The emergence of RAG recombinase to mediate V(D)J recombination has been considered an important milestone in the evolution of adaptive immunity in jawed vertebrates. In past decades, hypotheses, including the invasion of viral or bacterial genes and the transposition of mobile elements, have been proposed to shed light on the origin and evolution of the RAG genes and V(D)J recombination. In 2016, our discovery of the long-sought-after ProtoRAG transposon in the lancelet—a 'living fossil' of vertebrates—directly supported the hypothesis that the RAG recombinase complex originated from an ancestral RAG-like (RAGL) transposon [1]. A typical ProtoRAG contains a pair of recombination signal sequence (RSS)-like terminal inverted repeats (TIRs) and convergently orientated RAG1-like (RAG1L) and RAG2-like (RAG2L) genes [1,2]. After the discovery of ProtoRAG, RAGL transposons were recently found to be distributed widely in bilaterians [3,4]. However, the existence of the RAGL transposon in more primitive organisms and its continuous evolution remain to be further elucidated.

After searching for RAG1 or RAG2 homologs in >680,000 assembled genomes (Supplementary Table S1), 786 RAG1L and 191 RAG2L homologs were found in 173 species distributed broadly across the Eukaryota (Supplementary Fig. S1). Although most of the identified sequences were partial, some complete and potentially active RAGL transposons were found in some protostomes and cnidarians, as previously reported [3,4]. In Protostomia, RAGL genes are complete and tightly linked in several lophotrochozoans but are fragmented and uncoupled in ecdysozoans. A similar observation was found in the Cnidaria, as RAGLs are complete and probably active in several corals such as Fungia costulata and Fungia tenuis (Fig. 1a), but are fragmented and uncoupled in some other cnidarians. These observations suggest that the RAGL transposon emerged earlier than the divergence of bilaterians and non-bilaterians. Importantly, complete and tightly linked RAG1 and RAG2 homologs were found in the unicellular microalgae Aureococcus anophagefferens (Class Pelagophyceae, Phylum Ochrophyta, Kingdom Stramenopila). Fragments of RAG1L homologs were also identified in several other primitive eukaryotes (Supplementary Fig. S1), tracking the origin of RAGL homologs back to the early eukaryotes for the first time.

After cloning the ancient AanRAGL from A. anophagefferens (Fig. 1a), typical transposon features of AanRAGL elements were found, including a pair of asymmetric TIRs and the 5-bp target site duplications (TSDs). Similar to the TIRs in ProtoRAG and the RSSs in V(D)J recombination, the paired TIRs of AanRAGL are characterized by two conserved elements, a 13-bp element (CA-CACCGAACCT) and a 10-bp element (CCTCAA[C/T]A[C/T]G), which are separated by a pair of 4/13-bp space sequences (Fig. 1b). Bracketed by the pair of TIRs, two single-exon encoded genes similar to RAG1 and RAG2 were identified. However, unlike the other identified RAGL transposons, AanRAG2L is located upstream of AanRAG1L and transcribed toward AanRAG1L in the same direction (Fig. 1a). To reveal the transpositional activity of AanRAGL in vivo, flanking sequences of AanRAGL were cloned from the genome of the NCMA culture (CCMP 1984, source of the reference genome) using the splinkerette-PCR method. AanRAGL was found to have been polymorphically inserted into the host genome, which was shown in both the electrophoresis and the alignment of various flanking sequences (Fig. 1c and Supplementary Fig. S2). Among these polymorphic insertions, two paired flanking sequences were seamlessly joined in the reference genome, revealing a recent transposition event in the genome of our NCMA cultures but not in the reference genome (Fig. 1d, upper). This insertion was confirmed by cloning the intact RAGL transposon from the NCMA culture but not from another Institute of Oceanology Chinese Academy of Sciences culture (Fig. 1e). Moreover, another pair of flanking sequences without TIRs and TSDs can be well aligned to the reference genome (Fig. 1d, lower), indicating the loss of some AanRAGL copies in the reference genome due to transpositional activity. Thus, as early as the unicellular eukaryotes (median origin time 1552 Myr) [5], active RAGL
Figure 1. Evolution of the RAG1 and Transib homologs. (a) Schematic diagram showing the genomic organization of RAG-like (RAGL) homologs in Aureococcus anophagefferens (pelagophyte), Fungia costulata (stony coral), Branchiostoma belcheri (lancelet) and Mus musculus (mouse).
transposons have appeared and preserved many conserved characteristics of RAG homologs. The divergent gene direction of the AanRAG1/2 and some single RAG1L fragments in other primitive eukaryotes imply that the original RAGL transposons may have experienced gene insertion or inversion. Before the discovery of ProtoRAG in lancelets, Transibs were found to be widely distributed in protostomes and cnidarians, and RAGL transposons were thought to be derived from a Transib transposon by acquisition of a RAG2L gene [6–10]. The recent identification of RAGL transposons in protostomes and cnidarians suggests another possibility: that Transib arose from a RAGL transposon by loss of RAG2L [4]. To conclusively clarify the evolutionary relationship between RAGL and Transib transposons, we searched for Transib homologs as performed in RAGL and identified dozens of potential Transib-like proteins in bacteria (42), fungi (20), plants (39), and ctenophores (6), and hundreds in other metazoans. Analysis of the Transib and RAGL1L proteins showed that the average protein identities among RAGL1L core, Transib, combined RAG1L core and Transib were 36.75%, 34.28% and 26.22%, respectively (Supplementary Fig. S3a). RAG1L and Transib proteins were phylogenetically clustered into two separate clans after setting the root ahead of the bacterial Transib branch (Fig. 1f and Supplementary Fig. S3b). The pelagophyte AanRAG1L was in the root of the RAG1 clan and the phylogenetic relationship of RAG1 homologs was generally consistent with their host species (Fig. 1f and Supplementary Fig. S4a). In addition, RAG1Ls were expanded into several diverged copies in some protostomes, echinoderms and hemichordates, and appeared to evolve slowly in chordates. Similar observations were obtained from the analyses of RAG2 homologs (Supplementary Fig. S4b). These results suggest that the evolution of the RAGL transposons was mainly in a vertical manner and probably accompanied by limited HGT.

For the Transib clan, two major subgroups were gathered and defined as Subgroups I and II (Fig. 1f and Supplementary Fig. S3b). The Subgroup I Transibs transposons all contain a pair of asymmetric TIRs and show a closer relationship with RAG1L than those of Subgroup II (Fig. 1f and g). However, the Subgroup II Transib transposons, which include most of the previously identified sequences, contain both asymmetric and symmetric TIRs (Fig. 1f and g). Unlike the vertical evolution of the RAGL transposons, the Transib transposons may have experienced massive HGT events.
as the phylogenetic relationships of Transib transposases were not consistent with that of their host species.

Close to the root of the phylogenetic tree, RAGIL and Transib superfamilies share a common ancestor that may have descended from bacterial Transib. Analyses of the composed domains in these ancient RAGIL and Transib proteins further reveal their conservation and diversification during evolution. First, the RAG2-binding region in the zinc-binding domain (ZnB) of RAG1L (α11–α12 in RAG1 core) was previously found to have a common ancestor that may have shared a common ancestor that may have been lost in Transibs [10]. Here, we found that the ancient bacterial and Subgroup I Transibs but not Subgroup II Transibs harbored an equivalent α11–α12 region (Fig. 1h), suggesting that the RAG2-binding region in ZnB was a constitutive region in the Transib ancestor but was exclusively lost in Subgroup II Transibs. Second, both RAGIL and Subgroup I Transibs contain the CTT1 domain (previously called CTT*, type I C-terminal tail), whereas a structurally different CTT domain (CTT2, type II C-terminal tail) was found in Subgroup II Transibs (Fig. 1h). The CTT1 domain has been shown to be critical for the lancelet RAGL complex to interact with its intact TIRs [2,9], whereas the CTT2 domain may be important for the interaction with the ZnB domain of Transibs [10]. Thus, the CTT1 in RAGILs should be inherited from a constitutive region in Transib ancestors, whereas the diverged CTT2 in Subgroup II Transibs may be an adaptation to the loss of the α11–α12 region in their ZnB domain. Third, similar to the diverged NBD* in RAGIL proteins, an equivalent nonamer binding domain (NBD) with several conserved positions (GRP in RAG1 NBD, Supplementary Fig. S5) was found in both Subgroup I and II Transibs, suggesting an equivalent NBD domain in the ancestors of RAGIL and Transib. In addition to these domains, several short regions in RAGILs and Transibs also experienced specific gain or loss, such as the gain of the loop region between β1–β3 in the pre-RNase H (PreRNH) domain of RAGILs and the loss of partial α1 and α17 in Subgroup II Transibs (Supplementary Fig. S5). Overall, Subgroup I Transibs preserved some ancestral regions shared by RAGILs, which were exclusively lost or diverged in Subgroup II Transibs. Subgroup I Transibs may represent intermediates in the early evolution of Transib and RAGIL.

Finally, we proposed an updated evolutionary model to elucidate the origin and evolution of the RAGL and Transib transposons (Fig. 1i). After descending directly from a bacterial Transib ancestor, the vertebrate RAG genes and their homologs mostly evolved in a vertical manner beginning in euukaryotes, whereas the Transibs experienced massive HGT. As intermediates for the early evolution of Transib and RAGIL, one of the Subgroup I Transibs acquired RAG2L early in the euukaryotes to generate the ancient RAGL transposon. This ancient RAGL then underwent host domestication in a vertical manner, including domain gain and loss, gene duplication, transposon fossilization and key amino acid adaptation, to finally shape the RAG machinery in vertebrates.

DATA AVAILABILITY

The representative sequences annotated during the current study are provided online in Supplementary Data 2. All other relevant datasets could be provided upon reasonable request.

SUPPLEMENTARY DATA

Supplementary data are available at NSR online.

ACKNOWLEDGEMENTS

We thank Dr. Hui Huang for providing the coral samples (F. costulata and F. tenius) and Professor Shichun Sun for providing the Notopterus guniculatus samples. We thank Professor Pierre Pontarotti for providing several RAGL sequences that he previously reported in echinoderms and hemichordates.

FUNDING

This work was supported by the National Natural Science Foundation of China (31270888 and 31970852 to S.Y., 32170934 and 31800740 to X.T.), the Ministry of Science and Technology of the People’s Republic of China (2018YFD0900502 to X.T., 2021YFA0909800 to X.T.), the Guangdong Science and Technology Department (2017B030314021 to S.Y. and A.X.), and the Innovation Group Project of Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) (311021006 to S.Y.).

AUTHOR CONTRIBUTIONS

X.X., T.T. and S.Y. conceived of the study. X.T. prepared all the sequences data, except some Transib sequences collected by T.Z., and X.T. conducted all the analyses work. X.T., Z.H., F.C. and X.W. performed the experiments cooperatively. X.T. drafted the manuscript. S.Y. and A.X. edited and approved the submitted manuscript.

Conflict of interest statement. None declared.

This work was supported by the National Natural Science Foundation of China (31270888 and 31970852 to S.Y., 32170934 and 31800740 to X.T.), the Guangdong Key Laboratory of Pharmaceutical Functional Genes, Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), School of Life Sciences, Sun Yat-sen University, China; 2School of Life Sciences, Beijing University of Chinese Medicine, China; 3Center for Infection and Immunity, School of Medicine, Sun Yat-sen University, China; 4Laboratory for Marine Biology and Biotechnology, Qingdao National Laboratory for Marine Science and Technology, China and 5Shanghai Institute of Immunology, College of Basic Medical Sciences, Shanghai Jiao Tong University School of Medicine, China
*Corresponding authors.
E-mails: lssxal@mail.sysu.edu.cn;
yuanshch@mail.sysu.edu.cn

REFERENCES

1. Huang S, Tao X and Yuan S et al. Cell 2016; 166: 102–14.
2. Tao X, Yuan S and Chen F et al. Nat Sci Rev 2020; 7: 403–17.
3. Morales Poole JR, Huang S and Xu A et al. ImmunoGeNetics 2017; 69: 391–400.
4. Martin EC, Vicari C and Tsakou-Ngouafo L et al. Mobile DNA 2020; 11: 17.
5. Heidgas SB, Marin J and Suleski M et al. Mol Biol Evol 2019; 32: 835–45.
6. Kapitonov VV and Jurka J. PloS Biol 2005; 3: e181.
7. Kapitonov VV and Koonin EV. Biol Direct 2015; 10: 20.
8. Carmona LM and Schatz DG. FEBS J 2017; 284: 1590–605.
9. Zhang Y, Ching TC and Huang G et al. Nature 2019; 569: 79–84.
10. Liu C, Yang Y and Schatz DG. Nature 2019; 575: 540–44.