



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Origins of the RAG transposome and the MHC

Citation for published version:

Tsakou-Ngouafo, L, Paganini, J, Kaufman, J & Pontarotti, P 2020, 'Origins of the RAG transposome and the MHC', *Trends in Immunology*, vol. 41, no. 7, pp. 561-571. <https://doi.org/10.1016/j.it.2020.05.002>

Digital Object Identifier (DOI):

[10.1016/j.it.2020.05.002](https://doi.org/10.1016/j.it.2020.05.002)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Peer reviewed version

Published In:

Trends in Immunology

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



1

2 Origins of the RAG transposome and the MHC

3

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

5

6 1. Aix Marseille University IRD, APHM, MEPHI, IHU Méditerranée Infection, Marseille France

7 3 19-21 Boulevard Jean Moulin 13005 Marseille, France.

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

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

10 U. K.

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

12 Cambridge, U. K.

13 5. University of Edinburgh, Institute for Immunology and Infection Research, Charlotte Auerbach

14 Road, EH9 3FL, Edinburgh, U. K.

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

16 Corresponding authors: Pierre Pontarotti, pierre.pontarotti@univ-amu.fr, Jim

17 Kaufman, jim.kaufman@ed.ac.uk

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

19 convergent evolution

20 **Abstract**

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

35 **Presentation of the Hypotheses**

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

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

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

59 **excision** and DNA repair, and posit that the biochemical switch from the **RAG transposon**
60 insertion and excision to the RAG sequence-specific recombination was a simple functional shift.
61 These three properties would have increased the likelihood of RAG transposon being co-opted as
62 a major player modulating the somatic diversity of the antibodies and T cell receptors (TCRs).
63 Finally, we propose that the somatic receptor diversity orchestrated by RAG allowed the
64 emergence of the MHC peptide binding promiscuity and polymorphism. Many excellent papers
65 and reviews have described and proposed hypotheses about the origin and the evolution of the
66 adaptive immune system and the MHC, but here we focus on the origin of the **somatic**
67 **diversification** and its consequence on the evolution of the MHC.

68

69 **The origins of vertebrate adaptive immunity in metazoans**

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

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

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

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

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

105

106 **The next step: emergence of diversified receptors**

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

121 ***RAG origin***

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

226

227

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

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

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

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

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

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

276

277 **Concluding Remarks**

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

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

294

295 **References**

- 296 1. Flajnik, M.F. and Kasahara, M. (2010) Origin and evolution of the adaptive immune
297 system: genetic events and selective pressures. *Nat. Rev. Genet.* 11, 47-59
- 298 2. Kaufman, J. (2018) Unfinished Business: Evolution of the MHC and the Adaptive
299 Immune System of Jawed Vertebrates. *Annu. Rev. Immunol.* 36, 383-409
- 300 3. Pancer, Z. and Cooper, M.D. (2006) The evolution of adaptive immunity. *Annu. Rev.*
301 *Immunol.* 24, 497-518
- 302 4. Boehm, T. *et al.* (2018) Evolution of Alternative Adaptive Immune Systems in
303 Vertebrates. *Annu. Rev. Immunol.* 36, 19-42
- 304 5. Flajnik, M.F. (2018) A Convergent Immunological Holy Trinity of Adaptive Immunity in
305 Lampreys: Discovery of the Variable Lymphocyte Receptors. *J. Immunol.* 201, 1331-
306 1335
- 307 6. Buckley, K.M. and Rast, J.P. (2015) Diversity of animal immune receptors and the
308 origins of recognition complexity in the deuterostomes. *Dev. Comp. Immunol.* 49, 179-
309 189
- 310 7. Litman, G.W. *et al.* (2010) The origins of vertebrate adaptive immunity. *Nat. Rev.*
311 *Immunol.* 10, 543–553
- 312 8. Homa J. *et al.* (2013) Exposure to immunostimulants induces changes in activity and
313 proliferation of coelomocytes of *Eisenia andrei*. *J. Comp. Physiol. B.* 183, 313-322
- 314 9. Holm, K. *et al.* (2008) Induced cell proliferation in putative haematopoietic tissues of the
315 sea star, *Asterias rubens* (L.). *J. Exp. Biol.* 211, 2551-2558

- 316 10. Salamat, Z. and Sullivan, J.T. (2009) Involvement of protein kinase C signalling and
317 mitogen-activated protein kinase in the amebocyte-producing organ of *Biomphalaria*
318 *glabrata* (Mollusca). *Dev. Comp. Immunol.* 33, 725–727
- 319 11. Milutinović, B. and Kurtz, J. (2016) Immune memory in invertebrates. *Semin. Immunol.*
320 28, 328-342
- 321 12. Glover, L. *et al.* (2016) VEX1 controls the allelic exclusion required for antigenic
322 variation in trypanosomes. *Proc. Natl. Acad. Sci. USA.* 113, 7225–7230
- 323 13. Monahan, K. and Lomvardas, S. (2015) Monoallelic expression of olfactory receptors.
324 *Annu. Rev. Cell Dev. Biol.* 31, 721-740
- 325 14. Conticello, S.G. (2008) The AID/APOBEC family of nucleic acid mutators. *Genome*
326 *Biol.* 9, 229
- 327 15. Arakawa, H. *et al.* (2002) Requirement of the activation-induced deaminase (AID) gene
328 for immunoglobulin gene conversion. *Science* 295, 1301-1306
- 329 16. Rogozin, I.B. *et al.* (2007) Evolution and diversification of lamprey antigen receptors:
330 evidence for involvement of an AID-APOBEC family cytosine deaminase. *Nat. Immunol.*
331 8, 647–656
- 332 17. Liu, M.C. *et al.* (2018) Diversification of AID/APOBEC-like deaminases in metazoa:
333 multiplicity of clades and widespread roles in immunity. *Nat. Comm.* 9, 1948
- 334 18. Krishnan, A. *et al.* (2018) Diversification of AID/APOBEC-like deaminases in metazoa:
335 multiplicity of clades and widespread roles in immunity. *Proc. Natl. Acad. Sci. U S A.*
336 115, E3201-E3210

- 337 19. Thompson, C.B. (1995) New insights into V(D)J recombination and its role in the
338 evolution of the immune system. *Immunity* 3, 531-539
- 339 20. Sakano, H. *et al.* (1979) Sequences at the somatic recombination sites of immunoglobulin
340 light-chain genes. *Nature* 280, 288-294
- 341 21. McBlane, J.F. *et al.* (1995) Cleavage at a V(D)J recombination signal requires only
342 RAG1 and RAG2 proteins and occurs in two steps. *Cell* 83, 387-395
- 343 22. Agrawal, A. *et al.* (1998) Transposition mediated by RAG1 and RAG2 and its
344 implications for the evolution of the immune system. *Nature* 394, 744-751
- 345 23. Hiom, K. *et al.* (1998) DNA transposition by the RAG1 and RAG2 proteins: a possible
346 source of oncogenic translocations. *Cell* 94, 463-470
- 347 24. Huang, S. *et al.* (2016) Discovery of an Active RAG Transposon Illuminates the Origins
348 of V(D)J Recombination. *Cell* 166, 102-114
- 349 25. Lafaille, J.J. *et al.* (1989) Junctional sequences of T cell receptor gamma delta genes:
350 implications for gamma delta T cell lineages and for a novel intermediate of V-(D)-J
351 joining. *Cell* 59, 859-870
- 352 26. Fugmann, S.D. (2010) The origins of the Rag genes--from transposition to V(D)J
353 recombination. *Semin. Immunol.* 22, 10-16
- 354 27. Liu, K. and Wessler S.R. (2017) Transposition of Mutator-like transposable elements
355 (MULEs) resembles hAT and Transib elements and V(D)J recombination. *Nucleic Acids*
356 *Res.* 45, 6644-6655
- 357 28. Hickman, A.B. *et al.* (2018) Structural insights into the mechanism of double strand

- 358 break formation by Hermes, a hAT family eukaryotic DNA transposase. *Nucleic Acids*
359 *Res.* 46, 10286-10301
- 360 29. Colot, V. *et al.* (1998) Extensive, nonrandom diversity of excision footprints generated by
361 Ds-like transposon Ascot-1 suggests new parallels with V(D)J recombination. *Mol. Cell*
362 *Biol.* 18, 4337-4346
- 363 30. Bonatto, D. *et al.* (2005) In silico identification and analysis of new Artemis/Artemis-like
364 sequences from fungal and metazoan species. *Protein J.* 24, 399-411
- 365 31. Tiefenbach, T. and Junop, M. (2011) Pso2 (SNM1) is a DNA structure-specific
366 endonuclease. *Nucleic Acids Res.* 40, 2131-2139
- 367 32. Carmona, L.M. and Schatz, D.G. (2017) New insights into the evolutionary origins of the
368 recombination-activating gene proteins and V(D)J recombination. *FEBS J.* 284, 1590-
369 1605
- 370 33. Zhang, Y. *et al.* Transposon molecular domestication and the evolution of the RAG
371 recombinase. *Nature* 569, 79-84
- 372 34. Tsakou, L. *et al.* (2020) DDE transposon as public goods. *Evolutionary Biology*. Eds.,
373 Pierre Pontarotti, Springer International Publishing, in press
- 374 35. Baudry, C. *et al.* PiggyMac, a domesticated piggyBac transposase involved in
375 programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. *Genes Dev.*
376 23, 2478-2483
- 377 36. Cheng, C.Y. *et al.* (2016) The piggyBac transposon-derived genes TPB1 and TPB6
378 mediate essential transposon-like excision during the developmental rearrangement of
379 key genes in *Tetrahymena thermophila*. *Genes Dev.* 30, 2724-2736.

- 380 37. Rajaei, N. *et al.* (2014) Domesticated transposase Kat1 and its fossil imprints induce
381 sexual differentiation in yeast. *Proc. Natl. Acad. Sci. U S A.* 111, 15491-15496
- 382 38. Barsoum, E. *et al.* (2010) Alpha3, a transposable element that promotes host sexual
383 reproduction. *Genes Dev.* 24, 33-44
- 384 39. Lu, H. *et al.* (2007) Extent to which hairpin opening by the Artemis:DNA-Pkcs complex
385 can contribute to junctional diversity in V(D)J recombination. *Nucleic Acids Res.* 35,
386 6917-6923
- 387 40. Arensburger, P. *et al.* (2011) Phylogenetic and functional characterization of the hAT
388 transposon superfamily. *Genetics* 188, 45-57
- 389 41. Dupeyron, M. *et al.* (2019) Evolution of Mutator transposable elements across eukaryotic
390 diversity. *Mob DNA* 10, 12
- 391 42. Hencken, C.G. *et al.* (2012) Functional characterization of an active Rag-like transposase.
392 *Nat Struct. Mol. Biol.* 19, 834-836
- 393 43. Wallau, G.L. *et al.* (2012) Horizontal transposon transfer in eukarya: detection, bias, and
394 perspectives. *Genome Biol. Evol.* 4, 689-699
- 395 44. Dotto BR *et al.* (2018) HTT-DB: new features and updates. *Database (Oxford)* 1
- 396 45. Joly-Lopez, Z. *et al.* (2016) Phylogenetic and Genomic Analyses Resolve the Origin of
397 Important Plant Genes Derived from Transposable Elements. *Mol. Biol. Evol.* 33, 1937-
398 1956
- 399 46. Bouallègue, M. *et al.* (2017) Molecular Evolution of piggyBac Superfamily: from
400 Selfishness to Domestication. *Genome Biol. Evol.* 9, 323-339

- 401 47. Peccoud, J. *et al.* (2017) Massive horizontal transfer of transposable elements in insects.
402 *Proc. Natl. Acad. Sci. USA.* 114, 4721-4726
- 403 48. Morales Poole, J.R. *et al.* (2017) The RAG transposon is active through the deuterostome
404 evolution and domesticated in jawed vertebrates. *Immunogenetics* 69, 391-400
- 405 49. Martin, E.C. *et al.* (2020) Evidence for an ancient bilaterian origin of the RAG-like
406 transposon. *Mobile Elements*, in press
- 407 50. Aravin, A.A. *et al.* (2007) The Piwi-piRNA pathway provides an adaptive defense in the
408 transposon arms race. *Science* 318, 761-764
- 409 51. Lerat, E. *et al.* TEtools facilitates big data expression analysis of transposable elements
410 and reveals an antagonism between their activity and that of piRNA genes. *Nucleic Acids*
411 *Res.* 45, e17
- 412 52. Kapitonov, V.V. and Jurka, J. (2005) RAG1 core and V(D)J recombination signal
413 sequences were derived from Transib transposons. *PLoS Biol.* 3, e181
- 414 53. Kapitonov, V.V. and Koonin E.V. (2015) Evolution of the RAG1-RAG2 locus: both
415 proteins came from the same transposon. *Biol. Direct.* 10, 20
- 416 54. Callebaut, I. and Mornon, J.P. (1998) The V(D)J recombination activating protein RAG2
417 consists of a six-bladed propeller and a PHD fingerlike domain, as revealed by sequence
418 analysis. *Cell. Mol. Life Sci.* 54, 880-891
- 419 55. Hsu, E., and Lewis, S.M. (2015) The origin of V(D)J diversification. In: *Molecular*
420 *Biology of B cells*, Alt, F.W., Honjo, T., Radbruch, A., and Reth, M., eds. Elsevier,
421 Academic Press, Amsterdam, pp. 133-148.

- 422 56. Danchin E. *et al.* (2004) The major histocompatibility complex origin. *Immunol. Rev.*
423 198, 216-232
- 424 57. Spurgin, L.G. and Richardson, D.S. (2010) How pathogens drive genetic diversity: MHC,
425 mechanisms and misunderstandings. *Proc. Biol. Sci.* 277, 979–988
- 426 58. Radwan, J. *et al.* (2020) Advances in the evolutionary understanding of MHC
427 polymorphism. *Trends. Genet.* 36, 298-311
- 428 59. Kaufman, J. (2018) Generalists and Specialists: A New View of How MHC Class I Mol-
429 ecules Fight Infectious Pathogens. *Trends Immunol.* 39, 367–379
- 430 60. Chappell, P. *et al.* (2015) Expression levels of MHC class I molecules are inversely
431 correlated with promiscuity of peptide binding. *eLIFE* 4, e05345
- 432 61. Manczinger, M. *et al.* (2019) Pathogen diversity drives the evolution of generalist MHC-
433 II alleles in human populations. *PLoS Biol.* 17, e3000131
- 434 62. Vidović, D. and Matzinger, P. (1988) Unresponsiveness to a foreign antigen can be
435 caused by self-tolerance. *Nature* 336, 222-225
- 436 63. Nowak, M.A. *et al.* (1992) The optimal number of major histocompatibility complex
437 molecules in an individual. *Proc. Natl. Acad. Sci. U S A.* 89, 10896-10899
- 438 64. Borghans, J.A. *et al.* (2003) Thymic selection does not limit the individual MHC
439 diversity. *Eur. J. Immunol.* 33, 3353-3358
- 440 65. Teng, G. and Schatz, D.G. (2015) Regulation and Evolution of the RAG Recombinase.
441 *Adv. Immunol.* 128, 1-39
- 442 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

- 464 allorecognition. *Curr. Biol.* 21, R82-92
- 465 76. Saak, C.C. and Gibbs, K.A. (2016) The Self-Identity Protein IdsD Is Communicated
466 between Cells in Swarming *Proteus mirabilis* Colonies. *J. Bacteriol.* 198, 3278-3286
- 467 77. Wall, D. (2016) Kin Recognition in Bacteria. *Annu. Rev. Microbiol.* 70, 143-160
- 468 78. Boehm ,T. (2006) Quality control in self/nonself discrimination. *Cell* 125, 845-858
- 469 79. Ruff, J.S. *et al.* (2012) MHC signaling during social communication. *Adv. Exp. Med.*
470 *Biol.* 738, 290-313
- 471 80. Leinders-Zufall, T. *et al.* (2009) Structural requirements for the activation of vomeronasal
472 sensory neurons by MHC peptides. *Nat. Neurosci.* 12, 1551-1558
- 473 81. Flajnik, M.F. *et al.* (1991) Which came first, MHC class I or class II? *Immunogenetics* 33,
474 295-300
- 475 82. Credle, J.J. *et al.* (2005) On the mechanism of sensing unfolded protein in the
476 endoplasmic reticulum. *Proc. Natl. Acad. Sci. USA.* 102, 18773-18784
- 477 83. Karagöz, G.R.E. *et al.* (2017) An unfolded protein-induced conformational switch
478 activates mammalian IRE1. *eLIFE* 6, e30700
- 479 84. Dijkstra, J.M. and Yamaguchi. T. (2019) Ancient features of the MHC class II
480 presentation pathway, and a model for the possible origin of MHC molecules.
481 *Immunogenetics* 71, 233-249
- 482 85. Du Pasquier, L. (2000) The phylogenetic origin of antigen-specific receptors. *Curr. Top.*
483 *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

492 **Acknowledgements**

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

497

498 **Figure legends**

499

500 **Figure 1. Phylogenetic distribution: RAG in jawed vertebrates and the RAG-like**
501 **transposon.** On the consensus bilaterian tree is shown the presence of RAG-like transposons
502 [24,26,49,52,53] and RAG among clades sequenced in the databases. The comparative activity of
503 RAG in V(D)J recombination among the jawed vertebrates and the activity of RAG-like
504 transposons is adapted [24,88] showing that this biochemical switch would constitute an
505 unconstrained evolutionary step.

506

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

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

516 The integration of the RAG transposon into an ancestral TCR-like gene may have led to a
517 significantly increased probability of recognition by these original, non-diverse TCRs. As a result,
518 mutations in the ancestral MHC-like molecule may have led to a conformational ability to bind
519 new PAMPs; as a consequence, mutated MHC molecules could have then been recognized by
520 diverse TCRs thereafter, becoming evolutionarily selected via three mechanisms peptide-binding
521 promiscuity allelic polymorphism and as by the expansion into multigene families.

522 The expanded ability of this ancient MHC/TCR system to recognize new PAMPs would
523 presumably allow peptides to be bound, presented and recognized during an immune response.

524 **Box 1.**

525 **Brief overview of the adaptive immune system in vertebrates**

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

540

541 **Box 2**

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

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

554 It is important to note that the function of possible T- and B-cell lineages before the adaptive
555 immunity arose is entirely unclear. **Innate lymphoid cells** (ILCs) found in mammals are potential
556 candidates for the functions of non-adaptive T cells before adaptive immunity (although they
557 could also be a novelty of placental mammals), but system replacement might be more likely
558 [48]. If the first adaptive immune system was based on VLR, then in jawed vertebrates, a shift
559 occurred from the IgV-IgC innate immunity to the IgV-IgC adaptive immunity, followed by the
560 loss of the VLR-based adaptive immunity. If the first adaptive immune system was based on IgV-
561 IgC, then in cyclostomes the reverse may have occurred. In fact, such replacements have been
562 noted for natural killer (NK) cell receptors [2]: at least three families of NK cell receptors exist
563 with analogous functions: lectin-like receptors (overwhelmingly in rodents and to a lesser extent
564 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.

567 **Box 3**

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

569 The antibody and TCR genes of jawed vertebrates are assembled from variable (V), diversity (D),
570 and joining (J) gene segments during B- and T-lymphocyte development to give contiguous VJ
571 and VDJ sequences. The process to excise the DNA between the gene segments is initiated by the
572 RAG endonuclease. The RAG endonuclease specifically recognizes recombination signal
573 sequences (RSSs) that flank each gene segment. RSSs are composed of conserved heptamer and
574 nonamer sequences separated by a less conserved spacer sequence of either 12 or 23 bp (12RSS
575 and 23RSS). RAG-mediated DNA cleavage occurs preferentially in a complex containing a
576 12RSS and a 23RSS, involving a nick-hairpin mechanism.

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

583 The discovery of RSSs, along with the mechanism of RSS cleavage which is similar to several
584 cut-and-paste DNA transposases (DDE transposases) [20,21] resulted in the hypothesis that a
585 DDE transposon invaded an Ig-like gene, leading eventually to antibody/TCR gene
586 rearrangement [19]. This hypothesis was strengthened by the demonstration that RAG is capable
587 of DNA transposition [22,23]. The discovery of the Transib transposon in non-vertebrates, which
588 corresponds to the RAG1 core sequence and whose TIRs are similar to the RSSs supports this
589 hypothesis [52]. The finding of complete RAG transposons (formed by RAG1-like and RAG2-
590 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

596 **Box 4**

597 **The function and origin of MHC molecules**

598 The high polymorphism of classical MHC genes is generally accepted to be a consequence of a
599 molecular arms race between host and pathogens. However, the MHC can also be involved in
600 inbreeding avoidance behavior and kin-specific cooperation. Since kin selection and inbreeding
601 avoidance are universal phenomena [66-77], some authors have proposed that the immune
602 function of the MHC is a derived function [78,79]. However, even in the best-studied systems for
603 mate choice, evidence that MHC molecules participate and putative mechanisms remain unclear
604 [58,80].

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

618 A linked issue is whether primordial MHC genes and molecules were organized as in the class I
619 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].

628 **Highlights**

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

631

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

634

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

638

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

642

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

645

646

647 **Outstanding questions**

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

649 What are the functions of the RAG genes in invertebrates?

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

652

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

655

656

657 **Glossary**

658

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

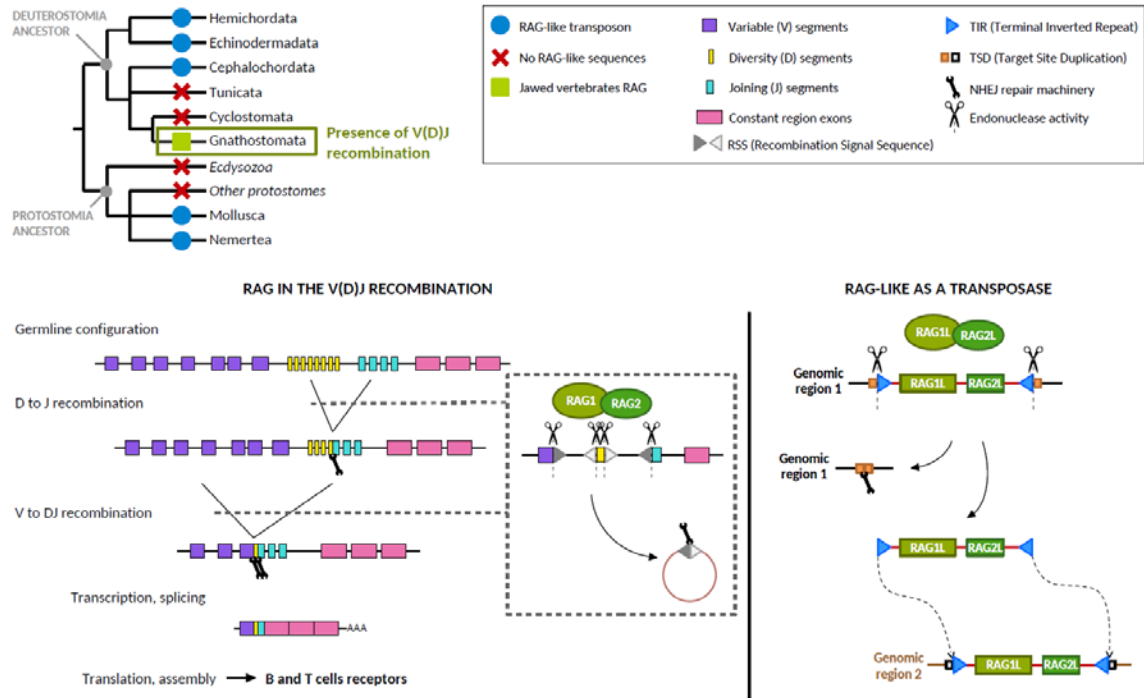
700 **Protostome:** a clade of animals including mainly the arthropods, annelids, and molluscs; sister
701 group of the deuterostomes with bilaterans

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

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

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

709 **Terminal deoxytransferase (TdT):** an enzyme that adds randomly adds nucleotides to
710 untemplated broken ends of DNA, particularly during somatic diversification of antibody and
711 TCR genes
712 **Toll-like receptor (TLR):** one class of PRRs involved in initiation of innate immune responses
713 **Transib:** the DDE transposon from protostomes whose transposase gene is closest to RAG1 and
714 whose TIR is similar to the RAG transposon and the V(D)J RSSs .
715



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).

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

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

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

735 The number of possible positions per gene V is about 250, in order to have a J sequence of at least
736 50 nucleotides [88]. We could estimate that 100 copies of V gene were present.

737 Number of possible transposition events on a V gene:

738 $10^{-4} \times 2.10^8 \times 10 \times 250 \times 100 = 5 \times 10^9$

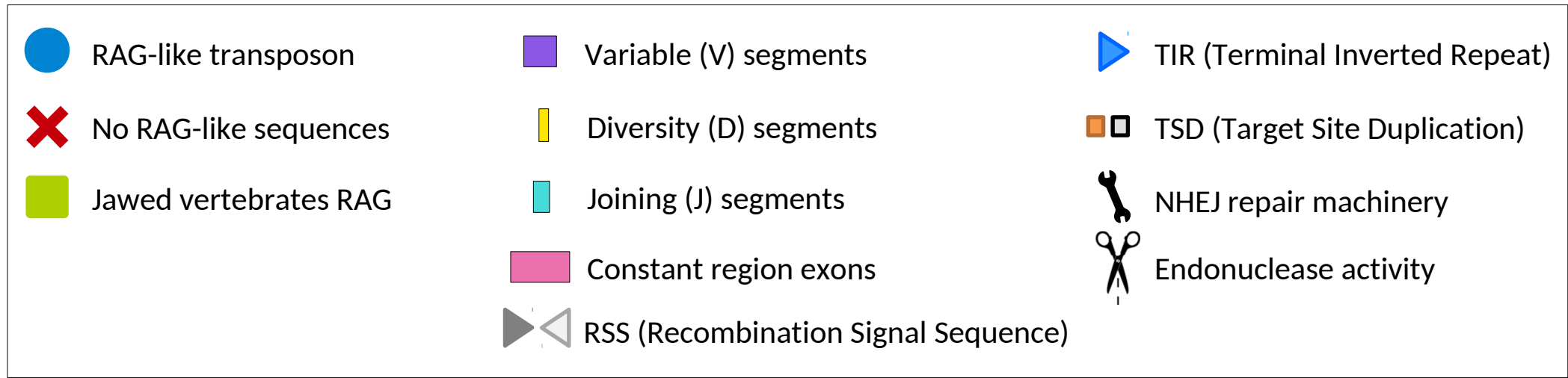
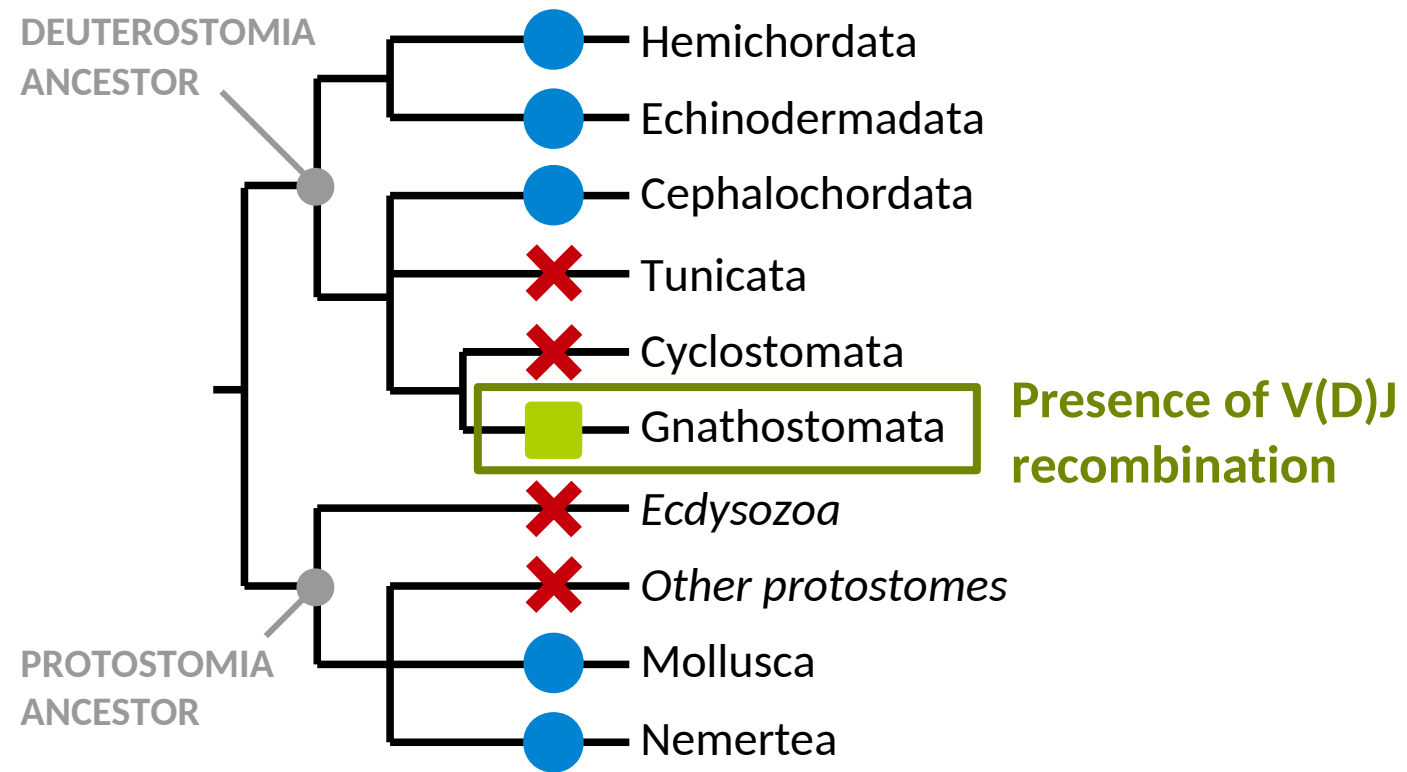
739 The size of a deuterostomian genome is in average 5.10^8

740

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

744 $1 - (499999999/500000000) > 5.10^9 = 1\%$

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

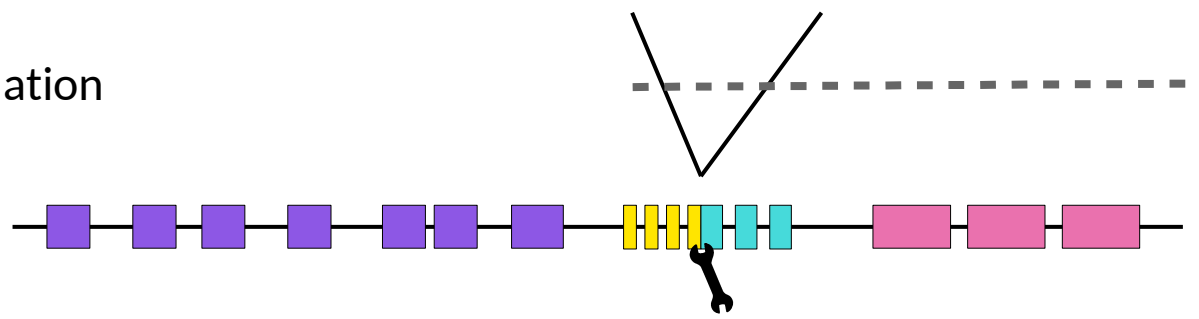


RAG IN THE V(D)J RECOMBINATION

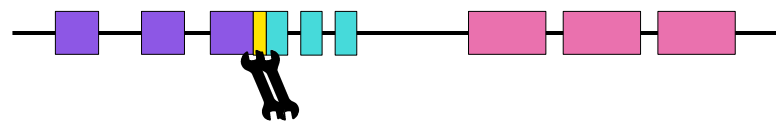
Germline configuration



D to J recombination



V to DJ recombination

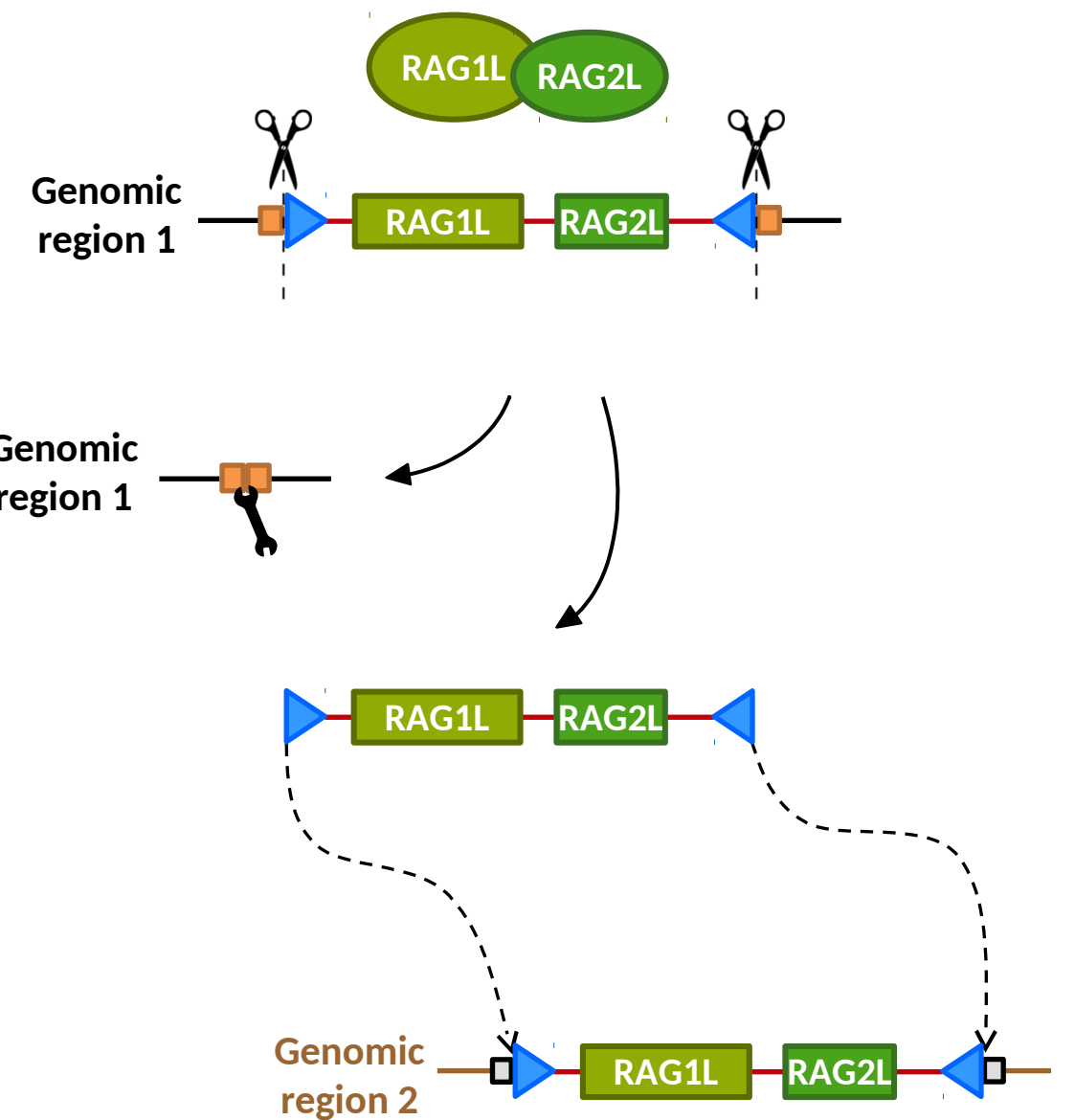


Transcription, splicing



Translation, assembly → B and T cells receptors

RAG-LIKE AS A TRANSPOSASE



Supplemental Information

Origins of the RAG transposome and the MHC

Tsakou-Ngouafo L¹, Paganini J², 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.