Origins of the RAG transposome and the MHC

Origins of the RAG transposome and the MHC

Tsakou L(1), Paganini J(2), Kaufman J(3,4,5), Pontarotti P(1,6)

1. Aix Marseille University IRD, APHM, MEPHI, IHU Méditerranée Infection, Marseille France 3 19-21 Boulevard Jean Moulin 13005 Marseille, France.

2. Xegen 15 rue de la République, 13420 Gemenos, France.

3. University of Cambridge, Department of Pathology, Tennis Court Road, CB2 1QP, Cambridge, U. K.

4.University of Cambridge, Department of Veterinary Medicine, Madingley Road, CB2 0ES, Cambridge, U. K.

5. University of Edinburgh, Institute for Immunology and Infection Research, Charlotte Auerbach Road, EH9 3FL, Edinburgh, U. K.

6. SNC5039 CNRS, 19-21 boulevard Jean Moulin,13005 Marseilles, France.

Corresponding authors: Pierre Pontarotti, pierre.pontarotti@univ-amu.fr, Jim Kaufman, jim.kaufman@ed.ac.uk

Key Words: hairpin, flanking, DDE transposon excision, Artemis, palindromic diversity, convergent evolution

Abstract

The appearance of adaptive immunity in vertebrates remains unclear, although many proposals have been made. In this speculative review, we describe the complex innate immune systems in place before the emergence of the vertebrates, and propose the existence of a molecule(s) on the surface of some cells able to present pathogen-associated molecular patterns (PAMPs) to a specific receptor(s) on other cells, much like molecules of the major histocompatibility complex (MHC) and T cell receptors (TCRs). Crucially, an MHC-like molecule with a mutation allowing it to recognize a new PAMP would be unlikely to be recognized by the specific TCR-like molecule, and so there would be no selection for the new MHC-like molecule whose gene would then be lost by neutral drift. The integration of the recombination activating gene (RAG) transposon in a TCR-like gene would have led to a significant increase in the recognition possibilities, so that new MHC-like variants could be recognized and selected, along with the new RAG/TCR-like system. The eventual consequence of this scenario would be the ability of the MHC to present many peptides, through multigene families, polymorphism of individual genes and an increase in peptide-binding repertoire (promiscuity).

Presentation of the Hypotheses

At the start of the investigation of the vertebrate adaptive immune system, two aspects were particularly impressive: first, B- and T-cell repertoire diversity and the generation of this diversity by recombination and second, the enormous polymorphism of molecules encoded by the major histocompatibility complex (MHC) and that they bind numerous peptides. It has now become clear that this molecular system, largely involving immunoglobulin (Ig) domains, may have already been in place in the lineage leading to jawed vertebrates, including cartilaginous and bony fish, amphibians, reptiles, birds and mammals [1,2] The discovery of a parallel system in jawless fish, based on leucine-rich repeats (LRRs) rather than Ig domains, suggests that the cellular system for vertebrates is common to both jawless and jawed vertebrates [3-5] (see Box 1).

In this opinion piece, we propose several hypotheses that together can explain the emergence of the **recombination activating gene (**RAG)-based adaptive immune system in a jawed vertebrate ancestor as a consequence of the evolution of several linked biological traits, using the consequence of this co-opting of traits for a proposal on the origin of the MHC polymorphism. We first discuss the concept that a complex innate immune system may have existed long before the emergence of the vertebrate ancestor, including large multigene families able to recognize foreign pathogens, cell proliferation and immune memory after pathogen contact, and pathogen defense using **AID/APOBEC-like** (activation-induced deaminase/apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) cytidine deaminase genes. We also present arguments supporting the possible presence of clonal expansion and **allelic exclusion**, with each clone expressing a member of the multigene family and recognizing a pathogen-associated pattern

Next, we describe that the **RAG DDE transposon** was active in organisms from the ancestor of the **bilaterians** to the ancestor of jawed vertebrates (Figure 1), note that the RAG DDE transposon belongs to a functional transposon family that allows **palindromic (P) diversity** after

excision and DNA repair, and posit that the biochemical switch from the **RAG transposon** insertion and excision to the RAG sequence-specific recombination was a simple functional shift. These three properties would have increased the likelihood of RAG transposon being co-opted as a major player modulating the somatic diversity of the antibodies and T cell receptors (TCRs).

Finally, we propose that the somatic receptor diversity orchestrated by RAG allowed the emergence of the MHC peptide binding promiscuity and polymorphism. Many excellent papers and reviews have described and proposed hypotheses about the origin and the evolution of the adaptive immune system and the MHC, but here we focus on the origin of the **somatic diversification** and its consequence on the evolution of the MHC.

The origins of vertebrate adaptive immunity in metazoans

In common with the vertebrate adaptive immune systems, other **metazoans** can have large multigene families able to recognize foreign pathogens. There is also evidence for cell proliferation after pathogen contact and immune memory. In addition, clonal expansion and allelic exclusion of receptors are present in some metazoans, and AID/APOBEC-like enzymes are widely present.

For the first three points, some non-vertebrate metazoan genomes display large multigene families involved in innate immunity, including those based on LRRs, such as toll-like receptors (TLRs) and other pathogen recognition receptors (PRRs) likely to recognize pathogen-associated molecular patterns (PAMPS), and those based on Ig-like domains, such as IgV-IgC receptors likely involved in natural killer activity [6,7]. Second, PAMP activation gives rise to cell activation in metazoans [4], but there are reports that PAMP activation gives rise to immune system cell proliferation [8-10]. Third, numerous studies have demonstrated various forms of immune memory in many non-vertebrate metazoans [for review, see ref. 11]. Although the

evidence is fragmentary, the existence of even a few examples shows that these biological traits exist outside of vertebrates and may have provided the basis for the vertebrate adaptive immune system.

For the last two points, clonal expression of receptors and allelic exclusion are common mechanisms in eukaryotes rather than mechanisms limited to the vertebrate adaptive immune system, like multigene family of olfactory receptors and/or the antigenic variation of variable surface glycoproteins (VSGs) in trypanosomes [12,13]. AID/APOBEC enzymes have several functions in vertebrates [14] including generating diversity of non-self recognition, producing point mutations (for instance, B-cell receptors in jawed vertebrates) and driving gene conversion mechanisms by DNA breakage followed by repair mechanisms that increase the probability of gene conversion in cyclostomes and some invertebrates [15,16]. Orthologues of this family are also found with similar activities in deuterostomes, and the AID/APOBEC-like cytidine deaminase is expressed preferentially in tissues undergoing constant direct interaction with potential pathogens, can be induced upon pathogen challenge and is involved in innate immunity acting on non-self-DNA [17,18].

Thus, in the pre-adaptive immune system, multigene families of PRRs and IgV-IgC receptors could have recognized PAMPs leading to cellular activation and proliferation, and immune memory. The generation of diversity for these multigene families could be driven by members of the AID/APOBEC family [as proposed by ref. 17], first involved in non-self-recognition with one family member co-opted during vertebrate evolution by shifting the mutagenic activity from non-self to self. The mechanisms for clonal expression and allelic exclusion would lead to each clone expressing a single member of the multigenic family recognizing particular PAMPs.

The next step: emergence of diversified receptors

As described above, two adaptive immune systems are found in vertebrates (see Box 1). In considering the origins of the adaptive immune system of jawless vertebrates, two potentially ancestral genes are found in various metazoans and could have given rise to the diversified **variable lymphocyte receptors** (VLRs): many proteins with LRR domains, most particularly the **toll-like receptors** (TLRs), and the AID/APOBEC-like enzymes. In contrast, the emergence of the adaptive immune system of jawed vertebrates is less clear, with plausible candidates for the receptors in metazoans but rather complex in terms of the generation of diversity (see Box 2). Antibody and TCR genes of jawed vertebrates are based on Ig domains assembled from separate variable (V), diversity (D) and joining (J) gene segments during B and T lymphocyte development to give contiguous VJ and VDJ sequences. The process is initiated by the RAG endonuclease involved in excision of DNA between the gene segments and continues by ubiquitously-expressed DNA repair enzymes (see Box 3). The appearance of RAG has long been considered a key evolutionary step that can explain the origin of the jawed vertebrate adaptive system [19,20].

RAG origin

The discovery of recombination signal sequences (RSSs) flanking the V, D and J gene segments, along with the mechanism of RSS cleavage which is similar to several cut-and-paste DNA transposases (DDE transposases) [20-23], resulted in the hypothesis (see Box 3) that a DDE transposon inserted into an Ig-like gene, leading eventually to antibody/TCR gene rearrangement [19].

The experimental analyses of the RAG transposon from amphioxus (a chordate from the sister group of vertebrates, see Figure 1) which has no known adaptive immune system shed light on the functional shift from a RAG transposon to the RAG sequence-specific recombination

activating system. First, the excision reaction is similar for the two endonucleases: the transposase recognizes **terminal inverted repeat** (TIR) sequences, and the co-opted endonuclease (RAG) recognizes TIR-like sequences (that is, the RSSs) [24]. Both involve a nick-hairpin mechanism characteristic of several DDE DNA transposases, including RAG/Transib (with Transib having only the RAG1 core, which is the endonuclease), HAT and Mutator [25-28]. After excision, the hairpin-tipped segments are processed by the evolutionarily conserved endonuclease Artemis, performing an asymmetric opening of hairpin and leading to palindromic P nucleotide variation (see Box 3) [24,29]. Other non-vertebrate species also have a RAG transposon that is likely to work in the similar manner as in amphioxus (Box 4 and below).

It should be noted that Artemis and all proteins involved in **non-homologous end joining** (NHEJ, a ubiquitous DNA repair pathway) are present in all metazoans [30], and that homologs performing a similar function are present in all eukaryotes, including PSO2 in yeast [31]. Thus, co-option of RAG is not just co-option of the transposon, but co-option of a whole system of transposition which includes the cellular proteins that the transposon interacts with to perform the transposition. In this view, the RAG transposome includes the DDE transposon (transposase /TIR), the Artemis nuclease and the cellular NHEJ enzymatic machinery.

There are differences between the RAG transposome (dependent on the RAG transposon, a piece of selfish DNA) and the RAG system (which has been “domesticated” for a useful function in the organism). One major difference is at the level of the flanking fragment, in which **terminal deoxytransferase** (TdT) adds N-nucleotides to the V, D, and J segments of the TCR and BCR genes during gene recombination, increasing **junctional diversity**. The TdT gene has a long phylogenetic history (P. Pontarotti, unpublished data), so it seems clear that the domesticated RAG system co-opted TdT. A second major difference is at the level of the excised fragment flanked by RSSs or TIRs. The domesticated RAG actively directs cleaved signal and coding ends into the NHEJ repair pathway for signal- and coding-joint formation. In contrast, the RAG

transposon strongly favors transposition, but allows some TIR-TIR joint formation [24,32,33]. It is possible that the ancestral transposase partially prevented the interaction between the TIR and the NHEJ repair pathway, and that the RAG in jawed vertebrates lost this property, although this remains unknown. In vitro approaches to study the mechanism revealed important amino acid positions in the RAG proteins involved in suppressing transposition [33], which is important to avoid harmful effects for the organism.

The biochemical functions of the DDE transposome and the vertebrate RAG system (a sequence-specific recombination activating system are similar; hence the biochemical shift from a transposome to a sequence-specific recombination activating system seems to constitute a relatively straightforward evolutionary step [34]. This idea is supported by the fact that many other DDE transposomes have been co-opted as sequence specific recombination activating systems [34], including Piggymac/TPB1/TPB2/TPB6 in ciliates [35,36], Kat 1 in yeast [37] and MATalpha3 in yeast [38].

The vertical evolution of the RAG transposon and the origin of RAG

From the concepts presented above, any DDE transposon capable of creating a hairpin in the region flanking the excised fragment could have been co-opted as RAG, since such DDE transposons are able to generate the P nucleotides involved in the generation of diversity [39]. One might wonder what the advantage of the RAG transposon might be, compared to these other transposons. The answer could come from the different evolutionary behavior of these transposons.

Phylogenetic analysis has been performed on hairpin-forming DDE transposons: HAT [40], Mutator [41], Transib [42] and other DDE transposons [43-47]. Such phylogenetic studies show that these DDE transposons have apparently evolved in a horizontal manner, which contrasts with the transposon RAG that evolved in a vertical manner. In contrast, the phylogenetic analysis of

RAG transposon and vertebrate RAG sequences, as well as sequences belonging to the RAG family with unknown status and fossilized RAG transposons, shows a sequence tree topology following the species phylogenetic tree [48,49]. The phylogenetic reconstruction also indicates that the RAG structure appeared at least at the origin of the bilaterians (animals including protostomes, deuterostomes and a few other groups, Figure 1). Therefore, the RAG transposon appears to have been active since its birth in the ancestor of the bilaterians and was co-opted as a specific endonuclease in the jawed vertebrate ancestor. The presence of the RAG transposon that was inherited in the genome from one generation to the next increased the likelihood that it would be co-opted compared to the other transposons that evolve(d) by horizontal transmission between individuals.

Horizontal transfer of DDE transposons may allow these transposable elements to enter naïve genomes which they invade by making copies of themselves and then escape before they become fully silenced by the **Piwi-piRNA pathway**, which is a host mechanism against transposable elements [50,51]. The RAG transposon is able to transpose within a genome (Huang et al., 2016, Morales Poole et al., 2017) [24,48], but to our knowledge, not between genomes of divergent species. Therefore, on the one hand, the RAG transposon seems to have lost the ability to transpose between species, and on the other hand, the RAG transposon seems to have evolved a mechanism to escape the Piwi-piRNA system of the host.

In this context, it should be noted that only one of the two subunits encoded by the RAG transposon comes from a transposon, while the other seems to have a host origin. The RAG1 subunit corresponds to the DDE transposase highly related to the **transib** (present in several protostomes), while the RAG2 in the RAG transposon came from a host genome [52,53]. Several sequence similarity analyses propose that a RAG-like open reading frame flanked by RSS-like TIRs captured a RAG2-like open reading frame of an ancestral protostome to give rise to the original RAG transposon [7,32,54]. Thus, the transposon domesticated a part of the host genome,

perhaps to evade the Piwi-piRNA of the host and avoid inactivation. However, it is also possible that the transposon was retained for an unknown reason, perhaps including another function for the host.

Consequently, we propose the following conjectural scenario to enhance the published model [26]: i) some time ago, there was an insertion of a complete RAG transposon (or possibly the corresponding **miniature inverted-repeat transposable element** (MITE, corresponding to the TIR of the RAG transposon)) that separated an IgV domain (already involved in immune recognition) into V and J segments; ii) after the insertion of the complete transposon, the transposase was lost, leaving the native TIRs between the V and J segments intact, while a transposase from another RAG transposon was used, and which in turn, lost its TIR; iii) The TIR-like sequence could be recognized by the RAG transposase and excised along with the internal sequence, leaving hairpin-tipped ends on the flanking segments. These segments could be processed via Artemis opening the hairpins asymmetrically followed by the DNA repair system leading to palindromic (P) diversity. The ability to generate diversity increased with the duplication of the VJ unit (V-TIR-TIR-J) and the co-option of a TdT gene. The system later became more complex, as described by others [55].

It should be noted that the transposon and its corresponding MITE had hundreds of millions of years to be inserted anywhere in the genome of many protostome lineages. Some of these events were likely to have been negatively selected, some were neutral, and it is possible that the insertion into a genetic system already involved in non-self-recognition was positively selected. We estimate the probability of a RAG transposon insertion in an ancestral V domain to give rise to a bona fide V-J module in some metazoans to be 99% (see table S1).

A third step: antibody/TCR receptor somatic diversity could drive the appearance of MHC promiscuity and polymorphisms

The classical class I/II genes of the MHC are highly polymorphic, encoding proteins that bind processed peptides within the cell, move to the cell surface and then interact with TCRs expressed on the surface of T-cells. Each MHC allelic form can bind many peptides, both self and non-self, with a specific amino acid motif. Most developing T-cells with TCRs that react with self-MHC molecules bound to self-peptides are eliminated during maturation in the thymus. During infection, both self and non-self-peptides are presented by MHC proteins, with non-self-peptides recognized by TCRs on T-cells, which activate the immune system to respond in a variety of ways. These MHC genes evolved in the ancestor of jawed vertebrates in roughly the same time window as the RAG/VDJ generation of somatic diversity [1,56]. Various hypotheses have been proposed for the origin of MHC genes (see Box 4). In this speculative review, we propose the scenario that the MHC evolved from **pathogen recognition receptors** (PRRs) from the innate immune system.

The first part of our hypothesis is that the ancestral MHC-like molecule could have bound some **pathogen associated molecular patterns** (PAMPs), presenting them to ancestral TCR-like molecules. The ancestral MHC-like molecule may have been limited to just a few pathogens, and each ancestral TCR-like molecule may have only recognized a particular class of PAMP bound to the ancestral MHC-like molecule. Thus, if a mutation of the ancestral MHC-like molecule allowed binding of a new PAMP, this combination might not be recognized by the ancestral BCR/TCR-like molecules (even if they were encoded by a multigene family); therefore, the new MHC-like molecule might not be selected and the mutant gene could be lost by **genetic drift**. In fact, if the new MHC-like molecule lost binding to the original PAMPs, it might be negatively selected.

In the second part of this hypothesis, the integration of the RAG transposon into ancestral BCR/TCR-like genes may have led to a significantly increased possibility of recognition; we will focus here only on TCRs as they interact with the MHC. As a result of the increased possibilities of recognition by the TCRs, mutations in the ancestral MHC-like molecule leading to the binding of new PAMPs could have been recognized by the TCRs and therefore been selected. Presumably, this expanded ability of this ancient MHC/TCR system to recognize new PAMPs would have eventually allowed peptides to be bound, presented and recognized during an immune response.

As a third part of this hypothesis, we posit that the ancient MHC molecule was selected to bind many peptides to allow the recognition of numerous pathogens, possibly via the appearance of allelic polymorphisms and peptide-binding promiscuity (as well as from the generation of multigene families). Both allelic polymorphisms and promiscuity are properties of MHC molecules encoded by a single gene, and both extend the number of peptides that can be bound, and thus, the number of pathogens that can be recognized [57-59]. If a particular MHC molecule only bound a limited number of peptides, then a new pathogen would not be recognized by the MHC/TCR system unless a mutation occurred in MHC genes; thus, such a mutation would be selected to deal with the new pathogen. However, the mutation might prevent the new molecule from binding the previously-bound peptides, so that the host would be vulnerable to the original pathogen still in the environment. In order to deal with both old and newly-arising pathogens, pathogen-mediated selection leads to allelic polymorphism [57-59].

Another way to increase the ability to recognize new pathogens would be to increase the range of peptides bound, and such promiscuity can be an important feature of MHC molecules [59-61]. A third way to increase recognition of new pathogens would be to increase the number of MHC genes, but the need to avoid recognition of self-peptides might limit the size of the MHC multigene family (although there are theoretical arguments to the contrary) [62-64].

Concluding Remarks

We propose a model whereby the ancestral MHC-like molecule had an innate immune function, but when ancestral TCR-like molecules began to diversify due to RAG domestication and thus increase their recognition potential, ancient MHC molecules might have increased their peptide-binding capacity through increased promiscuity. However, the peptide-binding capacity may have been still low compared to the recognition capacity of the TCR; therefore, allelic polymorphism may have evolved via pathogen-mediated selection. As this hypothesis begins with the recognition of PAMPs, for which LRR-containing molecules such as TLRs are major players, a similar scenario might be envisaged for the VLR system based on LRRs. Thus far, no equivalent of an MHC molecule in cyclostomes has been reported (Box 1), but some analogous molecule might be expected based on this model (see outstanding questions).

Transposable elements are usually considered to be egotistical pieces of DNA, although there is much research on their potential utility for the host organisms. The case of the RAG transposon is particularly spectacular: a small piece of DNA that has completely changed immunity in jawed vertebrates and indeed, the research work of many if not most immunologists (including the authors of this opinion article). It will be exciting to discover which other accidents of evolution have led to such enormous consequences.

References

1. Flajnik, M.F. and Kasahara, M. (2010) Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat. Rev. Genet.* 11, 47-59

2. Kaufman, J. (2018) Unfinished Business: Evolution of the MHC and the Adaptive Immune System of Jawed Vertebrates. *Annu. Rev. Immunol.* 36, 383-409

3. Pancer, Z. and Cooper, M.D. (2006) The evolution of adaptive immunity. *Annu. Rev. Immunol.* 24, 497-518

4. Boehm, T. *et al.* (2018) Evolution of Alternative Adaptive Immune Systems in Vertebrates. *Annu. Rev. Immunol.* 36, 19-42

5. Flajnik, M.F. (2018) A Convergent Immunological Holy Trinity of Adaptive Immunity in Lampreys: Discovery of the Variable Lymphocyte Receptors. *J. Immunol*. 201, 1331-1335

6. Buckley, K.M. and Rast, J.P. (2015) Diversity of animal immune receptors and the origins of recognition complexity in the deuterostomes. *Dev. Comp. Immunol.* 49, 179-189

7. Litman, G.W. *et al.* (2010) The origins of vertebrate adaptive immunity. *Nat. Rev. Immunol.* 10, 543–553

8. Homa J. *et al.* (2013) Exposure to immunostimulants induces changes in activity and proliferation of coelomocytes of *Eisenia andrei*. *J. Comp. Physiol. B.* 183, 313-322

9. Holm, K. *et al.* (2008) Induced cell proliferation in putative haematopoietic tissues of the sea star, *Asterias rubens* (L.). *J. Exp. Biol.* 211, 2551-2558

10. Salamat, Z. and Sullivan, J.T. (2009) Involvement of protein kinase C signalling and mitogen-activated protein kinase in the amebocyte-producing organ of *Biomphalaria glabrata* (Mollusca). *Dev. Comp. Immunol*. 33, 725–727

11. Milutinović, B. and Kurtz, J. (2016) Immune memory in invertebrates. *Semin. Immunol*. 28, 328-342

12. Glover, L. *et al.* (2016) VEX1 controls the allelic exclusion required for antigenic variation in trypanosomes. *Proc. Natl. Acad. Sci. USA*. 113, 7225–7230

13. Monahan, K. and Lomvardas, S. (2015) Monoallelic expression of olfactory receptors. *Annu. Rev. Cell Dev. Biol.* 31, 721-740

14. Conticello, S.G. (2008) The AID/APOBEC family of nucleic acid mutators. *Genome Biol*. 9, 229

15. Arakawa, H. *et al.* (2002) Requirement of the activation-induced deaminase (AID) gene for immunoglobulin gene conversion. *Science* 295, 1301-1306

16. Rogozin, I.B. *et al.* (2007) Evolution and diversification of lamprey antigen receptors: evidence for involvement of an AID-APOBEC family cytosine deaminase. *Nat. Immunol.* 8, 647–656

17. Liu, M.C. *et al.* (2018) Diversification of AID/APOBEC-like deaminases in metazoa: multiplicity of clades and widespread roles in immunity. *Nat. Comm.* 9, 1948

18. Krishnan, A. *et al.* (2018) Diversification of AID/APOBEC-like deaminases in metazoa: multiplicity of clades and widespread roles in immunity. *Proc. Natl. Acad. Sci. U S A.* 115, E3201-E3210

19. Thompson, C.B. (1995) New insights into V(D)J recombination and its role in the evolution of the immune system. *Immunity* 3, 531-539

20. Sakano, H. *et al.* (1979) Sequences at the somatic recombination sites of immunoglobulin light-chain genes. *Nature* 280, 288-294

21. McBlane, J.F. *et al.* (1995) Cleavage at a V(D)J recombination signal requires only RAG1 and RAG2 proteins and occurs in two steps. *Cell* 83, 387-395

22. Agrawal, A. *et al.* (1998) Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* 394, 744-751

23. Hiom, K. *et al.* (1998) DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. *Cell* 94, 463-470

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

25. Lafaille, J.J. *et al.* (1989) Junctional sequences of T cell receptor gamma delta genes: implications for gamma delta T cell lineages and for a novel intermediate of V-(D)-J joining. *Cell* 59, 859-870

26. Fugmann, S.D. (2010) The origins of the Rag genes--from transposition to V(D)J recombination. *Semin. Immunol.* 22, 10-16

27. Liu, K. and Wessler S.R. (2017) Transposition of Mutator-like transposable elements (MULEs) resembles hAT and Transib elements and V(D)J recombination. *Nucleic Acids Res.* 45, 6644-6655

28. Hickman, A.B. *et al.* (2018) Structural insights into the mechanism of double strand

break formation by Hermes, a hAT family eukaryotic DNA transposase. *Nucleic Acids Res.* 46, 10286-10301

29. Colot, V. *et al.* (1998) Extensive, nonrandom diversity of excision footprints generated by Ds-like transposon Ascot-1 suggests new parallels with V(D)J recombination. *Mol. Cell Biol.* 18, 4337-4346

30. Bonatto, D. *et al.* (2005) In silico identification and analysis of new Artemis/Artemis-like sequences from fungal and metazoan species. *Protein J.* 24, 399-411

31. Tiefenbach, T. and Junop, M. (2011) Pso2 (SNM1) is a DNA structure-specific endonuclease. *Nucleic Acids Res.* 40, 2131-2139

32. Carmona, L.M. and Schatz, D.G. (2017) New insights into the evolutionary origins of the recombination-activating gene proteins and V(D)J recombination. *FEBS J.* 284, 1590-1605

33. Zhang, Y. *et al.* Transposon molecular domestication and the evolution of the RAG recombinase. *Nature* 569, 79-84

34. Tsakou, L. *et al.* (2020) DDE transposon as public goods. *Evolutionary Biology.* Eds., Pierre Pontarotti, Springer International Publishing, in press

35. Baudry, C. *et al.* PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. *Genes Dev.* 23, 2478-2483

36. Cheng, C.Y. *et al.* (2016) The piggyBac transposon-derived genes TPB1 and TPB6 mediate essential transposon-like excision during the developmental rearrangement of key genes in *Tetrahymena thermophila*. *Genes Dev.* 30, 2724-2736.

37. Rajaei, N. *et al.* (2014) Domesticated transposase Kat1 and its fossil imprints induce sexual differentiation in yeast. *Proc. Natl. Acad. Sci. U S A.* 111, 15491-15496

38. Barsoum, E. *et al.* (2010) Alpha3, a transposable element that promotes host sexual reproduction. *Genes Dev.* 24, 33–44

39. Lu, H. *et al.* (2007) Extent to which hairpin opening by the Artemis:DNA-Pkcs complex can contribute to junctional diversity in V(D)J recombination. *Nucleic Acids Res.* 35, 6917-6923

40. Arensburger, P. *et al.* (2011) Phylogenetic and functional characterization of the hAT transposon superfamily. *Genetics* 188, 45-57

41. Dupeyron, M. *et al.* (2019) Evolution of Mutator transposable elements across eukaryotic diversity. *Mob DNA* 10, 12

42. Hencken, C.G. *et al.* (2012) Functional characterization of an active Rag-like transposase. Nat *Struct. Mol. Biol.* 19, 834-836

43. Wallau, G.L. *et al.* (2012) Horizontal transposon transfer in eukarya: detection, bias, and perspectives. *Genome Biol. Evol.* 4, 689-699

44. Dotto BR *et al.* (2018) HTT-DB: new features and updates. *Database (Oxford*) 1

45. Joly-Lopez, Z. *et al.* (2016) Phylogenetic and Genomic Analyses Resolve the Origin of Important Plant Genes Derived from Transposable Elements. *Mol. Biol. Evol.* 33, 1937-1956

46. Bouallègue, M. *et al.* (2017) Molecular Evolution of piggyBac Superfamily: from Selfishness to Domestication. *Genome Biol. Evol.* 9, 323-339

47. Peccoud, J. *et al.* (2017) Massive horizontal transfer of transposable elements in insects. *Proc. Natl. Acad. Sci. USA.* 114, 4721-4726

48. Morales Poole, J.R. *et al.* (2017) The RAG transposon is active through the deuterostome evolution and domesticated in jawed vertebrates. *Immunogenetics* 69, 391-400

49. Martin, E.C. *et al.* (2020) Evidence for an ancient bilaterian origin of the RAG-like transposon. *Mobile Elements*, in press

50. Aravin, A.A. *et al.* (2007) The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* 318, 761-764

51. Lerat, E. *et al.* TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. *Nucleic Acids Res.* 45, e17

52. Kapitonov, V.V. and Jurka, J. (2005) RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol*. 3, e181

53. Kapitonov, V.V. and Koonin E.V. (2015) Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. *Biol. Direct*. 10, 20

54. Callebaut, I. and Mornon, J.P. (1998) The V(D)J recombination activating protein RAG2 consists of a six-bladed propeller and a PHD fingerlike domain, as revealed by sequence analysis. *Cell. Mol. Life Sci.* 54, 880-891

55. Hsu, E., and Lewis, S.M. (2015) The origin of V(D)J diversification. In: Molecular Biology of B cells, Alt, F.W., Honjo, T., Radbruch, A., and Reth, M., eds. Elsevier, Academic Press, Amsterdam, pp. 133-148.

56. Danchin E. *et al.* (2004) The major histocompatibility complex origin. *Immunol. Rev.* 198, 216-232

57. Spurgin, L.G. and Richardson, D.S. (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc. Biol. Sci.* 277, 979–988

58. Radwan, J. *et al.* (2020) Advances in the evolutionary understanding of MHC polymorphism. *Trends. Genet.* 36, 298-311

59. Kaufman, J. (2018) Generalists and Specialists: A New View of How MHC Class I Molecules Fight Infectious Pathogens. *Trends Immunol*. 39, 367–379

60. Chappell, P. *et al.* (2015) Expression levels of MHC class I molecules are inversely correlated with promiscuity of peptide binding. *eLIFE* 4, e05345

61. Manczinger, M. *et al.* (2019) Pathogen diversity drives the evolution of generalist MHC-II alleles in human populations. *PLoS Biol.* 17, e3000131

62. Vidović, D. and Matzinger, P. (1988) Unresponsiveness to a foreign antigen can be caused by self-tolerance. *Nature* 336, 222-225

63. Nowak, M.A. *et al.* (1992) The optimal number of major histocompatibility complex molecules in an individual. *Proc. Natl. Acad. Sci. U S A.* 89, 10896-10899

64. Borghans, J.A. *et al.* (2003) Thymic selection does not limit the individual MHC diversity. *Eur. J. Immunol.* 33, 3353-3358

65. Teng, G. and Schatz, D.G. (2015) Regulation and Evolution of the RAG Recombinase. *Adv. Immunol.* 128, 1-39

66. Cardarelli, L. *et al.* (2015) Two Proteins Form a Heteromeric Bacterial Self-Recognition

443 Complex in Which Variable Subdomains Determine Allele-Restricted Binding. *Mbio.* 6,
444 e00251

445 67. De Tomaso A.W. (2018) Allorecognition and Stem Cell Parasitism: A Tale of
446 Competition, Selfish Genes and Greenbeards in a Basal Chordate. In: Pontarotti P. (eds)
447 Origin and Evolution of Biodiversity. Springer.

448 68. Espinosa, A. and Paz-Y-Miño-C, G. (2014) Evidence of Taxa-, Clone-, and Kin-
449 discrimination in Protists: Ecological and Evolutionary Implications. *Evol. Ecol.* 28,
450 1019-1029

451 69. Fujii, S. *et al.* (2016) Non-self- and self-recognition models in plant self-incompatibility.
452 *Nat. Plants* 2, 16130

453 70. Gruenheit, N. *et al.* (2017) A polychromatic 'greenbeard' locus determines patterns of
454 cooperation in a social amoeba. *Nat. Comm.* 8, 14171

455 71. Harada, Y. *et al.* (2008) Mechanism of self-sterility in a hermaphroditic chordate. *Science*
456 320, 548-550

457 72. Heller, J. *et al.* (2016) Greenbeard Genes Involved in Long-Distance Kind Discrimination
458 in a Microbial Eukaryote. *PLoS Biol* 14, e1002431

459 73. Kües, U. (2015) From two to many: Multiple mating types in Basidiomycetes. *Fungal*
460 *Biology Reviews* 29, 126-166

461 74. Paoletti, M. (2016) Vegetative incompatibility in fungi: From recognition to cell death,
462 whatever does the trick. *Fungal Biology Reviews* 30, 152-162

463 75. Rosengarten, R.D. and Nicotra, M.L. (2011) Model systems of invertebrate

allorecognition. *Curr. Biol.* 21, R82-92

76. Saak, C.C. and Gibbs, K.A. (2016) The Self-Identity Protein IdsD Is Communicated between Cells in Swarming Proteus mirabilis Colonies. *J. Bacteriol.* 198, 3278-3286

77. Wall, D. (2016) Kin Recognition in Bacteria. *Annu. Rev. Microbiol.* 70, 143-160

78. Boehm ,T. (2006) Quality control in self/nonself discrimination. *Cell* 125, 845-858

79. Ruff, J.S. *et al.* (2012) MHC signaling during social communication. *Adv. Exp. Med. Biol.* 738, 290-313

80. Leinders-Zufall, T. *et al.* (2009) Structural requirements for the activation of vomeronasal sensory neurons by MHC peptides. *Nat. Neurosci.* 12, 1551-1558

81. Flajnik, M.F. *et al.* (1991) Which came first, MHC class I or class II? *Immunogenetics* 33, 295-300

82. Credle, J.J. *et al.* (2005) On the mechanism of sensing unfolded protein in the endoplasmic reticulum. *Proc. Natl. Acad. Sci. USA.* 102, 18773-18784

83. Karagöz, G.R.E. *et al.* (2017) An unfolded protein-induced conformational switch activates mammalian IRE1. *eLIFE* 6, e30700

84. Dijkstra, J.M. and Yamaguchi. T. (2019) Ancient features of the MHC class II presentation pathway, and a model for the possible origin of MHC molecules. *Immunogenetics* 71, 233-249

85. Du Pasquier, L. (2000) The phylogenetic origin of antigen-specific receptors. *Curr. Top. Microbiol. Immunol.* 248, 160–185

484 86. Kaufman, J.F. *et al.* (1984) The class II molecules of the human and murine major
485 histocompatibility complex. *Cell* 36, 1-13

486 87. Xiao, J. *et al.* (2018) An Invariant Arginine in Common with MHC Class II Allows
487 Extension at the C-Terminal End of Peptides Bound to Chicken MHC Class I. *J.*
488 *Immunol.* 201, 3084-3095

489 88. Janeway, C.A. *et al.* (2001) Immunobiology: the immune system in health and disease. 5th
490 edition. New York: Garland Science

491

Acknowledgements

This work was supported by the French Government under the «Investissements d'avenir» (Investments for the Future) program managed by the Agence Nationale de la Recherche (ANR, fr: National Agency for Research), (reference: Méditerranée Infection 10-IAHU-03) to P. P., and by an Investigator Award from the Wellcome Trust to J. K. (110106/Z/15/Z).

Figure legends

Figure 1. Phylogenetic distribution: RAG in jawed vertebrates and the RAG-like transposon. On the consensus bilaterian tree is shown the presence of RAG-like transposons [24,26,49,52,53] and RAG among clades sequenced in the databases. The comparative activity of RAG in V(D)J recombination among the jawed vertebrates and the activity of RAG-like transposons is adapted [24,88] showing that this biochemical switch would constitute an unconstrained evolutionary step.

Figure 2. **Antibody/TCR receptor somatic diversity might drive the appearance of MHC promiscuity and polymorphisms.** We propose a hypothetical model whereby the ancestral MHC-like molecule bound certain PAMPs, presenting them to ancestral TCR-like molecules. The ancestral MHC-like molecule may have been limited to just a few pathogens, and each ancestral TCR-like molecule might have only recognized a particular class of PAMP bound to the ancestral MHC-like molecule.

A mutation in the ancestral MHC-like molecule may have allowed binding of a new PAMP, but this new combination could not be recognized by the ancestral TCR-like molecule. As a result, the new MHC-like molecule would be lost by genetic drift.

The integration of the RAG transposon into an ancestral TCR-like gene may have led to a significantly increased probability of recognition by these original, non-diverse TCRs. As a result, mutations in the ancestral MHC-like molecule may have led to a conformational ability to bind new PAMPs; as a consequence, mutated MHC molecules could have then been recognized by diverse TCRs thereafter, becoming evolutionarily selected via three mechanisms peptide-binding promiscuity allelic polymorphism and as by the expansion into multigene families. The expanded ability of this ancient MHC/TCR system to recognize new PAMPs would presumably allow peptides to be bound, presented and recognized during an immune response.

Box 1.

Brief overview of the adaptive immune system in vertebrates

The jawed vertebrate immune system is based on a complex cellular system made of T-cells and B-cells with immunoglobulin (Ig) domain-containing receptors and/or secreted proteins, including antibodies and both kinds of T-cell receptors (TCRs), those composed of α and β chains, and those composed of γ and δ chains. The generation of antigen receptor diversity is driven by the recombination activating genes, RAG1 and RAG2. Each unique receptor is expressed by a different cell clone through the action of allelic exclusion. In jawless fish (agnatha or cyclostomes), the other living vertebrate phylum, the receptors are based on the leucine-rich repeat (LRR) module, and include variable lymphocyte receptor-A (VLR-A), VLR-B and VLR-C. The diversity generation occurs via **gene conversion** driven by a protein of the AID-APOBEC family, but again, unique receptors are expressed by different clones with transcriptomic profiles much like jawed vertebrate lymphocytes: VLR-A like αβ T-cells, VLR-B like B-cells and VLR-C like γδ T-cells [4,32]. In jawed vertebrates, γδ-cells bind various cell surface molecules, but αβ TCRs recognize peptides bound specifically to MHC molecules; whether there is a functional equivalent of MHC molecules in jawless fish remains unclear.

Box 2

The next step: evolution of two systems of adaptive immunity in vertebrates

An important question concerns the origin of the complexity of cells involved in adaptive immunity. Both molecular systems with somatic diversification (VLR/AID and VDJ/RAG) could have been in place along with a pre-adaptive immune system [2,4,5]. Then the two molecular systems might have evolved in an independent manner in the two vertebrate lineages, jawless fish and jawed vertebrates. The mechanism of diversity generation is similar in both vertebrate lineages, starting with a DNA double-strand break (DSB) in the region involved in DNA recognition, followed by gene repair from either **non-homologous end-joining** (NHEJ) mechanisms or **gene conversion** [15,16]. The DSB in cyclostomes (and some jawed vertebrates) is due to an enzyme of the AID/APOBEC family and repair by gene conversion events, while the DSB in most jawed vertebrates is due to the RAG sequence-specific endonuclease and followed by DNA repair through a NHEJ mechanism.

It is important to note that the function of possible T- and B-cell lineages before the adaptive immunity arose is entirely unclear. **Innate lymphoid cells** (ILCs) found in mammals are potential candidates for the functions of non-adaptive T cells before adaptive immunity (although they could also be a novelty of placental mammals), but system replacement might be more likely [48]. If the first adaptive immune system was based on VLR, then in jawed vertebrates, a shift occurred from the IgV-IgC innate immunity to the IgV-IgC adaptive immunity, followed by the loss of the VLR-based adaptive immunity. If the first adaptive immune system was based on IgV-IgC, then in cyclostomes the reverse may have occurred. In fact, such replacements have been noted for natural killer (NK) cell receptors [2]: at least three families of NK cell receptors exist with analogous functions: lectin-like receptors (overwhelmingly in rodents and to a lesser extent in certain other mammals), Ig-like receptors of the KIR family (one or another of the KIR sub-

565 families, as in humans and other mammals) and a completely different family of Ig-like receptors
566 in bony fish.

Box 3

Emergence of rearranging B- and T-cell receptors and Brief history of the origin of RAG

The antibody and TCR genes of jawed vertebrates are assembled from variable (V), diversity (D), and joining (J) gene segments during B- and T-lymphocyte development to give contiguous VJ and VDJ sequences. The process to excise the DNA between the gene segments is initiated by the RAG endonuclease. The RAG endonuclease specifically recognizes recombination signal sequences (RSSs) that flank each gene segment. RSSs are composed of conserved heptamer and nonamer sequences separated by a less conserved spacer sequence of either 12 or 23 bp (12RSS and 23RSS). RAG-mediated DNA cleavage occurs preferentially in a complex containing a 12RSS and a 23RSS, involving a nick-hairpin mechanism.

After cleavage, the hairpin-tipped coding segments are cut by the Artemis endonuclease, joined imprecisely by the repair cell machinery to form a coding joint (CJ). The imprecise joins are due to the palindromic (P) diversity (due to Artemis), nucleotide deletion diversity and nucleotide (N) diversity (due to the terminal deoxynucleotidyl transferase, TdT), while the cleaved RSSs (and eliminated DNA segments) are joined precisely to form a signal joint (SJ). End-processing and joining are carried out by the NHEJ DNA repair pathway [for complete review, see ref 65].

The discovery of RSSs, along with the mechanism of RSS cleavage which is similar to several cut-and-paste DNA transposases (DDE transposases) [20,21] resulted in the hypothesis that a DDE transposon invaded an Ig-like gene, leading eventually to antibody/TCR gene rearrangement [19]. This hypothesis was strengthened by the demonstration that RAG is capable of DNA transposition [22,23]. The discovery of the Transib transposon in non-vertebrates, which corresponds to the RAG1 core sequence and whose TIRs are similar to the RSSs supports this hypothesis [52]. The finding of complete RAG transposons (formed by RAG1-like and RAG2-like sequences) in the genome of the protochordate amphioxus (*Branchiostoma belcheri*) [24]and

591 the hemichordate *Ptychodera flava* [48], as well as fossilized transposons in several
592 deuterostomes [26,48,53] and protostomes [49] indicates that the RAG transposon was present at
593 least as far back as the bilaterian ancestor, remained active in several lineages and was co-opted
594 as part of V(D)J recombination machinery in jawed vertebrates [48,49].

595

Box 4

The function and origin of MHC molecules

The high polymorphism of classical MHC genes is generally accepted to be a consequence of a molecular arms race between host and pathogens. However, the MHC can also be involved in inbreeding avoidance behavior and kin-specific cooperation. Since kin selection and inbreeding avoidance are universal phenomena [66-77], some authors have proposed that the immune function of the MHC is a derived function [78,79]. However, even in the best-studied systems for mate choice, evidence that MHC molecules participate and putative mechanisms remain unclear [58,80].

Various hypotheses have been proposed for the origin of MHC genes. One suggestion was that chaperone genes gave rise to the peptide-binding domains characteristic of MHC molecules [81]. Although subsequent structural analysis of HSP70 rules out the specific example suggested by these authors [2], it remains possible that a different ancient chaperone could be the ancestor. Another candidate is IRE1, which is involved as a sensor in the unfolded protein response, and has a structure and peptide binding properties like MHC molecules [82,83]. A recent suggestion is that the primordial MHC-like molecule evolved from a heavy chain-only antibody molecule that cycled between endosomal compartments and the surface [84]. Another suggestion is that NK cell receptor-ligand interactions allowed TCR-MHC interactions to evolve, with NK cells being potentially ancestral to T cells [85]. NK cells can recognize stressed cells without direct pathogen recognition. A specific scenario was recently suggested in which an NK cell receptor recognized an MHC-like molecule with a closed groove, which evolved into an MHC-like molecule with an open groove to detect proteins starting with leucine, which appear in stressed cells [2].

A linked issue is whether primordial MHC genes and molecules were organized as in the class I or class II systems. A scenario based on structure is that the original MHC molecule was a

620 homodimer of class II β-like chains, with gene duplication and divergence giving rise to
621 heterodimers of class II α-like and β-like chains, followed by an inversion leading to a class I
622 heavy-like chain and a β_2-microglobulin chain with a transmembrane region, and subsequent
623 mutation to give a class I-like molecule [2,86]. A scenario based on function suggested the
624 transfer of a peptide-binding region from a chaperone in front of an IgC-like region to produce a
625 class I-like heavy chain first [79]. Recent evidence for highly promiscuous peptide binding and
626 C-terminal protrusions of peptides from the groove of chicken class I molecules renders the
627 differences between class I and II molecules less clear [59,60,87].

Highlights

RAG evolved from a DDE transposon present in the ancestor of bilaterian animal; it evolved in a vertical manner and was domesticated as RAG in a jawed vertebrate ancestor.

This RAG-like transposon belonged to a transposon family that has the ability to create palindromic (P) diversity

A proposed model is that the jawed vertebrate ancestor possessed a complex and powerful innate immune system, where the pre-MHC molecule was able to bind and present certain PAMP molecules to a monomorphic non-rearranging TCR-like molecule.

The integration of the RAG transposon in the module of recognition of the TCR-like gene may have led to a significant increase of the recognition possibilities which presumably allowed new MHC-like variants to be selected.

Hypothetically, the increase in recognition possibilities may have also led to the appearance of MHC polymorphisms and an increase in peptide-binding repertoires (promiscuity).

Outstanding questions

What was the original function of the pre-MHC molecule and what was its origin**?**

What are the functions of the RAG genes in invertebrates?

Do any of the somatically diversified receptors in cyclostomes (lampreys and hagfish) recognize highly polymorphic cell surface molecules analogous to MHC molecules?

Do other coupled systems of highly polymorphic loci with somatically-diversified receptors exist amongst living organisms?

Glossary

AID/APOBEC deaminases (AADs): family of enzymes that convert cytidine to uridine in single-stranded nucleic acids. They are involved in numerous mutagenic processes, including those underpinning vertebrate innate and adaptive immunity

Allelic exclusion: a process by which only one allele of a gene is expressed while the other allele is silenced.

Bilaterian: metazoan animals that have a bilaterally symmetric body plan, including the protostomes and deuterostomes

Cyclostome: jawless fish (also known as agnathan); the sister group of jawed vertebrates

Deuterostome: a clade of animals including the jawed vertebrates , the jawless fish, cephalochordates (such as amphioxus), urochordates, hemichordates and echinoderms (such as sea urchins); the sister group of protosomes within bilaterans.

DDE transposon (also called class II transposon): a DNA fragment formed by two terminal inverted repeats surrounding a sequence coding for the transposase gene. The transposase gene is expressed and translated by the host cell, recognizes and cuts the TIR to excise the transposon. The broken chromosome ends are then repaired and the transposon will insert at another site in the genome.

Genetic drift: a mechanism of evolution in which allele frequencies of a population change over generations due to chance

Junctional diversity during somatic V(D)J recombination, during which the different variable segments of TCR and antibody genes are rearranged by introducing double-strand breaks between the required segments, which form hairpin loops at the ends. The hairpins are cleaved in an asymmetric manner by the Artemis enzyme, followed by joining of the broken genomic region with variable addition or subtraction of nucleotides to generate junctional diversity.

Metazoan: multicellular animals, as opposed to plants, fungi and various single-celled protists

Miniature Inverted-repeat Transposable Elements (MITEs): non-autonomous DDE transposon, which don't code for a transposase and thus must use a transposase encoded by another transposon

Non-homologous end joining (NHEJ) is a pathway that repairs double-strand breaks in DNA, with the ends of the breaks directly ligated without the need for a homologous template

Palindromic (P) diversity is due to nucleotides added during the V(D)J recombination or after transposon excision, due to asymmetric cleavage of the hairpin by the enzyme Artemis followed by normal cellular DNA repair mechanisms.

Pathogen-associated molecular patterns (PAMPs): molecules arising from and specific to pathogens (and other non-host organisms)

Pathogen recognition receptor (PRR): germline-encoded host receptors, which specifically detect molecules arising specifically from pathogens (PAMPs), other non-host molecules or host molecules in unusual locations

Piwi-interacting RNA (piRNA): family of small non-coding RNA molecules that interact with piwi-subfamily Argonaute proteins, forming piRNA complexes which are involved in the epigenetic and post-transcriptional silencing of transposable elements and the regulation of other genetic elements in germ line cells

Protostome: a clade of animals including mainly the arthropods, annelids, and molluscs; sister group of the deuterostomes with bilaterans

Recombination activating genes (RAGs) are two host genes located next to each other that encode RAG1 and RAG2 proteins, which as a complex initiates the rearrangement of gene segments of the genes encoding antibody and TCR molecules.

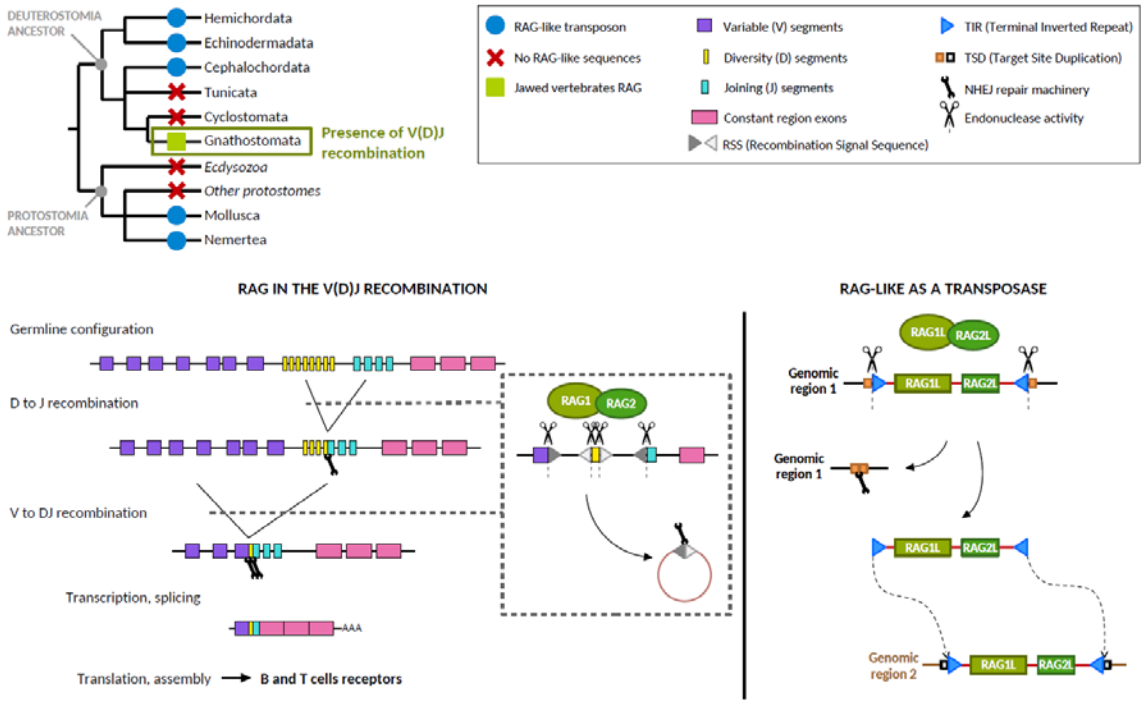
RAG DDE transposon: the RAG-like sequence found in non-vertebrates functioning as transposon

Somatic diversification: the process of mutation in somatic cells, for example genomic rearrangement

Terminal deoxytransferase (TdT): an enzyme that adds randomly adds nucleotides to untemplated broken ends of DNA, particularly during somatic diversification of antibody and TCR genes

Toll-like receptor (TLR): one class of PRRs involved in initiation of innate immune responses

Transib: the DDE transposon from protostomes whose transposase gene is closest to RAG1 and whose TIR is similar to the RAG transposon and the V(D)J RSSs .



716
717

718 **Figure 1 - Repartition and function of the jawed vertebrates RAG and the RAG-**
719 **like transposon.** On the consensus bilaterian tree is shown the presence of RAG-like
720 transposons [based on ref. 24, 26, 52 and 53] and RAG among clades sequenced in
721 the databases. The comparative activity of RAG in the jawed vertebrates V(D)J
722 recombination and the activity of RAG-like transposons [adapted from ref. 24 and
723 88] shows that this biochemical switch constitutes an easy evolutionary step.
724 Furthermore, the cuts and junctions happening in such processes create P and N
725 diversity (see text).

Table S1. Estimated Probability of the RAG transposon insertion in an ancestral V domain.

The average of transposition for a given DDE Transposon per genome is about 10^{-4}/year(Adrion et al; 2027) . The generation time is about 1 year in average for Deuterostomia (this is calculated on the average generation age of the deuterostomian) might be estimated: 10 including (TIR-------TIR) (MIR), likely more if we look at the *Ptychodera* genome [48].

The time of evolution in the deuterostomia lineage of the RAG transposon before its co-option as RAG VDJ recombinase was about 200 million years (the difference between the time appearance of the RAG transposon in the ancestor of deuterostomes and its co-option in the jawed vertebrate ancestor)[48]

The number of possible positions per gene V is about 250, in order to have a J sequence of at least 50 nucleotides [88]. We could estimate that 100 copies of V gene were present.

Number of possible transposition events on a V gene:

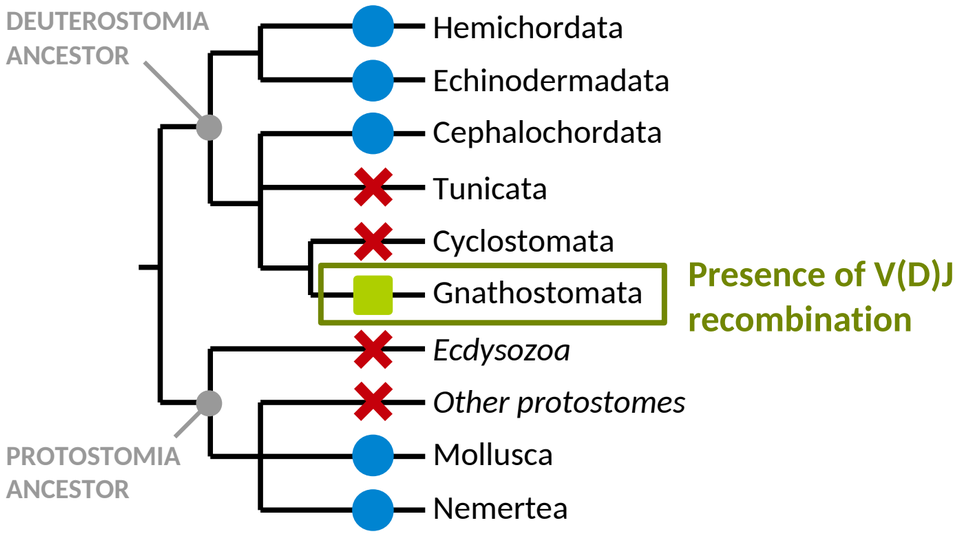
10^{-4} x2.10^{8} x10X 250x 100 = 5 x10^{9}

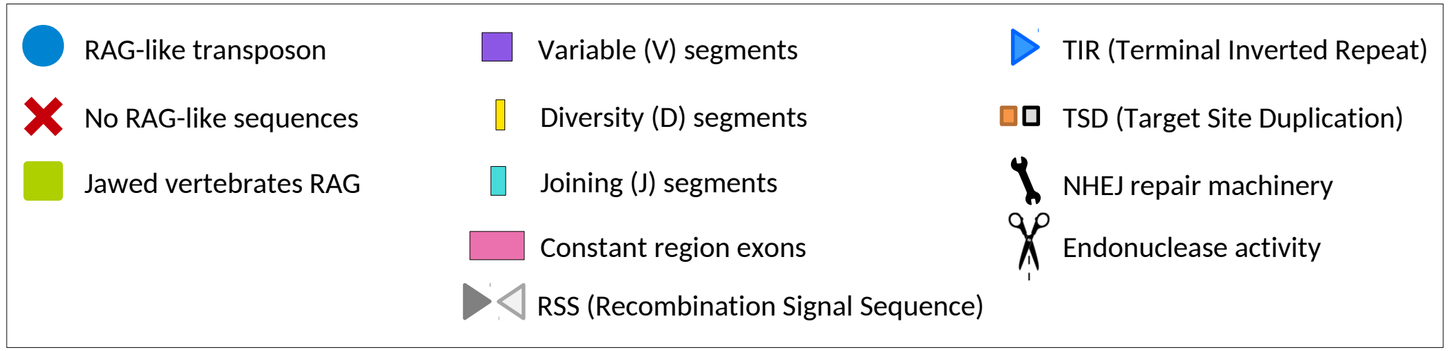
The size of a deuterostomian genome is in average 5.10^{8}

The probability of observing at least one event in 5.10^{9} repetitions is the complement of not observing any and as the events are independent (and follow the same distribution) the probability of not observing a single event in 5.10^{9} trials is the probability to do not observe it

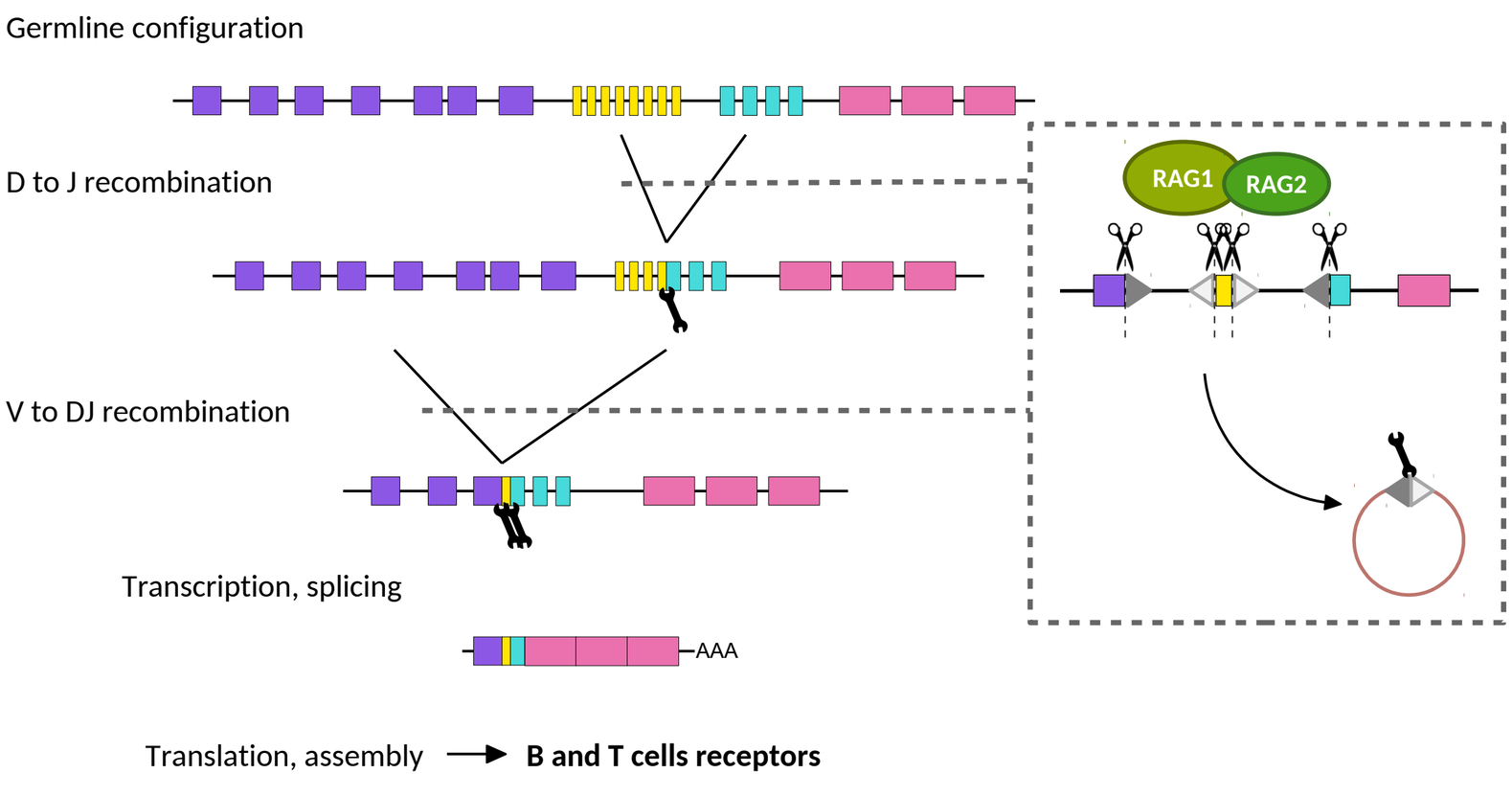
1 - (499999999/500000000)>5.10^{9} = 1%

Probability that the event happened might then be: 99%.

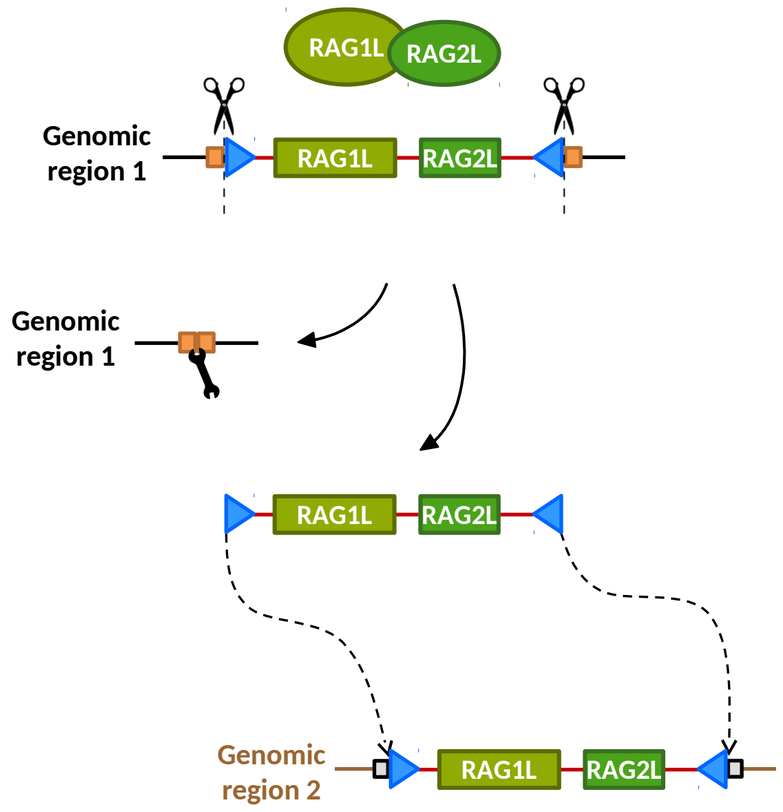




RAG IN THE V(D)J RECOMBINATION



RAG-LIKE AS A TRANSPOSASE



Supplemental Information

Origins of the RAG transposome and the MHC

Tsakou-Ngouafo L[1], Paganini J[2], Kaufman J[3,4,5], Pontarotti P[1,6]

1. Aix Marseille University IRD, APHM, MEPHI, IHU Méditerranée Infection, Marseille France 19-21 Boulevard Jean Moulin 13005 Marseille, France.
2. Xegen 15 rue de la République, 13420 Gemenos, France.
3. University of Cambridge, Department of Pathology, Tennis Court Road, CB2 1QP, Cambridge, U. K.
4. University of Cambridge, Department of Veterinary Medicine, Madingley Road, CB2 0ES, Cambridge, U. K.
5. University of Edinburgh, Institute for Immunology and Infection Research, Charlotte Auerbach Road, EH9 3FL, Edinburgh, U. K.
6. SNC5039 CNRS, 19-21 boulevard Jean Moulin,13005 Marseilles, France.

Correspondence: pierre.pontarotti@univ-amu.fr, jim.kaufman@ed.ac.uk

Supplemental Table 1.

Assuming

the average rate of transposition for a given DDE Transposon per genome is about 10^{-3}/year [89],

the generation time is about one year for deuterostomes on average, but the generation time would be ten years if we consider the *Ptychodera* genome [90,48],

the time of evolution in the deuterostome lineage of the RAG transposon before its co-option as RAG VDJ recombinase was about 200 million years (based on the difference between the appearance of the RAG transposon in the ancestor of deuterostomes and its co-option in the jawed vertebrate ancestor [48]),

the number of possible positions per gene V is about 250 (based on a V domain encoded by 300 nucleotides which is separated into a V gene segment followed by a J segment of at least 50 nucleotides [88]),

the number of V genes present in the ancestor when the RAG transposon was co-opted was 100 (based on the number of V genes per vertebrate locus and the number of TLR genes present in sea urchins [88,91,92]),

the average size of a deuterostome genome is 5×10^8 [93],

then

the number of possible transposition events on a V gene would be

$10^{-3} \times 10 \times (2 \times 10^8) \times 250 \times 100 = 5 \times 10^{10}$

so

the chance for a transposon to insert into a V gene would be

$(5 \times 10^{10}) / (5 \times 10^8) = 100$**.**

Since

the probability of observing at least one event in 100 repetitions
is the complement of not observing any
and, as the events are independent and follow the same distribution,
the probability of not observing a single event is

1 - (99/100) = 1%

then

the probability of the event happening would be 99%.

References

90. Thomas JA et al. 2010. A generation time effect on the rate of molecular evolution in invertebrates. Molec Biol Evol 27, 1173-1180.

91. Dishaw LJ et al. 2012. Brief Funct Genomics 11, 167-176.

92. Buckley KM, Rast JP. 2015. Diversity of animal immune receptors and the origins of recognition complexity in the deuterostomes. Dev Comp Immunol 49, 179-189.

93. Gregory TR et al. 2007. Eukaryotic genome size databases. Nucleic Acids Res 35, D332.