

Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes

Jin Sun^{1,2}, Yu Zhang³, Ting Xu², Yang Zhang⁴, Huawei Mu², Yanjie Zhang², Yi Lan¹, Christopher J. Fields⁵, Jerome Ho Lam Hui⁶, Weipeng Zhang¹, Runsheng Li², Wenyan Nong⁶, Fiona Ka Man Cheung⁶, Jian-Wen Qiu^{2*} and Pei-Yuan Qian^{1,7*}

Hydrothermal vents and methane seeps are extreme deep-sea ecosystems that support dense populations of specialized macrobenthos such as mussels. But the lack of genome information hinders the understanding of the adaptation of these animals to such inhospitable environments. Here we report the genomes of a deep-sea vent/seep mussel (*Bathymodiolus platifrons*) and a shallow-water mussel (*Modiolus philippinarum*). Phylogenetic analysis shows that these mussel species diverged approximately 110.4 million years ago. Many gene families, especially those for stabilizing protein structures and removing toxic substances from cells, are highly expanded in *B. platifrons*, indicating adaptation to extreme environmental conditions. The innate immune system of *B. platifrons* is considerably more complex than that of other lophotrochozoan species, including *M. philippinarum*, with substantial expansion and high expression levels of gene families that are related to immune recognition, endocytosis and caspase-mediated apoptosis in the gill, revealing presumed genetic adaptation of the deep-sea mussel to the presence of its chemoautotrophic endosymbionts. A follow-up metaproteomic analysis of the gill of *B. platifrons* shows methanotrophy, assimilatory sulfate reduction and ammonia metabolic pathways in the symbionts, providing energy and nutrients, which allow the host to thrive. Our study of the genomic composition allowing symbiosis in extremophile molluscs gives wider insights into the mechanisms of symbiosis in other organisms such as deep-sea tubeworms and giant clams.

The environment of deep-sea hydrothermal vents and methane seeps is characterized by darkness, lack of photosynthesis-derived nutrients, high hydrostatic pressure, variable temperatures and high concentrations of heavy metals and other toxic substances^{1,2}. Despite this hostile environment, these ecosystems support dense populations of macrobenthos which, with the help of chemoautotrophic endosymbionts, are fuelled by simple reduced molecules such as methane and hydrogen sulfide^{1–3}. Studying vent and seep organisms may answer fundamental questions about their origin and adaptation to the extreme environments^{1,4}. Deep-sea mussels (Mytilidae, Bathymodiolinae) often dominate at hydrothermal vents and cold seeps around the world. Owing to their ecological importance and remarkable biological characteristics, including their ability to survive an extended period under atmospheric pressure⁵, they are a recognized model for studies of symbiosis⁶, immunity⁷, adaptation to abiotic stress⁸, ecotoxicology⁹ and biogeography¹⁰.

Results and discussion

We sequenced the genomes of both *B. platifrons* Hashimoto and Okutani, 1994 (a deep-sea mussel, Fig. 1a) and *M. philippinarum* (Hanley, 1843) (a shallow-water mussel, Fig. 1b) in the family Mytilidae using a whole-genome shotgun approach and compared their features (Supplementary Note 1). Both assembled genomes were highly repetitive (Supplementary Fig. 2). The genome of *B. platifrons*

(1.64 Gb) was smaller than that of *M. philippinarum* (2.38 Gb), mainly because it has fewer repeats, particularly the Helitron-transposable elements and unclassified repeats that together contributed a 707 Mb difference between the two genomes (Supplementary Note 9). The N_{50} value (the shortest sequence length at 50% of the genome size) of the *B. platifrons* scaffolds and contigs was 343.4 Kb and 13.2 Kb, respectively; and the N_{50} value of the *M. philippinarum* scaffolds and contigs was 100.2 Kb and 19.7 Kb, respectively. The deep-sea and shallow-water mussel genome contained 96.3% and 93.7% complete and partial universal single-copy metazoan orthologous genes, respectively, indicating the completeness of the assembly and gene models (Supplementary Note 6). As with other bilaterians, the two mussel genomes contained all the expected Hox and ParaHox genes, as well as similar numbers of microRNAs, again suggesting that the assemblies encapsulate a nearly complete representation of genic information (Supplementary Note 13 and 14). Although the gills of *B. platifrons* contain endosymbiotic methane-oxidizing bacteria (MOB)¹¹, the host scaffolds did not contain bacterial nucleotide sequences, showing a lack of horizontal gene transfer from the symbiont. *B. platifrons* has a lower rate of heterozygosity compared to the other three bivalves with a sequenced genome (Supplementary Fig. 3a), potentially owing to recurrent population bottlenecks as a result of population extinction and re-colonization¹².

The *B. platifrons* genome had 33,584 gene models, 86.2% of which had transcriptome support, based on transcriptome data from three

¹Division of Life Science, Hong Kong University of Science and Technology, Hong Kong, China. ²Department of Biology, Hong Kong Baptist University, Hong Kong, China. ³Shenzhen Key Laboratory of Marine Bioresource and Eco-Environmental Science, College of Life Science, Shenzhen University, Shenzhen, China. ⁴Key Laboratory of Tropical Marine Bio-resources and Ecology, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou 510301, China. ⁵High Performance Computing in Biology, Roy J. Carver Biotechnology Centre, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. ⁶Simon F. S. Li Marine Science Laboratory, School of Life Sciences, Centre for Soybean Research, Partner State Key Laboratory of Agrobiotechnology, the Chinese University of Hong Kong, Hong Kong, China. ⁷HKUST-CAS Joint Laboratory, Sanya Institute of Deep Sea Science and Engineering, Sanya 572000, China. *e-mail: qiujuw@hkbu.edu.hk; boqianpy@ust.hk

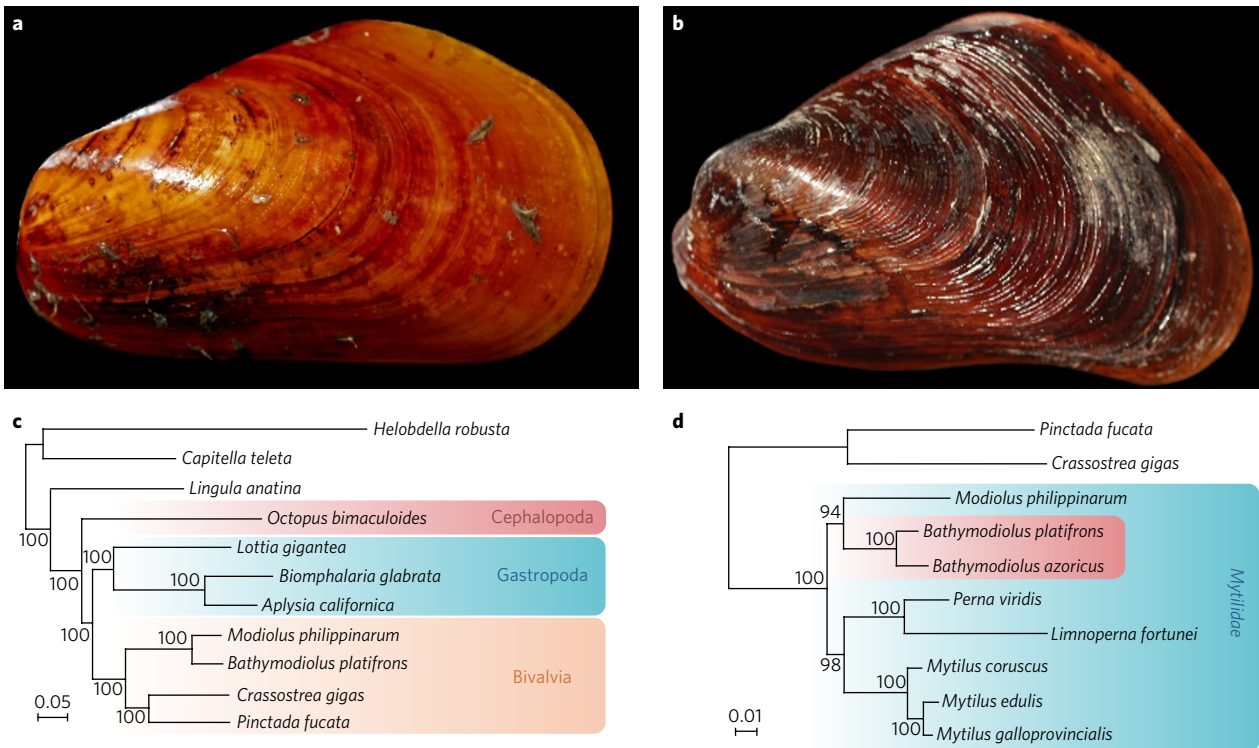


Figure 1 | Shell morphology and phylogenetic positions of *B. platifrons* and *M. philippinarum*. **a, b**, Shell of the deep-sea mussel *B. platifrons* (**a**, 10.7 cm) and shallow-water mussel *M. philippinarum* (**b**, 6.5 cm). **c**, Phylogenetic tree of 11 lophotrochozoan species based on genome data; the value of bootstrap support is shown at each node. The scale bar indicates 0.05 expected substitutions per site. **d**, Phylogenetic tree of Mytilidae based on transcriptome data from representative genera/species. The scale bar indicates 0.01 changes per site.

tissues in a previous study¹³, and our new transcriptome data from three additional tissues; whereas the *M. philippinarum* genome had 36,549 gene models, 86.0% of which had transcriptome support, based on our new transcriptome data from five tissues (Supplementary Note 5). A total of 5,527 *B. platifrons* gene models were validated by shotgun proteomics using gill samples (false discovery rate (FDR) < 0.01), and the gene-expression level correlated well ($P < 1 \times 10^{-5}$) with their corresponding protein abundance (Supplementary Note 20). Comparisons among five molluscan genomes revealed 6,883 shared protein domains and 474 protein domains unique to *B. platifrons* (Fig. 2a and Supplementary Table 14).

A phylogenetic tree constructed using 375 single-copy orthologues from 11 lophotrochozoan species (Fig. 1c) and another tree constructed using transcriptome sequences of representative members of Mytilidae (Fig. 1d) showed that, within the genera used, *Modiolus* is the closest shallow-water relative of *Bathymodiolus*. Using the lophotrochozoan tree as a reference, the time of divergence between *B. platifrons* and *M. philippinarum* was estimated to be around 110.4 million years ago (Ma), with a 95% confidence interval of 52.4–209.7 Ma (Fig. 2b), which is close to the upper age limit of deep-sea symbiotic mussels (102 Ma) previously estimated using five genes¹⁴. This result supports the hypothesis that the ancestors of modern deep-sea mussels might have experienced an extinction event during the global anoxia period that has been associated with the Palaeocene–Eocene thermal maximum at around 57 Ma¹⁵.

Gene-family analysis among the 11 lophotrochozoans revealed the expansion of 111 protein domains and contraction of 39 domains in *B. platifrons* (Fig. 2c and Supplementary Table 10). The number of contracted domains in the deep-sea mussel is the smallest among all the species examined. This indicated that adaptation of *B. platifrons* to the deep-sea chemosynthetic environment was mainly mediated by gene family/domain expansion, and the deep-sea mussel has retained most genes of its shallow-water ancestors.

Among the most numerous expanded gene families in *B. platifrons* are the HSP70 family (179 proteins) and the ABC transporters (393 proteins). HSP70 is important for protein folding, and thus high expression of these genes in the gill (Supplementary Fig. 15) could indicate their active response to stresses such as changing temperature, pH and hypoxia. However, because it took approximately three hours from the sample collection to preservation, the high expression of some of the stress-related genes could also be caused by the stress experienced during the sampling and transportation. Nevertheless, the expression level of HSP70 has been reported to positively correlate with the level of DNA-strand breakage in a deep-sea mussel¹⁶. Therefore, expansion of this gene family may provide additional genetic resources allowing deep-sea mussels to cope with abiotic stressors. ABC transporters are known to move toxic chemicals outside mussel gill epithelial cells¹⁷, forming the first line of defence against toxic chemicals in the vent/seep fluid. Many of the ABC transporters were highly expressed in the gill (Supplementary Fig. 16), indicating their active role in protecting the deep-sea mussel through detoxification.

Deep-sea mussel *Bathymodiolus* spp. obtain their symbionts from the environment, and house them in the vacuoles of gill epithelial cells¹⁸. Therefore, they must recognize the bacteria in the water column, incorporate them into the host cell, and control their population growth (Fig. 3). The target microbes carry signature molecules on their surface, which are recognized by the transmembrane pattern-recognition receptors (PRRs) of the host¹⁹. Among the PRRs in *B. platifrons*, there was a significant expansion of genes with an immunoglobulin (Ig) domain and genes with a fibrinogen domain, both of which are critical in allowing the sea anemone *Aiptasia* to recognize its dinoflagellate symbiont²⁰. In addition, peptidoglycan-recognition proteins (PGRPs) that can bind with peptidoglycan²¹ on the bacterial cell wall were also expanded in the genome of *B. platifrons*. Many of these PRR genes were also highly

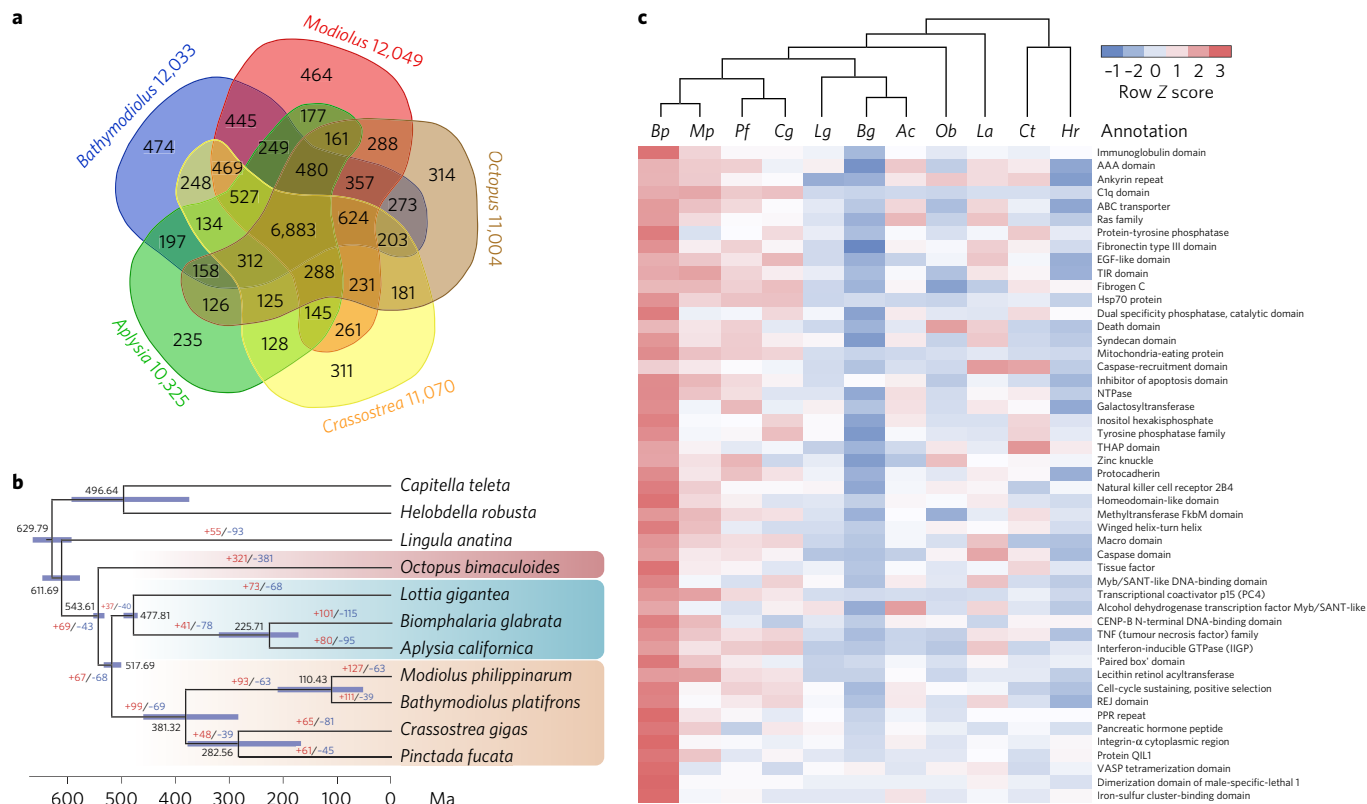


Figure 2 | Gene family representation analysis and divergence time of 11 lophotrochozoan species. **a**, Venn diagram showing the number of shared and unique Pfam functional domains in five molluscan species. **b**, Divergence times of 11 lophotrochozoan species with error bars (purple lines) indicating the 95% confidence interval, and the event of gene family expansions (red) and contractions (blue) analysed by counting Pfam domains. **c**, Heat map of major annotated Pfam domains that are expanded in *B. platifrons*, with multiple domains in a given gene being counted as one. *Ac*, *Aplysia californica*; *Bp*, *Bathymodiolus platifrons*; *Bg*, *Biomphalaria glabrata*; *Cg*, *Crassostrea gigas*; *Ct*, *Capitella teleta*; *Hr*, *Hellobdella robusta*; *La*, *Lingula anatina*; *Lg*, *Lottia gigantea*; *Mp*, *Modiolus philippinarum*; *Ob*, *Octopus bimaculoides*; *Pf*, *Pinctada fucata*.

expressed in the gill (Supplementary Note 18), consistent with the gill containing most of the MOB in the deep-sea mussels.

Endocytosis of exogenous bacteria is a critical step in establishing the host–symbiont relationship. Several observations showed molecular mechanisms of endocytosis of MOB by *B. platifrons*. (1) The Toll-like receptor 13 (TLR13) was found to be expanded in the genome of *B. platifrons*, and these receptors were highly expressed in the gill (Supplementary Fig. 20). TLR13 functions as an endosomal receptor in various groups of animals, which can sense microbes only after their internalization by phagocytosis²². (2) Two adhesion gene families (syndecan and protocadherin) were expanded in the genome of *B. platifrons*, and these families have been reported to mediate endocytosis in other species^{23,24} (Supplementary Figs 21 and 22). (3) Positive selection of six genes in the endocytosis pathway was shown by a branch-site evolutionary analysis (Supplementary Fig. 14 and Supplementary Note 17). Among the positively selected genes, a Wiskott–Aldrich syndrome protein and SCAR homologue (WASH) functions as a nucleation-promoting factor (NPF) to stimulate actin polymerization on endosomal membranes and consequently regulate vesicular trafficking²⁵.

Apoptosis plays multiple functions in host–symbiont interactions such as post-phagocytic winnowing in coral–dinoflagellate symbiosis²⁶ and symbiont recycling in insect–bacteria symbiosis²⁷. In *B. platifrons*, several families of the classic apoptosis system including TNF, DEATH, CARD and caspases were remarkably expanded and were highly expressed in the gill (Table 1 and Supplementary Note 18). As the central controllers of programmed cell death, caspases initiate the transduction of apoptotic signals by activation of members of the TNF superfamily and their DEATH receptor²⁸,

which can eventually result in the release of lysozymes to digest the symbionts²⁹. A TUNEL assay showed the presence of numerous apoptotic cells in deep-sea mussels³⁰ and in late-juvenile stages of vent tubeworms when they have acquired symbionts³¹, further substantiating the evolutionarily conserved role of apoptosis in establishing the host–symbiont relationship in various groups of invertebrates. In addition, as required to maintain homeostasis in symbiosis, *B. platifrons* has a complicated anti-apoptosis system with 130 inhibitors of apoptosis proteins (IAPs), which exceeds the number of IAPs in any other lophotrochozoan species available for comparison (Fig. 4). The high expression of many IAP genes is consistent with their role in maintaining the symbiont population in the gill of *B. platifrons* (Fig. 4b). Therefore, the expansion of the apoptosis- and anti-apoptosis-related gene families could contribute to the regulation of the symbiont population in *Bathymodiolus*.

To better understand the relationship between the host and its symbionts, 220 MOB proteins were identified through metaproteomics of the *B. platifrons* gill (Supplementary Note 20), leading to the discovery of three key metabolic pathways in the symbionts (Fig. 3). The methane-oxidation pathway, indicated by high expression of two signature enzymes of methanotrophy (that is, methane monooxygenase and methanol dehydrogenase)¹¹, fuels the host–symbiont relationship (Supplementary Fig. 28). The assimilatory sulfate reduction pathway, identified by several adenylyl-sulfate kinases, produces sulfur-containing amino acids³². The ammonia-assimilation pathway, identified by glutamate-ammonia ligase and glutamate synthase, detoxifies ammonia and provides the symbionts and host with the much needed organic nitrogen in the deep sea¹¹.

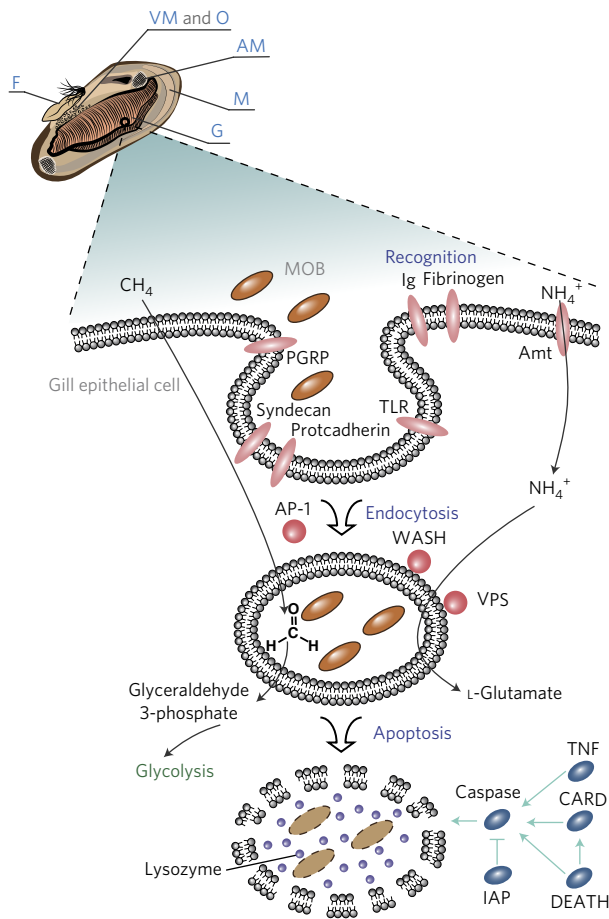


Figure 3 | A model of symbiosis between *B. platifrons* and methane oxidizing bacteria. A model of symbiosis between *B. platifrons* and MOB.

Top left, the diagram of the six tissues used for transcriptome sequencing, including the major MOB-containing gill. AM, adductor muscle; F, foot; G, gill; M, mantle; O, gonad; VM, visceral mass. Bottom, the host has mechanisms for MOB recognition, endocytosis and population control through apoptosis. The symbiont has various metabolic pathways (that is, methanotrophy and ammonia metabolism) that are beneficial to the host. AMT, ammonia transporter; AP-1, adaptor protein complex 1; CARD, caspase-recruitment-domain protein; IAP, inhibitor of apoptosis; Ig, immunoglobulin; LRP, low-density lipoprotein-receptor-related protein; PGRP, peptidoglycan recognition protein; TLR, Toll-like receptor; VPS, vacuolar protein-sorting protein; WASH, Wiskott-Aldrich syndrome protein and SCAR homologue.

Our study has provided genomic resources for understanding how the deep-sea mussel has adapted to extreme abiotic stresses and absence of photosynthesis-derived energy and nutrients in the chemosynthetic ecosystems. The general mechanisms of symbiosis revealed in our study are of relevance to other symbiotic organisms such as deep-sea tubeworms and clams, as well as shallow-water corals.

Methods

Mussel collection and DNA extraction. Individuals of *B. platifrons* were collected using the manned submersible *Jiaolong* from a methane seep (22° 06.921' N, 119° 17.131' E, 1,122 m deep; Supplementary Fig. 1) in the South China Sea in July 2013¹³. The mussels were kept in an enclosed sample chamber placed in the sample basket of the submersible, and it took approximately three hours from sampling to sample preservation. Once the samples were brought to the upper deck of the mothership, the adductor muscle of an individual was dissected, cut into small pieces and immediately fixed in RNAlater. The sample was then transported to Hong Kong Baptist University (HKBU) on dry ice and stored

at -80°C until use. DNA was extracted using the CTAB method, and its quality and quantity was checked with standard agarose gel electrophoresis and a Qubit Fluorometer, respectively.

The shallow-water horse mussels *M. philippinarum* were collected from a sandy shore at Ting Kok, Hong Kong (22° 46.989' N, 114° 21.389' E; Supplementary Fig. 1) during low tide in April 2014. The mussels were transported to HKBU and maintained in a seawater aquarium until use. The species identity was confirmed based on a morphological description³³. DNA from the adductor muscle was extracted using the DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany).

Genome sequencing and reads cleaning. For *B. platifrons*, nine libraries with insert sizes ranging from 180 bp to 16 kb were constructed using different preparation methods (see Supplementary Table 1a for details) and sequenced by Illumina HiSeq2000 and HiSeq2500 platforms. For *M. philippinarum*, six libraries with insert sizes ranging from 180 bp to 5 kb were constructed and sequenced by Illumina HiSeq2000 and HiSeq2500 platforms. Low-quality reads and sequencing-adaptor-contaminated reads were trimmed and error corrected. Reads of mate-pair libraries were de-duplicated and sorted.

Transcriptome sequencing. For *B. platifrons*, the mantle, foot and gill transcriptomes were generated from a single *B. platifrons* as previously reported¹³. The adductor muscle, visceral mass and ovary tissues from another *B. platifrons* collected from a hydrothermal vent at the Noho site in November 2015 were dissected on board and fixed in RNAlater, and their RNA was extracted using TRIzol and sent to BGI (Shenzhen) for sequencing using a eukaryotic mRNA enrichment and library preparation protocol. For *M. philippinarum*, RNA was extracted from the mantle, foot, gill, adductor muscle and visceral mass and sequenced using the same method as for *B. platifrons*. The RNA-sequencing (RNA-seq) data are summarized in Supplementary Table 5. Raw reads were trimmed and assembled using Trinity version 2.0.6³⁴.

Genome assembly. The genomes were assembled using Platanus version 1.21³⁵. Our genome-assembly pipeline involved assembling of DNA sequences using different parameters, merging results from different assemblies, removing redundant scaffolds, further scaffolding using transcripts, filling the gaps and correcting the errors (Supplementary Fig. 4).

Identification of repeats and transposable elements. Repeats and transposable elements were screened using the RepeatModeler pipeline version 1.0.8. RMBlast was used to search and construct a species-specific repeats library of the genome, and RepeatMasker³⁶ was used to identify, classify and mask the repeats with the species-specific repeats library and repeats in the RepBase with the default settings.

Genome annotation. Trinity was used to generate another version of transcripts using the genome-based mode. The transcriptome data from the *de novo* and genome-assisted approaches were further assembled using PASA pipeline version 2.0.2³⁷ with BLAT as the aligner. The two mussel genomes were annotated using MAKER version 3.0³⁸. MAKER was run twice, in order to train the gene models and increase the gene-predication accuracy. The gene models were further filtered by running OrthoMCL³⁹ with the protein models from six species of mollusks (*Crassostrea gigas*, *Pinctada fucata*, *Lottia gigantea*, *Aplysia californica*, *Biomphalaria glabrata*, and *Octopus bimaculoides*) and sequences that could be clustered with these molluscs were retained. Transcriptome reads were then mapped to the remaining 'orphan' genes, and only those supported by at least 10 parts per million (ppm) reads were retained. BLASTp was applied to search the predicted mussel protein models against the whole non-redundant database with an *E* value of 1×10^{-5} . The .xml format files were uploaded to BLAST2GO to obtain the GO and InterprotScan annotation. Pfam annotation was performed by searching the 16,295 Pfam entries (version 29.0) using a hidden Markov model (HMM)⁴⁰. The clean RNA-seq reads were quantified using kallisto⁴¹. Gene-expression levels were quantified based on transcripts per million (TPM).

Phylogenomics. Protein sequences of the nine published or public lophotrochozoan genomes (Supplementary Table 3) were downloaded and analysed together with data from the two mussels used in the phylogenomic analysis. OrthoMCL³⁹ was applied to determine and cluster gene families, and the BBH approach in BLASTp was used with a threshold value of 1×10^{-5} . This analysis resulted in 35,057 gene families among the 11 species. Only single-copy orthologues in each gene cluster were concatenated without partition information for the phylogenetic analysis using RaxML⁴² with the GTR + Γ model and slow bootstrapping.

Divergent time among the mollusks was estimated using MCMCTree, which applied the Bayesian estimation method for species divergence⁴³. The maximum-likelihood tree obtained using the 11 lophotrochozoan species was used to provide a reference tree topology, and the WAG + Γ protein substitution model was employed on each site. The tree was calibrated with the following time frames to constrain the age of the nodes between the species: minimum = 305.5 Ma and soft maximum = 581 Ma, for *C. teleta* and *H. robusta*⁴⁴; minimum = 168.6 Ma and soft

Table 1 | Distribution of protein Pfam domains associated with apoptosis among 11 lophotrochozoans.

Pfam	Domain	Bp	Mp	Pf	Cg	Lg	Bg	Ac	Ob	La	Ct	Hr
PF00656	Caspase	75*	42	45	38	18	17	21	38	71	42	15
PF00653	IAP repeat	130*	95	68	49	26	43	52	16	37	26	6
PF00452	Bcl-2	12	10	12	8	11	9	8	14	16	15	12
PF00020	TNFR_c6	37	40	27	28	31	5	18	12	49	28	12
PF00229	TNF	50*	33	25	29	11	4	9	8	24	22	5
PF05729	NACHT	245	204	352*	127	105	53	170	102	194	246	84
PF00931	NB-ARC	78	69	95	52	104*	27	55	47	88	77	37
PF00531	DEATH	155*	98	125	55	40	8	53	186*	126	37	23
PF01335	DED	26	18	63*	22	10	2	7	6	47	8	2
PF00619	CARD	133*	94	66	57	32	6	36	29	212*	197*	16
Total		941	703	878	465	388	174	429	458	864	698	212

Each number shows the number of proteins containing the specific domain. Protein family expansion is indicated by an asterisk, with a significant corrected *P* value ($P < 0.05$). In *B. platifrons*, five gene families are significantly expanded when compared to other species. In any of the other 10 species, two gene families are significantly expanded at the most. *Ac*, *Aplysia californica*; *Bg*, *Biomphalaria glabrata*; *Bp*, *B. platifrons*; *Cg*, *Crassostrea gigas*; *Ct*, *Capitella teleta*; *Hr*, *Hydrobia ulvae*; *La*, *Lingula anatine*; *Lg*, *Lottia gigantea*; *Mp*, *Modiolus philippinarum*; *Ob*, *Octopus bimaculoides*; *Pf*, *Pinctada fucata*. Domain abbreviations: IAP, inhibitor of apoptosis; Bcl-2, apoptosis regulator proteins, Bcl-2 family; TNFR, tumour necrosis factor receptor; TNF, tumour necrosis factor; NACHT, a domain that can be found in NAIP, CIITA, HET-E and TEPI proteins; NB-ARC, a nucleotide-binding adaptor shared by APAF-1, certain *R* gene products and CED-4; DED, death effector domain; and CARD, caspase recruitment domain.

maximum = 473.4 Ma, for *A. californica* and *B. glabrata*⁴⁴; minimum = 470.2 Ma and soft maximum = 531.5 Ma, for *A. californica* (or *B. glabrata*) and *L. gigantea*⁴⁴; minimum = 532 Ma and soft maximum = 549 Ma, for the first appearance of mollusks⁴⁵; and minimum = 550.25 Ma and soft maximum = 636.1 Ma, for the first appearance of Lophotrochozoa⁴⁵.

To understand the phylogenetic positions of these two mussels in the family Mytilidae better, transcriptome sequences from six additional mytilid species were downloaded from NCBI and re-assembled where appropriate, with details shown in the Supplementary Information. Protein sequences from each species were grouped into gene families using OrthoMCL³⁹ as described above. Only the one-to-one orthologues were aligned, trimmed and concatenated. Maximum-likelihood trees were built using RaxML⁴² with the GTR + Γ model and slow bootstrapping.

Gene-family evolution. We applied a sensitive HMM scanning method on known pfam functional protein domains to classify the gene families⁴⁶. Annotation of Pfam domains were performed on all 11 lophotrochozoan species after excluding protein domains of various transposable elements. A domain with multiple copies in a protein was counted only once. Thereafter, in order to identify gene-family expansion/contraction events in *B. platifrons*, a two-tailed Fisher's exact test was conducted. The Pfam domain counts in *B. platifrons* were compared against the background average domain counts of the 9 non-mussel mollusks (that is, species other than *B. platifrons* and *M. philippinarum*). In addition, OrthoMCL³⁹ was applied to cluster gene families and CAFÉ version 3.1 was applied to determine the significance of gene family expansion/contraction⁴⁷. For each gene family, CAFÉ generated a family-wide *P* value, with a significant *P* value indicating a possible gene-family expansion or contraction event. A branch-specific *P* value was also generated for each branch/node using the Viterbi method. A family-wide *P* value less than 0.01 and a branch/node Viterbi *P* value less than 0.001 was considered as a signature of gene family expansion/contraction for a specific gene family and specific species, respectively, as recommended for CAFÉ version 3.1 in ref. ⁴⁷.

Proteomics characterization. Gill samples from three individuals of *B. platifrons* fixed separately in RNAlater were used in proteomic analysis⁴⁸. In brief, the samples were homogenized in a lysis buffer containing 8 M urea and 40 mM HEPES (pH = 7.5), sonicated to break the cells up, centrifuged at 4 °C at 15,000g for 15 min. The supernatant that contained the proteins was collected, purified with a methanol/chloroform protein-precipitation method, and re-constituted in lysis buffer. Three samples were then pooled in equal amounts, reduced using dithiothreitol, alkylated using iodoacetamide, and digested using sequencing-grade Trypsin. The digested sample, containing approximately 1 mg peptides, was fractionated using a strong cation-exchange (SCX) column following the procedure described previously⁴⁸.

Each SCX fraction was re-dissolved in 0.1% formic acid and analysed on a LTQ-Orbitrap Elite mass spectrometer coupled to an Easy-nLC (Thermo Fisher, Bremen, Germany) equipped with an Acclaim PepMap RSLC C18 column (Thermo Fisher Scientific, Sunnyvale, California, USA).

The MS/MS data were uploaded into Mascot version 2.3.2 (Matrix Sciences, London, UK) to search against the deduced *B. platifrons* protein sequences, with their reversed sequences used to calculate the FDR. The searching parameters were as follows: peptide charge: +2, +3 and +4; fixed modification: carbamidomethyl (C); variable modification: oxidation (M) and deamidated (NQ); maximum missed

cleavage: 2; peptide mass tolerance: 4 ppm; and fragment mass tolerance: 0.05 Da. The FDR was dynamically controlled at 0.01.

Because the gills of *B. platifrons* contain MOB^{11,49} that belong to the family Methylococcaceae¹³, we also identified and quantified the bacterial proteins in the gill samples to gain insights into the symbiotic relationship between the bacteria and their host. Protein sequences from 24 published bacterial genomes in Methylococcaceae were downloaded and appended with the 1,402 bacterial that were identified in our previous meta-transcriptome analysis¹³. This created a database with 86,023 bacterial proteins. The database search criteria were identical to those described previously for the mussel proteins. The identified/matched protein sequences were further annotated by BLASTp against the KEGG database using an *E* value of 1×10^{-7} .

Determination of positively selected genes. Single-copy orthologues were selected among *B. platifrons*, *M. philippinarum*, *C. gigas* and *P. fucata* after running the OrthoMCL pipeline³⁹ as mentioned above. For each single-copy orthologue group, they were aligned, trimmed and concatenated. The phylogenetic tree was constructed using RaxML⁴² with the GTR + Γ model and rapid bootstrapping.

Signatures of positively selected genes along a specific branch can be detected by branch-site models implemented in the codeml module of the phylogenetic analysis by maximum likelihood (PAML) package version 4.8⁵⁰. The branch-site model requires a lineage designated as 'foreground' phylogeny (that is, *B. platifrons* in the former phylogenetic tree) and the rest of the lineages as 'background' phylogeny (that is, *M. philippinarum*, *C. gigas*, and *P. fucata* in the former phylogenetic tree). Alternative model (model = 2, NSsite = 2, fix_omega = 0) and null hypothesis (model = 2, NSsite = 2, fix_omega = 1 and omega = 1) likelihood values were tested using χ^2 analysis. Only genes that had Bayesian Empirical Bayes (BEB) sites > 90% and a corrected *P* value less than 0.1 were considered to have undergone positive selection. A KEGG-pathway enrichment was performed with a χ^2 test (if any of the components are over 5) or the Fisher's exact test (all other analyses).

Data availability. Raw reads, assembled genome sequences and annotation are accessible from NCBI under BioProject numbers PRJNA328542 and PRJNA328544, Sequence Read Archive accession numbers SRP078287 and SRP078294, and Whole Genome Shotgun project numbers MJUT00000000 and MJUU00000000. The genome assemblies and annotations are also available from the Dryad Digital Repository at <http://dx.doi.org/10.5061/dryad.h9942>.

Received 3 August 2016; accepted 16 February 2017; published 3 April 2017

References

1. Van Dover, C. L. *The Ecology of Deep-Sea Hydrothermal Vents* (Princeton Univ. Press, 2000).
2. Levin, L. A. Ecology of cold seep sediments: interactions of fauna with flow, chemistry and microbes. *Oceanogr. Mar. Biol.* **43**, 1–46 (2005).
3. Sibuet, M. & Olu, K. Biogeography, biodiversity and fluid dependence of deep-sea cold-seep communities at active and passive margins. *Deep Sea Res. II* **45**, 517–567 (1998).

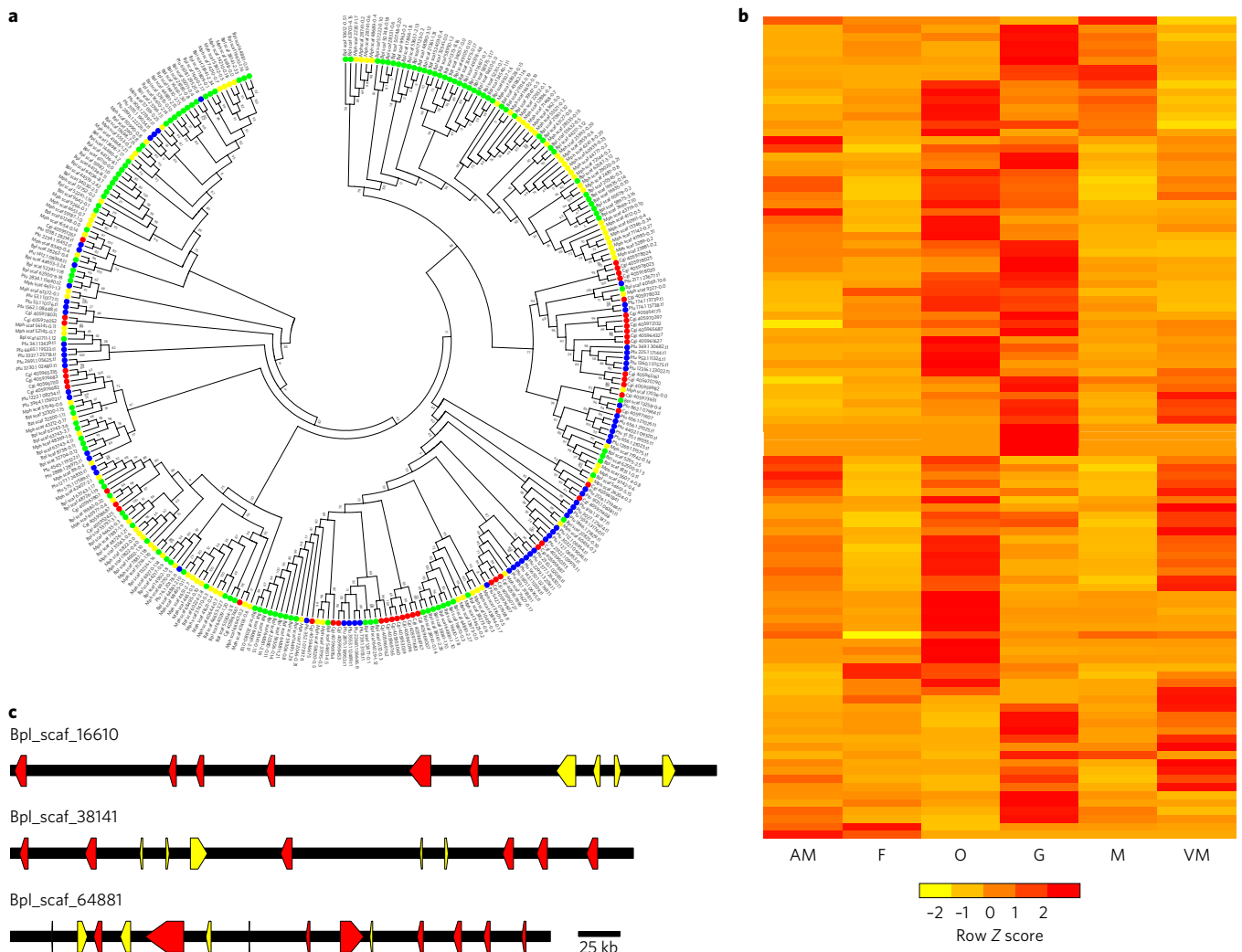


Figure 4 | Expansion of the IAP family in *B. platifrons*. **a**, Unrooted genealogy of IAP genes in *Bathymodiolus platifrons* (Bpl, green), *Crassostrea gigas* (Cgi, red), *Modiolus philippinarum* (Mph, yellow) and *Pinctada fucata* (Pfu, blue). **b**, Heat map of expression of IAP genes in six tissues. AM, adductor muscle; F, foot; G, gill; O, ovary; M, mantle; VM, visceral mass. **c**, Genomic arrangement of three IAP clusters. IPA genes are shown in red, whereas other genes are labelled in yellow. Scaffolds Bpl_scaf_16610 and Bpl_scaf_38141 each contain 6 IAP genes that are transcribed in the same direction, whereas scaffold Bpl_scaf_64881 contains 8 IAP genes that are transcribed in opposite directions.

- Little, C. T. S. & Vrijenhoek, R. C. Are hydrothermal vent animals living fossils? *Trends Ecol. Evol.* **18**, 582–588 (2003).
- Colaço, A. *et al.* LabHorta: a controlled aquarium system for monitoring physiological characteristics of the hydrothermal vent mussel *Bathymodiolus azoricus*. *ICES J. Mar. Sci.* **68**, 349–356 (2011).
- Dubilier, N., Bergin, C. & Lott, C. Symbiotic diversity in marine animals: the art of harnessing chemosynthesis. *Nat. Rev. Microbiol.* **6**, 725–740 (2008).
- 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).
- Boutet, I., Jollivet, D., Shillito, B., Moraga, D. & Tanguy, A. Molecular identification of differentially regulated genes in the hydrothermal-vent species *Bathymodiolus thermophilus* and *Paralvinella pandorae* in response to temperature. *BMC Genomics* **10**, 222 (2009).
- Bougerol, M., Boutet, I., LeGuen, D., Jollivet, D. & Tanguy, A. Transcriptomic response of the hydrothermal mussel *Bathymodiolus azoricus* in experimental exposure to heavy metals is modulated by the Pgm genotype and symbiont content. *Mar. Genomics* **21**, 63–73 (2015).
- Johnson, S., Won, Y.-J., Harvey, J. & Vrijenhoek, R. A hybrid zone between *Bathymodiolus* mussel lineages from eastern Pacific hydrothermal vents. *BMC Evol. Biol.* **13**, 21 (2013).
- DeChaine, E. G. & Cavanaugh, C. M. in *Molecular Basis of Symbiosis: Symbioses of Methanotrophs and Deep-Sea Mussels (Mytilidae: Bathymodiolinae)* (ed. Overmann, J.) 227–249 (Springer Berlin Heidelberg, 2006).
- Faure, B., Schaeffer, S. W. & Fisher, C. R. Species distribution and population connectivity of deep-sea mussels at hydrocarbon seeps in the Gulf of Mexico. *PLoS ONE* **10**, e0118460 (2015).
- Wong, Y. H. *et al.* High-throughput transcriptome sequencing of the cold seep mussel *Bathymodiolus platifrons*. *Sci. Rep.* **5**, 16597 (2015).
- Lorion, J. *et al.* Adaptive radiation of chemosymbiotic deep-sea mussels. *Proc. Biol. Sci.* **280**, 20131243 (2013).
- Olu-Le Roy, K., von Cosel, R., Hourdez, S., Carney, S. L. & Jollivet, D. Amphi-Atlantic cold-seep *Bathymodiolus* species complexes across the equatorial belt. *Deep Sea Res. I* **54**, 1890–1911 (2007).
- Pruski, A. M. & Dixon, D. R. Heat shock protein expression pattern (HSP70) in the hydrothermal vent mussel *Bathymodiolus azoricus*. *Mar. Environ. Res.* **64**, 209–224 (2007).
- Luckenbach, T. & Epel, D. ABCB- and ABCC-type transporters confer multitoxenobiotic resistance and form an environment-tissue barrier in bivalve gills. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **294**, R1919–R1929 (2008).
- Won, Y.-J. *et al.* Environmental acquisition of thiotrophic endosymbionts by deep-sea mussels of the genus *Bathymodiolus*. *Appl. Environ. Microbiol.* **69**, 6785–6792 (2003).
- Hoffmann, J. A., Kafatos, F. C., Janeway, C. A. & Ezekowitz, R. Phylogenetic perspectives in innate immunity. *Science* **284**, 1313–1318 (1999).
- Baumgarten, S. *et al.* The genome of *Aiptasia*, a sea anemone model for coral symbiosis. *Proc. Natl Acad. Sci. USA* **112**, 11893–11898 (2015).

21. Dziarski, R. & Gupta, D. The peptidoglycan recognition proteins (PGRPs). *Genome Biol.* **7**, 1–13 (2006).
22. Biondo, C. *et al.* The role of endosomal Toll-like receptors in bacterial recognition. *Eur. Rev. Med. Pharmacol. Sci.* **16**, 1506–1512 (2012).
23. Chen, K. & Williams, K. J. Molecular mediators for raft-dependent endocytosis of syndecan-1, a highly conserved, multifunctional receptor. *J. Biol. Chem.* **288**, 13988–13999 (2013).
24. de Beco, S., Gueudry, C., Amblard, F. & Coscoy, S. Endocytosis is required for E-cadherin redistribution at mature adherens junctions. *Proc. Natl Acad. Sci. USA* **106**, 7010–7015 (2009).
25. Derivery, E. *et al.* The Arp2/3 activator WASH controls the fission of endosomes through a large multiprotein complex. *Dev. Cell* **17**, 712–723 (2009).
26. Dunn, S. R. & Weis, V. M. Apoptosis as a post-phagocytic winnowing mechanism in a coral–dinoflagellate mutualism. *Environ. Microbiol.* **11**, 268–276 (2009).
27. Vigneron, A. *et al.* Insects recycle endosymbionts when the benefit is over. *Curr. Biol.* **24**, 2267–2273 (2014).
28. Romero, A., Novoa, B. & Figueras, A. The complexity of apoptotic cell death in mollusks: an update. *Fish Shellfish Immunol.* **46**, 79–87 (2015).
29. Detree, C. *et al.* Multiple I-type lysozymes in the hydrothermal vent mussel *Bathymodiulus azoricus* and their role in symbiotic plasticity. *PLoS ONE* **11**, e0148988 (2016).
30. Guezzi, H., *et al.* The potential implication of apoptosis in the control of chemosynthetic symbionts in *Bathymodiulus thermophilus*. *Fish Shellfish Immunol.* **34**, 1709 (2013).
31. Nussbaumer, A. D., Fisher, C. R. & Bright, M. Horizontal endosymbiont transmission in hydrothermal vent tubeworms. *Nature* **441**, 345–348 (2006).
32. Canfield, D. & Des Marais, D. Aerobic sulfate reduction in microbial mats. *Science* **251**, 1471–1473 (1991).
33. Lee, S. & Morton, B. in *The Malacofauna of Hong Kong and Southern China II: The Hong Kong Mytilidae* (eds Morton, B. & Dudgeon, D.) 49–76 (Hong Kong Univ. Press, 1985).
34. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
35. Kajitani, R. *et al.* Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**, 1384–1395 (2014).
36. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-3.0. (1996–2010).
37. Haas, B. J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
38. Cantarel, B. L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
39. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
40. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
41. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
42. Stamatakis, A., Ludwig, T. & Meier, H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21**, 456–463 (2005).
43. dos Reis, M. & Yang, Z. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Mol. Biol. Evol.* **28**, 2161–2172 (2011).
44. Benton, M. J., Donoghue, P. C. J. & Asher, R. J. in *The Timetree of Life: Calibrating and Constraining Molecular Clocks* (eds Hedges, S. B. & Kumar, S.) 35–86 (Oxford Univ. Press, 2009).
45. Benton, M. J. *et al.* Constraints on the timescale of animal evolutionary history. *Palaeontol. Electron.* **18**, 18.1.1FC (2015).
46. Albertin, C. B. *et al.* The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature* **524**, 220–224 (2015).
47. Han, M. V., Thomas, G. W. C., Lugo-Martinez, J. & Hahn, M. W. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecul. Biol. Evol.* **30**, 1987–1997 (2013).
48. Sun J. *et al.* First proteome of the egg perivitelline fluid of a freshwater gastropod with aerial oviposition. *J. Proteome Res.* **11**, 4240–4248 (2012).
49. Barry, J. P. *et al.* Methane-based symbiosis in a mussel, *Bathymodiulus platifrons*, from cold seeps in Sagami Bay, Japan. *Invertebr. Biol.* **121**, 47–54 (2002).
50. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

Acknowledgements

We thank the crew of research vessel (R/V) *Xiangyanghong 9* and operation team of the *Jiaolong*, and the crew of R/V *Kairei*, operation team of remotely operated vehicle (ROV) *Kaiko Mk-IV*, and on-board scientists during the cruise of KR15-17 for facilitating sampling of *B. platifrons*. We also thank M. Sun for her help with the figures and her continuous encouragement to the first author. This study was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (grant XDB06010102 to P.-Y.Q.), Hong Kong Baptist University (grant FRG2/14-15/002 to J.-W.Q.), and Scientific and Technical Innovation Council of Shenzhen and Guangdong Natural Science Foundation (grant 827000012, JCYJ20150625102622556, 2014A030310230 to Yu Z.). Genome sequencing was conducted by BGI (Shenzhen), Macrogen (Seoul), and the Roy J. Carver Biotechnology Center of University of Illinois Urbana–Champaign (UIUC). The mass spectrometry analysis was performed at the Instrumental Analysis & Research Center, Sun Yat-Sen University.

Author contributions

J.S., J.-W.Q. and P.-Y.Q. conceived and led the study. J.W.Q. and J.S. collected the samples. T.X. extracted the DNA and RNA, constructed the phylogenetic tree, and analysed the apoptosis-related gene families. J.S. and C.J.F. assembled the genome. J.S., C.J.F. and R.L. annotated the genomes. J.S. and Yang Z. analysed the immune-related gene families. Yanj Z. analysed the HSP70 gene family. J.H.L.H., W.N., and F.K.M.C. analysed microRNA and hox genes. Y.L. analysed the positively selected genes. H.M., Yu Z., and W.Z. analysed the proteome. J.S. performed the other bioinformatic analysis. All of the authors read, wrote and approved the manuscript.

Additional information

Supplementary information is available for this paper.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.-W.Q. and P.-Y.Q.

How to cite this article: Sun, J. *et al.* Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. *Nat. Ecol. Evol.* **1**, 0121 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Competing interests

The authors declare no competing financial interests.



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/>