Unexpected diversity displayed in cDNAs expressed by the immune cells of the purple sea urchin, *Strongylocentrotus purpuratus*

David P. Terwilliger, Katherine M. Buckley, Dhruti Mehta, Priya G. Moorjani and L. Courtney Smith


You might find this additional info useful...

Supplemental material for this article can be found at:
http://physiolgenomics.physiology.org/content/suppl/2006/08/16/26.2.134.DC1.html

This article cites 51 articles, 20 of which can be accessed free at:
http://physiolgenomics.physiology.org/content/26/2/134.full.html#ref-list-1

This article has been cited by 8 other HighWire hosted articles, the first 5 are:

*Genome-wide polymorphisms show unexpected targets of natural selection*
Melissa H. Pespeni, David A. Garfield, Mollie K. Manier and Stephen R. Palumbi
[Abstract] [Full Text] [PDF]

*Genome-wide polymorphisms show unexpected targets of natural selection*
Melissa H. Pespeni, David A. Garfield, Mollie K. Manier and Stephen R. Palumbi
*Proc. R. Soc. B*, October 12, 2011; .
[Abstract] [Full Text] [PDF]

*Diversification of innate immune genes: lessons from the purple sea urchin*
L. Courtney Smith
[Abstract] [Full Text] [PDF]

*Highly Variable Immune-Response Proteins (185/333) from the Sea Urchin, Strongylocentrotus purpuratus: Proteomic Analysis Identifies Diversity within and between Individuals*
Nolwenn M. Dheilly, Sham V. Nair, L. Courtney Smith and David A. Raftos
[Abstract] [Full Text] [PDF]

*Sequence Variations in 185/333 Messages from the Purple Sea Urchin Suggest Posttranscriptional Modifications to Increase Immune Diversity*
Katherine M. Buckley, David P. Terwilliger and L. Courtney Smith
[Abstract] [Full Text] [PDF]

Updated information and services including high resolution figures, can be found at:
http://physiolgenomics.physiology.org/content/26/2/134.full.html

Additional material and information about *Physiological Genomics* can be found at:
http://www.the-aps.org/publications/pg

This information is current as of May 6, 2012.
Unexpected diversity displayed in cDNAs expressed by the immune cells of the purple sea urchin, *Strongylocentrotus purpuratus*

David P. Terwilliger,1 Katherine M. Buckley,1 Dhruti Mehta,2 Priya G. Moorjani,3 and L. Courtney Smith1

1Department of Biological Sciences, George Washington University; 2Howard University College of Medicine; and 3Program in Genomics and Bioinformatics, George Washington University, Washington, District of Columbia

Submitted 23 January 2006; accepted in final form 24 April 2006

Terwilliger, David P., Katherine M. Buckley, Dhruti Mehta, Priya G. Moorjani, and L. Courtney Smith. Unexpected diversity displayed in cDNAs expressed by the immune cells of the purple sea urchin, *Strongylocentrotus purpuratus*. Physiol Genomics 26: 134–144, 2006; doi:10.1152/physiolgenomics.00011.2006.—We recently identified a unique family of transcripts, the 185/333 family, that comprise ~60% of the mRNAs induced by coelomocytes from the purple sea urchin in response to immunological challenge from lipopolysaccharide. An analysis of 81 full-length cDNAs revealed 67 unique nucleotide sequences encoding 64 different proteins. Diversity of the transcripts was based on 25 sequence blocks, or “elements,” which resulted in 22 different element patterns based on their presence or absence. Furthermore, there was a high level of nucleotide variation within elements, including single nucleotide polymorphisms and insertions/deletions, both of which resulted in amino acid sequence variability. The deduced 185/333 proteins contained an NH2-terminal leader, a glycine-rich region with an RGD motif, a histidine-rich region, and a COOH-terminal region. Two 185/333 genes, identified in the partially assembled *Strongylocentrotus purpuratus* genome, have two exons. The first encoded the leader, and the second encoded the remainder of the predicted protein. Estimates from quantitative PCR indicated that there were ~100 alleles in the diploid genome. These results suggested that the purple sea urchin may have mechanisms for generating high levels of diversity in response to immunological challenge that have not been considered previously.

Invertebrates were once thought to have simple immune systems; however, recent evidence suggests that both invertebrates and plants express families of immune response genes with high levels of sequence diversity (reviewed in Refs. 13, 22–25). The fresh water snail, *Biomphalaria glabrata*, expresses fibrinogen-related proteins (FREPs) that contain one or two highly variable immunoglobulin superfamily (IgSF) domains (1, 17, 21, 53, 54). The protocordate, *Branchiostoma floridae* (Amphioxus), expresses a family of chitin-binding genes that also have diversified IgSF variable domains (6–8). A single copy gene in *Drosophila* called the Down syndrome cell adhesion molecule (DSCAM), which is expressed in hemocytes, also contains multiple IgSF domains with high levels of sequence diversity generated by alternative splicing of 95 exons (47). The existence of highly variable defense response genes is not restricted to animals. Plant genomes contain large families of disease resistance (R) genes (reviewed in Ref. 25) that are closely linked and undergo both gene duplication and conversion events, which generate new members (20, 34, 35). Thus, in light of recent evidence and contrary to previous notions that innate immunity was invariant, both invertebrates and plants may be able to diversify their innate immune responses.

The identification of families of variable immune response genes has been facilitated by the characterization of expressed sequence tags (ESTs), which can provide a global picture of transcriptional activity after immune challenge. In the purple sea urchin, *Strongylocentrotus purpuratus*, EST studies have identified a variety of immune-related genes that are upregulated in response to lipopolysaccharide (LPS) (28, 42), including homologs of complement C3, factor B, and putative complement regulators (9, 27, 44; reviewed in Refs. 14, 38–40) as well as a unique family of highly variable transcripts (28). About 60% of the ESTs characterized by Nair et al. (28) showed significant similarity to each other and to an unknown sequence called DD185 (36) or EST333 (42). These transcripts, called 185/333, displayed high levels of diversity when the ESTs were aligned and compared. The 185/333 transcripts increased significantly within 6 h of bacterial challenge (36), suggesting that they encode proteins involved in the sea urchin immune response (28).

The current study provides an analysis of 81 full-length 185/333 cDNAs that displayed high levels of diversity based on the presence or absence of blocks of sequence called elements, in addition to small insertions/deletions (indels) and single nucleotide polymorphisms (SNPs) within the elements. The 185/333 genes are small and have two exons, and a preliminary estimate of gene copy number suggests that there are ~100 alleles per individual. In combination with the emerging data on immune diversity in snails, shrimp, flies, and Amphioxus (6–8, 10, 11, 24, 31, 47), these results suggest a paradigm shift in our understanding of the complexities of innate immunity in invertebrates (13, 23).

The cDNA sequences generated in this study were submitted to GenBank, with accession numbers from DQ183104 to DQ183184, and are also listed in Supplemental Table S1 (the online version of this article contains Supplemental Material; see Physiological Genomics web site).

**MATERIALS AND METHODS**

**Nucleic Acid Isolation**

An arrayed cDNA library was constructed from mRNA from coelomocytes pooled from five bacterially activated sea urchins (5, 28, 32, 33, 36), and filters were provided to us by Eric H. Davidson and colleagues (California Institute of Technology, Pasadena, CA). The library was constructed by use of conventional methods such that...
PCR amplification was not involved at any point in this process. Linkers with restriction sites were ligated onto the cDNAs followed by ligation to the pBK-CMV plasmid (Stratagene, La Jolla, CA) and bacterial transformation, and colonies were transferred directly into 384-well plates.

Genomic DNA (gDNA) from sperm was isolated from a single sea urchin as described (19, 44). A different sea urchin that had been injected with LPS (41) was used to isolate gDNA from coelomocytes. Coelomic fluid (600 μl) was withdrawn and diluted into ice-cold calcium/magnesium-free sea water with EDTA and imidazole (CMFSW-EI: 460 mM NaCl, 10.73 mM KCl, 7.04 mM Na2SO4, 2.38 mM NaHCO3, 30 mM EDTA, 50 mM imidazole, pH 7.4) to a final volume of 2 ml. Coelomocytes (1 x 106) were pelleted and resuspended in 200 μl of Tris-buffered saline (TBS: 200 mM Tris, pH 7.4; 500 mM NaCl) with 400 μg of RNaseA, and the gDNA was isolated using the DNeasy Tissue Kit (Qiagen, Valencia, CA) with minor changes. Proteinase K (company supplied) was added, mixed, and incubated at 55°C for 1 h, followed by the addition of AL buffer (company supplied). Two washes rather than one were performed with AW1 buffer (company supplied).

PCR and Cycle Sequencing

Templates for PCR were either 200 ng of cloned cDNA or 2 μg of gDNA. Reactions consisted of 1 μM each primer (Table 1), 0.5 μM each deoxynucleotide, 2 mM MgCl2, 1× company-supplied buffer, and 0.5 U/20 μl Tag DNA polymerase (Invitrogen, Carlsbad, CA) and were performed with an annealing temperature of 55°C and a 4-min extension.

Clones identified as 185/333 by subtracted-probe hybridization (28) were manually picked from the arrayed library plates and grown overnight in L broth with 25 μg/ml kanamycin, and the plasmids were isolated using Wizard Plus SV Minipreps DNA Isolation Kit (Invitrogen). Sequencing was performed directly from these plasmids with 3–7× coverage using primers listed in Table 1. DNA sequencing reactions employed the Big Dye Cycle Sequencing Kit (Applied Biosystems, Foster City, CA) and were analyzed on a Prism 377 DNA sequencer (Applied Biosystems). The error rate of the polymerase employed in the ABI kit (v.2.0 and v.3.0) is two to three errors for 100–150 kb.

Quantitative PCR

Using quantitative PCR (qPCR), we employed primers (Table 1) to amplify the 185/333 leader, as well as part of the actin coding region, from sperm gDNA isolated from a single sea urchin. gDNA was cleaned by the CTAB method (48), treated with proteinase K, digested for 2 h with HindIII, and subsequently purified with the Wizard SV Gel and PCR Clean-up System (Promega, Madison, WI). DNA concentration was measured using the QuantiT PicoGreen DNA Quantitation Assay (Invitrogen). Plates were read on a Wallac Victor microplate spectrophotometer (Perkin Elmer, Wellesley, MA) with 485-nm excitation and 535-nm emission for 1 s. Each qPCR reaction consisted of 1× Absolute QPCR SYBR Fluorescein Mix (ABgene, Surrey, UK), 500 nM each primer, and 10-fold serial dilutions of gDNA concentration ranging from 100 to 1 ng in 20 μl. A standard curve was generated using five 10-fold serial dilutions (10-7 to 103 plasmids/sample) of three mixed 185/333 genes cloned from coelomocyte gDNA. qPCR reactions were performed in triplicate in an iCycler (Bio-Rad Laboratories, Hercules, CA), using the following program: 95°C Taq activation for 15 min, followed by 40 cycles of 95°C for 15 s, 58°C (185/333 primers) or 63°C (actin primers) for 30 s, and 72°C for 30 s. Fluorescence data were collected during the annealing step of the amplification program. A melt curve was performed for all samples, and data analysis was done with the iCycler software (Bio-Rad Laboratories).

Southern Blots
gDNA was digested to completion with EcoRI, HindIII, or PstI, and Southern blots were performed as described (43, 44).

Several clones were used as templates for riboprobes (cDNA clones; Sp0254, Sp0323, Sp0329, Sp0419, Sp0641, Sp0653, Sp0761), which were synthesized and hybridized to blots as previously described (27). Clones of individual elements [F2R5–28 (element 1) and F5R9–17 (element 25)] were amplified from gDNA by PCR using the appropriate primers (Table 1) and cloned into pCR4-TOPO vector using TOPO-TA Cloning Kit for Sequencing (Invitrogen).

Bioinformatics

Sequence alignments were done with BioEdit (16). The features of the 185/333-deduced protein transcripts were predicted by ProtParam (http://www.expasy.org/tools/protparam.html), statistical analysis of protein sequences (SAPS; http://www.isrec.isb-sib.ch/software/SAPS_form.html; Ref. 3), the HMMTOP server (45, 46), and PredictProtein (http://cubic.bioc.columbia.edu/predictprotein; Ref. 37). For secondary structure predictions, only stretches of greater than or equal to four residues were considered significant. WinClada v.1.00.08 (29) was used to identify identical sequences for the nucleotide and the amino acid alignments.

Sequence Diversity

Diversity scores were calculated based on the method of Durbin et al. (12). Elements 11 and 15 were analyzed either with all sequences represented or as individual subelements. A diversity score (D) for each nucleotide position was calculated by summing the diversity

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence (5' → 3')</th>
<th>185/333 Annealing Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>185-5'UTR</td>
<td>TAGCATGCGAGAGCCT</td>
<td>5'UTR</td>
</tr>
<tr>
<td>185-3'UTR</td>
<td>AATTTCTACACCTCGGCGAC</td>
<td>3'UTR</td>
</tr>
<tr>
<td>185-LR1</td>
<td>ACTCTGACTACATSTGGCGYG</td>
<td>leader</td>
</tr>
<tr>
<td>185-F2</td>
<td>AAGGATTTCAATGAGAACTGAG</td>
<td>element 1</td>
</tr>
<tr>
<td>185-F5</td>
<td>GGACGCARGAMGATCTCT</td>
<td>element 25</td>
</tr>
<tr>
<td>185-F6</td>
<td>GAAAGAAGAAATTGATGGGTC</td>
<td>element 7</td>
</tr>
<tr>
<td>185-R5</td>
<td>ACTCTGACTACATGCGAGACGCAC</td>
<td>elements 4 and 6; element 5 (Sp0289 only)</td>
</tr>
<tr>
<td>185-R6</td>
<td>GGAAGATGGCAGTCTCTACT</td>
<td>element 7</td>
</tr>
<tr>
<td>185-R9</td>
<td>CTTGATGTGGAGAARATGGC</td>
<td>element 25</td>
</tr>
<tr>
<td>DMR</td>
<td>ATGACACCTCTACATGC</td>
<td>3'UTR</td>
</tr>
<tr>
<td>CyF</td>
<td>ACGAGCAATGTTTGGCCTCTTGTAT</td>
<td>actin</td>
</tr>
<tr>
<td>CyR</td>
<td>CGGTCTTCTGTCATCCATACCGCAGCA</td>
<td>actin</td>
</tr>
<tr>
<td>T7</td>
<td>TAAATGCTAGTATCAATTAGG</td>
<td>na</td>
</tr>
<tr>
<td>T3</td>
<td>ATTAACACCTCAGTAAAGGGA</td>
<td>na</td>
</tr>
</tbody>
</table>

na, Not applicable.
scores for each state (5 possible states for 4 nucleotides and a gap, or 21 possible states for 20 amino acids and a gap) present at a given position. The total diversity score for each element was the sum of the diversity scores for all nucleotide positions within the element divided by the length of the element, as follows:

$$ D = \frac{\sum_{i=1}^{n} \left( \sum_{j=0}^{l} \left( -l \ln \left( \frac{j}{n} \right) \right) \right)}{L} $$

where $D$ = diversity score for the element, $L$ = length of the element in either nucleotides or amino acids, $l$ = the specific alignment position (a subset of $L$), $i$ = the number of times the single state occurred, and $n$ = the number of sequences present at a given position in the alignment. The frequency and number of different nucleotides for the variable positions were kept constant among all variable positions. Sequences located 3' of the first stop codon were not included in the analysis. Preliminary analyses indicated that neither the length of the alignment nor the number of sequences analyzed influenced the diversity scores. The analysis was performed with an internal Perl (v.5.6) script. A two-tailed $t$-test was employed for statistical comparisons of diversity scores.

Molecular evolutionary genetics analysis (MEGA3; Ref. 18) was used to generate neighbor-joining phylogenetic trees employed in phylogenetic analysis by maximum likelihood (PAML; Ref. 50) to identify statistically significant nonsynonymous and synonymous (dn/ds) ratios.

**RESULTS**

A subset of ESTs ($n = 307$) originally identified by Nair et al. (28) was aligned to determine the extent of the sequence variability (data not shown). From those results, 81 cDNAs were chosen for full-length sequencing. An alignment of the cDNAs revealed a greater level of complexity than had been previously identified from the partial ESTs (Fig. 1). For complete alignments, see Supplemental Figs. S1 and S2.
Element Patterns

An optimal alignment of the 81 full-length cDNA sequences required additional gaps that defined 25 elements (Fig. 1A; for details, see Supplemental Figs. S1 and S2 and Supplemental Table S1). None of the 81 cDNAs possessed all elements; coding regions of individual cDNAs had between 6 [Sp0289 (pattern D1.1); see Fig. 1A] and 22 [Sp0313 (A1) and Sp0164 (C1)] nucleotides. On the basis of the presence or absence of the elements, 22 distinct element patterns were identified. Eight of these patterns were represented by multiple clones, defined as cDNA sets, and had between 2 and 24 members (Fig. 1A; Supplemental Table S1). The remaining 14 patterns were identified from single clones.

All of the cDNA sequences [except Sp0289 (D1.1); see below] contained a leader and elements 1, 7, 9, and 25a (Fig. 1A). When present, element 15 displayed significant variability in both nucleotide sequence and length (Fig. 2, A and B), and groups of cDNAs (Fig. 1A, left column, groups 1–7) could be categorized based on the presence, absence, or subtype of element 15. Group 1 had element 15a, and group 2 had element 15b, etc., whereas groups 6 and 7 were missing element 15. Among the 22 different patterns observed, groups of cDNAs were characterized by shared subsets of elements and had identical patterns of elements 13–25a. For example, group 1 shared a subset of elements (leader, 1, 5, 7, 9–11a, 14–16, 20–25a) and was the only group that had element 23 (Fig. 1A). cDNAs within groups 2, 3, 4, and 5 also shared unique subsets of elements, suggesting that suites of elements tended to appear together.

Types of Sequence Diversity

Three types of diversity. In addition to the differential presence of 25 elements, there was substantial nucleotide diversity within the elements that consisted of divergent subelements, indels, and SNPs (Fig. 2). Element 11 had two mutually exclusive, highly divergent subelements, 11a and 11b, that encoded very different amino acid sequences (Fig. 2, C and D). The variability among the five types of element 15 (a–e), based on length of the element, was suggestive of indels (Fig. 2, A and B). A third type of diversity was characterized by

Fig. 2. Three types of sequence diversity in 185/333 cDNAs. A and B: alignment of subelements 15a–15e (11 unique sequences of 72 cDNAs) shows length variations that may be due to indels (A). The corresponding amino acid sequences (B) show variable numbers of histidines. C and D: alignment of subelements 11a and 11b (3 unique sequences of 11) shows divergent nucleotide (C) and amino acid sequences (D). E: alignment of subelement 4 (21 unique sequences of 80) shows diversity based on single nucleotide polymorphisms (SNPs). Gray boxes indicate nonsynonymous nucleotide changes (55 within 66 nucleotides, 13 variant amino acid positions within a length of 22). G and H: alignment of the 3 types of element 25b (9 unique sequences of 81) shows both indels and SNPs.
SNPs as exemplified by element 4 (Fig. 2E). The amino acid variability in element 4 resulted from 55 nonsynonymous nucleotide substitutions (51% of the nucleotide variability in this element) and encoded 20 different deduced amino acid sequences (Fig. 2F).

Terminal element diversity. Element 25, which was present in all cDNAs, had both indels and SNPs and was divided into three subsets (elements 25a, -b, and -c) based on the location of the stop codon (Fig. 1A; Supplemental Figs. S1 and S2). SNPs at the 3’-end of elements 25a, -b, and -c sometimes resulted in stop codons (Fig. 1A; Supplemental Figs. S1 and S2). When clones had a stop codon at the end of element 25a, they also had stop codons at the 3’-ends of both 25b and 25c (Fig. 1A; Supplemental Figs. S1 and S2). Other clones that had the initial stop at the end of 25b also had a stop at the end of 25c. One clone, B4, had a single stop at the end of 25c (Fig. 1A). In contrast, D1.1 had a rare frame shift caused by the deletion of two nucleotides at position 286 that resulted in an early stop codon at nucleotide 362. Consequently, the putative protein would be truncated in element 5 with only 112 amino acids (Fig. 1A; Supplemental Figs. S1 and S2 and Supplemental Table S1).

Indels in element 25b were the basis for three sequence variants (Fig. 2, G and H; Supplemental Fig. S1). Type 25b-1 (present in 54 clones; translated in 21) encoded the full-length sequence (54 nt). Types 25b-2 (48 nt; present in 15 clones, translated in 14) and 25b-3 (39 nt; present in 12 clones, translated in 10) had in-frame deletions at two locations (Fig. 2, G and H; Supplemental Fig. S1 and Supplemental Table S1). These examples illustrate the different types of sequence variability observed throughout the 185/333 cDNA sequences, which resulted in significant variations in the encoded proteins (see below).

Quantitation of Sequence Diversity

Because nucleotide diversity was noted throughout the cDNA sequences, diversity was quantified for individual elements. The diversity scores were based on the frequency of different nucleotides present at each position within a given alignment (see Ref. 12). The maximum possible diversity score was 1.609 for an alignment in which every variable position in the alignment had all four nucleotides as well as a gap, or a score of 3.045 for 20 amino acids plus a gap. The score was 0 if all sequences were identical.

Modeled alignments. To interpret the biological correlations of realistic diversity scores for the 185/333 sequences, diversity scores were calculated for a series of alignments modeled after nucleotide and amino acid elements of known sequence variability (Fig. 3A). The variables employed to model the elements were derived from the average length and variability of the 185/333 elements. The alignments were varied by
changing the number of different bases or amino acids present at each variable position and/or by changing the number of sequences containing variations. A maximum of five different states were present at each variant position (4 nucleotides and a gap), and the 185/333 protein alignments displayed no more than five different amino acids per position (Supplemental Fig. S2).

Diversity scores for 185/333 elements. The modeled alignments provided a maximum diversity score of 0.5304 for an alignment in which 40% of the variable positions contained four different amino acids in 20% of the sequences (Fig. 3A). This value was similar to the highest observed amino acid diversity score among the elements, element 11 (0.5381; Fig. 3B), in which subelements 11a and 11b were significantly different from each other (Fig. 2A). The nucleotide diversity score for element 1 (0.1103) was slightly greater than that of a modeled element in which eight (10%) of the sequences had a single change at 30% of the positions (0.0954). In contrast, the nucleotide diversity score for element 15b (0.0136) was similar to a modeled alignment in which only one sequence (1%) had a single change in 30% of the positions (0.0197) (Fig. 3, A and B).

Diversity of cDNA sets. Nucleotide sequence diversity for members of cDNA sets (Fig. 3, C and D) showed that the average diversity scores were much lower than scores of individual elements, which were collected from across sets of cDNAs plus those from unique patterns (0.010 ± 0.007 vs. 0.073 ± 0.06; Fig. 3, B vs. D). This result suggested that the members of sets were more similar to each other than to members of other sets.

In a parallel analysis, diversity for each nucleotide position for the entire alignment was calculated using the approach of Wu and Kabat (49) and the method of Durbin et al. (12). Results were similar to the analysis of individual elements and showed that polymorphic nucleotides were present throughout the length of the cDNAs with no apparent hypervariable regions (Supplemental Fig. S3).

Nonsynonymous vs. Synonymous Nucleotide Substitutions

Diversity within the nucleotide sequence and the resulting changes in the amino acid sequence is typically analyzed through ratios of nonsynonymous to synonymous (dn/ds) nucleotide substitutions (50). Results from cDNA sets with four or more members indicated that patterns E2, D1, O1, and E3 appeared to be under positive selection for diversification (dn/ds > 1), whereas C1, A2, and E1 did not show positive selection (dn/ds < 1) (Table 2). In addition, the region spanning the leader and element 1 from all 81 cDNAs was long enough for analysis by PAML and had a dn/ds ratio of 1.18, indicating that it was also under positive selection for diversification. However, inspection of the sequence alignments for the leader and element 1 (Supplemental Figs. S1 and S2) revealed that the majority of the nonsynonymous changes were located in element 1. This was also evident from the ratios of nucleotide and amino acid diversity scores (see below; Fig. 3B), suggesting that the elevated dn/ds ratio was largely driven by element 1 rather than the leader. These results were in agreement with the dn/ds analysis of the same region in the ESTs (28) and suggested that a diversifying selection pressure may exist for these transcripts.

Sequence length requirements for PAML meant that individual elements were too short to be analyzed separately. As an alternative, ratios of the nucleotide and amino acid diversity scores of individual elements were used to estimate the impact of nucleotide substitutions on the corresponding changes in the amino acid sequence (Fig. 3B). Results ranged from 0 (all nucleotide changes resulting in synonymous amino acid changes) to 3 (all nucleotide changes resulting in nonsynonymous amino acid changes). The observed range among the elements was 0.7 (element 3) to 3.0 (elements 11b, 15b, 15c, and 24). Elements 3 and 19 exhibited the most synonymous nucleotide polymorphisms, while ratios of ≈2.5 were found for elements 1, 8, 9, 10, 11b, 14, 15b, 15c, and 24, in which ≈83% of the nucleotide substitutions resulted in nonsynonymous amino acid changes. These ratios reflected only the synonymous/nonsynonymous changes observed within the elements and were independent of the overall diversity scores or numbers of variable positions within the elements. For example, the elevated ratio for element 11b was due to an SNP in one sequence that resulted in a nonsynonymous change, which meant that all nucleotide changes for 11b resulted in a nonsynonymous change.

Ratios for nucleotide and amino acid diversity scores were also calculated for cDNA sets (Fig. 3D). The average ratio for the sets was similar to the average ratio for all the elements (1.80 vs. 1.88, respectively), although the range of scores was narrower (1.5–2.2). These ratios indicated that many of the mutations were nonsynonymous. However, the ratios did not correlate directly with the dn/ds values (Table 2), indicating that this measure only evaluates the relative proportion of synonymous to nonsynonymous changes and is not a direct measure of selection. Two comparable sets, E3 and A2, had similar nucleotide diversity scores, but E3 had a higher amino acid diversity score, which was reflected in the ratios for the two sets (2.0 for E3 compared with 1.6 for A2), suggesting that the number of nonsynonymous nucleotide substitutions differed between these two cDNA sets.

Table 2. Nucleotide diversity in cDNA sets that have identical element patterns

<table>
<thead>
<tr>
<th>Set Identifier</th>
<th>cDNAs per Set</th>
<th>Unique cDNA Sequences</th>
<th>dn/ds*</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2</td>
<td>24</td>
<td>15</td>
<td>1.2670</td>
</tr>
<tr>
<td>D1</td>
<td>13</td>
<td>10</td>
<td>1.3462</td>
</tr>
<tr>
<td>O1</td>
<td>8</td>
<td>8</td>
<td>1.0230</td>
</tr>
<tr>
<td>E3</td>
<td>5</td>
<td>4</td>
<td>1.7745</td>
</tr>
<tr>
<td>A2</td>
<td>4</td>
<td>4</td>
<td>0.8909</td>
</tr>
<tr>
<td>C1</td>
<td>7</td>
<td>3</td>
<td>0.7313</td>
</tr>
<tr>
<td>E1</td>
<td>3</td>
<td>3</td>
<td>0.2767</td>
</tr>
</tbody>
</table>

*Under positive selection (>1.0).

Structure. The deduced 185/333 proteins (with the exception of the truncated protein encoded by D1.1) shared an overall organization of an NH2-terminal hydrophobic leader followed by a glycine-rich (gly-rich) region, a histidine-rich (his-rich) region, and a COOH-terminal region (Fig. 1B). There were nine elements within the gly-rich region and 16 elements within the his-rich region. On the basis of variations in element patterns, the predicted sizes of the encoded proteins ranged from 23.1 kDa [Sp0726 (O2), 654 nt] to 55.3 kDa [Sp0313 (A1),
1,515 nt; Supplemental Table S1). None of the deduced 185/333 proteins contained cysteines or showed discernable secondary structure other than the hydrophobic leader that contained a short β-strand followed by a short α-helical region which extended into element 1. Another short α-helical region was identified in element 7 (Fig. 1B). With no known homologs, predictions of tertiary structure of these proteins could not be done.

Conserved motifs. There were a number of conserved motifs located in element 7 near the end of the gly-rich region including an RGD motif (Fig. 1B; Supplemental Fig. S2). Within the his-rich region, there were 11 patches of histidines, varying from 2 to 17 amino acids in length, interspersed with Gly, Arg, and Gln (Fig. 1B; Supplemental Fig. S2). Five categories of repeats were located throughout the sequence (Fig. 1B; Supplemental Fig. S2), and the full-length cDNAs showed additional variability within repeat types 1, 2, 4, and 5 compared with the repeats identified previously in the ESTs (28) (data not shown). Conserved N-linked glycosylation sites were identified in 16 locations, and although no sequence contained more than eight, most were positioned in the his-rich region and located at the start of each of the type 2 and type 3 tandem repeats (Fig. 1B; Supplemental Table S1). Seven conserved O-linked glycosylation sites were identified within a short region of nine amino acids in element 25a (Fig. 1B; Supplemental Fig. S2 and Supplemental Table S1). The deduced 185/333 proteins exhibited significant diversity in length and sequence based on element patterns, amino acid sequence variability, and positions for posttranslational modification.

Gene Structure

Genome blots. The diversity among the cDNA sequences suggested that the 185/333 gene might be a single copy with many exons encoding all possible elements, similar to Drosophila DSCAM (47). Alternatively, there might be a family of 185/333 genes, each one encoding a different variant, similar to snail FREPs (24). To address the question of whether the messages were derived by alternative splicing from a single copy gene or from multiple genes, genome blots were performed on sperm gDNA from three sea urchins (Fig. 4A).

The resulting banding patterns differed significantly among the three animals and the three restriction enzymes used to digest the DNA, indicating that the 185/333 locus was highly polymorphic. Surprisingly, six different riboprobes produced from full-length cDNAs gave identical banding patterns (Fig. 4A).

The seventh probe generated from Sp0641 (E5) identified the same bands but displayed extra bands in some lanes (Fig. 4B). Comparisons of the cDNA sequences that were used to make the probes indicated that they were members of three cDNA sets (E2, D1, O1; two clones were not fully sequenced) plus E5, which had a unique element pattern. Comparisons of the elements present in the riboprobes did not explain the extra bands or why all blots were identical. These results did not clarify how many 185/333 genes were present in the genome, because the larger bands could conceivably contain multiple genes or the entire set of bands in each lane could represent multiple exons from one or a few large genes.

To determine whether the 185/333 transcripts were encoded by a large gene with many exons, genome blots were analyzed with probes corresponding to element 1 or a portion of element 25. If the set of bands present on the blots represented separate exons, probes of individual elements should hybridize to a subset of those bands. However, the resulting banding patterns for both of the element probes were identical to the patterns observed for the six full-length probes (Fig. 4A). Furthermore, probes for both elements hybridized to DNA fragments as small as 1.55 kb, indicating that some or all of the genes were small.

PCR amplification of gDNA. In a parallel investigation, to understand the 185/333 gene structure, gDNA was isolated from coelomocytes or sperm from two sea urchins and analyzed by PCR with a pair of primers (Table 1) that amplified the region spanning elements 1–7. Because there were four type 1 repeats within this region (Fig. 1B; Supplemental Fig. S2), the reverse primer annealed to several sites within individual elements and amplified a variable number of bands depending on the number of elements present. Four bands were amplified from the control clone, A1 (Fig. 5, lane 4), which had all elements between 1 and 7, whereas only one band was amplified from B3, which was missing elements 5 and 6 (Fig. 5A).
Both sources of gDNA amplified fragments that were the same sizes as those amplified 3 or 4 fragments, respectively, as indicated by the arrows at (pattern B3 were used as controls.

1–7 were amplified by PCR from gDNA isolated from sperm or coelomocytes as those amplified from the cDNA templates (Fig. 5, lane 2). When gDNA was amplified from four different animals, results showed three or four bands of the same sizes as those amplified from the cDNA templates (Fig. 5, lanes 3 and 4). Although results from the genomic template did not rule out amplification from separate genes with variations in the presence or absence of elements 1–7, results were consistent with small genes lacking introns between elements 1 and 7 and were in agreement with the genome blots.

Although the PCR results indicated that there were no introns between elements 1 and 7, some elements, such as 5 and 6, were commonly missing (Fig. 1A). Consequently, the possibility of alternative splicing from splice sites within the open reading frame was investigated for all cDNAs. Consensus, canonical splice sites were identified at the boundaries of elements 4, 5, 6, 13, and 25b (Fig. 1B; Supplemental Fig. S1). The amino acids present at these sites in elements 4, 5, and 6 were Gly/Arg/Arg/Phe (Supplemental Fig. S2), and the codons for Arg/Phe provided canonical splice signals: GT-AG or AT-AG (Supplemental Fig. S1). The amino acids present at these sites in elements 4, 5, and 6 were Gly/Arg/Arg/Phe (Supplemental Fig. S2), and the codons for Arg/Phe provided canonical splice signals: GT-AG or AT-AG (Supplemental Fig. S1) (4).

Genome analysis. Bioinformatic approaches were used to obtain definitive information on the gene structure. Basic local alignment search tool (BLAST) searches of the partially assembled sea urchin genome (assembly 9/22/03) were performed using 185/333 cDNA sequences that identified two contigs. The gene on contig12117 (Fig. 6A) matched to the element pattern of E2, while the gene on contig12130 (Fig. 6B) matched D1. Both genes were organized similarly: the first exon encoded the leader followed by a small intron (400 bp), and the second exon encoded elements 1–25. These small genes were consistent with results from genome blots probed with single elements (Fig. 4) and with PCR amplification of elements 1–7 from genomic templates (Fig. 5).

Gene copy number. Because only two 185/333 genes were identified from the partially assembled sea urchin genome, qPCR with gDNA from an individual sea urchin was used to estimate the 185/333 gene copy number. This method amplified all copies of the leader derived from all alleles and pseudogenes, including identical copies. Results suggested the presence of 80–120 alleles in a diploid genome. To validate the qPCR assay, the number of actin alleles was quantified in an identical manner, and results showed the presence of 16 alleles per diploid genome, which was consistent with 5 actin genes and 3 pseudogenes (19).

DISCUSSION

The data presented here, in addition to preliminary reports (28, 36), show that the 185/333 cDNAs represent an unexpectedly diverse set of transcripts produced in response to immune challenge from bacteria or LPS. Optimal alignments of the full-length cDNAs reveal 22 different element patterns that can be categorized into seven cDNA groups based on the presence/absence of 25 elements and eight cDNA sets composed of multiple cDNAs with identical element patterns, but not identical sequence. Diversity is present throughout the length of the sequences; however, no hypervariable regions were detected. The level of diversity varies among the elements. Positive selection was observed within some cDNA sets as well as the sequence spanning the leader and element 1. The elevated level of nonsynonymous substitutions in the 185/333 sequences implies that this system is under pressure to diversify. On the basis of the timing of the expression in response to LPS and bacteria (28, 36), pathogen pressure may be the source of the underlying selection toward diversification. A number of conserved motifs are present in the deduced proteins including an RGD sequence. The 185/333 proteins lack cysteines and are not homologous to any known protein, making predictions of folding uninformative. The genes are small, with two exons and one intron, and ~100 alleles are present in a diploid genome of this outbred species.

Message Diversity

The primary result presented is the unexpected level of sequence diversity in a set of messages expressed by the immune cells of an invertebrate responding to an immune challenge. The diversity is achieved through a combination of differing element patterns and sequence variability within those elements that appear as indels of varying size and SNPs. Element 11, however, is composed of two subelements, 11a and 11b, that have such different sequences that they may actually be different elements. Because none of the cDNAs had both 11a and 11b, it was not possible to determine how these two regions aligned relative to each other and, therefore, were
aligned together. Alternatively, aligning them separately would have created a discontinuity in the alignment.

The presence of 11a or 11b or the lack of element 11 in some of the cDNAs could be viewed similarly to our treatment of element 15 as a means to categorize cDNAs into groups. The element patterns associated with a specific subelement 15 may be representative of subfamilies of genes that have been defined in other systems, such as the 13 subfamilies of the FREP genes (52, 53). The diversity of the sequences among the members of sets was significantly less than the diversity of individual elements taken from across sets and groups (Fig. 3). This indicates that sets of cDNAs with shared element patterns also tend to have sequences that are more similar and suggests that they may be derived from alleles of the same gene or from a subfamily of genes with the same element patterns but variant sequences.

From the estimate of ~100 alleles in the diploid genome, there may be copies of genes that are more similar to each other, and because our cDNAs were collected from a library constructed from five sea urchins, individuals may harbor different alleles of the same gene. This polymorphism within the population is evident from the variety of banding patterns among the individuals used for the genome blots (Fig. 4). The observed discrepancy between the qPCR estimates of gene copy number and the number of genes that were identified within the initial assembly of the sea urchin genome may be explained by problems with the assembly. It is possible that the similarity among the 185/333 trace sequences may have resulted in their assembly into a few “consensus” genes rather than into separate genes. It is noteworthy that the genes identified from the genome match to the two common cDNAs. Consequently, until a more accurate genome is assembled, the qPCR allele estimates are the most accurate gene number estimate. Future work will determine whether the number of genes within an individual can account for the diversity observed in the messages or whether additional posttranscriptional and posttranslational modifications are involved, including cryptic alternative splicing, RNA editing, variable glycosylation, or other modifications to the amino acid sequence.

**Gene Structure**

The discovery that the 185/333 genes are small, with two exons and a single intron, dispelled the notion that alternative splicing from a multi-exonic gene was the source of the element variability, although cryptic splicing has not been ruled out. A comparison of the two genes identified from the genome, patterns E2 and D1 (Fig. 1A), indicates that the major differences are due to the presence or absence of elements 6, 12, 15, and 16–19, plus the position of the stop codon in element 25. Sequence variations are present within elements shared between the two genes. These differences reside entirely within the second exon of each gene and consequently dictate the variation in the element patterns of cDNAs that would be transcribed from these genes. Preliminary data for >100 genes that have been cloned and sequenced from two individual sea urchins indicate that most of the 185/333 genes have the same structure: two exons and one intron (unpublished data). Consequently, the source(s) of the diversity in element patterns that are present within the second exon is unknown.

**Paradigm Shift**

The accepted paradigm for innate immunity states that broad recognition of pathogen-associated molecular patterns is mediated by germline-encoded, invariant receptors that function within innate immunity, while diversification of immune responses is limited to adaptive immunity. It was thought that the pattern recognition receptors had been selected over evolutionary time and were optimized for the efficient detection of broad classes of pathogens. However, recent studies, in addition to the results presented here, have shown that invertebrates, and perhaps all organisms, diversify their innate immune response to compensate for the appearance of new pathogen variants. Higher vertebrates accomplish this through mechanisms that modify complex but essentially single copy genes, such as immunoglobulins and T cell receptors that undergo somatic rearrangements mediated by recombinases and other enzymes. In a lower vertebrate, the sea lamprey, diversity has been identified in the variable lymphocyte receptors (VLRs), in which somatic diversification changes the numbers of leucine-rich repeats (LRRs) that are encoded by individual messages (2, 32). The DSCAM gene expressed in *Drosophila* hemocytes has 95 exons and produces long primary transcripts that are alternatively spliced to produce ~18,000 different proteins that may be involved in effective phagocytosis of microbes (47).

In other systems, diverse messages are transcribed from large gene families. These include the FREP genes in snails that are thought to diversify through alternative splicing and gene conversion (30, 51; reviewed in Ref. 24), the VCBP gene family in Amphioxus (6–8, 22, 23, 51), and the antimicrobial penaeidins in shrimp (10, 15, 31). Large families of genes involved in immune responses (R) are also found in higher plants. Closely linked R genes encode LRR-containing proteins, generate diversity through gene duplication and conversion, and have been described as a “patchwork” locus of similar genes (34). The close linkage enables the R loci to essentially act as a set of large repeats that promote gene duplication in addition to chromosomal mispairing during meiosis, which leads to unequal crossovers and results in expansion and contraction of the R loci in different haplotypes (34, 35). These mechanisms have resulted in 149–163 R genes in *Arabidopsis* (25, 26). Overall, these mechanisms support significant sequence variability while at the same time limiting sequence homogenization from excessive gene conversion. There are some similarities between the plant R gene system and the 185/333 genes in the purple sea urchin that also reveal a patchwork of elements that is made more complex through variations in the sequence of the elements themselves. Preliminary evidence indicates that the 185/333 genes are closely linked (unpublished data), and clues to the mechanisms of diversity in this gene family may be guided by what is known about the plant R genes.

The diverse transcripts that have been identified and characterized within various invertebrates and in higher plants show that a high level of diversity can be generated within the innate immune systems of these organisms. These results provide evidence that most, if not all, animals and plants have diversified immune responses that are a result of mechanisms...
IMMUNE DIVERSITY IN A SEA URCHIN

143

or combinations of mechanisms and may be unique to each species (13, 23, 25).

ACKNOWLEDGMENTS

We thank Drs. David Raftos, Virginia Brockton, Sham Nair, and Paul Gross for helpful suggestions during manuscript preparation. We also thank Drs. Sheri Church and Weigun Peng for assistance with diversity analysis, Claudio Gutman for sequencing assistance, and Heather Del Valle for bioinformatic assistance. Current address for P. G. Moorjani: Massachusetts General Hospital, Center for Human Genetic Research, Psychiatric and Neurodevelopmental Genetics Unit, Boston, Massachusetts.

GRANTS

This research was supported by the National Science Foundation (MCB-0077970 and MCB-0424235, to L. C. Smith).

REFERENCES

39. Smith LC, Chang I, Britten RJ, and Davidson EH. Sea urchin genes expressed in activated coelomocytes are identified by expressed sequence


