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

Kantamas Apitanyasai\(^1,2\), Shiao-Wei Huang\(^3\), Tze Hann Ng\(^1,2\), Shu-Ting He\(^1,2\), Yu-Hsun Huang\(^1,2\), Shen-Po Chiu\(^1\), Kuan-Chien Tseng\(^6\), Shih-Shun Lin\(^5\), Wen-Chi Chang\(^5\), James G. Baldwin-Brown\(^7\), Anthony D. Long\(^7\), Chu-Fang Lo\(^1,2\), Hon-Tsen Yu\(^2\) & Han-Ching Wang\(^1,2\)*

Using two advanced sequencing approaches, Illumina and PacBio, we derive the entire Dscam gene from an M2 assembly of the complete *Penaeus monodon* genome. The *P. monodon* Dscam (*PmDscam*) gene is \(\sim 266\) kbp, with a total of 44 exons, 5 of which are subject to alternative splicing. *PmDscam* has a conserved architectural structure consisting of an extracellular region with hypervariable Ig domains, a transmembrane domain, and a cytoplasmic tail. We show that, contrary to a previous report, there are in fact 26, 81 and 26 alternative exons in N-terminal Ig2, N-terminal Ig3 and the entirety of Ig7, respectively. We also identified two alternatively spliced exons in the cytoplasmic tail, with transmembrane domains in exon variants 32.1 and 32.2, and stop codons in exon variants 44.1 and 44.2. This means that alternative splicing is involved in the selection of the stop codon. There are also 7 non-constitutive cytoplasmic tail exons that can either be included or skipped. Alternative splicing and the non-constitutive exons together produce more than 21 million isoform combinations from one *PmDscam* locus in the *P. monodon* gene. A public-facing database that allows BLAST searches of all 175 exons in the *PmDscam* gene has been established at http://pmdscam.dbbs.ncku.edu.tw/.

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 plays several important roles in the development of the nervous system in insects\(^2–4\). Structurally, Dscam consists of 10 Ig domains and six fibronectin type III repeats connected to a transmembrane domain and a cytoplasmic tail. The Dscam gene is hypervariable, with three large tandem arrays located on the N-terminal Ig2, the N-terminal Ig3 and the entire Ig7 domain, with each array having many near-duplicate exons\(^3,5–7\). In *Drosophila*, this allows thousands of Dscam isoforms to be generated through mutually exclusive alternative splicing of the near-duplicate exons\(^8,9\). The resulting isoforms act as axon guidance receptors in the nervous system and also, at least in insects such as the mosquito, as immune receptors that are capable of recognizing diverse pathogens\(^2,3\). In some arthropods, Dscam plays an essential role in immunity by recognizing specific pathogens, and producing pathogen-specific isoforms in response to immune challenge\(^10–15\). 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 arthropods with 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 involved 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)

\(^1\)Department of Biotechnology and Bioindustry Sciences, National Cheng Kung University, Tainan, Taiwan. 
\(^2\)International Center for the Scientific Development of Shrimp Aquaculture, National Cheng Kung University, Tainan, Taiwan. 
\(^3\)Department of Life Sciences, National Taiwan University, Taipei, Taiwan. 
\(^4\)Department of Life Sciences, National Cheng Kung University, Tainan, Taiwan. 
\(^5\)Institute of Biotechnology, National Taiwan University, Taipei, Taiwan. 
\(^6\)Institute of Tropical Plant Sciences, National Cheng Kung University, Tainan, Taiwan. 
\(^7\)Department of Ecology and Evolutionary Biology, University of California, Irvine, Irvine, California, USA. 
*email: wanghc@mail.ncku.edu.tw*
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 both shrimp (*Litopenaeus vannamei* and *Penaeus monodon*) and crab (*Eriocheir sinensis*) express a unique tail-less form of Dscam that had neither a transmembrane domain nor cytoplasmic tail. Type III polyadenylation was thought to provide a mechanism that would generate both membrane-bound Dscam and tail-less Dscam.

In the present study, to expand upon and correct our previous understanding of shrimp Dscam, we used hybrid assembly and two advanced sequencing approaches, Illumina and PacBio, to construct an M2 assembly of the entire *Penaeus monodon* genome, from which we derive a draft of the Dscam gene. We show that in *Penaeus monodon* Dscam (*PmDscam*), the exons in Ig2, Ig3 and Ig7 are in fact derived from 26, 81 and 26 mutually exclusive alternative variants, respectively. Based on our new transcriptomics data, we were also able to show that *PmDscam* has a relatively complex cytoplasmic tail structure that is distinct from insect Dscam. Several highly conserved functional motifs were discovered in the cytoplasmic tail. In addition to our structural analysis of the *PmDscam* gene, we also found that most of the alternative exons in the gene were selected in both nervous and immune-related cells. We also show that the various alternatively spliced exons in the extracellular region together with the alternatively spliced and non-constitutive exons in the cytoplasmic tail are capable of generating over 21 million distinct protein isoforms.

**Results**

**Construction of the PmDscam gene from the sequencing and M2 assembly of the whole *P. monodon* genome.** The procedures illustrated in Fig. 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 S1; Fig. S1). After a draft *PmDscam* gene was derived from the polished M2 assembly, most of the remaining gaps in the *PmDscam* sequence were closed by PCR amplification Sanger sequencing (Fig. 1A). The cytoplasmic tail was determined as shown in Fig. 1B, and the complete *Penaeus monodon* Dscam gene was found to have a size of approximately 266 kbp (Fig. 2). Figure 2 also shows how the three platforms and the transcriptomics data were used to build this construction.

**Penaeus monodon Dscam gene organization.** While our previous study of *PmDscam* was based only on cDNA transcripts, here the assembled *P. monodon PmDscam* gene reveals for the first time the entire gene structure. The *PmDscam* gene contains a total of 44 exons (Fig. 3), with 137 exon variants that are subject to
mutually exclusive alternative splicing. Unfortunately, however, even with the PacBio data, Sanger sequencing, and the cDNA transcripts, we were unable to identify the 5′-UTR of Dscam that is presumably located in exon 1. This 5′-UTR has been identified in other crustacean species4,15, and it remains unclear why it could not be found.
in *PmDscam*. The *PmDscam* gene is organized into 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 being derived from the mutually exclusive splicing of 26, 81 and 26 variants, respectively (Fig. 3A). Meanwhile, the cytoplasmic tail has two alternatively spliced exons, exon 32 and exon 44, both of which are derived from two mutually exclusive variants (Fig. 3B). The mature mRNA thus consists of a protein with the same conserved structure that is seen in other arthropod25, i.e. 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).

**Analysis of *PmDscam* hypervariable regions.** First, to identify the hypervariable regions of Ig2, Ig3 and Ig7 in the *PmDscam* gene, we searched for the conserved amino acid sequences of isoform variants from each domain. Once identified, the multiple hypervariable exons variants were checked manually and a total of 26, 81 and 26 spliced forms of the exons variants encoding Ig2, Ig3 and Ig7 were detected, respectively. These numbers are in contrast to those in Chou et al.16, where the number of exon variants in Ig2, Ig3 and Ig7 were reported to be 28, 43 and 19, respectively, from cDNA cloning. The isoform sequences from each domain were aligned using Clustal Omega and Genedoc software, and the resulting amino acid sequences are shown in Fig. 4. 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 × 81 × 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 are folded into a horse-shoe conformation, with parts of Ig2 and Ig3 contributing to two composite surface epitopes, epitope I and epitope II21. Although these two epitopes are not well conserved in insects16, they are highly conserved among crustaceans. Epitope I is responsible for homophilic binding specificity, while it has been hypothesized that epitope II binds to non-Dscam ligands21. Here, we used PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred) to locate the two epitopes in the Ig2 (exon 4) and Ig3 (exon 6) variants. Epitope I and epitope II sequence logos for exon 4 and exon 6 were then generated using WebLogo (http://wrblogo.berkeley.edu/). In exon 4, the sequence of approximately 12 amino acids before the conserved residue 16I, and the 13 amino acids after the conserved residue 41 V were identified as part of epitope I and II, respectively (Fig. 5A). In exon 6, the 8 amino acids after the conserved residue 9 K(R) completed epitope I, and the 8 amino acids before the conserved L C motif completed epitope II (Fig. 5B).

**Expression of *PmDscam* isoform variants in hemocytes and nerve tissues.** To check whether all of the isoform variants derived from the three hypervariable regions (exons 4, 6 and 15) are actually expressed in shrimp, and also to investigate whether there might be any differences in their expression patterns in immune-related cells (hemocytes) versus nerve tissue, 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* gene database. As Fig. 6B–D shows, a small number of exon variants were not detected in either tissue. Among the exon 4 variants, isoform 1 and isoform 15 were not found in either hemocytes or nerve tissue (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). Curiously, we also note that the population distribution of the exon 15 isoforms was much more restricted in hemocytes than in nerve tissue (Fig. 6D). It remains unclear why these missing variants would fail to be expressed in one or both of these tissues.

**A complex cytoplasmic tail organization.** In our previous study16, 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, this earlier analysis contained several errors, and some of the downstream functional protein motifs were still missing. Here, using *P. monodon* Dscam protein sequences to search for additional putative exons against our transcriptomics database and then compared with *Drosophila* and *Daphnia*, 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* gene. The amino acid sequences of each cytoplasmic exon are shown in Table 1. Differences between the naming system used in Chou et al.16 and the exons in Fig. 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* Dscam8. Interestingly, mutually exclusive alternative splicing was also found in exon 44, with both of the two alternative exons containing the stop codon. In fact, the sequence for exon variant 44.2 is entirely contained within that of exon variant 44.1, and it is only because different reading frames are used to translate these two sequences that two distinct exons are expressed. Further, we found a special case that if exon 43 is included, it is always followed by exon variant 44.1, and the resulting nucleotide sequence will produce a stop codon in the very first amino acid of exon variant 44.1 (Fig. 7A). As noted previously16, in addition to the poly(A) tail that is located 364 nucleotides downstream of the 44.2 stop codon, there is also a stop codon and a poly(A) additional signal on the intron after exon 31 and before exon variant 32.1. When this intron is spliced and translation continues to the next exon (i.e. exon variant 32.1 or 32.2), the normal, membrane-bound form of Dscam is produced, but when this intron is included, 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 insects20,16,24. Bioinformatics analysis of exon organization in 20 *PmDscam* contigs containing the cytoplasmic tail 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). With the additional constraint that exon 43 is always followed by exon variant 44.1, this means that there must be at least 384
unique isoforms of the *PmDscam* cytoplasmic tail (i.e. $2^7 \times 3$, where the presence or absence of exons 32, 34, 36, 37, 39, 41 and 42 account for the seven powers of 2, and the three valid combinations of exons 43 and 44 account for the multiplicative factor of 3).

The transmembrane domain (TM) is located in either exon variant 32.1 or exon variant 32.2 (Fig. 7C; Table 1. 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 Fig. 7C and Table 1. Putative Scr homology 2 (SH2) binding motifs were predicted in exon variant 32.1, and exons 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 exon 43 and exon 44.

**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* gene.
difficulty in genome assembly and other genetic studies. 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, while the *Penaeus monodon* genome size was estimated to be ~2.1 Gb. In the present study, the *P. monodon* whole-genome sequence was assembled using state-of-the-art genomics techniques, including a combination of short read Illumina and long read PacBio sequencing and hybrid assembly. From this whole genome sequence, we obtained a *Penaeus monodon Dscam* (PmDscam) gene of ~266 kb that was subjected to correction and analysis (Fig. 2A).

We reported previously that PmDscam has a typical Dscam domain architecture similar to arthropod Dscam. 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. The diversity of the hypervariable regions, i.e. the Ig2, Ig3 and Ig7 domains, arises from mutually exclusive alternative splicing, which ensures that in mature mRNA there is only one exon variant selected from each array cluster. In the present study, we found that the PmDscam gene has a total of 44 exons, including three hypervariable regions in the extracellular region, i.e. the extracellular exon variant clusters 4, 6, 15, and two cytoplasmic tail variant exon clusters (32 and 44), each of which consists of two
mutually exclusive alternatively spliced variants (Fig. 3A,B). In contrast to our previous study, which reported finding 28, 43 and 19 alternative sequences for N-terminal Ig2, N-terminal Ig3 and the entirety of Ig7, respectively16, Fig. 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.16 study. The new sequencing methods used here have now corrected both of these errors.

Our present results also show that, compared to the three hypervariable regions in other arthropods, PmDscam has the highest number of total possible combinations3,8,15,22. That is, as noted above, since there are 54,756 possible combinations that can be generated by the extracellular region, and 384 more that can be produced by the cytoplasmic tail, PmDscam can express 54,756 tailless isoforms plus 21,026,304 isoforms (i.e.

---

**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 15 (Ig2-Ig7 domain) was sequenced. Red boxes represent detection of the isoform while green boxes represent non-detection of the isoform.
54,756 × 384) that are membrane-bound. Against this total of 21,081,060 isoforms, by comparison, there are only 30,600 Dscam isoforms in crab, 19,008 in Drosophila and 3,264 in Daphnia.

The presence of Dscam in both nerve cells and immune-related cells such as hemocytes implies it might have a role in both the nervous and immune systems. Assuming that these two roles are functionally distinct, we might further expect to see different populations of Dscam isoforms in these two tissues. However, when we compared the expression of the PmDscam hypervariable exons in hemocytes and nerves, we found that the expressed variants for exon 4 were very similar (Fig. 6B). Curiously, we also found that there was a higher level of amino acid similarity between the exon 4 variants than between the variants of the other two hypervariable exons (Fig. 4).

For exon 6, different isoforms were expressed even though the overall population diversity was similar (Fig. 6C). Lastly, we observed a high diversity of exon 15 variants in nerve tissues compared to hemocytes (Fig. 6D). Overall, P. monodon Dscam populations are therefore unlike those of Drosophila and Daphnia, both of which show less diversity in all three of the corresponding exons in their immune cells compared to their nervous systems.

The protein structure of Dscam's extracellular domain supports its involvement in binding interactions. Parts of the Ig2 and Ig3 domains form a horseshoe configuration which allows independent interactions on either side of the horseshoe. Surface epitope I is important for homophilic binding specificity and is made up of N-terminal sequences from exon 4 and exon 6, while epitope II, which may be involved in non-Dscam binding, is made up of C-terminal sequences from the same two exons. In PmDscam, the two epitopes (Fig. 5) presumably fulfill the same functions. However, we also note that the amino acid sequences of the PmDscam epitopes have a high similarity to those of EsDscam, suggesting that, as in crab, PmDscam may bind with specific pathogens and regulate phagocytosis.

Sequences derived from our transcriptomics data were used to determine the location of the unknown exons in the cytoplasmic tail of shrimp Dscam. Unlike Dscam from other arthropods, PmDscam not only has two alternative variants that encode for the transmembrane domain, but also two alternative variants that encode for the stop codon in the cytoplasmic tail (Fig. 7B).

Figure 7. Organization of the PmDscam cytoplasmic tail. Exon numbers were determined according to the location of the exon in the PmDscam gene. (A) Schematic diagram showing the cytoplasmic tail exon combinations of 20 PmDscam contigs obtained from transcriptomics data. (B) Summary of exon types. Constitutive exons, inserted or skipped exons and alternative exon variants 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.
Table 1. A comparison of the amino acid sequences of *Pm*Dscam cytoplasmic tail exons 31–44 with sequences of *D. melanogaster* exons 16–24 and *D. magna* exons 24–31. *Underlining: transmembrane domain (TM); bold: SH2 binding site; bolded italics: SH3 binding site; box: proline motif; italics: PDZ motif; and asterisk: stop codon.

| Shrimp exon # | Tail element | Species | Amino acid sequence* | D. melanogaster exon # (Identity[%]) | D. magna exon # (Identity[%]) |
|---------------|--------------|---------|----------------------|--------------------------------------|------------------------------|
| 31            | E0           | Pm, Lv  | VAEEYAVATTLTGT       | 16 (76.9)                            | 24 (84.6)                    |
| 32.1          | E1A          | Pm, Lv  | GTIAAPAREVPAGGDLPIYLNLNLIVPVVSAYVWVIALVJICGTYLGRWNTIK   | 17.1 (49.1)                     | 25 (50)                     |
| 32.2          | E1B          | Pm, Lv  | ATPLPTYSGDRTVLWPDNWPKWLDNLVYPVATIVYLVIGYVTVAYATRRKNGIENLR | 17.2 (38)                       | —                             |
| 33            | E2           | Pm, Lv  | EEEYQQQVYQQNASMPPTSMKRPFREELYGITPPNKLPPPGQSYNTCDRKR     | 18 (64.3)                       | 26 (55.4)                    |
| 34            | E3           | Pm, Lv  | GGGSGRGTIATWDPRPPYEEILSHLPPGRPRLGPPQPGSQRDLRS          | 19 (37.3)                       | 27 (37.4)                    |
| 35            | E4           | Pm, Lv  | GGDGEICPFAYTHHLGFRMEDPQQAGNNQPFTFHGNQGHSQIFHVNSAPRSM   | 20 (53.7)                       | 28 (67.9)                    |
| 36            | E5           | Pm, Lv  | PRHSGNYYSVCSGQTYCGHTPNQH           | 20 (14.8)                       | 29 (14.8)                    |
| 37            | E6           | Pm, Lv  | PRHSGNYYSCVAGYEYPGG           | 20 (25)                           | —                             |
| 38            | E7           | Pm, Lv  | FPSTYYSTVPGMDTASMNSNTSFSTPYDDDPARSDEEDQYGGSTYSGGYPYARIDVSQSGTAKRS  | 21 (29)                       | 29 (20.3)                    |
| 39            | E8           | Pm, Lv  | NGHHPYPAPVGGPQPSNHFRICK      | 22 (45.3)                       | 29 (38)                     |
| 40            | E9           | Pm, Lv  | RGSTSGAGQGSPERMDSCPALESPLDSSGLGSLNDSDNSTASNQFSEAECDHDLVQRNYG  | 23 (11.1)                       | 30 (11.1)                    |
| 41            | E10          | Pm, Lv  | RACACTKTP                                                                 | —                             | 29 (18.2)                    |
| 42            | E11          | Pm, Lv  | VKATSTEEMRKLKD                 | 23 (33.3)                       | 30 (60)                     |
| 43            | E12          | Pm, Lv  | KLNKT*                                                                        | —                             | —                             |
| 44.1          | E13A         | Pm, Lv  | NEAAAHQGNGGLRMSDEMN*          | 24 (26.9)                       | 31 (38.1)                    |
| 44.2          | E13B         | Pm, Lv  | EMRQLPTFRMEA*                                                              | —                             | —                             |

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. However, while inclusion or exclusion of exons in *Daphnia* can result in the absence of an ITIM motif and PDZ domain, splicing of *Pm*Dscam cytoplasmic tail exons results only in the absence of the ITAM motif and not the PDZ domain (Fig. 7C). In *Pm*Dscam, there is a PDZ domain in the C-terminal regions of exon 43, and the exon variants 44.1 and 44.2 (Fig. 6C, Table 1), suggesting that these mutually exclusive alternative PDZ domains might interact with different proteins located in various parts of the cellular membrane. Isomers with or without these motifs may have important differences in signaling capacity and in their ability to regulate the expression of surface membrane receptors.

**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 *Pm*Dscam 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 which of the hypervariable exons were used to produce a particular isoform. The sequence of this *Pm*Dscam gene as well as our in-house database should be useful resources for future research.

**Methods**

**Whole-genome sequencing.** To construct the complete Dscam gene (*Pm*Dscam) for the tiger shrimp *Penaeus monodon*, we first used a combination of traditional, next-generation, and new third-generation sequencing strategies to assemble a polished draft of the entire *P. monodon* genome (Fig. 1A). For the Illumina whole-genome sequencing, the standard phenol–chloroform procedure was used to extract genomic DNA from the muscle tissue of an adult female (F09) collected from the coastal waters of Taiwan. 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 S2). 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 S2). 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 S2).

**De novo genome assembly.** As Fig. 1A shows, for the preliminary genome assembly, we first assembled the Illumina short reads using two different programs, Allpaths-LG and Velvet, 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.
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., we first assembled approximately 140 Gb of Illumina reads.
ing to the manufacturer's protocol. First-strand cDNA synthesis was performed using SuperScript® II Reverse
RNA samples were isolated from hemocytes using REzolTM C&T reagent (Protech Technology, Taiwan) accord-
and to confirm the sequences, Sanger sequencing was performed using cDNA and genomic DNA samples. Total
merging had the ordinary cutoff been used. The M2 assembly was polished using one round of
Quiver
for seed contigs for merging) due to the low average contig size across the genome, which would have prevented
path} \text{hybrid-}
−1 \text{parameter (minimum size cutoff }
first draft M2 assembly (Fig. 1A; Table S1). Default merging parameters (python merge_wrapper.py ${hybrid-
most contiguous but more complete), and the result was then merged to the VELVET assembly to produce the
DBG2OLC assembly (most contiguous and least complete) was merged to the ALLPATHS assembly (the next
transcriptome sequencing and assembly.

| Primer     | Primer sequence (5′-3′) |
|------------|------------------------|
| D-F16      | 5′ ATGGGCACACTCTATATG 3′ |
| D-F24      | 5′ CTGATCTTCGCCCTCCTC 3′ |
| D-R30      | 5′ CAAGATGCGATAGTCAC 3′ |

| Intron/Exon confirmation |
|-------------------------|
| D-Ig4-F                 | 5′ TCGAGACCTGCTGACTGTG 3′ |
| D-Ig4-R                 | 5′ GTGCTAATGGCAAGACACG 3′ |
| D-Ig6-F                 | 5′ TCCAGTTCAACCAAAATGATGGA 3′ |
| D-Ig6-R                 | 5′ AACATGAGGGTGCACTTG 3′ |
| D-Ig8-F                 | 5′ CAGGCTTGAATGGCGA 3′ |
| D-Ig9-R                 | 5′ TTCCAAAGCGATGTAGCC 3′ |
| D-gFN31-F               | 5′ AGAACATGGCAGCTGTGGT 3′ |
| D-gFN31-R               | 5′ TCCACACCTGATGGTAAGGC 3′ |
| D-gFN36-F               | 5′ GACGCTTCAATTCTCTTG 3′ |
| D-gFN36-R               | 5′ TCCAGTGAGATCAGTAGG 3′ |

Table 2. Nucleotide sequences of the primers used.

data (obtained from the 180 bp insert library) using Platanus44; this assembly was then combined with all the
PacBio long reads using DBG2OLC45 to produce the hybrid assembly.

To obtain an optimum assembly that had both contiguity and completeness and could serve as a practi-
cal genome database, the three assemblies were sequentially merged using quickmerge46. 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 S1). Default merging parameters (python merge_wrapper.py $\text{hybrid-}
-path} \text{|selfpath} -hco 5 -c 1.5 -l 10000) were used, with the exception of the $-1$ 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 Quiver47 error cor-
correction and one round of Pilon48 error correction, again as described in Chakraborty et al.49. All available PacBio
data and all available non-matepair Illumina data were used for polishing. The polished M2 assembly of the P.
monodon genome was then used to produce a draft sequence of the P. monodon Dscam gene.

Next, in order to fill the gaps which were still found in some parts of the PmDscam gene (please see Fig. 2A)
and to confirm the sequences, Sanger sequencing was performed using cDNA and genomic DNA samples. Total
RNA samples were isolated from hemocytes using REzolTM C&T reagent (Protech Technology, Taiwan) accord-
ning 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 2). The PCR products were separated by agarose gel electrophoresis and puri-
fied 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.

Transcriptome sequencing and assembly. For the transcriptome sequencing, Penaeus monodon post-
larvae 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.2.1.149) with strand-specific mode (SS_lib_type RF). For functional classifica-
tion, annotations were determined using BLAST with the Flybase database, and analysis was conducted using

| Primer Primer sequence (5′-3′) |
| D-F16 | 5′ ATGGGCACACTCTATATG 3′ |
| D-F24 | 5′ CTGATCTTCGCCCTCCTC 3′ |
| D-R30 | 5′ CAAGATGCGATAGTCAC 3′ |

| Intron/Exon confirmation |
| D-Ig4-F | 5′ TCGAGACCTGCTGACTGTG 3′ |
| D-Ig4-R | 5′ GTGCTAATGGCAAGACACG 3′ |
| D-Ig6-F | 5′ TCCAGTTCAACCAAAATGATGGA 3′ |
| D-Ig6-R | 5′ AACATGAGGGTGCACTTG 3′ |
| D-Ig8-F | 5′ CAGGCTTGAATGGCGA 3′ |
| D-Ig9-R | 5′ TTCCAAAGCGATGTAGCC 3′ |
| D-gFN31-F | 5′ AGAACATGGCAGCTGTGGT 3′ |
| D-gFN31-R | 5′ TCCACACCTGATGGTAAGGC 3′ |
| D-gFN36-F | 5′ GACGCTTCAATTCTCTTG 3′ |
| D-gFN36-R | 5′ TCCAGTGAGATCAGTAGG 3′ |

Table 2. Nucleotide sequences of the primers used.

Scientific Reports | (2019) 9:16595 | https://doi.org/10.1038/s41598-019-52656-x
Identification of \( \text{Pm} \text{Dscam} \) hypervariable regions and sequence analysis. To obtain the hypervariable sequences of the \( \text{PmDscam} \) exons in Ig2, Ig3 and Ig7, we first searched the corrected M2 assembly to find the locations of the conserved amino acid sequences of previous known \( \text{PmDscam} \) isoform variants from each domain. 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 \( \text{PmDscam} \) gene sequences. Like the other \( \text{PmDscam} \) exons, the hypervariable region exons (i.e. exon 4, 6 and 15) were named according to their order of the location in the \( \text{PmDscam} \) gene.

Expression of \( \text{PmDscam} \) isoform variants in hemocytes and nerve tissues. To investigate the expression of the \( \text{PmDscam} \) hypervariable exons, hemocytes and nerve tissues were collected from ten individual shrimp. 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 REzol™ C&T reagent (Protech Technology, Taiwan) following the manufacturer’s instructions. The extracted mRNA was used as a template to synthesize first-strand cDNA with SuperScript® III 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 \( \text{PmDscam} \). The first amplification used the primers D-F16 and D-R30 (Table 2). 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 (Biomann). 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 2) 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 \( \text{PmDscam} \) gene database. Isoform sequences were aligned with Crustal Omega (http://www.ebi.ac.uk/uniprot/).

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

Received: 23 July 2019; Accepted: 17 October 2019;
Published online: 12 November 2019

References
1. Yamakawa, K. et al. 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. 7, 227–37 (1998).
2. Watson, F. L. et al. Extensive diversity of Ig-superfamily proteins in the immune system of insects. Science, 309, 1874–8 (2005).
3. Dong, Y., Taylor, H. E. & Dimopoulos, G. A, Dscam, a hypervariable immunoglobulin domain-containing receptor of the Anopheles gambiae innate immune system. PLoS Biol. 4, e229 (2006).
4. Armitage, S. A., Peuss, R. & Kurtz, J. Dscam and pancrustacean immune memory – a review of the evidence. Dev Comp Immunol. 48, 315–23 (2015).
5. Schmucker, D. et al. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell, 101, 671–84 (2000).
6. Brites, D. et al. The Dscam homologue of the crustacean Daphnia is diversified by alternative splicing like in insects. Mol Biol Evol. 25, 1429–39 (2008).
7. Ng, T. H., Chiang, Y. A., Yeh, Y. C. & Wang, H. C. Review of Dscam-mediated immunity in shrimp and other arthropods. Dev Comp Immunol. 46, 129–38 (2014).
8. Celotto, A. M. & Graveley, B. R. Alternative splicing of the Drosophila Dscam pre-mRNA is both temporally and spatially regulated. Genetics, 159, 599–608 (2001).
9. Schmucker, D. & Chen, B. Dscam and DSCAM: complex genes in simple animals, complex animals yet simple genes. Genes Dev. 23, 147–56 (2009).
10. Chou, P. H. et al. The putative invertebrate adaptive immune protein Lipopenaeus vannamei Dscam (LVdscam) is the first reported Dscam to lack a transmembrane domain and cytoplasmic tail. Dev Comp Immunol. 33, 1258–67 (2009).
11. Hsueh, H. Y. et al. Properties of Lipopenaeus vannamei Dscam (LVdscam) isoforms related to specific pathogen recognition. Fish Shellfish ImmunoL, 35, 1272–81 (2015).
12. Smith, P. H. et al. Alternative splicing of the Anopheles gambiae Dscam gene in diverse Plasmodium falciparum infections. Malar J, 10, 156 (2011).
13. Dong, Y., Cirimotich, C. M., Pike, A., Chandra, R. & Dimopoulos, G. Anopheles NF-κB-regulated splicing factors direct pathogen-specific repertoires of the hypervariable pattern recognition receptor AgDscam. Cell Host Microbe, 12, 521–30 (2012).
Acknowledgements
We warmly thank Mr. Paul Barlow, National Cheng Kung University for his helpful criticism of the manuscript. This study was supported financially by the Ministry of Science and Technology [MOST 106-2633-B-006-004; MOST 107-3017-F-006-001; MOST108-2314-B-006-096-MY3].

Author contributions
A.D.L., C.F.L., H.T.Y. and H.C.W. conceived and designed the experiments; K.A., S.W.H., T.H.N., S.T.H., Y.H.H., S.P.C. and J.G.B. performed the experiments and analyzed the data; S.W.H., K.C.T., S.S.L. and W.C.C. performed the bioinformatic analysis; K.A., S.W.H., T.H.N., H.T.Y. and H.C.W. wrote the paper. All authors read and approved the final manuscript.
Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-52656-x.
Correspondence and requests for materials should be addressed to H.-C.W.
Reprints and permissions information is available at www.nature.com/reprints.
Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019