An ancient evolutionary origin of the Rag1/2 gene locus

Sebastian D. Fugmann*, Cynthia Messier†, Laura A. Novack*, R. Andrew Cameron‡, and Jonathan P. Rast*§

*Laboratory of Cellular and Molecular Biology, National Institute on Aging, Baltimore, MD 21224; †Sunnybrook and Women’s Research Institute and Department of Medical Biophysics, University of Toronto, 2075 Bayview Avenue, Room S-1268, Toronto, ON, Canada M4N 3M5; and ‡Division of Biology, 156-29 California Institute of Technology, Pasadena, CA 91125

Edited by Masatoshi Nei, Pennsylvania State University, University Park, PA, and approved January 10, 2006 (received for review November 8, 2005)

The diversity of antigen receptors in the adaptive immune system of jawed vertebrates is generated by a unique process of somatic gene rearrangement known as V(D)J recombination. The Rag1 and Rag2 proteins are the key mediators of this process. They are encoded by a compact gene cluster that has exclusively been identified in animal species displaying V(D)J-mediated immunity, and no homologous gene pair has been identified in other organisms. This distinctly restricted phylogenetic distribution has led to the hypothesis that one or both of the Rag genes were coopted after horizontal gene transfer and assembled into a Rag1/2 gene cluster in a common jawed vertebrate. The hypothesis is further supported by (i) the highly conserved compact structure of the Rag gene locus, (ii) similarities of Rag1 to transposases and integrases, and (iii) the ability of recombinant Rag1/2 proteins to catalyze transposition reactions in vitro (4–6). In the extreme extension of this scenario, the prototypes of Rag1 and Rag2 encoded the transposase component of a mobile DNA element that jumped into the genome of a common ancestor of the modern jawed vertebrates. Then, as an intact transposon or through a nonautonomous element under its control, the proto-Rag genes reversedly disrupted a primordial antigen receptor gene locus (4, 7). Other variants of this hypothesis suggest that the Rag1 core region was derived from a transposable element early in animal evolution and that the Rag1/2 cluster may have been assembled much later in a jawed vertebrate ancestor (6). The evolutionary shift that presumably accompanied this event correlates with the phylogenetically inferred rapid appearance of the entire complex of Ig/TCR/MHC-mediated adaptive immunity (3, 8, 9). The suggested sudden emergence of this system has been referred to as an immunological “big bang” (10). This hypothesis is supported in the form of negative evidence by the inability to identify direct homologs of Rag1 and Rag2, and by the lack of functionally similar homologs of genes encoding TCR, Ig, and MHC Class I/II molecules from animals outside of the jawed vertebrates, including two completely sequenced urochordate genomes (11). Alternatively, the apparent phylogenetic discontinuity in adaptive immunity genes could be a consequence of gene loss and undersampling, and a longer and more gradual evolutionary process may underlie the emergence of the key elements of the vertebrate adaptive immune system (12).

Sea urchins are echinoderms, a sister group of the chordates. The Rag1/2 gene cluster is predicted to be missing in this phylum by evolutionary scenarios in which the locus was assembled as a consequence of horizontal gene transfer close to the time of the emergence of jawed vertebrate adaptive immunity. This study now reveals the unexpected presence of a Rag1/2-like cluster in the purple sea urchin with similarities on the levels of genomic structure and organization, regulation of expression, and properties of the encoded proteins.

Results

Sprag1L, a Rag1-Like Gene in the Purple Sea Urchin. A search of the Strongylocentrotus purpuratus trace sequence and assembly (11/23/04 release) made available by the Baylor College of Medicine Human Genome Sequencing Center (www.hgsc.bcm.tmc.edu/projects/seaurchin/) revealed sequences with similarity to the core region of Rag1. An independent computational analysis of these sequence fragments that relates them and Rag1 to the Transib transposase family was recently reported (6). The ORFs in nearly all of these elements (26 in 10 contigs) are fragmentary and disrupted by premature stop codons, and they are likely to be pseudogenes. Although it is presently unclear how these fragments originated, they do not seem to be overt transposon remnants, in that no terminal inverted repeats are evident and the regions with similarity to Rag1 tend to be very incomplete. One sequence element (contained in Supertig17631), however, maintains a large ORF that exhibits extensive similarity to the core region (amino acids 384–1,008) of mouse Rag1. This

© 2006 by The National Academy of Sciences of the USA
element, which we designate SpRag1L (for *S. purpuratus* Rag1-like), was chosen for more detailed analysis.

Expression analyses (described below) indicated that the SpRag1L gene is transcribed and spliced during embryogenesis and in a variety of adult tissues, including coelomocytes, blood-like cells that mediate immunity and wound healing (13). To determine the full-length transcript sequence, we used a PCR-based strategy. The sequence of the major cDNA transcript for the SpRag1L gene (GenBank accession no. DQ082723) consists of 3,363 nucleotides encoding a 983-aa polypeptide (see Fig. 1 and Figs. 6 and 7, which are published as supporting information on the PNAS web site; cDNA sequences from different animals varied in length within a repetitive region described below, but were otherwise nearly identical to the genome sequence). The similarity of SpRag1L to the core region of mouse *Rag1* is significant (31% amino acid identity), in particular with respect to residues that are important for nuclear localization in mouse *Rag1*. A conserved non-core putative zinc binding motif is indicated by gray shading. Numbers indicate amino acid position. Designations relative to the end of each sequence.

SpRag2L, a Rag2-Like Gene in the Purple Sea Urchin Closely Linked to SpRag1L. The presence of a rag2-like gene immediately raised the question whether a rag2-like gene is also present in the sea urchin genome. Initial BLAST searches of the sea urchin whole genome shotgun traces and genome assembly were not successful. Because Rag2 is known to be somewhat more divergent than Rag1 within vertebrates (20), we focused on the sequence downstream of SpRag1L to look for genes with low but distinct identity to rag2. Computational analysis revealed a predicted terminal exon in reverse orientation and 3,181 bp downstream of the SpRag1L termination codon (within the range of intergenic distances for vertebrate Rag genes). We determined the full cDNA sequence of this gene using a RACE strategy on gastrula and coelomocyte RNA. The gene encodes a 511-aa protein (GenBank accession no. DQ082724; Fig. 8, which is published as supporting information on the PNAS web site) whose predicted structure resembles Rag2. The first 424 aa are predicted to encode a six-bladed β-propeller (Fig. 2a). Like Rag2, this region matches the β-propeller of the galactose oxidase central domain SCOP profile (E-value 6.0 × 10⁻⁵) with each blade formed by a kelch repeat (PFAM match E-value = 1.4 × 10⁻²). As for vertebrate Rag2, a C-terminal plant homeodomain (PHD) domain was identified (amino acids 433–481; PFAM match E-value 3.0 × 10⁻⁹; Fig. 2b) (21, 22). A complete screen of all genes in the sea urchin genome (20,986 NCBI GNONOM gene models) using HMMER shows that, of 73 potential kelch-repeat sequences in BLASTP searches of all vertebrate proteins by using the non-PHD region (E-value 10⁻¹⁰), 13, 71 unique predicted protein structures, and 490 proteins with a putative PHD domain, only a single gene, namely the gene downstream of SpRag1L, contains both domains. Rag2 is the only gene in the human genome with this predicted domain structure (with a total of 71 predicted kelch-repeat/β-propeller-containing proteins and 490 proteins with a putative PHD domain, the gene downstream of SpRag1L, contains both domains). As for vertebrate Rag2, the interaction with Rag2, can also be readily identified. Although amino acid identity in the Rag1 nonamer-binding domain (NBD; amino acids 389–441) is lower (25%; Fig. 1, yellow shading), most of the basic residues implicated in DNA binding in this region are nonetheless conserved. Notably, sequence similarity extends also into the non-core region of Rag1. A 108-aa stretch shows 30% identity to a putative zinc-binding domain (ZBD) in mouse Rag1 (amino acid positions 101–223, Fig. 1) (18). The two stretches of sequence similarity with Rag1 are separated by a repetitive coding region (“unalignable region” in Fig. 1) containing 11 repeats of an 8-aa peptide (TAPLTPTA) corresponding to the position of the RING-finger domain present in all known vertebrate Rag1 proteins (19).
flanking genes. Unlike the Rag1 genes of most vertebrate classes and all Rag2 genes, the coding regions of both sea urchin genes are encoded in multiple exons. The genomic sequence of SpRag1L exon 3 contains a repetitive region with multiple variants of a 24-bp coding sequence. To address discrepancies between genomic PCR distance measurements and the genome sequence, we determined the sequence of this exon directly from the BAC clones, revealing a correctly phased ORF through this exon interspersed with 57 variants of the repeated 8 aa motif (see Fig. 10, which is published as supporting information on the PNAS web site).

The map that results from this BAC-based analysis is shown in Fig. 3a. Notably, rag1 of bony fish contains an intron in nearly the same position as intron 3 of SpRag1L (within 5 aa in a region of unsure sequence alignment), separating the nonamer-binding domain-like domain from the rest of the core region. The lack of an intron at this position in the Rag1 gene of a cartilaginous fish (20) and a phase difference at the splice site may suggest that the intron was acquired independently in the bony fishes and sea urchin genes. Nevertheless, each scenario requires two presumably rare events, either of intron loss or gain. The presence of multiple introns in the SpRagL-coding regions suggests that they represent functioning genes, as opposed to the remnants of a recently integrated mobile element.

**Coexpression of SpRag1L/2L Transcripts.** What is the function of the SpRag1L/2L gene pair? The SpRag1L/2L proteins may facilitate somatic rearrangement of yet unidentified genes in the sea urchin genome. Alternatively, these genes may perform a more basic function, such as excising mobile DNA elements from the genome; V(D)J recombination would then represent a highly specialized version of this function that was acquired later. As an initial step to gain insight into this question, we measured their transcription at different stages of embryonic development and in different adult tissues by real-time quantitative PCR (QPCR) using primers that are specific to the SpRag1L/2L genes and that amplify regions spanning introns. Expression of message from both genes is readily quantifiable in blastula and gastrula stage embryos. Lower levels of spliced message expression from both genes were evident in tissues including, but not limited to, coelomocytes (blood-like cells that are the source of the cDNA sequences described here). We used embryonic stages, in which the distribution of message prevalence is well characterized, to assess the relative expression levels of the SpRag1L/2L genes (Fig. 3b). Analysis of coelomocyte mRNA from a series of different individuals showed a similar degree of correlation between rag1 and rag2 expression levels (see Fig. 11, which is published as supporting information on the PNAS web site). Thus, like their vertebrate homologs, SpRag1L and SpRag2L are coexpressed, and expression levels tend to be correlated.
analysed by Western blotting by using a polyclonal anti-GST antiserum. Streptactin Sepharose beads were incubated with the crude lysates and was repeatedly lower compared with that of Rag2 from mouse or sea urchin. anti-strep antisera, respectively. Note that the expression level of shark Rag2 serve as negative controls. The amounts of the Rag and Rag-like proteins in the washed extensively with lysis buffer. The bound protein complexes were nated by a dash in expression vectors (expressing the GST and strep tag by themselves, desig- S indicated combinations of strep-tagged Rag1 and GST-tagged Rag2 from mouse (Mm, M. musculus), shark (C. leucas; Cp, C. plumbeus), and sea urchin (Sp, S. purpuratus) were coexpressed in 293T cells by transient transfection. Empty expression vectors (expressing the GST and strep tag by themselves, designated by a dash in a and b, respectively) and the PHD domain of mouse Rag2 serve as negative controls. The amounts of the Rag and Rag-like proteins in the crude cell lysates were determined by Western blot analysis with anti-GST and anti-strep antisera, respectively. Note that the expression level of shark Rag2 was repeatedly lower compared with that of Rag2 from mouse or sea urchin. (a) Glutathione Sepharose beads were incubated with the crude lysates and washed extensively with lysis buffer. The bound protein complexes were analyzed by Western blotting by using a polyclonal anti-strep antisemur. (b) Streptactin Sepharose beads were incubated with the crude lysates and washed extensively with lysis buffer. The bound protein complexes were analyzed by Western blotting by using a polyclonal anti-GST antisemur.

SpRag1L/2L Complex Formation. The physiological function of vertebrate Rag1 and Rag2 is to bind to recombination signal sequences (RSSs) that flank each Ig and TCR gene segment, and to cut DNA adjacent to these elements. This process requires a well orchestrated set of interactions between Rag1, Rag2, and a pair of RSSs. Because there is currently no evidence for V(D)J-type gene rearrangement in the sea urchin, the SpRag1L/2L gene products may function very differently compared with Rag1/2 of jawed vertebrates. The first step in V(D)J recombination is the formation of a stable Rag1–Rag2 complex. To test whether SpRag1L and SpRag2L share this property, we coexpressed them as strep- and GST-fusion proteins in 293T cells and carried out pull-down experiments with the respective cell lysates (Fig. 4). The protein amounts of SpRag1L and SpRag2L varied somewhat between individual transfections, as can be seen from the Western blot analysis of the crude lysates (Fig. 4 a and b). The analyses of the pull-down assay showed that SpRag1L interacts specifically with SpRag2L (Fig. 4 a, lane 5), whereas neither the GST tag alone (lane 1), nor the C terminus of mouse Rag2 (amino acids 384–527; lane 2), nor full-length mouse Rag2 (lane 3) were able to coprecipitate any SpRag1L protein. Interestingly Rag2 from sandbar shark (Carcharhinus plumbeus, Cp) also interacts with SpRag1L (Fig. 4 a, lane 4), suggesting that at least parts of the Rag1–Rag2 interaction surfaces are well conserved. We then coexpressed SpRag2L with SpRag1L as well as Rag1 from various vertebrate species to perform pull-down experiments with streptactin Sepharose, an affinity resin for the strep-tag (Fig. 4 b). Although all proteins were present in the crude lysate at comparable levels, the Western blot analysis of the complexes bound to the streptactin resin revealed that SpRag2L interacts with SpRag1L (Fig. 4 b, lane 4) and Rag1 from bull shark (Carcharhinus leucas, Cl; lane 3). In contrast, neither the strep tag alone, nor mouse Rag1 was able to bind any SpRag2L (Fig. 4 b, lanes 1 and 2). In summary, the pull-down assays indicate that SpRag1L and SpRag2L interact to form a heterodimeric complex, and furthermore, their ability to also interact with shark Rag1/2 provides initial evidence that this complex may resemble the vertebrate Rag1/Rag2 complex.

The minimal core domain of vertebrate Rag1 (mouse amino acids 384–1,008) used in the pull-down assay is required for all catalytic activity and binds specifically to RSS in dsDNA. A subdomain thereof, the central domain (cd, mouse amino acids 528–760) was recently reported to bind to ssDNA, which may resemble an unwound state of the DNA that is a transient intermediate in the cleavage reaction (24). The respective central domain of sea urchin Rag1L when purified as a recombinant protein from Escherichia coli shows similar DNA-binding properties (Fig. 12, which is published as supporting information on the PNAS web site). This observation suggests that the SpRag1L protein may use DNA as its substrate, the sequence of which may overlap with vertebrate RSSs similar to the case of the terminal repeats of the Rag1-related Transib transposase family (6). The identification of the cognate target motif of SpRag1L and the SpRag1L/2L complex will be important for future studies.

**Discussion**

In this study, we identify and characterize a gene pair, SpRag1/2L, in the sea urchin that is likely to share a common clustered ancestor with the Rag1/2 genes of jawed vertebrate adaptive immunity. This conclusion is based on similarities of the encoded proteins at the level of primary amino acid sequence, predicted protein structures, and their molecular interactions. Additional support is provided by the conserved genomic organization of the gene locus. These findings immediately raise the question whether the SpRagL locus encodes functional genes or inactive pseudogenes, or represents a remnant of a recently integrated mobile DNA element. The presence of complex exon/intron structure, readily detectable spliced messages, and their regulated expression suggests that SpRag1/2L are authentic genes. Furthermore, the absence of detectable terminal inverted repeats surrounding the genes further suggests that this locus did not result from a recent integration of a mobile DNA element. The most parsimonious explanation of our observations requires that the jawed vertebrate and sea urchin clusters diverged from a Rag1/2-like gene pair that was present in the common ancestor of living deuterostomes (Fig. 5). In contrast, no Rag1/2 homologs are evident in the fully sequenced genome of *Ciona intestinalis* (11), a urochordate. The reduced size (~150 Mbp vs. ~800 Mbp for *S. purpuratus*) and gene composition of the *Ciona* genome are consistent with a scenario of significant gene loss in this lineage (25), and it is plausible that the absence of a Rag cluster in this group is a consequence of this process.

One controversial question still remains: what are the origins of the Rag gene cluster and how does this relate to the evolutionary acquisition of a V(D)J recombination system? A currently widely held model suggests that Rag1 or both Rag genes were derived by horizontal gene transfer of a transposon, a mobile DNA element. In particular, a recent report suggested that a Transib transposon encoding a Rag1-like transposase may have been the origin of the minimal core region of Rag1 (6). The origin of Rag2, whether a part of the transposon or a preexisting gene by itself, remains more elusive and speculative in this model. Alternatively, the cluster may have originated by a more typical vertical evolutionary process, and be related to transposons only indirectly. Whatever the precise details of the
Intron transposons (i.e., before the divergence of the lineages that led to the sea urchins and vertebrates). The encoded proteins are likely to have carried out a "primitive" function that was different from its current role in vertebrates. Thus, an explanation of the origin of jawed vertebrate adaptive immunity. Along with recent findings such as a novel adaptive immune system mediated by lymphocyte-like cells in jawless vertebrates (27), this report presents possibilities for gaining deeper understanding of the emergence of jawed vertebrate adaptive immunity.

Materials and Methods

Animals. S. purpuratus specimens were obtained from Westwind Sealab Supplies (Victoria, BC, Canada) and from Patrick Leahy of the Caltech Kerchhoff Marine Laboratory (Newport Beach, CA). For some PCR analyses, DNA from the single male animal that is the subject of genomic sequencing was used as template.

Library Screening, Long-Range PCR, and RACE PCR. BAC clones were obtained by screening library filters representing ~12× genomic coverage from the single male animal that is the subject of the Sea Urchin Genome Project (28). BAC clone DNA was isolated by using the NucleoBond AX BAC Maxi kit (Clontech, BD Biosciences), and genomic distances were determined by long-range PCR (Expand Long Template PCR system; Roche Applied Sciences). RT-PCR and a RACE strategy was used to isolate transcript sequence (GeneRacer kit; Invitrogen).

To verify the genomic organization of the SpRagL locus, to establish linkage between regions encoded in different contigs, and to refine the genomic sequence in unresolved regions, we isolated and characterized two BAC clones from an S. purpuratus BAC library (Bac149P17, 150 kb; Bac78F1, 40 kb); BAC clones were obtained from the Genomics Technology Facility, Beckman Institute, California Institute of Technology, Pasadena, CA. Importantly, this library was constructed by using DNA from the same single male animal that is also the basis of the Sea Urchin Genome Project (28). Long-range PCR measurements of distances between exons and introns and to flanking genes on template from the two BAC clones spanning the entire SpRagL locus were used to confirm the 11/23/04 and subsequent genome sequence assemblies, resolve discrepancies, and bridge exons in separate contigs.

Expression Analysis by Quantitative PCR. RNA samples were isolated by using Trizol (Invitrogen) and RNeasy Microkit columns (Qiagen, Valencia, CA) with DNase treatment before reverse transcription. First-strand cDNA was synthesized from random hexamers by using TaqMan Reverse Transcription Reagents (Applied Biosystems). Quantitative PCR was carried out as described (29). Independent primer sets were used to confirm expression profiles. Measurements were made in quadruplicate on an ABI7000 real-time PCR machine by using SYBR green chemistry (Applied Biosystems). All primer sets spanned introns. The primer amplification efficiencies used to calculate transcript levels (1.86 for SpRag1L and 1.90 for SpRag2L) were measured by generating a standard curve with 50–8,000 copies of linearized cDNA plasmid. Prevalence was normalized to parallel 18S rRNA measurements. Each PCR well contained cDNA template generated from ~15 ng of total RNA (five embryo equivalents).

GST and Streptactin Pull-Down. The expression vectors for mouse Rag1 and Rag2 (pEBB-NSRAG1, pEBG-23C, and pEBG-PHDR2) have been described (30). The cDNAs of SpRag1L and bull shark Rag1 were cloned as BamHI/XhoI fragments into pEBB-NS (30) to create pEBBNS-spR1L and pEBBNS-clRag1, respectively. Similarly, the SpRag2L and sandbar shark Rag2 cDNAs were cloned as BamHI/NotI fragments into pEBG (31).
to generate pEBG-SpRag2L and pEBG-cpRag2, respectively. The pull-down assays were performed as described (30). Briefly, 293T cells were transfected with combinations of the Rag1 and Rag2 expression vectors by calcium phosphate precipitation. After 48 h, cells were harvested, lysed by sonication, and spun to remove insoluble debris. Lysates were then incubated with glutathione Sepharose beads (Amersham Pharmacia Biotech) or streptactin Sepharose beads (IBA), respectively. After extensive washing, bound proteins were eluted by boiling in SDS sample buffer and subsequently analyzed by Western blot by using anti-GST (Pharmingen) and anti-strep (IBA) antisera.

Expression of Recombinant Proteins and Gel-Shift Analysis. The respective bacterial expression vectors pMH6-mmR1cd and pMH6-spR1cd were individually transformed into the Rosetta E. coli strain (Novagen). Liquid cultures were grown at 16°C, and protein expression was induced at a cell density of OD\(_{600}\) = 0.8 by adding isopropyl \(\beta\)-D-thiogalactoside (IPTG) to a final concentration of 0.25 mM. After 16 h, the bacteria were harvested, and recombinant proteins were purified according to ref. 24. Gel-shift assays were performed as described (24).

Sequence Analysis and Structure Prediction. BLAST analyses were performed locally on the 11/23/04 Sea Urchin genome assembly, subsequently assemblies, and WGS trace sequences by using the BLASTALL program suite (32). Amino acid positions in the text are designated relative to Rag1/2 sequences of \textit{Mus musculus}, Protein Information Resource (PIR) accession nos. B33754 and P21784, respectively. Secondary structure prediction was performed by using GOR (31). HMMER 2.3.2 (33), and RSI-BLAST comparisons were made with the SMART prediction suite (34). A nonstringent search for Kelch- and PHD-containing proteins from the NCBI sea urchin gene models collected all HMMER matches to the Kelch1 (PF01344), Kelch2 (PF07646), and PHD (PF00628) PFAM profiles with \(E\)-value cut-off of 10.

Detailed experimental procedures and oligonucleotide sequences can be found in Supporting Experimental Procedures, which is published as supporting information on the PNAS web site.

We thank Gary W. Litman, Eric H. Davidson, Ellen V. Rothenberg, Michael J. Pazin, Michele K. Anderson, and F. Nina Papavassiliou for comments on the manuscript, and L. Courtney Smith and Susanna M. Lewis for helpful discussions. We are grateful to Darrell Norton, C. Titus Brown, and Gail Mueller for technical help. We thank Samuel Schluter (University of Arizona, Tucson) for providing the shark Rag1 and Rag2 cDNAs. In addition, we thank David G. Schatz and Eric H. Davidson for their support and stimulating discussions. This work was supported by a Canada Foundation for Innovation grant and funds from the Sunnybrook and Women’s Research Institute (to J.P.R.), and by the Intramural Research Program of the National Institute on Aging/National Institutes of Health.