The gene structure and super-hypervariability of the complete *Penaeus monodon* Dscam gene

Kantamas Apitanyasai

National Cheng Kung University

Shiao-Wei Huang

National Taiwan University

Tze Hann Ng

National Cheng Kung University

Shu-Ting He

National Cheng Kung University

Yu-Hsun Huang

National Cheng Kung University

Shen-Po Chiu

National Cheng Kung University

Kuan-Chien Tseng

National Cheng Kung University

Shih-Shun Lin

National Taiwan University

Wen-Chi Chang

National Cheng Kung University

James G. Baldwin-Brown

University of California

Anthony D. Long

University of California

Chu-Fang Lo

National Cheng Kung University

Hon-Tsen Yu

National Taiwan University

Han-Ching Wang (✉ wanghc@mail.ncku.edu.tw)

National Cheng Kung University

Research article

Keywords: Dscam, genome, immunoglobulin family, hypervariability, shrimp

Posted Date: May 24th, 2019

DOI: <https://doi.org/10.21203/rs.2.9774/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background In pancrustaceans, the Down syndrome cell adhesion molecule (Dscam) is an extraordinarily complex, single-locus gene, with the potential for generating thousands of isoforms by combining alternative splicing exons. In the present study, we used two advanced sequencing approaches, Illumina and PacBio, with hybrid assembly to analyze the entire Dscam genomic structure in *Penaeus monodon*. **Results** The *P. monodon* Dscam (*PmDscam*) genome was ~250 kbp, with a total of 175 constitutive and alternative splicing exons. Analysis of *PmDscam* cDNA and genomics revealed a conserved architectural structure consisting of an extracellular region with hypervariable Ig domains, a transmembrane domain, and a cytoplasmic tail. While the number of splicing exon variants in N-terminal Ig2, N-terminal Ig3 and the entirety of Ig7 were previously reported to be 28, 43 and 19, we now show that there are in fact 26, 81 and 26 alternative exons in these regions, respectively. We also identified two alternative variants of two exons in the cytoplasmic tail, as well as 7 cytoplasmic tail elements that can either be included or skipped. The presence of three stop codon sites in the cytoplasmic tail region means that alternative splicing is involved in the selection of the stop codon. **Conclusions** In total, alternative splicing provides for 54,756 potential combinations in the extracellular region, plus 512 potential combinations in the cytoplasmic tail, all derived from one *PmDscam* genome locus. We have also established a public-facing *PmDscam* genome database (<http://pmdscam.dpbs.ncku.edu.tw/>) to facilitate future research on characterizing the involvement of Dscam in pancrustacean immunity.

Background

Dscam belongs to the immunoglobulin (Ig) superfamily gene, and it was first identified in the human chromosome in relation to the development of neuronal connectivity (1). This gene also has important roles in nervous system development in insects (2-4). The typical structure of Dscam consists of 10 Ig domains and six fibronectin type III repeats connected to a transmembrane domain and a cytoplasmic tail (5). The Dscam gene structure is hypervariable, with three large tandem arrays located on the N-terminal of Ig2, the N-terminal of Ig3 and the entire Ig7 domain, with each array having many near-duplicate exons (3,5-7). In *Drosophila*, thousands of Dscam isoforms can be generated through mutually exclusive alternative splicing of these duplicated exons (8-9). The resulting isoforms function as an axon guidance receptor in the nervous system and also, at least in some insects, such as mosquito, as immune receptors that are capable of recognizing diverse pathogens (2-3,5). In some arthropods, Dscam has an essential role in immunity through specific recognition, with pathogen-specific isoforms produced in response to immune challenges (3,10-15). Dscam is also potentially able to generate a specific, long-lasting immune response, and with its hypervariability, it has been hypothesized to be an ortholog of antibody genes in vertebrates (16-17). Functionally, Dscam provides an “immunological memory” and supports a novel immune mechanism (“innate immunity with specificity” or “immune priming”) which allows the innate immune system to exhibit characteristics of adaptive immunity (18-20).

Dscam protein forms a horse-shoe shaped structure comprised of the first four extracellular Ig domains, with two surface epitopes, epitope I and epitope II, formed by part of the Ig2 and Ig3 domains. Epitope I is engaged in homophilic binding specificity, whereas epitope II is hypothesized to be involved in pathogen recognition (21-22). Originally, Dscam was thought to occur only as a membrane-bound form with a transmembrane domain (TM) and a cytoplasmic tail, and although Dscam can be secreted from cells in *Drosophila*, this can only be achieved by proteolytic activity. Surprisingly, however, it was subsequently found that shrimp (*Litopenaeus vannamei* and *Penaeus monodon*) and crab (*Eriocheir sinensis*) both expressed a unique tail-less form of Dscam that had neither a transmembrane domain nor cytoplasmic tail (10,16,23,24). Type III polyadenylation was hypothesized to generate membrane-bound Dscam and tail-less Dscam (16).

In the present study, to expand upon and correct our previous understanding of shrimp Dscam, we used two advanced sequencing approaches, Illumina and PacBio, with hybrid assembly, to analyze the Dscam genomic structure in *Penaeus monodon*. We show that *Penaeus monodon* Dscam (*PmDscam*) in fact has at least 26, 81 and 26 alternative spliced forms of the exons encoding Ig2, Ig3 and Ig7, respectively. Our new transcriptomics data also reveals a relatively complex *PmDscam* cytoplasmic tail structure that is distinct from insect Dscam. Several highly conserved functional motifs were discovered in the cytoplasmic tail. In addition to our analysis of the *PmDscam* genomic structure, we also found that most exons in the genome were selected in both nervous and immune related-cells. In different combinations, the various alternatively spliced exons in the extracellular regions as well as the cytoplasmic tail could generate up to 20 million distinct protein isoforms. Taken together, these findings highlight the substantial diversification of Dscam structure. We also provide a draft of the complete genome of tiger shrimp Dscam, which is accessible via our public-facing *PmDscam* database.

Results

Whole genome sequencing and genome assembly

The procedures illustrated in Figure 1 produced a first draft M2 assembly which had the highest contiguity of any assembly that we generated, with an N50 of 5.1 kb in 2.2 million contigs. The final assembly size was 2.6 Gb (Table S2; Figure S1). After gaps in the *PmDscam* sequences were closed by PCR amplification and sequenced using the Sanger sequencing platform, a final corrected M2 assembly was produced (Fig. 1A). The completely constructed draft of the *Penaeus monodon* Dscam genome has a size of approximately 260 kbp (Fig. 2). Figure 2 also shows how the three platforms and the transcriptomics data contributed to this construction.

Penaeus monodon Dscam gene organization

Previously we reported the full-length cDNA of *PmDscam* (16). Here an assembled *P. monodon* *PmDscam* genome reveals the complete *PmDscam* gene structure. Excluding exon 1, the *PmDscam* gene contains 175 exons (Fig. 3). 38 of these exons are constitutive and 137 use alternative splicing. Analysis of the genomic sequences in combination with the cDNAs revealed that the *PmDscam* gene (Fig. 3) has a much more complex cytoplasmic tail than other pancrustacean and arthropod species, *e.g.*, *Daphnia* and *Drosophila*, while the overall gene organization is otherwise similar. Unfortunately, however, we were

unable to identify the 5'-UTR of Dscam located in exon 1 even though this has been identified in other crustacean species (4, 15). The gene organization of *PmDscam* consists of two main parts: the extracellular region (Fig. 3A) and the cytoplasmic tail (Fig. 3B). The extracellular region of *PmDscam* has three alternatively spliced exons, with exons 4, 6 and 15 consisting of 26, 81 and 26 variable exons, respectively (Fig. 3A). Meanwhile, the cytoplasmic tail has two alternatively spliced exons, exon 32 and 44, which contain 2 variable exons each (Fig. 3B). The mature mRNA consists of 44 exons or less and each cDNA sequence contains only one of each of the variable exons (Fig. 3C, upper panel) to encode a protein that shares a conserved structure with other arthropods (25). The *PmDscam* gene thus generates a protein that includes immunoglobulin (Ig) domains, fibronectin type III repeats (FNIII) and a transmembrane domain (TM) in the configuration 9(Ig)-4(FNIII)-Ig-2(FNIII)-TM-cytoplasmic tail (Fig. 3C, lower panel). In addition, as a result of alternative splicing, the hypervariable sites exon 4, exon 6 and exon 15 which code for the N-terminal of Ig2 and Ig3 and the entire Ig7, respectively, produce a large number of different isoform variations. The cytoplasmic tail also contains two alternative exons: exon 32 and exon 44, which code for the transmembrane domain (TM) and stop codon domain, respectively (Fig. 3C, lower panel).

Identification of *PmDscam* hypervariable regions

To identify the hypervariable sequences of Ig2, Ig3 and Ig7 in the *PmDscam* genome, the conserved amino acid sequences of isoform variants from each domain were searched for in the genome. The multiple hypervariable exons were corrected by eye and a total of 26, 81 and 26 spliced forms of the exons encoding Ig2, Ig3 and Ig7 were detected, respectively. These numbers are in contrast to those in Chou *et al.* (2011), where the number of exon variants in Ig2, Ig3 and Ig7 were reported to be 28, 43 and 19, respectively. The isoform sequences from each domain were aligned using Clustal Omega and Genedoc software, and the resulting amino acid sequences are shown in Figure 4. Each of the detected hypervariable regions show several conserved amino acids, some partly conserved amino acids and a number of variable amino acid sequences. Assuming that these alternative variants can be selected independently, then the extracellular region of *PmDscam* can potentially generate at least 54,756 different unique isoforms ($26 \times 81 \times 26 = 54,756$). We note that one of the Ig7 variants has an abnormal length (Fig. 4C), although the significance of this, if any, is unclear.

The first four Ig domains of Dscam were reported to have a horse-shoe conformation, and parts of Ig2 and Ig3 contribute to two composite surface epitopes, epitope I and epitope II (21). Although, these two epitopes are not well conserved in insects (21), they are highly conserved among crustaceans (15). Epitope I is responsible for homophilic binding specificity, while epitope II was hypothesized to bind to non-Dscam ligands (21). Here, we used PSIPRED (<http://bioinf.cs.ucl.ac.uk/psipred>) to identify the two epitopes located in Ig2- (exon 4) and Ig3-spliced (exon 6) variants. The epitope I and epitope II sequence logos in exon 4 and exon 6 were then generated using WebLogo (<http://wrblgo.berkeley.edu/>). In exon 4, the sequence of approximately 12 amino acids before conserved residues 16I, and 13 amino acids after conserved residues 41W were considered to belong to epitope I and II, respectively (Fig. 5A). In exon 6, 8 amino acids after conserved residues 9K(R) were considered to belong to epitope I, and 8 amino acids before the conserved LLC motif were considered to belong to epitope II (Fig. 5B).

Detection of *PmDscam* isoform variants expressed in different tissues

To confirm whether the isoform variants of the three hypervariable exons (exon 4, 6 and 15) obtained from the genome sequence are actually present in shrimp, amplicons spanning the hypervariable exons were amplified from hemocytes and nerve tissue from ten individual shrimp using gene specific primers (Fig. 6A). After cloning and sequencing, the obtained nucleotide sequences were BLASTed against our *PmDscam* genome database. Almost all the isoform variants of exon 4, 6 and 15 found in the *PmDscam* genome were expressed (Fig. 6B-6D). Among the exon 4 variants, isoform 1 and isoform 15 were not found in either hemocytes or nerve tissue, while isoform 19 was absent only from nerve (Fig. 6B). For exon 6, isoforms 10, 38, 51, 52, 70 and 72 were absent from both hemocytes and nerve (Fig. 6C), while isoforms 4, 7, 10, 15 and 16 of the exon 15 domain were also absent from both tissues (Fig. 6D). Although, interestingly, fewer exon 15 isoforms were detected in hemocytes than in nerve tissue (Fig. 6D), these results suggest that at least most of the isoform variants of the hypervariable exons can be found in shrimp.

A complex cytoplasmic tail organization

In our previous study (16), although we successfully identified several cytoplasmic tail isoforms of *PmDscam*, we were only able to identify *PmDscam* element 0 to element 8 (with elements 0-5 corresponding to exons 31-38; the numbering of the elements corresponds to the exons in *Daphnia* Dscam). However, the earlier analysis contained several errors, and some of the downstream functional protein motifs were still missing. Here, using *Drosophila* and *Daphnia* Dscam protein sequences to search for additional putative elements against our transcriptomics sequence, we were able to identify the cytoplasmic tail of *PmDscam* from exon 31 to the stop codon in exon 44 (Fig. 7A). We named these exons according to the order in which they are located in the *PmDscam* genome. The amino acid sequences of each cytoplasmic element are shown in Table 2. Differences between the naming system used in Chou *et al.* (2011) and the exons in Figure 7 include: exons 36, 37 and 38, which were previously thought to be variants C, B and A of element 5, respectively, and the amino acid sequences from exon 39 to exon 44, which were grouped together as element 8. Two alternative kinds of transmembrane domain were found in exon 32; this is like *Drosophila* but unlike *Daphnia* Dscam (6). Interestingly, mutually exclusive alternative splicing was also found in exon 44, with both of the two alternative exons containing the stop codon. In fact, based on the *PmDscam* genomic sequence, exon 44.1 and exon 44.2 are located in the same area but the amino acids of each element are translated with different reading frames, and this results in the expression of two different elements. Further, we found a rare case that if exon 43 is included, it is always followed by exon 44.1, and the resulting nucleotide sequence will produce a stop codon in the very first amino acid of exon 44.1 (Fig. 7A). As noted previously (16), there is a poly(A) site on exon 31. When translation continues to the next exon (i.e. exon 32.1 or 32.2), the normal, membrane-bound form of Dscam is produced, but when this poly(A) tail is added, it results in the production of the tail-less form of *PmDscam*. This tail-less form has been found in several crustaceans, but not in insects (10,16,24). Bioinformatics analysis of exon organization in 20 *PmDscam* contigs

found that exons 31, 33, 35, 38 and 40 are constitutively expressed, while, exons 34, 36, 37, 39, 41, 42 and 43 can be either included or excluded (Fig. 7B). The combination of the seven “optional” exons plus the two mutually exclusive exons (i.e. 32 and 44) can potentially generate at least 512 (i.e. 2⁹) unique isoforms of the *PmDscam* cytoplasmic tail.

The transmembrane domain (TM) is located in either exon 32.1 or exon 32.2 (Fig. 7C; Table 2). The other functional motifs of Dscam, which are highly conserved among crustaceans and insects, were predicted with the simple modular architecture research tool (SMART) version 4.0 and are also shown in Figure 7C and Table 2. Putative Scr homology 2 (SH2) binding motifs were predicted in exons 32.1, 33, 34, 36 and 38, while putative Scr homology 3 (SH3) binding motifs were predicted in exons 33, 34 and 40. An immunoreceptor tyrosine-based activation motif, ITAM (consensus: YXXL) was predicted in exon 34. A polyproline motif was predicted in exon 40, and Zo-1 protein (PDZ) domain motifs were predicted in exons 43, 44.1 and 44.2. However, we were unable to identify an immunoreceptor tyrosine-based inhibitory motif (ITIM) in any of the *PmDscam* exon variants.

Table 2 also shows the result of aligning the amino acid sequence of the *PmDscam* cytoplasmic tail against the cytoplasmic tail domains of both *Drosophila melanogaster* (AF260530) and *Daphnia magna* (ACC65887). *PmDscam* exons 31-44 correspond to exons 16-24 of *D. melanogaster* Dscam and exons 24-31 of *D. magna* Dscam. *PmDscam* exon 39 can be found in *D. magna* but not in *D. melanogaster*, while *PmDscam* exons 32.2 and 37 are absent from *D. magna*. In contrast to other crustacean and insect Dscams, we note that exons 43 and 44.2 have so far been found only in shrimp. We further note that exons 41 and 43 were found in *P. monodon* and not in *L. vannamei* Dscam. The *PmDscam* cytoplasmic tail includes important protein motifs that correspond to those in *Drosophila* and *Daphnia* Dscam, even though many of the amino acid sequences in each exon share a percent identity of less than 50% (Table 2). Taken together, *PmDscam* exhibits a cytoplasmic tail arrangement that is the most complex to have so far been reported in any arthropod. Information on both the nucleotide and amino acid sequences of the extracellular region and cytoplasmic tail of *PmDscam* is now publicly accessible from our shrimp Dscam in-house database (<http://pmdscam.dbbs.ncku.edu.tw/>).

The *PmDscam* ORF

The complete full-length *PmDscam*, including both the extracellular region and the cytoplasmic tail, is shown in Figure 8. The open reading frame (ORF) of *PmDscam* contains 6,135 bp encoding a predicted protein of 2,045 amino acid residues, although the lengths of the nucleotide and amino acid sequences vary as a result of the alternative splicing of hypervariable exons. The putative signal peptide predicted by Signal P3.0 domain analysis is located at the N-terminus. Domain homology analysis using SMART software showed that the deduced amino acid sequence contained ten tandem repeat immunoglobulin domains (Ig), six fibronectin type III domains (FNIII) and thirteen elements in the cytoplasmic tail. The hypervariable sequences in Ig2, Ig3 and Ig7 are indicated. The conserved cell attachment RGD motif (Arg-Gly-Asp) is located between the Ig6 and Ig7 domains at amino acids 595 to 597. The mutually exclusive alternative splicing elements 1 and 13 in the cytoplasmic tail are also indicated.

Discussion

During the past decade, several approaches, including BAC end sequencing, linkage map construction, transcriptome sequencing and whole-genome sequencing, have been used to investigate the genome and genetic properties of crustaceans (26-27). However, the large and highly repetitive sequences of the crustacean genome cause difficulty in genome assembly and other genetic studies (26,28). Furthermore, crustacean genomes show substantial variations in size. For example, the genomes of caridean shrimp (*Exopalaemon carinicauda*) and white shrimp (*Litopenaeus vannamei*) are 5.73 and 2.3 Gb, respectively (28-29), while the *Penaeus monodon* genome size was estimated to be ~2.1 Gb. In the present study, the *P. monodon* whole-genome sequence analysis was conducted using state-of-the-art genomics techniques, including a combination of short read Illumina and long read PacBio sequencing and hybrid assembly. A *Penaeus monodon* Dscam (*PmDscam*) genome, ~250 kb, was assembled, corrected and analyzed (Fig. 2A).

We reported previously (16) that *PmDscam* has a typical Dscam domain architecture similar to arthropod Dscam (9). The extracellular region has 10 immunoglobulin domains and six fibronectin III domains, i.e., [Ig1-Ig9]-[FNIII 1-FNIII 4]-[Ig10]-[FNIII 5-FNIII 6], with half of the second and third Ig domains and the entire Ig7 domain encoded by arrays of near-duplicate exons. The FNIII6 of the extracellular region is followed by a transmembrane domain and a cytoplasmic tail (5-6). Diversity of the hypervariable regions, i.e. the Ig2, Ig3 and Ig7 domains, occurs through mutually exclusive alternative splicing which ensures that in mature mRNA, there is only one exon selected from each array cluster (7). In the present study, we found that the *PmDscam* genome has a total of 175 exons, with five variable regions: the extracellular exon clusters 4, 6, 15 and two cytoplasmic tail exon clusters (32 and 44), which had two alternative splicing exons each (Fig. 3A, 3B). In contrast to our previous study, which reported finding 28, 43 and 19 alternative sequences for N-terminal Ig2, N-terminal Ig3 and the entire Ig7, respectively (16), figure 4 shows that the correct numbers are in fact 26, 81 and 26. There are two reasons for these discrepancies. In the previous study, isoforms with only a single amino acid difference were counted as distinct isoforms even though they were more likely to have resulted from sequencing errors. This would have artificially inflated the earlier figure. Conversely, a number of isoforms were simply not found in the Chou *et al.* (2011) study. The new sequencing methods used here have now corrected both of these errors.

In insects, hypervariability is also produced by mutually exclusive RNA splicing that occurs in clusters of alternative splicing exons (2,5). In a comparison of hypervariable exons among arthropods, *PmDscam* had the most multiple exon variants (3,8,15,22). In *PmDscam*, we found that each alternative splicing exon has a different level of conservation: exon 4 variants have a higher similarity of amino acids in each variant compared to exons 6 and 15 (Fig. 4). Based on alternative splicing in those hypervariable exons, we infer that there are at least 54,756 and 512 possible combinations for the extracellular region and cytoplasmic tail, respectively. It is noteworthy that *P. monodon* can generate the most Dscam isoforms among currently known arthropod species. For example, *P. monodon* can generate more Dscam isoforms than crab (30,600) *Drosophila* (19,008) and *Daphnia* (3,264) (6,8,15).

Since the presence of Dscam in nerve cells and immune-related cells or hemocytes implies it might have a role in both the nervous and immune systems (2,4,30), we investigated the population distribution of the *PmDscam* hypervariable exons which encode for Ig2, Ig3 and Ig7 in both hemocytes and in nerve tissues. The populations of exon 4 (Fig. 6B) and exon 6 (Fig. 6C) variants were similar in hemocytes and nerve tissue, whereas there was a higher diversity of exon 15 variants in nerve tissues compared to hemocytes (Fig. 6D). In this regard, *P. monodon* Dscam populations are unlike those of *Drosophila* and *Daphnia*, both of which show less diversity in their immune Dscam isoform populations than in the Dscam populations of their nervous systems (2,6).

Based on protein structure, the extracellular domain was reported to be involved in binding interactions. Parts of the Ig2 and Ig3 domains form a horseshoe configuration with a dependent interaction on either side of the horseshoe (21). The N-terminal sequences of exon 4 and exon 6 contribute to composite surface epitope I, whereas the C-terminal sequences of exon 4 and exon 6 contribute to epitope II (21). Epitope I is important for homophilic binding specificity, whereas epitope II may be involved in non-Dscam binding (21). The two epitopes located in the exon 4 and exon 6 spliced variants were also detected in *PmDscam* (Fig. 5), suggesting that it may function as proposed by Meijers *et al.* (2007). In addition, sequences of those two epitopes had a high similarity in amino acid sequences when compared to *EsDscam*, suggesting that as in crab (15), *PmDscam* may bind with specific pathogens and regulate phagocytosis.

Transcriptomics were used to determine the unknown exon sequences in the cytoplasmic tail of shrimp Dscam. Here, unlike Dscam from other arthropods, *PmDscam* had not only two alternative exons that encoded for transmembrane domains, but also two alternative exons that encoded for stop codons in the cytoplasmic tail (Fig. 7B). Several functional conserved domains among arthropod Dscam were discovered, including SH2-binding motif, SH3-binding motif, ITAM motif, polyproline motif and PDZ motif (Fig. 7C; Table 2). These small binding motifs are involved in specific protein-protein interactions in cellular signal transduction (31-32). The SH2/SH3-binding motif interacts with Dock then activates axon guidance in *Drosophila* (5). The ITAM motif was reported to be involved in downstream protein tyrosine kinase (PTK)-mediated immunoreceptor signaling after ligand binding and it regulates the expression of surface membrane receptors (6,33). The PDZ motif determines which exons are present on the cytoplasmic tail (34). However, no immune tyrosine-based inhibition motif (ITIM) (I/S/V/LXYXXV/L) was found in *PmDscam*. The ITIM motif is also missing from crab Dscam (23,35), and indicates that these two crustaceans may have only positive transmembrane signaling. In *Daphnia*, the cytoplasmic tail can include or exclude the ITIM or ITAM motif, implying variable signal capacity (6). Like other arthropod Dscams, the RGD (Arg-Gly-Asp) motif that is recognized by integrin family members (36) was also present between Ig6 and Ig7 in the *PmDscam* extracellular region. Additionally, we found that alternative splicing also produced exon variability in the cytoplasmic tail (Fig. 7B) that is similar to that found in other arthropod Dscams (37-39). Depending upon the RNA splicing, exons in the cytoplasmic tail can be either excluded or included, which can affect both the length and the frame shift of *PmDscam*'s reading frame. Similarly, in *Daphnia*, if exon 30 was excluded, the reading frame for exon 31 was shifted, whereas exclusion of exon 27 did not affect the reading frame (6). However, while inclusion or exclusion of exons in *Daphnia* Dscam can result in the absence of an ITIM motif and PDZ domain (6), splicing of *PmDscam* cytoplasmic tail exons results only in the absence of the ITAM motif and not the PDZ domain (Fig. 6C). In *PmDscam*, the PDZ domain is located in the C-terminal region of exon 43, exon 44.1 and exon 44.2 (Fig. 6C, Table 2), suggesting that these alternative PDZ domains may interact with different proteins that are located in various parts of the cellular membrane (39). Isoforms with or without these motifs may have important differences in signaling capacity and regulation of expression of surface membrane receptors (40). Although most of the alternative splicing exons in the *PmDscam* and *Drosophila* Dscam have a relatively low amino-acid identity (Table 2), there is also a high (>50%) level of amino-acid conservation in 3 out of 5 of the constitutive domains, suggesting that shrimp and insect Dscam might share a common ancestor. Finally, we would note that *PmDscam* has a great diversity of isoforms, and its complex cytoplasmic tail structure enables >20 million isoforms via alternative splicing of the extracellular regions and cytoplasmic tail. This is a considerably greater number of isoforms than any other crustacean or arthropod (5-6,10,15,23).

Conclusions

Combining all the data obtained from genomics, transcriptomics and cDNA, we successfully generated an in-house database (<http://pmdscam.dbbs.ncku.edu.tw/>) of *PmDscam* which was sufficient to support BLAST function ability for nucleotide and amino acids sequences of the extracellular regions and cytoplasmic tail. This database should be useful for researchers who need to identify isoforms of each hypervariable exons. We are confident that this *PmDscam* genome as well as our in-house database will be useful resources for research into the involvement of Dscam in pancrustacean immunity.

Methods

Whole-genome sequencing

To construct the complete Dscam genome (*PmDscam*) for the tiger shrimp *Penaeus monodon*, we first used a combination of traditional, next-generation, and new third-generation sequencing strategies to assemble a draft genome (Fig. 1A). For the Illumina whole-genome sequencing, genomic DNAs were extracted from the muscle tissue of an adult female (F09) collected from the coastal waters of Taiwan, following the standard phenol-chloroform procedure. Using the standard operating protocol provided by Illumina (San Diego, CA, USA), two different types of insert library for sequencing were constructed: paired-end libraries for small inserts (180, 350, and 500 bp), and mate-pair libraries for large inserts (2, 5, and 8 kb) (Table S1). Paired-end sequencing was performed using the Illumina HiSeq platform, and a total of 585.60 Gb of raw reads (293.03 Gb from the small insert libraries and 292.57 Gb from the large insert libraries) were generated (Table S1). After quality control removing low-quality reads as well as PCR-replicates and adapter sequences, we obtained 486.22 Gb (224.06X of genome coverage) of clean data for subsequent assembling.

In addition, to improve the assembly quality and increase the scaffold N50, we adopted PacBio (Pacific Biosciences) single-molecule real-time sequencing strategy. Pleopod genomic DNA (F40) was extracted using the Blood and Cell Culture DNA Midi Kit (Qiagen) for construction of a 20-kb insert-size library. A total of 29 SMRTcells were sequenced on the PacBio RS II platform, producing ~17.9 Gb of long reads data with a read length N50 of 11.6 kb (mean 9.14 kb) (Table S1).

De novo genome assembly

As Figure 1A shows, for the preliminary genome assembly, we first assembled the Illumina short reads using two different programs, *Allpaths-LG* (41) and *Velvet* (42), separately. The ALLPATHS assembly had a higher N50 length (6,606 bp vs. 2,458 bp) and a much lower contig number (251,428 vs. 2,003,807) than the VELVET assembly, but the total contig length (1,101,722,092 bp) was only half of the VELVET assembly (2,167,365,623 bp). The VELVET assembly contig length was very close to the full length of the *P. monodon* genome (~2.17 Gb) as estimated by flow cytometry (43).

To improve the scaffold N50, a third assembly was produced. This was a hybrid assembly combining both the Illumina short reads and PacBio long reads data. However, due to computational limitations, not all Illumina data were used for this assembly. Following Chakraborty *et al.* (2016), we first assembled approximately 140 Gb of Illumina data (obtained from the 180 bp insert library) using *Platanus* (44); this assembly was then combined with all the PacBio long reads using *DBG2OLC* (45) to produce the hybrid assembly.

To obtain an optimum assembly that had both contiguity and completeness and could serve as a practical genome database, the three assemblies were sequentially merged using *quickmerge* (46). For this process, the DBG2OLC assembly (most contiguous and least complete) was merged to the ALLPATHS assembly (the next most contiguous but more complete), and the result was then merged to the VELVET assembly to produce the first draft M2 assembly (Fig. 1A; Table S2). Default merging parameters (python `merge_wrapper.py` \${hybridpath} \${selfpath} -hco 5 -c 1.5 -l 10000) were used, with the exception of the -l parameter (minimum size cutoff for seed contigs for merging) due to the low average contig size across the genome, which would have prevented merging had the ordinary cutoff been used. The M2 assembly was polished using one round of *Quiver* (47) error correction and one round of *Pilon* (48) error correction, again as described in Chakraborty *et al.* (2016). All available PacBio data and all available non-matepair Illumina data were used for polishing.

In order to fill the gaps which were found in some parts of the genome and to confirm the sequences, Sanger sequencing was performed using cDNA and genomic DNA samples. Total RNA samples were isolated from hemocytes using RNeasy C&T reagent (Qiagen) according to the manufacturer's protocol. First-strand cDNA synthesis was performed using SuperScript® II Reverse Transcriptase (Invitrogen) according to the manufacturer's instructions. Genomic DNA was extracted from the pleopods of individual shrimp using a DNA extraction kit (GeneReach Biotechnology Corp.). The hemocyte cDNA and pleopod genomic DNA were used as templates for PCR amplification of the exon and intron fragments using gene specific primers (Table 1). The PCR products were separated by agarose gel electrophoresis and purified prior to cloning. The purified DNA fragments were cloned into RBC T&A cloning vector (RBC Bioscience, Taiwan) and sequenced using M13F and M13R universal primers. The resulting Sanger sequences were then merged with the first draft M2 assembly to produce the corrected M2 assembly (Fig. 1A).

Transcriptome sequencing and assembly

For the transcriptome sequencing, *Penaeus monodon* postlarvae were challenged with Nidovirus. Pooled stomach samples were taken from the postlarvae in both the control and Nidovirus-infected group at 48 h post infection. A RNeasy Mini Kit (Qiagen) was used to extract the total RNA following the manufacturer's instructions. Quantification and quality control of the RNA samples were determined by an RNA 6000 Nano kit with an Agilent2100 Bioanalyzer (Agilent Technologies Inc.). Paired-end sequencing was performed on an Illumina NextSeq500 (Genomics BioSci & TechCo.), and the paired-end reads were assembled using Trinity (v.2.1.1; 49) with strand-specific mode (SS_lib_type RF). For functional classification, annotations were determined using BLAST with the Flybase database, and analysis was conducted using PANTHER (50). For the gene-to-gene correlation network, annotations were determined using BLAST with the NCBI-PM and EMBL-CDS databases, and analysis was conducted using the ContigViews (51) web server.

The transcriptomics database was used to search for the remaining exons located in the cytoplasmic tail region. To obtain the sequence of the cytoplasmic tail, several conserved sequences of *PmDscam* (Table S3; 16) were first used to search against the transcriptomics database. Then, all of the nucleotides were translated to amino acid sequences, and BLASTed against the NCBI database. The obtained sequences were analyzed and identified as both nucleotide and amino acid sequences in each exon. Finally, the *PmDscam* genome database was searched for the nucleotide sequences of each exon to find the location of those exons on the *PmDscam* genome (Fig. 1B). The corresponding sequences have been uploaded to NCBI database (under progress), and the total exons sequences for *PmDscam* is already uploaded in our in-house database (please see section 3.5).

Identification of *PmDscam* hypervariable regions and sequence analysis

To obtain the hypervariable sequences of the *PmDscam* exons Ig2, Ig3 and Ig7, we first searched the corrected M2 assembly to find the locations of the conserved amino acid sequences of previous known *PmDscam* isoform variants from each domain (16). To ensure that every potential isoform variant was included, we then aligned all matching variants and used the conserved sequences from each variable region as a guide to search for all the possible exons in the *PmDscam* genome sequences. The exons from each hypervariable region were named according to the order of the location in the *PmDscam* genome sequence.

Diversity of hypervariable regions in immune-related tissues

To investigate the expression of hypervariable exons in shrimp, hemocytes and nerve tissues were collected from ten individual shrimp and used to amplify the hypervariable regions of *PmDscam*. For the hemocyte samples, hemolymph was drawn from the ventral sinus using a sterile 1-ml syringe with anticoagulant solution and centrifuged at 10,000 g for 1 min at 4°C to separate the hemocytes. Then, for both the hemocytes and excised nerve tissue, total RNA was extracted from each sample using RNeasy C&T reagent (Qiagen) following the manufacturer's instructions. The extracted

mRNA was used as a template to synthesize first-strand cDNA with SuperScript® II Reverse Transcriptase (Invitrogen) according to the manufacturer's instructions. To obtain the cDNA sequence of the Ig2, Ig3 and Ig7 variable exons, we performed the polymerase chain reaction (PCR) using 2 nested sets of oligonucleotide primer pairs specific to *PmDscam*. The first amplification used the primers D-F16 and D-R30 (Table 1). The PCR reaction mixture contained 0.2 mM dNTP, 1.5 mM MgCl₂, 0.2 µM of each primer and 2X Taq DNA Polymerase Mastermix-RED (Biomax). The PCR reaction was carried out as follows: 94° C for 5 min, then 35 cycles of 94° C for 30 sec, 55° C for 30 sec, 72° C for 2 min, followed by a final extension at 72° C for 10 min. The PCR product was then diluted and used as the template for the second amplification of the nested PCR with the primers D-F24 and D-R30 (Table 1) in the presence of 1 unit of Takara Ex taq polymerase (Takara). The PCR reaction was carried out as described above. The PCR products were purified and cloned into RBC T&A cloning vector (RBC Bioscience, Taiwan). Individual colonies (n=20) containing insert fragments from each sample were selected randomly and sequenced using M13F and M13R universal primers. BLAST was used to check that the obtained sequences corresponded to our *PmDscam* genome database. Isoform sequences were aligned with Crustal Omega (<http://www.ebi.ac.uk/uniprot/>).

The *PmDscam* Database

The *PmDscam* database was constructed on a LAMP (Linux+Apache+MySQL+PHP) system. The web interface is written in PHP. BLAST algorithms (52), including blastn, blastp and blastx, were used for sequence alignment, with the e-value set to 10e-10 as default. There are 175 *P. monodon* Dscam exons in the *PmDscam* database. Users can input multiple sequences in FASTA format to perform an analysis. All the blast results for each sequence will be shown.

Abbreviations

cDNA: Complementary DNA

DNA: Deoxyribonucleic acid

Dscam: Down syndrome cell adhesion molecule

FNIII: Fibronectin type III repeats

Hcy: Hemocytes

Ig: Immunoglobulin

ITAM: An immunoreceptor tyrosine-based activation motif

ITIM: An immunoreceptor tyrosine-based inhibitory motif

kbp: Kilobase pairs

Nev: Nerve

ORF: Open reading frame

PCR: Polymerase chain reaction

PDZ: Zo-1 protein

RNA: Ribonucleic acid

SH2: Putative Scr homology2

SH3: Putative Scr homology3

TM: Transmembrane domain

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported financially by the Ministry of Science and Technology [MOST 106-2633-B-006-004; MOST 106-2313-B-006 -007 -MY3; MOST 107-3017-F-006-001; 107-2313-B-006 -006 -MY3].

Authors' contributions

ADL, CFL, HTY and HCW conceived and designed the experiments; KA, SWH, THN, STH, YHH, SPC and JGB performed the experiments and analyzed the data; SWH, KCT, SSL and WCC performed the bioinformatic analysis; KA, SWH, THN, HTY and HCW wrote the paper. All authors read and approved the final manuscript.

Acknowledgments

We warmly thank Mr. Paul Barlow, National Cheng Kung University for his helpful criticism of the manuscript.

References

1. Yamakawa K, Huot YK, Haendelt MA, Hubert R, Chen XN, Lyons GE, Korenberg JR. DSCAM: a novel member of the immunoglobulin superfamily maps in a Down syndrome region and is involved in the development of the nervous system. *Hum Mol Genet.* 1998;7(2):227-37.
2. Watson FL, Püttmann-Holgado R, Thomas F, Lamar DL, Hughes M, Kondo M, Rebel VI, Schmucker D. Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science.* 2005;309(5742):1874-8.
3. Dong Y, Taylor HE, Dimopoulos G. AgDscam, a hypervariable immunoglobulin domain-containing receptor of the *Anopheles gambiae* innate immune system. *PLoS Biol.* 2006;4(7):e229.
4. Armitage SA, Peuss R, Kurtz J. Dscam and pancrustacean immune memory – a review of the evidence. *Dev Comp Immunol.* 2015;48(2):315-23.
5. Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL. *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell.* 2000;101(6):671-84.
6. Brites D, McTaggart S, Morris K, Anderson J, Thomas K, Colson I, Fabbro T, Little TJ, Ebert D, Du Pasquier L. The Dscam homologue of the crustacean *Daphnia* is diversified by alternative splicing like in insects. *Mol Bio Evol.* 2008;25(7):1429-39.
7. Ng TH, Chiang YA, Yeh YC, Wang HC. Review of Dscam-mediated immunity in shrimp and other arthropods. *Dev Comp Immunol.* 2014;46(2):129-38.
8. Celotto AM, Graveley BR. Alternative splicing of the *Drosophila* Dscam pre-mRNA is both temporally and spatially regulated. *Genetics.* 2001;159(2):599-608.
9. Schmucker D, Chen B. Dscam and DSCAM: complex genes in simple animals, complex animals yet simple genes. *Genes Dev.* 2009;23(2):147-56.
10. Chou PH, Chang HS, Chen IT, Lin HY, Chen YM, Yang HL, Wang HC. The putative invertebrate adaptive immune protein *Litopenaeus vannamei* Dscam (LvDscam) is the first reported Dscam to lack a transmembrane domain and cytoplasmic tail. *Dev Comp Immunol.* 2009;33(12):1258-67.
11. Hung HY, Ng TH, Lin JH, Chiang YA, Chuang YC, Wang HC. Properties of *Litopenaeus vannamei* Dscam (LvDscam) isoforms related to specific pathogen recognition. *Fish Shellfish Immunol.* 2013;35(4):1272-81.
12. Smith PH, Mwangi JM, Afrane YA, Yan G, Obbard DJ, Ranford-Cartwright LC, Little TJ. Alternative splicing of the *Anopheles gambiae* Dscam gene in diverse *Plasmodium falciparum* infections. *Malar J.* 2011;10:156.
13. Dong Y, Cirimotich CM, Pike A, Chandra R, Dimopoulos G. Anopheles NF- κ B-regulated splicing factors direct pathogen-specific repertoires of the hypervariable pattern recognition receptor AgDscam. *Cell Host Microbe.* 2012;12(4):521-30.
14. Ng TH, Hung HY, Chiang YA, Lin JH, Chen YN, Chuang YU, Wang HC. WSSV-induced crayfish Dscam shows durable immune behavior. *Fish Shellfish Immunol.* 2014;40(1):78-90.
15. Li XJ, Yang L, Li D, Zhu YT, Wang Q, Li WW. Pathogen-Specific Binding Soluble Down Syndrome Cell Adhesion Molecule (Dscam) Regulates Phagocytosis via Membrane-Bound Dscam in Crab. *Front Immunol.* 2018;9:801.

16. Chou PH, Chang HS, Chen IT, Lee CW, Hung HY, Wang HC. *Penaeus monodon* Dscam (PmDscam) has a highly diverse cytoplasmic tail and is the first membrane-bound shrimp Dscam to be reported. *Fish Shellfish Immunol.* 2011;30(4-5):1109-23.
17. Brites D, Du Pasquier L. Somatic and Germline Diversification of a Putative Immunoreceptor within One Phylum: Dscam in Arthropods. *Results Probl Cell Differ.* 2015;57:131-58.
18. Kurtz J. Specific memory within innate immune system. *Trends Immunol.* 2005;26(4):186-92.
19. Kurtz J, Armitage SA. Alternative adaptive immunity in invertebrates. *Trends Immunol.* 2006;27(11):493-6.
20. Milutinović B, Kurtz J. Immune memory in invertebrates. *Semin. Immunol.* 2016;28(4):328-42.
21. Meijers R, Puettmann-Holgado R, Skinotis G, Liu JH, Walz T, Wang JH, Schmucker D. Structural basis of Dscam isoform specificity. *Nature.* 2007;449(7161):487-91.
22. Brites D, Encinas-Viso F, Ebert D, Pasquier LD, Haag CR. Population genetics of duplicated alternatively spliced exons of the Dscam gene in *Daphnia* and *Drosophila*. *PLoS One.* 2011;6(12): e27947.
23. Jin XK, Li WW, Wu MH, Guo XN, Li S, Yu AQ, Zhu YT, He L, Wang Q. Immunoglobulin superfamily protein Dscam exhibited molecular diversity by alternative splicing in hemocytes of crustacean, *Eriocheir sinensis*. *Fish Shellfish Immunol.* 2013;35(3):900-9.
24. Wang J, Wang L, Gao Y, Jiang Q, Yi Q, Zhang H, Zhou Z, Qiu L, Song L. A tailless Dscam from *Eriocheir sinensis* diversified by alternative splicing. *Fish Shellfish Immunol.* 2013;35(2):249-61.
25. Armitage SAO, Kurtz J, Brites D, Dong Y, Pasquier LD, Wang HC. Dscam1 in pancrustacean immunity: current status and a look to the future. *Front Immunol.* 2017;8:662.
26. Zhang L, Yang C, Zhang Y, Li L, Zhang X, Zhang Q, Xiang J. A genetic linkage map of Pacific white shrimp (*Litopenaeus vannamei*): sex-linked microsatellite markers and high recombination rates. *Genetica.* 2007;131(1):37-49.
27. Zhao C, Zhang X, Liu C, Huan P, Li F, Xiang J, Huang C. BAC end sequencing of Pacific white shrimp *Litopenaeus vannamei*: a glimpse into the genome of Penaeid shrimp. *Chin J Oceanol Limnol.* 2012;30(3):456–70.
28. Yu Y, Zhang X, Yuan J, Li F, Chen X, Zhao Y, Huang L, Zheng H, Xiang J. Genome survey and high-density genetic map construction provide genomic and genetic resources for the Pacific White Shrimp *Litopenaeus vannamei*. *Sci Rep.* 2015;5:15612.
29. Yuan J, Gao Y, Zhang X, Wei J, Liu C, Li F, Xiang J. Genome Sequences of Marine Shrimp *Exopalaemon carinicauda* Holthuis Provide Insights into Genome Size Evolution of Caridea. *Mar Drugs.* 2017;15(7):213.
30. Brites D, Brena C, Ebert D, Du Pasquier L. More than one way to produce protein diversity: duplication and limited alternative splicing of an adhesion molecule gene in basal arthropods. *Evolution.* 2013;67(10):2999–3011.
31. Cohen GB, Ren R, Baltimore D. Modular binding domains in signal transduction proteins. *Cell.* 1995;80(2):237-48.
32. Pawson T. Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems. *Cell.* 2004;116(2):191-203.
33. Barrow AD, Trowsdale J. You say ITAM and I say ITIM, let's call the whole thing off: the ambiguity of immunoreceptor signaling. *Eur J Immunol.* 2006;36(7):1646-53.
34. Sheng M, Sala C. PDZ domains and the organization of supramolecular complexes. *Annu Rev Neurosci.* 2001;24:1-29.
35. Li W, Tang X, Chen Y, Sun W, Liu Y, Gong Y, Wen X, Li S. Characterize a typically Dscam with alternative splicing in mud crab *Scylla paramamosain*. *Fish Shellfish Immunol.* 2017;71:305-318.
36. Ruoslahti E, Pierschbacher MD. New perspectives in cell adhesion: RGD and integrins. *Science.* 1987;238(4826):491-7.
37. Wang J, Ma X, Yang JS, Zheng X, Zugates CT, Lee CHJ, Lee T. Transmembrane/juxtamembrane domain-dependent Dscam distribution and function during mushroom body neuronal morphogenesis. *Neuron.* 2004;43(5):663–72.
38. Yu HH, Yang JS, Wang J, Huang Y, Lee T. Endodomain diversity in the *Drosophila* Dscam and its roles in neuronal morphogenesis. *J Neurosci.* 2009;29(6):1904-14.
39. Fanning AS, Anderson JM. PDZ domains: fundamental building blocks in the organization of protein complexes at the plasma membrane. *J Clin Invest.* 1999;103(6):767-72.
40. Indik ZK, Park JG, Hunter S, Schreiber AD. Structure/function relationships of Fc gamma receptors in phagocytosis. *Semin Immunol.* 1995;7(1):45-54.

41. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB. ALLPATHS: *De novo* assembly of whole-genome shotgun microreads. *Genome Res.* 2008;18(5):810-20.
42. Zerbino DR, Birney E. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* 2008;18(5):821–9.
43. Huang SW, Lin YY, You EM, Liu TT, Shu HY, Wu KM, Tsai SF, Lo CF, Kou GH, Ma GC, Chen M, Wu D, Aoki T, Hirono I, Yu HT. Fosmid library end sequencing reveals a rarely known genome structure of marine shrimp *Penaeus monodon*. *BMC Genomics.* 2011;12:242.
44. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, Kohara Y, Fujiyama A, Hayashi T, Itoh T. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* 2014;24(8):1384-95.
45. Ye C, Hill CM, Wu S, Ruan J, Ma ZS. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third-generation sequencing technologies. *Sci. Rep.* 2016;6:31900.
46. Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. Contiguous and accurate *de novo* assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res.* 2016;44(19):e147.
47. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods.* 2013;10:563–569.
48. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One.* 2014;9(11):e112963.
49. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thomson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, *et al.* Trinity: reconstructing a full-length transcriptome without a genome from RNA–Seq data. *Nat Biotechnol.* 2011;29(7):644–652.
50. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.* 2017;45(1):183-189.
51. Liu LY, Tseng HJ, Lin CP, Lin YY, Huang YH, Huang CK, Chang TH, Lin SS. High-throughput transcriptome analysis of the leafy flower transition of *Catharanthus roseus* induced by peanut witches'-broom phytoplasma infection. *Plant Cell Physiol.* 2014;55(5):942–57.
52. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinform.* 2009;10:421.

Tables

Table 1

Nucleotide sequence of the primers used.

Primer	Sequence (5'-3')	GenBank accession number
Ig domain variant		
D-F16	5' ATGGGCACTACCTATATG 3'	
D-F24	5' CTGATCTTCCCTCCCTTC 3'	
D-R30	5' CAAGATCGCGATAGTCAC 3'	
Introns/Exons confirmation	5' TCGAGACTGTGCTCACTGT 3'	
D-Ig4-F	5' GTGTCAATGCGAAGAACAGC 3'	
D-Ig4-R	5' TGCAGTTCACACAAATGATGGA 3'	
D-Ig6-F	5' AACAAATGAGGGTGCCATTG 3'	
D-Ig6-R	5' CACGCTGGATTGTGGAAC 3'	
D-Ig8-F	5' TTCCAAAAGCATTGGTAGCC 3'	
D-Ig9-R	5' CAGGCAACCGCGTAGACTTC 3'	
D-gFNIII2-F	5' CTACCGTCAGGAAGTGCACG 3'	
D-gFNIII3-R	5' GAAGCCTGAGATTCCTGAAG 3'	
D-gFNIII6-F	5' TCCAGTGAGAGTCAATGTAG 3'	
D-gE0-R	5' TCTCTGCTGTAGTCGTCATC 3'	
D-EL1(B)-F	5' CTCAACATCGCAGTCTAAAG 3'	
D-EL1(B)-R	5' CCCAGTTATTGCCACTATCG 3'	
D-I-E1B-F	5' CCCAGTTATTGCCACTATCG 3'	
D-I-E2-R	5' TCTGCCCTTATGCTACCTTC 3'	
D-I-E4-F	5' GGGTATGGCCACAAGTATAG 3'	
D-I-E4-R	5' AACCGTGGAAGCACCTCT 3'	
D-gE9-F	5' TTGGTTTGTGCACAATGTCT 3'	
D-gE10-R	5' CCAAGAGCACAGAGGAGATG 3'	
D-gE11-F	5' GGTCTTATTCAGTTTCCTCG 3'	
D-gE12A-R		

Table 2

A comparison of the amino acid sequences of *PmDscam* cytoplasmic tail exons 31-44 with sequences of *D.Melanogaster* exons 16-24 and *D. Magna* exons 24-31.

Shrimp exon #	Species	Variant	Amino acid sequence	<i>D.Melanogaster</i> exon # (Identity(%))	<i>D. ex</i> (lc
31	Pm, Lv		VAEYEVATLTLTG	16 (76.9)	24
32.1	Pm, Lv	A	GTIAPAREVPAFGAGDLPIY <u>LNLNLIVPVSAVVIVLAIV</u> IIC YLRGR NTPIK	17.1 (49.1)	25
32.2		B	ATLPPTVSDSRVTWLPDWWPKWLDLNLVLPVIATIVVIIVIGIVVICVAVTRRKNGIENLR	17.2 (38)	-
33	Pm, Lv		EEVYQQYQYNASMPPPSTMDKRHPGFREELG YIPPP NRKL PPVPG SQYNTCDRIKR	18 (64.3)	26
34	Pm, Lv		GGGSGRGTHATWD PRRPM YEELSLHPPPGRRIP LG PPQPL GSQDTLRS	19 (37.3)	27
35	Pm, Lv		GGDDEICPYATFHLLGFREEMDPQQAGNNFQTFPHQNGHGSQQHFVN SPAS SRSM	20 (53.7)	28
36	Pm, Lv		PRHGSGNYSCVSGDY TCG HTPN EG HQ	20 (14.8)	29
37	Pm, Lv		PRHGSGNYSCVAGEYGP GG	20 (25)	-
38	Pm, Lv		PPSSTYYSTVPGDMTASRMSNSTFSPT YDDP ARSDEESDQYGGSTYSGGGPYARAIDSVSQSGTAKRLS	21 (29)	29
39	Pm, Lv		NGGHPPGAPVSGPQPSNHRFICK	-	29
40	Pm, Lv		RGSTSGSAGQGSPEPPPPP PPRNG DLPLDSSGLGSSLNDSNNSTASNQFSEACDHDLVQRNYG	22 (45.3)	29
41	Pm		RHCAQTKP	23 (11.1)	30
42	Pm, Lv		VKATKSTEEMRKLLDK	23 (53.3)	30
43	Pm		<i>KLNKT*</i>	-	-
44.1	Pm, Lv	A	NEAAAHIQNGGLRMVSD EMNNV*	24 (26.9)	31
44.2	Pm, Lv	B	EMKQLPT FKMEA*	-	-

Underlining: transmembrane domain (TM); bold: SH2 binding site; bolded italics: SH3 binding site; box: polyproline motif; italics: PDZ motif; and asterisk: stop codon.

Figures

Figure 1

Strategies and genomic sequencing methods. (A) Construction and characterization of the complete PmDscam genome. (B) Analysis of the cytoplasmic tail and location of the cytoplasmic tail exons.

Figure 2

Schematic diagram of the *Penaeus monodon* Dscam genome structure. (A) PmDscam genomic DNA spans 266 kbp. PacBio and Illumina sequencing were used to characterize the entire genome, with Sanger sequencing used to fill some gaps and confirm sequences. Transcriptomics data were used to identify the cytoplasmic tail and some parts of the extracellular region. The gap near the N-terminal corresponds to a part of the 5'-UTR that we were unable to find in the genome. Samples used for sequencing were derived from both genomic DNA (black) and cDNA (dark grey) samples. Parts of the genome containing repeat unknown sequences (N) are light grey. (B) Location of primers for PCR amplification and Sanger sequencing of the PmDscam genome. Samples were extracted from both genomic DNA and cDNA.

Figure 3

Organization of the PmDscam gene. The PmDscam genome structure consists of 175 exons: 38 exons constitutively spliced (black lines) and 137 exons alternatively spliced (color lines). (A) The PmDscam extracellular region is encoded by exon 1 to exon 31. The variable regions are exon 4 (red), exon 6 (blue) and exon 15 (green), which contain 26 variants, 81 variants 26 variants, respectively. The dashed line represents exon 1, which was not identified in the PmDscam genome. (B) The cytoplasmic tail is encoded by exon 31 to exon 44. The variable regions are exon 32 and exon 44, with each of these two exons containing two variants. After RNA splicing, each transcript contains one alternative variant from each of these exons. (C) PmDscam mRNA (upper panel) contains both constitutive exons (white) and mutually exclusive alternative splicing exons. Alternatively spliced exons encode the N-terminal half of Ig2 (red), the N-terminal half of Ig3 (blue), the entire Ig7 (green), the transmembrane domain (purple) and exon 44 in the cytoplasmic tail, respectively. The PmDscam protein structure (lower panel) is comprised of the extracellular region, which contains 10 immunoglobulin (Ig) domains and six fibronectin type 3 (FNIII) domains, and the cytoplasmic tail.

(A)

(B)

(C)

Figure 4

Multiple amino acid sequence alignments of each of the PmDscam extracellular variable regions. (A) 26 variants encode the N-terminal Ig2 domain in the Ig2 exon 4 cluster. (B) 81 variants encode the N-terminal Ig3 domain in the Ig3 exon 6 cluster. (C) 26 variants encode the entire Ig7 domain in the Ig7 exon 15 cluster. The total number of amino acids for each isoform is indicated on the right. Identical (black) and similar (grey and light grey) amino acids are indicated. Exon#: the exon numbers correspond to the exon's location in the PmDscam genome.

(A) The conservation of exon 4 in *P. monodon* (B) The conservation of exon 6 in *P. monodon*

Figure 5

Differential sequence conservation of epitopes I and II of PmDscam. (A) Sequence logo representation of the conservation of exon 4 variants in *P. monodon*. (B) Sequence logo representation of the conservation of exon 6 variants in *P. monodon*. Bits on the y-axis indicate units of evolutionary conservation.

Figure 6
PmDscam isoform expression in hemocytes and nerve tissue. (A) Location of primers for PCR amplification and sequencing of PmDscam cDNA domain structure. (B) Exon 4 variants (C) exon 6 variants and (D) exon 15 variants detected in hemocyte (hcy) and nerve (nev) cDNA. Hemocyte and nerve samples were collected from 10 individual shrimp for total RNA extraction and cDNA synthesis. Twenty individual cDNA clones were obtained from each sample and their exon 4-exon 6 (Ig2-Ig7 domain) was sequenced. Red boxes represent detection of the isoform while, green boxes represent non-detection of the isoform.

Figure 7
Organization of the element of the PmDscam cytoplasmic tail. E indicates the element's exon number according to its location in the PmDscam genome. (A) Schematic diagram showing the cytoplasmic tail exon combinations of 20 PmDscam contigs obtained from transcriptomics data. (B) Summary of exons

types. Constitutive exons, insertion or skip exons and alternative exons are shown as light grey boxes, dark grey boxes and black boxes, respectively. (C) Cytoplasmic tail exons showing locations of common functional domains and motifs. Asterisks indicate a stop codon.

Figure 8

The amino acid sequence of a PmDscam isoform. Putative signal peptides are in italics. Ig domains are shaded in light grey. The variable sequences in Ig2, Ig3 and Ig7 are in bold. FNIII domains are boxed. A conserved RGD motif is indicated by underlining. In the cytoplasmic tail the constitutive and optional domains are shaded black, while the exclusive alternatively spliced domains are shaded dark gray and bold. Element 1(A) [E1(A)] is the transmembrane domain (i.e. Exon 32.1). The asterisk indicates the stop codon.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplement1.doc](#)