Pre-Cambrian origin of *envelope*-carrying retrotransposons 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 germline genome. One rare example of ERVs that additionally carry *envelope* genes, are *Ty3/gypsy* errantiviruses in Drosophila. Though they are structurally analogous to 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* in invertebrate 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 and glycoprotein B from herpesviruses, and both types are equally abundant and widespread. 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

Long terminal repeat (LTR) retrotransposons are one of the most abundant classes of transposons that are ubiquitously present in eukaryotic genomes. They replicate by reverse-transcribing the messenger RNA and inserting the DNA copy into the genome. They essentially carry two open reading frames (ORFs), namely, *gag* and *pol*. *gag* encodes for a structural protein that packages RNA into capsids while pol encodes a multi-functional protein that contains the reserve transcriptase, RNase H and integrase domains. *Gag* and *pol* ORFs are flanked by 5' and 3' LTRs that act as a promoter for transcription as well as a primer for reverse transcription. Although intact copies can continue to replicate, individual insertions acquire mutations and degrade overtime once they become inactive. Some retrotransposon remnants remain in the genome to provide functional ORFs and regulatory sequences of host gene expression.

It is widely accepted that retroviruses are evolutionarily sister to LTR retrotransposons among other retroelements (Xiong and Eickbush 1988; Xiong and Eickbush 1990). In addition to *pol* and *gag*, retroviruses additionally carry the *envelope* gene, which encodes a membrane fusion protein essential for infectivity. Although they typically horizontally transmit between individual animals, they can occasionally infect the germline cells of the host, vertically transmit thereafter, and become known as endogenous retroviruses (ERVs). A well-studied example of such endogenization events is found in Koala, in which nearly identical copies of an exogenous lymphoma-causing retrovirus are found in the genomes of geographically distinct Koala populations (Tarlinton et al. 2006; Zheng et al. 2020). Although individual ERV insertions are subject to genetic drift and degrade overtime after they acquire mutations, some insertions remain infectious and continue to make new insertions in the germline genome (Ribet et al. 2008; Miyazawa et al. 2010; Anai et al. 2012; Verneret et al. 2025). However, it is unclear how long such infectious ERVs vertically transmit in a long evolutionary time. When homologous ERVs are found in distantly related species, it can be either seen that they have been inherited from their common ancestor or recently transmitted from one species to another through horizontal transfer. The mechanism of such horizontal transfer events is even less clear although it is often hypothesised that exogenous retroviruses that are borne by one species infect the other species that shares the habitat (Hayward et al. 2020; Simpson et al. 2022; Mottaghinia et al. 2024).

envelope-carrying ERVs are also found in invertebrate genomes, to which no exogenous retroviruses are currently known. Some *gypsy* ERVs or better known as *gypsy* LTR retrotransposons in Drosophila carry an *envelope* gene (Song et al. 1994; Pelisson et al. 2002). These so-called errantiviruses behave like bona fide retroviruses, making virus-like particles in the ovarian somatic cells and infecting the germline cells (Song et al. 1997; Leblanc et al. 2000). Therefore, errantiviruses replicate within the genome of the same animal as other retrotransposons do but through an intra-gonadal infection mechanism. Errantiviruses are widespread in Drosophila genomes and form a monophyletic group among other Drosophila *gypsy* elements based on the phylogeny of the *pol* ORF, suggesting that a subtype of *gypsy* acquired the *envelope* gene in the common ancestor of Drosophila and has been vertically transmitting thereafter (Bargues and Lerat 2017). It is postulated that the

ancient *gypsy* retrotransposon in Drosophila acquired the *envelope* gene from an insect-specific DNA virus, baculovirus, through recombination (Malik et al. 2000; Rohrmann and Karplus 2001). However, *Ty3/gypsy* LTR retrotransposons with *envelope* genes have been also sporadically documented in non-insect animals, such as in nematodes, rotifers, and tunicates, obscuring their true origin (Volff et al. 2004; Rodriguez et al. 2017; Sacco et al. 2022).

Here we comprehensively searched for *envelope*-carrying *Ty3/gypsy* retrotransposons in invertebrate metazoan genomes. We found elements with uninterrupted ORFs flanked by putative LTR sequences in nearly every major phylum, including ancient animals such as cnidarians, ctenophores, and tunicates. Furthermore, most of these elements are present in multiple copies in their respective genomes, indicating recent transposition. Phylogenetic and structural analyses of the *pol* and *envelope* genes show that they have co-evolved with their host species for a long evolutionary time. Therefore, we propose that the acquisition of the *envelope* gene to retrotransposons occurred anciently in metazoans, and they have been vertically transmitting since pre-Cambrian era.

Results

The approach to study the phylogeny and evolution of LTR retrotransposons, is conventionally to mine the genome for sequences which share homology to retroviral *gag*, *pol*, or *envelope (env)* genes. Prior studies on ERVs and retrotransposons have assembled trees based on reverse transcriptase (RT) or Integrase (IN) from *pol* to obtain insights (Xiong and Eickbush 1990; Wang and Han 2021). One can also investigate synteny of loci between different hosts, estimating an age of origin for those elements (Bonnaud et al. 2005; Ueda et al. 2020).

This study aims to investigate the evolution of *env* genes acquired by active *Ty3/gypsy* retrotransposons, and to explore structural variations within those elements. The conventional approach, therefore, needs to be altered to approach this. We only consider intact copies of *env*-containing retrotransposons; those that carry uninterrupted open reading frames (ORFs) for all three retroviral proteins (Fig 1A). We directed our search within a diverse array of hosts, and where possible, focused on such elements that are multiplied in the genome to understand relationships of currently active elements. This approach enables us to study structure and phylogeny more accurately, as structural predictions will be more confident on non-degraded elements. Additionally, the study enables the discovery of new elements that are active today, potentially providing insights on more recent acquisition events. The study of syntenic loci here is inappropriate, as the genomic location of active transposons will likely never be the same in evolutionarily distant hosts. Like most studies of repeat elements, the sequences are highly diverse, and we therefore supplement phylogenetic evidence with structural predictions.

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). The first tBLASTn search identified elements outside Diptera, 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 (Table S1, summarised in Fig 1). The majority of the newly identified genomic copies carry putative LTRs and tRNA primer binding sequences (1151 out of 1512 elements, Table S1), suggesting their recent integrations. Additionally, most of the annotated elements sit within contigs larger than 0.1 mega bases (1434 elements, Table S1), indicating their true origins in their respective genomes. We ran multiple sequence alignments using the conserved Integrase (IN) and Reverse Transcriptase (RT) domains of Pol and assembled maximum probability trees to infer phylogenetic relationships of the identified elements. ORFs of annotated Drosophila *gypsy* errantiviruses were also included in the trees to aid classifications. We performed a similar analysis on vertebrate genomes using known vertebrate retroviral *gag*, *pol*, and *env* sequences as bait in addition to Drosophila errantiviruses. We did not identify any intact ERVs in vertebrate

genomes that were homologous to invertebrate elements, nor did we find elements in invertebrate genomes that predicted homology to vertebrate elements. There is therefore a clear phylogenetic separation between invertebrate elements and vertebrate ERVs, despite structural similarities. We refer to the genomic copies that are homologous to Drosophila *gypsy* as errantiviruses in this study. The investigation on the vertebrate ERVs will be reported elsewhere, as it within a different scope to errantiviruses.

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

We found that intact copies of errantiviruses carrying all three retroviral ORFs 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). Many carry ENV proteins that are homologous to glycoprotein F1, commonly found in paramyxoviruses, such as Nipah, Hendra and Respiratory Syncytial Viruses (McLellan et al. 2013; Dang et al. 2019), but is also carried in insect baculoviruses (and Drosophila errantiviruses) (Malik et al. 2000; Rohrmann and Karplus 2001). Other elements we identified 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). 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. 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, Table S1), a hallmark feature of recent transposition (Cordaux and Batzer 2009; Gagnier et al. 2019), indicating that most elements that we have annotated likely behave as bona-fide retrotransposons (Fig 1C). 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 phylogenetic tree assembled on the intact errantivirus *pol* genes shows congruence, meaning that the tree assembled on *pol* RT closely reflects that of *pol* IN. Errantiviruses within most clades in one tree are represented in the same or closely related clades in the other (Fig S2B). Importantly, the errantiviruses found in phyla outside of insects, are in a different clade to those found within, which we call “ancient errantiviruses” and “insect errantiviruses”, respectively (Fig 2 and S2A). Notably however, few elements found in insects sit within clades of ancient errantiviruses, rather than other insect ones (Fig 2). 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 (Bargues and Lerat 2017) (Fig 2). Though, one element from a cnidarian species *Paleozoanthus reticulatus* clusters together with insect elements, in a clade adjacent to DM176 and Idefix (Fig 2), its origin is unclear as the element is found on a short genomic contig (Table S1).

Many of the newly identified "insect errantiviruses" sit in separate clades to the Annotated Drosophila *gypsy* elements (Fig 2, tree positions I3 - I6 in Fig S2A). These "insect errantivirus" 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 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 *gypsy* elements to carry either Serine or Lysine tRNA PBS (Terzian et al. 2001; Stefanov et al. 2012). Notably, most discrete clades comprised elements from the same insect orders, suggesting long term vertical transmission and divergence within related species (Fig S3).

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 S4C). 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 S4B). Discrete clades are consistently made up of elements in hosts from the same, or closely related phyla (Fig S4B), for example, tree position A5 consists of elements found in Thecostraca, Annelida and Mollusca, each of which form their own discrete clade (Fig S4B). This indicates that they were either present in their common ancestor or inherited more recently within the same animal group. Given that neighbouring clades consist of elements found in related animal phyla, it is more likely that they are ancient in nature. For example, clades that consist of insect elements are flanked by clades of other arthropod elements (Fig S4B, tree position A2). Clades containing Annelida elements often cluster with or next to mollusc elements (Fig S4B, tree positions A5-A7). This congruence between clades and species is highly indicative of long-term vertical transfer, as known in non *env*-containing retrotransposons (Sormacheva et al. 2012; Wanner and Faulk 2021). Within larger clades containing a wider variety of host orders such as tree position A9, the separation is still evident and strongly supported clades only consist of elements from the same order of host (Fig S4B). Furthermore, clades formed on the trees using *pol* proteins tend to carry the same type of *env* (Fig 3A and S4A). For example, within A6 in the "ancient errantivirus" group, elements that carry HSV/gB-type *env* are in a separate clade to those carrying F-type *env* (Fig S4B). Though some clades appear to be a mix of both, elements with HSV/gB-type *env* still tend to cluster together, indicating that they have stayed with *pol* for long evolutionary periods of time.

<u>ENV glycoproteins carried by errantiviruses are highly structured, and their divergence follows that of *pol*</u>

As previously mentioned, we found that errantiviruses can carry HSV/gB-type or glycoprotein F-type *env*. While most HSV/gB-type *env* is found in the "ancient errantivirus" clade, one clade in "insect errantiviruses" exclusively consists of hymenopteran errantiviruses which carry both HSV/gB-type and F-type *env* genes (Fig 3A-B). 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 and F-type *env* have been together with their *pol* genes at least throughout the evolution of hymenopteran species (Fig 3B).

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 (Jumper et al. 2021).

F glycoproteins analysed are highly structured and consistently form four domains that are highly conserved across all elements (Fig 4A). They also predicted a highly conserved Furin cleavage site and hydrophobic fusion peptides in the ectodomain (Fig 4A 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 (McLellan et al. 2013) (Fig 4C). Most errantiviruses also have an α helix at the C terminus of the ectodomain, referred to as "Domain A" (Fig 4A and 4C). The predicted 3D trimeric structure also reveals two potential configurations of the glycoprotein, based on domain 4: either a long α helix, or two shorter helices. These long and short barrel structures resemble pre- and post-fusion structures of glycoprotein F described in prior studies (McLellan et al. 2013). 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*-F, 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 4B). Importantly, the ENV ectodomains found in "ancient errantiviruses" cluster separately to those found in "insect errantiviruses", including those insect elements that present themselves with ancient errantiviruses in the RT tree (Fig 4B). This suggests that *env*-F 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 5A). Like F, the trimeric structures of HSV/gB-type ENV in errantiviruses are also highly structured. They carry five 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 5B). 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 (Heldwein et al. 2006; Connolly et al. 2011). Notably, errantiviruses found in insect orders that carry HSV/gB-type *env*, only carry one of these fusion loops (Fig S6). The structure of the HSV/gB-type ENV carried by these errantiviruses are stabilized by disulphide bonds between cysteine residues, which we refer to as cysteine bridges (Lopper and Compton 2002; Vollmer et al. 2020). These bridges are highly conserved but appear to vary the most within domains I and II of the HSV/gB ectodomain (Fig 5C and S6). Additionally, in some cases, a cysteine bridge within domain I is located near the fusion loops (Fig 5C 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 (Tree positions A3 and A9 in fig S4B) all carry configuration 'C1c', despite being found in different clades on the tree, even though they each cluster themselves separately from other elements (Fig S4B, 5D and S6). Similarly, errantiviruses carrying HSV/gB-type *env* 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 5D 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 5D). Though, in those cases, they appear to continue to separate based on *pol* RT (Fig S7). For example, Nematoda elements with configuration 'C2a' and 'C2c' are found in a different clade to those with 'C1a' and 'C1d' (Fig 5D). Although, it should be noted that elements with 'C1a' in Nematoda are found in two different clades (Fig 5D).

As mentioned previously, the clades that form on RT mostly separate elements that carry HSV/gB-type ENV from F-type ENV within discrete clades (Fig S4B). 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-type ENV, 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 F-type ENV (Fig 3B). 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 3B). Some species, such as a hymenopteran *Lasioglossum lativentre*, carry elements with HSV/gB-type, as well as those with F-type ENV (Fig 3B and Table S1). Despite being found in the genome of the same species, the *pol* of errantiviruses with F- and HSV/gB-type ENV do not stay together, further supporting their more ancient origin (Fig 3B). It should be noted that some clades, such as ones in tree position A2 within "ancient

errantiviruses", have elements that carry F- and HSV/gB-type ENV sitting together within the same discrete clade, suggestive of more recent recombination (Fig S4B).

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

To further characterize the errantiviruses identified in this study, we analysed the region between the IN 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" (Malik and Eickbush 2001). We therefore characterised structures within the bridge region using the 3D fold arrangement of the core β-sheet structure using Alphafold 2.

We found various structures within the bridge region, mostly consisting of core β-sheets resembling the 3D fold arrangement of RNase H domains but we also found other uncommon substructures (Fig 6A). Almost all errantiviruses that sit within the "insect errantivirus" group have duplicated the RNase H domain similarly to vertebrate retroviruses, with the C terminal copy presenting itself as a "pseudo" RNase H domain (Fig 6A-B). This "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 6A).

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 tf1 pol* (Fig 6B-C), indicating it as a feature of highly ancient transposons.

The other structures ("Mini" domains 2, 3 and "Pseudo" domain 2) present themselves less frequently as outliers, and are restricted to specific clades, or even solitary elements (highlighted in yellow and dark blue in Fig 6B). One notable outlier is an insect element within the "ancient errantivirus" group that carries a variation of a "pseudo" RNase H domain (position 7, noted as "pseudo" domain 2 in Fig 6). This same domain can be found in a phylogenetic group in the "insect errantivirus" group (position 3 in Fig 6B). Furthermore, some elements in this "insect errantivirus" group carry a "mini" domain as well (position 3 in Fig 6B). Another is that some elements within the "insect errantivirus" group only carry a solitary RNase H domain (positions 4 and 5 in Fig 6B). It should be noted that these outliers are mostly specific to small phylogenetic groups/clades, and likely also arose anciently.

Active errantiviruses are sourced from various lineages of *Ty3/gypsy* retrotransposons

To further investigate the origins of *env*-containing errantiviruses, we assembled a phylogenetic tree from the RT of representative errantiviruses and included non-*env*-containing retrotransposons that fall under the broader

umbrella of "*Ty3/gypsy*" elements. This includes transposons such as *MAG* and *Tor*, and Chromoviruses such as *Maggy*, which are annotated in the Gyspy DB (https://gydb.org). We expected that *env*-containing errantiviruses would form a monophyletic clade in the tree if they arose from a specific non-*env*-containing *Ty3/gypsy* element.

We firstly find that most non-*env*-containing *Ty3/gypsy* elements sit in the 'ancient errantivirus' clade, with the exception of the *412/MDG1* group - specifically the Drosophila *gypsy* elements *BLOOD* and *Gypsy-28_Dan* - which cluster with Drosophila and other insect errantiviruses, consistent with a previous study (Bargues and Lerat 2017) (Fig 7). Additionally, we find that the non-*env*-containing elements rarely regularly form discrete clades without *env*-containing errantiviruses. For example, *Tor4b* is found sister to Tunicata elements carrying F-type *env*, and *Cer1* was found sister to a Platyhelminthes errantivirus carrying HSV/gB-type *env*. Interestingly, Chromoviridae, which are found in Fungi and Plants and are representative of highly ancient, widespread *Ty3/gypsy* elements (Gorinsek et al. 2004; Kordis 2005), are found to cluster within the tree, nested with "ancient" errantivirus clades (Fig 7). They sit within a clade that includes Tunicata errantivirus carrying HSV/gB-type *env*.

The annotated *Ty3/gypsy* elements present themselves in the tree, spread between and within different "ancient" errantiviruses from different hosts, with little pattern to their spread regarding host or *env* type. Notably, a clade of errantiviruses from Cnidaria, Mollusca and Annelida hosts all carry the chromodomain, and the clade sits basally to known chromoviruses, but sister to non-chromodomain-containing errantiviruses (Fig 7, Fig S7). We also observe outliers, where only one or two errantiviruses carry chromodomains within the whole clade. For example, Annelida_Polychaeta_Dgyr_Errantivirus_1 contains a chromodomain, but is located in a clade with non-chromodomain-containing errantiviruses (Fig S7). These observations rule out a single origin of *env*-containing errantiviruses and instead suggest that multiple lineages of ancient *Ty3/gypsy* elements contributed to the current distribution and diversity of errantiviruses.

Discussion

We found copies of *env*-containing *Ty3/gypsy* retrotransposons (errantiviruses) in the genomes of nearly every major metazoan order. They are intact, carrying uninterrupted ORFs of all three retroviral proteins and are flanked by LTRs. Additionally, most errantiviruses identified in this study have multiple copies in their respective genomes, likely indicating recent transposition. When we further characterized these elements, we found that they carry different ENV proteins; ENV resembling glycoprotein F from paramyxoviruses, and ENV resembling glycoprotein B from herpesviruses. Despite great divergences in the primary sequence, the structural domains of these proteins are highly conserved. We therefore propose that the term 'errantivirus' should be used to refer to such *env*-containing *Ty3/gypsy* elements found outside *Drosophila* as well.

We found that these errantiviruses are highly ancient, being found in animals such as cnidarians, ctenophores and tunicates – lineages that diverged early in metazoan evolution. We identified an "ancient errantivirus" group, which sits separately to elements found in insects (Fig 2). These ancient elements are also structurally distinct, with the bridge region between RT and IN of *pol* resembling that of yeast *Ty3* retrotransposons, rather than that of insects. Additionally, *pol* genes of non-*env*-containing *Ty3/gypsy* elements from diverse eukaryotic taxa form clades with the "ancient errantivirus" group (Fig 7). Although these patterns are consistent with their ancient nature, the specific origins of *env*-containing errantiviruses in metazoans remain elusive. Errantiviruses from the same or closely related phyla tend to cluster together in the *pol* RT phylogenetic tree (Fig S4). However, the pattern becomes convoluted when highly ancient errantiviruses are compared, as elements from distantly related phyla often cluster together without representation from intermediate animal lineages (Fig S4 and Fig 7). Many of these clades tend to have longer branches, suggesting that they are distantly related and diverged a long time ago. For example, within tree position A10 (Fig S4), 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.

The errantiviruses we identified form multiple clades with previously characterized non-*env*-carrying *Ty3/gyspy* elements in the *pol* tree (Fig 7). Additionally, some errantiviruses carry chromodomains, but do not always cluster together with other chromodomain carrying retrotransposons. This pattern could be explained by different lineages of *Ty3/gyspy* frequently acquiring *env* from other errantiviruses. This would explain annotated *Ty3/gyspy* elements forming clades with both HSV/gB- and F-type *env*-containing errantiviruses. Another explanation could be that widespread horizontal transfer is occurring between distantly related species. As the species diverge from each other, different types of *env* could be acquired, lost and regained multiple times. However, neither of these explanations account for the congruence seen between *pol* and the host, at least at the

level of animal classes and orders (Fig S4), nor the structural divergence of HSV/gB-type ENV (Fig 5 and S6). The spread of errantiviruses is indicative of multiple lineages of retrotransposon contributing to the current picture of errantiviruses, but the congruence supports their ancient nature. One plausible explanation is that different lineages of *Ty3/gyspy* within the last common ancestor acquired *env*, which were inherited and diverged with the host species, infrequently recombining with other elements within the same/similar hosts.

Acquisition of ENV proteins by errantiviruses

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 (Malik et al. 2000; Rohrmann and Karplus 2001). While *Ty3/gypsy* 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* indicates that they have likely co-evolved (Fig 4). This, along with the presence of errantiviruses in ancient host orders such as Cnidaria, Ctenophora and Tunicata, suggests that *env*-carrying *Ty3/gypsy* 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 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* genes with HSV/gB-type *env* mostly cluster within their own clades and are not mixed with *pol* with F-type *env*, 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*, likely occurred in the last common ancestor of metazoans.

The survival mechanism of errantiviruses in the genome

In Drosophila, errantiviruses occupy the somatic cell niche in the ovaries that surround the germline, and the expression of *env* allows virus-like particles to infect the germline to create new copies (Song et al. 1994). As an adaptation, Drosophila species have duplicated the germline transposon defence mechanism to target errantiviruses in the somatic cells of the gonad (Brennecke et al. 2007). 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 (Ribet et al. 2008). However, these additional ORFs and genes in endogenous retroviral elements are constantly subject to evolutionary pressure from their hosts (Wang and Han 2020; Wang et al. 2020). 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 (Heldwein et al. 2006; Connolly et al. 2011; McLellan et al. 2013). While entry of HSV also requires three other glycoproteins, gD, gH and gL, gB is specifically required for membrane fusion (Johnson et al. 2011; Jambunathan et al. 2021). 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.

In summation, the *env* gene appears to have been anciently acquired by *Ty3/gypsy* elements, and has been retained by those elements, diverging as their host species evolved. Today, we see a snapshot of errantiviruses widespread throughout metazoan genomes, continuing to create new insertions to survive long term.

Materials and methods

Genome assemblies used in the study

Reference genome sequences were retrieved from NCBI and grouped into the following insect orders and other metazoan phyla: Diptera, Lepidoptera, Coleoptera, Hymenoptera, Hemiptera, Polyneoptera, and Palaeoptera for insects; 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. The list of the genome assemblies used can be found in Supplementary Table S4.

Identification of active genomic copies of *Ty3/gypsy* errantiviruses

We ran a total of five iterations of tBLASTn (ncbi-blast-2.9.0) to search for the genomic copies of *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 GAG, POL and ENV protein sequences of the Dipteran *gypsy* errantiviruses from the RepBase (https://www.girinst.org/repbase/) for the first tBLAST search. The bait sequences for all tBLASTn iterations can be found at https://github.com/RippeiHayashi/errantiviruses_2025/tree/main. We ran tBLASTn with default parameters (-evalue 10). We applied different criteria to identify ERVs from species outside insects 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.

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 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 (https://www.benchling.com/). The identity of GAG ORFs were further confirmed by HHpred (https://toolkit.tuebingen.mpg.de/tools/hhpred). 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 945 POL ORFs from the first tBLASTn search 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 152 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 (Sacco et al. 2022). 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 (Dezordi et al. 2020) 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 451 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 34 additional baits to cover all the major clades in the phylogenetic tree for a third tBLASTn search, which identified 61 new elements. We repeated tBLASTn runs five times in total; the fourth and fifth runs used 17 and 7 new baits and identified 49 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*. Secondly, we ran HHpred from hh-suite 3.3.0 (Steinegger et al. 2019) 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 (1386 out of 1512 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. 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 (Wilhelm and Wilhelm 2001). 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 1151 out of 1512 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, S4 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 (Das and Georgiadis 2004), 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) (Engelman and Cherepanov 2014), 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) 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-type and HSV/gB-type 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 Integrase 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 4 and S2, 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 86 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, the RNase H and chromodomains using Alphafold 2

We took representatives of the F-type and HSV/gB-type ENV proteins and the region of POL between RT and Integrase, that we termed the bridge region, using branch definitions from the phylogenetic trees of *pol* RT. We also took the C-terminal half of POL, including the Integrase domain, from representative errantiviruses, investigated in Figure 7. We predicted their structures when chromodomains were detected by hhpred (see

above). 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 (Mirdita et al. 2022). 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). 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). Structures of all predicted chromodomains in POL were confirmed by Alphafold.

Figure legends

Figure 1 – The widespread nature of intact GAG-POL-ENV elements across metazoan genomes. (A) Open reading frame arrangement of intact errantivirus copies. **(B)** Summary of Errantiviruses found in Metazoan phyla with type of ENV protein indicated. **(C)** 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 – Phylogeny of Reverse Transcriptase (RT) from all intact metazoan errantiviruses. Maximum likelihood tree constructed on representative *pol* RT proteins of intact errantiviruses (see methods) and rooted to the midpoint. The "ancient errantiviruses" and "insect errantiviruses" phylogenetic groups are 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.

Figure 3 – *env* classifications of intact metazoan errantiviruses. (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) (**) (purple). Clades containing hymenopteran elements in "Insect errantiviruses" that carry either F or HSV/gB *env* are annotated as 1 and 2. **(B)** 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.

Figure 4 – Structure and phylogeny of glycoprotein F-type ENV found in intact errantiviruses. (A) Domain arrangement of glycoprotein F ectodomain in Errantiviruses. **(B)** A Sankey plot of *pol* RT and glycoprotein F-type ENV ectodomain maximum likelihood trees for representative errantiviruses, showing a phylogenetic congruence between them. Bootstrap values for nodes are given. The "ancient errantivirus" group is highlighted in green. **(C)** Monomeric and trimeric structures of the "pre-fusion configuration" and the "post fusion configuration". Schematic cartoons of F-type ENV ectodomain domain 4 are shown in **(C')** and **(C'')** Respectively. Structures are taken from Diptera_Sciomyzidae_Cmar_errantivirus_19 and Branchipoda_Diplostraca_Dcar_errantivirus_3 for "pre fusion" and "post fusion" structures in **(C)**.

Figure 5 – Structure of glycoprotein HSV/gB-type ENV found in intact errantiviruses. (A) 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-type ENV 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 on structure of

the errantivirus HSV/gB-type ENV ectodomain. Shown are variable cysteine bridges in the errantivirus HSV/gB-type ENV ectodomains, in domain I and II (Indicated with orange bar. Proximity of cysteine bridge to fusion loops of domain I are indicated for structures within C1, C2 and C3. Structures 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-type ENV of different metazoan orders. Positions for cysteine bridge configurations within Nematoda and Platyhelminthes shown on Simple schematic. Exact positions of all cysteine bridges can be found in Supplementary Figure S6.

Figure 6 - RNase H domains found in the bridge region of intact errantiviruses. (A) 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. **(B)** 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). **(C)** The structure of the bridge region of yeast *Schizosaccharomyces pombe Ty3* element *tf1* predicted by Alphafold 2.

Figure 7 – Trees of known *Ty3/gypsy* elements and newly identified Errantiviruses. Midpoint rooted *pol* RT maximum likelihood tree of representative errantiviruses inclusive of non-*env Ty3/gypsy pol* RT. Host Order and *env* type given for each clade, as well as solitary elements on the tree are shown. Errantiviruses containing chromodomains are marked in yellow. The "ancient errantivirus" group is coloured in green and "insect errantiviruses" in black. Known *Ty3/gypsy* elements includes Gypsy Group 1 OSIRIS (Gypsy-21_DAn), Gypsy Group 1 OSVALDO (Dypsy-3_DEl), Gypsy Group 2 BICA (Gypsy-6_DFi), Gypsy Group 2 BLASTOPIA (Gypsy-32_DWil), Gypsy Group 2 MDG3 (INVADER3, MDG3_DM), Gypsy Group 2 MICROPIA (Gypsy-7_DSe), Gypsy Group 2 SACCO (Gypsy-10_DAn), Gypsy Group 3 412/MDG1 (BLOOD, Gypsy-28_Dan), Gypsy Group 3 CHIMPO (Gypsy-7_DAn), Gypsy Group 3 17.6 (DM176, IDEFIX, ZAM), Gypsy Group 3 GYPSY (Gypsy-5_Dgri, Gypsy-Gypsy), Mag (Mag_As, MAG_Bmor), Cer (Cer1, Cer4), Tor (Tor1, Tor2, Tor4b), Athila/Tat (ATHILA0p1), and Chromoviridae (ATHILA0p1, CRM_AAM94350.1, REINA, Saccharomyces_Ty3, SCL213_Galadriel, Skipper*, Spombe_Tf1_AAA35339.1, SUSHI).

Supplementary Figure S1 – Identification of errantiviruses 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 errantiviruses 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, F-type and HSV/gB-type *env* ectodomains.

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) (purple). **(B)** A Sankey plot of *pol* RT and *pol* IN 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 – Insect errantivirus clades. A maximum likelihood tree of *pol* RT for the "insect errantivirus" group in isolation, with host phylogeny and *env* type, and discrete phylogenetic groups (I1-I6) indicated. Host orders (column 1) are annotated for Diptera (red), Lepidoptera (orange), Coleoptera (yellow), Hymenoptera (green), Hemiptera (cyan), Polyneoptera (pink) and Palaeoptera (purple). ENV types (column 2) are annotated for Glycoprotein F (F) (orange), and glycoprotein B (HSV) (purple).

Supplementary Figure S4 – 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) are 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) (purple). **(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 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 ectodomains of representative 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 α helix structures are shown upstream and downstream of the Furin cleavage site.

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 – Presence of chromodomain in "ancient" errantiviruses – Maximum likelyhood tree of RT from representative errantiviruses along with known *Ty3/gypsy* elements. ENV type (*) annotated for Glycoprotein F (F) (orange), and glycoprotein B (HSV/gB) (purple). Chromodomain (**) annotated in purple.

Supplementary Table S1 – Master errantivirus table. Table of all 1512 newly identified *env*-containing errantiviruses 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 S2 – Structural Analysis of errantiviruses. Lists of representative errantiviruses taken for structural analyses guided by Alphafold 2: RNase H domains found in the bridge region of *pol* (RNase H), F-type ENV ectodomain (F), HSV/gB-type ENV ectodomain, with cysteine bridge configuration indicated (HSV), and chromodomains found at the C-terminus of *pol* (chromodomain).

Supplementary Table S3 – Node information for errantiviruses. List of errantiviruses and phylogenetic node where they are found for errantivirus RT and IN (IN-RT Nodes Errantivirus Fig S2) and RT and ENV from representative F-type errantiviruses (RT-ENV Nodes F-type Fig 4). Data was then used for congruence plots.

Supplementary Table S4 – Genome assemblies analysed in this study. List of genome assemblies with species name, size of assembly, class, family and number of tBLASTn hits for errantivirus ORFs. Genomes that we used to annotate errantiviruses are indicated. Tabs are individually detailed for Diptera, Lepidoptera, Coleoptera, Hymenoptera, Hemiptera, Polyneoptera, Palaeoptera, non-insect Arthropoda, non-arthropod Protostomia, Cnidaria, Ctenophora, Porifera, and non-vertebrate Deuterostomia.

Acknowledgements

We acknowledge Julius Brennecke, Kirsten Senti and Alex Greenwood for valuable comments in preparation of the manuscript. This work was 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_2025/tree/main): the code used to annotate and analyse genomic copies of errantiviruses, genomic sequences and open reading frames of the newly identified errantiviruses, pdb files for the predicted structures of ENV proteins, the bridge region and the C-terminal half

of POL, 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

Anai Y, Ochi H, Watanabe S, Nakagawa S, Kawamura M, Gojobori T, Nishigaki K. 2012. Infectious endogenous retroviruses in cats and emergence of recombinant viruses. *J Virol* **86**: 8634-8644.

Bargues N, Lerat E. 2017. Evolutionary history of LTR-retrotransposons among 20 Drosophila species. *Mob DNA* **8**: 7.

Bonnaud B, Beliaeff J, Bouton O, Oriol G, Duret L, Mallet F. 2005. Natural history of the ERVWE1 endogenous retroviral locus. *Retrovirology* **2**: 57.

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* **128**: 1089-1103.

Connolly SA, Jackson JO, Jardetzky TS, Longnecker R. 2011. Fusing structure and function: a structural view of the herpesvirus entry machinery. *Nat Rev Microbiol* **9**: 369-381.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**: 691-703.

Dang HV, Chan YP, Park YJ, Snijder J, Da Silva SC, Vu B, Yan L, Feng YR, Rockx B, Geisbert TW et al. 2019. An antibody against the F glycoprotein inhibits Nipah and Hendra virus infections. *Nat Struct Mol Biol* **26**: 980-987.

Das D, Georgiadis MM. 2004. The crystal structure of the monomeric reverse transcriptase from Moloney murine leukemia virus. *Structure* **12**: 819-829.

Dezordi FZ, Vasconcelos CRDS, Rezende AM, Wallau GL. 2020. 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.

Engelman A, Cherepanov P. 2014. Retroviral Integrase Structure and DNA Recombination Mechanism. *Microbiol Spectr* **2**: 1-22.

Gagnier L, Belancio VP, Mager DL. 2019. Mouse germ line mutations due to retrotransposon insertions. *Mob DNA* **10**: 15.

Gorinsek B, Gubensek F, Kordis D. 2004. Evolutionary genomics of chromoviruses in eukaryotes. *Mol Biol Evol* **21**: 781-798.

Hayward JA, Tachedjian M, Kohl C, Johnson A, Dearnley M, Jesaveluk B, Langer C, Solymosi PD, Hille G, Nitsche A et al. 2020. Infectious KoRV-related retroviruses circulating in Australian bats. *Proc Natl Acad Sci U S A* **117**: 9529-9536.

Heldwein EE, Lou H, Bender FC, Cohen GH, Eisenberg RJ, Harrison SC. 2006. Crystal structure of glycoprotein B from herpes simplex virus 1. *Science* **313**: 217-220.

Jambunathan N, Clark CM, Musarrat F, Chouljenko VN, Rudd J, Kousoulas KG. 2021. 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**.

Johnson DC, Wisner TW, Wright CC. 2011. Herpes simplex virus glycoproteins gB and gD function in a redundant fashion to promote secondary envelopment. *J Virol* **85**: 4910-4926.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**: 583-589.

Kordis D. 2005. A genomic perspective on the chromodomain-containing retrotransposons: Chromoviruses. *Gene* **347**: 161-173.

Leblanc P, Desset S, Giorgi F, Taddei AR, Fausto AM, Mazzini M, Dastugue B, Vaury C. 2000. Life cycle of an endogenous retrovirus, ZAM, in Drosophila melanogaster. *J Virol* **74**: 10658-10669.

Lopper M, Compton T. 2002. Disulfide bond configuration of human cytomegalovirus glycoprotein B. *J Virol* **76**: 6073-6082.

Malik HS, Eickbush TH. 2001. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Res* **11**: 1187-1197.

Malik HS, Henikoff S, Eickbush TH. 2000. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res* **10**: 1307-1318.

McLellan JS, Ray WC, Peeples ME. 2013. Structure and function of respiratory syncytial virus surface glycoproteins. *Curr Top Microbiol Immunol* **372**: 83-104.

Mirdita M, Schutze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2022. ColabFold: making protein folding accessible to all. *Nat Methods* **19**: 679-682.

Miyazawa T, Yoshikawa R, Golder M, Okada M, Stewart H, Palmarini M. 2010. Isolation of an infectious endogenous retrovirus in a proportion of live attenuated vaccines for pets. *J Virol* **84**: 3690-3694.

Mottaghinia S, Stenzel S, Tsangaras K, Nikolaidis N, Laue M, Müller K, Hölscher H, Löber U, McEwen GK, Donnellan SC et al. 2024. A recent gibbon ape leukemia virus germline integration in a rodent from New Guinea. *Proc Natl Acad Sci U S A* **121**: e2220392121.

Pelisson A, Mejlumian L, Robert V, Terzian C, Bucheton A. 2002. Drosophila germline invasion by the endogenous retrovirus gypsy: involvement of the viral env gene. *Insect Biochem Mol Biol* **32**: 1249-1256.

Ribet D, Harper F, Dupressoir A, Dewannieux M, Pierron G, Heidmann T. 2008. An infectious progenitor for the murine IAP retrotransposon: emergence of an intracellular genetic parasite from an ancient retrovirus. *Genome Res* **18**: 597-609.

Rodriguez F, Kenefick AW, Arkhipova IR. 2017. LTR-Retrotransposons from Bdelloid Rotifers Capture Additional ORFs Shared between Highly Diverse Retroelement Types. *Viruses* **9**.

Rohrmann GF, Karplus PA. 2001. Relatedness of baculovirus and gypsy retrotransposon envelope proteins. *BMC Evol Biol* **1**: 1.

Sacco MA, Lau J, Godinez-Vidal D, Kaloshian I. 2022. Non-canonical nematode endogenous retroviruses resulting from RNA virus glycoprotein gene capture by a metavirus. *J Gen Virol* **103**.

Simpson J, Kozak CA, Boso G. 2022. Cross-species transmission of an ancient endogenous retrovirus and convergent co-option of its envelope gene in two mammalian orders. *PLoS Genet* **18**: e1010458.

Song SU, Gerasimova T, Kurkulos M, Boeke JD, Corces VG. 1994. An env-like protein encoded by a Drosophila retroelement: evidence that gypsy is an infectious retrovirus. *Genes Dev* **8**: 2046-2057.

Song SU, Kurkulos M, Boeke JD, Corces VG. 1997. Infection of the germ line by retroviral particles produced in the follicle cells: a possible mechanism for the mobilization of the gypsy retroelement of Drosophila. *Development* **124**: 2789-2798.

Sormacheva I, Smyshlyaev G, Mayorov V, Blinov A, Novikov A, Novikova O. 2012. Vertical evolution and horizontal transfer of CR1 non-LTR retrotransposons and Tc1/mariner DNA transposons in Lepidoptera species. *Mol Biol Evol* **29**: 3685-3702.

Stefanov Y, Salenko V, Glukhov I. 2012. Drosophila errantiviruses. *Mob Genet Elements* **2**: 36-45.

Steinegger M, Meier M, Mirdita M, Vohringer H, Haunsberger SJ, Soding J. 2019. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**: 473.

Tarlinton RE, Meers J, Young PR. 2006. Retroviral invasion of the koala genome. *Nature* **442**: 79-81.

Terzian C, Pélisson A, Bucheton A. 2001. Evolution and phylogeny of insect endogenous retroviruses. *BMC Evol Biol* **1**: 3.

Ueda MT, Kryukov K, Mitsuhashi S, Mitsuhashi H, Imanishi T, Nakagawa S. 2020. Comprehensive genomic analysis reveals dynamic evolution of endogenous retroviruses that code for retroviral-like protein domains. *Mob DNA* **11**: 29.

Verneret M, Leroux C, Faraut T, Navratil V, Lerat E, Turpin J. 2025. A genome-wide study of ruminants uncovers two endogenous retrovirus families recently active in goats. *Mob DNA* **16**: 4.

Volff JN, Lehrach H, Reinhardt R, Chourrout D. 2004. Retroelement dynamics and a novel type of chordate retrovirus-like element in the miniature genome of the tunicate Oikopleura dioica. *Mol Biol Evol* **21**: 2022-2033.

Vollmer B, Pražák V, Vasishtan D, Jefferys EE, Hernandez-Duran A, Vallbracht M, Klupp BG, Mettenleiter TC, Backovic M, Rey FA et al. 2020. The prefusion structure of herpes simplex virus glycoprotein B. *Sci Adv* **6**.

Wang J, Han GZ. 2020. Frequent Retroviral Gene Co-option during the Evolution of Vertebrates. *Mol Biol Evol* **37**: 3232-3242.

-. 2021. A Sister Lineage of Sampled Retroviruses Corroborates the Complex Evolution of Retroviruses. *Mol Biol Evol* **38**: 1031-1039.

Wang W, Zhao H, Han GZ. 2020. Host-Virus Arms Races Drive Elevated Adaptive Evolution in Viral Receptors. *J Virol* **94**.

Wanner NM, Faulk C. 2021. Suggested Absence of Horizontal Transfer of Retrotransposons between Humans and Domestic Mammal Species. *Genes (Basel)* **12**.

Wilhelm M, Wilhelm FX. 2001. Reverse transcription of retroviruses and LTR retrotransposons. *Cell Mol Life Sci* **58**: 1246-1262.

Xiong Y, Eickbush TH. 1988. Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. *Mol Biol Evol* **5**: 675-690.

-. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J* **9**: 3353-3362.

Zheng H, Pan Y, Tang S, Pye GW, Stadler CK, Vogelnest L, Herrin KV, Rideout BA, Switzer WM. 2020. Koala retrovirus diversity, transmissibility, and disease associations. *Retrovirology* **17**: 34.

Figure 1

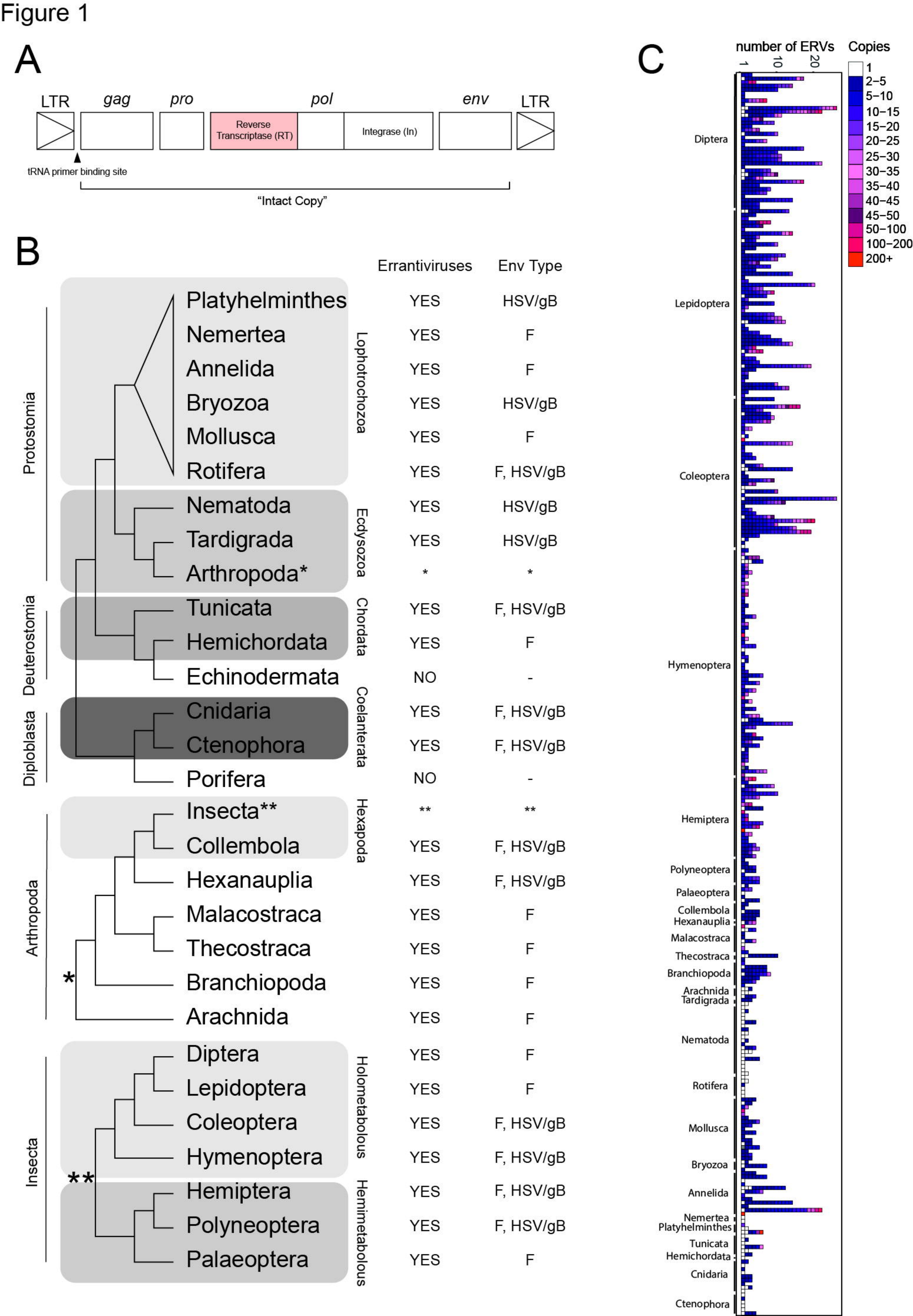

A
LTR
gag
pro
pol
env
LTR
Reverse Transcriptase (RT)
Integrase (In)
tRNA primer binding site
"Intact Copy"
B
Errantiviruses
Env Type
Platyhelminthes YES HSV/gB
Nemertea YES F
Annelida YES F
Bryozoa YES HSV/gB
Mollusca YES F
Rotifera YES F, HSV/gB
Nematoda YES HSV/gB
Tardigrada YES HSV/gB
Arthropoda* * *
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
Protostomia
Deuterostomia
Diploblasta
Arthropoda
Insecta
Lophotrochozoa
Ecdysozoa
Chordata
Coelanterata
Hexapoda
Holometabolous
Hemimetabolous
C
number of ERVs
Copies
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+
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

Figure 2

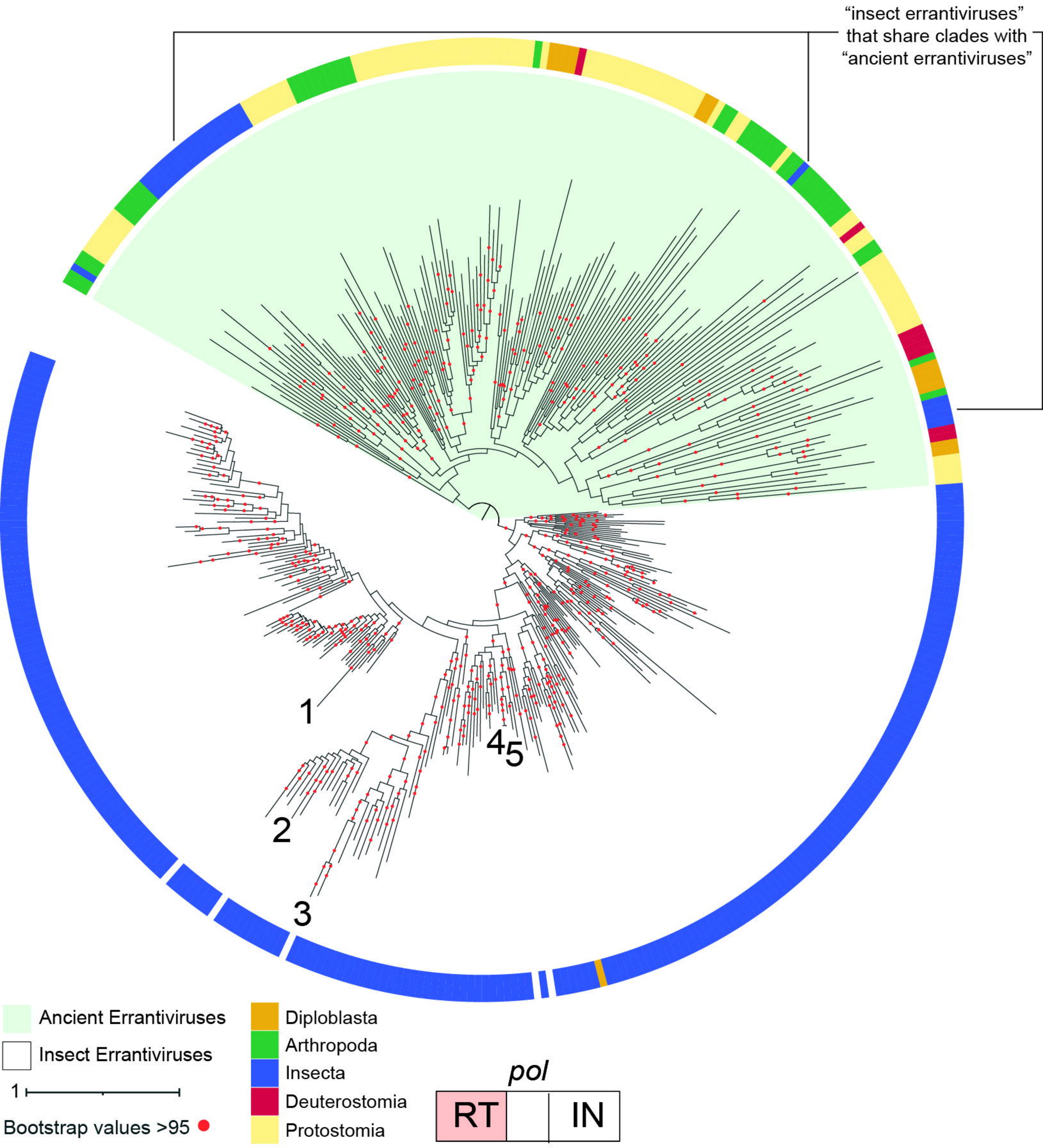
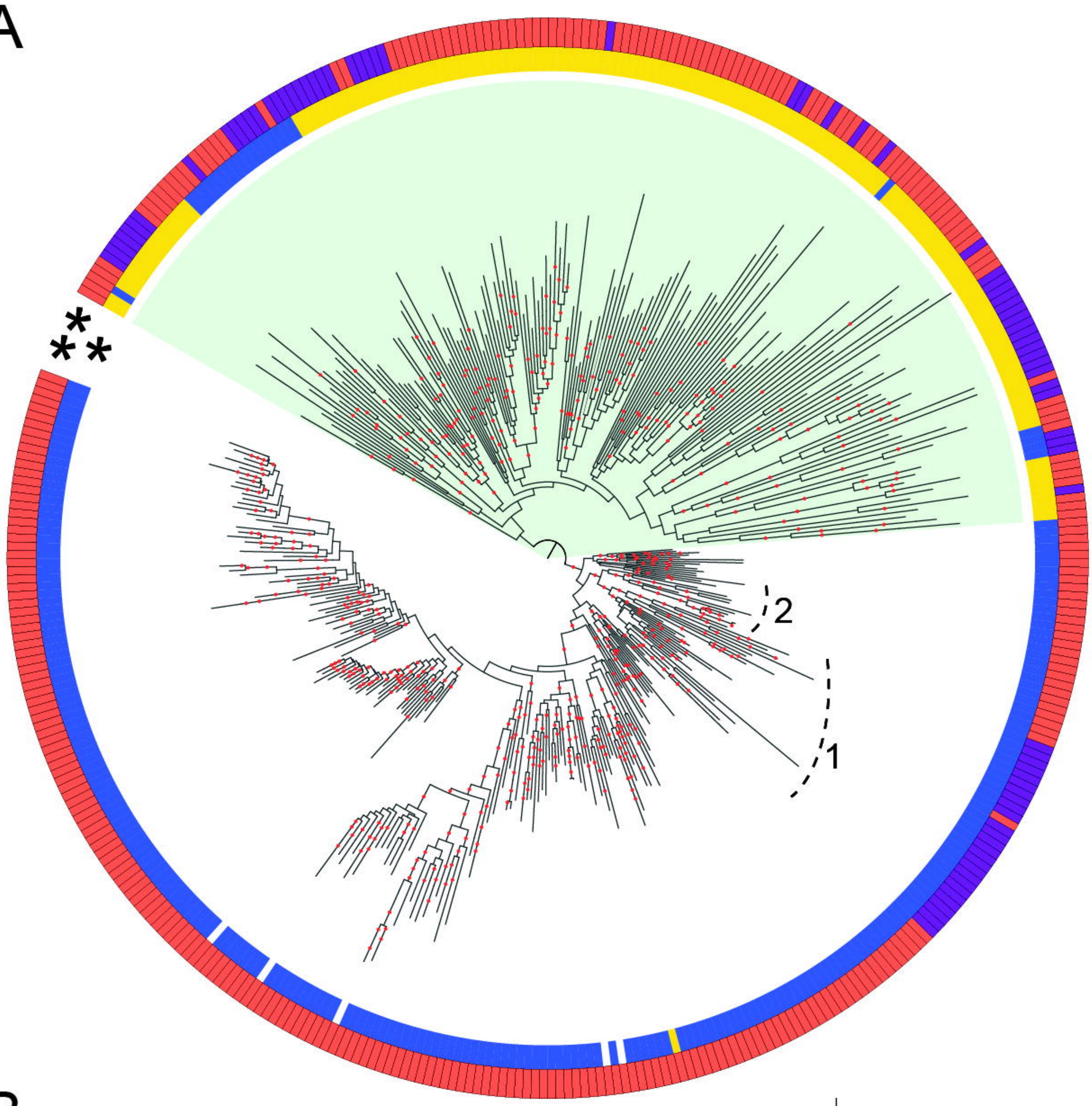
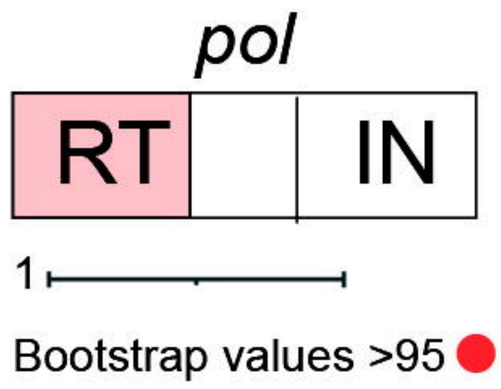

"insect errantiviruses"
that share clades with
"ancient errantiviruses"
1
2
3
4
5
Ancient Errantiviruses
Insect Errantiviruses
1
Bootstrap values >95
Diploblasta
Arthropoda
Insecta
Deuterostomia
Protostomia
pol
RT
IN

Figure 3

A





Bootstrap values >95

*
Insects
Other Animals

**
F
HSV/gB

Ancient Errantiviruses
Insect Errantiviruses

B

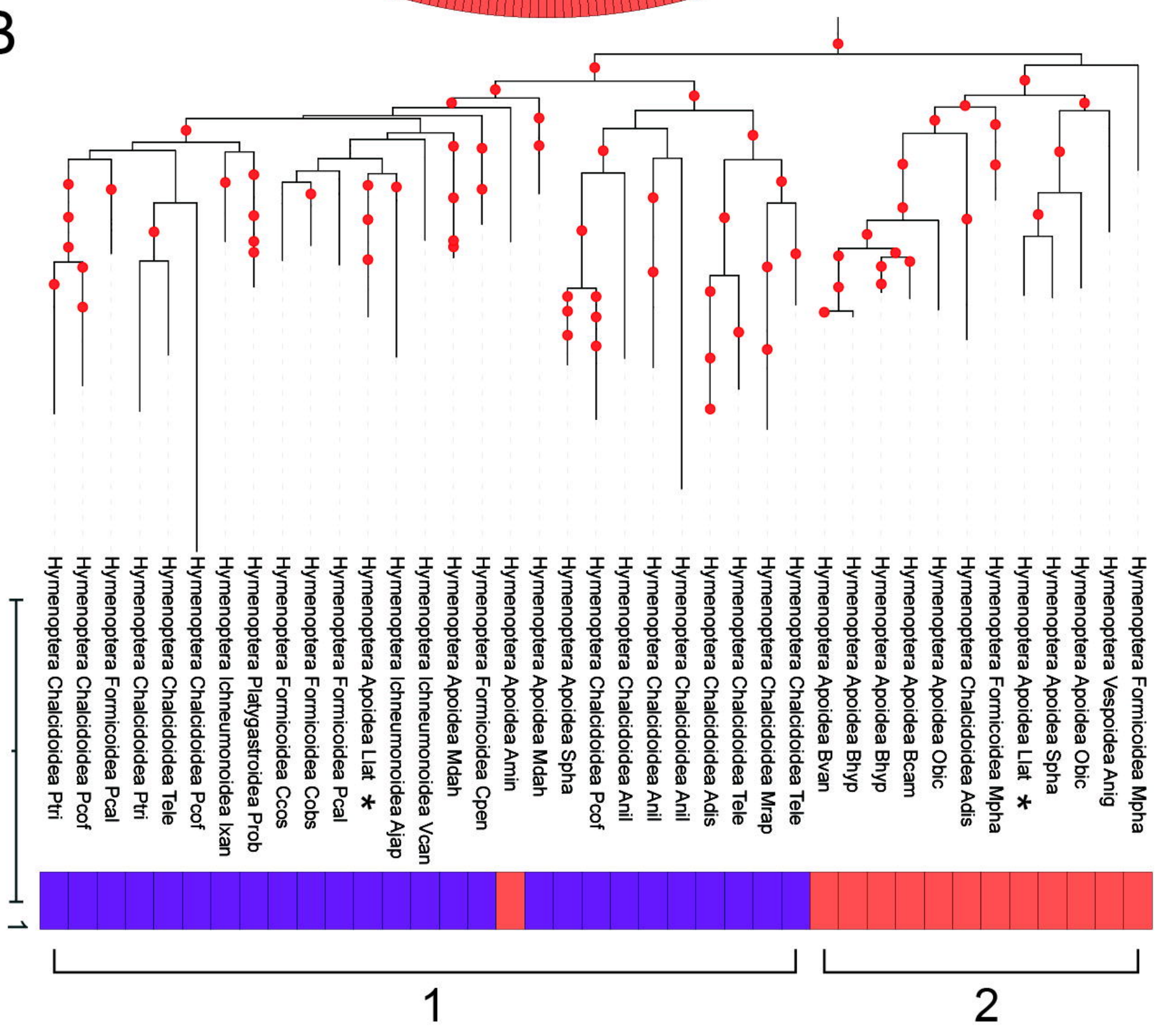
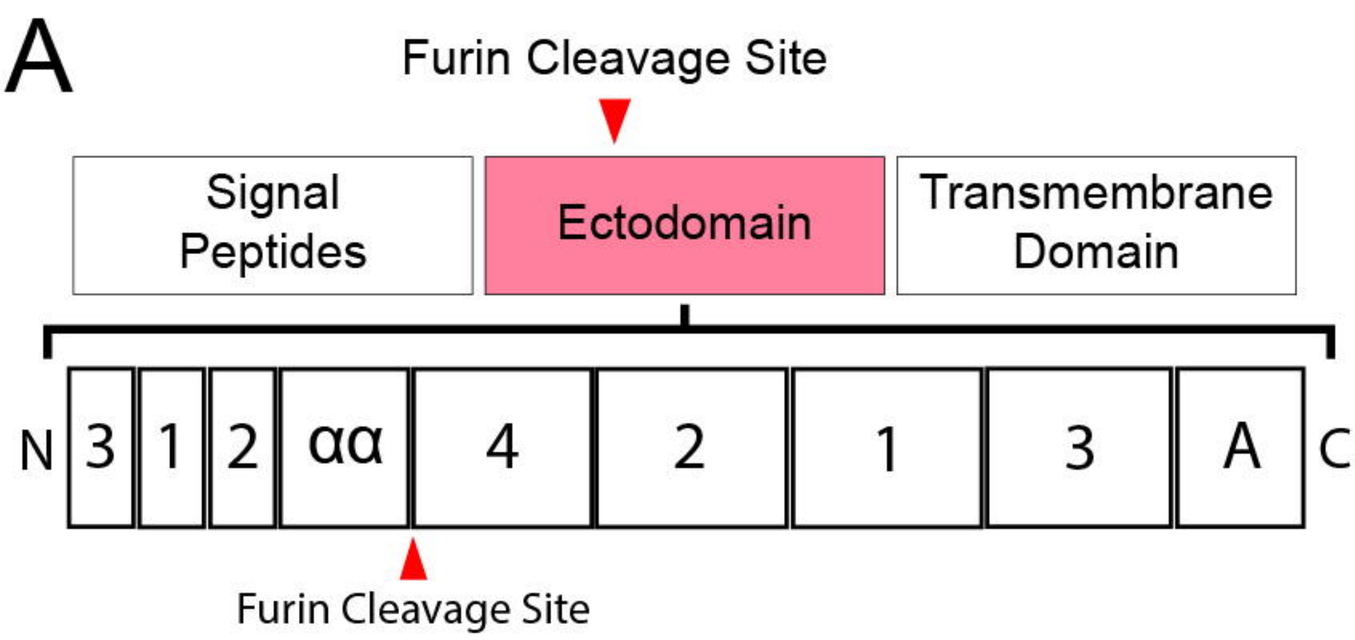
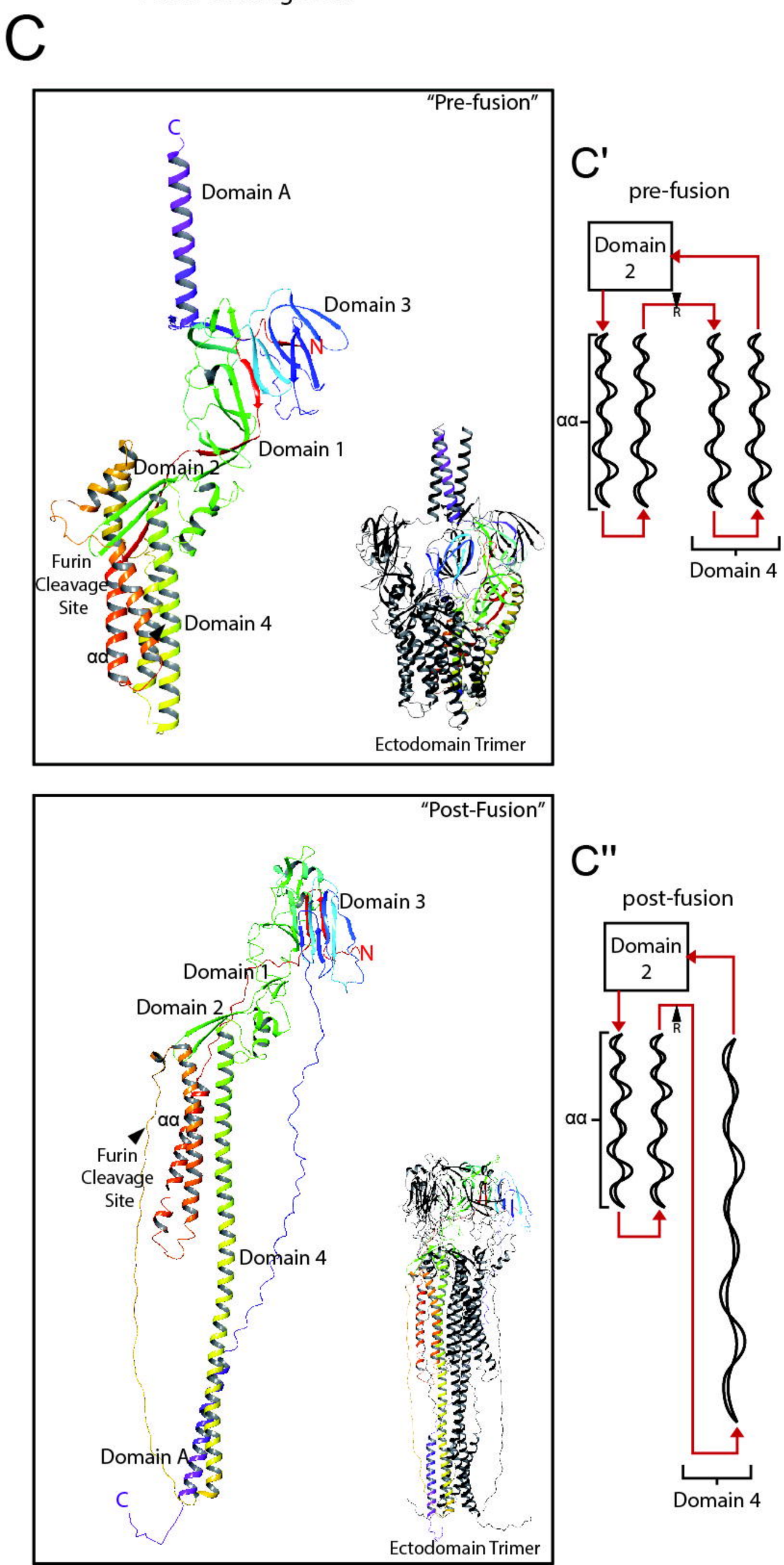


Figure 4


A
Furin Cleavage Site
Signal Peptides
Ectodomain
Transmembrane Domain
N
3
1
2
αα
4
2
1
3
A
C
Furin Cleavage Site


C
"Pre-fusion"
C
Domain A
Domain 3
N
Domain 1
Domain 2
Furin Cleavage Site
Domain 4
αα
Ectodomain Trimer
C'
pre-fusion
Domain 2
R
αα
Domain 4
"Post-Fusion"
Domain 3
N
Domain 1
Domain 2
αα
Furin Cleavage Site
Domain 4
Domain A
C
Ectodomain Trimer
C"
post-fusion
Domain 2
R
αα
Domain 4

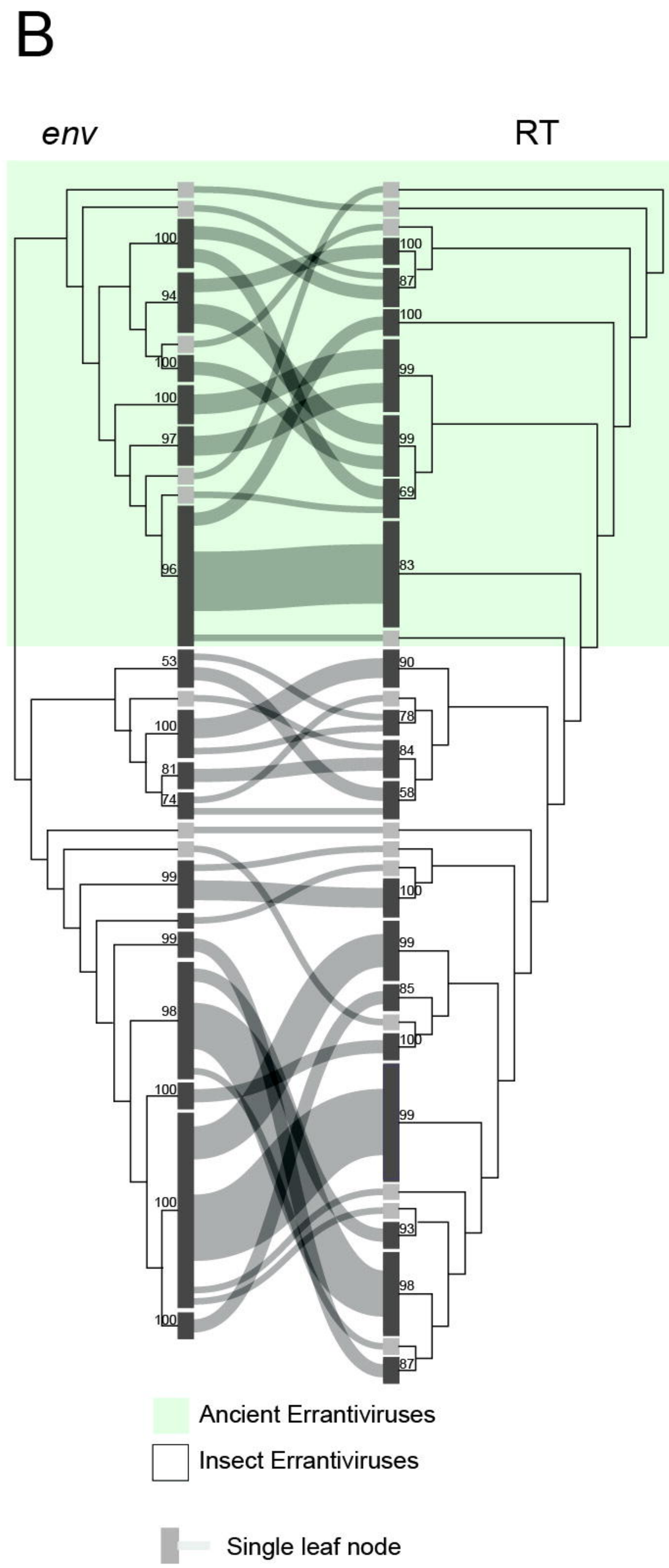

B
env
RT
Ancient Errantiviruses
Insect Errantiviruses
Single leaf node

Figure 5

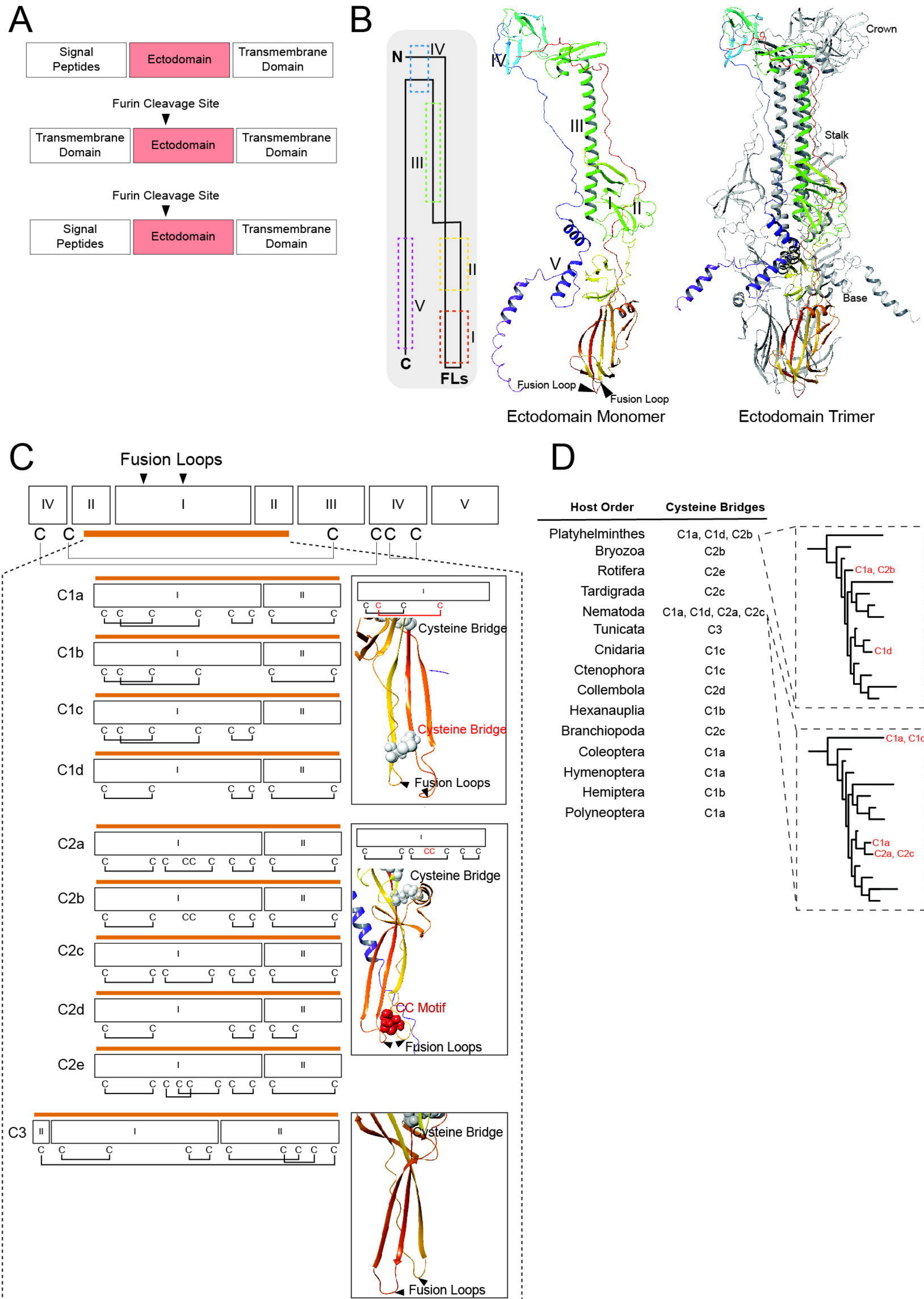
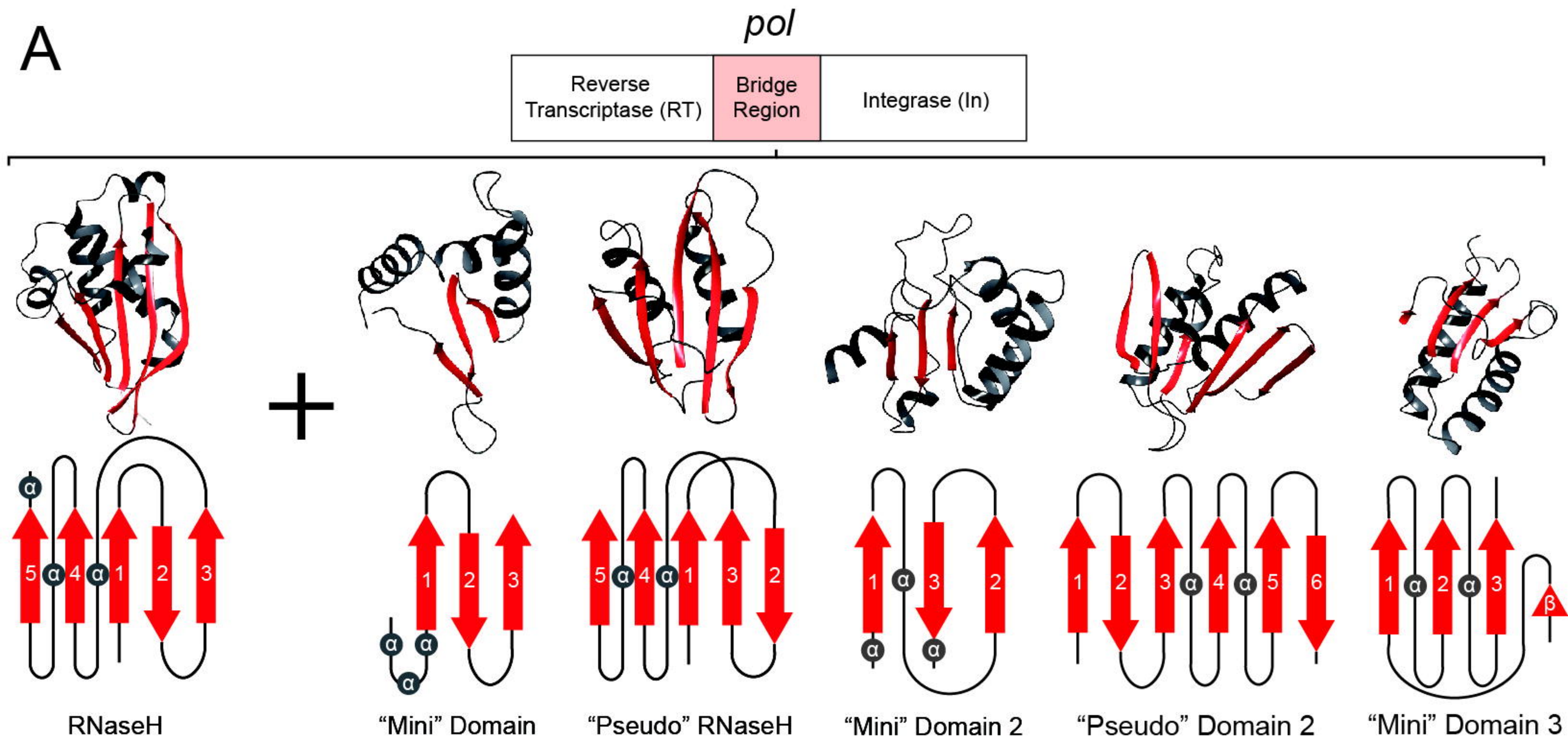
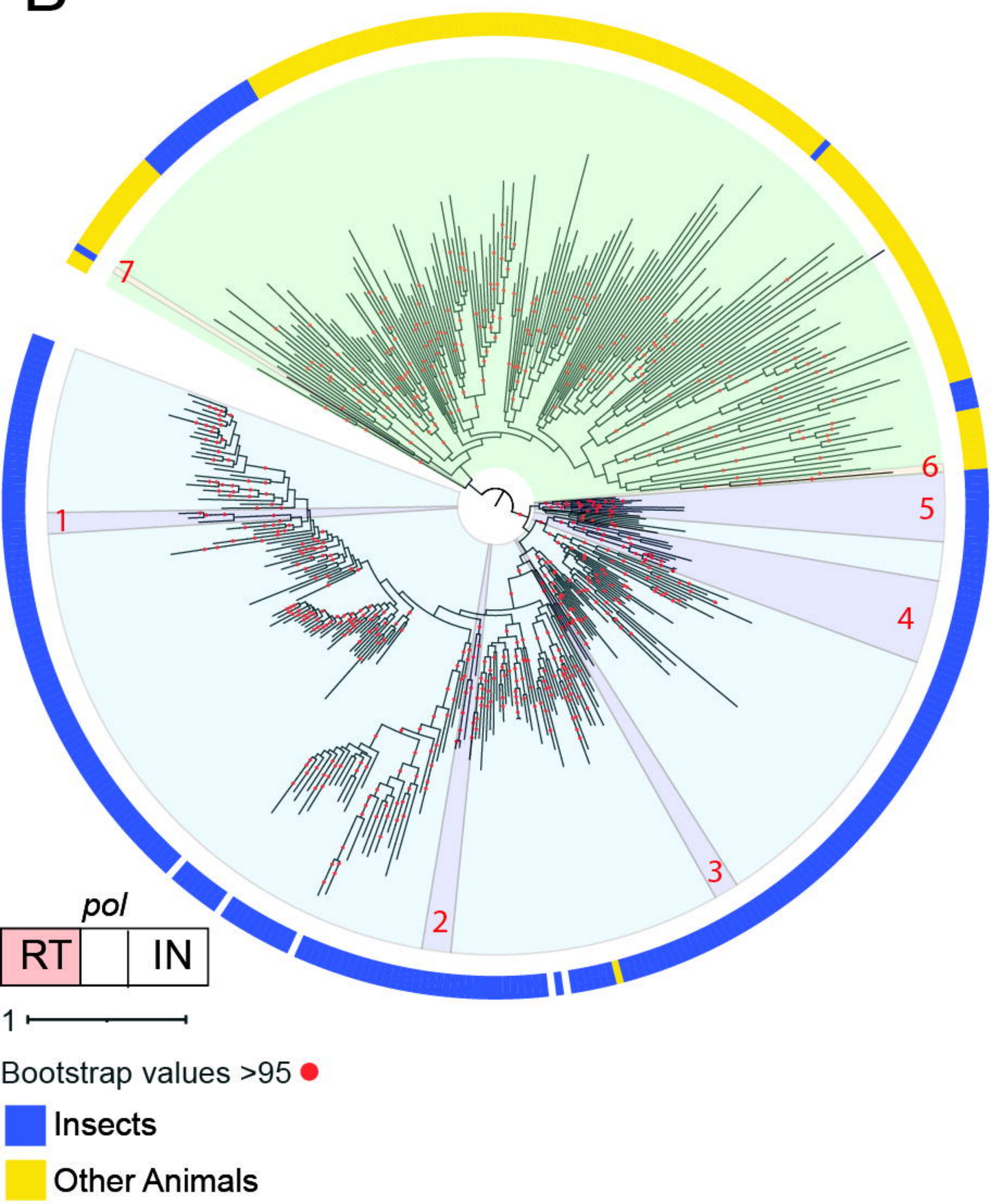

A
Signal Peptides
Ectodomain
Transmembrane Domain
Furin Cleavage Site
Transmembrane Domain
Ectodomain
Transmembrane Domain
Furin Cleavage Site
Signal Peptides
Ectodomain
Transmembrane Domain
B
N
C
FLs
Crown
Stalk
Base
Fusion Loop
Fusion Loop
Ectodomain Monomer
Ectodomain Trimer
C
Fusion Loops
C1a
C1b
C1c
C1d
C2a
C2b
C2c
C2d
C2e
C3
Cysteine Bridge
Cysteine Bridge
Fusion Loops
Cysteine Bridge
CC Motif
Fusion Loops
Cysteine Bridge
Fusion Loops
D
Host Order
Cysteine Bridges
Platyhelminthes C1a, C1d, C2b
Bryozoa C2b
Rotifera C2e
Tardigrada C2c
Nematoda C1a, C1d, C2a, C2c
Tunicata C3
Cnidaria C1c
Ctenophora C1c
Collembola C2d
Hexanauplia C1b
Branchiopoda C2c
Coleoptera C1a
Hymenoptera C1a
Hemiptera C1b
Polyneoptera C1a
C1a, C2b
C1d
C1a, C1d
C1a
C2a, C2c

Figure 6

A



B



"Ancient" Errantiviruses

RNase H + "Mini" domain

RNase H + "Mini" domain 2[6]

RNase H + "Pseudo" domain 2[7]

Insect Errantiviruses

RNase H + "Pseudo" RNase H

RNase H + "Pseudo" RNase H + "Pseudo" RNase H[2]

RNase H + "Mini" domain 2[1]

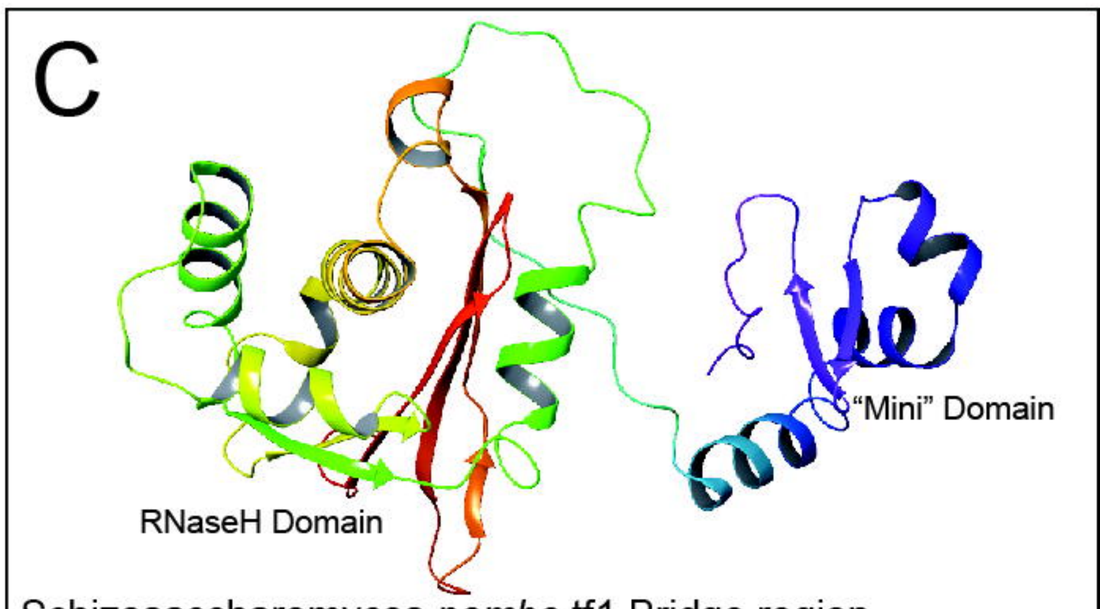
RNase H + "Pseudo" domain 2[3]

RNase H + "Mini" domain[3]

RNase H + "mini" domain 3[1,4]

RNase H only[4,5]

C



Schizosaccharomyces *pombe* tf1 Bridge region

Figure 7

Newly Identified *env* containing "errantiviruses"
Newly Identified *env* containing "errantiviruses" carrying chromodomain
Known Ty3/Gypsy elements that do not contain *env*
Known Ty3/Gypsy elements that contain *env*
"Ancient" Errantiviruses
"Insect" Errantiviruses
Malacostraca (F), Hexanauplia (F)
Hemiptera (F), Polyneoptera (HSV+F), Palaeoptera (F), Coleoptera (HSV+F)
Branchiopoda (HSV), Collembola (HSV), Annelida (HSV)
Nematoda (HSV)
Annelida (F)
Annelida (F), Thecostraca (F), Mollusca (F)
Cnidaria (HSV), Thecostraca (F), Platyhelminthes (HSV), Bryozoa (HSV)
Cnidaria (F), Mollusca (F), Annelida (F)
Annelida (F), Hexanauplia (HSV), Rotifera (F)
Rotifera (F), Arachnids (F)
Tardigrada (HSV+F), Collembola (F), Branchiopoda (F)
Collembola Entomobryomorpha Scur (F)
Gypsy-3 DEI
Gypsy-21 DAn
REINA
SCL213 Galadriel
CRM AAM94350.1
Spombe Tf1 AAA35339.1
SUSHI
Maggy
Chromoviridiae
Saccharomyces Ty3
Tunicata Ascidiacea Clep (HSV)
Gypsy-7 DAn
Tunicata Ascidiacea Atur (HSV)
Tunicata Ascidiacea Clep (HSV)
Tunicata Ascidiacea Clep (HSV)
Cnidaria (HSV), Annelida (F)
Thecostraca (F), Annelida (F), Hemichordata (F)
Nematoda (HSV), Annelida (F)
Thecostraca (F), Diptera (F), Bryozoa (HSV)
Annelida (F), Nematoda (HSV)
Annelida Polychaeta Slim (F)
Ctenophora (F)
Tor4b
Tunicata (F)
Nemertea Pilidiophora Llon (F)
Bryozoa Gymnolaemata Bsto (HSV)
Skipper *
Annelida Polychaeta Himp (F)
ATHILA0p1
Ctenophora (HSV+F)
Collembola (F), Hemiptera (HSV+F)
Gypsy-6 DFi
Gypsy-7 DSe
Gypsy-32 DWil
MDG3 DM
INVADER3
Mollusca Bivalvia Dpol (F)
Cer1
Platyhelminthes Rhabditophora Mlig (HSV)
Gypsy-10 DAn
MAG Bmor
Mag As
Cer4
Bryozoa Gymnolaemata Cpal (HSV)
Tor2
Tor1
Tunicata Appendicularia Odio (F)
Annelida (F), Branchiopoda (F)
Collembola (F)
Hemiptera (F)
Hemiptera (F), Polyneoptera (F)
Diptera (F), Coleoptera (F), Hymenoptera (F), Polyneoptera (F), Palaeoptera (F), BLOOD, Gypsy-28 DAn
Hymenoptera (HSV+F)
Hymenoptera (F)
Diptera (F), Lepidoptera (F)
Coleoptera (F), Hymenoptera (F)
Lepidoptera (F)
Coleoptera Bupresrtidae Acya (F)
Coleoptera (F)
Diptera (F), IDEFIX, DM176
Diptera (F), Lepidoptera (F), Coleoptera (F), Hymenoptera (F), Hemiptera (F), Cnidaria* (F)
Lepidoptera (F), Coleoptera (F)
Coleoptera Bupresrtidae Acya (F)
Diptera (F), Lepioptera (F), Coleoptera (F)
Coleoptera (F)
Diptera (F)
Diptera (F), ZAM
Diptera (F), GYPSY-Gypsy, GYPSY gypsy-5 DGri
Diptera (F)

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**

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

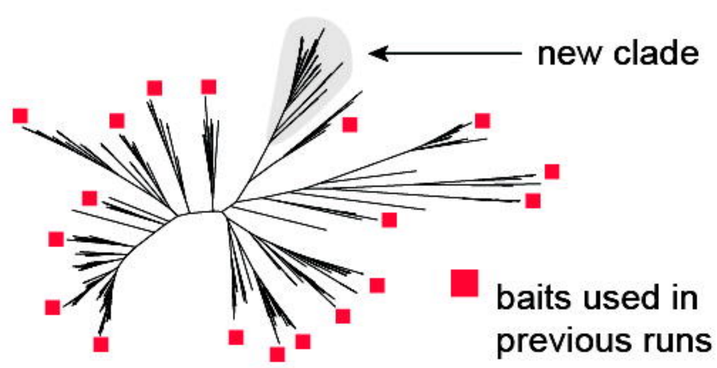
- Only the GAG-POL-ENV elements were annotated in the insect genomes.
- For non-insect Arthropoda and other Protostomia genomes, ENV was predicted from the downstream region of GAG-POL.
- For non-vertebrate Deuterostomia, Cnidaria and Ctenophora genomes, ENV was predicted from the downstream region of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

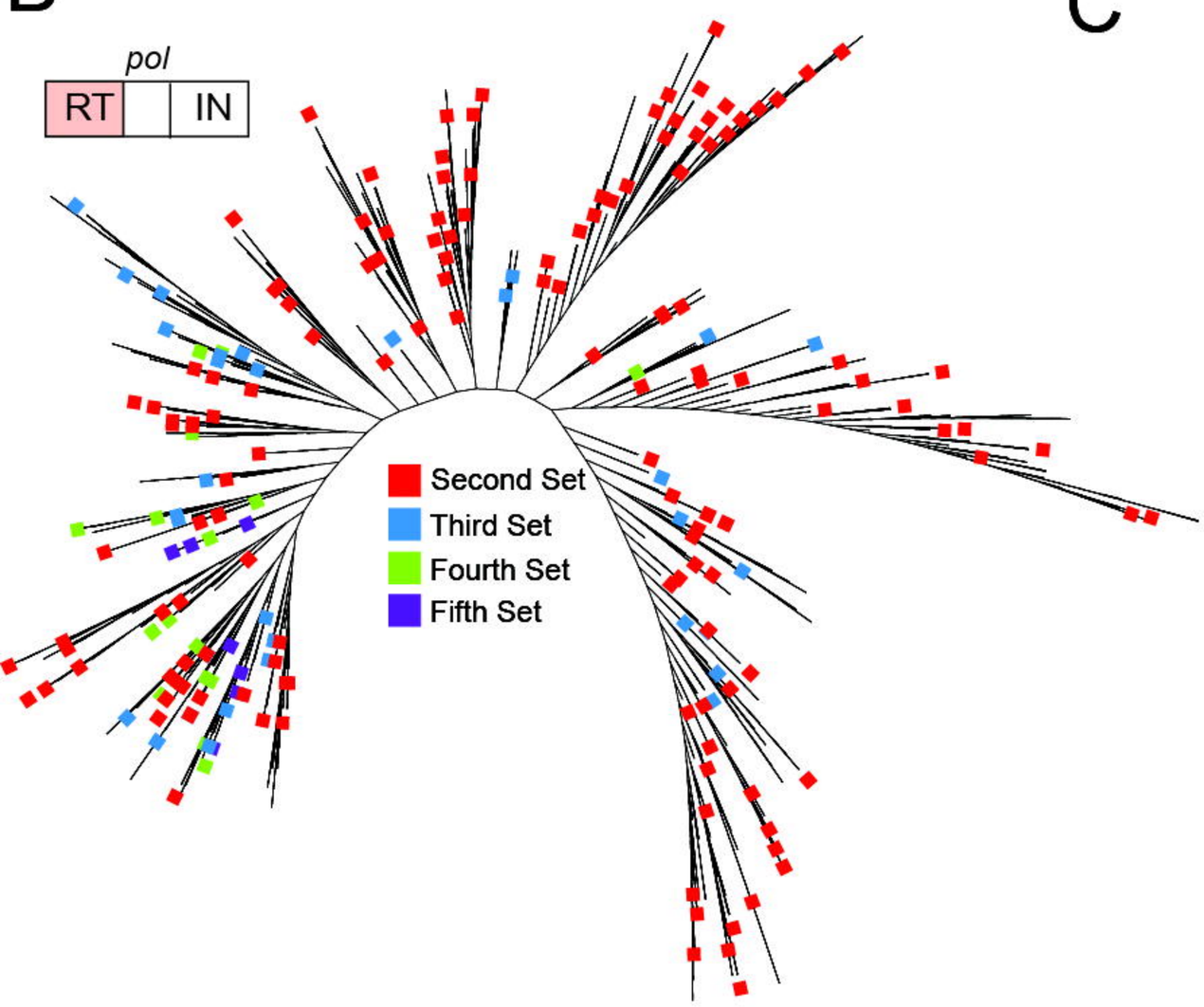
Step 4. curate the identified ERVs by domain structure prediction using hmmscan and HHpred
***All** three ORFs are required to predict homology to known viral proteins listed below*
GAG and POL: homologous to retroviral or Ty3/gypsy GAG
POL: homologous to Ty3/gypsy POL
ENV: homologous to paramyxovirus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

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

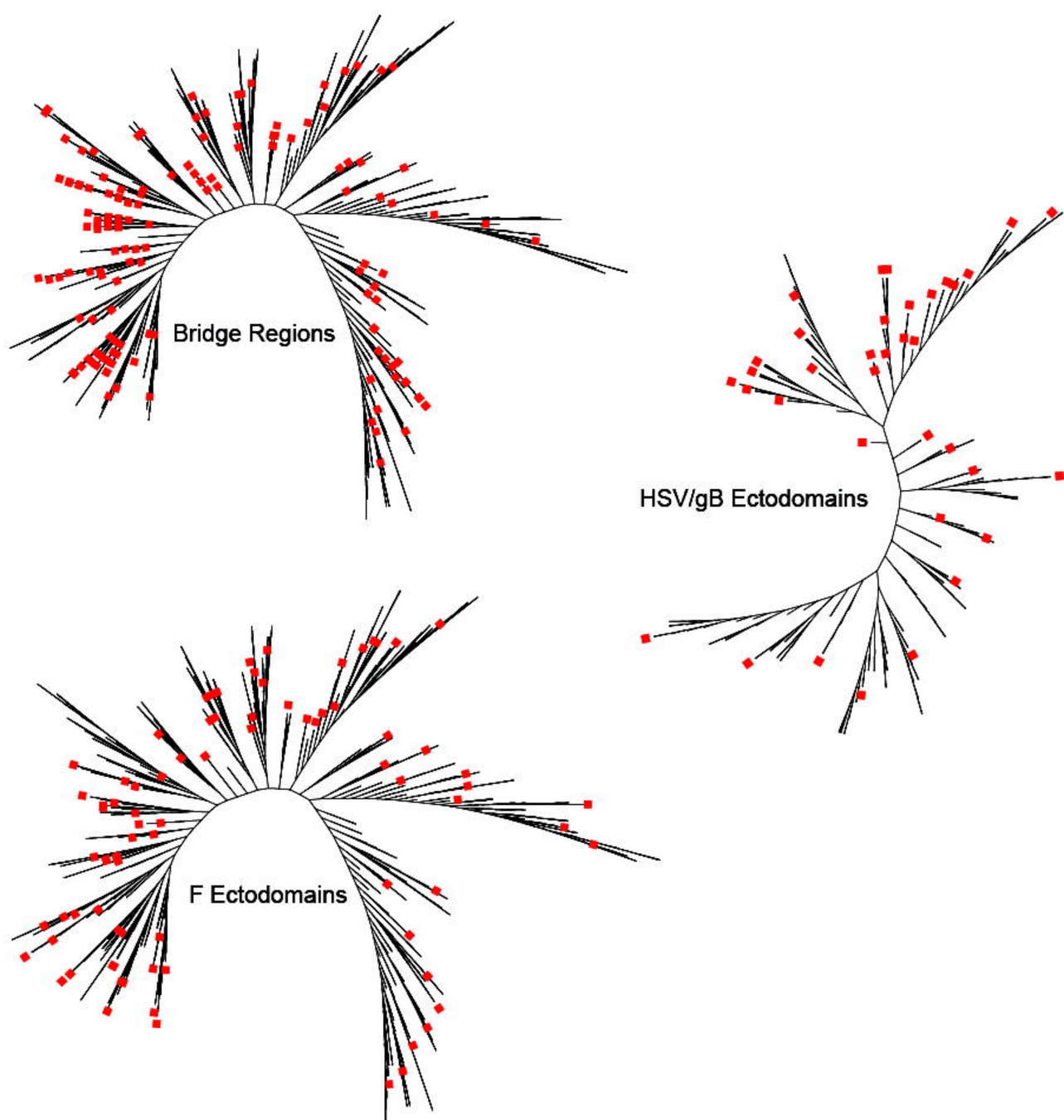
tBLASTn interations	*identified ERVs*
1st	945
2nd	451
3rd	61
4th	49
5th	6



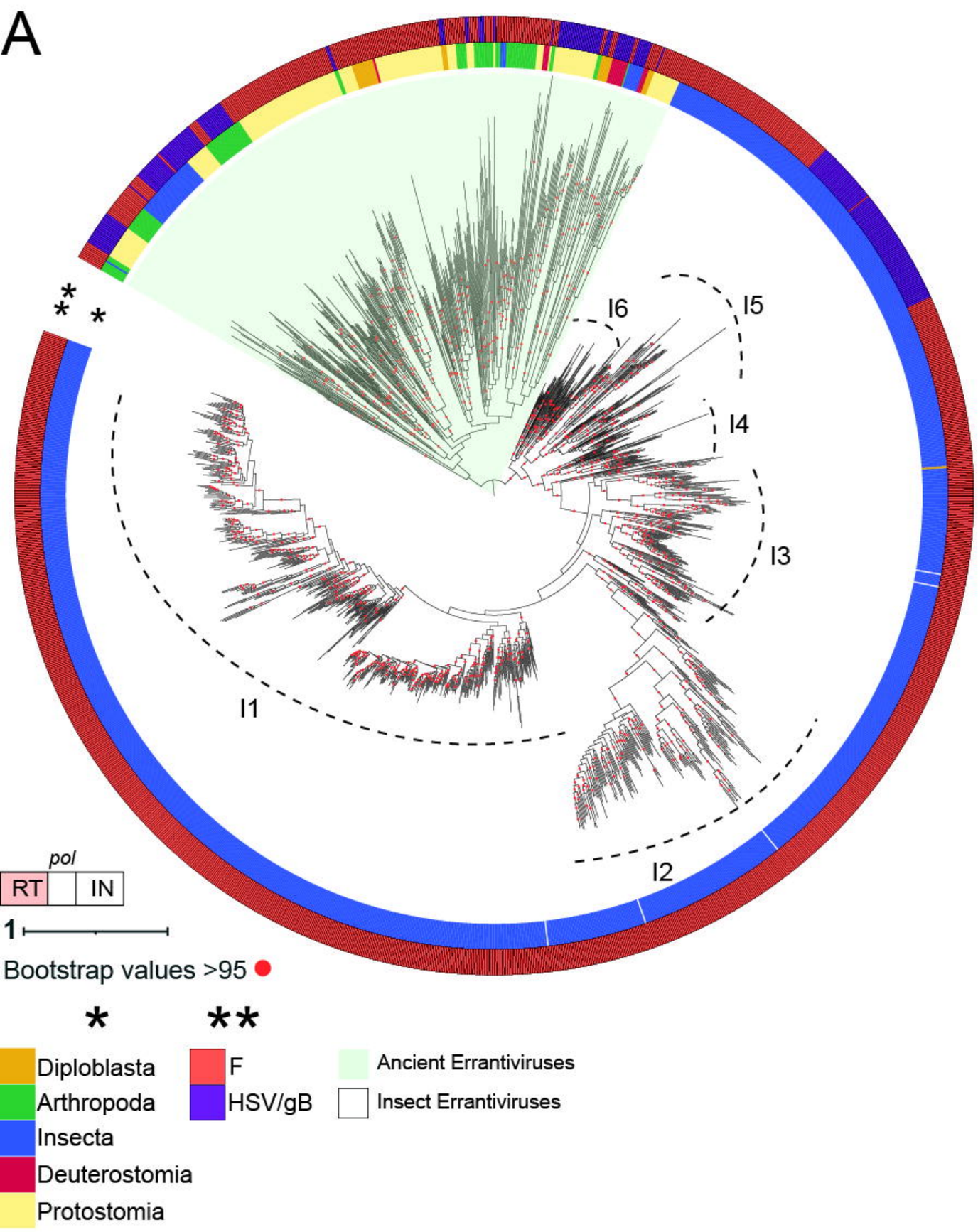
B

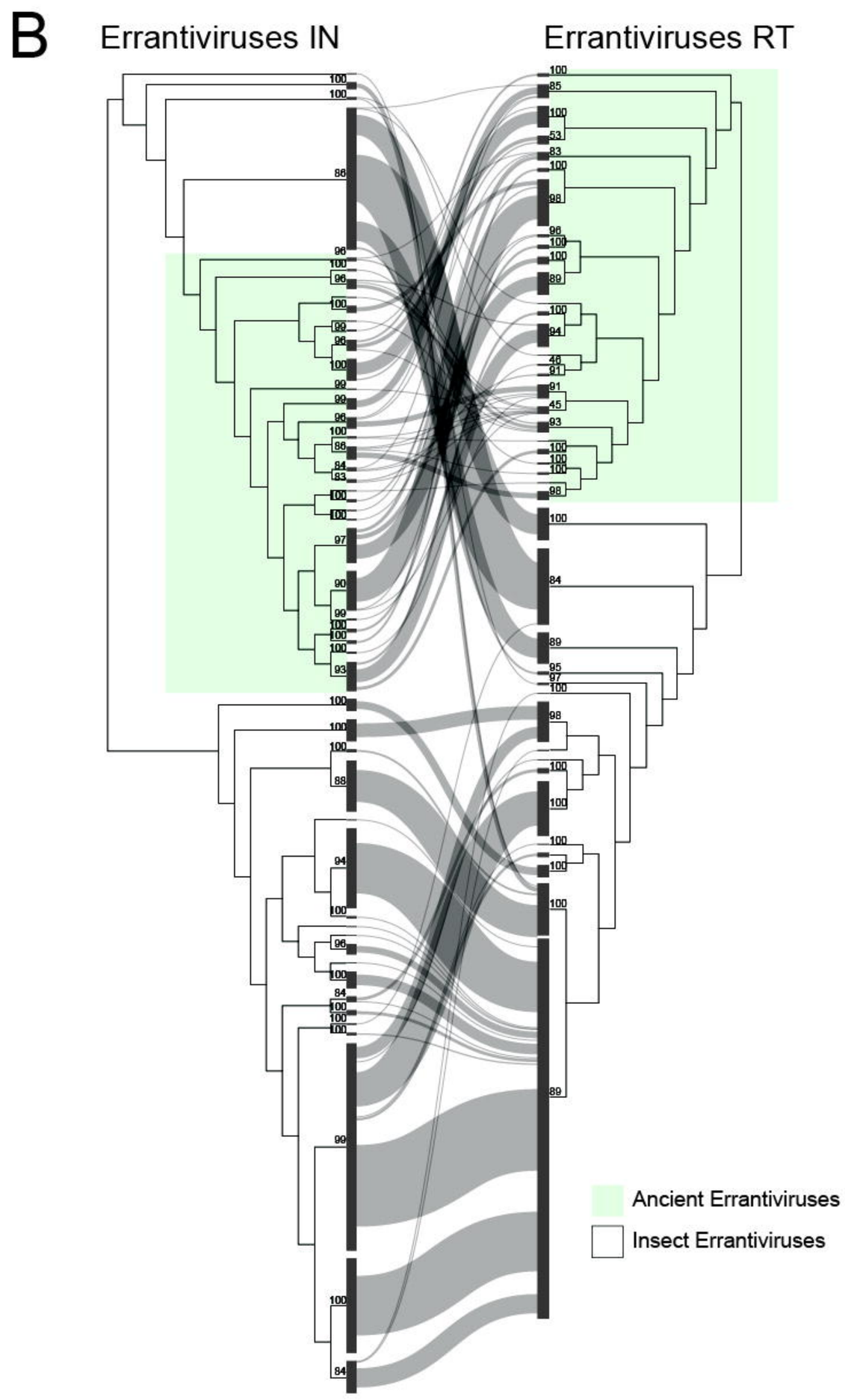


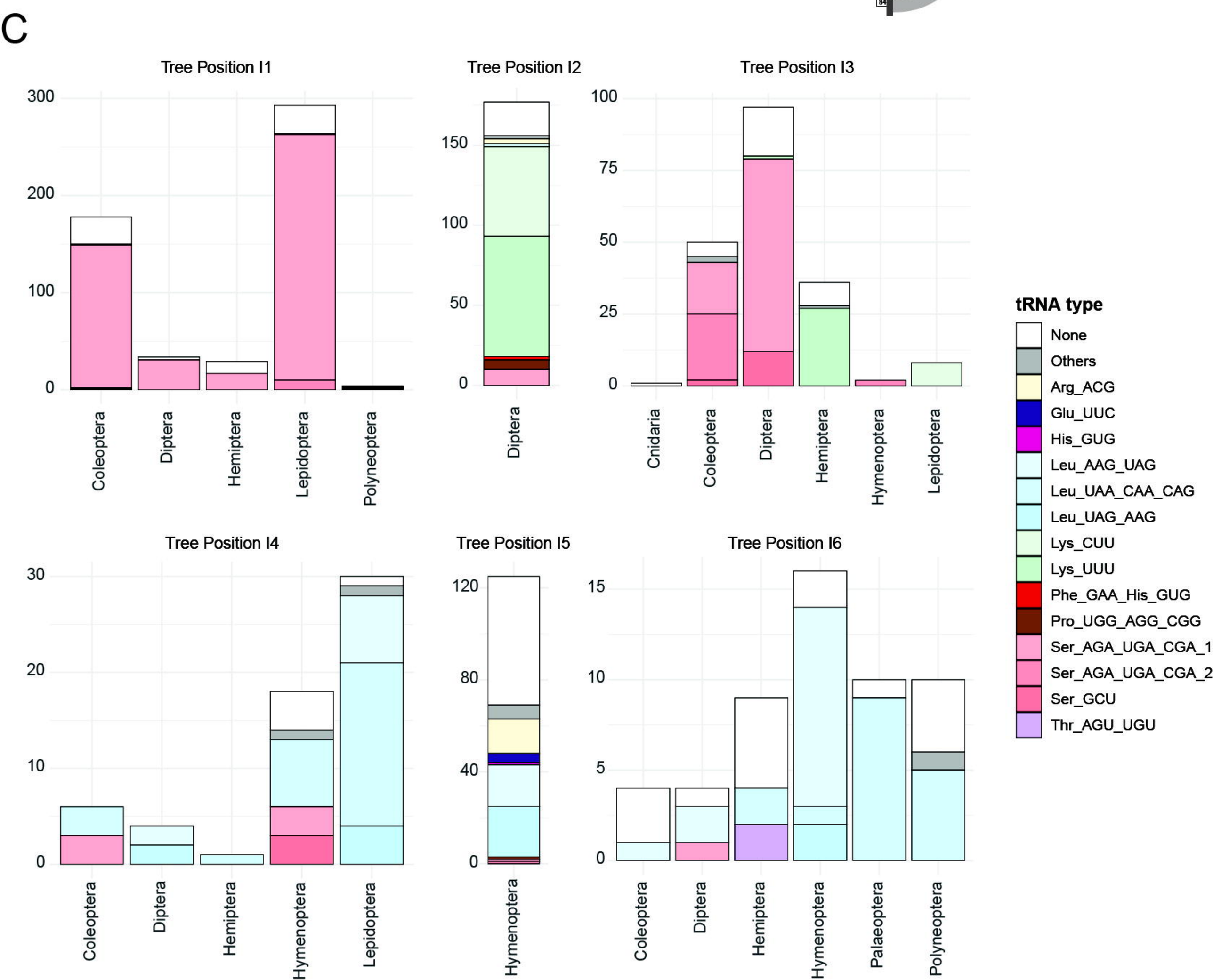
C



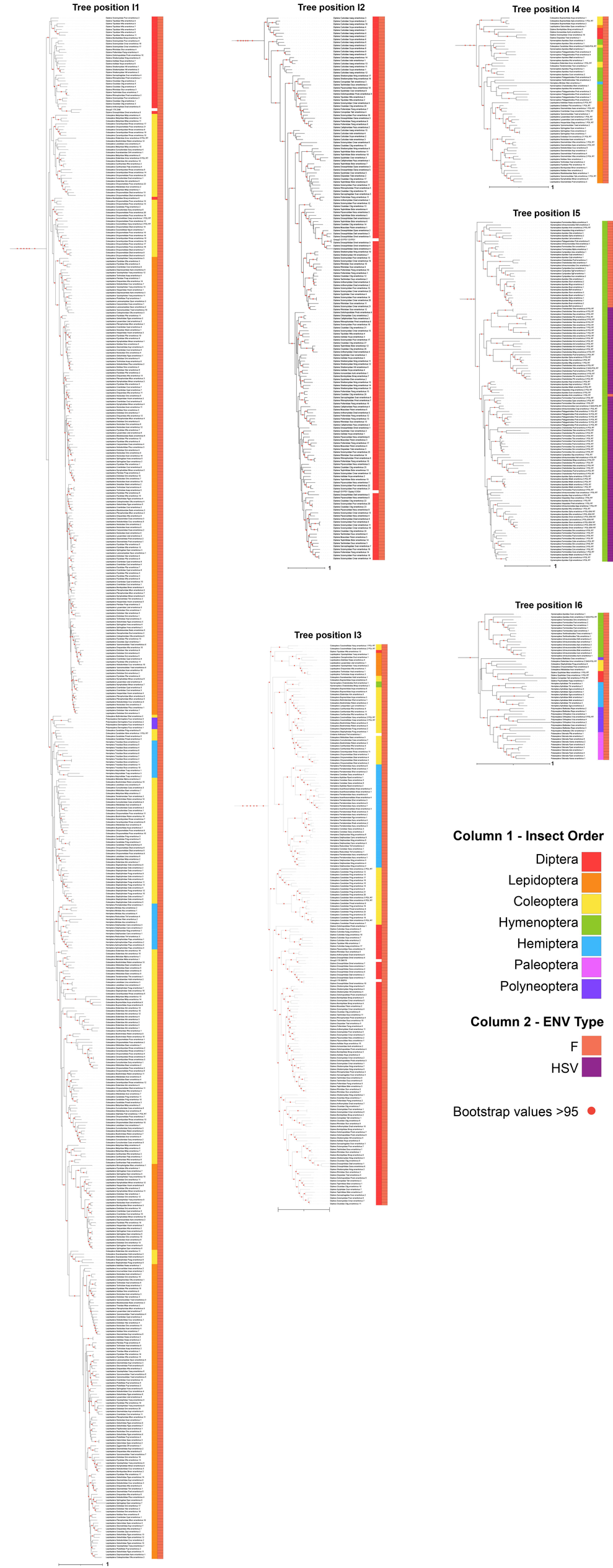
Supplementary Figure S2



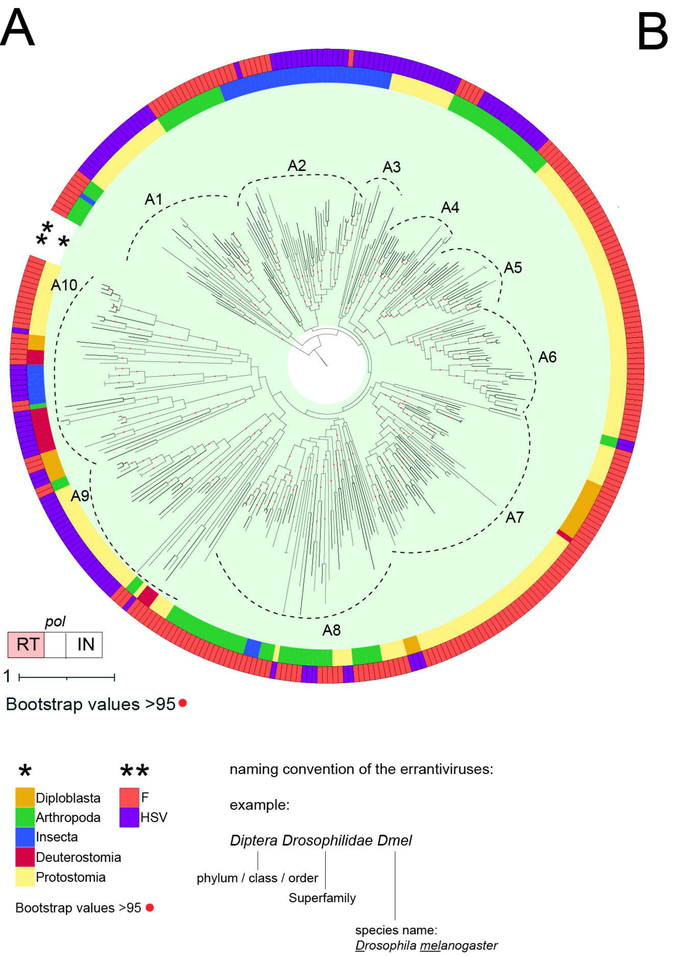


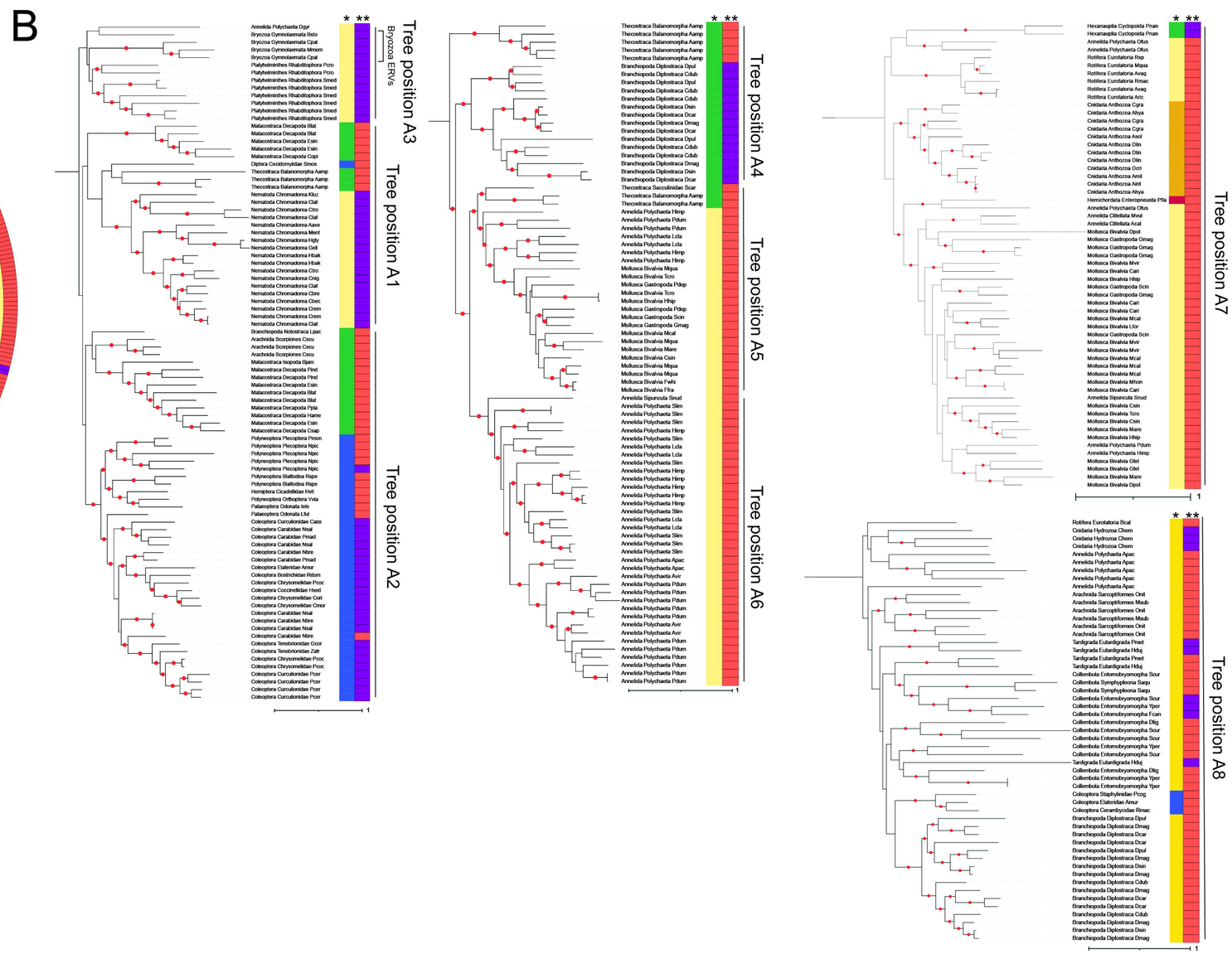


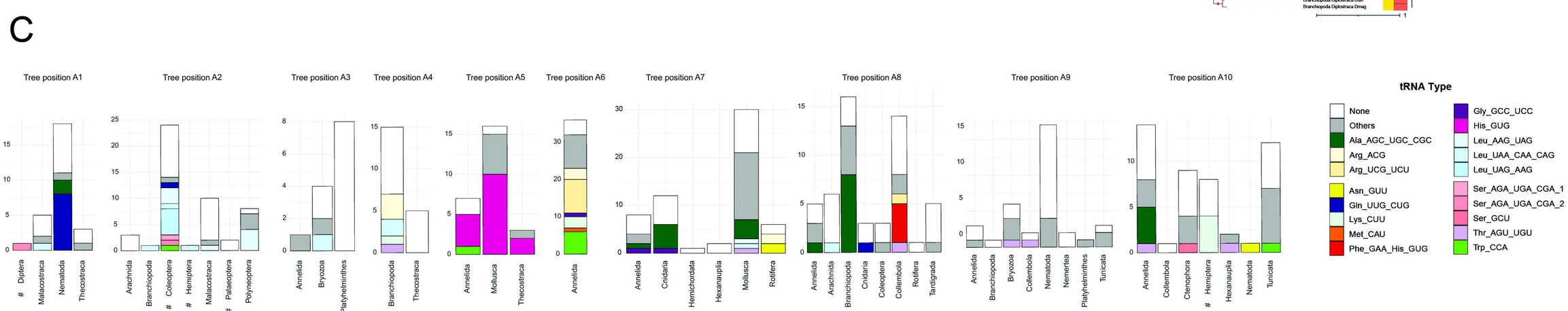
Supplementary Figure S3



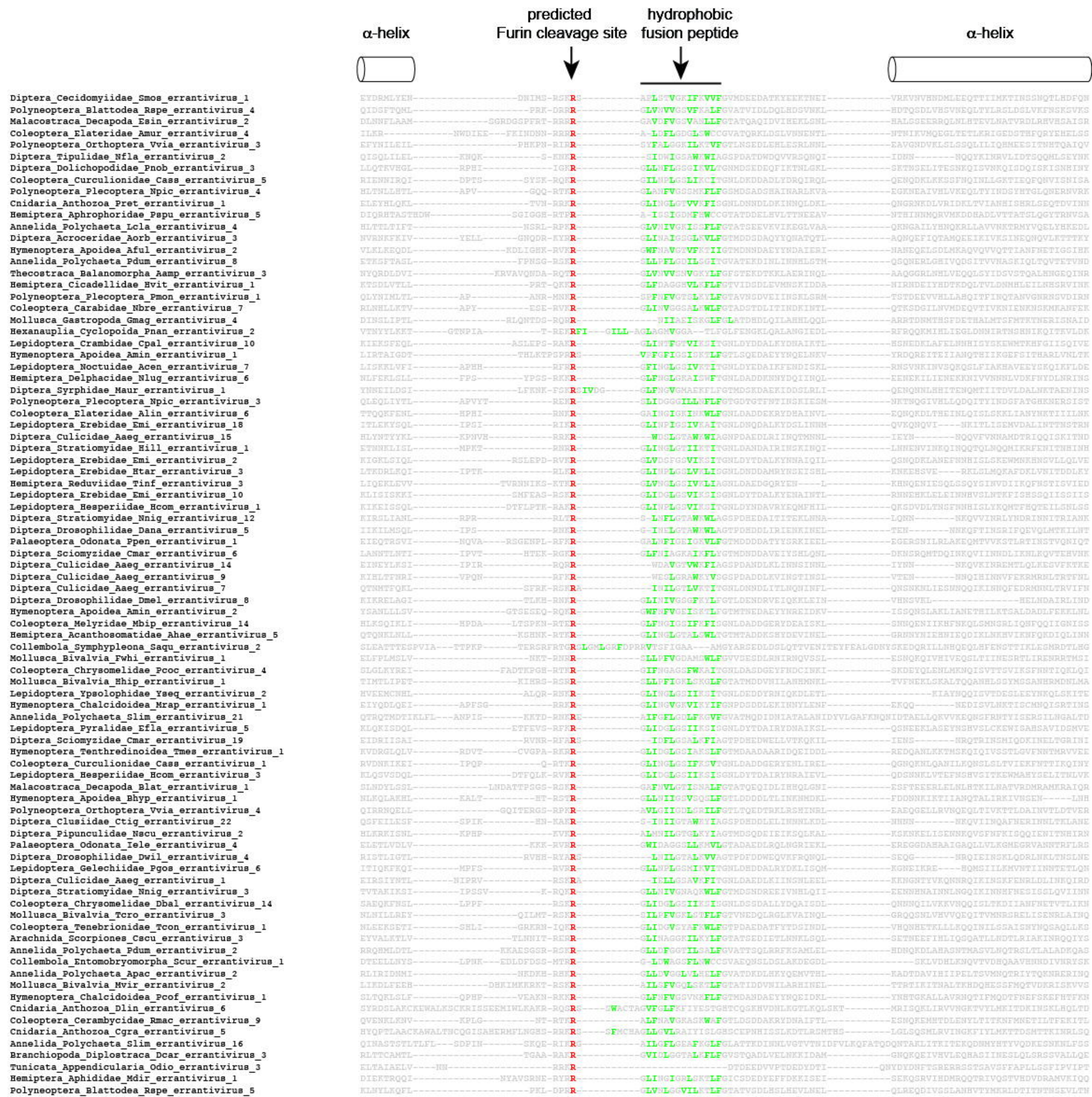
Supplementary Figure S4



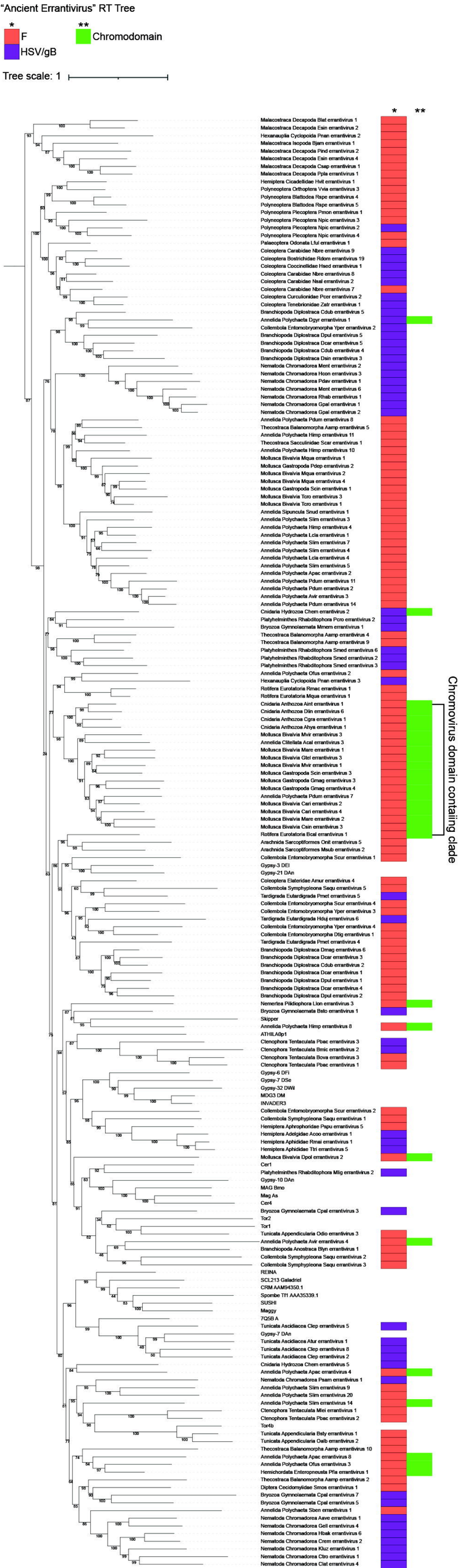




Supplementary Figure S5



Supplementary Figure S6

Supplementary Figure S7

"Ancient Errantivirus" RT Tree