Sequence Variations in *185/333* Messages from the Purple Sea Urchin Suggest Posttranscriptional Modifications to Increase Immune Diversity[1]

Katherine M. Buckley, David P. Terwilliger,[2] and L. Courtney Smith[3]

The *185/333* gene family is highly expressed in two subsets of immune cells in the purple sea urchin in response to immune challenges. The genes encode a surprisingly diverse set of transcripts, which is a function of the variable presence or absence of blocks of shared sequences, known as elements that generate element patterns. Diversity is also the result of a significant level of point mutations. Together, variable element patterns and single nucleotide polymorphisms result in many unique transcripts. The *185/333* genes only have two exons, with the variable element patterns encoded entirely within the second exon. The diversity of the gene family may be the result of frequent recombination among the *185/333* genes that generates a mosaic distribution of element sequences among the genes. A comparative analysis of the sequences for the genes and messages from individual sea urchins indicates that these two sequence sets have largely different nucleotide sequences and appear to use different element patterns. Furthermore, the nucleotide substitution patterns between genes and messages reveal a strong bias toward transitions, particularly cytidine to uridine conversions. These data are consistent with cytidine deaminase activity and may represent a novel form of immunological diversification in an invertebrate immune response system. ***The Journal of Immunology,*** **2008, 181: 8585–8594.**

Molecular diversity among immune response proteins has been observed in nearly all organisms, including vertebrates, invertebrates, and plants (reviewed in Refs. 1, 2). In addition to the well-understood mechanisms in vertebrate somatic gene rearrangement of the Ig gene family (3), novel approaches to increasing immune system diversity have been identified in cyclostomes and a surprising number of invertebrate species that do not have mammal-like adaptive immune functions (2, 4). Variable lymphocyte receptor genes are composed of multiple cassettes of leucine-rich repeats that are somatically assembled through a process that employs short sequence repeats that flank the cassettes (5). This assembly is mediated by a putative AID[4] (activation-induced cytidine deaminase)/APOBEC (apolipoprotein B mRNA editing catalytic component)-like cytidine deaminase (6). The freshwater snail, *Biomphalaria glabrata*, which is an intermediate host for the human parasite *Schistosoma mansoni*, expresses fibrinogen-related proteins (FRePs) in response to parasite infection (7–9). *FReP* genes undergo low-frequency somatic diversification using a small set of source genes and point mutations (10). In amphioxus, five gene families encode the Ig domain-containing variable chitin-binding proteins (VCBP) (11). Each locus has many alleles that yield high levels of protein diversity among individuals. In insects, a novel system exists for generating immune diversity using the Down syndrome cell adhesion molecule (*Dscam*) gene, which also encodes Ig domains (12–14). *Dscam* is a single-copy locus in which 4 of the 24 exons exhibit substantial duplication that results in differential exon arrays. Although all of the exons in the arrays are transcribed, all but one of each exon variant are spliced out through mutually exclusive alternative splicing. The number of exons in the *Dscam* arrays, plus the combinatorial nature of the splicing, yields a potential transcript repertoire of 38,016 unique transcripts from a single gene (15). There are ~16,000 unique *Dscam* transcripts expressed in mosquito hemocytes that encode proteins putatively involved in microbial phagocytosis (12). In higher plants, a large number of disease resistance genes are organized genomically in gene clusters, an arrangement that promotes diversification through sequence exchange via unequal crossovers and gene conversion (1, 16). The presence of environmental or pathogen stressors increases the frequency with which these diversification events occur in plants (17–19). In all, these studies suggest that mechanisms to diversify immune genes have evolved independently in multiple species, underscoring the role that this diversity putatively plays in defending the host against complex pathogen pressures.

The genome of the purple sea urchin, *Strongylocentrotus purpuratus*, has provided insight into immunological diversification mechanisms (20, 21). This species has long been a model for molecular and developmental biology in part based on its phylogenetic position at the base of the deuterostome clade. Sea urchins are members of the echinoderm phylum, which is the sister phylum to chordates. The echinoid class of echinoderms includes the sea urchins and sand dollars, while the other classes include sea cucumbers, brittle stars, sea stars, and crinoids or sea lilies. The close

Department of Biological Sciences, George Washington University, Washington, DC 20052

Received for publication July 22, 2008. Accepted for publication October 13, 2008.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] This research was supported by funding from the National Science Foundation (MCB-0424235) (to L.C.S.) and a Weintraub Fellowship from the George Washington University (to K.M.B.).

[2] Current address: College of Osteopathic Medicine, Touro University, Henderson, NV 89014.

[3] Address correspondence and reprint requests to Dr. L. Courtney Smith, 340 Lisner Hall, 2023 G Street NW, Washington, DC 20052. E-mail address: csmith@gwu.edu

[4] Abbreviations used in this paper: AID, activation-induced cytidine deaminase; APOBEC, apolipoprotein B mRNA editing catalytic component; CR, coding region; EST, expressed sequence tag; gDNA, genomic DNA; FReP, fibrinogen-related protein; NLR, NOD-like receptor; PAMP, pathogen-associated molecular pattern; Polμ, DNA polymerase μ; SNP, single nucleotide polymorphism; SRCR, scavenger receptor cysteine-rich; VCBP, variable chitin-binding proteins.

Table I. *185/333 terminology*

Term	Definition
Coelomocyte	Sea urchin immune cells located primarily in the body cavity or coelom
Diversity score	A measure of the entropy of the alignment; based on the number and frequency of different nucleotides or amino acids at a given position in the alignment
Elements	Blocks of shared sequence found in multiple *185/333* sequences
Element patterns	A specific set of elements
Element subtype	Element Ex15 is composed of element subtypes that differ in length
Intron types α–ϵ	There are five phylogenetic categories of intron sequence
Major element patterns	Common element patterns identified multiple times in different animals
Minor element patterns	Rare element patterns that are only found once, or in single animals
Orphan messages	Messages from which the source gene could not be determined
Percentage of variable positions	The percentage of nonconserved positions in an alignment
Source gene	The gene from which a message was most likely transcribed based on the fewest number of substitutions and insertions/deletions
Unique element sequences	Specific element sequences that are different from all others
Unshared polymorphisms	SNPs present in only a single sequence
Variant element patterns	Element patterns that contain premature stop codons due to a frameshift or point mutation

phylogenetic relationship between echinoderms and chordates makes investigations of the sea urchin immune system evolutionarily relevant. Inspection of the *S. purpuratus* genome has identified a previously unknown level of immune system complexity (20, 21). The genome contains homologs of the complement cascade components (reviewed in Ref. 22), as well as a number of large gene families that encode TLRs, NOD-like receptors (NLRs), and scavenger receptor cysteine-rich (SRCR) gene families (20). In vertebrates, these genes encode proteins that interact with pathogens either directly or indirectly. Many of the sea urchin immune gene families exhibit significant expansion in the *S. purpuratus* genome compared with their vertebrate counterparts. The sea urchin TLR gene family is composed of 222 gene models, compared with between 1 and 20 genes found in genomes of other animals (23). It has been suggested that the multiplicity of this gene family may increase the spectrum of recognizable pathogens (20) or enhance the specificity with which pathogens may be recognized (24). Similarly, there are 203 NLR gene models and 1095 SRCR domains within the *S. purpuratus* genome, as compared with ~20 NLR genes and 81 SRCR domains in humans (20). Thus, the sea urchin, which occupies a unique phylogenetic position as an invertebrate at the base of the deuterostome clade, contains a complex immune system that is characterized in part by a number of large gene families.

The *185/333* gene family is another major component of the purple sea urchin immune response (25–29). Originally identified as 60% of the expressed sequence tags (ESTs) isolated from coelomocytes (immune cells; Table I) following challenge with LPS (27), the *185/333* sequences are very unusual. They are homologous to only two previously uncharacterized sea urchin sequences, and may be specific to echinoids (Refs. 25, 27 and D. A. Raftos, unpublished observation). Optimal alignment of the *185/333* sequences requires the insertion of large gaps, which define blocks of sequence known as "elements". The presence and absence of these elements define "element patterns" (see Fig. 1 and Table I) (27, 29). Elements range in size from 12 to 357 nucleotides, and each element is variable in sequence. Additionally, the *185/333* sequences have six different types of repeats that are present in both tandem and interspersed orientations. The significant diversity in the *185/333* sequences is primarily the result of the element patterns, and secondarily the result of point mutations.

Although *185/333* expression has been observed in immunoquiescent sea urchins, there is a striking increase in expression following immune challenge with either whole bacteria (30), LPS (27), dsRNA, or β-1,3-glucan (28). In response to challenge with various pathogen-associated molecular patterns (PAMPs) and sham-injected injury controls that received artificial coelomic fluid, there is a change in the dominant element patterns of the messages and an increase in the number of minor patterns (28). Furthermore, different suites of 185/333 proteins have been observed by two-dimensional electrophoresis following challenge with LPS vs peptidoglycan (50). These data suggest that the sea urchin may be able to discriminate among different PAMPs by modulating the specific *185/333* genes expressed, although the mechanisms are unknown.

The function of the 185/333 proteins is currently unknown. The predicted proteins have a leader sequence, a glycine-rich region, a conserved RGD motif, a histidine-rich region, and do not contain any cysteine residues (27, 29). The proteins localize to cytoplasmic vesicles of two subsets of coelomocytes, small phagocytes, and polygonal cells, and despite the apparent lack of transmembrane region, they are also present on the surface of small phagocytes (31). The 185/333 proteins are associated with small phagocytes that are present in small cellular clots, which likely lead to syncytia formation (31). Preliminary data suggest that they may bind bacterial cell surfaces (our unpublished observations). Significant increases in protein and transcript expression following immune challenge, the observed diversity in the genes (25, 26), transcripts (28, 29), and proteins and the 185/333 proteins are expressed primarily by the immune cells suggest an immunological role (31). Early work on diverse immune response systems in other invertebrates, such as the VCBPs in protochordates (11) and FRePs in snails (7), similarly did not demonstrate function. However, like the FRePs and VCBPs, the general characteristics of the *185/333* system strongly suggests that that 185/333 proteins are involved in immune function.

The diversity among the *185/333* genes may have been driven by pathogen pressure. Detailed analysis of the repeats suggests that the evolution of the genes and the source of the diversity may be the result of a high rate of gene recombination (25). No recombination hotspots have been identified, and results suggest that recombination may occur at any point along the gene. Molecular clock analysis of the sequences suggests that the current *185/333* genes recombine rapidly and are the result of a recent diversification of the gene family. This frequent recombination generates mosaic or hybrid genes from parent sequences, which are speculated to be the basis for the broad and diversifying 185/333 protein repertoire.

The *185/333* sequences are extremely diverse. From 872 gene and message sequences isolated from 16 animals, 477 unique coding regions (CRs) with 51 distinct element patterns have been identified (26, 28, 29). Previous work, which analyzed messages and genes independently, has shown that all of the genes and half of the messages have CRs with stop codons in one of three positions in the terminal element (26, 29). The remaining *185/333* messages encode truncated proteins due to either a frame shift leading to missense sequence and a premature stop codon or a single nucleotide polymorphism (SNP) that introduces a premature stop codon (28).

In the present study, we have compared the *185/333* genes to the messages from three animals to understand the means by which the *185/333* system diversifies. Notably, we found that, although the repertoires of both genes and messages were diverse (26, 28, 29), very few genes and messages from individual sea urchins had identical sequences. That the gene and message sequences from individual animals differed so strikingly led us to analyze further the differences that characterized these two sets of sequences. Both the sequences of the elements and the types of element patterns in the genes were different from those in the messages. Despite the sequence differences, the gene from which each message was most likely transcribed was estimated based on the fewest number of changes between the two sequences. Results of this analysis suggested that most of the messages were the product of one gene per individual. The rest of the *185/333* genes were either expressed at very low levels or were not expressed in response to the immune challenges that have been employed. Detailed sequence comparisons between the genes and their assumed transcripts indicated a strong preference for cytidine to uridine substitutions, a change that is consistent with cytidine deaminase activity (32). Overall, the data indicate that the *185/333* gene and message sequences are different and suggest an unknown mechanism of posttranscriptional RNA editing that may be a previously unidentified mechanism for generating invertebrate immunological diversity.

Materials and Methods

Sequences and animals

The *185/333* gene sequences from animals designated 2, 4, and 10 employed in this study are available from GenBank (accession nos. EF607618–EF607793) (26). Genes from animal 1 were isolated and sequenced as in Ref. 28 and are reported here with GenBank accession nos. EU401669–EU401677 (supplemental File 1).[5] Genes from animals 1, 2, and 4 were isolated from genomic DNA (gDNA) obtained from coelomocytes; genes from animal 10 were from sperm-derived gDNA (26). Transcript sequences from animals 1, 2, and 4 were described in Ref. 28. The GenBank accession numbers of the *185/333* mRNA sequences can be found in supplemental File 2. From animal 1, *185/333* transcripts were isolated before and after challenge with LPS. *185/333* transcripts were isolated before and after sham injections with artificial coelomic fluid to measure the injury response in animal 4. Animal 2 was the source for *185/333* messages isolated before and after three separate challenges with LPS, β-1,3-glucan, and dsRNA (28).

Sea urchins 1, 2, 4, and 10 were outbred individuals that were obtained and maintained as described in Ref. 33. Briefly, sea urchins were collected from the nearshore waters off the coastline of southern California by SCUBA divers and maintained in the marine aquarium facility at George Washington University. There is an estimated 4% difference between individual *S. purpuratus* animals (34), and it is unlikely that animals 1, 2, 4, and 10 were related. Based on the size of the animals, we estimated that they were likely more than 5 years old, and perhaps as old as 50 years (21, 35).

[5] The online version of this article contains supplemental material.

Bioinformatics

Sequences were aligned according to the cDNA-based alignment (26, 28, 29). BioEdit (36) was used to manipulate and optimize the alignments. PERL scripts were used to calculate diversity scores and diversity characteristics (26, 29). Diversity scores are a measure of entropy (37) and are a function of the number and frequency of nucleotides in each position within the alignment. The maximum possible score occurs given an equal distribution of each of the four nucleotides among the sequences. Diverse sequences have higher diversity scores; conserved sequences receive a diversity score of 0. The diversity score of an alignment is the average score of each of the alignment positions. The gene from which each message was transcribed was estimated using a PERL script. Sequences from each gene and message were compared in a pairwise fashion to calculate the number of changes required to generate each message from each gene. Nucleotide substitutions were scored as 1, and gaps were scored as 10 regardless of the length of the gap. For each message, the gene with the lowest score was considered to be the most likely source. Messages for which the lowest gene score was >50 were considered to be "orphans" with no known source gene. This cutoff was chosen based on a histogram of the lowest gene scores (supplemental File 3).

Results

The *185/333* sequences contain six types of repeats (25, 27) that facilitate at least two different alignments (26, 29). Sequences can be aligned according to the locations of gaps used for the cDNAs, which has been called the cDNA-based alignment (Fig. 1) (29), or based on the locations of the repeats as the primary determinant in defining the element borders (26). Comparisons between the two alignments did not indicate significant differences with regard to the level of sequence diversity, the number of elements or element patterns, or the identification of gene recombination (25). Because the transcripts have been analyzed previously using the cDNA-based alignment (28, 29), that alignment is used here.

Unique sequences

In previous studies, *185/333* genes were cloned from gDNA isolated from three animals (animals 1, 2, and 4) and messages were isolated from the same three animals before and following challenge with various PAMPs (26, 28). The sequences of the genes and mRNAs were characterized separately in previous work (26, 28); however, the results presented here are the first comparison between the genes and messages from the three animals. The sequences were analyzed to determine how many unique sequences (those that did not have identical nucleotide sequence with any other gene or message) were found among the genes and the messages, and how many unique messages were present before and after immune challenge to facilitate subsequent comparisons between the messages and the genes.

The genes and messages from animals 1 and 4 were largely unique, both with respect to messages isolated before vs after challenge and when genes and messages were compared. The majority (62–86%) of the messages collected from animals 1 and 4 before and after immune challenge were unique (Table II). All but two of the *185/333* genes from animals 1 and 4 had unique CRs (4-1520 and 4-1532; supplemental File 1). None of the CRs from the genes isolated from animal 1 matched identically to any of the messages. Animal 4 had CRs from two genes that were identical to six messages (Table II and supplemental File 1) (26). The genes and messages from animals 1 and 4, however, constituted smaller datasets than those from animal 2, which was more exhaustively sampled. It was therefore possible that the lack of identity between the gene and message sequences was due to incomplete sampling rather than a true discord between the sequences. Based on the larger datasets, a more thorough analysis of the genes and messages from animal 2 was undertaken.

Animal 2 was challenged separately with three PAMPs (LPS, β-1,3-glucan, and dsRNA) over a period of about 4 years (28).

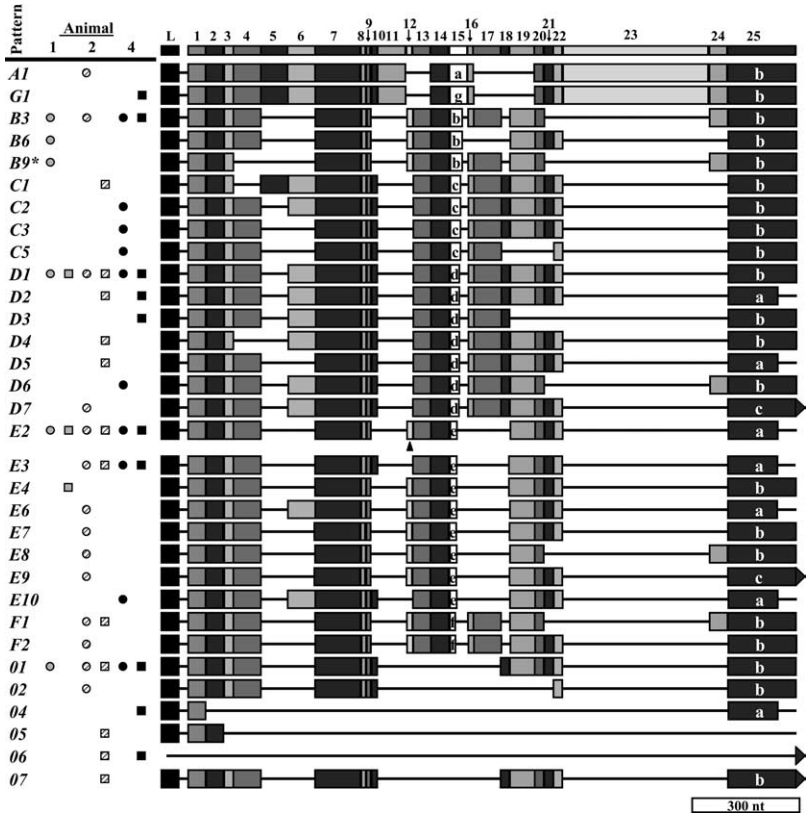


FIGURE 1. Element patterns found in the genes and messages. Element patterns identified in genes are indicated by circles; patterns identified in messages are indicated with squares. The animal (1, 2, 4) from which each element pattern was isolated is indicated. A consensus sequence that includes all possible elements is shown at the top. The arrowhead below element pattern *E2* indicates the position of the stop codon in the element pattern variant *E2.1* (28). The subtype of elements Ex15 and Ex25 is indicated by the letters in the boxes. *B9* is a newly described element pattern that is associated with intron ε (not shown) (26). The intron (not shown) is located between the leader (L) and the first element.

Unique messages constituted 56–95% of the messages sequenced, depending on the immune challenge (Table II). Although no identical messages were found both before and after challenge with the same PAMP, a few identical messages were identified multiple times during the different immune challenges of animal 2. The suites of messages expressed before and after challenge were both equally diverse and were also largely different from one another. There were 51 unique CRs of 53 cloned genes from animal 2 (Tables II and III and supplemental File 1) (26). Comparison between the sequences of the genes and messages indicated that only five sequences were common to both genes and messages (Table III). The extensive analysis of animal 2 indicated that not only was a single animal capable of expressing at least 89 unique *185/333* transcripts under immunoquiescent and postchallenge conditions, but that these transcripts were distinct from the relatively large number of unique alleles characterized from animal 2. This surprising result, that the *185/333* messages generally differed from the genes in individual sea urchins, prompted further investigation into the specific differences that characterized the two sets of sequences.

Table II. *185/333 messages isolated before and after different challenges have different sequences than the* 185/333 *genes*

CR Sequences	Animal 1		Animal 2		Animal 4	
	Unique	Total	Unique	Total	Unique	Total
Genes	9	9	51[a]	53	29	30
Messages						
Pre-LPS	12	15	9	10		
Post-LPS	18	32	12	12		
All LPS	29 (1)[b]	47	21 (0)	22		
Pre-β-1,3-glucan			11	15		
Post-β-1,3-glucan			18	35		
All β-1,3-glucan			28 (1)	50		
Pre-dsRNA			21	32		
Post-dsRNA			27	44		
All dsRNA			45 (3)	76		
Pre-aCF[c]					6	8
Post-aCF					37	42
All aCF					43 (0)	50
Total messages	29	47	89 (5)	148	43	50
Genes plus messages	38	56	135	235	70	80
Sequences shared between genes and messages	0		5		2	

[a] Animal 2 had 51 genes with unique CRs; 2 shared exons but differed in the intron.
[b] Numbers in parentheses indicate the number of sequences that are shared pre- and postchallenge with specific PAMPs or injury.
[c] Artificial coelomic fluid (aCF) was injected as an injury control.

Element patterns

Gene element patterns. Changes in transcript element patterns have been used to infer changes in *185/333* gene expression before

Table III. *Few* 185/333 *genes and messages have identical coding regions*

Animal	Genes[a]	Messages	Element Pattern[b]
2	2-118	CG2-2425[c]	*E2*
	2-119	CG2-2404	
	2-059	CG2-2401	*01*
		CG2-2447	
		LPS2-2403	
	2-028	CG2-2444	*D1*
	2-057		
	2-107	LPS2-2404	*E3*
	2-103	LPS2-2405	*F1*
4	4-1550	4-2419	*E2*
		4-2422	
		4-2426	
		4-2429	
	4-1503	4-2432	*01*
		4-2425	

[a] See supplemental Files 1 and 2 for information on gene and message sequences.
[b] Element patterns correspond to those shown in Fig. 1.
[c] Messages names that start with "CG" were isolated after challenge with dsRNA; messages isolated following challenge with LPS start with "LPS".

Table IV. *Element patterns of the 185/333 genes and messages*

	Animal 1	Animal 2	Animal 4
No. of element patterns (unique genes)	6 (9)	14 (53)	10 (29)
No. of element patterns (unique messages)	3 (29)	12 (89)	10 (43)
No. of different element patterns in genes plus message	7	21	15
Element patterns shared between genes and messages[a]	*D1, E2*	*D1, E2, E3, 01, F1*	*D1, E2, E3, 01, B3*

[a] See Fig. 1 for element patterns.

vs after challenge with PAMPs (28). Similarly, this approach has been used here to compare message element patterns to gene element patterns for three individual sea urchins to determine whether the two data sets only differed in sequences as described above, or whether the differences extended to element patterns. For animal 1, six element patterns were found among nine genes (Table IV) including *E2*, which was the most common element pattern identified from the messages (28, 29), and *D1*, which was most common from the genes (26). In addition, two new element patterns, *B9ε* and *B3δ*, were observed among the genes cloned from animal 1 (Supplemental File 1). The 53 genes from animal 2 had 14 different element patterns, while animal 4 had 10 element patterns from 29 genes (Table IV). There were fewer element patterns than unique sequences because some genes with the same element pattern had different sequences. In total, 21 element patterns were identified from 92 unique *185/333* genes from three animals.

Message element patterns. From the 29 unique messages isolated from animal 1, three element patterns were identified (Table IV). The messages from animal 2 had 12 different element patterns from 89 unique messages, and animal 4 had 10 element patterns from 43 distinct messages. Element patterns *D1* and *E2* were the only patterns shared among all three animals. Similar to the genes, there were many messages that shared element patterns but had different sequences. Although numerous element patterns were identified from genes and messages, it was surprising that very few of the patterns were shared between the two sets of sequences. From animal 1, only patterns *E2* and *D1* were observed in both genes and messages. Five element patterns were shared between genes and messages for animals 2 and 4 (Table IV). The message repertoires were characterized by major and minor patterns (Table I) (28). The major patterns were common to both genes and messages and were present in messages isolated from multiple animals following various antigenic challenges. It has been suggested that certain element patterns may be common among individual sea urchins (e.g., *01, D1,* and *E2*, which have been identified in genes and messages of every animal analyzed) but that each individual may also maintain a repertoire of less common, minor patterns (26, 28). When the frequencies of the element patterns were considered, however, it was apparent that the two groups of sequences were also different. For example, 81% of messages isolated following various Ag challenges were either *E2* or an *E2* variant, as compared with 9% of genes. Therefore, although the major element patterns were represented in both genes and messages, their relative frequencies differed considerably. Furthermore, the minor element patterns were less likely to be shared.

Variant element patterns. Previous analysis of the *185/333* transcripts showed that a number of variant element patterns had premature stop codons (28). The most common example was variant *E2.1*, which had the same elements as *E2* messages, but included a SNP in element Ex10 that altered a histidine codon to a stop. In addition to *E2.1*, 11 other element pattern variants were identified. Half of the element patterns from nine sea urchins had messages with frame shifts and early stop codons that encoded truncated proteins (28). No variant element patterns, including *E2.1*, have been observed among the sequenced *185/333* genes (26), suggesting that variant element patterns are a feature unique to the *185/333* messages.

Sequence diversity of genes and messages

The sequence diversity of the *185/333* system has been well documented (26–29) and is thought to stem, at least in part, from frequent recombination among members of the *185/333* gene family (25). Diversity of genes and messages has been illustrated through 1) entropy-based diversity scores of both the full-length sequences and the individual elements that are calculated using the frequency of each nucleotide present in an alignment position (37), 2) the percentage of variable positions, 3) the number of SNPs that are present only in a single sequence (unshared polymorphisms), and 4) the average number of unique element sequences (26). To characterize the sequence differences between the genes and messages from individual animals, the attributes listed above were calculated for the two datasets separately and in combination for each animal both as complete alignments and as individual elements.

Alignment diversity. Diversity scores and variable positions: Results from these analyses supported the previous observation that full sequences of both the genes and messages from individual animals were diverse (Fig. 2*A*) (26, 28). Diversity scores for the alignments of either genes or messages ranged from 0.0397 to 0.2745 (messages from animals 1 and 4, respectively; Fig. 2*A*). However, when the gene and message sequences were combined into a single alignment, the diversity score was consistently higher than would be expected given the diversity scores of gene and message alignments alone (Fig. 2*A*). Similarly, although between 5.8% and 16.7% of the alignment positions were variable when genes and messages were analyzed as distinct groups (messages from animals 1 and 2, respectively; Fig. 2*B*), when the two groups were combined, the percentage of variable positions increased. These data suggested that, although animals maintained diverse gene and message repertoires, the two sets of sequences were different from one another.

Unshared polymorphisms: The number of SNPs observed in single sequences was calculated for genes and messages from each animal, as well as for sequences from all animals combined, to assess the frequency of random point mutations. An increased frequency of random point mutations present in the messages may suggest nondirected posttranscriptional editing. Results show that there were fewer unshared SNPs per sequence in animal 2 compared with animals 1 and 4 (Fig. 2*B*). However, the low numbers of sequences collected from animals 1 and 4 artificially inflated the numbers of unshared SNPs. Single messages with novel element patterns from animals 1 and 4 were the source of 92% and 75% of the unshared SNPs, respectively. When the outlier messages were removed from the data for animals 1 and 4, reanalysis indicated that results were similar to the results for animal 2 (not shown). When the sequences from animal 2 were analyzed, there was no

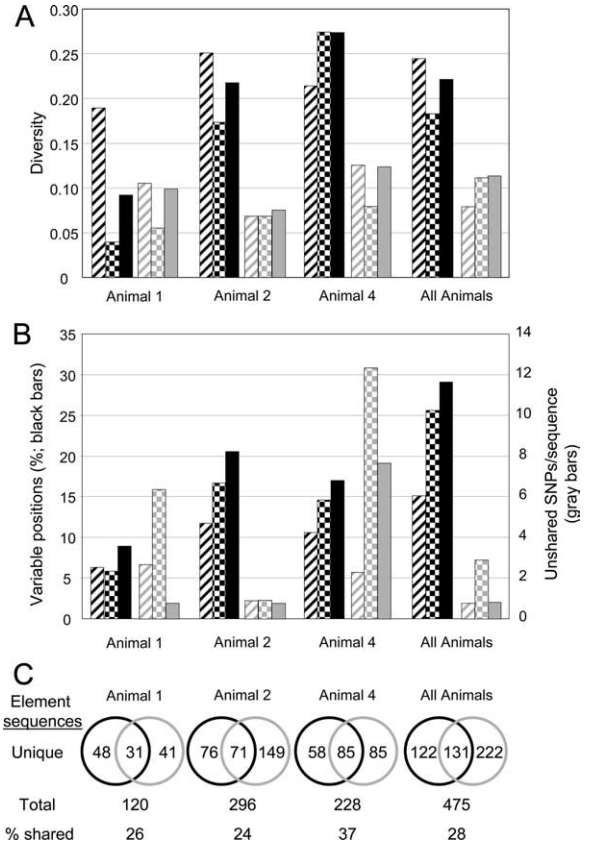


FIGURE 2. Diversity characteristics of the genes and messages. *A*, Diversity of both full-length sequences and elements of genes and messages is similar. Diversity scores (26, 29, 37) were calculated from alignments of the full-length *185/333* genes and messages, as well as the average diversity of individual elements from the genes and messages. Diversity of the full-length gene sequences is indicated by bars with black diagonal lines, messages by bars with black checkered boxes, and the two datasets combined as solid black bars. The average diversity of the elements from the genes is indicated by bars with gray diagonal lines, messages as bars with gray checkered boxes, and the combined data sets as solid gray bars. The *t* tests indicated that the diversity of the elements isolated from genes vs messages was not statistically different ($p > 0.05$). *B*, The number of variable positions and average number of unshared SNPs per sequence are similar among genes and messages. The percentage of positions that are variable is indicated with black bars. The average number of unshared SNPs per sequence is shown with gray bars. Diversity values calculated using the genes only are shown with diagonal lines and messages only are shown with checkered boxes. Diversity characteristics calculated using the gene and message sequences together are indicated with solid bars. The percentage of variable positions or unshared SNPs in genes and messages were not statistically different, as determined using two-tailed *t* tests ($p > 0.05$). *C*, Less than 37% of element sequences are common to *185/333* genes and messages. The total number of unique element sequences from the genes (black circles) and messages (gray circles) and the number of element sequences common to both genes and messages were calculated and are shown in the intersection of the circles. The total number of element sequences and the percentage shared are indicated.

difference in the number of unshared SNPs between the genes and messages. Overall, there was no significant difference in the number of unshared SNPs between genes and messages, suggesting that random posttranscriptional modifications do not appear to be altering the *185/333* messages.

Element diversity. Diversity of individual element sequences was also analyzed to circumvent the problem of the gaps introduced to compensate for the variations in element patterns (26, 29). Diversity scores and the number of unique sequences were calculated for each element (Fig. 2, *A* and *C*). In agreement with previous observations (26, 29), the average diversity scores of the elements were lower than the scores of the full-length alignment because element alignments eliminated the problems introduced by the large gaps required to align full-length sequences. No significant difference was observed between the genes and messages with regard to the diversity scores or number of unique element sequences. When the element sequences from genes and messages were compared, an average of 28% were shared among both genes and messages (Fig. 2*C*). Thus, although some element sequences were shared among genes and messages, variations in element patterns plus variations in element sequences resulted in very few shared full-length sequences.

Table V. *The majority of the genes were likely transcribed from a few genes*

Animal	Source Gene	No. Messages	Pattern[a]	Mean Score
2	2-063	104 (51)[b]	*E2*	17.4[f]
	2-095	2	*E2*	11.5
	2-119	1	*E2*	1.0
	2-090	2	*E9*	28.5
	2-107	5	*E3*	1.4
	2-073	1	*E7*	16.0
	2-052	2	*D1*	5.0
	2-057	1	*D1*	14.0
	2-077	1	*D1*	14.0
	2-065	2	*01*	33.0
	2-059	1	*01*	1.0
	2-103	1	*F1*	2.0
	Too short[c]	4	n/a	n/a
	Exact[d]	8	n/a	0
	Orphan[e]	13	n/a	75.0
4	4-1550	26 (24)	*E2*	7.0[g]
	4-1532	2	*E3*	11.0
	4-1503	2	*01*	2.0
	4-2412	2	*D1*	4.0
	4-1543	2	*D1*	16.0
	4-1501	1	*D1*	5.0
	4-1538	1	*D1*	4.0
	4-2410	1	*D1*	5.0
	4-1548	2 (1)	*B3*	9.0
	Too short[c]	2	n/a	n/a
	Exact[d]	6	n/a	0
	Orphan[e]	3	n/a	96.0
1	1-1517	44 (22)	*E2*	33.7[h]
	1-1502	1	*D1*	3.0
	Orphan[e]	2	n/a	73.0

[a] See Fig. 1.
[b] The numbers in parentheses indicate the number of unique messages. A lack of parenthetical values indicates that all of the messages were unique.
[c] Messages <200 bp were considered to be too short for informative analysis.
[d] Messages that matched exactly to gene sequences; see Table 2.
[e] Scores >50 (see supplemental File 3) were not considered to be viable matches.
[f] The range of the scores for gene 2-063 and its corresponding messages was 15–46 with a SD of 3.9.
[g] The range of the scores for gene 4-1550 and its corresponding messages was 31–45 with a SD of 5.0.
[h] The range of the scores for gene 1-1517 and its corresponding messages was 1–18 with a SD of 2.4.

Matching messages to genes

Although very few of the message sequences matched exactly to genes from within individual animals (Table III), it was of interest to identify from which gene each message was most likely transcribed. To this end, messages were compared in a pairwise manner to genes from each individual animal and scored for substitutions and gaps. For each message, the gene with the lowest score was considered to be the most likely source (Table V). Of the 245 messages analyzed, 221 were most likely the products of only 24

genes (Table V). The remaining 24 messages included 6 that were too short (<200 nucleotides) for meaningful analysis, and 18 for which the optimal source gene had a score of >50 (supplemental File 3), and therefore they were considered to be orphans and derived from genes that had not been cloned and sequenced. For each animal, most of the messages appeared to have been transcribed from a single gene (Table V). For example, 70% of the messages from animal 2 were most similar to a single *E2* gene. Similarly, 54% of the messages from animal 4 and 94% of messages isolated from animal 1 were also likely the products of *E2* genes. From animals 1 and 2, no messages were identical to the genes from which most of the messages were most likely transcribed (genes 1-1517 and 2-063, respectively; Table V). In contrast, four messages from animal 4 that were isolated following sham injection were identical to gene 4-1550 (element pattern *E2*; supplemental File 1), which was the most likely source for most of the messages from animal 4 (Tables II and IV). Previous work predicted that gene expression changed in response to challenge with PAMPs and was the source of sequence changes in the messages (28). However, more detailed analysis of the data presented herein predict that a large majority of the messages isolated both before and after challenge were most likely transcribed from a single gene. This unexpected result, given the gene family size (25, 29) and that was observed in each of the three animals, prompted us to characterize further the differences between the genes and messages.

All of the messages from animal 2 that were most likely transcribed from gene 2-063 were isolated either before or after challenge with either dsRNA or β-1,3-glucan; none was isolated either before or after challenge with LPS. All 10 of the messages collected before LPS challenge were classified as orphans (Table V). Of the 12 messages isolated following LPS challenge, 3 were identical to known *185/333* genes (Table III), 5 were likely the product of gene 2-107, and the remaining 4 were similar in sequence to four different genes (Table V). Most of the messages isolated before and after dsRNA and β-1,3-glucan challenge were derived from gene 2-063 (Table V). Similarly, most of the messages isolated before and after injury from animal 4 and LPS challenge from animal 1 were likely derived from single genes. Therefore, despite the extraordinary diversity observed within both the gene family and transcript repertoire of single animals, it appears that only a small fraction of the genes were highly expressed under the immunological challenges employed.

To better characterize the specific sequence differences between the genes and messages, each message was compared with the gene from which it was most likely transcribed, and the number of each type of nucleotide substitution was counted (Fig. 3). The overwhelming majority of the changes (73%) were transitions (either a purine substituted for a purine or a pyrimidine in place of a pyrimidine). In all, the predicted changes between the genes and messages had a 3:1 transition/transversion ratio. This is much higher than the expected 1:2 ratio that would occur if transitions and transversions (a purine substituted for a pyrimidine, or vice versa) occurred randomly (there are four possible transitions and eight possible transversions). Notably, 30% of the substitutions resulted from a cytidine in the gene and a uridine in the message. This percentage of cytidine to uridine transitions is significantly higher ($p < 0.001$) than expected given typical RNA polymerase activity, and is consistent with the activity of cytidine deaminases (32).

185/333 sequence diversity occurs throughout the length of the sequence; there are no hypervariable regions (26, 29). Likewise, the differences between the genes and messages were located throughout the *185/333* sequence, with no obvious pattern of changes (Fig. 4). However, there were certain positions within the alignment in which most of the messages were different from their

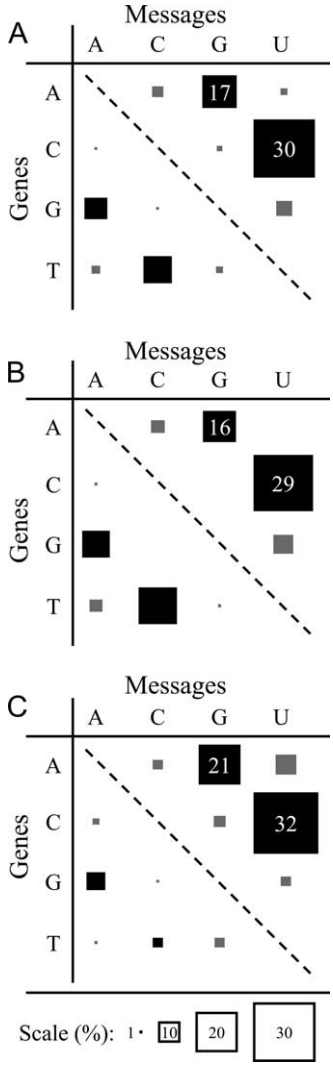


FIGURE 3. Nucleotide differences between the genes and messages suggest a strong cytidine to uridine preference. Messages were compared with the genes from which they were most likely transcribed, and the number of each type of nucleotide substitution was determined. The size of each box indicates the percentage of total substitutions for three analyses: (*A*) when both genes and messages were included, (*B*) when only those messages were included that were likely transcribed from gene 2-063, or (*C*) when only those messages that were not transcribed from gene 2-063 were included. The size of the boxes is scaled to percentage, as shown at the bottom of the figure. Missing boxes indicate that those transitions or transversions did not occur. Gray boxes indicate transversions; black boxes indicate transitions.

corresponding gene (Fig. 4). Three categories of positions showed different numbers of messages with altered nucleotides. The top category included 17 positions in which more than 110 messages (of 207 total) differed from the genes from which they were most likely transcribed. Positions in the medium category were such that between 24 and 50 of the messages differed from their corresponding genes. The bottom category was defined as those positions in which <15 of the messages were different than the genes. Therefore, although differences between the gene and message sequences occurred throughout, certain positions appeared to be altered more frequently than others.

Diversity of sperm- vs coelomocyte-derived genes

Because the differences between genes and messages were heavily biased toward cytidine to uridine transitions, we investigated

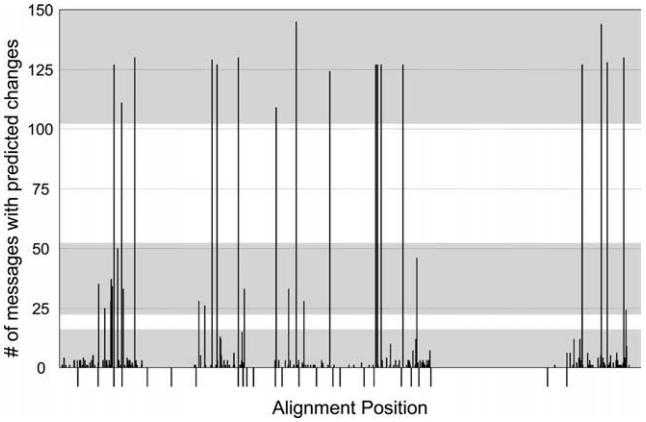


FIGURE 4. Locations of variant nucleotide positions in messages relative to their corresponding gene. For each alignment position, the number of messages with a different nucleotide than the gene from which they were most likely transcribed is indicated by bars. The number of messages with substitutions relative to the corresponding gene for each position could be categorized into three levels: low (<12), medium (25–50), and high (>110), as indicated by the gray shading. The short bars below the *x*-axis indicate the locations of the element borders. A total of 217 variant nucleotide positions were observed. For most of these positions (89 positions), a single message was different from its corresponding gene. However, for 17 positions, >100 messages differed from their corresponding genes.

whether the discrepancy between the gene and message sequences was the result of an unknown mechanism generating somatic diversification that occurred specifically in coelomocytes. *185/333* genes have been analyzed from coelomocyte gDNA (animals 1, 2, and 4), as well as gDNA isolated from sperm (animal 10) (26). The diversity of the genes isolated from these two tissue types was analyzed to address this question (supplemental File 4). No significant differences were observed in the average diversity scores for elements or for full-length sequences, nor for the percentage of variable positions for genes isolated from either tissue source. To understand further the relationship between genes isolated from sperm and coelomocytes, pairs of putative homologous genes were identified via the pairwise comparison technique that was used to correlate gene and message sequences. Analysis of the difference between pairs of genes revealed a slight transition bias (a 5:4 transition/transversion ratio; data not shown) in which most changes were either adenine to guanine or guanine to adenine. The strong preference for cytidine to uridine transitions identified in comparisons between genes and messages (Fig. 4) was not observed when gene repertoires from sperm and coelomocytes, as well as genes from different individuals, were compared. The diversity of the members of the *185/333* gene family was not elevated in genes isolated from coelomocyte gDNA and suggested that a coelomocyte-specific somatic diversification mechanism may not occur.

Discussion

The data presented herein indicate that the *185/333* genes and messages isolated from single animals are largely different and suggest a putative posttranscriptional editing mechanism that favors transition substitutions. In addition to sequence differences, the genes and messages share the same major element patterns, but have different minor element patterns. Although the sequence diversity within the message repertoire is similar to that of the genes, the increased diversity of both full-length and element alignments of genes plus messages compared with separate alignments indicates that the sequences of the genes and messages differ. Despite the large size and diversity of the gene family, it appears that only a few genes are the source of most of the messages. This implies that most of the sequenced *185/333* genes may not be not expressed under the immunological challenges that have been employed to date.

The results presented herein rely on accurate sequences of the *185/333* messages and genes. Because the *185/333* sequences were amplified by PCR before cloning, and were analyzed by cycle sequencing, previous studies have investigated two possible sources of errors that would artificially inflate sequence diversity: *Taq*-induced errors (38) and template switching (39, 40). Given the conditions used to amplify and clone the *185/333* messages, the error rate of the *Taq* polymerases was previously analyzed by amplification, cloning, and sequencing a single *185/333* cDNA, which was determined to be one error per 18 clones (28). Similarly, the error rate for the cloned and sequenced *185/333* genes was determined to be a maximum of one error in 15 clones (26). The error rate of the cycle sequencing method was tested by sequencing a single clone multiple times and was determined to be <0.005% (26). Therefore, it is unlikely that the disparity observed between the gene and message sequences is driven by *Taq*-induced errors alone.

Template switching is another possible source of error, and it is a process by which artificial, recombinant DNA molecules are generated during PCR amplification when multiple homologous templates are present in the reaction (39, 40). Because the *185/333* gene family and transcript repertoire are both composed of pools of very similar sequences, it was important to determine the role that template switching might play during the amplification of these sequences. The conditions for PCR in which the *185/333* genes were originally amplified were mimicked, and the rate of template switching for a pair of clones was assessed using pairs of nested primers and measuring how frequently their orientation with respect to each other switched in the amplicons that were generated (26). Template switching occurred in <1 in 1100 amplicons, which would be undetectable in the ~100 clones isolated from each animal. Consequently, template switching is unlikely to be a source of sequence diversity for the set of *185/333* genes and messages analyzed herein and in previous studies (25). Because the rates of *Taq*-induced misincorporation and template switching are below thresholds that would affect the sequence diversity, this indicates that the observed differences among sequences presented herein are reliable.

The size of the *185/333* gene family has been estimated to be between 80 and 120 alleles using three different approaches: quantitative PCR with gDNA as the template (26), a statistical estimation based on the number of genes sequenced from three animals vs the frequency that duplicate genes were identified (25), and a rough analysis of numbers of *185/333*$^+$ bacterial artificial chromosomes (BACs), including estimates of the average size of the BAC inserts, genome coverage by the BAC library, average size of the *185/333* genes, and distance between clustered genes (our unpublished results). If each allele has a unique sequence, only about half of the alleles from animal 2 have been sequenced (53 unique alleles) (26). It is possible that the genes from which most of the sequenced messages in animal 2 were transcribed have not been cloned, and that this may account for the discord between the gene and message sequences. However, based on the total number of different *185/333* sequences isolated from animal 2, this hypothesis may not be feasible. In addition to the 53 unique genes cloned from animal 2, 89 additional unique transcripts were identified, for a total of 135 unique sequences, which is more than the estimated number of alleles in the *185/333* gene family (25, 29). If about half of the *185/333* genes from animal 2 have been isolated, then there was an ~50% possibility that the sequenced genes were the source

of most of the messages. Given that eight unique *E2* genes were isolated from animal 2 (26), and if the genes were cloned in the same proportion in which they exist in the genome, we can estimate that there may be between 12 and 18 *E2* alleles in an individual animal (based on a gene family size of 80–120 alleles and a range of 40–60% chance that the correct gene was isolated). However, 54 unique transcripts with either *E2* or *E2.1* element patterns were sequenced, which is many more than the estimated number of *E2* alleles. Although posttranscriptional editing could increase the sequence diversity of the messages, it is not likely to change the element patterns of either the genes or the messages. Thus, that more unique *E2* transcripts were identified than the estimated number of *E2* genes within an individual sea urchin supports the hypothesis of posttranscriptional diversification of the messages rather than purely gene-encoded diversity.

Generating the diversity

Cytidine deaminase activity. Analysis of the differences between the messages and the genes from which they were likely transcribed shows a strong bias toward transitions, particularly cytidine to uridine substitutions. This type of change is often the result of deamination reactions that modify cytidine by a hydrolytic deamination at the C4 and C6 positions of the purine base (reviewed in Ref. 41). The most commonly known form of cytidine to uridine editing occurs in the vertebrate mRNA that encodes apolipoprotein B and changes a glutamine (CAA) to a stop (UAA), which truncates the protein (42, 43) (reviewed in Ref. 32). This editing reaction is mediated in part by the APOBEC-1 protein (44), which associates with other proteins, including the APOBEC complementation factor, to form an editing complex known as the editosome (41, 45). Three mammalian proteins are related to APOBEC-1, with the most interesting from the view point of immune function being AID. AID is involved in generating Ig diversity through somatic hypermutation and class switch recombination (reviewed in Ref. 46), and it has also been implicated in the assembly of cyclostome variable lymphocyte receptor genes (6). Additionally, all major lineages of higher plants exhibit cytidine to uridine editing in mitochondrial and chloroplast mRNAs and tRNAs (47). The mechanisms for mRNA editing in plants remain largely unknown (41).

Analysis of the sea urchin genome reveals a few annotated homologs of the RNA editing proteins (21). Several cytidine deaminases have been annotated, but phylogenetic analysis suggests that these proteins are not homologous to either AID or members of the APOBEC enzyme family (20). BLAST searches of the sea urchin EST sequences with the mouse APOBEC-1 sequence (GenBank accession no. NM_031159.3) did not result in any significant matches, suggesting that homologs to APOBEC-1, if they exist within the *S. purpuratus* genome, are not commonly or highly expressed. In contrast, there are sea urchin gene models that encode APOBEC complementation factor-like proteins (SPU_011837 and SPU_011875). Because the RNA editing proteins are small and rapidly evolving, they may be too divergent for phylogenetic analysis of the sequences to suggest a function (20). Given that the *185/333* messages appear to differ from the genes both before and after immune challenge suggests that this diversification may be constitutive, rather than a specific response to immune challenge, as in AID. If constitutive RNA editing is involved in this system, this suggests that a non-AID-like deaminase may be involved. Thus, although some of the mRNA editing components have been identified in the sea urchin, it is not clear whether any have deaminase-like function.

Low-fidelity polymerases. In addition to possible RNA editing activity, other proteins may be involved in diversification of the *185/333* messages during transcription. One such enzyme identified in the sea urchin genome is the TdT/polymerase μ (TdT/Polμ; SPU_009980) homolog (20, 24). In higher vertebrates, TdT/Polμ is involved in Ig diversification and has been implicated in low-fidelity DNA replication (48, 49). Although no significant matches to TdT/Polμ have been identified among the currently available sea urchin ESTs, expression may be restricted to specific times or tissues or may be below detection.

The diversity of *185/333* genes isolated from coelomocyte gDNA is comparable to the diversity of sperm-derived *185/333* genes, which suggests that deaminases do not somatically diversify the *185/333* genes. Despite the fact that variability among *S. purpuratus* individuals is estimated to be 4% (21, 34), comparison of the genes isolated from sperm vs coelomocyte gDNA from different animals did not reveal a cytidine to uridine transition bias similar to that observed between the genes and messages from single animals. Based on the data presented herein, message diversification appears to be the result of an unknown posttranscriptional RNA editing mechanism that, in many cases, alters specific nucleotides that change specific codons resulting either in the incorporation of a different amino acid into the encoded protein or in a stop. Changes also include small insertions/deletions that alter the reading frame and result in missense amino acid incorporation typically leading to a stop. It is of note that half of the messages encode truncated proteins (28) of which at least some are expressed (Dheilly et al., submitted), suggesting that the benefits of diversification through this method may be limited by the introduction of premature stop codons that may alter protein function. Alternatively, given that as many as 55% of the *E2* messages are altered to the *E2.1* variant with a single early stop codon (28), it is possible that the N-terminal, glycine-rich region of the truncated protein may have a function that is independent of the histidine-rich C terminus.

Conclusions: two levels of diversity

The diversity of the *185/333* genes may be, in part, the result of frequent recombination among the genes (25). This recombination may be facilitated by a variety of repeats within the genes, by di- and tri-nucleotide repeats that flank the genes and are present in the short intergenic region, and by the close linkage among the genes within the genome (26). This putative propensity for recombination has been proposed as a major diversification system acting within the *185/333* gene family to generate a large number of unique genes from a limited number of element sequences. As a result, the members of the *185/333* gene family have a mosaic or patchwork appearance of elements. It is therefore unexpected that a second level of diversification may be involved in this system in the form of edited messages (and perhaps poor fidelity transcription), and that, despite maintaining a large and diverse gene family, only a few of the genes may be transcribed. Future analysis of the *cis*-promoters, predictions of binding sites for transcription factors, and the regulation of *185/333* gene expression may show that most of the genes may not be expressed, but may only function as a source for gene recombination. The *185/333* system in the sea urchin therefore represents an intriguing example of invertebrate immunological diversity that appears to be generated not only at the genomic level, but also posttranscriptionally by acting on the mRNA sequences.

Acknowledgments

We thank Dr. David Raftos and the anonymous reviewers for input that resulted in improvements to this manuscript.

Disclosures

The authors have no financial conflicts of interest.

References

1. McDowell, J. M., and S. A. Simon. 2008. Molecular diversity at the plant-pathogen interface. *Dev. Comp. Immunol.* 32: 736–744.
2. Pancer, Z., and M. D. Cooper. 2006. The evolution of adaptive immunity. *Annu. Rev. Immunol.* 24: 497–518.
3. Tonegawa, S. 1983. Somatic generation of antibody diversity. *Nature* 302: 575–581.
4. Flajnik, M. F., and L. Du Pasquier. 2004. Evolution of innate and adaptive immunity: can we draw a line? *Trends Immunol.* 25: 640–644.
5. Nagawa, F., N. Kishishita, K. Shimizu, S. Hirose, M. Miyoshi, J. Nezu, T. Nishimura, H. Nishizumi, Y. Takahashi, S. Hashimoto, et al. 2007. Antigen-receptor genes of the agnathan lamprey are assembled by a process involving copy choice. *Nat. Immunol.* 8: 206–213.
6. Rogozin, I. B., L. M. Iyer, L. Liang, G. V. Glazko, V. G. Liston, Y. I. Pavlov, L. Aravind, and Z. Pancer. 2007. Evolution and diversification of lamprey antigen receptors: evidence for involvement of an AID-APOBEC family cytosine deaminase. *Nat. Immunol.* 8: 647–656.
7. Adema, C. M., L. A. Hertel, R. D. Miller, and E. S. Loker. 1997. A family of fibrinogen-related proteins that precipitates parasite-derived molecules is produced by an invertebrate after infection. *Proc. Natl. Acad. Sci. USA* 94: 8691–8696.
8. Leonard, P. M., C. M. Adema, S. M. Zhang, and E. S. Loker. 2001. Structure of two *FREP* genes that combine IgSF and fibrinogen domains, with comments on diversity of the *FREP* gene family in the snail *Biomphalaria glabrata. Gene* 269: 155–165.
9. Zhang, S. M., P. M. Leonard, C. M. Adema, and E. S. Loker. 2001. Parasite-responsive IgSF members in the snail *Biomphalaria glabrata*: characterization of novel genes with tandemly arranged IgSF domains and a fibrinogen domain. *Immunogenetics* 53: 684–694.
10. Zhang, S. M., C. M. Adema, T. B. Kepler, and E. S. Loker. 2004. Diversification of Ig superfamily genes in an invertebrate. *Science* 305: 251–254.
11. Cannon, J. P., R. N. Haire, and G. W. Litman. 2002. Identification of diversified genes that contain immunoglobulin-like variable regions in a protochordate. *Nat. Immunol.* 3: 1200–1207.
12. Dong, Y., H. E. Taylor, and G. Dimopoulos. 2006. AgDscam, a hypervariable immunoglobulin domain-containing receptor of the *Anopheles gambiae* innate immune system. *PLoS Biol.* 4: e229.
13. Graveley, B. R., A. Kaur, D. Gunning, S. L. Zipursky, L. Rowen, and J. C. Clemens. 2004. The organization and evolution of the dipteran and hymenopteran Down syndrome cell adhesion molecule (*Dscam*) genes. *RNA* 10: 1499–1506.
14. Schmucker, D., J. C. Clemens, H. Shu, C. A. Worby, J. Xiao, M. Muda, J. E. Dixon, and S. L. Zipursky. 2000. *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* 101: 671–684.
15. Watson, F. L., R. Puttmann-Holgado, F. Thomas, D. L. Lamar, M. Hughes, M. Kondo, V. I. Rebel, and D. Schmucker. 2005. Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science* 309: 1874–1878.
16. Meyers, B. C., A. Kozik, A. Griego, H. Kuang, and R. W. Michelmore. 2003. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell* 15: 809–834.
17. Boyko, A., P. Kathiria, F. J. Zemp, Y. Yao, I. Pogribny, and I. Kovalchuk. 2007. Transgenerational changes in the genome stability and methylation in pathogen-infected plants: (virus-induced plant genome instability). *Nucleic Acids Res.* 35: 1714–1725.
18. Kovalchuk, I., O. Kovalchuk, V. Kalck, V. Boyko, J. Filkowski, M. Heinlein, and B. Hohn. 2003. Pathogen-induced systemic plant signal triggers DNA rearrangements. *Nature* 423: 760–762.
19. Lucht, J. M., B. Mauch-Mani, H. Y. Steiner, J. P. Metraux, J. Ryals, and B. Hohn. 2002. Pathogen stress increases somatic recombination frequency in *Arabidopsis*. *Nat. Genet.* 30: 311–314.
20. Hibino, T., M. L. Coll, C. Messier, A. C. Majeske, D. P. Terwilliger, K. M. Buckley, V. Brockton, S. Nair, K. Berney, S. D. Fugmann, et al. 2006. The immune gene repertoire encoded in the purple sea urchin genome. *Dev. Biol.* 300: 349–365.
21. Sodergren, E., G. M. Weinstock, E. H. Davidson, R. A. Cameron, R. A. Gibbs, R. C. Angerer, L. M. Angerer, M. I. Arnone, D. R. Burgess, R. D. Burke, et al. 2006. The genome of the sea urchin *Strongylocentrotus purpuratus. Science* 314: 941–952.
22. Smith, L. C., L. A. Clow, and D. P. Terwilliger. 2001. The ancestral complement system in sea urchins. *Immunol. Rev.* 180: 16–34.
23. Roach, J. C., G. Glusman, L. Rowen, A. Kaur, M. K. Purcell, K. D. Smith, L. E. Hood, and A. Aderem. 2005. The evolution of vertebrate Toll-like receptors. *Proc. Natl. Acad. Sci. USA* 102: 9577–9582.
24. Rast, J. P., L. C. Smith, M. Loza-Coll, T. Hibino, and G. W. Litman. 2006. Genomic insights into the immune system of the sea urchin. *Science* 314: 952–956.
25. Buckley, K. M., S. Munshaw, T. B. Kepler, and L. C. Smith. 2008. The *185/333* gene family is a rapidly diversifying host-defense gene cluster in the purple sea urchin. *Strongylocentrotus purpuratus. J. Mol. Biol.* 379: 912–928.
26. Buckley, K. M., and L. C. Smith. 2007. Extraordinary diversity among members of the large gene family, *185/333*, from the purple sea urchin, *Strongylocentrotus purpuratus. BMC Mol. Biol.* 8: 68.
27. Nair, S. V., H. Del Valle, P. S. Gross, D. P. Terwilliger, and L. C. Smith. 2005. Macroarray analysis of coelomocyte gene expression in response to LPS in the sea urchin: identification of unexpected immune diversity in an invertebrate. *Physiol. Genomics* 22: 33–47.
28. Terwilliger, D. P., K. M. Buckley, V. Brockton, N. J. Ritter, and L. C. Smith. 2007. Distinctive expression patterns of *185/333* genes in the purple sea urchin, *Strongylocentrotus purpuratus*: an unexpectedly diverse family of transcripts in response to LPS, β-1,3-glucan, and dsRNA. *BMC Mol. Biol.* 8: 16.
29. Terwilliger, D. P., K. M. Buckley, D. Mehta, P. G. Moorjani, and L. C. Smith. 2006. Unexpected diversity displayed in cDNAs expressed by the immune cells of the purple sea urchin, *Strongylocentrotus purpuratus. Physiol. Genomics* 26: 134–144.
30. Rast, J. P., Z. Pancer, and E. H. Davidson. 2000. New approaches towards an understanding of deuterostome immunity. *Curr. Top. Microbiol. Immunol.* 248: 3–16.
31. Brockton, V., J. H. Henson, D. A. Raftos, A. J. Majeske, Y. O. Kim, and L. C. Smith. 2008. Localization and diversity of 185/333 proteins from the purple sea urchin: unexpected protein-size range and protein expression in a new coelomocyte type. *J. Cell Sci.* 121: 339–348.
32. Chester, A., J. Scott, S. Anant, and N. Navaratnam. 2000. RNA editing: cytidine to uridine conversion in apolipoprotein B mRNA. *Biochim. Biophys. Acta* 1494: 1–13.
33. Gross, P. S., L. A. Clow, and L. C. Smith. 2000. SpC3, the complement homologue from the purple sea urchin, *Strongylocentrotus purpuratus*, is expressed in two subpopulations of the phagocytic coelomocytes. *Immunogenetics* 51: 1034–1044.
34. Britten, R. J., A. Cetta, and E. H. Davidson. 1978. The single-copy DNA sequence polymorphism of the sea urchin *Strongylocentrotus purpuratus. Cell* 15: 1175–1186.
35. Ebert, T. A. 1967. Negative growth and longevity in the purple sea urchin *Strongylocentrotus purpuratus* (Stimpson). *Science* 157: 557–558.
36. Hall, T. A. 1999. BioEdit: a user friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* 41: 95–98.
37. Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. 1998. *Biological Sequence Analysis, Probability Models of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge, U.K.
38. Cline, J., J. C. Braman, and H. H. Hogrefe. 1996. PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res.* 24: 3546–3551.
39. Odelberg, S. J., R. B. Weiss, A. Hata, and R. White. 1995. Template-switching during DNA synthesis by *Thermus aquaticus* DNA polymerase I. *Nucleic Acids Res.* 23: 2049–2057.
40. Thompson, J. R., L. A. Marcelino, and M. F. Polz. 2002. Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by "reconditioning PCR". *Nucleic Acids Res.* 30: 2083–2088.
41. Gott, J. M., and R. B. Emeson. 2000. Functions and mechanisms of RNA editing. *Annu. Rev. Genet.* 34: 499–531.
42. Chen, S. H., G. Habib, C. Y. Yang, Z. W. Gu, B. R. Lee, S. A. Weng, S. R. Silberman, S. J. Cai, J. P. Deslypere, M. Rosseneu, et al. 1987. Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. *Science* 238: 363–366.
43. Powell, L. M., S. C. Wallis, R. J. Pease, Y. H. Edwards, T. J. Knott, and J. Scott. 1987. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell* 50: 831–840.
44. Davidson, N. O., T. L. Innerarity, J. Scott, H. Smith, D. M. Driscoll, B. Teng, and L. Chan. 1995. Proposed nomenclature for the catalytic subunit of the mammalian apolipoprotein B mRNA editing enzyme: APOBEC-1. *RNA* 1: 3.
45. Backus, J. W., and H. C. Smith. 1992. Three distinct RNA sequence elements are required for efficient apolipoprotein B (apoB) RNA editing in vitro. *Nucleic Acids Res.* 20: 6007–6014.
46. Durandy, A. 2003. Activation-induced cytidine deaminase: a dual role in class-switch recombination and somatic hypermutation. *Eur. J. Immunol.* 33: 2069–2073.
47. Freyer, R., M. C. Kiefer-Meyer, and H. Kossel. 1997. Occurrence of plastid RNA editing in all major lineages of land plants. *Proc. Natl. Acad. Sci. USA* 94: 6285–6290.
48. Kunkel, T. A., and K. Bebenek. 2000. DNA replication fidelity. *Annu. Rev. Biochem.* 69: 497–529.
49. Ruiz, J. F., O. Dominguez, T. Lain de Lera, M. Garcia-Diaz, A. Bernad, and L. Blanco. 2001. DNA polymerase μ, a candidate hypermutase? *Philos. Trans. R. Soc. London B Biol. Sci.* 356: 99–109.
50. Dheilly, N. M., S. V. Nair, L. C. Smith, and D. A. Raftos. Highly variable immune response: proteins from the sea urchin, *Strongylocentrotus purpuratus*: proteomic analysis of diversity within and between individuals. *J. Immunol.* In press.

Supplemental File 1: *185/333* genes used in the comparison of genes and messages

Genbank Accession Number	Clone Number	Element Pattern	Animal
EU401670	1-1502	*D1*α	1
EU401671	1-1503	*01*δ	1
EU401672	1-1506	*E2*δ	1
EU401673	1-1517	*E2*δ	1
EU401674	1-1522	*B3*δ	1
EU401675	1-1526	*D1*α	1
EU401676	1-1527	*D1*α	1
EU401677	1-1538	*B6*ε	1
EU401669	1-1540	*B9*ε	1
EF607711	2-028	*D1*α	2
EF607728	2-052	*D1*α	2
EF607731	2-057	*D1*α	2
EF607732	2-059	*01*δ	2
EF607734	2-063	*E2*δ	2
EF607736	2-065	*01*δ	2
EF607742	2-073	*E7*δ	2
EF607744	2-077	*D1*α	2
EF607752	2-090	*E9*δ	2
EF607756	2-095	*E2*δ	2
EF607682	2-103	*F1*	2
EF607686	2-107	*E3*δ	2
EF607695	2-118	*E2*δ	2
EF607696	2-119	*E2*δ	2
EF607760	4-11501	*D1*α	4
EF607761	4-11503	*01*δ	4
EF607767	4-11520	*E3*δ	4
EF607773	4-11532	*E3*δ	4
EF607778	4-11538	*D1*α	4
EF607780	4-11543	*D1*α	4
EF607784	4-11548	*B3*β	4
EF607785	4-11550	*E2*δ	4
EF607786	4-12410	*D1*α	4
EF607787	4-12412	*D1*α	4

Table 1: Characteristics of the *185/333* mRNA sequences used in the analysis.

Accession	Clone #	Pattern	Length	Animal	Time*	Most Likely Gene
EF066304	1-1504	E2	935	1	Pre-challenge	1-1517
EF066307	1-1505	E2.1	894	1	Pre-challenge	1-1517
EF066305	1-1512	E2.1	935	1	Pre-challenge	1-1517
EF066303	1-1514	E2.1	934	1	Pre-challenge	1-1517
EF066301	1-1515	E2	934	1	Pre-challenge	1-1517
EF066298	1-1523	E2	934	1	Pre-challenge	1-1517
EF066299	1-1528	E2.1	934	1	Pre-challenge	1-1517
EF066309	1-1532	E2.1	934	1	Pre-challenge	1-1517
EF066327	1-1533	E2	934	1	Pre-challenge	1-1517
EF066300	1-1534	E2.1	934	1	Pre-challenge	1-1517
EF066302	1-1535	E2.1	934	1	Pre-challenge	1-1517
EF066324	1-1536	E2.1	934	1	Pre-challenge	1-1517
EF066308	1-1539	E2.1	934	1	Pre-challenge	1-1517
EF066306	1-1547	E2.1	935	1	Pre-challenge	1-1517
EF066323	1-1549	E2.1	934	1	Pre-challenge	1-1517
EF066267	1-2402	E2	935	1	Post LPS	1-1517
EF066232	1-2404	E2	935	1	Post LPS	1-1517
EF066260	1-2405	E2	935	1	Post LPS	1-1517
EF066262	1-2406	E2	935	1	Post LPS	1-1517
EF066247	1-2407	E2	935	1	Post LPS	1-1517
EF066253	1-2412	E4	920	1	Post LPS	1-1517
EF066251	1-2413	E2	935	1	Post LPS	1-1517
EF066277	1-2414	D1	1136	1	Post LPS	1-1502
EF066237	1-2416	E2	935	1	Post LPS	1-1517
EF066248	1-2417	E2	935	1	Post LPS	1-1517
EF066264	1-2418	E2	935	1	Post LPS	1-1517
EF066246	1-2420	E2	935	1	Post LPS	1-1517
EF066243	1-2421	E2	935	1	Post LPS	1-1517
EF066254	1-2422	E2	935	1	Post LPS	1-1517
EF066282	1-2424	E5.1	846	1	Post LPS	Orphan
EF066265	1-2425	E2	935	1	Post LPS	1-1517
EF066255	1-2426	E2	935	1	Post LPS	1-1517
EF066259	1-2427	E2	935	1	Post LPS	1-1517
EF066256	1-2428	E2	935	1	Post LPS	1-1517
EF066238	1-2429	E5.1	846	1	Post LPS	Orphan
EF066258	1-2430	E2	935	1	Post LPS	1-1517
EF066249	1-2431	E2	935	1	Post LPS	1-1517
EF066252	1-2432	E2	935	1	Post LPS	1-1517
EF066257	1-2433	E2	935	1	Post LPS	1-1517
EF066241	1-2434	E2	935	1	Post LPS	1-1517
EF066235	1-2435	E2	935	1	Post LPS	1-1517
EF066234	1-2436	E2	935	1	Post LPS	1-1517

Accession	Clone #	Pattern	Length	Animal	Time*	Most Likely Gene
EF066263	1-2437	E2	935	1	Post LPS	1-1517
EF066250	1-2439	E2	935	1	Post LPS	1-1517
EF066236	1-2440	E2	935	1	Post LPS	1-1517
EF066266	1-2441	E2	894	1	Post LPS	1-1517
EF066242	1-2442	E2	935	1	Post LPS	1-1517
EF066294	2-1501	C1	1151	2	Pre-challenge	Orphan
EF066288	2-1502	C1	1151	2	Pre-challenge	Orphan
EF066285	2-1505	C1	1151	2	Pre-challenge	Orphan
EF066286	2-1506	C1	1151	2	Pre-challenge	Orphan
EF066322	2-1507	C1	1151	2	Pre-challenge	Orphan
EF066291	2-1508	C1	1151	2	Pre-challenge	Orphan
EF066289	2-1509	C1	1151	2	Pre-challenge	Orphan
EF066326	2-1510	C1	1151	2	Pre-challenge	Orphan
EF066297	2-1511	C1	1151	2	Pre-challenge	Orphan
EF066287	2-1514	C1	1151	2	Pre-challenge	Orphan
EF066273	2-2401	E3	932	2	Post LPS	2-107
EF066261	2-2403	01	830	2	Post LPS	**2-059**
EF066268	2-2404	E3	932	2	Post LPS	**2-107**
EF066228	2-2405	F1	1031	2	Post LPS	**2-103**
EF066275	2-2406	E2	917	2	Post LPS	2-119
EF066272	2-2407	E2	932	2	Post LPS	2-107
EF066271	2-2408	E2	932	2	Post LPS	2-107
EF066274	2-2409	E2	930	2	Post LPS	2-107
EF066278	2-2411	E2	932	2	Post LPS	2-107
EF066279	2-2413	F1.1	1031	2	Post LPS	2-103
EF066284	2-2414	D2.1	1134	2	Post LPS	2-077
EF066230	2-2415	D2	1142	2	Post LPS	2-057
EF066210	2-1502	E2	935	2	Pre-challenge	2-063
EF066208	2-1503	E2.1	935	2	Pre-challenge	2-063
EF066213	2-1509	E2.1	934	2	Pre-challenge	2-063
EF066218	2-1511	E2	935	2	Pre-challenge	2-063
EF066207	2-1513	E2.1	935	2	Pre-challenge	2-063
EF066216	2-1518	E2	934	2	Pre-challenge	2-063
EF066220	2-1519	E2	935	2	Pre-challenge	2-063
EF066211	2-1523	E2.1	935	2	Pre-challenge	2-063
EF066215	2-1524	E2.1	934	2	Pre-challenge	2-063
EF066209	2-1531	E2.1	933	2	Pre-challenge	2-063
EF066212	2-1533	E2.1	935	2	Pre-challenge	2-063
EF066222	2-1536	E2.1	935	2	Pre-challenge	2-063
EF066214	2-1540	E2.3	934	2	Pre-challenge	Orphan
EF066217	2-1546	E2	935	2	Pre-challenge	2-063
EF066219	2-1548	E2	935	2	Pre-challenge	2-063
EF066143	2-2403	E2	934	2	Post Lam	2-063
EF066149	2-2404	E2	935	2	Post Lam	2-063
EF066159	2-2405	E2	932	2	Post Lam	2-063

Accession	Clone #	Pattern	Length	Animal	Time*	Most Likely Gene
EF066147	2-2406	E2	933	2	Post Lam	2-063
EF066110	2-2409	E2	933	2	Post Lam	2-063
EF066140	2-2410	E2	934	2	Post Lam	2-063
EF066146	2-2411	E2	935	2	Post Lam	2-063
EF066053	2-2412	E2	935	2	Post Lam	2-063
EF066090	2-2413	E2	935	2	Post Lam	2-063
EF066148	2-2415	E2	935	2	Post Lam	2-063
EF066096	2-2416	E2	935	2	Post Lam	2-063
EF066108	2-2417	E2	935	2	Post Lam	2-063
EF066058	2-2418	E2	935	2	Post Lam	2-063
EF066095	2-2419	E2	919	2	Post Lam	2-095
EF066151	2-2420	E2	935	2	Post Lam	2-063
EF066100	2-2421	E2	935	2	Post Lam	2-063
EF066062	2-2422	E2	934	2	Post Lam	2-063
EF066043	2-2423	E2.1	934	2	Post Lam	2-063
EF066156	2-2424	E2	934	2	Post Lam	2-063
EF066102	2-2425	E2	935	2	Post Lam	2-063
EF066145	2-2426	E2	933	2	Post Lam	2-063
EF066138	2-2427	E2	934	2	Post Lam	2-063
EF066154	2-2430	E2	934	2	Post Lam	2-063
EF066133	2-2431	E2	935	2	Post Lam	2-063
EF066069	2-2432	E2	934	2	Post Lam	2-063
EF066155	2-2434	E2	934	2	Post Lam	2-063
EF066044	2-2436	E2.1	934	2	Post Lam	2-063
EF066144	2-2437	E2	935	2	Post Lam	2-063
EF066072	2-2438	E2	934	2	Post Lam	2-063
EF066153	2-2439	E2	935	2	Post Lam	2-063
EF066142	2-2440	E2	934	2	Post Lam	2-063
EF066157	2-2442	E2	934	2	Post Lam	2-063
EF066160	2-2445	E2	935	2	Post Lam	2-063
EF066150	2-2446	E2	935	2	Post Lam	2-063
EF066086	2-2448	E2	935	2	Post Lam	2-063
EF065967	2-1501	E2.1	934	2	Pre-challenge	2-063
EF065968	2-1503	E2.1	934	2	Pre-challenge	2-063
EF065969	2-1505	E2.1	934	2	Pre-challenge	2-063
EF065989	2-1506	E2	935	2	Pre-challenge	2-063
EF066033	2-1507	E2	934	2	Pre-challenge	2-063
EF065970	2-1508	E2.1	934	2	Pre-challenge	2-063
EF065971	2-1509	E2.1	934	2	Pre-challenge	2-063
EF065972	2-1510	E2.1	933	2	Pre-challenge	2-063
EF066034	2-1511	E2	934	2	Pre-challenge	2-063
EF065992	2-1514	E2	919	2	Pre-challenge	2-095
EF065993	2-1516	E2	934	2	Pre-challenge	2-063
EF065966	2-1517	E2.1	931	2	Pre-challenge	2-063
EF065973	2-1519	E2.1	933	2	Pre-challenge	2-063

Accession	Clone #	Pattern	Length	Animal	Time*	Most Likely Gene
EF065974	2-1521	E2.1	934	2	Pre-challenge	2-063
EF065975	2-1522	E2.1	934	2	Pre-challenge	2-063
EF065976	2-1523	E2.1	934	2	Pre-challenge	2-063
EF065977	2-1527	E2.1	934	2	Pre-challenge	2-063
EF065994	2-1528	E2	934	2	Pre-challenge	2-063
EF066035	2-1529	C1	1150	2	Pre-challenge	Orphan
EF065995	2-1532	E2	929	2	Pre-challenge	2-090
EF065978	2-1533	E2.1	934	2	Pre-challenge	2-063
EF065979	2-1535	E2.1	934	2	Pre-challenge	2-063
EF065980	2-1536	E2.1	934	2	Pre-challenge	2-063
EF066039	2-1537	E2.1	934	2	Pre-challenge	2-063
EF065982	2-1538	E2.1	934	2	Pre-challenge	2-063
EF065983	2-1539	E2.1	934	2	Pre-challenge	2-063
EF066027	2-1540	D1	1135	2	Pre-challenge	2-052
EF066032	2-1541	E2	934	2	Pre-challenge	2-063
EF065997	2-1542	E2	934	2	Pre-challenge	2-063
EF065984	2-1543	E2.1	934	2	Pre-challenge	2-063
EF065985	2-1546	E2.1	934	2	Pre-challenge	2-063
EF065986	2-1547	E2.1	934	2	Pre-challenge	2-063
EF065878	2-2401	07	830	2	Post dsRNA	**2-059**
EF065865	2-2403	E2	934	2	Post dsRNA	2-063
EF065910	2-2404	E2	916	2	Post dsRNA	**2-119**
EF065836	2-2405	E2.1	934	2	Post dsRNA	2-063
EF065837	2-2406	E2.1	934	2	Post dsRNA	2-063
EF065838	2-2407	E2.1	934	2	Post dsRNA	2-063
EF065839	2-2408	E2.1	934	2	Post dsRNA	2-063
EF065917	2-2409	D1	1135	2	Post dsRNA	2-052
EF065840	2-2410	E2.1	934	2	Post dsRNA	2-063
EF065855	2-2411	05	135	2	Post dsRNA	2-082
EF065841	2-2412	E2.1	934	2	Post dsRNA	2-063
EF065842	2-2413	E2.1	934	2	Post dsRNA	2-063
EF065864	2-2414	E2	934	2	Post dsRNA	2-063
EF065856	2-2415	05	135	2	Post dsRNA	2-082
EF065916	2-2416	E2.5	934	2	Post dsRNA	2-063
EF065907	2-2418	E2.6	930	2	Post dsRNA	Orphan
EF065843	2-2419	E2.1	934	2	Post dsRNA	2-063
EF065844	2-2420	E2.1	934	2	Post dsRNA	2-063
EF065860	2-2421	D4	1063	2	Post dsRNA	2-065
EF065845	2-2422	E2.1	934	2	Post dsRNA	2-063
EF065914	2-2423	06	79	2	Post dsRNA	2-063
EF065846	2-2424	E2.1	919	2	Post dsRNA	2-090
EF065906	2-2425	E2	916	2	Post dsRNA	**2-119**
EF065868	2-2426	E2	934	2	Post dsRNA	2-063
EF065847	2-2427	E2.1	934	2	Post dsRNA	2-063
EF065848	2-2428	E2.1	934	2	Post dsRNA	2-063

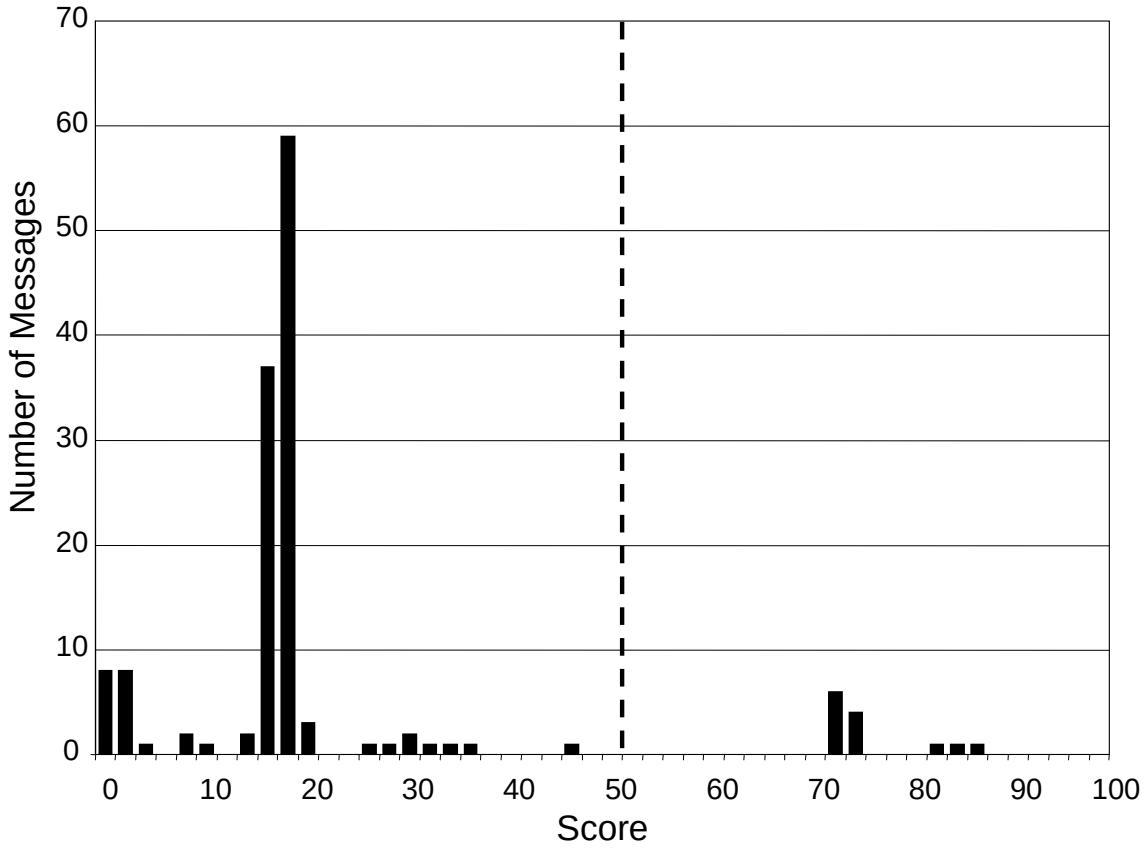
Accession	Clone #	Pattern	Length	Animal	Time*	Most Likely Gene
EF065849	2-2429	E2.1	934	2	Post dsRNA	2-063
EF065829	2-2430	E2	934	2	Post dsRNA	2-063
EF065850	2-2431	E2.1	934	2	Post dsRNA	2-063
EF065911	2-2432	E2	931	2	Post dsRNA	2-063
EF065876	2-2435	07	829	2	Post dsRNA	2-059
EF065851	2-2436	E2.1	934	2	Post dsRNA	2-063
EF065908	2-2437	E2	931	2	Post dsRNA	2-073
EF065861	2-2438	D5	1068	2	Post dsRNA	2-065
EF065832	2-2439	E2	934	2	Post dsRNA	2-063
EF065857	2-2440	05	135	2	Post dsRNA	2-082
EF065918	2-2441	E2.1	934	2	Post dsRNA	2-063
EF065833	2-2442	E2	934	2	Post dsRNA	2-063
EF065874	2-2443	E2	934	2	Post dsRNA	2-063
EF065863	2-2444	D1	1135	2	Post dsRNA	**2-057**
EF065875	2-2445	E2	934	2	Post dsRNA	2-063
EF065853	2-2446	E2.1	934	2	Post dsRNA	2-063
EF065877	2-2447	07	829	2	Post dsRNA	**2-059**
EF065835	2-2448	E2.1	933	2	Post dsRNA	2-063
EF065781	4-1504	D1	1135	4	Pre-challenge	4-1501
EF065777	4-1507	E2	935	4	Pre-challenge	4-1550
EF065793	4-1510	E2.1	934	4	Pre-challenge	4-1550
EF065778	4-1519	E2.1	934	4	Pre-challenge	4-1550
EF065776	4-1522	E2	994	4	Pre-challenge	4-1543
EF065775	4-1529	E2.1	934	4	Pre-challenge	4-1550
EF065780	4-1539	E2.1	933	4	Pre-challenge	4-1550
EF065779	4-1549	E2.1	934	4	Pre-challenge	4-1550
EF065741	4-2401	E2	934	4	Post aCF	4-1550
EF065723	4-2402	04	176	4	Post aCF	4-1550
EF065748	4-2403	E2.1	934	4	Post aCF	4-1550
EF065773	4-2404	D1	1135	4	Post aCF	4-2410
EF065720	4-2405	E2	934	4	Post aCF	4-1550
EF065721	4-2406	G1	1411	4	Post aCF	Orphan
EF065732	4-2407	E2	934	4	Post aCF	4-1550
EF065727	4-2408	E2	934	4	Post aCF	4-1550
EF065751	4-2409	E2.1	934	4	Post aCF	4-1550
EF065772	4-2411	E2	934	4	Post aCF	4-1550
EF065746	4-2413	D1	1135	4	Post aCF	4-2412
EF065760	4-2416	03.1	534	4	Post aCF	4-1532
EF065731	4-2417	E2	934	4	Post aCF	4-1550
EF065735	4-2418	E2	934	4	Post aCF	4-1550
EF065771	4-2419	E2	934	4	Post aCF	4-1550
EF065765	4-2420	D1	1144	4	Post aCF	4-1538
EF065730	4-2421	E2	934	4	Post aCF	4-1550
EF065734	4-2422	E2	934	4	Post aCF	4-1550
EF065754	4-2423	E2.1	934	4	Post aCF	4-1550

Accession	Clone #	Pattern	Length	Animal	Time*	Most Likely Gene
EF065768	4-2424	E3	931	4	Post aCF	4-1532
EF065761	4-2425	01	826	4	Post aCF	4-1503
EF065739	4-2426	E2	934	4	Post aCF	4-1550
EF065729	4-2428	E2	934	4	Post aCF	4-1550
EF065726	4-2429	E2	934	4	Post aCF	4-1550
EF065740	4-2430	D3	970	4	Post aCF	4-1543
EF065759	4-2432	01	826	4	Post aCF	4-1503
EF065722	4-2433	B3	1048	4	Post aCF	4-1548
EF065755	4-2434	E2.1	934	4	Post aCF	4-1550
EF065719	4-2435	E2	934	4	Post aCF	4-1550
EF065747	4-2437	E2	934	4	Post aCF	4-1550
EF065724	4-2438	D1	1138	4	Post aCF	4-1543
EF065733	4-2439	D2	1069	4	Post aCF	Orphan
EF065763	4-2440	01	826	4	Post aCF	4-1503
EF065725	4-2441	B3	1048	4	Post aCF	4-1548
EF065764	4-2442	06	79	4	Post aCF	4-1550
EF065762	4-2443	01.3	826	4	Post aCF	4-1503
EF065742	4-2444	E2	934	4	Post aCF	4-1550
EF065736	4-2445	E2	934	4	Post aCF	4-1550
EF065756	4-2446	E2.1	934	4	Post aCF	4-1550
EF065728	4-2447	E2	934	4	Post aCF	4-1550
EF065774	4-2448	E2	911	4	Post aCF	4-1550
EF065749	4-2450	D1	1135	4	Post aCF	4-2412

*LPS = lipopolysaccharide; aCF = artificial coelomic fluid; Lam = β,1-3-glucan

Sequences in bold are exact matches to their corresponding genes (see Table 2)

Supplemental File 3: Histogram of the most likely gene scores for the messages isolated from animal 2. For each *185/333* message isolated from animal 2, the gene with the lowest score was considered to be the gene from which the message was most likely transcribed. A histogram of these scores for every message/gene combination from animal 2 is shown. The values on the y-axis indicate the number of messages for which the most likely gene had a specific score (shown on the x-axis). The bin size for the histogram was 2. The dotted line indicates the score above which the messages were considered to be orphans.



Supplemental File 4: The diversity of the genes derived from coelomocytes is similar to those isolated from sperm. Diversity scores for the complete alignments and individual elements, and the percent of variable positions from the entire alignment were calculated from genes isolated three animals. Genes from animals 2 (bars with diagonal stripes) and 4 (horizontally striped bars) were cloned from gDNA isolated from coelomocytes, whereas genes from animal 10 (black bars) were isolated from sperm gDNA (29).