

# SCIENTIFIC REPORTS



OPEN

## High-throughput transcriptome sequencing of the cold seep mussel *Bathymodiolus platifrons*

Received: 15 October 2014

Accepted: 08 October 2015

Published: 23 November 2015

Yue Him Wong<sup>1,\*</sup>, Jin Sun<sup>2,\*</sup>, Li Sheng He<sup>3</sup>, Lian Guo Chen<sup>1</sup>, Jian-Wen Qiu<sup>2</sup> & Pei-Yuan Qian<sup>1</sup>

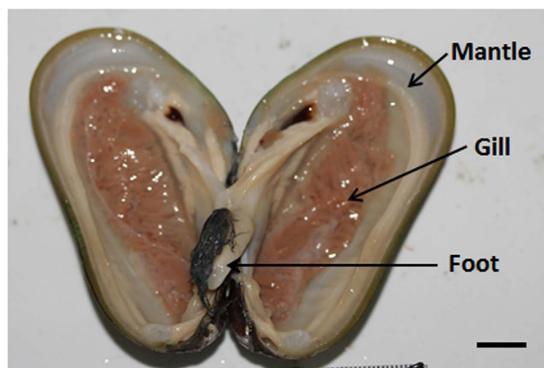
**Bathymodiolid mussels dominate hydrothermal vents, cold methane/sulfide-hydrocarbon seeps, and other sites of organic enrichment.** Here, we aimed to explore the innate immune system and detoxification mechanism of the deep sea mussel *Bathymodiolus platifrons* collected from a methane seep in the South China Sea. We sequenced the transcriptome of the mussels' gill, foot and mantle tissues and generated a transcriptomic database containing 96,683 transcript sequences. Based on GO and KEGG annotations, we reported transcripts that were related to the innate immune system, heavy metal detoxification and sulfide metabolic genes. Our in-depth analysis on the isoforms of peptidoglycan recognition protein (PGRP) that have different cellular location and potentially differential selectivity towards peptidoglycan (PGN) from gram-positive and gram-negative bacteria were differentially expressed in different tissues. We also reported a potentially novel form of metallothionein and the production of phytochelatin in *B. platifrons*, which has not been reported in any of its coastal relative *Mytilus* mussel species. Overall, the present study provided new insights into heavy metal and sulfide metabolism in *B. platifrons* and can be served as the basis for future molecular studies on host-symbiont interactions in cold seep mussels.

In deep sea hydrothermal vent and hydrocarbon seep ecosystems, chemosynthetic microbes are the primary producers<sup>1</sup>. Symbiosis between vent or seep macro-faunal such as mussels, snails, shrimp, tubeworms and crabs and chemosynthetic microbes is a common adaptive mechanism<sup>2–4</sup>. In these symbiotic systems, the microbial biomass serves as either the major or the sole food source of the host<sup>5</sup>. Vent and seep effluents are also known to enrich in heavy metal, sulfide and different hydrocarbon species<sup>6,7</sup>. Hence, in addition to acquisition of chemosynthetic microbes as symbionts, both vent and seep macro-fauna have to adapt to a highly toxic chemical environment. Among different deep sea macro-fauna, Bathymodiolid mussels represent one of the highly specialized animals to vent and seep ecosystems. These deep sea mussels are evolved with the mechanisms to acquire special nutritional advantage from chemosynthetic bacteria and to tolerate a range of highly toxic chemicals. These adaptive features have enabled Bathymodiolid mussels to flourish and dominate hydrothermal vents, cold methane/sulfide-hydrocarbon seeps, and other sites of organic enrichment (e.g., sunken wood and whale bones) in the Atlantic, Pacific, and Indian Oceans<sup>8</sup>.

Bathymodiolid mussels are capable of acquiring chemoautotrophic bacteria as their major nutritional food source<sup>9</sup>. Instead of a vertical transmission of microbial symbionts, Bathymodiolid mussels actively acquire thiotrophic and/or methanotrophic gamma-proteobacteria starting from juvenile stage<sup>10</sup>. These bacterial symbionts are maintained in bacteriocytes, a type of hemocytes with specialized cellular compartment for the storage of symbiotic bacteria<sup>10</sup>. Bacteriocytes may be absorbed via phagocytosis occasionally for nutritional purposes<sup>10</sup>. To date, it remains unclear how Bathymodiolid mussels distinguish

<sup>1</sup>Division of Life Science, School of Science, the Hong Kong University of Science and Technology, Hong Kong S.A.R.

<sup>2</sup>Department of Biology, Hong Kong Baptist University, Hong Kong S.A.R. <sup>3</sup>Sanya Institute of Deep-sea Science and Engineering, Chinese Academy of Sciences, Hainan, People Republic of China. \*These authors contributed equally to this work. Correspondence and requests for materials should be addressed to J.-W.Q. (email: qiuwj@hkbu.edu.hk) or P.-Y.Q. (email: boqianpy@ust.hk)



**Figure 1.** A photograph of *Bathymodiolus platifrons* with arrows showing the gill, foot and mantle. Scale bar = 1 cm.

pathogens from symbionts and how symbionts avoid triggering adverse immune responses from the host.

Bathymodiolid mussels have also been reported to accumulate high concentration of heavy metals, and tolerate high levels of toxic hydrocarbons and sulfide<sup>11,12</sup>. However, majority of the studies concerning heavy metal accumulation and sulfide detoxification focused on vent mussels. For instance, Metallothionein genes from several vent dwelling *Bathymodiolus* species have been reported<sup>13</sup>. A recent high-throughput sequencing analysis of the vent mussel *Bathymodiolus azoricus* has revealed an extensive collection of innate immune transcripts in the gill<sup>14</sup>. However, the metabolism potential of the vent mussel in term of heavy metal and sulfide detoxification were not explored.

The objective of this study was to generate a comprehensive transcriptome database for the methane seep mussel *Bathymodiolus platifrons* (see Fig. 1). To gain an insight into the adaptive features of the seep mussel, we focused on genes related to immune function and detoxification. We perform in-depth analysis on these functional genes and analyzed their gene expression pattern in the gill, mantle and foot. This work extended our understanding on the mechanisms of detoxification and introduced new perspective on the mechanism of symbiont acquisition in Bathymodiolid mussels.

## Results

**De novo assembly and functional annotation of *B. platifrons* transcriptome.** The mitochondrial COI and NADH4 gene sequences of the specimen had highest similarity (99% and 100%, respectively) to the respective *B. platifrons* sequences from Okinawa Trough, Hatoma Knoll<sup>15</sup>. We produced five Gbp clean data (approximately 55 million clean reads) from each of the three tissues (Table S1). Over 98% clean Illumina reads in all three tissues exceeded Q20, indicating high quality of the sequencing data. The raw sequencing data have been submitted to the Short Read Archive (SRA) of NCBI under accession number SRP035485. The final transcriptome generated from Trinity *de novo* assembling contained 96,683 transcript sequences. The length of the transcripts ranged from 200 to 24,272 bp, with an average length of 673 bp and N50 value of 983 bp. The statistics for the data output and *de novo* assemblies are summarized in Table 1. The length distribution of all transcripts is shown in Figure S1. In total, 40,935 (42.3%) transcripts had at least one significant hit in the NCBI non-redundant (nr), Swiss-Prot, COG and KEGG protein database. The functional annotation results are shown in Table 1. The COG and KEGG functional categories for the annotated transcripts are shown in Fig. 2A,B, respectively. There were 14,656 (39.5%) nr-annotated transcripts assigned to major Gene Ontology (GO) categories, i.e. ‘Biological Process’, ‘Cellular Component’, and ‘Molecular Function’ (Fig. 2C). ESTscan and BLAST search of all the aforementioned protein databases resulted in the prediction of 44,633 protein coding transcripts (Table 1).

To evaluate the comprehensiveness of the current *B. platifrons* transcriptome assembly, we downloaded 458 single-copy genes (the core proteins) from the Core Eukaryotic Gene Mapping Approach (CEGMA) dataset<sup>16</sup> and using tBLASTn with a evalue cut-off of  $10^{-3}$  to assay the completeness of our transcriptome assembly. All the 458 core proteins have at least one match in the current *B. platifrons* transcriptome (Supplemental data file 1). The average percentage of identity of alignment was 70% (S.D. 12.7%). This result indicated the high quality and the comprehensiveness of our assembled transcriptome.

The best hit of the majority of the annotated transcripts (15,005 out of 37,039) are *Crassostrea gigas* sequences in the nr database. This is reasonable as *C. gigas* is thus far the only bivalve whose genome has been sequenced (Fig. 2D). There were 1,402 bacterial transcripts (Fig. 2D, indicated by an asterisk). Approximately 0.2%, 0.09% and 0.12% of the clean Illumina reads from the gill, foot and mantle, respectively, were mapped to these bacterial transcripts. The LCA assignment algorithm assigned 225 bacterial transcripts to Methylococcaceae and 367 to Bradyrhizobiaceae (Fig. 2E). In general, transcripts assigned to Methylococcaceae were highly expressed in the gill while those assigned to Bradyrhizobiaceae

<b><i>De novo</i> assembly by Trinity</b>	
Total base (bp)	65,038,187
Total number of transcripts	96,683
Number of singletons	68,714
Number of gene clusters	27,969
Mean length of transcripts (bp)	673
N50 (bp)	982
Transcripts size range (bp)	200–24,272
<b>Functional annotation</b>	
Total number of transcripts annotated by public databases	40,935
NCBI non-redundant (nr) database (e-value < 1 e <sup>-5</sup> )	37,039
SwissProt (e-value < 1 e <sup>-5</sup> )	27,690
KEGG	23,524
COG	12,036
Gene Ontology	14,656
Coding sequence prediction	
CDS predicted from BLAST results	37,376
CDS predicted by ESTScan	7,257
Total CDS predicted	44,633
Mean length of CDS (aa)	213

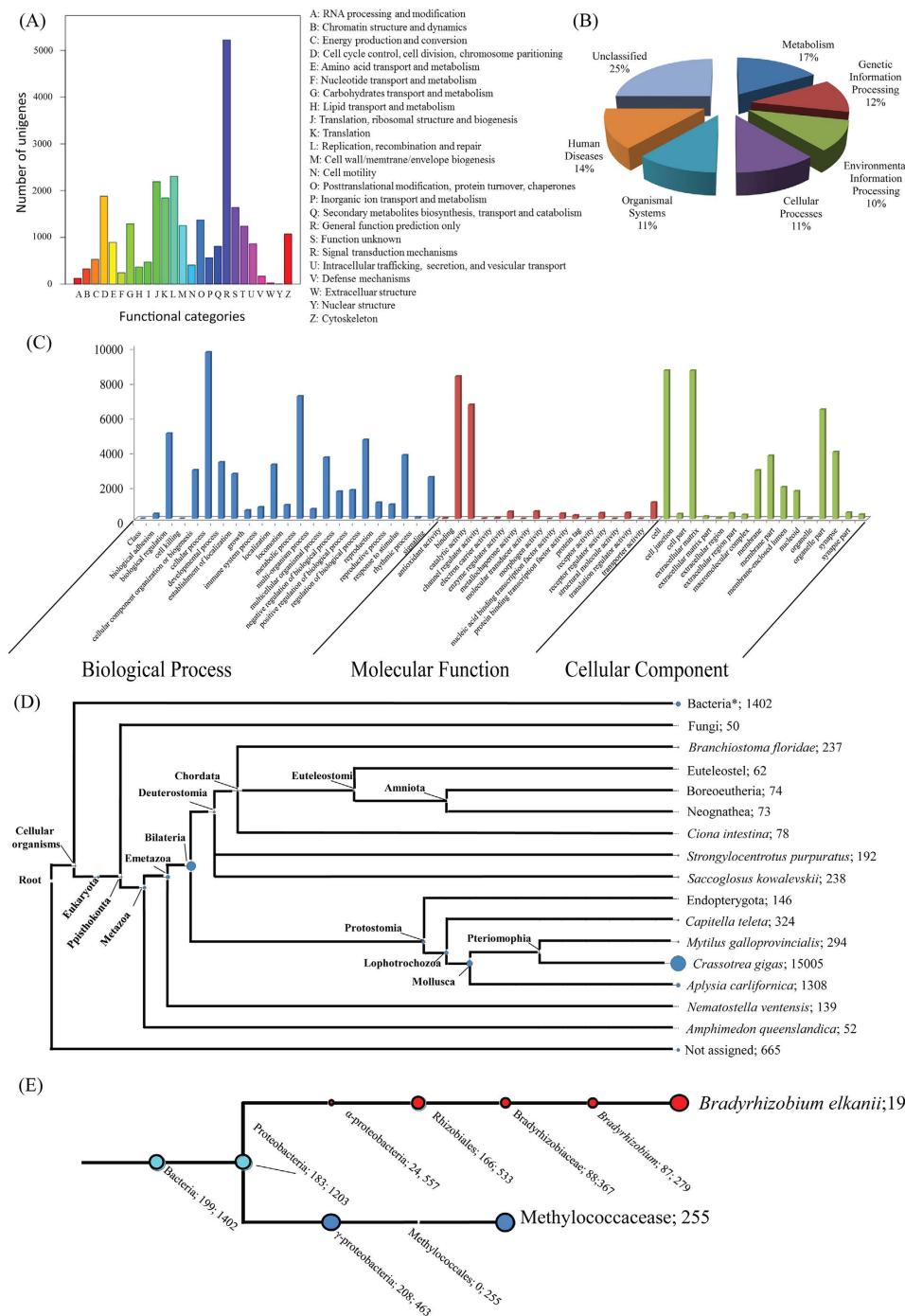
**Table 1.** Summary of assembling and functional annotation of *B. platifrons* transcriptome.

were highly expressed in the foot (Supplemental data file 2). Among those 225 bacterial transcripts to Methylococcaceae, metabolic functional genes that play essential role in methane oxidation, sulfur reduction and nitrite/nitrate reduction were recovered (Fig. S2).

**Function of dominant transcripts.** Close to half of the dominant transcripts (top 5% transcripts in term of total expression level) have no match to proteins in any public database (Fig. S3B). The ten most enriched KEGG pathways and GO Biological Process (GOBP) terms among the dominant transcript list are shown in Fig. 3C,D, respectively. The full list of enriched KEGG pathways and GOBP terms are shown in Table S2 and S3 respectively. In both KO and GOBP enrichment analyses, terms related to *de novo* synthesis (KEGG pathways: “splicesosome”, “RNA transport” and “ribosome”; GOBP term: “translation”) were highly enriched in the dominant transcript list (Fig. S3C & S3D). Oxidative phosphorylation pathway was the most enriched KEGG pathway in the dominant transcript list (Fig. S3C), indicating a high level of energy requirement and active engagement of the mussel in *de novo* synthesis.

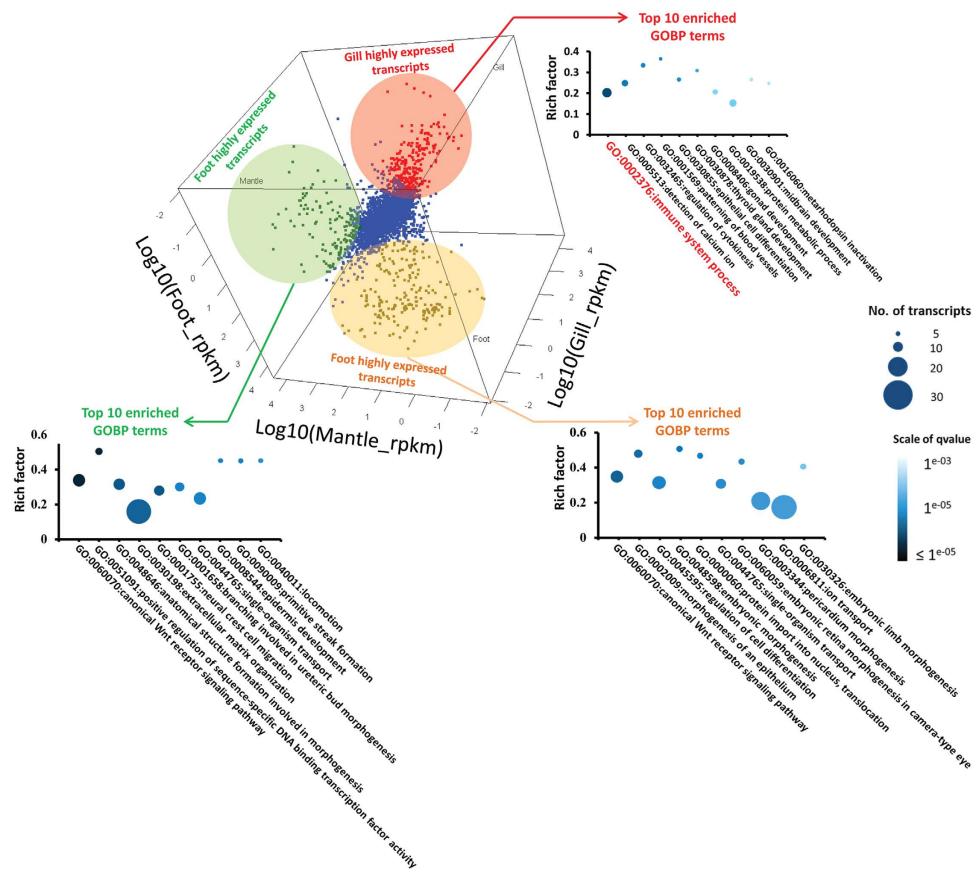
**Differential expression analysis.** We then compared the gene expression profiles of the three examined tissues to identify differentially expressed genes. As quantification accuracy is positively correlated with sequence coverage<sup>17</sup>, only 7,304 transcripts that have at least 750 clean reads (equivalent to 5 ppm total mapped reads) mapped to all three examined tissues in summation were considered in the differential expression analysis. In each pairwise comparison, differentially expressed transcripts was defined as log<sub>2</sub>(ratio) values greater than 2 SD from the mean. The mean, SD and cutoff log<sub>2</sub>(ratio) values for differentially expressed transcripts in each pairwise comparison were listed in Table 2. We focused on transcripts that shown distinctively higher expression level in one of the examined tissues when compared with the other two. We found that transcripts that were dominantly expressed in gill [transcripts highly expressed in the gill in both gill-to-mantle (G-to-M) and gill-to-foot (G-to-F) comparisons] are enriched with genes involved in immune system (*GO:0002376-immune system process*) (see Fig. 4, the GO term is highlighted in red color). Both foot (transcripts highly expressed in the foot in G-to-F and foot-to-mantle (F-to-M) comparisons) and mantle dominant transcripts (transcripts highly expressed in the mantle in both G-to-M and F-to-M comparisons) were enriched with genes with developmental functions/regulations (Fig. 4). In accord to the biomineralization function of mantle in bivalves<sup>18</sup>, the term *ion transport* (*GO:0006811*) was found to be enriched in the mantle dominant transcript list. More detail analyses on mantle dominant transcripts were reported in Sun *et al.* (Submitted).

As we were aiming at generating a database for *B. platifrons*, we did not have replications on our RNAseq data. Hence, we verified the tissue dominant pattern of gene expression detected from RNAseq using qRT-PCR. Different from the RNAseq data, the qRT-PCR experiments were conducted with 3



**Figure 2. A summary of the functional annotation of *B. platifrons* transcriptome (the overall transcriptome). Functional classification of annotated transcripts by (A) COG, (B) KEGG (C) Gene Ontology. (D) LCA assignment of annotated transcripts, (E) LCA assignment of bacterial transcripts. The size of circle is proportional to the number of transcript assigned to the taxon. The number behind each taxon name indicates the number of transcript assigned.**

biological replicates (i.e. 3 individuals) and 2 experimental repeats per biological replicate. We randomly selected 10 transcripts for qRT-PCR quantification from each of the above tissue dominant transcript list. In general, the qRT-PCR results were in good agreement to the result of RNAseq quantification, as indicated by positive linear correlations and rather high Pearson product-moment correlation coefficients ( $R^2$  for Gill-to-Foot: 0.9582;  $R^2$  for Gill-to-Mantle: 0.9697;  $R^2$  for Foot-to-Mantle: 0.8532; see Fig. S3). This result indicated the high reliability of our RNAseq-based comparison of tissue dominant gene expression patterns. Annotation information and the relative expression level as calculated from RNAseq and qRT-PCR results were provided in Table S4.

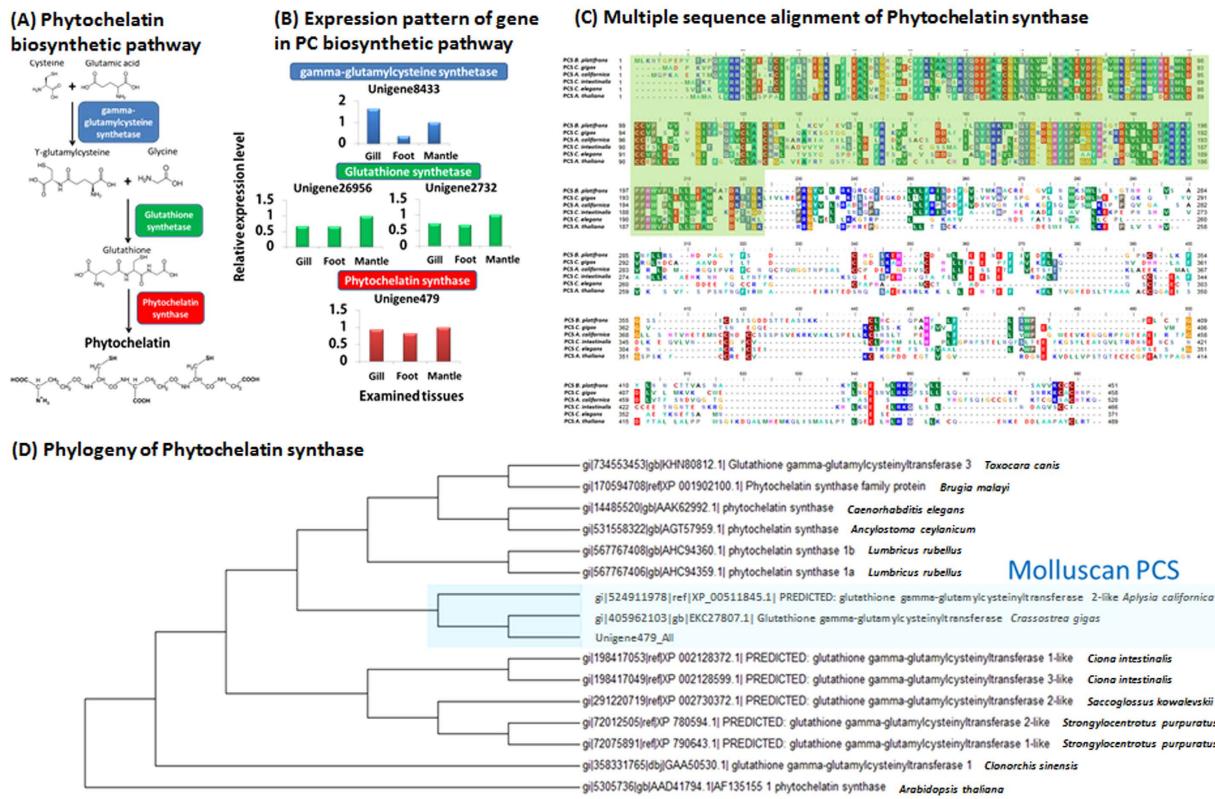


**Figure 3. Enrichment analysis of transcripts predominantly expressed in the gill, foot and mantle.** In the 3D scatterplot, transcripts predominantly expressed in the gill are red dots and highlighted in pale red circle; transcripts predominantly expressed in the foot are orange dots and highlighted in pale orange circle; transcripts predominantly expressed in the mantle were marked as green dot and highlighted in pale green circle. The top 10 most enriched GO Biological Process terms for each tissue specific dominant transcript list were indicated by arrows of the corresponding color.

	Pairwise comparison		
	Gill-to-Foot	Gill-to-Mantle	Foot-to-Mantle
log <sub>2</sub> (ratio) mean	0.05	0.06	-0.02
S.D.	2.55	2.22	1.72
2 S.D.	5.09	4.44	3.43
log <sub>2</sub> (ratio) cutoff for transcripts that were			
highly expressed in the gill	5.14 (mean + 2 S.D.)	4.50 (mean + 2 S.D.)	NA
highly expressed in the foot	-5.04 (mean - 2 S.D.)	NA	3.41 (mean + 2 S.D.)
highly expressed in the mantle	NA	-4.37 (mean - 2 S.D.)	-3.45 (mean - 2 S.D.)

**Table 2. Identification of differentially expressed transcripts in each pairwise comparison.**

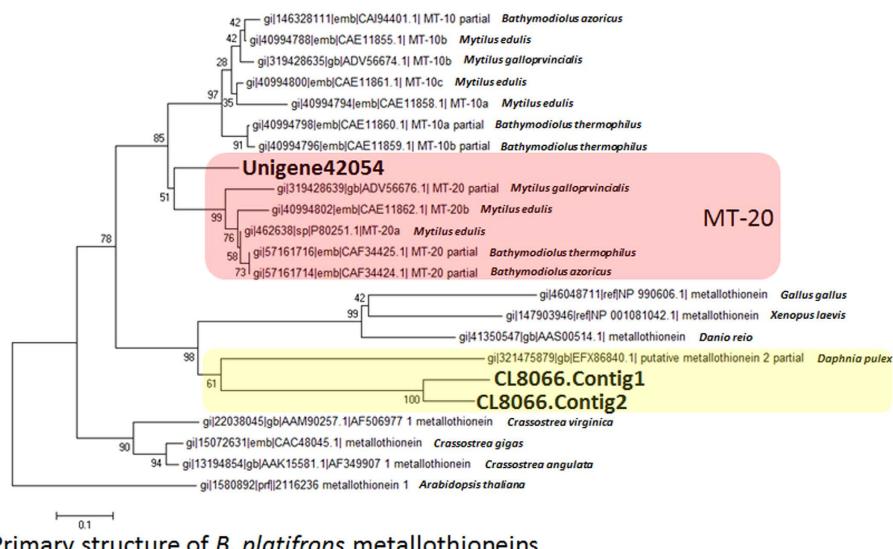
**Phytocochelin synthase and dominant expression of metallothioneins.** Phytocochelin, the oligomer of glutathione is one of the major molecules directly involved in metal ion detoxification<sup>19</sup>. We detected one transcript (Unigene479\_All) best matched to glutathione gamma-glutamylcysteinyltransferase 3, or phytocochelin synthase (PSC) from the Pacific Oyster *Crassostrea gigas*. Unigene479\_All did not show substantial variation in the three examined tissue. Alignment of *B. platifrons* PSC protein sequence with the orthologs from mollusks (*C. gigas*, *Aplysia californica*), urochordate (*Ciona intestinalis*), nematode (*C. elegans*), and plant (*Arabidopsis thaliana*) revealed that the N-terminal PC synthase domain in *B. platifrons* PSC (pfam domain no. PF05023, HMM prediction eval 4.1e<sup>-79</sup>) is highly conserved, but the C-terminal region is highly variable (see Fig. 4C). This result is consistent with the alignment result of



**Figure 4. *B. platifrons* phytochelatin synthase.** (A) The biosynthetic pathway of phytochelatin and (B) the expression pattern of genes involved in phytochelatin biosynthesis. (C) Protein sequence alignment of Unigene479\_All to phytochelatin synthase/Glutathione gamma-glutamylcysteinyltransferase from the Pacific Oyster *Crassostrea gigas* (EKC27807.1), the nematode *Caenorhabditis elegans* (AAK62992.1), the sea hare *Aplysia californica* (XP\_005110845.1), the tunicate *Ciona intestinalis* (XP\_005110845.1), and the model plant species *Arabidopsis thaliana* (AAD41794.1). The highly conserved N-terminal phytochelatin synthase domain was highlighted in pale green box. (D) Maximum likelihood phylogenetic tree of phytochelatin synthase with 1000 bootstrap replications. Molluscan PCS cluster were highlighted in pale blue box.

PCSs from plant, yeast and nematode<sup>20</sup>. Searching through the current transcriptome assembly, we found that the enzymes involved in major reactions in biosynthesis of phytochelatin were detected but did not show drastic expressional variation (not exceeding a two-fold difference) in the three examined tissues (see Fig. 4A,B). Nonetheless, our data suggested that the mussel is able to produce PC in all examined tissues. PCS sequences were also recovered in *B. azoricus* transcriptome (Fig. S5), as revealed by Blastp search of Unigene479 against the DeepSeaVent database, the customized protein database derived from *B. azoricus* gill transcriptome<sup>21</sup>. *B. platifrons* PCS clustered with PCS from *C. gigas* and *Aplysia californica* in the phylogenetic analysis (Fig. 4D), suggesting that the PCS in both *Bathymodiolus* mussel transcriptomes are unlikely to be originated from parasitic organisms.

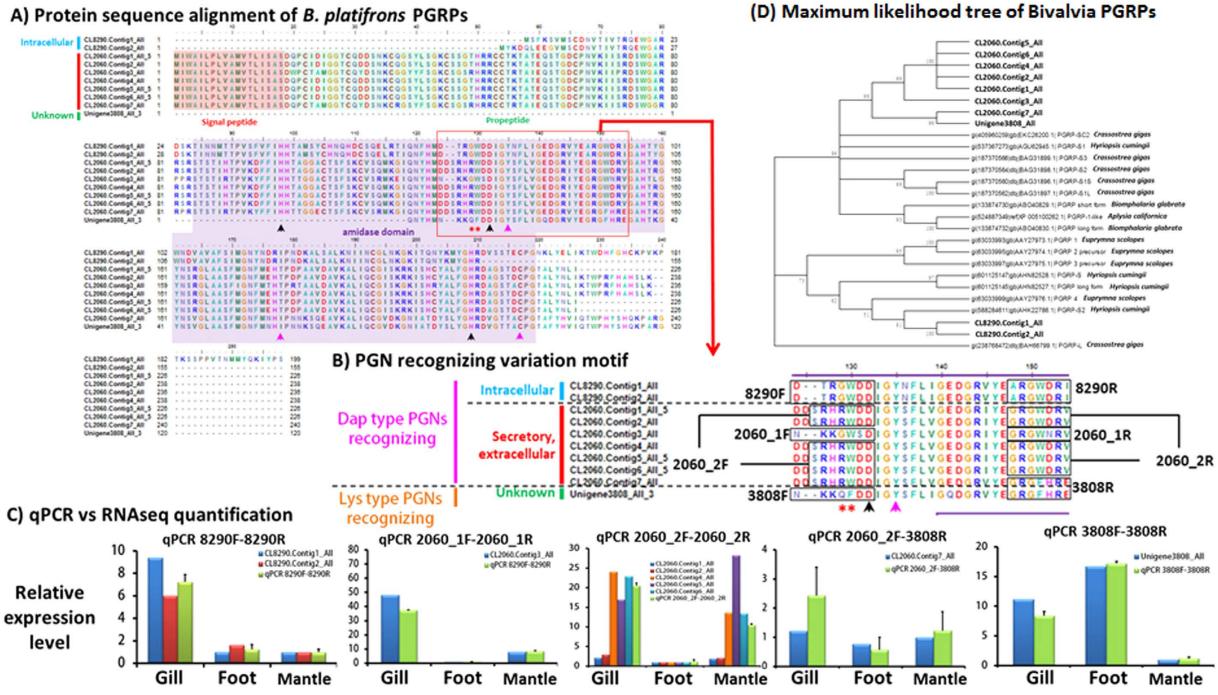
We then examined metallothionein (MT), a well-studied metal chelating protein. All three MT transcripts (Unigene42054, CL8066.Contig1, CL8066.Contig2) were predominantly expressed in all the three examined *B. platifrons* tissues (Table S5). Former studies on the MT from *B. azoricus* and *B. thermophilus* reported two MT variants, MT-10 and MT-20<sup>17</sup>. Our phylogenetic analysis suggested that, while Unigene42054 is homologous to the MT-20 cluster, CL8066.Contig1 and CL8066.Contig2 are clearly non-homologous to either *Mytilidae* or vertebrate MTs (Fig. 5A). In fact, while CL8066.Contig1\_All were poorly aligned to *Daphnia pulex* MT (bitscore 33.5, Evalue 1.8), CL8066.Contig2\_All has no significant match in nr (Table S5). However, the metallothionein-like proteins encoded by these two transcripts contain seven Cys-Xaa-Cys motifs (indicated by pale red boxes in Fig. 5B), suggesting their potential role in metal ion binding. On the other hand, *B. platifrons* MT-20 (product of Unigene42054) contains nine of Cys-Xaa-Cys motifs (indicated by pale red boxes in Fig. 5B) that are conserved within *Mytilidae*. Comparing to the MT-20 from *B. thermophilus* and *B. azoricus*, however, *B. platifrons* MT-20 as well as *M. edulis* MT-20 has one less Cys-Xaa-Cys motif at the 55<sup>th</sup>–57<sup>th</sup> residue (indicated by pale green box in Fig. 5B).

(A) Phylogeny of *B. platifrons* metallothioneins(B) Primary structure of *B. platifrons* metallothioneins

**Figure 5. Sequence analysis of metallothioneins (MT) detected in *B. platifrons* transcriptome. (A)** A maximum likelihood tree of MT with 1000 bootstrap replications. The MT-20 clade, where Unigene42054\_All situated, was highlighted in pale red box; CL2066.Contig1 and CL2066.Contig2, which clustered with *Daphnia pulex* MT, were highlighted in yellow box. **(B)** Sequence alignment of, in the upper part, Unigene42054\_All with MT-20 from *Mytilus edilis* (CAE11862.1), *Bathymodiolus azoricus* (CAF34424.1) and *B. thermophiles* (CAF34425.1), and, in the lower part, CL2066.Contig1 and CL2066.Contig2. Note that the Cys-Xaa-Cys motifs in the MT sequences were highlighted in pale red boxes.

**Sulfur metabolism.** We detected four transcripts of sulfite reductase (three alpha and one beta component), which catalyzes the oxidation of sulfide to sulfite, and two of sulfite oxidase, which catalyzes the oxidation of sulfite to sulfate, in the current *B. platifrons* transcriptome. Both sulfite oxidase transcripts were predominantly expressed in gill (Table 2). Based on the nr annotation, three of the sulfite reductase transcripts (Unigene8776\_All, Unigene19303\_All and Unigene58781\_All) were from eukaryotic source while the remaining one transcripts (Unigene46612\_All) was from methane-oxidizing bacteria. We also detected three transcripts annotated as sulfide:quinone oxidoreductase (SQR), a membrane bound mitochondrial protein catalyzing the formation of polysulfide from sulfide using a quinone molecule<sup>22</sup>. Two of the SQR transcripts shown a relative higher expression in gill, followed by foot and then mantle. The transcripts of taurine transporter (TAUT), which was implicated in sulfide detoxification and supplying sulfur to thiotrophic symbionts in Bathymodiolid mussels<sup>23,24</sup>, exhibited higher expression in the mantle rather than in the gill.

**Expression patterns of genes implicated in immune response.** We found a repertoire of immune system genes that were known to be up-regulated in response to experimental introduction of virulent factors such as peptidoglycans, lipopolysaccharides and beta-glucans, in *C. gigas* or coastal mussels *Mytilus* spp. (Table S6). The list included pattern recognizing proteins peptidoglycan recognition protein (PGRP), Toll-like receptor and bactericidal permeability increasing protein (BPI), downstream transcription factor Lipopolysaccharide-induced tumor necrosis factor-alpha factor (LITAF), inflammatory cytokines interleukin-17 (IL-17) and tumor necrosis factor (TNF), antimicrobial peptides big defensin (BD) and macin. Certain transcripts of PGRP, BPI, TLR, IL-17, BD and macin exhibited distinctive tissue expression pattern (at least three fold differences) in either one of the three pairwise comparisons (gill-to-foot, gill-to-mantle, foot-to-mantle) (highlighted by colored box in Table S6). To further investigate the potential molecular mechanisms of the host-symbiont interaction, in the following paragraphs, we focused on the PGRP, one of the crucial microbial associated molecule pattern recognition proteins, for sequence homology analysis.



**Figure 6. Sequence analysis of peptidoglycan recognition protein isoforms detected in *B. platifrons* transcriptome.** (A) Sequence alignment of all PGRP transcripts in *B. platifrons* transcriptome. N-terminal signal peptides were highlighted in pale red box; propeptide sequences were highlighted in pale green box. amidase domains were highlighted in pale purple box. Black arrows indicated residues essential for zinc-binding function. Purple arrows indicated residues essential for hydrolytic amidase activity. Red asterisks indicated the two residues that determine the binding selectivity of PGRPs to Lys-type and Dap-type PGNs. (B) The PGN recognizing variation motif as signified by red outline box in (A). The black, purple arrows and red asterisks indicated the same functional residues as stated above. Note the G or RW residues are selective to Dap-type PGN while QF residues are selective to Lys-type PGN. Gene specific primer pairs targeting different PGRP isoforms was indicated in black boxes. (C) RNAseq quantifications versus qRT-PCR results. All qRT-PCR results were shown in light green bar in all charts while other color bars indicated the RNAseq quantification of different PGRP transcripts in different tissues. (D) Maximum likelihood tree of Bivalvia PGRPs with 1000 bootstrap replications.

**Differential expression of peptidoglycan recognition protein isoforms.** In the current *B. platifrons* transcriptome, 11 transcripts were matched to PGRP sequences (Table S6). Two of the 11 transcripts were grouped under the cluster CL8290 and eight under CL2060. All the PGRPs encoded by the transcripts in CL2060 were predicted to contain an N-terminal signal peptide, followed by a propeptide and then by the amidase domain (Fig. 6A). This result indicated the differences in cellular localization of different PGRPs isoforms. It is unknown at this stage if the PGPR encoded by Unigene3808 is secretory or not as Unigene3808 is a 3' partial sequence. Based on sequence homology, however, this transcript was likely to be derived from the same gene as the contigs from CL2060 (Fig. 6A). Multiple sequence alignment with insect and vertebrate PGRPs (see Fig. S8) revealed that all the amino acid residues essential for zinc binding (black arrow in Fig. 6A)<sup>25</sup> and amidase activity (purple arrow in Fig. 6A)<sup>26</sup> were conserved in all the 11 PGRP transcripts. Several insect based studies revealed two variable residues (indicated by red asterisk in Fig. 6A,B) that might affect peptidoglycan (PGN) binding selectivity towards Lysine-type (Lys-type) and diaminopimelic acid-type (Dap-type) PGNs<sup>27–32</sup>. We found that these two transcript variants also exist among *B. platifrons* PGRPs. For the PGRPs encoded by transcripts from CL8290 and CL2060, the two residues corresponding to the two variable residues in insect PGRPs were Gly/Arg-Trp. For PGRP encoded by Unigene3808\_All, the two corresponding residues were Gln-Phe. These suggested that PGRPs encoded by transcripts from CL8290 and CL2060 might favor the binding of the Dap-type PGNs<sup>28</sup> while that from Unigene3808\_All might favor binding of Lys-type PGNs (indicated by red asterisk in Fig. 6B)<sup>30</sup>. Differential expression analysis based on our RNAseq data revealed that Unigene3808 had highest expression level in the foot while transcripts from CL8290 and CL2060 showed higher expression in the gill and mantle. Such differential expression pattern of these PGRP transcripts were confirmed by qRT-PCR assays (Fig. 6C) using primers specifically recognizing PGRP isoforms with Gly-Trp, Arg-Trp or Gln-Phe residues (Fig. 6B). Taken together, our results indicate that

isoforms of PGRP may have differential selectivity towards different types of PGNs, and with different abundance in different tissues.

## Discussion

Our primary objective was to produce a comprehensive transcriptome database for each examined tissue for the methane seep dwelling *B. platifrons*. Hence, we pooled the total RNA of each tissue from three individuals during the cDNA library preparation step to maximize the sequencing coverage with the same amount of output. Because there were no replicates, our differential gene expression comparisons among the three tissues only targeted genes that were remarkably different. As quantification accuracy is positively correlated with sequence coverage<sup>17</sup>, transcripts with low coverage (with less than 5 ppm total mapped reads) were filtered out. Similar filtering approach (10 ppm total mapped reads as cutoff) was adopted in a coral transcriptomic study<sup>33</sup>. Our qRT-PCR assays results of 30 differential expressed transcripts shown good agreement with the RNAseq results, indicating that our approach on differential tissue expression analysis was valid.

In the current study, we unexpectedly recovered 1,402 microbial transcripts in our *B. platifrons* transcriptome assembly, in which up to 16% of these transcripts were likely to be originated from the methanotroph symbiont. Bacterial genes from the bacterial family Methylococcaceae were also detected in a transcriptomic study of *B. azoricus*<sup>14</sup>. In prokaryotes, mRNA is mature upon transcription<sup>34</sup> and only a small fraction of mRNAs is polyadenylated at any given time<sup>35</sup>. Instead of stabilizing the RNA molecules, polyadenylation of RNA promotes degradation<sup>36</sup>. Hence, the microbial transcripts in *B. platifrons* transcriptome may represent a fraction of predominant bacterial mRNA that have high turnover rate. While the discovery of transcripts related to methane oxidation was not surprising, the gill specific expression of denitrification genes was unexpected. This may indicate that the methanotrophic endosymbionts might be actively catalyzing the production of ammonium via denitrification. A previous study has reported Bathymodiolid mussels to actively assimilate ammonium and free amino acids but not nitrate<sup>37</sup>. The authors of that study estimated that, in the presence of methane, the acquisition rate of ammonium correlated with the carbon fixation rate<sup>37</sup>. Our transcriptomic data suggested that, in addition to acquiring host ammonia waste, the methanotrophic symbiont may also be able to produce ammonia as their nitrogen source. On the other hand, substantial number of foot dominant bacterial transcripts was assigned to the nitrogen-fixing bacterial family Bradyrhizobiaceae. To date, no reports have described endo- or ectosymbiosis in the foot of any *Bathymodiolus* mussel. Dang *et al.*<sup>38</sup> recovered diverse nifH and nifH-like gene sequences from the surface sediments of the methane seep in the Okhotsk Sea. In addition, Rhizobia have been isolated from organic rich sapropel layers in the Mediterranean seabed<sup>39</sup>. These studies suggested the presence of nitrogen-fixing bacteria in the sediment of the methane seep and these nitrogen-fixing bacteria might be attached to the surface of the foot tissue of the mussel during sediment digging. Nonetheless, the presence of Bradyrhizobiaceae bacteria in the foot has to be validated by 16S amplicon pyrosequencing.

In the present study, we found that a number of pattern recognition proteins such as PGRP, BPI, Toll-like receptors were highly expressed in the gill compared to the foot and mantle. Yet, our RNAseq data showed that different isoforms of the same gene the expression of immune system genes can exhibit highly variable expression pattern. For instance, while all *B. platifrons* PGRP isoforms contain amidase domain, they may have different binding affinity toward Gram-positive bacterial (Lys-type) and Gram-negative and Gram-positive bacilli (Dap-type) PGNs. Our RNAseq data indicated that the Dap-type PGNs binding PGRPs have a much higher expression level in the gill compared to the foot. The Lys-type PGNs binding PGRP, on the other hand, exhibited a slightly higher expression level in the foot compared to the gill. It is worth noting that the selectivity of PGRPs to Lys-type and Dap-type PGNs may not be absolute<sup>26,32</sup>. At this stage, we cannot confirm if different PGRPs isoforms in *B. platifrons* really exhibit selectivity toward different types of PGN. Nevertheless, we speculate that the expression patterns of different PGRP isoforms were related to the gram-negative methanotrophic symbiont in the gill of *B. platifrons*. In insects, the PGRP-LB has been shown to be a key factor in suppressing virulent bacteria. It has been suggested that the outer membrane protein OmpA of the endosymbiotic bacteria *Sodalis glossinidius* can interact with the Tse-Tse fly PGRP-LB (one of the long PGRP splice variants). Binding of OmpA to PGRP-LB lead to suppression of the immunodeficiency (Imd) pathway<sup>40</sup> and thereby the insect host immune response. Given that 1) PGRP-LB is required in insect-symbiont interaction, 2) the C-terminal sequence of major outer membrane protein (MopB) of the methanotrophic bacteria *Methylococcus capsulatus* shows significant identity to the OmpA family<sup>41</sup>, and 3) these PGRPs were dominantly expressed in the gill, we propose that *B. platifrons* PGRPs is a good candidate gene for studying the mechanism of immunotolerance of the *Bathymodiolus* mussel host to the methane-oxidizing bacteria.

The endemism of *B. platifrons* in both hydrothermal vent and methane seeps in the Pacific Ocean<sup>8</sup> suggested that this Bathymodiolid species has the metabolic capacity to metabolize a range of toxic chemical substances in these habitats. Distinctively high heavy metal concentration is one of the signposts of hydrothermal vent and methane seep habitats<sup>6,7</sup>. High concentration of heavy metal ions promotes the production of reactive oxygen species, which could be hazardous to cellular metabolic reactions<sup>42</sup>. Hemostasis of heavy metal ions is therefore of crucial importance in maintaining a stable intracellular chemical environment. Up-regulation of MTs, a class of ubiquitously expressing low molecular weight cysteine-rich metal ion binding proteins implicated in heavy metal homeostasis<sup>42</sup>, has been documented

in various bivalves after exposure to heavy metals<sup>43–45</sup>. The dominant expression of MT in *B. platifrons* indicated high heavy metal level in the seep environment. A group of MT-like proteins, which contains seven Cys-Xaa-Cys motifs but did not show any conservation (except the Cys-Xaa-Cys motifs) to any known MTs, was recovered from the current *B. platifrons* transcriptome. Further *in vitro* experiments are needed to determine the affinity of this MT-like protein toward different metal ions.

One of the surprising finding in term of heavy metal detoxification was the discovery of glutathione (GSH) gamma-glutamylcysteinyltransferase, or phytochelatin synthase (PCS). Phytochelatin(PC) contains repeats of Y-GluCys dipeptide followed by a terminal Gly. In the presence of metal ion, PCS catalyzed the transpeptidation of the Y-GluCys moiety of GSH onto a second GSH molecule to form repeats of PC, or a Y-GluCys dipeptide-metal complex<sup>20,46</sup>. *In vivo* study using PC from plants demonstrated that PC preferably chelates Cd, but can also form complex with Ag and Cu<sup>47</sup>. PCS has been mainly characterized in plant, yeast and parasitic nematodes<sup>47–49</sup>. Hence, we did suspect that the recovery of PCS in *B. platifrons* transcriptome might have been due to the contamination by small parasitic organisms such as nematodes. These small organisms were hard to be seen and removed during sampling. Recently, however, PCS has been reported in several molluscan species, including the Pacific Oyster *C. gigas* genome<sup>50</sup>. More importantly, sequence alignment as well as phylogenetic analysis suggested that the PCS in *B. platifrons* transcriptome clustered with the PCS from *C. gigas*. In fact, PCS was also expressed in the gill of *B. azoricus* (as indicated by the result of reciprocal blast search, Fig. S6). This suggested that *B. azoricus* also has the potential to produce PC. This result further supported that the PCS sequence in *B. platifrons* transcriptome was not from a contamination source. There is so far no report of PCS gene or production of PC in any member of Mytilidae. This may imply that either PCS gene is specific to Bathymodiolid mussels among Mytilidae or previous high-throughput studies on coastal member of Mytilidae (such as *Mytilus* spp.) have missed out PCS gene expression, owing to limitations on sequencing or annotation. Detection of PC in *B. platifrons* tissues by analytical chemistry method is required to validate if the mussel is really producing PC *in vivo*. It will also be essential to examine if PC biosynthesis increase in response to heavy metal administration.

Hydrogen sulfide is another major component in vent and seep effluent<sup>8,9</sup>. The compound is highly inhibitive to aerobic respiration. While the blood of tubeworm and another bivalve were capable of buffering the acidifying effect of hydrogen sulfide<sup>51,52</sup>, the blood of Bathymodiolid mussels did not show such buffering characteristics, implying that there is no sulfide-binding protein in its blood<sup>53</sup>. Instead, it was suggested that sulfide oxidase activity in both the symbiont-containing gill tissue and peripheral lining of other symbiont-free tissues serves as “peripheral defense” line to oxidize any sulfide that enters the cells to sulfite<sup>53</sup>. In the current study, we found that, except sulfite oxidase, the expression of mitochondrial enzymes that participate in sulfide oxidation or sulfide detoxification did not show drastic differential expression among the three examined tissues. This finding is in accordance with the existence of a peripheral defense line against sulfide in both symbiont-containing and symbiont-free tissues. Comparing to sulfite reductase, transcripts of SQR showed a much higher expression value (rpkm). SQR catalyzes the formation of polysulfide from sulfide. SQR in lugworms and other marine invertebrates that inhabit high sulfide environments<sup>54</sup> avoid elevation of intracellular sulfide concentrations. Abundant expression of SQR in *B. platifrons* suggested that the mussel may mainly remove sulfide by the mean of SQR catalyzing polysulfide formation.

High levels of thiotaurine has been found in the tissues of vent and seep invertebrates, leading to the hypothesis that the unusual amino acid is involved in the detoxification of sulfide<sup>55–57</sup>. In the thiotrophic symbiont harboring *B. septemfierum*, the expression level of the taurine transporter TAUT was higher in the gill compared to the foot and mantle<sup>58</sup>. This result has led Kioto *et al.*<sup>58</sup> to propose that the mussel gill epithelial cells supply sulfur to the thiotrophic symbionts with thiotaurine and hypotaurine by TAUT. In *B. platifrons* from Sagami Bay, however, although the mussel did not contain thiotrophic symbionts, TAUT expression also increased in response to a continuous exposure to sulfide (at 1 atm)<sup>59</sup>, further suggesting that the involvement of TAUT in sulfide detoxification. In the current study, however, the expression level of TAUT was higher in the mantle than in the gill. Such inconsistent finding might be due to a lower sulfide level in the methane seep environment.

## Conclusion

We presented a transcriptomic database for the cold seep mussel *B. platifrons*. We highlighted a repertoire of genes involved in the innate immune system, heavy metal binding and sulfide detoxification. The expression patterns of these genes in the gill, foot and mantle were also reported. In particular, we reported a potentially novel metal binding MT-like protein and the possibility of PC production in *B. platifrons*. Our in-depth analysis of PGRP, BD and macin demonstrated that different isoforms of the same genes may exhibit a very different tissue expression pattern. These expressional differences may be related to functional variation among isoforms. Overall, we believe that our transcriptomic data will serve as an important reference for future research on adaptation Bathymodiolid mussels and host-symbiont interaction of Bathymodiolid mussels.

## Materials and Methods

**Sample collection and preparation for Illumina sequencing.** Mussel samples were collected on June 19, 2013 using the manned submersible Jiaolong. The sampling site was a cold seep located at

a depth of 1122 m on the continental slope of the South China Sea ( $22^{\circ} 06.921' N$ ,  $119^{\circ} 17.131' E$ ). Upon arrival at the sea surface, the gill, the foot and the mantle tissues (Fig. 1) from three specimens of *B. platifrons* were dissected and stored individually in RNAlater (Life Technologies, USA). The shell lengths of the three specimens were 7.4 cm, 8.1 cm and 7.6 cm. One of the three *B. platifrons* was genotyped by sequencing the mitochondrial genes COI and NADH4 using previously reported primers<sup>15</sup>. These sequences were deposited in NCBI with the accession numbers KJ174328 and KJ174329. Around 1 g of gill, 4 g of adductor foot muscle and 4 g of mantle tissues from each mussel were used for RNA extraction. Each tissue sample was placed into a sterile ( $180^{\circ} C$  overnight) mortar that had been pre-chilled with liquid nitrogen. Additional liquid nitrogen was poured onto the tissue sample. A pestle pre-chilled in liquid nitrogen was used to crack the frozen tissues into fine powder. TRIzol reagent (Invitrogen, USA) was then added and the sample was mixed using the pestle. RNA extraction was performed according to the manufacturer's protocol excluding the RNA precipitation step where 1/2 volume of 3 M sodium acetate and 1/2 volume of pre-chilled isopropanol were added to the retrieved aqueous phase solution. The RNA pellets were dissolved in diethylpyrocarbonate (DEPC)-treated water and stored in  $-80^{\circ} C$  until cDNA synthesis. Total RNA (10  $\mu g$ ) from each tissue was re-precipitated by adding 1/5 volume of 10 M lithium chloride at  $-20^{\circ} C$  for one hour. The RNA pellets were re-dissolved in DEPC-treated water. mRNA was isolated with Oligo(dT) beads and fragmented into small pieces to avoid priming bias. Using these short fragments as templates, random hexamers were used to synthesize double-stranded cDNA using SuperScript® Double-Stranded cDNA Synthesis Kit (Invitrogen, USA). The synthesized cDNA was subjected to end-repair, phosphorylation, 3' adenylation, and ligation to sequencing adapters using Illumina Truseq RNA Sample Preparation kit v2 (Illumina, USA).

**Illumina sequencing, de novo assembly and functional annotation of *B. platifrons* transcriptomes.** Illumina sequencing, *de novo* assembly and functional annotation were performed by the Beijing Genomics Institute, Hong Kong. Illumina sequencing was performed using a HiSeq™ 2000 (Illumina, USA) in paired-end mode with a read length of 100 bp. Five Gbp clean data was targeted for each tissue library. Briefly, cDNA libraries from the three tissues were sequenced separately. Raw reads with adaptor sequences, or with more than 5% unknown nucleotides, or with more than 20% low quality bases (base quality  $\leq 10$ ) were removed using the NGS QC toolkit package (version 2.3)<sup>60</sup>. All of the clean reads from each examined tissue were firstly assembled independently using Trinity assembly algorithm<sup>61</sup>. The assembling parameters were set as “seqType=fq, min\_contig\_length=100, min\_kmer\_cov=2”, with the rest being default parameter. All the sequences from all three tissue specific transcriptomes were then taken into further process of redundancy removing using CD-HIT-EST v4.6<sup>62</sup> with a sequence identity threshold of 99% in every 1000 bp. Gene family clustering analysis was performed using MC-UPGMA v1.0.0<sup>63</sup>, where non-redundant sequences with similarity between each other exceeding 70% were assigned into one gene family cluster. The prefix “CL” was given to these transcript sequences, e.g. CL1190.Contig1\_All, CL1190.Contig2\_All, CL1190.Contig3, etc. Non-redundant sequences that did not form any cluster remain as singletons and their prefix remained as Unigene.

**Functional annotation and protein coding sequence prediction.** All the transcript sequences were searched against protein databases including NCBI non-redundant (nr) database (as of October, 2013), Swiss-Prot, KEGG and COG (e-value  $< 0.00001$ ) using BLASTx (Version: BLAST-2.2.28+)<sup>64</sup>. Gene Ontology terms (as of October 20, 2013) information for the annotated transcripts were retrieved from the NCBI database (as of 20<sup>th</sup> October, 2013)<sup>65</sup>. Taxonomical assignment of annotated transcripts was performed using the LCA assignment algorithm in MEGAN v5.2.3)<sup>66</sup> based on the top 20 hits for each transcript in the nr database. For the protein coding sequence prediction, BLASTx alignment results were used to decide the frame direction and the coding region of the transcript. For transcripts that could not be orientated by BLASTx alignment, ESTScan was performed. The coding region of all of the transcripts was then translated into amino acid sequences based on the standard codon table. Protein domain discovery was performed using the software InterProScan 5 (v5.8–49)<sup>67</sup>, with Pfam as the database and an evalue cutoff  $\leq 0.01$ . To evaluate the comprehensiveness of the current *B. platifrons* transcriptome assembly, we downloaded the 458 single-copy genes (core proteins) from the Core Eukaryotic Gene Mapping Approach (CEGMA) dataset<sup>16</sup>. We used tBLASTN with a evalue cut-off of  $10^{-3}$  to assay the completeness of both assembly. The same procedure was conducted to assess the completeness of a decapod genome<sup>68</sup>.

**Expression level and differential expression analysis.** Clean reads from each tissue specific read library were aligned to the whole body transcriptome using Bowtie2<sup>69</sup>. The number of aligned reads per transcript was counted using SAMtools (v2.0.1)<sup>70</sup>. Transcript expression levels in each tissue are presented as FPKM (fragments per kilobase of transcript per million mapped reads)<sup>71</sup>. To investigate the dominant gene function in each tissue, transcripts were sorted based on their total expression level (summation of expression value in all three examined tissues). Cumulative hypergeometric distribution tests were conducted to extract significantly over-represented KEGG terms in the list of the top 5% of the expressed transcripts<sup>72</sup>. Enrichment for a functional annotation (such as KEGG or GO) term was determined based on the P value  $\leq 0.001$ . The degree of enrichment of an enriched term was indicated as a Rich factor, which was calculated as the number of transcripts with the enriched term in the list

divided by the total number of transcripts with the enriched term. The closer the rich factor to 1, the higher the fold-enrichment of the enriched term.

As quantification accuracy is positively correlated with sequence depth<sup>17</sup>, prior to the differential expression analysis, transcripts that have less than 750 clean reads (equivalent to 5 ppm total mapped reads) mapped to the three examined tissues were filtered out. This resulted in 7,304 transcripts for differential expression analysis. Although such transcript filtering parameter was far more stringent comparing to a transcriptomic study conducted using the mussel *Perna viridis*<sup>73</sup>, applying such parameter can ensure high quantification accuracy among abundantly expressed transcripts. Similar filtering approach (10 ppm total mapped reads as cutoff) was adopted in a coral transcriptomic study<sup>33</sup>. Three pairwise comparisons were performed: Gill vs Foot, Gill vs Mantle and Foot vs Mantle. In each comparison, the log<sub>2</sub> RPKM ratios of each transcript were calculated. The mean and standard deviation (SD) of the log<sub>2</sub>(ratio) in each comparison were then computed. To ensure a high accuracy in identification of differentially expressed transcripts, differential expressions were defined as 2 SD from the mean log<sub>2</sub>(ratio) value in each pairwise comparison. Such stringent criterion for identification of differentially expressed genes is rather common in microarray-based studies<sup>74,75</sup>. GO Biological function (GOBP) terms enrichment analysis in these tissue specific dominant gene lists were performed using cumulative hypergeometric distribution tests as described. Enrichment for a GOBP term was determined based on the P value  $\leq 0.0001$ . The Rich factor of an enriched term was calculated as described.

**Validation of differentially expressed genes by semi-quantitative real-time PCR.** To evaluate the reliability of our differential expression analysis, primer pairs specific to 30 selected transcripts, as well as to PGPR mentioned in the result part (see Fig. 6), were designed (see result part on the criteria of selection of these transcripts) for qRT-PCR assay. The 18S rRNA transcript was used as an internal standard within each sample. All gene specific primer sequences are listed in Table S7.

The cDNA samples prepared for qRT-PCR assays were derived from the three specimens that were used for preparing cDNA for Illumina sequencing. Briefly, total RNA from each specimen was prepared using TRIzol as described above. DNA contaminant was removed using the Turbo DNA-free kit (Ambion, USA). The double-stranded cDNA was synthesized by using the Superscript double-stranded cDNA synthesis kit (Invitrogen, USA) with the random hexamer primer according to the manufacturer's protocol. For each transcript, duplicated qRT-PCR assays were performed on each tissue from each mussel specimen (i.e. 3 biological specimens X 2 technical repeats per tissue). All qRT-PCR assays were carried out using iTaqSYBR Green Supermix with ROX (BioRad Life Science, United States) and were run on a Stratagene mx3000p PCR machine (Agilent Technologies, Santa Clara, CA, United States). The results were normalized using the housekeeping gene 18S. The relative expression levels were calculated using the  $2^{-\Delta\Delta CT}$  method<sup>76</sup>. To compare the qRT-PCR result against RNAseq quantification of the selected transcripts, the differential expression level between the gill and the foot (Gill-to-Foot ratio), the gill and the mantle (Gill-to-Mantle ratio), and the foot and the mantle (Foot-to-Mantle ratio) from both quantification methods were calculated and then log<sub>2</sub> transformed. For each comparison, the log<sub>2</sub> transformed differential expression level of each transcript from both quantification methods were plotted on a scatterplot and the Pearson's r ( $R^2$ ) value was calculated by excel.

**Sequence analysis.** All sequence alignment and phylogenetic tree construction were performed in the software MEGA 6<sup>77</sup>. Sequence alignment of PGPRs, PCSs and MTs was performed by the built-in alignment program MUSCLE<sup>78</sup>. Maximum likelihood tree of MTs, PCS and Bivalvia PGPRs were constructed based on a G + I Jones-Taylor-Thornton (JTT) model, with 1000 bootstraps. Prediction of signal peptide and propeptide sequences of PGPRs was performed by ProP 1.0 server<sup>79</sup>.

## References

1. Jannasch, H. W. & Mottl, M. J. Geomicrobiology of deep-sea hydrothermal vents. *Science* **229**, 717–725 (1985).
2. Snelgrove, P. V. R. & Smith, C. G. A riot of species in an environmental calm: the paradox of the species-rich deep-sea floor. In: Gibson R. N., Barnes M. & Atkinson R. J. A. editors. *Oceanogr. Mar. Biol.* **40**, 311–342 (2003).
3. De Chaine, E. G., Bates, A. E., Shank, T. M. & Cavanaugh, C. M. Off-axis symbiosis found: Characterization and biogeography of bacterial symbionts of Bathymodiolid mussels from Lost City hydrothermal vents. *Environ. Microbiol.* **8**, 1902–1912 (2006).
4. Huber, J. A. et al. Microbial population structures in the deep marine biosphere. *Science* **318**, 97–100 (2007).
5. Sibuet, M. & Olu, K. Biogeography, biodiversity and fluid dependence of deep-sea cold-seep communities at active and passive margins. *Deep-Sea Res.* **45**, 517–567 (1998).
6. Johnson, K. S., Childress, J. J., Hessler, R. R., Sakamoto-Arnold, C. M. & Beehler, C. L. Chemical and biological interactions in the Rose Garden hydrothermal vent field, Galapagos spreading center. *Deep-Sea Res.* **35**, 1723–1744 (1988).
7. Gieskes, J. et al. A study of the chemistry of pore fluids and authigenic carbonates in methane seep environments: Kodiak Trench, Hydrate Ridge, Monterey Bay, and Eel River Basin. *Chem. Geol.* **220**, 329–345 (2005).
8. Jones, W. J. et al. Evolution of habitat use by deep-sea mussels. *Mar. Biol.* **148**, 841–851 (2006).
9. Duperron, S. et al. Dual symbiosis in a *Bathymodiolus* sp. mussel from a methane seep on the Gabon continental margin (Southeast Atlantic): 16S rRNA phylogeny and distribution of the symbionts in gills. *Applied. Environ. Microbiol.* **71**, 1694–1700 (2005).
10. Won, Y. J. et al. Environmental acquisition of thiotrophic endosymbionts by deep-sea mussels of the genus *Bathymodiolus*. *Applied. Environ. Microbiol.* **69**, 6785–6792 (2003).
11. Cosson, R. P. et al. Spatial variation of metal bioaccumulation in the hydrothermal vent mussel *Bathymodiolus azoricus*. *Mar. Environ. Res.* **65**, 405–415 (2008).

12. Powell, M. A. & Somero, G. N. Adaptations to sulfide by hydrothermal vent animals: sites and mechanisms of detoxification and metabolism. *Biol. Bull.* **171**, 274–290 (1986).
13. Hardivillier, Y. et al. Do organisms living around hydrothermal vent sites contain specific metallothioneins? The case of the genus *Bathymodiolus* (Bivalvia, Mytilidae). *Comp. Biochem. Physiol. C* **139**, 111–118 (2004).
14. Bettencourt, R. et al. High-throughput sequencing and analysis of the gill tissue transcriptome from the deep-sea hydrothermal vent mussel *Bathymodiolus azoricus*. *BMC Genomics* **11**, 559 (2010).
15. Barry, J. P. et al. Methane-based symbiosis in a mussel, *Bathymodiolus platifrons*, from cold seeps in Sagami Bay, Japan. *Invertebr. Biol.* **121**, 47–54 (2002).
16. Parra, G., Bradnam, K. & Korf, I. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
17. Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* **8**, 469–477 (2011).
18. Wada, K. & Fujinuki, T. in *The Mechanisms of Mineralization in Invertebrates and Plants*, Biominerization in bivalve molluscs with emphasis on the chemical composition of the extrapallial fluid, 175–190 (University Of South Carolina Press, 1976).
19. Cobbett, C. S. Phytochelatins and their role in heavy metal detoxification. *Plant Physiol.* **123**, 825–33 (2000).
20. Clemens, S., Kim, E. J., Neumann, D. & Schroeder, J. I. Tolerance to toxic metals by a gene family of phytochelatin synthases from plants and yeast. *EMBO J.* **18**, 3325–3333 (1999).
21. The DeepSeaVent database, an exhaustive collection of genes expressed in the deep-sea hydrothermal vent mussel *Bathymodiolus azoricus*. Available at: <http://transcriptomics.biocant.pt:8080/deepSeaVent/?rvn=1#> (Accessed at 3<sup>rd</sup> Jan 2015).
22. Arieli, B., Shahak, Y., Taglicht, D., Hauska, G. & Padan, E. Purification and characterization of sulfide-quinone reductase, a novel enzyme driving anoxygenic photosynthesis in *Oscillatoria limnetica*. *J. Biol. Chem.* **269**, 5705–5711 (1994).
23. Inoue, K. et al. Possible role of a taurine transporter in the deep-sea mussel *Bathymodiolus septendierum* in adaptation to hydrothermal vents. *FEBS Lett.* **582**, 1542–1546 (2008).
24. Koito, T. et al. Effect of long-term exposure to sulfides on taurine transporter gene expression in the gill of the deep-sea mussel *Bathymodiolus platifrons*, which harbors a methanotrophic symbiont. *Fisheries Sci.* **76**, 381–388 (2010).
25. Cheng, X., Zhang, X., Pflugrath, J. W. & Studier, F. W. The structure of bacteriophage T7 lysozyme, a zinc amidase and an inhibitor of T7 RNA polymerase. *Proc. Natl. Acad. Sci. USA* **91**, 4034–4038 (1994).
26. Mellroth, P., Karlsson, J. & Steiner, H. A scavenger function for a *Drosophila* peptidoglycan recognition protein. *J. Biol. Chem.* **278**, 7059–7064 (2003).
27. Michel, T., Reichhart, J. M., Hoffmann, J. A. & Royet, J. *Drosophila* Toll is activated by Gram-positive bacteria through a circulating peptidoglycan recognition protein. *Nature* **414**, 756–759 (2001).
28. Gottar, M. et al. The *Drosophila* immune response against Gram-negative bacteria is mediated by a peptidoglycan recognition protein. *Nature* **416**, 640–644 (2002).
29. Lim, J. H. et al. Structural basis for preferential recognition of diaminopimelic acid-type peptidoglycan by a subset of peptidoglycan recognition proteins. *J. Biol. Chem.* **281**, 8286–8295 (2006).
30. Park, J. W. et al. Clustering of peptidoglycan recognition protein-SA is required for sensing lysine-type peptidoglycan in insects. *Proc. Natl. Acad. Sci. USA* **104**, 6602–6607 (2007).
31. Royet, J. & Dziarski, R. Peptidoglycan recognition proteins: pleiotropic sensors and effectors of antimicrobial defences. *Nat. Reviews Microbiol.* **5**, 264–277 (2007).
32. Gelius, E., Persson, C., Karlsson, J. & Steiner, H. A mammalian peptidoglycan recognition protein with N-acetylmuramoyl-L-alanine amidase activity. *Biochem. Biophys. Res. Com.* **306**, 988–994 (2003).
33. Meyer, E., Aglyamova, G. V. & Matz, M. V. Profiling gene expression responses of coral larvae (*Acropora millepora*) to elevated temperature and settlement inducers using a novel RNA-Seq procedure. *Mol. Ecol.* **20**, 3599–3616 (2011).
34. Proudfoot, N. J., Furger, A. & Dye, M. J. Integrating mRNA processing with transcription. *Cell* **108**, 501–512, (2002).
35. Dreyfus, M. & Régnier, P. The poly (A) tail of mRNAs: bodyguard in eukaryotes, scavenger in bacteria. *Cell* **111**, 611–613 (2002).
36. Blum, E., Carposis, A. J. & Higgins, C. F. Polyadenylation promotes degradation of 3'-Structured RNA by the *Escherichia coli* mRNA degradosome *in vitro*. *J. Biol. Chem.* **274**, 4009–4016 (1999).
37. Lee, R. W., Thuesen, E. V., Childress, J. J. & Fisher, C. R. Ammonium and free amino acid uptake by a deep-sea mussel (*Bathymodiolus* sp., undescribed) containing methanotrophic bacterial symbionts. *Mar. Biol.* **113**, 99–106 (1992).
38. Dang, H., Luan, X., Zhao, J. & Li, J. Diverse and novel nifH and nifH-like gene sequences in the deep-sea methane seep sediments of the Okhotsk Sea. *Applied Environ. Microbiol.* **75**, 2238–2245 (2009).
39. Süss, J. et al. Widespread distribution and high abundance of *Rhizobium radiobacter* within Mediterranean subsurface sediments. *Environ. Microbiol.* **8**, 1753–1763 (2006).
40. Weiss, B. L., Wu, Y., Schwank, J. J., Tolwinski, N. S. & Aksoy, S. An insect symbiosis is influenced by bacterium-specific polymorphisms in outer-membrane protein A. *Proc. Natl. Acad. Sci. USA* **105**, 15088–15093 (2008).
41. Fjellbirkeland, A., Bemanian, V., McDonald, I. R., Murrell, J. C. & Jensen, H. B. Molecular analysis of an outer membrane protein, MopB, of *Methylococcus capsulatus* (Bath) and structural comparisons with proteins of the OmpA family. *Arch. Microbiol.* **173**, 346–351 (2000).
42. Kägi, J. H. & Vallee, B. L. Metallothionein: a cadmium and zinc-containing protein from Equine Renal Cortex II. *J Biol Chem* **236**, 2435–2442 (1961).
43. Viarengo, A. et al. Role of metallothioneins in Cu and Cd accumulation and elimination in the gill and digestive gland cells of *Mytilus galloprovincialis* lam. *Mar. Environ. Res.* **16**, 23–36 (1985).
44. Géret, F., Jouan, A., Turpin, V., Bebianno, M. J. & Cosson, R. P. Influence of metal exposure on metallothionein synthesis and lipid peroxidation in two bivalve mollusks: the oyster (*Crassostrea gigas*) and the mussel (*Mytilus edulis*). *Aquat. Living Resour.* **15**, 61–66 (2002).
45. Hamza-Chaffai, A., Amiard, J. C., Pellerin, J., Joux, L. & Berthet, B. The potential use of metallothionein in the clam *Ruditapes decussatus* as a biomarker of *in situ* metal exposure. *Comp. Biochem. Physiol. C* **127**, 185–197 (2000).
46. Rauser, W. E. Phytochelatins and related peptides: structure, biosynthesis, and function. *Plant Physiol.* **109**, 1141–49 (1995).
47. Maitani, T., Kubota, H., Sato, K. & Yamada, T. The composition of metals bound to class III metallothionein (phytochelatin and its desglycyl peptide) induced by various metals in root cultures of *Rubia tinctorum*. *Plant Physiol.* **110**, 1145–50 (1996).
48. Ha, S. B. et al. Phytochelatin synthase genes from *Arabidopsis* and the yeast *Schizosaccharomyces pombe*. *Plant Cell.* **11**, 1153–1163 (1999).
49. Vatamianiuk, O. K., Bucher, E. A., Ward, J. T. & Rea, P. A. A new pathway for heavy metal detoxification in animals—phytochelatin synthase is required for cadmium tolerance in *Caenorhabditis elegans*. *J. Biol. Chem.* **276**, 20817–20 (2001).
50. Zhang et al. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **490**, 49–54 (2012).
51. Arp, A. J. & Childress, J. J. Sulfide binding by the blood of the hydrothermal vent tube worm, *Riftia pachyptila*. *Science* **219**, 295–297. (1983).
52. Arp, A. J., Childress, J. J. & Flslser, C. R., Jr. Metabolic and blood gas transport characteristics of the hydrothermal vent bivalve *Calyptogena magnifica*. *Physiol. Zool.* **57**, 648–662. (1984).

53. Powell, M. A. & Somero, G. N. Adaptations to sulfide by hydrothermal vent animals: sites and mechanisms of detoxification and metabolism. *Biol. Bull.* **171**, 274–290 (1986).
54. VÖLkel, S. & Grieshaber, M. Sulphide oxidation and oxidative phosphorylation in the mitochondria of the lugworm. *J. Exp. Biol.* **200**, 83–92 (1997).
55. Brand, G. L. *et al.* Hypotaurine and thiotaurine as indicators of sulfide exposure in bivalves and vestimentiferans from hydrothermal vents and cold seeps. *Mar. Ecol.* **28**, 208–218 (2007).
56. Rosenberg, N. K., Lee, R. W. & Yancey, P. H. High contents of hypotaurine and thiotaurine in hydrothermal-vent gastropods without thiop trophic endosymbionts. *J. Exp. Zool. A.* **305**, 655–662 (2006).
57. Pruski, A. M. & Fiala-Médioni, A. Stimulatory effect of sulphide on thiotaurine synthesis in three hydrothermal-vent species from the East Pacific Rise. *J. Exp. Biol.* **206**, 2923–2930 (2003).
58. Koito, T. *et al.* Decline in taurine transporter mRNA and thioautotrophic bacterial 16S rDNA levels after transplantation of the hydrothermal-vent mussel *Bathymodiolus septendierum* to a non-vent position. *Cah. Biol. Mar.* **51**, 429–433 (2010).
59. Koito, T. *et al.* Effect of long-term exposure to sulfides on taurine transporter gene expression in the gill of the deep-sea mussel *Bathymodiolus platifrons*, which harbors a methanotrophic symbiont. *Fisheries Sci.* **76**, 381–388 (2010).
60. Patel, R. K. & Jain, M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS One.* **7**, e30619 (2012).
61. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
62. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics.* **22**, 1658–1659 (2006).
63. Loewenstein, Y., Portugal, E., Fromer, M. & Linial, M. Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space. *Bioinformatics.* **24**, i41–i49 (2008).
64. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
65. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
66. Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N. & Schuster, S. C. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* **21**, 1552–1560 (2011).
67. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics.* **30**, 1236–1240 (2014).
68. Kenny, N. J. *et al.* Genomic sequence and experimental tractability of a new decapod shrimp model, *Neocardina denticulata*. *Mar. Drugs.* **12**, 1419–1437 (2014).
69. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods.* **9**, 357–359 (2012).
70. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* **26**, 841–842 (2010).
71. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods.* **5**, 621–628 (2008).
72. Huttenhower, C. *et al.* Exploring the human genome with functional maps. *Genome Res.* **19**, 1093–1106 (2009).
73. Leung, P. T. *et al.* De novo transcriptome analysis of *Perna viridis* highlights tissue-specific patterns for environmental studies. *BMC Genomics.* **15**, 804 (2014).
74. Han, H. *et al.* Identification of differentially expressed genes in pancreatic cancer cells using cDNA microarray. *Cancer Res.* **62**, 2890–2896 (2002).
75. Yang, I. V. *et al.* Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol.* **3**, 1–0062 (2002).
76. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(2Delta Delta C(T)) Method. *Methods.* **25**, 402–408 (2001).
77. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729 (2013).
78. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–97 (2004).
79. Duckert, P., Brunak, S. & Blom, N. Prediction of proprotein convertase cleavage sites. *Protein Eng. Des. Sel.* **17**, 107–112 (2004).

## Acknowledgements

We thank Chinese Ocean Mineral Resources Research and Development Association (COMAR) and National Deep Sea Center of China for inviting J.-W. Qiu to participate in the cruise, the captain and crew of R/V Xiangyanghong 9, and the operation team of the Jiaolong for facilitating the sampling. This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (project number: XDB06010102), China Mineral Resources Research and Development Association (COMRRDA12SC01), the Research Grant Council of HKSAR government (662413) and Hong Kong Baptist University (project number: FRG2/13–14/009).

## Author Contributions

Y.H.W., J.S., J.-W.Q. and P.-Y.Q. conceived and designed the project. J.-W.Q. collected samples. Y.H.W. and J.S performed the bioinformatics analysis. Y.H.W., S.J., L.S.H. and L.G.C. wrote the paper. J.-W.Q. and P.-Y.Q. supervised the work.

## Additional Information

**Accession Codes:** The raw sequencing data has been submitted to the Short Read Archive (SRA) of NCBI under the accession number SRP035485.

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Wong, Y. H. *et al.* High-throughput transcriptome sequencing of the cold seep mussel *Bathymodiolus platifrons*. *Sci. Rep.* **5**, 16597; doi: 10.1038/srep16597 (2015).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>