Molecular Immunology 117 (2020) 131–138

Contents lists available at ScienceDirect

Molecular Immunology

journal homepage: www.elsevier.com/locate/molimm

Short communication

Analyses of RAG1 and RAG2 genes suggest different evolutionary rates in the Cetacea lineage

Bruna C. Dias, Mariana F. Nery*

Laboratory of Evolutionary Genomics, State University of Campinas, Department of Genetics, Evolution, Microbiology and Immunology, Campinas, São Paulo, Brazil

ARTICLE INFO

Keywords:
Evolution
Molecular evolution
Colonization
Ecological change
Immune system

ABSTRACT

V(D)J recombination is a process of somatic recombination catalyzed by proteins encoded by RAG1 and RAG2 genes, both restricted to the genome of jawed vertebrates. Their proteins constitute the enzymatic core of V(D)J recombination machinery and are crucial for jawed vertebrate adaptive immunity. Mammals possess great ecological diversity, and their complex evolutionary history associated with radiation to different environments presented many distinct pathogenic challenges from these different habitats. Cetaceans comprise a mammalian order of fully aquatic mammals that have arisen from a complete terrestrial ancestor and, accordingly, was confronted with challenges from changing environmental pathogens while they transitioned from land to sea. In this study we undertook molecular evolutionary analyses of RAG1 and RAG2 genes, exploring the possible role of natural selection acting on these genes focusing on the cetacean lineage. We performed phylogenetic reconstructions on IQ-TREE, together with selection analyses in the codeml program of the PAML package, and in the FITMODEL program for codon evolution and switching on both the RAG1 and RAG2 genes. Our findings demonstrate that RAG1 and RAG2 remained fairly conserved among tetrapods, with purifying selection acting on both genes, with evidence for a few punctuated shifts in nucleotide substitution rates of both genes along tetrapod evolution. We demonstrate differential evolution in the closely linked genes RAG1 and RAG2 specifically in cetaceans.

1. Introduction

The adaptive immune system became possible after the acquisition of a retroposon that, millions of years ago, invaded the genome of an early vertebrate, as only vertebrates have both of the elements of the retroposon: two sites of recognition signal sequences (RSSs) and the presence of recombination-activation genes (Schatz et al., 1989; Janeway, 2001). These genes, known as RAG1 and RAG2, encode a site-specific recombinase that acts on germline gene segments to produce all immunoglobulin molecules and T cell receptors of the adaptive immune system (Schatz et al., 1989; Janeway, 2001). The RAG transposon domestication model predicts a critical divergence during chordate evolution in which, in jawed vertebrates, the RAG transposase acquired properties of a recombinase, whereas in amphioxus, transposase functions were retained (Zhang et al., 2019).

The jawed vertebrate adaptive immune system relies on a diverse array of immunoglobulins (Ig) and T-cell antigen receptors (TCRs) for specific recognition of antigens (Agrawal et al., 1998; Teng and Schatz, 2015). This system consists of cells that provide pathogen specific immunity to the host through somatic rearrangement of antigen receptor genes (Thompson, 1995). Each Ig or TCR polypeptide consists of a constant region and a variable region, and it is in the variable region that the antigen recognition and specificity is determined (Schatz, 2004). In the germ line, the variable region is encoded by non-contiguous gene portions split into V (variable), J (joining) and, in some cases, D (diversity) segments. Each of these gene segments is joined in a site-specific recombination reaction, known as V(D)J recombination, to form the exon that encodes the antigen-binding portion of the polypeptide (Agrawal et al., 1998; Schatz, 2004).

The discovery of RAG genes is considered a hallmark of adaptive immunity, as their proteins constitute the enzymatic core of V(D)J recombination machinery (Fugmann et al., 2006; Kapitonov and Koonin, 2015; Poole et al., 2017). The RAG1-RAG2 complex catalyzes random assembly of V, D, and J gene segments that are present in the genome in numerous copies and generate the enormous variety of the assembled antibodies and antigen receptors (Gellert, 2002; Kapitonov and Koonin, 2015). RAG1 and RAG2 genes are closely linked in the human genome, located just 8 kilobases apart from one another, and they are convergently transcribed (Oettinger et al., 1990; Greenhalgh et al., 1993).

Among jawed vertebrates, mammals possess great ecological

* Corresponding author.
E-mail address: mariananery@gmail.com (M.F. Nery).

https://doi.org/10.1016/j.molimm.2019.10.014
Received 18 July 2019; Received in revised form 14 October 2019; Accepted 23 October 2019

diversity, and their complex evolutionary history associated with their radiation to different environments, might have depicted many distinct pathogenic challenges from different habitats (Tian et al., 2018). The aquatic environment was one of these habitats colonized by different extant mammalian lineages five times. Cetaceans comprise a mammalian order of fully aquatic mammals, that originated about 50 million years ago in the Eocene epoch from a terrestrial ancestor (Thewissen et al., 2009). In addition to anatomical and physiological innovations required for life in water, cetaceans must have been confronted with challenges from changing environmental pathogens while they transitioned from land to sea (Shen et al., 2012; Ishengoma and Agaba, 2017). These challenges exerted intensified selection pressure on the genomes of colonizing species, especially on those genes and gene families related to the immune system, as already reported (e.g. Haldane, 1949; Wlasiuk and Nachman, 2010; Areal et al., 2011; Ishengoma and Agaba, 2017).

In this context, cetaceans comprise ideal candidate taxa to study the molecular evolutionary mechanisms behind the vertebrate immune system, since this lineage has experienced a radical habitat change during its evolutionary history. Accordingly, the aim of this study was to investigate the molecular evolution of the RAG1 and RAG2 genes in a phylogenetic framework, exploring the possible role of natural selection acting on these genes focusing on cetaceans in comparison to their terrestrial counterparts.

2. Material and methods

2.1. Phylogenetic reconstruction

Vertebrate RAG1 and RAG2 coding sequences were retrieved from GenBank (NCBI) and Ensembl public databases. All accession numbers are depicted in Supplementary Table 1. A total of 49 sequences of terrestrial vertebrates, 7 other marine mammals and 8 cetacean sequences were selected to perform the analyses. Nucleotide and amino acid sequences were aligned using MUSCLE software (Edgar, 2004). We used the PAL2NAL program to generate a codon alignment, and this alignment was used to estimate the type and rate of nucleotide substitutions in coding DNA (Suyama et al., 2006). The phylogenetic tree was reconstructed using IQ-TREE (Nguyen et al., 2015) software using a maximum likelihood approach, and implemented the fast method of substitution model selection with ModelFinder program (Kalyaanamoorthy et al., 2017). The ultrafast bootstrap approximation (UFBoot) (Minh et al., 2013) with 1000 replicates, also was implemented in the IQ-TREE software package. The models MGK + F3 × 4 + G4 and GY + F1 × 4 + G4 were used for RAG1 and RAG2 respectively as selected by Bayesian Information criterion (BIC).

2.2. Selection analyses

To test for signatures of positive selection and to infer sites under positive selection, we used the codeml program in the PAML 4.7 package (Yang, 2007) and FITMODEL (Guindon et al., 2004), which implements Markov-modulated models of codon evolution or switching site evolution models on RAG1 and RAG2 genes separately.

In PAML, we applied branch models, which allow the ω ratio to vary among branches in the phylogeny and are useful for detecting positive selection acting on particular lineages (Yang, 1998; Yang and Nielsen, 1998). Within branch model strategy, we estimated the likelihood of the free-ratio model, which assumes an independent ω ($=d_N/d_S$) ratio for each branch; the one-ratio model, which estimates a unique ω value for all branches along the tree; and the two-ratio model, which estimates one ω for the Cetacea lineage and another for the rest of the phylogeny. Also, we tested the branch-site model, that attempts to detect positive selection that affects only a few sites along a specified branch (Zhang et al., 2005). The branch-site analysis divides the tree into foreground branches (Cetacea), where sites may be under positive selection, and background branches where positive selection is absent (the rest of the phylogeny) (Yang and Nielsen, 2002; Yang et al., 2005; Zhang et al., 2005). Under this model, sites are categorized into four classes 0, 1, 2a, and 2b. Site class 0 includes codons that evolve under purifying selection on both the foreground and background branches, with $0 < \omega_0 < 1$. In site class 1, codons evolve neutrally in all lineages, with $\omega_1 = 1$, whereas in classes 2a and 2b, positive selection is allowed on the foreground branches with $\omega_2 > 1$, but not on the background branches. This model is compared with the corresponding null hypothesis of neutral evolution, where ω_2 is fixed to 1. If the null hypothesis is rejected by the likelihood ratio test (LRT), a Bayes empirical Bayes approach is used to calculate the posterior probabilities that each site has evolved under positive selection on the foreground lineage (Yang et al., 2005).

To further investigate the molecular evolution of RAG1 and RAG2 genes, we performed likelihood analyses under a nested set of codon-substitution models with FITMODEL version 0.5.3 (Guindon et al., 2004). We used models M0 and M3, and the switching models M3 + S1 and M3 + S2. Model M0 assumes that all sites in a sequence alignment are subject to the same selection process (homogeneous). As implemented in FITMODEL, under the M3 model, variation in selective constraint across sites is modeled as three rate ratio classes with ω1, ω2 and ω3. Switching was modeled as a time-reversible Markov process with three additional parameters: the overall rate of interchange among rate ratio classes (δ), a coefficient for shifts between ω1 and ω3 (α), and a coefficient for shifts between ω2 and ω3 (β). The S1 model implemented in FITMODEL imposes equal switching rates among ω1, ω2 and ω3 rate ratio classes ($\alpha = \beta = 1$), and the S2 model allows α and β to vary freely accounting for unequal rates of switches between selection classes (Guindon et al., 2004).

For both PAML and FITMODEL nested likelihood ratio tests (LRTs) were performed for model comparisons. For PAML models, the LRTs were performed between free-ratio model vs. one-ratio model and two-ratio model vs. one-ratio model (Table 1). For FITMODEL the comparisons were performed between the following models: no rate heterogeneity vs. variation across sites (M0 vs. M3), variation across sites without vs. with switching among substitution rate ratio classes (M3 vs. M3 + S1), and equal switching rates vs. class-dependent switching rates across branches (M3 + S1 vs. M3 + S2) (Table 2). The chi-square test was employed to estimate the statistical difference ($P < 0.05$). Degrees of freedom for each test were equal the difference in the number of parameter estimated for the models under comparison.

Additionally, we used the HyPhy package (Pond and Frost, 2005) in the DataMonkey Server (Weaver et al., 2018) to implement the RELAX model (Wertheim et al., 2014), which tests whether the strength of natural selection has been relaxed or intensified along a specified set of test branches.

Also, it is important to consider that GC-biased gene conversion (gBGC) is a recombination-associated process that causes variation in GC content and has an important influence on substitution patterns, leading to sequence accelerated evolution (Galtier et al., 2009). To test for gBGC on our alignment, we implemented the program phastBias (Capra et al., 2013) available as part of the PHAST software package (Hubisz et al., 2011) that uses a hidden Markov model and statistical phylogenetic models that consider the influence of both natural selection and gBGC on substitution rates and patterns.

3. Results

3.1. Phylogenetic analyses of RAG1 and RAG2 genes

Mammalian RAG1 and RAG2 phylogenetic trees resulted in similar topologies, with occasional differences in some phylogenetic relationships. In both trees (Figs. 1 and 2), mammals form a monophyletic clade, with the lineage of cetaceans being part of Cetartiodactyla, and sister clade of Hippopotamidae. In RAG1, both cetaceans and bovids

Table 1
Likelihood analyses of the branch models and branch-site models in the PAML program for RAG1 and RAG2. Abbreviations of table's cells are as follows: likelihood value (lnL), omega value (ω), number of parameters (np), likelihood ratio test (LRT), degrees of freedom (df) and the P values.

PAML							
BRANCH MODEL							
	Model	lnL	ω	np	LRT	df	P value
RAG1	One-ratio model	−57620.68	0.09074	126	—	—	—
	Two-ratio model	−57221.58	various	249	798.2	123	0
	Free-ratio model	−57599.56	0.08868; 0.22258	127	42.24	1	0
RAG2	One-ratio model	−26095.59	0.13293	126	—	—	—
	Two-ratio model	−25925.59	various	249	340	123	0
	Free-ratio model	−26095.29	0.13250; 0.16241	127	0.6	1	0.43
BRANCH-SITE MODEL							
	Model	lnL	np	LRT	df	P value	
RAG1	Null Model	−56293.55	128	—	—	—	
	Model A	−56293.55	129	0	1	1	
RAG2	Null Model	−25656.34	128	—	—	—	
	Model A	−25655.46	129	1.76	1	0.18	

Table 2
Likelihood analyses and P values of the models in the FITMODEL program for RAG1 and RAG2 sequence data. Abbreviations of table's cells are as follows: likelihood value (lnL), omega value (ω), parameter estimates values (p), switching rates values (R) and P values.

Fitmodel				
	M0	M3	M3 + S1	M3 + S2
RAG1				
lnL	−65937.20	−62608.04	−62441.29	−62381.36
ω1 ω2 ω3	0.3	0.0 0.3 0.9	0.0 0.3 1.1	0.0 0.3 1.3
p1 p2 p3	1.0	0.3 0.3 0.2	0.4 0.3 0.2	0.4 0.3 0.1
R12 R13 R23	—	—	0.15 0.15 0.15	0.19 0.00 1.18
P value	—	0	P < 0.001	P < 0.001
RAG2				
lnL	−31291.80	−30688.42	−30635.45	−30629.07
ω1 ω2 ω3	0.3	0.0 0.3 0.9	0.0 0.3 1.1	0.0 0.7 2.4
p1 p2 p3	1.0	0.3 0.3 0.2	0.4 0.3 0.2	0.1 0.5 0.3
R12 R13 R23	—	—	0.15 0.15 0.15	0.39 0.00 0.70
P value	—	0	P < 0.001	P < 0.001

clade are closer to Camelidae, while in RAG2 they are closer to Suidae. Tree lengths indicate that RAG1 and RAG2 experienced accelerated evolutionary rates in the ancestral lineage of all tetrapods, after their separation from fishes. The lengths of tree branches also show that both genes had greatly accelerated evolutionary rates in the Pinnipedia lineage, with most evolutionary modifications happening in *Arctocephalus* for both genes (Supplementary material, Figs. 1 and 2).

3.2. Identifying sites under positive selection

For RAG1, the model that best fit our data on branch model was the two-ratio model, where the Cetacea lineage has a greater ω value (0.22) when compared to the rest of the tree (0.08), suggesting that cetaceans accumulated more modifications on the RAG1 sequence throughout their evolutionary history (Table 1). The RELAX algorithm implemented on DataMonkey identified a significant relaxation in the selection pressure in the cetacean lineage for this gene, that could explain this evolutionary acceleration. For RAG2, the two-ratio model did not fit the data better, a result that matches the result of RELAX, which did not identify a significant relaxation in the cetacean lineage when compared to other lineages. The results from the branch-site model were not significant for both genes (Table 1), not being able to identify sites under positive selection in cetacean lineage.

3.3. Shift in the site-specific selection process

We carried out maximum-likelihood analysis on FITMODEL software under a set of branch-site codon substitution models in order to investigate the substitution process on RAG1 and RAG2 genes. We performed likelihood analyses under a nested set of codon-substitution models (M3, M0, M3 + S1, M3+s2) (Guindon et al., 2004). Table 2 shows that log likelihoods improved significantly as parameters were added to the nested substitution models ($P < 0.001$). These results suggest that M3 + S2 (unequal switching rates among three ω rate ratio classes) is the best codon substitution model for both RAG1 and RAG2 genes. Under this model, the substitution rate ratio estimated for the three classes were ω1 = 0.0, ω2 = 0.3 and ω3 = 1.3 (Table 2). The switching rate (represented by the R letters on Table 2) between ω2 and ω3 (R23 = 1.18 for RAG1 and R23 = 0.70 for RAG2) was significantly higher than the switching rates between ω1 and ω2 and between ω1 and ω3 (Table 2). These results implies that site-specific shifts between moderate purifying selection (ω2) and relaxed selection (ω3) occurred more frequently than shifts that involves the most highly constrained rate ratio classes for both RAG1 and RAG2 genes. With an ω considerably lower than one for almost all the rate ratio classes, excluding ω3 for both genes, the sum of the parameter estimates (p1, p2 and p3) values of the M3 + S2 model suggests that most sites are under purifying selection for both genes (RAG1 = 80% and RAG2 = 90%; Table 2).

No position in our alignment was identified with probability of being in one of the gBGC states, thus rejecting the action of GC-biased gene conversion in RAG1/RAG2 genes in cetaceans.

4. Discussion

This study represents the first molecular evolution analyses of RAG1 and RAG2 genes focusing on cetaceans. The phylogenetic trees generated depicted short branches for both genes among all cetacean species (Figs. 1 and 2). On both trees, the phylogeny is fully resolved for Cetacea, and forms a definite clade for mammals with the lineage of cetaceans being part of Cetartiodactyla and sister clade of Hippopotamidae. Considering RAG1, Cetartiodactyla clade is better supported in relation to RAG2 phylogeny. For RAG1 gene tree, cetaceans and bovids are closer to camelids. This result does not support previous phylogenetic studies using RAG1 gene, which places cetaceans and bovids closer to suids (Waddell and Shelley, 2003). On the other hand, our RAG2 phylogenetic analyses show cetaceans and bovids closer to suids instead of camelids, which is the most accepted position of this phylogeny considering previous phylogenetic studies (e.g. Waddell and Shelley, 2003; Gatesy et al., 2013). On a previous study of extant

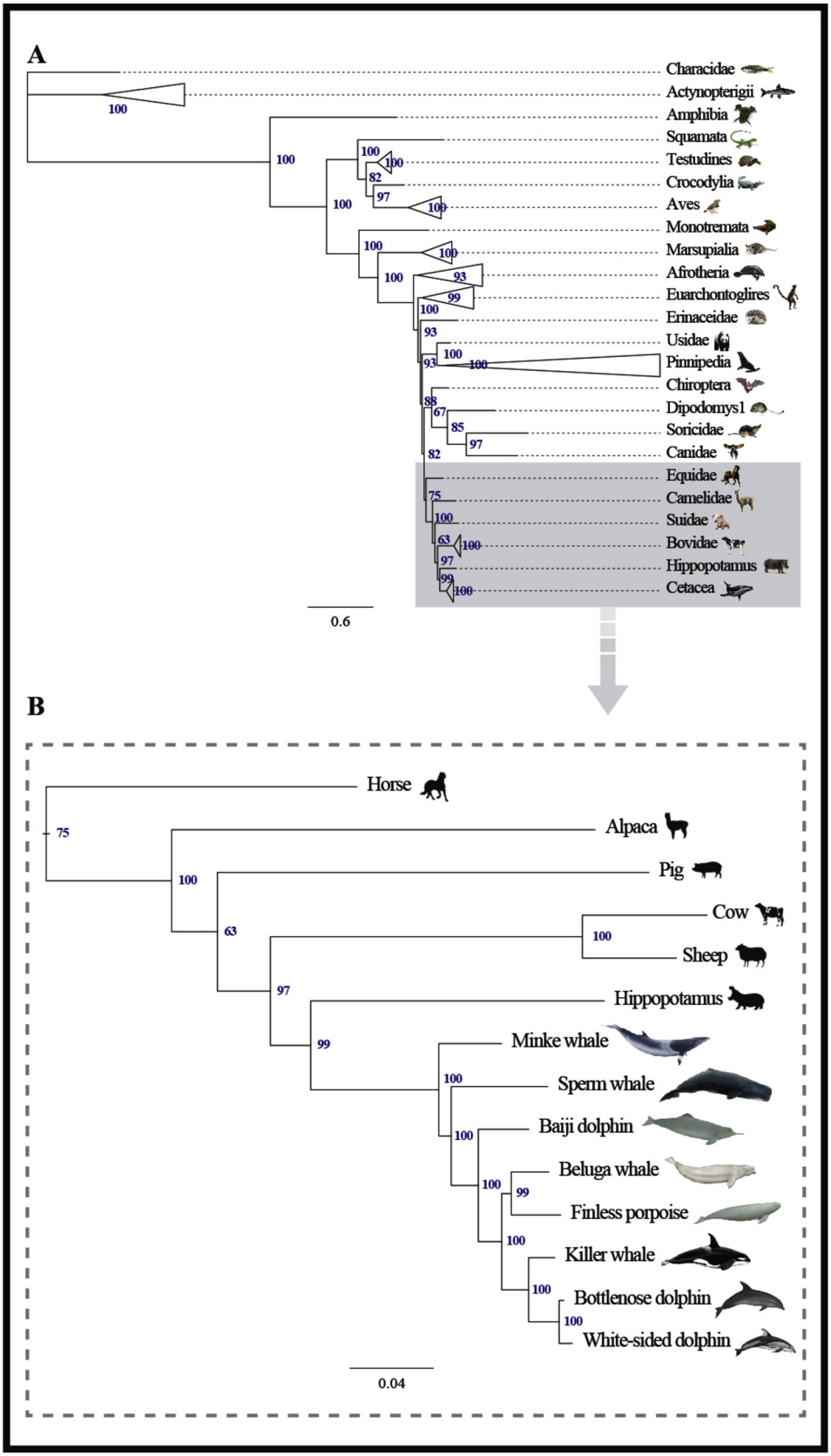

Fig. 1. Phylogenetic relationships recovered through maximum likelihood analyses from the molecular data of RAG1 gene. (A) Phylogenetic relationships among all the species in the study. The gray chart represents the Euungulata clade. (B) Phylogenetic relationships at the species level of the Euungulata clade, with emphasis on cetaceans. The numbers in blue represent the bootstrap values that support each node of the phylogenies. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

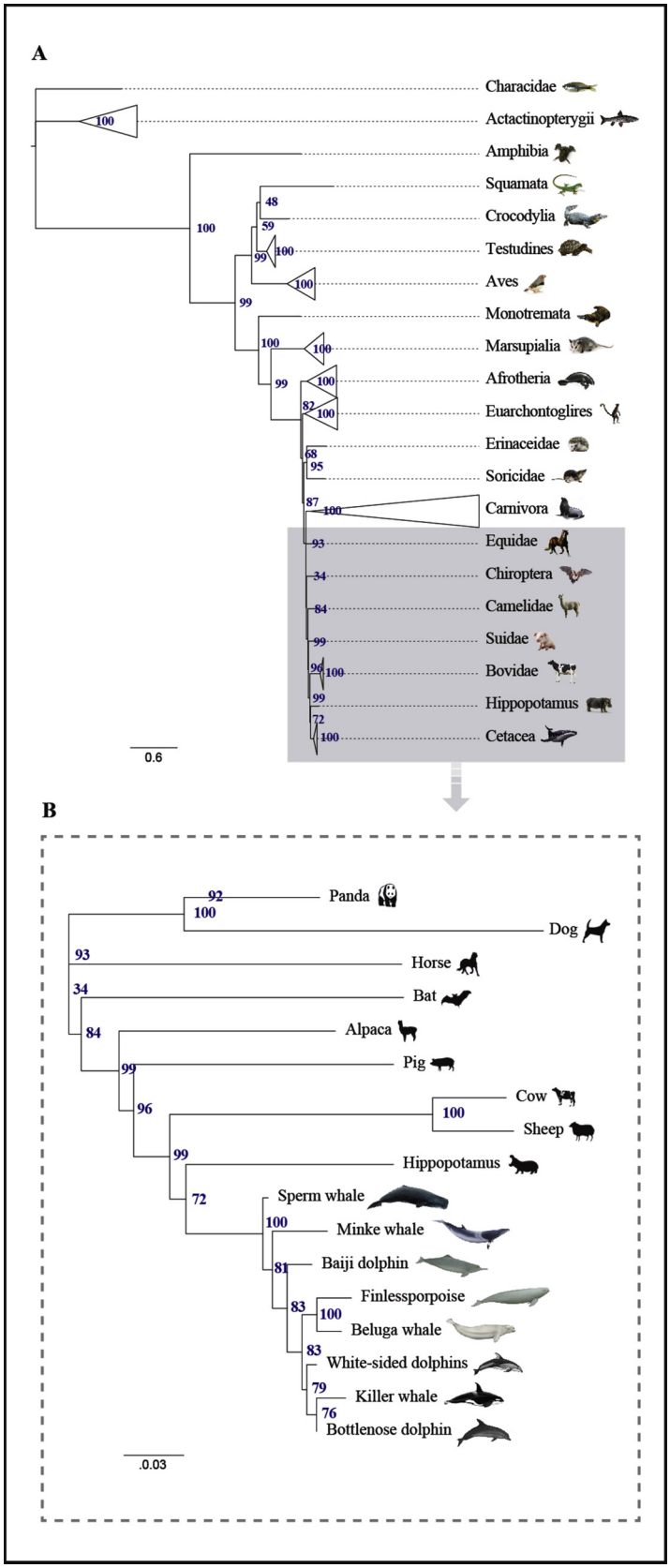

Fig. 2. Phylogenetic relationships recovered through maximum likelihood analyses from the molecular data of RAG2 gene. (A) Phylogenetic relationships among all the species in the study. (B) Phylogenetic relationships at the species level of the Euungulata clade, with emphasis on cetaceans, and the insertion of microbats closely related to the Artiodactyls. The numbers in blue represent the bootstrap values that support each node of the phylogenies. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

cetaceans including RAG1 gene, the topology derived from Bayesian analyses of nucleotide resembled the Cetacea topology generated by our analyses, with the same relationship found between whales and dolphins (McGowen et al., 2009).

The ω ratio is a measure of natural selection acting on the protein. In short, values of $\omega < 1$, $= 1$, and > 1 indicates negative purifying selection, neutral evolution, and positive selection respectively. However, the ratio averaged over all sites and all lineages is almost never > 1, since positive selection is unlikely to affect all sites over prolonged time (Yang, 1997). Our PAML results for RAG1 indicated a statistically higher ω value for cetaceans (0.22) when compared to the rest of the phylogeny (0.08), which means that this lineage accumulated more modifications throughout its evolutionary history, i.e. this gene on this branch experienced an acceleration in the rate of evolution. This result corroborates with the result from the RELAX model that identified a relaxation in the purifying selection pressure for RAG1 in the cetacean lineage, suggesting that such relaxation may have allowed the acceleration on the evolutionary rate in cetaceans, as seen with their higher ω value. Taken together, the results corroborate a relaxation on the selective pressure of RAG1 gene rather than a stronger and direct action of positive selection. For RAG2, the two-ratio model did not fit better the data, a result that agrees with the RELAX result, which did not identify a significant relaxation along the cetacean lineage when compared to other lineages. It is possible that other lineages of our set of species have experienced an acceleration in the evolution rate of this gene, since the free-ratio model better fits the data when compared to the one-ratio model.

On FITMODEL, a posterior probability greater than 90% for selection class ω3 are considered to have evolved under relaxed selection, whereas a posterior probability lower than 20% for selection class ω3 is considered to be subject to purifying selection. For RAG1, under the M3 + S2 model, only one site (270) presented ω3 posterior probability of 95% and thus, was identified as evolving under relaxed selection through the phylogeny. The core RAG1 domain contains the nonamer binding domain (NBD), and two domains namely the central with zinc finger B (ZFB) region and C-terminal domains. The site 270 is located in the CDN region, which possess sequence signals critical for nuclear localization, zinc coordination, and interactions with nucleic acid, and is conserved from the sea urchin to human (Arbuckle et al., 2011; Kumar et al., 2015). Mutations in the first 90 amino acids of RAG2 severely inhibit basic recombination reaction, formation of signal joints by deletion and formation of both signal and coding joints by inversional recombination (Cuomo and Oettinger, 1994). The first 90 amino acids of RAG2 sequences on our FITMODEL analyses was conserved and had no sites evolving under relaxed selection. A total of 24 sites were under purifying selection (ω3 < 20%), and the rest had posterior probabilities ranging from 21% to 89%, whereas the first 63 amino acids had the same ω3 values (33%), confirming their conservation. A total of four sites had a posterior probability greater than 90% for ω3 in RAG2 analyses (113, 124, 397, 428). The site 428 is found in the C-terminus of the protein (Cuomo and Oettinger, 1994). None of the estimates values for both genes were near 1, which would suggest neutrality.

Although RAG1 and RAG2 genes are highly conserved, previous comparisons of RAG2 amino acids in frogs, mammals and chickens indicated that RAG2 is less conserved than RAG1 (Greenhalgh et al., 1993). Our FITMODEL analyses for RAG2 from all tetrapods confirm this statement, considering the four sites found to be under relaxed selection compared to one site of RAG1.

RAG1 appears to have originated from a TE of the Transib family and is able to mediate low levels of recombination in the absence of RAG2, since this gene contain all the essential domains and activities needed to bind and cleave DNA, placing RAG2 in the role of an accessory or regulatory factor (Ji et al., 2010; Teng et al., 2015; Carmona et al., 2016). Evidence supports a model for RAG evolution in which a Transib transposon captured a RAG2-like open reading frame in an early deuterostome to give rise to the original RAG transposon, which in turn gave rise to RAG1, RAG2 and RSSs in jawed vertebrates and RAG1L and RAG2L transposable elements and gene pairs in invertebrates (Kapitonov and Jurka, 2005; Carmona and Schatz, 2017). Despite this RAG1 mediating ability, it is still proposed that RAG2 plays a critical role in the establishment and evolution of V(D)J recombination (Carmona et al., 2016). It is believed that, when acquiring a RAG2-like element, both genes were able to provide functional advantages that allowed for the evolution of adaptive immune system of early jawed vertebrates (Carmona et al., 2016). A more recent study proposes that a modular design of the RAG complex - with largely autonomous catalytic cores, swappable DNA binding modules and a RAG2 accessory subunit - facilitated the adaptation of RAG family enzymes to changing host environments and functional demands, which includes the adaptations in jawed vertebrates that led to a 'tamed' RAG recombinase that possesses coupled cleavage activity, adherence to the RSSs 12/23 rule and suppressed transposition activity (Zhang et al., 2019). Thus, despite closely linked in the genome and despite working together for generating diversity in the adaptive immune system, RAG1 and RAG2 genes have different evolutionary origins, function separately and thus, may have different evolutionary rates.

Drastic environmental changes such as the transition from a terrestrial to marine habitat should select for numerous evolutionary adaptations (Chikina et al., 2016). The oceans harbor an enormous diversity and number of viruses and prokaryotes, which could frequently become a threat to marine mammals (Whitman et al., 1998; Suttle, 2007). The abundance of viruses exceeds that of bacteria and archaea by approximately 15-fold (Bettarel et al., 2000), and estimates of cell density, volume, and carbon indicate that prokaryotes are ubiquitous in marine environment (Ducklow and Carlson, 1992; Simon, 1994). The RAG1 gene structure is not conserved in fishes but it is highly conserved among tetrapods (Kumar et al., 2015), and one could hypothesize that this gene behave differently in different environments, being less evolutionarily conserved in the aquatic environment, considering the greater number and diversity of pathogens in this habitat.

5. Conclusion

In summary, our findings demonstrate that RAG1 and RAG2 genes remained fairly conserved among tetrapods, with purifying selection happening on both RAG1 (80% of sites under purifying selection) and RAG2 (90% of sites under purifying selection) genes, and evidence for a few punctuated shifts in nucleotide substitution rates of both genes along tetrapod evolution. When considering only the cetacean lineage, RAG1 gene shows an accelerated rate of evolution with a relaxation of the selective pressure, while for RAG2 this relaxation was not observed and no specific acceleration on evolutionary rate. These results demonstrate differential evolution happening in the closely linked genes RAG1 and RAG2 in cetaceans, with RAG1 being less conserved when compared to other mammals of the phylogeny. It is important to note that this is only a brief part of the whole story. Since RAG genes act on DNA substrates and on a complex panorama of recombination signals, future work focusing on locus subject to the action of RAG genes, such as Ig and TCR genes, will be important to further clarify how molecular evolution acts on immunological genes during the occupation of new environments.

Author contributions

MFN and BCD designed the study. BD organized the database, performed the statistical analyses and draft the manuscript. MN coordinated the study, revised and edited the manuscript. All authors read and approved the final and submitted manuscript.

Declaration of Competing Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

This research was financially supported by the São Paulo Research Foundation (FAPESP), process number 2015/18269-1 and a Scholarship granted to BD (2017/14831-2). We thank Mr. Lucas Canesin and Dr. Lucas Freitas for their bioinformatics assistance in the analyses. We also thank Ms. Érica Souza, member of the Laboratory of Evolutionary Genomics of the State University of Campinas, for her guidance and contributions to this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.molimm.2019.10.014.

References

Agrawal, A., Eastman, Q.M., Schatz, D.G., 1998. Implications of transposition mediated by V(D)J-recombination proteins RAG1 and RAG2 for origins of antigen-specific immunity. Nature 394, 744–751.
Arbuckle, J.L., Rahman, N.S., Zhao, S., Rodgers, W., Rodgers, K.K., 2011. Elucidating the domain architecture and functions of non-core RAG1: the capacity of a noncore zinc-binding domain to function in nuclear import and nucleic acid binding. BMC Biochem. 12, 23.
Areal, H., Abrantes, J., Esteves, P.J., 2011. Signatures of positive selection in Toll-like receptor (TLR) genes in mammals. BMC Evol. Biol. 11, 368.
Bettarel, Y., Sime-Ngando, T., Amblard, C., Laveran, H., 2000. A comparison of methods for counting viruses in aquatic systems. Appl. Environ. Microbiol. 66, 2283–2289.
Capra, J.A., Hubisz, M.J., Kostka, D., Pollard, K.S., Siepel, A., 2013. A model-based analysis of GC-Biased gene conversion in the human and chimpanzee genomes. PLoS Genet. 9, 8.
Carmona, L.M., Schatz, D.G., 2017. New insights into the evolutionary origins of the recombination-activating gene proteins and V(D)J recombination. FEBS J. 284, 1590–1605.
Carmona, L.M., Fugmann, S.D., Schatz, D.G., 2016. Collaboration of RAG2 with RAG1-like proteins during the evolution of V(D)J recombination. Genes Dev. 30, 909–917.
Chikina, M., Robinson, J.D., Clark, N.L., 2016. Hundreds of genes experienced convergent shifts in selective pressure in marine mammals. Mol. Biol. Evol. 33, 2182–2192.
Cuomo, C.A., Oettinger, M.A., 1994. Analysis of regions of RAG-2 important for V(D)J recombination. Nucleic Acids Res. 22, 1810–1814.
Ducklow, H.W., Carlson, C.A., 1992. Oceanic bacterial production. Adv. Microb. Ecol. 12, 113–181.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.
Fugmann, S.D., Messier, C., Novack, L.A., Cameron, R.A., Rast, J.P., 2006. An ancient evolutionary origin of the Rag1/2 gene locus. Proc. Nat. Acad. Sci. 103, 3728–3733.
Galtier, N., Duret, L., Glémin, S., Ranwez, V., 2009. GC-biased gene conversion promoted the fixation of deleterious amino acid changes in primates. Trends Genet. 25, 1–5.
Gatesy, J., Geisler, J.H., Chang, J., Buell, C., Berta, A., Meredith, R.W., Springer, M.S., McGowen, M.R., 2013. A phylogenetic blueprint for a modern whale. Mol. Phylogenet. Evol. 66, 479–506.
Gellert, M., 2002. V(D)J recombination: RAG proteins, repair factors, and regulation. Annu. Rev. Biochem. 71, 101–132.
Greenhalgh, P., Olesen, C.E., Steiner, L.A., 1993. Characterization and expression of recombination activating genes (RAG-1 and RAG2) in *Xenopus laevis*. J. Immunol. 151, 3100–3110.
Guindon, S., Rodrigo, A.G., Dyer, K.A., Huelsenbeck, J.P., 2004. Modeling the site-specific variation of selection patterns along lineages. Proc. Nat. Acad. Sci. 101, 12957–12962.
Haldane, J.B.S., 1949. Disease and evolution. Ric. Sci. Suppl. A 19, 68–76.
Hubisz, M.J., Pollard, K.S., Siepel, A., 2011. PHAST and RPHAST: phylogenetic analysis with space/time models. Briefings Bioinf. 12, 41–51.
Ishengoma, E., Agaba, M., 2017. Evolution of toll-like receptors in the context of terrestrial ungulates and cetaceans diversification. BMC Evol. Biol. 171, 54.
Janeway Jr., C.A., 2001. How the immune system works to protect the host from infection: a personal view. Proc. Nat. Acad. Sci. 98, 7461–7468.
Ji, Y., Resch, W., Corbett, E., Yamane, A., Casellas, R., Schatz, D.G., 2010. The in vivo pattern of binding of RAG1 and RAG2 to antigen receptor loci. Cell. 141, 419–431.
Kalyaanamoorthy, S., Minh, B.Q., Wong, T.F.K., von Haeseler, A., Jermiin, L.S., 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 587–589.
Kapitonov, V.V., Jurka, J., 2005. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol. 3, e181.
Kapitonov, V.V., Koonin, E.V., 2015. Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. Biol. Direct 10, 20.
Kumar, A., Bhandari, A., Sarde, S.J., Muppavarapu, S., Tandon, R., 2015. Understanding V(D)J recombination initiator RAG1 gene using molecular phylogenetic and genetic variant analyses and upgrading missense and non-coding variants of clinical importance. Biochem. Biophys. Res. Comm. 462, 301–313.
McGowen, M.R., Spaulding, M., Gatesy, J., 2009. Divergence date estimation and a comprehensive molecular tree of extant cetaceans. Mol. Phylogenet. Evol. 53, 891–906.
Minh, B.Q., Nguyen, M.A., Von Haeseler, A., 2013. Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. 30, 1188–1195.
Nguyen, L.T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQTREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274.
Oettinger, M.A., Schatz, D.G., Gorka, C., Baltimore, D., 1990. RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. Science. 248, 1517–1523.
Pond, S.L.K., Frost, S.D.W., 2005. A genetic algorithm approach to detecting lineage specific variation in selection pressure. Mol. Biol. Evol. 22, 478–485.
Poole, J.R., Huang, S.F., Xu, A., Bayet, J., Pontarotti, P., 2017. The RAG transposon is active through the deuterostome evolution and domesticated in jawed vertebrates. Immunogenetics 69, 391–400.
Schatz, D.G., 2004. Antigen receptor genes and the evolution of a recombinase. Semin. Immunol. 16, 245–256.
Schatz, D.G., Oettiger, M.A., Baltimore, D., 1989. The V(D)J recombination activating gene, RAG-1. Cell 59, 1035–1048.
Shen, T., Xu, S., Wang, X., Yu, W., Zhou, K., Yang, G., 2012. Adaptive evolution and functional constraint at TLR4 during the secondary aquatic adaptation and diversification of cetaceans. BMC Evol. Biol. 12, 39.
Simon, M., 1994. Diel variability of bacterioplankton biomass production and cell multiplication in Lake Constance. Arch. Hydrobiol. 130, 283–302.
Suttle, C.A., 2007. Marine viruses — major players in the global ecosystem. Nature Rev. Microbiol. 5, 801–812.
Suyama, M., Torrents, D., Bork, P., 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34, W609–W612.
Teng, G., Schatz, D.G., 2015. Regulation and evolution of the RAG recombinase. Adv. Immunol. 128, 1–39.
Teng, M.W., Ngiow, S.F., Ribas, A., Smyth, M.J., 2015. Classifying cancers based on T-cell infiltration and PD-L1. Cancer Res. 75, 2139–2145.
Thewissen, J.G.M., Cooper, L.N., George, J.C., Bajpai, S., 2009. From land to water: the origin of whales, dolphins, and porpoises. Evo. Edu. Outreach. 2, 272–288.
Thompson, C.B., 1995. New insights into V(D)J recombination and its role in the evolution of the immune system. Immunity. 3, 531–539.
Tian, R., Chen, M., Chai, S., Rong, X., Chen, B., Ren, W., Xu, S., Yang, G., 2018. Divergent selection of pattern recognition receptors in mammals with different ecological characteristics. J. Mol. Evol. 86, 138–149.
Waddell, P.J., Shelley, S., 2003. Evaluating placental inter-ordinal phylogenies with novel sequences including RAG1, gamma-fibrinogen, ND6, and mt-tRNA, plus MCMC-driven nucleotide, amino acid, and codon models. Mol. Phylogenet. Evol. 28, 197–224.
Weaver, S., Shank, S.D., Spielman, S.J., Li, M., Muse, S.V., Pond, S.L.K., 2018. Datamonkey 2.0: a modern web application for characterizing selective and other evolutionary processes. Mol. Biol. Evol. 35, 773–777.
Wertheim, J.O., Murrell, B., Smith, M.D., Pond, S.L.K., Scheffler, K., 2014. RELAX: detecting relaxed selection in a phylogenetic. Mol. Biol. Evol. 32, 820–832.
Whitman, W.B., Coleman, D.C., Wiebe, W.J., 1998. Prokaryotes: the unseen majority. Proc. Natl. Acad. Sci. U.S.A. 95, 6578–6583.
Wlasiuk, G., Nachman, M.W., 2010. Adaptation and constraint at toll-like receptors in primates. Mol. Biol. Evol. 27, 2172–2186.
Yang, Z., Nielsen, R., 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19, 908–917.
Yang, Z., Wong, W.S., Nielsen, R., 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. 22, 1107–1118.
Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556.
Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.
Yang, Z.H., 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15, 568–573.
Yang, Z.H., Nielsen, R., 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. 46, 409–418.
Zhang, J., Nielsen, R., Yang, Z., 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22, 2472–2479.
Zhang, Y., Cheng, T.C., Huang, G., Lu, Q., Surleac, M.D., Mandell, J.D., Pontarotti, P., Petrescu, A.J., Xu, A., Xiong, Y., Schatz, D.G., 2019. Transposon molecular domestication and the evolution of the RAG recombinase. Nature 569, 79–84.

Glossary

MGK + F3 × 4 + G4 model of substitution

MGn: onsynonymous/synonymous (dn/ds) rate ratio with additional transition/transversion (ts/tv) rate ratio;

+*F3×4:* unequal nucleotide frequencies and unequal nucleotide frequencies over three codon positions;
+*G4:* discrete Gamma model with default 4 rate categories.

GY+F1×4+G4 model of substitution

GYn: onsynonymous/synonymous and transition/transversion rate ratios;
+*F1×4u:* nequal nucleotide frequencies but equal nucleotide frequencies over three codon positions;
+*G4:* discrete Gamma model with default 4 rate categories.
Markov-modulated Markov models of codon evolution: this means that the future of the process (that is, the various states possibly reached and their probabilities of occurrence) depends only on the present state, not on past. Markov processes can be in discrete time, when states are assigned to successive "steps," or "generations," or in continuous time, when the time to next event is an exponential random variable.

Abbreviations

LRTs: likelihood ratio tests LRTs;
TE: transposable element;
RAG1: recombination activation gene 1;
RAG2: recombination activation gene 2;