Intact genomic copies of envelope carrying retroelements are found widespread across 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 multiple intact copies 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, and likely share a common and highly ancient origin.

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.

Though Drosophila errantiviruses share structural similarities to vertebrate ERVs, their exact origin is poorly understood. It has been suggested that Ty3 elements acquired the *env* gene from infecting baculoviruses in an ancestral Drosophila species [11,12]. However, a preliminary search revealed that errantiviruses may be more widespread than previously thought. This combined with their unique cell niche in the Drosophila gonad prompted us to further explore their origins, as well as how widespread they are.

Results

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 1610 elements that carry nearly intact open reading frames of all three genes (Fig 1A, Table S1). A majority of the newly identified genomic copies carry putative LTRs and tRNA primer binding sequences (1205 elements, summarised in Fig S1B and Table S2), 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 were also included in the tree to aid classifications. We refer to elements found within and outside Vertebrata as vERVs and *gypsy* errantiviruses, respectively.

We found that intact copies of errantiviruses 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). We similarly find that intact copies of *env*-containing vERVs are found in all vertebrate orders including jawless fish, which is consistent with previous studies[13,14]. In both cases, each element has multiple intact copies in their respective host genomes (Table S1), indicating that they have likely been actively transposing recently.

We assembled phylogenetic trees using the sequences of the *pol* Integrase and RT domain, and identified discrete clades based on distance and supported by bootstrap values greater than 95 (Fig S1A, S2A) When we investigated where the elements in the Integrase tree are represented in RT, we found that the trees assembled are congruent with each other (Fig S2B, Table S3). For example, RT clade A only consists of elements found in Integrase clade A. In almost all cases, the clades in one tree are represented by similar clades in the other. RT clade N, for example, mostly consists of elements from Integrase clade O, but also contains elements from clade P, as well as elements that are not present in a clade but sit between Integrase clades N, O, P and Q. This congruence suggests that the Pol sequences analysed in this study have not undergone recent degradation, further confirming the 'intactness' of these elements.

We then characterised the type of tRNA primer binding sites (PBS) that are found in the errantiviruses that we identified in this study. Drosophila errantiviruses have been identified to have Serine or Lysine tRNAs priming the elements [15,16]. In our investigation we find that insect elements in Integrase clades A to N predominantly carry the same $tRNA^{Lys}$ and $tRNA^{Ser}$ PBS (Fig S1B). Clades A and H, which contain Diptera and Hemiptera elements respectively, mostly have $tRNA^{Lys}$ PBS, while the rest predominantly use Serine. The exception, however, is clade N, which consists of elements carrying $tRNA^{Leu}$, $tRNA^{Arg}$, $tRNA^{Glu}$, $tRNA^{His}$ and $tRNA^{Pro}$ in addition to $tRNA^{Ser}$ PBS. The tRNA primer binding site for Leucine is the most prominent in this clade and is found in elements from insect orders (Table S2). The $tRNA^{Ser}$ PBS in this clade is found in Hymenoptera and Diptera, while the rest outside this clade are only found within Hymenoptera. Outside of insects, in Integrase clades O and P, the tRNA PBS becomes more diverse, appearing to reflect the phylogeny of the elements. However, there appears to be some consistency in the primers. For example, Branchiopoda elements in clade O all use $tRNA^{Ala}$, Coleoptera elements in the same clade continue to use mainly $tRNA^{Ser}$ and $tRNA^{Leu}$, and Mollusca elements mainly use $tRNA^{His}$ (Table S2, Fig S1B). Some non-insect Arthropoda elements as well as two elements from Mollusca within these clades also use $tRNA^{Leu}$.

Elements in Vertebrate clades Q-V also have a large variety of tRNA PBS, but follow a similar pattern to clade O. For example, $tRNA^{Pro}$, $tRNA^{Gly}$ and $tRNA^{Gln}$ are mainly used by mammalian elements with a few elements from the other orders also using the same, while $tRNA^{Phe}$ and $tRNA^{Ser}$ are mainly used by elements in fish. While the type of tRNA PBS carried by an element tends to overall be roughly respective of their host species and their branch, the pattern is not consistent enough to draw a conclusion as to whether they reflect the overall divergence of the errantiviruses and vERVs on the tree.

The phylogenetic tree assembled using the sequences of the *pol* Integrase domains appears to separate into clades that reflect the hosts that they are found in (Fig 1A). For example, Integrase clades A, B and C exclusively contain dipteran elements, whereas clades F and J are exclusive to Lepidoptera. Some clades

91 contain a mix of hosts, though they are normally related. The Integrase clade E consists of mostly
92 coleopteran elements, but also contains some elements found in other holometabolous insect orders
93 Hymenoptera and Diptera as well. The notable exceptions are Integrase clades O and P, where elements
94 found in non-insect Arthropoda, non-Arthropoda Protostomia and ancient orders are found. Though, within
95 each respective clade, they still appear to separate based on their host (Fig S3C-D). Gamma-retroviruses
96 are found in Integrase clade S, while other known retroviruses, including alpha-, beta-, delta-, lenti- and
97 formy-retroviruses, are all found in Q. The more recently identified loki-retroviruses[17] are found in clade R
98 (Fig 1A, Table S3). Known dipteran errantiviruses Gypsy-Gypsy, and Gypsy-5-Dgri sit within clade A,
99 Gypsy-17.6 and Idefix sit within clade B, and Gypsy-ZAM sits within clade C (Fig 1A, Table S3). This
100 arrangement is consistent with previous studies that have classified retroviruses and Drosophila *gypsy*
101 elements [17-19].

102 To further explore the congruence between host phylogeny, we specifically identified discrete clades for
103 Arthropoda (A1-A24) and Vertebrata (V1-V12) in the *pol* Integrase tree and investigated the host makeup
104 of those clades (Fig S3A-B). The balloon plot is presented as a proportion of representation. A data point
105 of 0.5 would indicate that 50% of the elements from that host are represented in that clade. We found that
106 most elements within Arthropoda and Vertebrata clades are formed by elements either from single or
107 multiple related orders or classes (Fig S3B). For example, clades V3 and V5 only contain elements found
108 in mammals, and V1 contains elements from birds, but also contains elements from Mammals and
109 Reptiles. Within Arthropoda, the pattern is similar, though it should be noted that some of the elements
110 found in insects that sit within clades that contain non-insect Arthropoda carry a different type of *env* than
111 the others, namely, an *env* that resembles that from herpesviruses.

112 We further examined the Env protein found in these intact copies of errantiviruses and vERVs, as well as
113 the divergence between Env and Integrase. We firstly found that the errantiviruses identified in this study
114 carry two types of Env; those resembling fusion glycoproteins from Nipah viruses (referred to as
115 glycoprotein F), and those resembling glycoprotein B from herpesviruses (referred to as HSV) (Fig 2,
116 S4A-B). The vERVs identified also carry glycoprotein F, but most carry glycoprotein GP2, a typical
117 vertebrate retrovirus Env. While individual clades contain a mixture of Env representation, the congruence
118 between Integrase and Env appears to hold (Fig 2, Fig S4, Table S5). For example, Integrase clade O
119 contains both elements that carry HSV and F (Fig 2C), meaning that similar Pol proteins carry different
120 Env. However, when individual trees are made using only HSV or F, elements appear to form similar
121 clades between Integrase and Env (Fig S4), meaning that similar Pol proteins carry similar Env genes
122 individually for HSV or F. For example, clade H in the Integrase tree for HSV consists of elements only
123 from clade H in the Env HSV tree (Fig S4B). Similar congruence can be seen between Integrase and Env
124 for glycoprotein F (Fig S4A), though elements that carry GP2 are less congruent than either F or HSV (Fig
125 S4C).

126 We subsequently characterised the region between the Integrase and RT domains of Pol, referred to as the
127 "bridge region", for representative elements within the Integrase/RT phylogenetic trees. The bridge 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" [20]. We therefore characterised structures within the bridge region using the 3D fold arrangement of the core β-sheet structure. Aside from a few elements in insects, nearly all elements have duplicated the RNase H domain, which have undergone degradation and structural changes. We mainly found three types of structures widespread in the bridge region of Pol; the typical RNase H/Tether arrangement, a "pseudo" RNase H domain, and a "mini" domain (Fig 3C). The "pseudo" 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. The "mini" domain consists of three small core β-sheets. In addition, we found three other structures that are less commonly found in the bridge region (Fig S5A). The separation of clades in the Integrase tree appears to agree with the domain configurations of the bridge region of the elements, further supporting our clade definition. For example, all representatives taken from Integrase clade A (Fig 1) carry the "pseudo" domain (Fig 3A). Furthermore, Integrase clades O and P, which contains a wide variety of host species, mostly carry the "mini" domain. The exception is Integrase clade N, whose elements carry either the "pseudo" domain, or "pseudo" domain 3 (Fig 3A, Fig S5), though within the clade, the elements are still separated (Fig 3A).

The "pseudo" RNase H domain is mostly specific to insects only, whereas vertebrate elements are the only ones to have two RNase H-like/Tether domains within their bridge region (Fig 3). However, most vertebrate elements additionally carry the "mini" domain, which almost all elements in ancient animals, some elements in insects, and even yeast Ty3 elements carry in the bridge region of their Pol (Fig S5B). While insect elements mainly carry a "Pseudo" RNase H, some can also carry other uncommon domains, and they overall have greater variability in the structures present in the "bridge region". It should be noted that one Cnidaria element that sits in a clade with insects on the Integrase and RT trees carries the "pseudo" domain, and that the annotated Loki-retrovirus "loki-xtr" only carries one RNase H domain and the "mini" domain (Table S6).

Our findings show that multiple intact copies of errantiviruses and *env*-containing vERVs are ubiquitously present in metazoan genomes, indicating recent activity. Additionally, the errantiviruses identified in this study adhere to their hosts, and the phylogeny of ORFs for *env* and *pol* resemble each other. Finally, vERVs and errantiviruses share common features such as structures in the bridge domain of Pol, the structure of Env as well as their host-adhering nature.

Discussion

We found that *env*-containing errantiviruses are more widespread than previously thought. Genomes from nearly every major order contain multiple intact copies, with uninterrupted ORFs and intact flanking LTRs, which is suggestive of recent transposition. Though they are recent, they still appear to reflect the phylogeny of their hosts, meaning that errantiviruses found from one order are more like elements found either within the same order, or closely related ones, though there are exceptions which will be discussed shortly. Non-insect errantiviruses are structurally distinct from insect errantiviruses, which is reflected in

their own separate clades. For example, they do not carry a "pseudo" RNase H domain within the bridge region of Pol, but rather carry a "mini" domain. However, there are still insect errantiviruses present in these distinct clades, carrying same structural features as non-insect errantiviruses, with their Pol resembling those from ancient animals more than they do other insects. These findings suggest that the errantiviruses found in insects may have more ancient origins than previously thought, in addition to being highly widespread. 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 [11,12]. While Ty3 elements may have obtained the *env* gene via recombination, it is unlikely that baculoviruses are the source for the *env* gene found in errantiviruses, as they are more widespread than insects, and can carry different types of *env* genes as well. Additionally, whether errantiviruses originally acquired HSV or F type *env* is also unclear, as non-insect clades which include elements found in ancient animals, carry both types of *env*. Insect errantivirus clade N (Fig 2) also contains both types of *env*, but the elements that carry HSV type *env* are distinct from the other errantiviruses within this clade, further indicating that recombination events may have happened, but they are not happening frequently enough to randomise their spread.

There are a few ways to explain the widespread and multicopy nature of errantiviruses in metazoans, one of which is that these errantiviruses frequently undergo horizontal transfer. Drosophila errantiviruses have been previously suggested to horizontally transfer between species [19,21] and would explain their widespread presence across metazoans. It has also been suggested that some LTR retrotransposons in bivalves can horizontally transfer between different species and can even jump across phyla as well [22]. While horizontal transfer of errantiviruses between distantly related species cannot be excluded, it is likely to be a rare occurrence, given the congruence of the phylogeny of *int* with that of their hosts. Additionally, some of the different elements that cluster together are found in hosts that are magnitudes of order apart from each other, and it is unlikely that they can easily transfer. Furthermore, it is unlikely that horizontal transfer followed by re-entry into the germline can happen frequently enough to explain their widespread nature.

Alternatively, the widespread nature of these errantiviruses can be explained via an "endogenized retrovirus model", in which exogenous retroviruses enter the germline by chance, and become fixed in the population, where they remain as evolutionary relics of past infection, degrading over time [23,24]. Those exogenous viruses would likely be species specific, given that they are mostly unlikely to freely move away from their natural host species, often referred to as a species barrier [25,26]. This model is supported by studies that have identified exogenous virus counterparts of endogenous retroviral copies, for gibbon ape leukaemia viruses and koala retroviruses [27,28]. For example, the koala gamma retrovirus KoRV-A is the endogenous form of the exogenous KoRV-B retrovirus [29]. This model would explain the phylogeny of the errantiviruses following that of their hosts. However, exogenous retroviruses are widely considered to be vertebrate-specific[30]. We find it unlikely that elements that are as abundant and ubiquitous as we observed in invertebrate hosts remain undetected to date. Furthermore, some insect elements resemble those from ancient animals such as non-insect arthropods, Nematoda, Platyhelminthes and Thecostraca (Fig S3C-D),

202 which would indicate that the potential exogenous viruses are not limited to one host species, contrary to
203 the "endogenized retrovirus model".

204 Another possibility that explains both the phylogeny as well as the widespread and multicopy nature of
205 these elements, is that they are highly ancient germline transposons that continue to remain active. In this
206 scenario, the last common ancestor of metazoans would have had this element in their genome, which
207 remained active as species diverged. As long as one of the transposon copies remains active, it would be
208 vertically inherited, and diverge along with the genome of the host species, whereas older and inactive
209 copies would be degraded over time. Additionally, it is possible that *env* was present in the original LTR
210 retrotransposon, or an ancient recombination event is responsible for the presence of the *env* gene.

211 One question that arises, is why these elements continue to retain *env* over such long periods of time,
212 remaining potentially infectious when expressed. In Drosophila, errantiviruses occupy the somatic cell
213 niche in the ovaries that surround the germline, and the expression of *env* allows the Ty3 virus-like
214 particles to infect the germline to create new copies [9]. As an adaptation, Drosophila species have
215 duplicated the germline transposon defence mechanism to target errantiviruses in the somatic cells of the
216 gonad [31] . It is possible that this is more universal, and that *env*-containing retrotransposons occupy the cell
217 niches surrounding the germline, and that expression of *env* allows them to continue to infect the germline.
218 For example, murine IAP retrotransposons that carry *env* genes also remain infectious, similar to
219 Drosophila errantiviruses [32]. Given the widespread nature of intact copies of *env* containing retroelements
220 identified in this study, it may, therefore, also be possible that a defence mechanism against them is more
221 common.

222 We finally found that vERVs are similarly as widespread as errantiviruses, and like them, also carry
223 multiple intact copies in their respective genomes. While retroviruses are known to have duplicated the
224 RNase H domain and degrade the catalytic residues to form a 'tether' domain [20], duplications of RNase H
225 have frequently occurred in errantiviruses as well, and have diverged into a "pseudo" domain. Although
226 the "tether" domain is strictly found in vERVs, the duplication of the RNase H domain alone may not
227 separate retroviruses from other errantiviruses. We also identified a "mini" domain towards the C-terminus
228 of the bridge region of vERVs, which is a feature shared with errantiviruses in ancient animals, and even
229 yeast Ty3 elements. More strikingly, some vERVs carry F-type *env* proteins, which are carried by a
230 majority of the errantiviruses identified in this study. These elements form a separate clade, along with the
231 recently identified loki-retroviruses, which were also found to carry the same *env* type [17]. One possible
232 explanation for this is that vERVs undergo recombination events with other viruses. However, given that
233 the F-type vERVs form their own clade, it is unlikely that this happens frequently or recently, else they
234 would be more widespread across the phylogenetic tree. Another possibility is that the divergence of
235 vERVs and errantiviruses occurred after the acquisition of *env*, which would imply that they are equally
236 ancient, and may have a common evolutionary ancestor and origin.

237

Acknowledgements

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

Author contributions

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

Data Availability

The following data are available from the git repository (https://github.com/RippeiHayashi/errantiviruses_ERVs): the code used to annotate and analyse genomic copies of errantiviruses and vERVs, genomic sequences and open reading frames of the newly identified ERVs, pdb files for the predicted structures of 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.

257 **Materials and methods**

258 Genome assemblies used in the study

259 Reference genome sequences were retrieved from NCBI and grouped into the following insect orders,
260 vertebrate classes and other metazoan phyla: Diptera, Lepidoptera, Coleoptera, Hymenoptera, Hemiptera,
261 Polyneoptera, and Palaeoptera for insects; jawless-fish, fish, Amphibian, reptile, bird, and mammal for
262 vertebrates; Arachnida, Branchiopoda, Chilopoda, Collembola, Diplopoda, Hexanauplia, Malacostraca,
263 Merostomata, Ostracoda, and Thecostraca for Arthropoda; Acanthocephala, Annelida, Brachiopoda,
264 Bryozoa, Dicyemida, Mollusca, Nemertea, Onychophora, Orthonectida, Phoronida, Platyhelminthes,
265 Priapulida, Rotifera, and Tardigrada for Protostomia; Cephalochordata, Echinodermata, Hemichordata, and
266 Tunicata for Deuterostomia; Cnidaria, Ctenophora, and Porifera. We obtained lists of reference genomes in
267 each category in May 2023 using the Entrez Direct command line tool except for the vertebrate genomes
268 which we used the list from the previous work by Wang et al. [17]. The list of the genome assemblies used
269 can be found in Supplementary table S7.

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

272 We ran two iterations of tBLASTn (ncbi-blast-2.9.0) to search for the genomic copies of vERVs and *gypsy*
273 errantiviruses, firstly using the baits of known elements, and secondly using the baits identified from the
274 first iteration. We used two sets of sequences for the first tBLAST search. The first set included GAG,
275 POL and ENV protein sequences of the Dipteran *gypsy* errantiviruses from the RepBase
276 (https://www.girinst.org/repbase/). The second set included GAG, POL and ENV protein sequences of
277 selected vertebrate retroviruses that were used in the previous study [17]. We applied different criteria to
278 identify ERVs from species outside insects and vertebrates as we did not have prior knowledge of whether
279 or what kind of ERVs are found in those species. We briefly summarise our approach below.

280 We first used open reading frames (ORFs) of vertebrate retroviruses as baits for the vertebrate genomes
281 and those of the Dipteran errantiviruses for the other genomes to run tBLASTn with default parameters.
282 Genomic regions that showed homology to the POL baits with a bit score of greater than 50 or to the GAG
283 and ENV baits with a bit score of greater than 30 were individually pooled. Genomic regions that had hits
284 from multiple baits were merged for their coordinates using bedtools/2.28.0. Regions that were larger than
285 1500 nucleotides for POL and 300 nucleotides for GAG or ENV were collected. We then looked for
286 regions that carry GAG, POL and ENV sequences that are next to each other in this order for insect
287 genomes, and POL and ENV for the other genomes. We allowed up to 500 nucleotides of gaps between
288 them for insect and vertebrate elements while we allowed up to 1000 nucleotides for the other genomes.
289 For GAG-POL-ENV elements, we excluded regions of smaller than 4000 nucleotides or larger than 10000
290 nucleotides. For POL-ENV elements, we excluded regions of smaller than 3000 nucleotides or larger than
291 10000 nucleotides.

292 We subsequently checked whether the regions contain uninterrupted ORFs by making translated sequences
293 using transeq from EMBOSS-6.6.0 and excluding ORFs with more than five stop codons. We clustered all
294 the identified GAG, POL and ENV from each genome by Clustal Omega
295 (https://www.ebi.ac.uk/Tools/msa/clustalo/) and identified genomic regions that harbour all three ORFs
296 that are duplicated in the genome. These processes eliminated genomic copies of ERVs that are being
297 degraded.

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

307 We took confident 1128 POL ORFs from the first tBLASTn search and made a multiple sequence
308 alignment of the conserved regions of the Integrase domain using mafft-7.505 with the option '—auto',
309 and inferred the phylogenetic relationships using iqtree-1.6.12 with the option '-m rtREV+R4 -bb 1000'
310 (see more details in the section below). We took 178 representative POL from the phylogenetic tree and
311 used the POL and the corresponding GAG and ENV ORFs from the same genomic regions for a second
312 tBLASTn search. We repeated the above steps after combining the results of the two tBLASTn searches.
313 We did not find any ERVs in Nematoda genomes that have GAG, POL and ENV in this order. A previous
314 study reported non-canonical Nemetoda ERVs that have flipped the order of ENV and POL [33]. Thus, we
315 searched for regions that have ENV and POL in this order allowing gaps of up to 2000 nucleotides
316 between them. GAG was subsequently predicted from 5000 nucleotides upstream to the predicted ENV-
317 POL. We extended the search for potential GAG-ENV-POL insertions to other Protostomia genomes
318 outside Arthropoda and found similar elements in Bryozoa and Platyhelminthes. Furthermore, we found
319 that the glycoprotein gene present in mosquito *BEL/pao* retrotransposons [34] are also found next to *gypsy*
320 POL genes. The glycoprotein gene resembles the *envelope* genes from Chuviridae and is structurally
321 similar to glycoprotein B from Herpesviruses. We identified ten mosquito *BEL/pao* retrotransposons that
322 carry this glycoprotein from RepBase (BEL-199_AA, BEL-279_AA, BEL-652_AA, BEL-801_AA, BEL-
323 804_AA, BEL-879_AA, BEL-51_AnFu, BEL-56_AnFu, BEL-81_AnFu, and BEL-71_CQ) and used them
324 as baits for a tBLASTn search against Arthropoda genomes. The tBLASTn search identified 114 *gypsy*
325 elements with homologous glycoprotein genes from Coleoptera, Hymenoptera, Polyneoptera, and
326 Hexaneuplia species.

327

328 We ran Clustal Omega to select ORFs that are found multiple times in the genome for insects and
329 vertebrates while we focused on finding intact ORFs for the other genomes (see below) regardless of the
330 copy number. We additionally ran tBLASTn using the fully annotated ORFs as baits to interrogate the
331 copy number of ERVs in the genome where they are found. The results are summarised in Supplementary
332 table x where we counted the number of tBLASTn hits with a more than 90% identity to the bait
333 sequences. A majority of newly identified ERVs (1516 out of 1610 elements) carry ORFs that are found at
334 least twice in the genome (Table S1), further validating our approach of finding active or recently active
335 insertions.

336 Following, we further curated newly identified ERVs. Firstly, we used hmmer 3.3.1 (http://hmmer.org/) to
337 make sure that the ERVs that we identified are classified as Ty3/gypsy. We compared POL ORFs from the
338 newly identified ERVs against the POL ORFs of known retroelements that are available from the Gypsy
339 Database 2.0 (https://gydb.org/index.php/Main_Page). The Gypsy Database 2.0 collects hmm profiles of
340 Ty1/Copia, Bel/Pao, Ty3/Gypsy, and Retrovirus POL proteins. The hmmer search classified all the newly
341 identified ERVs into Ty3/Gypsy or Retroviruses. Secondly, we ran HHpred from hh-suite 3.3.0 [35] or the
342 online tool of HHpred (https://toolkit.tuebingen.mpg.de/tools/hhpred) to predict GAG, ENV, and POL
343 ORFs using the default search parameters. For the command line HHpred, we first made multiple sequence
344 alignments using hhblits and the Uniclust database UniRef30/2022_02 (https://uniclust.mmseqs.com/).
345 This was followed by hhsearch to predict domain structures using the collection of pdb70 from March
346 2022 (https://wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite_dbs/). We used default settings
347 to run hhblits and hhsearch. Furthermore, we reannotated ORFs in ERVs except for those from insects
348 when they are only partially predicted homology to baits in the tBLASTn search. We extended the
349 predicted ORFs in the genomic region to upstream until a start codon is found, and to downstream until a
350 stop codon is found. We used benchling to facilitate the identification of continuous and intact ORFs.
351 ERVs that predicted all GAG, ENV and POL ORFs were taken for further analyses.

352 Phylogenetic analysis of POL and ENV domains

353 We obtained multiple sequence alignments of the RT and Integrase domains of POL as follows. For the
354 RT domain, we first took the HHpred alignments of POL against the RT domains of Moloney Murine
355 Leukaemia Virus (Protein Data bank entry 4MH8_A) and Human Immunodeficiency Virus type 1 (Protein
356 Data bank entry 4G1Q_B), spanning regions that correspond to their fingers, palm, thumb and connection
357 subdomains [36], namely the amino acid positions of 28 to 426 and 7 to 405, respectively. However, this
358 process caused small truncations at either ends of the domain because of poor alignments at the individual
359 substructures. To overcome this issue and to facilitate sequence alignments, we extended the N- and C-
360 termini by 10 and 15 aa, respectively, after merging the two predicted regions of RT domains. We then ran
361 mafft-7.505 to generate a multiple sequence alignment, from which we manually trimmed the either ends
362 to have intact fingers, palm, thumb and connection subdomains for all POL ORFs. Similarly, we took
363 HHpred alignments of POL against the Integrase domains of Human Spumaretrovirus (Protein Data bank
364 entry 3OYM_A) and Simian T-lymphotropic Virus type 1 (Protein Data bank entry 7OUH_B), spanning

365 regions that correspond to their N-terminal domain (NTD) and catalytic core domain (CCD) [37], namely the
366 amino acid positions of 54 to 273 and 8 to 197, respectively. We used the predicted Integrase domains with
367 15 and 30 aa extensions at the N- and C-termini, respectively, for generating a multiple sequence
368 alignment before manually trimming the ends. We inspected the multiple sequence alignments using
369 Mview (https://www.ebi.ac.uk/Tools/msa/mview/) and removed ERVs that contained truncated RT or
370 Integrase domains from the analyses. We included ORFs from 5 Drosophila errantiviruses (DM176,
371 IDEFIX, ZAM, GYPSY, and Gypsy-5_DGri) and 23 representative vertebrate retroviruses (BLV, FLV,
372 HTLV-2, KoRV, LDV, MMTV, MuLV, REV, RSV, SRV-1, BFV, EFV, FFV, SFV, HIV-1, EIAV, FIV,
373 SnRV, Loki-Ofa, Loki-Pma, Loki-Sri, Loki-Str, and Loki-Xtr [17]) in the multiple sequence alignments.

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

399 1610 ERVs from all phyla combined retained intact Integrase and RT domains in POL and one of the three
400 types of ERVs, and thus were used for further analyses. We also included the above mentioned 5
401 Drosophila errantiviruses and 23 vertebrate retroviruses for phylogenetic analyses. We ran iqtree-1.6.12

402 with the option '-m rtREV+R4 -bb 1000' to generate the phylogenetic trees of all Integrase and RT
403 domains, those from ERVs with each of the three types of ENVs, and the trees of each ENV type.

404 Clades are then defined mainly based on how closely they cluster together on the tree, as well as the
405 relative distance from neighbouring clades. In some cases, multiple clades are combined based on their
406 relative position to each other from the root. Clade definitions are summarised in Supplementary Table S3
407 and represented in Figure 1. We then utilize these clade definitions to analyse congruency between
408 Integrase and host phylogeny, Integrase and Env, and Integrase and RT, (Table S3-5). We also use these
409 clade definitions to select random representative ERVs for structural prediction of the RNase H domain
410 using Alphafold 2, discussed later (Table S6).

411 For our analysis on host phylogeny, we separately define clades that contain ERVs found in Arthropoda or
412 Vertebrata using a similar approach, but this time excluding clades that do not contain either. We labelled
413 these clades A1-A24 for Arthropoda, and V1-V2 for vertebrates (Table S4). We then constructed a balloon
414 plot of clade against host species using the ggballoon plot package in R. The size of each data point was
415 made proportional to species representation. For example, a balloon size of 0.5 for an order in one clade,
416 would mean that 50% of the elements found in that order are found in that given clade.

417 For our analysis on Integrase and RT, as well as Integrase and Env, we constructed balloon plots in a
418 similar manner, but the size of the balloon is simply proportional to the total number of elements
419 represented in the point, and not as a proportion.

420 <u>Identification of putative long terminal repeats and transfer RNA (tRNA) primer binding sequences</u>

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

438 GCF_016086655.3, GCF_024679095.1, GCF_019059575.1 for Arthropoda, GCF_000326865.2,
439 GCF_020536995.1, GCF_000002985.6, GCF_000715545.1 for Protostomia, GCF_000222465.1 and
440 GCF_022113875.1 for Cnidaria, GCF_013122585.1, GCF_000003605.2, GCF_000003815.2, and
441 GCF_000002235.5 for Deuterostomia, and GCF_015237465.2, GCF_000001405.40, GCF_017654675.1,
442 GCF_900496995.4, GCF_900700375.1, and GCF_010993605.1 for Vertebrata. The minimum
443 requirements of the priming events are incompletely understood, but the length of PBS for
444 retrotransposons is generally between 8 and 18 nucleotides and the PBS is positioned near the 3' end of the
445 5' LTR [38]. We chose the 12 nucleotides at the 3' end of tRNAs as putative primer sequences and searched
446 for PBS in the 100 nucleotide-window centred at the 3' end of the 5' LTR using bowtie-1.2.3 allowing up
447 to two mismatches. We added "CCA" to the 3' end of the tRNA sequence when it is not present in the
448 genomic sequences. We successfully identified putative PBSs in 1205 out of 1610 ERVs that we identified
449 in this study. We grouped ERVs based on the type of tRNA used for the PBS and summarised the results
450 in Supplementary figure S1 and Supplementary table S2.

451 Structure prediction of the RNase H domains using Alphafold 2

452 Using clade definitions from the phylogenetic trees of Integrase and RT, we selected representative ERVs
453 from each clade for the structural prediction of the RNase H domain. We used the colabfold wrap v1.4.0 of
454 Alphafold 2 (https://github.com/YoshitakaMo/localcolabfold/releases/tag/v1.4.0), and used MMSeqs2 for
455 the database search and clustering of protein sequences before performing the structural prediction [39]. We
456 then visualised the RNase H domains using Maestro from Schrödinger
457 (https://newsite.schrodinger.com/platform/products/maestro/), and characterised visible structures by the
458 orientation and order of β sheets present, grouping them by letter (Table S6). Once the first set of proteins
459 were analysed, elements with unique structures were further analysed by characterising RNase H domains
460 from the remaining ERVs in that clade.

461

462

463

464 **Figure Legends**

465 Figure 1 – Phylogenetic relationships between *gypsy* errantiviruses and vertebrate endogenous retroviruses
466 (vERVs). (A) Maximum likelihood tree constructed based on *pol* Integrase proteins of *gyspy* errantiviruses
467 and vERVs, with host phyla annotated in colour. (B) Distribution of vERVs/*gypsy* errantiviruses in
468 metazoa with *env* type indicated.

469 Figure 2 – Representation of *env* type in phylogenetic tree of Integrase in *gypsy* errantiviruses and vERVs.
470 (A) Maximum likelihood tree constructed based on *pol* Integrase proteins of *gyspy* errantiviruses and
471 vERVs, with Env type annotated in colour. (B-D) Close up view of Integrase clades N, O and P
472 respectively. Host species for sub-branches within clade annotated, with *env* type annotated for mixed sub-
473 branches.

474 Figure 3 – Domain architecture and 3D structure of Ribonuclease H (RNase H) in *gypsy* errantiviruses and
475 vERVs. (A) Maximum likelihood tree constructed based on *pol* Integrase proteins of *gyspy* errantiviruses
476 and vERVs with RNase H domain architecture annotated in colour. V – Vertebrates, I – Insects, AA –
477 Ancient Animals. Integrase clades N, O and P are also shown in grey. (B) Summary of bridge region
478 RNase H domains found in representative elements in Insects, Ancient Animals and Vertebrates. (C)
479 Alphafold-predicted 3D structure of canonical RNase H/Tether, 'mini' RNase H and 'pseudo' RNase H
480 domains, with topology maps of α helices (grey) and β sheets (red). (D) Distribution of shared RNase H
481 domain architectures in the bridge region of elements within insects, vertebrates, and ancient animals.
482 Structures for Alphafold taken from Annelida_Polychaeta_Apac_errantivirus_7,
483 Bryozoa_Gymnolaemata_Cpal_errantivirus_3 and Coleoptera_Coccinellidae_Npum_errantivirus_1 for
484 RNase H/Tether, 'mini' RNase H and 'pseudo' RNase H domains respectively.

485

486 Supplementary Figure S1 – Bootstrap values and tRNA primer binding sites of *gypsy* errantiviruses and
487 vERVs. (A) Maximum likelihood tree constructed using *pol* Integrase proteins. Red triangles represent
488 branches/nodes with bootstrap values >95. The tree and the clade definition are the same as in Fig 1A. (B)
489 Type of tRNAs carried by *gypsy* errantiviruses and vERVs in discrete clades from the tree.

490 Supplementary Figure S2 – Phylogenetic congruence between Integrase and RT proteins of vERVs and
491 *gypsy* errantiviruses. (A) Maximum likelihood tree constructed based on Pol RT proteins of *gypsy*
492 errantiviruses and vERVs. Bootstrap values >95 annotated on nodes with red triangles. (B) Balloon plot of
493 number of elements shared between clades in Integrase and RT. Size of balloon is proportional to number
494 of elements shared between discrete clades of Integrase and RT.

495 Supplementary Figure S3 – Phylogenetic relationships between *gyspy* errantiviruses/vERVs and their
496 hosts. (A) Arthropoda (A1-A24) and Vertebrate (V1-V12) specific clades on maximum likelihood tree
497 constructed based on Integrase proteins. The same tree was used in Figure 1A. (B) Balloon plot
498 representing Arthropoda/Vertebrate clade to order congruence. Size of balloon represented in proportion of

499 total elements from an order, where 1 represents a balloon that contains all the elements from that order.
500 (C-D) Close ups of clades O and P from Figure 1A. (E) Expanded and detailed colour code for host
501 phylum annotation in maximum likelihood trees.

502 Supplementary Figure S4 – Phylogenetic congruence between *pol* Integrase and Env for *gypsy*
503 errantiviruses and vERVs. Trees for Integrase and Env with clade definitions matching the annotation in
504 the balloon plot of number of elements shared between clades in Integrase and Env for elements containing
505 type F (A), GP2 (B) and HSV (C). Size of balloon is proportional to number of elements shared between
506 discrete clades of Integrase and Env. (A') Maximum likelihood tree constructed on *int* proteins of elements
507 carrying EnvF. (A") Maximum likelihood tree constructed on Env type F. (A"') Balloon plot of Integrase
508 vs Env for F type elements. (B') Maximum likelihood tree constructed on Integrase proteins of elements
509 carrying Env HSV. (B") Maximum likelihood tree constructed on Env type HSV. (B") Balloon plot of
510 Integrase vs Env for HSV type elements. (C') Maximum likelihood tree constructed on *int* proteins of
511 elements carrying Env GP2. (C") Maximum likelihood tree constructed on Env type GP2. (C"') Balloon
512 plot of Integrase vs Env for GP2 type elements.

513 Supplementary Figure S5 – 3D structure of uncommon and yeast RNase H-like domains in *gypsy*
514 errantiviruses. (A) Alphafold-predicted 3D structure of uncommonly found RNase H-like domains with
515 topology maps of α helices (grey) and β sheets (red). (B) 3D structures and topology maps of 'mini' RNase
516 H domains in yeast Ty3/*gypsy* elements. Structures for Alphafold taken from
517 Annelida_Polychaeta_Slim_errantivirus_9, Diptera_Cecidomyiidae_Smos_errantivirus_1 and
518 Hymenoptera_Apoidea_Obic_errantivirus_2 for "Mini" domain 2, "Pseudo" domain 2 and "Pseudo"
519 domain 3 respectively.

520

521 Supplementary table S1 – Master *gypsy* errantivirus and vERV table. Information on name, host
522 classification, genome assembly, genomic coordinates, *env* type, type of tRNA used as primer, LTR
523 coordinates and the number of mismatches between annotated tRNA sequences and the tRNA primer
524 binding sequence in tab 1. Copy number of *gypsy* errantivirus and vERV open reading frames in tab 2,
525 tRNA classification guide in tab 3.

526 Supplementary table S2 – Raw data of Integrase clade definition against tRNA primer classification for
527 supplementary figure S1B.

528 Supplementary table S3 – Clade definitions of Integrase and RT maximum likelihood trees. Clade
529 definitions for Integrase in tab 1. Clade definitions for RT in tab 2. Raw balloon plot data for
530 supplementary figure S2B of Integrase vs RT in tab 3.

531 Supplementary table S4 – Raw data for supplementary figure S3. Arthropoda clade definitions for
532 Integrase in tab 1. Raw balloon plot table of host order/class against Arthropoda clade in tab 2, with total
533 number of elements from order/class in tree given. Vertebrate clade definitions for Integrase in tab 3. Raw

534 balloon plot table of host order/class against Vertebrate clade in tab 4, with total number of elements from
535 order/class in tree given.

536 Supplementary table S5 – Raw data for supplementary figure 4. Clade definitions for Integrase carrying F
537 type Env in tab 1. Clade definitions for Env F in tab 2. Raw balloon plot of Integrase vs Env for F carrying
538 ERVs in tab 3. Clade definitions for Integrase carrying HSV type Env in tab 4. Clade definitions for Env
539 HSV in tab 5. Raw balloon plot of Integrase vs Env for HSV carrying ERVs in tab 6. Clade definitions for
540 Integrase carrying GP2 type Env in tab 7. Clade definitions for Env GP2 in tab 8. Raw balloon plot of
541 Integrase vs Env for GP2 carrying ERVs in tab 9.

542 Supplementary table S6 – RNase H classification based on arrangement of β sheets for representative
543 elements.

544 Supplementary table S7 – Total list and details of analysed genomes. Diptera genomes in tab 1.
545 Lepidoptera genomes in tab 2. Coleoptera genomes in tab 3. Hymenoptera genomes in tab 4. Hemiptera
546 genomes in tab 5. Polyneoptera genomes in tab 6. Palaeoptera genomes in tab 7. Non-insect Arthropoda
547 genomes in tab 8. Protostomia genomes excluding Arthropoda and Nematoda in tab 9. Nematoda genomes
548 in tab 10. Cnidaria genomes in tab 11. Ctenophora genomes in tab 12. Porifera genomes in tab 13.
549 Deuterostomia genomes excluding vertebrates in tab 14. Vertebrate genomes in tab 15. Numbers of
550 predicted ORFs are shown. ORFs that are found downstream of predicted POL were also counted for
551 Protostomia, Cnidaria, Ctenophora, Porifera and Deuterostomia. Genomes that we used to annotate vERVs
552 and errantiviruses are indicated.

553

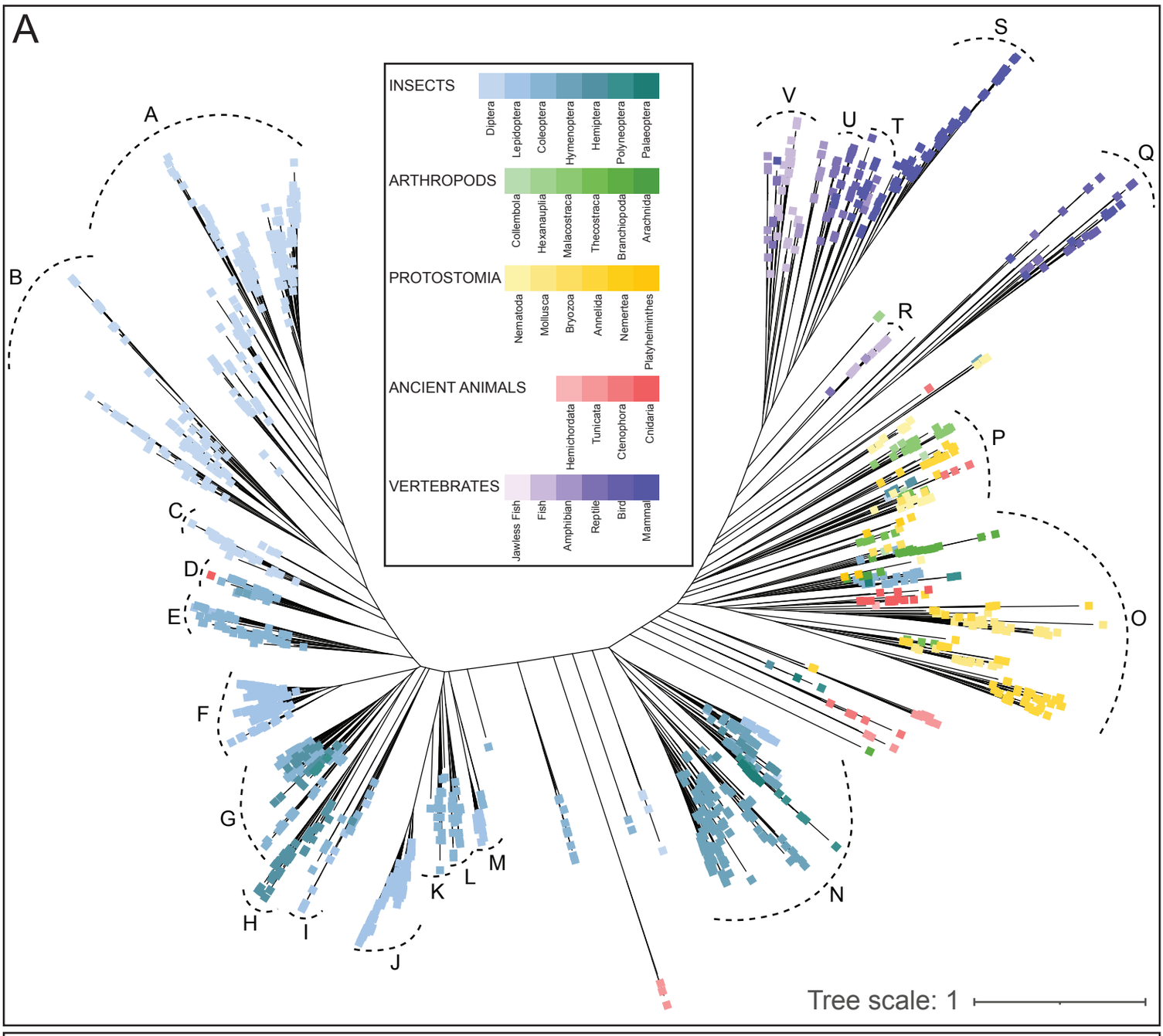
554 References

555 1 Johnson, W. E. Endogenous Retroviruses in the Genomics Era. *Annu Rev Virol* **2**, 135-159
556 (2015). https://doi.org:10.1146/annurev-virology-100114-054945
557 2 Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on
558 host biology. *Nat Rev Genet* **13**, 283-296 (2012). https://doi.org:10.1038/nrg3199
559 3 Zhuo, X., Rho, M. & Feschotte, C. Genome-wide characterization of endogenous retroviruses
560 in the bat Myotis lucifugus reveals recent and diverse infections. *J Virol* **87**, 8493-8501
561 (2013). https://doi.org:10.1128/JVI.00892-13
562 4 Metegnier, G. *et al.* Comparative paleovirological analysis of crustaceans identifies multiple
563 widespread viral groups. *Mob DNA* **6**, 16 (2015). https://doi.org:10.1186/s13100-015-0047-3
564 5 Dewannieux, M. & Heidmann, T. Endogenous retroviruses: acquisition, amplification and
565 taming of genome invaders. *Curr Opin Virol* **3**, 646-656 (2013).
566 https://doi.org:10.1016/j.coviro.2013.08.005
567 6 Coffin, J. M., Hughes, S. H. & Varmus, H. E. (1997).
568 7 Pelisson, A., Mejlumian, L., Robert, V., Terzian, C. & Bucheton, A. Drosophila germline
569 invasion by the endogenous retrovirus gypsy: involvement of the viral env gene. *Insect*
570 *Biochem Mol Biol* **32**, 1249-1256 (2002). https://doi.org:10.1016/s0965-1748(02)00088-7
571 8 Keegan, R. M., Talbot, L. R., Chang, Y. H., Metzger, M. J. & Dubnau, J. Intercellular viral
572 spread and intracellular transposition of Drosophila gypsy. *PLoS Genet* **17**, e1009535 (2021).
573 https://doi.org:10.1371/journal.pgen.1009535
574 9 Song, S. U., Gerasimova, T., Kurkulos, M., Boeke, J. D. & Corces, V. G. An env-like protein
575 encoded by a Drosophila retroelement: evidence that gypsy is an infectious retrovirus.
576 *Genes Dev* **8**, 2046-2057 (1994). https://doi.org:10.1101/gad.8.17.2046
577 10 Brasset, E. *et al.* Viral particles of the endogenous retrovirus ZAM from Drosophila
578 melanogaster use a pre-existing endosome/exosome pathway for transfer to the oocyte.
579 *Retrovirology* **3**, 25 (2006). https://doi.org:10.1186/1742-4690-3-25
580 11 Malik, H. S., Henikoff, S. & Eickbush, T. H. Poised for contagion: evolutionary origins of the
581 infectious abilities of invertebrate retroviruses. *Genome Res* **10**, 1307-1318 (2000).
582 https://doi.org:10.1101/gr.145000
583 12 Rohrmann, G. F. & Karplus, P. A. Relatedness of baculovirus and gypsy retrotransposon
584 envelope proteins. *BMC Evol Biol* **1**, 1 (2001). https://doi.org:10.1186/1471-2148-1-1
585 13 Hayward, A., Cornwallis, C. K. & Jern, P. Pan-vertebrate comparative genomics unmasks
586 retrovirus macroevolution. *Proc Natl Acad Sci U S A* **112**, 464-469 (2015).
587 https://doi.org:10.1073/pnas.1414980112
588 14 Wang, J. & Han, G. Z. Genome mining shows that retroviruses are pervasively invading
589 vertebrate genomes. *Nat Commun* **14**, 4968 (2023). https://doi.org:10.1038/s41467-023-
590 40732-w
591 15 Stefanov, Y., Salenko, V. & Glukhov, I. Drosophila errantiviruses. *Mob Genet Elements* **2**, 36-
592 45 (2012). https://doi.org:10.4161/mge.19234
593 16 Terzian, C., Pélisson, A. & Bucheton, A. Evolution and phylogeny of insect endogenous
594 retroviruses. *BMC Evol Biol* **1**, 3 (2001). https://doi.org:10.1186/1471-2148-1-3
595 17 Wang, J. & Han, G. Z. A Sister Lineage of Sampled Retroviruses Corroborates the Complex
596 Evolution of Retroviruses. *Mol Biol Evol* **38**, 1031-1039 (2021).
597 https://doi.org:10.1093/molbev/msaa272
598 18 Gifford, R., Kabat, P., Martin, J., Lynch, C. & Tristem, M. Evolution and distribution of class II-
599 related endogenous retroviruses. *J Virol* **79**, 6478-6486 (2005).
600 https://doi.org:10.1128/JVI.79.10.6478-6486.2005
601 19 Bargues, N. & Lerat, E. Evolutionary history of LTR-retrotransposons among 20 Drosophila
602 species. *Mob DNA* **8**, 7 (2017). https://doi.org:10.1186/s13100-017-0090-3

603 20 Malik, H. S. & Eickbush, T. H. Phylogenetic analysis of ribonuclease H domains suggests a
604 late, chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Res* **11**,
605 1187-1197 (2001). https://doi.org:10.1101/gr.185101
606 21 Loreto, E. L., Carareto, C. M. & Capy, P. Revisiting horizontal transfer of transposable
607 elements in Drosophila. *Heredity (Edinb)* **100**, 545-554 (2008).
608 https://doi.org:10.1038/sj.hdy.6801094
609 22 Metzger, M. J., Paynter, A. N., Siddall, M. E. & Goff, S. P. Horizontal transfer of
610 retrotransposons between bivalves and other aquatic species of multiple phyla. *Proc Natl*
611 *Acad Sci U S A* **115**, E4227-E4235 (2018). https://doi.org:10.1073/pnas.1717227115
612 23 Grandi, N. & Tramontano, E. Human Endogenous Retroviruses Are Ancient Acquired
613 Elements Still Shaping Innate Immune Responses. *Front Immunol* **9**, 2039 (2018).
614 https://doi.org:10.3389/fimmu.2018.02039
615 24 Johnson, W. E. Origins and evolutionary consequences of ancient endogenous retroviruses.
616 *Nat Rev Microbiol* **17**, 355-370 (2019). https://doi.org:10.1038/s41579-019-0189-2
617 25 Louz, D., Bergmans, H. E., Loos, B. P. & Hoeben, R. C. Cross-species transfer of viruses:
618 implications for the use of viral vectors in biomedical research, gene therapy and as live-
619 virus vaccines. *J Gene Med* **7**, 1263-1274 (2005). https://doi.org:10.1002/jgm.794
620 26 Subbarao, K. The Critical Interspecies Transmission Barrier at the Animal¯Human Interface.
621 *Trop Med Infect Dis* **4** (2019). https://doi.org:10.3390/tropicalmed4020072
622 27 Zheng, H. *et al.* Koala retrovirus diversity, transmissibility, and disease associations.
623 *Retrovirology* **17**, 34 (2020). https://doi.org:10.1186/s12977-020-00541-1
624 28 Mottaghinia, S. *et al.* A recent gibbon ape leukemia virus germline integration in a rodent
625 from New Guinea. *Proc Natl Acad Sci U S A* **121**, e2220392121 (2024).
626 https://doi.org:10.1073/pnas.2220392121
627 29 Tarlinton, R. E., Meers, J. & Young, P. R. Retroviral invasion of the koala genome. *Nature* **442**,
628 79-81 (2006). https://doi.org:10.1038/nature04841
629 30 Zheng, J., Wei, Y. & Han, G. Z. The diversity and evolution of retroviruses: Perspectives from
630 viral "fossils". *Virol Sin* **37**, 11-18 (2022). https://doi.org:10.1016/j.virs.2022.01.019
631 31 Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon
632 activity in Drosophila. *Cell* **128**, 1089-1103 (2007). https://doi.org:10.1016/j.cell.2007.01.043
633 32 Ribet, D. *et al.* An infectious progenitor for the murine IAP retrotransposon: emergence of
634 an intracellular genetic parasite from an ancient retrovirus. *Genome Res* **18**, 597-609 (2008).
635 https://doi.org:10.1101/gr.073486.107
636 33 Sacco, M. A., Lau, J., Godinez-Vidal, D. & Kaloshian, I. Non-canonical nematode endogenous
637 retroviruses resulting from RNA virus glycoprotein gene capture by a metavirus. *J Gen Virol*
638 **103** (2022). https://doi.org:10.1099/jgv.0.001739
639 34 Dezordi, F. Z., Vasconcelos, C. R. D. S., Rezende, A. M. & Wallau, G. L. In and Outs of
640 Chuviridae Endogenous Viral Elements: Origin of a Potentially New Retrovirus and Signature
641 of Ancient and Ongoing Arms Race in Mosquito Genomes. *Front Genet* **11**, 542437 (2020).
642 https://doi.org:10.3389/fgene.2020.542437
643 35 Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein
644 annotation. *BMC Bioinformatics* **20**, 473 (2019). https://doi.org:10.1186/s12859-019-3019-7
645 36 Das, D. & Georgiadis, M. M. The crystal structure of the monomeric reverse transcriptase
646 from Moloney murine leukemia virus. *Structure* **12**, 819-829 (2004).
647 https://doi.org:10.1016/j.str.2004.02.032
648 37 Engelman, A. & Cherepanov, P. Retroviral Integrase Structure and DNA Recombination
649 Mechanism. *Microbiol Spectr* **2**, 1-22 (2014). https://doi.org:10.1128/microbiolspec.MDNA3-
650 0024-2014.
651 38 Wilhelm, M. & Wilhelm, F. X. Reverse transcription of retroviruses and LTR retrotransposons.
652 *Cell Mol Life Sci* **58**, 1246-1262 (2001). https://doi.org:10.1007/PL00000937

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

Figure 1



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, HSV
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, HSV
Collembola (Arthropoda, *)	Yes	F
Hexanauplia (Arthropoda, *)	Yes	F, HSV
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, HSV
Hymenoptera (Insecta, ***)	Yes	F, HSV
Hemiptera (Insecta, ***)	Yes	F, HSV
Polyneoptera (Insecta, ***)	Yes	F, HSV
Palaeoptera (Insecta, ***)	Yes	F

Figure 2

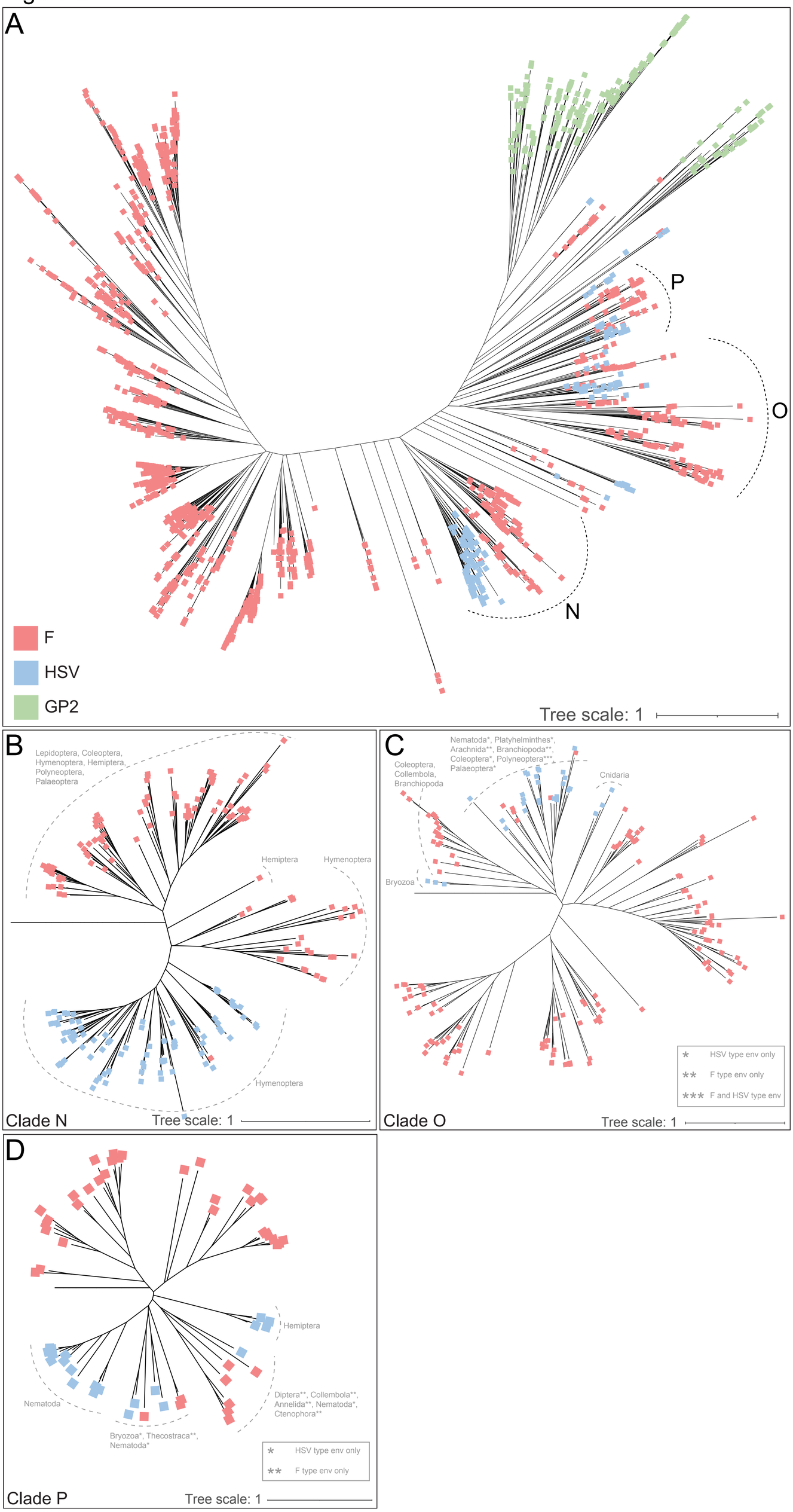
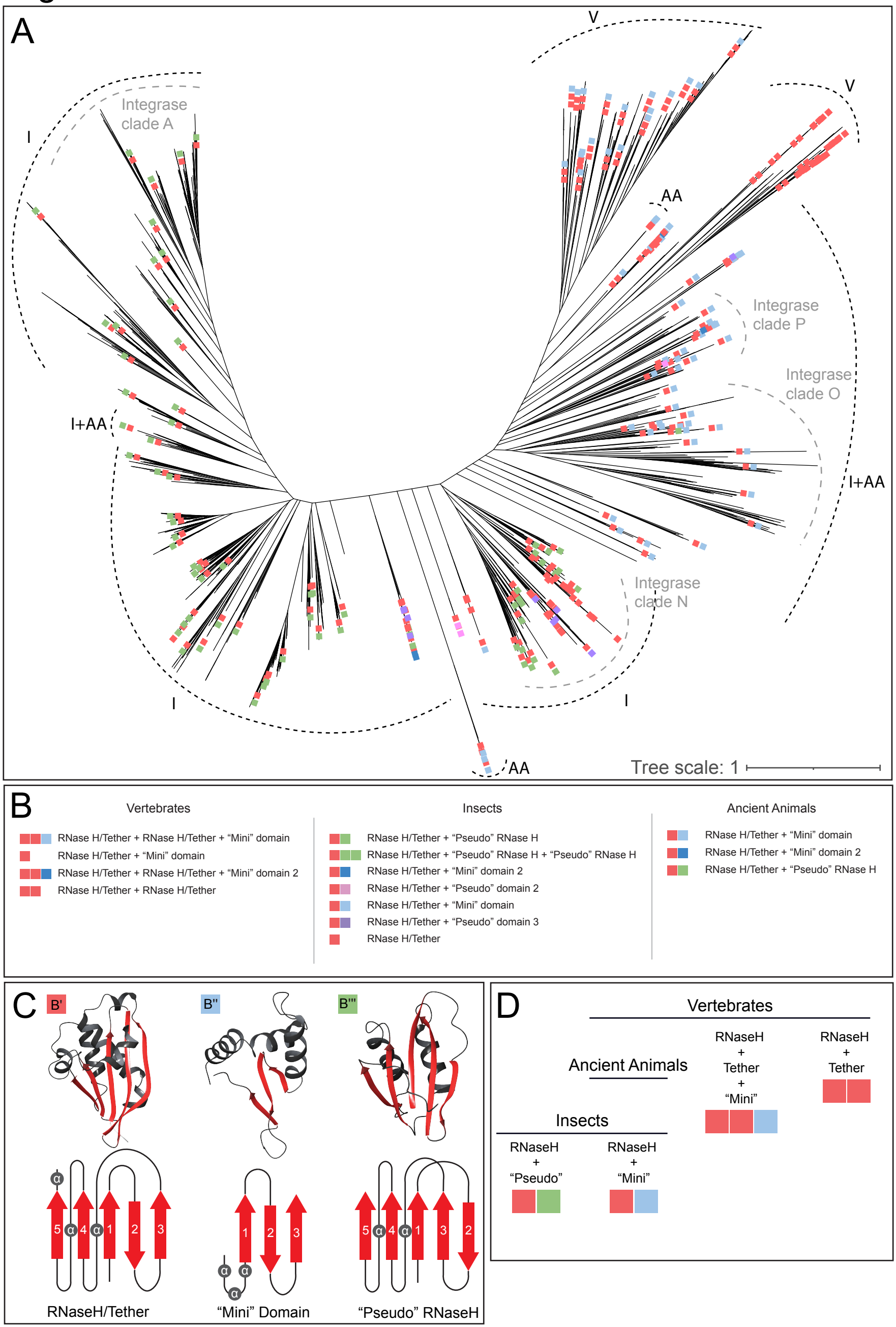
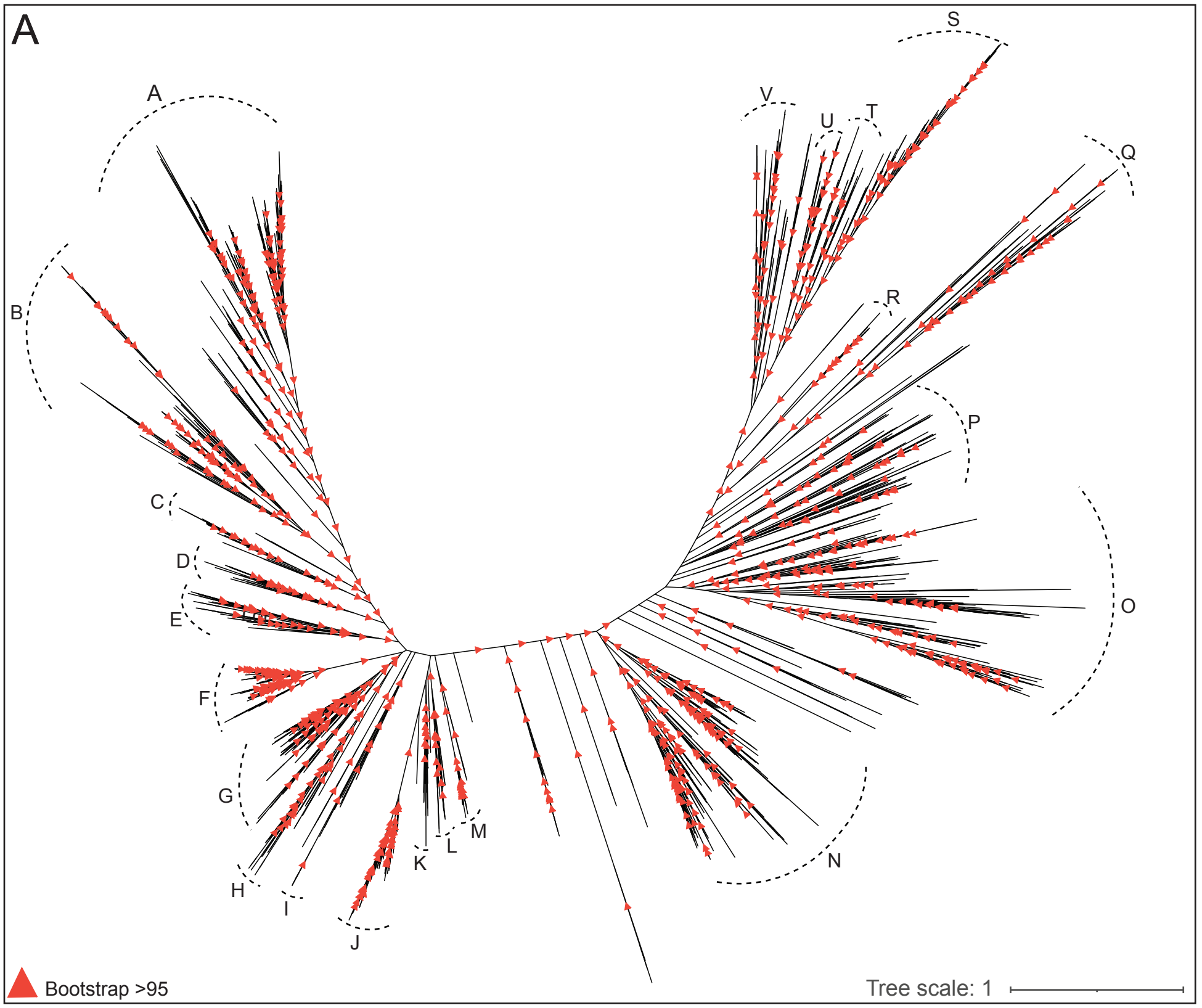
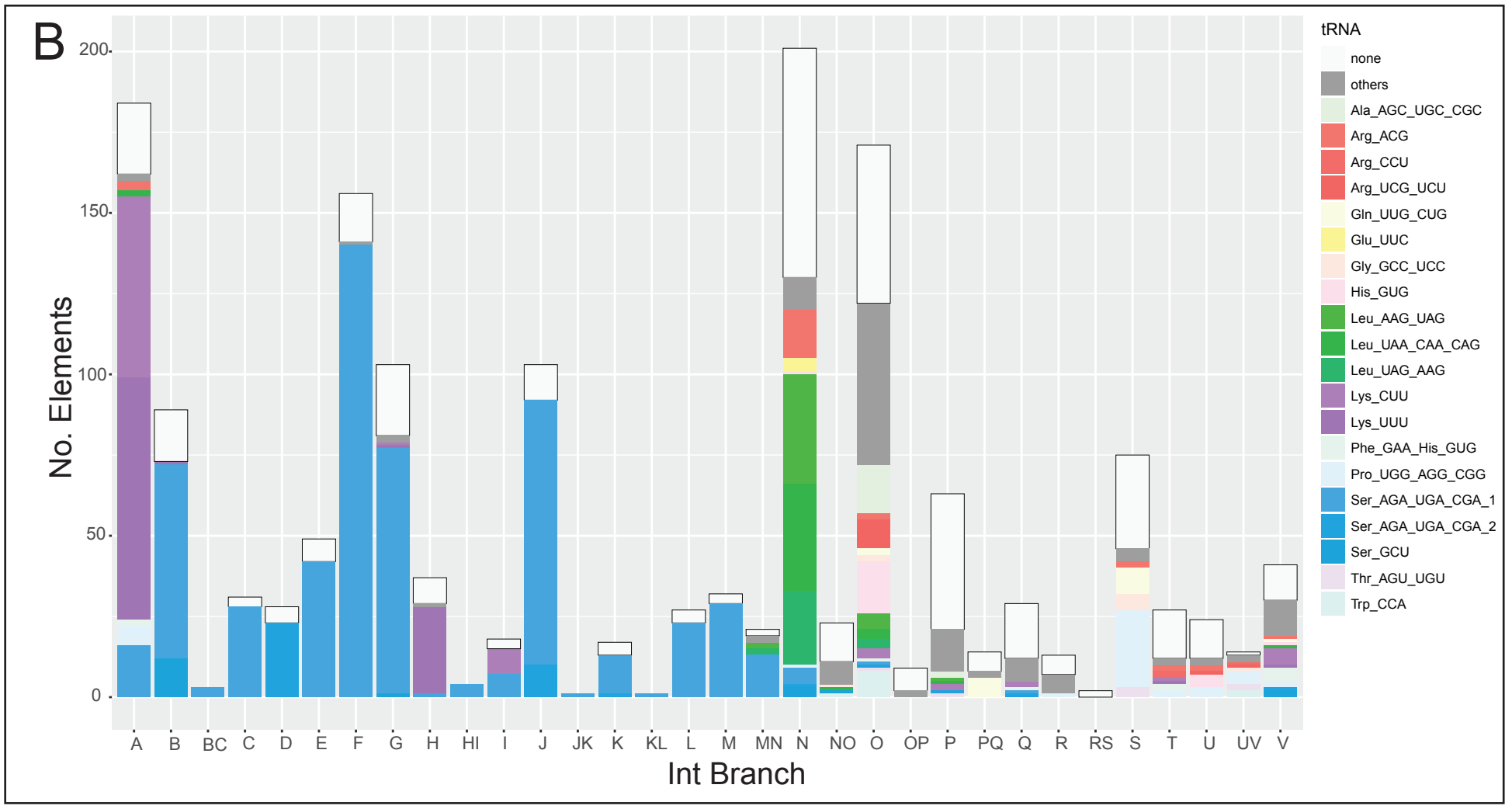

A
P
O
N
F
HSV
GP2
Tree scale: 1
B
Lepidoptera, Coleoptera, Hymenoptera, Hemiptera, Polyneoptera, Palaeoptera
Hemiptera
Hymenoptera
Hymenoptera
Clade N
Tree scale: 1
C
Nematoda*, Platyhelminthes*, Arachnida**, Branchiopoda**, Coleoptera*, Polyneoptera***, Palaeoptera*
Coleoptera, Collembola, Branchiopoda
Cnidaria
Bryozoa
* HSV type env only
** F type env only
*** F and HSV type env
Clade O
Tree scale: 1
D
Hemiptera
Nematoda
Diptera**, Collembola**, Annelida**, Nematoda*, Ctenophora**
Bryozoa*, Thecostraca**, Nematoda*
* HSV type env only
** F type env only
Clade P
Tree scale: 1

Figure 3

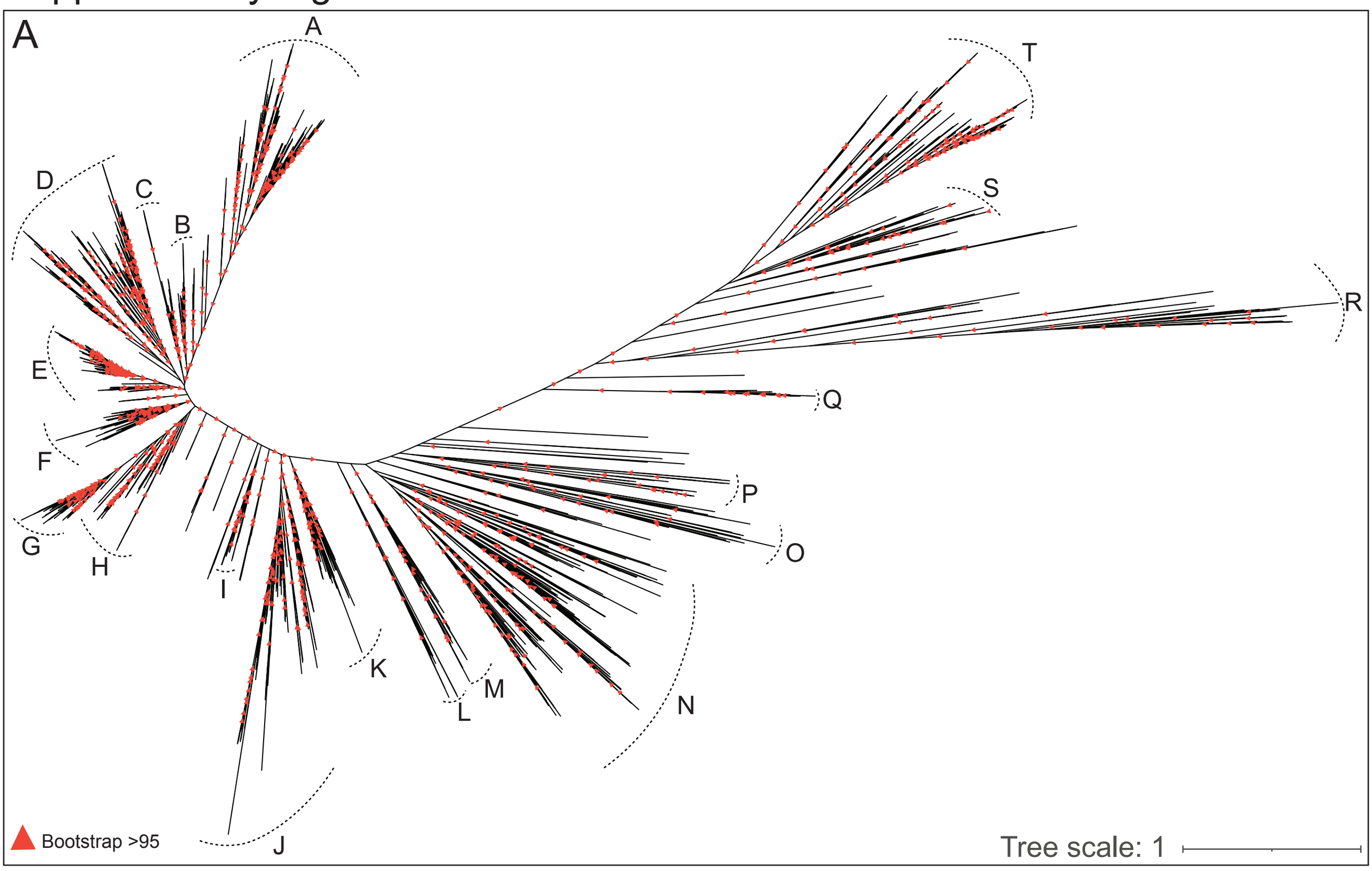

A
V
V
Integrase clade A
I
AA
Integrase clade P
Integrase clade O
I+AA
I+AA
Integrase clade N
I
I
AA
Tree scale: 1
B
Vertebrates
RNase H/Tether + RNase H/Tether + "Mini" domain
RNase H/Tether + "Mini" domain
RNase H/Tether + RNase H/Tether + "Mini" domain 2
RNase H/Tether + RNase H/Tether
Insects
RNase H/Tether + "Pseudo" RNase H
RNase H/Tether + "Pseudo" RNase H + "Pseudo" RNase H
RNase H/Tether + "Mini" domain 2
RNase H/Tether + "Pseudo" domain 2
RNase H/Tether + "Mini" domain
RNase H/Tether + "Pseudo" domain 3
RNase H/Tether
Ancient Animals
RNase H/Tether + "Mini" domain
RNase H/Tether + "Mini" domain 2
RNase H/Tether + "Pseudo" RNase H
C
B'
B''
B'''
RNaseH/Tether
"Mini" Domain
"Pseudo" RNaseH
D
Vertebrates
Ancient Animals
RNaseH + Tether + "Mini"
RNaseH + Tether
Insects
RNaseH + "Pseudo"
RNaseH + "Mini"

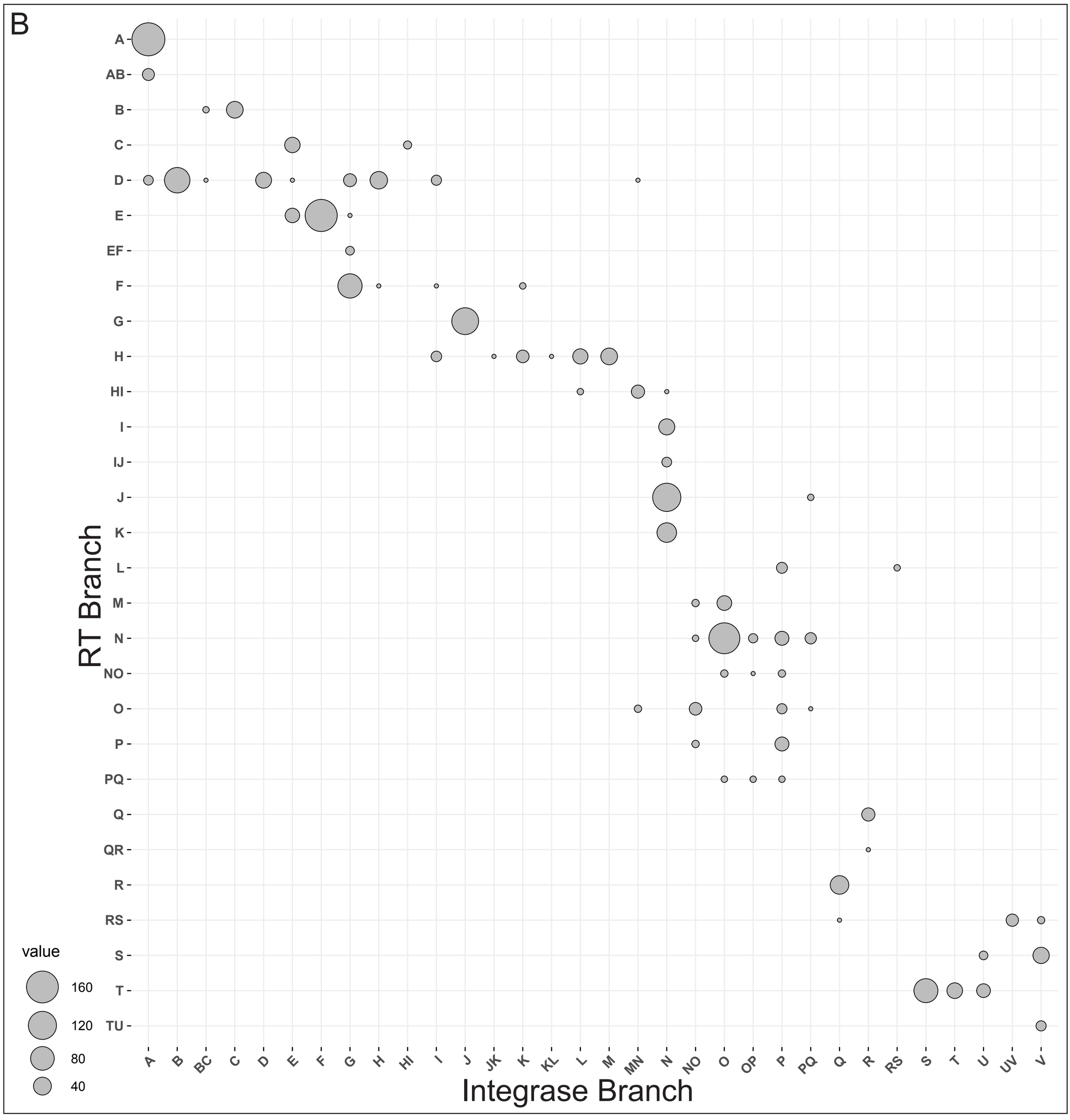
Supplementary Figure S1



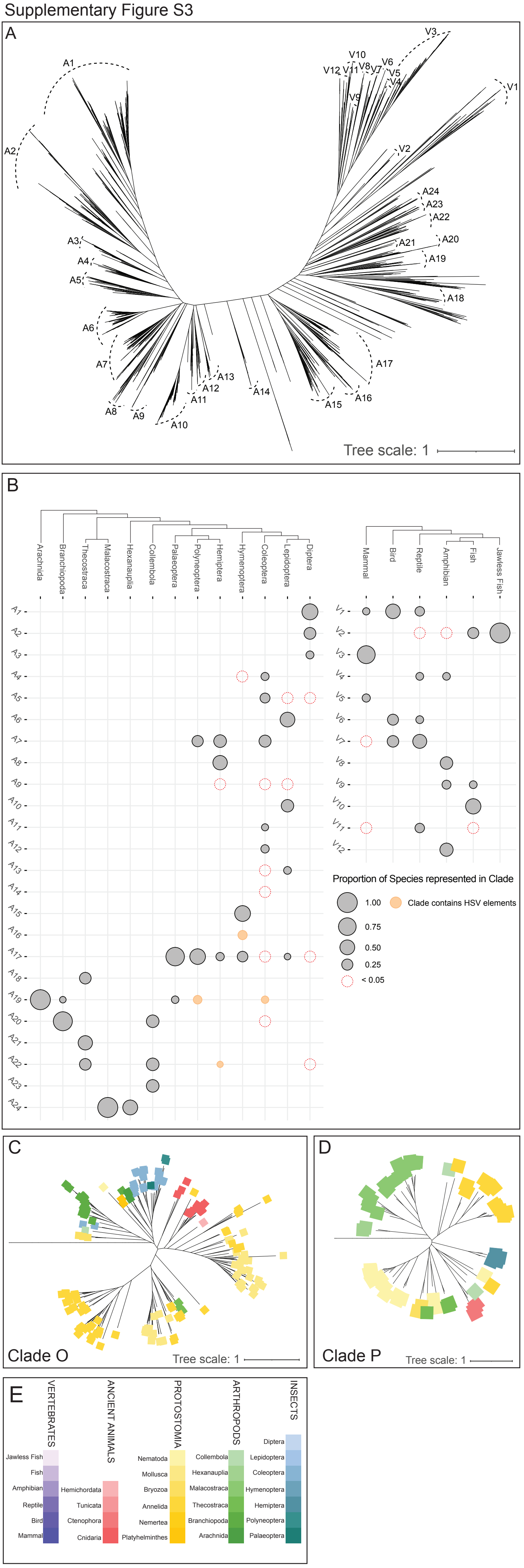


Supplementary Figure S2

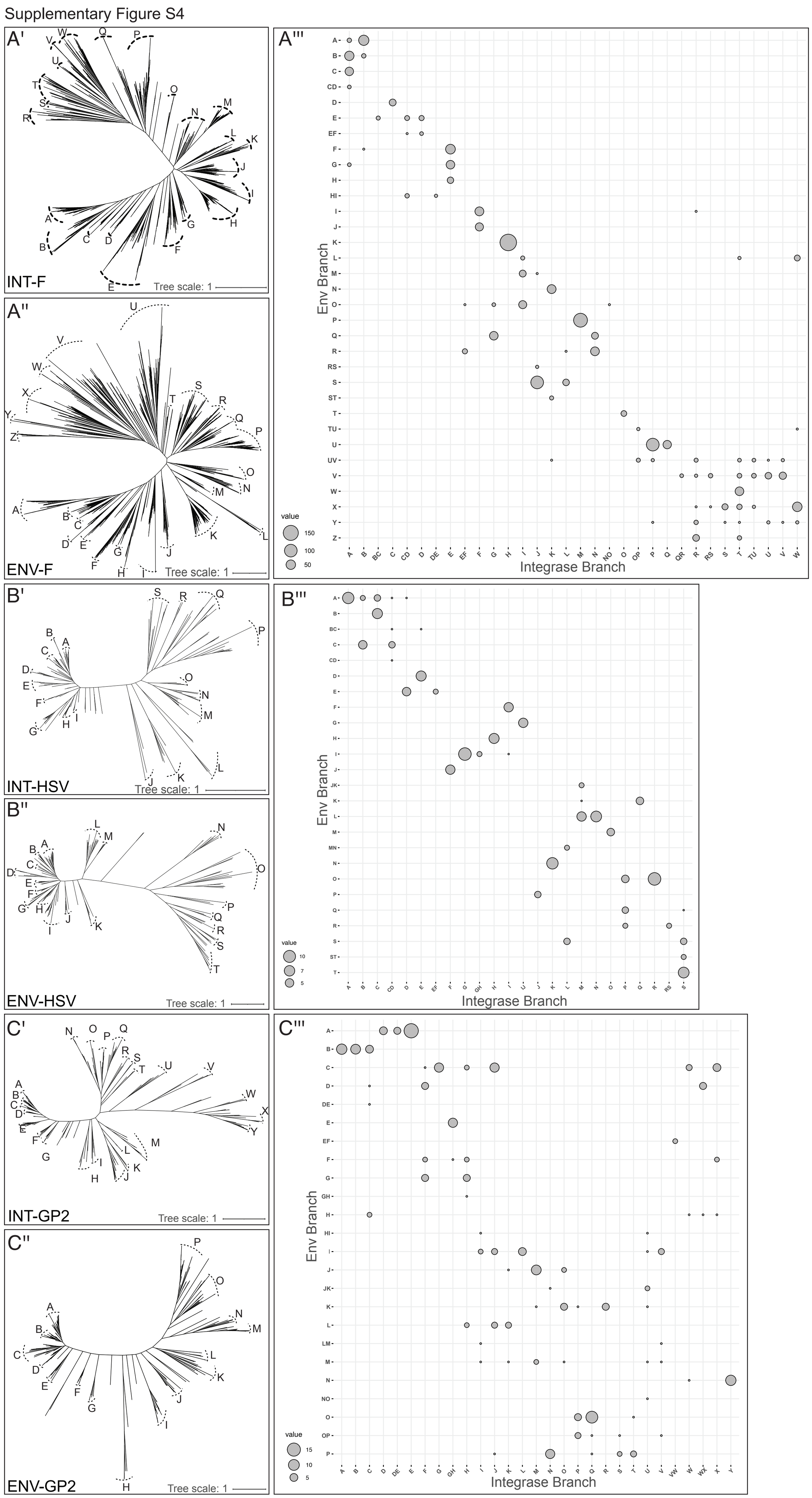
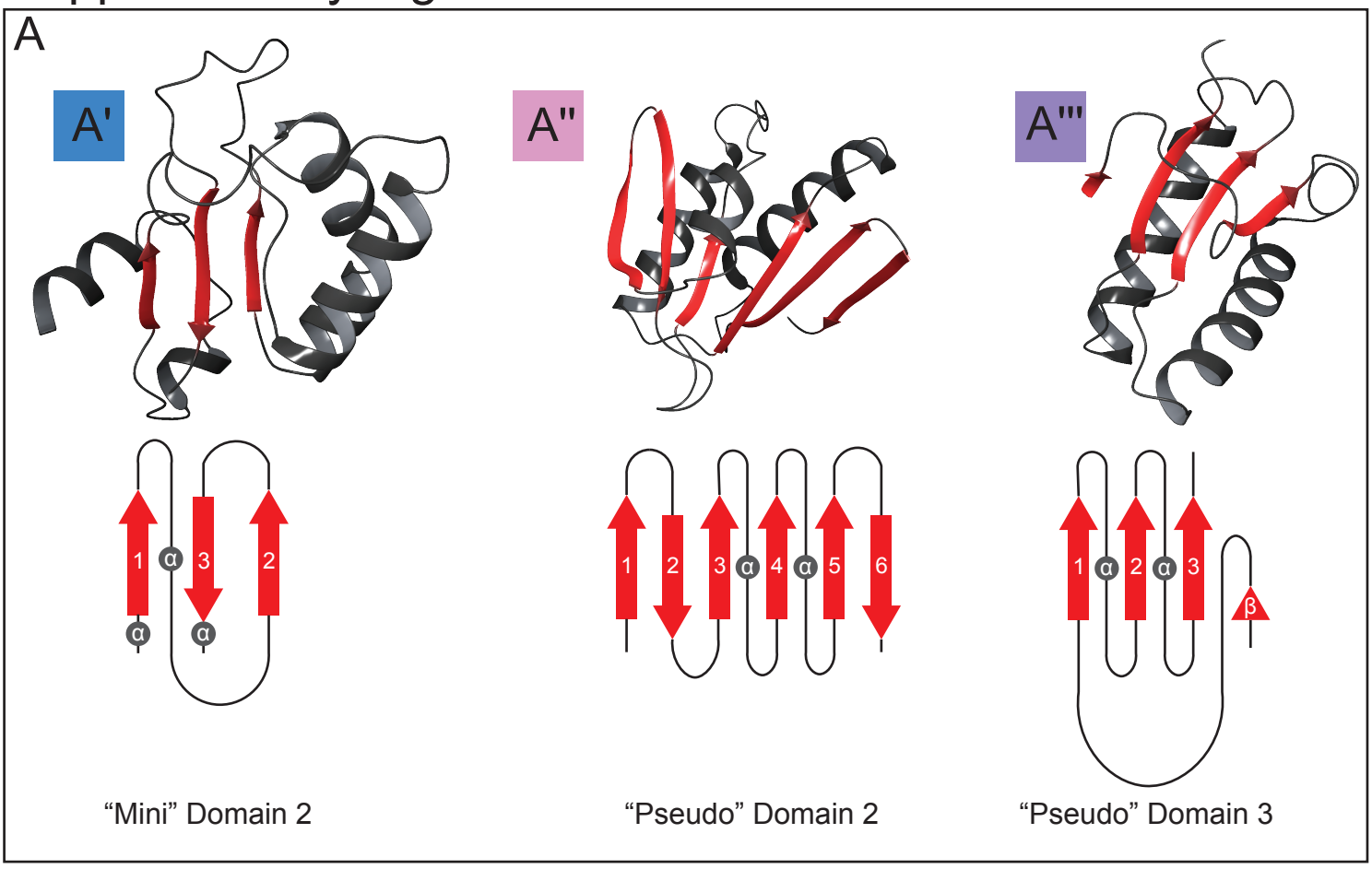
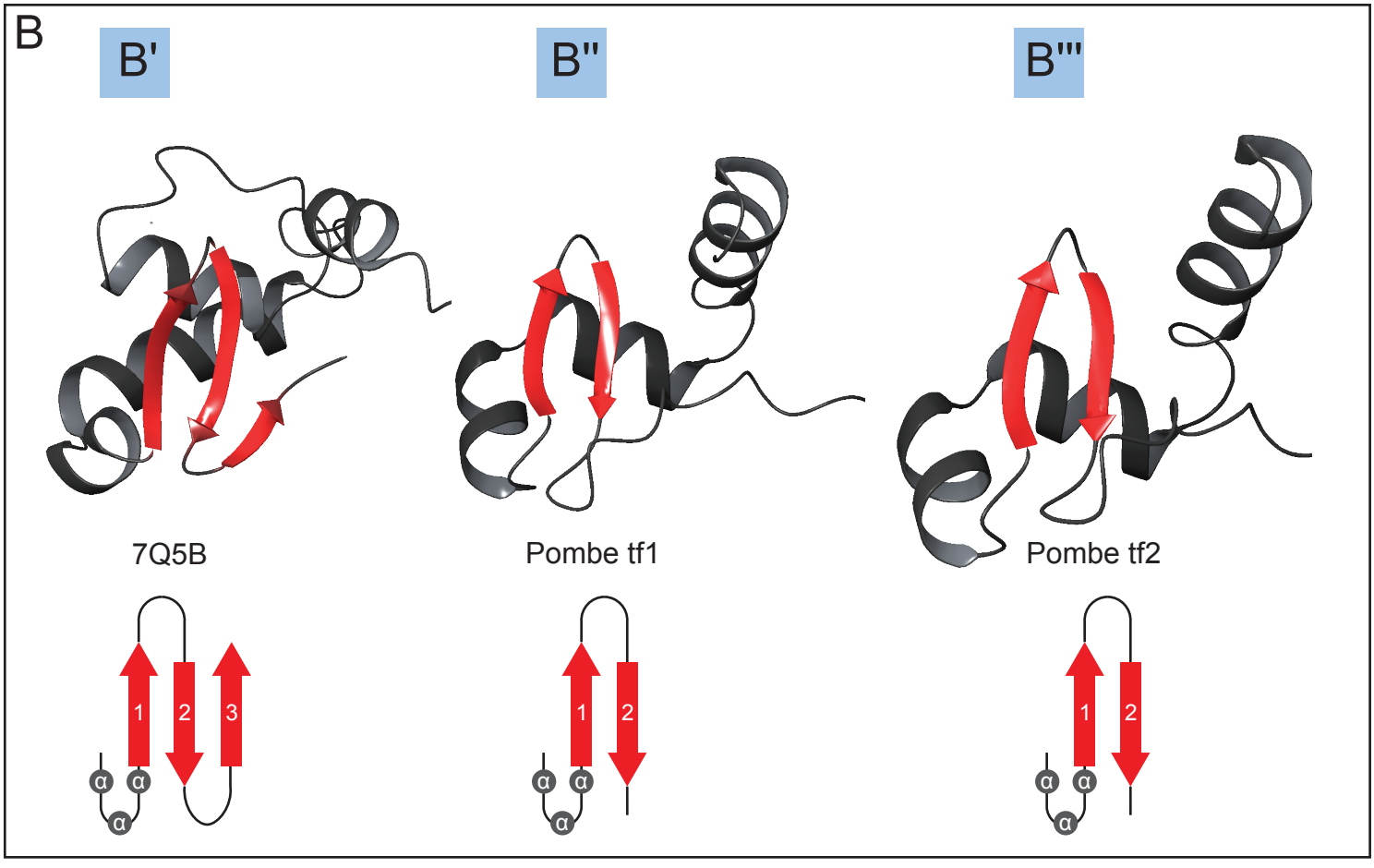




Supplementary Figure S3



Supplementary Figure S4



Supplementary Figure S5