Ancient origins of endogenous retroviruses in Metazoans

Shashank Chary[1] and Rippei Hayashi[1, 2 #]

1. The John Curtin School of Medical Research, The Australian National University, 131 Garran Road, Acton, ACT 2601, Australia

2. The Shine-Dalgarno Centre for RNA Innovation, The Australian National University, 131 Garran Road, Acton, ACT 2601, Australia

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

Abstract

Endogenous retroviruses (ERVs) are generally seen as remnants of past retroviral infections, but some can make new germline insertions that are transmitted across generations. ERVs can carry membrane fusion glycoproteins, or *env* genes, which allow them to remain infectious. We systematically searched for such active ERVs containing *env* in metazoan genomes and found that they are ubiquitously present across taxa including ancient animals such as Cnidarians, Ctenophores and Tunicates. Phylogenetic analyses revealed that the divergence of ERVs, including those found in vertebrate and insect genomes, largely follow the divergence of host organisms, indicating that they share a common and highly ancient origin, and that they survive over time primarily through vertical transfer.

Introduction

Retroviruses have two mechanisms through which they survive; most commonly, they exogenously infect other hosts, transmitting horizontally[1]. However, they can occasionally infect the germline cells of the host; where genetic information is stored for the next generation, and integrate into the genome [1,2]. In this way, they survive by transmitting vertically down into the next generation. These proviral copies in the germline genome are referred to as endogenous retroviruses (ERVs) [1,3-5]. ERVs are best characterised in vertebrates and make up 5 to 10% of their genomes. An intact copy of an ERV would contain viral *gag*, *pol* and *env* genes, and be flanked by long terminal repeats (LTRs), broadly classified under class I LTR retrotransposons [6].

LTR retrotransposons are ubiquitously found in higher eukaryotes. One of the best studied examples is Ty3/Gypsy. Interestingly, a branch of *gypsy* transposons in Drosophila species carry *env* genes and are referred to as errantiviruses[7,8]. When the defence mechanism against transposable elements is affected, errantiviruses are expressed in the somatic cells of the female ovaries, forming virus like molecules that infect the adjacent germline cells[9,10]. This means that if the errantivirus remains active in the somatic cells, it can maintain the generation of new copies that can be vertically transmitted to the next generation.

Although vertebrate ERVs and Drosophila errantiviruses are structurally similar, the exact origins of vertebrate ERVs that are vertically inherited are not well understood. This combined with the unique cell niche of errantiviruses that allows their long-term survival in Drosophila prompted us to question their origins, as well as how widespread they are.

Results and Discussion

We systematically searched for homologous elements of vertebrate ERVs and errantiviruses in the referenced genomes across metazoan phyla using the annotated protein coding sequences of *gag, pol* and *env* for tBLASTn searches (see methods). We focused our search on the genomic copies that are recently acquired to understand relationships of currently active elements. We identified 1496 elements that carry nearly intact open reading frames of all three genes (Supplementary table 1). Most of the newly identified genomic copies carry putative LTRs and tRNA priming sequences (1150 elements, summarised in Fig S5B and Supplementary table 5), suggesting their recent integrations. We ran multiple sequence alignments using the conserved Integrase and Reverse Transcriptase (RT) domains of Pol as well as the Env proteins, and assembled maximum probability trees to infer phylogenetic relationships of the identified elements. ORFs of selected vertebrate retroviruses and Drosophila errantiviruses are also included in the tree to aid classifications.

We found that errantiviruses are more ancient and widespread than previously thought. Multiple intact copies are found not only across all insect orders including Palaeoptera and Polyneoptera, but also in most major phyla across metazoans. This includes species from ancient phyla such as Cnidaria, Tunicates and Ctenophora, but excludes Echinodermata and Porifera (Fig 1B). ERVs containing *env* genes are found in all vertebrate orders, including jawless fish, with their genomes containing multiple intact ERV copies. We

refer to elements found within and outside Vertebrata as vERVs and *gypsy* errantiviruses, respectively. The separation and the origins of the two classes will be discussed shortly.

We found that both vERVs and *gypsy* errantiviruses have likely co-evolved with their host species. When the phylogenetic tree of the *pol* Integrase domains of all identified elements is grouped into branches, most vERVs form their own branches (Fig S1). These branches (V1-V9) individually contain elements from evolutionarily related species. For example, the vertebrate branches V8 and V9 only contain mammalian elements while V6 contains elements from reptiles and birds, and V5 contains elements from fish, amphibians and reptiles (Fig S1B). Importantly, the divergence of the branches roughly follows the divergence of species, as the branches that are next to each other in the tree tend to contain elements from related species classes (Fig S1B). Gamma-retroviruses are found in V9 while other known retroviruses, including alpha-, beta-, delta-, lenti- and formy-retroviruses, are all found in V1. The more recently identified loki-retroviruses are found in V2.

Similarly, *gypsy* errantiviruses from insects and other Arthropoda species mostly cluster together in the phylogenetic tree of *pol* Integrase (Fig 1A, S1B). Like vERVs, most branches (A1-A20) are formed by elements either from single or multiple related orders or classes (Fig S1B). For example, elements from hemimetabolous insects (Hemiptera, Polyneoptera and Palaeoptera) tend to cluster together as seen in branches A7, A14 and A18, and with elements from the ancient holometabolous insect order Hymenoptera in the A14 or Arthropoda classes Arachnida and Branchiopoda in the A18 (Fig S1B). Known dipteran errantiviruses Gypsy-Gypsy, and Gypsy-5-Dgri sit within clade A, Gypsy-17.6 and Idefix sit within clade B, and Gypsy-ZAM sits within clade C (Fig 1A), all of which are Dipteran branches.

Species from other Protostomia and Deuterostomia phyla also carry *gypsy* errantiviruses. These include Molluscs, Annelids, Nematodes, Platyhelminthes, Hemichordates and Tunicates. Furthermore, non-bilaterian animals Cnidarians and Ctenophores also carry *gypsy* errantiviruses (Fig 1B, Fig S1C-E). These observations, as well as the divergence patterns of vERVs and *gypsy* errantiviruses in Arthropoda, strongly suggest that the ancestors of these elements existed before the split of bilaterian and non-bilaterian animals in the pre-Cambrian era and likely co-evolved together with the host organisms.

We further examined the divergence of Env and Pol proteins to inspect the relationship between vERVs and *gypsy* errantiviruses. We found that they carry three distinct types of Env proteins. Most errantiviruses identified in this study carry fusion glycoprotein F/Nipahvirus-like *env*, however, elements from more ancient animals such as Cnidarias, Ctenophores and Tunicates also carry glycoproteins that are related to those from Herpesviruses (HSV) (Fig 1B, Fig 2). In addition, Platyhelminthes, Nematoda and Bryozoa exclusively carry *gypsy* errantiviruses with HSV-type *env* (Fig 1B). They are also flipped, in that the arrangement of their ORFs is *gag-env-pol*, which is structurally analogous to the recently identified Nematoda endogenous retroviruses[11]. ERVs found in vertebrate genomes mostly carry glycoprotein GP2, but we also found elements in jawless fish, fish, amphibians, and reptiles that carry fusion glycoprotein F (Fig 1B, Fig 2). These elements cluster together with loki-retroviruses (Fig S1A), that also carry the F-type *env* [12]. F-type *env* being found in both vERVs and invertebrate *gypsy* errantiviruses may suggest their

common origin. Alternatively, the F-type *env* gene could have been acquired multiple times in evolution. The latter possibility is less likely, as we found that the phylogeny of F-type *env* genes is largely congruent with the phylogeny of *pol* genes that they carry (Fig S2A). When all vERVs and errantiviruses with the F-type *env* are grouped based on the branches of the *env* and *pol* Integrase trees in which they are found, elements from the same branch in one tree tend to form a single branch, or multiple neighbouring branches (Fig S2A). Further supporting their monophyletic relationships, there are no major discordances in the distance of neighbouring branches between the two trees. F-type *env*-carrying vERVs and loki retroviruses cluster together with elements from Tunicates, ancient Protostomia and non-bilaterian phyla both in the *env* and *pol* Integrase trees (Fig S2, Table S3). Interestingly, several insect errantiviruses are also found in the same branch, indicating that this branch contains elements that share traits of ancient ancestry elements. Trees of *env* and *pol* Integrase that only assembled elements that carry either HSV-type *env* or GP2-type *env* also largely agree with each other (Fig S2B-C). These observations suggest that different types of *env* genes were likely acquired in ERVs and errantiviruses only once in evolution.

The *pol* protein of vertebrate retroviruses has duplicated the Ribonuclease H (RNase H) domain between RT and Integrase. The N-terminal copy has been partially degraded and is called the "tether" domain[13]. We found that most insect errantiviruses have also acquired a second RNase H-like domain, located C-terminal to the functionally conserved RNase H domain (see methods). The second RNase H-like domain retains the barrel-like structure made of five core β-sheets, but the order of the second and the third β-sheets is flipped (Fig 3B-C). We refer to this domain as "pseudo" RNase H domain. We also found that most vERVs carry an additional small structure, consisting of three core β-sheets, located C-terminally to the RNase H domains (Fig 3B-C). We name this small structure "mini" RNase H domain. Interestingly, many of the ancient *gypsy* errantiviruses as well as several insect errantiviruses also carry structurally similar "mini" domains. Despite being highly divergent at the primary sequence level, structural arrangements of the "pseudo" and "mini" domains are highly conserved, particularly for the core β-sheets, suggesting their functional importance. The "mini" domain being shared between vERVs and *gypsy* errantiviruses further supports the idea that they share a common ancestor between them.

The errantiviruses and vERVs we identified all have multiple intact copies in the genome, meaning that there has not been enough time for these insertions to degrade. If the widespread presence of these elements had been a result of horizontal transfer, it is unlikely that the elements would follow the phylogeny of their hosts as closely as we observed. Furthermore, we find it unlikely that an endogenous virus can horizontally transfer to another species, and then invade into the germline to create multiple intact copies so often. For similar reasons, it is unlikely that the elements that we identified are exogenous in nature. We thus propose that *gypsy* errantiviruses as well as vERVs are highly ancient, and have survived over time by creating new copies in the germline, transferring vertically.

We also propose that the vERVs and errantiviruses that we identified in this study may have a common origin, as demonstrated by them both carrying type F *env*, as well as the "mini" RNase H domain. One explanation for elements carrying F-type *env*, may be recombination events between the vERVs and other

infecting viruses. Though we find it more likely that a recombination event happened in the common ancestor leading to some elements carrying F and some carrying GP2. While F-type *env* is not seen in mammals or birds, it is still present in more ancient orders of reptile, amphibians, fish and jawless fish. It was originally suggested, that *gypsy* errantiviruses may have acquired its *env* gene from baculoviruses, some of which carry an F-type ENV protein Ld130 [14,15]. However, in this scenario, it is more likely that baculoviruses acquired the *env* gene from errantiviruses given their more ancient origins and broader host range.

Most animals have defence mechanisms that keep transposable elements silenced (e.g., KRAB Zinc Finger and piRNA pathway)[16-18]. In Drosophila, *gypsy* errantiviruses are expressed in the ovarian somatic niche, and invade the germline cells using *env*. In this way, they can evade the germline defence pathway. Over time, this has led to the formation of a somatic defence pathway in Drosophila animals to deal with this errantivirus[19]. We therefore speculate that the elements we identified may have found a similar cell niche within the gonad, that allows them to invade the germline.

Acknowledgements

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

Author contributions

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

Data Availability

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

Conflicts of interests

The authors declare no conflicts of interests.

Materials and methods

Genome assemblies used in the study

Reference genome sequences were retrieved from NCBI and grouped into the following insect orders, vertebrate classes and other metazoan phyla: Diptera, Lepidoptera, Coleoptera, Hymenoptera, Hemiptera, Polyneoptera, and Palaeoptera for insects; jawless-fish, fish, Amphibian, reptile, bird, and mammal for vertebrates; Arachnida, Branchiopoda, Chilopoda, Collembola, Diplopoda, Hexanauplia, Malacostraca, Merostomata, Ostracoda, and Thecostraca for Arthropoda; Acanthocephala, Annelida, Brachiopoda, Bryozoa, Dicyemida, Mollusca, Nemertea, Onychophora, Orthonectida, Phoronida, Platyhelminthes, Priapulida, Rotifera, and Tardigrada for Protostomia; Cephalochordata, Echinodermata, Hemichordata, and Tunicata for Deuterostomia; Cnidaria, Ctenophora, and Porifera. We obtained lists of reference genomes in each category in May 2023 using the Entrez Direct command line tool except for the vertebrate genomes which we used the list from the previous work by Wang et al. [12]. The list of the genome assemblies used can be found in Supplementary table 7.

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

We ran two iterations of tBLASTn (ncbi-blast-2.9.0) to search for the genomic copies of vERVs and *gypsy* errantiviruses, firstly using the baits of known elements, and secondly using the baits identified from the first iteration. We used two sets of sequences for the first tBLASTn search. The first set included GAG, POL and ENV protein sequences of the Dipteran *gypsy* errantiviruses from the RepBase (https://www.girinst.org/repbase/). The second set included GAG, POL and ENV protein sequences of selected vertebrate retroviruses that were used in the previous study [12]. We applied different criteria to identify ERVs from species outside insects and vertebrates as we did not have prior knowledge of whether or what kind of ERVs are found in those species. We briefly summarise our approach below.

We first used open reading frames (ORFs) of vertebrate retroviruses as baits for the vertebrate genomes and those of the Dipteran errantiviruses for the other genomes to run tBLASTn with default parameters. 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 and vertebrate 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. Number of

uninterrupted ORFs per genome was counted after the second iteration of tBLASTn and summarised in Supplementary table 7. 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 duplicated in the genome. These processes eliminated genomic copies of ERVs that are being degraded.

The above process yielded abundant putative intact copies of ERVs from insect and vertebrate genomes whereas we took additional steps for the other genomes. We extended the genomic regions of predicted POL-ENV to upstream by 2000 nucleotides and manually annotated putative GAG ORFs using the online sequence viewer Benchling (https://www.benchling.com/) for the Arthropoda genomes outside insects as well as other Protostomia genomes. 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. Where possible only the duplicated copies of ERVs were taken for further analyses outside insects and vertebrates as well.

We took confident 1128 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 178 representative POL from the phylogenetic tree and used the POL and the corresponding GAG and ENV ORFs from the same genomic regions for a second tBLASTn search. We repeated the above steps after combining the results of the two tBLASTn searches. We did not find any ERVs in Nematoda genomes that have GAG, POL and ENV in this order, however, a previous study reported non-canonical Nemetoda ERVs that have flipped the order of ENV and POL[12]. 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 Protostomia genomes outside Arthropoda and found similar elements in Bryozoa and Platyhelminthes. We ran Clustal Omega to select ORFs that are found multiple times in the genome for insects and vertebrates while we focused on finding intact ORFs for the other genomes (see below) regardless of the copy number. We additionally ran tBLASTn using the fully annotated ORFs as baits to interrogate the copy number of ERVs in the genome where they are found. The results are summarised in Supplementary table 1, where we counted the number of tBLASTn hits with a more than 90% identity to the bait sequences. A majority of newly identified ERVs (1409 out of 1496 elements) carry ORFs that are found at least twice in the genome, further validating our approach of finding active or recently active insertions.

Following, we further curated newly identified ERVs. Firstly, we used hmmer 3.3.1 (http://hmmer.org/) to make sure that the ERVs that we identified are classified as Ty3/gypsy. We compared POL ORFs from the newly identified ERVs against the POL ORFs of known retroelements that are available from the Gypsy Database 2.0 (https://gydb.org/index.php/Main_Page). The Gypsy Database 2.0 collects hmm profiles of

Ty1/Copia, Bel/Pao, Ty3/Gypsy, and Retrovirus POL proteins. The hmmer search classified all the newly identified ERVs into Ty3/Gypsy or Retroviruses. Secondly, we ran HHpred from hh-suite 3.3.0 [20] 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. Furthermore, we reannotated ORFs in ERVs except for those from insects when they are only partially predicted homology to baits in the tBLASTn search. We extended the predicted ORFs in the genomic region to upstream until a start codon is found, and to downstream until a stop codon is found. We used benchling to facilitate the identification of continuous and intact ORFs. ERVs that predicted all GAG, ENV and POL ORFs were taken for further analyses except for one element found in Hemichordata, which only had POL and ENV. We included the Hemichordate element in the analyses because of its unique place in evolution.

Phylogenetic analysis of POL and ENV domains

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 [21], 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) [22], 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 additionally included ORFs from 5 Drosophila errantiviruses (DM176, IDEFIX, ZAM, GYPSY, and Gypsy-5_DGri) and 23 representative vertebrate retroviruses (BLV, FLV, HTLV-2, KoRV, LDV, MMTV, MuLV, REV, RSV, SRV-1, BFV, EFV, FFV, SFV, HIV-1, EIAV, FIV, SnRV, Loki-Ofa, Loki-Pma, Loki-Sri, Loki-Str, and Loki-Xtr [12]) in the multiple sequence alignments.

We identified ERVs that carry three distinct types of ENV or Fusogen proteins, namely glycoprotein F/Nipahvirus type (F-type), Herpes simplex virus type (HSV-type), and glycoprotein 2 type (GP2-type), and separately generated multiple sequence alignments. Briefly, we took HHpred alignments of F-type ENV against metapneumovirus fusion protein (Protein Data bank entries 7SEM_F and 5WB0_F), RSV F glycoprotein (6VKE_F and 6Q0S_F), parainfluenza virus fusion protein (4GIP_F and 2B9B_B), Newcastle disease virus (1G5G_E and 3MAW_A), paramyxovirus (1ZTM_C), canine Distemper virus (6XYE_B), and Nipah virus fusion glycoprotein (5EVM_D) and merged the predicted regions. We subsequently ran mafft-7.505 to generate a multiple sequence alignment, from which we manually removed ENVs with large truncations and trimmed sequences at either ends that are less conserved. We then ran a second multiple sequence alignment only including the trimmed and conserved regions of ENVs. We included F-type ENVs from SnRV and the above mentioned Loki retroviruses in the alignment. We took a similar approach for the HSV-type ENV. We took HHpred alignments of Human herpesvirus 5 prefusion glycoprotein B (7KDP_B and 5C6T_A), Herpes simplex virus type I glycoprotein B (5V2S_A), Human alphaherpesvirus 3 glycoprotein B (6VN1_C and 2GUM_C), and Suid alphaherpesvirus 1 glycoprotein B, and merged the predicted regions. We then ran mafft-7.505 once to generate a multiple sequence alignment, which we further trimmed either ends to obtain alignments from a highly conserved region. HHpred found homology to Syncytin-2 (7OIX_A) for all the GP2-type ENV ORFs that we identified. Some GP2-type ENV ORFs additionally showed homology to the N-terminal receptor binding domain of Feline Leukemia Virus glycoprotein (1LCS_B), Murine Leukemia virus glycoprotein 70 (1AOL_A), and the Human endogenous retrovirus EnvP(b)1 (6W5Y_B). We first aligned the full length ORFs of the GP2-type ENVs, and trimmed either ends to have C-terminal regions that are homologous to Syncytin-2. ENVs with large truncations are removed. We then performed a second alignment using a set of intact and trimmed GP2-type ENV ORFs. Among the above mentioned ORFs from the representative vertebrate retroviruses, we included BLV, FLV, HTLV-2, KoRV, LDV, MMTV, MuLV, REV, RSV and SRV-1 in the alignment of the GP2-type ENVs. We found that two Ctenophora errantiviruses that we identified, namely, Ctenophora_Tentaculata_Pbac_errantivirus_3 and Ctenophora_Tentaculata_Pbac_errantivirus_5 carry both the HSV-type and F-type ENV ORFs with HSV-type ENV being at the upstream immediately downstream of POL. In this case we only included the HSV-type ENV ORFs in the analysis.

1496 ERVs from all phyla combined retained intact Integrase and RT domains in POL and one of the three types of ERVs, and thus were used for further analyses. We also included the above mentioned 5 Drosophila errantiviruses and 23 vertebrate retroviruses for phylogenetic analyses. 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, those from ERVs with each of the three types of ENVs, and the trees of each ENV type. Branches of the tree of all Integrase domains that support bootstrap values of >95 are shown in Supplementary Figure 5.

Branches are then defined mainly based on how closely they cluster together on the tree, as well as the relative distance from neighbouring branches. In some cases, multiple branches are combined based on their relative position to each other from the root. Branch definitions are summarised in Supplementary tables 3 and 4. We then utilize these branch definitions to analyse congruency between *int* and RT, *int* and *env*, and *int* and host phylogeny, generating balloon plots using ggballoonplot from ggpubr 0.6.0 and ggplot2 3.4.2 packages in R to represent the number of shared elements between branches. The raw data for these tables is also summarised in Supplementary tables 3 and 4. We also used the branch definitions for *int* and RT from Supplementary table 4, to select representative ERVs for structural prediction of the RNase H domain using Alphafold 2, discussed later (Supplementary table 6).

For our analysis on host phylogeny, we separately defined branches that contain ERVs found in Arthropoda or vertebrates using a similar approach, but this time excluding branches that do not contain either. These Arthropoda and vertebrate branch definitions are summarised in Supplementary table 2. We labelled these clades A1-A20 for Arthropoda, and V1-V9 for vertebrates. We then constructed a balloon plot of branch against host species. The size of each data point was made proportional to species representation in a branch. For example, a balloon of 0.5 for an order in one clade, would mean that 50% of the elements found in that order are found in that given clade.

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

Long terminal repeats (LTRs) of ERVs were identified as follows. Firstly, we fetched the genomic coordinates of ERV insertions and extended 3000 nucleotides at either ends from the Start codon of GAG and the Stop codon of ENV ORFs (POL was taken for the GAG-ENV-POL ERVs) for all species categories except for Cnidaria, Ctenophora and Deuterostomia where we extended 5000 nucleotides at either ends from the Start and Stop codons of POL. We retrieved the genomic sequence of the extended ERVs and ran BLASTn against the same sequence to identify repeated sequences. We considered repeated regions of greater than 99% identity and the tBLASTn bit score between 200 and 5000 as putative LTRs. We excluded LTRs whose distance between them is shorter than 5000 nucleotides or longer than 14000 nucleotides. Following, we searched for putative tRNA primer binding sequences (PBS). tRNA sequences were retrieved from genomic annotations from representative genomes in respective species categories except for Ctenophora where we took tRNA sequences from Protostomia, Cnidaria and Deuterostomia because there were no annotated genomes available at the time of the study. Representative genomes that we used to retrieve tRNA sequences are: GCF_000001215.4 and GCF_013141755.1 for Diptera, GCF_014905235.1 and GCF_905147365.1 for Lepidoptera, GCF_000002335.3 and GCF_917563875.1 for Coleoptera, GCF_003254395.2 and GCF_016802725.1 for Hymenoptera, GCF_014356525.2 and GCF_020184175.1 for Hemiptera, GCF_002891405.2 and GCF_021461395.2 for Polyneoptera, GCF_921293095.1 for Palaeoptera, GCF_000671375.1, GCF_020631705.1, GCF_002217175.1, GCF_016086655.3, GCF_024679095.1, GCF_019059575.1 for Arthropoda, GCF_000326865.2, GCF_020536995.1, GCF_000002985.6, GCF_000715545.1 for Protostomia, GCF_000222465.1 and GCF_022113875.1 for Cnidaria, GCF_013122585.1, GCF_000003605.2, GCF_000003815.2, and

GCF_000002235.5 for Deuterostomia, and GCF_015237465.2, GCF_000001405.40, GCF_017654675.1, GCF_900496995.4, GCF_900700375.1, and GCF_010993605.1 for Vertebrata. The minimum requirements of the priming events are incompletely understood, but the length of PBS for retrotransposons is generally between 8 and 18 nucleotides and the PBS is positioned near the 3' end of the 5' LTR [23]. 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 1150 out of 1496 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 Supplementary table 1. We also plotted a graph to represent the types tRNA present in branches on the *int* tree, summarised in Supplementary table 5 and represented in Supplementary figure 5.

<u>Structure prediction of the "bridge region" using Alphafold 2</u>

Using branch definitions from the phylogenetic trees of *int* and RT, we selected representative ERVs from each branch to predict the structure of the region of POL that spans between the RT and Integrase domains. We refer to this region of POL as "bridge region". We found that ERV POL proteins invariantly carry a highly conserved RNase H domain that is located immediately C-terminal to the RT domain. We defined the start of this conserved RNase H domain as the N-terminal end of the bridge region. The C-terminal end of the bridge region was defined right before the start of N-terminal domain of the Integrase, which carries two conserved Histidine and Cytosine residues at the catalytic core [22]. We identified the conserved RNase H domain and the Integrase domain using HHpred alignments to the structure of the budding yeast Ty3/gypsy POL, 7Q5B_C. We then used the colabfold v1.4.0 [24] wrap of Alphafold 2 [25] to predict the structure of the bridge regions. We used Maestro (https://www.schrodinger.com/products/maestro) to visualise the structures, and characterised visible structures by the orientation and order of β sheets present, grouping them by letter. Once the first set of proteins were analysed, elements with unique structures were further analysed by characterising the bridge regions from the remaining ERVs in that branch. The results are summarised in Supplementary table 6. Representative structures of the bridge region and the topology of β sheets and α helices are shown in Figure 3 and Supplementary figure 4.

Figure Legends

Figure 1 – Phylogenetic relationships between *gypsy* errantiviruses and vertebrate endogenous retroviruses (vERVs). (A) Maximum likelihood tree constructed based on *int* proteins of *gyspy* errantiviruses and vERVs, with host species annotated in color. (B) Distribution of vERVs/*gypsy* errantiviruses in metazoa with *env* type indicated.

Figure 2 – Representation of *env* type in phylogenetic tree of *int* in *gypsy* errantiviruses and vERVs.

Figure 3 – Domain architecture and 3D structure of Ribonuclease H (RNase H) in *gypsy* errantiviruses and vERVs. (A) Representation of RNase H domain architecture in phylogenetic tree of *gypsy* errantiviruses and vERVs. (B) Alphafold-predicted 3D structure of canonical RNase H, 'mini' RNase H and 'pseudo' RNase H domains, with topology maps of α helices (grey) and β sheets (red). (C) Distribution of RNase H domain architectures within insects, vertebrates, and ancient animals.

Supplementary Figure 1 – Phylogenetic relationships between *gyspy* errantiviruses/vERVs and their hosts. (A) Arthropoda (Red) (A1-A20) and Vertebrate (Blue) (V1-V9) specific branches on maximum likelihood tree constructed based on *int* protein. (B) Balloon plot representing Arthropoda/Vertebrate clade to order congruence. Size of balloon represented in proportion of total elements from an order, where 1 represents a balloon that contains all the elements from that order. (C-E) Close ups of clades O-Q from Figure 1A. (F) Expanded and detailed colour code for host organism data represented in maximum likelihood trees.

Supplementary Figure 2 – Phylogenetic congruence between *int* and *env* for *gypsy* errantiviruses and vERVs. Trees for *int* and *env* with branches/discrete clades annotated in colour, matching the annotation in the balloon plot of number of elements shared between clades in *int* and *env* for elements containing type F (Fig S2A), GP2 (Fig S2B) and HSV (Fig S2C). Size of balloon is proportional to number of elements shared between discrete branches of *int* and *env*.

Supplementary Figure 3 – Phylogenetic congruence between *int* and RT proteins of vERVs and *gypsy* errantiviruses. (A) Maximum likelihood tree constructed based on RT proteins of *gypsy* errantiviruses and vERVs. (B) Balloon plot of number of elements shared between clades in *int* and RT. Size of balloon is proportional to number of elements shared between discrete branches of *int* and RT.

Supplementary Figure 4 – 3D structure of uncommon and yeast RNase H-like domains in *gypsy* errantiviruses. (A) Alphafold-predicted 3D structure of uncommonly found RNase H-like domains with topology maps of α helices (grey) and β sheets (red). (B) 3D structures and topology maps of 'mini' RNase H domains in yeast Ty3/*gypsy* elements

Supplementary figure 5 – Bootstrap values and tRNA primers of *gypsy* errantiviruses and vERVs. (A) Maximum likelihood tree constructed using *int* proteins. Red circles represent branches/nodes with bootstrap values >95. (B) Type of tRNAs carried by *gypsy* errantiviruses and vERVs in discrete branches from the *int* tree.

Supplementary table 1 – Master *gypsy* errantivirus and vERV table. Information on name, host classification, genome assembly, genomic coordinates, *env* type, type of tRNA used as primer, LTR coordinates and the number of mismatches between annotated tRNA sequences and the tRNA primer binding sequence in tab 1. Copy number of *gypsy* errantivirus and vERVs in tab 2, tRNA classification guide in tab 3.

Supplementary table 2 – Raw data for supplementary figure 1. Arthropoda branch definitions for *int* in tab 1. Raw balloon plot table of host order/class against Arthropoda branch in tab 2, with total number of elements from order/class in tree given. Vertebrate branch definitions for *int* in tab 3. Raw balloon plot table of host order/class against Vertebrate branch in tab 2, with total number of elements from order/class in tree given.

Supplementary table 3 – Raw data for supplementary figure 2. Branch definitions for *int* carrying F type *env* in tab 1. Branch definitions for *env* F in tab 2. Raw balloon plot of *int* vs *env* for F carrying ERVs in tab 3. Branch definitions for *int* carrying GP2 type *env* in tab 4. Branch definitions for *env* GP2 in tab 5. Raw balloon plot of *int* vs *env* for GP2 carrying ERVs in tab 6. Branch definitions for *int* carrying HSV type *env* in tab 7. Branch definitions for *env* HSV in tab 8. Raw balloon plot of *int* vs *env* for HSV carrying ERVs in tab 9.

Supplementary table 4 – Branch definitions of *int* and RT maximum likelihood trees. Branch definitions for *int* in tab 1. Branch definitions for RT in tab 2. Raw balloon plot data for supplementary figure 3B of *int* vs RT in tab 3.

Supplementary table 5 – Raw data for supplementary figure 5B. tRNA primer and *int* branch definition for each *gypsy* errantivirus and vERV.

Supplementary table 6 – RNase H domain information given. Names of representative elements taken for RNase H analysis with letter denoting type of domain found. (A) RNase H, (B) mini–RNase H, (C) outlier type 1 (dark blue in supplementary figure 4). (D) pseudo-RNase H, (E) outlier type 2 (pink in supplementary figure 4), (F) outlier type 3 (purple in supplementary figure 4).

Supplementary table 7 – Total list and details of analysed genomes. Diptera genomes in tab 1. Lepidoptera genomes in tab 2. Coleoptera genomes in tab 3. Hymenoptera genomes in tab 4. Hemiptera genomes in tab 5. Polyneoptera genomes in tab 6. Palaeoptera genomes in tab 7. Non-insect Arthropoda genomes in tab 8. Protostomia genomes excluding Arthropoda and Nematoda in tab 9. Nematoda genomes in tab 10. Cnidaria genomes in tab 11. Ctenophora genomes in tab 12. Porifera genomes in tab 13. Deuterostomia gemomes excluding vertebrates in tab 14. Vertebrate genomes in tab 15. Numbers of predicted ORFs are shown. ORFs that are found downstream of predicted POL were also counted for Protostomia, Cnidaria, Ctenophora, Porifera and Deuterostomia. Genomes that we used to annotate vERVs and errantiviruses are indicated.

References

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

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

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

4 Metegnier, G. *et al.* Comparative paleovirological analysis of crustaceans identifies multiple widespread viral groups. *Mob DNA* **6**, 16 (2015). https://doi.org/10.1186/s13100-015-0047-3

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

6 Boeke, J. D. & Stoye, J. P. in *Retroviruses* (eds J. M. Coffin, S. H. Hughes, & H. E. Varmus) (1997).

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

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

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

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

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

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

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

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

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

16 Yang, P., Wang, Y. & Macfarlan, T. S. The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. *Trends Genet* **33**, 871-881 (2017). https://doi.org/10.1016/j.tig.2017.08.006

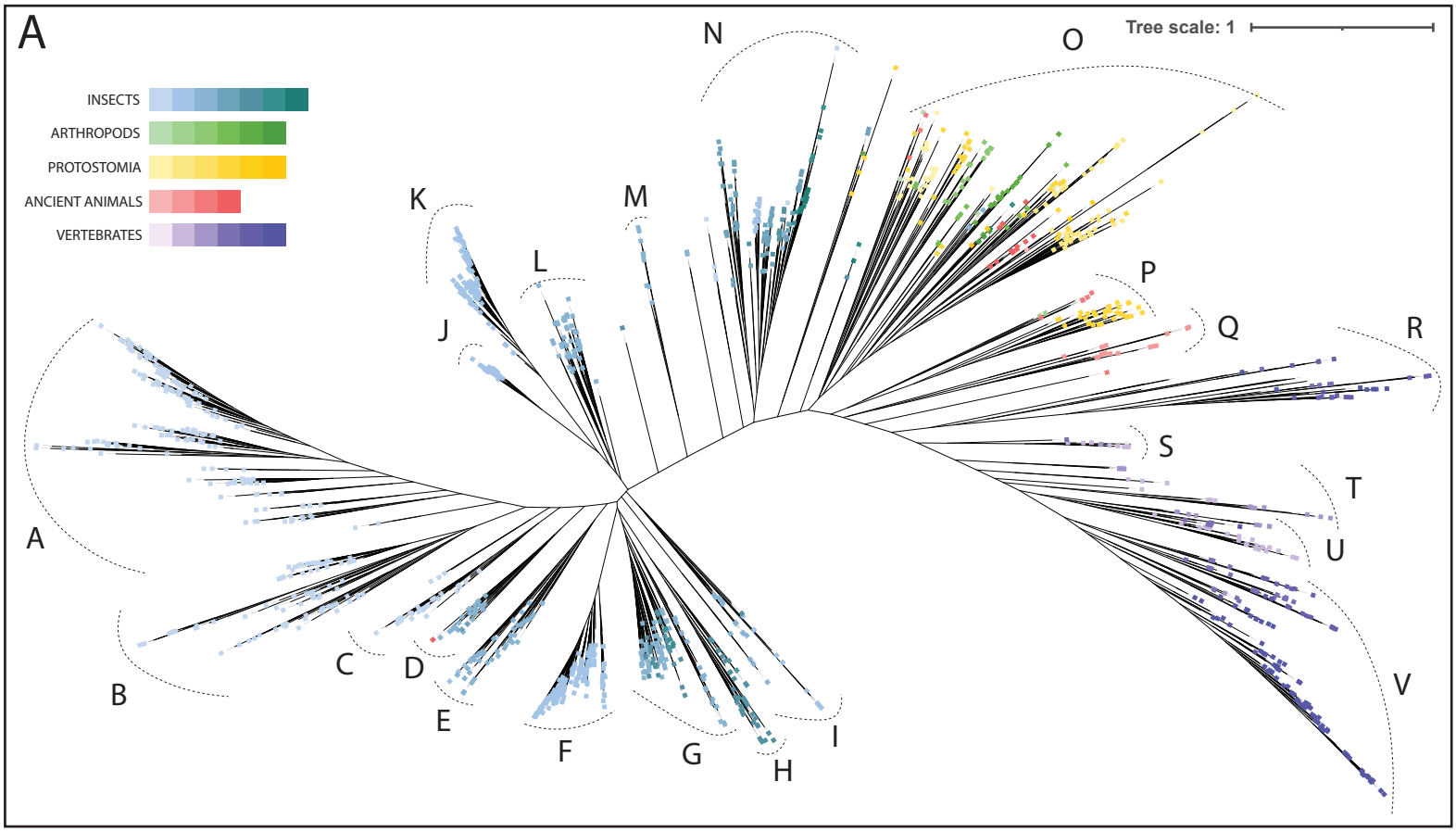
17 Czech, B. *et al.* piRNA-Guided Genome Defense: From Biogenesis to Silencing. *Annu Rev Genet* **52**, 131-157 (2018). https://doi.org/10.1146/annurev-genet-120417-031441

18 Yamashiro, H. & Siomi, M. C. PIWI-Interacting RNA in Drosophila: Biogenesis, Transposon Regulation, and Beyond. *Chem Rev* **118**, 4404-4421 (2018). https://doi.org/10.1021/acs.chemrev.7b00393

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

493 20 Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein
494 annotation. *BMC Bioinformatics* **20**, 473 (2019). https://doi.org/10.1186/s12859-019-3019-7
495 21 Das, D. & Georgiadis, M. M. The crystal structure of the monomeric reverse transcriptase
496 from Moloney murine leukemia virus. *Structure* **12**, 819-829 (2004).
497 https://doi.org/10.1016/j.str.2004.02.032
498 22 Engelman, A. & Cherepanov, P. Retroviral Integrase Structure and DNA Recombination
499 Mechanism. *Microbiol Spectr* **2**, 1-22 (2014). https://doi.org/10.1128/microbiolspec.MDNA3-
500 0024-2014.
501 23 Wilhelm, M. & Wilhelm, F. X. Reverse transcription of retroviruses and LTR retrotransposons.
502 *Cell Mol Life Sci* **58**, 1246-1262 (2001). https://doi.org/10.1007/PL00000937
503 24 Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679-
504 682 (2022). https://doi.org/10.1038/s41592-022-01488-1
505 25 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**,
506 583-589 (2021). https://doi.org/10.1038/s41586-021-03819-2

507

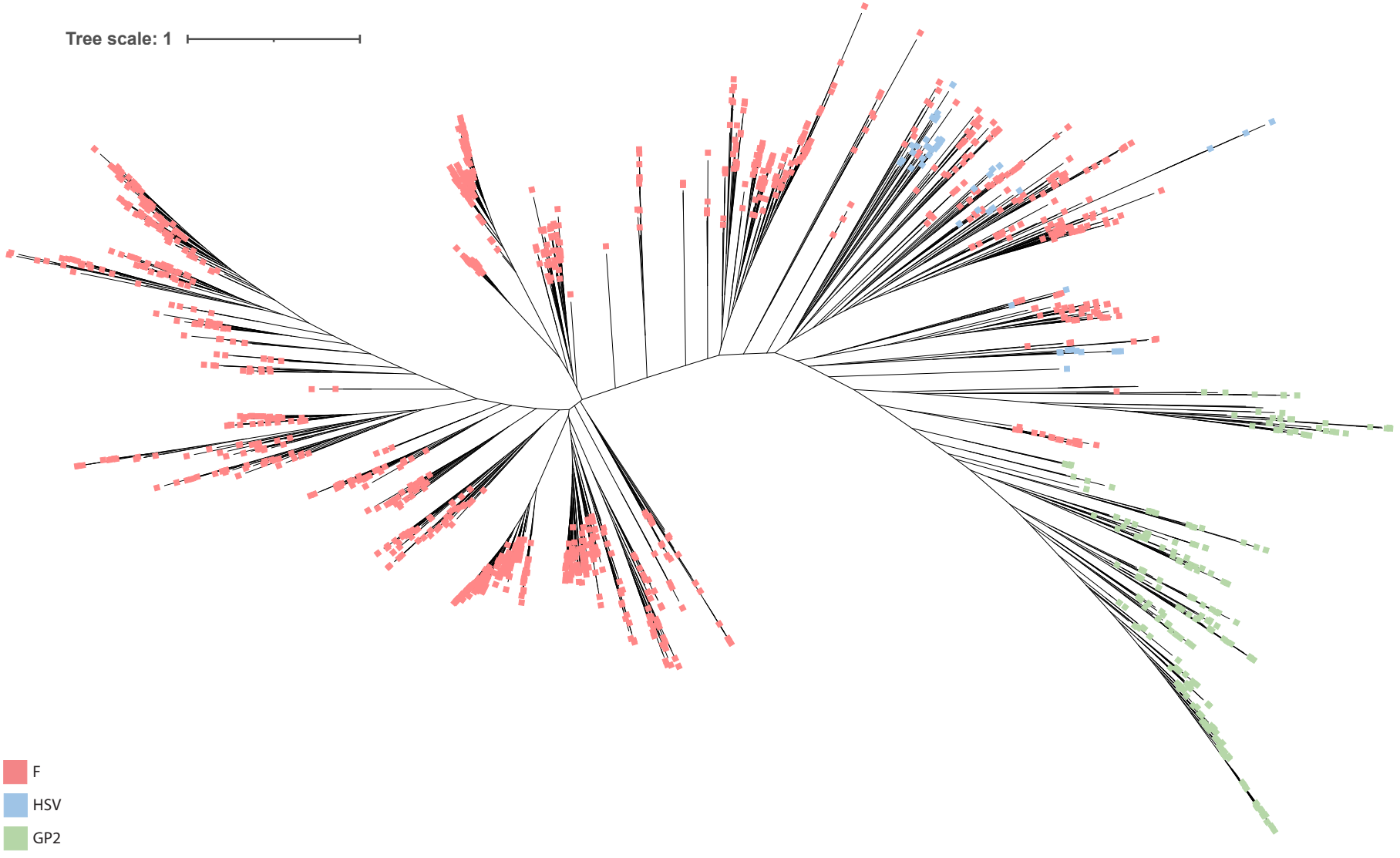


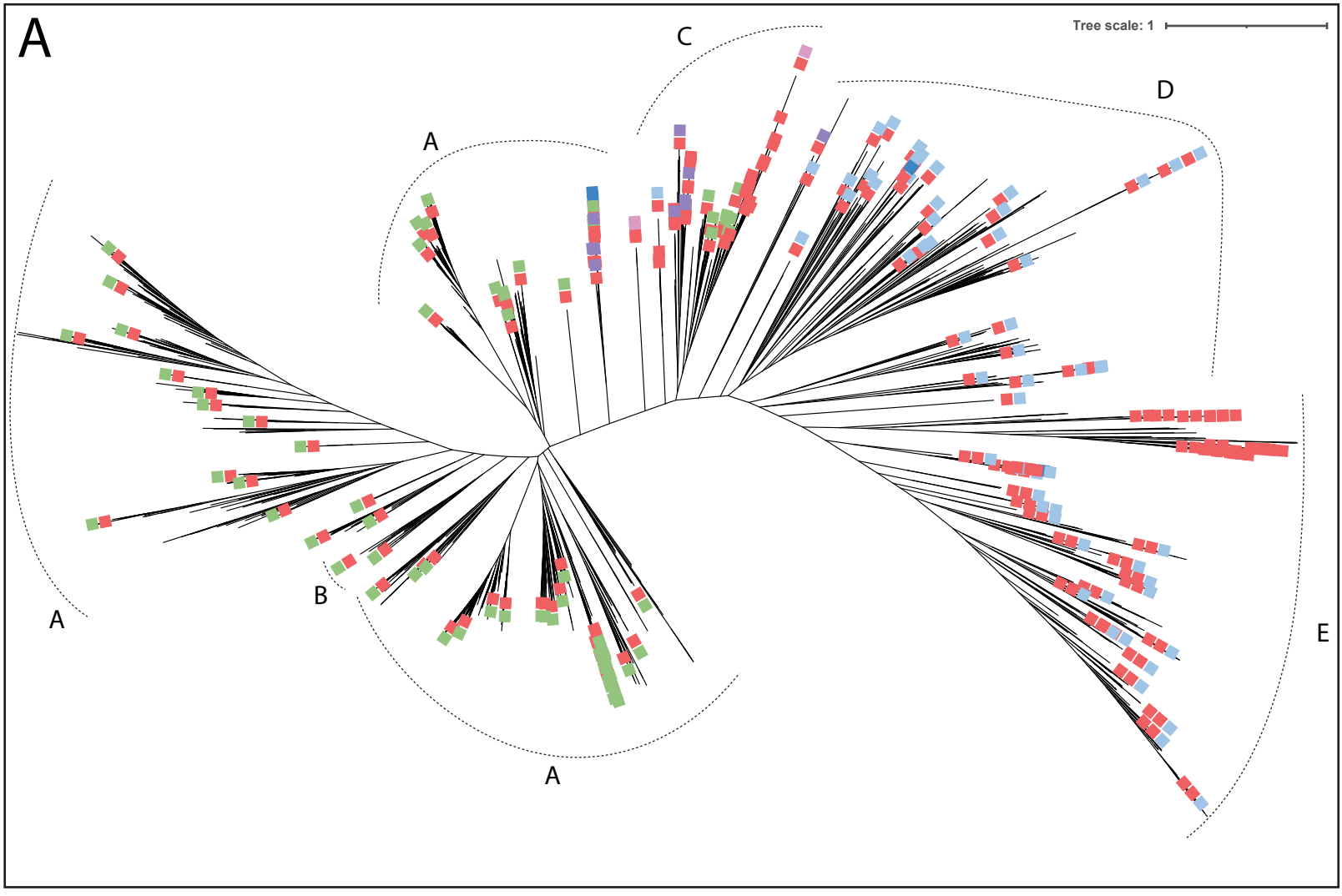
B

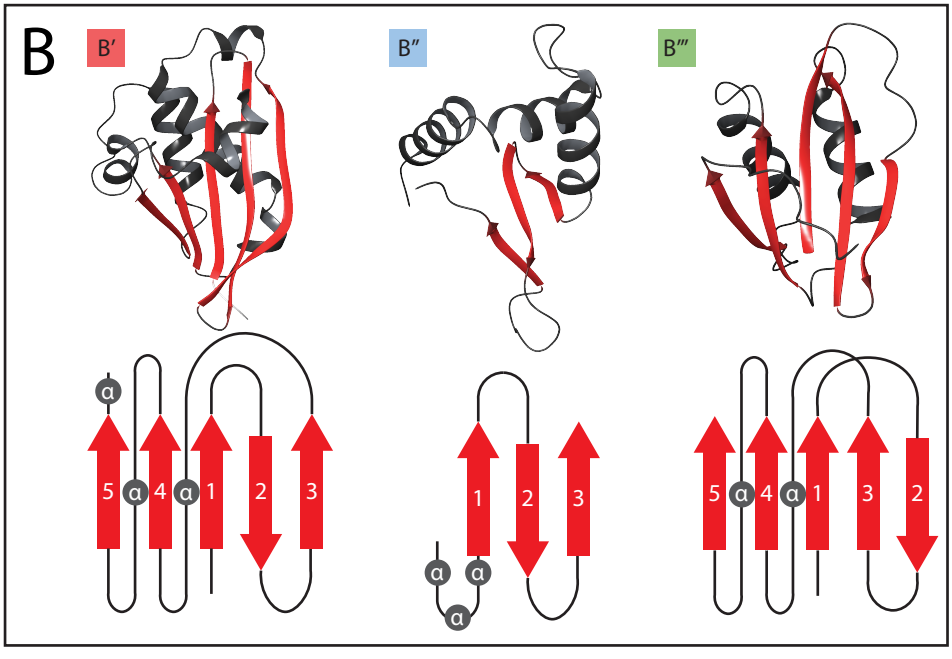
	GAG-POL-ENV	Env Type
Platyhelminthes (Lophotrochozoa, Protostomia)	Yes	HSV[1]
Nemertea (Lophotrochozoa, Protostomia)	Yes	F
Annelida (Lophotrochozoa, Protostomia)	Yes	F
Bryozoa (Lophotrochozoa, Protostomia)	Yes	HSV[1]
Mollusca (Lophotrochozoa, Protostomia)	Yes	F
Nematoda (Ecdysozoa, Protostomia)	Yes	HSV[1]
Arthropoda* (Ecdysozoa, Protostomia)	Yes	F
Vertebrata** (Chordata, Deuterostomia)	Yes	GP2, F
Tunicata (Chordata, Deuterostomia)	Yes	F, HSV
Hemichordata (Deuterostomia)	Yes	F
Echinodermata (Deuterostomia)	No	-
Cnidaria (Coelanterata, Diploblasta)	Yes	F, HSV
Ctenophora (Coelanterata, Diploblasta)	Yes	F, HSV
Porifera (Diploblasta)	No	-

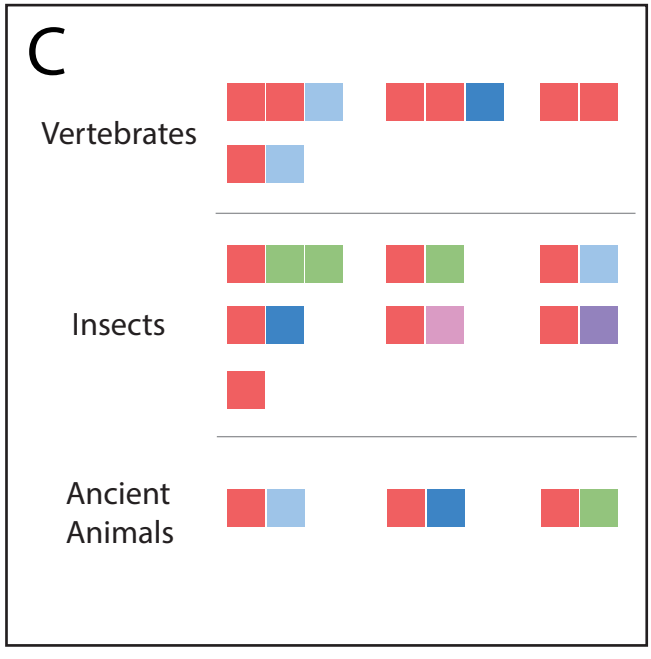
1 - Orientation of ORFs is GAG-ENV-POL

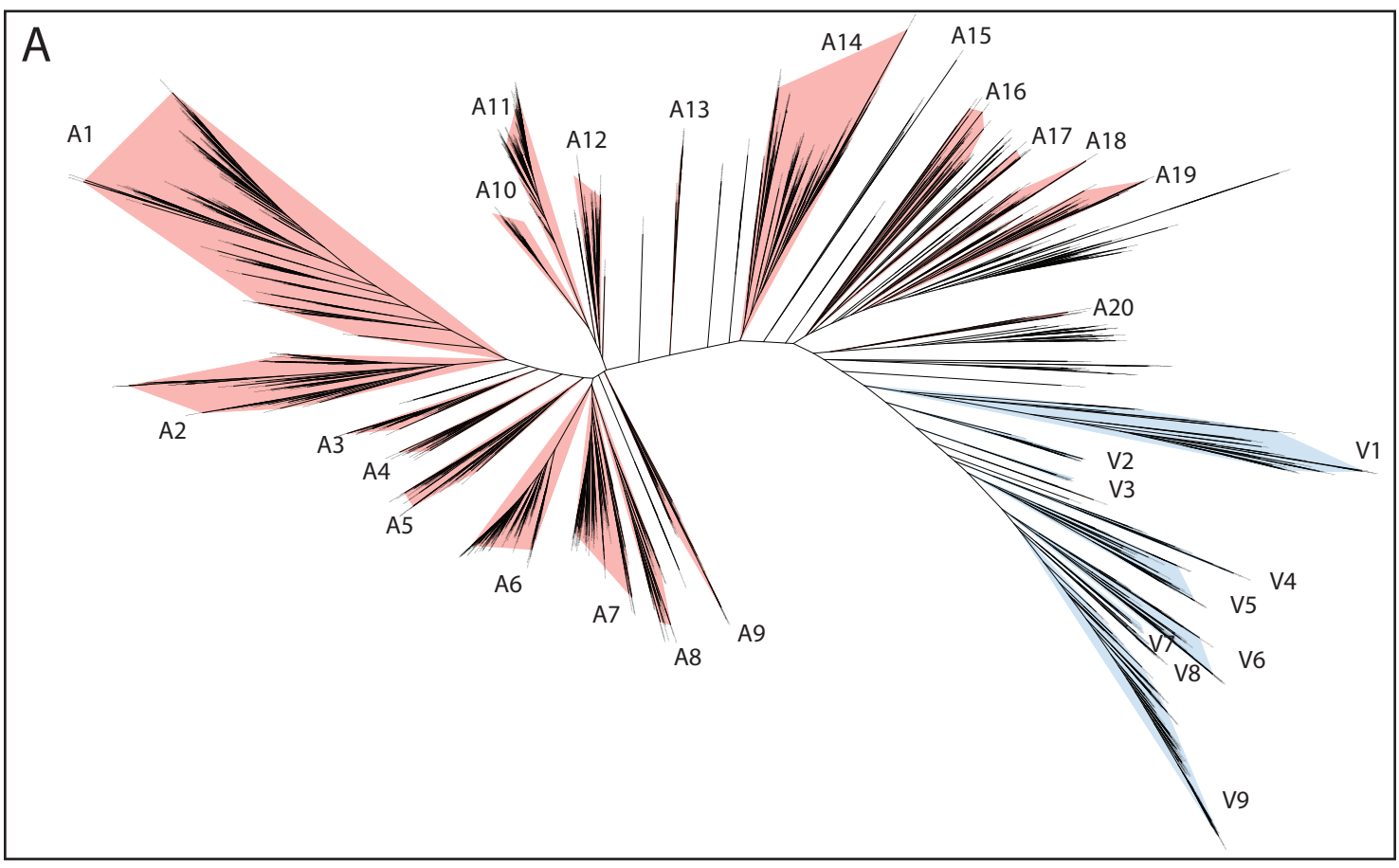
	GAG-POL-ENV	Env Type
Insecta*** (Arthropoda)	Yes	F
Collembola (Arthropoda)	Yes	F
Hexanauplia (Arthropoda)	Yes	F
Malacostraca (Arthropoda)	Yes	F
Thecostraca (Arthropoda)	Yes	F
Branchiopoda (Arthropoda)	Yes	F
Arachnida (Arthropoda)	Yes	F
Mammals (Vertebrata)	Yes	GP2
Birds (Vertebrata)	Yes	GP2
Reptiles (Vertebrata)	Yes	GP2, F
Amphibians (Vertebrata)	Yes	GP2, F
Fish (Vertebrata)	Yes	GP2, F
Jawless Fish (Vertebrata)	Yes	F
Diptera (Insecta)	Yes	F
Lepidoptera (Insecta)	Yes	F
Coleoptera (Insecta)	Yes	F
Hymenoptera (Insecta)	Yes	F
Hemiptera (Insecta)	Yes	F
Polyneoptera (Insecta)	Yes	F
Palaeoptera (Insecta)	Yes	F

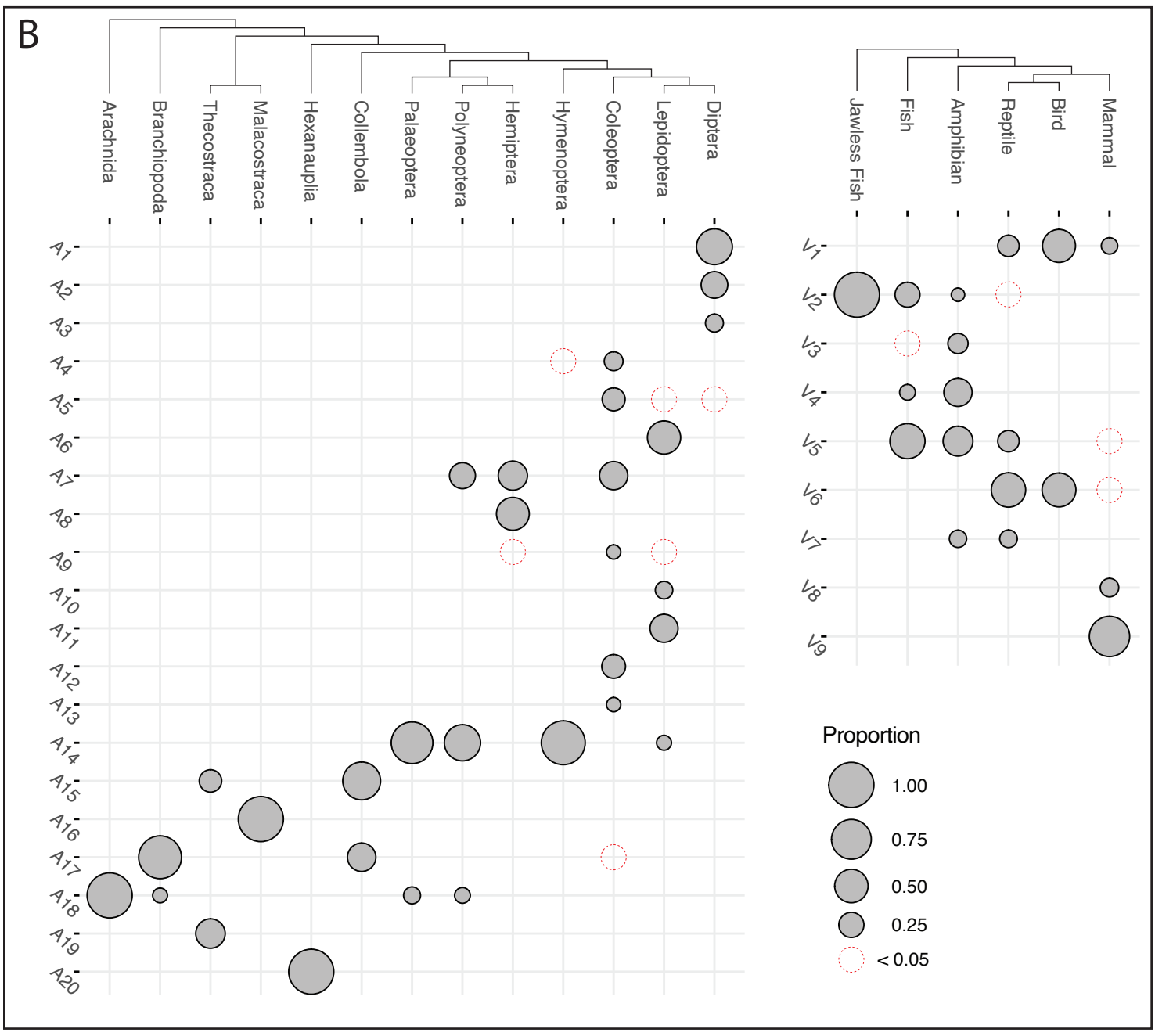

Tree scale: 1
F
HSV
GP2

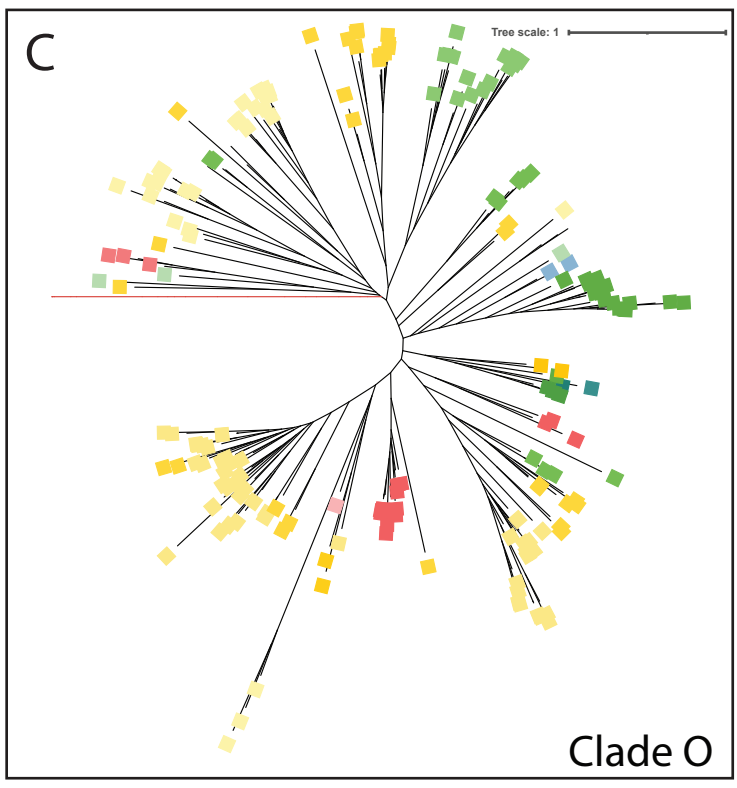

A
Tree scale: 1
A
C
D
A
B
A
E

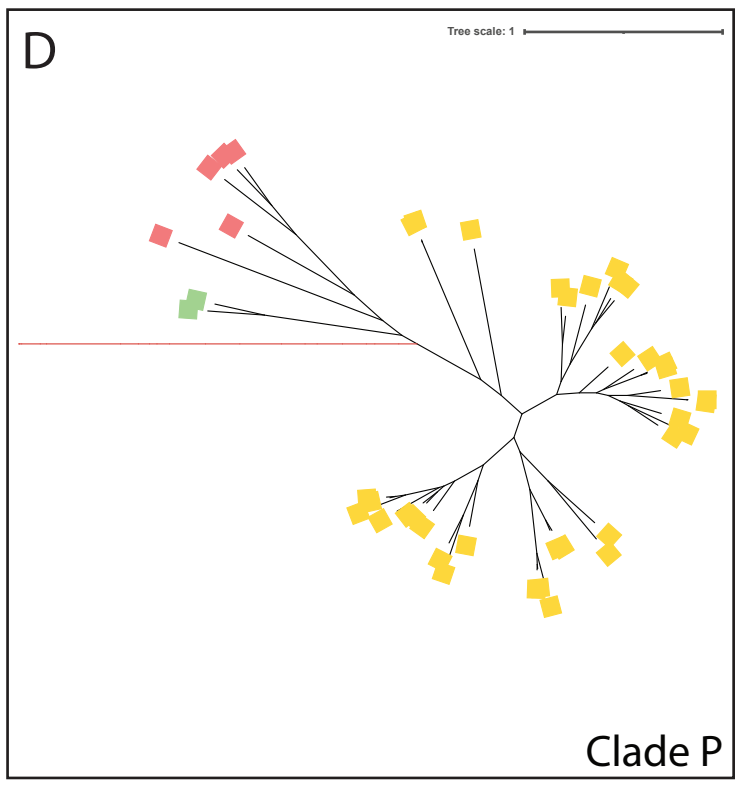

B
B′
B″
B‴
α
5
α
4
α
1
2
3
1
2
3
α
α
α
5
α
4
α
1
3
2

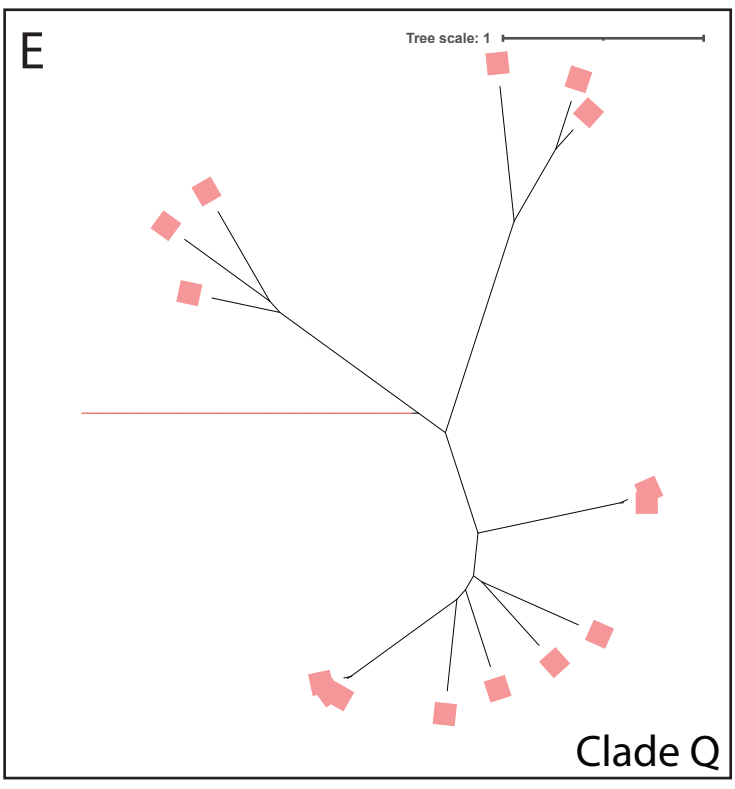

C
Vertebrates
Insects
Ancient Animals

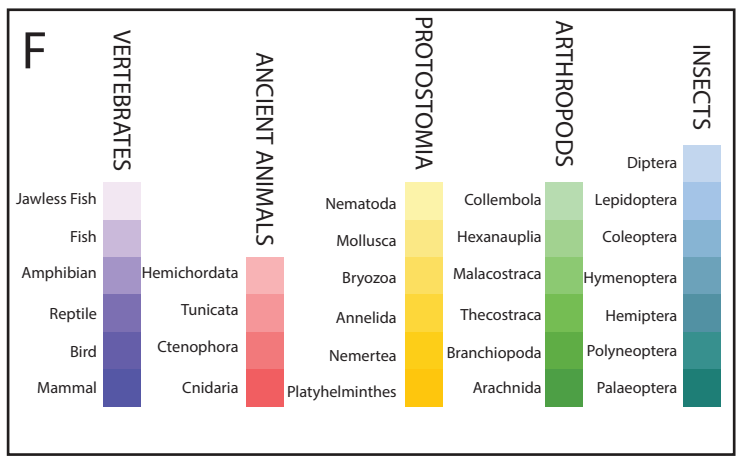

A
A1
A2
A3
A4
A5
A6
A7
A8
A9
A10
A11
A12
A13
A14
A15
A16
A17
A18
A19
A20
V1
V2
V3
V4
V5
V6
V7
V8
V9

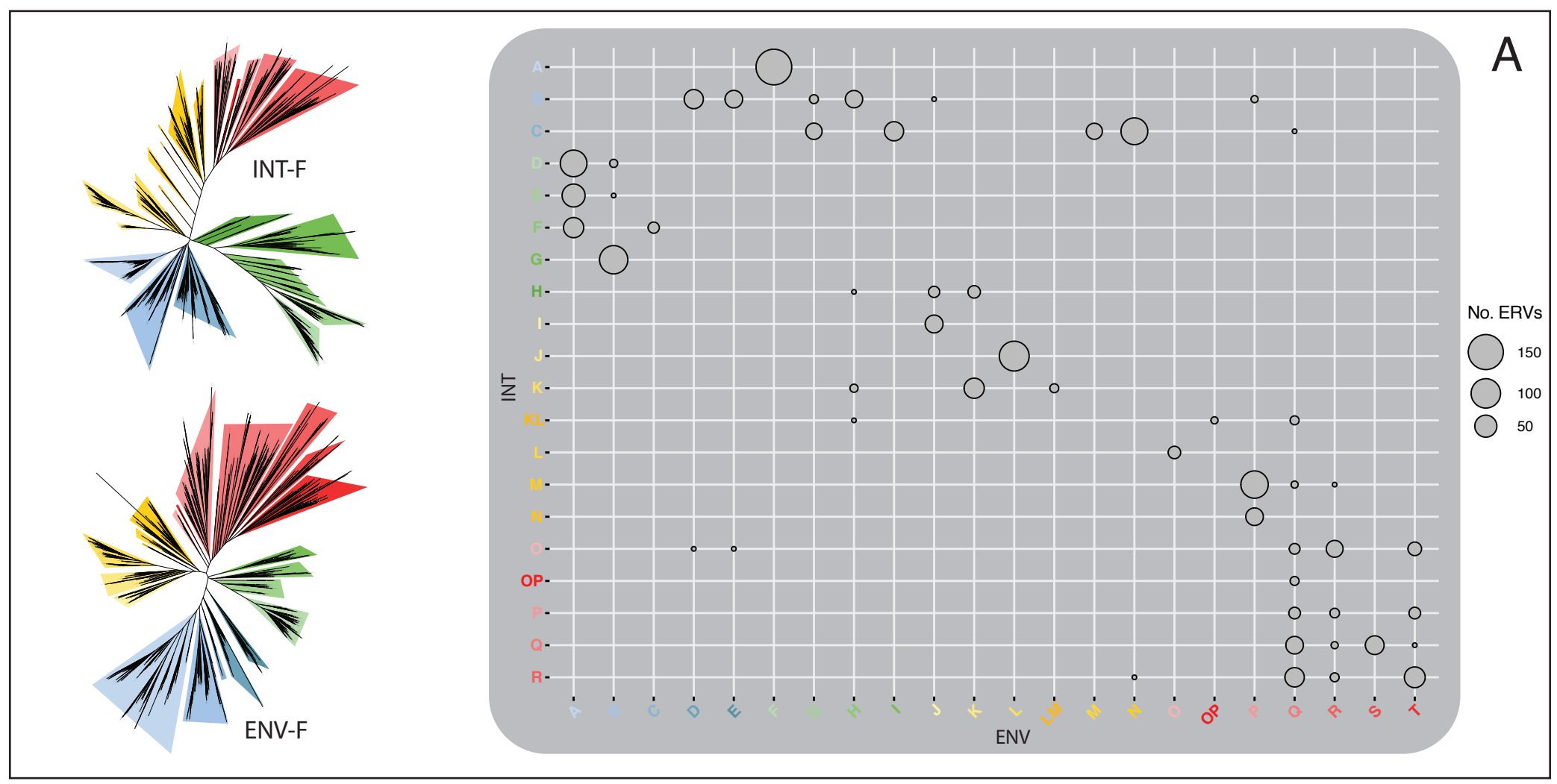

B
Arachnida
Branchiopoda
Thecostraca
Malacostraca
Hexanauplia
Collembola
Palaeoptera
Polyneoptera
Hemiptera
Hymenoptera
Coleoptera
Lepidoptera
Diptera
Jawless Fish
Fish
Amphibian
Reptile
Bird
Mammal
A1
A2
A3
A4
A5
A6
A7
A8
A9
A10
A11
A12
A13
A14
A15
A16
A17
A18
A19
A20
V1
V2
V3
V4
V5
V6
V7
V8
V9
Proportion
1.00
0.75
0.50
0.25
< 0.05

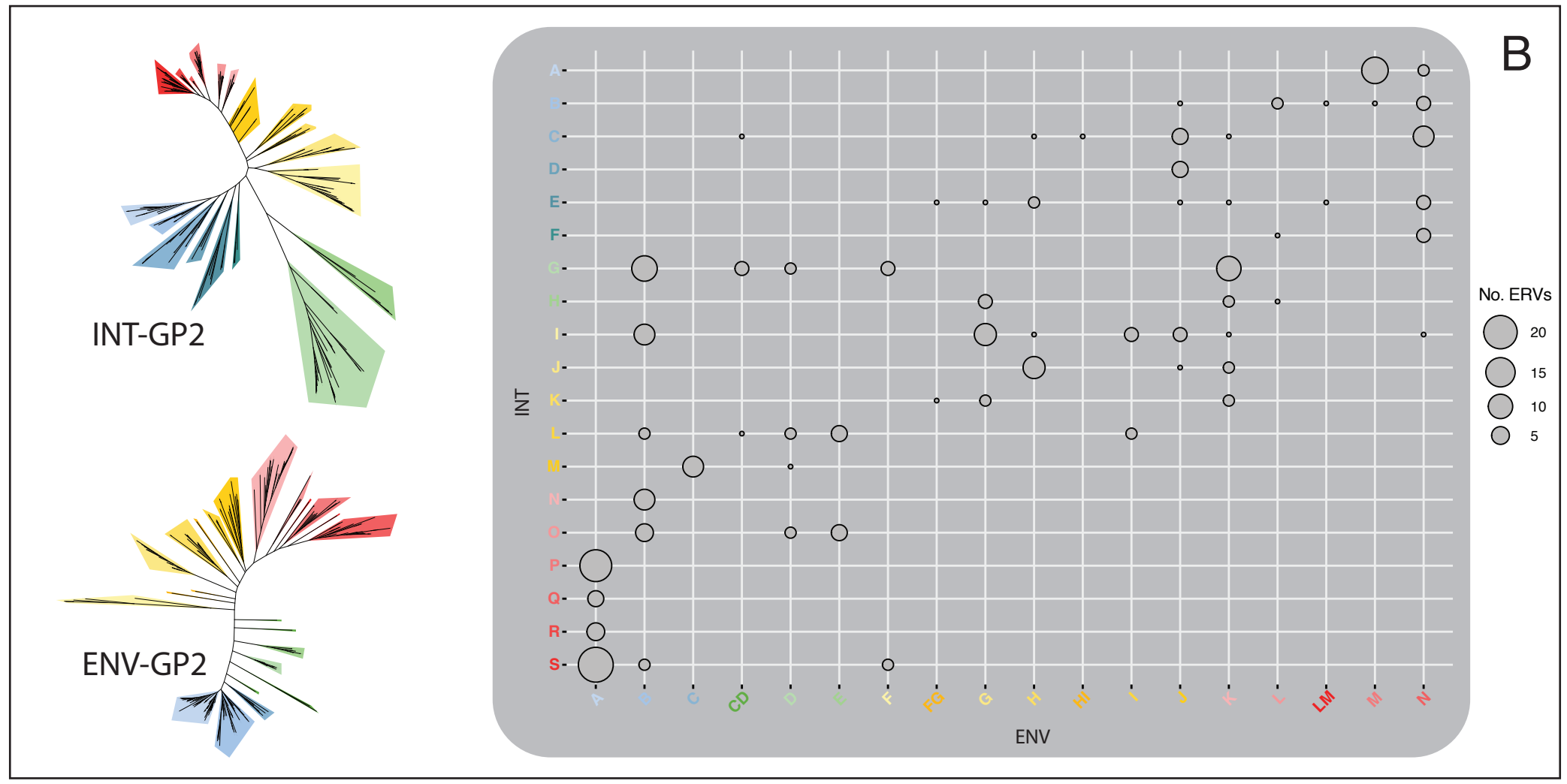

C
Tree scale: 1
Clade O

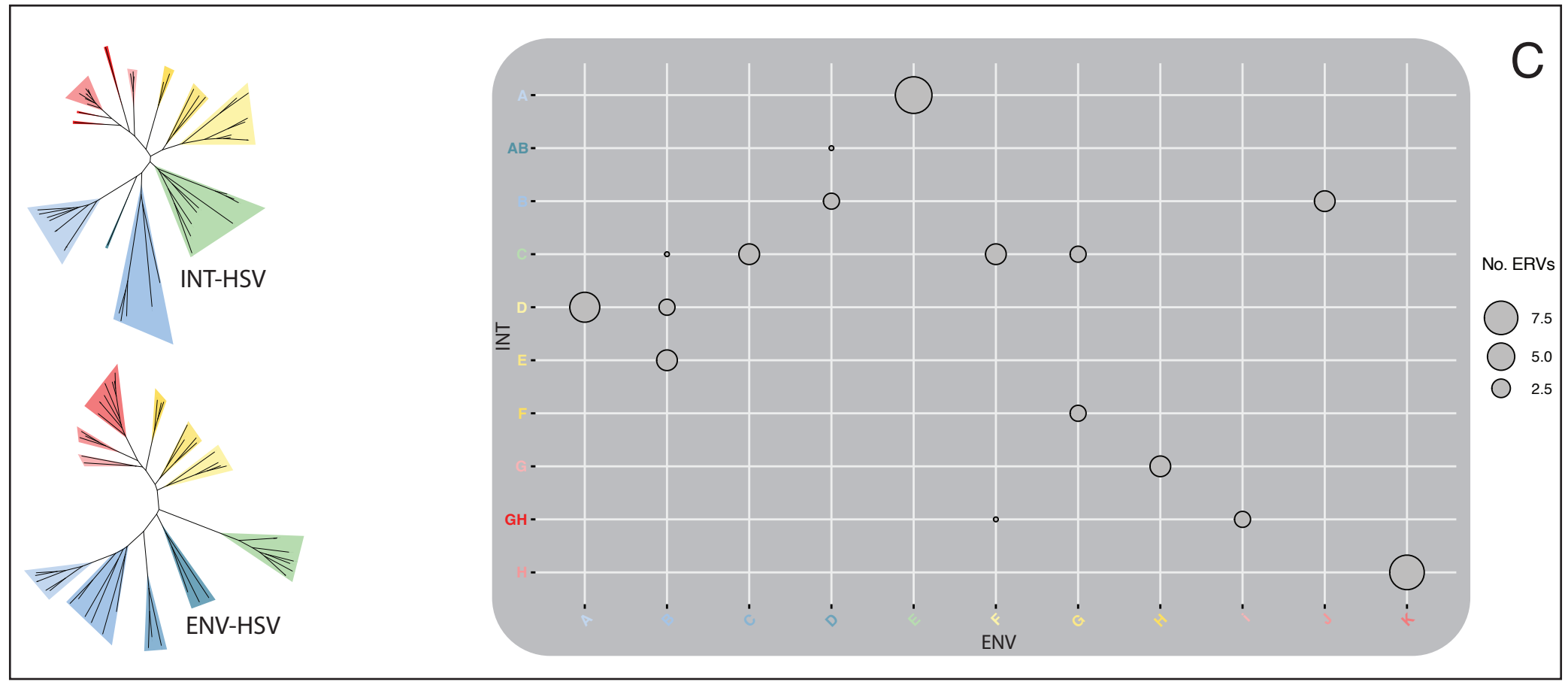

D
Tree scale: 1
Clade P

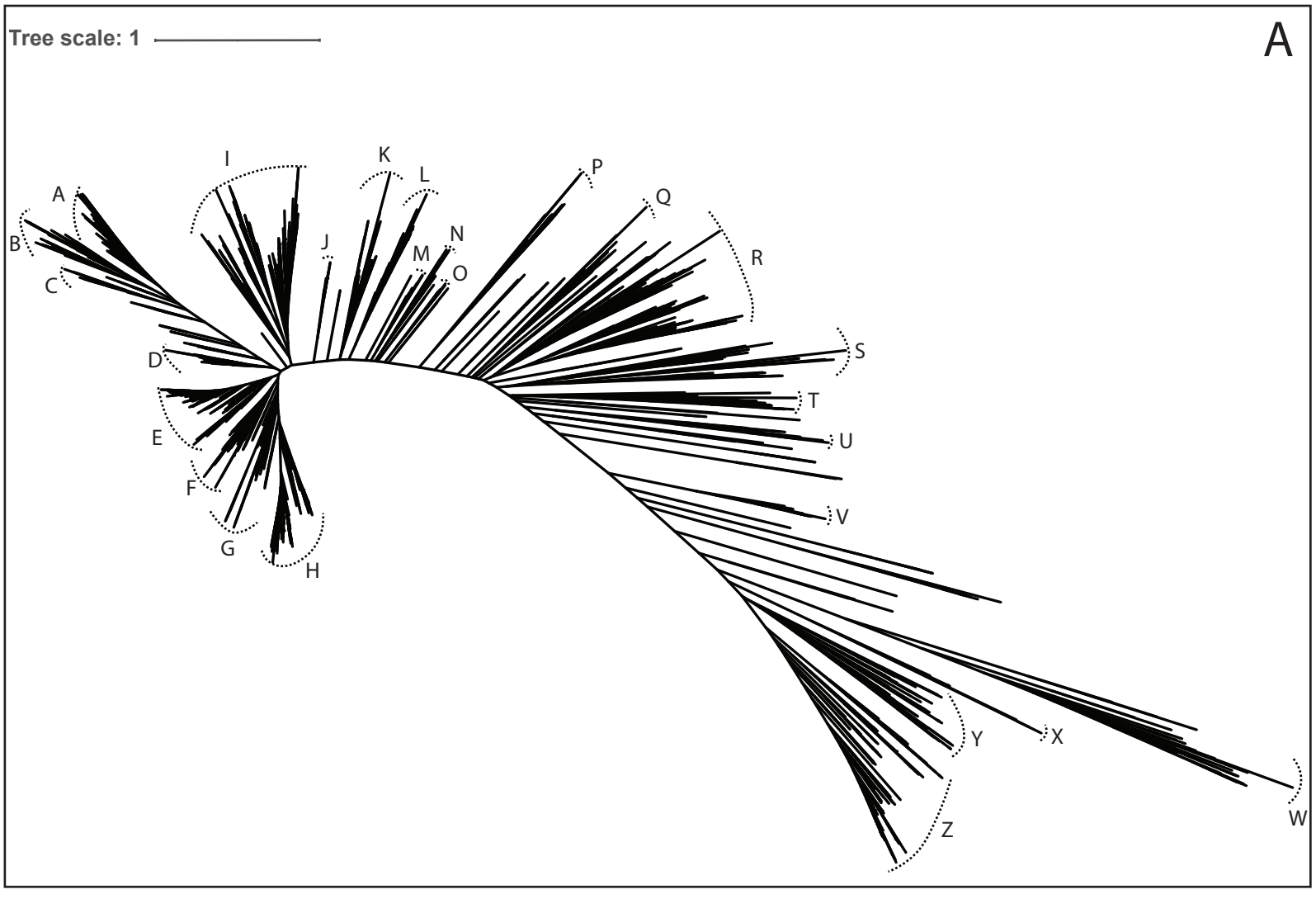

E
Tree scale: 1
Clade Q

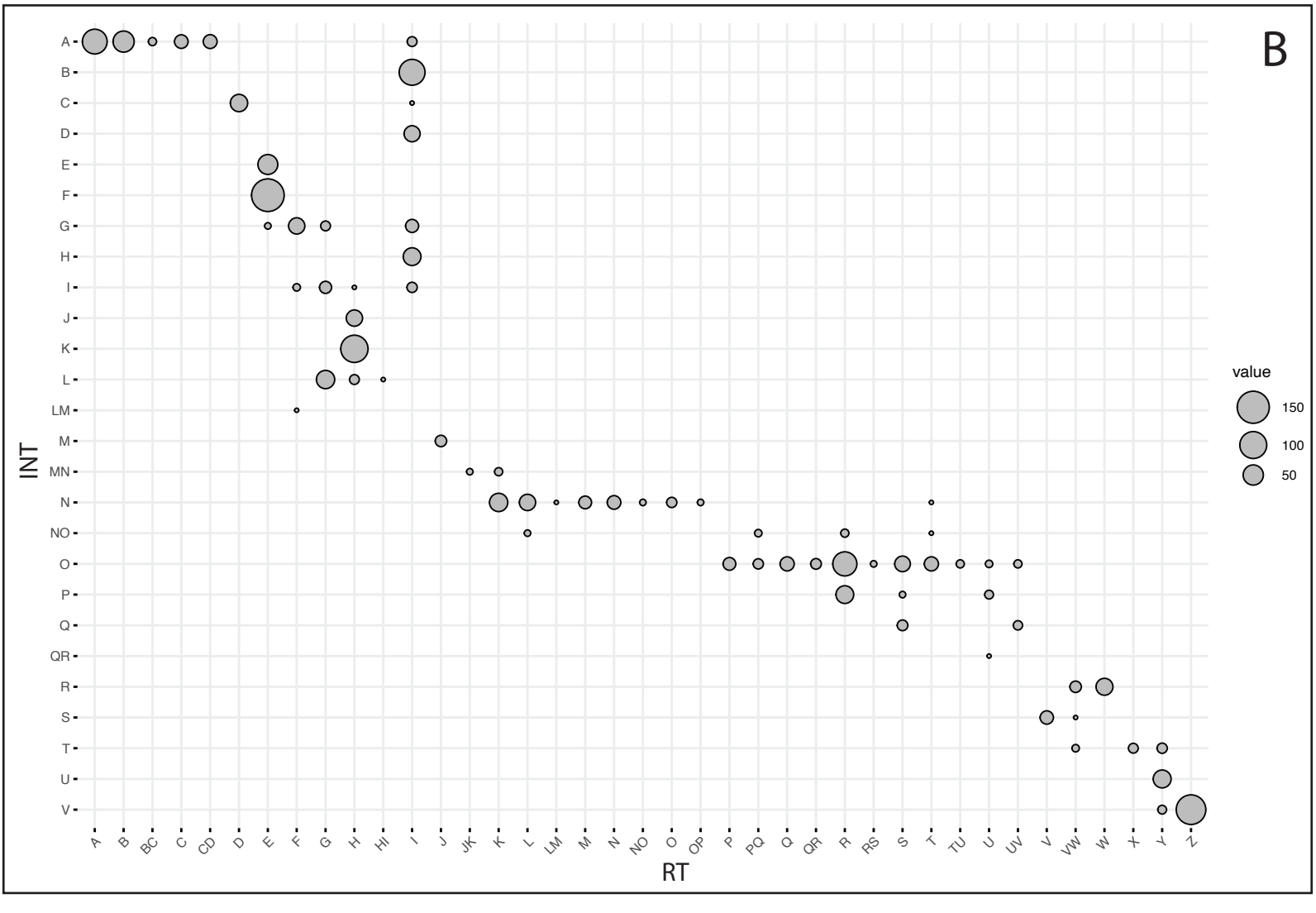

F
VERTEBRATES
Jawless Fish
Fish
Amphibian
Reptile
Bird
Mammal
ANCIENT ANIMALS
Hemichordata
Tunicata
Ctenophora
Cnidaria
PROTOSTOMIA
Nematoda
Mollusca
Bryozoa
Annelida
Nemertea
Platyhelminthes
ARTHROPODS
Collembola
Hexanauplia
Malacostraca
Thecostraca
Branchiopoda
Arachnida
INSECTS
Diptera
Lepidoptera
Coleoptera
Hymenoptera
Hemiptera
Polyneoptera
Palaeoptera

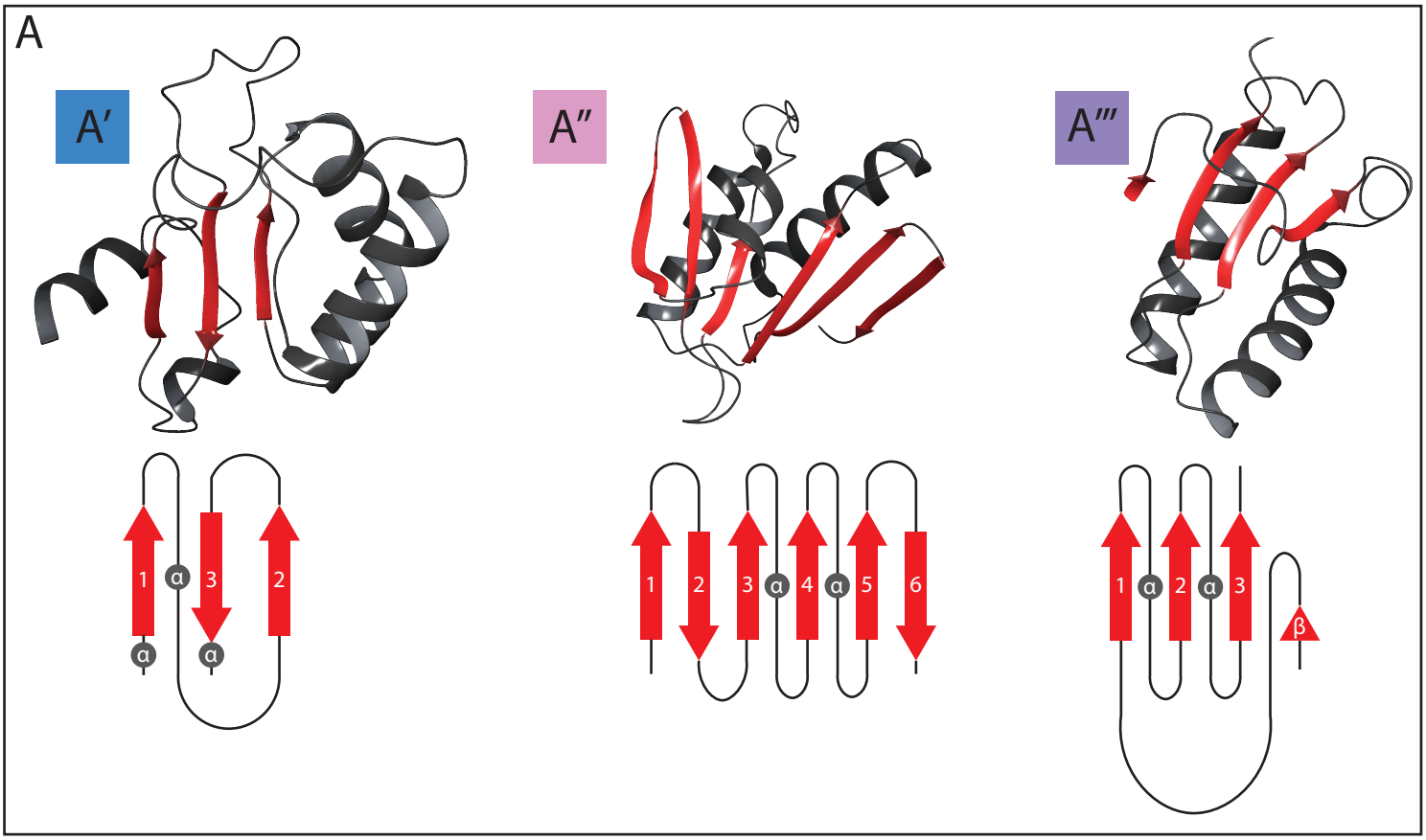

A
INT-F
ENV-F
INT
ENV
No. ERVs
150
100
50

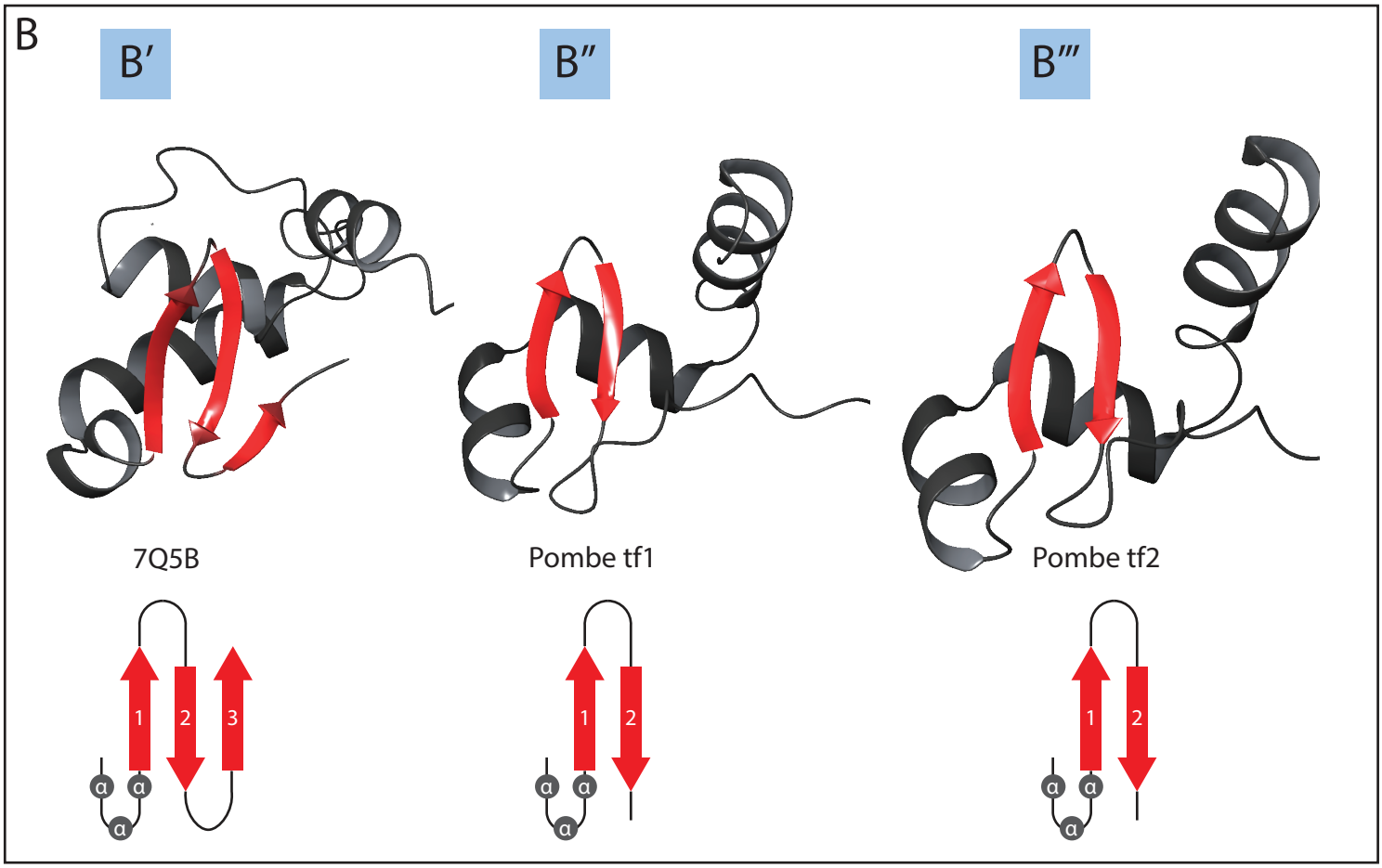

B
INT-GP2
ENV-GP2
INT
ENV
No. ERVs
20
15
10
5

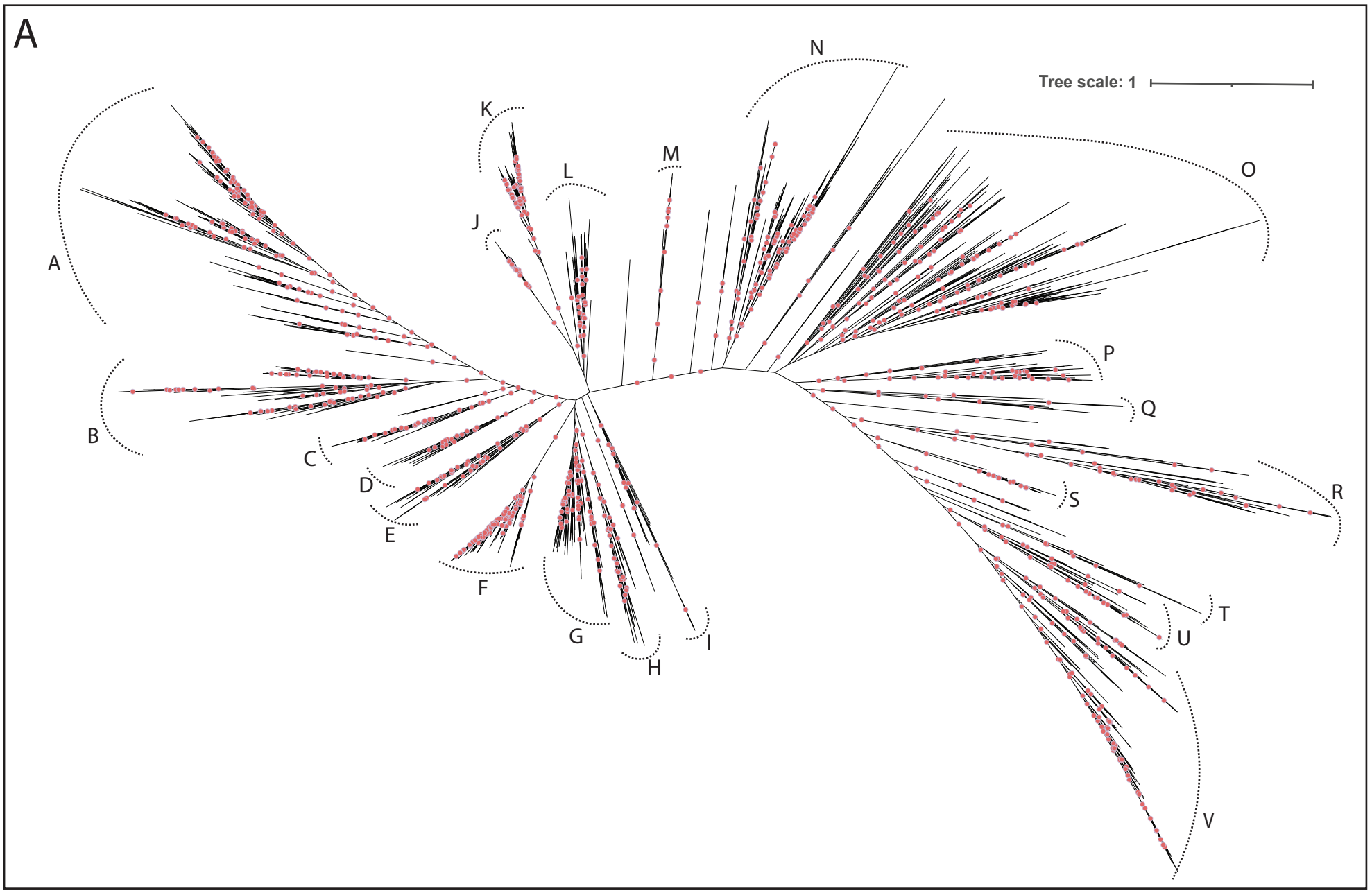

C
INT-HSV
ENV-HSV
INT
ENV
No. ERVs
7.5
5.0
2.5

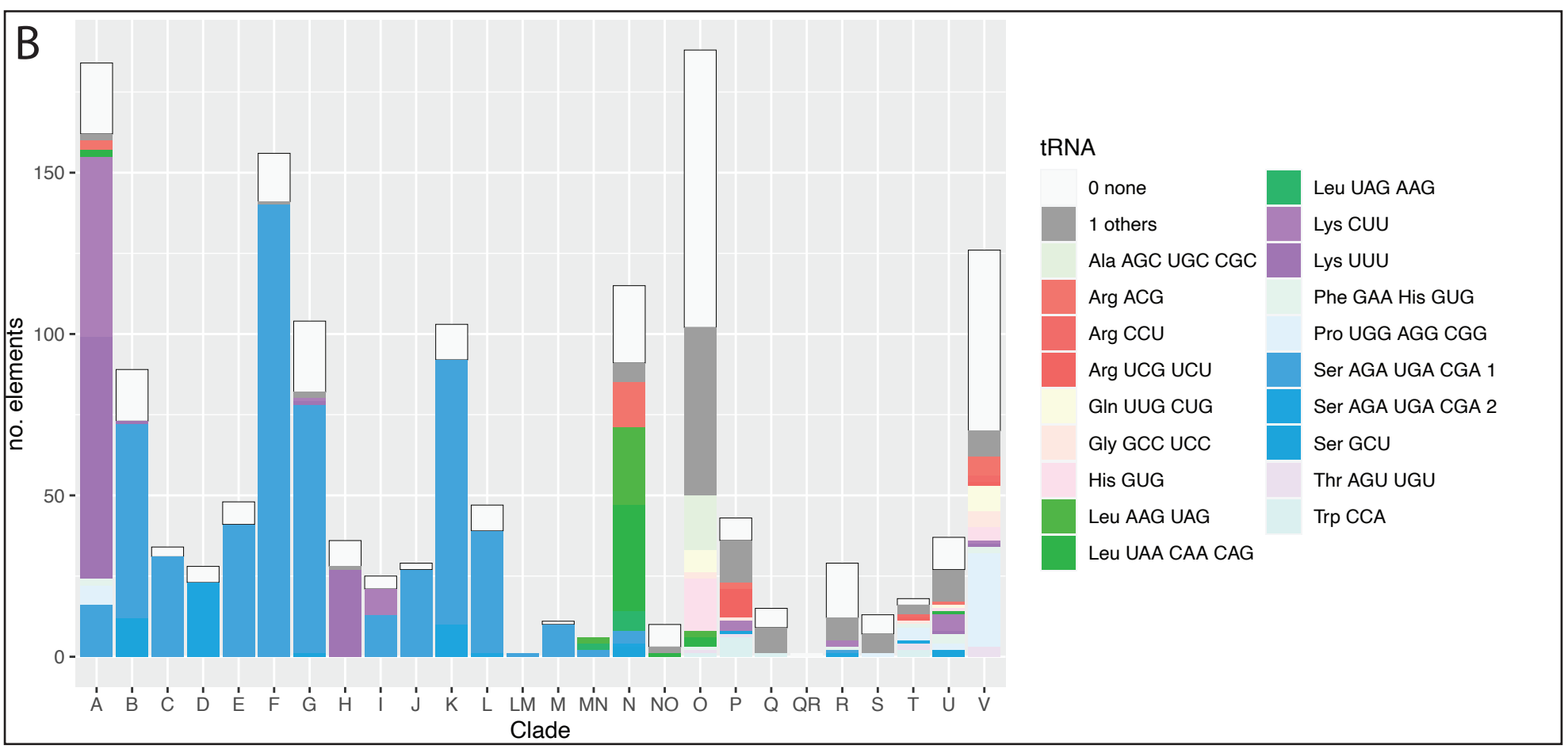

A
Tree scale: 1
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
X
Y
Z
W


B
INT
RT
A
B
C
D
E
F
G
H
I
J
K
L
LM
M
MN
N
NO
O
P
Q
QR
R
S
T
U
V
BC
CD
HI
JK
OP
PQ
RS
TU
UV
VW
W
X
Y
Z
value
150
100
50


A
A′
A″
A‴
1
α
3
2
α
α
1
2
3
α
4
α
5
6
1
α
2
α
3
β


B
B′
B″
B‴
7Q5B
Pombe tf1
Pombe tf2
1
2
3
α
α
α
1
2
α
α
α
1
2
α
α
α


A
Tree scale: 1
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V


B
no. elements
0
50
100
150
A B C D E F G H I J K L LM M MN N NO O P Q QR R S T U V
Clade
tRNA
0 none
1 others
Ala AGC UGC CGC
Arg ACG
Arg CCU
Arg UCG UCU
Gln UUG CUG
Gly GCC UCC
His GUG
Leu AAG UAG
Leu UAA CAA CAG
Leu UAG AAG
Lys CUU
Lys UUU
Phe GAA His GUG
Pro UGG AGG CGG
Ser AGA UGA CGA 1
Ser AGA UGA CGA 2
Ser GCU
Thr AGU UGU
Trp CCA