

The RAG transposon is active through the deuterostome evolution and domesticated in jawed vertebrates.

Jose Ricardo Morales Poole¹, Sheng Feng Huang², Anlong Xu^{2,3}, Pierre Pontarotti¹

Jose Ricardo Morales Poole and Sheng Feng Huang contributed equally to the work.

¹ Aix Marseille Université, CNRS, Centrale Marseille, I2M UMR 7373, équipe évolution biologique modélisation, 13453, Marseille, France.

² State Key Laboratory of Biocontrol, Guangdong Key Laboratory of Pharmaceutical Functional Genes, School of Life Sciences, Sun Yat-sen University, Guangzhou, 510275, People's Republic of China.

³ Beijing University of Chinese Medicine, Dong San Huan Road, Chao-yang District, Beijing, 100029, People's Republic of China.

Abstract

RAG1 and RAG2 are essential subunits of the V(D)J recombinase required for the generation of the variability of antibodies and T-cell receptors in jawed vertebrates. It was demonstrated that the amphioxus homologue of RAG1-RAG2 is encoded in an active transposon, belonging to transposase DDE superfamily. We show here that the RAG transposon has been active through the deuterostome evolution and is still active in several lineages. The RAG transposon corresponds to several families present in deuterostomes. RAG1-RAG2 V(D)J recombinase evolved from one of them, partially due to the new ability of the transposon to interact with the cellular reparation machinery. Considering the fact that the RAG transposon survived millions of years in many different lineages, in multiple copies and that DDE transposases evolved many times their association with proteins involved in repair mechanisms, we propose that the apparition of V(D)J

recombination machinery could be a predictable genetic event.

Introduction

The recombination-activating gene products known as RAG1 and RAG2 proteins constitute the enzymatic core of the V(D)J recombination machinery of jawed vertebrates. The RAG1-RAG2 complex catalyzes random assembly of variable, diverse and joining gene segments that are present in the jawed vertebrate genome in numerous copies and together with hyper-mutation generate the enormous diversity of the assembled antibodies and T-cell receptors. Therefore, RAG1-RAG2 role in the V(D)J rearrangement of antigen receptors is crucial for the jawed vertebrate adaptive immunity¹. Concerning the origins of RAG1-RAG2 remains elusive for more than 30 years as the genes were only found in jaw vertebrates². On the other hand, striking similarities between RAG1 and DDE transposase has been noted: common reaction chemistry for DNA cleavage, similar organization of protein domain structure and sequence similarities between recombination signal sequences (RSSs) and terminal inverted repeat (TIRs) targeted by transposases^{3,4}. The hypothetical transposon ancestry of RAG was further supported upon the demonstration of RAG1-RAG2 mediated transposition *in vitro*^{5,6} and *in vivo*⁷⁻¹⁰, though the efficiency of such reactions *in vivo* is highly disfavored comparing to recombination.

A next step in the understanding of RAG1-RAG2 recombinase evolution was the discovery of a RAG1-RAG2-like locus in purple sea urchin genome, which genes for both proteins are oriented in close proximity in a head-to-head manner as RAG1-RAG2 locus in vertebrates. However this locus lacks TIR and thus does not show typical features of a transposon¹¹.

Due to the similarity between RAG1 and *Transib* transposon (a family from the DDE transposon superfamily) and the fact that RAG2 lacks similarity to any known transposon protein, even though harbors Kelch-like repeats and PHD domain as other eukaryotic proteins, led several authors to propose that a *Transib*-like transposon joined the deuterostomian ancestor genome followed by exon shuffling events bringing *Transib* and the ancestor of RAG2 together⁴. As a result, the RAG1-

RAG2 locus was then recruited for an unknown function. A second, much more recent recruitment as RAG1-RAG1 V(D)J recombinase most likely occurred at the base of the jawed vertebrate evolution. Kapitonov and Koonin¹² went a step further and provided *in silico* evidences that RAG1 and RAG2 subunits of the V(D)J recombinase evolved from two proteins encoded in a single transposon as they found three sequences that could correspond to fossilized RAG1-RAG2 transposon (including TIRs) in one starfish genome. A major step in the understanding of the RAG1-RAG2 evolution was reported by our group¹³ showing for the first time the presence of an active RAG transposon in the cephalochordate *Branchiostoma belcheri* named ProtoRAG. The full length ProtoRAG transposon is bound by 5 bp target site duplications (TSDs) and a pair of terminal inverted repeats (TIRs) resembling V(D)J recombination signal sequences (RSSs). Between the TIRs reside tail-to-tail oriented, intron-containing and co-transcribed, RAG1-like and RAG2-like genes. The RAG transposon has been recently active in amphioxus as shown by indel polymorphisms. Furthermore the amphioxus RAG1-RAG2-like proteins together could mediate TIR-dependent transposon excision, host DNA recombination, transposition and even signal joint formation at low frequency, using reaction mechanisms similar to those used by vertebrate RAGs¹³. Here we bring more information about the evolution of RAG transposons. We show that beside *B. belcheri*, an active RAG transposon is found in the hemichordate *Ptychodera flava*, that several fossilized transposons are found in several deuterostomes species suggesting that RAG transposon has been active through the history of the deuterostome lineage.

Results

Description of an active RAG transposon in *P. flava* and many fossilized transposons in deuterostomes

Due to the discovery of an active RAG transposon in amphioxus *B. belcheri*, we screened all the available genome and EST projects using the query sea urchin RAG1L and RAG2L sequences. Many hits in several deuterostomians species were found, hits are found in protostomians but they

79 show low similarity and correspond to the transib transposons¹⁴ and the chapaev transposon family
80 ¹⁵. The family reported by Panchin and Moroz as well as many other families were found during our
81 survey. However the connection between these families and the RAG1-RAG2 is not clear even if
82 they are related.

83 Among the hits found in deuterostomes, one of them corresponds to a complete transposon and
84 other several fossilized transposons (see Supplementary Table 1) in the hemichordate *P. flava*. In
85 other deuterostome species, we evidenced RAG1L-RAG2L structures without TIRs but with many
86 fragment copies of the RAG1L-RAG2L locus, some species with an incomplete transposon with
87 TIR and RAGL sequences and many other copies of RAG1L-RAG2L fragments. The presence of
88 TIR on many of these copies might indicate that they correspond to fossilized transposons.

89 Transcribed sequences database are available for several deuterostomes and in most of the case
90 RAG1L and RAG2L transcripts are found, complete or incomplete, thus revealing the
91 domestication of the transposon or their activity.

92 Based on the phylogeny of the RAG1L and RAG2L protein sequences, we can find several RAG
93 families in *P. flava*. Among them, B and C families have unambiguous TIR and TSD structure. The
94 phylogenetic analysis is shown in Figure 2 and supplementary Table 1 and described in the
95 phylogenetic relation between the RAG families session. Two copies of B family show a TSD-
96 5TIR-RAG1L-RAG2L-3TIR-TSD structure. While one of these copies encodes a complete RAG1L
97 and RAG2L protein, the other one has suffered different level of pseudogenization. But its presence
98 confirm beyond doubt that the authentic RAG transposon appears in this family. In the other hand,
99 the C family has four copies with 5TIR-(RAG1L-RAG2L)-3TIR structure, two of them even
100 containing the correct TSD in addition of 12 5/3TIR. All this copy seems to be inactivated
101 (Supplementary Table 1). We failed to find TSD-TIR structure for other RAG-like families (A,
102 unclassified families) in *P. flava*, this could be due to the poor genome assembly or to the fact that
103 some families have become inactive. Anyway, this finding is sufficient to prove that multiple
104 families of RAG transposon have been and are thriving in *P. flava*. Moreover, we found several

fossilized transposons in the case of *Patiria minata* as partially described in 2015¹², a 5TIR-
RAG1L_fragment-3TIR structure containing TSD and no RAG2L protein, a 5TIR adjacent to
RAG1L structure (TSD-5TIR-RAG1L) and other 12 5/3TIR sequences. These structures indicate
that RAG was an active transposon during the echinoderm evolution. Afterwards a comparative
sequence analysis was made in *B. belcheri*, *Branchiostoma floridae*, *P. flava* (Pfl) and *P. minata*
(Pmi) TIR sequences (Figure 2) showing no identity between different *Transib*, vertebrate RSS and
amphioxus, Pmi and Pfl species if we exclude the first CAC nucleotides. Nonetheless both
sequences analyzed in amphioxus, shares TIR similarity, suggesting a possibly common origin of
RAG transposon between this two species of amphioxus. However, there is no identity between B
and E RAG transposon families in *P. flava*, suggesting despite the similarity between RAG-like
proteins of both families, no TIR similarity between each other, as they may be not reactive or
functionally compatible. Previously, an equivalent of RSS nonamer, a stretch of nine highly
conserved nucleotides has been found in the amphioxus ProtoRAG TIR, though this ProtoRAG
nonamer have no similarity with the nonamer found in RSS¹³. However, there are no such nonamer
or equivalently conserved oligomer found in *P. minata* and *P. flava* B and C ProtoRAG family. All
this suggests that the nonamer structure is not important in echinoderms and hemichordates phyla,
but became important in amphioxus and vertebrates.

The species tree in Figure 1 (see also Supplementary Table 1) shows a summary of RAG1L-RAG2L
sequences distribution in deuterostome according to the available data. When genomic and
transcription data are available species name appears in red while when only genomic data are
available the species names are colored in blue and by last only available expressed sequence data
corresponds to the species name colored in black. We predict that the transposon is active if bona
fide sequences are present in the genome in several copies and fragments and if the putative
transposons are transcribed as in the case for *P. flava* RAGL-B and *B. belcheri*. In the other hand *P.*
miniata seems not to be transcribed since only fossilized transposons are found in the genome. In
two species of sea urchin, *Eucidaris tribuloides* and *Lytechinus variegatus*, no transcribed

131 sequences are found, but many copies of RAG1L-RAG2L are present on the genome without TIRs
132 indicating that might be fossilized transposons that became inactivated by the loss of the TIR
133 sequences.

134 The case of *S. purpuratus* is more difficult to understand: the published RAG1L-RAG2L locus ¹¹
135 renamed here RAG1L-RAG2L B1, was believed to be domesticated as the coding sequence is not
136 interrupted by stop codons and therefore RAG1L and RAG2L could be functional proteins has no
137 TIR sequences and both RAG1L and RAG2L are transcribed ¹¹. However, we found many
138 fragments highly similar to this sequence in the *S. purpuratus* genome which could reveal a recent
139 transposition event followed by the domestication of one of its copies (see supplementary data and
140 figure 2). We find another RAGL copy arose from a duplication event occurred at the origin of the
141 echinoderms, named RAG1L-B2. The RAG1L-B2 copy is only found fragmented with multiple
142 recent copies in the genome whereas is complete as RAG1L transcript. A possible explanation for
143 this second locus could be the existence of an active transposon with the genome sequence not well
144 assembled or otherwise a domesticated or recent fossilized transposon. For most of the species we
145 do not have information at the genomic level, but if we find RAGL transcript, this sequence could
146 correspond to an active transposon, domesticated transposon or recent pseudogene. Anyway, this
147 shows that the transposon has been present in their ancestors.

148

149 **Features of the proteins encoded by the RAG-like proteins**

150 In ambulacraria (echinoderm and hemichordate) the deuterostome RAG1-like, 816-1136 aa-long
151 shares around 26.52% sequence identity between RAG1L-B family and vertebrate RAG1, around
152 33.21% between the orthologous RAGL-A family and the vertebrate RAG1 and only 27.79%
153 between RAG1L-A and RAG1L-B, while inside RAG1L-B family are sharing 48.75% of sequence
154 identity and only 20.13% respect to *Transib* transposase in terms of core region. As regards to
155 RAG1 lancelet, 30.47% and 37.62% sequence identity are shared with A and B families
156 respectively and only 27.45% with RAG1 vertebrate (see Supplementary Figure 2A). Clusters of

high identity are found between RAG1L and vertebrate RAG1 along much of their length, suggesting conservation of multiple functional elements. Vertebrate RAG1 uses four acidic residues to coordinate critical divalent cations at the active site¹⁶ and all four are conserved in RAG1L (Supplementary Figure 1A, red highlight). In addition, many cysteine and histidine residues that coordinate zinc ions and play a critical role in proper folding of RAG1¹⁷, are conserved between RAG1L and vertebrate RAG1 (Supplementary Figure 1A, * and # symbols). However, RAG1L does not share much identity with vertebrates RAG1 in the region corresponding to the nonamer binding domain, consistent with the fact that RAG transposons TIRs have no clear similarity to the RSS nonamer. In fact, different families of RAG1-like have little similarity to each other in the putative nonamer binding domain, consistent with the fact that different ProtoRAG families have very different TIR sequences and no obvious nonamer regions, excluding the amphioxus TIR. Finally, there are also some RAG1-like specific conserved regions (see Supplementary Figure 1, underlined by *). Should be noted that PflRAG1L-A share many sites with jawed vertebrate RAG1. RAG2L 366-535aa long, shares weak sequence identity between B family and vertebrate RAG2 (18.69%) and between B family and lancelet RAG2L (25.02%). In the other hand RAG2L-B family shares around 45.90% while lancelet RAG2L shares only 20.24% identity with RAG2 vertebrate (supplementary Figure 2B). However, the N-terminal six-bladed β -propeller domain (six Kelch-like repeats), which is conserved in both vertebrate RAG2 and ProtoRAG RAG2L, can be discerned in RAG2L. Strikingly, amphioxus RAG2L lacks the entire RAG2 C-terminal region, including the PHD domain as shown previously¹³. However, this PHD domain is present in all other echinoderm and hemichordate RAG2 proteins (Supplementary Figure 1B). Thus, the absence of this region in amphioxus RAG transposon might be a secondary loss.

179

180 **Phylogenetic relation between the RAG families**

181 The phylogenetic analysis with the complete RAG sequences from the available deuterostome
182 species are shown in Figure 3A and 3B and synthesized in Table 1. At least two sub-families have

183 been present in the ancestral deuterostome, named RAGL-B and RAGL-A. Other families as
184 RAGL-C have not been included in the phylogenetic history as they are found only in one species
185 (Table 1).

186 In the case of the orthologous relation found between RAG1L-A of *P. flava* (hemichordate) and
187 vertebrates RAG1 recombinase, we can observe that RAGL-A was lost in many lineages excluding
188 hemichordate and jawed vertebrates. RAGL-B conversely, is lost in tunicates and in vertebrate
189 lineage but conserved in several lineages as cephalochordates, hemichordates and echinoderms.
190 Furthermore the phylogenetic analysis shows that RAGL-B has been duplicated in the echinoderm
191 ancestor after the hemichordate/echinoderm split and both copies have been kept (even if most of
192 them have been inactivated) in most of the echinoderm species (Table 1 and Figure 3C).

193

194 **RAG transposon has been active during the deuterostome evolution**

195 The Figure 4 shows the RAG transposon activity summary during deuterostome evolution. The
196 transposon has been active in the deuterostome ancestor and in the branch that leads to the common
197 ancestor of chordate, still active in cephalochordate and domesticated as a RAG1-RAG2 V(D)J
198 recombinase in the common ancestor of jawed vertebrate. The transposon has been lost in the
199 Petromyzon lineage (the class Myxini genome is not available and therefore we cannot state for this
200 important phylogenetic phyla). The transposon has been active in the branch originated from the
201 node between deuterostome and ambulacraria antecessors. It remains active in hemichordates inside
202 the subphylum of Enteropneusta (at least on the *P. flava* lineage) but is lost in the other
203 enteropneusts as *S. kowalevskii*. Unfortunately we do not have genome information for the other
204 hemichordate subphyla: Pterobranchia. In the case of the echinoderm lineage, the transposon has
205 been present in the echinoderm common ancestor, in the branch leading to the common ancestor of
206 crinoid, in the clade formed by the sea urchin and holothuroids and in the clade formed by
207 starfishes/ophiuroids. It has been then lost in the crinoid lineage. The transposon has been active in
208 the branch that goes from the common ancestor of echinoderm to the common ancestor of sea

urchin/Holothuroids and starfishes/brittle stars. Concerning the Asteroidea/Ophiuroidea group, the transposon has been active in their common ancestor and has been active in the Ophiure lineage in particular in *O. spicalatus* where the transposon is likely to be active or has lost its activity recently. The transposon seems to have been inactivated in the starfish lineage but fragments showing similarities to RAG1-L and/or RAG2-L transposons are found in this species. Furthermore transposons are clearly found fossilized in *P. miniata*. In the case of sea urchin/holothuroids group, it seems that the transposon has been active in their common ancestor and inactivated in the holothurian lineage. By last transposon seems to be active in some sea urchin lineages as in *E. tribuloide* but much less in others.

Discussion

In this report we show that a RAG transposon has been present in the deuterostome common ancestors and was active since then in some lineages, fossilized later during evolution and domesticated at least in the case of jawed vertebrates. The structural and regulatory features that cause the jawed vertebrate RAG V(D)J recombinase to favor deletional/inversional recombination over transposition as in the case of the RAG transposase¹³ is not yet resolved. It could be explained by how the cleaved ends and particularly the signals ends are processed. The RAG V(D)J recombinase binds signal ends tightly as expected for a transposase but it has acquired the possibility to give up these end efficiently to the non-homologous end joining machinery. This allows recombination and prevents the propagation¹. Thus the jawed vertebrate V(D)J recombinase differs from the current RAG transposon, as well as its transposon precursor, in how it interfaces with the DNA repair apparatus. This new property occurred likely in the jawed vertebrate common ancestor.

DDE transposases have been shown to interact with repair proteins. For example the Sleeping Beauty transposase interacts directly with the Ku70 repair protein¹⁸ and the pogo transposase of *D. melanogaster* interacts with the proliferating cell nuclear antigen (PCNA), a key protein for DNA

235 replication and repair¹⁹. Therefore, the associations of DDE transposon with DNA repair and
 236 replication factors appear to evolve in a convergent manner²⁰. This characteristic and the fact that
 237 the transposon survived during millions of years in multiple copies in different lineages increased
 238 the probability of the co-option of the RAG transposon as V(D)J recombinase. Therefore the
 239 apparition of V(D)J recombination machinery in the jawed vertebrate phyla could be labeled as a
 240 predictable genetic events.

241 Our results could also explain better the origins of the T-cell receptor and B-cell receptor gene
 242 organization. The earlier proposed scenario^{4, 21} involved an insertion of the RAG transposon into
 243 the ancestral IG/TCR V-gene, prior to the externalization of the RAG1-RAG2 complex while
 244 leaving the RSS-like TIR within the IG/TCR V-gene. This was followed by duplication of this new
 245 genetic structure: **VRSS-RSSJ**. The RAG transposon was then co-opted as V-J recombinase and the
 246 system started to work. However, this scenario explains the V-J structure IG light chain, TCR alpha
 247 and gamma chain but not the VDJ organization of IG heavy chain or TCR beta and delta chains.
 248 This is how we explain this: while one RAG was domesticated (likely RAGL-A orthologue), other
 249 RAG transposons (likely RAGL-B orthologue) were still active as one of them split the **VRSS-**
 250 **RSSJ** copy and gave rise to **VRSS-RSSDRSS-RSSJ**. RAG transposase became then extinct and
 251 finally was lost during vertebrate evolution.

252

253 **Material and methods**

254 **Identification of RAG1 and RAG2-like sequence in different data bases**

255 RAG1-RAG2-like locus identified in the echinoderm *Strongylocentrotus purpuratus* and in the
 256 vertebrate genome were used as a protein sequence to perform a TBLASTN-based search against
 257 the NCBI nr protein, transcriptome shotgun assembly (TSA) and the WGS database (as of June
 258 2016)²². These retrieved sequences were extracted and translated by ExPASy Translate tool.
 259 Potential open reading frames of RAG1-RAG2 elements used in this study were predicted using
 260 FGENESH²³ with the sea urchin organism specific gene-finding parameters. The mRNA sequences

261 were then assembled into contigs by CAP3²⁴.

262

263 **Phylogenetic analysis**

264 The alignment and trees were constructed using MEGA6 (complete deletion, WAG with Freqs. (+F)
265 correction model, 1000 bootstrap replicates²⁵). Thus, whether are active, fossilized or domesticated
266 were classified into families. Short sequence copies, were analyzed one by one with the reference
267 data set as it is not informative to compare sequences that do not overlap.

268

269 **Sequence searches for TIR and TSD motifs**

270 We used three methods to search target site duplication (TSD) and terminal invert repeat (TIR)
271 sequences. In the first method, the upstream and downstream 20 Kb of sequence flanking the
272 RAG1-RAG2-like sequences were extracted and separated into a set of small fragments (using a
273 window size of 60 bp and a step size of 1 bp). In the first method, each upstream fragment was
274 compared with each downstream fragment for 4-6 bp TSDs and possible TIRs using a custom Perl
275 script. We required 40% identity for potential TIR pairs, and allowed only one mismatch for TSD
276 pairs. In the second method, all upstream fragments were compared against all downstream
277 fragments using BLAST. We required a minimum e-value of 100 and sequence identity of 40% in
278 the BLAST search. However, these two methods failed to work well and provided no reliable
279 results. Therefore, we turn to the e third method. In this method, we posited that if there are multiple
280 copies of ProtoRAG transposons in the genome assembly, comparison between these copies could
281 help to determine their terminal sequences (TIR, etc.).

282

283 We focused on finding more complete elements that contain both TIR and RAG gene fragments,
284 such as “5TIR-RAGs-3TIR”, “5TIR-RAGs” and “RAGs-3TIR”.

285 Here is our procedure:

1. First we identified all genomic regions containing RAG1/2 fragments by using TBLASTN and the amphioxus and vertebrate RAG1/2 proteins as queries;
2. The region containing RAG1/2 plus upstream 20kb and downstream 20kb was extracted, which we called the RAG region;
3. Because there should a clear border between the ProtoRAG and the host DNA, we can determine the potential 5' and 3'-terminal of the ProtoRAG transposon by comparing RAG regions with each other by using BLASTN (see the figure below);
4. Finally, we examine the potential 5/3-terminal sequences of the RAG regions. Most of them have been destroyed and therefore no detectable TIRs, but there are several of them show clear and intact TIR structure.
5. And the TSD if presents, should be right next to the TIR sequences.

Therefore, the sequences containing the RAG1/2-like fragments and the 20 Kb flanking regions were compared to each other and also to the whole genome assembly using BLAST. The terminal sequences were analyzed using a custom Perl script and then subjected to manual inspection.

Summary of data availability

In order to detect the absence or presence of a given structure in the genome or transcriptome, we need to extract all the available taxonomic information from the NCBI database. It has to be noted that even if the sequence for a given genome is not complete, when RAG1L-RAG2L seems to be an active transposon, we should find an active or at least a fossilized transposons (in several copies). Focusing on the genome database we can find species as *Parastichopus parvimensis*, *Acanthaster planci*, *Ophiothrix spiculata*, *Petromyzon marinus*, *Branchiostoma belcheri*, *Oikopleura dioica*, *Botryllus schlosseri* & *Ciona savignyi*. Transcript sequences can be provided for *Saccoglossus kowalevskii*, *Anneissia japonica*, *Psathyrometra fragilis*, *Abyssocucumis albatrossi*, *Sclerodactyla briareus*, *Apostichopus japonicus*, *Parastichopus californicus*, *Echinarachnius parma*, *Evechinus*

313 *chloroticus*, *Paracentrotus lividus*, *Sphaerechinus granularis*, *Arbacia punctulata*, *Henricia* sp. AR-
 314 2014, *Echinaster spinulosus*, *Peribolaster folliculatus*, *Leptasterias* sp. AR-2014, *Pisaster*
 315 *ochraceus*, *Marthasterias glacialis*, *Asterias rubens*, *Asterias forbesi*, *Asterias amurensis*, *Luidia*
 316 *clathrata*, *Patiria pectinifera* & *Ophiocoma echinata*. Finally, together with genomic information
 317 and transcript expression we have *Ptychodera flava*, *Eucidaris tribuloides*, *Strongylocentrotus*
 318 *purpuratus*, *Lytechinus variegatus*, *Patiria miniata*, *Homo sapiens*, *Mus musculus*, *Gallus gallus*,
 319 *Xenopus tropicalis*, *Latimeria chalumnae*, *Danio rerio*, *Carcharhinus leucas*, *Carcharhinus*
 320 *plumbeus*, *Branchiostoma floridae* and *Ciona intestinalis*.

321

322 **References**

- 323 1. Teng G., Schatz D.G., 2015. Regulation and Evolution of the RAG Recombinase. *Adv.*
 324 *Immunol.* **128**, 1-39.
- 325 2. Danchin E. *et al.*, 2004. The major histocompatibility complex origin. *Immunol. Rev.* **198**,
 326 216-232.
- 327 3. Kapitonov V.V., Jurka J., 2005. RAG1 core and V(D)J recombination signal sequences were
 328 derived from Transib transposons. *PLoS Biol.* **3**, 998-1011.
- 329 4. Fugmann S.D., 2010. The origins of the Rag genes from transposition to V(D)J
 330 recombination. *Semin. Immunol.* **22**, 10-16.
- 331 5. Agrawal A. *et al.*, 1998. Transposition mediated by RAG1 and RAG2 and its implications
 332 for the evolution of the immune system. *Nature* **394**, 744-751.
- 333 6. Hiom K. *et al.*, 1998. DNA transposition by the RAG1 and RAG2 proteins: a possible
 334 source of oncogenic translocations. *Cell* **94**, 463-470.
- 335 7. Chatterji M. *et al.*, 2006. Mobilization of RAG-generated signal ends by transposition and
 336 insertion in vivo. *Mol. Cell. Biol.* **26**, 1558-1568.
- 337 8. Curry, J.D. *et al.*, 2007. Chromosomal reinsertion of broken RSS ends during T cell

338 development. *J. Exp. Med.*, 204, 2293-2303.

339 9. Ramsden D.A. *et al.*, 2010. Weed, B.D., Reddy, Y.V. V(D)J recombination: Born to be wild.
340 *Semin Cancer Biol.* **20**, 254-260.

341 10. Vanura K. *et al.*, 2007. In vivo reinsertion of excised episomes by the V(D)J recombinase: a
342 potential threat to genomic stability. *PLoS Biol.* **5**,
343 <http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0050043>

344 11. Fugmann S.D. *et al.*, 2006. An ancient evolutionary origin of the Rag1/2 gene locus. *Proc.*
345 *Natl. Acad. Sci. USA* **103**, 3728-3733.

346 12. Kapitonov V.V., Koonin E.V., 2015. Evolution of the RAG1-RAG2 locus: both proteins
347 came from the same transposon. *Biol. Direct*,
348 <http://biologydirect.biomedcentral.com/articles/10.1186/s13062-015-0055-8>.

349 13. Huang S. *et al.*, 2016. Discovery of an Active RAG Transposon Illuminates the Origins of
350 V(D)J Recombination. *Cell*. **166**, 102-114.

351 14. Panchin Y., Moroz L.L., 2008. Molluscan mobile elements similar to the vertebrate
352 recombination-activating genes *Biochem Biophys Res Commun.* **369**, 818–823.

353 15. Kapitonov V.V., Jurka J., 2007. Chapaev-a novel superfamily of DNA transposons. *Rebase*
354 *Reports.* **7**, 777–777.

355 16. Ru H. *et al.*, 2015. Molecular Mechanism of V(D)J Recombination from Synaptic RAG1-
356 RAG2 Complex Structures. *Cell*. **163**, 1138-1152.

357 17. Kim M.S. *et al.*, 2015. Crystal structure of the V(D)J recombinase RAG1-RAG2. *Nature*
358 **518**, 507-511.

359 18. Izsvák Z. *et al.*, 2004. Healing the wounds inflicted by sleeping beauty transposition by
360 double-strand break repair in mammalian somatic cells. *Mol Cell*. **13**, 279-290.

361 19. Warbrick E. *et al.*, 1998. PCNA binding proteins in *Drosophila melanogaster*: the analysis

- 362 of a conserved PCNA binding domain. *Nucleic Acids Res.* 26, 3925-3932.
- 363 20. Feschotte C., Pritham E.J, 2007. DNA transposons and the evolution of eukaryotic genomes.
364 *Annu. Rev. Genet.* **41**, 331-368.
- 365 21. Koonin E.V., Krupovic M., 2015. Evolution of adaptive immunity from transposable
366 elements combined with innate immune systems. *Nat. Rev. Genet.* **16**, 184-192.
- 367 22. Altschul S.F. *et al.*, 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.
- 368 23. Solovyev V. *et al.*, 2006. Automatic annotation of eukaryotic genes, pseudogenes and
369 promoters. *Genome Biol.* **7**, S10.1-S10.12.
- 370 24. Huang X., Madan A., 1999. CAP3: A DNA sequence assembly program. *Genome Res.* **9**,
371 868-877.
- 372 25. Tamura K., *et al.*, 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0.
373 *Mol. Biol. Evol.* **30**, 2725-2729.

374

375 We thank the EBM laboratory for advice and Olivier Loison for editing the manuscript.

376

377 **Author's contribution: JRMP, PP and SFH conceived the project and design the study. JRMP,**
378 **PP and SFH analyzed the results. JRMP, PP, SFH and ALX wrote the manuscript.**

379

380 The authors declare no competing financial interest

381 Correspondence and requests for materials should be addressed to **pierre.pontarotti@univ-amu.fr**
382 and **morales.poole@gmail.com**

383

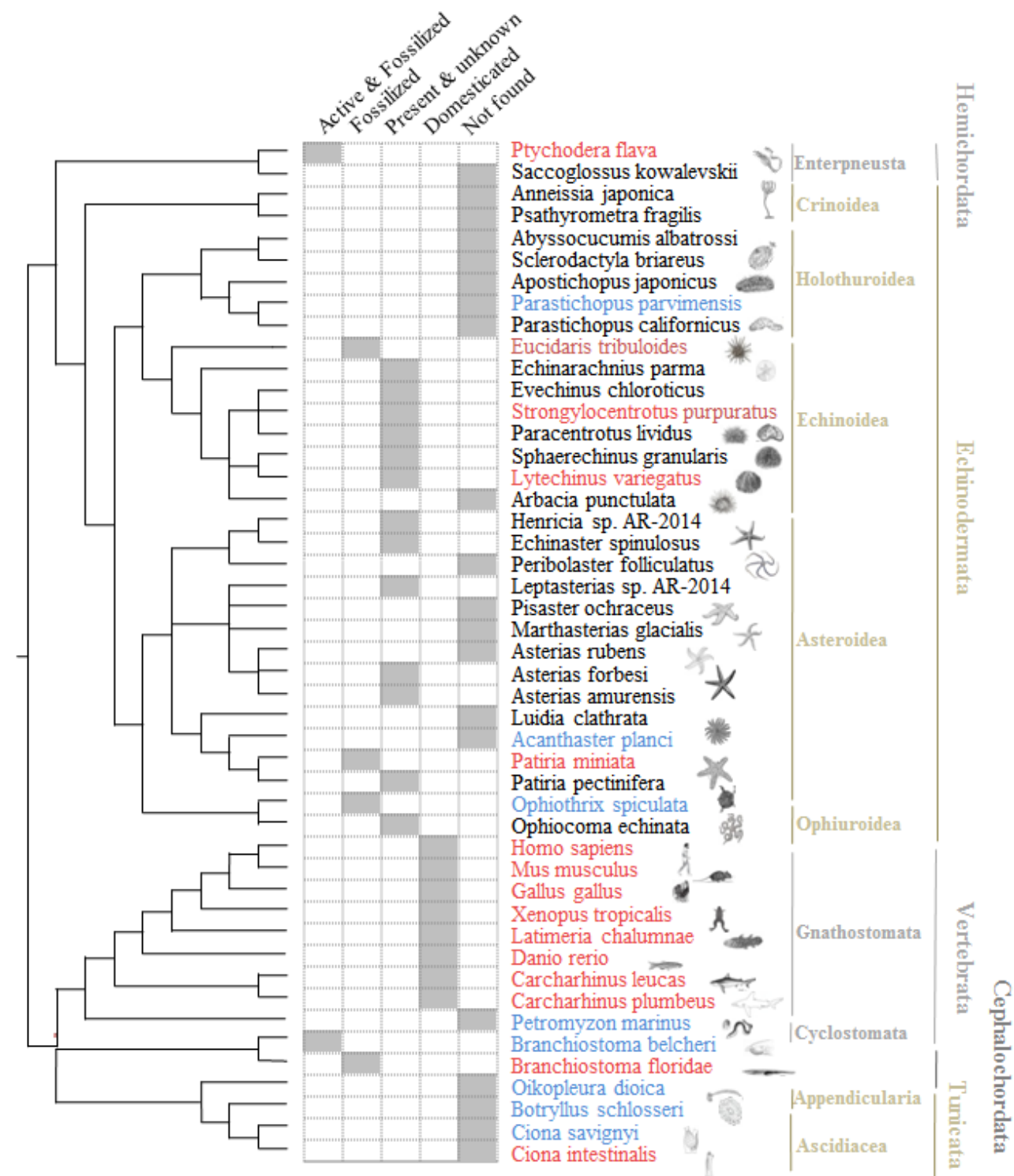
384

385

386

387

388 **Figures and tables**



389

390 **Figure 1 | Distribution of the RAG1-RAG2 sequences in deuterostomes.** Only species for which
391 the genomic and/or transcription data are available are represented in the phylogenetic tree. Species
392 are colored in red when genomic and transcription data are available, in blue when only genomic
393 data are available and species are colored in black only when expressed sequence data are available.

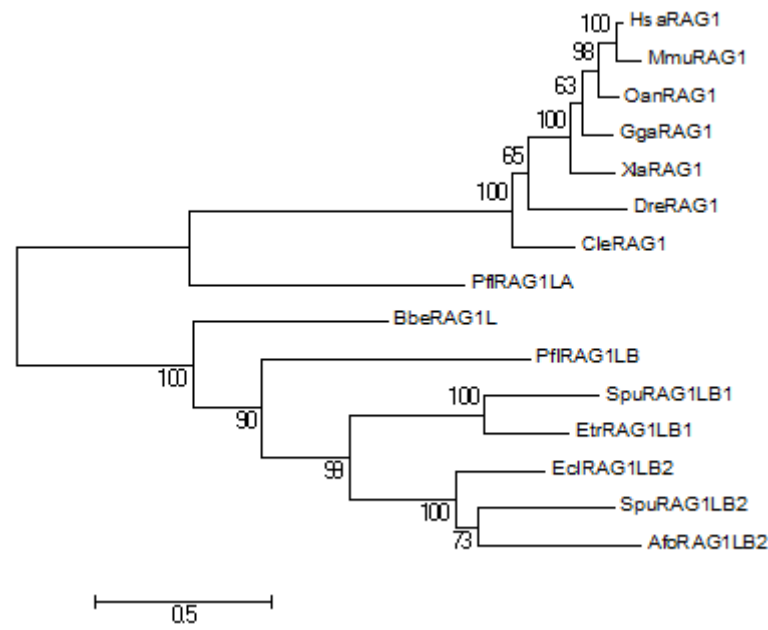
394

```

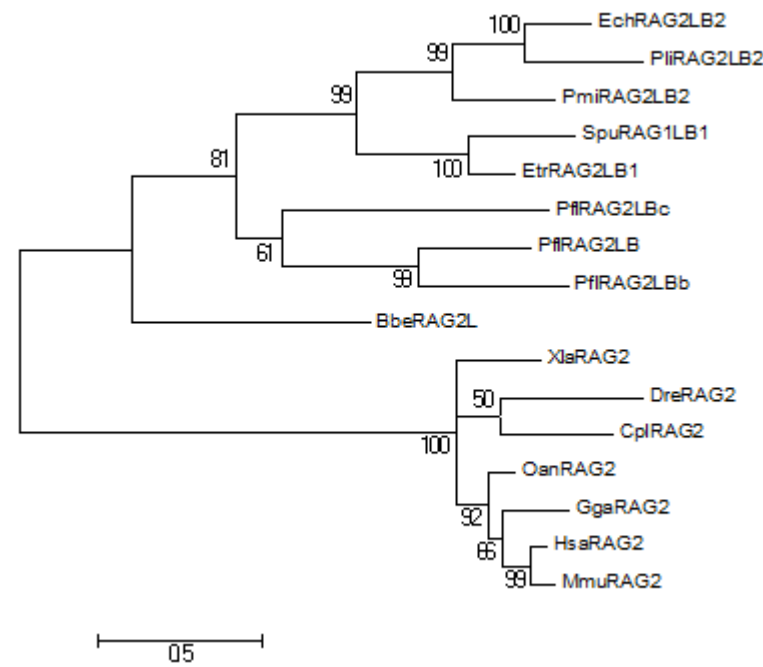
      *
ccgCTGCGtgc : 11 : Pfl_B_BCFJ01017854_BCFJ01052781
gccCAATGtgc : 11 : Pfl_B_BCFJ01094280
cacTGTGG--- : 8 : Pfl_B_BCFJ01052780
cacCATCCgta : 11 : Pfl_C_BCFJ01036631
aacCCCGGctg : 11 : Pfl_C_BCFJ01107546
gtcTGGCA--- : 8 : Pfl_C_BCFJ01102604
ctcGGGTG--- : 8 : Pfl_C_BCFJ01084502
ttaCCTTC--- : 8 : Pfl_C_BCFJ01046932
tgcTGCCA--- : 8 : Pfl_C_BCFJ01016857
tcgCAGTG--- : 8 : Pfl_C_BCFJ01107167
gtgCAATG--- : 8 : Pfl_C_BCFJ01047137
tgcGGGCG--- : 8 : Pfl_C_BCFJ01031953
---CGCATtcg : 8 : Pfl_C_BCFJ01070588
---GGGTGcca : 8 : Pfl_C_BCFJ01150129
---GGGCGgtc : 8 : Pfl_C_BCFJ01103012
---CAGTGtgg : 8 : Pfl_C_BCFJ01107168
---CGCGCtcg : 8 : Pfl_C_BCFJ01287958
---CACGGgta : 8 : Pfl_C_BCFJ01048214
---CGCATtcg : 8 : Pfl_C_BCFJ01083599
----- : - : blank
cgtCCAGGgtc : 11 : Pmi_JH779599
aacCCATAccg : 11 : Pmi_JH774215
ctcTGTTAtat : 11 : Pmi_JH780459
gagTTTAG--- : 8 : Pmi_JH769343
gatTTTAG--- : 8 : Pmi_JH774292
tgtTCATG--- : 8 : Pmi_JH782081
gcgATGTG--- : 8 : Pmi_JH775549
aagCGGGA--- : 8 : Pmi_JH781149
---CGTGGcat : 8 : Pmi_AKZP01156453
---TTCAGcaa : 8 : Pmi_JH771625
---CATATgca : 8 : Pmi_AKZP01165822
---GTGCGgga : 8 : Pmi_AKZP01162400

```

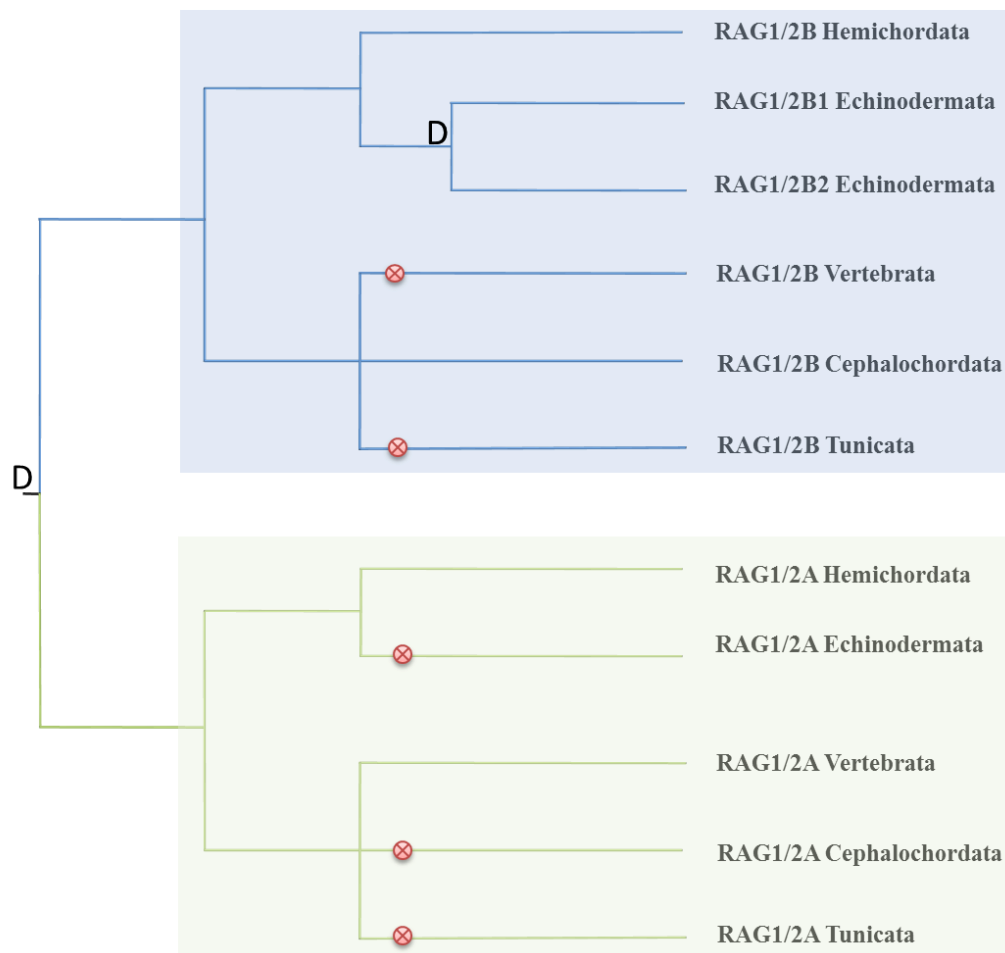




404

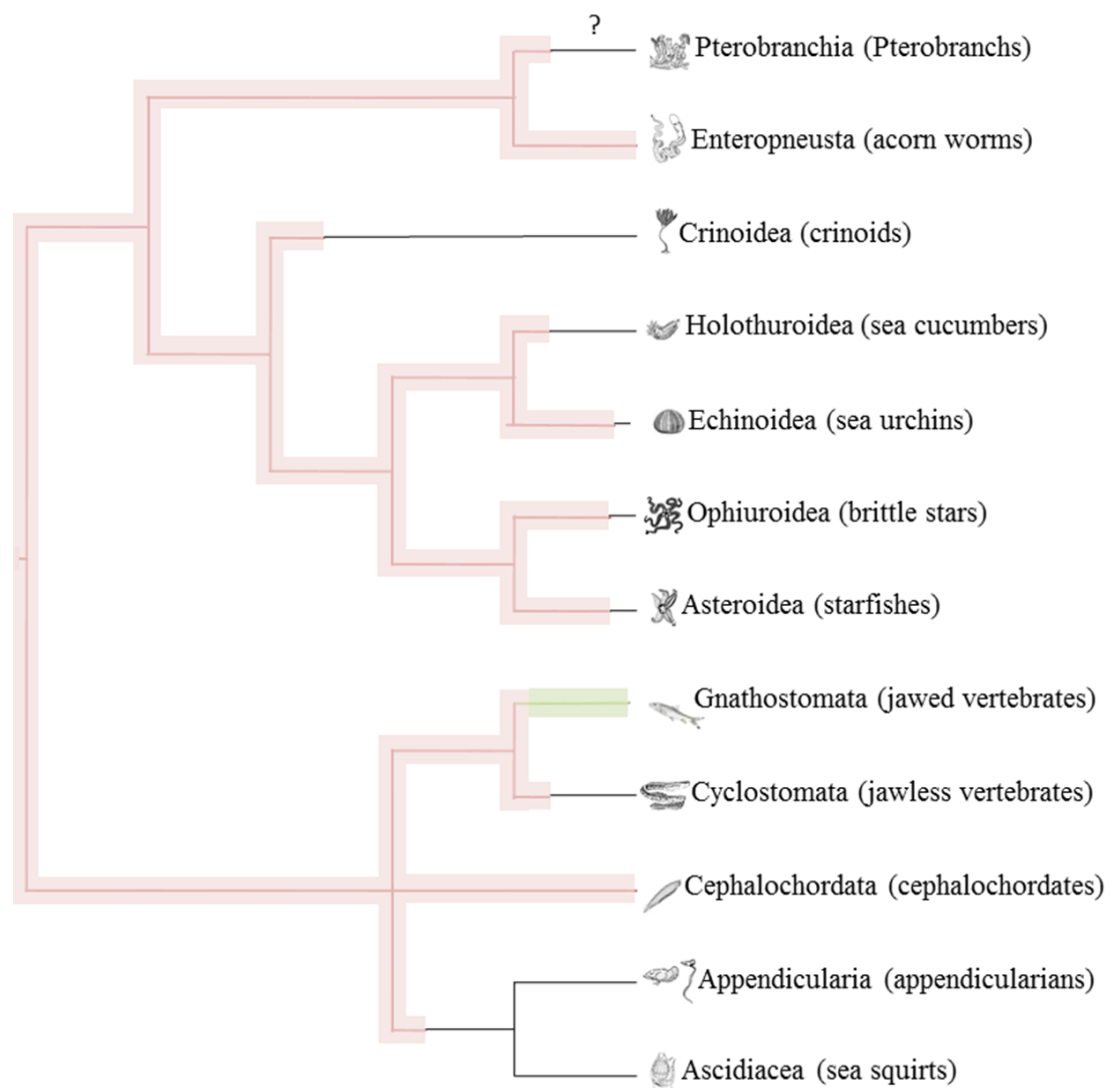


405



406

407 **Figure 3 | Phylogenetic tree with RAG1 and RAG2 complete sequence and outline of the**
408 **duplication (D) and lost (X).** Phylogenetic tree with RAG1 (A) and RAG2 (B) complete
409 sequence. Outline of the duplication and loss of the RAG transposon (C).



410

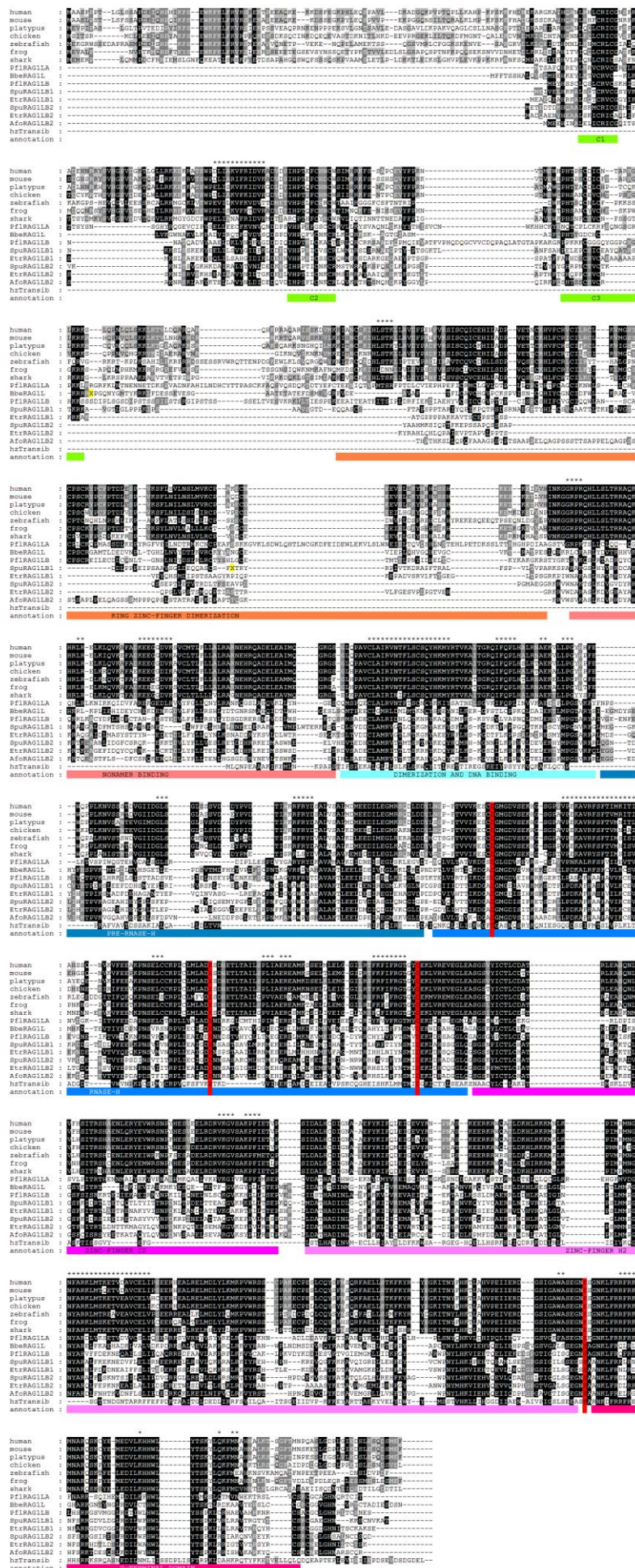
411 **Figure 4 | Evolution of the RAG transposon.** Transposon activity is indicated in bold pink and

412 V(D)J recombinase activity is indicated in bold green.

413

	<i>P. flava</i>	<i>L. variegatus</i>	<i>E. tribuloides</i>	<i>P. lividus</i>	<i>S. purpuratus</i>	<i>E. parma</i>	<i>P. pectinifera</i>	<i>A. aminumensis</i>	<i>Leptasterias</i> sp.	<i>A. forbesi</i>	<i>E. chloronotus</i>	<i>P. minutata</i>	<i>O. spiculata</i>	<i>Hemictia</i> sp. AR-2014	<i>S. granularis</i>	<i>E. spinulosus</i>	<i>O. echinata</i>	<i>A. japonica</i>	<i>P. fragilis</i>	<i>A. albatrossi</i>	<i>S. briareus</i>	<i>A. japonicus</i>	<i>P. parvimonensis</i>	<i>A. californicus</i>	<i>P. punctulata</i>	<i>P. polliculatus</i>	<i>M. glaciatus</i>	<i>A. rubens</i>	<i>L. clathrata</i>	<i>A. planici</i>
RAG1B1-like																														
RAG1B2-like																														
RAG1B-like	*																													
Other families	**																													
Not phylogenetically assigned																														
Not found																														
RAG2B1-like																														
RAG2B2-like																														
RAG2B-like	***																													
Other families	****																													
Not phylogenetically assigned																														
Not found																														

Table 1 | Presence of RAG subfamilies in the different species. Sequences were classified through phylogenetic analysis. Short sequence copies, were analyzed one by one against the reference data described in Figure 3A and 3B as it is not informative to compare sequences that do not overlap. The classification as B family (or A family labeled with “**”) is straightforward as it is based on orthologous relationships between different phyla (differences between echinoderms and hemichordates for example). Inside B family, two groups named B1 and B2 are found in several echinoderms. If an echinoderm sequence is classify as B family, but not as B1 or B2 we call it B-like (RAG1Bd-like is labeled with “*” while RAG2Bb-like and RAG2Bc-like are labeled with “***”). We have two specific cases, C family only found in *P. flava* (RAG1 labeled with “**”) and RAG2 labeled with “****”) and X family in *Ophiotrix spiculata*. The rest of species if they do not belong to A or B family, are not phylogenetically assigned due to the fact that none enough phylogenetic signals are available.



show slightly more similar to vertebrate RAG1, and those regions were labeled with “*”. GenBank accessions for mouse RAG1, shark RAG1, lancelet RAG2L and sea urchin RAG1L are NP_033045, XP_007886047, KJ748699 and NP_001028179, respectively.

(B) Protein alignment of RAG2L with vertebrate RAG2. Color shading shows the conservation of physiochemical properties. The N-terminal amino acid sequence can be grouped into Kelch-like repeats. The central conserved GG motifs of the six Kelch-like repeats are underlined in red. The plant homeodomain (PHD) is also underlined below the alignment. GenBank accessions for mouse RAG2, shark RAG2, lancelet RAG2L and sea urchin RAG2L are NP_033046, XP_007885835, KJ748699 and NP_001028184, respectively.

Percent Identity	Matrix - created by Clustal2.1										
1: transib-1_HM	100.00	18.97	21.85	19.39	19.54	19.60	19.93	18.49	18.40	19.83	21.75
2: BbeRAG1L	18.97	100.00	39.96	35.03	39.49	37.53	40.20	37.88	30.47	27.55	27.35
3: PflRAG1LB	21.85	39.96	100.00	35.39	40.37	40.84	41.04	39.00	28.49	26.54	27.30
4: SpuRAG1LB1	19.39	35.03	35.39	100.00	62.11	43.13	46.87	42.06	27.34	24.74	24.69
5: EtrRAG1LB1	19.54	39.49	40.37	62.11	100.00	45.70	47.65	43.71	28.79	26.60	26.77
6: AfoRAG1LB2	19.60	37.53	40.84	43.13	45.70	100.00	56.55	54.70	26.78	26.87	25.86
7: EclRAG1LB2	19.93	40.20	41.04	46.87	47.65	56.55	100.00	58.77	27.55	28.35	26.75
8: SpuRAG1LB2	18.49	37.88	39.00	42.06	43.71	54.70	58.77	100.00	26.80	27.03	25.79
9: PflRAG1LA	18.40	30.47	28.49	27.34	28.79	26.78	27.55	26.80	100.00	33.26	33.15
10: HsaRAG1	19.83	27.55	26.54	24.74	26.60	26.87	28.35	27.03	33.26	100.00	64.41
11: CleRAG1	21.75	27.35	27.30	24.69	26.77	25.86	26.75	25.79	33.15	64.41	100.00

Percent Identity	Matrix - created by Clustal2.1									
1: HsaRAG2	100.00	55.34	19.93	17.10	19.21	16.34	19.74	18.62	16.75	18.54
2: CplRAG2	55.34	100.00	20.54	17.05	20.21	16.75	20.16	19.05	19.90	21.04
3: BbeRAG2L	19.93	20.54	100.00	29.50	28.61	21.05	23.91	27.46	20.53	25.07
4: PflRAG2LBc	17.10	17.05	29.50	100.00	29.80	28.67	29.26	30.77	27.63	27.82
5: PflRAG2LB	19.21	20.21	28.61	29.80	100.00	29.50	29.61	34.13	33.33	29.48
6: StrRAG2LB1	16.34	16.75	21.05	28.67	29.50	100.00	64.16	41.29	37.20	37.01
7: EtrRAG2LB1	19.74	20.16	23.91	29.26	29.61	64.16	100.00	44.74	42.64	39.91
8: PmiRAG2LB2	18.62	19.05	27.46	30.77	34.13	41.29	44.74	100.00	53.79	49.33
9: EchRAG2LB2	16.75	19.90	20.53	27.63	33.33	37.20	42.64	53.79	100.00	59.09
10: PliRAG2LB2	18.54	21.04	25.07	27.82	29.48	37.01	39.91	49.33	59.09	100.00

Figure S2 | Percent Identity Matrix of RAG1 (S2A) and RAG2 (S2B). In order to provide a multiple alignment, Clustal-Omega requires a guide tree which defines the order in which sequences/profiles are aligned. A guide tree in turn is constructed, based on a distance matrix. Conventionally, this distance matrix is comprised of all the pair-wise distances of the sequences. The distance measure Clustal-Omega uses for pair-wise distances of un-aligned sequences is the k-

459 tuple measure. By default, the distance matrix is used internally to construct the guide tree and is
460 then discarded. By specifying, the internal distance matrix can be written to file.

461

462 **Table S1 | RAGL distribution in non-chordate genome and expressed sequence.** Distribution in
463 the cephalordate phyla: *B. belcheri* and *B. floridae* available in ¹³.

464