Data Article

Data on the evolutionary history of the V(D)J recombination-activating protein 1 – RAG1 coupled with sequence and variant analyses

Abhishek Kumar\textsuperscript{a,b}, Anita Bhandari\textsuperscript{c}, Sandeep J. Sarde\textsuperscript{a,d}, Sekhar Muppavarapu\textsuperscript{d}, Ravi Tandon\textsuperscript{e}

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

\textbf{ARTICLE INFO}

\textbf{Article history:}
Received 18 November 2015
Received in revised form 5 May 2016
Accepted 12 May 2016
Available online 20 May 2016

\textbf{ABSTRACT}

RAG1 protein is one of the key component of RAG complex regulating the V(D)J recombination. There are only few studies for RAG1 concerning evolutionary history, detailed sequence and mutational hotspots. Herein, we present out datasets used for the recent comprehensive study of RAG1 based on sequence, phylogenetic and genetic variant analyses (Kumar et al., 2015)\textsuperscript{[1]}. Protein sequence alignment helped in characterizing the conserved domains and regions of RAG1. It also aided in unraveling ancestral RAG1 in the sea urchin. Human genetic variant analyses revealed 751 mutational hotspots, located both in the coding and the non-coding regions. For further analysis and discussion, see (Kumar et al., 2015)\textsuperscript{[1]}. © 2016 Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

\textbf{Specifications Table}

| Subject area                     | Biology                  |
|----------------------------------|--------------------------|
| More specific subject area       | Molecular evolution and bioinformatics |
| Type of data                     | Tables, figures          |
How data was acquired

Retrieved from public databases

Data format

Analyzed data

Experimental factors

RAG1 sequences were retrieved from ENSEMBL and/or NCBI database.

Experimental features

RAG1 protein alignment using Muscle tool and edited in the GeneDoc

RAG1 Variants were analyzed with SIFT, Polyphen & rSNPbase

Data source location

Germany

Data accessibility

Data is with this article

Value of the data

- Protein sequence analysis data reveal that SpRAG1L possesses only 19–20% identities with vertebrate RAG1, which helped us in deriving an ancestral RAG1 protein in sea urchin. This approach can be used to detect origins for different proteins.

- Protein sequence alignment locates two major domains and several regions of RAG1, which suggested that these fragments were conserved from sea urchin to human. This hints evolutionary conservation of protein domains in the protein of interest and their ancestors.

- Data on the genetic variant analysis suggests that human RAG1 gene has 751 variants.

- Furthermore, there are 267 missense variants of human RAG1 causing change in amino acids including 140 deleterious mutations. These variant data serve as the mutational hotspots within the coding region of human RAG1. Assessment of mutational hotspots for any protein is critically important for understanding its function and roles in diseases.

- Additionally, 284 non-coding variants were identified with 94% regulatory in nature, which are often called as regulatory SNP (rSNP). These data are source of regulatory implications flanking any given gene.

1. Data

Table 1 lists all RAG1 sequences used in Kumar et al. [1] and these sequences are used for constructing protein sequence alignment of RAG1 (Fig. S1). This protein alignment is the basis for the Figs. 2 and 3 and Table 1–5 of Kumar et al. [1]. Details of human RAG1 variants are summarized in the Table S1 and regulatory SNPs in the Table S2. These two supplementary tables are primary data for variant analyses described in Fig. 4 and Tables 2–5 of Kumar et al. [1].

2. Experimental design, materials and methods

Using the BLAST homology detection tool [2], we extracted RAG1 gene from vertebrate genomes listed either in Ensembl release 77 [3] or NCBI. To ensure accuracy of gene structures, we combined the gene predictions of the Ensembl [3] and AUGUSTUS tool [4]. We used human RAG1 as the standard sequence for intron position mapping and numbering of intron positions, followed by suffixes a–c for their location as reported previously [5]. We aligned selected RAG1 protein sequences using MUSCLE tool [6] with and we manually adjusted alignment with GENEDOC tool [7]. We reconstructed a phylogenetic tree with maximum likelihood method, based on the JTT matrix-based model [8] with 1000 bootstrap replicates. We imported all consensus trees to MEGA 6 software [9], 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 two genome browsers namely, NCBI map viewer [10] and ENSEMBL genome browser [11,12]. Furthermore, we generated human RAG1 variants from 1092 human genomes from 14 different populations available in 1000 genomes project [13]. We analyzed the impact assessments of missense variants on the human RAG1 protein using SIFT [14] and PolyPhen V2 [15] tools, as described previously [16–19]. We detected regulatory nature of
| Name      | Organism     | Species                | Accession id          | Chromosomal localization       |
|-----------|--------------|------------------------|-----------------------|---------------------------------|
| HsapRAG1  | Human        | Homo sapiens           | ENSG000000166349      | Chromosome 11: 36,332,259-36,614,706 |
| MmusRAG1  | Mouse        | Mus musculus           | ENSMUSG000000061311   | Chromosome 2: 101,638,282-101,649,501 |
| RnorRAG1  | Rat          | Rattus norvegicus      | ENSRNOG000000064630   | Chromosome 3: 97,866,048-97,877,145 |
| TgutRAG1  | Zebrafinch   | Taeniopygia guttata    | ENSTJUG00000010147    | Chromosome 5: 17,596,747-17,599,869 |
| MgalRAG1  | Turkey       | Meleagris gallopavo    | ENSMAG000000015794    | Chromosome 5: 19,778,620-19,781,748 |
| PsnRAG1   | Turtle       | Pelodiscus sinensis    | ENSPSG00000001811     | Scaffold JH209124.1: 1,890,899-1,894,018 |
| DrefRAG1  | Zebrafish    | Danio rerio            | ENSDARG000000062122   | Chromosome 25: 9,231,637-9,238,142 |
| TrubRAG1  | Fugu         | Takifugu rubripes      | ENSTRU00000001340     | scaffold: 302: 189,544-193,510 |
| TnigRAG1  | Tetraodon    | Tetraodon nigroviridis | ENSTNG00000012168     | Chromosome 13: 5,598,243-5,602,176 |
| OniRAG1   | Tilapia      | Oreochromis niloticus  | ENSONIG000000014593   | Scaffold GL831142.1: 1,924,501-1,931,477 |
| GmorRAG1a | Cod          | Gadus morhua           | ENSMOC00000003395     | Gene: 2196: 249,630-253,939 |
| XmacRAG1  | Platyfish    | Xiphophorus maculatus  | ENSMXAG00000000820    | Scaffold JH556735.1: 897,221-901,222 |
| GacuRAG1  | Stickelback  | Gasterosteus aculeatus | ENSCAGC000000011665   | groupXIX: 14,493,756-14,497,767 |
| OlatRAG1  | Medaka       | Oryzias latipes        | ENSORTL000000011969   | Chromosome 6: 17,343,305-17,347,405 |
| LchaRAG1  | Coelancanth  | Latimeria chalumnae    | ENSLACG00000004406    | Scaffold JH126568.1: 121,273-124,451 |
| VpaRAG1   | Alpaca       | Vicugna pacos          | ENSVPAG00000008826    | Gene: 2429: 269,595-273,365 |
| AcarR1    | Anole lizard | Anolis carolinensis    | ENSAACG00000005096    | Chromosome 1: 53,518,235-53,521,375 |
| DnoRAG1   | Armadillo    | Dasypus novemcinctus   | ENSDNOG00000006294    | Scaffold JH582431.1: 4,276,543-4,279,674 |
| OgarRAG1  | Bushbaby     | Otolemur garnettii     | ENSOAG000000027393    | Scaffold GL873520.1: 63,167,240-63,168,052 |
| FcatRAG1  | Cat          | Felis catus            | ENSFACG00000002908    | Chromosome D1: 92,125,946-92,129,077 |
| AmeRAG1   | Cave fish    | Astyanax mexicanus     | ENSAMXGM00000007587   | Scaffold KB871579.1: 5,211,103-5,217,994 |
| PtroRAG1  | Chimpanzee   | Pan troglodytes        | ENSPTRG00000003512    | Chromosome 11: 36,559,562-36,571,320 |
| BtauRAG1  | Cow          | Bos taurus             | ENSBTAG000000040293   | Chromosome 15: 67,827,233-67,830,364 |
| ClamRAG1  | Dog          | Canis lupus familiaris | ENSFCAG000000006807   | Chromosome 18: 31,631,533-31,634,664 |
| TtruRAG1  | Dolphin      | Tursiops truncatus     | ENSTRTRG00000014075   | scaffold: 11071: 196,070-199,540 |
| ApalRAG1  | Duck         | Anas platyrhynchos     | ENSAPLG00000011756    | Scaffold KB742537.1: 887,774-890,899 |
| LaiRAG1   | Elephant     | Loxodonta africana     | ENSLAFG00000023173    | SuperContig: scaffold: 21: 12,902,400-12,905,531 |
| MftrRAG1  | Ferret       | Mustela putorius furo  | ENSMFPUC000001963     | Scaffold GL896949.1: 10,184,818-10,187,949 |
| FalbRAG1  | Flycatcher   | Ficedula albicollis    | ENSFALG00000014372    | Scaffold JH603235.1: 4,349,497-4,397,619 |
| NleuRAG1  | Gibbon       | Nomascus leucogenys    | ENSNELAG00000017951   | SuperContig: GL392764.1: 51,275,048-51,286,754 |
| GgorRAG1  | Gorilla      | Gorilla gorilla        | ENSGG00000001312      | Chromosome 11: 37,187,229-37,198,984 |
| CporRAG1  | Guinea Pig   | Cavia porcellus        | ENSCP00000004516      | scaffold: 92: 2,485,274-2,488,405 |
| EcabRAG1  | Horse        | Equus caballus         | ENSECA00000021936     | Chromosome 12: 3,025,356-3,033,251 |
| PcapRAG1  | Hyrax        | Procavia capensis      | ENSPCAP00000001732    | Chromosome 12: 3,025,356-3,033,251 |
| MmulRAG1  | Macaque      | Macaca mulatta         | ENSMUMG00000018267    | Chromosome 12: 99,857,897-99,869,953 |
| CjcaRAG1  | Marmoset     | Callichthix jacchus    | ENSCJAG00000011082    | Scaffold GL430055: 356,167-359,258 |
| MlucRAG1  | Microbat     | Myotis lucifugus       | ENSMULG0000000544     | Gene: 3288: 841,983-845,201 |
| MmurtRAG1 | Mouse Lemur  | Microctes murinus      | ENSMIMG00000008611    | Chromosome 5: 272,756,599-272,759,742 |
| MdomRAG1  | Oppossum     | Monodelphis domestica  | ENSMODG000000024470    | Chromosome 5: 272,756,599-272,759,742 |
| Name   | Organism          | Species                          | Accession id                  | Chromosomal localization       |
|--------|-------------------|----------------------------------|-------------------------------|--------------------------------|
| AmelRAG1 | Panda             | Ailuropoda melanoleuca          | ENSAMEG000000019378           | Scaffold GL193442.1: 461,741-464,872 |
| SscrRAG1 | Pig               | Sus scrofa                       | ENSSSCG00000026145            | Chromosome 2: 26,730,010-26,738,216 |
| OanaRAG1 | Platypus          | Ornithorhynchus anatinus         | ENSOANG000000011770           | Chromosome 3: 11,364,602-11,365,783 |
| OcinRAG1 | Rabbit            | Oryctolagus cuniculus            | ENSOCUG000000006989           | Chromosome 1: 175,828,096-175,831,224 |
| OarRAG1 | Sheep             | Ovis aries                       | ENSOARG000000010441           | Chromosome 15: 65,210,839-65,213,970 |
| SaraRAG1 | Shrew             | Sorex araneus                    | ENSSARG000000010950           | GeneScaffold_5915: 66,723-69,956   |
| LocRAG1 | Spotted gar       | Lepisosteus oculatus             | ENSLOCG00000001283           | Chromosome LG27: 1,403,519-1,420,074 |
| ItriRAG1 | Squirrel          | Ictidomys tridecemlineatus      | ENSSTOG000000025584           | Scaffold JH393343.1: 1,817,576-1,820,707 |
| TsyRAG1 | Tarsier           | Tarsius syrichta                 | ENSTSYG00000007158           | scaffold_7240: 21,771-24,902      |
| SharRAG1 | Tasmanian devil  | Sarcophilus harrisii             | ENSSHAG000000014085           | Scaffold GL864890.1: 1,391,366-1,394,509 |
| TbeRAG1 | Tree Shrew        | Tupaia belangeri                | ENSTBE00000003010             | GeneScaffold_4067: 865-4,810       |
| MeuRAG1 | Wallaby           | Macropus eugenii                | ENSEMUG00000003165           | Scaffold77145: 3,005-5,812        |
| XtroRAG1 | Xenopus           | Xenopus tropicalis               | ENSXETP00000016443/XP_002937338 | Scaffold GL172917.1: 903,952-910,208 |
| SpuRAG1L | Sea urchin        | Strongylocentrotus purpuratus    | AAZ23546.1*                  | NA                              |

* From NCBI.
non-coding variants using the rSNPbase (this database provides reliable, and comprehensive regulatory annotations [20] and such variants are called regulatory SNP or rSNP).

Acknowledgments

We thank Faculty Recharge Programme of the University Grant Commission (UGC-FRP) and UPE II (Pl#15) for the financial support to RT.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2016.05.021.

References

[1] A. Kumar, A. Bhandari, S.J. Sarde, S. Muppavarapu, R. Tandon, Understanding V(D)J recombination initiator RAG1 gene using molecular phylogenetic and genetic variant analyses and upgrading missense and non-coding variants of clinical importance, Biochem. Biophys. Biophys. Res. Commun. 462 (2015) 301–313.

[2] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[3] P. Flicek, I. Ahmed, M.R. Amode, D. Barrett, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, C. Garcia-Giron, L. Gordon, T. Hourlier, S. Hunt, T. Juettemann, A.K. Kahari, S. Keenan, M. Komorowska, E. Kulesha, I. Longden, T. Maurel, W.M. McLaren, M. Moffato, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H.S. Riat, G.R. S. Ritchie, M. Ruffler, M. Schuster, D. Sheppard, D. Sobral, K. Taylor, A. Thomann, S. Trevanion, S. White, S.P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, J. Harrow, J. Herrero, T.J.P. Hubbard, N. Johnson, R. Kinsella, A. Parker, G. Spudich, A. Yates, A. Zadissa, S.M.J. Searle, Ensembl 2013, Nucleic Acids Res. 41 (2013) D48–D55.

[4] 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.

[5] A. Kumar, H. Ragg, Ancestry and evolution of a secretory pathway serpin, BMC Evol. Biol. 8 (2008) 250.

[6] R.C. Edgar, Muscle: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (2004) 1792–1797.

[7] I. Longden, T. Maurel, W.M. McLaren, M. Muffato, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H.S. Riat, G.R. S. Ritchie, M. Ruffler, M. Schuster, D. Sheppard, D. Sobral, K. Taylor, A. Thomann, S. Trevanion, S. White, S.P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, J. Harrow, J. Herrero, T.J.P. Hubbard, N. Johnson, R. Kinsella, A. Parker, G. Spudich, A. Yates, A. Zadissa, S.M.J. Searle, Ensembl 2013, Nucleic Acids Res. 41 (2013) D48–D55.
[18] A. Kumar, A. Bhandari, S.J. Sarde, C. Goswami, Genetic variants and evolutionary analyses of heparin cofactor II, Immunobiology 219 (2014) 713–728.
[19] A. Kumar, S.J. Sarde, A. Bhandari, Revising angiotensinogen from phylogenetic and genetic variants perspectives, Biochem. Biophys. Res. Commun. 446 (2014) 504–518.
[20] L. Guo, Y. Du, S. Chang, K. Zhang, J. Wang, rSNPBase: a database for curated regulatory SNPs, Nucleic Acids Res. 42 (2014) D1033–D1039.