Biochemical and Biophysical Research Communications 462 (2015) 301–313

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

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

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

Abhishek Kumar [a, b, *], Anita Bhandari [c], Sandeep J. Sarde [a, d], Sekhar Muppavarapu [d], Ravi Tandon [e, **]

[a] Department of Genetics & Molecular Biology in Botany, Institute of Botany, Christian-Albrechts-University at Kiel, Kiel, Germany
[b] Division of Molecular Genetic Epidemiology, German Cancer Research Center, Heidelberg, Germany
[c] Molecular Physiology, Institute of Zoology, Christian-Albrechts-University at Kiel, Kiel, Germany
[d] Agrigenomics Program, Christian-Albrechts-University at Kiel, Kiel, Germany
[e] School of Biotechnology, Jawaharlal Nehru University, New Delhi, India

ARTICLE INFO

Article history:
Received 8 April 2015
Available online 12 May 2015

Keywords:
RAG1
Ancestral locus
Phylogenetic analysis
RAG1 variants
Missense variants
Non-coding variants

ABSTRACT

The recombination-activating genes (RAGs) encode for V(D)J recombinases responsible for rearrangements of antigen-receptor genes during T and B cell development, and RAG expression is known to correlate strictly with the process of rearrangement. There have been several studies of RAG1 illustrating biochemical, physiological and immunological properties. Hitherto, there are limited studies on RAG1 focusing molecular phylogenetic analyses, evolutionary traits, and genetic variants in human populations. Hence, there is a need of a comprehensive study on this topic. In the current report, we have shed light into insights of evolutionary traits and genetic variants of human RAG1 gene using 1092 genomes from human populations. Syntenic analyses revealed that two RAG genes are physically linked and conserved on the same locus in head-to-head orientation from sea urchin to human for about 550 MY. Spliceosomal introns have been in invaded in fishes and sea urchin, whereas gene structures of RAG1 gene from tetrapods remained single exon architecture. We compiled 751 genetic variants in human RAG1 gene using 1092 human genomes; where major stockholders of variant classes are 79% single nucleotide polymorphisms (SNPs), 12.2% somatic single nucleotide variants (somatic SNVs) and 6.8% deletion. Out of 267 missense variants, 140 are deleterious mutations. We identified 284 non-coding variants with 94% regulatory in nature.

1. Introduction

Robustness of the adaptive immune system are dependent on the highly coordinated assembly of the genes encoding immunoglobulin and T cell receptor subunits in a process known as V(D)J recombination [1,2]. During the recombination process of immunoglobulin (Ig) or T-cell receptor (TCR) gene, there is a combination of variable (V), diversity (D) and joining (J) gene segments resulting into V–D–J exon. These recombination processes are controlled by two recombination activating genes (RAGs) in a cell lineage and stage specific manner [1,2]. Initially, these two RAG genes were deciphered during process of identifying the 'recombinase' responsible for V(D)J recombination [1,2]. The discovery of RAG is considered to be the hallmark of adaptive immunity, and RAG homologues were found in many jawed vertebrates. This precipitous procurement of RAGs assisted the immune system to achieve and maintain with diversity and this process is often called as 'the immunological big bang' [3,4]. Due to several permutation and combinations of three gene segments, V(D)J recombination events lead to the wide arrays of sequence diversity in the antigen receptor repertoire. The human RAG1 protein consists of 1043 residues, spanning two domains, the N-terminal non-core domain (NCD) and the C-terminal core RAG domain and these two domains have several conserved regions and are essential for recombination activity of the RAG1 [5].

* Corresponding author. Division of Molecular Genetic Epidemiology, German Cancer Research Center, Heidelberg, Germany.
** Corresponding author. School of Biotechnology, Jawaharlal Nehru University, New Delhi, India.
E-mail addresses: abhishek.abhishekkumar@gmail.com (A. Kumar), ravitandon@jnu.ac.in (R. Tandon).

http://dx.doi.org/10.1016/j.bbrc.2015.04.125
0006-291X/

Despite several works were done to elucidate the mechanisms coordinating the adaptive immune system, there are many questions unanswered including the origin of RAG proteins, which give rise to adaptive immune system. In the current study, we aimed to carry out a comprehensive study of RAG1 using the sequence, molecular phylogeny and genetic variants of RAG1. We provide several vignettes of RAG1 gene including comprehensive and updated germline variants catalogue of RAG1 gene with implications of missense and non-coding variants.

2. Materials and methods

2.1. Mining RAG1 gene and gene structure prediction

RAG1 gene from vertebrate genomes were mined via Ensembl release 77 (October 2014) [6] using BLAST as listed in Table S1. To ensure accuracy of gene structures, predictions of the Ensembl [6] were coupled with predictions of AUGUSTUS gene prediction tool [7]. Human RAG1 was used as standard sequence for intron position mapping and numbering of intron positions, followed by suffixes a–c for their location as reported previously [8].

2.2. Sequence alignment, phylogenetic and synteny analyses

We created alignment using MUSCLE alignment program [9] with default setting and this alignment was manually adjusted with GENEDOC tool [10]. We reconstructed a phylogenetic tree with maximum likelihood method, based on the JTT matrix-based model [11] with 1000 bootstrap replicates. We imported all consensus trees to MEGA 6 software [12], where we edited and visualized these trees as per requirement. To detect the orthologs of RAG1 gene, we analyzed micro-synteny across different genomes using NCBI mapviewer [13] and ENSEMBL genome browser [14,15].

2.3. Computing catalogue of RAG1 variants in 1092 human genomes

RAG1 variants were generated from 1092 human genomes from 14 different populations available in 1000 genomes project [16]. Impact assessments of missense variants on the human RAG1 protein were computed using two “start of the art” tools, namely SIFT [17] and PolyPhen V2 [18], as described previously [19–22]. Regulatory nature of non-coding variants was detected using rSNPbase, which provides reliable and comprehensive regulatory annotations [23] and such variants are called regulatory SNP (rSNP).

3. Results

3.1. Spliceosomal introns invade RAG1 gene in fishes and sea urchin

RAG1 gene from tetrapods are intron-less such as a single coding-exon with size 3132 base pair (bp) and 3126 bp in human and duck, respectively (Fig. 1). However, this gene structure is not conserved in fishes. Coelacanth has two exons E1′ and E2′ of size 246 and 2882, respectively, invaded by an intron of size 39 bp. In contrast, ray-finned fishes possess three exons and two introns gene structure with first exon E1 is ranged from 327 to 338 bp, while exon E2 is ranged from 1112 to 1133 bp and exon E3 is sized from 1739 to 1754 bp. The introns I1 and I2 are ranged from 75 bp (in *Fugu*) to 92 (in stickleback) and 706 bp (in *Fugu*) to 848 bp (in Atlantic cod, platyfish and zebrafish), respectively. These two introns are inserted at position 103c and 476a (human RAG1 residue numbering) in the conserved domain in ray-finned fishes, while RAG1 of coelacanth depicts intron at the position 82a (Fig. 2). Ancestral SpRAG1L gene from sea urchin possesses four exon/three intron gene structure. Exons of SpRAG1L are named as e1–e4 with sizes 188, 528, 248, and 1800 bp, respectively. Whereas introns of SpRAG1L are called i1–i3 with size 1600, 118 and 585 bp, respectively. This study suggests that introns are invaded in the RAG1 gene from fishes and sea urchins.

3.2. RAG1 protein is highly conserved as illustrated by protein domains analyses

RAG1 protein is highly conserved in vertebrates as human RAG1 shares 75/88, 80/88, 69/82, and 58/70 percentage identities/similarities with turkey, turtle, coelacanth and platyfish, respectively (Table 1). Ancestral SpRAG1L protein from sea urchin shares lower conservation levels with vertebrate RAG1 as SpRAG1L protein shares only 19–20% and 33–35% identities and similarities, respectively (Table 1 and Fig. S1). Human RAG1 protein is 1043 amino acid long, which comprises several key domains or regions (Figs. 2 and S1). The size of the N-terminal region is 86 amino acid and it is not highly conserved. However, it possesses a single patch of amino acids, which is highly conserved in the region of 14–40. RAG1 protein from sea urchin lacks this N-terminal region. Up next is the recently identified CND (87–217 residues), that possesses sequence signals, which are critical for nuclear localization, zinc coordination, and interactions with nucleic acid [5]. This domain is conserved from sea urchin to human, with few indels, such as a large deletion in SpRAG1L protein with comparisons to vertebrates at the position 120 (human RAG1 amino acid numbering) and ray-finned fishes specific deletion of 4–5 residues at the position 187. In the region with 285–380 residues, there is a conserved zinc dimerization domain (ZDD), with two zinc fingers, namely RING, which demonstrates E3 ubiquitin ligase activity and zinc finger A (ZFA) (Figs. 2 and S1). The ZDD domain is conserved in vertebrates but not significantly conserved in the SpRAG1L, especially these two zinc fingers appears to be missing in the SpRAG1L (Fig. S1). The core RAG1 domain contains following regions: the nonamer binding domain (NBD) [24], and two domains namely the central with ZFB region and C-terminal domains [25] in the residue ranges of 389–464, 528–760, and 761–979, respectively. The entire core RAG1 domain is conserved in all species analyzed at higher level than the N-terminal regions (large segments with >85% identities [red color] in Fig. S1). This gives rise in the percentage sequence identities as human core RAG1 shares 89, 92, 84, 75 and 27 against core RAG1 of turkey, turtle, coelacanth platyfish and sea urchin, respectively. The recombination signal sequence (RSS) and the acidic residues in the active site [26,27] are also conserved from sea urchin to human (Fig. S1). Overall, protein domains of RAG1 protein is conserved from origin in sea urchin to human with few segmental exceptions in SpRAG1L (Fig. S1). Phylogenetic history of RAG1 is shown in Fig. 3A.

3.3. Human RAG1 locus is evolutionary conserved over period of ~550 MY

Two RAG genes are physically linked in the head-to-head orientation in the human chromosome 11 (Fig. 3B). This genomic locus is conserved in several mammals such as mouse (chromosome 2), rat (chromosome 3), pig (chromosome 2) and opossum (chromosome 5). Birds also possess this fragment as in duck (KB742537.1), chicken (chromosome 5), flycatcher (Scaffold JH603235.1), turkey (chromosome 5), and zebra finch (chromosome 5). Same chromosomal maps were found in two reptiles including anole lizard (Chromosome 1) and Chinese softshell turtle (Scaffold_JH209124.1) Frogs also maintained this locus as shown for *Xenopus* (Scaffold_GL172917.1). Fishes retained this RAG1–RAG2 locus such as coelacanth (Scaffold_JH126568.1) and also in several

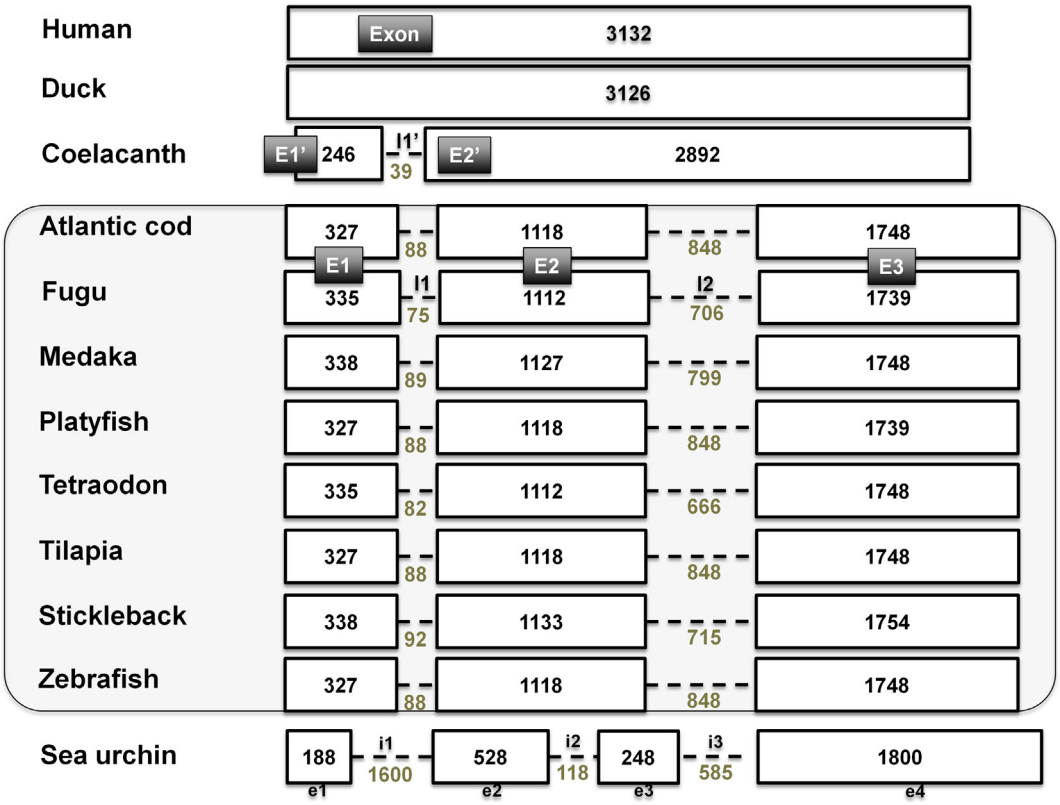

Fig. 1. Gene structure patterns depict that spliceosomal intron insertions in RAG1 genes from fishes and sea urchin. Variations in gene structures of RAG1 in different vertebrates illustrates that one, two and three introns are inserted in ray-finned fishes, coelacanth and sea urchin, respectively. However, RAG1 gene from tetrapods is intron-less.

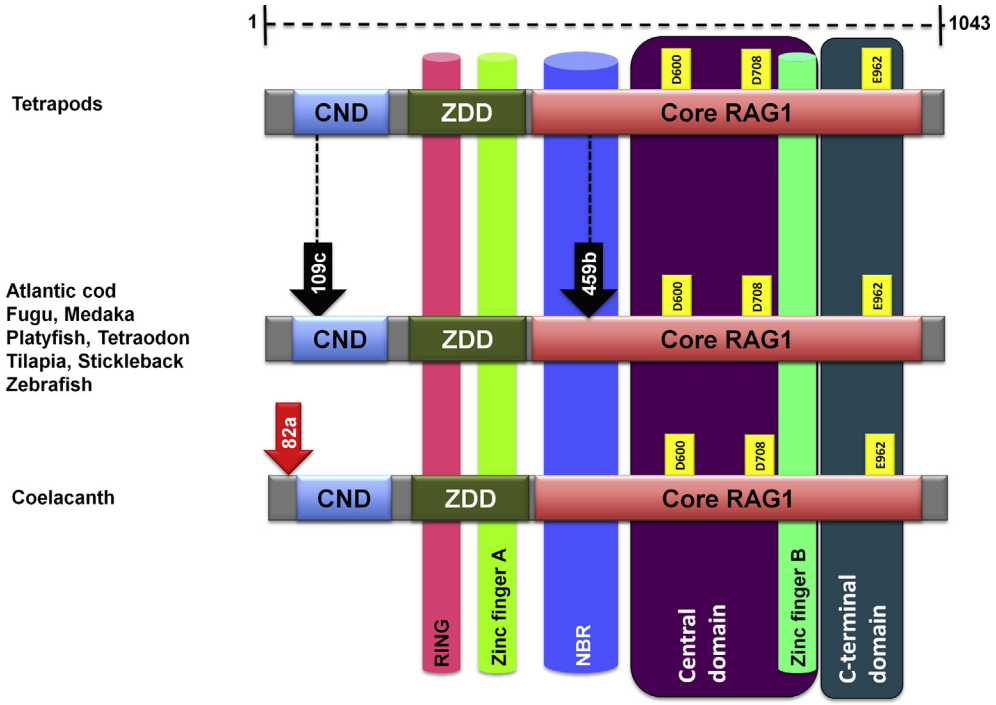

Fig. 2. Overview of protein domain organization of RAG1 corroborates that RAG1 is composed of the N-terminal non-core domain (NCD) and the core RAG1 domain. RAG1 protein have two novel intron inserted at position 103c and 476a in the conserved domain in ray-finned fishes, while RAG1 of coelacanth depicts intron at the position 82a (positioning according to human RAG1 protein).

Table 1
Ancestral RAG1 protein from sea urchin is highly divergent in comparisons to vertebrate RAG1 protein as illustrated by percentage identities and similarities values of RAG1 proteins from selected organisms.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	
1. Human		82	75	70	74	80	69	53	56	56	56	57	58	20	**% Sequence Identities**
2. Opossum	90		74	69	72	79	67	53	56	55	55	56	56	19	
3. Turkey	86	85		70	74	82	66	51	56	55	54	56	57	19	
4. *Xenopus*	83	82	82		68	72	66	52	55	55	54	56	55	19	
5. Lizard	85	83	85	81		78	67	51	55	54	54	56	56	19	
6. Turtle	88	87	90	83	87		69	52	57	56	55	57	57	19	
7. Coelacanth	82	80	80	79	80	80		54	58	56	57	58	58	19	
8. Cod	67	67	66	66	66	66	69		66	68	70	71	67	19	
9. Zebrafish	72	71	72	71	72	72	73	77		70	71	73	70	19	
10. Tetraodon	70	69	69	69	69	69	71	78	81		77	78	76	18	
11. Stickleback	71	70	70	70	70	69	71	80	82	86		84	79	19	
12. Tilapia	70	69	70	69	71	70	72	81	84	87	90		82	20	
13. Platyfish	70	68	70	69	70	69	71	76	80	84	85	87		19	
14. Strongylocentrotus	35	35	35	35	35	35	34	34	35	34	34	35	34		
	% Sequence Similarities														

ray-finned fishes for example, Atlantic cod (GeneScaffold_2196), medaka (chromosome 6), platyfish (Scaffold_JH556735.1), *Takifugu* (Scaffold_302), *Tetraodon* (chromosome 13) tilapia (Scaffold_GL831142.1), stickleback (group XIX), and zebrafish (chromosome 25). Notably, we have not detected RAG genes in two lamprey genomes. However, we found homologs of RAG genes as SpRAG1L and SpRAG2L in sea urchin genome, localized in same loci. Nevertheless, flanking genes cannot be confirmed, as majority of these genes are hypothetical or unknown genes in sea urchin. RAG1 gene from coelacanth, ray-finned fishes and sea urchin possess introns which are designed by + sign followed by number of introns such as 1, 2 and 3 introns (I), respectively. In the nutshell, RAG1/2 locus is maintained in all vertebrates after origin in sea urchin, at ~550 MY (Fig. 3B).

3.4. Variants of RAG1 gene encoded by 1092 human genomes

We computed variations in the RAG1 gene in 1092 human genomes from 14 different populations and details are provided in Fig. 4 with majority of these are SNPs. There are 751 variations (Table S2) in total with major components of 594 SNPs, 92 somatic SNVs, and 51 deletions, shared in eight variant types (Fig. 4). Top six

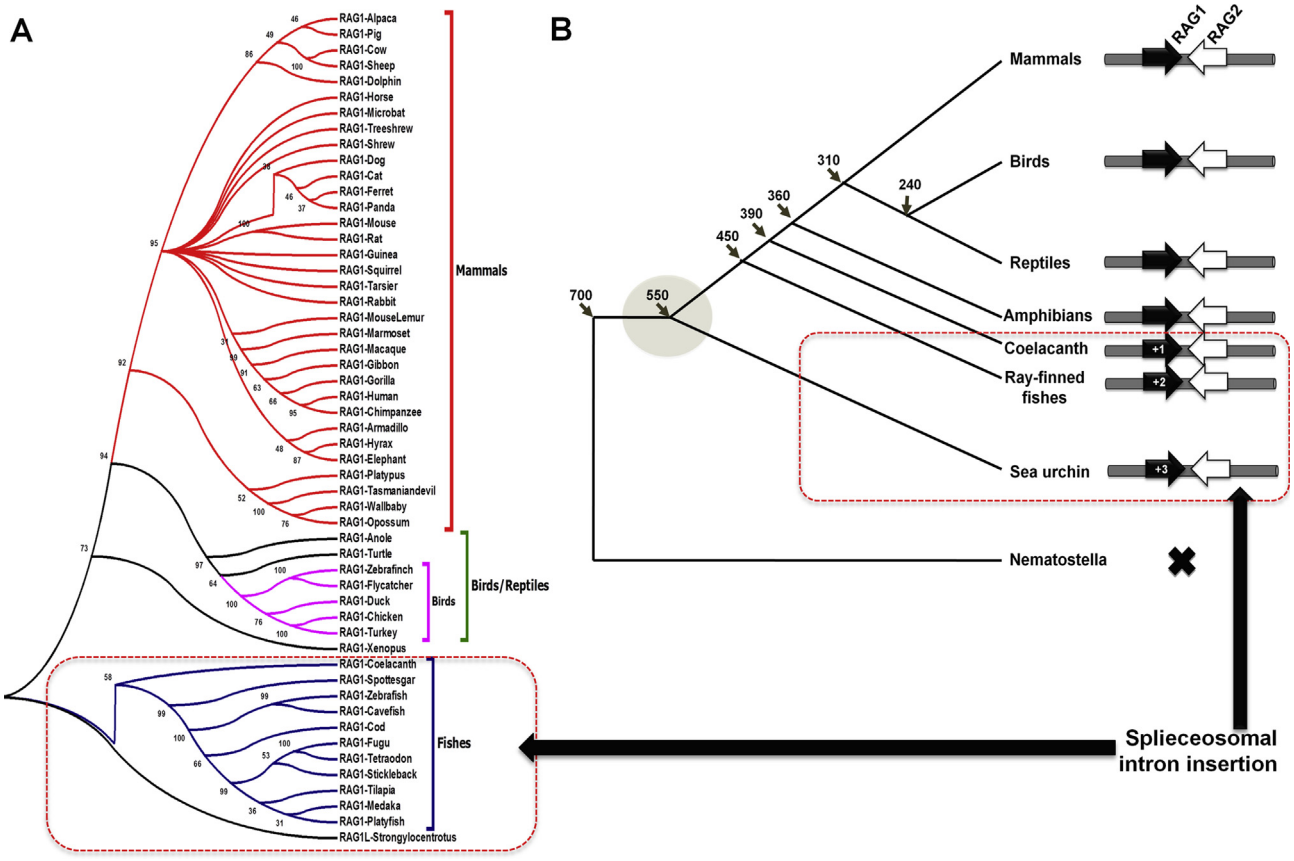

Fig. 3. Phylogenetic and syntenic analyses of RAG1 reveal points of spliceosomal intron insertion. **A**. Evolutionary history of RAG1 from selected vertebrates and sea urchin. **B**. Synteny analysis reveals that RAG1–RAG2 locus is conserved from sea urchin to human, dating ~550 MY old.

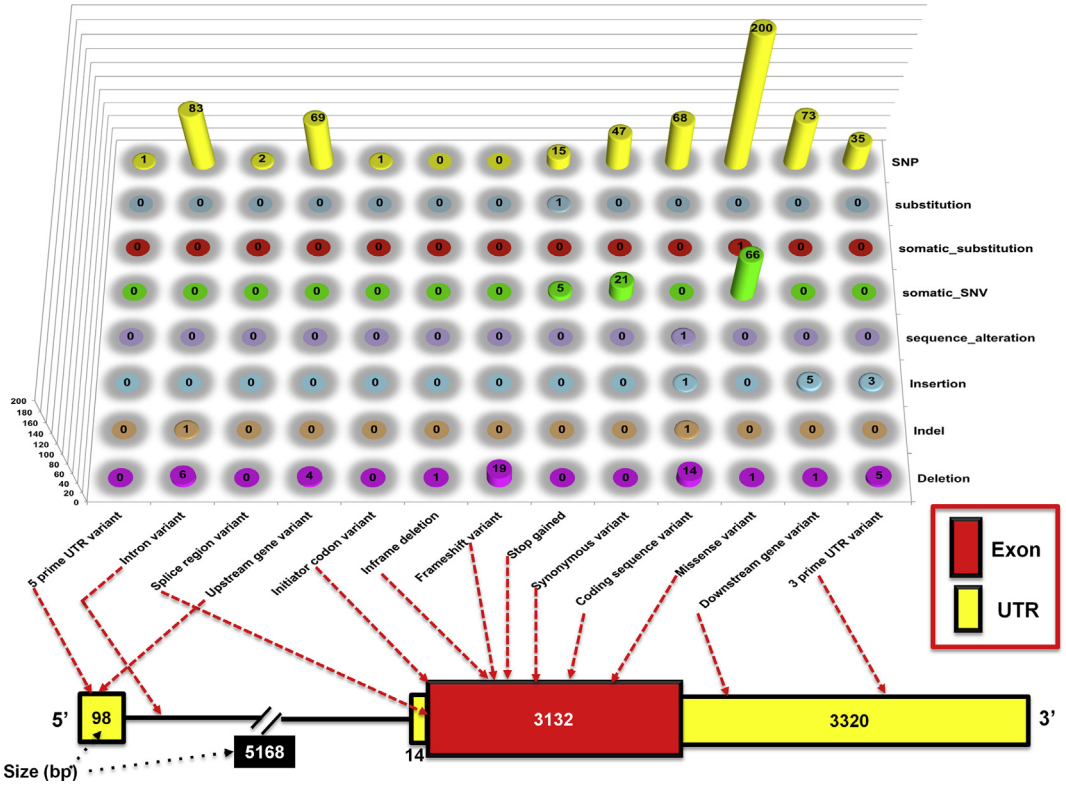

Fig. 4. Overview of genetic variants in the human RAG1 gene derived using 1092 genomes from 14 different human populations.

variant types (based on location) were missense variants (37.7%) and intron variants (12%), coding sequence variants (11.3%), downstream gene variants (10.5%), upstream gene variants (9.7%) and synonymous variants (9%). However, majority of variants were localized on the single coding exons and have implications on proteins such as missense variants (267) coding sequence variants (85), synonymous variants (68), stop gained (21), frameshift variants (19), splice region variants (2), initiator codon variant (1) and inframe deletion (1). Tables 2–5 provide details of these variants. Out of 751 RAG1 variants, 44.2% were validated variations.

3.5. Characterization of missense variants and disease co-relations

We further investigated 267 missense variations to examine what are the critical changes that can cause alteration in both RAG1 amino acid sequence and thus secondary structural elements (compiled in Table 2). Out of 267 missense variants, 140 are deleterious to the RAG1 protein as predicted using at least one of SIFT [17] and PolyPhen V2 [18] tools in Table 2. Herein, we describe roles of these variants in the different domains and regions of human RAG1 protein.

The N-terminal end before CND: The N-terminal end before CND has 18 missense variants including 11 deleterious variants (I23F, L32V, F33L, R34L, P42H, E48D, K49N, D51Y, H84R, K86M and PK85Q).

CND domain: Twenty-four variants were found in the CND region, of which 11 are deleterious in nature, namely G99S, L107V, S169L, H171Q, C176F, S181I, F187I, S188N, C192Y, E193K and S210F.

Region between CND and ZDD: The region joining CND and ZDD domains has 18 missense variants and half of these are deleterious, namely L221H, A243T, R244G R247G, R247C, R251G, S259I, N268S and I272K.

ZDD domain: We found 18 variants in the ZDD domain. RING finger region of ZDD has 10 missense variants, 50% of which are deleterious variants such as D302E, T306I, R314W, C328Y and P329S. The region between RING finger and zinc finger A has four missense variants, however none of these four are deleterious to RAG1 protein. The zinc finger A (ZFA) has 5 missense mutations, of which 40% are deleterious in nature, which included M355T and N364S.

Core RAG1 domain: The core RAG1 domain is largest region of the RAG1 protein and this domain harbors 177 human missense variants and 53% of there are deleterious to human RAG1 (Table 2). The NBR region of the core domain possesses 37 missense variants including 21 deleterious variants, predicted jointly by SIFT [17] and PolyPhen V2 [18] tools. These critical mutational positions are I385T, R394W, R396H, R396L, R396H (2 variants) R396C, T403S, R404Q, R404W, R410Q, D429G (2 variants), V433M, M435V A444V (2 variants), R449K, A456D, G460A and P467L.The region between NBR and core central domain harbors 15 mutations of which 11 are deleterious, namely, G496W, R507W (3 variants), R507Q, F520L, W522C (2 variants), P525S, P525H and P525L. Furthermore, the core central (CD) domain contains 53 mutations. Out of 53 missense variants, 24 are deleterious and following amino acid changes are resultants of these genetic variants – S546F, K558N, R559S (2 variants), R561H, R561C, R561S, A565D, A569S, D576V, M581I, M605T, K621R, K621N, A622P, R624H, R624C, R624H, R624C, R624H, R624C R699W, R716W and G720D. Additionally, zinc finger B region contains 13 mutations including five deleterious variants namely, R737H (2 variants), A740S, H753L and H753L. The C-terminal end of the core RAG1 domain has 59 missense variants including 31 positions deleterious to the RAG1 protein and these positions are E773G, R776L, R776W (2 variants), R778Q (2 variants), K780N, V782D, I810V, N823T, E827Q, R841W (2 variants), K843E, N855I, V866M, L885R, L885R, D887N Y912F, Y912C (2 variants), R918H, T925M, I956T, E965V, N968K, K969T and R975Q.

Table 2
Summary of missense RAG1 variants in the 1092 human genomes from 14 ethnicities. Deleterious missense variants are marked in bold and italics as based on SIFT value up to 0.05 plus PolyPhen-V2 score >0.5 and only by one of these two methods, respectively.

Mutants	Localization	Variation ID	AC	Codons	Alleles	MA	GFreq	SIFT	PolyPhen	Disease association
R314W	ZDD: within RING	RAG1base_RAG1_DNA:g.7232C > T	Y	CGG, TGG	C/T	–	–	1DEL(0)	627POD(0.626)	CHIDG
R778Q	Core RAG1; C-terminal domain	RAG1base_RAG1_DNA:g.8625G > A	R	CGG, CAG	G/A	–	–	21DEL(0.02)	877POD(0.876)	CHIDG
R778Q	Core RAG1; C-terminal domain	COSM1168721	R	CGG, CAG	G/A	–	–	21DEL(0.02)	877POD(0.876)	CHIDG
R975Q	Core RAG1; C-terminal domain	rs150739647	R	CGG, CAG	G/A	–	–	1DEL(0)	916PRD(0.915)	CHIDG
R507W	Core RAG1; region between NBR and CD-core	rs104894298	Y	CGG, TGG	C/T	–	–	1DEL(0)	950PRD(0.949)	CHIDG, OS
R507W	Core RAG1; region between NBR and CD-core	RAG1base_HSRAG1:g.1631C > T	Y	CGG, TGG	C/T	–	–	1DEL(0)	871POD(0.87)	CHIDG, OS
R507W	Core RAG1; region between NBR and CD-core	RAG1base_RAG1_DNA:g.7811C > T	Y	CGG, TGG	C/T	–	–	1DEL(0)	871POD(0.87)	CHIDG, OS
R737H	Core RAG1; CD-core: ZFB	RAG1base_RAG1_DNA:g.8502G > A	R	CGT, CAT	G/A	–	–	1DEL(0)	976PRD(0.975)	CHIDG, OS
R737H	Core RAG1; CD-core: ZFB	rs104894286	R	CGT, CAT	G/A	–	–	1DEL(0)	976PRD(0.975)	CHIDG, OS
C328Y	ZDD: within RING	RAG1base_HSRAG1:g.1095G > A	R	TGT, TAT	G/A	–	–	1DEL(0)	855POD(0.854)	OS
R410Q	Core RAG1; within NBR	rs199474684	R	CGG, CAG	G/A	–	–	1DEL(0)	910PRD(0.909)	OS
D429G	Core RAG1; within NBR	rs104894292	R	GAT, GGT	A/G	–	–	1DEL(0)	910PRD(0.909)	OS
D429G	Core RAG1; within NBR	RAG1base_HSRAG1:g.1398A > G	R	GAT, GGT	A/G	–	–	1DEL(0)	910PRD(0.909)	OS
M435V	Core RAG1; within NBR	rs141524540	R	ATG, GTG	A/G	–	–	1DEL(0)	985PRD(0.984)	OS
A444V	Core RAG1; within NBR	RAG1base_HSRAG1:g.1443C > T	Y	GCG, GTG	C/T	–	–	11DEL(0.01)	806POD(0.805)	OS
A444V	Core RAG1; within NBR	rs199474685	Y	GCG, GTG	C/T	–	–	21DEL(0.02)	985PRD(0.984)	OS
L454Q	Core RAG1; within NBR	rs199474677	W	CTG, CAG	T/A	–	–	–	–	OS
R474H	Core RAG1; within NBR	RAG1base_HSRAG1:g.1533G > A	R	CGT, CAT	G/A	–	–	–	–	OS
R474H	Core RAG1; within NBR	rs199474686	R	CGT, CAT	G/A	–	–	–	–	OS
R474H	Core RAG1; within NBR	RAG1base_RAG1_DNA:g.7713G > A	R	CGT, CAT	G/A	–	–	–	–	OS
W522C	**Core RAG1; region between NBR and CD-core**	**rs193922461**	**K**	**TGG, TGT**	**G/T**	–	–	**1DEL(0)**	**976PRD(0.975)**	**OS**
W522C	**Core RAG1; region between NBR and CD-core**	**RAG1base_HSRAG1:g.1678G > T**	**K**	**TGG, TGT**	**G/T**	–	–	**1DEL(0)**	**976PRD(0.975)**	**OS**
R561H	**Core RAG1; CD-core**	**RAG1base_RAG1_DNA:g.7974G > A**	**R**	**CGC, CAC**	**G/A**	–	–	**1DEL(0)**	**989PRD(0.988)**	**OS**
R561C	**Core RAG1; CD-core**	**RAG1base_HSRAG1:g.1793C > T**	**Y**	**CGC, TGC**	**C/T**	–	–	**1DEL(0)**	**993PRD(0.992)**	**OS**
R561S	**Core RAG1; CD-core**	**COSM415610**	**M**	**CGC, AGC**	**C/A**	–	–	**11DEL(0.01)**	**455POD(0.454)**	**OS**
E669G	Core RAG1; CD-core	RAG1base_HSRAG1:g.2118A > G	R	GAG, GGG	A/G	–	–	–	–	OS
R699W	Core RAG1; CD-core	**rs199474676**	**Y**	**CGG, TGG**	**C/T**	–	–	**1DEL(0)**	**916PRD(0.915)**	**OS**
H753L	Core RAG1; CD-core: ZFB	**rs199474687**	**W**	**CAT, CTT**	**A/T**	–	–	**1DEL(0)**	**995PRD(0.994)**	**OS**
H753L	**Core RAG1; CD-core: ZFB**	**RAG1base_HSRAG1:g.2370A > T**	**W**	**CAT, CTT**	**T/A**	–	–	**1DEL(0)**	**999PRD(0.998)**	**OS**
L885R	**Core RAG1; C-terminal domain**	**RAG1base_RAG1_DNA:g.8946T > G**	**K**	**CTG, CGG**	**T/G**	–	–	**1DEL(0)**	**976PRD(0.975)**	**OS**
L885R	**Core RAG1; C-terminal domain**	**RAG1base_HSRAG1:g.2766T > G**	**K**	**CTG, CGG**	**T/G**	–	–	**1DEL(0)**	**927PRD(0.926)**	**OS**
L885R	**Core RAG1; C-terminal domain**	**rs199474691**	**K**	**CTG, CGG**	**T/G**	–	–	**1DEL(0)**	**878POD(0.877)**	**OS**
Y912C	**Core RAG1; C-terminal domain**	**RAG1base_HSRAG1:g.2847A > G**	**R**	**TAC, TGC**	**A/G**	–	–	**21DEL(0.02)**	**878POD(0.877)**	**OS**
Y912C	**Core RAG1; C-terminal domain**	**rs104894290**	**R**	**TAC, TGC**	**A/G**	–	–	**21DEL(0.02)**	**878POD(0.877)**	**OS**
S401P	Core RAG1; within NBR	rs199474682	Y	TCG, CCG	T/C	–	–	–	–	OS
R624C	**Core RAG1; CD-core**	**rs199474688**	**Y**	**CGT, TGT**	**C/T**	–	–	**1DEL(0)**	**976PRD(0.975)**	**OS**
R624C	**Core RAG1; CD-core**	**RAG1base_RAG1_DNA:g.8162C > T**	**Y**	**CGT, TGT**	**C/T**	–	–	**1DEL(0)**	**976PRD(0.975)**	**OS**
E669G	Core RAG1; CD-core	rs199474689	R	GAG, GGG	A/G	–	–	–	–	OS
R699W	Core RAG1; CD-core	COSM287899	Y	CGG, TGG	C/T	–	–	–	–	OS
R737H	Core RAG1; CD-core: ZFB	RAG1base_HSRAG1:g.2322G > A	R	CGT, CAT	G/A	–	–	–	–	OS
R737H	Core RAG1; CD-core: ZFB	COSM1470480	R	CGT, CAT	G/A	–	–	–	–	OS
E722K	Core RAG1; CD-core: ZFB	*rs28933392*	*R*	*GAG, AAG*	*G/A*	–	–	*71TOL(0.07)*	***986PRD(0.985)***	***SCID***
V433M	**Core RAG1; within NBR**	**rs199474679**	**R**	**GTG, ATG**	**G/A**	–	–	**1DEL(0)**	**644POD(0.643)**	**OS, SCID**
R559S	**Core RAG1; CD-core**	**RAG1base_RAG1_DNA:g.7969G > T**	**K**	**AGG, AGT**	**G/T**	–	–	**1DEL(0)**	**985PRD(0.984)**	**OS, SCID**
R559S	**Core RAG1; CD-core**	**rs199474681**	**K**	**AGG, AGT**	**G/T**	–	–	**1DEL(0)**	**932PRD(0.931)**	**OS, SCID**
R624H	**Core RAG1; CD-core**	**RAG1base_HSRAG1:g.1983G > A**	**R**	**CGT, CAT**	**G/A**	–	–	**1DEL(0)**	**965PRD(0.964)**	**OS, SCID**
R624H	**Core RAG1; CD-core**	**RAG1base_RAG1_DNA:g.8163G > A**	**R**	**CGT, CAT**	**G/A**	–	–	**1DEL(0)**	**965PRD(0.964)**	**OS, SCID**
R841W	**Core RAG1; C-terminal domain**	**RAG1base_RAG1_DNA:g.8813C > T**	**Y**	**CGG, TGG**	**C/T**	–	–	**1DEL(0)**	**965PRD(0.964)**	**T-CMVA**
R841W	**Core RAG1; C-terminal domain**	**rs104894287**	**Y**	**CGG, TGG**	**C/T**	–	–	**1DEL(0)**	**965PRD(0.964)**	**T-CMVA**
Q981P	**C-terminal end after Core RAG1**	**RAG1base_RAG1_DNA:g.9234A > C**	**M**	**CAG, CCG**	**A/C**	–	–	**1DEL(0)**	**976PRD(0.975)**	**T-CMVA**
Q19R	N-terminal end before CND	rs150201913	R	CAG, CGG	A/G	–	–	–	–	–
I23F	**N-terminal end before CND**	**COSM231244**	**W**	**ATT, TTT**	**A/T**	–	–	**31DEL(0.03)**	**705POD(0.704)**	–
L32V	**N-terminal end before CND**	**rs149364682**	**S**	**CTG, GTG**	**C/G**	–	–	**1DEL(0)**	**998PRD(0.997)**	–
F33L	**N-terminal end before CND**	**TMP_ESP_11_36594953**	**M**	**TTC, TTA**	**C/A**	–	–	**1DEL(0)**	**963PRD(0.962)**	–
R34L	**N-terminal end before CND**	**COSM368513**	**K**	**CGG, CTG**	**G/T**	–	–	**1DEL(0)**	**963PRD(0.962)**	–
R34Q	N-terminal end before CND	TMP_ESP_11_36594955	R	CGG, CAG	G/A	–	–	–	–	–
R34W	*N-terminal end before CND*	*COSM1353713*	*Y*	*CGG, TGG*	*C/T*	–	–	***1DEL(0)***	*3BEN(0.002)*	–
R36I	*N-terminal end before CND*	*COSM1507897*	*K*	*AGA, ATA*	*G/T*	–	–	*51DEL(0.05)*	*263BEN(0.262)*	–
P42H	**N-terminal end before CND**	**COSM186659**	**M**	**CCT, CAT**	**C/A**	–	–	**1DEL(0)**	**994PRD(0.993)**	–
E48D	**N-terminal end before CND**	**COSM1507896**	**W**	**GAA, GAT**	**A/T**	–	–	**1DEL(0)**	**993PRD(0.992)**	–
K49N	**N-terminal end before CND**	**COSM257851**	**K**	**AAG, AAT**	**G/T**	–	–	**1DEL(0)**	**993PRD(0.992)**	–
D51V	N-terminal end before CND	rs147486240	W	GAT, GTT	A/T	–	–	–	–	–
D51Y	**N-terminal end before CND**	**TMP_ESP_11_36595005**	**K**	**GAT, TAT**	**G/T**	–	–	**1DEL(0)**	**953PRD(0.952)**	–
V65A	N-terminal end before CND	rs143654819	Y	GTC, GCC	T/C	–	–	–	–	–
L80F	N-terminal end before CND	COSM1287527	K	TTG, TTT	G/T	–	–	–	–	–
H84R	**N-terminal end before CND**	**rs150199231**	**R**	**CAC, CGC**	**A/G**	**G**	**0.005**	**21DEL(0.02)**	**972PRD(0.971)**	–
K86M	**N-terminal end before CND**	**TMP_ESP_11_36595111**	**W**	**AAG, ATG**	**A/T**	–	–	**1DEL(0)**	**893POD(0.892)**	–

Table 2 (*continued*)

Mutants	Localization	Variation ID	AC	Codons	Alleles	MA	GFreq	SIFT	PolyPhen	Disease association
D93H	CND	COSM1507895	S	GAC, CAC	G/C	–	–	1001TOL(1)	1BEN(0)	–
E95G	CND	rs142057334	R	GAG, GGG	A/G	–	–	–	–	–
G99S	**CND**	**rs138676205**	**R**	**GGC, AGC**	**G/A**	**A**	**0.003**	**1DEL(0)**	**989PRD(0.988)**	–
L107V	**CND**	**rs144019501**	**S**	**CTT, GTT**	**C/G**	–	–	**1DEL(0)**	**989PRD(0.988)**	–
R112C	CND	rs146457887	Y	CGC, TGC	C/T	–	–	–	–	–
R112C	CND	COSM136829	Y	CGC, TGC	C/T	–	–	511TOL(0.51)	1BEN(0)	–
E122D	CND	COSM317943	S	GAG, GAC	G/C	–	–	–	–	–
K136Q	CND	RAG1base_RAG1_DNA:g.6698A > C	M	AAA, CAA	A/C	–	–	291TOL(0.29)	1BEN(0)	–
G139V	CND	rs140648865	K	GGC, GTC	G/T	–	–	1001TOL(1)	1BEN(0)	–
R142Q	CND	TMP_ESP_11_36595279	R	CGA, CAA	G/A	–	–	–	–	–
K144N	CND	rs144430517	K	AAG, AAT	G/T	–	–	–	–	–
W151R	CND	COSM376240	Y	TGG, CGG	T/C	–	–	–	–	–
A156V	CND	rs1801203	Y	GCC, GTC	C/T	–	–	611TOL(0.61)	1BEN(0)	–
S169L	**CND**	**rs4151027**	**Y**	**TCG, TTG**	**C/T**	**T**	**0.001**	**1DEL(0)**	**910PRD(0.909)**	–
H171Y	CND	TMP_ESP_11_36595365	Y	CAC, TAC	C/T	–	–	–	–	–
H171Q	**CND**	**TMP_ESP_11_36595367**	**S**	**CAC, CAG**	**C/G**	–	–	**1DEL(0)**	**874POD(0.873)**	–
C176F	**CND**	**rs149229197**	**K**	**TGC, TTC**	**G/T**	–	–	**1DEL(0)**	**910PRD(0.909)**	–
S181I	**CND**	**TMP_ESP_11_36595396**	**K**	**AGC, ATC**	**G/T**	–	–	**1DEL(0)**	**874POD(0.873)**	–
F187I	**CND**	**TMP_ESP_11_36595413**	**W**	**TTT, ATT**	**T/A**	–	–	**1DEL(0)**	**760POD(0.759)**	–
S188N	**CND**	**COSM1353714**	**R**	**AGC, AAC**	**G/A**	–	–	**1DEL(0)**	**874POD(0.873)**	–
C192Y	**CND**	**RAG1base_RAG1_DNA:g.6867G > A**	**R**	**TGT, TAT**	**G/A**	–	–	**1DEL(0)**	**874POD(0.873)**	–
E193K	**CND**	**rs34841221**	**R**	**GAG, AAG**	**G/A**	–	–	**1DEL(0)**	**910PRD(0.909)**	–
S210F	**CND**	**rs147440161**	**Y**	**TCC, TTC**	**C/T**	–	–	**1DEL(0)**	**910PRD(0.909)**	–
S210F	CND	COSM108879	Y	TCC, TTC	C/T	–	–	–	–	–
R218H	Region between CND and ZDD	rs202178215	R	CGT, CAT	G/A	–	–	–	–	–
R219Q	Region between CND and ZDD	rs147055289	R	CGG, CAG	G/A	–	–	–	–	–
R219W	Region between CND and ZDD	TMP_ESP_11_36595509	Y	CGG, TGG	C/T	–	–	–	–	–
L221H	**Region between CND and ZDD**	**rs138419861**	**W**	**CTC, CAC**	**T/A**	–	–	**1DEL(0)**	**745POD(0.744)**	–
Q242R	Region between CND and ZDD	rs76897604	R	CAA, CGA	A/G	G	0.003	–	–	–
A243T	**Region between CND and ZDD**	**COSM223454**	**R**	**GCA, ACA**	**G/A**	–	–	**1DEL(0)**	**913PRD(0.912)**	–
R244G	**Region between CND and ZDD**	**rs199474683**	**R**	**AGA, GGA**	**A/G**	–	–	**11DEL(0.01)**	**627POD(0.626)**	–
R247G	**Region between CND and ZDD**	**rs147203889**	**S**	**CGT, GGT**	**C/G**	–	–	**1DEL(0)**	**644POD(0.643)**	–
R247C	**Region between CND and ZDD**	**COSM1507894**	**Y**	**CGT, TGT**	**C/T**	–	–	**1DEL(0)**	**910PRD(0.909)**	–
R247H	Region between CND and ZDD	rs4151029	R	CGT, CAT	G/A	A	0.001	–	–	–
H249R	Region between CND and ZDD	rs3740955	R	CAC, CGC	A/G	A	0.409	–	–	–
R251G	**Region between CND and ZDD**	**rs144616804**	**R**	**AGA, GGA**	**A/G**	–	–	**1DEL(0)**	**644POD(0.643)**	–
S259I	**Region between CND and ZDD**	**TMP_ESP_11_36595630**	**K**	**AGC, ATC**	**G/T**	–	–	**1DEL(0)**	**644POD(0.643)**	–
M263L	Region between CND and ZDD	COSM336373	W	ATG, TTG	A/T	–	–	–	–	–
A267T	Region between CND and ZDD	rs148393376	R	GCC, ACC	G/A	–	–	–	–	–
N268S	**Region between CND and ZDD**	**TMP_ESP_11_36595657**	**R**	**AAC, AGC**	**A/G**	–	–	**1DEL(0)**	**847POD(0.846)**	–
I272K	**Region between CND and ZDD**	**TMP_ESP_11_36595669**	**W**	**ATA, AAA**	**T/A**	–	–	**1DEL(0)**	**847POD(0.846)**	–
S275N	Region between CND and ZDD	TMP_ESP_11_36595678	R	AGT, AAT	G/A	–	–	–	–	–
D302E	**ZDD: within RING**	**rs4151030**	**M**	**GAC, GAA**	**C/A**	**A**	**0.017**	**11DEL(0.01)**	**918PRD(0.917)**	–
T306I	**ZDD: within RING**	**rs145962212**	**Y**	**ACC, ATC**	**C/T**	–	–	**11DEL(0.01)**	**918PRD(0.917)**	–
R314W	ZDD: within RING	rs121918568	Y	CGG, TGG	C/T	–	–	–	–	–
R319T	ZDD: within RING	rs139883723	S	AGA, ACA	G/C	–	–	–	–	–
V323I	ZDD: within RING	rs144391165	R	GTC, ATC	G/A	–	–	1DEL(0)	1BEN(0)	–
G325D	ZDD: within RING	rs201462389	R	GGC, GAC	G/A	A	0.001	1DEL(0)	1BEN(0)	–
C328Y	ZDD: within RING	rs121918571	R	TGT, TAT	G/A	–	–	–	–	–
P329S	**ZDD: within RING**	**rs181967562**	**Y**	**CCC, TCC**	**C/T**	**T**	–	**1DEL(0)**	**855POD(0.854)**	–
C335F	ZDD; region between RING and ZFA	TMP_ESP_11_36595858	K	TGC, TTC	G/T	–	–	–	–	–
D339V	ZDD; region between RING and ZFA	COSM428954	W	GAC, GTC	A/T	–	–	131TOL(0.13)	393BEN(0.392)	–
V350I	ZDD; region between RING and ZFA	rs200758244	R	GTC, ATC	G/A	A	0.001	131TOL(0.13)	393BEN(0.392)	–
L354Q	ZDD; at the start of ZFA	COSM1507893	W	CTG, CAG	T/A	–	–	–	–	–
L354R	ZDD; at the start of ZFA	COSM1317316	K	CTG, CGG	T/G	–	–	–	–	–
M355T	**ZDD; within ZFA**	**rs151077440**	**Y**	**ATG, ACG**	**T/C**	–	–	**1DEL(0)**	**913PRD(0.912)**	–
N364S	**ZDD; within ZFA**	**rs186342720**	**R**	**AAT, AGT**	**A/G**	**G**	**0.001**	**1DEL(0)**	**939PRD(0.938)**	–
N364D	ZDD; within ZFA	TMP_ESP_11_36595944	R	AAT, GAT	A/G	–	–	–	–	–
I385T	**Core RAG1; within NBR**	**rs141049427**	**Y**	**ATT, ACT**	**T/C**	–	–	**11DEL(0.01)**	**919PRD(0.918)**	–
R394W	**Core RAG1; within NBR**	**RAG1base_RAG1_DNA:g.7472C > T**	**Y**	**CGG, TGG**	**C/T**	–	–	**1DEL(0)**	**939PRD(0.938)**	–
R396H	Core RAG1; within NBR	RAG1base_RAG1_DNA:g.7479G > A	R	CGC, CAC	G/A	–	–	–	–	–
R396C	Core RAG1; within NBR	RAG1base_HSRAG1:g.1298C > T	Y	CGC, TGC	C/T	–	–	–	–	–
R396H	**Core RAG1; within NBR**	**COSM304336**	**R**	**CGC, CAC**	**G/A**	–	–	**1DEL(0)**	**874POD(0.873)**	–
R396L	**Core RAG1; within NBR**	**RAG1base_HSRAG1:g.1299G > T**	**K**	**CGC, CTC**	**G/T**	–	–	**1DEL(0)**	**910PRD(0.909)**	–
R396H	**Core RAG1; within NBR**	**RAG1base_HSRAG1:g.1299G > A**	**R**	**CGC, CAC**	**G/A**	–	–	**1DEL(0)**	**760POD(0.759)**	–
R396H	**Core RAG1; within NBR**	**rs104894291**	**D**	**CGC, CAC**	**G/T/A**	–	–	**1DEL(0)**	**874POD(0.873)**	–
R396C	**Core RAG1; within NBR**	**RAG1base_RAG1_DNA:g.7478C > T**	**Y**	**CGC, TGC**	**C/T**	–	–	**1DEL(0)**	**874POD(0.873)**	–
R396C	Core RAG1; within NBR	rs104894289	Y	CGC, TGC	C/T	–	–	–	–	–
L402P	Core RAG1; within NBR	COSM1353716	Y	CTG, CCG	T/C	–	–	–	–	–
T403S	**Core RAG1; within NBR**	**rs202189218**	**W**	**ACT, TCT**	**A/T**	–	–	**1DEL(0)**	**127BEN(0.126)**	–
R404Q	**Core RAG1; within NBR**	**RAG1base_RAG1_DNA:g.7503G > A**	**R**	**CGG, CAG**	**G/A**	–	–	**1DEL(0)**	**847POD(0.846)**	–
R404W	**Core RAG1; within NBR**	**RAG1base_RAG1_DNA:g.7502C > T**	**Y**	**CGG, TGG**	**C/T**	–	–	**1DEL(0)**	**985PRD(0.984)**	–
R410Q	Core RAG1; within NBR	RAG1base_HSRAG1:g.1341G > A	R	CGG, CAG	G/A	–	–	111TOL(0.11)	393BEN(0.392)	–

(*continued on next page*)

Table 2 (*continued*)

Mutants	Localization	Variation ID	AC	Codons	Alleles	MA	GFreq	SIFT	PolyPhen	Disease association
V433M	Core RAG1; within NBR	RAG1base_HSRAG1:g.1409G > A	R	GTG, ATG	G/A	–	–	–	–	–
V433A	Core RAG1; within NBR	TMP_ESP_11_36596152	Y	GTG, GCG	T/C	–	–	–	–	–
M435V	Core RAG1; within NBR	RAG1base_HSRAG1:g.1415A > G	R	ATG, GTG	A/G	–	–	–	–	–
R449K	**Core RAG1; within NBR**	**RAG1base_RAG1_DNA:g.7638G > A**	**R**	**AGG, AAG**	**G/A**	–	–	**21DEL(0.02)**	**985PRD(0.984)**	–
R449K	Core RAG1; within NBR	rs4151031	R	AGG, AAG	G/A	A	0.004	–	–	–
A456D	**Core RAG1; within NBR**	**rs201779957**	**M**	**GCC, GAC**	**C/A**	–	–	**41DEL(0.04)**	**975PRD(0.974)**	–
G460A	**Core RAG1; within NBR**	**rs191898756**	**S**	**GGA, GCA**	**G/C**	**C**	**0.001**	**11DEL(0.01)**	**993PRD(0.992)**	–
P467L	**Core RAG1; within NBR**	**rs142486791**	**Y**	**CCA, CTA**	**C/T**	–	–	**31DEL(0.03)**	**985PRD(0.984)**	–
R474C	Core RAG1; within NBR	rs199474678	Y	CGT, TGT	C/T	–	–	–	–	–
R474S	Core RAG1; within NBR	RAG1base_RAG1_DNA:g.7712C > A	M	CGT, AGT	C/A	–	–	–	–	–
M487V	Core RAG1; region between NBR and CD-core	*rs112047157*	*R*	*ATG, GTG*	*A/G*	–	–	*221TOL(0.22)*	***963PRD(0.962)***	–
A493T	Core RAG1; region between NBR and CD-core	*TMP_ESP_11_36596331*	*R*	*GCC, ACC*	*G/A*	–	–	*91TOL(0.09)*	***805POD(0.804)***	–
G496W	**Core RAG1; region between NBR and CD-core**	**COSM314667**	**K**	**GGG, TGG**	**G/T**	–	–	**41DEL(0.04)**	**957PRD(0.956)**	–
R497K	Core RAG1; region between NBR and CD-core	COSM1353717	R	AGA, AAA	G/A	–	–	–	–	–
R507Q	**Core RAG1; region between NBR and CD-core**	**rs143969029**	**R**	**CGG, CAG**	**G/A**	**A**	**0.001**	**1DEL(0)**	**871POD(0.87)**	–
P515L	Core RAG1; region between NBR and CD-core	COSM1263747	Y	CCA, CTA	C/T	–	–	–	–	–
F520L	**Core RAG1; region between NBR and CD-core**	**rs61758790**	**K**	**TTT, TTG**	**T/G**	–	–	**1DEL(0)**	**965PRD(0.964)**	–
P525S	**Core RAG1; region between NBR and CD-core**	**rs4151032**	**Y**	**CCT, TCT**	**C/T**	**T**	**0.003**	**1DEL(0)**	**927PRD(0.926)**	–
P525H	**Core RAG1; region between NBR and CD-core**	**COSM1507892**	**M**	**CCT, CAT**	**C/A**	–	–	**1DEL(0)**	**976PRD(0.975)**	–
P525L	**Core RAG1; region between NBR and CD-core**	**COSM1263748**	**Y**	**CCT, CTT**	**C/T**	–	–	**1DEL(0)**	**965PRD(0.964)**	–
S546F	**Core RAG1; CD-core**	**rs148235228**	**Y**	**TCC, TTC**	**C/T**	–	–	**1DEL(0)**	**965PRD(0.964)**	–
V548L	Core RAG1; CD-core	COSM352369	K	GTG, TTG	G/T	–	–	–	–	–
D554N	Core RAG1; CD-core	COSM167756	R	GAC, AAC	G/A	–	–	–	–	–
K558N	**Core RAG1; CD-core**	**rs150790148**	**S**	**AAG, AAC**	**G/C**	–	–	**1DEL(0)**	**993PRD(0.992)**	–
R559S	Core RAG1; CD-core	RAG1base_HSRAG1:g.1789G > T	K	AGG, AGT	G/T	–	–	–	–	–
R561C	Core RAG1; CD-core	RAG1base_RAG1_DNA:g.7973C > T	Y	CGC, TGC	C/T	–	–	–	–	–
R561C	Core RAG1; CD-core	rs104894285	Y	CGC, TGC	C/T	–	–	1001TOL(1)	18BEN(0.017)	–
R561H	Core RAG1; CD-core	rs104894284	R	CGC, CAC	G/A	–	–	–	–	–
R561H	Core RAG1; CD-core	RAG1base_HSRAG1:g.1794G > A	R	CGC, CAC	G/A	–	–	–	–	–
D563Y	Core RAG1; CD-core	COSM1507891	K	GAT, TAT	G/T	–	–	–	–	–
A565D	**Core RAG1; CD-core**	**RAG1base_RAG1_DNA:g.7986C > A**	**M**	**GCT, GAT**	**C/A**	–	–	**11DEL(0.01)**	**806POD(0.805)**	–
A569S	**Core RAG1; CD-core**	**rs138101978**	**K**	**GCT, TCT**	**G/T**	–	–	**1DEL(0)**	**921PRD(0.92)**	–
D576V	**Core RAG1; CD-core**	**rs142333735**	**W**	**GAC, GTC**	**A/T**	–	–	**1DEL(0)**	**975PRD(0.974)**	–
D576Y	Core RAG1; CD-core	COSM186661	K	GAC, TAC	G/T	–	–	–	–	–
M581I	**Core RAG1; CD-core**	**COSM163910**	**S**	**ATG, ATC**	**G/C**	–	–	**1DEL(0)**	**575POD(0.574)**	–
D585G	Core RAG1; CD-core	COSM84060	R	GAC, GGC	A/G	–	–	–	–	–
S601T	Core RAG1; CD-core	rs145951370	W	TCT, ACT	T/A	–	–	–	–	–
S601P	Core RAG1; CD-core	COSM1507890	Y	TCT, CCT	T/C	–	–	–	–	–
C602W	Core RAG1; CD-core	RAG1base_RAG1_DNA:g.8098T > G	K	TGT, TGG	T/G	–	–	–	–	–
M605T	**Core RAG1; CD-core**	**COSM338957**	**Y**	**ATG, ACG**	**T/C**	–	–	**11DEL(0.01)**	**870POD(0.869)**	–
K621R	**Core RAG1; CD-core**	**COSM1353719**	**R**	**AAG, AGG**	**A/G**	–	–	**11DEL(0.01)**	**950PRD(0.949)**	–
K621N	**Core RAG1; CD-core**	**COSM13915**	**K**	**AAG, AAT**	**G/T**	–	–	**21DEL(0.02)**	**985PRD(0.984)**	–
A622P	**Core RAG1; CD-core**	**RAG1base_RAG1_DNA:g.8156G > C**	**S**	**GCA, CCA**	**C/G**	–	–	**1DEL(0)**	**957PRD(0.956)**	–
A622T	Core RAG1; CD-core	rs148380512	R	GCA, ACA	G/A	A	0.001	–	–	–
R624H	**Core RAG1; CD-core**	**rs199474680**	**R**	**CGT, CAT**	**G/A**	–	–	**1DEL(0)**	**965PRD(0.964)**	–
R624C	**Core RAG1; CD-core**	**RAG1base_HSRAG1:g.1982C > T**	**Y**	**CGT, TGT**	**C/T**	–	–	**1DEL(0)**	**976PRD(0.975)**	–
S637N	Core RAG1; CD-core	COSM138808	R	AGC, AAC	G/A	–	–	–	–	–
N650T	Core RAG1; CD-core	COSM328851	M	AAC, ACC	A/C	–	–	–	–	–
M661I	Core RAG1; CD-core	COSM1353721	R	ATG, ATA	G/A	–	–	411TOL(0.41)	2BEN(0.001)	–
T670M	Core RAG1; CD-core	rs139863630	Y	ACG, ATG	C/T	–	–	–	–	–
T670M	Core RAG1; CD-core	COSM186663	Y	ACG, ATG	C/T	–	–	–	–	–
I679V	Core RAG1; CD-core	rs143227621	R	ATT, GTT	A/G	G	–	–	–	–
R699Q	Core RAG1; CD-core	TMP_ESP_11_36596950	R	CGG, CAG	G/A	–	–	–	–	–
G709D	Core RAG1; CD-core	**RAG1base_RAG1_DNA:g.8418G > A**	**R**	**GGC, GAC**	**G/A**	–	–	**31DEL(0.03)**	**455POD(0.454)**	–
G709C	Core RAG1; CD-core	COSM1492530	K	GGC, TGC	G/T	–	–	–	–	–
V715L	Core RAG1; CD-core	COSM322973	K	GTG, TTG	G/T	–	–	–	–	–
R716W	Core RAG1; CD-core	RAG1base_RAG1_DNA:g.8438C > T	Y	CGG, TGG	C/T	–	–	–	–	–
R716W	**Core RAG1; CD-core**	**rs199776076**	**Y**	**CGG, TGG**	**C/T**	–	–	**1DEL(0)**	**916PRD(0.915)**	–
G720D	**Core RAG1; CD-core**	**TMP_ESP_11_36597013**	**R**	**GGC, GAC**	**G/A**	–	–	**1DEL(0)**	**916PRD(0.915)**	–
E722K	Core RAG1; CD-core: ZFB	RAG1base_HSRAG1:g.2276G > A	R	GAG, AAG	G/A	–	–	–	–	–
I729L	Core RAG1; CD-core: ZFB	*rs201192233*	*M*	*ATT, CTT*	*A/C*	*C*	*0.001*	*71TOL(0.07)*	***986PRD(0.985)***	–
C730F	Core RAG1; CD-core: ZFB	*RAG1base_RAG1_DNA:g.8481G > T*	*K*	*TGT, TTT*	*G/T*	–	–	*1001TOL(1)*	***523POD(0.522)***	–
L732F	Core RAG1; CD-core: ZFB	RAG1base_RAG1_DNA:g.8486C > T	Y	CTT, TTT	C/T	–	–	–	–	–
A740S	**Core RAG1; CD-core: ZFB**	**COSM926739**	**K**	**GCC, TCC**	**G/T**	–	–	**1DEL(0)**	**878POD(0.877)**	–

Table 2 (*continued*)

Mutants	Localization	Variation ID	AC	Codons	Alleles	MA	GFreq	SIFT	PolyPhen	Disease association
F746C	Core RAG1; CD-core: ZFB	TMP_ESP_11_36597091	K	TTC, TGC	T/G	–	–	–	–	–
R759C	Core RAG1; CD-core: ZFB	COSM145840	Y	CGT, TGT	C/T	–	–	–	–	–
S771T	Core RAG1; C-terminal domain	TMP_ESP_11_36597165	W	TCT, ACT	T/A	–	–	–	–	–
E773G	**Core RAG1; C-terminal domain**	**TMP_ESP_11_36597172**	**R**	**GAA, GGA**	**A/G**	–	–	**11DEL(0.01)**	**805POD(0.804)**	–
R776W	**Core RAG1; C-terminal domain**	**rs121918572**	**Y**	**CGG, TGG**	**C/T**	–	–	**1DEL(0)**	**976PRD(0.975)**	–
R776L	**Core RAG1; C-terminal domain**	**COSM386382**	**K**	**CGG, CTG**	**G/T**	–	–	**1DEL(0)**	**976PRD(0.975)**	–
R776Q	Core RAG1; C-terminal domain	RAG1base_RAG1_DNA:g.8619G > A	R	CGG, CAG	G/A	–	–	–	–	–
R776W	**Core RAG1; C-terminal domain**	**COSM355977**	**Y**	**CGG, TGG**	**C/T**	–	–	**1DEL(0)**	**995PRD(0.994)**	–
R778G	Core RAG1; C-terminal domain	RAG1base_RAG1_DNA:g.8624C > G	S	CGG, GGG	G/C	–	–	–	–	–
R778Q	Core RAG1; C-terminal domain	rs121918569	R	CGG, CAG	G/A	–	–	–	–	–
R778W	Core RAG1; C-terminal domain	RAG1base_RAG1_DNA:g.8624C > T	Y	CGG, TGG	C/T	–	–	–	–	–
K780N	**Core RAG1; C-terminal domain**	**COSM1507889**	**W**	**AAA, AAT**	**A/T**	–	–	**1DEL(0)**	**523POD(0.522)**	–
V782D	**Core RAG1; C-terminal domain**	**rs200300629**	**W**	**GTC, GAC**	**T/A**	–	–	**21DEL(0.02)**	**963PRD(0.962)**	–
G802S	Core RAG1; C-terminal domain	rs151322536	R	GGC, AGC	G/A	–	–	–	–	–
I810V	**Core RAG1; C-terminal domain**	**rs61752933**	**R**	**ATC, GTC**	**A/G**	–	–	**1DEL(0)**	**950PRD(0.949)**	–
K820R	Core RAG1; C-terminal domain	rs2227973	R	AAG, AGG	A/G	G	0.246	–	–	–
N823T	**Core RAG1; C-terminal domain**	**rs147656090**	**M**	**AAT, ACT**	**A/C**	–	–	**1DEL(0)**	**965PRD(0.964)**	–
E827Q	**Core RAG1; C-terminal domain**	**COSM428955**	**S**	**GAG, CAG**	**G/C**	–	–	**1DEL(0)**	**965PRD(0.964)**	–
R841W	Core RAG1; C-terminal domain	RAG1base_HSRAG1:g.2633C > T	Y	CGG, TGG	C/T	–	–	–	–	–
K843E	**Core RAG1; C-terminal domain**	**rs186717025**	**R**	**AAG, GAG**	**A/G**	**G**	**0.001**	**11DEL(0.01)**	**932PRD(0.931)**	–
R851K	Core RAG1; C-terminal domain	*rs142345523*	*R*	*AGG, AAG*	*G/A*	–	–	***11DEL(0.01)***	*88BEN(0.087)*	–
N855I	**Core RAG1; C-terminal domain**	**rs199474690**	**W**	**AAC, ATC**	**A/T**	–	–	**1DEL(0)**	**871POD(0.87)**	–
N855I	**Core RAG1; C-terminal domain**	**RAG1base_HSRAG1:g.2676A > T**	**W**	**AAC, ATC**	**T/A**	–	–	**1DEL(0)**	**871POD(0.87)**	–
A857S	Core RAG1; C-terminal domain	TMP_ESP_11_36597423	K	GCC, TCC	G/T	–	–	–	–	–
V866M	**Core RAG1; C-terminal domain**	**COSM1127793**	**R**	**GTG, ATG**	**G/A**	–	–	**11DEL(0.01)**	**976PRD(0.975)**	–
D867N	Core RAG1; C-terminal domain	COSM428956	R	GAT, AAT	G/A	–	–	–	–	–
A868V	Core RAG1; C-terminal domain	*rs193922462*	*Y*	*GCA, GTA*	*C/T*	–	–	*231TOL(0.23)*	*1BEN(0)*	–
V869I	Core RAG1; C-terminal domain	rs201313833	R	GTT, ATT	G/A	–	–	–	–	–
E876K	Core RAG1; C-terminal domain	*rs145772007*	*R*	*GAG, AAG*	*G/A*	–	–	***11DEL(0.01)***	*9BEN(0.008)*	–
E876K	Core RAG1; C-terminal domain	COSM1470479	R	GAG, AAG	G/A	–	–	–	–	–
R878I	Core RAG1; C-terminal domain	COSM377837	–	AGG, ATC	GG/TC	–	–	–	–	–
E880K	Core RAG1; C-terminal domain	rs4151033	R	GAG, AAG	G/A	A	0.013	–	–	–
D887N	**Core RAG1; C-terminal domain**	**rs4151034**	**R**	**GAT, AAT**	**G/A**	**A**	**0.003**	**1DEL(0)**	**976PRD(0.975)**	–
W896R	Core RAG1; C-terminal domain	RAG1base_RAG1_DNA:g.8978T > C	Y	TGG, CGG	T/C	–	–	–	–	–
R897Q	Core RAG1; C-terminal domain	COSM276989	R	CGA, CAA	G/A	–	–	–	–	–
Y912F	**Core RAG1; C-terminal domain**	**COSM314666**	**W**	**TAC, TTC**	**A/T**	–	–	**1DEL(0)**	**927PRD(0.926)**	–
R918H	**Core RAG1; C-terminal domain**	**TMP_ESP_11_36597607**	**R**	**CGT, CAT**	**G/A**	–	–	**21DEL(0.02)**	**878POD(0.877)**	–
T925M	**Core RAG1; C-terminal domain**	**rs144893101**	**Y**	**ACG, ATG**	**C/T**	–	–	**1DEL(0)**	**976PRD(0.975)**	–
K926T	Core RAG1; C-terminal domain	COSM284453	M	AAG, ACG	A/C	–	–	–	–	–
Y931S	Core RAG1; C-terminal domain	rs75591129	M	TAT, TCT	A/C	–	–	–	–	–
E948D	Core RAG1; C-terminal domain	TMP_ESP_11_36597698	W	GAA, GAT	A/T	–	–	–	–	–
I956T	**Core RAG1; C-terminal domain**	**rs182385524**	**Y**	**ATT, ACT**	**T/C**	**C**	**0.001**	**1DEL(0)**	**950PRD(0.949)**	–
I956T	Core RAG1; C-terminal domain	RAG1base_RAG1_DNA:g.9159T > C	Y	ATT, ACT	T/C	–	–	–	–	–
E965V	**Core RAG1; C-terminal domain**	**COSM335997**	**W**	**GAG, GTG**	**A/T**	–	–	**1DEL(0)**	**977PRD(0.976)**	–
N968K	**Core RAG1; C-terminal domain**	**rs193922463**	**M**	**AAC, AAA**	**C/A**	–	–	**21DEL(0.02)**	**988PRD(0.987)**	–
K969T	**Core RAG1; C-terminal domain**	**COSM72375**	**M**	**AAA, ACA**	**A/C**	–	–	**11DEL(0.01)**	**523POD(0.522)**	–
R973C	Core RAG1; C-terminal domain	COSM926744	Y	CGC, TGC	C/T	–	–	–	–	–
R973H	Core RAG1; C-terminal domain	RAG1base_RAG1_DNA:g.9210G > A	R	CGC, CAC	G/A	–	–	121TOL(0.12)	1BEN(0)	–
R975W	Core RAG1; C-terminal domain	rs121918570	Y	CGG, TGG	C/T	–	–	–	–	–
R975W	Core RAG1; C-terminal domain	*RAG1base_RAG1_DNA:g.9215C > T*	*Y*	*CGG, TGG*	*C/T*	–	–	***11DEL(0.01)***	*5BEN(0.004)*	–
R975Q	Core RAG1; C-terminal domain	RAG1base_HSRAG1:g.3036G > A	R	CGG, CAG	G/A	–	–	–	–	–
Q981P	C-terminal end after Core RAG1	rs104894288	M	CAG, CCG	A/C	–	–	–	–	–
K992R	**C-terminal end after Core RAG1**	**RAG1base_RAG1_DNA:g.9267A > G**	**R**	**AAA, AGA**	**A/G**	–	–	**1DEL(0)**	**976PRD(0.975)**	–
K992E	**C-terminal end after Core RAG1**	**RAG1base_RAG1_DNA:g.9266A > G**	**R**	**AAA, GAA**	**A/G**	–	–	**1DEL(0)**	**976PRD(0.975)**	–
M1006V	C-terminal end after Core RAG1	rs139113046	R	ATG, GTG	A/G	G	0.001	–	–	–
H1009Y	C-terminal end after Core RAG1	*TMP_ESP_11_36597879*	*Y*	*CAT, TAT*	*C/T*	–	–	***1DEL(0)***	*9BEN(0.008)*	–
H1009R	C-terminal end after Core RAG1	rs200043512	R	CAT, CGT	A/G	–	–	–	–	–
P1021H	C-terminal end after Core RAG1	COSM1223038	M	CCT, CAT	C/A	–	–	–	–	–
S1024G	**C-terminal end after Core RAG1**	**COSM926745**	**R**	**AGC, GGC**	**A/G**	–	–	**21DEL(0.02)**	**719POD(0.718)**	–
G1030V	**C-terminal end after Core RAG1**	**TMP_ESP_11_36597943**	**K**	**GGC, GTC**	**G/T**	–	–	**11DEL(0.01)**	**977PRD(0.976)**	–
D1039E	**C-terminal end after Core RAG1**	**TMP_ESP_11_36597971**	**K**	**GAT, GAG**	**T/G**	–	–	**11DEL(0.01)**	**977PRD(0.976)**	–
M1041L	C-terminal end after Core RAG1	COSM322972	W	ATG, TTG	A/T	–	–	–	–	–
PK85Q	**N-terminal end before CND**	**RAG1base_RAG1_DNA:g.6546_6548delCTA**	–	**CCTAAG, CAG**	**CTA/–**	–	–	**21DEL(0.02)**	**932PRD(0.931)**	–

CHIDG – Combined cellular and humoral immune defects with granulomas; SCID – Severe combined immunodeficiency autosomal recessive T-cell-negative/B-cell-negative/NK-cell-positive; OS – Omenn syndrome; T-CMVA –– Alpha/beta T-cell lymphopenia, with gamma/delta T-cell expansion, severe cytomegalovirus infection and autoimmunity; DEL – deleterious; PRD – probably damaging; TOL – tolerated; BEN – benign; POD – possibly damaging; Gfre – Global frequency; MA – Minor allele; AC – Ambiguity code.

C-terminal end after Core RAG1: The C-terminal end after Core RAG1 (actual C-terminal end of RAG1) has 13 missense variants, which include 7 deleterious mutations, such as Q981P, K992R, K992E, S1024G, G1030V, D1039E and PK85Q.

About 50 missense mutations in RAG1 protein causes a spectrum of immunological dysregulation namely combined cellular and humoral immune defects with granulomas (CHIDG), severe combined immunodeficiency autosomal recessive T-cell-negative/

Table 3
Summary of synonymous RAG1 variants in the 1092 human genomes from 14 ethnicities.

Mutants.	Localization	Variation ID	Ambiguity code	Codons	Alleles	Minor allele	Global frequency
H20	N-terminal end before CND	rs138801620	Y	CAC, CAT	C/T	T	0.001
K49	N-terminal end before CND	COSM1298010	R	AAG, AAA	G/A	–	–
S62	N-terminal end before CND	rs201625282	K	TCT, TCG	T/G	–	–
P63	N-terminal end before CND	rs34357808	R	CCA, CCG	A/G	G	0.011
N94	CND	COSM926729	Y	AAC, AAT	C/T	–	–
N94	CND	rs148700564	Y	AAC, AAT	C/T	–	–
A101	CND	rs4151025	R	GCG, GCA	G/A	A	0.009
L107	CND	TMP_ESP_11_36595175	K	CTT, CTG	T/G	–	–
P152	CND	TMP_ESP_11_36595310	K	CCG, CCT	G/T	–	–
I161	CND	TMP_ESP_11_36595337	Y	ATC, ATT	C/T	–	–
S169	CND	COSM343989	S	TCG, TCC	G/C	–	–
S169	CND	COSM171937	R	TCG, TCA	G/A	–	–
P197	CND	TMP_ESP_11_36595445	R	CCG, CCA	G/A	–	–
N199	CND	rs148288583	Y	AAC, AAT	C/T	–	–
N199	CND	COSM1353715	Y	AAC, AAT	C/T	–	–
P206	CND	rs200182366	M	CCC, CCA	C/A	A	0.004
S210	CND	rs4151028	H	TCC, TCA	C/T/A	–	–
A217	CND	rs141561979	Y	GCC, GCT	C/T	–	–
R219	Region between CND and ZDD	COSM396370	R	CGG, CGA	G/A	–	–
L236	Region between CND and ZDD	COSM1263746	M	CTC, CTA	C/A	–	–
I266	Region between CND and ZDD	rs141654332	Y	ATC, ATT	C/T	T	0.001
A280	Region between CND and ZDD	TMP_ESP_11_36595694	R	GCA, GCG	A/G	–	–
P284	Region between CND and ZDD	TMP_ESP_11_36595706	R	CCA, CCG	A/G	–	–
C293	ZDD: within RING	rs188125314	Y	TGC, TGT	C/T	T	–
V356	ZDD; within ZFA	TMP_ESP_11_36595922	R	GTG, GTA	G/A	–	–
I376	ZDD; within ZFA	COSM926732	M	ATC, ATA	C/A	–	–
R396	Core RAG1; within NBR	TMP_ESP_11_36596042	S	CGC, CGG	C/G	–	–
R412	Core RAG1; within NBR	TMP_ESP_11_36596090	R	AGG, AGA	G/A	–	–
V433	Core RAG1; within NBR	rs145877904	S	GTG, GTC	G/C	–	–
I457	Core RAG1; within NBR	rs138387845	Y	ATC, ATT	C/T	–	–
N508	Core RAG1; region between NBR and CD-core	TMP_ESP_11_36596378	Y	AAT, AAC	T/C	–	–
E510	Core RAG1; region between NBR and CD-core	COSM186660	R	GAG, GAA	G/A	–	–
Q523	Core RAG1; region between NBR and CD-core	COSM269995	R	CAG, CAA	G/A	–	–
L526	Core RAG1; region between NBR and CD-core	rs143289774	R	CTG, CTA	G/A	A	0.001
T533	Core RAG1; CD-core	TMP_ESP_11_36596453	W	ACT, ACA	T/A	–	–
G543	Core RAG1; CD-core	rs17853743	R	GGA, GGG	A/G	–	–
T555	Core RAG1; CD-core	rs141265699	Y	ACC, ACT	C/T	–	–
L566	Core RAG1; CD-core	TMP_ESP_11_36596550	Y	TTG, CTG	T/C	–	–
I577	Core RAG1; CD-core	COSM1475405	Y	ATC, ATT	C/T	–	–
G606	Core RAG1; CD-core	COSM1353718	R	GGA, GGG	A/G	–	–
G606	Core RAG1; CD-core	COSM13951	W	GGA, GGT	A/T	–	–
D607	Core RAG1; CD-core	rs183806098	Y	GAC, GAT	C/T	T	0.001
G613	Core RAG1; CD-core	COSM394912	K	GGG, GGT	G/T	–	–
Q639	Core RAG1; CD-core	rs144475142	R	CAG, CAA	G/A	–	–
G696	Core RAG1; CD-core	rs144503943	S	GGC, GGG	C/G	–	–
R699	Core RAG1; CD-core	COSM363644	K	CGG, CGT	G/T	–	–
L721	Core RAG1; CD-core: ZFB	TMP_ESP_11_36597017	Y	CTC, CTT	C/T	–	–
V727	Core RAG1; CD-core: ZFB	rs148763154	S	GTC, GTG	C/G	–	–
E770	Core RAG1; C-terminal domain	TMP_ESP_11_36597164	R	GAG, GAA	G/A	–	–
S771	Core RAG1; C-terminal domain	rs142419805	Y	TCT, TCC	T/C	–	–
G781	Core RAG1; C-terminal domain	COSM1353722	S	GGG, GGC	G/C	–	–
E817	Core RAG1; C-terminal domain	rs61758791	R	GAA, GAG	A/G	–	–
K847	Core RAG1; C-terminal domain	COSM1507888	R	AAA, AAG	A/G	–	–
P848	Core RAG1; C-terminal domain	COSM369807	W	CCA, CCT	A/T	–	–
A857	Core RAG1; C-terminal domain	rs141560248	Y	GCC, GCT	C/T	T	0.001
S875	Core RAG1; C-terminal domain	rs200264827	S	TCC, TCG	C/G	G	0.001
C905	Core RAG1; C-terminal domain	rs138119069	Y	TGC, TGT	C/T	–	–
Q917	Core RAG1; C-terminal domain	rs150721661	R	CAG, CAA	G/A	A	0.005
T936	Core RAG1; C-terminal domain	rs199921936	M	ACC, ACA	C/A	–	–
A958	Core RAG1; C-terminal domain	rs138510915	R	GCA, GCG	A/G	–	–
A960	Core RAG1; C-terminal domain	rs1980131	R	GCA, GCG	A/G	G	0.070
R975	Core RAG1; C-terminal domain	COSM1507887	M	CGG, AGG	C/A	–	–
Y985	C-terminal end after Core RAG1	rs139084848	Y	TAT, TAC	T/C	C	0.001
K992	C-terminal end after Core RAG1	COSM1263745	R	AAA, AAG	A/G	–	–
Y997	C-terminal end after Core RAG1	TMP_ESP_11_36597845	Y	TAC, TAT	C/T	–	–
Y1001	C-terminal end after Core RAG1	TMP_ESP_11_36597857	Y	TAC, TAT	C/T	–	–
S1024	C-terminal end after Core RAG1	COSM1289550	Y	AGC, AGT	C/T	–	–

B-cell-negative/NK-cell-positive (SCID), omenn syndrome (OS) and alpha/beta T-cell lymphopenia, with gamma/delta T-cell expansion, severe cytomegalovirus infection and autoimmunity (T-CMVA) [28–31]. These 50 missense variants were picked by this study (Table 2), 80% of which are predicted to deleterious in nature. Majority of these variants are in the core RAG1 domain, except for R314W, which is in the RING finger region and Q981P in the C-terminal end (Figs. 2 and S1). These diseases have different

Table 4
Summary of frameshift variants of RAG1 in the 1092 human genomes from 14 ethnicities.

Mutants	Localization	Variation ID	Alleles
K86	N-terminal end before CND	RAG1base_HSRAG1:g.368_369delAA	AA/–
K86	N-terminal end before CND	RAG1base_RAG1_DNA:g.6548_6549delAA	AA/–
K86	N-terminal end before CND	TMP_ESP_11_36595109	AA/–
I113	CND	rs4151026	–/C
M173	CND	RAG1base_RAG1_DNA:g.6811delT	T/–
M173	CND	RAG1base_HSRAG1:g.631delT	T/–
L240	Region between CND and ZDD	rs146046446	T/–
A255	Region between CND and ZDD	RAG1base_RAG1_DNA:g.7057delA	A/–
S259	Region between CND and ZDD	RAG1base_HSRAG1:g.887delA	A/–
K354	ZDD; at the start of ZFA	RAG1base_HSRAG1:g.1173delT	T/–
D382	ZDD; within ZFA	RAG1base_HSRAG1:g.1258delA	A/–
C470	Core RAG1; within NBR	RAG1base_HSRAG1:g.1521_1522delGC	GC/–
I537	Core RAG1; CD-core	RAG1base_HSRAG1:g.1723_1735del	TATTGATGGGCTG/–
P649	Core RAG1; CD-core	RAG1base_RAG1_DNA:g.8239delT	T/–
H668	Core RAG1; CD-core	RAG1base_RAG1_DNA:g.8294delC	C/–
L690	Core RAG1; CD-core	RAG1base_HSRAG1:g.2182_2189delAATGCTTG	AATGCTTG/–
V715	Core RAG1; CD-core	TMP_ESP_11_36596996	G/–
T865	Core RAG1; C-terminal domain	TMP_ESP_11_36597448	TG/–
S875	Core RAG1; C-terminal domain	RAG1base_RAG1_DNA:g.8915delT	T/–

phenotypes [28–31] and some of these are involved in more than one disease (Table 2), which reflected that these variants are "hotspots" of RAG1 functionality.

Furthermore, 67 synonymous variants are distributed over entire RAG1 gene (Table 3). A total of 19 frameshift variants were also found (Table 4). There are 21 stop-gained variants and one variant each within categories inframe deletion and initiator codon variants (Table 5).

3.6. Characterization of non-coding variants

A total of 284 non-coding variants were identified (Fig. 4, Table S2), which included 89 intronic variants are localized in the single intron, 77 downstream coding variants, 73 upstream coding variants, and 43 UTR variants. The regulatory variant rSNPdb database [23] predicted 94% of non-coding variants of RAG1 as regulatory SNP rSNP (Table S2), added by different approaches as such all rSNP are RNA binding protein mediated regulation, while 40% are proximal transcriptional regulators, 20.4% SNPs are under linkage disequilibrium (LD, r2 > 0.8) and 11.3% SNPs are involved in distal transcriptional regulation (Table S2). However, only five SNPs were expression quantitative trait loci (eQTL/eQ), regulating mRNA expressions of two genes namely RAG1 and C11orf74/LOC119710 on the chromosome 11. These five SNPs were three upstream (rs12277745, rs3758873 and rs4150999), one 3′ UTR (rs4151047) and one intronic (rs4151001) variants.

4. Discussion

Herein, we described RAG1 from evolutionary, protein domains and genetic variant perspectives. Eukaryotic genomes are always influx of several types of gene rearrangements, such as inversions, translocations, duplications, and transpositions. Gene structural changes of RAG1 are clear representations of these influxes, which are not limited to genome compacted fishes as demonstrated in previous studies [20–22,32–34], reflecting several mechanisms are operational that can lead into intron insertions. First report of RAG1 intron was in *Takifugu* and it was amplified using degenerated PCR

Table 5
Summary of inframe deletion, initiator codon and stop gained variants of RAG1 in the 1092 human genomes from 14 ethnicities.

Mutants	Localization	Variation ID	Ambiguity code	Codons	Variation type	Alleles
A979-	C-terminal end after Core RAG1	RAG1base_RAG1_DNA:g.9227_9229delGCC	–	GCC, –	Inframe deletion	GCC/–
M1V	N-terminal end before CND	rs200575481	R	ATG, GTG	Initiator codon variant	A/G
S26*	N-terminal end before CND	COSM339567	M	TCA, TAA	Stop gained	C/A
R108*	CND	rs193922464	Y	CGA, TGA	Stop gained	C/T
R142*	CND	RAG1base_RAG1_DNA:g.6716C > T	Y	CGA, TGA	Stop gained	C/T
*Q248**	*CND*	*RAG1base_RAG1_DNA:g.7034C > T*	*Y*	*CAG, TAG*	*Stop gained*	*C/T*
*Y333**	*ZDD; region between RING and ZFA*	*RAG1base_RAG1_DNA:g.7291T > A*	*W*	*TAT, TAA*	*Stop gained*	*A/T*
Y589*	Core RAG1; CD-core	RAG1base_RAG1_DNA:g.8059C > G	S	TAC, TAG	Stop gained	G/C
Y589*	Core RAG1; CD-core	RAG1base_HSRAG1:g.1879C > G	S	TAC, TAG	Stop gained	G/C
E646*	Core RAG1; CD-core	COSM1353720	K	GAA, TAA	Stop gained	G/T
E665*	Core RAG1; CD-core	COSM186662	K	GAG, TAG	Stop gained	G/T
E774*	Core RAG1; C-terminal domain	rs104894282	K	GAA, TAA	Stop gained	G/T
E774*	Core RAG1; C-terminal domain	RAG1base_HSRAG1:g.2432G > T	K	GAA, TAA	Stop gained	G/T
RK829S*	Core RAG1; C-terminal domain	RAG1base_RAG1_DNA:g.8779_8780delinsTT	–	AGGAAA, AGTTAA	Stop gained	GA/TT
*L872**	*Core RAG1; C-terminal domain*	*RAG1base_HSRAG1:g.2727T > A*	*W*	*TTA, TAA*	*Stop gained*	*A/T*
*R897**	*Core RAG1; C-terminal domain*	*RAG1base_RAG1_DNA:g.8981C > T*	*Y*	*CGA, TGA*	*Stop gained*	*C/T*
*R897**	*Core RAG1; C-terminal domain*	*RAG1base_HSRAG1:g.2801C > T*	*Y*	*CGA, TGA*	*Stop gained*	*C/T*
C900*	Core RAG1; C-terminal domain	RAG1base_HSRAG1:g.2812C > A	M	TGC, TGA	Stop gained	C/A
Y938*	Core RAG1; C-terminal domain	rs104894283	K	TAT, TAG	Stop gained	T/G
Y938*	Core RAG1; C-terminal domain	RAG1base_HSRAG1:g.2926T > G	K	TAT, TAG	Stop gained	T/G
W959*	Core RAG1; C-terminal domain	RAG1base_HSRAG1:g.2988G > A	R	TGG, TAG	Stop gained	G/A
E965*	Core RAG1; C-terminal domain	COSM257852	K	GAG, TAG	Stop gained	G/T
E1042*	C-terminal end after Core RAG1	COSM212763	K	GAA, TAA	Stop gained	G/T

[35]. Intron insertions in RAG1 gene in vertebrate and invertebrates species lead into of different fates as previously described for the CCM3 gene [19].

We identified the ancestral loci of RAG1 gene using synteny analyses. This is also known from a recent study of sea urchin [36]. We made attempts to extract RAG1 gene like fragments from two known lamprey genomes, but lampreys don't have any hits. We corroborate that jawless fishes have most likely lost RAG genes. Several studies have been carried out about the origin of RAG genes, including in sharks [37] and by detection of only core central fragment in amphioxus [38]. Additionally, it is also postulated that fungal *Transib* transposon [39] are ancestors of RAG1 as they have conserved active sites of RAG1. This is agreeable explanation, however, the first evidence of loci conservation comes from sea urchin. Notably, these studies provide hints that acquisition of RAG genes is not as abrupt as it was previously assumed. Transposable elements are accountable for bringing various types of diversities and are responsible for diversities of adaptive immunity as mediated by *Transib* transposon [39]. Similarly, analogous system is recently reported in microbes as *CRISPR-Cas* systems [40]. There are advantages of transposon mediated diversities, as it can be easily wipe out when not required, probably lampreys use this system of wiping out.

Total 751 germline variants were identified in this study including 267 missense variants with 140 being deleterious mutations. This included about known 50 disease variants, which causes disruption of recombination activities of RAG1, leading into four known syndromes. The current study almost triples the missense mutational hotspots of RAG1 within the coding region. Additionally, we have identified 286 non-coding variants, which are also critical along with missense variants. As it is becoming clear that non-coding regions of human genome are also regulatory in the post-ENCODE era [41] and discovery of several types of non-coding RNA based gene expression regulators in human genome such as microRNA [42] and long non-coding RNA [43]. These variants setup a platform for further clinical studies related to diseases, which are caused by RAG1.

Conflict of interest

None.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.bbrc.2015.04.125.

Transparency document

Transparency document associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.bbrc.2015.04.125.

References

[1] M.A. Oettinger, D.G. Schatz, C. Gorka, D. Baltimore, RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination, Science 248 (1990) 1517–1523.
[2] D.G. Schatz, M.A. Oettinger, D. Baltimore, The V(D)J recombination activating gene, RAG-1, Cell 59 (1989) 1035–1048.
[3] S.F. Schluter, R.M. Bernstein, H. Bernstein, J.J. Marchalonis, 'Big Bang' emergence of the combinatorial immune system, Dev. Comp. Immunol. 23 (1999) 107–111.
[4] J.J. Marchalonis, S.F. Schluter, R.M. Bernstein, V.S. Hohman, Antibodies of sharks: revolution and evolution, Immunol. Rev. 166 (1998) 103–122.
[5] J.L. Arbuckle, N.S. Rahman, S. Zhao, W. Rodgers, K.K. Rodgers, Elucidating the domain architecture and functions of non-core RAG1: the capacity of a non-core zinc-binding domain to function in nuclear import and nucleic acid binding, BMC Biochem. 12 (2011) 23.
[6] P. Flicek, I. Ahmed, M.R. Amode, D. Barrell, K. Beal, et al., Ensembl 2013, Nucleic Acids Res. 41 (2013) D48–D55.
[7] M. Stanke, B. Morgenstern, AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints, Nucleic Acids Res. 33 (2005) W465–W467.
[8] A. Kumar, H. Ragg, Ancestry and evolution of a secretory pathway serpin, BMC Evol. Biol. 8 (2008) 250.
[9] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (2004) 1792–1797.
[10] K.B. Nicholas, H.B. Nicholas Jr., D.W.I. Deerfield, GeneDoc: analysis and visualization of genetic variation, EMBNEWNEWS 4 (1997) 14.
[11] D.T. Jones, W.R. Taylor, J.M. Thornton, The rapid generation of mutation data matrices from protein sequences, Comput. Appl. Biosci. 8 (1992) 275–282.
[12] K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, MEGA6: molecular evolutionary genetics analysis version 6.0, Mol. Biol. Evol. 30 (2013) 2725–2729.
[13] T.G. Wolfsberg, Using the NCBI map viewer to browse genomic sequence data, Curr. Protoc. Bioinforma. 5 (2010) 1–25 [Chapter 1, Unit 1].
[14] P. Flicek, B.L. Aken, B. Ballester, K. Beal, E. Bragin, et al., Ensembl's 10th year, Nucleic Acids Res. 38 (2010) D557–D562.
[15] T.J. Hubbard, B.L. Aken, S. Ayling, B. Ballester, K. Beal, et al., Ensembl 2009, Nucleic Acids Res. 37 (2009) D690–D697.
[16] G.R. Abecasis, A. Auton, L.D. Brooks, M.A. DePristo, R.M. Durbin, et al., An integrated map of genetic variation from 1,092 human genomes, Nature 491 (2012) 56–65.
[17] P.C. Ng, S. Henikoff, SIFT: predicting amino acid changes that affect protein function, Nucleic Acids Res. 31 (2003) 3812–3814.
[18] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, et al., A method and server for predicting damaging missense mutations, Nat. Methods 7 (2010) 248–249.
[19] A. Kumar, A. Bhandari, C. Goswami, Surveying genetic variants and molecular phylogeny of cerebral cavernous malformation gene, CCM3/PDCD10, Biochem. Biophys. Res. Commun. 455 (2014) 98–106.
[20] A. Kumar, A. Bhandari, S.J. Sarde, C. Goswami, Sequence, phylogenetic and variant analyses of antithrombin III, Biochem. Biophys. Res. Commun. 440 (2013) 714–724.
[21] A. Kumar, A. Bhandari, S.J. Sarde, C. Goswami, Genetic variants and evolutionary analyses of heparin cofactor II, Immunobiology 219 (2014) 713–728.
[22] A. Kumar, S.J. Sarde, A. Bhandari, Revising angiotensinogen from phylogenetic and genetic variants perspectives, Biochem. Biophys. Res. Commun. 446 (2014) 504–518.
[23] L. Guo, Y. Du, S. Chang, K. Zhang, J. Wang, rSNPBase: a database for curated regulatory SNPs, Nucleic Acids Res. 42 (2014) D1033–D1039.
[24] F.F. Yin, S. Bailey, C.A. Innis, M. Ciubotaru, S. Kamtekar, et al., Structure of the RAG1 nonamer binding domain with DNA reveals a dimer that mediates DNA synapsis, Nat. Struct. Mol. Biol. 16 (2009) 499–508.
[25] J.L. Arbuckle, L.A. Fauss, R. Simpson, L.M. Ptaszek, K.K. Rodgers, Identification of two topologically independent domains in RAG1 and their role in macromolecular interactions relevant to V(D)J recombination, J. Biol. Chem. 276 (2001) 37093–37101.
[26] D.R. Kim, Y. Dai, C.L. Mundy, W. Yang, M.A. Oettinger, Mutations of acidic residues in RAG1 define the active site of the V(D)J recombinase, Genes Dev. 13 (1999) 3070–3080.
[27] S.D. Fugmann, I.J. Villey, L.M. Ptaszek, D.G. Schatz, Identification of two catalytic residues in RAG1 that define a single active site within the RAG1/RAG2 protein complex, Mol. Cell 5 (2000) 97–107.
[28] J.P. de Villartay, A. Lim, H. Al-Mousa, S. Dupont, J. Dechanet-Merville, et al., A novel immunodeficiency associated with hypomorphic RAG1 mutations and CMV infection, J. Clin. Investig. 115 (2005) 3291–3299.
[29] A. Villa, C. Sobacchi, L.D. Notarangelo, F. Bozzi, M. Abinun, et al., V(D)J recombination defects in lymphocytes due to RAG mutations: severe immunodeficiency with a spectrum of clinical presentations, Blood 97 (2001) 81–88.
[30] O. Alsmadi, A. Al-Ghonaium, S. Al-Muhsen, R. Arnaout, H. Al-Dhekri, et al., Molecular analysis of T-B-NK+ severe combined immunodeficiency and Omenn syndrome cases in Saudi Arabia, BMC Med. Genet. 10 (2009) 116.
[31] C. Schuetz, K. Huck, S. Gudowius, M. Megahed, O. Feyen, et al., An immunodeficiency disease with RAG mutations and granulomas, N. Engl. J. Med. 358 (2008) 2030–2038.
[32] H. Ragg, A. Kumar, K. Köster, C. Bentele, Y. Wang, et al., Multiple gains of spliceosomal introns in a superfamily of vertebrate protease inhibitor genes, BMC Evol. Biol. 9 (2009) 208.
[33] A. Kumar, A. Bhandari, R. Sinha, P. Goyal, A. Grapputo, Spliceosomal intron insertions in genome compacted ray-finned fishes as evident from phylogeny of MC receptors, also supported by a few other GPCRs, PloS One 6 (2011) e22046.
[34] A. Kumar, A. Bhandari, S.J. Sarde, C. Goswami, Molecular phylogeny of C1 inhibitor depicts two immunoglobulin-like domains fusion in fishes and ray-finned fishes specific intron insertion after separation from zebrafish, Biochem. Biophys. Res. Commun. 450 (2014) 219–226.
[35] B. Venkatesh, Y. Ning, S. Brenner, Late changes in spliceosomal introns define clades in vertebrate evolution, Proc. Natl. Acad. Sci. U. S. A. 96 (1999) 10267–10271.

[36] S.D. Fugmann, C. Messier, L.A. Novack, R.A. Cameron, J.P. Rast, An ancient evolutionary origin of the Rag1/2 gene locus, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 3728–3733.
[37] G.W. Litman, J.P. Cannon, L.J. Dishaw, Reconstructing immune phylogeny: new perspectives, Nat. Rev. Immunol. 5 (2005) 866–879.
[38] Y. Zhang, K. Xu, A. Deng, X. Fu, A. Xu, et al., An amphioxus RAG1-like DNA fragment encodes a functional central domain of vertebrate core RAG1, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) 397–402.
[39] V.V. Kapitonov, J. Jurka, RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons, PLoS Biol. 3 (2005) e181.
[40] R. Barrangou, L.A. Marraffini, CRISPR-Cas systems: prokaryotes upgrade to adaptive immunity, Mol. Cell 54 (2014) 234–244.
[41] Consortium EP, An integrated encyclopedia of DNA elements in the human genome, Nature 489 (2012) 57–74.
[42] H.S. Soifer, J.J. Rossi, P. Saetrom, MicroRNAs in disease and potential therapeutic applications, Mol. Ther. 15 (2007) 2070–2079.
[43] J.L. Rinn, H.Y. Chang, Genome regulation by long noncoding RNAs, Annu. Rev. Biochem. 81 (2012) 145–166.