Immunologic Research
https://doi.org/10.1007/s12026-023-09451-8

ORIGINAL ARTICLE

Evolutionary preservation of CpG dinucleotides in RAG1 may elucidate the relatively high rate of methylation-mediated mutagenesis of RAG1 transposase

Mariam M. Fawzy[1] · Maiiada H. Nazmy[1] · Azza A. K. El-Sheikh[2] · Moustafa Fathy[1]

Received: 10 August 2023 / Accepted: 24 December 2023

Abstract

Recombination-activating gene 1 (RAG1) is a vital player in V(D)J recombination, a fundamental process in primary B cell and T cell receptor diversification of the adaptive immune system. Current vertebrate RAG evolved from RAG transposon; however, it has been modified to play a crucial role in the adaptive system instead of being irreversibly silenced by CpG methylation. By interrogating a range of publicly available datasets, the current study investigated whether RAG1 has retained a disproportionate level of its original CpG dinucleotides compared to other genes, thereby rendering it more exposed to methylation-mediated mutation. Here, we show that 57.57% of RAG1 pathogenic mutations and 51.6% of RAG1 disease-causing mutations were associated with CpG methylation, a percentage that was significantly higher than that of its RAG2 cofactor alongside the whole genome. The CpG scores and densities for all RAG ancestors suggested that RAG transposon was CpG denser. The percentage of the ancestral CpG of RAG1 and RAG2 were 6% and 4.2%, respectively, with no preference towards CG containing codons. Furthermore, CpG loci of RAG1 in sperms were significantly higher methylated than that of RAG2. In conclusion, RAG1 has been exposed to CpG mediated methylation mutagenesis more than RAG2 and the whole genome, presumably due to its late entry to the genome later with an initially higher CpG content.

Keywords CpG methylation · RAG · V(D)J recombination · Transposition · Mutagenesis

Introduction

Significant numbers of primary immunodeficiencies, such as severe combined immunodeficiency (SCID) and Omenn syndrome (OS), result from mutations in the recombination-activating gene (*RAG*). RAG1 and RAG2 proteins, encoded by the *RAG1* and *RAG2* genes, are critical for V(D)J recombination process to recombine V (Variable), D (Diversity), and J (Joining) gene segments at antigen receptor loci and in turn, generate a vast array of the productive immunoglobulin, or T cell receptor exons during lymphocyte development. The V, D, or J gene segments are abutted by DNA recombination signal sequences (RSSs) and are specifically recognized by RAG1 [1]. RAG1 (within the RAG1:RAG2 complex) binds to the RSSs to induce double DNA breaks (DDB) next to the coding segments to generate coding ends and joint ends. Both ends are processed by a non-homologous end joining (NHEJ) pathway [2]. RAG1 is proposed to have evolved from a RAG transposon that entered the vertebrate genome through horizontal gene transfer 500 million years ago [3] and underwent a domestication process to

Mariam M. Fawzy and Maiiada H. Nazmy contributed equally to this work (co-first authors).

✉ Moustafa Fathy
mostafa_fathe@minia.edu.eg

Mariam M. Fawzy
mariam.mahrous@minia.edu.eg

Maiiada H. Nazmy
maiiada_nazmy@minia.edu.eg

Azza A. K. El-Sheikh
aaelsheikh@pnu.edu.sa

[1] Department of Biochemistry, Faculty of Pharmacy, Minia University, Minia 61519, Egypt

[2] Basic Health Sciences Department, College of Medicine, Princess Nourah bint Abdulrahman University, 11671 Riyadh, Saudi Arabia

Published online: 19 January 2024

generate RAG recombinase with diminished transposition activity to perform the highly specialized function for a powerful adaptive immune system [4]. Being a former transposon, RAG1 is predicted to have been exposed to methylation during evolution as one main role of DNA methylation is inducing transcriptional silencing of transposable elements (TEs) that pose a continuous threat to the genome stability due to its intrinsic mobile nature [5–7].

DNA methylation occurs predominately, but not exclusively, in the CpG dinucleotides, where a methyl group is introduced in the 5-carbon of Cytosine, followed by spontaneous hydrolytic deamination reaction converting cytosine into thymine (T) [8]. Cytosine (C) and guanine (G) bases account for 40% of the human genome. Although the hypothetical expected content for CpG dinucleotides is 0.04, the actual value is between 0.008 and 0.01 [1–3]. This discrepancy is chiefly mediated by DNA methylation-mediated mutagenesis. Indeed, 70–80% of CpG dinucleotides in the human genome are 5-methylated [9].

This study investigated to what extent methylation-induced mutagenesis has contributed to *RAG1* disease-causing mutations by exploring online resources and checking the inheritance of methylation-mediated mutations in *RAG1* compared to *RAG2*. Furthermore, it aimed to test the hypothesis that the high mutation rate of *RAG1* is because it still has many of its original CpG dinucleotides and, thus, is more prone to methylation-mediated mutagenesis compared to other genes. This purpose was achieved by analyzing the CpG densities and scores in the ancestral genes of *RAG1* and checking the ancestral CpG in RAG1/RAG2 coding sequences of other vertebrates. Finally, the study checked if the relatively high CpG density of *RAG1* caused preference towards CG-containing codons when compared to *RAG2* and the whole genome.

Materials and methods

Software

The following software were used in the current study: MegaX (Molecular Evolutionary Genetics Analysis) [10] 64-bit, Excel, Expasy translate online tool.

Mining and investigation of RAG1 and RAG2 mutations in publicly available data

Online repositories

The present study was concerned with substitutional point mutations identified through the coding sequences of *RAG1* and *RAG2* in the National Center for Biotechnology Information (NCBI) section of (ClinVar). The analysis excluded frameshift mutations. Mutations caused by CpG methylation (these were converted only to TpG or CpA) were identified by manual mapping of each mutation on *RAG1* and *RAG2* coding sequences downloaded from ensemble and aligned with its protein translate (ExPASy – Online Translate tool) to confirm mutation position and codon change. The clinical significance of mutations in NCBI was as pathogenic(P), likely pathogenic (LP), benign (B), likely benign (LB), uncertain significance (Un.S), and conflicting interpretation of pathogenicity (CP). The website uses the term 'conflicting interpretation of pathogenicity' (CP) to describe mutations having conflicting data from different submitters. Our analysis included pathogenic, likely pathogenic mutations and 'CP' mutations submitted as P or LP more times than other interpretations. Supplementary tables S1 and S2 for *RAG1* and *RAG2* contain codon change, methylation status and clinical significance for each mutation. Finally, the percentage of (CpG) methylation-mediated mutations among all mutations linked to pathogenicity was calculated for *RAG1* and *RAG2*.

Published clinical data

The current study examined scientific journals for published clinical data of patients with RAG mutations [11–80] to explore the incidence of CpG methylation-mediated mutations. This examination included patients' ethnicity that duplicated mutations in patients of the same population were counted as one mutation unless patients belonged to unrelated families. Furthermore, we analyzed the data displayed by Lawless et al. [49] who innovated a tool referred to as the average mutation rate residue frequency (MRF) to predict the likelihood of clinically related mutations in *RAG1* and *RAG2*. The higher the MRF score was, the higher the possibility of occurrence of clinically related mutations. They displayed MRF for a list of *RAG1* and *RAG*2 mutations. The current study categorized mutations into CpG and non-CpG methylation-induced mutations and calculated the average MRF and *p*-value.

Investigation of the methylation ratio of CpG loci in RAG1 and RAG2 by using the online available bisulfite seq data analysis of spermatozoa cells

To check if the methylation-mediated mutations are inherited, this study examined the methylation in the germ cell (sperm). The current study determined methylated CpG loci of RAG1 and RAG2 by using the online available bisulfite sequence analysis, which depends on using bisulfite before the high-throughput sequencing for the differentiation between methylated and non-methylated cytosine. Bisulfite converts the non-methylated cytosine to uracil, which then is read as thymine through sequencing, while methylated cytosine remains unchanged [81].

The genome data viewer displayed the methylation pattern of only one sample via the link https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1127119, and UCSC (University of California Santa Cruz) epigenome viewer Human chr11:36589563-36601310 - UCSC Genome Browser v309 (epigenomebrowser.org). For sampling, testis spermatozoa primary cells Donor 390ATA mapped by Illumina Bisulfite-Seq read and mappings were processed into graphs of methylation proportions. The genome viewer included the whole genome of the sample with an option to zoom in on the required position. The present study interpreted the methylation percent of the CpG loci from the genome browser, and calculated the percentage of each category, mean, and *p*-value between methylation ratios in RAG1 and RAG2.

Estimation of CpG density and CpG score in the ancestral RAG genes

As the proposed ancestors for RAGs, the sequences of the potential RAG relative: *Hztransib*, *spRAG1L*, *spRAG2L*, and *ProtoRAG* were downloaded from NCBI, and their CpG densities were estimated. The CpG density was identified by calculating the total number of CG dinucleotides in the gene and comparing this to the overall gene length using the (Len) function in Microsoft Excel. CpG density was expressed as a percentage.

CpG score is a known tool used to assess the extent of genome exposure to DNA methylation [82, 83]. The higher the CpG score, the lower exposure to DNA methylation is. The CpG score was calculated by dividing the observed CpG density by the expected CpG density (G+C)/2)2.

Examination of ancestral CpGs for RAG1 and RAG2

The *RAG1* cDNAs from different species were aligned and used to estimate the number of CpGs in the ancestral gene. The “Fasta” files for *RAG1/RAG2* coding sequences of 40 species were downloaded from the Ensemble. Species were selected to be representative of all categories: birds, reptiles, rodents, primates, mammals, and fish. The alignment results were exported from MegaX to Excel and then printed. CpG dinucleotides were checked throughout all 40 species, and we marked the ones where it is likely to have mutated via methylation-mediated mutagenesis. If there was just one or two CpGs at a position in all 40 species, this was likely to have arisen rather than to have been mutated in the remaining 39 species. Therefore, the ancestral CpG dinucleotides were marked when detected in 3 or more species with seven or more TG/CA dinucleotides in other species or when CpG dinucleotides were detected in 10 or more species in a position. Finally, the percentage of these columns to the total number of nucleotides in the coding sequence was estimated and represented the ancestral CpG density.

Analysis of CG containing codons in RAG1 and RAG2 compared with that in the human genome

RAG1 is a DNA-binding protein, so it is expected to have high Arginine “R” residues. Six codons encode Arginine: four CGXs (X is G, C, A or T), AGA, and AGG. We aimed to investigate whether most arginine residues are encoded by CGX, which may be the reason for RAG1’s high CpG density. The sequence analysis website (Codon Usage Calculator - Free Online Analysis Tool - BiologicsCorp) verified the fraction of each arginine codon in RAG1.and results were compared with that in RAG2 and the human genome. Additionally, any other CG-containing codons were checked in both RAG1 and RAG2. Generally, CG-containing codons are CGT, CGC, CGA, and CGG for arginine; GCG for alanine; TCG for serine; CCG for proline; and ACG for threonine amino acids. These codons were also checked in the human genome using another sequence analysis tool called GenScript Codon Usage Frequency Table (chart) (https://www.genscript.com/tools/codon-frequency-table).

Statistical analysis

An independent sample *t*-test was used to compare the means between the two groups. The *Z*-score test was used to compare between two proportions.

Results

Pathogenic RAG1 mutations were associated with high CpG methylation status

RAG1 and RAG2 mutations are reportedly frequent in multiple primary immunodeficiencies. Mining the ClinVar tool at the NCBI database identified 393-point mutations at the coding sequence of RAG1, from whom 59 mutations were linked to disease pathogenicity. Stratifying pathogenic *RAG1* mutations revealed that 33/59 (55.9%) mutations were purely pathogenic, 17/59 (28.8%) were likely pathogenic, 8/59 (13.56%) were reported to be pathogenic or likely pathogenic, and only 1/59 (1.69%) mutation conflicting with pathogenicity (Fig. 1a). To investigate if RAG1 mutations with their related pathogenicity might be caused by CpG methylation, mapping the *RAG1* Open Reading Frame (ORF) for CpG methylation was manually performed after excluding mutations marked with mixed or conflicting pathogenicity. 57.57% (19/33) of the purely pathogenic *RAG1* mutations had at least one C: G > T: G or C: G > C: A change

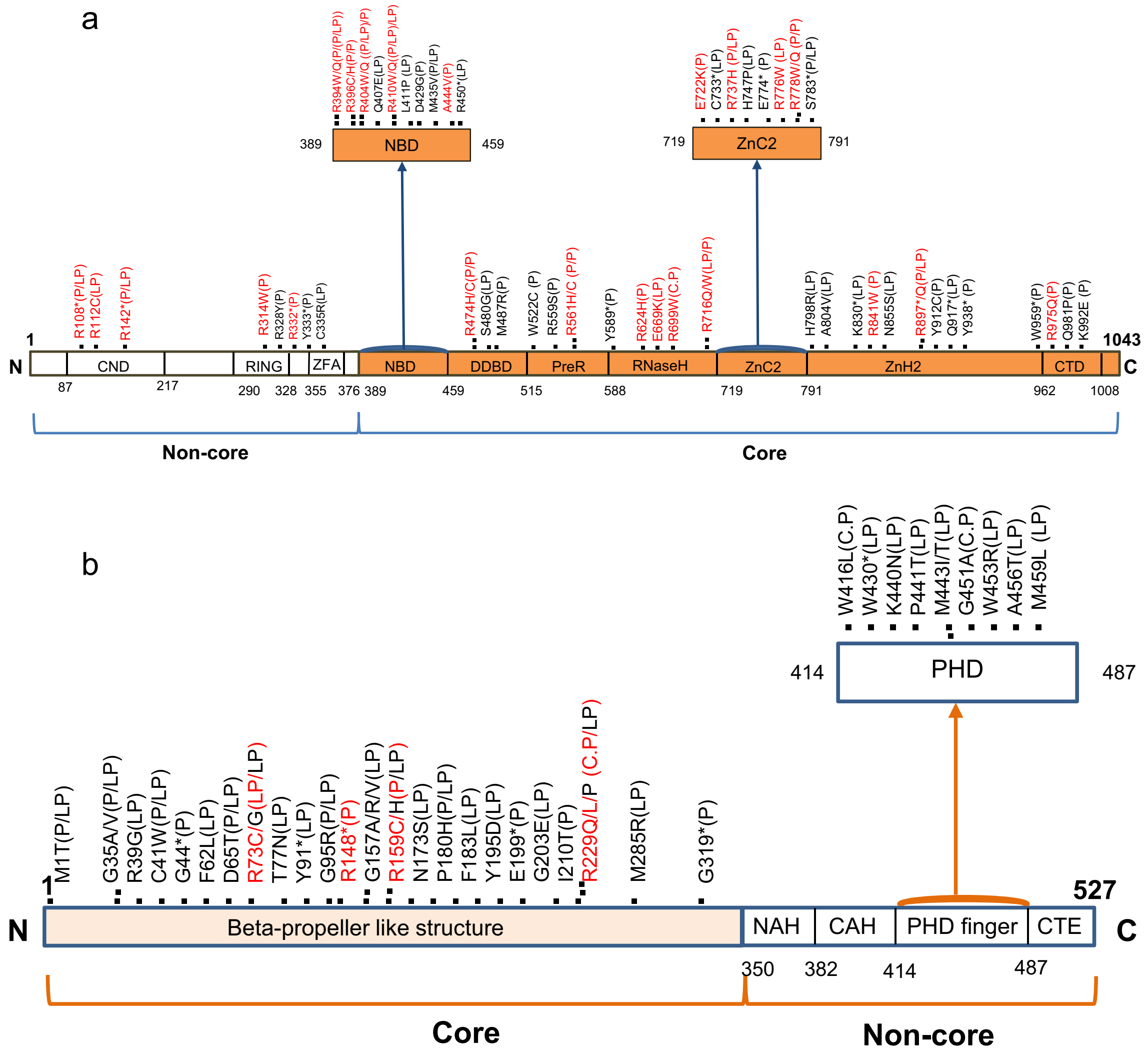


Fig. 1 Point mutations with pathogenicity of RAG1 and RAG2 proteins. **a** shows 59-point mutations with pathogenicity over the RAG1 protein main domains, 50.8% of them are methylation-mediated mutations marked with red font. (noncore domains: CND, central non-core domain; RING, really interesting new gene, ZFA, zinc finger A, Core domains: NBD, nonamer binding domain; DDBD, dimerization and DNA binding domain; PreR, Pre-RNAse H; CTD, C-terminal domain); **b** shows 41-point mutations with pathogenicity spreading over the RAG2 protein main domains, 12.19% of them are methylation-mediated mutations. (non-core domains: NAH; N-terminal acidic hinge, CAH, C-terminal acidic hinge, CTE; C-terminal extension)

compared to 23.5% (4/17) mutations assigned as likely pathogenic (z score = 2.2882, p-value = 0.02) suggesting higher frequency of CpG methylation in *RAG1* mutations that had clear pathogenic role across different disease etiologies.

The coding sequence of RAG2 had 216-point mutations, 41 of which were described to have pathogenicity. Again, stratifying pathogenic RAG2 mutations revealed that 7/41 (17.07%) mutations were purely pathogenic, 26/41 (63.4%) were likely pathogenic, 5/41 (12.19%) were reported to be pathogenic or likely pathogenic, and 3/38 (7.89%) mutation conflicting with pathogenicity (Fig. 1b). Unlike RAG1, only 28.57% (2/7) and 7.69% (2/26) of the pure and likely pathogenic *RAG2* mutations, respectively, were linked to CpG methylation (z score = −1.5, p-value = 0.13) probably highlighting other causative factors that are linked with RAG2 pathogenicity. As expected, the percentage of CpG methylation-mediated pathogenic mutations in *RAG1* was significantly higher than that of *RAG2* (50.8% versus 12.19%, z score = 3.9857, p-value < 0.0001).

To validate this data, we investigated papers reporting the role of RAG1 and RAG2 in primary immunodeficiencies and their methylation status. In line with the NCBI data, 51.64% (110 out of 213) of the total number of mutations in cases with immunodeficiencies or autoimmunity had CpG mutations mediated by methylation in *RAG1* coding sequence compared to 28.86% (28/97) in *RAG2* coding sequence (z

score = 3.74, *p* value = 0.0002). Noteworthy, the CpG mutations mediated by methylation in the human genome were about 31.5%, a percentage lower than CpG mutations mediated by methylation in the *RAG1* coding sequence (*z* score = 3.25, *p* value = 0.001).

To investigate the probability of mutation for each amino acid residue in RAG1 and RAG2 proteins, Lawless and his colleagues innovated the average mutation rate residue frequency (MRF) tool by multiplying residue frequency by mutation rate per residue [49]. They found a positive correlation between MRF and the prediction of *RAG1*, but not *RAG2*, mutation-related pathogenicity. In 66 pathogenic *RAG1* mutations analyzed by the same research group, we identified 27 CpG methylation-mediated mutations and 39 non-CpG mutations with an average MRF of 0.04 and 0.02, respectively (*t*-value = 7.41515, *p* < 0.0001). 23/27 CpG-related *RAG1* mutations (85.18%) had an MRF maximum value (MRFmax) of 0.043 compared to the 0/41 *RAG2* mutation. In conclusion, the high frequency of CpG mutations mediated by methylation in the *RAG1* (but not in the whole genome or RAG2) coding sequence suggests a role of DNA methylation in the pathogenesis of RAG1 mutation.

To discern whether high CpG frequency in the *RAG1* coding sequence is inherited or acquired, *RAG1* methylation status in sperm samples should be checked. Typically, methylation needs to occur in the germ cells/progenitor cells for the mutations to happen and is manifested subsequently in lymphocytes. Searching the Genome Expression Omnibus (GEO) database for sperm bisulfite sequencing identified a study that includes one sample (donor 390ATA, accession: GSM1127119). Analyzing *RAG1* and *RAG2* methylation status in that study showed that around 90.19% of *RAG1* and 55.56% of *RAG2* CpG loci had a high methylation level. The methylation ratio of CpG loci in the RAG1 sequence was significantly higher than that in RAG2 (*p*-value = 0.001), suggesting that the high methylation ratio in the *RAG1* sequence compared to *RAG2* might be inherited.

The CpG dinucleotides frequency was higher in the ancestral RAG transposons

RAG1 evolved from RAG transposon and was introduced into the vertebrate genome half a billion years ago. Investigation of the methylation status in *RAG* ancestors might explain the higher CpG density shown in human *RAG1*. *Transib*, *SpRAG1L* and *SpRAG2L* (in sea Urchin), and the *ProtoRAGs BbRAG1L* and *BbRAG2L* (in Lancelet) are DNA transposal superfamilies that had a sequence similarity to human *RAG*. As in Table 1, the CpG density (and CpG score) for *Helicoverpa Zea Transib*, *SpRAG1L*, *SpRAG2L*, *BbRAG1L*, and *BbRAG2L* were 3% (0.83), 3.27% (0.48), 2.5% (0.5), 2.7% (0.48), and 2.7% (0.49), respectively, compared to 1.6% (0.27) and 0.56% (0.12) in human *RAG1* and *RAG2*. High CpG densities and scores for all *RAG* relatives in comparison with the current *RAG* suggest that the original *RAG* transposons, from which the current *RAG* evolved, had high CpG density and were exposed to mutagenic deamination until they reached the current CpG density of 1.6% in *RAG1*.

Table 1 The number and percentage of CpG and CpG score in the ancestral RAGs

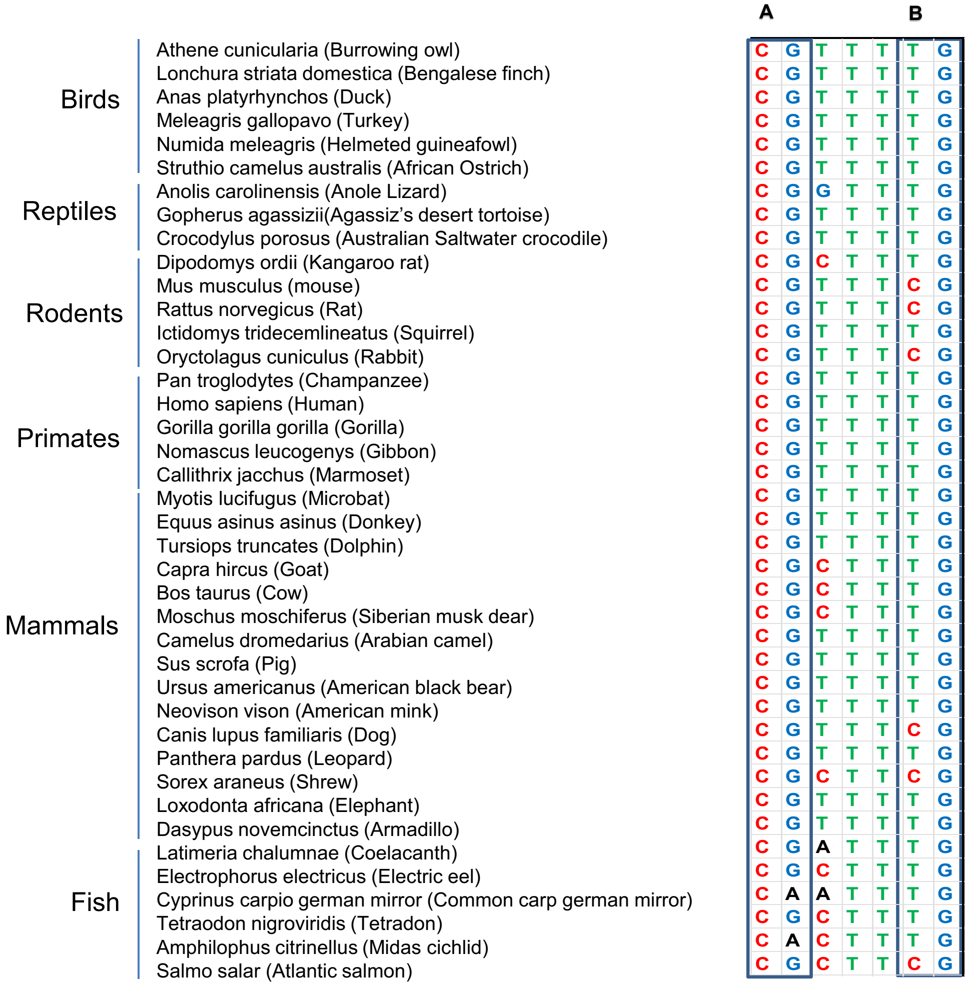
Name	Number of CpG	CpG density	CpG score (CpG o/e ratio)
Helicoverpa Zea Transib	44	3%	0.83
(Sea Urchin) SpRAG1L	98	3.27%	0.48
SpRAG2L	38	2.5%	0.5
Lancelet ProtoRAG	128	2.7%	0.48
BbRAG1L	93	2.7%	0.49
BbRAG2L	35	3.14%	0.48
Homo Sapiens RAG 1	51	1.6%	0.27
Homo Sapiens RAG 2	9	0.56%	0.12

After assessing the DNA methylation frequency in *RAG* ancestors, the study performed a subsequent comparative analysis of the CpG frequency in *RAG1* and *RAG2* sequences in other vertebrates. One hundred eighty-nine conserved/mutant CG dinucleotides were identified across the aligned *RAG1* coding sequences in 40 species, 23 of whom were conserved. CG > TG and CG > CA nucleotide changes were observed 95 and 56 times, respectively, while 15 CG > CA or TG loci were detected (Fig. 2). The vertebral CpG density in the *RAG1* coding sequence, denoted by the total number of CpG loci (189) divided by the 3132 nucleotides in the *RAG1* coding sequence, was 6.03% compared to only 4.2% (67 out of 1584) in the *RAG2* coding sequence (*z* score = 2.583, *p* value = 0.00988) (Fig. 3; supplementary table S3) further confirming the higher abundance of CpG in vertebral *RAG1* sequence.

High abundance of CpG in RAG1 did not confer preference towards CG-containing codons

RAG1 directly binds and cleaves DNA at the border of signal sequences, while RAG2 does not have a DNA binding affinity but instead forms a RAG1-RAG2 complex. One explanation of the RAG1 DNA binding activity is the higher Arginine amino acid content compared to that in RAG2. Arginine can be encoded by six codons (four CGX codons, AGA, or AGG), so the relatively large numbers of CpG dinucleotides in RAG1 may induce a preference towards CGX codons encoding arginine. To test this hypothesis, arginine-encoding sequences in *RAG1* and *RAG2* were counted. Table 2 shows that the RAG1 protein includes 66 arginine residues: 31 CGX codons (46%) and 36 AGA/AGG codons (54%), whereas the human genome

Fig. 2 The representative figure for alignment of *RAG1* coding sequences from 40 different species categorized into birds, reptiles, rodents, primates, mammals, and fish. Columns A, B, and C represent columns with ancestral CpGs through the species. Column A has 5 CG and 27 CA, and column B has 6 CG and 34 TG

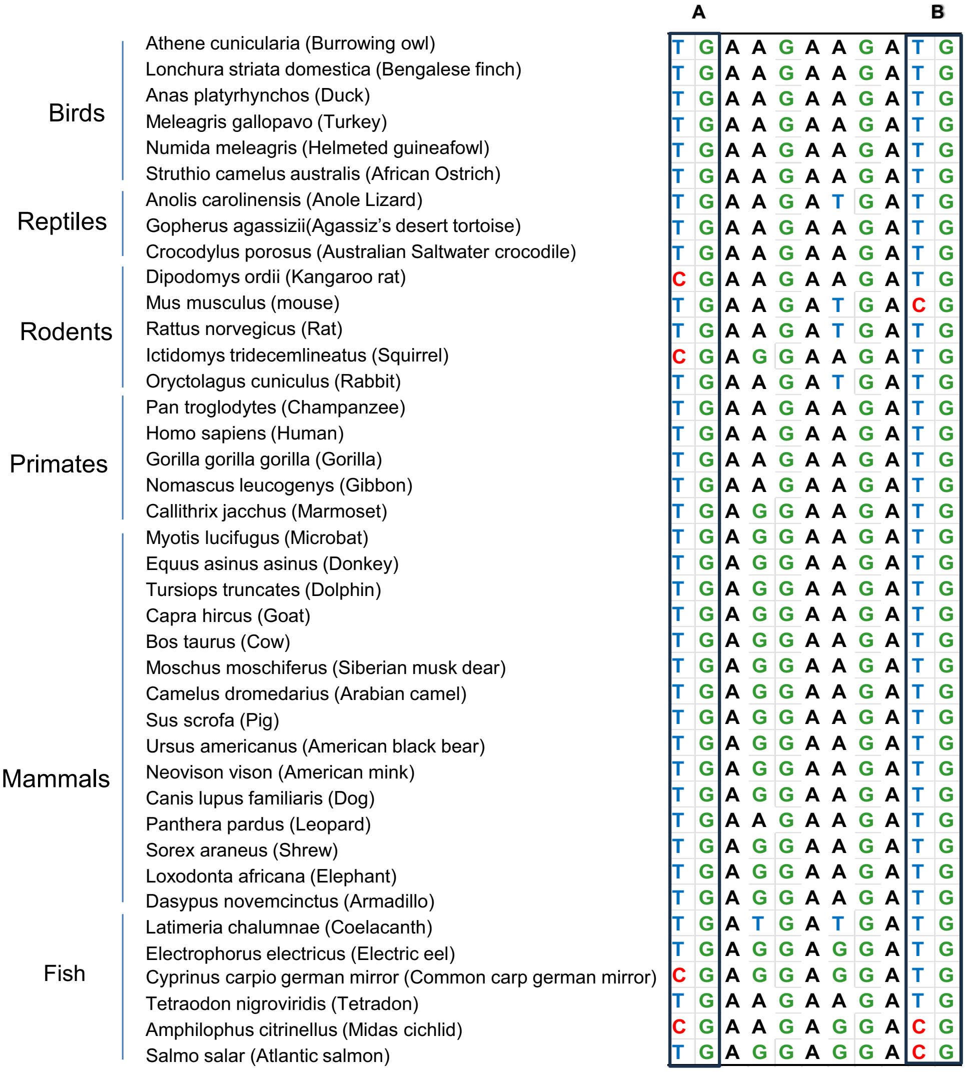


		A					B	
Birds	Athene cunicularia (Burrowing owl)	C	G	T	T	T	T	G
	Lonchura striata domestica (Bengalese finch)	C	G	T	T	T	T	G
	Anas platyrhynchos (Duck)	C	G	T	T	T	T	G
	Meleagris gallopavo (Turkey)	C	G	T	T	T	T	G
	Numida meleagris (Helmeted guineafowl)	C	G	T	T	T	T	G
	Struthio camelus australis (African Ostrich)	C	G	T	T	T	T	G
Reptiles	Anolis carolinensis (Anole Lizard)	C	G	G	T	T	T	G
	Gopherus agassizii(Agassiz's desert tortoise)	C	G	T	T	T	T	G
	Crocodylus porosus (Australian Saltwater crocodile)	C	G	T	T	T	T	G
Rodents	Dipodomys ordii (Kangaroo rat)	C	G	C	T	T	T	G
	Mus musculus (mouse)	C	G	T	T	T	C	G
	Rattus norvegicus (Rat)	C	G	T	T	T	C	G
	Ictidomys tridecemlineatus (Squirrel)	C	G	T	T	T	T	G
	Oryctolagus cuniculus (Rabbit)	C	G	T	T	T	C	G
Primates	Pan troglodytes (Champanzee)	C	G	T	T	T	T	G
	Homo sapiens (Human)	C	G	T	T	T	T	G
	Gorilla gorilla gorilla (Gorilla)	C	G	T	T	T	T	G
	Nomascus leucogenys (Gibbon)	C	G	T	T	T	T	G
	Callithrix jacchus (Marmoset)	C	G	T	T	T	T	G
Mammals	Myotis lucifugus (Microbat)	C	G	T	T	T	T	G
	Equus asinus asinus (Donkey)	C	G	T	T	T	T	G
	Tursiops truncates (Dolphin)	C	G	T	T	T	T	G
	Capra hircus (Goat)	C	G	C	T	T	T	G
	Bos taurus (Cow)	C	G	C	T	T	T	G
	Moschus moschiferus (Siberian musk dear)	C	G	C	T	T	T	G
	Camelus dromedarius (Arabian camel)	C	G	T	T	T	T	G
	Sus scrofa (Pig)	C	G	T	T	T	T	G
	Ursus americanus (American black bear)	C	G	T	T	T	T	G
	Neovison vison (American mink)	C	G	T	T	T	T	G
	Canis lupus familiaris (Dog)	C	G	T	T	T	C	G
	Panthera pardus (Leopard)	C	G	T	T	T	T	G
	Sorex araneus (Shrew)	C	G	C	T	T	C	G
	Loxodonta africana (Elephant)	C	G	T	T	T	T	G
	Dasypus novemcinctus (Armadillo)	C	G	T	T	T	T	G
Fish	Latimeria chalumnae (Coelacanth)	C	G	A	T	T	T	G
	Electrophorus electricus (Electric eel)	C	G	C	T	T	T	G
	Cyprinus carpio german mirror (Common carp german mirror)	C	A	A	T	T	T	G
	Tetraodon nigroviridis (Tetradon)	C	G	C	T	T	T	G
	Amphilophus citrinellus (Midas cichlid)	C	A	C	T	T	T	G
	Salmo salar (Atlantic salmon)	C	G	C	T	T	C	G

contains 60% CGX and 40% AGA/AGG (GenScript website). Other amino acids encoded by XCG codons, including alanine, serine, proline, and threonine, were also analyzed for their CG content. Once more, only 2/68 of alanine-encoded codons were GCG (2.9%), 2/85 serine-encoded codons were TCG (2.3%), 2/52 proline-encoded codons were CCG (3.8%), while 2/40 (5%) threonine-encoded codons were ACG. Interestingly, the prevalence of these amino acids CG-containing-codons was lower than the corresponding human genome. Regarding *RAG2*, 8/18 of arginine-encoding codons were CGX (44.45%), 10/18 were AGA/AGG (55.55%), while alanine, serine, proline, and threonine amino acids were neither encoded by GCG, TCG, CCG, nor ACG in RAG2 protein. In conclusion, high CpG content in the *RAG1* coding sequence did not confer any codon-usage bias in the relevant protein.

Discussion

RAG-mediated V(D)J recombination is essential for durable adaptive immunity. Therefore, mutations in the human RAG genes are correlated with a significant number of immunodeficiencies. In this study, we tried to determine whether the high mutation rate of *RAG1* was because it retained many of its original CpG and, consequently, was more exposed to methylation-mediated mutagenesis than other genes. This study is the first to check the extent of CpG methylation contribution in RAG disease-causing mutations. A review of NCBI-identified pathogenic mutations in the coding sequences of *RAG1* demonstrated that CpG methylation was the causative for 57.57% of these



		A									B	
Birds	Athene cunicularia (Burrowing owl)	T	G	A	A	G	A	A	G	A	T	G
	Lonchura striata domestica (Bengalese finch)	T	G	A	A	G	A	A	G	A	T	G
	Anas platyrhynchos (Duck)	T	G	A	A	G	A	A	G	A	T	G
	Meleagris gallopavo (Turkey)	T	G	A	A	G	A	A	G	A	T	G
	Numida meleagris (Helmeted guineafowl)	T	G	A	A	G	A	A	G	A	T	G
	Struthio camelus australis (African Ostrich)	T	G	A	A	G	A	A	G	A	T	G
Reptiles	Anolis carolinensis (Anole Lizard)	T	G	A	A	G	A	T	G	A	T	G
	Gopherus agassizii(Agassiz's desert tortoise)	T	G	A	A	G	A	A	G	A	T	G
	Crocodylus porosus (Australian Saltwater crocodile)	T	G	A	A	G	A	A	G	A	T	G
Rodents	Dipodomys ordii (Kangaroo rat)	C	G	A	A	G	A	A	G	A	T	G
	Mus musculus (mouse)	T	G	A	A	G	A	T	G	A	C	G
	Rattus norvegicus (Rat)	T	G	A	A	G	A	T	G	A	T	G
	Ictidomys tridecemlineatus (Squirrel)	C	G	A	G	G	A	A	G	A	T	G
	Oryctolagus cuniculus (Rabbit)	T	G	A	A	G	A	T	G	A	T	G
Primates	Pan troglodytes (Champanzee)	T	G	A	A	G	A	A	G	A	T	G
	Homo sapiens (Human)	T	G	A	A	G	A	A	G	A	T	G
	Gorilla gorilla gorilla (Gorilla)	T	G	A	A	G	A	A	G	A	T	G
	Nomascus leucogenys (Gibbon)	T	G	A	A	G	A	A	G	A	T	G
	Callithrix jacchus (Marmoset)	T	G	A	G	G	A	A	G	A	T	G
Mammals	Myotis lucifugus (Microbat)	T	G	A	G	G	A	A	G	A	T	G
	Equus asinus asinus (Donkey)	T	G	A	G	G	A	A	G	A	T	G
	Tursiops truncates (Dolphin)	T	G	A	G	G	A	A	G	A	T	G
	Capra hircus (Goat)	T	G	A	G	G	A	A	G	A	T	G
	Bos taurus (Cow)	T	G	A	G	G	A	A	G	A	T	G
	Moschus moschiferus (Siberian musk dear)	T	G	A	G	G	A	A	G	A	T	G
	Camelus dromedarius (Arabian camel)	T	G	A	G	G	A	A	G	A	T	G
	Sus scrofa (Pig)	T	G	A	G	G	A	A	G	A	T	G
	Ursus americanus (American black bear)	T	G	A	G	G	A	A	G	A	T	G
	Neovison vison (American mink)	T	G	A	G	G	A	A	G	A	T	G
	Canis lupus familiaris (Dog)	T	G	A	G	G	A	A	G	A	T	G
	Panthera pardus (Leopard)	T	G	A	A	G	A	A	G	A	T	G
	Sorex araneus (Shrew)	T	G	A	G	G	A	A	G	A	T	G
	Loxodonta africana (Elephant)	T	G	A	G	G	A	A	G	A	T	G
	Dasypus novemcinctus (Armadillo)	T	G	A	G	G	A	A	G	A	T	G
Fish	Latimeria chalumnae (Coelacanth)	T	G	A	T	G	A	T	G	A	T	G
	Electrophorus electricus (Electric eel)	T	G	A	G	G	A	G	G	A	T	G
	Cyprinus carpio german mirror (Common carp german mirror)	C	G	A	G	G	A	G	G	A	T	G
	Tetraodon nigroviridis (Tetradon)	T	G	A	A	G	A	A	G	A	T	G
	Amphilophus citrinellus (Midas cichlid)	C	G	A	A	G	A	G	G	A	C	G
	Salmo salar (Atlantic salmon)	T	G	A	G	G	A	G	G	A	C	G

Fig. 3 The representative figure for alignment of *RAG2* coding sequences from 40 different species categorized into birds, reptiles, rodents, primates, mammals, and fish. Columns A, B, and C represent columns with ancestral CpGs through the species. Column A has 4 CG and 36 TG, and column B has 3 CG and 37 TG

mutations. Even after including others described as likely pathogenic and conflicting interpretations with pathogenicity (specifically that has been submitted as pathogenic/likely pathogenic more often than other clinical significance interpretations), the percentage remained high (50.8%) and was significantly higher than 12.19%, the percent of *RAG2* point mutations caused by CpG methylation to the whole RAG2 mutations linked to pathogenicity (p-value < 0.0001).

Then, this study inspected published clinical cases with immunodeficiencies and found that 51.6% and 28.86% of patients with *RAG1* and *RAG2* mutations, correspondingly, had CpG methylation-mediated mutagenesis (p-value=0.0002). Additionally, the RAG1 percentage is significantly higher than 31.5%, the percentage of the disease-causing methylation-mediated mutations in the human genome (p-value = 0.001) [84]. These findings agreed with the MRF values calculated by Lawless et al. [49] and analyzed in the present work. They used population genetics data from about 146,000 individuals for minor variant analysis. To validate the calculated scores of MRF, they used 44 previously identified pathogenic variants stated in patients and recombination activity scores from 110 mutated RAG1/2. Likewise, they compared probabilities with 98 currently reported diseased cases in humans. They also used a genome sequence dataset of 558 patients with primary immunodeficiency/wild-type RAG as negative controls. They found a positive correlation between the MFR values and pathogenicity prediction of RAG1 and not RAG2 mutations [49].

Although CpG methylation-mediated mutation is known to be inherited [85], we had to confirm the methylation status at the CpG loci in RAG1 and RAG2 in specific way. After examining the methylation level through the online available bisulfite seq analysis of one spermatozoa sample, methylation levels for the CpG loci observed in *RAG1* were significantly higher than in *RAG2* (p-value = 0.001), which might explain the higher mutation levels in *RAG1* than in *RAG2* in

Table 2 The percentage of CG containing codons (with red font) compared to non-CG-containing codons for amino acids: arginine, alanine, serine, proline, and threonine in RAG1, RAG2, and the human genome

Amino acid codons		RAG1			RAG2			Human genome		
		Percentage		Number	Percentage		Number	Percentage		Number
Arginine	CGT	12.12	45.5	8	11.11	44.45	2	8	60	93,458
	CGC	6		4	22.22		4	19		217,130
	CGA	6		4	5.5		1	11		126,113
	CGG	21.2		14	5.5		1	21%		235,938
	AGA	19.7	54.5	13	44.44	55.55	8	20	40	228,151
	AGG	34.8		23	11.11		2	20		227,281
Alanine	GCT	35.3		24	30		6	26		370,873
	GCC	35.3		24	40		8	40		567,930
	GCA	26.5		18	30		6	23		317,338
	GCG	2.9		2	Zero		Zero	11		150,708
Serine	TCT	26.4		23	31		13	18		291,040
	TCC	23		20	16.67		7	22		346,943
	TCA	15		13	21.4		9	15		233,110
	TCG	2.3		2	Zero		Zero	6		89,429
	AGT	17.24		15	16.67		7	15		237,404
	AGC	16		14	14.29		6	24		385,113
Proline	CCT	32.7		17	31		9	28		343,793
	CCC	15.4		8	27.6		8	33		397,790
	CCA	48		25	41.4		12	27		331,944
	CCG	3.8		2	Zero		Zero	11		139,414
Threonine	ACT	37.5		15	37		13	24		255,582
	ACC	43.5		17	17.14		6	36		382,050
	ACA	15		6	45.7		16	28		294,223
	ACG	5		2	Zero		Zero	12		123,533

agreement with Zhou et al. who found a positive correlation between the methylation level and the mutational rate in the human germline (sperms and oocytes) when analyzed by whole genome bisulfite sequencing during the development stage [86].

Following that, the study undertook to identify the reason for the relatively high incidence of CpG methylation-mediated mutations in *RAG1* by examining its evolutionary roots. Since identifying V(D)J recombination, the standard features between this process and cut-and-paste transposition have had significant attention [87]. RAG cleaves adjacent to RSS, which is reminiscent of inverted terminal repeats (TIRs) targeted by transposases, but instead of NHEJ, the transposon is inserted into the target DNA generating characteristic edges called target site duplication (TSD), whose length is variable according to the TE. These features, along with the discovery of RAG-mediated transposition, strengthen the hypothesis of V(D)J recombination evolution known as transposon/split receptor gene, which assumed that *RAG1*, *RAG2*, and the gene segments of antigen receptor loci have originated from the TE containing *RAG1*-like (*RAG1L*) and *RAG2L* genes flanked with TIRs [87–89]. Vertebrates' RAG emerged by horizontal transfer to the genome of jawed vertebrates as a RAG transposon at the time of the emanation of their complex adaptive immune system [90–92]. Believed all adaptive immunity components arose about 500 million years ago after the division of jawless vertebrates without any known source in the ancient species. So, it was called the immunological "big bang" [93]. Sequence similarities were identified between the current *RAG1* and *Transib* family of TEs. *RAG1* consists of an active core region and a regulatory non-core region (Fig. 1). *Transib* resembles only the core region; its TIRs are like RSS, especially the heptamer. *Hztransib* is the active member of this family and has *a RAG*-like transposition manner, including the five bp TSDs left after transposition, so it is considered a precursor for RAG transposon [91]. However, the *Transib* family lacks *RAG2L*. Both *RAG1L* and *RAG2L* are in the purple sea urchin (*Strongylocentrotus purpuratus (SP)*), which has an established evolution relation with the human genome [94]. *SPRAG1L* is like core *RAG1* in addition to the N-terminal RING domain of the non-core, and *SPRAG2L* is like *RAG2*. Unlike other transposons, they have no TIRs or TSDs [92].

The *ProtoRAG* from the Chinese lancelets (*Branchistoma belcheri (Bb)*) consists of *BbRAG1L* and *BbRAG2L* genes, which have structural similarities with *RAG1* and *RAG2*, respectively. Additionally, they are convergently transcribed (as in the case of *RAG1* and *RAG2*) and flanked with RSS-like TIRs. They left five bp TSDs after transposition. The structural similarity between *BbRAG1L* and *RAG1* exceeds the core and extends to include the RING/Zinc finger in the non-core region's N-terminal region [92]. Many studies suggested models to illustrate how *RAG* has evolved from these transposons [4, 95–97], starting with Hztransib as the most ancient ancestor, passing with modifications and acquisition of *RAG2L* to get *SPRAGL* or *ProtoRAG*, which underwent further changes to diminish the transposition activity and yield the current sophisticated immune system. The process of *RAG* transposon adaptation over the years, giving rise to the existing RAG, is called Transposon domestication [98]. The methylation status of the ancestral RAGs was checked by calculating their CpG densities and scores. Zhou and his colleagues suggested the CpG score as an indicator of the rate of germline CpG mutations through evolution. Additionally, they found an inverse correlation between the age of TEs and CpG density, in turn, CpG score [99].

Table 1 shows that CpG densities and scores for all RAG relatives were higher than that of the human RAG1, confirming that the original RAG1 transposon had higher CpG density when entered our genome. It is worth mentioning that CpG scores for *SpRAG2L* (0.5) and *BbRAG2L* (0.49) are slightly higher than that for *SpRAG1L* (0.48) and *BbRAG2L* (0.48), which is quite different from the present case of the human *RAG* where *RAG2* CpG score (0.12) is lower than half that of *RAG1*(0.27) and even lower than that of the whole genome (0.2-0.25). This lower-than-expected CpG score of *RAG2* is against the assumption that RAG2 entered the genome later than RAG1 and the hypothesis of Kapitonov and Koonin about the common transposon from which both proteins evolved [3] unless *RAG2* was exposed exclusively to extensive CpG methylation mediated mutagenesis during evolution.

Besides, the CpG conversion into CpA, TpG, or both was identified in the aligned 40 RAG coding sequences from different vertebrates to investigate the original number of CpG dinucleotides in the ancient vertebrate *RAG*. *RAG1* had a higher fraction of ancestral (CpG)s (6%) than *RAG2* (4.2%) and the current *RAG1* CpG fraction (1.6%). Remarkably, we found the expected CpG content for *RAG1* and *RAG2* to be 6% and 4.5%, respectively, based on the (C + G) percent in *RAG1* (48.7%) and *RAG2* (42.5%).

Lastly, the study tried to check if there is a relation between the comparatively high CpG density in RAG1 and the abundance of CG-containing codons. Non-even usage for the synonymous codons is observed in different species and is known as codon usage bias (CUB) [100]. Preference to specific codons rather than others of the same amino acid is affected mainly by mutation and natural selection [101, 102]. Checking the CG-containing codons has revealed that the relatively high CpG density in *RAG1*, compared to *RAG2* and the whole genome, was not related to the high arginine residues present in such a DNA-binding protein like *RAG1*, as 54% of arginine residues were encoded with AGA/AGG, while the other four codons (CGG, CGC, CGA, and CGT) encoded only 46%. However, most arginine codons (60%) in the human genome were CGX, while only 40% were AGA/AGG.

Conclusions

Disease causing methylation-mediated mutations occurred more frequently in RAG1 coding sequence compared to *RAG2* and the human genome. *RAG1* had higher CpG density and CpG score than *RAG2* and the human genome, so it seemed that *RAG1* kept most of its original CpG dinucleotides. Further research should be done to discover the exact mechanism behind the extremely high methylation rate experienced by *RAG2* during evolution.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1007/s12026-023-09451-8.

Acknowledgements We would like to present our thanks to Dr. Joan Boyes, associate professor in the Faculty of Biological Sciences, Leeds University UK, and to Mr. Wael Aziz, lead software developer, Leeds UK.

Author contribution Conceptualization, M. M. F., M. H. N., and M. F.; methodology, M. M. F. and A. A. K. E.; software and validation, M. M. F. and M. F.; data curation and formal analysis, M. M. F.; investigation and resources, M. H. N. and M. F.; writing—original draft preparation, M. M. F and M. H. N.; writing—review and editing, M. F, A. A. K. E. and M. H. N.; visualization, project administration, and supervision, M. H. N. and M. F.; funding acquisition, A. A. K. E. All authors have read and agreed to the published version of the manuscript.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB). This research is supported by the scientific mission system, Ministry of Higher Education, Egypt.

Data availability Data generated or analyzed during this study are provided in full within the published article and its supplementary materials.

Compliance with ethical standards

Ethics approval or informed consent No human subjects or experimental animals involved, not applicable.

Conflict of interest The authors declare no competing interests.

References

1. Ramsden DA, Baetz K, Wu GE. Conservation of sequence in recombination signal sequence spacers. Nucleic Acids Res. 1994;22(10):1785–96. https://doi.org/10.1093/nar/22.10.1785.
2. Chang HHY, et al. Non-homologous DNA end joining and alternative pathways to double-strand break repair. Nat Rev Mol Cell Biol. 2017;18(8):495–506. https://doi.org/10.1038/nrm.2017.48.
3. Kapitonov VV, Koonin EV. Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. Biol Direct. 2015;10:20. https://doi.org/10.1186/s13062-015-0055-8.
4. Zhang Y, et al. Transposon molecular domestication and the evolution of the RAG recombinase. Nature. 2019;569(7754):79–84. https://doi.org/10.1038/s41586-019-1093-7.
5. Ben-Hattar J, Jiricny J. Methylation of single CpG dinucleotides within a promoter element of the Herpes simplex virus tk gene reduces its transcription in vivo. Gene. 1988;65(2):219–27. https://doi.org/10.1016/0378-1119(88)90458-1.
6. Watt F, Molloy PL. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. Genes Dev. 1988;2(9):1136–43. https://doi.org/10.1101/gad.2.9.1136.
7. Iguchi-Ariga SM, Schaffner W. CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. Genes Dev. 1989;3(5):612–9. https://doi.org/10.1101/gad.3.5.612.
8. Murphy KM, Travers P, Walport M. in NCBI bookshelf. London: Garland Pub; 2007.
9. Jabbari K, Bernardi G. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. Gene. 2004;333:143–9. https://doi.org/10.1016/j.gene.2004.02.043.
10. Kumar S, et al. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018;35(6):1547–9. https://doi.org/10.1093/molbev/msy096.
11. Abolhassani H, et al. A hypomorphic recombination-activating gene 1 (RAG1) mutation resulting in a phenotype resembling common variable immunodeficiency. J Allergy Clin Immunol. 2014;134(6):1375–80. https://doi.org/10.1016/j.jaci.2014.04.042.
12. Allan J, et al. The structure of histone H1 and its location in chromatin. Nature. 1980;288(5792):675–9. https://doi.org/10.1038/288675a0.
13. Alsmadi O, et al. Molecular analysis of T-B-NK+ severe combined immunodeficiency and Omenn syndrome cases in Saudi Arabia. BMC Med Genet. 2009;10:116. https://doi.org/10.1186/1471-2350-10-116.
14. Asai E, et al. Analysis of mutations and recombination activity in RAG-deficient patients. Clin Immunol. 2011;138(2):172–7. https://doi.org/10.1016/j.clim.2010.11.005.
15. Avila EM, et al. Highly variable clinical phenotypes of hypomorphic RAG1 mutations. Pediatrics. 2010;126(5):e1248–52. https://doi.org/10.1542/peds.2009-3171.
16. Bai X, et al. Clinical, immunologic, and genetic characteristics of RAG mutations in 15 Chinese patients with SCID and Omenn syndrome. Immunol Res. 2016;64(2):497–507. https://doi.org/10.1007/s12026-015-8723-4.
17. Baumann M, et al. Regulation of V(D)J recombination by nucleosome positioning at recombination signal sequences. Embo J. 2003;22(19):5197–207. https://doi.org/10.1093/emboj/cdg487.
18. Cassani B, et al. Defect of regulatory T cells in patients with Omenn syndrome. J Allergy Clin Immunol. 2010;125(1):209–16. https://doi.org/10.1016/j.jaci.2009.10.023.
19. Cavadini P, et al. AIRE deficiency in thymus of 2 patients with Omenn syndrome. J Clin Invest. 2005;115(3):728–32. https://doi.org/10.1172/JCI23087.
20 Chen K, et al. Autoimmunity due to RAG deficiency and estimated disease incidence in RAG1/2 mutations. J Allergy Clin Immunol. 2014;133(3):880-2 e10. https://doi.org/10.1016/j.jaci.2013.11.038.
21. Chi ZH, et al. Targeted high-throughput sequencing technique for the molecular diagnosis of primary immunodeficiency disorders. Medicine. 2018;97(40):e12695. https://doi.org/10.1097/MD.0000000000012695. (**Baltimore**).
22. Chou J, et al. A novel homozygous mutation in recombination activating gene 2 in 2 relatives with different clinical phenotypes: Omenn syndrome and hyper-IgM syndrome. J Allergy Clin Immunol. 2012;130(6):1414–6. https://doi.org/10.1016/j.jaci.2012.06.012.
23. Corneo B, et al. Three-dimensional clustering of human RAG2 gene mutations in severe combined immune deficiency. J Biol Chem. 2000;275(17):12672–5. https://doi.org/10.1074/jbc.275.17.12672.
24. Corneo B, et al. Identical mutations in RAG1 or RAG2 genes leading to defective V(D)J recombinase activity can cause either T-B-severe combined immune deficiency or Omenn syndrome. Blood. 2001;97(9):2772–6. https://doi.org/10.1182/blood.v97.9.2772.
25. Crestani E, et al. RAG1 reversion mosaicism in a patient with Omenn syndrome. J Clin Immunol. 2014;34(5):551–4. https://doi.org/10.1007/s10875-014-0051-2.
26. Dalal I, et al. Novel mutations in RAG1/2 and ADA genes in Israeli patients presenting with T-B-SCID or Omenn syndrome. Clin Immunol. 2011;140(3):284–90. https://doi.org/10.1016/j.clim.2011.04.011.
27. De Ravin SS, et al. Hypomorphic Rag mutations can cause destructive midline granulomatous disease. Blood. 2010;116(8):1263–71. https://doi.org/10.1182/blood-2010-02-267583.
28. de Villartay JP, et al. A novel immunodeficiency associated with hypomorphic RAG1 mutations and CMV infection. J Clin Invest. 2005;115(11):3291–9. https://doi.org/10.1172/JCI25178.
29. Dhingra N, et al. Severe combined immunodeficiency caused by a new homozygous RAG1 mutation with progressive encephalopathy. Hematol Oncol Stem Cell Ther. 2014;7(1):44–9. https://doi.org/10.1016/j.hemonc.2013.11.001.
30. Ehl S, et al. A variant of SCID with specific immune responses and predominance of gamma delta T cells. J Clin Invest. 2005;115(11):3140–8. https://doi.org/10.1172/JCI25221.
31. Erman B, et al. Investigation of genetic defects in severe combined immunodeficiency patients from Turkey by targeted sequencing. Scand J Immunol. 2017;85(3):227–34. https://doi.org/10.1111/sji.12523.
32. Fazlollahi MR, et al. Clinical, laboratory, and molecular findings for 63 patients with severe combined immunodeficiency: a decade's experience. J Investig Allergol Clin Immunol. 2017;27(5):299–304. https://doi.org/10.18176/jiaci.0147.
33. Felgentreff K, et al. Clinical and immunological manifestations of patients with atypical severe combined immunodeficiency. Clin Immunol. 2011;141(1):73–82. https://doi.org/10.1016/j.clim.2011.05.007.
34. Geier CB, et al. Leaky RAG deficiency in adult patients with impaired antibody production against bacterial polysaccharide

antigens. PLoS ONE. 2015;10(7):e0133220. https://doi.org/10.1371/journal.pone.0133220.

35. Gennery AR, et al. Omenn's syndrome occurring in patients without mutations in recombination activating genes. Clin Immunol. 2005;116(3):246–56. https://doi.org/10.1016/j.clim.2005.04.014.
36. Gomez CA, et al. Mutations in conserved regions of the predicted RAG2 kelch repeats block initiation of V(D)J recombination and result in primary immunodeficiencies. Mol Cell Biol. 2000;20(15):5653–64. https://doi.org/10.1128/mcb.20.15.5653-5664.2000.
37. Gruber TA, et al. Clinical and genetic heterogeneity in Omenn syndrome and severe combined immune deficiency. Pediatr Transplant. 2009;13(2):244–50. https://doi.org/10.1111/j.1399-3046.2008.00970.x.
38 Henderson LA, et al. Expanding the spectrum of recombination-activating gene 1 deficiency: a family with early-onset autoimmunity. J Allergy Clin Immunol. 2013;132(4):969-71 e1-2. https://doi.org/10.1016/j.jaci.2013.06.032.
39. Hill DA, et al. Risk of non-Hodgkin lymphoma (NHL) in relation to germline variation in DNA repair and related genes. Blood. 2006;108(9):3161–7. https://doi.org/10.1182/blood-2005-01-026690.
40. Jenuwein T, Allis CD. Translating the histone code. Science. 2001;293(5532):1074–80. https://doi.org/10.1126/science.1063127.
41. John T, et al. Unrelated Hematopoietic Cell Transplantation in a Patient with Combined Immunodeficiency with Granulomatous Disease and Autoimmunity Secondary to RAG Deficiency. J Clin Immunol. 2016;36(7):725–32. https://doi.org/10.1007/s10875-016-0326-x.
42. Karaca NE, et al. Diverse phenotypic and genotypic presentation of RAG1 mutations in two cases with SCID. Clin Exp Med. 2009;9(4):339–42. https://doi.org/10.1007/s10238-009-0053-1.
43. Kato M, et al. Omenn syndrome–review of several phenotypes of Omenn syndrome and RAG1/RAG2 mutations in Japan. Allergol Int. 2006;55(2):115–9. https://doi.org/10.2332/allergolint.55.115.
44. Khan TA, et al. Novel RAG1 mutation and the occurrence of mycobacterial and Chromobacterium violaceum infections in a case of leaky SCID. Microb Pathog. 2017;109:114–9. https://doi.org/10.1016/j.micpath.2017.05.033.
45. Ktiouet S, et al. Omenn syndrome due to mutation of the RAG2 gene. J Eur Acad Dermatol Venereol. 2009;23(12):1449–51. https://doi.org/10.1111/j.1468-3083.2009.03232.x.
46. Kuijpers TW, et al. Idiopathic CD4+ T lymphopenia without autoimmunity or granulomatous disease in the slipstream of RAG mutations. Blood. 2011;117(22):5892–6. https://doi.org/10.1182/blood-2011-01-329052.
47. Kuo TC, Schlissel MS. Mechanisms controlling expression of the RAG locus during lymphocyte development. Curr Opin Immunol. 2009;21(2):173–8. https://doi.org/10.1016/j.coi.2009.03.008.
48. Kutukculer N, et al. Novel mutations and diverse clinical phenotypes in recombinase-activating gene 1 deficiency. Ital J Pediatr. 2012;38:8. https://doi.org/10.1186/1824-7288-38-8.
49. Lawless D, et al. Predicting the occurrence of variants in RAG1 and RAG2. J Clin Immunol. 2019;39(7):688–701. https://doi.org/10.1007/s10875-019-00670-z.
50. Lee YN, et al. A systematic analysis of recombination activity and genotype-phenotype correlation in human recombination-activating gene 1 deficiency. J Allergy Clin Immunol. 2014;133(4):1099–108. https://doi.org/10.1016/j.jaci.2013.10.007.
51. Lev A, et al. Characterizing T cells in SCID patients presenting with reactive or residual T lymphocytes. Clin Dev Immunol. 2012;2012:261470. https://doi.org/10.1155/2012/261470.
52 Luger K, et al. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature. 1997;389(6648):251–60. https://doi.org/10.1038/38444.
53. Luk ADW, et al. Family history of early infant death correlates with earlier age at diagnosis but not shorter time to diagnosis for severe combined immunodeficiency. Front Immunol. 2017;8:808. https://doi.org/10.3389/fimmu.2017.00808.
54. Matangkasombut P, et al. Lack of iNKT cells in patients with combined immune deficiency due to hypomorphic RAG mutations. Blood. 2008;111(1):271–4. https://doi.org/10.1182/blood-2007-06-096487.
55. Meshaal S, et al. Mutations in recombination activating gene 1 and 2 in patients with severe combined immunodeficiency disorders in Egypt. Clin Immunol. 2015;158(2):167–73. https://doi.org/10.1016/j.clim.2015.04.003.
56. Meshaal SS, et al. Phenotypical heterogeneity in RAG-deficient patients from a highly consanguineous population. Clin Exp Immunol. 2019;195(2):202–12. https://doi.org/10.1111/cei.13222.
57. Noordzij JG, et al. The immunophenotypic and immunogenotypic B-cell differentiation arrest in bone marrow of RAG-deficient SCID patients corresponds to residual recombination activities of mutated RAG proteins. Blood. 2002;100(6):2145–52.
58. Patiroglu T, Akar HH, Van Der Burg M. Three faces of recombination activating gene 1 (RAG1) mutations. Acta Microbiol Immunol Hung. 2015;62(4):393–401. https://doi.org/10.1556/030.62.2015.4.4.
59. Safaei S, et al. IL7R and RAG1/2 genes mutations/polymorphisms in patients with SCID. Iran J Allergy Asthma Immunol. 2011;10(2):129–32. https://doi.org/10.1002/ijaai.129132.
60. Schroder C, et al. Evaluation of RAG1 mutations in an adult with combined immunodeficiency and progressive multifocal leukoencephalopathy. Clin Immunol. 2017;179:1–7. https://doi.org/10.1016/j.clim.2016.12.013.
61. Schuetz C, et al. An immunodeficiency disease with RAG mutations and granulomas. N Engl J Med. 2008;358(19):2030–8. https://doi.org/10.1056/NEJMoa073966.
62. Schuetz C, et al. SCID patients with ARTEMIS vs RAG deficiencies following HCT: increased risk of late toxicity in ARTEMIS-deficient SCID. Blood. 2014;123(2):281–9. https://doi.org/10.1182/blood-2013-01-476432.
63. Schwarz K, et al. RAG mutations in human B cell-negative SCID. Science. 1996;274(5284):97–9. https://doi.org/10.1126/science.274.5284.97.
64. Sharapova SO, et al. Molecular characteristics, clinical and immunologic manifestations of 11 children with Omenn syndrome in East Slavs (Russia, Belarus, Ukraine). J Clin Immunol. 2016;36(1):46–55. https://doi.org/10.1007/s10875-015-0216-7.
65. Sharapova SO, et al. The clinical and genetic spectrum of 82 patients with RAG deficiency including a c. 256_257delAA founder variant in Slavic countries. Front Immunol. 2020;11:900.
66. Sheehan WJ, et al. Novel presentation of Omenn syndrome in association with aniridia. J Allergy Clin Immunol. 2009;123(4):966–9. https://doi.org/10.1016/j.jaci.2008.12.007.
67. Shen J, et al. A Novel RAG1 Mutation in a compound heterozygous status in a child with Omenn syndrome. Front Genet. 2019;10:913. https://doi.org/10.3389/fgene.2019.00913.
68. Signorini S, et al. Intrathymic restriction and peripheral expansion of the T-cell repertoire in Omenn syndrome. Blood. 1999;94(10):3468–78.

69. Sobacchi C, et al. RAG-dependent primary immunodeficiencies. Hum Mutat. 2006;27(12):1174–84. https://doi.org/10.1002/humu.20408.
70. Stanhope-Baker P, et al. Cell type–specific chromatin structure determines the targeting of V(D)J recombinase activity in vitro. Cell. 1996;85:887–97.
71. Strauss KA, et al. Clinical application of DNA microarrays: molecular diagnosis and HLA matching of an Amish child with severe combined immune deficiency. Clin Immunol. 2008;128(1):31–8. https://doi.org/10.1016/j.clim.2008.02.016.
72. Tabori U, et al. Detection of RAG mutations and prenatal diagnosis in families presenting with either T-B- severe combined immunodeficiency or Omenn's syndrome. Clin Genet. 2004;65(4):322–6. https://doi.org/10.1111/j.1399-0004.2004.00227.x.
73. Tirosh I, et al. Recombination activity of human recombination-activating gene 2 (RAG2) mutations and correlation with clinical phenotype. J Allergy Clin Immunol. 2019;143(2):726–35. https://doi.org/10.1016/j.jaci.2018.04.027.
74. Villa A, et al. Partial V(D)J recombination activity leads to Omenn syndrome. Cell. 1998;93(5):885–96. https://doi.org/10.1016/s0092-8674(00)81448-8.
75. Villa A, et al. V(D)J recombination defects in lymphocytes due to RAG mutations: severe immunodeficiency with a spectrum of clinical presentations. Blood. 2001;97(1):81–8. https://doi.org/10.1182/blood.v97.1.81.
76. Walter JE, et al. Broad-spectrum antibodies against self-antigens and cytokines in RAG deficiency. J Clin Invest. 2015;125(11):4135–48. https://doi.org/10.1172/JCI80477.
77. Xiao Z, et al. A novel missense RAG-1 mutation results in T-B-NK+ SCID in Athabascan-speaking Dine Indians from the Canadian Northwest Territories. Eur J Hum Genet. 2009;17(2):205–12. https://doi.org/10.1038/ejhg.2008.150.
78. Yancopoulos GD, Alt FW. Developmentally controlled and tissue-specific expression of unrearranged VH gene segments. Cell. 1985. 40: 271–281. J Immunol. 2012;188(1):10–20.
79. Zhang J, et al. Novel RAG1 mutation in a case of severe combined immunodeficiency. Pediatrics. 2005;116(3):e445–9. https://doi.org/10.1542/peds.2005-0369.
80. Zhang ZY, et al. Clinical characteristics and molecular analysis of three Chinese children with Omenn syndrome. Pediatr Allergy Immunol. 2011;22(5):482–7. https://doi.org/10.1111/j.1399-3038.2010.01126.x.
81. Shiraishi M, Hayatsu H. High-speed conversion of cytosine to uracil in bisulfite genomic sequencing analysis of DNA methylation. DNA Res. 2004;11(6):409–15. https://doi.org/10.1093/dnares/11.6.409.
82. Suzuki MM, et al. CpG methylation is targeted to transcription units in an invertebrate genome. Genome Res. 2007;17(5):625–31. https://doi.org/10.1101/gr.6163007.
83. Zeng J, Yi SV. DNA methylation and genome evolution in honeybee: gene length, expression, functional enrichment covary with the evolutionary signature of DNA methylation. Genome Biol Evol. 2010;2:770–80. https://doi.org/10.1093/gbe/evq060.
84. Cooper DN, Youssoufian H. The CpG dinucleotide and human genetic disease. Hum Genet. 1988;78(2):151–5. https://doi.org/10.1007/bf00278187.
85. Cooper DN, et al. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. Hum Genomics. 2010;4(6):406–10. https://doi.org/10.1186/1479-7364-4-6-406.
86. Zhou Y, et al. The impact of DNA methylation dynamics on the mutation rate during human germline development. G3. 2020;10(9):3337–46. https://doi.org/10.1534/g3.120.401511. (**Bethesda**).
87. Jones JM, Gellert M. The taming of a transposon: V(D)J recombination and the immune system. Immunol Rev. 2004;200:233–48. https://doi.org/10.1111/j.0105-2896.2004.00168.x.
88. Thompson CB. New insights into V(D)J recombination and its role in the evolution of the immune system. Immunity. 1995;3(5):531–9. https://doi.org/10.1016/1074-7613(95)90124-8.
89. Fugmann SD. The origins of the Rag genes—from transposition to V(D)J recombination. Semin Immunol. 2010;22(1):10–6. https://doi.org/10.1016/j.smim.2009.11.004.
90. Giorgetti OB, et al. Origin and evolutionary malleability of T cell receptor α diversity. Nature. 2023;619(7968):193–200. https://doi.org/10.1038/s41586-023-06218-x.
91. Kapitonov VV, Jurka J. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol. 2005;3(6):e181. https://doi.org/10.1371/journal.pbio.0030181.
92. Fugmann SD, et al. An ancient evolutionary origin of the Rag1/2 gene locus. Proc Natl Acad Sci U S A. 2006;103(10):3728–33. https://doi.org/10.1073/pnas.0509720103.
93. Marchalonis JJ, et al. Phylogenetic emergence and molecular evolution of the immunoglobulin family. Adv Immunol. 1998;70:417–506. https://doi.org/10.1016/s0065-2776(08)60392-2.
94. Davidson EH. The sea urchin genome: where will it lead us? Science. 2006;314(5801):939–40. https://doi.org/10.1126/science.1136252.
95. Camargo MM, Nahum LA. Adapting to a changing world: RAG genomics and evolution. Hum Genomics. 2005;2(2):132–7. https://doi.org/10.1186/1479-7364-2-2-132.
96. Liu C, et al. Structural insights into the evolution of the RAG recombinase. Nat Rev Immunol. 2022;22(6):353–70. https://doi.org/10.1038/s41577-021-00628-6.
97. Teng G, Schatz DG. Regulation and evolution of the RAG recombinase. Adv Immunol. 2015;128:1–39. https://doi.org/10.1016/bs.ai.2015.07.002.
98. Zhang Y, et al. Transposon molecular domestication and the evolution of the RAG recombinase. Nature. 2019;569:79–84.
99. Zhou W, et al. DNA methylation enables transposable element-driven genome expansion. Proc Natl Acad Sci U S A. 2020;117(32):19359–66. https://doi.org/10.1073/pnas.1921719117.
100. Parvathy ST, Udayasuriyan V, Bhadana V. Codon usage bias. Mol Biol Rep. 2022;49(1):539–65. https://doi.org/10.1007/s11033-021-06749-4.
101. Hershberg R, Petrov DA. Selection on codon bias. Annu Rev Genet. 2008;42:287–99. https://doi.org/10.1146/annurev.genet.42.110807.091442.
102. Duret L. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev. 2002;12(6):640–9. https://doi.org/10.1016/s0959-437x(02)00353-2.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Disclaimer/Publisher's note The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Springer and/or the editor(s). Springer and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.