

Evolution and diversification of lamprey antigen receptors: evidence for involvement of an AID-APOBEC family cytosine deaminase

Igor B Rogozin^{1,2,6}, Lakshminarayan M Iyer^{1,6}, Lizhi Liang³, Galina V Glazko⁴, Victoria G Liston⁵, Youri I Pavlov⁵, L. Aravind¹ & Zeev Pancer³

The variable lymphocyte receptors (VLRs) of jawless vertebrates such as lamprey and hagfish are composed of highly diverse modular leucine-rich repeats. Each lymphocyte assembles a unique VLR by rearrangement of the germline gene. In the lamprey genome, we identify here about 850 distinct cassettes encoding leucine-rich repeat modules that serve as sequence templates for the hypervariable VLR repertoires. The data indicate a gene conversion-like process in VLR diversification. Genomic analysis suggested a link between the VLR and platelet glycoprotein receptors. Lamprey lymphocytes express two putative deaminases of the AID-APOBEC family that may be involved in VLR diversification, as indicated by *in vitro* mutagenesis and recombination assays. Vertebrate acquired immunity could have therefore originated from lymphocyte receptor diversification by an ancestral AID-like DNA cytosine deaminase.

Antigen receptors of jawless vertebrates are fundamentally different from those of jawed vertebrates. Instead of immunoglobulins, the variable lymphocyte receptors (VLRs) of jawless fish are composed of highly diverse leucine-rich repeats (LRRs). In both vertebrate clades, however, lymphocytes bear somatically rearranged receptors of sufficient diversity to 'anticipate' newly encountered antigenic determinants, such as the anthrax spore coat BclA glycoprotein¹. In gnathostomes, assembly of repertoires that can potentially exceed 10¹⁴ different B cell receptors or T cell receptors occurs mainly by joining of immunoglobulin gene fragments encoding variable, diversity and joining segments by means of the recombinase-activating gene-encoded proteins RAG1 and RAG2 (ref. 2). Immunoglobulins are further diversified by somatic hypermutation, gene conversion or both of these processes that are mediated by the enzyme activation-induced cytosine deaminase (AID)^{3,4}.

Lamprey and hagfish represent the only two extant taxa of jawless vertebrates. Two types of VLRs, VLRA and VLRB, have been identified in hagfish; each has unique sequence features, and they are encoded by two separate loci⁵. Phylogenetic analysis of agnathan VLR genes has indicated that the single lamprey VLR is the ortholog of hagfish VLRB, whereas no ortholog of VLRA has been identified among about 18,000 expressed sequence tags derived from lamprey lymphocytes^{5,6}. The germline VLR genes of both agnathans do not encode

functional proteins but instead encode only portions of the amino (N) and carboxyl (C) termini of the mature VLRs; the sequences encoding those portions are separated by noncoding intervening regions. In lymphocytes, the germline VLR genes are assembled by somatic DNA rearrangement into mature VLR genes that encode the functional receptors. Each lymphocyte assembles a mature VLR gene with a unique diversity region that encodes a highly variable set of LRR modules: a 30- to 38-residue N-terminal LRR (LRRNT), one 18-residue LRR (LRR1), up to seven 24-residue LRRs (LRRV), one terminal 24-residue LRRV with a distinct sequence 'signature' (LRRVe), one 13-residue truncated LRR (connecting peptide (CP)), and a 51- to 66-residue C-terminal LRR (LRRCT)^{1,5,6}.

The mechanism of assembly of mature VLR genes in agnathan lymphocytes is unknown, but it has been predicted that the large repertoires can result from combinatorial assembly of diverse LRR module gene sequences contained in cassette arrays that flank the germline VLR genes^{1,5,6}. Here we show that each mature VLR is assembled from multiple genomic cassettes by a gene conversion-like process. We also present evidence of possible involvement of a DNA cytosine-deaminase member of the AID-APOBEC family in VLR assembly and that VLRs and the vertebrate platelet glycoprotein receptors may share a common evolutionary origin.

¹National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA. ²Institute of Cytology and Genetics, Novosibirsk 630090, Russia. ³Center of Marine Biotechnology, University of Maryland Biotechnology Institute, Baltimore, Maryland 21202, USA. ⁴Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, New York 14642, USA. ⁵Eppley Institute for Research in Cancer and Allied Diseases, Nebraska Medical Center, Omaha, Nebraska 68198, USA. ⁶These authors contributed equally to this work. Correspondence should be addressed to Z.P. (pancer@comb.umbi.umd.edu).

Received 14 February; accepted 4 April; published online 29 April 2007; doi:10.1038/ni1463

Table 1 Genomic VLRA and VLRB cassettes

Cassette type	Number	VLRA comments	Number	VLRB comments
3'LRRNT	13	Partially overlap germline 5'LRRNT	37	33 partially overlap germline 5'LRRNT
3'LRRNT-LRR1	30	Partially overlap germline 5'LRRNT; 24 include 3'LRRNT-5'LRR1; 6 include LRRV	68	64 partially overlap germline 5'LRRNT; 51 include 5'LRRV; 10 include LRRV
LRR1-LRRV	22	17 include 3'LRR1-5'LRRV; 5 include 3'LRR1-LRRV	N/A	N/A
LRRV	165	27 complete modules, 299 partial; 10 include 3'LRRV-LRRV-5'LRRV, 16 include LRRV-5'LRRV, 125 include 3'LRRV-5'LRRV, 7 only 3'LRRV, 6 only 5'LRRV, 1 only LRRV	199	301 complete modules, 261 partial; 130 contain 2-8 LRRV, 61 include LRRVe; 180 include 3'LRR1 or 3'LRRV; 81 include 5'LRRV; 10 include 3'LRRV-5'LRRV, 1 only 3'LRRV
LRRVe	57	Only 3'LRRV-5'LRRVe	64	29 complete LRRVe; 35 only 5'LRRVe; 37 include 3'LRRV-5'LRRVe
CP	52	3 include 3'LRRVe-CP; 42 include 3'LRRVe-CP-5'LRRCCT; 7 include CP-5'LRRCCT; 22 overlap germline 3'LRRCCT	28	27 include 3'LRRVe-CP-5'LRRCCT, partially overlap germline 5'LRRCCT
5'LRRCCT	54	Partially overlap germline 3'LRRCCT beginning at LRRCCT residues 1, 45 or 66	58	57 partially overlap germline 5'LRRCCT; 56 partially overlap germline 3'LRRCCT
Total	393		454	

Partial module sequences are presented as their 5' and 3' portions (all sequences are labeled as the module encoded). N/A, not applicable.

RESULTS

Assembly of mature VLRA and VLRB genes

We assembled a 'draft' of the sea lamprey (*Petromyzon marinus*) genome from sequences in the 'trace' archive of the National Center for Biotechnology Information. This draft included 70–80% of the coding portion of the genome. Arrays of tens of cassettes flanking both sides of the VLR genes have been identified before, each encoding one to three different types of LRR modules^{5,6}. To identify all available sequences encoding LRR cassettes corresponding to VLRA, we systematically scanned the genome using lamprey and hagfish VLR sequences as 'queries'. Our analysis demonstrated the presence of two types of genomic cassettes: 454 that exclusively encoded lamprey VLRB modules and 393 cassettes that exclusively encoded modules related to the hagfish VLRA (Supplementary Figs. 1 and 2 online); no modules were common to both VLR types. The VLRA-related cassettes resided in contiguous segments spanning about 2.2 megabases, at an average distance of 5.6 kilobases, and the VLRB-related cassettes resided in contiguous segments of about 2.1 megabases, spaced (on average) every 4.6 kilobases.

Using the VLRA cassettes as 'queries' to search against the database of about 18,000 lymphocyte expressed sequence tags^{5,6}, we identified one lamprey mature VLRA transcript. We also identified ten germline VLRA gene transcripts, indicating the gene was transcribed before its assembly into mature VLRA, as noted before for the preassembled lamprey VLRB and both types of hagfish VLR genes^{1,5}. The genomic organization of lamprey VLRA was similar to that of hagfish VLRA (Supplementary Fig. 3 online). The germline gene contained only two of the coding portions of mature VLRA; one encoded the signal peptide and the LRRNT, and the other was 3' sequence encoding the C-terminal portion of the LRRCCT and the C terminus. The latter included a potential glycosylphosphatidylinositol membrane-anchorage motif, which is a characteristic mode by which VLRA attach to the cell surface^{5,6}. A 210-nucleotide noncoding intervening region separated the two coding portions. Expression of VLRA was detectable by RT-PCR only in lymphocyte samples, showing highly diverse transcripts of mature VLRA as well as germline transcripts from the preassembled gene. The abundance of VLRA transcripts was lower than that of VLRB, and immune stimulation with a mixture of antigen and mitogen did not affect VLRA expression (data not shown), whereas mature VLRB transcripts are highly inducible by immune stimulation^{1,6}.

We categorized the genomic VLRA and VLRB cassettes into six to seven groups encoding different types of LRR modules, each corresponding to a distinct segment of the mature VLRA (Table 1). Most of the LRR modules encoded by the VLRA cassettes were incomplete, whereas VLRB cassettes usually encoded one or more complete modules and were flanked by sequences encoding additional module portions. The most abundant modules encoded by the cassettes were LRRV and LRRVe, 513 of which were encoded by 295 VLRA cassettes and 820 of which were encoded by 351 VLRB cassettes. The 5' sequences encoding the relevant portions of LRRCCT were also abundant in cassettes, with 54 VLRA cassettes encoding N-terminal LRRCCT modules that represented three distinct subgroups, each spanning a different portion of the LRRCCT module, and 58 VLRB cassettes encoding N-terminal LRRCCT modules.

The coding portions from these cassettes were 'tiled over' libraries of mature VLR sequences as sets of consecutive alignments along a minimal length of 30 nucleotides without any mismatch (Supplementary Figs. 4 and 5 online). Altogether these cassette alignments accounted for 87% of the sequences in 194 cloned mature VLRA genes and 93% of the sequences in 586 mature VLRB genes. We made an independent analysis of a set of 50 mature VLRB clones derived from liver DNA of the adult sea lamprey donor for the genome sequence project. Although the liver is a leukocyte-rich organ, we did not find mature VLRA amplicons in this DNA sample. The amount of coverage and mismatches among the 50 mature VLRB sequences 'tiled over' cassettes from the same animal was not different from those among 586 sequences derived from unrelated animals, indicating that our analysis was not biased because of genetic polymorphism among the landlocked population of sea lamprey in the Great Lakes of North America (our source for samples), which are notably less polymorphic than the anadromous sea lamprey⁶. We therefore merged these sets into a data set of 636 mature VLRB sequences (Supplementary Fig. 5).

VLRB rearrangement–intermediate clones, which are genes that have undergone cassette insertions but retain intact portions of the intervening region and are therefore nonfunctional, have been analyzed before. Among the VLRB rearrangement intermediates, cases of cassette insertions into the 5' or 3' coding portions of the preassembled gene have been noted¹. Among VLRA rearrangement intermediates, we identified two cases of insertions only in the 3' portion of

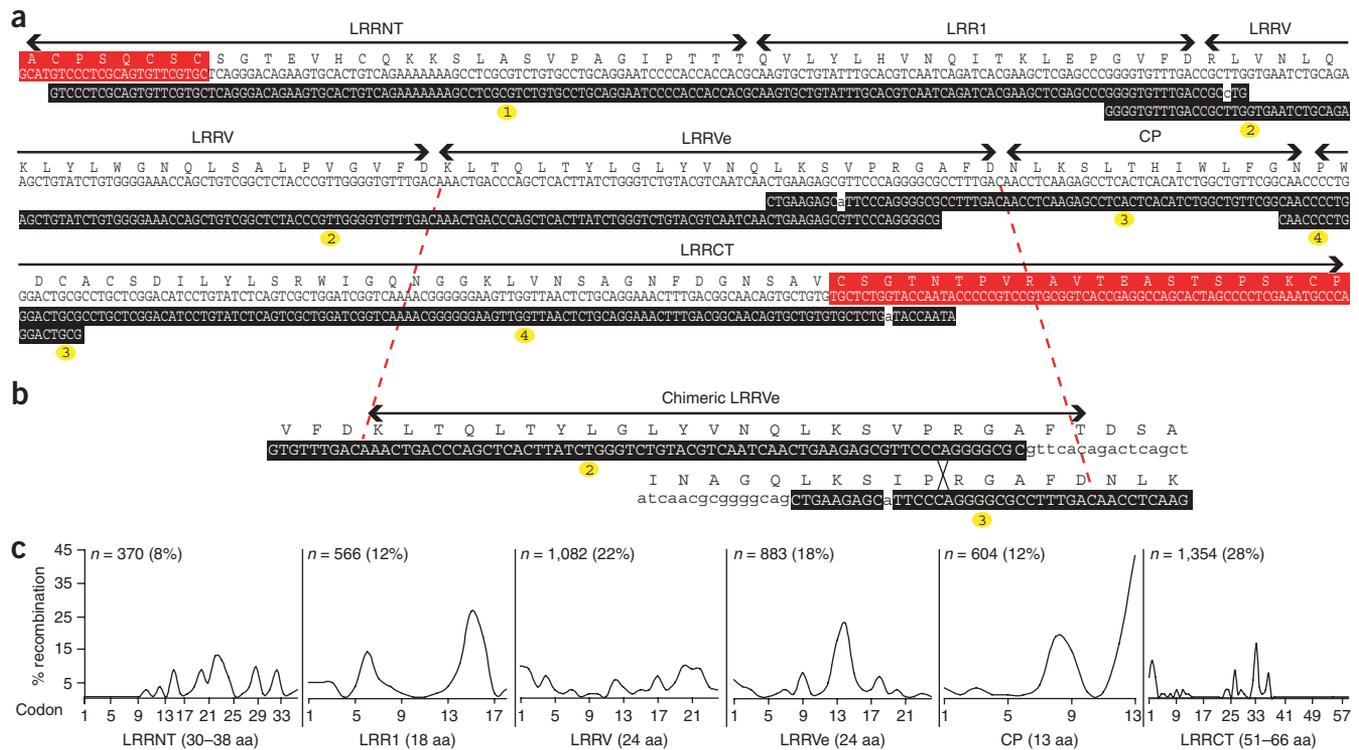


Figure 2 Stepwise assembly of mature *VLRB* genes from genomic cassettes. **(a)** Assembly of a representative *VLRB* (PmVLRB.4.2; GenBank accession code, 50086788) from four cassettes. The nucleotide and translated sequences are presented as described in **Figure 1a**. **(b)** A chimeric LRRVe module consisting of portions encoded by cassettes 2 and 3. **(c)** Distribution of cassette-recombination sites among sequences encoding constituent LRR modules, from 636 mature *VLRB* genes (presented as described in **Fig. 1c**): 517 clones published before¹; 69 clones from lymphocyte RT-PCR products of unstimulated animals 1, 3 and 4 (ref. 6); and 50 clones from a liver genomic DNA sample of the sea lamprey donor used for the genome sequence project.

or more that we used in our analysis. We found no evidence of diversification by means of template-independent DNA polymerases, such as terminal deoxynucleotidyl transferase, which generates junctional diversity in jawed vertebrate immunoglobulins⁸. Furthermore, none of the lamprey *VLR* genes or cassettes bore known recombination signals, and we identified no homologs of genes encoding RAG1, RAG2 or Transib transposons in the sea lamprey genome, although homologs of genes encoding these recombine proteins of unknown function have been reported for cnidarians and echinoderms^{9,10}.

Lamprey cytosine deaminases

We identified genes encoding two members of the AID-APOBEC cytosine deaminase family in the lamprey genome, called 'PmCDA1' and 'PmCDA2' here (for *P. marinus* cytosine deaminase; **Supplementary Fig. 6** online). The PmCDA1 polypeptide of 208 residues is encoded by a single exon, whereas the 331-residue PmCDA2 is encoded by four exons. Expression of each gene was detectable only in lymphocytes from blood and hematopoietic tissues, as indicated by

coexpression of the lymphocyte-specific marker *VLRB*⁶ (**Fig. 3a**). The two lamprey cytosine deaminases featured the characteristic deaminase 'HxE-PCxxC' (where 'x' is any amino acid) zinc-coordination motif¹¹, and we identified a C-terminal DNA-binding 'AT-hook' motif¹² in PmCDA2 (**Supplementary Figs. 6 and 7** online).

The lamprey cytosine deaminases emerged in phylogenetic trees as the closest 'sister group' of the gnathostome AID-APOBEC family of cytosine deaminases, with over 97% 'bootstrap' support for such grouping by various tree construction methods (**Fig. 3b**). We detected no AID-APOBEC members in invertebrate chordate genomes such as ciona or amphioxus. A sequence-structure analysis of the deaminase superfamily showed unique features shared by the AID-APOBEC proteins and PmCDA1 and PmCDA2. A conserved tryptophan residue in the C-terminal α -helix and a three-residue insert N-terminal to the 'PCxxC' motif distinguished the vertebrate AID-APOBEC clade from all other known deaminases, including the Tad deaminases, ADAR, Cdd1-like cytidine deaminases and deoxycytidylate deaminases¹¹. Further analysis showed that the deaminase superfamily could be

Table 2 Homogenous and chimeric LRR modules encoded by mature *VLRA* and *VLRB*

		LRRNT	LRR1	LRRV	LRRVe	CP	LRRCT	Total (%)
<i>VLRA</i>	Chimeric ^a	157	144	376	110	48	71	906 (86)
	Homogenous	11	11	9	0	120	0	151 (14)
<i>VLRB</i>	Chimeric ^a	572	252	404	476	214	549	2,467 (74)
	Homogenous	0	330	194	6	344	0	874 (26)

Samples included 194 mature *VLRA* sequences and 636 mature *VLRB* sequences.

^aLRR modules encoded by mature *VLR* genes composed of two or more modules encoded by genomic cassettes.

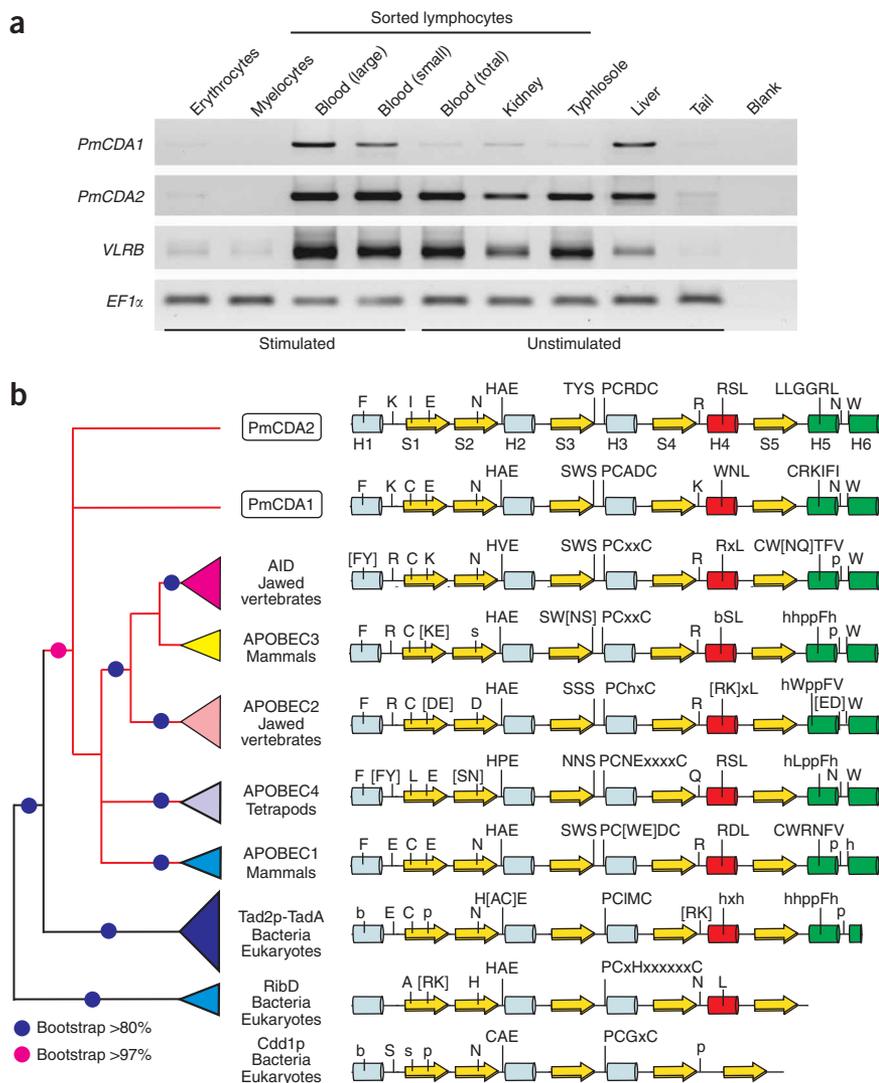
Figure 3 The sea lamprey cytosine deaminases.

(a) Distribution of *PmCDA1* and *PmCDA2* transcripts detected by RT-PCR of samples from sorted lymphocytes from hematopoietic tissues, the large and small blood lymphocyte fractions, sorted erythrocytes and macrophages, and liver and tail tissue; samples were obtained from untreated animals (Unstimulated) or after weekly stimulations for 4 weeks with an antigen-mitogen mixture (Stimulated). *VLRB*, lymphocyte-specific marker; *EF1 α* (elongation factor 1 α ; GenBank accession code, 51788791), cDNA control; Blank, no cDNA. (b) Maximum-likelihood tree of representative cytosine deaminases, rooted with the RibD-like deaminase; Cdd1p (bottom) represents the cytidine deaminase clade. Proportional triangles indicate phyletic spread. Cylinders indicate solved or predicted α -helices (H1–H6); arrows indicate β -strands (S1–S5; numbered for *PmCDA2*). The distinctive α -helix 4 is red; helices conserved among the TadA-Tad2p family and the AID-APOBEC family are green. There is a shared α -helix between strands 4 and 5 and the catalytic 'HxE' motif. Upper-case letters indicate distinguishing residues; brackets surround similar substitutions (b, big; h, hydrophobic; p, polar, s, small).

categorized into two distinct clades. One, whose members included the AID-APOBEC deaminases, usually featured an 'HxE' motif in the catalytic active site and shared a conserved α -helix between β -strands 4 and 5 of the core deaminase fold, making β -strand 5 parallel to β -strand 4 in the secondary structure. In the second clade, members typically had a 'CxE' motif in the active site and had no helix between β -strands 4 and 5 (Fig. 3b and Supplementary Fig. 7). In the former clade, we established a specific relationship between the AID-APOBEC family and the Tad2p-like tRNA adenosine deaminases, which are orthologs of the prokaryotic TadA¹¹. Members of the AID-APOBEC family and the Tad2p family shared two distinct C-terminal helices and several critical residues that included a highly conserved cysteine in β -strand 1 and, in α -helix 5, a strongly conserved motif 'hhppFh' (where 'h' is a hydrophobic residue, 'p' is a polar residue and 'F' is phenylalanine). The crystal structure of TadA in complex with its tRNA substrate showed that the phenylalanine residue in α -helix 5 was critical for the base-stacking interaction with its RNA substrate¹³. This comparison suggested that the Tad2p-like deaminases and AID-APOBEC proteins interact with their nucleic acid substrates in a similar way. The ancestral AID-APOBEC protein thus seems to have emerged early in vertebrate radiation, most probably from the Tad2p lineage.

PmCDA1-induced mutagenesis

Expression of human AID in *Escherichia coli* generates DNA cytosine-deamination sites that cause transition mutations of C:G nucleotide pairs to A:T pairs¹⁴. We therefore expressed *PmCDA1* in *E. coli* and tested its potential to confer rifampicin resistance through mutations in the RNA polymerase gene *rpoB*¹⁵. Expression of *PmCDA1* induced a higher frequency of rifampicin-resistant mutants than did a catalytically inactive *PmCDA1* with alanine residues replacing H66, E68,



C97 and C100. We found a 75-fold increase in the rate of rifampicin-resistance mutations when *PmCDA1* was expressed in an 'ung⁻' strain deficient in uracil-DNA glycosylase, an enzyme required for DNA repair (Supplementary Table 1 online). We then sequenced the *rpoB* gene from rifampicin-resistant ung⁻ colonies expressing *PmCDA1* and found dC-to-dT or dG-to-dA transitions (Supplementary Table 2 online), consistent with mutagenic activity by DNA cytosine deamination. We also noted DNA deamination activity of *PmCDA1* in a cell-free transcription-translation assay as mutations induced in a coexpressed *lacZ* construct (Supplementary Fig. 8 online). We transformed *E. coli* cells with *PmCDA1*-treated *lacZ* plasmid and assayed the cells for α -complementation of β -galactosidase. The assay showed a sevenfold increase in the number of mutant colonies over the number of colonies transformed with a *lacZ* plasmid treated with the catalytically inactive *PmCDA1* (42×10^{-3} per 10^3 versus 6×10^{-3} per 10^3). The *PmCDA1*-induced mutations were mostly C:G-to-A:T transitions that clustered in the first 300 bases of *lacZ* (Supplementary Table 2), similar to the mutagenic activity *in vitro* of human AID¹⁶.

The lamprey *PmCDA1* was highly mutagenic when expressed in ung⁻ yeast. Expression of *PmCDA1* induced canavanine resistance due to mutations in the arginine-permease gene *can1* (Supplementary Table 2), as well as in two other reverse-mutation assays¹⁷

Table 3 Spectra of *can1* mutations in *ung⁻ Saccharomyces cerevisiae*

	Spontaneous mutations ^a	PmCDA1-induced mutations	
G→A	14 (56%)	34 (33%)	
C→T	7 (29%)	68 (65%)	
G→T	0	2 (2%)	
Other single substitutions	4 (15%)	0	
Dinucleotide substitutions	0	1	
			'Fold increment' ^b
ABC:GVT	3 (13%)	44 (42%)	2.4 (0.006)
ABC ^u M:KGV ^u T	2 (9%)	37 (35%)	2.9 (0.013)
AKCM:KGMT	1 (4%)	31 (30%)	2.7 (0.019)
AKC:GMT	2 (9%)	35 (33%)	2.5 (0.024)
ANCM:KGNT	2 (9%)	40 (38%)	2.3 (0.041)

Mutation frequencies are corrected to represent equal proportions of the four bases. Nucleotide ambiguities: 'K' is G or T; 'M' is A or C; 'B' is C or G or T; 'V' is A or C or G; 'N' is A or T or C or G; only underlined positions are considered to be potential cytosine-deamination sites.

^aExpression of catalytically inactive PmCDA1 was used to estimate the background of spontaneous mutations. ^bPmCDA1-induced mutations at a mutable motif relative to the average at other C or G sites; *P* values in parentheses (*P* ≤ 0.01 is considered significant).

(Supplementary Table 3 online). Furthermore, expression of PmCDA1 in wild-type yeast diploids induced an increase in the rate of intragenic recombination of more than 20-fold over the rate induced by the catalytically inactive PmCDA1 or the rate induced by PmCDA1 expression in *ung⁻* yeast diploids (Supplementary Table 4 online). Thus, for enhancement of gene conversion in yeast, both the mutagenic activity of PmCDA1 and the endogenous uracil-DNA glycosylase were required, most probably to trigger DNA strand breaks at sites of cytosine deamination.

The spectrum of *can1* mutations induced in yeast expressing PmCDA1 was distinctively different from spontaneously occurring mutations, with fewer mutations in A:T base pairs and a lower frequency of transversions versus transitions (Table 3 and Supplementary Table 3 online). We therefore used a Monte Carlo simulation¹⁸ to analyze all possible sequence variants surrounding the sites of PmCDA1-induced *can1* mutations. We found five motifs that represented the mutational context; the motif with the best 'score' was ABC:GVT (mutable positions underlined; 'B' indicates T or C or G, and 'V' indicates A or G or C), which was 2.1-fold over-represented among the PmCDA1-induced *can1* mutations but rare among spontaneous mutations. Analysis of mismatches between mature *VLR* genes and corresponding genomic cassettes showed similar occurrence of mismatched A:T and C:G sites in *VLRA* cassettes (436 versus 483), whereas mismatched C:G sites in *VLRB* cassettes were 1.5-fold more abundant than mismatched A:T sites (901 versus 623). Among mature *VLRB* sequences, we noted a 1.8-fold greater abundance of mismatches in A sites relative to T sites (404 versus 220), whereas mismatched C:G sites were unbiased. Germline C:G-to-A:T sites mutations are characteristic of somatic hypermutation 'hotspots' in genes encoding immunoglobulins in jawed vertebrates^{18,19} resulting from AID activity⁴. These data could therefore suggest the presence of potential cytosine-deamination sites in *VLRB* cassettes.

To test that hypothesis, we analyzed the data sets of genomic cassettes aligned with mature *VLR* sequences for the occurrence of C:G mismatches in the context of the ABC:GVT motif. We noted a highly significant correlation for the mutable motif at C:G mismatch sites between *VLRB* cassettes and mature genes (*P* < 0.001) for

sequences from both the genome project donor and unrelated animals, and this correlation 'held' after elimination of CpG dinucleotides that undergo spontaneous mutations in vertebrate genomes (Table 4). The mutable bases in the ABC:GVT motif were rarely associated with the second position in *VLRB* codons (10%), at which all nucleotide substitutions result in amino acid substitutions, and the mutable motif was also rare among codons encoding solvent-exposed residues in LRRV and LRRVe modules, which are predicted to form the antigen contact interface of *VLRB*^{1,7}. In chicken immunoglobulins, the AID-mutable motif 'WRCH: DGYW' (mutable positions underlined; 'W' indicates A or T, 'R' indicates A or G, 'H' indicates A or C or T, 'D' indicates A or G or T, and 'Y' indicates C or T) is rare among sequences encoding antigen-contact residues²⁰, whereas the mutational 'hotspots' in mammalian antibodies are typically associated with regions of antigen contact²¹. Our analysis thus suggested that somatic mutations have only minor involvement in the diversification of potential antigen-contact residues in lamprey VLRLs, as in the case of chicken immunoglobulins, for which gene conversion is the chief mechanism of diversification^{3,20}.

The origin of cyclostome VLRLs

The lamprey LRRCT modules were extremely diverse in amino acid sequence, with most of the diversity mapping to a 5- to 14-residue insert located between the α -helix and the first β -strand of the module (Fig. 4a). This distinctive insert was absent from the LRRCT of other animal LRR-containing proteins such as the Toll family, CD180 or the expanded family of amphioxus LRR proteins²². The only other known protein with such an insert is the platelet glycoprotein receptor GPIIb α , which has a 17-residue insert in its LRRCT. Vertebrate GPIIb α , GPIIb β and GPIX are similarly structured proteins with flanking LRRNT and LRRCT modules and a central region with one (GPIIb β or GPIX) or eight (GPIIb α) 24-residue LRRV equivalents. Like the VLRLs, GPIIb α has a C-terminal stalk and a serine-threonine-rich region, followed by a transmembrane domain and a cytoplasmic tail²³.

The lamprey genome had three orthologs encoding the components of the platelet glycoprotein receptor complex, GPIIb α , GPIIb β and GPIX, all of which are LRR-containing proteins closely related in sequence to the VLRLs. The lamprey glycoprotein receptor complex genes were expressed in lymphocytes and possibly also in thrombocytes, as indicated by several expressed sequence tags from our enriched lymphocyte library^{5,6} (Fig. 4b). Comparison of the structure

Table 4 Occurrence of the ABC:GVT motif at VLR mismatch sites

	C:G mismatch	ABC:GVT motif	<i>P</i> value
<i>VLRA</i>			
Single mismatch	483	72 (12%)	0.125
Single except CpG	379	61 (16%)	0.102
Multiple mismatches ^a	39	6 (15%)	0.331
Multiple except CpG	30	4 (13%)	0.659
<i>VLRB</i>			
Single mismatch	901	229 (25%)	<0.001*
Single except CpG	724	197 (27%)	<0.001*
Multiple mismatches	198	66 (33%)	<0.001*
Multiple except CpG	155	54 (35%)	<0.001*

Mismatch sites between cassettes and 194 mature *VLRA* sequences and 636 mature *VLRB* sequences. In ABC:GVT, only the underlined positions are considered potential cytosine-deamination sites. The percent of total mismatches is in parentheses. The fraction of C:G bases in ABC:GVT mutable motifs are as follows: *VLRA*, 0.114; *VLRB*, 0.102. *, *P* ≤ 0.05 (significant).

^aSame mismatch site found in two or more cassettes.

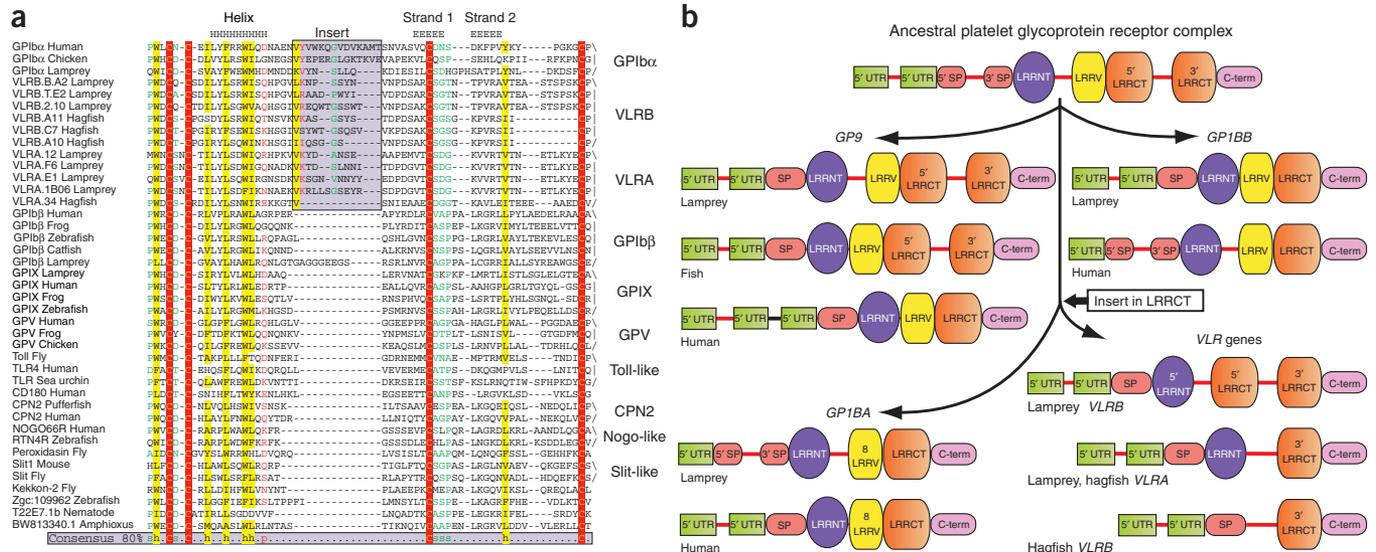


Figure 4 Evolutionary link between agnathan VLRs and vertebrate platelet receptor glycoproteins. **(a)** Alignment of metazoan LRRCT domains, showing a unique insert (boxed) in the LRRCT of the VLRs and GPIb α . Residues: hydrophobic (A, C, F, I, L, M, V, W, Y), yellow highlighting; small (A, G, S, V, C, D, N, P, T), green letters; polar (S, T, E, D, K, R, N, Q, H, C), red highlighting; conserved cysteine residues, red letters. Genus and species: human, *Homo sapiens*; chicken, *Gallus gallus*; lamprey, *Petromyzon marinus*; hagfish, *Eptatretus stoutii*; frog, *Xenopus laevis*; zebrafish, *Danio rerio*; catfish, *Ictalurus punctatus*; fly, *Drosophila melanogaster*; sea urchin, *Strongylocentrotus purpuratus*; pufferfish, *Takifugu rubripes*; mouse, *Mus musculus*; nematode *Caenorhabditis elegans*; amphioxus, *Branchiostoma floridae*. **(b)** The ancestral vertebrate platelet glycoprotein receptor gene complex and evolution of the VLR genes. Cyclostome VLR genes are predicted to have diverged from a GPIb α -like gene, based on a unique insert in the sequence encoding LRRCT, a conserved intron in the 5' untranslated region (UTR) and similarly positioned introns and intervening regions in GPIX and germline VLRB. GPIb α has eight LRRV modules. SP, signal peptide. Split and incomplete module sequences are presented as their 5' and 3' portions (all sequences are labeled as the module encoded); red lines, conserved positions of introns or intervening regions; black line in human GP9, newly identified intron. The models are based on genomic and cDNA sequences (GenBank accession codes in parentheses): GP9 (encoding GPIX), lamprey (51795833), zebrafish (32340789) and human (4504077); GPIBB (encoding GPIβ), lamprey (51799529) and human (4504073); GPIBA (encoding GPIb α), lamprey (51789956) and human (47419932).

of lamprey and gnathostome glycoprotein receptor complex genes indicated the ancestral form probably contained an intron in the 5' untranslated region and two large introns in the coding region (Fig. 4b). Notably, all agnathan VLR genes contained an intron in the 5' untranslated region, and internal introns of the lamprey platelet glycoprotein receptor complex genes approximately corresponded to the positions of the two intervening regions in the lamprey VLRB gene. We did not detect platelet glycoprotein receptor complex family members in the genomes of the sea urchin, ciona or amphioxus, suggesting the family arose in the vertebrate ancestor.

DISCUSSION

Various LRR-containing proteins have been recruited for microbial recognition on many occasions in plants and animals, in some cases with enormous lineage-specific expansions. These include hundreds of cytoplasmic and extracellular plant disease-resistance proteins²⁴, a vastly expanded arsenal of 203 Nod–NACHT–LRR proteins and 222 Toll-like receptors in the sea urchin²⁵ and, in amphioxus, about 90 Nod–NACHT–LRR proteins, about 50 Toll-like receptors and several hundreds of cell surface and secreted forms of LRR-containing proteins²². The animal extracellular versions of LRR proteins resemble VLRs in having the typical LRRNT, LRR and LRRCT modules, but otherwise are not closely related to the VLRs. Instead, our analysis has indicated that VLRs are related to the vertebrate-specific platelet glycoprotein complex hemostatic receptors, based on gene structure and on the distinct LRRCT insert. Notably, the crystal structure of human GPIb α in complex with its ligand, the vWA domain of von Willebrand factor, shows that the LRRCT insert forms an extended hairpin that projects across the concave surface in contact

with the von Willebrand factor²⁶. The highly variable inserts in the LRRCT of VLRs may therefore similarly interact with antigens, regulating the size and affinity of ligands accommodated by the concave surface. The similarity between VLRs and platelet glycoprotein receptor complex suggested a possible scenario for the origin of these receptors. Before the vertebrate radiation, the ancestral platelet glycoprotein receptor diversified into at least three members, GPIb α , GPIβ and GPIX, that were adapted for hemostatic function. In cyclostomes, a duplication of GPIb α that already featured an LRRCT insert may have generated the ancestral VLR, which was 'recruited' as a lymphocytic receptor. Given that VLRs are glycosylphosphatidylinositol-anchored proteins, they may form a complex with the platelet glycoprotein receptor complex orthologs, thus enabling signaling through their cytoplasmic tail, as occurs among members of the platelet glycoprotein complex²³. Recruitment of LRR proteins as lymphocyte antigen receptors and their mechanism of somatic diversification thus seem to be convergent innovations of the cyclostomes.

The availability of a nearly complete sea lamprey genome sequence allowed us to make a genome-wide analysis of the building blocks of VLRs. About 2,000 highly variable genomic sequences encoding LRR modules of different types provide sequence templates for almost all VLR diversity. The lower boundary of agnathan VLR repertoires has been estimated to be 10^{14} to 10^{17} different receptors of each type, assuming that complete LRR modules are the building blocks of mature VLRs¹. However, analysis of 2,235 cassette insertion sites in VLRA and 4,859 in VLRB indicated that 74–86% of the sequences encoding modules in mature VLRs are chimeric; such hybrid module sequences have also been identified in mature VLRB of the Japanese

lamprey (*Lethenteron japonicum*)²⁷. It is therefore possible that much larger VLR repertoires are generated by intramodular recombination among the sequences encoding 763 modules in 393 genomic VLRA cassettes and the sequences encoding 1,106 modules in 454 VLRB cassettes of the sea lamprey.

Assembly of mature VLR genes occurs through conversion events along relatively short regions of sequence homology, comparable in size to gene-conversion tracts of 8–200 nucleotides in sequences of chicken immunoglobulin pseudogenes²⁸. In the lamprey, insertion of cassette sequences begins with homology-based pairing between the coding portions in cassettes and in preassembled VLR genes or in previously inserted elements, as in a typical gene-conversion process²⁸. Cassette insertions terminate uniquely, however, by pairing between noncoding flanking regions of the cassettes and short homologous sequences along the intervening regions of germline VLR genes, as evident among rearrangement intermediate clones of both VLR types, in which insertions often terminate with noncoding flanking regions derived from the donor loci¹.

Assembly of VLRB in the Japanese lamprey has been proposed to occur by a 'copy-choice' mechanism²⁷. According to the model, the short homology regions between cassette-encoded modules and preassembled gene portions can be used to copy donor modules into the assembling gene. In the fission yeast *Schizosaccharomyces pombe*, 'copy-choice' recombination mediates mating-type switching after nicking of the 'imprinted' allele by a sequence-specific single-strand nuclease²⁹, and 'copy choice' can explain *in vitro* replication slippage that causes deletion of DNA sequences flanked by short direct repeats³⁰. However, analysis of ten sea lamprey VLRB rearrangement intermediates in which only deletions of germline intervening regions occurred without cassette insertions showed no obvious pattern in the context or distribution of deletion sites¹, and no flanking repeats have been reported in 41 such clones from the Japanese lamprey²⁷. A mechanism other than 'copy choice' may have therefore generated such deletions.

Here we have demonstrated two AID-APOBEC family members expressed specifically in lamprey lymphocytes, thus expanding this family of proteins beyond the jawed vertebrates. Our data have indicated that PmCDA1 induced mutations in yeast genes in a mutable context that was common among VLRB cassettes, accounting for about 27% of the C:G mismatches between cassettes and mature VLRB sequences. Furthermore, expression of PmCDA1 in yeast induced intragenic mitotic recombination events that typically occur by gene conversion; expression of human AID in yeast generates a similar 'recombinogenic' phenotype³¹. We therefore speculate that cytosine-deamination sites along homology-paired cassettes and coding and noncoding portions in VLR genes may produce DNA breaks, with the potential to initiate the process of insertion of cassette sequences into the assembling VLR genes; such breaks can also trigger the DNA deletions in intervening regions of VLR genes. AID-induced DNA strand breaks are a requisite for gnathostome class-switch recombination and for immunoglobulin gene conversion^{3,4}, the process that generates most of the diversity in antibodies of chicken, rabbits, cattle, swine and horses^{32–35}. Future studies should characterize the possible function(s) of the lamprey AID-APOBEC members in the assembly of VLRA and during additional waves of diversification that may be required to mount such remarkably diverse VLR repertoires, perhaps analogous to the waves of antibody affinity maturation in B cells of jawed vertebrates.

Our phylogenetic analysis has suggested that the AID-APOBEC family emerged at the 'dawn' of vertebrate radiation. Ancestral

vertebrates could have therefore used deaminase-mediated mutagenesis as protective means against various infectious agents^{36,37}. Such mutagenic activity could have also resulted in somatic mutations in genes encoding immune receptors of the phagocytes, some of which could enhance microbial recognition, thus paving the way for a primordial mechanism of deaminase-mediated receptor diversification, as hypothesized before^{8,38}. In our scenario for the origin of vertebrate acquired immunity, we propose that lymphocytes evolved in the ancestral vertebrate to serve as the platform for somatic receptor diversification. The new class of proliferation-competent lymphocytes most probably stemmed from the nonproliferating phagocytes of ancestral vertebrates; the phagocytic activity of the B cells of fish and frogs provides support for this view³⁹. To confer effective immunity, the newly emerged lymphocytes acquired an exclusive potential for clonal expansion after antigen stimulation through their uniquely diversified receptors. Subsequently, different types of variable receptors and recombination strategies were selected in the jawless and jawed vertebrates. Nonetheless, the high selective value of the Cambrian innovation of long-lived proliferation-competent lymphocytes bearing rearranging antigen receptors is attested by the conservation of both VLRA and VLRB in lamprey and hagfish, two cyclostome lineages that split about 500 million years ago⁴⁰, and by the conservation of rearranging immunoglobulins for about 450 million years in all the gnathostomes⁴¹.

METHODS

Sample preparation and PCR amplification. RNA samples originated from four unstimulated animals and four that were stimulated by weekly intraperitoneal injection for 4 weeks of 1×10^7 live *E. coli*, 1×10^7 sheep erythrocytes, 50 μ g phytohemagglutinin and 25 μ g pokeweed mitogen. Lymphocytes were sorted from the blood, kidneys and typhlosoles; erythrocytes and macrophages were sorted from the blood as described⁶. Livers and tails were also collected from these animals. Animal procedures were approved by the Institutional Animal Care and Use Committee of the University of Maryland Biotechnology Institute.

Expand High Fidelity (Roche) was used for PCR; *PmCDA1* was amplified in 35 cycles, *PmCDA2* was amplified in 30 cycles, and *VLRB* and *EF1 α* were amplified in 25 cycles. Mature VLRA amplicons were excised from gels and were cloned from lymphocyte RT-PCR products and from genomic PCR products. Additional VLRB clones originated from stored cDNA of unstimulated animals 1, 3 and 4 (ref. 6) and from a liver genomic DNA sample of the sea lamprey donor that was used for the genome sequence project. Lamprey cytosine deaminases were cloned from lymphocyte RT-PCR products with primers PmCDA1.F1 plus PmCDA1.R1, and PmCDA2.F2 plus PmCDA2.R2 (primers, **Supplementary Table 5** online).

Expression of recombinant PmCDA1. For bacterial expression, the *PmCDA1* ORF was amplified with primers PmCDA1.F and PmCDA1.R, which introduced 5' *Nde*I and 3' *Xho*I restriction sites for cloning into pET-24b (Novagen). Catalytically inactive PmCDA1 was generated by replacement of H66, E68, C97 and C100 with alanine residues by oligonucleotide mutagenesis. For yeast expression, *PmCDA1* was amplified with PmCDA-N and PmCDA-C, which introduced 5' *Not*I and 3' *Spe*I restriction sites for cloning into pESC-LEU (Stratagene).

PmCDA1 was expressed in *E. coli* Rosetta DE3 (Novagen) and in the *ung*⁻ derivative NR16930 (*ung152:Tn10 Tet^R*). Rifampicin selection was accomplished with 0.1 mM isopropylthiogalactoside for expression induction; 40 *rpoB* mutants were sequenced as described¹⁴. An RTS 100 *E. coli* HY kit (Roche) was used for *in vitro* transcription and translation of PmCDA1. Reaction volumes of 33 μ l contained 400 ng pET-19b plasmid, encoding β -galactosidase (insert control 69676; Novagen), and 100 ng PCR-amplified T7 cassettes, encoding PmCDA1.6H, or catalytically inactive PmCDA1.6H, or the empty pET-24b as control. After 12 h at 20 °C, the *lacZ* plasmids were extracted by phenol-chloroform and were introduced by electroporation into *E. coli* NR16930 cells. After 5 h of culture at 37 °C, the plasmids were purified

again and were used to transform TOP10 bacteria (Invitrogen) expressing T7 RNA polymerase under control of the *tac* promoter in a pACYC184 construct. The *lacZ* mutants were quantified with an α -complementation assay on isopropylthiogalactoside–5-bromo-4-chloro-3-indoyl- β -D-galactoside plates. For PmCDA1 treatment, 42 colonies were white for every 1,000 blue colonies; for treatment with the catalytically inactive PmCDA1, 6 colonies were white for every 1,000 blue colonies; and for treatment with the control pET-24b, 8 colonies were white for every 1,000 blue colonies. Fifteen white clones from the PmCDA1 treatment and seven from the treatment with catalytically inactive PmCDA1 were sequenced with the primers T7 and lacZ.R (nucleotides 1564–1583 of *lacZ*).

PmCDA1 was expressed in yeast strain 1B-D770 and its *ung*⁻ derivative. From the *ung*⁻ strain, 92 PmCDA1-induced *can1* mutants were sequenced, and from those that expressed the catalytically inactive PmCDA1, 26 *can1* mutants were sequenced. Yeast cultures, sequencing of *can1* mutants, and the adenine- and tryptophan-reversion assays of nonsense mutations in suppressor genes encoding tRNA were done as described^{17,31}. Mutation rates were determined by fluctuation tests with nine to twelve independent bacterial or yeast cultures, assuming constant mutation rates during PmCDA1 induction. Rates of recombination induced by PmCDA1 expression were determined in the yeast wild-type diploid strain YUNI500 and in the YUNI501 *ung*⁻ strain as described³¹, except that Noble agar (Sigma) was used.

Sea lamprey genome sequence. A ‘draft’ genome sequence was assembled at a coverage of 5.9 \times from 9,424,475 sequences in the sea lamprey Ensembl archive (ftp://ftp.ensembl.org/pub/traces/petromyzon_marinus), with 6,472,038,425 ‘Q20 bases’ (sequencing base ‘calls’ with an error rate of less than 1%) with the parallel contig assembly program PCAP⁴². The assembly size of about 1 gigabases instead of the predicted 2.3 gigabases was due to the 56% content of DNA repeats, which the assembly program stacked or discarded. The quality of the draft assembly was estimated with 3,958 unique sequences assembled from about 18,000 sea lamprey expressed sequence tags^{5,6}. In searches using the MEGABLAST program, 87% of the queries retrieved ‘hits’ with expectation scores (*E* values) of 10⁻⁵⁰ or less, 80% of the queries retrieved ‘hits’ with *E* values of 10⁻⁷⁵ or less, and 72% of the queries retrieved ‘hits’ with *E* values of 10⁻¹⁰⁰ or less.

Cataloging of VLR cassettes. Cassettes were extracted from 160 VLRA contiguous segments and 163 VLRB contiguous segments with the TBLASTX program⁴³ using nucleotide sequences encoding all the LRR modules from our VLR data sets as ‘queries’. For the removal of non-VLR sequences, searches with the BLASTN program⁴³ and TBLASTN program⁴³ were used with cassettes as ‘queries’ against the data sets of mature VLRA and VLRB and against eukaryotic genomes in GenBank. Only cassettes with significant ‘best hits’ with VLR were retained (*E* value \leq 0.05). Cassettes were ‘tiled over’ mature VLR sequences and were classified per type (Table 1 and Supplementary Figs. 2 and 3 online). Same-type cassette protein sequences were aligned with the T-COFFEE program⁴⁴ and were used to create sequence ‘logos’ (<http://weblogo.berkeley.edu>). Nucleotide alignments were then constructed based on protein alignments.

Sequence analysis and phylogenetic trees. The sea lamprey cytosine deaminases and previously identified family members were used as ‘queries’ for searches with the PSI-BLAST program⁴⁵ against GenBank databases, with a profile threshold of 0.01 and iteration until convergence. The resulting proteins were clustered with BLASTCLUST (<ftp://ftp.ncbi.nih.gov/blast>). Multiple alignment of the cytosine-deaminase ‘superfamily’ was constructed with structural alignments using the JPRED⁴⁵, PSI-BLAST and DALI⁴⁶ programs, to combine family alignments obtained with the T-COFFEE program⁴⁴ with adjusted secondary structure predictions. For phylogenetic analyses, neighbor-joining trees were constructed with the WEIGHBOR program⁴⁷, which provides corrections for long-branch attraction. Maximum-likelihood trees were constructed with the maximum likelihood distance matrices as input, followed by local rearrangement by the Protml program⁴⁸ and the Molphy program (<http://www.ism.ac.jp/ismilib/softother.e.html>). ‘Bootstraps’ were calculated with the relative estimate of logarithmic likelihood with 10,000 replicates.

Analysis of recombination events in VLR genes and mutation spectra. Recombination sites were analyzed in alignments of mature VLR sequences

‘tiled over’ genomic cassettes (Supplementary Fig. 3 online). The FASTA program (<ftp://ftp.virginia.edu/pub/fasta>) was used for database searches and alignments. An alignment length of 30 nucleotides or more without mismatch was used; alignments were then extended in the 5’ and 3’ directions with 10 nucleotides or less, allowing three mismatches or less. Intramodular recombination was defined for modules in mature VLR genes that formed from two or more cassettes. Sequence mismatches between mature VLR genes and cassettes were considered potential somatic mutations if no contradictions with other cassettes were found. The amount of coverage and mismatches in alignments of 50 mature VLRB genes ‘tiled’ to cassettes derived from the same animal was compared with that of 586 sequences derived from unrelated animals with the Fisher exact test. The fraction of single mismatches was not different for the sets of sequences; *P* = 0.90 for all single C:G mismatches and *P* = 0.89 when CpG oligonucleotides were excluded. There was also no difference in the percent of chimeric and homogenous modules for the two sets (*P* = 0.97).

Mutation spectra were calculated for concatenated mature VLRA or VLRB sequences; the potential somatic mutations were merged into one spectrum per VLR type. Correlations between the distribution of mutable motifs and mutations along target sequences were measured with the Monte Carlo CONSEN program¹⁸. The probability ‘*P*_{W \leq Wrandom}’ was calculated by comparison of the mutational spectra (‘W’) with 10,000 sets of randomized mutations (‘Wrandom’). Correlations with *P* values of 0.05 or less were considered significant. A Monte Carlo modification of the Pearson χ^2 test of spectra homogeneity⁴⁹ was used for comparison of the frequencies of different types of substitutions and distributions of mutations across target sequences.

Accession codes. GenBank: PmCDA1, PmCDA2, VLRA and VLRB from unstimulated larvae, EF094560–EF094823; VLRB from the genome project donor, EF464171–EF464219; VLR cassettes, EF528588–EF529434.

Note: Supplementary information is available on the Nature Immunology website.

ACKNOWLEDGMENTS

We thank the Genome Sequencing Center at Washington University for public access to the lamprey genome sequences; E.R. Mardis for genomic DNA from the sea lamprey donor of the genome sequence project; S. Kozmin (National Institute of Environmental Health Sciences) for the *E. coli* Rosetta *ung*⁻ strain, A. Lada (Saint Petersburg University in Russia) for helping measure PmCDA1-induction of recombination in yeast; and M.D. Cooper (University of Alabama at Birmingham), M.F. Flajnik (University of Maryland, Baltimore) and M. Diaz (National Institute of Environmental Health Sciences) for discussions. Supported by the National Library of Medicine–National Institutes of Health–Department of Health and Human Services Intramural Research Program (I.B.R., L.M.I. and L.A.) and the National Science Foundation (MCB-0614672 to Z.P.). This is contribution 07-165 from the Center of Marine Biotechnology.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at <http://www.nature.com/natureimmunology/>
Reprints and permissions information is available online at <http://ngp.nature.com/reprintsandpermissions>

1. Alder, M.N. *et al.* Diversity and function of adaptive immune receptors in a jawless vertebrate. *Science* **310**, 1970–1973 (2005).
2. Oettinger, M.A., Schatz, D.G., Gorka, C. & Baltimore, D. RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. *Science* **248**, 1517–1523 (1990).
3. Arakawa, H., Hauschild, J. & Buerstedde, J.M. Requirement of the activation-induced deaminase (AID) gene for immunoglobulin gene conversion. *Science* **295**, 1301–1306 (2002).
4. Muramatsu, M. *et al.* Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102**, 553–563 (2000).
5. Pancer, Z. *et al.* Variable lymphocyte receptors in hagfish. *Proc. Natl. Acad. Sci. USA* **102**, 9224–9229 (2005).
6. Pancer, Z. *et al.* Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. *Nature* **430**, 174–180 (2004).
7. Kim, H.M. *et al.* Structural diversity of the hagfish variable lymphocyte receptors. *J. Biol. Chem.* **282**, 6726–6732 (2007).

8. Schatz, D.G. Antigen receptor genes and the evolution of a recombinase. *Semin. Immunol.* **16**, 245–256 (2004).
9. Fugmann, S.D., Messier, C., Novack, L.A., Cameron, R.A. & Rast, J.P. An ancient evolutionary origin of the Rag1/2 gene locus. *Proc. Natl. Acad. Sci. USA* **103**, 3728–3733 (2006).
10. Kapitonov, V.V. & Jurka, J. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol.* **3**, e181 (2005).
11. Conticello, S.G., Thomas, C.J., Petersen-Mahrt, S.K. & Neuberger, M.S. Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. *Mol. Biol. Evol.* **22**, 367–377 (2005).
12. Aravind, L. & Landsman, D. AT-hook motifs identified in a wide variety of DNA-binding proteins. *Nucleic Acids Res.* **26**, 4413–4421 (1998).
13. Losey, H.C., Ruthenburg, A.J. & Verdine, G.L. Crystal structure of *Staphylococcus aureus* tRNA adenosine deaminase TadA in complex with RNA. *Nat. Struct. Mol. Biol.* **13**, 153–159 (2006).
14. Petersen-Mahrt, S.K., Harris, R.S. & Neuberger, M.S. AID mutates *E. coli* suggesting a DNA deamination mechanism for antibody diversification. *Nature* **418**, 99–103 (2002).
15. Gariyban, L. *et al.* Use of the rpoB gene to determine the specificity of base substitution mutations on the *Escherichia coli* chromosome. *DNA Repair (Amst.)* **2**, 593–608 (2003).
16. Bransteitter, R., Pham, P., Calabrese, P. & Goodman, M.F. Biochemical analysis of hypermutational targeting by wild type and mutant activation-induced cytidine deaminase. *J. Biol. Chem.* **279**, 51612–51621 (2004).
17. Mayorov, V.I. *et al.* Expression of human AID in yeast induces mutations in context similar to the context of somatic hypermutation at G-C pairs in immunoglobulin genes. *BMC Immunol.* **6**, 10 (2005).
18. Rogozin, I.B., Pavlov, Y.I., Bebenek, K., Matsuda, T. & Kunkel, T.A. Somatic mutation hotspots correlate with DNA polymerase η error spectrum. *Nat. Immunol.* **2**, 530–536 (2001).
19. Milstein, C., Neuberger, M.S. & Staden, R. Both DNA strands of antibody genes are hypermutation targets. *Proc. Natl. Acad. Sci. USA* **95**, 8791–8794 (1998).
20. Rogozin, I.B., Sredneva, N.E. & Kolchanov, N.A. Somatic hypermutagenesis in immunoglobulin genes. III. Somatic mutations in the chicken light chain locus. *Biochim. Biophys. Acta* **1306**, 171–178 (1996).
21. Wagner, S.D., Milstein, C. & Neuberger, M.S. Codon bias targets mutation. *Nature* **376**, 732 (1995).
22. Pancer, Z. & Cooper, M.D. The evolution of adaptive immunity. *Annu. Rev. Immunol.* **24**, 497–518 (2006).
23. Canobbio, I., Balduini, C. & Torti, M. Signalling through the platelet glycoprotein Ib-V-IX complex. *Cell. Signal.* **16**, 1329–1344 (2004).
24. Meyers, B.C., Kozik, A., Griego, A., Kuang, H. & Michelmore, R.W. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell* **15**, 809–834 (2003).
25. Hibino, T. *et al.* The immune gene repertoire encoded in the purple sea urchin genome. *Dev. Biol.* **300**, 349–365 (2006).
26. Huizinga, E.G. *et al.* Structures of glycoprotein Ibx and its complex with von Willebrand factor A1 domain. *Science* **297**, 1176–1179 (2002).
27. Nagawa, F. *et al.* Antigen-receptor genes of the agnathan lamprey are assembled by a process involving copy choice. *Nat. Immunol.* **8**, 206–213 (2007).
28. McCormack, W.T. & Thompson, C.B. Chicken IgL variable region gene conversions display pseudogene donor preference and 5' to 3' polarity. *Genes Dev.* **4**, 548–558 (1990).
29. Arcangioli, B. & de Lahondes, R. Fission yeast switches mating type by a replication-recombination coupled process. *EMBO J.* **19**, 1389–1396 (2000).
30. Viguera, E., Canceill, D. & Ehrlich, S.D. Replication slippage involves DNA polymerase pausing and dissociation. *EMBO J.* **20**, 2587–2595 (2001).
31. Poltoratsky, V.P., Wilson, S.H., Kunkel, T.A. & Pavlov, Y.I. Recombinogenic phenotype of human activation-induced cytosine deaminase. *J. Immunol.* **172**, 4308–4313 (2004).
32. Di Noia, J.M. & Neuberger, M.S. Immunoglobulin gene conversion in chicken DT40 cells largely proceeds through an abasic site intermediate generated by excision of the uracil produced by AID-mediated deoxycytidine deamination. *Eur. J. Immunol.* **34**, 504–548 (2004).
33. Butler, J.E. Immunoglobulin diversity, B-cell and antibody repertoire development in large farm animals. *Rev. Sci. Tech.* **17**, 43–70 (1998).
34. Reynaud, C.A., Anquez, V., Grimal, H. & Weill, J.C. A hyperconversion mechanism generates the chicken light chain preimmune repertoire. *Cell* **48**, 379–388 (1987).
35. Thompson, C.B. & Neiman, P.E. Somatic diversification of the chicken immunoglobulin light chain gene is limited to the rearranged variable gene segment. *Cell* **48**, 369–378 (1987).
36. Rogozin, I.B., Basu, M.K., Jordan, I.K., Pavlov, Y.I. & Koonin, E.V. APOBEC4, a new member of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases predicted by computational analysis. *Cell Cycle* **4**, 1281–1285 (2005).
37. Gourzi, P., Leonova, T. & Papavasiliou, F.N. A role for activation-induced cytidine deaminase in the host response against a transforming retrovirus. *Immunity* **24**, 779–786 (2006).
38. Flajnik, M.F. Comparative analyses of immunoglobulin genes: surprises and portents. *Nat. Rev. Immunol.* **2**, 688–698 (2002).
39. Li, J. *et al.* B lymphocytes from early vertebrates have potent phagocytic and microbicidal abilities. *Nat. Immunol.* **7**, 1116–1124 (2006).
40. Kuraku, S. & Kuratani, S. Time scale for cyclostome evolution inferred with a phylogenetic diagnosis of hagfish and lamprey cDNA sequences. *Zool. Sci.* **23**, 1053–1064 (2006).
41. Litman, G.W., Cannon, J.P. & Rast, J.P. New insights into alternative mechanisms of immune receptor diversification. *Adv. Immunol.* **87**, 209–236 (2005).
42. Huang, X., Wang, J., Aluru, S., Yang, S.P. & Hillier, L. PCAP: a whole-genome assembly program. *Genome Res.* **13**, 2164–2170 (2003).
43. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
44. Notredame, C., Higgins, D.G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
45. Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, M. & Barton, G.J. JPred: a consensus secondary structure prediction server. *Bioinformatics* **14**, 892–893 (1998).
46. Holm, L. & Sander, C. Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.* **20**, 478–480 (1995).
47. Bruno, W.J., Socci, N.D. & Halpern, A.L. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.* **17**, 189–197 (2000).
48. Hasegawa, M., Kishino, H. & Saitou, N. On the maximum likelihood method in molecular phylogenetics. *J. Mol. Evol.* **32**, 443–445 (1991).
49. Adams, W.T. & Skopek, T.R. Statistical test for the comparison of samples from mutational spectra. *J. Mol. Biol.* **194**, 391–396 (1987).