Innate immune complexity in the purple sea urchin: diversity of the Sp185/333 system

L. Courtney Smith*

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

INTRODUCTION

SEA URCHINS AND ECHINODERMSThe phylum Echinodermata is phylogenetically positioned within the deuterostome lineage of animals that includes the Chordata and a few minor invertebrate phyla. There are five extant classes of echinoderms, of which the echinoid class includes sea urchins and sand dollars (Figure 1). Immunologists initially assumed that all animals had adaptive immune functions like mammals and worked toward demonstrating these characteristics in many species of animals including invertebrates. Initial investigations to determine the immune capabilities in echinoderms employed allograft rejection assays (Hildemann and Dix, 1972; Karp and Hildemann, 1976; Coffaro and Hinegardner, 1977; Coffaro, 1980; Varadarajan and Karp, 1983). However, the rejection kinetics of first set, second set, and third party allografts in the sea urchin, Lytechinus pictus (Coffaro and Hinegardner, 1977; Coffaro, 1979), showed that the response did not demonstrate specific immune recognition and immune memory and was therefore solely innate (Smith and Davidson, 1992). The initial assumptions that immunologists held about universal functions of adaptive immunity were very wrong.

The cellular mediators of immunity in echinoderms are coelomocytes that are present in the fluid that fills the coelomic cavity (Booolootian and Geise, 1958; Johnson, 1969a; Smith, 1981; reviewed in Smith et al., 2010). Coelomocyte immune function was first recognized by Metchnikoff (1893) when he inserted rose prickles, glass rods, and bacteria into the blastocoel of larval sea stars and observed either encapsulation or phagocytosis by the blastocoelar cells (larval equivalents of coelomocytes). More recent analysis of the clearance of foreign cells and particles from adult and larval echinoderms shows that it is swift, efficient, and is mediated by the phagocytic cells in the coelomic and blastocoelar cavities (Reinisch and Bank, 1971; Yui and Bayne, 1983; Plytycz and Seljelid, 1993; Silva, 2000; Furukawa et al., 2009). In the purple

The California purple sea urchin, Strongylocentrotus purpuratus, is a long-lived echinoderm with a complex and sophisticated innate immune system. There are several large gene families that function in immunity in this species including the Sp185/333 gene family that has ~50 (±10) members. The family shows intriguing sequence diversity and encodes a broad array of diverse yet similar proteins. The genes have two exons of which the second encodes the mature protein and has repeats and blocks of sequence called elements. Mosaics of element patterns plus single nucleotide polymorphisms-based variants of the elements result in significant sequence diversity among the genes yet maintains similar structure among the members of the family. Sequence of a bacterial artificial chromosome insert shows a cluster of six, tightly linked Sp185/333 genes that are flanked by GA microsatellites. The sequences between the GA microsatellites in which the Sp185/333 genes and flanking regions are located, are much more similar to each other than are the sequences outside the microsatellites suggesting processes such as gene conversion, recombination, or duplication. However, close linkage does not correspond with greater sequence similarity compared to randomly cloned and sequenced genes that are unlikely to be linked. There are three segmental duplications that are bounded by GAT microsatellites and include three almost identical genes plus flanking regions. RNA editing is detectible throughout the mRNAs based on comparisons to the genes, which, in combination with putative post-translational modifications to the proteins, results in broad arrays of Sp185/333 proteins that differ among individuals. The mature proteins have an N-terminal glycine-rich region, a central RGD motif, and a C-terminal histidine-rich region. The Sp185/333 proteins are localized to the cell surface and are found within vesicles in subsets of polygonal and small phagocytes. The coelomocyte proteome shows full-length and truncated proteins, including some with missense sequence. Current results suggest that both native Sp185/333 proteins and a recombinant protein bind bacteria and are likely important in sea urchin innate immunity.

Keywords: echinoid, invertebrate, evolution, innate immunity, coelomocyte, microsatellites, RNA editing, gene family

Sp185/333 elements result in significant sequence diversity among the genes yet maintains similar structure among the members of the family. Sequence of a bacterial artificial chromosome insert shows a cluster of six, tightly linked Sp185/333 genes that are flanked by GA microsatellites. The sequences between the GA microsatellites in which the Sp185/333 genes and flanking regions are located, are much more similar to each other than are the sequences outside the microsatellites suggesting processes such as gene conversion, recombination, or duplication. However, close linkage does not correspond with greater sequence similarity compared to randomly cloned and sequenced genes that are unlikely to be linked. There are three segmental duplications that are bounded by GAT microsatellites and include three almost identical genes plus flanking regions. RNA editing is detectible throughout the mRNAs based on comparisons to the genes, which, in combination with putative post-translational modifications to the proteins, results in broad arrays of Sp185/333 proteins that differ among individuals. The mature proteins have an N-terminal glycine-rich region, a central RGD motif, and a C-terminal histidine-rich region. The Sp185/333 proteins are localized to the cell surface and are found within vesicles in subsets of polygonal and small phagocytes. The coelomocyte proteome shows full-length and truncated proteins, including some with missense sequence. Current results suggest that both native Sp185/333 proteins and a recombinant protein bind bacteria and are likely important in sea urchin innate immunity.
sea urchin, there are three classes of coelomocytes: phagocytes, spherule cells, and vibratile cells (Smith et al., 2010; Figure 2). The phagocytes are composed of three morphotypes that vary in size and shape. The large phagocytes include the polygonal and discoidal cells that have different morphologies when spread on glass (Henson et al., 1992, 1999; Edds, 1993). The cytoskeleton of the polygonal phagocytes is shaped by parallel bundles of actin that give the cell an angular, polygonal shape. Discoidal phagocytes, which are slightly smaller than the polygonal cells, have bundles of actin that are aligned radially from the nucleus to the periphery resulting in a disk-like or “fried egg” morphology (Figures 2 and 3). Small phagocytes are significantly smaller than the large phagocytes and have perpetual filopodial morphology (Gross et al., 2000; Brockton et al., 2008; Figures 2–4). Large phagocytes, on the other hand, readily modify their morphology and appear as lamellipodial or filopodial depending on the microenvironmental conditions of the coelomic cavity (Edds, 1977, 1979, 1993; Henson et al., 1999). The spherule cells include cells with colorless inclusions and cells with red spherules that contain echinochrome A, which has antibacterial and antifungal activity (Johnson and Chapman, 1970; Messer and Wardlaw, 1980; Service and Wardlaw, 1984; Calestani et al., 2003). The vibratile cells have a single flagellum, are highly motile, and may be involved with clotting reactions (Johnson, 1969b; Bertheussen and Seijelid, 1978; Sacchi and Smith, unpublished). Immune functions mediated by coelomocytes include phagocytic activities of the large phagocytes, encapsulation by many types of cells, degranulation of echinochrome A from red spherule cells in response to bacterial contact (Johnson, 1969b; Johnson and Chapman, 1970) and injury (Coffaro and Hinegardner, 1977), and secretion of the complement C3 homolog by a subset of the discoidal cells (Gross et al., 2000; reviewed in Smith et al., 2010). Unfractionated coelomocytes from the sea urchin Paracentrotus lividus, show cytotoxic activity against red blood cells and K562 tumor cells, a function that may be mediated by the colorless spherule cells that is augmented by the presence of phagocytes (Arizza et al., 2007).
Early work to analyze gene expression in coelomocytes under conditions of immune challenge identified homologs of complement components including putative complement regulatory proteins, in addition to genes encoding transcription factors, lectins, ions channels, a Tie receptor homolog, lysosomal enzymes, cytoskeletal proteins, folding chaperones, mitochondrial enzymes, proteins that function in RNA splicing, signaling pathways, and secretion, plus a large number of unknowns (Smith et al., 1996, 1998; Al-Sharif et al., 1998; Rast et al., 2000; Multerer and Smith, 2004; Nair et al., 2005; Stevens et al., 2010). Many of the expressed sequence tags (ESTs) correlate with gene models annotated in the genome of the purple sea urchin, although the major finding from genome annotation was the striking level of complexity and sophistication of this invertebrate innate immune system (Hibino et al., 2006). For example, large gene families that are homologous to those in vertebrates that function in innate immune detection, such as the Toll-like receptors, NOD-like receptors and scavenger receptors are greatly expanded in the purple sea urchin (Rast et al., 2006; Rast and Messier-Solek, 2008). One gene family that is highly up-regulated in response to immune challenge is the Sp185/333 family (Rast et al., 2000; Nair et al., 2005). This family has been estimated to have about 50 ± 10 gene loci using a variety of methods including quantitative PCR of genomic DNA, a statistical estimate based on the number of unique genes cloned and sequenced from three sea urchins, and an estimate from screens of two arrayed bacterial artificial chromosome (BAC) libraries (Terwilliger et al., 2006; Buckley et al., 2008a; Ghosh et al., 2010). BLAST analysis of 1025 randomly chosen cDNA clones that were up-regulated in response to lipopolysaccharide (LPS) showed that 73% matched to two sequences on GenBank; DD185 (Rast et al., 2000) and EST333 (Smith et al., 1996). Hence the name for the set of cDNAs, Sp185/333, which was chosen because none of the sequences matched to proteins with known function and homologs have not been identified in organisms outside of echinoids.

THE Sp185/333 GENES

GENE STRUCTURE

Initial speculation on the Sp185/333 gene structure was based on the extraordinary structure of the cDNAs (Nair et al., 2005). Optimal amino acid alignments of the translated cDNAs required artificial insertions of large gaps (see below), which seemed consistent with extensive alternative splicing of transcripts from one or a few large genes with many exons, similar to that for Dscam in arthropods (Schmucker et al., 2000; Watson et al., 2005; Dong et al., 2006; Brites et al., 2008; Schmucker and Chen, 2009). This prediction of large genes with many exons might have been consistent with Southern blots if all of the positive bands had been large (Figure 5). However, genomic DNA digested with PstI released small fragments that were hybridized by probes from both the 5’ and 3’ ends of the cDNAs (Terwilliger et al., 2006). Furthermore, small amplicons were produced by PCR of genomic DNA using primers that hybridized to the ends of the cDNAs (Buckley and Smith, 2007). These results suggested that the cDNAs were expressed from many small genes rather than a few large genes with many exons. Accordingly, preliminary genome sequences from the purple sea urchin confirmed that the Sp185/333 genes were small (1.2–2 kb) and had only two exons (Terwilliger et al., 2006; Figure 6). The first exon is very short (51 or 54 nucleotides),
encodes the hydrophobic leader, and is followed by an intron of 380–413 nucleotides. The second exon encodes the mature protein and has the same unusual structure as the cDNAs, which is evident from alignments of the genes (Figure 6). Our initial speculation that alternative splicing generated the structures of the cDNAs is clearly improbable given that the genes only have two exons. Furthermore, splicing signals that might function within the intron or second exon to generate alternatively spliced transcripts are not present in the Sp185/333 sequences (Buckley et al., 2009).

TWO ALIGNMENTS
The gaps that are required to align optimally the Sp185/333 genes and cDNAs define recognizable blocks of sequence called elements (Figure 6). There are 25–27 different elements depending on the alignment that are shared among genes and cDNAs. They are variably present or absent within different genes and cDNAs resulting in recognizable mosaics of elements, called element patterns. Note that elements are not the result of alternative splicing during processing of the mRNAs, but are present within the second exon of the genes. Based on alignments of 121 unique gene sequences (of 171 sequenced genes), 31 different element patterns have been identified from three sea urchins (Buckley and Smith, 2007). When both genes and cDNA sequences from 16 individuals are evaluated together, 51 different element patterns can be identified. Although each type of element has a recognizable sequence, there are variants of each type that show sequence diversity. There can be between 1 and 28 (average = 11/element) different sequence versions for a given element based on genes cloned from individual animals (Buckley et al., 2008a). In addition to the elements present in both genes and cDNAs, there are three types of element specific to (A) which are defined by the location of three possible stop codons. A common, single nucleotide RNA edit in element pattern E2 alters a glycine codon to a stop (X) in element 13 and is denoted as an E2.1 pattern. Repeats, shown at the bottom of (A, B), are shown as different colors and occur as tandem repeats and interspersed tandem repeats. The genes have two exons (brackets at the top of each alignment), of which the first encodes the leader and the second encodes the mature protein including all of the elements. A single intron of ~400 nt is positioned between the leader and the first element and is not shown. Modified from Ghosh et al. (2010) with permission from Elsevier.
to the complexity of the elements and element patterns, there are six types of repeats in the second exon of the genes (Nair et al., 2005; Terwilliger et al., 2006; Buckley et al., 2008a). Depending on the gene, there are two to four imperfect type 1 repeats near the 5’ end of the second exon that are direct, tandem repeats (Figure 6). Toward the 3’ end of the exon, there are five additional types of repeats that occur in an interspersed and partially duplicated pattern. Because of the repeats, the Sp185/333 sequences can be aligned in two different ways that are equally optimal. Initially, the alignment was done with cDNA sequences and no attempt was made to correlate the elements with the repeats (Terwilliger et al., 2006, 2007; Figure 6A). The second alignment was done with both cDNA and gene sequences and efforts were made to make the repeats and elements correspond as much as possible (Buckley and Smith, 2007; Figure 6B). The two alignments are somewhat similar for the type 1 tandem repeats, but are quite different in the region of the interspersed repeats near the 3’ end of the exon with different numbers of interspersed repeats in the two alignments. Comparisons between the two alignments shows that some of the sequences could be interdigitated, essentially collapsing some of the repeats together, which reduced their numbers in the repeat-based alignment. (Compare the interspersed repeats shown in Figures 6A,B. For a detailed illustration, see Buckley and Smith, 2007 and Figure 7 within that publication.) Although the number of elements in the two alignments differ because of the collapse of some of the repeats (25 vs. 27 elements), the overall length of the two alignments is similar and both are about equally optimal (Buckley and Smith, 2007).

ORIGINS OF GENE STRUCTURE

The extraordinary sequence diversity of the Sp185/333 gene family is due primarily to the mosaic element patterns in the second exon. Sequence variations within individual elements adds a second level of sequence diversity to the genes. The diverse structure of the genes is extraordinary and comparisons of full-length sequences among three sea urchins has not identified an identical shared gene (Buckley et al., 2008a). Yet surprisingly, 28% of individual element sequences are shared among genes from two or more animals (Figure 7). This paradox of element sequences shared among genes, but genes that are not shared among individuals, is quite remarkable and hints at an intriguing mechanism to generate the diversity within this gene family. Initial efforts to determine whether the type 1 repeats might be involved in generating diversity predicted three theoretical ancestral type 1 repeats (5’ end of the second exon; Buckley et al., 2008a). A computational estimate of the evolutionary history suggests that the extant type 1 repeats may have been derived from these ancestral sequences through modification by recombination, duplication, deletion, and point mutations. In the repeat-based alignment, the five types of interspersed repeats in the 3’ end of the exon are present in a block that is repeated twice with an extra type 4 repeat suggesting origins from duplication events (Figure 6B). Because there is no correlation between the patterns of tandem type 1 repeats and the patterns of the interspersed repeats in the 3’ end of the exon, this suggests swift recombination among genes plus recombination among the repeats within the genes. Several examples of putative gene recombination could be identified because different regions of a few genes matched optimally to corresponding regions of two other genes. Because the recombination rate among Sp185/333 genes may be quite high, identifying the ancestral genes that contributed sequence to a given recombinant gene is quite difficult. Consequently, sequences evaluation to identify recombination sites or hotspots using incongruence length difference (ILD; Michevich and Ferris, 1981) and incongruence permutation (IP; Huelsenbeck and Bull, 1996) both show that recombination is likely to occur at any position within and between elements and repeats and that there are no recombination hotspots (Buckley et al., 2008a). The level of putative gene recombination among members of the Sp185/333 gene family is swift, appearing surprisingly similar to somatic recombination of gene segments that is known to occur for T cell receptor genes (Stiu et al., 1984). Given that these genes appear to encode proteins with immune effector function, ongoing diversification, based in part by gene recombination, will be advantageous for the host in the arms race with pathogens.

AGE ESTIMATE FOR THE GENE FAMILY

Given the apparently swift recombination of the Sp185/333 genes, molecular clock analysis of the evolutionary history of the Sp185/333 family suggests that the last common ancestral Sp185/333 sequence of the extant genes may have been present in the population of purple sea urchins 2.7–10 million years ago (Ghosh et al., 2010). This time frame overlaps with the estimated time period when Strongylocentrotus purpuratus diverged from its sister species (Grula et al., 1982; Biermann, 1998; Lee, 2003; Buckley et al., 2008a). When trace sequences of 185/333 genes from the genomes of two other sea urchins, S. franciscanus and Allocentrotus fragilis are compared to the Sp185/333 sequences from S. purpuratus, molecular clock analysis shows that the divergence times for these three gene families appear older than the
three species of sea urchins in which they function. This apparent paradox is another illustration of a rapid rate of evolution of the 185/333 genes in all three species, a characteristic that is consistent with immune function, participation in an arms race with pathogens, and mechanisms to promote sequence diversification. All three of these echinoid species live in the marine environment at the intersection of water and sediment. This habitat is estimated to have $10^7$–$10^8$ microbes per milliliter of sea water with two or three orders of magnitude more in the sediment (Massana et al., 1997; Llobet-Brossa et al., 1998; Whitman et al., 1998). It is likely that some of the microbes that co-habit with sea urchins are pathogens or opportunists and function as selection pressure to drive the diversification of the Sp185/333 gene family and the complexity of the innate immune system in these invertebrates.

**SIZE AND ORGANIZATION OF THE Sp185/333 GENE FAMILY**

The three different approaches used to estimate the size of the Sp185/333 family included qPCR of genomic DNA, a computational likelihood method to evaluate sequenced genes, and screens of two BAC libraries. Results from each evaluation was in general agreement and provided a family size of $\sim$50–10 genes (Buckley et al., 2008a; Ghosh et al., 2010). Quantitative PCR of genomic DNA samples from three sea urchins indicated 40–60 genes (Terwilliger et al., 2006). A computational likelihood method evaluates the probability of sequencing the same gene twice given the total number of genes cloned and sequenced and resulted in an estimate of 45–71 genes (Buckley et al., 2008a). Screens of two BAC libraries for clones with Sp185/333 sequences estimated a haploid genomic region of 200–250 kb in which the genes might be located, given the fold coverage of the genome estimated a haploid genomic region of 200–250 kb in which the genes might be located, given the fold coverage of the genome estimated a haploid genomic region of 200–250 kb in which the genes might be located, given the fold coverage of the genome estimated a haploid genomic region of 200–250 kb in which the genes might be located, given the fold coverage of the genome estimated a haploid genomic region of 200–250 kb in which the genes might be located, given the fold coverage of the genome estimated a haploid genomic region of 200–250 kb in which the genes might be located, given the fold coverage of the genome estimated a haploid genomic region of 200–250 kb in which the genes might be located, given the fold coverage of the genome. If gene duplication and/or gene conversion functions in generating the diversity observed within the Sp185/333 gene family, closely linked genes would be expected to show more sequence similarity than unlinked genes. However, the diversity of the linked genes on the 7096 BAC is not different from the diversity of 121 unique genes that have been randomly isolated and sequenced from three sea urchins (Miller et al., 2010). Proximity is not associated with sequence similarity, which also holds true for the three linked D1 genes in the BAC compared to randomly cloned and sequenced D1 genes from three different animals. This suggests that if gene conversion is involved with sequence diversification, genes may be sharing sequences from the entire family rather than just those that are most tightly linked and/or that conversion tracts may be shorter than the length of a gene.

With the availability of the sequence for the intergenic regions between the Sp185/333 genes on the BAC assembly, additional oddities, and similarities were discovered. All six genes are flanked by microsatellites on both the 5’ side (30–60 GA repeats) and the 3’ side (140–165 GA repeats; Figure 8; Miller et al., 2010). In addition, there are GAT microsatellites that surround the D1 genes and are positioned at the edges of three $\sim$13.5 kb tandem segmental duplications. These duplications have very low sequence diversity including the three D1 genes that are 99.7% identical. This not only suggests a recent duplication event of the regions bounded by GAT microsatellites, but it also infers the potential of an unknown
number of identical Sp185/333 genes in any given genome. Because previous estimates of family size (see above) have assumed that all genes have different sequences, the demonstration of genes on the 7096 BAC that are almost identical indicates that our estimates of family size are likely to be too low.

Speculation on diversification mechanisms that act on the Sp185/333 gene family is based on a number of unusual aspects of the clustered genes including sequence similarity, repeats, tight linkage, and microsatellites. Instability has been documented for regions of genomes that harbor tandem gene arrays, microsatellites, minisatelites, and other types of repeats (Gendrel et al., 2006; Wilson et al., 2000; Trifonov, 2003; Bagshaw et al., 2008; Smith, 2010; Despons et al., 2011). It is noteworthy that D1 is the most common element pattern identified from randomly cloned and sequenced genes (Buckley and Smith, 2007), which corresponds with multiple D1 genes on the BAC insert that are positioned within segmental duplications. This suggests that the D1 genes in other regions of the genome may also be surrounded by GAT microsatellites that drive segmental duplications. The sequence between the GA microsatellites, which includes the Sp185/333 genes and their flanking regions, has much lower sequence diversity than the intergenic regions positioned outside of the microsatellites (Miller et al., 2010). The sequence conservation may be driven by the presence of the microsatellites on either side of the genes plus element sequences that are shared within the second exons. The shared elements may promote gene recombination, conversion, and crossover, whereas the microsatellites may block the progression of DNA strand exchange during gene conversion, thereby protecting the region of the genome harboring clusters of Sp185/333 genes from sequence homogenization (Gendrel et al., 2006; Miller et al., 2010). This model predicts that the exchanged sequences would be limited to the regions located between microsatellites, which would show higher sequence conservation. In general, the genomic structure of the family predicts mechanisms for sequence homogenization through gene duplication and/or conversion, yet maintains sequence diversity among the members of the family. Other families of clustered genes tend to generate pseudogenes, such as primate odorant receptors, sea urchin TLRs, and fish aquaporins (Gilad et al., 2003; Rast and Messier-Solek, 2008; Dong et al., 2009; Tingaud-Sequeira et al., 2010) however, only a single pseudogene has been identified from the 171 sequenced Sp185/333 genes (Buckley and Smith, 2007). The Sp185/333 gene family appears to employ intriguing mechanisms for sharing sequences, maintaining diversification, and blocking or correcting disruption of the open reading frame.

THE DIVERSITY OF THE Sp185/333 mRNAs
The striking up-regulation of Sp185/333 gene expression in response to a variety of immune challenges was the initial basis for our interest in investigating this immune response system. The genes are expressed in coelomocytes within 3–6 h of challenge from LPS, β-1,3-glucan, double stranded RNA (dsRNA), peptidoglycan (PGD), and heat-killed marine bacteria (Rast et al., 2006; Nair et al., 2005; Terwilliger et al., 2007). Blastocoelar cells, larval phagocytes in S. purpuratus, express the Sp185/333 genes when cultured with marine microbes (Ghosh et al., 2010). The coding regions of the Sp185/333 mRNAs range in size from 0.16 to 1.5 kb, depending on the element pattern. In addition to the wide variety of element patterns encoded in the genes, half of the mRNAs show changes to the reading frame, compared to the genes, and appear as single nucleotide polymorphisms (SNPs) that encode non-synonymous amino acids or introduce early stop codons. Furthermore, small indels shift the reading frame to mis-sense sequence, typically ending with an early stop (Terwilliger et al., 2007; Ghosh et al., 2010). Unexpectedly, rather than being degraded as non-functional mRNAs, at least some are translated because both truncated and missense Sp185/333 proteins are produced (Dheilly et al., 2009). The most common element pattern observed in the cDNAs is E2, which is also present in a truncated version, E2.1, that has a SNP in element 13 that changes a glycine to a stop and deletes about half of the protein (Terwilliger et al., 2006, 2007; see Figure 6A). Prior to immune challenge, 81% of the sequenced cDNAs encode a truncated protein with the E2.1 pattern, whereas after challenge this is decreased to 58% (Ghosh et al., 2011). This variation can be interpreted as directed or regulated RNA editing rather than a change in gene expression because no gene has been identified with a stop codon in element 13.

Although many mRNAs encode truncated and missense proteins (Terwilliger et al., 2007), this is strikingly different from the sequenced genes of which all but one (of 171) have perfect open reading frames (Buckley and Smith, 2007). More detailed comparisons of genes and mRNAs from three individual sea urchins show a number of additional unexpected findings (Buckley et al., 2008b). The sequences of the genes and mRNAs both show high sequence diversity, which is present throughout and is not concentrated in diversity hotspots. However, for individual animals, the level of sequence diversity for genes is different from that for mRNAs. Similarly, the level of sequence diversity for elements that are shared between genes and mRNAs is also different. This means that the most variable nucleotide positions in the genes do not correspond with the most variable positions in the mRNAs. Very few of the mRNAs isolated from an individual animal match identically to any of the genes cloned from the same animal. Furthermore, estimates to identify the most likely gene as the source for sets of messages that share the same element pattern (based on the fewest number of substitutions and indels) show that 67–98% of the messages from individual sea urchins match best to a single gene with an E2 element pattern. Comparisons between the variable positions in the gene and its most likely set of transcribed messages indicate that 73% of the changes are transitions, significantly higher than the expected 33% that would occur randomly. Furthermore, 30% of these transitions change a cytidine in the gene to uridine in the mRNA. This is consistent with RNA editing by a cytidine deaminase, a family of enzymes for which several gene models have been identified in the genome (Hibino et al., 2006). In addition to possible RNA editing, transcription of the Sp185/333 genes may be performed by a low fidelity polymerase, such as polμ, which has also been annotated in the genome (Hibino et al., 2006).

THE Sp185/333 PROTEINS
STRUCTURE
The structure of the Sp185/333 proteins, as deduced from the genes and mRNAs, have an N-terminal hydrophobic leader that is
encoded by the first exon and is assumed to be cleaved during processing. The mature proteins are encoded by the second exon and have a generic structure even though there is significant sequence diversity among them (Terwilliger et al., 2006; Figure 9). The N-terminal region of the mature, full-length proteins is glycine-rich and includes two to four imperfect tandem repeats, and the central region has an integrin binding motif; arginine–glycine–aspartic acid (RGD). In addition to the interspersed, imperfect, tandem repeats, the C-terminal region is histidine-rich with patches of poly-histidines that are interspersed with proline, glycine, arginine, and glutamine, and patches of acidic amino acids composed of short stretches of glutamic acid often combined with aspartic acid (Terwilliger et al., 2006). Up to three stop codons are present in conserved positions at the 3′ end of the genes, and alter the length of the C-terminus for different proteins. Conserved, predicted N-linked glycosylation sites are present throughout the length of the proteins, and O-linked sites are predicted in those sequences that employ the third stop codon and have the longest C-terminal region (Terwilliger et al., 2006). The deduced proteins have no recognizable transmembrane region, no conserved motif for a glycosylphosphatidylinositol (GPI) linkage, and no apparent means to be associated with membranes. The native proteins have no predictable secondary structure or folding characteristics, and consequently may be present in an extended, unfolded, or intrinsically disordered conformation. The amino acid sequence predictions from the genes indicate that the proteins have no cysteines, although frame shifts introduced by RNA editing predicts missense cysteines in some truncated proteins (Terwilliger et al., 2006, 2007). The deduced protein contains a leader (L), a glycine-rich region (orange line), and a histidine-rich region (blue line). Symbols indicate the presence of an RGD motif within element 7 (red star with a yellow center), N-linked glycosylation sites (green heptagons), an O-linked glycosylation site (pink heptagon), patches of acidic amino acids (red arrows), and histidines (purple arrows). Modified from Terwilliger et al. (2006).

**Table 1 | Mass spectrometric identification of native Sp185/333 proteins from coelomic fluid**.

<table>
<thead>
<tr>
<th>Sp185/333 sequence</th>
<th>Element</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFNERREKENDTERGQGGFGGRPGMMGMPGP</td>
<td>1–2</td>
</tr>
<tr>
<td>RGRGGGFFGGRPGGMGMPGQDGPGMPG</td>
<td>1–2–3</td>
</tr>
<tr>
<td>RFDGEPASAPQM[EJ]GRRQQNGPMGGR</td>
<td>4</td>
</tr>
<tr>
<td>MGGRhnstPeFGGSRPDBGAG</td>
<td>6–7</td>
</tr>
<tr>
<td>(Overlap with above)FGGSRPDBGAGGRRPGQGGR</td>
<td>6–7</td>
</tr>
<tr>
<td>RGDGEETDAQQIGDGLGPGDFGHPHIHJR</td>
<td>7–8–9–10</td>
</tr>
<tr>
<td>PQTQDRNRLVSATKAAMRM</td>
<td>13–14</td>
</tr>
<tr>
<td>MAVLTLATMAATTIIATTQVKT</td>
<td>14–15</td>
</tr>
<tr>
<td>KFGGDHPFGR</td>
<td>23</td>
</tr>
<tr>
<td>adVEIAVNEEDVN</td>
<td>25b</td>
</tr>
</tbody>
</table>

*Amino acid sequences of peptides identified by LC–MS/MS from gel slices after one dimensional separation of coelomic fluid proteins from three different sea urchins as reported in Dheilly et al. (2009). Modified from Table 1 in Dheilly et al. (2009).

1 Element in which the sequence is present. For description of elements, see Buckley and Smith (2007), Terwilliger et al. (2007), and Figure 6A.

2 Exact matches were not found from the translated cDNAs reported by Terwilliger et al., 2007, see Supplemental Table 2). Lower case letters indicate mismatched amino acids between proteomic result and deduced sequence.

3 Missense sequence matches to the amino acid sequence predicted from one cDNA (GenBank accession number; EF065834) reported by Terwilliger et al. (2007).

4 Missense sequence matches to amino acid sequences predicted from five cDNAs (GenBank accession numbers; EF066834, EF065907, EF066020, EF066222, EF066214) reported by Terwilliger et al. (2007).

**DIVERSITY**

The combination of the mosaic structure of the elements plus RNA editing and the introduction of early stop codons results in a deduced size range of the Sp185/333 proteins of 4–55 kDa (Terwilliger et al., 2006, 2007). This includes varying numbers of histidine and acidic patches, the presence or absence of the RGD motif, and the deletion of the entire C-terminal histidine-rich region. However, the observed size range of the native proteins is unexpectedly large relative to deduced predictions, and likely results in part, from multimerization (Brockton et al., 2008). The presence of truncated and missense proteins in coelomic fluid is demonstrated on two dimensional (2D) Western blots by fewer spots for antisera that recognize the central and C-terminal regions of the proteins (Dheilly et al., 2009; Figure 10). Multimerization and putative post-translational modifications are likely underlying mechanisms for generating up to 264 different Sp185/333-positive (Sp185/333+) spots for individual sea urchins (Figure 11). Variations in the arrays of Sp185/333 proteins evaluated from different animals are striking (Figure 12) and suggestive of variable gene expression, perhaps differences in the numbers and/or versions of the genes in different individuals, in addition to mRNA editing and post-translational modifications to the proteins. It is apparent that the array of proteins from individual sea urchins shows more
FIGURE 10 | Truncated Sp185/333 proteins are present in the coelomic fluid. Enlarged regions of three different two dimensional Western blots of coelomic fluid proteins from the same sea urchin were analyzed with different anti-Sp185/333 sera: anti-66, anti-68, or anti-71. Anti-66 recognizes AHAQRDFNERRGKENDTER from element 1; anti-68 recognizes GRRGDEETDAQQIGDGLC from element 7; anti-71 recognizes TEEGSPRRGDRPPYGNR from element 25 (see Figure 6A for element positions). Decreasing numbers of spots in blots analyzed with antisera that recognize peptides in more C-terminal regions of the proteins suggests that many are either truncated or have missense sequence toward the C-terminus. Reprinted from Dheilly et al. (2009) with permission from the American Association of Immunologists, Inc., copyright 2009.

FIGURE 11 | A single sea urchin can have as many as 264 spots that are Sp185/333+. Coelomic fluid proteins (200 μg) from sea urchin 12 (see Figure 12) were separated by 2D electrophoresis and transferred to a filter. The filter was immunostained with an equal mixture of the three anti-Sp185/333 sera (see legend for Figure 10) and exposed to autoradiographic film for 1, 5, or 10 min. The different exposures were merged to give a final composite image. Isoelectric points (pI) are shown at the top and the molecular weight standards (kDa) are shown to the left. Reprinted from Dheilly et al. (2009) with permission from the American Association of Immunologists, Inc., copyright 2009.

variation than can likely be encoded by a gene family of \(\sim 50\) members.

**EXPRESSION**

Subsets of the polygonal and small phagocytes express Sp185/333 proteins, which are present in the trans cisternae of the Golgi Apparatus, in perinuclear vesicles, and are associated with the membrane of transport vesicles that appear to fuse with the plasma membrane and are present on the extracellular surface of the small phagocytes (Brockton et al., 2008; Dheilly et al., 2011; Figures 3 and 4). It is noteworthy and although the Sp185/333 proteins do not have a predicted transmembrane region, they are also associated with the membranes of internal organelles and vesicles, and with the plasma membrane. This localization may be mediated by interactions between the RGD motif and one or more of the integrin family members in the purple sea urchin (Whittaker et al., 2006) that are expressed in coelomocytes. A second sea urchin species, *Heliocidaris erythrogramma*, expresses He185/333 proteins, which are very similar to those in *S. purpuratus*, and are present in phagocytes and in colorless spherule cells (Dheilly et al., 2011).

**FUNCTION**

The putative immune function of the Sp185/333 proteins has been predicted based on the diversity of the genes, mRNAs, and proteins, and the striking up-regulation of gene expression in response to immunological challenge (Rast et al., 2000; Nair et al., 2005; Brockton et al., 2008; Dheilly et al., 2009). The amount of Sp185/333 protein in coelomic fluid increases in response to injection of LPS, which is reflected by an increase in the number of Sp185/333+ cells (Brockton et al., 2008; Dheilly et al., 2009). These changes appear to be due, in part, to an increase in the number of small phagocytes that includes an increase in cells that are Sp185/333+. On the other hand, while there is no increase in the number of polygonal cells in response to immune challenge, there is an increase in the percentage of Sp185/333+ polygonal cells. Organs also have Sp185/333+ cells including esophagus, gut, testis, ovary, pharynx, and axial organ (Majeske and Smith, unpublished). It is noteworthy that the number of Sp185/333+ cells in the axial organ increases significantly in response to injections of LPS, whereas the other tissues do not show a similar response. The axial organ is located at the axis of the somewhat spherical echinoid (Hyman, 1955) and has no clearly defined function. However, cells in the axial organ of the sea star, *Asterias rubens*, proliferate in response to LPS and concanavalin A (Holm et al., 2008). Some gut associated cells in the sea urchin *H. erythrogramma*, are He185/333+ and may be phagocytes (Dheilly et al.,
To date, it appears Sp185/333+ cells are present throughout echinoid tissues, but it is not known whether they are wandering coelomocytes or stromal cells of the organs. The increase in Sp185/333+ cells and the increase in protein content post-immune challenge is consistent with an immune function of this protein family. Speculation that the Sp185/333 proteins have antibacterial activity has been based on the diversity of the proteins and their expression kinetics in response to challenge from bacterial molecular patterns or from heat-killed marine gram negative bacterial species, *Vibrio diazotrophicus*. Native Sp185/333 proteins bind tightly to *V. diazotrophicus* (Figure 13), are not eluted by high or low pH, and in some cases, show an increase in size upon binding the target cells (Schrankel and Smith, unpublished). Perhaps the oligomerization that is commonly observed on Western blots is an aspect of binding to targets. Because predictions of Sp185/333 folding suggests that they may be unordered and unfolded, this has led to questions of whether the proteins conform to the target for binding and whether binding is specific. Testing these functions with a recombinant Sp185/333 protein shows that it binds to *V. diazotrophicus* and Baker’s yeast in the absence of the coelomic fluid, but does not bind to the gram positive species, *Bacillus cereus* or *B. subtilis*. Binding to *V. diazotrophicus* appears to be saturable (Figure 14) and preliminary competition assays suggest that the binding site may be specific (Lun and Smith, unpublished). Future evaluation of different versions of recombinant Sp185/333 proteins with different element patterns, different numbers of repeats and histidine patches (see Figure 9), or that are truncated may have a variety of anti-pathogen functions besides binding such as opsonization and/or agglutination of bacteria to augment phagocytosis, plus bacteriostatic or bactericidal activity.

CONCLUSION

The rapid evolution of pathogens imposes significant burdens on metazoan immune systems to prevent or combat effectively all possible infections. Vertebrate adaptive immune systems employ somatic recombination of the immunoglobulin gene family (Hozumi and Tonegawa, 1976; Brack et al., 1978; Weigert et al., 1978) or assembly of the variable lymphocyte receptor genes (Nagawa et al., 2007; Rogozin et al., 2007; Kishishita et al., 2010) as mechanisms that enable specific recognition and neutralization of pathogens. Until recently, immune diversification mechanisms were thought to be restricted to the vertebrates and that invertebrates relied on innate immunity based entirely on receptors and antimicrobial peptides with broad specificities for pathogen detection and antibacterial capabilities. However, this may not be adequate for host protection, and several highly variable immune response gene families have been identified in a few invertebrate groups that may significantly expand capabilities for non-self detection and protection against pathogen attack (reviewed in Ghosh et al., 2011). The central question that emerges from evidence of invertebrate immune gene diversity is – what are the mechanisms that function to enhance diversification in these groups? DSCAM variability in arthropods is based on extensive alternative splicing (Watson et al., 2005; Brites et al., 2008, 2010) and fibrinogen related proteins

![FIGURE 13](image1.png) Native Sp185/333 proteins bind *Vibrio diazotrophicus* (Vd), a gram negative marine bacterial species. Whole coelomic fluid (wCF) lysate from a sea urchin was incubated with Vd. Bacteria were pelleted, washed, and analyzed by Western blot using equal amounts of all three anti-Sp185/333 sera (see legend for Figure 10). wCF and Vd alone are shown for comparison. Lane 1, wCF; lane 2, wCF proteins bound to Vd; lane 3, Vd. Protein standard is to the left. Unpublished figure provided by Catherine Schrankel.

![FIGURE 14](image2.png) A recombinant Sp185/333 protein (rSp0032) binds *Vibrio diazotrophicus* (Vd). Biotinylated rSp0032 incubated with 10⁶ Vd cells and post-labeled with Neutravidin–FITC shows increased binding with increased protein concentration within the gate area (R2). Binding plateaus at about 400 μM of rSp0032 indicating the saturation point. MFI, mean fluorescence intensity. Results from flow cytometry (inserted images) show fluorescence (X-axis; fluorescent events measuring FITC associated with bacterial cells (Y-axis; side scatter or cell counts) for increasing concentrations of rSp0032. Unpublished figure provided by Catherine Schrankel.
(FRePs) in mollusks (Stout et al., 2009; Hanington and Zhang, 2010; Loker, 2010). Other invertebrate immune genes are single copy, such as *fulc* in compound tunicates, in which the diversity is encoded in the many hundreds of alleles in the population and is enhanced by alternative splicing (De Tomaso et al., 2005; McKitrick and De Tomaso, 2010). Except for alternative splicing of *Dscam* mRNAs, diversification mechanisms beyond gene duplication are not understood for invertebrate immunity, and mechanisms to generate and/or maintain diversity may be different for each group of organisms or even for individual species. The beginnings for understanding invertebrate immune diversification has come from investigations of the FReP3 gene subfamily (Loker et al., 2004; Zhang et al., 2004). The number of FReP3 loci in the genome of an individual snail is insufficient to account for the level of sequence diversity observed in the FReP3 mRNAs. Consequently, FReP3 genes may undergo mutations and/or somatic recombinatorial diversity, which is predicted to result from the concatenation of pairs of gene segments in the hemocytes.

Whether sea urchins have specific mechanisms to manipulate and diversify the members of the Sp185/333 gene family, or whether diversification is the result of genomic instability and DNA repair, or both, is not known. However, multiple levels sequence diversification in the Sp185/333 system appears to act on the genes, the mRNAs, and the proteins (Figure 15). Gene diversification may occur because region(s) of the genome that harbor clusters of Sp185/333 genes (assuming that they have structure similar to that observed for the 7096 BAC sequence; Figure 8) are highly unusual due of the preponderance of various sized repeats. These include (i) large repeats in the form of segmental duplications and the genes themselves based on their sequence similarity and tight linkage, (ii) medium repeats that appear as shared elements and both tandem and interspersed repeats within the genes, and (iii) small repeats or microsatellites that tightly flank the genes in addition to simple repeats that also appear within the genes (Buckley and Smith, 2007; Miller et al., 2010). This range of clustered repeats may act together, perhaps synergistically, to drive genomic instability that appears as gene duplication, deletion, conversion, and recombination. This may change the sequences of the genes including changes in the mosaic patterns of the elements plus SNPs within the elements. In addition, the Sp185/333 family may show size variation among individuals because the genomic structure that harbors the family may promote both meiotic mispairing of genes on sister chromatids and drive unequal crossovers. On the other hand, instability and diversification may be controlled so that gene or regional sequence homogenization from gene conversion is blocked from progressing through the microsatellites. Furthermore, only a single pseudogene and no gene fragments have been detected from the 171 Sp185/333 genes that have been cloned and sequenced, which suggests mechanisms to either block pseudogene formation or to correct or delete them once formed.

The diversity that is encoded within the members of the gene family is expanded by apparent RNA editing that changes the nucleotide sequence at single positions to alter codons and insert stops, in addition to introducing small indels that change the reading frame to encode missense sequence (Figure 15). Although RNA editing is typically observed for a single or a few nucleotide positions in other systems (Chan et al., 1997; Maas and Rich, 2000), editing of the Sp185/333 mRNAs is a combination of random edits throughout the entire length plus site specific edits to certain nucleotides. The overall result broadens the array of expressed full-length proteins to include some that are altered at single amino acids, some with missense sequence, and many that are truncated. This is a noteworthy deviation from the central dogma in biology in which the proteins are only partially encoded by the genes.

The Sp185/333 proteins are larger than predicted from the cDNAs and genes suggesting post-translational diversification (Figure 15). This is evident for both native proteins and for individual recombinant proteins expressed in bacteria or insect cells (Brockton et al., 2008). The proteins may multimerize with themselves and/or with other Sp185/333 variants, both full-length and truncated, and possibly with other proteins in the coelomic fluid. Post-translational modifications may also act to diversify the proteins, perhaps as an underlying mechanism to produce a preponderance of acidic proteins (Dheilly et al., 2009) rather than the predicted full spectrum of isoelectric points. Given this dizzying array of putative diversification mechanisms, one must keep in mind that the proteins functioning in the immune...
response of the sea urchin must withstand selection pressures from pathogens and not be altered to autoreactive forms. The arms race between pathogen virulence and host self protection takes different forms in this coevolutionary process and its direction depends on the capabilities of the pathogens to evade or counteract the immune system and the diversification mechanisms employed by the host to identify, deter, or destroy the invaders.

ACKNOWLEDGMENTS

The author is appreciative of the preliminary and unpublished results provided by Katherine Buckley, Catherine Schrankel, and Cheng Man Lun. Julie Ghosh, Cheng Man Lun, and three reviewers kindly provided improvements to the manuscript. Lauren Sherman, Preethi Golconda, and Hung-Yen Chou helped with the early organization of the information for this review. Work on sea urchin immunology is supported by the National Science Foundation.

REFERENCES


Sea urchin DNA sequence variation and reduced interspecies differences of the less variable DNA sequences. Evolution 36, 655–676.


Smith, L. C., Shih, C. S., and Dachenhaus, S. G. (1998). Coelomocytes express SpBf, a homologue of factor B, the second component in the


Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 December 2011; accepted: 21 March 2012; published online: 12 April 2012.


This article was submitted to Frontiers in Molecular Innate Immunity, a specialty of Frontiers in Immunology.

Copyright © 2012 Smith. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.