

Origins of the RAG Transposome and the MHC

Louis Tsakou-Ngouafo, Julien Paganini, Jim Kaufman, Pierre Pontarotti

▶ To cite this version:

Louis Tsakou-Ngouafo, Julien Paganini, Jim Kaufman, Pierre Pontarotti. Origins of the RAG Transposome and the MHC. Trends in Immunology, 2020, 41 (7), pp.561-571. 10.1016/j.it.2020.05.002 . hal-03371051

HAL Id: hal-03371051 https://hal.science/hal-03371051v1

Submitted on 8 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Trends in Immunology Origins of the RAG transposome and the MHC --Manuscript Draft--

Manuscript Number:	TREIMM-D-20-00031R1
Article Type:	Opinion
Keywords:	hairpin; flanking; DDE transposon excision; Artemis; Palindromic diversity; convergent evolution
Corresponding Author:	Pierre Pontarotti CNRS marseilles, FRANCE
First Author:	Louis Tsakou
Order of Authors:	Louis Tsakou
	julien paganini
	Jim Kaufman, prof
	Pierre Pontarotti
Abstract:	How innate immunity gave rise to adaptive immunity in vertebrates remains unknown. We propose an evolutionary scenario beginning with pathogen-associated molecular pattern(s) (PAMPs) being presented by molecule(s) on one cell to specific receptor(s) on other cells, much like MHC molecules and T cell receptors (TCRs). In this model, mutations in MHC-like molecule(s) that bound new PAMP(s) would not be recognized by original TCR-like molecule(s), and new MHC-like gene(s) would be lost by neutral drift. Integrating recombination activating gene (RAG) transposon(s) in a TCR-like gene would result in greater recognition diversity, with new MHC-like variants recognized and selected, along with a new RAG/TCR-like system. MHC genes would be selected to present many peptides, through multigene families, allelic polymorphism, and peptide-binding promiscuity.

1	
2	Origins of the RAG Transposome and the MHC
3	
4	Louis Tsakou (1), Julien Paganini (2), Jim F. Kaufman J(3,4,5)*, Pierre Pontarotti (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 Marseilles, France.
16	*Corresponding authors: Pierre Pontarotti, pierre.pontarotti@univ-amu.fr, Jim Kaufman,
17	jim.kaufman@ed.ac.uk
18	Key Words: hairpin, flanking, DDE transposon excision, Artemis, palindromic diversity,
19	convergent evolution

20 Abstract. How innate immunity gave rise to adaptive immunity in vertebrates remains 21 unknown. We propose an evolutionary scenario beginning with pathogen-associated 22 molecular pattern(s) (PAMPs) being presented by molecule(s) on one cell to specific 23 receptor(s) on other cells, much like major histocompatibility complex (MHC) molecules 24 and T cell receptors (TCRs). In this model, mutations in MHC-like molecule(s) that 25 bound new PAMP(s) would not be recognized by original TCR-like molecule(s), and new 26 MHC-like gene(s) would be lost by neutral drift. Integrating recombination activating 27 gene (RAG) transposon(s) in a TCR-like gene would result in greater recognition 28 diversity, with new MHC-like variants recognized and selected, along with a new 29 RAG/TCR-like system. MHC genes would be selected to present many peptides, through 30 multigene families, allelic polymorphism, and peptide-binding promiscuity.

31 Presentation of the Evolutionary Hypotheses

32 When the vertebrate adaptive immune system was first investigated, two aspects were particularly 33 impressive: i) B- and T-cell repertoire diversity and the generation of this diversity by 34 recombination, and ii) the enormous polymorphism of molecules encoded by the major 35 histocompatibility complex (MHC), and these proteins bind numerous peptides. It has now 36 become clear that this molecular system, largely involving immunoglobulin (Ig) domains, may 37 have already been in place in the lineage leading to jawed vertebrates, including cartilaginous and 38 bony fish, amphibians, reptiles, birds and mammals [1,2] The discovery of a parallel system in 39 jawless fish, based on leucine-rich repeats (LRRs) rather than Ig domains, suggests that the 40 vertebrate cellular system is common to both jawless and jawed vertebrates [3-5] (see Box 1).

41 In this opinion piece, we propose several hypotheses that together, can explain the emergence of 42 the recombination activating gene (RAG)-based adaptive immune system in a jawed vertebrate 43 ancestor following the evolution of several linked biological traits. We use the evolutionary 44 consequence of co-opting genetic traits to propose the origin of MHC polymorphism. We first 45 discuss the concept that a complex innate immune system may have existed long before the 46 emergence of the vertebrate ancestor. This may have included large multigene families able to 47 recognize foreign pathogens, cell proliferation, and immune memory following pathogen contact, 48 well defense **AID/APOBEC-like** (activation-induced as as pathogen via 49 deaminase/apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) cytidine 50 deaminase genes. We also present arguments supporting the possible presence of clonal 51 expansion and allelic exclusion, with each clone expressing a member of the multigene family, 52 and recognizing a pathogen-associated molecular pattern (PAMP). We then describe the RAG 53 DDE transposon, which belongs to a functional transposon family that allows palindromic (P) 54 diversity after excision and DNA repair, recognized to have been active in organisms -- from the 55 ancestor of the bilaterians (animals including protostomes, deuterostomes and a few other 56 groups), to the ancestor of jawed vertebrates (Figure 1). We posit that the biochemical switch from the RAG transposon insertion and excision to the RAG sequence-specific recombination 57 58 was a relatively simple functional shift. These three properties would have increased the 59 likelihood of RAG transposon being co-opted as a major player in gene recombination, 60 modulating the somatic diversity of antibodies (immunoglobulins) and T cell receptors (TCR). 61 Finally, we propose that the somatic receptor diversity orchestrated by RAG allowed the 62 emergence of MHC peptide-binding promiscuity and polymorphism. Many excellent papers and 63 reviews have described and proposed hypotheses on the origin and evolution of the adaptive 64 immune system and the MHC, although here, we focus on the origin of somatic diversification 65 and its consequence on the evolution of MHC molecules.

66

67 The origins of vertebrate adaptive immunity in metazoans

In common with vertebrate adaptive immune systems, other **metazoans** can harbor 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.

To discuss the former first three points, some non-vertebrate metazoan genomes display large multigene families involved in innate immunity, including those based on LRRs -- such as tolllike receptors (TLRs) and other **pathogen recognition receptors** (PRRs) likely to recognize 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 cell proliferation [8-10]. Third, numerous studies have demonstrated various forms of immune memory in many non80 vertebrate metazoans (for review, see [11]). Although the evidence is fragmentary, the existence 81 of even a few examples shows that these biological traits exist outside of vertebrates and may 82 have provided the basis for the vertebrate adaptive immune system.

83 For the last two points, clonal expression of receptors and allelic exclusion are common 84 mechanisms in eukaryotes, rather than mechanisms limited to the vertebrate adaptive immune 85 system, such as the multigene family of olfactory receptors and/or the antigenic variation of 86 variable surface glycoproteins (VSGs) in trypanosomes [12,13]. AID/APOBEC enzymes have 87 several functions in vertebrates [14], including generating diversity of non-self recognition, 88 producing point mutations (for instance, B-cell receptors in jawed vertebrates), and driving gene 89 conversion mechanisms by DNA breakage, followed by repair mechanisms that increase the 90 probability of gene conversion in cyclostomes and some invertebrates [15,16]. Orthologs of this 91 family are also found with similar activities in deuterostomes, and the AID/APOBEC-like 92 cytidine deaminase is expressed preferentially in tissues undergoing constant direct interaction 93 with potential pathogens, can be induced upon pathogen challenge, and is involved in innate 94 immunity acting on non-self-DNA [17,18].

95 Thus, in the pre-adaptive immune system, multigene families of PRRs and IgV-IgC receptors 96 could have recognized PAMPs leading to cellular activation and proliferation, and immune 97 memory. The generation of diversity for these multigene families could have been driven by 98 members of the AID/APOBEC family [as proposed in [17]], first involved in non-self-recognition 99 with one family member co-opted during vertebrate evolution by shifting the mutagenic activity 100 from non-self to self. In this model, the mechanisms for clonal expression and allelic exclusion 101 would lead to each clone expressing a single member of the multigenic family recognizing 102 particular PAMPs.

104 The next step: emergence of diversified receptors

105 As described above, two adaptive immune systems are found in vertebrates (see Box 1). In 106 considering the origins of the adaptive immune system of jawless vertebrates, two potentially 107 ancestral genes are found in various metazoans and could have given rise to the diversified 108 variable lymphocyte receptors (VLRs); namely, many proteins with LRR domains --109 particularly TLRs and AID/APOBEC-like enzymes. In contrast, the emergence of the adaptive 110 immune system of jawed vertebrates is less clear, with plausible candidates for the receptors in 111 metazoans, but rather complex in terms of the generation of diversity (see Box 2). Antibody and 112 TCR genes of jawed vertebrates are based on Ig domains assembled from separate variable (V), 113 diversity (D) and joining (J) gene segments during B and T lymphocyte development to give 114 contiguous VJ and V(D)J sequences. The process is initiated by the RAG endonuclease involved 115 in excision of DNA between the gene segments and is continued by ubiquitously-expressed DNA 116 repair enzymes (see Box 3). The appearance of RAG has long been considered a key evolutionary 117 step that can explain the origin of the jawed vertebrate adaptive system [19,20].

118 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, eventually leading to antibody/TCR gene rearrangement [12] [19].

124 The experimental analyses of the RAG transposon from amphioxus (a chordate from the sister 125 group of vertebrates, see Figure 1), devoid of any known adaptive immune system, shed light on 126 the functional shift from a RAG transposon to the RAG sequence-specific recombination 127 activating system. First, the excision reaction is similar for the two endonucleases: the 128 transposase recognizes terminal inverted repeat (TIR) sequences, and the co-opted 129 endonuclease (RAG) recognizes TIR-like sequences (i.e. the RSSs) [24]. Both involve a nick-130 hairpin mechanism characteristic of several DDE DNA transposases, including RAG/Transib 131 (with Transib having only the RAG1 core, i.e. the endonuclease, present in several protostomes), 132 HAT, and Mutator [25-28]. After excision, the hairpin-tipped segments are processed by the 133 evolutionarily conserved endonuclease Artemis, performing an asymmetric opening of the 134 hairpin, and leading to palindromic P nucleotide variation (see Box 3) [24,29]. Other non-135 vertebrate species also bear a RAG transposon that is likely to work in a similar manner to that of 136 amphioxus (Box 4 and below).

Of note, 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**. Preliminary work suggests that the TdT gene has a long phylogenetic history, although this remains to be fully validated; in this scenario, the domesticated RAG system would have co-opted TdT. A second major difference 151 is at the level of the excised fragment flanked by RSSs or TIRs. The domesticated RAG actively 152 directs cleaved signal and coding ends into the NHEJ repair pathway for signal- and coding-joint 153 formation. In contrast, the RAG transposon strongly favors transposition, but allows some TIR-154 TIR joint formation [24,32,33]. It is possible that the ancestral transposase partially prevented the 155 interaction between the TIR and the NHEJ repair pathway, and that the RAG in jawed vertebrates 156 lost this property, although this remains unknown. In vitro approaches to study this mechanism 157 revealed important amino acid positions in RAG proteins involved in suppressing transposition 158 [33], which would be important for avoiding harmful effects in the organism.

The biochemical functions of the DDE transposome and the vertebrate RAG system (a sequencespecific 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].

166 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 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 behaviors 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

175 that these DDE transposons have apparently evolved in a horizontal manner, which contrasts with 176 the RAG transposon, which evolved in a vertical manner. In contrast, the phylogenetic analysis of 177 RAG transposon and vertebrate RAG sequences, as well as sequences belonging to the RAG 178 family with unknown status and fossilized RAG transposons, shows a sequence tree topology 179 following the species phylogenetic tree [48,49]. The phylogenetic reconstruction also indicates 180 that the RAG structure appeared at least at the origin of the bilaterians. Therefore, the RAG 181 transposon appears to have been active since its birth in the bilaterians' ancestor and co-opted as a 182 specific endonuclease in the jawed vertebrate ancestor. The presence of the RAG transposon that 183 was inherited in the genome from one generation to the next may have increased the likelihood 184 that it would be co-opted compared to other transposons that evolve(d) by horizontal transmission 185 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** -- a host mechanism against transposable elements [50,51]. The RAG transposon is able to transpose within a genome [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, it 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 **Transib**, while the RAG2 in the RAG transposon derived 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 from 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

201 was retained for an unknown reason, perhaps including another function for the host.

202 Consequently, we propose the following conjectural scenario to enhance the published model 203 [26]: i) some time ago, there was an insertion of a complete RAG transposon (or possibly the 204 corresponding miniature inverted-repeat transposable element (MITE, i.e. the TIR of the RAG 205 transposon) that separated an IgV domain (already involved in immune recognition) into V and J 206 segments; ii) after insertion of the complete transposon, the transposase was lost, leaving the 207 native TIRs between the V and J segments intact, while a transposase from another RAG 208 transposon was used, and which in turn, lost its TIR; iii) The TIR-like sequence could be 209 recognized by the RAG transposase and excised along with the internal sequence, leaving 210 hairpin-tipped ends on the flanking segments. These segments could then be processed via 211 Artemis, opening the hairpins asymmetrically, followed by the DNA repair system leading to 212 palindromic (P) diversity. The ability to generate diversity would have then increased with the 213 duplication of the VJ unit (V-TIR-TIR-J) and the co-option of a TdT gene. The system later 214 became more complex, as described by others [55].

Of note, the transposon and its corresponding MITE would have 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 neutral, and it is possible that the insertion into a genetic system was positively selected that was already engaged in non-self-recognition. 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 in supplementary material).

221

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

225 The classical class I/II genes of the MHC are highly polymorphic, encoding proteins that bind 226 processed peptides within the cell, move to the cell surface, and then interact with TCRs 227 expressed on the surface of T-cells. Each MHC allelic form can bind many peptides, both self and 228 non-self, with a specific amino acid motif. Most developing T-cells bearing TCRs that react with 229 self-MHC molecules bound to self-peptides, are eliminated during maturation in the thymus. 230 During infection, both self and non-self-peptides are presented by MHC proteins, with non-self-231 peptides recognized by TCRs on T-cells, which activate the immune system to respond in a 232 variety of ways. These MHC genes evolved in the ancestor of jawed vertebrates in roughly the 233 same time window as the RAG/V(D)J generation of somatic diversity [1,56]. Various hypotheses 234 have been proposed for the origin of MHC genes (see Box 4). In this speculative review, we also 235 propose a scenario whereby MHC genes evolved from PRRs from the innate immune system.

236 The first part of our hypothesis is that the ancestral MHC-like molecule could have bound some 237 PAMPs, presenting them to ancestral TCR-like molecules. The ancestral MHC-like molecule may 238 have been limited to just a few pathogens, and each ancestral TCR-like molecule may have only 239 recognized a particular class of PAMP bound to the ancestral MHC-like molecule. Thus, if a 240 mutation of the ancestral MHC-like molecule allowed binding of a new PAMP, this combination 241 might not be recognized by the ancestral BCR/TCR-like molecules (even if they were encoded by 242 a multigene family); therefore, the new MHC-like molecule might not be selected and the mutant 243 gene could be lost by genetic drift. In fact, if the new MHC-like molecule lost binding to the 244 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.

253 As a third part of this hypothesis, we posit that the ancient MHC molecule was selected to bind 254 many peptides to allow the recognition of numerous pathogens, possibly via the appearance of 255 allelic polymorphisms and peptide-binding promiscuity (as well as from the generation of 256 multigene families). Both allelic polymorphisms and promiscuity are properties of MHC 257 molecules encoded by a single gene, and both extend the number of peptides that can be bound, 258 and thus, the number of pathogens that can be recognized [57-59]. If a particular MHC molecule 259 only bound a limited number of peptides, then a new pathogen would not be recognized by the 260 MHC/TCR system unless a mutation occurred in MHC genes; thus, such a mutation would be 261 selected to deal with the new pathogen. However, the mutation might prevent the new molecule 262 from binding previously-bound peptides, so that the host would be vulnerable to the original 263 pathogen still in the environment. In order to deal with both old and newly-arising pathogens, 264 pathogen-mediated selection leads to allelic polymorphism [57-59]. Another way to increase 265 recognition of new pathogens would be to increase the range of peptides bound, and such 266 promiscuity can be an important feature of MHC molecules [59-61]. A third way to increase 267 recognition of new pathogens would be to increase the number of MHC genes, but the need to 268 avoid recognition of self-peptides might limit the size of the MHC multigene family (although 269 there are theoretical arguments to the contrary) [62-64].

270

271 Concluding Remarks

272 We propose a model whereby the ancestral MHC-like molecule harbored an innate immune 273 function, but when ancestral TCR-like molecules began to diversify due to RAG domestication, 274 thus increasing their recognition potential, ancient MHC molecules might have increased their 275 peptide-binding capacity through increased promiscuity. However, the peptide-binding capacity 276 may have still been low, compared to the recognition capacity of the TCR; therefore, allelic 277 polymorphism may have evolved via pathogen-mediated selection. As this hypothesis begins with 278 the recognition of PAMPs, for which LRR-containing molecules such as TLRs are major players, 279 a similar scenario might be envisaged for the VLR system, based on LRRs. Thus far, no 280 equivalent of an MHC molecule in cyclostomes has been reported (Box 1), but some analogous 281 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 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, also the research 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.

289 References

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

310	10.	Salamat, Z. and Sullivan, J.T. (2009) Involvement of protein kinase C signalling and
311		mitogen-activated protein kinase in the amebocyte-producing organ of Biomphalaria
312		glabrata (Mollusca). Dev. Comp. Immunol. 33, 725–727
313	11.	Milutinović, B. and Kurtz, J. (2016) Immune memory in invertebrates. Semin. Immunol.
314		28, 328-342
315	12.	Glover, L. et al. (2016) VEX1 controls the allelic exclusion required for antigenic
316		variation in trypanosomes. Proc. Natl. Acad. Sci. USA. 113, 7225-7230
317	13.	Monahan, K. and Lomvardas, S. (2015) Monoallelic expression of olfactory receptors.
318		Annu. Rev. Cell Dev. Biol. 31, 721-740
319	14.	Conticello, S.G. (2008) The AID/APOBEC family of nucleic acid mutators. Genome
320		<i>Biol.</i> 9, 229
321	15.	Arakawa, H. et al. (2002) Requirement of the activation-induced deaminase (AID) gene
322		for immunoglobulin gene conversion. Science 295, 1301-1306
323	16.	Rogozin, I.B. et al. (2007) Evolution and diversification of lamprey antigen receptors:
324		evidence for involvement of an AID-APOBEC family cytosine deaminase. Nat. Immunol.
325		8,647–656
326	17.	Liu, M.C. et al. (2018) Diversification of AID/APOBEC-like deaminases in metazoa:
327		multiplicity of clades and widespread roles in immunity. Nat. Comm. 9, 1948
328	18.	Krishnan, A. et al. (2018) Diversification of AID/APOBEC-like deaminases in metazoa:
329		multiplicity of clades and widespread roles in immunity. Proc. Natl. Acad. Sci. U S A.
330		115, E3201-E3210

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

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

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

395 47. Peccoud, J. et al. (2017) Massive horizontal transfer of transposable elements in insects. 396 Proc. Natl. Acad. Sci. USA. 114, 4721-4726 397 48. Morales Poole, J.R. et al. (2017) The RAG transposon is active through the deuterostome 398 evolution and domesticated in jawed vertebrates. Immunogenetics 69, 391-400 399 49. Martin, E.C. et al. (2020) Evidence for an ancient bilaterian origin of the RAG-like 400 transposon. Mobile Elements, in press 401 50. Aravin, A.A. et al. (2007) The Piwi-piRNA pathway provides an adaptive defense in the 402 transposon arms race. Science 318, 761-764 403 51. Lerat, E. et al. TEtools facilitates big data expression analysis of transposable elements 404 and reveals an antagonism between their activity and that of piRNA genes. Nucleic Acids 405 Res. 45, e17 406 52. Kapitonov, V.V. and Jurka, J. (2005) RAG1 core and V(D)J recombination signal 407 sequences were derived from Transib transposons. PLoS Biol. 3, e181 408 53. Kapitonov, V.V. and Koonin E.V. (2015) Evolution of the RAG1-RAG2 locus: both 409 proteins came from the same transposon. Biol. Direct. 10, 20 410 54. Callebaut, I. and Mornon, J.P. (1998) The V(D)J recombination activating protein RAG2 411 consists of a six-bladed propeller and a PHD fingerlike domain, as revealed by sequence 412 analysis. Cell. Mol. Life Sci. 54, 880-891 413 55. Hsu, E., and Lewis, S.M. (2015) The origin of V(D)J diversification. In: Molecular 414 Biology of B cells, Alt, F.W., Honjo, T., Radbruch, A., and Reth, M., eds. Elsevier, 415 Academic Press, Amsterdam, pp. 133-148.

- 416 56. Danchin E. *et al.* (2004) The major histocompatibility complex origin. *Immunol. Rev.*417 198, 216-232
- 418 57. Spurgin, L.G. and Richardson, D.S. (2010) How pathogens drive genetic diversity: MHC,
 419 mechanisms and misunderstandings. *Proc. Biol. Sci.* 277, 979–988
- 420 58. Radwan, J. *et al.* (2020) Advances in the evolutionary understanding of MHC
 421 polymorphism. *Trends. Genet.* 36, 298-311
- 422 59. Kaufman, J. (2018) Generalists and Specialists: A New View of How MHC Class I Mol423 ecules 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
- 426 61. Manczinger, M. *et al.* (2019) Pathogen diversity drives the evolution of generalist MHC427 II alleles in human populations. *PLoS Biol.* 17, e3000131
- 428 62. Vidović, D. and Matzinger, P. (1988) Unresponsiveness to a foreign antigen can be
 429 caused by self-tolerance. *Nature* 336, 222-225
- 430 63. Nowak, M.A. *et al.* (1992) The optimal number of major histocompatibility complex
 431 molecules in an individual. *Proc. Natl. Acad. Sci. U S A.* 89, 10896-10899
- 432 64. Borghans, J.A. *et al.* (2003) Thymic selection does not limit the individual MHC
 433 diversity. *Eur. J. Immunol.* 33, 3353-3358
- 434 65. Teng, G. and Schatz, D.G. (2015) Regulation and Evolution of the RAG Recombinase.
 435 *Adv. Immunol.* 128, 1-39
- 436 66. Cardarelli, L. *et al.* (2015) Two Proteins Form a Heteromeric Bacterial Self-Recognition

- 437 Complex in Which Variable Subdomains Determine Allele-Restricted Binding. *Mbio.* 6,438 e00251
- 439 67. De Tomaso A.W. (2018) Allorecognition and Stem Cell Parasitism: A Tale of
 440 Competition, Selfish Genes and Greenbeards in a Basal Chordate. In: Pontarotti P. (eds)
 441 Origin and Evolution of Biodiversity. Springer.
- 442 68. Espinosa, A. and Paz-Y-Miño-C, G. (2014) Evidence of Taxa-, Clone-, and Kin443 discrimination in Protists: Ecological and Evolutionary Implications. *Evol. Ecol.* 28,
 444 1019-1029
- 445 69. Fujii, S. *et al.* (2016) Non-self- and self-recognition models in plant self-incompatibility.
 446 *Nat. Plants* 2, 16130
- 447 70. Gruenheit, N. *et al.* (2017) A polychromatic 'greenbeard' locus determines patterns of
 448 cooperation in a social amoeba. *Nat. Comm.* 8, 14171
- 449 71. Harada, Y. *et al.* (2008) Mechanism of self-sterility in a hermaphroditic chordate. *Science*450 320, 548-550
- 451 72. Heller, J. *et al.* (2016) Greenbeard Genes Involved in Long-Distance Kind Discrimination
 452 in a Microbial Eukaryote. *PLoS Biol* 14, e1002431
- 453 73. Kües, U. (2015) From two to many: Multiple mating types in Basidiomycetes. *Fungal*454 *Biology Reviews* 29, 126-166
- 455 74. Paoletti, M. (2016) Vegetative incompatibility in fungi: From recognition to cell death,
 456 whatever does the trick. *Fungal Biology Reviews* 30, 152-162
- 457 75. Rosengarten, R.D. and Nicotra, M.L. (2011) Model systems of invertebrate

458 allorecognition. Curr. Biol. 21, R82-92 459 76. Saak, C.C. and Gibbs, K.A. (2016) The Self-Identity Protein IdsD Is Communicated 460 between Cells in Swarming Proteus mirabilis Colonies. J. Bacteriol. 198, 3278-3286 461 77. Wall, D. (2016) Kin Recognition in Bacteria. Annu. Rev. Microbiol. 70, 143-160 462 78. Boehm ,T. (2006) Quality control in self/nonself discrimination. Cell 125, 845-858 463 79. Ruff, J.S. et al. (2012) MHC signaling during social communication. Adv. Exp. Med. 464 Biol. 738, 290-313 465 80. Leinders-Zufall, T. et al. (2009) Structural requirements for the activation of vomeronasal 466 sensory neurons by MHC peptides. Nat. Neurosci. 12, 1551-1558 467 81. Flajnik, M.F. et al. (1991) Which came first, MHC class I or class II? Immunogenetics 33, 468 295-300 469 82. Credle, J.J. et al. (2005) On the mechanism of sensing unfolded protein in the 470 endoplasmic reticulum. Proc. Natl. Acad. Sci. USA. 102, 18773-18784 471 83. Karagöz, G.R.E. et al. (2017) An unfolded protein-induced conformational switch 472 activates mammalian IRE1. eLIFE 6, e30700 84. Dijkstra, J.M. and Yamaguchi. T. (2019) Ancient features of the MHC class II 473 474 presentation pathway, and a model for the possible origin of MHC molecules. 475 Immunogenetics 71, 233-249 476 85. Du Pasquier, L. (2000) The phylogenetic origin of antigen-specific receptors. Curr. Top. 477 Microbiol. Immunol. 248, 160–185

478	86. Kaufman, J.F. et al. (1984) The class II molecules of the human and murine major
479	histocompatibility complex. Cell 36, 1-13
480	87. Xiao, J. et al. (2018) An Invariant Arginine in Common with MHC Class II Allows
481	Extension at the C-Terminal End of Peptides Bound to Chicken MHC Class I. J.
482	Immunol. 201, 3084-3095
483	88. Janeway, C.A. et al. (2001) Immunobiology: the immune system in health and disease. 5 th
484	edition. New York: Garland Science
485 486 487	89. Adrion JR, Song MJ, Schrider DR, Hahn MW, Schaack S.Genome-Wide Estimates of Transposable Element Insertion and Deletion Rates in Drosophila Melanogaster. <i>Genome Biol Evol.</i> 2017 May 1;9(5):1329-1340.
488	
489	

490 Acknowledgements

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

496 Figure legends

497

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

505 Figure 2. Antibody/TCR receptor somatic diversity might drive the appearance of MHC 506 promiscuity and polymorphisms. We propose a hypothetical model whereby the ancestral 507 MHC-like molecule bound certain PAMPs, presenting them to ancestral TCR-like molecules. The 508 ancestral MHC-like molecule may have been limited to just a few pathogens, and each ancestral 509 TCR-like molecule might have only recognized a particular class of PAMP bound to the ancestral 510 MHC-like molecule. A mutation in the ancestral MHC-like molecule may have allowed binding 511 of a new PAMP, but this new combination could not be recognized by the ancestral TCR-like 512 molecule. As a result, the new MHC-like molecule would be lost by genetic drift. The integration 513 of the RAG transposon into an ancestral TCR-like gene may have led to a significantly increased 514 probability of recognition by these original, non-diverse TCRs. Thus, mutations in the ancestral 515 MHC-like molecule may have led to a conformational ability to bind new PAMPs; consequently, 516 mutated MHC molecules could have then been recognized by diverse TCRs thereafter, becoming 517 evolutionarily selected via three mechanisms, peptide-binding promiscuity, allelic polymorphism, 518 and expansion into multigene families. The expanded ability of this ancient MHC/TCR system to 519 recognize new PAMPs would presumably allow peptides to be bound, presented, and recognized 520 during an immune response.

521 **Box 1.** Brief overview of the adaptive immune system in vertebrates

522 The jawed vertebrate immune system is based on a complex cellular system comprising T-cells 523 and B-cells with immunoglobulin (Ig) domain-containing receptors, and/or secreted proteins, 524 including antibodies, and both kinds of T-cell receptors (TCR: those composed of α and β chains, 525 and those composed of γ and δ chains). The generation of antigen receptor diversity is driven by 526 the recombination activating genes, RAG1 and RAG2. Each unique receptor is expressed by a 527 different cell clone through the action of allelic exclusion. In jawless fish (agnatha or 528 cyclostomes), the other living vertebrate phylum, the receptors are based on the leucine-rich 529 repeat (LRR) module, and include variable lymphocyte receptor-A (VLR-A), VLR-B and VLR-C. 530 The generation of diversity occurs via gene conversion driven by a protein from the AID-531 APOBEC family, but again, unique receptors are expressed by different clones with 532 transcriptomic profiles, much like jawed vertebrate lymphocytes: VLR-A, like $\alpha\beta$ T-cells, VLR-533 B, like B-cells, and VLR-C, like $\gamma\delta$ T-cells [4,32]. In jawed vertebrates, $\gamma\delta$ T-cells bind various 534 cell surface molecules, but $\alpha\beta$ TCRs recognize peptides bound specifically to MHC molecules; 535 whether there is a functional equivalent of MHC molecules in jawless fish remains unclear.

537 Box 2

538 Evolution of two systems of adaptive immunity in vertebrates

539 An important question concerns the origin of the complexity of cells involved in adaptive 540 immunity. Both molecular systems with somatic diversification (VLR/AID and VDJ/RAG) could 541 have been in place along with a pre-adaptive immune system [2,4,5]. Then, the two molecular 542 systems might have evolved in an independent manner in the two vertebrate lineages, jawless fish 543 and jawed vertebrates. The mechanism of diversity generation is similar in both vertebrate 544 lineages, starting with a DNA double-strand break (DSB) in the region involved in DNA 545 recognition, followed by gene repair from either non-homologous end-joining (NHEJ) 546 mechanisms or gene conversion [15,16]. The DSB in cyclostomes (and some jawed vertebrates) 547 is due to an enzyme from the AID/APOBEC family and repair by gene conversion events, while 548 the DSB in most jawed vertebrates is due to the RAG sequence-specific endonuclease, followed 549 by DNA repair through a NHEJ mechanism.

550 It is important to note that the function of possible T- and B-cell lineages prior to the time 551 adaptive immunity arose is entirely unclear. Innate lymphoid cells (ILCs) found in mammals are 552 potential candidates for the functions of non-adaptive T cells before adaptive immunity (although 553 they could also be a novelty of placental mammals), but system replacement might be more likely 554 [48]. If the first adaptive immune system was based on VLR, then, in jawed vertebrates, a shift 555 might have occurred from IgV-IgC innate immunity, to IgV-IgC adaptive immunity, followed by 556 the loss of VLR-based adaptive immunity. If the first adaptive immune system was based on IgV-557 IgC, then, in cyclostomes, the reverse may have occurred. In fact, such replacements have been 558 noted for natural killer (NK) cell receptors [2]: at least three families of NK cell receptors exist 559 with analogous functions: lectin-like receptors (overwhelmingly in rodents and to a lesser extent 560 in certain other mammals), Ig-like receptors of the KIR family (one or another of the KIR subfamilies, as in humans and other mammals) and a completely different family of Ig-like receptorsin bony fish.

Box 3. Emergence of Rearranging B- and T-cell receptors and Brief History of the Origin of RAG

565 The antibody and TCR genes of jawed vertebrates are assembled from variable (V), diversity (D), 566 and joining (J) gene segments during B- and T-lymphocyte development to give contiguous VJ 567 and V(D)J sequences. The process to excise the DNA between the gene segments is initiated by 568 the RAG endonuclease. The RAG endonuclease specifically recognizes recombination signal 569 sequences (RSSs) that flank each gene segment. RSSs are composed of conserved heptamer and 570 nonamer sequences separated by a less conserved spacer sequence of either 12 or 23 bp (12RSS 571 and 23RSS). RAG-mediated DNA cleavage occurs preferentially in a complex containing a 572 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 [65]].

The discovery of RSSs, along with the mechanism of RSS cleavage (similar to several cut-andpaste DNA transposases (DDE transposases)) [20,21] led to the hypothesis that a DDE transposon invaded an Ig-like gene, eventually resulting in 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 (corresponding 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 the hemichordate *Ptychodera flava* [48], as well as fossilized transposons in several deuterostomes [26,48,53] and protostomes
[49], indicates that the RAG transposon was present at least as far back as the bilaterian ancestor,
remained active in several lineages, and was co-opted as part of V(D)J recombination machinery
in jawed vertebrates [48,49].

591

592 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 in this process, and the putative mechanisms, remain unclear [58,80].

600 Various hypotheses have been proposed for the origin of MHC genes. One suggestion was that 601 chaperone genes gave rise to the peptide-binding domains characteristic of MHC molecules [81]. 602 Although subsequent structural analysis of HSP70 rules out the specific example suggested by 603 these authors [2], it remains possible that a different ancient chaperone could be the ancestor. 604 Another candidate is IRE1, which is involved as a sensor in the unfolded protein response, and 605 has a structure and peptide-binding properties like MHC molecules [82,83]. A recent suggestion 606 was that the primordial MHC-like molecule evolved from a heavy chain-only antibody molecule 607 that cycled between endosomal compartments and the surface [84]. Another hypothesis was that 608 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 bearing leucine as a starting amino acid, and which appear in stressed cells [2].

614 A linked issue is whether primordial MHC genes and molecules were organized as in the class I 615 or class II systems. A scenario based on structure is that the original MHC molecule was a 616 homodimer of class II β -like chains, with gene duplication and divergence giving rise to 617 heterodimers of class II α -like and β -like chains, followed by an inversion leading to a class I 618 heavy-like chain and a β_2 -microglobulin chain with a transmembrane region, and a subsequent 619 mutation to give rise to a class I-like molecule [2,86]. A scenario based on function suggested the 620 transfer of a peptide-binding region from a chaperone in front of an IgC-like region to produce a 621 class I-like heavy chain first [79]. Recent evidence for highly promiscuous peptide binding and 622 C-terminal protrusions of peptides from the groove of chicken class I molecules renders the 623 differences between class I and II molecules less clear [59,60,87].

624 Glossary

625

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

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

631 Bilaterian: metazoan animals with a bilaterally symmetric body plan, including the protostomes632 and deuterostomes.

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

634 Deuterostome: clade of animals including jawed vertebrates, jawless fish, cephalochordates

635 (such as amphioxus), urochordates, hemichordates and echinoderms (such as sea urchins); the636 sister group of protostomes within bilaterians.

637 DDE transposon (aka class II transposon): DNA fragment formed by two terminal inverted 638 repeats surrounding a sequence coding for the transposase gene. The transposase gene is 639 expressed and translated by the host cell, recognizes and cuts the TIR to excise the transposon. 640 The broken chromosome ends are then repaired and the transposon inserts at another site in the 641 genome.

642 Genetic drift: mechanism of evolution in which allele frequencies of a population change over643 generations due to chance.

644 Junctional diversity: during somatic V(D)J recombination, different variable segments of TCR

645 and antibody genes are rearranged by introducing double-strand breaks between the required

646 segments, which form hairpin loops at the ends. The hairpins are cleaved in an asymmetric

647 manner by Artemis, followed by joining of the broken genomic region with variable addition or

648 subtraction of nucleotides to generate junctional diversity.

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

650 Miniature Inverted-repeat Transposable Elements (MITEs): non-autonomous DDE
651 transposon; does not code for a transposase and thus, must use a transposase encoded by another
652 transposon.

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

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

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

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

663 **Piwi-interacting RNA (piRNA)**: family of small non-coding RNA molecules that interact with 664 piwi-subfamily Argonaute proteins, forming piRNA complexes which are involved in the 665 epigenetic and post-transcriptional silencing of transposable elements, and the regulation of other 666 genetic elements in germline cells.

667 Protostome: clade of animals including mainly arthropods, annelids, and molluscs; sister group668 of the deuterostomes with bilaterians.

669 **Recombination activating genes (RAGs)** are two host genes located next to each other that 670 encode RAG1 and RAG2 proteins; these, as a complex, initiate the rearrangement of gene

671 segments encoding antibody and TCR molecules.

672 **RAG DDE transposon:** RAG-like sequence found in non-vertebrates, functioning as transposon.

673 Somatic diversification: the process of mutation in somatic cells, e.g. genomic rearrangement

674 Terminal deoxytransferase (TdT): enzyme that randomly adds nucleotides to untemplated

broken ends of DNA, particularly during somatic diversification of antibody and TCR genes

- 676 Toll-like receptor (TLR): one class of PRRs involved in initiating innate immune responses
- 677 Transib: DDE transposon from protostomes whose transposase gene is closest to RAG1, and
- 678 whose TIR is similar to the RAG transposon, and the V(D)J RSSs .

Highlights

RAG evolved from a DDE transposon present in the ancestor of bilaterian animals; 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 with an 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 recognition module of the TCR-like gene may have led to a significant increase in recognition possibilities, presumably allowing 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 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?

Figure 1



RAG IN THE V(D)J RECOMBINATION



- Variable (V) segments
- Diversity (D) segments
- Joining (J) segments
- Constant region exons
- RSS (Recombination Signal Sequence)

- TIR (Terminal Inverted Repeat)
- **TSD** (Target Site Duplication)



- NHEJ repair machinery
- Endonuclease activity





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

The average transposition for a given DDE transposon per genome is approximately **10⁻⁴**/year [89]. The generation time is approximately 1 year in average for Deuterostomia (calculated on the average generation age of the deuterostomian) and might be estimated as **10**, including (TIR-----TIR) (MIR), and potentially 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 V(D)J recombinase was approximately **200 million years** (the difference between the time the RAG transposon appeared in the ancestor of deuterostomes, and the time of its co-option in the jawed vertebrate ancestor)[48]. The number of possible positions per gene V is approximately **250**, to have a J sequence of at least 50 nucleotides [88]. We could estimate that **100** copies of V genes were present. Thus, the number of possible transposition events on a V gene might be:

 $10^{-4} \text{ x } 2(10)^8 \text{ x } 10 \text{ x } 250 \text{ x } 100 = 5 \text{ x10}^9.$

The size of a deuterostomian genome is on 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 of not observing it:

 $1 - (499999999/50000000) > 5(10)^9 = 1\%.$

Thus, the probability that the event happened might then be: 99%.

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

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

89. Adrion JR, Song MJ, Schrider DR, Hahn MW, Schaack S.Genome-Wide Estimates of Transposable Element Insertion and Deletion Rates in Drosophila Melanogaster. *Genome Biol Evol*. 2017 May 1;9(5):1329-1340.