruses

We assessed the properties of individual full-length elements for signatures of putative function (Fig. 7). With the exception of the *gag* gene, we identified intact ORFs in several reference copies and most of our non-reference sequenced proviruses. A reading frame for the *pol* gene was present in six proviruses; of these, all contained apparent RT, RNaseH, and integrase domains without any changes that would obviously be alter function. Likewise, an *env* ORF was present among seven proviruses, of which all but one contained the above mentioned functional domains (the SU-TM cleavage site is disrupted in the chr5:10,128,780 provirus: RRKA). Comparison of the rate of nonsynonymous (d_N) to synonymous (d_S) nucleotide substitutions for the seven intact *env* reading frames revealed an average d_N/d_S ratio of 0.525, indicating moderate purifying selection ($p=0.02$, Nei-Gojobori method). The hydrophobicity plot of each *env* ORF was

in agreement with that of the CfERVFc(a)1_{CON} provirus, with predicted segments for a fusion peptide, TM region, and ISD. Comparison to the *pol* and *env* translated products that would be predicted from the CfERVFc1(a)_{CON} inferred the individual proviruses shared 98.4% to 99.3% (Pol) and 98% to 99.6% (Env) amino acid identity, respectively, and each was distinct from the inferred consensus.

No complete *gag* reading frame was observed. Particularly when compared to *pol* and *env*, the *gag* gene had incurred a number of inactivating mutations, including shared frameshifts leading to premature stops. The longest *gag* reading frames (chr3:82,194,219 and chr26:35,982,438) both possessed a premature stop within the first zinc finger domain of the nucleocapsid. The only obvious gene inactivation in the latter provirus was the terminal frameshift in *gag*, a domain with roles in the encapsidation of viral genomic RNAs [54]. Thus, absence of both zinc finger domains and the N-terminal myristoylation site should interfere with canonical Gag functions, regardless of the presence of intact matrix and capsid domains. Excluding the frameshift leading to the abortive stop in those proviruses, the translated Gag would have respectively shared 97.8% and 98% amino



acid identity to the CfERV-Fc1(a)_{CON} Gag. Though none of the identified CfERV-Fc1(a) proviruses have retained complete reading frames for all genes, this finding does not exclude the possibility that rare intact proviruses remain to be identified, or that a putative infectious variant could be generated via recombination of co-packaged RNAs.

The majority of the CfERV-Fc1(a) proviruses could be assigned to one of two proposed subgroups based on the presence of a common deletion within the *env* gene (Fig. 7). The deletion spans a 1073 bp region of *env* (referred to here as *env*_{Δ1073}), removing the internal majority portions of SU and TM (see Additional file 9: Figure S3; including the putative receptor binding domain, motifs involved in SU-TM interactions, and transmembrane domain). Eight proviruses possessed the *env*_{Δ1073} deletion, including the duplicated locus. The prevalence of the *env*_{Δ1073} deletion was skewed toward proviruses that harbored multiple inactivating mutations, while only one possessed a retained ORF (chr11:12,752,994, *pol*), and proviruses with the *env*_{Δ1073} deletion had a greater number of LTR-LTR differences (mean of 8.17 vs 2.22, $p=0.022$ one sided *t* test), consistent with the older status of most of these loci. Additionally, the *env*_{Δ1073} 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 Fig. 5), suggesting its formation early in CfERV-Fc1(a) evolution (at least 8.7 mya; Fig. 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; Fig. 6). The *env*_{Δ1073} 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 [55, 56]. 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. We propose the *env*_{Δ1073} allele spread via template-switching of co-packaged *env*_{Δ1073} RNAs. Any of the above scenarios would result in the spread of an otherwise defective *env* gene. In contrast, all but two (chr4:22,610,555, chr33:22,146,581) of the most recently integrated proviruses contained an uninterrupted *env* reading frame. In addition to the *env*_{Δ1073} deletion, unique *env* deletions were present in two other elements; a 1702 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*_{Δ1073} 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). 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 [1]. Molecular signatures of *trans* complementation may be interpreted from the presence of inherited changes among multiple elements, particularly ones that would render a provirus defective [57, 58].

We observed evidence for the mobilization of CfERV-Fc1(a) copies via complementation. For example, examination of the proviral gene regions revealed inherited frameshift-causing indels and common premature stops that were variably present among the majority of elements (a total of 12 of the 19 proviruses; see Fig. 7). At least three distinct frameshifts leading to a stop within *gag* were shared over several elements (from the Fc1(a)_{CON} start, bp 882: chr4:22,610,555, chr11:12,752,994, chr12:869,873; bp 1911: chr17:9,744,973, chr33:22,146,581; bp 2203: chr3:82,194,219, chr26:35,982,438, and the duplicated chr3:219,396 and chrUn_JH373247:11,035 insertions). Proviruses also shared unique deletions leading to abortive stops within *pol* (near Fc1(a)_{CON} bp 3988: chr1:48,699,324, and chr3:82,194,219). In addition to the common *env*_{Δ1073} frameshift deletion, putative in-frame *pol* deletions were also present (Fc1(a)_{CON} bp 5263 Δ3 bp: chr3:82,194,219; chrUn_AAEX03024336:1; bp 5705 Δ27 bp: chr5:24,576,900, chrUn_AAEX03024336:1). Two proviruses contained a shared stop within *env* (Fc1(a)_{CON} bp 6240: chr3:82,194,219, chr6:47,934,941). The provirus on chromosome 3 possessed a total of four of the above changes differentially shared with other proviruses in *gag*, *pol*, and *env*; these were the only defective changes present within the element. While successive conversion events of the provirus from existing loci cannot be ruled out, this provirus appears to be a comparatively young element (only found in Old World wolves and dogs), which more likely suggests formation of the element via multiple intermediate variants. No other provirus contained multiple common indels.

We did not find evidence for expansion of the lineage via retrotransposition in *cis*, during which new insertions are generated in an intracellular process akin to the

retrotransposition of long interspersed elements [59]. Such post-insertion expansion is typically accompanied by a loss of the viral *env* gene, particularly within recently mobilized insertions (as interpreted, for example, by the derived phylogenetic placement), whereas *gag* and *pol* are retained. Our data suggest this scenario is unlikely given the absence of a functional *gag* gene and presence of a conserved *env* ORF in several elements, particularly young ones. In this regard, *cis* retrotransposition tends to facilitate rapid *env*-less copy expansion and therefore tends to occur among derived copies of a given lineage [60], and our data suggest the opposite regarding older (loss of *env*) and younger (*env* present) CfERV-Fc1(a) proviruses.

Discussion

Mammalian genomes are littered with the remnants of retroviruses, the vast majority of which are fixed among species and present as obviously defective copies [18, 39]. However, the genomes of several species harbor ERVs whose lineages contain relatively intact loci and are sometimes polymorphic, despite millions of years since integration [18, 39]. Such ERVs have the potential to express proviral-derived products or to alter the expression of host encoded genes, especially for intact ERVs or insertions near host genes. In particular, ERV expression from relatively recent integrants has been linked to disease (reviewed in [39, 61]). However, there is also growing evidence that many fixed loci have been functionally co-opted by the host and play a role in host gene regulation (reviewed in [62]). Illustrating both bursts of activity and putative extinction, our findings present a comprehensive assessment of the evolutionary history of a single retroviral lineage through the genomic surveys of nine globally distributed canid species, some represented by multiple subpopulations.

Relative to other animal models, ERV-host relationships within the dog have been understudied. Until now, reports of canine ERVs have been from analysis of a single genome assembly or limited screening of reference loci [11, 63, 64]. To further investigate a subset of apparent recent germline integrants [11] we surveyed the level of polymorphism and possible mechanisms of spread of

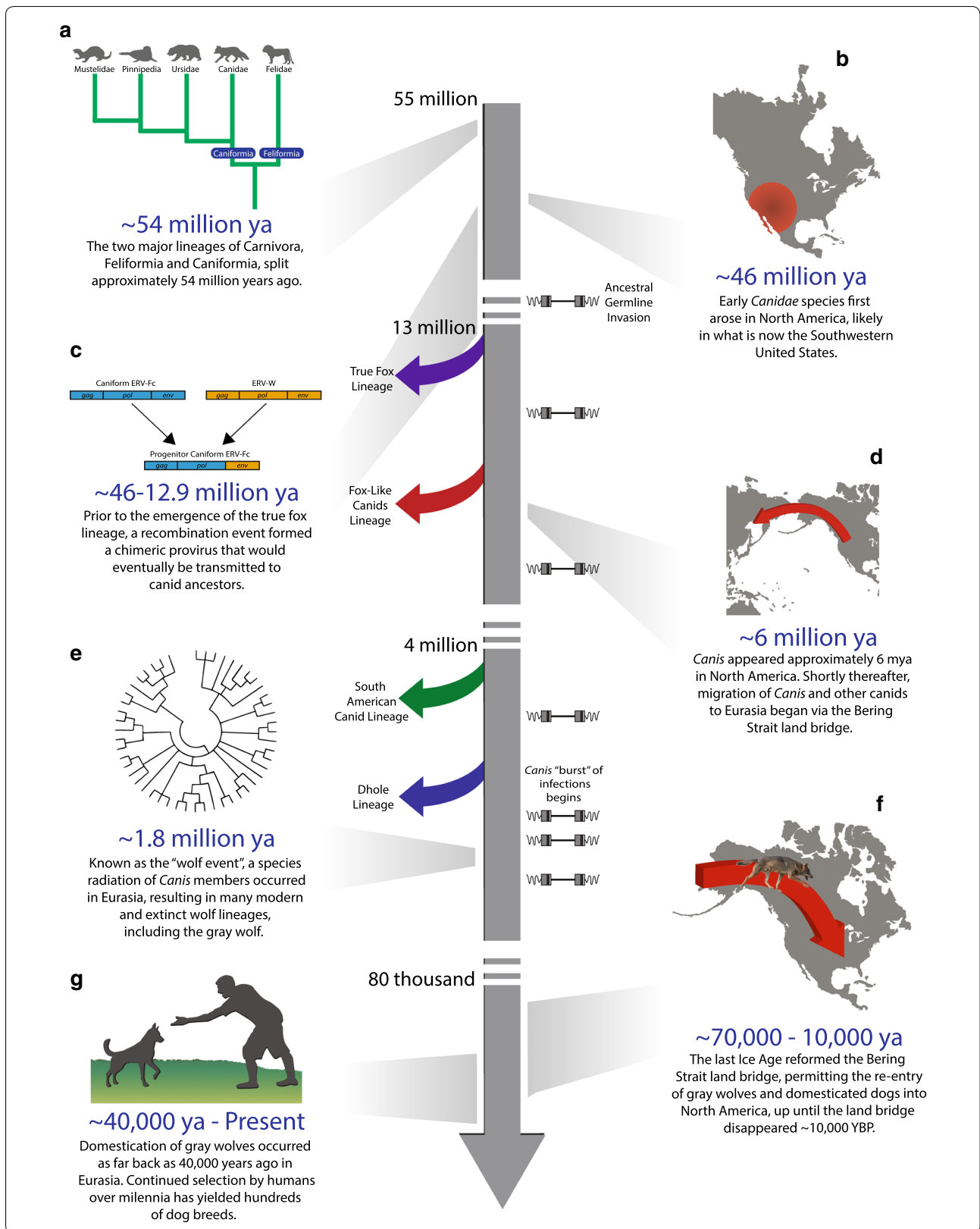
the γ -like ERV-Fc1(a) lineage across a diverse set of canid species. Our exhaustive analysis of CfERV-Fc1(a) loci is the first population-level characterization of a recently active ERV group in canids. We uncovered and genotyped numerous polymorphic sites, which include insertions missing from the dog reference genome assembly that contain ORFs, display high LTR identities, and have derived placements within a representative phylogeny, which are all characteristics of relatively young elements.

Although permutations indicated that CfERV-Fc1(a) insertions are significantly depleted within and near genes (Additional file 5: Figure S2), insertions were located with dog gene models, which raises the possibility of biological effects. For example, two intronic LTRs were fixed in all canids: one within *AIG1*, a transmembrane hydrolase involved in lipid metabolism [65]; the other in the diffuse panbronchiolitis region *DPCR1* of the dog major histocompatibility complex 1 [66]. Other intronic insertions were fixed in samples following the splits of the true and dog-like foxes. These included genes with homologs involved in tumor suppression (*OPCML*), cell growth regulation (*CDKL3*), DNA repair (*FANCL*), and innate immunity (*TMED7-TICAM2*). An exonic *Canis*-specific solo LTR was located at chr1:107,628,579 within the 3' UTR of *BCAT2*, an essential gene in metabolizing mitochondrial branched-chain amino acids. In humans, altered expression of *BCAT2* is implicated in tumor growth and nucleotide biosynthesis in some forms of pancreatic cancer [67–69]. The same LTR is situated ~550 bp upstream of *FUT2*, a fucosyltransferase involved ABH blood group antigen biosynthesis in mucosal secretions [70, 71]. *FUT2* variants affect secretion status and have been implicated in intestinal microbiota composition [72], viral resistance [73], and slowed progression of HIV [74]. Though connections between LTR presence and physiology are yet to be determined, these findings will inform future investigations into the potential effect of CfERVs on host biology.

CfERV-Fc1(a) integrants endogenized canid ancestors over a period of several millions of years (Fig. 8b–e). This activity included bouts of infectious activity/mobilization inferred from the last 20.4 my to within 1.6 mya, the latter of which are only present in *Canis* sub-populations.

(See figure on next page.)

Fig. 8 History of CfERV-Fc1(a) germline invasion in the Canidae. A timeline of major events in canid or CfERV-Fc1(a) evolutionary history relative to estimated insertion events. At the approximate time point, branching events of the major canid lineages are indicated by arrows along the timeline with colors matching Fig. 1. Indicated by proviruses to the right of the timeline are estimated insertion times based on genotyping data from Fig. 5. **a** Based on its presence in all canids, the recombination event that formed the provirus (**b**), which infected canid ancestors occurred sometime between the split of the major Caniform lineages (**a**) and the origins of canids 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 [35], **d, e** by [113], **f** by [114], and **g** by [44]



The mutation rate we used to obtain these estimated timeframes (1.33×10^{-9} changes per site per year [44]) coincides with those from two other ancient genome analyses, which utilized ancient DNA to calibrate wolf and dog mutation rates [75, 76]. However, our rate is substantially slower than those used previously to date reference CfERV-Fc1(a) members including 2.2×10^{-9} (as an “average” mammalian neutral substitution rate) [11] and the faster rate of 4.5×10^{-9} (as has been reported for the mouse) [10]. Applying those substitution rates to our data would infer much younger integration times of 11.85 mya to <0.91 mya and 6.1 mya to <0.48 mya, respectively. We note the precision in ERV-Fc1(a) age estimations using this method is subject to the accuracy of the inferred background mutation rate, but may also be skewed by other factors. For example, 12 of the 69 LTR-LTR base changes occur at CpG sites. Methylation may make these positions hypermutable, and contribute to an over-estimated age. Other possibilities, such as post-insertion sequence exchange between LTRs, also cannot be conclusively ruled out. Therefore, we interpret our estimations as broad formation times only.

Due to their complete absence of LTR divergence, the youngest CfERV-Fc1(a) ages are bounded to the estimate of 1.64 my, using the dog substitution rate. We employed an alternative approach that makes use of LTRs that shared haplotypes [5] to narrow the age estimations to ~547,220 and 410,415 years, again, as inferred from the time estimated to accrue one mutation across multiple identical LTRs (respectively across three and four LTRs per haplotype). For comparison, applying the average mammalian and mouse substitution rates to the same data would place either event respectively at 303,251 and 161,734 years ago (no change over three LTRs) and 227,438 and 121,300 years ago (no change over four LTRs). Both estimates are consistent with CfERV-Fc1(a) circulation after the estimated emergence of the gray wolf species 1.1 mya and pre-dating the split of the New and Old World gray wolves [48] (Fig. 8f). The branching patterns observed within our LTR phylogeny are consistent with these findings, implying bursts of replication from closely related variants now recorded in clusters of LTR haplotypes. In this regard, our findings suggest bouts of infection from multiple circulating viruses over a relatively short evolutionary time period.

CfERV-Fc1(a) activity coincided with major speciation events in canine evolution (Fig. 8b–e). Taking into consideration the above approaches for age estimations, we refined the dating of endogenization events by integrating inferred ages with that of orthologous presence/absence patterns across numerous canid lineages, many of which are recently diverged clades. The analysis served two purposes. First, we made use of the tenet that ERV

integration is permanent and the likelihood of two independent integration events at the same locus is negligible. In this way, the presence of an ERV insertion that is shared between individuals or species supports its origin in a common ancestor. Therefore, integration prior to or following the split of two or more species is supported by virtue of insertion presence/absence of occupied loci across those species. Second, the analysis allowed us to infer insertion genotypes across highly diverse canid representatives, thus providing the means to gauge the collective patterns of individual CfERV-Fc1(a) loci among contemporary animals to infer putative sub-population or species-specific integrants.

Comparisons of the approximate insertion dates discussed above in combination with estimated species split times would place the earliest CfERV-Fc1(a) germline invasions prior to or near the estimated divergence of the *Canidae* from now extinct ancestors (14.15 mya) [35], followed by invasions after the split of the true fox (12.9 mya) [35] and fox-like canid lineages (8.7 mya) [36]. Subsequent insertions also occurred prior to the split of the South American canid and wolf lineages (3.97 mya) [36]. According to this timeframe, and consistent with the detection of some young proviral insertions private to gray wolves and dogs alone (Fig. 5), the most recent invasions would have occurred around the time of the branching event that gave rise to gray wolves (1.10 mya) [36]. Based on the lack of observed dog-specific loci, our data suggests that CfERV-Fc1(a) replication ceased in wolf ancestors prior to domestication, which is estimated to have begun around 40,000 years ago [44] (Fig. 8g), but does not rule out continued activity. Analysis of additional genomes, particularly from gray wolves, should clarify the presence of such variants in future analysis.

CfERV-Fc1(a) activity included the spread of defective recombinants. Our comparative analysis of nucleotide differences shared among the proviruses supports a scenario in which CfERV-Fc1(a) members proliferated in canine ancestors via complementation. Patterns of discreet, shared changes among distinct elements in all viral genes were observed (*i.e.*, premature stops and common base changes, indels, in addition to the *env*_{Δ1073} segment; Fig. 7), consistent with the spread of mutations present from existing Fc1(a) copies, probably via co-packaging of the defective viral genomes. Of the 19 proviruses analyzed in full, the majority displayed shared discreet stops or the *env*_{Δ1073} deletion, in addition to in-frame indels. This pattern is consistent with the hypothesis that degradation of ERV genomes, particularly involving the loss of *env*, offers an evolutionary benefit to the host by preventing the potential horizontal spread of infectious viruses between individuals, as has been suggested [60, 77]. Similar patterns of recurrent *env* deletions have

also been described in the majority of HERV-W copies in humans [47]. The presence of intact *env* genes, and sequence signatures of selective pressure retained within those *env* reading frames, suggests involvement of Fc1(a) *env* leading to the putative formation of recombinant proviruses, rather than having been intracellularly retrotransposed (in *cis*) that would not require a functional Env. Altogether such patterns of reinfection may have predominantly occurred within a given individual, as none of these mechanisms explicitly requires (but does not rule out) spread to other individuals within the population; indeed concurrent reinfection of a single individual may also lead to unique proviruses later transmitted to offspring [78]. Several retroviruses, including HIV, have been shown to be capable of co-packaging RNA from other retroviruses, even ones with low sequence homology [54]. These findings suggest complementation was a predominant form of proliferation for the observed CfERV-Fc1(a) loci. In theory, a functional provirus could arise in a spontaneous recombinant, raising the possibility of bursts of amplification to come. Indeed, all viral genes in our consensus appear to be intact, illustrative that few changes would be required to generate a putatively infectious virus.

Patterns of shared sequence changes, such as premature stops and in-frame shifts, indicate that the oldest inherited change involved an in-frame shift in the *pol* gene (from the Fc1(a)_{CON} start, bp 5705 Δ 27 bp). Aside from the *env* _{Δ 1073} deletion, all other common changes were present in the lineage that led to the majority of young insertions (Fig. 6). Among the earliest inferred changes were premature stops in *gag* (CfERV-Fc1(a)_{CON} bp 882 and 2203, respectively) and *env* (CfERV-Fc1(a)_{CON} bp 6240), typically in elements within a *Canis*-specific subclade. Another inherited mutation is shared by the chr17:9,744,973 and chr33:22,146,581 proviruses as a third distinct stop in *gag*. LTR dating is limited, however based on its restriction to *Canis* members it likely originated within the last 2.74 my [36]. 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 source loci.

The apparent absence of any infectious retrovirus among canines is peculiar, particularly as individuals are likely to be challenged from viruses infecting prey species. Among mammals, the evolution and history of ERV-Fc included the generation of multiple recombinants and spread by cross-species transmission including to carnivores. Reflected in the ERV fossil record of the domestic dog genome is an expansion of the relatively young ERV-Fc1 that was generated from recombination with

the *env* of a distinct lineage closely related to ERV-W. The resulting virus would likely have altered pathogenic properties, particularly given the presence of a 'new' *env* in the chimera. Possibly, it was the acquisition of this *env* that allowed the virus to access and subsequently expand within the canid as a host.

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 [39]). 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 that have replicative activity, as well as proteins derived from related ERV members, are capable of driving aberrant cellular proliferation, tumorigenesis, and inciting immune responses [39]. It is well-known that canine cell lines are permissive for replication of retroviruses that infect other host species including human [79], a property possibly reflecting the loss of the antiviral factor TRIM5 α in canines [80]. While there have been reports of retroviral activities and particles displaying characteristic γ -like features in canine leukemias and lymphomas [26–32], those findings have not been substantiated. A recent report confirmed transcriptional activity from at least one γ -like CfERV group [non-Fc1(a)] in canine tissues and cell lines [64]. We have also preliminarily demonstrated expression of CfERV-Fc1(a) proviruses in canine tissues and tumor-derived cell lines (Jarosz and Halo, unpublished data). 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.

Conclusions

We identified, characterized, and genotypes numerous polymorphic CfERV-Fc1(a) insertions, including several absent from the canine reference genome. The discovered elements include proviruses that contain open reading frames and that have high-LTR identities, suggesting that they are relatively young insertions. Using these proviral sequences, we characterized a new CfERV-Fc1(a) consensus which includes an intact Env gene. The presence of disruptive mutations shared among elements indicates that ERV-Fc spread by *trans* complementation of defective proviruses. Comparison across related species indicates that multiple circulating variants that infected canid ancestors over the past 20 million years.

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 [36, 44, 45, 48, 81–84] (Additional file 1: 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. Single nucleotide variant (SNV) genotypes of sequenced samples were determined using Genome Analysis Toolkit (GATK) version 3.7 [85]. Information corresponding to all samples and sources of raw data is detailed in Additional file 1: Table S1.

Identification of annotated CfERV1 reference insertions

The dog ERV-Fc1(a) lineage is classified in Repbase as 'CfERV1' derived (Repbase update 10.08) [86]. We therefore mined the CanFam3.1 RepeatMasker output for elements classified as 'CfERV1_LTR' and 'CfERV1-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 ± 1 kb of each locus was extracted and used in BLAT searches against the other available carnivoran reference assemblies corresponding to ferret (*MusPut-Fur1.0*) [87], panda (*BGI_Shenzhen1.0*) [88], and cat (*Felis_catus_8.0*) [89] 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 Ref. [10]. 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) [37], which processed BAM alignment files from samples indicated in Additional file 1: 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 with 'CfERV1' reference coordinates based on RepeatMasker annotations of CanFam3.1. Only deletion

calls corresponding to sizes of a solo LTR (400–500 bp) or a full-length provirus (7–9 kb) were considered for further analysis.

Identification of non-reference of CfERV-Fc1(a) insertions

LTR-genome junctions corresponding to non-reference variants were assembled from supporting Illumina reads [9, 38], with modifications as follows. The chromosomal positions of candidate non-reference ERVs were first identified using the program RetroSeq [90]. Individual BAM files were queried using RetroSeq discovery to identify ERV-supporting discordant read pairs with one read aligned to the sequences corresponding to 'CfERV1' and 'CfERV1_LTR' from RepBase [86]. Individual BAM files were merged for subsequent steps using GATK as described [9]. RetroSeq call was run on the merged BAM files requiring ≥ 2 supporting read pairs for a call and output calls of levels 6, 7, and 8 further assessed, resulting in 2381 candidate insertions. Output calls within ± 500 bp of an annotated CfERV from the above queried classes were excluded to eliminate false calls of known loci. ERV-supporting read pairs and split reads within a 200 bp window of the call breakpoint were subjected to *de novo* assembly using the program CAP3 [91]. Output contigs were filtered to identify ERV-genome junctions requiring ≥ 30 bp of assembled LTR-derived and genomic sequence in the form of (i) one LTR-genome junction, (ii) linked assemblies of 5' and 3' LTR junctions, or (ii) a fully resolved LTR (~ 457 bp) with clear breakpoints that mapped to CanFam3.1. Contigs that contained putative CfERV junctions were then aligned back to the reference to precisely map the insertion position of each call. Assembly comparisons were visualized using the program Miropeats [92].

Validations and allele screening

For validating non-reference calls, primers were designed to flank the predicted insertion within ~ 200 bp based on the breakpoint position for a given site. Genomic DNA from a subset of samples with predicted insertion variants was used for validations. DNA with limited material was subjected to whole genome amplification (WGA) from ~ 10 ng genomic DNA according to the manufacturer's protocol (Repli-G, Qiagen). For each sample, WGA DNA was diluted 1:20 in nuclease free water and 1 μ L was utilized per PCR reaction. Two PCR reactions were run for each site in standard conditions using Taq polymerase (Invitrogen): one reaction utilized primers flanking each candidate call to detect the empty or solo LTR alleles; the second was to detect the presence of a proviral junction, utilizing the appropriate flanking primer paired with a primer within the CfERV-Fc1(a) proviral 5'UTR (near base ~ 506 from the start of the

Rebase F1 consensus element). Sanger sequencing was performed on at least one positive sample. When detected, provirus insertions were amplified in overlapping fragments from a single sample in a Picomaxx reaction per the manufacturer's instructions (Stratagene) and sequenced to $\geq 4 \times$ across the full element. A consensus was then constructed for each insertion based on the Sanger reads obtained from each site. The sequence of the chr5:78,331,579 provirus could not be fully resolved using Sanger reads and was completed using PCR-free PacBio sequencing reads obtained from Zoey, a Great Dane breed dog. All sequences corresponding to non-reference solo-LTR insertions and all sequenced proviral elements have been made available in Additional file 3: Table S2 and proviral sequences have been deposited in GenBank under accessions MK039120-MK039127.

Genomic distribution

The positions of the reference and non-reference insertions were intersected with Ensembl dog gene models (Release 81; ftp.ensembl.org/pub/release-81/gtf/canis_familiaris/). Intersections were performed using bedtools [93] with window sizes of 0, 5, 10, 25, 50, and 100 kb. To assess significant enrichment of insertions relative to genic regions, we performed one thousand permutations of randomly shuffled insertion positions, intersected the new positions with genes, and calculated the number of insertions intersecting genes within the varying window sizes as above. p values were calculated as the number of permuted insertion sets out of one thousand that intersected with less than or equal to the number of genes observed in the true insertion set.

Dating of individual proviruses

A molecular clock analysis based on LTR divergence was used to estimate times of insertion [9, 10, 42]. For 7 non-reference and 8 reference proviruses that had 5' and 3' LTRs present, the nucleotide differences between those LTRs was calculated, treating gaps > 2 bp as single changes. The total number of changes was then divided by the LTR length (e.g. 457 bp), and the percent divergence normalized to the inferred canine background mutation rate of 1.3×10^{-9} changes per site per year [44] to obtain age estimations in millions of years for individual insertions. The provirus at chr17:97,449,73 was excluded from the analysis due to truncation of its 3' LTR. We extended LTR dating to estimate times of formation for identical LTR groups that included solo LTRs using a modification of the above approach as described elsewhere [5]. Briefly, the total length in bp of the LTRs making up each cluster was collectively added and the age estimate obtained by the percent divergence for a single base pair to have been introduced along the total

length utilizing the same mutation rate of 1.3×10^{-9} changes per site per year.

In silico genotyping

We genotyped 145 insertions (89 reference and 56 non-reference insertions) utilizing whole genome Illumina reads and reconstructed alleles corresponding to the empty and occupied sites. Genotyping was performed on 332 individuals including the 101 samples utilized for discoveries of polymorphic variants [36, 44, 48, 81–84, 94–103] (Additional file 6: Table S4). Reference insertions were deemed to be suitable for genotyping based on manual assessment for the presence of paired TSDs and uninterrupted flanking sequence. Sites associated with duplication events were identified by comparison of flanking regions and TSD presence, and insertions within encompassing duplication (proviruses at chr3:219,396 and chrUn_JH373247:11,035), or situated within duplicated pre-insertion segments (chrUn_AAEX03025486:2349) were excluded, as were sites with single assembled junctions (chr13:20,887,612; chr27:44,066,943; Additional file 3: Table S2). The sequences from validated and completely assembled LTRs were utilized for allele reconstruction of non-reference sites. For example, the validated sequences for the non-reference solo LTRs at chr2:32,863,024 (8 bp LTR extension) and chr32:7,493,322 (associated with deletion of reference sequence) were included for genotyping of alternate alleles. For sites with linked, but non-resolved, 5' and 3' assembled junctions (i.e., missing internal sequence), we substituted the internal portion of each element from the Rebase CfERV1 consensus (see Additional file 3: Table S2), and used the inferred sequence for allele reconstruction. Insertion and pre-insertion alleles were then recreated based on ± 600 bp flanking each insertion point relative to the CanFam3.1 reference, accounting for each 5 bp TSD pair. For each sample, genotype likelihoods were then assessed at each site based on re-mapping of those reads to either allele, with error probabilities based on read mapping quality [38, 104], excluding sites without re-mapped reads for a given sample. Read pairs for which both reads mapped to the internal portion of the element were excluded to avoid false positive calls potentially introduced by non-specific alignment. The pipeline for genotyping is available at <https://github.com/KiddLab/insertion-genotype>. The genotyped samples were sorted by ancestral population, and allele frequencies estimated for the total number of individuals per population genotyped at each locus (Additional file 7: Table S5).

Admixture

A sample set containing only dogs and wolves were previously genotyped at approximately 7.6 million SNPs determined to capture genetic diversity across canids [44]. Using Plink [105], sites were filtered to remove those with missing genotypes in at least ten percent of samples, those in LD with another SNP within 50 bp ($-indep\text{-pairwise } 50 \ 10 \ 0.1$), and randomly thinned to 500,000 SNPs. To reduce the bias of relatedness, the sample set was further filtered to remove duplicates within a single modern breed, leaving 254 samples (Additional file 10: Table S7). Identification of wolf samples with high dog ancestry was made through five independent ADMIXTURE [106] analyses of the thinned SNP set with random seeds for K values 2 through 6. Since we aimed to discern cfERV-Fc1(a) insertions that may be dog-specific (*i.e.* having occurred since domestication), we removed any gray wolf that had high dog ancestry from further analysis. To do this, we calculated average dog ancestry within gray wolves at $K=3$ across all runs, which was the K value with the lowest cross validation error rate. Wolves with greater than 10% dog ancestry (an Israeli (isw01) and Spanish (spw01) wolf) were excluded from subsequent species and sub-population assessments.

Phylogenetic analysis

Nucleotide alignments were performed using MUSCLE [107] followed by manual editing in BioEdit [108] for intact CfERV-Fc1(a) LTRs from 19 proviral elements and 142 solo-LTRs. Of non-reference elements, the solo LTR with a 388 bp internal deletion at chr22:57,677,068 was excluded, as was the 141 bp truncated solo LTR at chr5:80,814,713. We also excluded partially reconstructed insertions corresponding to ‘one-sided’ assemblies or sites with linked 5′ and 3′ assembled junctions but that lacked internal resolution (Additional file 1: Table S1). A maximum likelihood (ML) phylogeny was reconstructed from the LTR alignment using FastTree [109] and the (GTR+CAT) model [generalized time reversible (GTR) model of nucleotide substitution plus “CAT” rate approximation]. Sites containing missing data or alignment gaps were removed from the analysis. To infer the robustness of inferred splits in the phylogeny, local support values were calculated using the ML-based approach implemented in FastTree, wherein the Shimodaira-Hasegawa test is applied to the three alternate topologies (NNIs) around each node. The average d_N/d_S ratio for intact env genes was determined using the codeml program in the PAML software package (version 4.8) [110] based on a Neighbor-Joining tree. Statistical significance was determined using the Nei–Gojori

method [111] implemented in MEGA7 [112] with a null hypothesis of strict neutrality ($d_N = d_S$).

Additional files

Additional file 1: 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.

Additional file 2: 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 5 bp 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).

Additional file 3: 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.

Additional file 4: 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.

Additional file 5: 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.

Additional file 6: 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 Additional file 1: Table S1

Additional file 7: 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 “-”.

Additional file 8: 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.

Additional file 9: 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/>).

Additional file 10: 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 Additional file 1: Table S1.

Abbreviations

BAM: binary alignment/map; CERV: canine endogenous retrovirus; d_n : rate of nonsynonymous substitutions; d_s : rate of synonymous substitutions; ERV: endogenous retrovirus; LTR: long terminal repeat; mya: million years ago; ORF: open reading frame; RT: reverse transcriptase; SNV: single nucleotide variant; TSD: target site duplication; WGA: whole genome amplification.

Authors' 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.

Author details

¹ Department of Biological Sciences, Bowling Green State University, Bowling Green, OH 43403, USA. ² Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI 48109, USA. ³ Centre for Virus Research, University of Glasgow, Glasgow G12 8QQ, Scotland, UK. ⁴ Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, 100 Washtenaw Ave., Ann Arbor, MI 48109, USA.

Acknowledgements

We thank John Coffin, Michael Freeman, Welkin Johnson, John Moran, 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 Marqués-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).

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in NCBI Sequence Read Archive (SRA, <https://www.ncbi.nlm.nih.gov/sra>) or are included in this published article and its supplementary information files. New proviral sequences have been deposited in GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>) under accessions MK039120-MK039127.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Funding

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.

Publisher's Note

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

Received: 17 January 2019 Accepted: 28 February 2019

Published online: 07 March 2019

References

- Boeke J, Stoye J. Retrotransposons, endogenous retroviruses, and the evolution of retroelements. In: Coffin J, Hughes S, Varmus H, editors. *Retroviruses*. New York: CSHL Press; 1997. p. 343–435.
- Troyer JL, Pecon-Slattery J, Roelke ME, Black L, Packer C, O'Brien SJ. Patterns of feline immunodeficiency virus multiple infection and genome divergence in a free-ranging population of African lions. *J Virol*. 2004;78(7):3777–91.
- Lober U, Hobbs M, Dayaram A, Tsangaras K, Jones K, Alquezar-Planas DE, Ishida Y, Meers J, Mayer J, Quedenau C, et al. Degradation and remobilization of endogenous retroviruses by recombination during the earliest stages of a germ-line invasion. *Proc Natl Acad Sci U S A*. 2018;115(34):8609–14.
- Tarlinton RE, Meers J, Young PR. Retroviral invasion of the koala genome. *Nature*. 2006;442(7098):79–81.
- Ishida Y, Zhao K, Greenwood AD, Roca AL. Proliferation of endogenous retroviruses in the early stages of a host germ line invasion. *Mol Biol Evol*. 2015;32(1):109–20.
- Roca AL, Pecon-Slattery J, O'Brien SJ. Genomically intact endogenous feline leukemia viruses of recent origin. *J Virol*. 2004;78(8):4370–5.
- Elleder D, Kim O, Padhi A, Bankert JG, Simeonov I, Schuster SC, Wittekindt NE, Motameny S, Poss M. Polymorphic integrations of an endogenous gammaretrovirus in the mule deer genome. *J Virol*. 2012;86(5):2787–96.
- Kamath PL, Elleder D, Bao L, Cross PC, Powell JH, Poss M. The population history of endogenous retroviruses in mule deer (*Odocoileus hemionus*). *J Hered*. 2014;105(2):173–87.
- Wildschutte JH, Williams ZH, Montesin M, Subramanian RP, Kidd JM, Coffin JM. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. *Proc Natl Acad Sci U S A*. 2016;113(16):E2326–34.
- Diehl WE, Patel N, Halm K, Johnson WE. Tracking interspecies transmission and long-term evolution of an ancient retrovirus using the genomes of modern mammals. *Elife*. 2016;5:e12704.
- Barrio AM, Ekerljung M, Jern P, Benachenhou F, Sperber GO, Bongcam-Rudloff E, Blomberg J, Andersson G. The first sequenced carnivore genome shows complex host-endogenous retrovirus relationships. *PLoS ONE*. 2011;6(5):e19832.
- Zhuo X, Rho M, Feschotte C. Genome-wide characterization of endogenous retroviruses in the bat *Myotis lucifugus* reveals recent and diverse infections. *J Virol*. 2013;87(15):8493–501.
- Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science*. 2016;351(6277):1083–7.
- Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, Firth A, Singer O, Trono D, Pfaff SL. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*. 2012;487(7405):57–63.
- Rebollo R, Romanish MT, Mager DL. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu Rev Genet*. 2012;46:21–42.
- Lavialle C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, Vernochet C, Heidmann T. Paleovirology of 'syncytins', retroviral env genes

- exapted for a role in placentation. *Philos Trans R Soc Lond B Biol Sci.* 2013;368(1626):20120507.
17. Nethe M, Berkhout B, van der Kuyl AC. Retroviral superinfection resistance. *Retrovirology.* 2005;2:52.
 18. Stoye JP. Studies of endogenous retroviruses reveal a continuing evolutionary saga. *Nat Rev Microbiol.* 2012;10(6):395–406.
 19. Blanco-Melo D, Gifford RJ, Bieniasz PD. Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors. *Elife.* 2017;6:e22519.
 20. Weiss RA, Stoye JP. Virology. Our viral inheritance. *Science.* 2013;340(6134):820–1.
 21. Lynch VJ, Leclerc RD, May G, Wagner GP. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat Genet.* 2011;43(11):1154–9.
 22. Benit L, Calteau A, Heidmann T. Characterization of the low-copy HERV-Fc family: evidence for recent integrations in primates of elements with coding envelope genes. *Virology.* 2003;312(1):159–68.
 23. Vargiu L, Rodriguez-Tome P, Sperber GO, Cadeddu M, Grandi N, Blikstad V, Tramontano E, Blomberg J. Classification and characterization of human endogenous retroviruses; mosaic forms are common. *Retrovirology.* 2016;13:7.
 24. Jern P, Sperber GO, Blomberg J. Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. *Retrovirology.* 2005;2:50.
 25. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ 3rd, Zody MC, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature.* 2005;438(7069):803–19.
 26. Ghernati I, Corbin A, Chabanne L, Auger C, Magnol JP, Fournel C, Monier JC, Darlix JL, Rigal D. Canine large granular lymphocyte leukemia and its derived cell line produce infectious retroviral particles. *Vet Pathol.* 2000;37(4):310–7.
 27. Modiano JF, Breen M, Burnett RC, Parker HG, Inusah S, Thomas R, Avery PR, Lindblad-Toh K, Ostrander EA, Cutter GC, et al. Distinct B-cell and T-cell lymphoproliferative disease prevalence among dog breeds indicates heritable risk. *Cancer Res.* 2005;65(13):5654–61.
 28. Modiano JF, Getzy DM, Akol KG, Van Winkle TJ, Cockerell GL. Retrovirus-like activity in an immunosuppressed dog: pathological and immunological findings. *J Comp Pathol.* 1995;112(2):165–83.
 29. Onions D. RNA-dependent DNA polymerase activity in canine lymphosarcoma. *Eur J Cancer.* 1980;16(3):345–50.
 30. Perk K, Safran N, Dahlberg JE. Propagation and characterization of novel canine lentivirus isolated from a dog. *Leukemia.* 1992;6(Suppl 3):155S–7S.
 31. Safran N, Perk K, Eyal O, Dahlberg JE. Isolation and preliminary characterization of a novel retrovirus isolated from a leukaemic dog. *Res Vet Sci.* 1992;52(2):250–5.
 32. Tomley FM, Armstrong SJ, Mahy BW, Owen LN. Reverse transcriptase activity and particles of retroviral density in cultured canine lymphosarcoma supernatants. *Br J Cancer.* 1983;47(2):277–84.
 33. Stocking C, Kozak CA. Murine endogenous retroviruses. *Cell Mol Life Sci.* 2008;65(21):3383–98.
 34. Macdonald DW, Sillero-Zubiri C. The biology and conservation of wild canids. New York: Oxford University Press; 2004.
 35. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: a resource for time-lines, timetrees, and divergence times. *Mol Biol Evol.* 2017;34(7):834–45.
 36. Koepfli KP, Pollinger J, Godinho R, Robinson J, Lea A, Hendricks S, Schweizer RM, Thalmann O, Silva P, Fan Z, et al. Genome-wide evidence reveals that African and Eurasian golden jackals are distinct species. *Curr Biol.* 2015;25(16):2158–65.
 37. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbil JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics.* 2012;28(18):i333–9.
 38. Wildschutte JH, Baron A, Diroff NM, Kidd JM. Discovery and characterization of Alu repeat sequences via precise local read assembly. *Nucleic Acids Res.* 2015;43(21):10292–307.
 39. Jern P, Coffin JM. Effects of retroviruses on host genome function. *Annu Rev Genet.* 2008;42:709–32.
 40. Grandi N, Tramontano E. HERV envelope proteins: physiological role and pathogenic potential in cancer and autoimmunity. *Front Microbiol.* 2018;9:462.
 41. Maruggi G, Porcellini S, Facchini G, Perna SK, Cattoglio C, Sartori D, Ambrosi A, Schambach A, Baum C, Bonini C, et al. Transcriptional enhancers induce insertional gene deregulation independently from the vector type and design. *Mol Ther.* 2009;17(5):851–6.
 42. Johnson WE, Coffin JM. Constructing primate phylogenies from ancient retrovirus sequences. *Proc Natl Acad Sci.* 1999;96(18):10254–60.
 43. Hughes JF, Coffin JM. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. *Proc Natl Acad Sci U S A.* 2004;101(6):1668–72.
 44. Botigue LR, Song S, Scheu A, Gopalan S, Pendleton AL, Oetjens M, Taravella AM, Seregely T, Zeeb-Lanz A, Arbogast RM, et al. Ancient European dog genomes reveal continuity since the Early Neolithic. *Nat Commun.* 2017;8:16082.
 45. Pendleton AL, Shen F, Taravella AM, Emery S, Veeramah KR, Boyko AR, Kidd JM. Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. *BMC Biol.* 2018;16(1):64.
 46. Flockerzi A, Burkhardt S, Schempp W, Meese E, Mayer J. Human endogenous retrovirus HERV-K14 families: status, variants, evolution, and mobilization of other cellular sequences. *J Virol.* 2005;79(5):2941–9.
 47. Grandi N, Cadeddu M, Blomberg J, Tramontano E. Contribution of type W human endogenous retroviruses to the human genome: characterization of HERV-W proviral insertions and processed pseudogenes. *Retrovirology.* 2016;13(1):67.
 48. Fan Z, Silva P, Gronau I, Wang S, Armero AS, Schweizer RM, Ramirez O, Pollinger J, Galaverni M, Ortega Del-Vecchio D, et al. Worldwide patterns of genomic variation and admixture in gray wolves. *Genome Res.* 2016;26(2):163–73.
 49. Belshaw R, Watson J, Katzourakis A, Howe A, Woolven-Allen J, Burt A, Tristem M. Rate of recombinational deletion among human endogenous retroviruses. *J Virol.* 2007;81(17):9437–42.
 50. Stankiewicz P, Lupski JR. Molecular-evolutionary mechanisms for genomic disorders. *Curr Opin Genet Dev.* 2002;12(3):312–9.
 51. Hughes JF, Coffin JM. Human endogenous retroviral elements as indicators of ectopic recombination events in the primate genome. *Genetics.* 2005;171(3):1183–94.
 52. Sinha A, Johnson WE. Retroviruses of the RDR superinfection interference group: ancient origins and broad host distribution of a promiscuous Env gene. *Curr Opin Virol.* 2017;25:105–12.
 53. Cianciolo GJ, Copeland TD, Oroszlan S, Snyderman R. Inhibition of lymphocyte proliferation by a synthetic peptide homologous to retroviral envelope proteins. *Science.* 1985;230(4724):453–5.
 54. Ali LM, Rizvi TA, Mustafa F. Cross- and co-packaging of retroviral RNAs and their consequences. *Viruses.* 2016;8(10):276.
 55. Clough JE, Foster JA, Barnett M, Wichman HA. Computer simulation of transposable element evolution: random template and strict master models. *J Mol Evol.* 1996;42(1):52–8.
 56. Nascimento FF, Rodrigo AG. Computational evaluation of the strict master and random template models of endogenous retrovirus evolution. *PLoS ONE.* 2016;11(9):e0162454.
 57. Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, Tristem M. Long-term reinfection of the human genome by endogenous retroviruses. *Proc Natl Acad Sci U S A.* 2004;101(14):4894–9.
 58. Belshaw R, Katzourakis A, Paces J, Burt A, Tristem M. High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection. *Mol Biol Evol.* 2005;22(4):814–7.
 59. Ostertag EM, Kazazian HH Jr. Biology of mammalian L1 retrotransposons. *Annu Rev Genet.* 2001;35:501–38.
 60. Magiorkinis G, Gifford RJ, Katzourakis A, De Ranter J, Belshaw R. Env-less endogenous retroviruses are genomic superspreaders. *Proc Natl Acad Sci U S A.* 2012;109(19):7385–90.
 61. Mager DL, Stoye JP. Mammalian endogenous retroviruses. *Microbiol Spectr.* 2015;3(1):MDNA3-0009-2014.
 62. Frank JA, Feschotte C. Co-option of endogenous viral sequences for host cell function. *Curr Opin Virol.* 2017;25:81–9.
 63. Jo H, Choi H, Choi MK, Song N, Kim JH, Oh JW, Seo K, Seo HG, Chun T, Kim TH, et al. Identification and classification of endogenous retroviruses in the canine genome using degenerative PCR and in silico data analysis. *Virology.* 2012;422(2):195–204.
 64. Tarlinton RE, Barfoot HK, Allen CE, Brown K, Gifford RJ, Emes RD. Characterisation of a group of endogenous gammaretroviruses in the canine genome. *Vet J.* 2013;196(1):28–33.

65. Parsons WH, Kolar MJ, Kamat SS, Cognetta AB 3rd, Hulce JJ, Saez E, Kahn BB, Saghatelian A, Cravatt BF. AIG1 and ADTRP are atypical integral membrane hydrolases that degrade bioactive FAHFs. *Nat Chem Biol*. 2016;12(5):367–72.
66. Yan J, Chen G, Zhao X, Chen F, Wang T, Miao F. High expression of diffuse panbronchiolitis critical region 1 gene promotes cell proliferation, migration and invasion in pancreatic ductal adenocarcinoma. *Biochem Biophys Res Commun*. 2018;495(2):1908–14.
67. Mayers JR, Torrence ME, Danai LV, Papagiannakopoulos T, Davidson SM, Bauer MR, Lau AN, Ji BW, Dixit PD, Hosios AM, et al. Tissue of origin dictates branched-chain amino acid metabolism in mutant Kras-driven cancers. *Science*. 2016;353(6304):1161–5.
68. Dey P, Baddour J, Muller F, Wu CC, Wang H, Liao WT, Lan Z, Chen A, Gutschner T, Kang Y, et al. Genomic deletion of malic enzyme 2 confers collateral lethality in pancreatic cancer. *Nature*. 2017;542(7639):119–23.
69. Ananieva EA, Wilkinson AC. Branched-chain amino acid metabolism in cancer. *Curr Opin Clin Nutr Metab Care*. 2018;21(1):64–70.
70. de Mattos LC. Structural diversity and biological importance of ABO, H, Lewis and secretor histo-blood group carbohydrates. *Rev Bras Hematol Hemoter*. 2016;38(4):331–40.
71. Ferrer-Admetlla A, Sikora M, Laayouni H, Esteve A, Roubinet F, Blancher A, Calafell F, Bertranpetit J, Casals F. A natural history of FUT2 polymorphism in humans. *Mol Biol Evol*. 2009;26(9):1993–2003.
72. Le Pendu J, Ruvoen-Clouet N, Kindberg E, Svensson L. Mendelian resistance to human norovirus infections. *Semin Immunol*. 2006;18(6):375–86.
73. Thorven M, Grahn A, Hedlund KO, Johansson H, Wahlfrid C, Larson G, Svensson L. A homozygous nonsense mutation (428G → A) in the human secretor (FUT2) gene provides resistance to symptomatic norovirus (GGII) infections. *J Virol*. 2005;79(24):15351–5.
74. Kindberg E, Hejdeman B, Bratt G, Wahren B, Lindblom B, Hinkula J, Svensson L. A nonsense mutation (428G → A) in the fucosyltransferase FUT2 gene affects the progression of HIV-1 infection. *AIDS*. 2006;20(5):685–9.
75. Skoglund P, Ersmark E, Palkopoulou E, Dalen L. Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Curr Biol*. 2015;25(11):1515–9.
76. Frantz LA, Mullin VE, Pionnier-Capitan M, Lebrasseur O, Ollivier M, Perri A, Linderholm A, Mattiangeli V, Teasdale MD, Dimopoulos EA, et al. Genomic and archaeological evidence suggest a dual origin of domestic dogs. *Science*. 2016;352(6290):1228–31.
77. Lober U, Hobbs M, Dayaram A, Tsangaras K, Jones K, Alquezar-Planas DE, Ishida Y, Meers J, Mayer J, Quedenau C, et al. Degradation and remobilization of endogenous retroviruses by recombination during the earliest stages of a germ-line invasion. *Proc Natl Acad Sci U S A*. 2018;115:8609–14.
78. Young GR, Eksmond U, Salcedo R, Alexopoulou L, Stoye JP, Kassiotis G. Resurrection of endogenous retroviruses in antibody-deficient mice. *Nature*. 2012;491(7426):774–8.
79. Fadel HJ, Poeschla EM. Retroviral restriction and dependency factors in primates and carnivores. *Vet Immunol Immunopathol*. 2011;143(3–4):179–89.
80. Sawyer SL, Emerman M, Malik HS. Discordant evolution of the adjacent antiretroviral genes TRIM22 and TRIM5 in mammals. *PLoS Pathog*. 2007;3(12):e197.
81. Auton A, Rui Li Y, Kidd J, Oliveira K, Nadel J, Holloway JK, Hayward JJ, Cohen PE, Grealley JM, Wang J, et al. Genetic recombination is targeted towards gene promoter regions in dogs. *PLoS Genet*. 2013;9(12):e1003984.
82. Decker B, Davis BW, Rimbault M, Long AH, Karlins E, Jagannathan V, Reiman R, Parker HG, Drogemuller C, Corneveaux JJ, et al. Comparison against 186 canid whole-genome sequences reveals survival strategies of an ancient clonally transmissible canine tumor. *Genome Res*. 2015;25(11):1646–55.
83. Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva PM, Galaverni M, Fan Z, Marx P, Lorente-Galdos B, et al. Genome sequencing highlights the dynamic early history of dogs. *PLoS Genet*. 2014;10(1):e1004016.
84. Marsden CD, Ortega-Del Vecchyo D, O'Brien DP, Taylor JF, Ramirez O, Vila C, Marques-Bonet T, Schnabel RD, Wayne RK, Lohmueller KE. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proc Natl Acad Sci USA*. 2016;113(1):152–7.
85. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297–303.
86. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*. 2005;110(1–4):462–7.
87. Peng X, Alfoldi J, Gori K, Einfeld AJ, Tyler SR, Tisoncik-Go J, Brawand D, Law GL, Skunca N, Hatta M, et al. The draft genome sequence of the ferret (*Mustela putorius furo*) facilitates study of human respiratory disease. *Nat Biotechnol*. 2014;32(12):1250–5.
88. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, et al. The sequence and de novo assembly of the giant panda genome. *Nature*. 2010;463(7279):311–7.
89. Pontius JU, Mullikin JC, Smith DR, Agencourt Sequencing T, Lindblad-Toh K, Gnerre S, Clamp M, Chang J, Stephens R, Neelam B, et al. Initial sequence and comparative analysis of the cat genome. *Genome Res*. 2007;17(11):1675–89.
90. Keane TM, Wong K, Adams DJ. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics*. 2013;29(3):389–90.
91. Huang X, Madan A. CAP3: a DNA sequence assembly program. *Genome Res*. 1999;9(9):868–77.
92. Parsons JD. Miropeats: graphical DNA sequence comparisons. *Comput Appl Biosci*. 1995;11(6):615–9.
93. Quinlan AR. BEDTools: the Swiss-army tool for genome feature analysis. *Curr Protoc Bioinform*. 2014;47:11–2 **11–34**.
94. Kim RN, Kim DS, Choi SH, Yoon BH, Kang A, Nam SH, Kim DW, Kim JJ, Ha JH, Toyoda A, et al. Genome analysis of the domestic dog (Korean Jindo) by massively parallel sequencing. *DNA Res*. 2012;19(3):275–87.
95. Vamathevan JJ, Hall MD, Hasan S, Woollard PM, Xu M, Yang Y, Li X, Wang X, Kenny S, Brown JR, et al. Minipig and beagle animal model genomes aid species selection in pharmaceutical discovery and development. *Toxicol Appl Pharmacol*. 2013;270(2):149–57.
96. Owczarek-Lipska M, Jagannathan V, Drogemuller C, Lutz S, Glanemann B, Leeb T, Kook PH. A frameshift mutation in the cubilin gene (CUBN) in Border Collies with Imlerslund-Grasbeck syndrome (selective cobalamin malabsorption). *PLoS ONE*. 2013;8(4):e61144.
97. Wang GD, Zhai W, Yang HC, Fan RX, Cao X, Zhong L, Wang L, Liu F, Wu H, Cheng LG, et al. The genomics of selection in dogs and the parallel evolution between dogs and humans. *Nat Commun*. 2013;4:1860.
98. Kim HM, Cho YS, Kim H, Jho S, Son B, Choi JY, Kim S, Lee BC, Bhak J, Jang G. Whole genome comparison of donor and cloned dogs. *Sci Rep*. 2013;3:2998.
99. Li Y, Wu DD, Boyko AR, Wang GD, Wu SF, Irwin DM, Zhang YP. Population variation revealed high-altitude adaptation of Tibetan mastiffs. *Mol Biol Evol*. 2014;31(5):1200–5.
100. Zhang W, Fan Z, Han E, Hou R, Zhang L, Galaverni M, Huang J, Liu H, Silva P, Li P, et al. Hypoxia adaptations in the grey wolf (*Canis lupus chanco*) from Qinghai-Tibet Plateau. *PLoS Genet*. 2014;10(7):e1004466.
101. Wang GD, Zhai W, Yang HC, Wang L, Zhong L, Liu YH, Fan RX, Yin TT, Zhu CL, Poyarkov AD, et al. Out of southern East Asia: the natural history of domestic dogs across the world. *Cell Res*. 2016;26(1):21–33.
102. Robinson JA, Ortega-Del Vecchyo D, Fan Z, Kim BY, vonHoldt BM, Marsden CD, Lohmueller KE, Wayne RK. Genomic flatlining in the endangered island fox. *Curr Biol*. 2016;26(9):1183–9.
103. Liu D, Xiong H, Ellis AE, Northrup NC, Rodriguez CO Jr, O'Regan RM, Dalton S, Zhao S. Molecular homology and difference between spontaneous canine mammary cancer and human breast cancer. *Cancer Res*. 2014;74(18):5045–56.
104. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987–93.
105. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559–75.
106. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;19(9):1655–64.

107. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004;32(5):1792–7.
108. Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser.* 1999;41:95–8.
109. Price MN, Dehal PS, Arkin AP. FastTree 2-approximately maximum-likelihood trees for large alignments. *PLoS ONE.* 2010;5(3):e9490.
110. Xu B, Yang Z. PAMLX: a graphical user interface for PAML. *Mol Biol Evol.* 2013;30(12):2723–4.
111. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol.* 1986;3(5):418–26.
112. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol.* 2016;33(7):1870–4.
113. Wang X, Tedford RH. *DOGS: their fossil relatives and evolutionary history.* New York, Chichester, West Sussex: Columbia University Press; 2008.
114. Koblmüller S, Vilà C, Lorente-Galdos B, Dabad M, Ramirez O, Marques-Bonet T, Wayne RK, Leonard JA. Whole mitochondrial genomes illuminate ancient intercontinental dispersals of grey wolves (*Canis lupus*). *J Biogeogr.* 2016;43(9):1728–38.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

