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

Physiol. Genomics 26:134-144, 2006. doi:10.1152/physiolgenomics.00011.2006

Physiological Genomics publishes results of a wide variety of studies from human and from informative model systems with techniques linking genes and pathways to physiology, from prokaryotes to eukaryotes. It is published quarterly in January, April, July, and October by the American Physiological Society, 9650 Rockville Pike, Bethesda MD 20814-3991. ISSN: 1094-8341, ESSN: 1531-2267. Visit our website at http://www.the-aps.org/.

Physiol Genomics 26: 134–144, 2006;
doi:10.1152/physiolgenomics.00011.2006.

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 Smith[1]

[1]*Department of Biological Sciences, George Washington University;* [2]*Howard University College of Medicine; and* [3]*Program 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 NH_2-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.

coelomocytes; innate immunity; echinoderm

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 protochordate, *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

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: L. C. Smith, 340 Lisner Hall, Dept. of Biological Sciences, George Washington Univ., 2023 G St. NW, Washington, DC 20052 (e-mail: csmith@gwu.edu).

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 Na_2SO_4, 2.38 mM $NaHCO_3$, 30 mM EDTA, 50 mM imidazole, pH 7.4) to a final volume of 2 ml. Coelomocytes (1×10^6) 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 $MgCl_2$, 1× company-supplied buffer, and 0.5 U/20 μl *Taq* 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 *Hin*dIII, and subsequently purified with the Wizard SV Gel and PCR Clean-up System (Promega, Madison, WI). DNA concentration was measured using the Quant-iT 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 (10^7-10^3 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 *Eco*RI, *Hin*dIII, or *Pst*I, 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 1. *Primers used in PCR and sequencing*

Primer	Sequence (5′ → 3′)	*185/333* Annealing Site
185-5′UTR	TAGCATCGGAGAGACCT	5′UTR
185-3′UTR	AAATTCTACACCTCGGCGAC	3′UTR
185-LR1	ATCRTYGCCATYSTGGCYG	leader
185-F2	AAGMGATTWCAATGAACKRCGAG	element 1
185-F5	GGAACYGARGAMGGATCTC	element 25
185-F6	GAAGAAGAAACTGATGCTGCC	element 7
185-R5	ACTCTGTACTGCGGAGAGCCGAC	elements 4 and 6; element 5 (*Sp0289* only)
185-R6	GCAGCATCAGTTTCTTCKTCTC	element 7
185-R9	CTTHARGTGGTGAARATGTCG	element 25
DMR	ATGCACCTTTCACTGGC	3′UTR
CyF	ACGACGATGTTGCCCCTCTTGTCAT	actin
CyR	GCTGTCCTTCTGTCCCATACCGACCA	actin
T7	TAATACGACTCACTATAGGG	na
T3	ATTAACCCTCACTAAAGGGA	na

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

$$D = \frac{\sum_{l=0}^{L}\left(\sum \frac{i}{n}\cdot\left[-ln\left(\frac{i}{n}\right)\right]\right)_l}{L}$$

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.

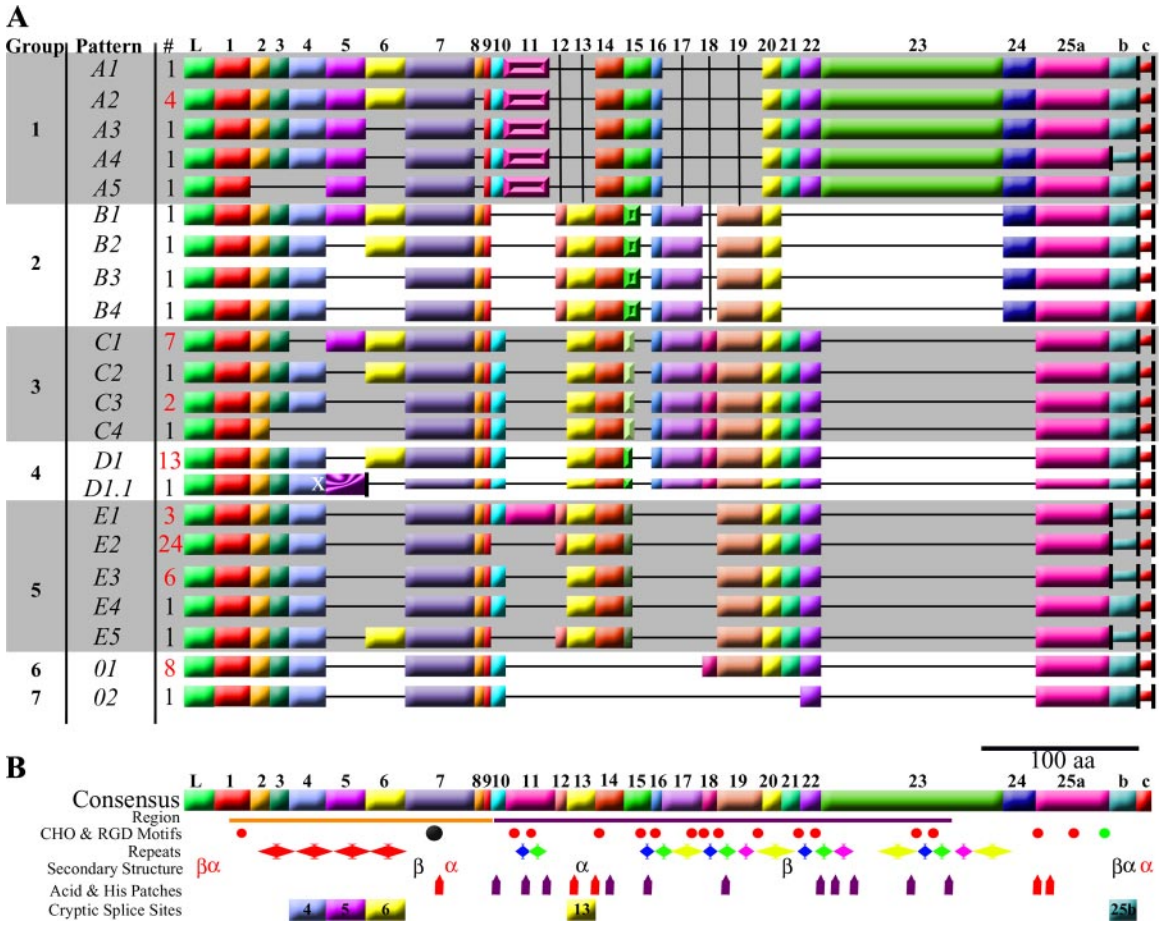

Fig. 1. Diversity of *185/333* transcripts. *A*: full-length sequences from 81 cDNA clones were manually aligned, and gaps (horizontal lines) were inserted to optimize the alignment. Gaps identified separate elements (numbered at *top*), which are represented by colored boxes. The cDNAs are organized into groups (numbered at *left*, identified by gray background shading) based on element 15. *Group 1* is defined by subelement 15a, *group 2* by subelement 15b, *group 3* by subelement 15c, *group 4* by subelement 15d, and *group 5* by subelement 15e. *Groups 6* and *7* do not have element 15. Groups are also indicated by pattern designations: *group 1* is *pattern A*, *group 2* is *pattern B*, etc. Element 25 was subdivided into 3 subelements, 25a, -b, and -c, based on the location of the stop codon (black vertical lines). A frame shift (white X) in element 4 of clone *D1.1* leads to an early stop codon (black vertical lines) in element 5 (black/purple). The remainder of the *D1.1* sequence is shown as smaller blocks to show that the sequence is present but to indicate that this region may not be translated. Eight sets of cDNAs (*E2*, *D1*, *C1*, *01*, *E3*, *A2*, *E1*, *C3*) are composed of multiple members that have identical element patterns (#). *B*: representation of the protein consensus sequence. The deduced protein is separated into a glycine-rich region (gly-rich; orange line) and histidine-rich region (his-rich; purple line). Symbols indicate the presence of an RGD motif in element 7 (black circle), N-linked glycosylation sites (red circles) and O-linked glycosylation sites (green circle), 5 types of repeats (colored diamonds: type 1 = red, type 2 = blue, type 3 = green, type 4 = purple, and type 5 = yellow), secondary structure predictions (α-helices and β-strands), patches of acidic amino acids (red vertical bars), patches of histidines (purple vertical bars), and 5 elements surrounded by cryptic splice signals. Drawn to scale.

Element Patterns

An optimal alignment of the 81 full-length cDNA sequences required additional gaps that defined 25 elements (Fig. 1*A*; 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. 1*A*)] and 22 [*Sp0313* (*A1*) and *Sp0164* (*C1*)] elements and ranged in length from 336 [*Sp0289* (*D1.1*)] to 1,485 [*Sp0313* (*A1*)] 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. 1*A*; 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. 1*A*). When present, element 15 displayed significant variability in both nucleotide sequence and length (Fig. 2, *A* and *B*), and groups of cDNAs (Fig. 1*A*, *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. 1*A*). 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

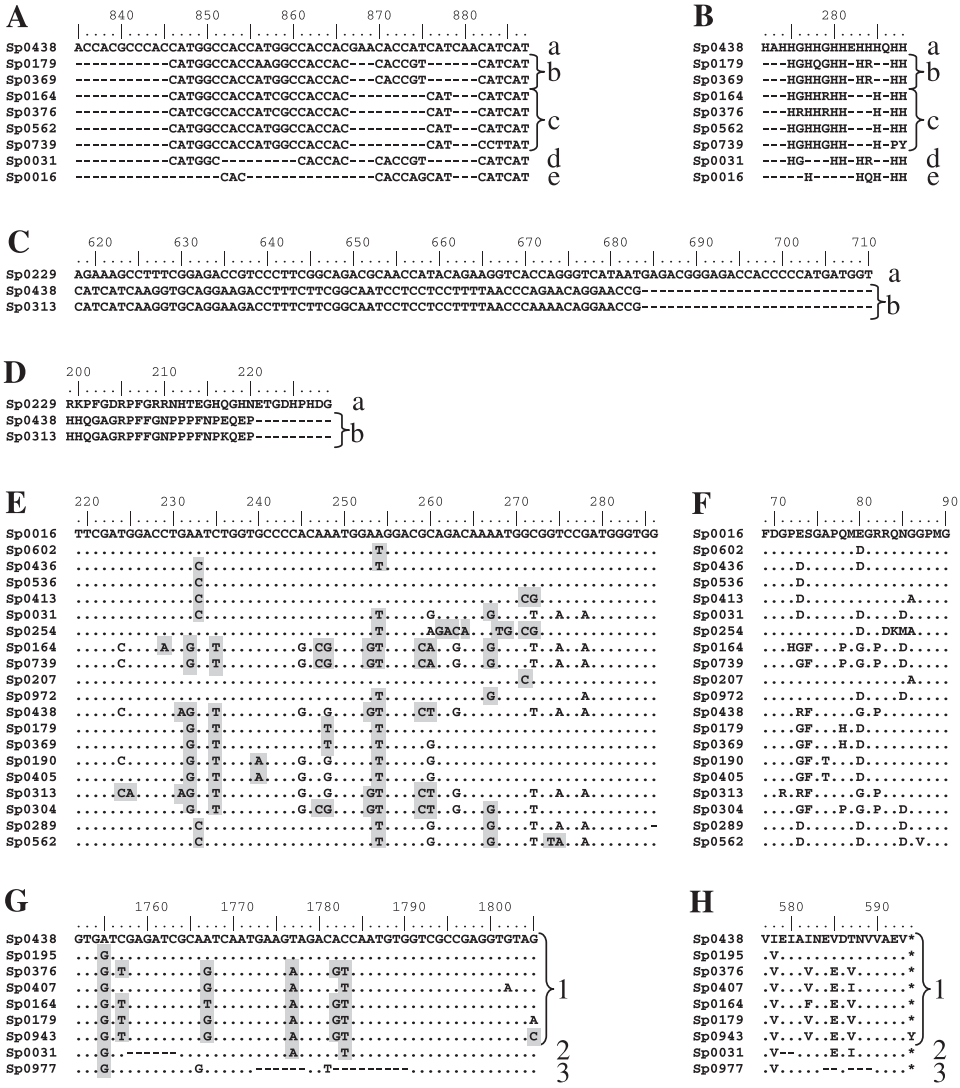

Fig. 2. Three types of sequence diversity in *185/333* cDNAs. *A* and *B*: alignment of subelements 15a–15e (11 unique sequences of 72 cDNAs) shows length variations that may be due to indels (*A*). The corresponding amino acid sequences (*B*) show variable numbers of histidines. *C* and *D*: alignment of subelements 11a and 11b (3 unique sequences of 11) shows divergent nucleotide (*C*) and amino acid sequences (*D*). *E* and *F*: alignment of element 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. 2*E*). 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. 2*F*).

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. 1*A*; Supplemental Figs. S1 and S2). SNPs at the 3′-end of elements 25a, -b, and -c sometimes resulted in stop codons (Fig. 1*A*; 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. 1*A*; 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. 1*A*). 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. 1*A*; 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. 3*A*). 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

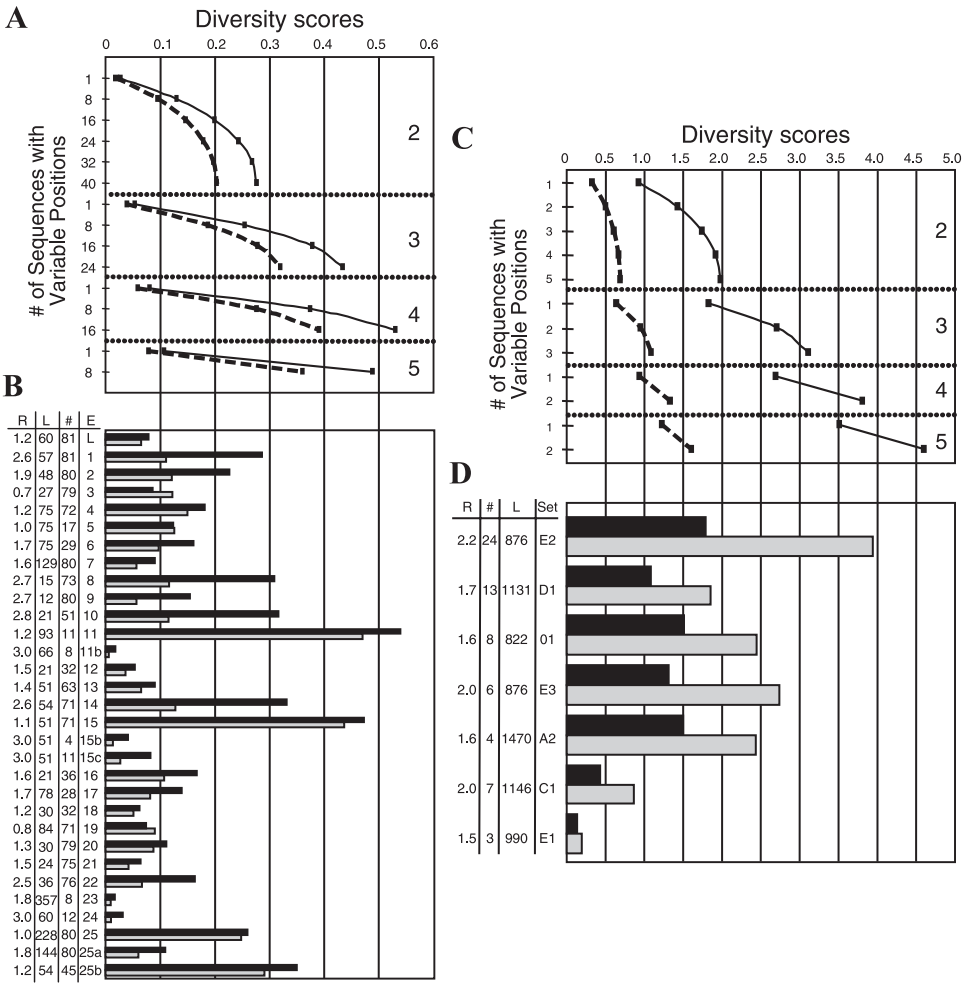

Fig. 3. Diversity scores for elements and cDNA sets. The parameters of the modeled alignments are based on the length and variability of the actual *185/333* elements. Nucleotide-modeled alignments are composed of 80 sequences that are 75 nucleotides long, of which 30% of the positions are variable. Amino acid-modeled alignments are composed of 80 sequences that are 25 amino acids long, with 40% variable positions. *A*: diversity scores for alignments of modeled nucleotide (dashed lines) and amino acid (solid lines) matrices. The no. of different bases present at each variable position is shown within the graph, at *right*. *B*: diversity scores of individual *185/333* elements. The diversity scores for alignments of nucleotides (black bars) and amino acids (gray bars) of *185/333* elements are shown on the *x*-axis. Table at *left* of graph indicates element (E) no. (L = leader), the length (*L*) of the element in nucleotides, and the number (#) of cDNAs in which a given element occurs. Ratios (*R*) of amino acid diversity vs. nucleotide diversity estimate nonsynonymous substitutions in the nucleotide sequences. Elements with diversity scores of 0 are omitted (11a, 15a, 15d, and 15e). *C*: diversity scores for modeled alignments (nucleotides, dashed lines; amino acids, solid lines) are based on cDNA sets that have identical element patterns. The number of different nucleotides present at each variable position is shown within the graph, at *right*. *D*: diversity scores of the cDNA sets that have identical element patterns (nucleotides, black bars; amino acids, gray bars). The cDNA set identifier (see Fig. 1*A*) is shown on the *y*-axis. *L*, #, and *R* are as in *B*.

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. 3*A*). This value was similar to the highest observed amino acid diversity score among the elements, element 11 (0.5381; Fig. 3*B*), in which subelements 11a and 11b were significantly different from each other (Fig. 2*A*). 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*, *01*, 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. 3*B*), 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.

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

Set Identifier	cDNAs per Set	Unique cDNA Sequences	dn/ds*
E2	24	15	1.2670
D1	13	10	1.3462
01	8	8	1.0230
E3	5	4	1.7745
A2	4	4	0.8909
C1	7	3	0.7313
E1	3	3	0.2767

*Under positive selection (>1.0).

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. 3*B*). 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 $\geq 83\%$ 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. 3*D*). 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.

185/333 Proteins

Structure. The deduced 185/333 proteins (with the exception of the truncated protein encoded by *D1.1*) shared an overall organization of an NH_2-terminal hydrophobic leader followed by a glycine-rich (gly-rich) region, a histidine-rich (his-rich) region, and a COOH-terminal region (Fig. 1*B*). 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 [*Sp0126* (*02*), 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. 1*B*). 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. 1*B*; 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. 1*B*; Supplemental Fig. S2). Five categories of repeats were located throughout the sequence (Fig. 1*B*; 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. 1*B*; Supplemental Table S1). Seven conserved O-linked glycosylation sites were identified within a short region of nine amino acids in element 25a (Fig. 1*B*; 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. 4*A*). 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. 4*A*). The seventh probe generated from *Sp0641* (*E5*) identified the same bands but displayed extra bands in some lanes (Fig. 4*B*). Comparisons of the cDNA sequences that were used to make the probes indicated that they were members of three cDNA sets (*E2*, *D1, 01*; 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. 4*A*). 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. 1*B*; 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.

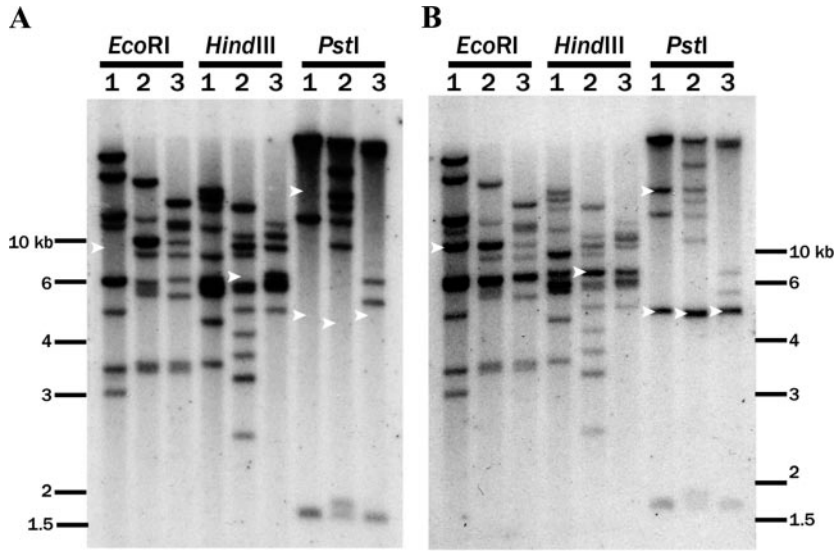

Fig. 4. Genome blots indicate that *185/333* genes are small. Genomic DNA (gDNA) from 3 sea urchins (*1–3*) was analyzed under high stringency with riboprobes produced from 7 full-length cDNAs and from elements 1 and 25b. *A*: identical banding pattern was observed for 6 full-length riboprobes and 2 single-element riboprobes. The full-length cDNA templates were *Sp0419* and *Sp0653* (same element pattern as *E2*), *Sp0761* (*D1*), and *Sp0254* (*01*), whereas *Sp0323* and *Sp0329* were not fully sequenced. *B*: the same pattern plus a few extra bands (white arrowheads) were identified by the riboprobe generated from *Sp0641* (*E5*). Size standards are shown to the side of each blot.

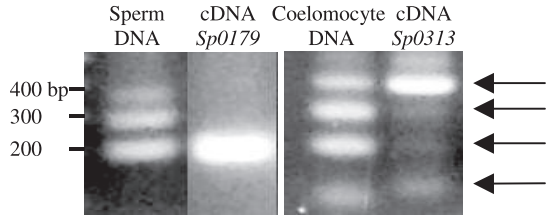

Fig. 5. Elements 1–7 of *185/333* genes are not separated by introns. Elements 1–7 were amplified by PCR from gDNA isolated from sperm or coelomocytes (different animals) with primers *185F2* and *185R5* (Table 1). Two cDNAs were used as controls. *Sp0313* (*pattern A1*) has elements 1–7, whereas *Sp0179* (*pattern B3*) is missing elements 5 and 6. The gDNA from 2 individuals amplified 3 or 4 fragments, respectively, as indicated by the arrows at *right*. Both sources of gDNA amplified fragments that were the same sizes as those from the cDNAs. Size standards are shown at *left*.

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. 1*A*). Consequently, the possibility of alternative splicing from cryptic splice sites within the open reading frame was investigated for all cDNAs. Consensus, canonical cryptic splice sites were identified at the boundaries of elements 4, 5, 6, 13, and 25b (Fig. 1*B*; 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. 6*A*) matched to the element pattern of *E2*, while the gene on *contig12130* (Fig. 6*B*) 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

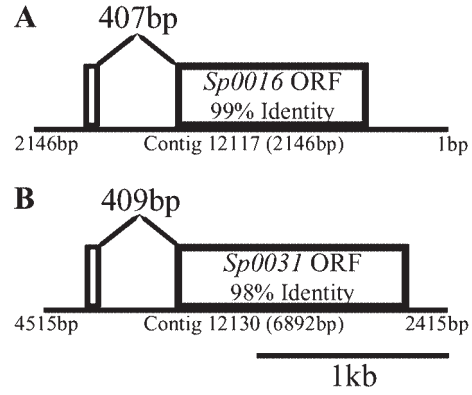

Fig. 6. *185/333* genes have 2 exons and 1 intron. Two contigs with full-length *185/333* genes were retrieved from the partially assembled sea urchin genome (assembly 09/22/03; **http://hgsc.bcm.tmc.edu/projects/seaurchin**) with matches to element patterns defined by *E2* (*A*) and *D1* (*B*). Horizontal line represents the contig from the genome database, while the boxes represent the 2 exons. Percent identities between the exons and the corresponding cDNA sequences are indicated within each box. The matches are not 100%, because the mRNA and gDNA were collected from different sea urchins.

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 most 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. 1*A*), 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

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

1. **Adema CM, Hertel LA, Miller RD, and Loker ES.** 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, 1997.
2. **Adler NM, Rogozin IB, Iyer LM, Glazko GV, Cooper MD, and Pancer Z.** Diversity and function of adaptive immune receptors in a jawless vertebrate. *Science* 310: 1970–1973, 2005.
3. **Brendel V, Bucher P, Nourbakhsh IR, Blaisdell BE, and Karlin S.** Methods and algorithms for statistical analysis of protein sequences. *Proc Natl Acad Sci USA* 89: 2002–2006, 1992.
4. **Burset M, Seledtsov IA, and Solovyev VV.** Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res* 28: 4364–4375, 2000.
5. **Cameron RA, Mahairas G, Rast JP, Martinez P, Biondi TR, Swartzell S, Wallace JC, Poustka AJ, Livingston BT, Wray GA, Ettensohn CA, Lehrach H, Britten RJ, Davidson EH, and Hood L.** A sea urchin genome project: sequence scan, virtual map, and additional resources. *Proc Natl Acad Sci USA* 97: 9514–9518, 2000.
6. **Cannon JP, Haire RN, and Litman GW.** Identification of diversified genes that contain immunoglobulin-like variable regions in a protochordate. *Nat Immun* 3: 1200–1207, 2002.
7. **Cannon JP, Haire RN, Rast JP, and Litman GW.** The phylogenetic origins of the antigen-binding receptors and somatic diversification mechanisms. *Immunol Rev* 200: 12–22, 2004.
8. **Cannon JP, Haire RN, Schnitker N, Mueller MG, and Litman GW.** Individual protochordates have unique immune-type receptor repertoires. *Curr Biol* 14: R465–R466, 2004.
9. **Clow LA, Raftos DA, Gross PS, and Smith LC.** The sea urchin complement homologue, SpC3, functions as an opsonin. *J Exp Biol* 207: 2147–2155, 2004.
10. **Cuthbertson BJ, Bullesbach EE, Fievet J, Bachere E, and Gross PS.** A new class (penaeidin class 4) of antimicrobial peptides from the Atlantic white shrimp (*Litopenaeus setiferus*) exhibits target specificity and an independent proline-rich-domain function. *Biochem J* 381: 79–86, 2004.
11. **Cuthbertson BJ, Yang Y, Bachere E, Bullesbach EE, Gross PS, and Aumelas A.** Solution structure of synthetic penaeidin-4 with structural and functional comparisons with penaeidin-3. *J Biol Chem* 280: 16009–16018, 2005.
12. **Durbin R, Eddy S, Krogh A, and Mitchison G.** *Biological Sequence Analysis, Probability Models of Proteins and Nucleic Acids*. Cambridge, UK: Cambridge Univ. Press, 1998.
13. **Flajnik MF and Du Pasquier L.** Evolution of innate and adaptive immunity: can we draw a line? *Trends Immunol* 25: 640–644, 2004.
14. **Gross PS, Al-Sharif WZ, Clow LA, and Smith LC.** Echinoderm immunity and the evolution of the complement system. *Dev Comp Immunol* 23: 429–442, 1999.
15. **Gross PS, Bartlett TC, Browdy CL, Chapman RW, and Warr GW.** Immune gene discovery by expressed sequence tag analysis of hemocytes and hepatopancreas in the Pacific White Shrimp, *Litopenaeus vannamei*, and the Atlantic White Shrimp, *L. setiferus*. *Dev Comp Immunol* 25: 565–577, 2001.
16. **Hall TA.** BioEdit: a user friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* 41: 95–98, 1999.
17. **Hertel LA, Adema CM, and Loker ES.** Differential expression of FREP genes in two strains of *Biomphalaria glabrata* following exposure to the digenetic trematodes *Schistosoma mansoni* and *Echinostoma paraensei*. *Dev Comp Immunol* 29: 295–303, 2005.
18. **Kumar S, Tamura K, and Nei M.** MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* 5: 150–163, 2004.
19. **Lee JJ, Shott RJ, Rose SJ 3rd, Thomas TL, Britten RJ, and Davidson EH.** Sea urchin actin gene subtypes. Gene number, linkage and evolution. *J Mol Biol* 172: 149–176, 1984.
20. **Lehmann P.** Structure and evolution of plant disease resistance genes. *J Appl Genet* 43: 403–414, 2002.
21. **Leonard PM, Adema CM, Zhang SM, and Loker ES.** 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, 2001.
22. **Litman GW, Cannon JP, and Dishaw LJ.** Reconstructing immune phylogeny: new perspectives. *Nat Rev Immunol* 5: 866–879, 2005.
23. **Litman GW, Cannon JP, and Rast JP.** New insights into alternative mechanisms of immune receptor diversification. *Adv Immunol* 87: 209–236, 2005.
24. **Loker ES, Adema CM, Zhang SM, and Kepler TB.** Invertebrate immune systems—not homogeneous, not simple, not well understood. *Immunol Rev* 198: 10–24, 2004.
25. **Meyers BC, Kaushik S, and Nandety RS.** Evolving disease resistance genes. *Curr Opin Plant Biol* 8: 129–134, 2005.
26. **Mondragon-Palomino M, Meyers BC, Michelmore RW, and Gaut BS.** Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. *Genome Res* 12: 1305–1315, 2002.
27. **Multerer KA and Smith LC.** Two cDNAs from the purple sea urchin, *Strongylocentrotus purpuratus*, encoding mosaic proteins with domains found in factor H, factor I, and complement components C6 and C7. *Immunogenetics* 56: 89–106, 2004.
28. **Nair SV, Del Valle H, Gross PS, Terwilliger DP, and Smith LC.** 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, 2005.
29. **Nixon KC.** *WinClada ver.1.00.08*. Ithica, NY: Nixon K. C., 1999–2002.
30. **Nowak TS and Loker ES.** *Echinostoma paraensei*: differential gene transcription in the sporocyst stage. *Exp Parasitol* 109: 94–105, 2005.
31. **O'Leary NA and Gross PS.** Genomic structure and transcriptional regulation of the penaeidin gene family from *Litopeneaus vannamei*. *Gene* 371: 75–83, 2006.
32. **Pancer Z, Amemiya CT, Ehrhardt GR, Ceitlin J, Gartland GL, and Cooper MD.** Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. *Nature* 430: 174–180, 2004.
33. **Pancer Z, Rast JP, and Davidson EH.** Origins of immunity: transcription factors and homologues of effector genes of the vertebrate immune system expressed in sea urchin coelomocytes. *Immunogenetics* 49: 773–786, 1999.
34. **Parniske M, Hammond-Kosack KE, Golstein C, Thomas CM, Jones DA, Harrison K, Wulff BB, and Jones JD.** Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the Cf-4/9 locus of tomato. *Cell* 91: 821–832, 1997.
35. **Parniske M and Jones JD.** Recombination between diverged clusters of the tomato Cf-9 plant disease resistance gene family. *Proc Natl Acad Sci USA* 96: 5850–5855, 1999.
36. **Rast JP, Pancer Z, and Davidson EH.** New approaches towards an understanding of deuterostome immunity. *Curr Top Microbiol Immunol* 248: 3–16, 2000.
37. **Rost B, Yachdav G, and Liu J.** The PredictProtein server. *Nucleic Acids Res* 32: W321–W326, 2004.
38. **Smith LC.** The complement system in sea urchins. *Adv Exp Med Biol* 484: 363–372, 2001.
39. **Smith LC.** Thioester function is conserved in SpC3, the sea urchin homologue of the complement component C3. *Dev Comp Immunol* 26: 603–614, 2002.
40. **Smith LC, Azumi K, and Nonaka M.** Complement systems in invertebrates. The ancient alternative and lectin pathways. *Immunopharmacology* 42: 107–120, 1999.
41. **Smith LC, Britten RJ, and Davidson EH.** Lipopolysaccharide activates the sea urchin immune system. *Dev Comp Immunol* 19: 217–224, 1995.
42. **Smith LC, Chang L, Britten RJ, and Davidson EH.** Sea urchin genes expressed in activated coelomocytes are identified by expressed sequence

tags. Complement homologues and other putative immune response genes suggest immune system homology within the deuterostomes. *J Immunol* 156: 593–602, 1996.
43. **Smith LC, Shih CS, and Dachenhausen SG.** Coelomocytes express SpBf, a homologue of factor B, the second component in the sea urchin complement system. *J Immunol* 161: 6784–6793, 1998.
44. **Terwilliger DP, Clow LA, Gross PS, and Smith LC.** Constitutive expression and alternative splicing of the exons encoding SCRs in *Sp152*, the sea urchin homologue of complement factor B. Implications on the evolution of the Bf/C2 gene family. *Immunogenetics* 56: 531–543, 2004.
45. **Tusnady GE and Simon I.** The HMMTOP transmembrane topology prediction server. *Bioinformatics* 17: 849–850, 2001.
46. **Tusnady GE and Simon I.** Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol* 283: 489–506, 1998.
47. **Watson FL, Puttmann-Holgado R, Thomas F, Lamar DL, Hughes M, Kondo M, Rebel VI, and Schmucker D.** Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science* 309: 1874–1878, 2005.
48. **Winnepenninckx B, Backeljau T, and De Wachter R.** Extraction of high molecular weight DNA from molluscs. *Trends Genet* 9: 407, 1993.
49. **Wu TT and Kabat EA.** An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J Exp Med* 132: 211–250, 1970.
50. **Yang Z.** PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13: 555–556, 1997.
51. **Yu C, Dong M, Wu X, Li S, Huang S, Su J, Wei J, Shen Y, Mou C, Xie X, Lin J, Yuan S, Yu X, Yu Y, Du J, Zhang S, Peng X, Xiang M, and Xu A.** Genes "waiting" for recruitment by the adaptive immune system: the insights from amphioxus. *J Immunol* 174: 3493–3500, 2005.
52. **Zhang SM, Adema CM, Kepler TB, and Loker ES.** Diversification of Ig superfamily genes in an invertebrate. *Science* 305: 251–254, 2004.
53. **Zhang SM, Leonard PM, Adema CM, and Loker ES.** 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, 2001.
54. **Zhang SM and Loker ES.** The FREP gene family in the snail *Biomphalaria glabrata*: additional members, and evidence consistent with alternative splicing and FREP retrosequences. Fibrinogen-related proteins. *Dev Comp Immunol* 27: 175–187, 2003.