Pre-Cambrian origin of *envelope*-carrying endogenous retroviruses in metazoans

Shashank Chary[1] and Rippei Hayashi[1 #]

1. The Shine-Dalgarno Centre for RNA Innovation, The John Curtin School of Medical Research, The Australian National University, 131 Garran Road, Acton, ACT 2601, Australia

correspondence: rippei.hayashi@anu.edu.au.

Abstract

Retrotransposons or endogenous retroviruses (ERVs) essentially carry open reading frames of *gag* and *pol*, which are utilized to selfishly replicate themselves in the host's germline genome. One rare example of ERVs that additionally carry *envelope* genes, are Ty3/*gypsy* errantiviruses in Drosophila. Though they are structurally analogous to vertebrate retroviruses, their evolutionary relationship as well as the origin of *envelope* genes remain elusive. We systematically searched for intact *envelope*-containing ERVs that are homologous to Ty3/*gypsy* and vertebrate retroviruses in metazoan genomes and found that they are widespread across taxa including ancient animals such as Cnidarians, Ctenophores and Tunicates. Most elements are multiplied in their respective genomes, implying recent germline expansion. *Envelope* genes are classified into those that resemble glycoprotein F from paramyxoviruses (F-type), glycoprotein B from herpesviruses (HSV/gB-type), and GP2 glycoprotein from retroviruses and Ebolavirus (GP2-type). Both F-type and HSV/gB-type ERVs are ubiquitously found in invertebrate genomes while vertebrate genomes carry F-type and GP2-type ERVs. Interestingly, vertebrate ERVs share a distinct and structurally conserved domain of POL with Ty3/*gypsy* elements from insects and ancient animals, suggesting that they share common origins. Furthermore, phylogenetic and structural analyses revealed that *envelope* genes have largely diverged with *pol* genes as well as with the host organisms throughout their evolutionary history and recombined infrequently, suggesting that the *envelope* acquisition to ERVs is ancient and likely dates to before the split of bilaterian and non-bilaterian animals in Pre-Cambrian era.

Introduction

Retroviruses have two mechanisms through which they survive; most commonly, they exogenously infect other hosts, transmitting horizontally [1]. However, they can occasionally infect the germline cells of the host; where genetic information is stored for the next generation and integrate into the genome [1,2]. They become fixed in the genome, vertically transferring to the next generation and become known as endogenous retroviruses (ERVs) [1-4]. They are well characterised in vertebrates, making up 5-10% of their genomes. The best-known example of this are Koala retroviruses (KoRVs), in which nearly identical copies of an exogenous lymphoma-causing retrovirus are found in the genomes of geographically distinct Koala populations [5,6]. Though, it should be noted that such germline infection events have rarely been identified, and the exact exogenous origins of many ERVs is not known.

Once a virus becomes endogenized, they have two possible fates. Most commonly, proviral copies in the genome acquire deletions and frameshifts, degrading over time, until only fragments remain [7]. Some retroviral genes are even co-opted by the host to assist in biological functions [8]. It is for this reason that ERVs tend to be thought of as ancient relics of past retroviral infection [9,10]. However, some ERVs retain functionality of their retroviral proteins, and can endogenously create new copies in the genome, keeping their code intact. This requires the activity of the retroviral genes *gag* and *pol*, which are usually flanked by long terminal repeats (LTRs). These active intact ERVs are more broadly classified as class I LTR retrotransposons [11].

Retrotransposons usually exist in the genome as 'selfish' genes, endogenously creating new copies in the germline to survive long term [2,12]. Sometimes, they can be lost from the genome, if they do not remain active enough, or if they acquire random mutations that affects their ability to transpose. The study of intact retrotransposons therefore provides an interesting opportunity, as they can provide great insights into the evolution of retroviruses. It has been postulated that intact ERVs and retrotransposons are the result of more recent retroviral infection [13].

The Ty3/*gypsy* family of retrotransposons is one of the largest and best studied families of LTR retrotransposons, present ubiquitously in the genomes of most eukaryotes. Their origin, likely to be highly ancient [14], has been suggested to be linked to that of vertebrate retroviruses as well. In fact, the *gag* and *pol* genes from vertebrate retroviruses are most closely linked to those of Ty3 elements [15-17]. A recent study identified a new class of ERV in vertebrates called Lokiviruses [18], which has been suggested to be a link between the two, further supporting the idea that the *gag* and *pol* genes of Ty3 retrotransposons and vertebrate retroviruses have a common origin.

One specific branch of Ty3 elements in Drosophila species carry an open reading frame (ORF) for the retroviral *envelope* or *env* gene in addition to *gag* and *pol*, often referred to as 'errantiviruses' [19,20]. Genetic evidence in Drosophila shows that these errantiviruses behave as bona-fide retrotransposons, creating new germline copies that are vertically inherited [21,22]. Interestingly, errantiviruses are expressed in the somatic cells of the female ovaries, forming virus like particles (VLPs) that infect the adjacent germline cells [21,22].

Essentially, errantiviruses can actively infect other cells, rather than maintain copies in the same cell as typical retrotransposons. This means that if the errantivirus remains active in the somatic cells, it can still maintain the generation of new copies that can be vertically transmitted to the next generation, endogenously via cell-to-cell transmission.

Errantiviruses are an interesting branch of retrotransposons, as they are the most 'intact' elements compared to retroviruses, given the presence of *env*. But although they share structural similarities, the exact origins are poorly understood. It has been suggested that Ty3 elements acquired the *env* gene from infecting baculoviruses in an ancestral Drosophila species via recombination events [23,24]. Errantiviruses are only known to be found in dipteran genomes, identified recently in mosquito species [25]. However, Baculoviruses are also known to infect non-dipteran insect orders, most famously Lepidoptera species [26]. Additionally, the origins for the *env* gene in baculoviruses is also poorly understood, leading to ambiguity in the true origins of these elements. Furthermore, LTR retrotransposons with *env* genes have been sporadically documented in other animals, such as in nematodes and rotifers [27,28]. However, their origins are even less clear, due to a lack of systematic studies to mine genomes for such elements.

Errantiviruses may therefore be more widespread than previously thought, and the search for intact errantivirus elements in metazoan orders allows us a unique opportunity to study the acquisition of retroviral *env* genes by retrotransposons over time.

Results

We systematically searched for homologous elements of Dipteran Ty3/*gypsy* errantiviruses in the referenced genomes across metazoan phyla using the annotated protein coding sequences of *gag*, *pol* and *env* for tBLASTn searches (see methods and Fig S1). We additionally used the protein coding sequences of known vertebrate retroviruses to search for homologous retroelements in vertebrate genomes. We took genomic copies that carry all three open reading frames (ORFs). Where possible, we focused our search on elements that are multiplied in the genome to understand relationships of currently active elements. The first tBLASTn search identified elements outside Diptera and Vertebrata. Therefore, we extended the search using the representative of newly identified elements as baits. We repeated this process and examined whether the additional baits uncovered elements that were not represented in the previous tBLASTn searches by identifying new clades in the phylogenetic tree of POL (see methods and Fig S1). After a five iteration of tBLASTn searches, we no longer discovered elements that formed a new POL phylogenetic clade (Fig S1B). These steps allowed us to obtain a comprehensive set of elements in metazoan genomes that are homologous to Dipteran Ty3/*gypsy* errantiviruses and vertebrate retroviruses (Table S1, summarised in Fig 1). Most of the newly identified genomic copies carry putative LTRs and tRNA primer binding sequences (1308 out of 1765 elements, Table S1), suggesting their recent integrations. Additionally, most of the annotated elements sit within contigs larger than 0.1 mega bases (1649 elements, Table S1), indicating their true origins in their respective genomes. We ran multiple sequence alignments using the conserved Integrase (INT) and Reverse Transcriptase (RT) domains of Pol and assembled maximum probability trees to infer phylogenetic relationships of the identified elements.

ORFs of selected annotated Drosophila Ty3 elements and vertebrate retroviruses were also included in the trees to aid classifications. We refer to elements found in non-vertebrate metazoan phyla as "errantiviruses", and elements found in vertebrate genomes as "vERVs".

Multiple intact elements of *env*-carrying Ty3 retrotransposons are found widespread across metazoan species

We found that intact errantiviruses are found not only outside of insect orders but are ubiquitously present across most major metazoan phyla. This includes species from ancient phyla such as Cnidaria, Tunicates and Ctenophora, but excludes Echinodermata and Porifera (Fig 1B and Table S1). This range of hosts, which includes ancient phyla, suggests that the elements identified in this study are highly ancient. This is also supported by the phylogenetic tree assembled on *pol* RT and INT proteins. Firstly, the phylogenetic tree assembled on *pol* RT reflects that of *pol* INT, with errantiviruses within most clades in one tree mapping to the same or closely related clades in the other (Fig S2B). Importantly, errantiviruses found in phyla outside of insects, cluster separately to those found within, which we call "ancient errantiviruses" and "insect errantiviruses", respectively (Fig 1C and S2A). Notably however, few elements found in insects sit within clades of ancient errantiviruses, rather than other insect ones (Fig 1C). Annotated Drosophila *gypsy* elements included in the tree, namely, Gypsy-ZAM, Gypsy-Gypsy/Gypsy-5_DGri and Gypsy-DM176/Idefix present themselves in different clades while they cluster together with other insect errantiviruses, in an arrangement with previous studies that have classified Drosophila *gypsy* elements [29] (Fig 1C). Though, one element from a cnidarian species *Paleozoanthus reticulatus* clusters together with insect elements, in a clade adjacent to DM176 and Idefix (Fig 1A), its origin is unclear as the element is found on a short genomic contig (Table S1). While we have potentially identified new classifications of insect errantiviruses, given new clades (tree positions I3 - I6 in Fig S2A) that are formed between known insect *gypsy* and "ancient errantiviruses", the separation between "insect errantiviruses" and "ancient errantiviruses" is still highly indicative of their ancient age.

Within "insect errantiviruses", the clades are strongly supported by bootstrap values greater than 95 and are additionally supported by the annotations of the tRNA primer binding site (PBS) found in each element. For example, Tree position I1 within "insect errantiviruses" mainly carry $tRNA^{Ser}$, whereas Tree position I2, consisting of only dipteran elements, predominantly use $tRNA^{Lys}$, though few use $tRNA^{Ser}$ or $tRNA^{Pro}$ (Fig S2A and S2C). Elements in insects within positions I4, I5 and I6 that predominantly carry $tRNA^{Leu}$ primer binding sites, in addition to Serine and other uncommon tRNAs (Fig S2C). This is mostly consistent with past works, that have identified Drosophila errantiviruses to carry either Serine or Lysine tRNA PBS [30,31].

The "ancient errantivirus" group also consists of clades strongly supported by bootstrap values greater than 95, though the tRNA PBS content within the clades is much more divergent (Fig S3C). Notably however, errantiviruses found in insects that sit within this group, in tree positions A1, A2 and A7 continue to carry predominantly $tRNA^{Ser}$, $tRNA^{Lys}$ and $tRNA^{Leu}$ (Fig S3B). The clades that are formed in this group also reflect their ancient nature, as they consistently are made up of elements in hosts from the same, or closely

related phyla (Fig S3B). For example, tree position A5 consists of elements found in Thecostraca, Annelida and Mollusca, each of which form their own discrete clade (Fig S3B). Even within larger clades, the separation is evident. For example, tree position A9 consists of elements in a wide variety of hosts, but strongly supported clades only consist of elements from the same order of host (Fig S3B). This congruence between clades and species is highly indicative of long-term vertical transfer, as known in non *env*-containing retrotransposons [32,33].

Most elements that were identified have multiple intact copies in their respective genomes (as defined as copies with >98% nucleotide identity across >98% of the three ORFs, see methods, Fig 1C, Table S1), a hallmark feature of recent transposition [34,35], indicating that most elements that we have annotated likely behave as bona-fide retrotransposons. Activity of these elements are highly varied, as the copy number can be as little as one intact copy, to hundreds of intact copies in the genome (Fig 1C, Table S1). Importantly, multi-copy ERVs are present in nearly all animal phyla where we see intact elements, suggesting that these elements are actively transposing in a wide range of metazoan species today.

The *env* ORF downstream of *pol* in many of the elements resembles glycoprotein F1, found more commonly in paramyxoviruses, such as Nipah, Hendra and Respiratory Syncytial Viruses [36,37], and is also carried in insect baculoviruses [23,24]. However, some errantiviruses instead carry *env* that resembles glycoprotein B (gB), famously found as part of the glycoprotein complex in Herpes Simplex Viruses (HSV). Both types of *env* glycoproteins are carried by errantiviruses found in ancient hosts such as rotifers, tunicates, cnidarians and ctenophores, as well as in hosts from insect and arthropod orders (Fig 1B and S3B). Though, errantiviruses found in Nematoda, Bryozoa and Platyhelminthes only carry HSV/gB-type *env*, and the ORFs are arranged as GAG-ENV-POL rather than GAG-POL-ENV. Furthermore, clades formed on the trees using *pol* proteins tend to carry the same type of *env* (Fig S2A and S3A). For example, within A6 in the "ancient errantivirus" group, elements that carry HSV/gB are in a separate clade to those carrying F (Fig S3B). Though some clades appear to be a mix of both, the HSV/gB elements still tend to cluster together, indicating that they have stayed with *pol* for long evolutionary periods of time.

<u>The divergence of *env* glycoproteins carried by errantiviruses follows that of *pol*</u>

We found that errantiviruses can carry HSV/gB-type *env* in addition to glycoprotein F, in both "insect errantiviruses" and "ancient errantiviruses" (Fig 2A). One clade that exclusively consists of hymenopteran errantiviruses carry both HSV/gB-type and F-type *env* genes (Fig 2A). Interestingly, elements with HSV/gB-type *env* and F-type *env* separate at the base of this clade while they both are found in a wide range of species superfamily within Hymenoptera, suggesting that both HSV/gB-type *env* and F-type *env* have been together with their *pol* genes at least throughout the evolution of hymenopteran species.

To further characterise the ENV proteins carried by errantiviruses, we identified the ectodomains within the *env* ORF, which is flanked by signal peptides and a transmembrane domain, using DeepTMHMM (https://doi.org/10.1101/2022.04.08.487609). We then analysed trimeric configurations of these ectodomains from representative errantiviruses from the phylogenetic tree using Alphafold 2 [38].

F glycoproteins analysed are highly structured, and consistently form 4 domains that are highly conserved across all elements (Fig S4A). They also predicted a highly conserved Furin cleavage site and hydrophobic fusion peptides in the ectodomain (Fig S4A and S5). The 3D fold of the ectodomain structures forms a "head" (domains 1-3) and a "stalk" (domain 4), strongly resembling 3D structures of F ectodomains characterised in previous studies, such as those from RSVs [36] (Fig S4D). Most errantiviruses also have an alpha helix at the C terminus of the ectodomain, referred to as "Domain A" (Fig S4A and S4D). The predicted 3D trimeric structure also reveals two potential configurations of the glycoprotein, based on domain 4: either a long alpha helix, or two shorter helices (Fig S4D). These long and short barrel structures resemble pre- and post-fusion structures of glycoprotein F described in prior studies [36]. These highly structured domains in the intact *env* ORFs in these errantiviruses demonstrates that *env* has remained intact and could suggest that their function also remains intact as well.

To explore the phylogenetic relationships between *pol* and *env*, we generated multiple sequence alignments of the ectodomains and RT domains from representative elements and assembled phylogenetic trees. We found that the ENV ectodomains cluster similarly to RT in *pol*, and ENV proteins tend to form clades within the same or closely related species as they do for RT (Fig S4B and S4C). Importantly, the ENV ectodomains found in "ancient errantiviruses" cluster separately to those found in "insect errantiviruses". Additionally, insect elements that present themselves with ancient errantiviruses in the RT tree, are similarly positioned in clades within the ENV tree (Fig S4C). This suggests that *env* genes have stayed with *pol*, and the acquisition of *env* likely occurred anciently.

The ectodomain of errantiviruses that carry HSV/gB-type *env* are either flanked by signal peptides and a transmembrane domain, or by two transmembrane domains and the Furin cleavage site is not always conserved (Fig 2B). Like F, the trimeric structures of HSV/gB found in errantiviruses are also highly structured. They carry 5 highly conserved and distinct domains (designated as domain I – V) arranged in a hairpin configuration, that fold into a trimeric structure with a base, stalk and crown (Fig 2B). Additionally, domain I carries highly conserved "fusion loops" consisting of hydrophobic amino acid residues. These domains have also been characterized in Herpes Simplex Viruses in prior studies [39,40]. Notably, errantiviruses found in insect orders that carry HSV/gB, only carry one of these fusion loops (Fig S6).

The structure of the HSV/gB carried by these errantiviruses are stabilized by disulphide bonds between Cysteine residues, which we refer to as Cysteine bridges [41,42]. These bridges are highly conserved but appear to vary the most within domains I and II of the HSV/gB ectodomain (Fig 2C and S6). Additionally, in some cases, a Cysteine bridge within domain I is located near the fusion loops (Fig 2C and S6). The configuration of Cysteine bridges appears to be specific to phyla, though this does not seem to be reflected by their positions in the tree based on *pol* RT. For example, Bryozoa elements in tree position A3, and Bryozoa elements in tree position A8, both carry configuration C1c, though they each cluster themselves separately from other elements (Fig S3B, 2D and S6). Errantiviruses carrying HSV/gB found in insect phyla mostly have the same configuration regardless of their position in the *pol* RT tree. For example, insect elements within the "ancient errantivirus" group, and insect elements within the "insect errantivirus"

group, all have the same Cysteine bridge configuration in HSV/gB (Fig 2A, 2D and S6). There are, however, cases where more than one type of configuration can be found within the same phyla, namely Nematoda and Platyhelminthes (Fig 2D). Though, in those cases, they appear to continue to separate based on *pol* RT (Fig S6). For example, Nematoda elements in Position A2 have configurations C1a or C1d, while those in Position A8 have configuration C2a or C2c (Fig S6).

As mentioned above, the clades that form on RT mostly separate elements that carry HSV/gB from F within discrete clades (Fig S3B). This, with the fact that the same substructures can be found in different orders, for example in Nematoda, Bryozoa and Platyhelminthes, suggests that their origin is likely ancient. The overall specificity of certain residues to host orders further suggests that recombination events would have only occurred between elements within the same animal order. Furthermore, the overall separation of distinct Cysteine bridge configurations of HSV/gB, that are represented in the *pol* RT tree within ancient phyla, suggests that those recombination events occurred anciently. While insect elements are separated on the tree, ERVs with HSV/gB-type ENV tend to cluster separately from those with F-type ENV. For example, in the "insect errantivirus" group, the clade that almost exclusively carries HSV/gB-type ENV sits sister to another clade that only carries glycoprotein F like ENV (Fig 2A'). Not only do they separate from each other, but HSV/gB-type ENV can be found in a variety of superfamilies within Hymenoptera, which suggests their ancient origin (Fig 2A). Some species, such as a hymenopteran *Lasioglossum lativentre*, carry elements with HSV/gB, as well as those with F (Table S1). Despite being found in the genome of the same species, the *pol* of errantiviruses with F and HSV/gB do not stay together, further supporting their more ancient origin (Fig 2A'). It should be noted that some clades, such as ones in tree position 1 within "ancient errantiviruses", have elements that carry HSV/gB and F sitting together within the same discrete clade, suggestive of more recent recombination (Fig S3B).

<u>The RNase H Domains within the bridge region of *pol* separates "insect errantiviruses" from "ancient errantiviruses"</u>

To further characterize the errantiviruses identified in this study, we analysed the region between the INT and RT domains of Pol, referred to as the "bridge region", for representative elements. This region is known to carry a ribonuclease H (RNase H) domain, which in vertebrate retroviruses has duplicated, and the N-terminal copy, while partially degraded, maintains the same structure and 3D fold arrangement, referred to as "tether" [43]. We therefore characterised structures within the bridge region using the 3D fold arrangement of the core β-sheet structure using Alphafold 2.

We found that almost all errantiviruses that sit within the "insect errantivirus" group have similarly duplicated the RNase H domain, with the C terminal copy presenting itself as a "pseudo" RNase H domain (Fig 3A-B). The "pseudo" RNase H domain retains the barrel-like structure made of five core β-sheets, but the order of the second and the third β-sheets is flipped (Fig 3B).

On the other hand, "ancient errantiviruses" carry an RNase H domain, and a C-terminal "mini" domain. This consists of three small core β-sheets and α-helix residues at its N terminus. This "mini" domain is also

carried by insect elements in this group, further suggesting that those insect elements are equally ancient as the other errantiviruses in this group. Furthermore, this “mini” domain is also found in the bridge region of yeast S. *pombe* Ty3 *pol* (Fig 3), indicating it as a feature of highly ancient transposons.

We also identified a few uncommon structures, that are found in some phylogenetic groups in the tree, all of which sit at the C terminus of the bridge region (Fig 3B). One notable outlier is an insect element within the “ancient errantivirus” group that carries a variation of a “pseudo” RNase H domain (position 6, noted as “pseudo” domain 2 in Fig 3). This same domain can be found in a phylogenetic group in the “insect errantivirus” group (position 3 in Fig 3A). Furthermore, some elements in this “insect errantivirus” group carry a “mini” domain as well (position 3 in Fig 3A). Another is that some elements within the “insect errantivirus” group only carry a solitary RNase H domain (positions 4 and 5 in Fig 3A). It should be noted that these outliers are mostly specific to small phylogenetic groups/clades, and likely also arose somewhat anciently.

Vertebrate genomes carry intact copies of *env*-carrying ERVs

After finding that *env*-carrying retrotransposons are widespread in invertebrate metazoan genomes, and that ancient errantiviruses present themselves as distinct to insect errantiviruses, we extended the search for intact retroelements to vertebrate genomes. We identified intact elements with the same approach as errantiviruses, and assembled a phylogenetic tree using the sequences of *pol* RT and INT. We also verified that the assembled tree on RT reflects the tree of INT; elements in one clade are found in the same or similar clades on the other tree (Fig S7A).We included the sequences of Epsilon, Gamma, Foamy, Delta, Lenti, Alpha, Beta and Loki viruses, as well as the sequences of SNRV and ERV-L to aid the classification of the genomic copies that we identify.

Like errantiviruses, every class within vertebrates carry intact copies of *env*-containing retroviruses endogenously within their genomes, referred to herein as “vERVs” (Fig 4). We classified the vERVs based on where annotated vertebrate retroviruses sit and found that each retrovirus strongly forms its own clade with corresponding vERVs, supported by bootstrap values greater than 95 (Fig 4A). The exceptions are Foamy, Delta and Lentiviruses, which do not form clades with vERVs, and rather form solitary clades on their own. Regardless, the classification and phylogeny of each vertebrate retroviral clade agrees with previous analysis of vertebrate retroviruses [18,44] (Fig 4B). In addition to existing retrovirus classification, we identify new clades of ERVs that sit basal to known retroviral clades (labelled as vERV 1-4) (Fig 4A). Furthermore, many vERVs that sit in the same large clade as gamma retroviruses are highly widespread and divergent, forming their own clades without known exogenous virus relatives, namely, the clades in vertebrate subgroups V5, V6, V7 and V8 (Fig 4A and S7D). Importantly, most vERV clades are found in a wide range of hosts, though they mostly cluster according to the host phylogeny. For example, gamma vERVs in subgroups V4 to V7, and V8 exclusively consist of mammalian and fish elements, respectively (Fig S7D). When multiple vertebrate classes are represented in the same subgroups of gamma vERVs, elements from evolutionarily related hosts tend to cluster together. For example, subgroup V6 consists of bird and reptile vERVs, while Amphibian and retile elements cluster together in subgroups V5 and V7 (Fig

S7D). Like errantiviruses, this is indicative of long-term vertical transfer. Furthermore, many of the elements identified carry intact LTRs, as well as tRNA PBS, further supporting the intactness of the elements. Though the tRNAs are highly varied and divergent, like those found in “ancient errantiviruses” (Fig S7E). Additionally, many of these elements have multiple intact copies in their respective genome suggesting recent transpositions (Fig 4D).

We further characterised the vERVs that we identified in this study, by analysing the bridge region of *pol*. As previously mentioned, vertebrate retroviruses have duplicated the RNase H domain, and the N-terminal copy, while partially degraded, maintains the same structure and 3D fold arrangement, referred to as “tether” [43]. Our analysis similarly finds that all vERVs carry both RNase H and tether, however, many of the elements additionally carry the “mini” domain identified in “ancient errantiviruses” and yeast Ty3 elements (summarised in Fig 4A-B, and an example of vERV “mini” RNase H domains shown in Fig 4F). This strongly supports the idea that many of these vERVs are highly ancient, and potentially share a common origin with Ty3 elements. One area of note, however, is that some vERVs in gamma retroviruses that do not carry the “mini” domain and form their own discrete clade with KoRV-A, potentially indicating that they have emerged more recently (Fig S7D).

We finally characterised the *env* ORF carried by these vERVs. Most carry Glycoprotein 1 (GP1) and Glycoprotein 2 (GP2), resembling *env* structures from other retroviruses as well as Ebolaviruses [45-47]. All elements analysed carry intact GP2, however we found structural variations within the GP1 domain in vERVs, which we analysed using Alphafold 2. Mainly, we find that GP1 from vERVs either carry a receptor binding domain (RBD) and a C domain, or they lack the RBD entirely (Fig 4A-C). All have a highly conserved Furin cleavage site upstream of GP2, and most carry two Cysteine residues upstream of the C domain (Fig 4C and S8). Additionally, we found other highly conserved Cysteine residues that likely form Cysteine bridges which are shared between groups of elements (Fig S8). Interestingly, some of the vERVs instead carry GP1 that resembles the structure of Syncytin-2 (Fig S7C-D). Notably, Syncytin-2 like GP1 domains are not restricted to a single order, being found in reptiles and mammals (Fig S7D). These highly consistent structures suggest that *env* in vERVs that we identified has not undergone degradation, and its function has remained intact as well. The presence or absence of RBD in these elements is mostly consistent with the formations of the clades, though some large clades are mixed in representation (Fig 4A-B and S7D). There is some ambiguity in these mixed clades, as it may suggest that the RBD has been lost frequently in different clades over time. However, even within those mixed clades, elements with only the C domain and elements with both RBD and C domains tend to form their own distinct clades. For example, within vertebrate group V5, elements with RBD and C form their own clade separate to elements with C only, suggesting their relatively ancient origin (Fig S7D). Vertebrate group V8 shows a greater degree of mixing suggestive of more recent recombination, though the outliers have larger branch lengths, and are not supported by bootstrap values greater than 95, indicating greater divergence within *pol* (Fig S7D).

Additionally, Cysteine bridges that are formed within GP1 are conserved between elements from the same vertebrate classes. For example, of the three distinct Cysteine bridges within the C domain of GP1 structures that only have the C domain, one of them (marked as "1" in Fig S8) are predominantly present in Amphibian and fish vERVs while no bird or reptile elements carry it. Similarly, one Cysteine bridge in the C domain within the RBD + C group (marked as "4" in Fig S8) is mainly carried by mammalian, bird and reptile vERVs while no fish or Amphibian vERVs carry it. These observations further suggest that the *env* genes in vERVs have stayed with *pol*, and that recombination events only occurred between closely related species or occurred anciently.

Lokivirus and SNRV vERVs do not carry GP1 and GP2 but carry *env* that resembles glycoprotein F instead. They are found in more ancient vertebrate classes, including reptiles, amphibians, fish and jawless fish, sit distantly in a sister clade to ERV-L, and carry RNase H, tether and a "mini" domain in their bridge region, highly suggestive of an ancient origin (Fig 4A-B). The ectodomain of the F glycoprotein carried by these vERVs retains the highly structured domains found in errantivirus F glycoproteins, as well as the highly conserved Furin cleavage site (Fig 4F and S5), suggesting that its function is intact.

We assembled a *pol* RT tree to include both errantiviruses and vERVs, and find that vERVs, "ancient errantiviruses" and "insect errantiviruses" are clearly distinct from one another. However, Lokivirus vERVs sit as a sister clade to ancient errantiviruses (Fig 4E-F). The commonality in features, such as the classification of *env* and the "mini" domain present in the bridge region of *pol*, is highly suggestive of a common origin between the two. Furthermore, phylogenetic analysis of the *env* F ectodomain clusters vertebrate *env* F with that of ancient errantiviruses, specifically that of tunicate *env* F (Fig S4C), strongly suggesting that the origin of *env* in these retroelements is highly ancient, and potentially shared between ancient animals and vertebrates.

Discussion

Endogenous Ty3 retroelements carrying *env* genes were only previously identified in Dipteran genomes. However, we found that genomes from nearly every major order carry such elements, and they are fully intact with uninterrupted ORFs and LTRs. Additionally, many of the elements annotated have multiple intact copies in their respective genomes, indicating that they have been transposing recently. The intactness of these elements is further supported by structural analysis of the *env* glycoproteins, where we found that highly conserved structural domains are still retained, despite divergences in the primary sequence.

We found that these errantiviruses are likely to be highly ancient, being found in ancient animals such as Cnidaria, Ctenophora and Tunicata. Phylogenetically, they are found within an "ancient errantivirus" group that sits separately to elements found in insects. These ancient elements are also structurally distinct, with the bridge region between INT and RT of *pol* resembling that of yeast Ty3 retrotransposons, rather than that of insects. Furthermore, the divergence of *env* genes appear to follow that of *pol*, with few exceptions, indicating that the *env* genes are similarly ancient, and that recombination events are likely rarely

occurring. Clades on the *pol* tree also tend to consist of elements from the same, or closely related hosts, rather than consist of elements within highly divergent host species, further supporting their ancient nature, and indicating that they have mainly been vertically transmitting over time.

Finally, we found that *env*-containing endogenous retroviruses are similarly found in vertebrate genomes, and the same observations made in errantiviruses mostly apply to these elements as well. The recently identified vertebrate Lokiviruses sit as a sister clade to "ancient errantiviruses" and shares structural features with them, suggesting their common origins.

It was originally speculated that errantiviruses acquired *env* from baculoviruses in insects, some of which carry an F-type ENV protein Ld130, potentially via recombination events [23,24]. While Ty3 elements may have obtained the *env* gene via recombination, it is unlikely that baculoviruses are the source for the *env* gene found in errantiviruses, as they are far more widespread than insects. The congruence of phylogeny between *env* and *pol* indicate that they have likely co-evolved (Fig S4). This, alongside the presence of errantiviruses in ancient host orders such as Cnidaria, Ctenophora and Tunicata, suggests that *env*-carrying Ty3 elements are highly ancient transposons, rather than the result of a recent recombination event with an exogenous virus. HSV/gB-type *env* in errantiviruses appears to be equally as ancient, as it is similarly found in ancient animal orders (Fig 1). The arrangement of Cysteine bridges that stabilize the structure of the ectodomain, are consistent between elements within the same order, even though *pol* presents itself separately on the tree (Fig 2 and S6). This implies that the *env* has diverged with the host, but likely undergoes recombination events with other elements within the same species. Though, *pol* with HSV/gB-type *env* mostly cluster within their own clades, and are not mixed with *pol* carrying glycoprotein F, indicating that recombination events are not happening frequently enough to randomise their spread, and likely occurred anciently. The origins and acquisition, therefore, of either type of *env* protein, likely occurred in the last common ancestor of metazoans.

Finding intact, multicopy elements in such a widespread manner brings into question the origins of these unique retrotransposons. One explanation could be that errantiviruses frequently undergo horizontal transfer, with shorter periods of vertical transmission. Drosophila errantiviruses have been previously suggested to horizontally transfer between species [29,48]. It has also been suggested that some LTR retrotransposons in bivalves can horizontally transfer between different species and can even jump across phyla as well [49]. While horizontal transfer of errantiviruses between distantly related species cannot be excluded, it is likely to be a rare occurrence, given that *pol* from the same order tend to form their own discrete clades (Fig S3 and S7). In instances where clades are formed from different phyla of hosts, they are magnitudes of order apart from each other, and it is unlikely that they can easily transfer. Additionally, many of these clades tend to have much longer branches, and it is therefore possible that they are still distantly related and diverged a long time ago. For example, within tree position A10 (Fig S3), a clade consisting of elements from Hexanauplia and Annelida are separated by longer branch lengths. The reason for these types of clades, and as to why we do not see elements in species between those distantly related orders, is likely due to the frequent loss of these retrotransposons from the genome. In such cases, distantly

related species have seen such retrotransposon copies being active, which are vertically inherited, and diverge along with the genome of the host species, while other species do not see this activity, and those inactive copies would be degraded over time.

In Drosophila, errantiviruses occupy the somatic cell niche in the ovaries that surround the germline, and the expression of *env* allows the Ty3 virus-like particles to infect the germline to create new copies [21]. As an adaptation, Drosophila species have duplicated the germline transposon defence mechanism to target errantiviruses in the somatic cells of the gonad [50]. It is possible that this is more universal, and that *env*-containing retrotransposons occupy the cell niches surrounding the germline, and that expression of *env* allows them to continue to infect the germline. For example, murine IAP retrotransposons that carry *env* genes also remain infectious, like Drosophila errantiviruses [51]. However, there are evolutionary costs associated with keeping additional ORFs, and genes in endogenous retroviral elements, as they are constantly subject to evolutionary pressure from their hosts [52,53]. It is therefore significant that the *env* gene is as highly conserved as we have observed, and that it does not seem to have been subject to frequent loss and acquisition. Though, the mechanisms of transmission via the action of this *env* gene remain elusive for most of the elements we identified, except for those in Drosophila species.

However, the structures formed by *env*, for both HSV/gB and glycoprotein F, form the same functional structures as those found in exogenous viruses, required for cell entry [36,39,40]. While entry of HSV also requires three other glycoproteins, gD, gH and gL, gB is specifically required for membrane fusion [54,55]. The mechanism of cell entry, therefore, is unknown for these errantiviruses that carry gB, and necessitates further investigation regarding their expression. Though, given their highly structured configuration, it is likely that they retain membrane fusion capabilities. Additionally, given the widespread nature of these *env*-containing elements, it may also be possible that somatic transposon defence mechanisms against them are more common.

We found that vERVs are similarly as widespread as errantiviruses, found ubiquitously in vertebrate classes. Many of the vERVs that we have identified in this study tend to form clades with known exogenous and endogenous retroviruses, and their phylogeny is consistent with previous studies [18,44]. However, many vERVs form their own clades, without known/annotated retroviruses. Interestingly, most of the vERVs have multiple intact copies in their respective genomes, meaning that they have been recently transposing as well (Fig 4).

Of the known classifications, only vERVs resembling Epsilon viruses or SNRVs are found specific to one vertebrate order, while the rest are found across multiple. This may be an indication that these vERVs have existed for long periods of time, remaining in the germline, and may not be derived from their exogenous counterparts. This is supported by the observation that vERVs found within the same vertebrate classes tend to cluster together, even within larger clades that have a mix of representation from different classes (Fig S7). Additionally, structural features of the vERVs tend to be different than those of the known counterparts. For example, MuERV-L, is not known to carry *env* genes [56,57], but vERVs in the same discrete clade do (Fig 4). Furthermore, these elements are found in reptiles, mammals and birds (Fig S7).

This magnitude of difference between the orders indicate that it is unlikely that they are sourced from the same exogenous virus. Additionally, the bridge region of most vERVs carries an RNase H domain, a tether domain, as well as a "mini" domain. The "mini" domain is not shared by many of the annotated retrovirus in this study (Fig 4). Though, the same "mini" domain is found in yeast Ty3, as well as "ancient errantiviruses", potentially supporting their common origin. Some vERVs do not have the "mini" domain either, but they are restricted to discrete clades. An example of this is a discrete clade of vERVs that resemble the Koala endogenous retrovirus; KoRV-A (Fig S7).

Evidence suggests that KoRV-A is the result of a recent endogenization event of the exogenous retrovirus KoRV-B [5,6]. Though, we find that KoRV-A forms a clade with vERVs found in Rodentia, Pilosa, Chiroptera and Perissodactyla, and not marsupials, it is still possible that they are the result of recent exogenous integrations. However, recent integrations of exogenous viruses cannot solely be responsible for all the genomic copies identified in this study, given the multicopy nature of these elements, and there is no evidence of exogenous counterparts to many of the clades of vERVs (Fig 4). It is possible that elements can be shared via horizontal transfer, potentially mediated by exogenous viruses. However, given the separation of *pol* clades between different classes of vertebrates, such horizontal transfer is either limited between closely related species, or does not happen frequently enough to randomize the distribution of *pol*. The relative frequencies of endogenization and germline expansion, and how they contribute to the evolution of vERVs remains as an open question. Identification of more exogenous retroviruses and better understanding of how vERVs make new germline copies will help address this fundamental question in the future.

As seen in errantiviruses with HSV/gB-type ENV, structural features of GP1 vary within large clades of the vERV *pol* RT tree (Fig 4 and S7). But they are still consistent in smaller distinct clades, indicating that *pol* and *env* are co-evolving in vERVs. For example, some discrete clades of gamma retroviruses carry GP1 that keeps both RBD and the C domain, and others only carry the C domain (Fig S7). Additionally, some vERVs carry GP1 that resembles syncytin-2. This is the first known characterisation of such intact endogenous viral copies, and most do not cluster with known exogenous retroviruses either. Notably, vERVs carrying Syncytin-2 are found in different vertebrate orders, potentially indicating its ancient origin. Importantly, GP1 and GP2 are structurally intact in most vERVs, and likely continue to be functional as membrane fusion proteins.

The clade of newly identified vERVs resembling SNRVs and Lokiviruses do not carry GP1/GP2, but instead carry glycoprotein F, consistent with previous studies [18] (Fig 4). Additionally, this clade carries the "mini" domain, which appears to be a hallmark of an ancient endogenous element (Fig 4). Lokivirus-like vERVs are also found in more ancient orders of vertebrates, Fish, Amphibians and Jawless fish. The structural features of Lokiviruses are shared with "ancient errantiviruses", and they indeed appear to be closely related to them, based on the phylogeny of their ENV as well as POL proteins (Fig S4). These findings strongly indicate that vERVs and Errantiviruses have a common origin which is highly ancient.

Our study sheds light on the potential origins of *env* containing endogenous retroelements. We propose that these *env* containing retrotransposons are highly ancient in nature, and in cases where they remain active, they diverge with their hosts, infrequently recombining with other elements within the same hosts. While the origin of the *env* genes is still not entirely known, we suggest that both HSV/gB-type and F-type *env* were anciently acquired before the split of bilaterian and non-bilaterian animals in Pre-Cambrian era, while GP1+GP2 *env* were later acquired in the common ancestor of vertebrates.

Materials and methods

Genome assemblies used in the study

Reference genome sequences were retrieved from NCBI and grouped into the following insect orders, vertebrate classes and other metazoan phyla: Diptera, Lepidoptera, Coleoptera, Hymenoptera, Hemiptera, Polyneoptera, and Palaeoptera for insects; jawless-fish, fish, Amphibian, reptile, bird, and mammal for vertebrates; Arachnida, Branchiopoda, Chilopoda, Collembola, Diplopoda, Hexanauplia, Malacostraca, Merostomata, Ostracoda, and Thecostraca for Arthropoda; Acanthocephala, Annelida, Brachiopoda, Bryozoa, Dicyemida, Mollusca, Nemertea, Onychophora, Orthonectida, Phoronida, Platyhelminthes, Priapulida, Rotifera, and Tardigrada for Protostomia; Cephalochordata, Echinodermata, Hemichordata, and Tunicata for Deuterostomia; Cnidaria, Ctenophora, and Porifera. We obtained lists of reference genomes in each category in May 2023 using the Entrez Direct command line tool except for the vertebrate genomes which we used the list from the previous work by Wang et al. [18]. The list of the genome assemblies used can be found in Supplementary table S4.

Identification of active genomic copies of vertebrate endogenous retroviruses (vERVs) and Ty3/*gypsy* errantiviruses

We ran a total of five iterations of tBLASTn (ncbi-blast-2.9.0) to search for the genomic copies of vERVs and Ty3/*gypsy* errantiviruses, firstly using the baits of known elements, and subsequently using the elements that were identified after each iteration. We chose new baits that were not represented in the phylogenetic tree of the POL (see below for the detail) from the previous iteration and repeated the process until we no longer found new elements. We used two sets of sequences for the first tBLAST search. The first set included GAG, POL and ENV protein sequences of the Dipteran *gypsy* errantiviruses from the RepBase (https://www.girinst.org/repbase/). The second set included GAG, POL and ENV protein sequences of selected vertebrate retroviruses that were used in a previous study [18]. The bait sequences for all tBLASTn iterations can be found at https://github.com/RippeiHayashi/errantiviruses_ERVs_2024/tree/main. We applied different criteria to identify ERVs from species outside insects and vertebrates as we did not have prior knowledge of whether or what kind of ERVs are found in those species. We briefly summarise our approach below.

We first used open reading frames (ORFs) of vertebrate retroviruses as baits for the vertebrate genomes and those of the Dipteran errantiviruses for the other genomes to run tBLASTn with default parameters (-evalue 10). Genomic regions that showed homology to the POL baits with a bit score of greater than 50 or

to the GAG and ENV baits with a bit score of greater than 30 were individually pooled. Genomic regions that had hits from multiple baits were merged for their coordinates using bedtools/2.28.0. Regions that were larger than 1500 nucleotides for POL and 300 nucleotides for GAG or ENV were collected. Following, for insect genomes we looked for regions that carry GAG, POL, ENV sequences that are next to each other in this order allowing up to 500 nucleotides of gaps between them as well as excluding regions of smaller than 4000 nucleotides or larger than 10000 nucleotides. For the other genomes we allowed up to 1000 nucleotides of gaps between POL and ENV and collected regions that had both ORFs that are not smaller than 3000 nucleotides or larger than 10000 nucleotides.

We subsequently checked whether the regions contain uninterrupted ORFs by making translated sequences using transeq from EMBOSS-6.6.0 and excluding ORFs with more than five stop codons. We clustered all the identified GAG, POL and ENV from each genome by Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) and identified genomic regions that harbour all three ORFs that are multiplied in the genome. We also searched for ORFs regardless of the copy number when there were no multiplied elements in the genome. When genomic copies were abundantly found in multiple closely related species, we only annotated the elements from representative genomes. Genomes that were annotated and the number of predicted ORFs per genome are summarised in Supplementary table S4.

The above process yielded abundant putative intact copies of ERVs from insect genomes whereas we took additional steps for the other genomes. For vertebrate genomes, we extended the genomic regions of predicted POL-ENV to upstream by 2000 nucleotides and manually annotated putative GAG ORFs using the online sequence viewer Benchling (https://www.benchling.com/). The identity of GAG ORFs were further confirmed by HHpred (https://toolkit.tuebingen.mpg.de/tools/hhpred). For Arthropoda genomes outside insects as well as other Protostomia genomes, we extended the genomic regions of predicted GAG-POL to downstream by 3000 nucleotides and manually annotated ENV ORFs using Benchling. For Cnidaria, Ctenophora and Deuterostomia genomes outside vertebrates we used transeq to annotate putative GAG and ENV ORFs from 5000 nucleotides downstream and 3000 nucleotides upstream to the predicted POL ORFs, respectively.

We took confident 1128 POL ORFs from the first tBLASTn search, including both invertebrate and vertebrate elements, and made a multiple sequence alignment of the conserved regions of the Integrase domain using mafft-7.505 with the option '--auto', and inferred the phylogenetic relationships using iqtree-1.6.12 with the option '-m rtREV+R4 -bb 1000' (see more details in the section below). We took 178 representative POL from the phylogenetic tree and used the POL and the corresponding GAG and ENV ORFs from the same genomic regions for a second tBLASTn search. We repeated the above steps after combining the results of the two tBLASTn searches. We did not find any ERVs in Nematoda genomes that have GAG, POL and ENV in this order while the previous study reported non-canonical Nemetoda ERVs that have flipped the order of ENV and POL [27]. Thus, we searched for regions that have ENV and POL in this order allowing gaps of up to 2000 nucleotides between them. GAG was subsequently predicted from 5000 nucleotides upstream to the predicted ENV-POL. We extended the search for potential GAG-ENV-

POL insertions to other non-insect genomes and found similar elements in Bryozoa and Platyhelminthes. We also found ENV-GAG-POL elements in Arachnida genomes (*Medioppia subpectinata* and *Oppia nitens*). These elements were recovered through the search of ENV-POL elements because GAG ORF was short. We did not extend our search to find other arrangements of ORFs.

Additionally, we found that the glycoprotein gene present in mosquito *BEL/pao* retrotransposons [25] are also found next to *gypsy* POL genes. The glycoprotein gene resembles the *envelope* genes from Chuviridae and is structurally similar to glycoprotein B from Herpesviruses. We identified ten mosquito *BEL/pao* retrotransposons that carry this glycoprotein from RepBase (BEL-199_AA, BEL-279_AA, BEL-652_AA, BEL-801_AA, BEL-804_AA, BEL-879_AA, BEL-51_AnFu, BEL-56_AnFu, BEL-81_AnFu, and BEL-71_CQ) and used them as baits for a tBLASTn search against Arthropoda genomes. The tBLASTn search identified 114 *gypsy* elements with homologous glycoprotein genes from Coleoptera, Hymenoptera, Polyneoptera, and Hexaneuplia species. Combined together, the second iteration of tBLASTn search identified 491 new elements. The phylogenetic analysis of POL from the elements identified in the first and second tBLASTn runs revealed clades that only contained new elements from the second tBLASTn run, suggesting that the first two tBLASTn runs were not comprehensive. Therefore, we took 42 additional baits to cover all the major clades in the phylogenetic tree for a third tBLASTn search, which identified 87 new elements across metazoan phyla. We repeated tBLASTn runs five times in total; the fourth and fifth runs used 21 and 7 new baits and identified 53 and 6 new elements, respectively. We stopped the tBLASTn search after five iterations because the phylogenetic analysis of POL did not reveal new clades.

Following, we further curated newly identified ERVs. Firstly, we used hmmer 3.3.1 (http://hmmer.org/) to make sure that the ERVs that we identified are classified as Ty3/*gypsy*. We compared POL ORFs from the newly identified ERVs against the POL ORFs of known retroelements that are available from the Gypsy Database 2.0 (https://gydb.org/index.php/Main_Page). The Gypsy Database 2.0 collects hmm profiles of Ty1/Copia, Bel/Pao, Ty3/*gypsy*, and Retrovirus POL proteins. The hmmer search classified all the newly identified ERVs into Ty3/*gypsy* or Retroviruses. Secondly, we ran HHpred from hh-suite 3.3.0 [58] or the online tool of HHpred (https://toolkit.tuebingen.mpg.de/tools/hhpred) to predict GAG, ENV, and POL ORFs using the default search parameters. For the command line HHpred, we first made multiple sequence alignments using hhblits and the Uniclust database UniRef30/2022_02 (https://uniclust.mmseqs.com/). This was followed by hhsearch to predict domain structures using the collection of pdb70 from March 2022 (https://wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite_dbs/). We used default settings to run hhblits and hhsearch. ENV and POL ORFs with large truncations were removed after the domain structure prediction. ERVs that predicted all GAG, ENV and POL ORFs were taken for further analyses.

We additionally ran BLASTn using the genomic regions that cover all three ORFs as baits to interrogate the copy number of ERVs in the genome where they are found. The results are summarised in Supplementary table S1 where we counted the number of BLASTn hits with a more than 98% identity in more than 98% of their length to the bait sequences. A majority of newly identified ERVs (1586 out of

1765 elements) carry ORFs that are found at least twice in the genome, further validating our approach of finding active or recently active insertions.

Identification of putative long terminal repeats and transfer RNA (tRNA) primer binding sequences

Long terminal repeats (LTRs) of ERVs were identified as follows. Firstly, we fetched the genomic coordinates of ERV insertions and extended 3000 nucleotides at either ends from the Start codon of GAG and the Stop codon of ENV ORFs (POL was taken for the GAG-ENV-POL ERVs) for all species categories except for Cnidaria, Ctenophora and Deuterostomia where we extended 5000 nucleotides at either ends from the Start and Stop codons of POL. We retrieved the genomic sequence of the extended ERVs and ran BLASTn against the same sequence to identify repeated sequences. We considered repeated regions of greater than 99% identity and the tBLASTn bit score between 200 and 5000 as putative LTRs. We excluded LTRs whose distance between them is shorter than 5000 nucleotides or longer than 14000 nucleotides. Following, we searched for putative tRNA primer binding sequences (PBS). tRNA sequences were retrieved from genomic annotations from representative genomes in respective species categories except for Ctenophora where we took tRNA sequences from Protostomia, Cnidaria and Deuterostomia because there were no annotated genomes available at the time of the study. Representative genomes that we used to retrieve tRNA sequences are: GCF_000001215.4 and GCF_013141755.1 for Diptera, GCF_014905235.1 and GCF_905147365.1 for Lepidoptera, GCF_000002335.3 and GCF_917563875.1 for Coleoptera, GCF_003254395.2 and GCF_016802725.1 for Hymenoptera, GCF_014356525.2 and GCF_020184175.1 for Hemiptera, GCF_002891405.2 and GCF_021461395.2 for Polyneoptera, GCF_921293095.1 for Palaeoptera, GCF_000671375.1, GCF_020631705.1, GCF_002217175.1, GCF_016086655.3, GCF_024679095.1, GCF_019059575.1 for Arthropoda, GCF_000326865.2, GCF_020536995.1, GCF_000002985.6, GCF_000715545.1 for Protostomia, GCF_000222465.1 and GCF_022113875.1 for Cnidaria, GCF_013122585.1, GCF_000003605.2, GCF_000003815.2, and GCF_000002235.5 for Deuterostomia, and GCF_015237465.2, GCF_000001405.40, GCF_017654675.1, GCF_900496995.4, GCF_900700375.1, and GCF_010993605.1 for Vertebrata. The minimum requirements of the priming events are incompletely understood, but the length of PBS for retrotransposons is generally between 8 and 18 nucleotides and the PBS is positioned near the 3' end of the 5' LTR [59]. We chose the 12 nucleotides at the 3' end of tRNAs as putative primer sequences and searched for PBS in the 100 nucleotide-window centred at the 3' end of the 5' LTR using bowtie-1.2.3 allowing up to two mismatches. We added “CCA” to the 3' end of the tRNA sequence when it is not present in the genomic sequences. We successfully identified putative PBSs in 1308 out of 1765 ERVs that we identified in this study. We grouped ERVs based on the type of tRNA used for the PBS and summarised the results in Figures S2, S3 and S7 as well as Supplementary table S1.

Phylogenetic analysis of POL

We obtained multiple sequence alignments of the RT and Integrase domains of POL as follows. For the RT domain, we first took the HHpred alignments of POL against the RT domains of Moloney Murine Leukaemia Virus (Protein Data bank entry 4MH8_A) and Human Immunodeficiency Virus type 1 (Protein

Data bank entry 4G1Q_B), spanning regions that correspond to their fingers, palm, thumb and connection subdomains [60], namely the amino acid positions of 28 to 426 and 7 to 405, respectively. However, this process caused small truncations at either ends of the domain because of poor alignments at the individual substructures. To overcome this issue and to facilitate sequence alignments, we extended the N- and C-termini by 10 and 15 aa, respectively, after merging the two predicted regions of RT domains. We then ran mafft-7.505 to generate a multiple sequence alignment, from which we manually trimmed the either ends to have intact fingers, palm, thumb and connection subdomains for all POL ORFs. Similarly, we took HHpred alignments of POL against the Integrase domains of Human Spumaretrovirus (Protein Data bank entry 3OYM_A) and Simian T-lymphotropic Virus type 1 (Protein Data bank entry 7OUH_B), spanning regions that correspond to their N-terminal domain (NTD) and catalytic core domain (CCD) [61], namely the amino acid positions of 54 to 273 and 8 to 197, respectively. We used the predicted Integrase domains with 15 and 30 aa extensions at the N- and C-termini, respectively, for generating a multiple sequence alignment before manually trimming the ends. We inspected the multiple sequence alignments using Mview (https://www.ebi.ac.uk/Tools/msa/mview/) and removed ERVs that contained truncated RT or Integrase domains from the analyses. We included ORFs from 5 Drosophila errantiviruses (DM176, IDEFIX, ZAM, GYPSY, and Gypsy-5_DGri) and 27 representative vertebrate retroviruses (sequences used in the previous study [18]: BLV, FLV, HTLV-2, KoRV, LDV, MMTV, MuLV, REV, RSV, SRV-1, WDSV, MuERV-L, BFV, EFV, FFV, SFV, HIV-1, EIAV, FIV, SnRV, Loki-Ofa, Loki-Pma, Loki-Sri, Loki-Str, and Loki-Xtr; these two sequences were additionally used in this study: epsilonretrovirus, WEHV1 (AAD30048.1) and avian gammaretrovirus, DIAV (AGV92859.1)) in the multiple sequence alignments.

We ran iqtree-1.6.12 with the option '-m rtREV+R4 -bb 1000' to generate the phylogenetic trees of all Integrase and RT domains, and those from ERVs found in vertebrate and invertebrate genomes, respectively. We then selected representative elements from clades, taking into consideration the relative distances between nodes in the same clade. For example, if the distance between two nodes is less than 0.1, meaning that there is little – no variations between their sequences, only one node in that position is taken. Additionally, we made sure that the representatives captured the diversity of *env* found within the clade. For example, if a clade contains both F and HSV elements, at least one for both is taken, regardless of distance. Furthermore, the phylogeny of the hosts was represented as fully as possible. For example, if a clade has elements from Diptera and Coleoptera, at least one from both is taken regardless of distance. Once representatives were collected, the tree was then accordingly pruned. It should be noted that the final pruned tree of representatives was only used as a figure aid for clarity (Fig 1-4), and all analyses were done on the full INT and RT phylogenetic trees. We used a similar approach as detailed to select representatives for structural analysis of *env* ectodomains, as well as the *pol* bridge region (Table S2). If any outliers were found, remaining elements from the same clade were also analysed, to check whether structural inconsistencies were consistent across the clade.

For congruence plots in figures S2, S4 and S7, we grouped elements based on which node they sit in the tree, ensuring that the node is well supported by bootstrap values greater than 95. We used the networkD3 package in R to generate Sankey plots to visualise the representation of elements from one tree in another. The node definitions are summarized in Table S3.

Further, we took 88 representatives of the F-type ENV based on their positions in the corresponding POL phylogenetic tree for the analysis. We identified the signal peptides and transmembrane regions by DeepTMHMM (https://dtu.biolib.com/DeepTMHMM) and obtained the ectodomains that are in between them. We subsequently ran mafft-7.505 to generate a multiple sequence alignment.

Structure prediction of ENV proteins and the RNase H domains using Alphafold 2

We took 88, 32, 130, and 243 representatives of the F-type, HSV/gB-type, and GP2-type ENV proteins and the region of POL between INT and RT, that we termed the bridge region, using branch definitions from the phylogenetic trees of *pol* RT. We used the colabfold wrap v1.4.0 of Alphafold 2 (https://github.com/YoshitakaMo/localcolabfold/releases/tag/v1.4.0), and used MMSeqs2 for the database search and clustering of protein sequences before performing the structural prediction [62]. We used the whole ectodomains (see above) of the F-type and HSV/gB-type ENV proteins and predicted the trimer structures using the multimer mode of Alphafold 2 (AlphaFold2-multimer-v2). For the GP2-type ENV proteins, we first defined the ectodomain that is in between the predicted signal peptides and transmembrane regions and identified the GP1 domain between the signal peptides and the putative Furin cleavage site. We predicted monomer structures of the GP1 domains and the bridge regions. We then visualised and characterised visible structures and disulfide bonds or Cysteine bridges using Maestro (Schrödinger Release 2024-3) and grouped them by common substructures (Table S2).

Figure Legends

Figure 1 – Phylogeny and copy number of invertebrate metazoan errantiviruses. (A) Open reading frame arrangement of intact errantivirus copies. **(B)** Summary of Errantiviruses found in Metazoan phyla with type of *env* protein indicated. **(C)** Maximum likelihood tree constructed on representative *pol* RT proteins of intact errantiviruses, with “ancient errantiviruses” and “insect errantiviruses” phylogenetic groups highlighted in green and white respectively. Positions of known dipteran errantiviruses, Gypsy_ZAM (1), Gypsy_Gypsy5_Dgri (2), Gypsy_Gypsy (3), Gypsy_Idefix (4), and Gypsy_DM176 (5) are indicated. Host classifications of errantiviruses are annotated in colour, non-arthropod Protostomia (yellow), Deuterostomia (red), Diploblasta (orange), non-insect Arthropoda (green), Insecta (blue). One cnidarian element within the group of insect errantiviruses is marked by an asterisk. **(D)** Copy number of intact errantivirus copies in metazoans. Column represents a genome assembly, each tile representative of annotated errantivirus, colour of tile represents copy number.

Figure 2 – Representation of *env* in errantiviruses, and structures of errantivirus glycoprotein B. (A) Shown is a *pol* RT maximum likelihood tree of the representative errantiviruses (see methods) with host (*) and type of *env* protein (**) indicated. Colours represent non-insect animals (*) (yellow), insects (*)

(blue), glycoprotein F (F) (**) (orange), and glycoprotein B (HSV/gB) (**) (blue). Clades containing hymenopteran elements in “Insect errantiviruses” that carry either F or HSV/gB *env* are annotated as 1 and 2. **(A')** Close up of clades 1 and 2 in “Insect errantiviruses”. F-type and HSV/gB-type errantiviruses that are both found in *Lasioglossum lativentre* are marked by asterisks. **(B)** Structural arrangements of Ectodomains, Transmembrane domains and signal peptides found in HSV/gB-type errantiviruses. **(B')** Shown are the arrangement of domains of the HSV/gB ectodomain, and Alphafold 2-predicted 3D monomer and trimer structures of the ectodomain. The structure is taken from Hemiptera_Pemphigidae_Elan_errantivirus_1. **(C)** Fusion loops and conserved Cysteine bridges are indicated on structure of the errantivirus HSV/gB ectodomain. **(C')** Shown are variable Cysteine bridges in the errantivirus HSV/gB ectodomains. Cysteine bridge configurations of domains I and II, with proximity of Cysteine bridge to fusion loops of domain I are indicated. Structures are taken from Cnidaria_Hydrozoa_Chem_errantivirus_5, Bryozoa_Gymnolaemata_Cpal_errantivirus_2, and Tunicata_Ascidiacea_Clep_errantivirus_1, for C1, C2 and C3 configurations, respectively. **(D)** Shown is the summary of the Cysteine bridge configurations found in HSV/gB of different metazoan orders. Exact positions of the Cysteine bridges can be found in Supplementary Figure S6.

Figure 3 – RNase H domains found in the bridge region of errantiviruses. (A) Shown is the *pol* RT maximum likelihood tree of representative errantiviruses, with RNase H domain architecture in bridge region of *pol* highlighted. “Insect errantiviruses” with RNase H + pseudo RNase H (blue), “insect errantiviruses” with bridge region outliers (dark blue), “ancient errantiviruses” with RNase H + mini domain (green), “ancient errantiviruses” with bridge region outliers (yellow). Insect and non-insect elements are annotated as blue and yellow, respectively. **(A')** RNase H domain configurations found in bridge region of “Insect errantiviruses” and “Ancient errantiviruses”, with outliers indicated, and the structure of yeast Schizosaccharomyces *pombe* Ty3 bridge region predicted by Alphafold 2 is shown. **(B)** Shown are Alphafold 2-predicted structures of domains found in the bridge regions of errantiviruses, with topology maps of α helices (grey) and β sheets (red). Structures are taken from Annelida_Polychaeta_Apac_errantivirus_7, Bryozoa_Gymnolaemata_Cpal_errantivirus_3, Coleoptera_Coccinellidae_Npum_errantivirus_1, Annelida_Polychaeta_Slim_errantivirus_9, Diptera_Cecidomyiidae_Smos_errantivirus_1 and Hymenoptera_Apoidea_Obic_errantivirus_2 for RNase H/Tether, ‘mini’ RNase H, ‘pseudo’ RNase H domains, “Mini” domain 2, “Pseudo” domain 2 and “Pseudo” domain 3, respectively.

Figure 4 – Widespread presence of *env*-containing endogenous retroviruses in Vertebrate classes.

(A) Maximum likelihood tree of *pol* RT from intact *env*-containing endogenous retroviruses (vERVs) in vertebrate classes. Classification of clades highlighted on tree, and detailed guide to clades in panel B. Host phylogeny annotated (*); mammals (orange), amphibians (blue), birds (pink), fish (green), reptiles (yellow), jawless fish (purple), known exogenous retroviruses (white). Bridge region structures annotated (**); RNase H + RNase H/Tether (orange), RNase H + RNase H/Tether + mini domain (blue). ENV structure and type annotated (***); GP1 with receptor binding domain (RBD) and C domain (yellow), GP1

with C domain only (blue), Mix of GP1 with RBD+C and GP1 with C only (teal), Glycoprotein F (grey). **(B)** Summary of vERV classification, with colours annotated to match tree classification in panel A. The bridge region architecture, *env* structure, and host phylogeny of clades are shown. **(C)** Structure of Glycoprotein 1 (GP1) domain of *env* with Alphafold 2-predicted structure of the receptor binding domain (RBD), C domain, and conserved Cysteine residues highlighted. Shown are the GP1 structures of Bird_Passeriformes_Nchr_vERV_2 (C only) and Amphibian_Anura_Rmar_vERV_3 (RBD + C). **(D)** Genomic copy number of vERVs. Columns represent individual genome assemblies grouped by vertebrate classes, each tile representative of annotated vERV, and colour of tile represents copy number. **(E)** *pol* RT maximum likelihood tree of representative errantiviruses and vERVs, with "insect errantiviruses" (white), "ancient errantiviruses" (green), "lokiviruses" (magenta) and "Vertebrate ERVs" (indigo) highlighted, with *env* types and bridge region structures (red – RNase H, Blue – mini domain, green – pseudo RNase H) indicated below. Host phylogeny indicated on tree, Insects (blue), non-insect Arthropoda (green), non-arthropod Protostomia (yellow), Diploblasta (orange), non-vertebrate Deuterostomia (red), vertebrates (purple). **(F)** Positional relationship of Lokiviruses to Errantiviruses, with Alphafold 2-predicted 3D structures of the bridge region and *env* ectodomain. Both structures are taken from fish_Cyprinodontiformes_Fhet_vERV_1. **(G)** Alphafold 2-predicted 3D structures of bridge region domains in known vertebrate retroviruses. Shown are structures from MMTV, FLV and SNRV.

Supplementary Figure S1 – Identification of errantiviruses and vERVs in metazoan genomes. **(A)** Computational approach to identify and annotate *env*-containing retroelements in metazoan genomes. **(B)** A pruned maximum likelihood tree of *pol* RT from both errantiviruses and vERVs with protein baits for consecutive tBLASTn searches annotated in colour, showing a comprehensive coverage of the baits in the tree. **(C)** *pol* RT maximum likelihood trees with representative elements taken for Alphafold 2 analysis highlighted in red for bridge domains (All RT), F *env* ectodomains (errantiviruses RT only), HSV/gB *env* ectodomains (HSV/gB-type RT only) and GP1 (vERVs RT only).

Supplementary Figure S2 – Complete errantivirus tree and annotations of tRNA primer binding sites of insect errantiviruses. (A) Complete Maximum likelihood tree constructed on *pol* RT proteins of intact errantiviruses, with discrete phylogenetic groups within "insect errantiviruses" indicated (I1 - I6). "ancient errantivirus" group highlighted in green, "insect errantivirus group highlighted in white. Host phylogeny (*) annotated for insects (blue), non-insect arthropods (green), non-arthropod Protostomia (yellow), Deuterostomia (Red) and Diploblasta (Orange). ENV type (**) annotated for Glycoprotein F (F) (orange), and glycoprotein B (HSV/gB) (blue). **(B)** A Sankey plot of *pol* RT and *pol* INT maximum likelihood trees for errantiviruses, showing an overall phylogenetic congruence between them. **(C)** Shown are annotations of tRNA primer binding sites of "insect errantiviruses" within phylogenetic groups of *pol* RT, demonstrating that the types of tRNAs used reflects the separation of *pol* RT. Number of elements are indicated on Y axis. Sequence information of the tRNA PBS can be found in Table S1.

Supplementary Figure S3 – Ancient errantivirus clades. **(A)** Maximum likelihood tree of *pol* RT for the "ancient errantivirus" group in isolation, with host phylogeny and *env* type, and discrete phylogenetic

groups (A1-A11) indicated. Host phylogeny (*) is annotated for insects (blue), non-insect arthropods (green), non-arthropod Protostomia (yellow), Deuterostomia (Red) and Diploblasta (Orange). ENV type (**) annotated for Glycoprotein F (F) (orange), and glycoprotein B (HSV) (blue). **(B)** Detailed view of phylogenetic groupings in "ancient errantiviruses" with host (*) and *env* (**) annotated. Bryozoa elements that are found in distinct clades of the tree are highlighted. Each element is named for the host species and simplified to Phylum/Class/Order, Superfamily, Genus and Species. For example, elements found in Diptera Drosophilidae *Drosophila melanogaster* are named as Diptera Drosophilidae Dmel. **(C)** tRNA annotations of "ancient errantiviruses" within phylogenetic groups of *pol* RT. Insect orders in the groups A1, A2 and A7 are marked by hash. Number of elements indicated on Y axis.

Supplementary Figure S4 – Glycoprotein F structures and Phylogeny. (A) Architecture of glycoprotein F ectodomain in Errantiviruses. **(B)** A Sankey plot of *pol* RT and glycoprotein F *env* ectodomain maximum likelihood trees for representative errantiviruses, showing a phylogenetic congruence between them. "ancient errantivirus" group is highlighted in green, and vERVs are highlighted in purple. **(C)** Detailed view of maximum likelihood trees of *pol* RT and *env* F ectodomains of representative elements for errantiviruses and vERVs. Host species annotated for insects (blue), vertebrates (purple), and other animals (yellow). **(D)** Alphafold 2-prediction of glycoprotein F ectodomains. Monomeric and trimeric structures of the "pre-fusion configuration" and the "post fusion configuration" are shown in **(D')** and **(D'')**, respectively. Schematic cartoons of *env* F ectodomain domain 4 are shown in **(D''')**. Structures are taken from Diptera_Sciomyzidae_Cmar_errantivirus_19 and Branchipoda_Diplostraca_Dcar_errantivirus_3 for "pre fusion" and "post fusion" structures in **(D')** and **(D'')**, respectively.

Supplementary Figure S5 – Conserved predicted Furin cleavage sites and hydrophobic fusion peptides of glycoprotein F. Shown is a multiple sequence alignment of the middle part of the representative ectodomains of glycoprotein F. Arginine (R) residues in the putative Furin cleavage site and hydrophobic residues (Leu, Ile, Val, Phe, and Trp) in the fusion peptides are highlighted in red and green, respectively. Furin cleavage sites were predicted using ProP-1.0 (https://services.healthtech.dtu.dk/services/ProP-1.0/). Conserved alpha helix structures are shown upstream and downstream of the Furin cleavage site. Sequences from vertebrate elements are marked by asterisks.

Supplementary Figure S6 – Conservation of Cysteine bridges in errantivirus HSV/gB ectodomains. Shown are alignments of domain I and II of all errantivirus HSV/gB ectodomains with Cysteine residues highlighted in red and hydrophobic residues in the fusion loop peptides highlighted in green. Sequences from different metazoan orders are separately aligned. Conserved Cysteine bridges within and across the alignments are numbered, and the tree positions of ERVs are shown where "insect errantiviruses" and "ancient errantiviruses" are highlighted in white and green, respectively, as in Fig 2A. ERVs that were used for the structural analysis are marked by asterisks. The structure of Hymenoptera_Formicoidea_Tbic_errantivirus_1 is shown to highlight the position of the Cysteine bridge in domain I of insect errantivirus HSV/gB ectodomain and the absence of the fusion loop II. Note that the

single HSV/gB-type ENV found in Annelida errantiviruses did not predict a trimer structure, therefore, it is not included in the structure classification.

Supplementary Figure S7 – Structural characterisation of vertebrate vERVs. (A) Phylogenetic congruence of *pol* RT and *pol* INT maximum likelihood trees for vertebrate endogenous retroviruses (vERVs). **(B)** *pol* RT maximum likelihood tree of vERVs with discrete phylogenetic groups (V1 – V18) annotated. **(C)** Alphafold 2-predicted 3D structures of GP1 outlier domains; Syncytin-2, expanded C domain, and extra domain. Structures are taken from Mammal_Diprotodontia_Neug_vERV_2, Bird_Galliformes_Csqu_vERV_2 and Amphibian_Gymnophonia_Rbiv_vERV_1 for Syncytin-2 like, expanded C domain, and extra domain GP1 structures, respectively. **(D)** A detailed view of phylogenetic groupings within vERVs is shown. Bridge region structures (*) annotated for RNase H + Tether + Mini (blue) and RNase H + Tether (orange). Host species (**) annotated for mammals (orange), birds (pink), reptiles (yellow), amphibians (blue), fish (green) and jawless fish (purple) as well as known exogenous retroviruses (white). ENV annotated for GP1 with RBD and C domains (yellow), GP1 with C domain only (blue), GP1 with Syncytin-2 (dark blue), and glycoprotein F (grey). Expanded C domain and extra domain annotated with green and red markers respectively. **(E)** tRNA annotations of vERVs within phylogenetic groups of *pol* RT. Number of elements indicated on Y axis.

Supplementary Figure S8 – Conservation of Cysteine residues in GP1 C domains. Shown are alignments of the representative C domains within GP1 structures analysed with Alphafold 2. Only the C domain sequences are shown for the RBD + C group. Cysteine residues (red), Cysteine bridges marked for conserved groups, and Furin cleavage site (cyan) are highlighted. Furin cleavage sites were predicted using ProP-1.0 (https://services.healthtech.dtu.dk/services/ProP-1.0/). Positions of numbered Cysteine bridges within each GP1 structure are consistent across different alignments. The conserved “CxxC” motif is also indicated. Tree positions of elements annotated with the vERV *pol* RT tree are shown. vERV names are coloured according to classes of their origin; fish (green), Amphibians (blue), reptiles (yellow), birds (magenta) and mammals (orange).

Supplementary Table 1 – Master ERV table. Table of all 1765 newly identified *env*-containing retroelements identified in this study, with host phylogeny, genome assembly name and coordinates, contig size, type of ENV, copy number, tRNA PBS type, 3' and 5' LTR coordinates and PBS mismatches. The second tab summarises the tRNA PBS types identified in this study and their 12 nucleotides sequences.

Supplementary Table 2 – Structural Analysis of errantiviruses and vERVs. RNase H domains found in the bridge region of *pol* for representative elements taken for Alphafold 2 analysis (RNase H). List of representative elements taken for Alphafold 2 prediction of *env* F ectodomain, and further analysis (F). List of representative elements taken for Alphafold 2 prediction of GP1 with domain architecture indicated (GP1). List of representative elements taken for Alphafold 2 prediction of HSV/gB ectodomain, with Cysteine bridge configuration indicated (HSV).

Supplementary Table 3 – Node information for errantiviruses and vERVs. List of errantiviruses and phylogenetic node where they are found for errantivirus RT and INT (INT-RT Nodes Errantivirus FigS2), vERV RT and INT (INT-RT Nodes vERVs FigS7), as well as F representatives INT and ENV (INT-ENV Nodes F-type FigS4). Data was then used for congruence plots in Fig S2, S4, and S7.

Supplementary Table 4 – Genome assemblies analysed in this study. List of genome assemblies with species name, size of assembly, class, family and number of tBLASTn hits for vERV/errantivirus ORFs. Genomes that we used to annotate vERVs and errantiviruses are indicated. Tabs are individually detailed for Diptera, Lepidoptera, Coleoptera, Hymenoptera, Hemiptera, Polyneoptera, Palaeoptera, non-insect Arthropoda, non-arthropod Protostomia, Cnidaria, Ctenophora, Porifera, non-vertebrate Deuterostomia, and Vertebrata.

Acknowledgements

We acknowledge Julius Brennecke, Kirsten Senti and Alex Greenwood for valuable comments in preparation of the manuscript. This work is supported by the Australian Research Council (DP210102385).

Author contributions

S.C. and R.H. analysed and interpreted the data and wrote the manuscript. The idea of this work was conceived by R.H., and S.C. and R.H. together made key findings.

Data Availability

The following data are available from the git repository (https://github.com/RippeiHayashi/errantiviruses_ERVs_2024/tree/main): the code used to annotate and analyse genomic copies of errantiviruses and vERVs, genomic sequences and open reading frames of the newly identified ERVs, pdb files for the predicted structures of ENV proteins and the POL bridge region, sequences of GAG, POL and ENV open reading frames that are used for tBLASTn search, multiple sequence alignments and the tree files that are generated in this study for the structural and phylogenetic analyses.

Conflicts of interests

The authors declare no conflicts of interests.

References

1 Johnson, W. E. Endogenous Retroviruses in the Genomics Era. *Annu Rev Virol* **2**, 135-159 (2015). https://doi.org:10.1146/annurev-virology-100114-054945

2 Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat Rev Genet* **13**, 283-296 (2012). https://doi.org:10.1038/nrg3199

3 Zhuo, X., Rho, M. & Feschotte, C. Genome-wide characterization of endogenous retroviruses in the bat Myotis lucifugus reveals recent and diverse infections. *J Virol* **87**, 8493-8501 (2013). https://doi.org:10.1128/JVI.00892-13

4 Dewannieux, M. & Heidmann, T. Endogenous retroviruses: acquisition, amplification and taming of genome invaders. *Curr Opin Virol* **3**, 646-656 (2013). https://doi.org:10.1016/j.coviro.2013.08.005

5 Tarlinton, R. E., Meers, J. & Young, P. R. Retroviral invasion of the koala genome. *Nature* **442**, 79-81 (2006). https://doi.org:10.1038/nature04841

6 Zheng, H. *et al.* Koala retrovirus diversity, transmissibility, and disease associations. *Retrovirology* **17**, 34 (2020). https://doi.org:10.1186/s12977-020-00541-1

7 Zheng, J., Wei, Y. & Han, G. Z. The diversity and evolution of retroviruses: Perspectives from viral "fossils". *Virol Sin* **37**, 11-18 (2022). https://doi.org:10.1016/j.virs.2022.01.019

8 Gasparotto, E. *et al.* Transposable Elements Co-Option in Genome Evolution and Gene Regulation. *Int J Mol Sci* **24** (2023). https://doi.org:10.3390/ijms24032610

9 Naville, M. & Volff, J. N. Endogenous Retroviruses in Fish Genomes: From Relics of Past Infections to Evolutionary Innovations? *Front Microbiol* **7**, 1197 (2016). https://doi.org:10.3389/fmicb.2016.01197

10 Grandi, N. & Tramontano, E. Human Endogenous Retroviruses Are Ancient Acquired Elements Still Shaping Innate Immune Responses. *Front Immunol* **9**, 2039 (2018). https://doi.org:10.3389/fimmu.2018.02039

11 Coffin, J. M., Hughes, S. H. & Varmus, H. E. (1997).

12 Platt, R. N., Vandewege, M. W. & Ray, D. A. Mammalian transposable elements and their impacts on genome evolution. *Chromosome Res* **26**, 25-43 (2018). https://doi.org:10.1007/s10577-017-9570-z

13 Mason, A. S., Fulton, J. E., Hocking, P. M. & Burt, D. W. A new look at the LTR retrotransposon content of the chicken genome. *BMC Genomics* **17**, 688 (2016). https://doi.org:10.1186/s12864-016-3043-1

14 Sandmeyer, S., Patterson, K. & Bilanchone, V. Ty3, a Position-specific Retrotransposon in Budding Yeast. *Microbiol Spectr* **3**, MDNA3-0057-2014 (2015). https://doi.org:10.1128/microbiolspec.MDNA3-0057-2014

15 Doolittle, R. F., Feng, D. F., Johnson, M. S. & McClure, M. A. Origins and evolutionary relationships of retroviruses. *Q Rev Biol* **64**, 1-30 (1989). https://doi.org:10.1086/416128

16 Xiong, Y. & Eickbush, T. H. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J* **9**, 3353-3362 (1990). https://doi.org:10.1002/j.1460-2075.1990.tb07536.x

17 Hayward, A. Origin of the retroviruses: when, where, and how? *Curr Opin Virol* **25**, 23-27 (2017). https://doi.org:10.1016/j.coviro.2017.06.006

18 Wang, J. & Han, G. Z. A Sister Lineage of Sampled Retroviruses Corroborates the Complex Evolution of Retroviruses. *Mol Biol Evol* **38**, 1031-1039 (2021). https://doi.org:10.1093/molbev/msaa272

19 Pelisson, A., Mejlumian, L., Robert, V., Terzian, C. & Bucheton, A. Drosophila germline invasion by the endogenous retrovirus gypsy: involvement of the viral env gene. *Insect Biochem Mol Biol* **32**, 1249-1256 (2002). https://doi.org:10.1016/s0965-1748(02)00088-7

20 Keegan, R. M., Talbot, L. R., Chang, Y. H., Metzger, M. J. & Dubnau, J. Intercellular viral spread and intracellular transposition of Drosophila gypsy. *PLoS Genet* **17**, e1009535 (2021). https://doi.org:10.1371/journal.pgen.1009535

21 Song, S. U., Gerasimova, T., Kurkulos, M., Boeke, J. D. & Corces, V. G. An env-like protein encoded by a Drosophila retroelement: evidence that gypsy is an infectious retrovirus. *Genes Dev* **8**, 2046-2057 (1994). https://doi.org:10.1101/gad.8.17.2046

22 Brasset, E. *et al.* Viral particles of the endogenous retrovirus ZAM from Drosophila melanogaster use a pre-existing endosome/exosome pathway for transfer to the oocyte. *Retrovirology* **3**, 25 (2006). https://doi.org:10.1186/1742-4690-3-25

23 Malik, H. S., Henikoff, S. & Eickbush, T. H. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res* **10**, 1307-1318 (2000). https://doi.org:10.1101/gr.145000

24 Rohrmann, G. F. & Karplus, P. A. Relatedness of baculovirus and gypsy retrotransposon envelope proteins. *BMC Evol Biol* **1**, 1 (2001). https://doi.org:10.1186/1471-2148-1-1

25 Dezordi, F. Z., Vasconcelos, C. R. D. S., Rezende, A. M. & Wallau, G. L. In and Outs of Chuviridae Endogenous Viral Elements: Origin of a Potentially New Retrovirus and Signature of Ancient and Ongoing Arms Race in Mosquito Genomes. *Front Genet* **11**, 542437 (2020). https://doi.org:10.3389/fgene.2020.542437

26 Jehle, J. A. *et al.* Molecular identification and phylogenetic analysis of baculoviruses from Lepidoptera. *Virology* **346**, 180-193 (2006). https://doi.org:10.1016/j.virol.2005.10.032

27 Sacco, M. A., Lau, J., Godinez-Vidal, D. & Kaloshian, I. Non-canonical nematode endogenous retroviruses resulting from RNA virus glycoprotein gene capture by a metavirus. *J Gen Virol* **103** (2022). https://doi.org:10.1099/jgv.0.001739

28 Rodriguez, F., Kenefick, A. W. & Arkhipova, I. R. LTR-Retrotransposons from Bdelloid Rotifers Capture Additional ORFs Shared between Highly Diverse Retroelement Types. *Viruses* **9** (2017). https://doi.org:10.3390/v9040078

29 Bargues, N. & Lerat, E. Evolutionary history of LTR-retrotransposons among 20 Drosophila species. *Mob DNA* **8**, 7 (2017). https://doi.org:10.1186/s13100-017-0090-3

30 Stefanov, Y., Salenko, V. & Glukhov, I. Drosophila errantiviruses. *Mob Genet Elements* **2**, 36-45 (2012). https://doi.org:10.4161/mge.19234

31 Terzian, C., Pélisson, A. & Bucheton, A. Evolution and phylogeny of insect endogenous retroviruses. *BMC Evol Biol* **1**, 3 (2001). https://doi.org:10.1186/1471-2148-1-3

32 Sormacheva, I. *et al.* Vertical evolution and horizontal transfer of CR1 non-LTR retrotransposons and Tc1/mariner DNA transposons in Lepidoptera species. *Mol Biol Evol* **29**, 3685-3702 (2012). https://doi.org:10.1093/molbev/mss181

33 Wanner, N. M. & Faulk, C. Suggested Absence of Horizontal Transfer of Retrotransposons between Humans and Domestic Mammal Species. *Genes (Basel)* **12** (2021). https://doi.org:10.3390/genes12081223

34 Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**, 691-703 (2009). https://doi.org:10.1038/nrg2640

35 Gagnier, L., Belancio, V. P. & Mager, D. L. Mouse germ line mutations due to retrotransposon insertions. *Mob DNA* **10**, 15 (2019). https://doi.org:10.1186/s13100-019-0157-4

36 McLellan, J. S., Ray, W. C. & Peeples, M. E. Structure and function of respiratory syncytial virus surface glycoproteins. *Curr Top Microbiol Immunol* **372**, 83-104 (2013). https://doi.org:10.1007/978-3-642-38919-1_4

37 Dang, H. V. *et al.* An antibody against the F glycoprotein inhibits Nipah and Hendra virus infections. *Nat Struct Mol Biol* **26**, 980-987 (2019). https://doi.org:10.1038/s41594-019-0308-9

38 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021). https://doi.org:10.1038/s41586-021-03819-2

39 Heldwein, E. E. *et al.* Crystal structure of glycoprotein B from herpes simplex virus 1. *Science* **313**, 217-220 (2006). https://doi.org:10.1126/science.1126548

40 Connolly, S. A., Jackson, J. O., Jardetzky, T. S. & Longnecker, R. Fusing structure and function: a structural view of the herpesvirus entry machinery. *Nat Rev Microbiol* **9**, 369-381 (2011). https://doi.org:10.1038/nrmicro2548

41 Lopper, M. & Compton, T. Disulfide bond configuration of human cytomegalovirus glycoprotein B. *J Virol* **76**, 6073-6082 (2002). https://doi.org:10.1128/jvi.76.12.6073-6082.2002

42 Vollmer, B. *et al.* The prefusion structure of herpes simplex virus glycoprotein B. *Sci Adv* **6** (2020). https://doi.org:10.1126/sciadv.abc1726

43 Malik, H. S. & Eickbush, T. H. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Res* **11**, 1187-1197 (2001). https://doi.org:10.1101/gr.185101

44 Gifford, R., Kabat, P., Martin, J., Lynch, C. & Tristem, M. Evolution and distribution of class II-related endogenous retroviruses. *J Virol* **79**, 6478-6486 (2005). https://doi.org:10.1128/JVI.79.10.6478-6486.2005

45 Lee, J. E. *et al.* Structure of the Ebola virus glycoprotein bound to an antibody from a human survivor. *Nature* **454**, 177-182 (2008). https://doi.org:10.1038/nature07082

46 Hötzel, I. Deep-Time Structural Evolution of Retroviral and Filoviral Surface Envelope Proteins. *J Virol* **96**, e0006322 (2022). https://doi.org:10.1128/jvi.00063-22

47 Hogan, V. & Johnson, W. E. Unique Structure and Distinctive Properties of the Ancient and Ubiquitous Gamma-Type Envelope Glycoprotein. *Viruses* **15** (2023). https://doi.org:10.3390/v15020274

48 Loreto, E. L., Carareto, C. M. & Capy, P. Revisiting horizontal transfer of transposable elements in Drosophila. *Heredity (Edinb)* **100**, 545-554 (2008). https://doi.org:10.1038/sj.hdy.6801094

49 Metzger, M. J., Paynter, A. N., Siddall, M. E. & Goff, S. P. Horizontal transfer of retrotransposons between bivalves and other aquatic species of multiple phyla. *Proc Natl Acad Sci U S A* **115**, E4227-E4235 (2018). https://doi.org:10.1073/pnas.1717227115

50 Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* **128**, 1089-1103 (2007). https://doi.org:10.1016/j.cell.2007.01.043

51 Ribet, D. *et al.* An infectious progenitor for the murine IAP retrotransposon: emergence of an intracellular genetic parasite from an ancient retrovirus. *Genome Res* **18**, 597-609 (2008). https://doi.org:10.1101/gr.073486.107

52 Wang, J. & Han, G. Z. Frequent Retroviral Gene Co-option during the Evolution of Vertebrates. *Mol Biol Evol* **37**, 3232-3242 (2020). https://doi.org:10.1093/molbev/msaa180

53 Wang, W., Zhao, H. & Han, G. Z. Host-Virus Arms Races Drive Elevated Adaptive Evolution in Viral Receptors. *J Virol* **94** (2020). https://doi.org:10.1128/JVI.00684-20

54 Johnson, D. C., Wisner, T. W. & Wright, C. C. Herpes simplex virus glycoproteins gB and gD function in a redundant fashion to promote secondary envelopment. *J Virol* **85**, 4910-4926 (2011). https://doi.org:10.1128/JVI.00011-11

55 Jambunathan, N. *et al.* Two Sides to Every Story: Herpes Simplex Type-1 Viral Glycoproteins gB, gD, gH/gL, gK, and Cellular Receptors Function as Key Players in Membrane Fusion. *Viruses* **13** (2021). https://doi.org:10.3390/v13091849

56 Kigami, D., Minami, N., Takayama, H. & Imai, H. MuERV-L is one of the earliest transcribed genes in mouse one-cell embryos. *Biol Reprod* **68**, 651-654 (2003). https://doi.org:10.1095/biolreprod.102.007906

57 Ribet, D. *et al.* Murine endogenous retrovirus MuERV-L is the progenitor of the "orphan" epsilon viruslike particles of the early mouse embryo. *J Virol* **82**, 1622-1625 (2008). https://doi.org:10.1128/JVI.02097-07

58 Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019). https://doi.org:10.1186/s12859-019-3019-7

59 Wilhelm, M. & Wilhelm, F. X. Reverse transcription of retroviruses and LTR retrotransposons. *Cell Mol Life Sci* **58**, 1246-1262 (2001). https://doi.org:10.1007/PL00000937

60 Das, D. & Georgiadis, M. M. The crystal structure of the monomeric reverse transcriptase from Moloney murine leukemia virus. *Structure* **12**, 819-829 (2004). https://doi.org:10.1016/j.str.2004.02.032

61 Engelman, A. & Cherepanov, P. Retroviral Integrase Structure and DNA Recombination Mechanism. *Microbiol Spectr* **2**, 1-22 (2014). https://doi.org:10.1128/microbiolspec.MDNA3-0024-2014.

62 Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679-682 (2022). https://doi.org:10.1038/s41592-022-01488-1

Figure 1

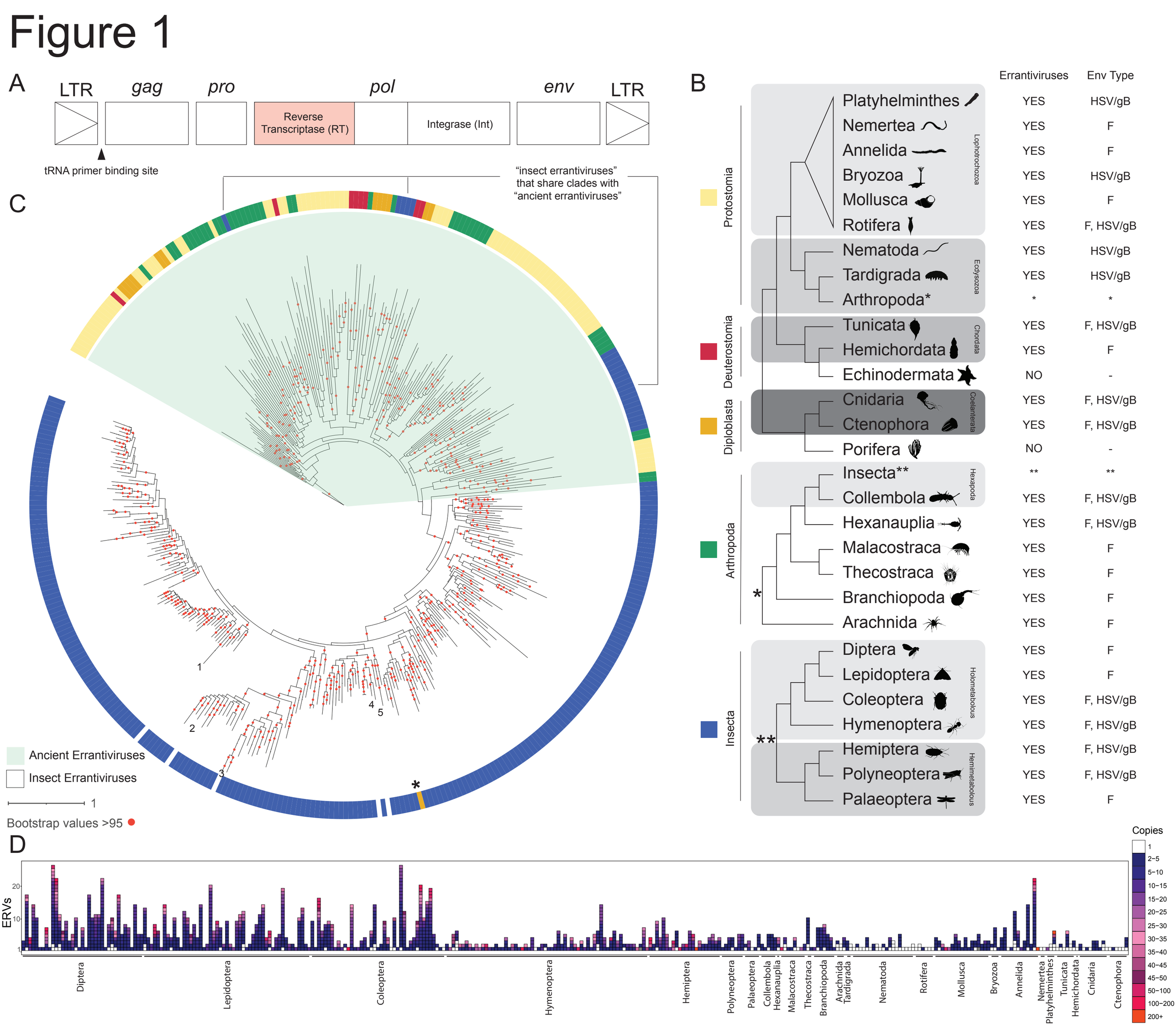
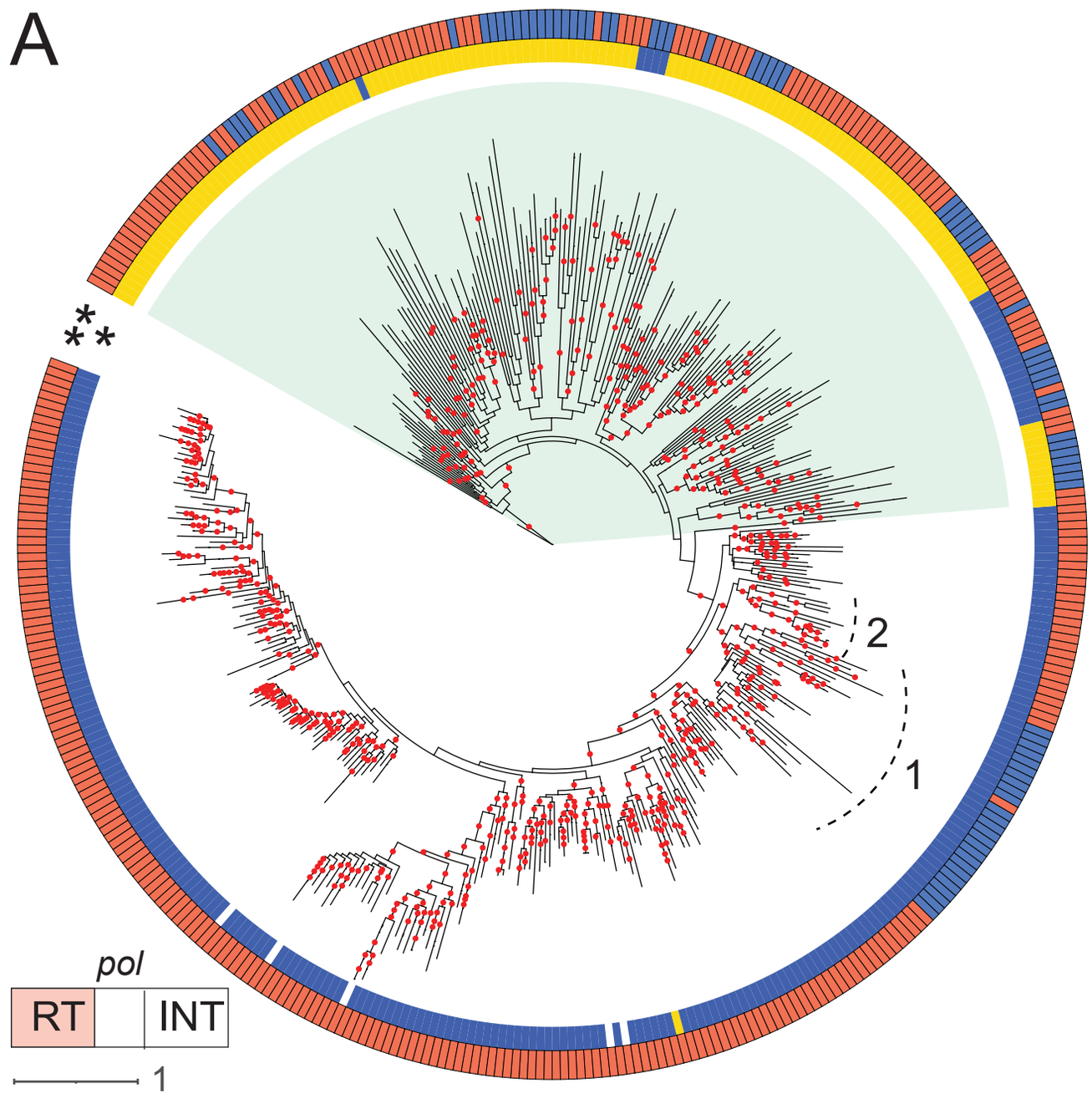

A
LTR
gag
pro
pol
env
LTR
Reverse Transcriptase (RT)
Integrase (Int)
tRNA primer binding site
B
Errantiviruses
Env Type
Protostomia
Deuterostomia
Diploblasta
Arthropoda
Insecta
Lophotrochozoa
Ecdysozoa
Chordata
Coelanterata
Hexapoda
Holometabolous
Hemimetabolous
Platyhelminthes YES HSV/gB
Nemertea YES F
Annelida YES F
Bryozoa YES HSV/gB
Mollusca YES F
Rotifera YES F, HSV/gB
Nematoda YES HSV/gB
Tardigrada YES HSV/gB
Arthropoda* * *
Tunicata YES F, HSV/gB
Hemichordata YES F
Echinodermata NO -
Cnidaria YES F, HSV/gB
Ctenophora YES F, HSV/gB
Porifera NO -
Insecta** ** **
Collembola YES F, HSV/gB
Hexanauplia YES F, HSV/gB
Malacostraca YES F
Thecostraca YES F
Branchiopoda YES F
Arachnida YES F
Diptera YES F
Lepidoptera YES F
Coleoptera YES F, HSV/gB
Hymenoptera YES F, HSV/gB
Hemiptera YES F, HSV/gB
Polyneoptera YES F, HSV/gB
Palaeoptera YES F
C
"insect errantiviruses" that share clades with "ancient errantiviruses"
1
2
3
4
5
Ancient Errantiviruses
Insect Errantiviruses
1
Bootstrap values >95
D
Copies
ERVs
20
10
1
Diptera
Lepidoptera
Coleoptera
Hymenoptera
Hemiptera
Polyneoptera
Palaeoptera
Collembola
Hexanauplia
Malacostraca
Thecostraca
Branchiopoda
Arachnida
Tardigrada
Nematoda
Rotifera
Mollusca
Bryozoa
Annelida
Nemertea
Platyhelminthes
Tunicata
Hemichordata
Cnidaria
Ctenophora
1
2–5
5–10
10–15
15–20
20–25
25–30
30–35
35–40
40–45
45–50
50–100
100–200
200+

Figure 2

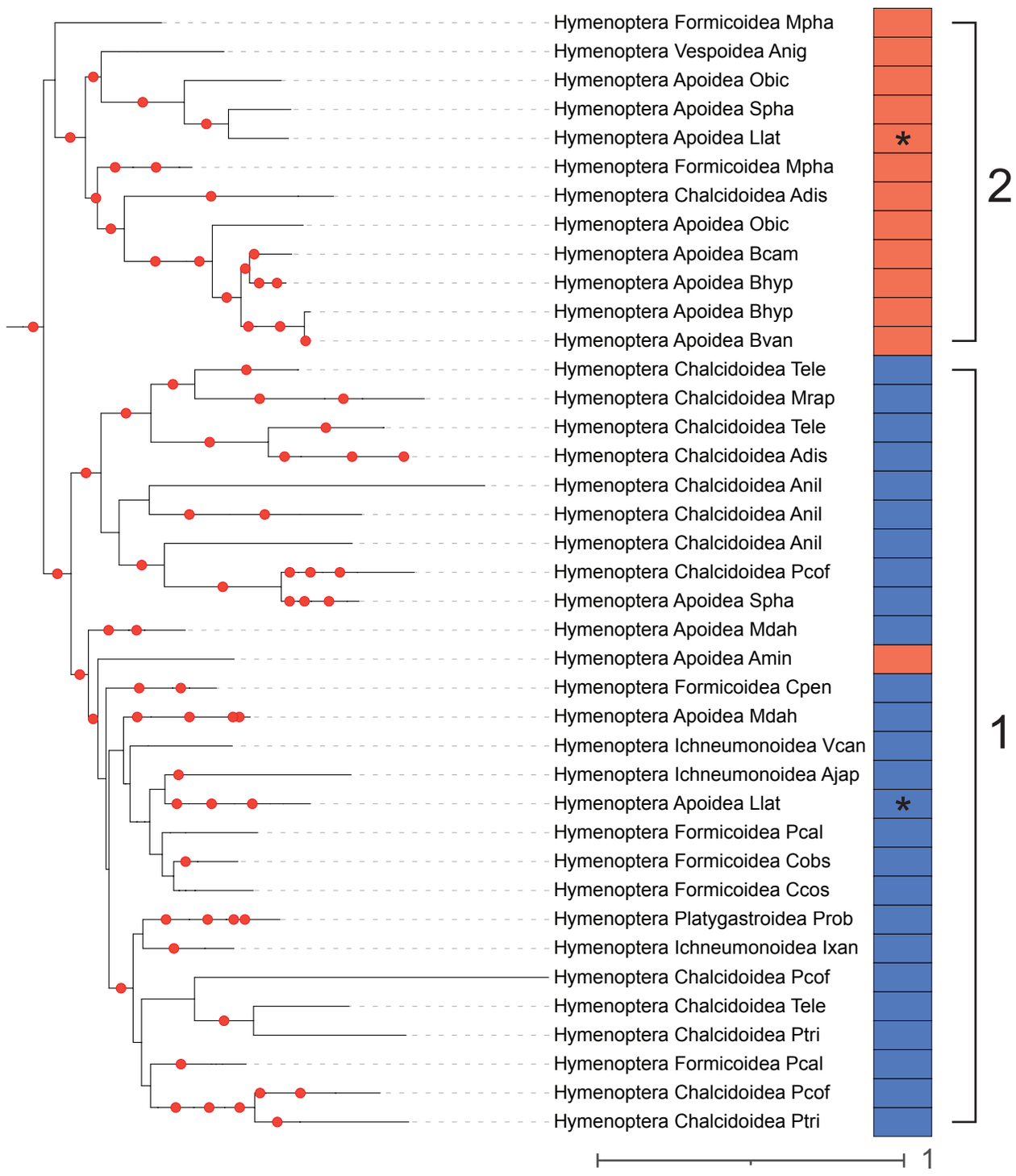
A



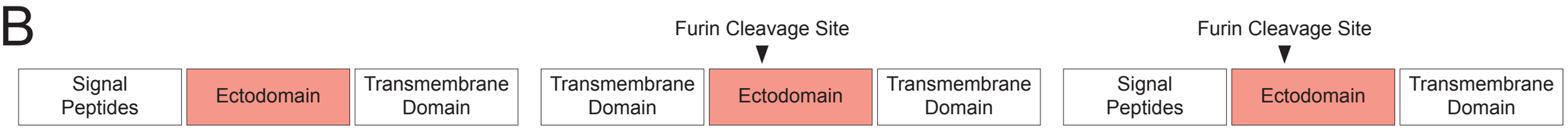
Bootstrap values >95

* Insects
Other Animals
** F
HSV/gB
Ancient Errantiviruses
Insect Errantiviruses

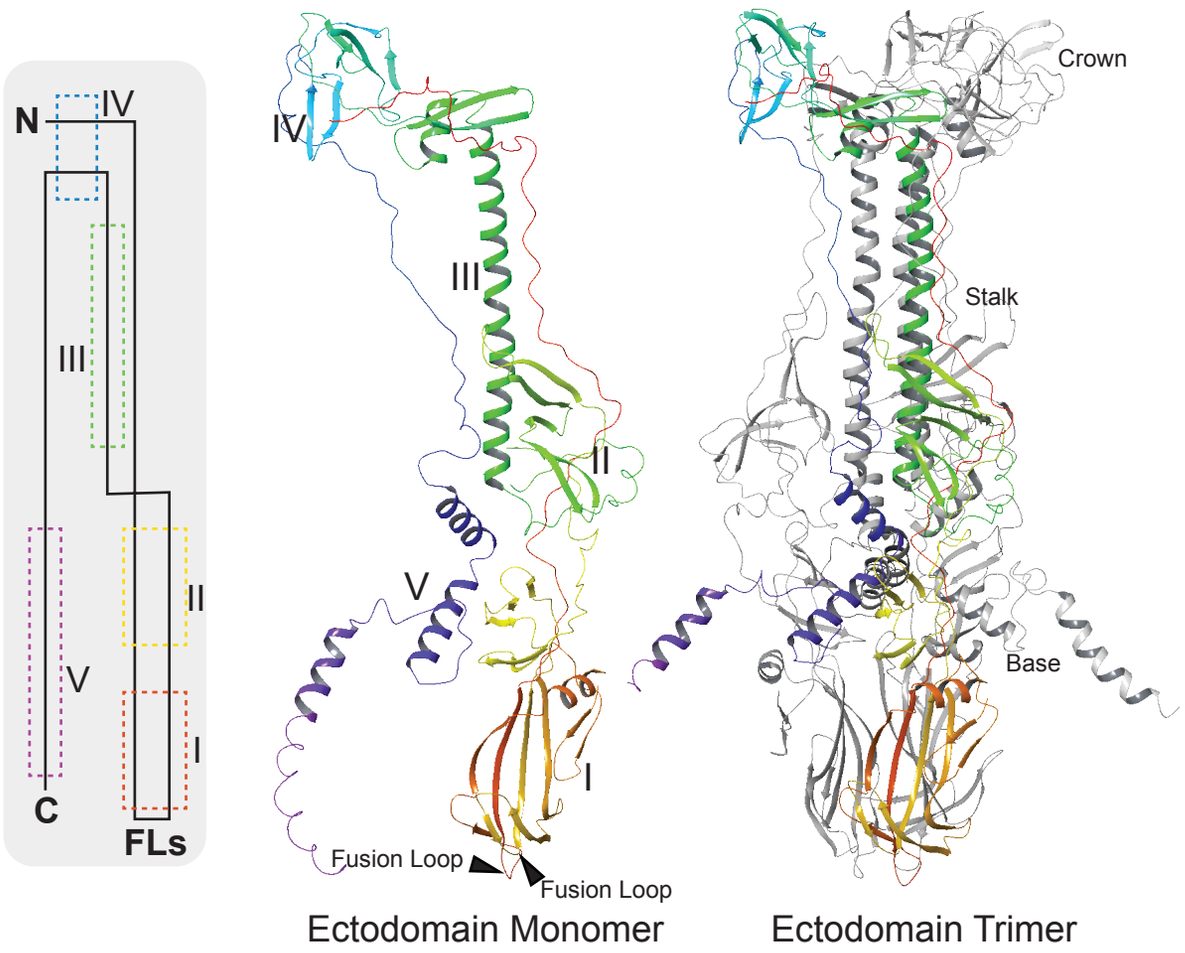
A'



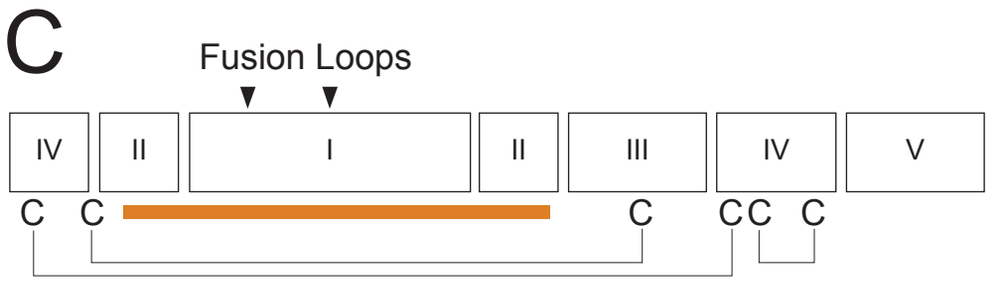
B



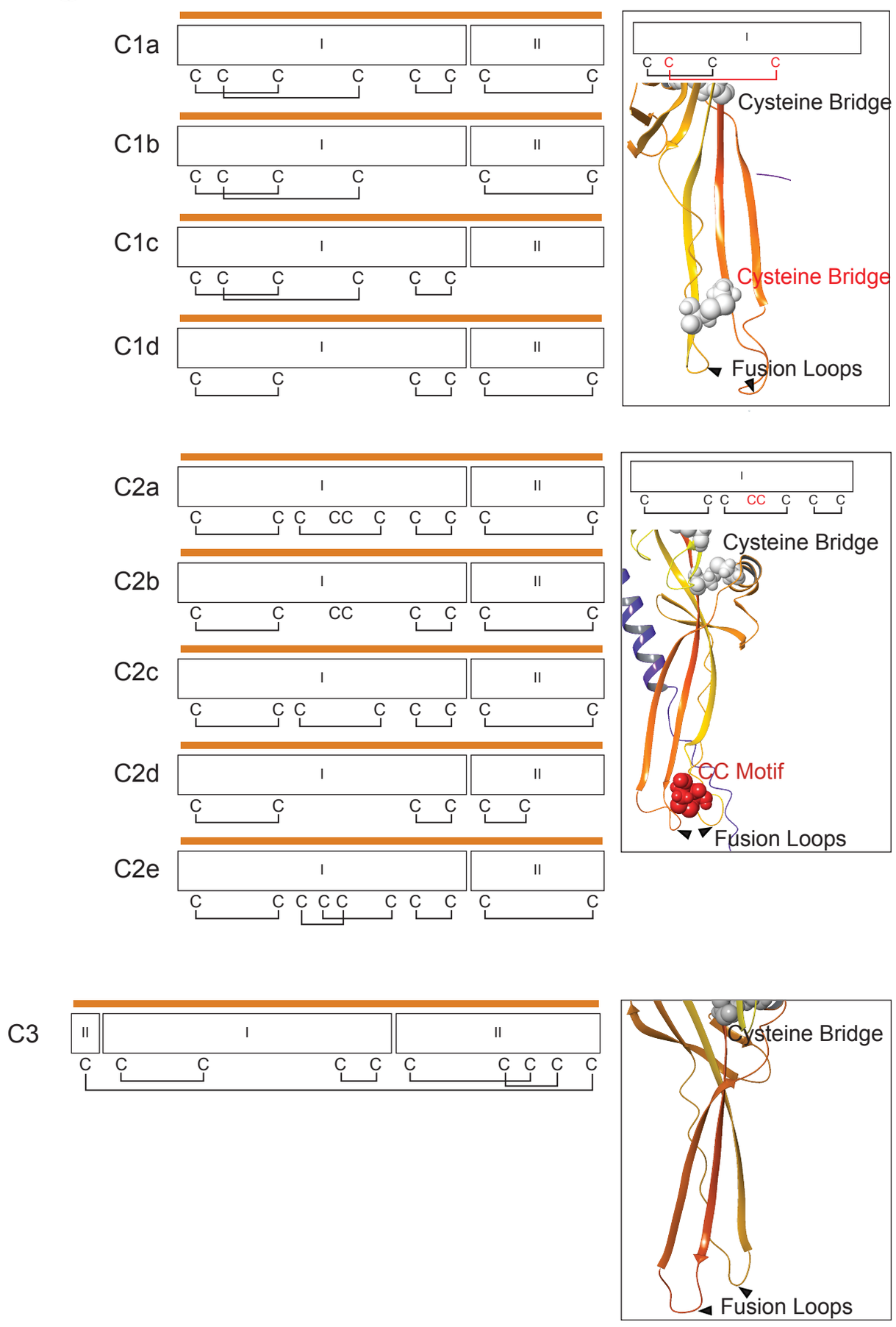
B'



C



C'



D

Host Order	Cysteine Bridges
Platyhelminthes	C1a, C1d, C2b
Bryozoa	C2b
Rotifera	C2e
Nematoda	C1a, C1d, C2a, C2c
Tardigrada	C2c
Tunicata	C3
Cnidaria	C1c
Ctenophora	C1c
Collembola	C2d
Hexanauplia	C1b
Branchiopoda	C2c
Coleoptera	C1a
Hymenoptera	C1a
Hemiptera	C1b
Polyneoptera	C1a

Figure 3

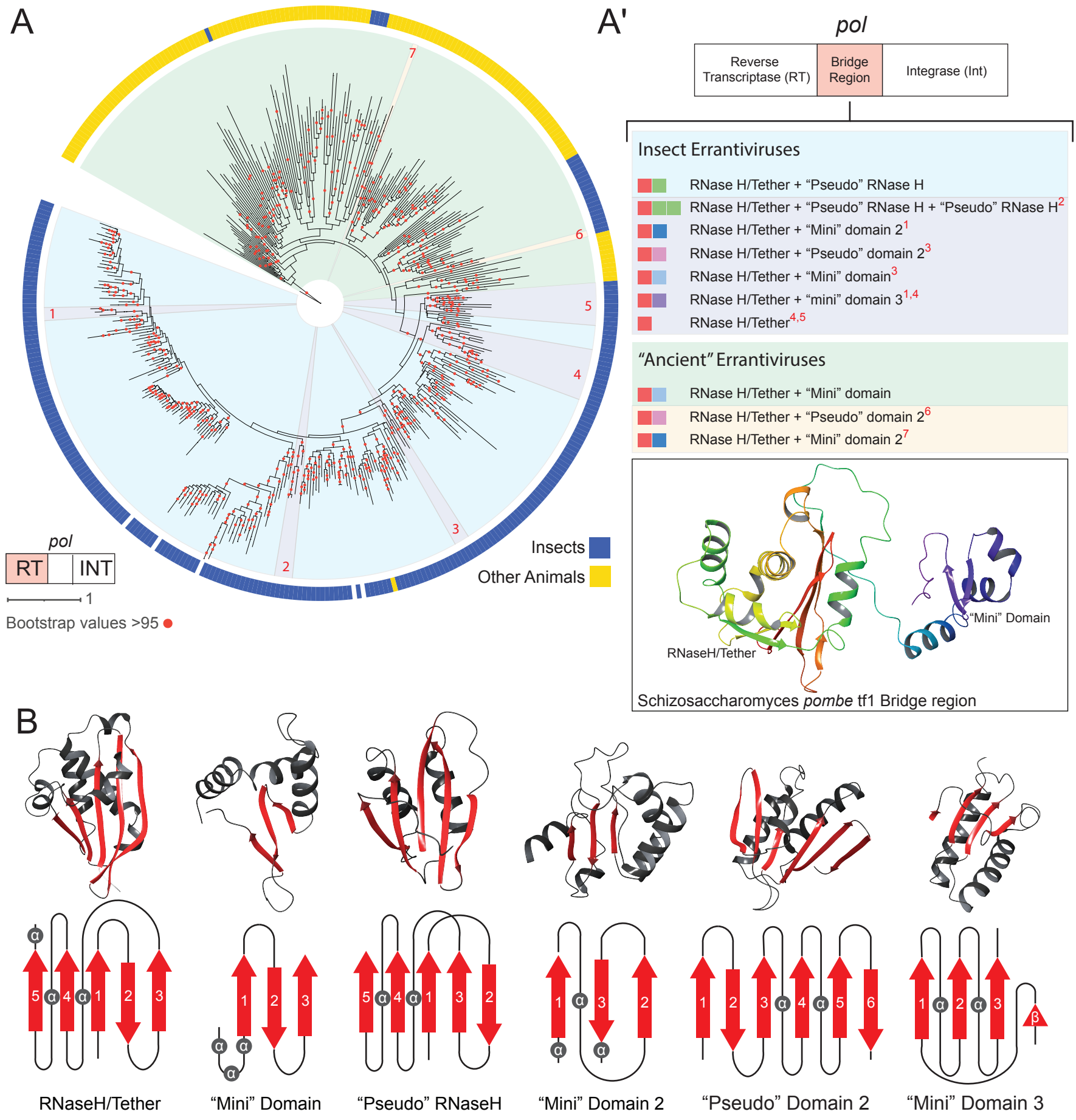
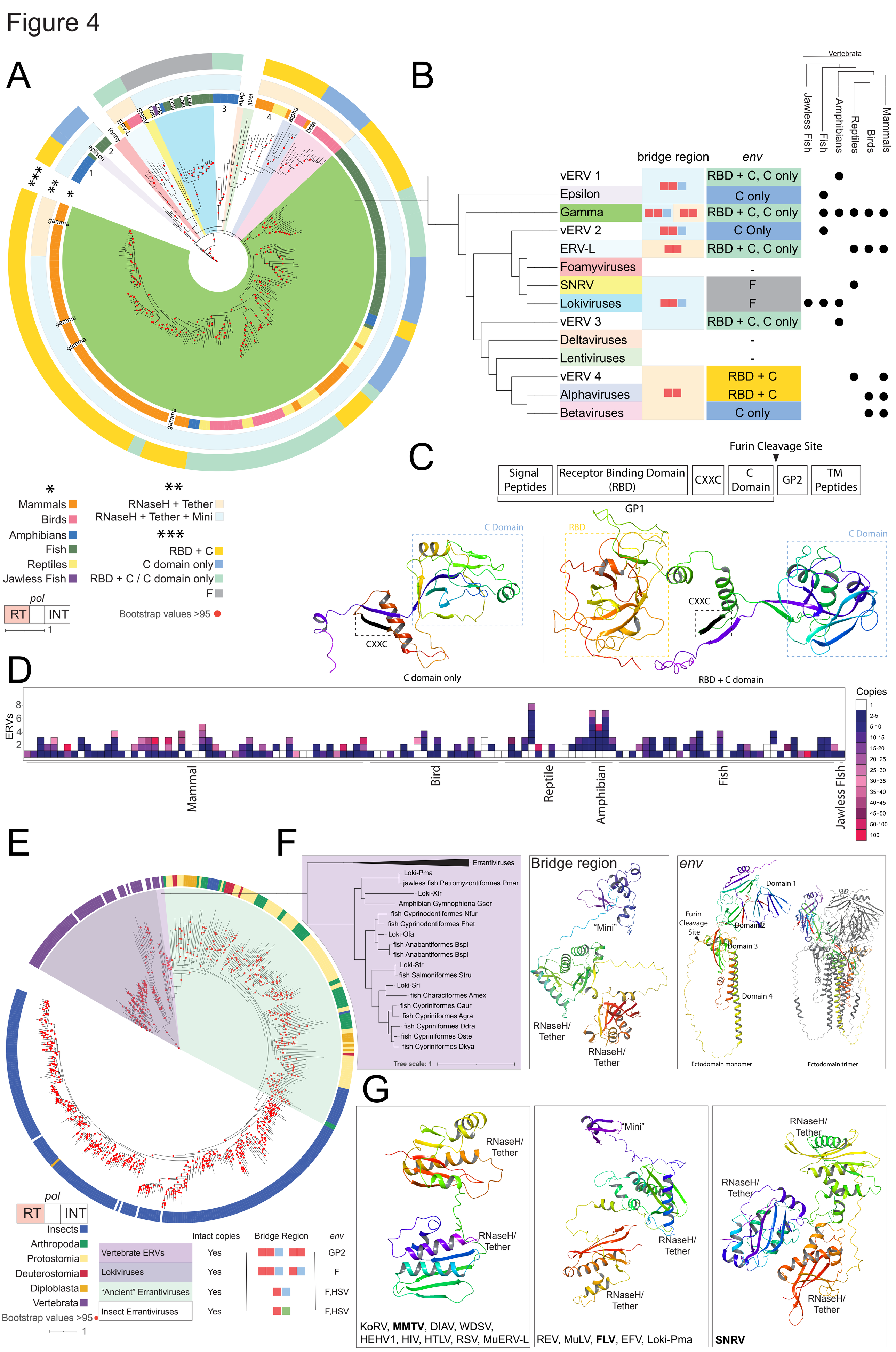

A
A'
pol
Reverse Transcriptase (RT)
Bridge Region
Integrase (Int)
Insect Errantiviruses
RNase H/Tether + "Pseudo" RNase H
RNase H/Tether + "Pseudo" RNase H + "Pseudo" RNase H2
RNase H/Tether + "Mini" domain 21
RNase H/Tether + "Pseudo" domain 23
RNase H/Tether + "Mini" domain3
RNase H/Tether + "mini" domain 31,4
RNase H/Tether4,5
"Ancient" Errantiviruses
RNase H/Tether + "Mini" domain
RNase H/Tether + "Pseudo" domain 26
RNase H/Tether + "Mini" domain 27
RNaseH/Tether
"Mini" Domain
Schizosaccharomyces pombe tf1 Bridge region
pol
RT
INT
1
Bootstrap values >95
Insects
Other Animals
B
RNaseH/Tether
"Mini" Domain
"Pseudo" RNaseH
"Mini" Domain 2
"Pseudo" Domain 2
"Mini" Domain 3

Figure 4


A
B
Vertebrata
Jawless Fish
Fish
Amphibians
Reptiles
Birds
Mammals
bridge region
env
vERV 1
Epsilon
Gamma
vERV 2
ERV-L
Foamyviruses
SNRV
Lokiviruses
vERV 3
Deltaviruses
Lentiviruses
vERV 4
Alphaviruses
Betaviruses
RBD + C, C only
C only
RBD + C, C only
C Only
RBD + C, C only
-
F
F
RBD + C, C only
-
-
RBD + C
RBD + C
C only
C
Furin Cleavage Site
Signal Peptides
Receptor Binding Domain (RBD)
CXXC
C Domain
GP2
TM Peptides
GP1
C Domain
RBD
C Domain
CXXC
CXXC
C domain only
RBD + C domain
Mammals
Birds
Amphibians
Fish
Reptiles
Jawless Fish
RT
INT
Bootstrap values >95
RNaseH + Tether
RNaseH + Tether + Mini
RBD + C
C domain only
RBD + C / C domain only
F
D
ERVs
Copies
Mammal
Bird
Reptile
Amphibian
Fish
Jawless FIsh
E
F
Errantiviruses
Loki-Pma
jawless fish Petromyzontiformes Pmar
Loki-Xtr
Amphibian Gymnophiona Gser
fish Cyprinodontiformes Nfur
fish Cyprinodontiformes Fhet
Loki-Ofa
fish Anabantiformes Bspl
fish Anabantiformes Bspl
Loki-Str
fish Salmoniformes Stru
Loki-Sri
fish Characiformes Amex
fish Cypriniformes Caur
fish Cypriniformes Agra
fish Cypriniformes Ddra
fish Cypriniformes Oste
fish Cypriniformes Dkya
Tree scale: 1
Bridge region
"Mini"
RNaseH/ Tether
RNaseH/ Tether
env
Domain 1
Domain 2
Domain 3
Domain 4
Furin Cleavage Site
Ectodomain monomer
Ectodomain trimer
G
RNaseH/ Tether
RNaseH/ Tether
KoRV, MMTV, DIAV, WDSV, HEHV1, HIV, HTLV, RSV, MuERV-L
"Mini"
RNaseH/ Tether
RNaseH/ Tether
REV, MuLV, FLV, EFV, Loki-Pma
RNaseH/ Tether
RNaseH/ Tether
RNaseH/ Tether
SNRV
pol
RT
INT
Insects
Arthropoda
Protostomia
Deuterostomia
Diploblasta
Vertebrata
Bootstrap values >95
Intact copies
Bridge Region
env
Vertebrate ERVs
Yes
GP2
Lokiviruses
Yes
F
"Ancient" Errantiviruses
Yes
F,HSV
Insect Errantiviruses
Yes
F,HSV

Supplementary Figure S1

A

Step 1. identify GAG, POL, ENV open reading frames (ORFs) by tBLASTn
using GAG, POL and ENV from RepBase Dipteran errantiviruses as baits against **1901 invertebrate genome assemblies**
using GAG, POL and ENV from RefSeq vertebrate retroviruses as baits against **747 vertebrate genome assemblies**

Step 2. identify loci where GAG, POL and ENV ORFs are next to one another

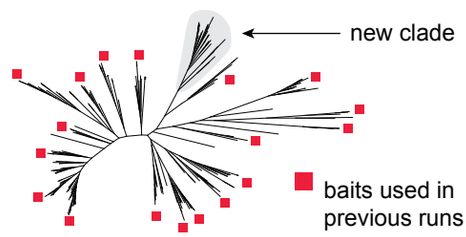
- Only the GAG-POL-ENV elements were annotated in the insect and vertebrate genomes.
- For non-insect Arthropoda and other Protostomia genomes, ENV was predicted from the downstream region of GAG-POL.
- For non-vertebrate Deuterostomia, Cnidaria and Ctenophora genomes, ENV was predicted from the downstream region of POL.
- GAG-ENV-POL elements were found in Nematoda, Bryozoa and Platyhenminthes genome.
- ENV-GAG-POL elements were uniquely found in Arachnida genome.

Step 3. cluster ERVs per genome using clustal omega and pick representative genomic copies

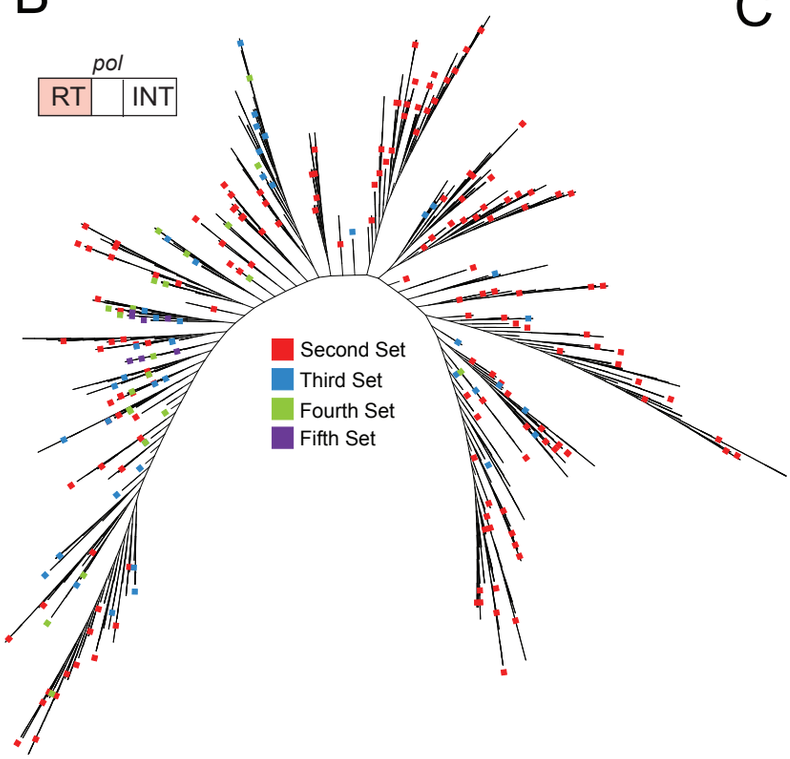
Step 4. curate the identified ERVs by domain structure prediction using hmmscan and HHpred
***All** three ORFs are required to predict homology to known viral proteins listed below*
GAG and POL: homologous to Ty3/gypsy or retroviral GAG and POL
ENV: homologous to paramyxovirus glycoprotein F (F-type), herpesvirus glycoprotein B (HSV/gB-type), or retrovirus glycoprotein (GP2-type)

Step 5. repeat the process until all clades in the phylogenetic tree of POL are represented by tBLASTn baits

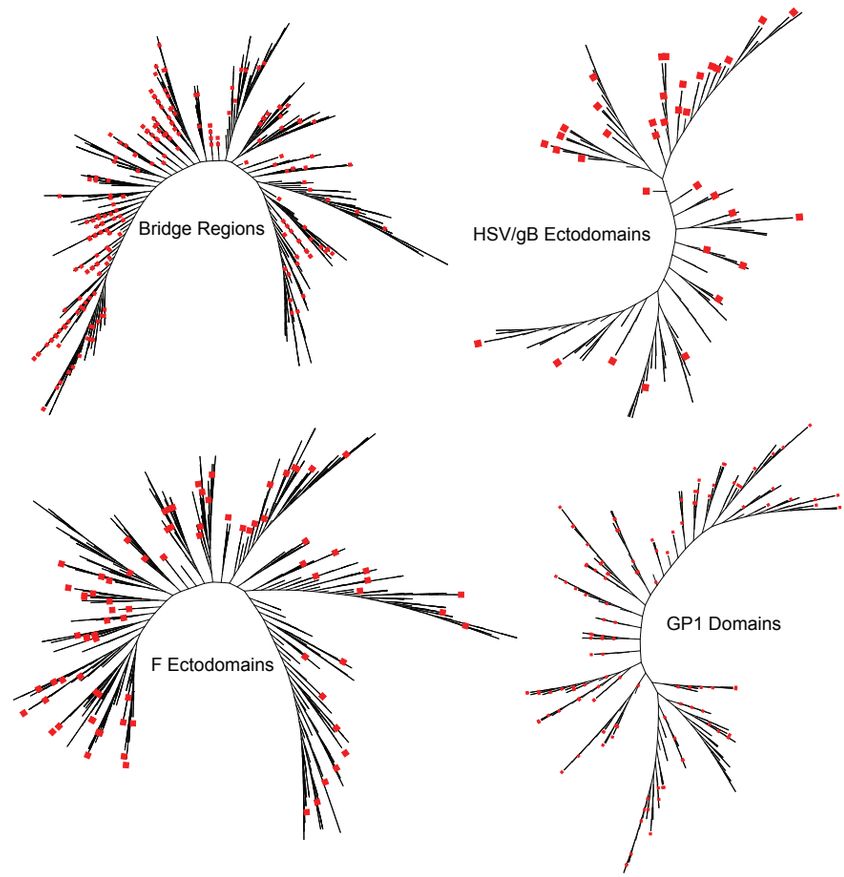
tBLASTn interations	*identified ERVs*
1st	1128
2nd	491
3rd	87
4th	53
5th	6



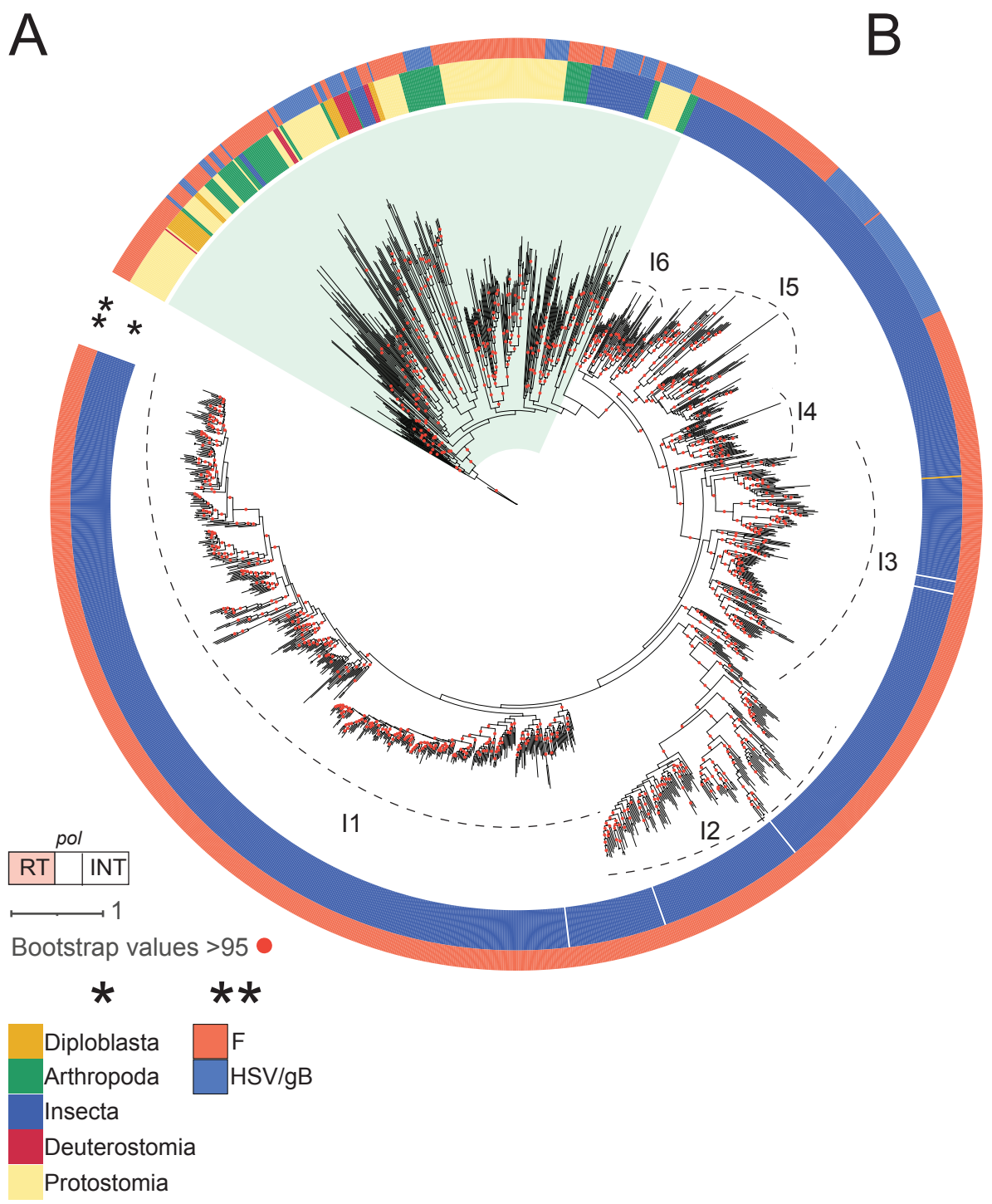
B

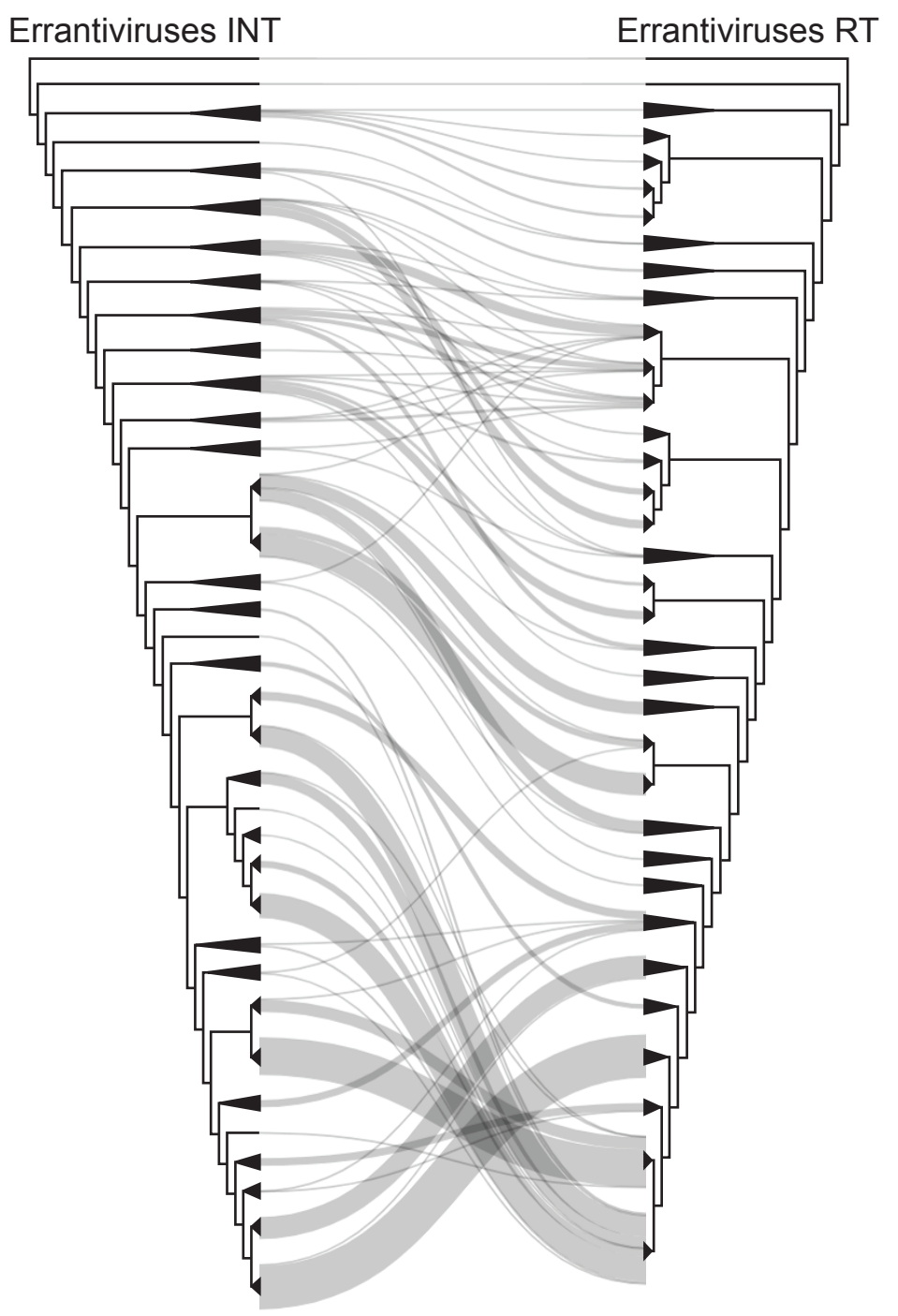


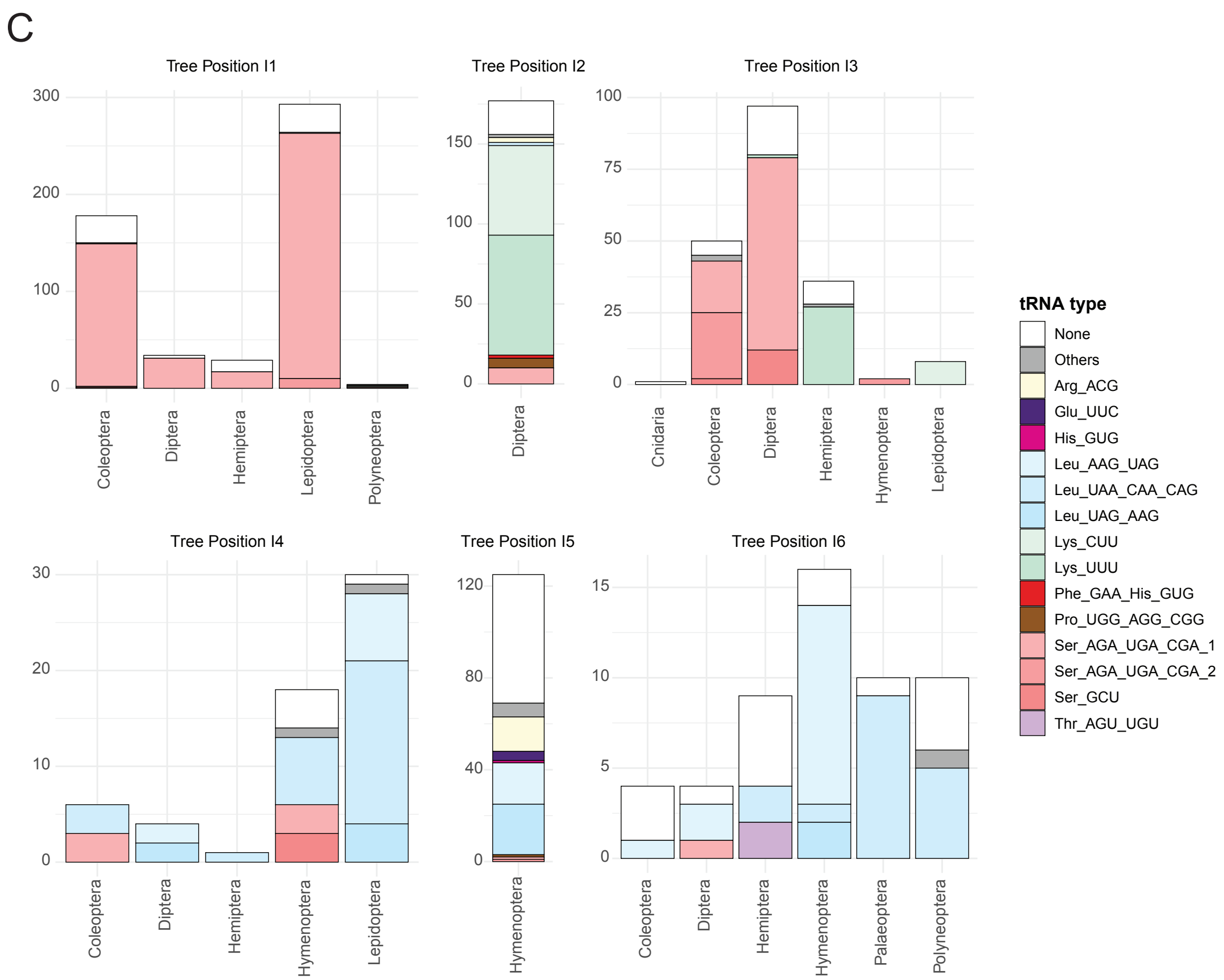
C



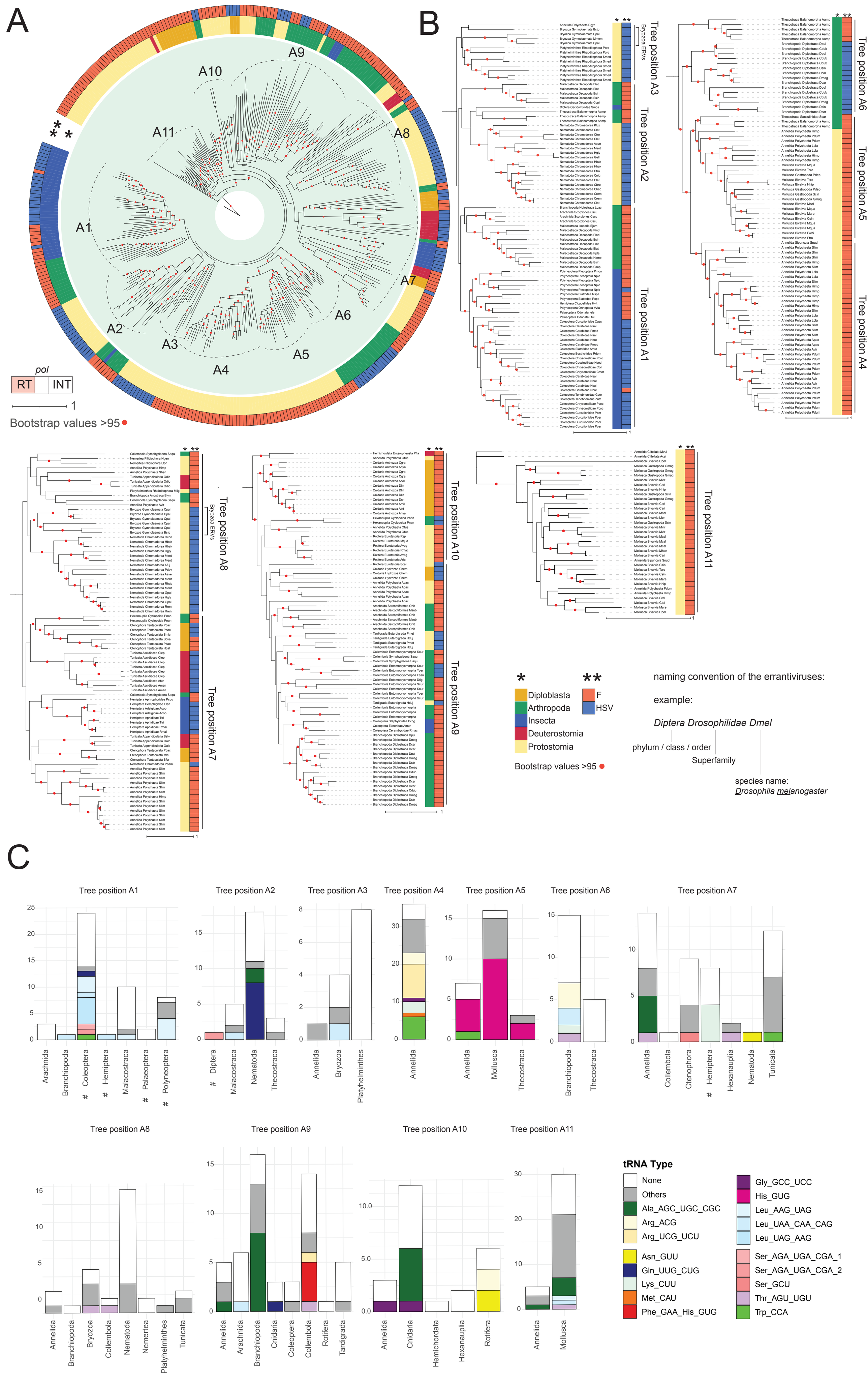
Supplementary Figure S2



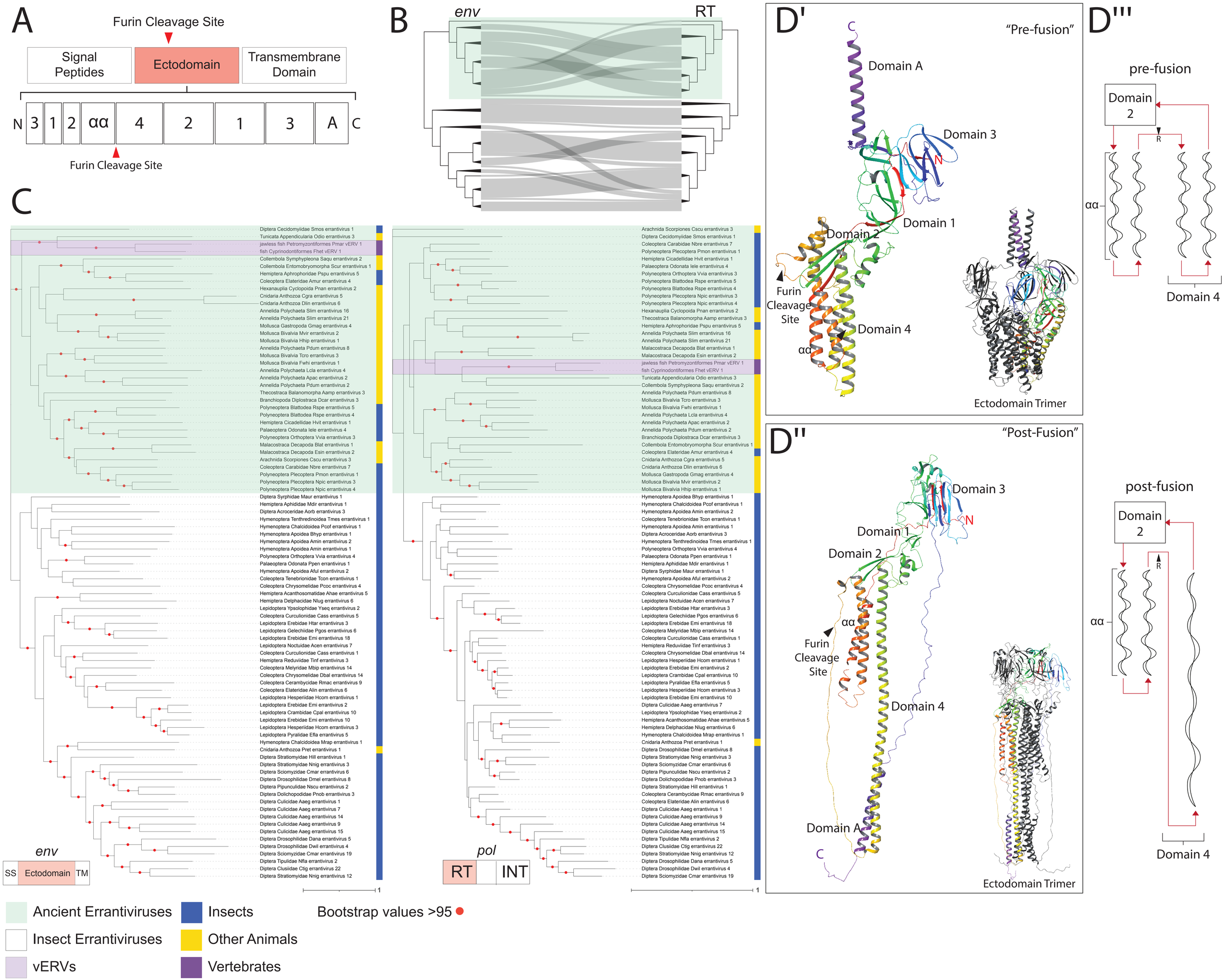




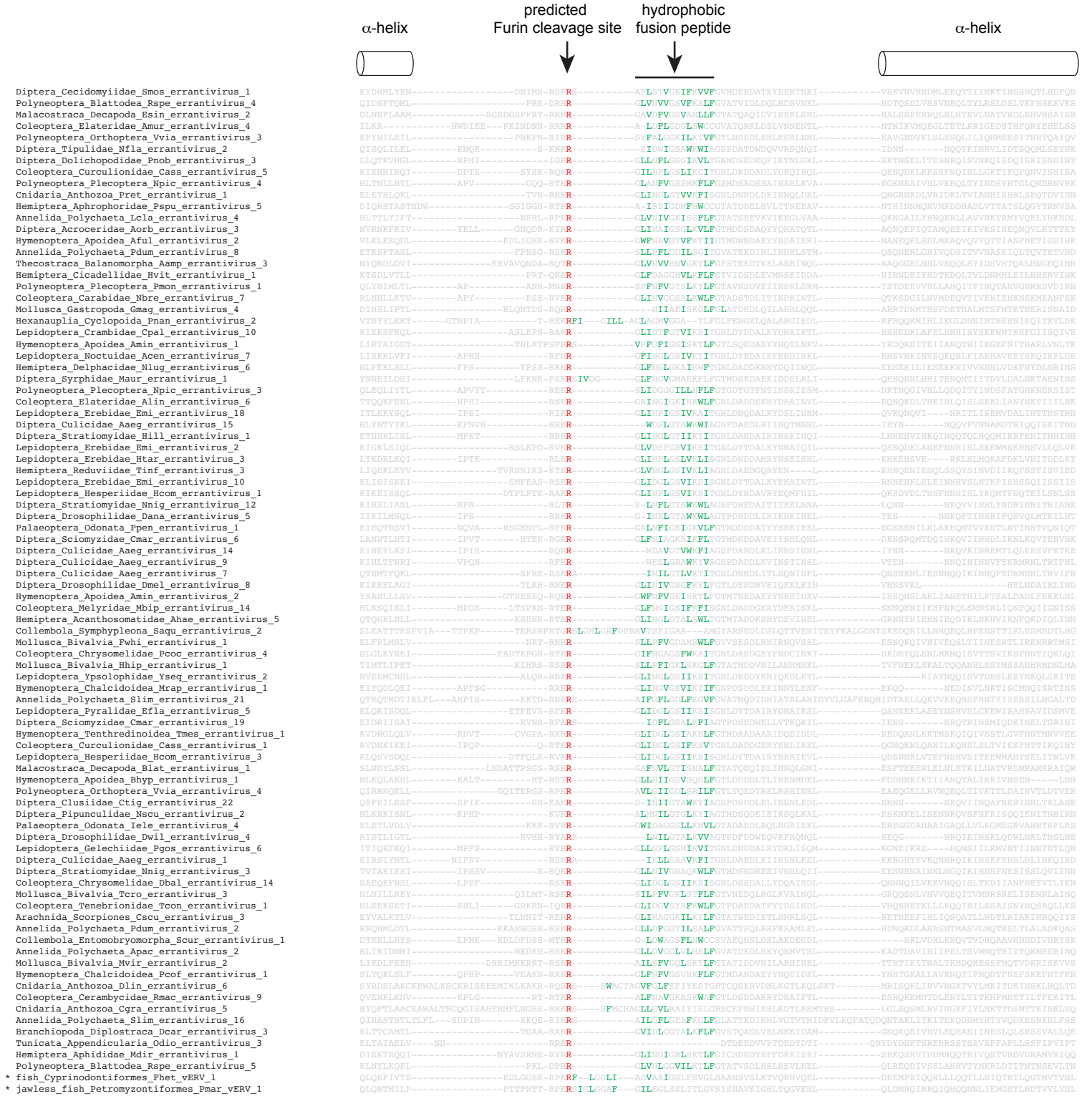
Supplementary Figure S3



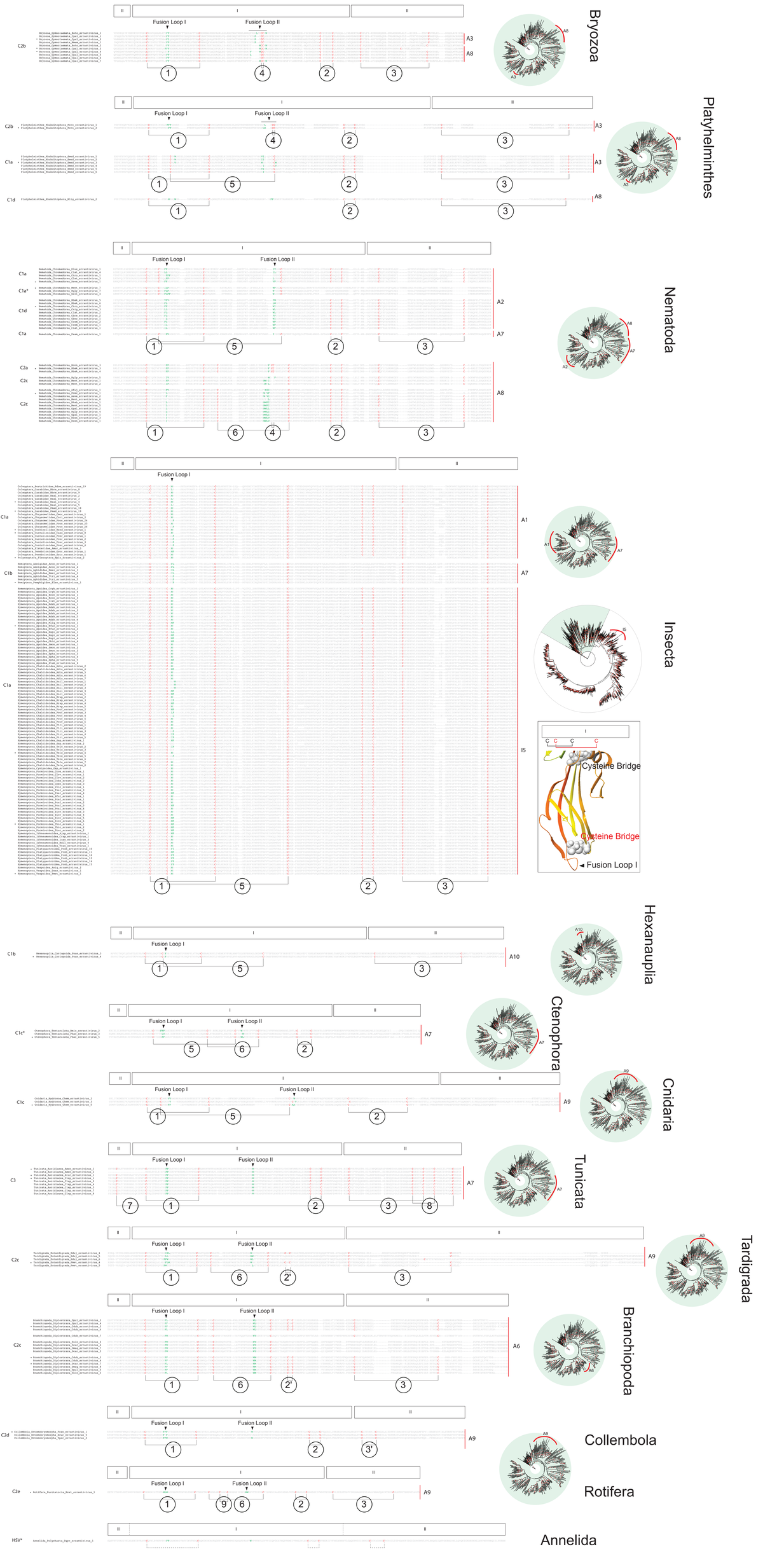
Supplementary Figure S4



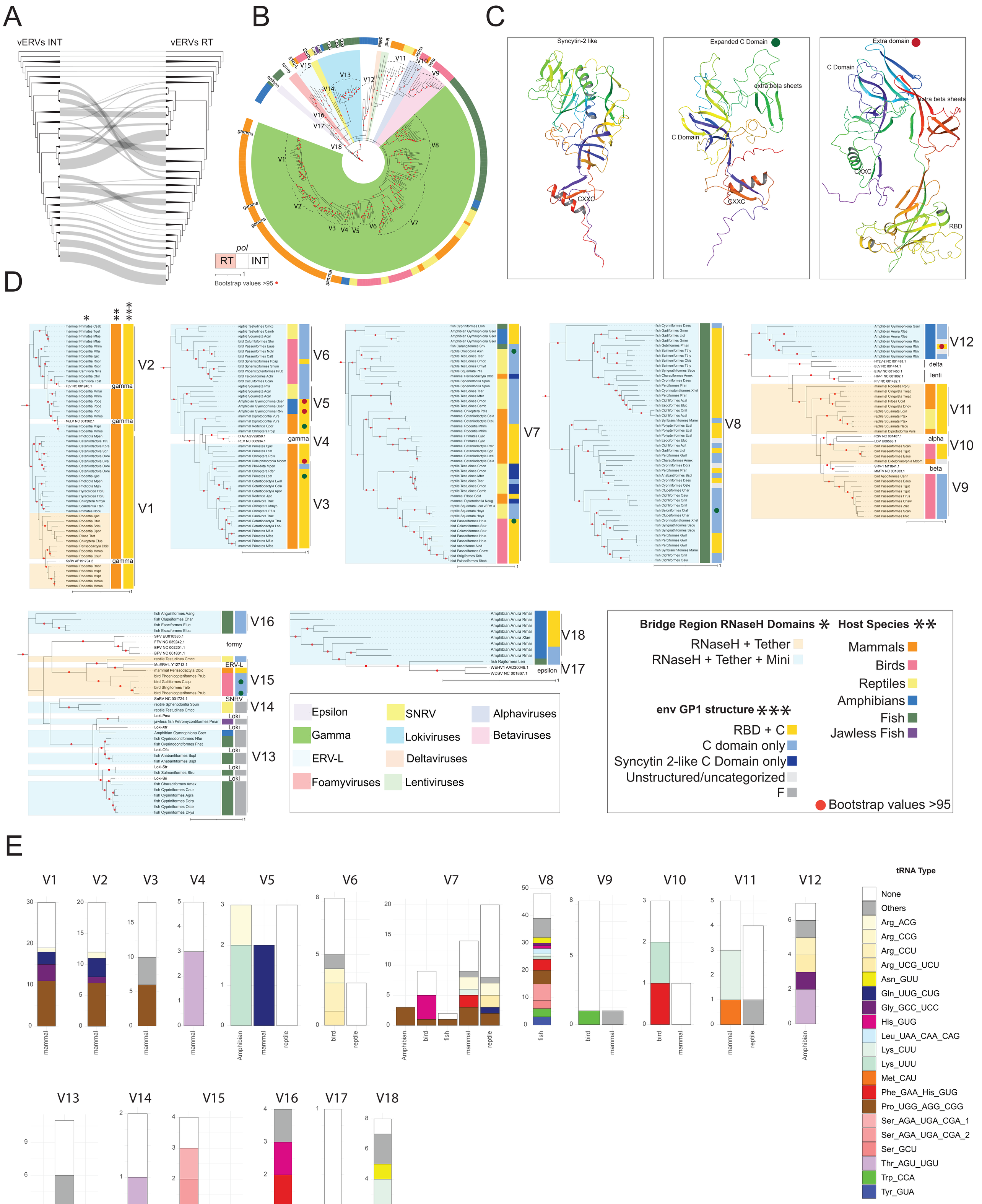
Supplementary Figure S5



Supplementary Figure S6

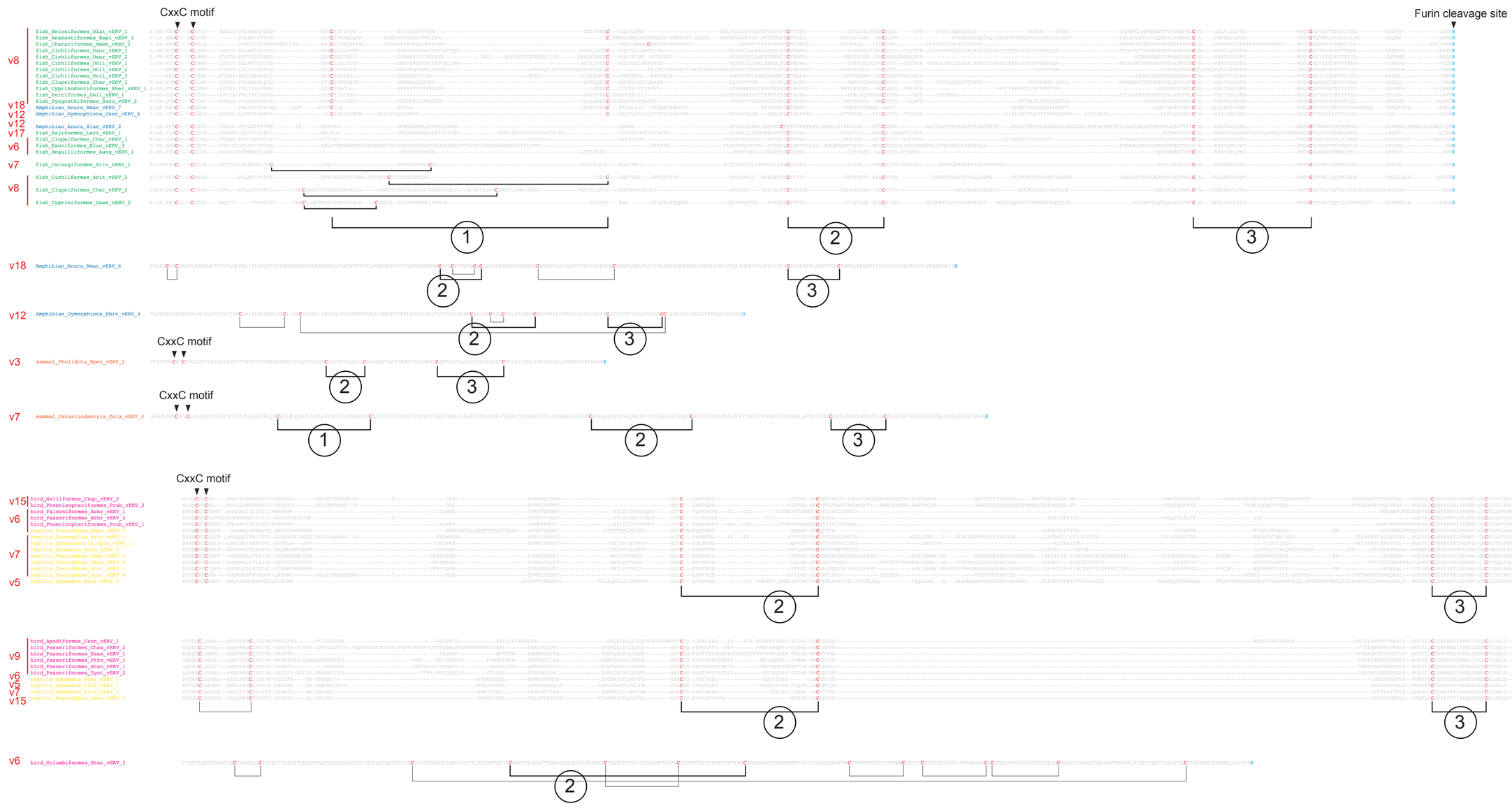


Supplementary Figure S7

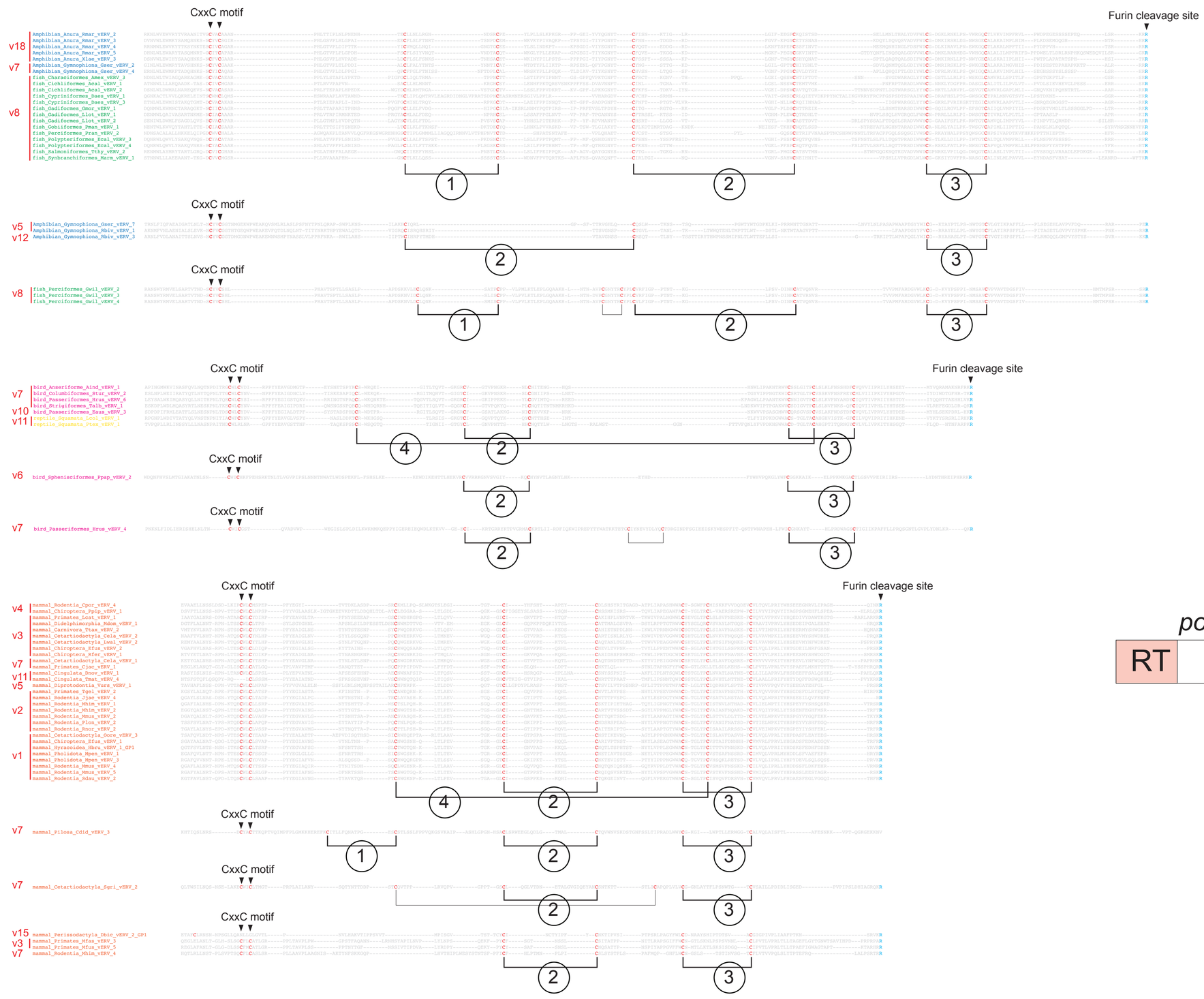


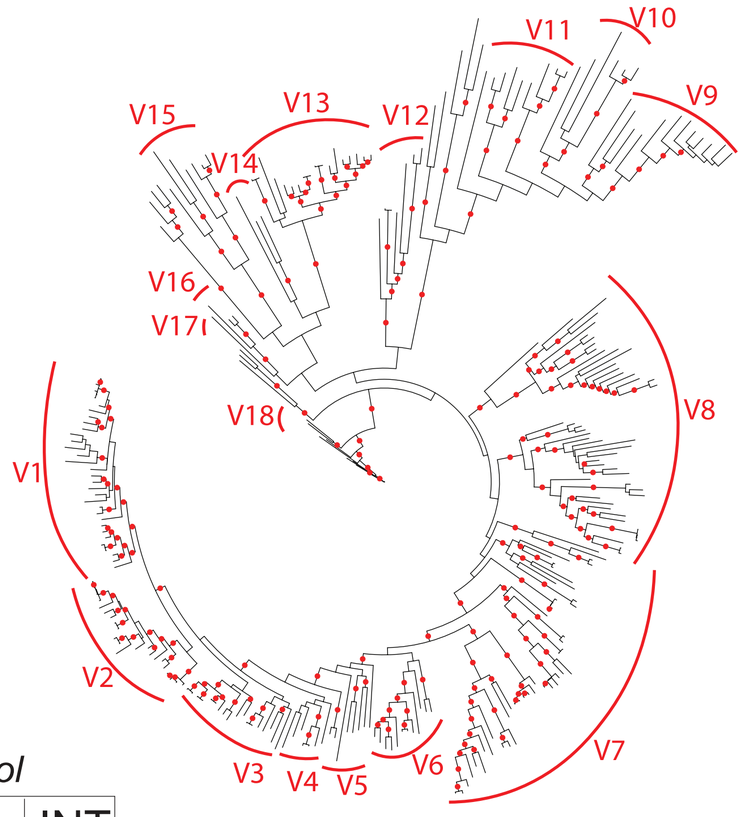
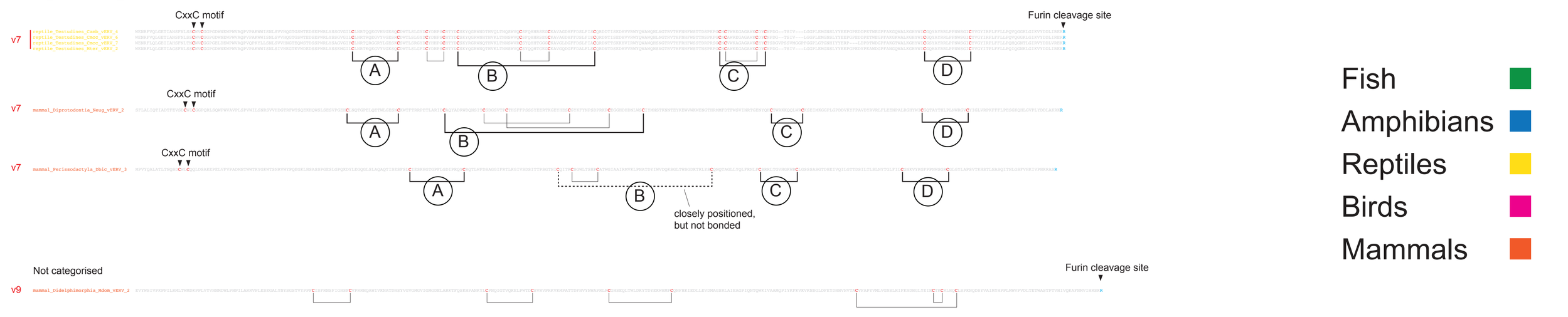
Supplementary Figure S8

C only



RBD + C





Syncytin-2