# Genomic Insights Into Chemosynthetic Symbiosis in a Deep-Sea Hydrothermal Vent Mussel

**Kai Zhang**
The Hong Kong University of Science and Technology https://orcid.org/0000-0002-3225-5994
**Yao Xiao**
The Hong Kong University of Science and Technology
**Jin Sun**
Ocean University of China
**Ting Xu**
The Hong Kong University of Science and Technology
**Kun Zhou**
The Hong Kong University of Science and Technology
**Yick Hang Kwan**
The Hong Kong University of Science and Technology
**Jianwen Qiu** ( qiujw@hkbu.edu.hk )
Hong Kong Baptist University
**Pei-Yuan Qian**
The Hong Kong University of Science and Technology

**Research Article**

**Keywords:** Deep-sea, Mussel holobionts, Chemosynthetic symbiosis, Hydrothermal vents, Hologenome, Bacterial endosymbiont, Adaptation, Holobiont defense.

**Posted Date:** January 7th, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-1220069/v1

# Genomic insights into chemosynthetic symbiosis in a deep-sea hydrothermal vent mussel

Kai Zhang [1,2,†], Yao Xiao [1,2,†], Jin Sun [3], Ting Xu[1,2,4], Kun Zhou[1,2], Yick Hang Kwan [1,2], Jian-Wen Qiu [2,4*], Pei-Yuan Qian [1,2*]

[1] Department of Ocean Science and Hong Kong Branch of the Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), The Hong Kong University of Science and Technology, Hong Kong, China;

[2] Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Guangzhou 511458, China;

[3] Institute of Evolution & Marine Biodiversity, Ocean University of China, Qingdao, 266003, China

[4] Department of Biology, Hong Kong Baptist University, Hong Kong, China

***Correspondence**: qiujw@hkbu.edu.cn; boqianpy@ust.hk

†These authors contributed equally to this work.

**Abstract**

**Background:** Symbiosis with chemosynthetic bacteria has allowed many invertebrates to flourish in 'extreme' deep-sea chemosynthesis-based ecosystems, such as hydrothermal vents and cold seeps. Bathymodioline mussels are considered as models of deep-sea animal-bacteria symbiosis, but the diversity of molecular mechanisms governing host-symbiont interactions remains understudied owing to the lack of hologenomes. In this study, we adopted a total hologenome approach in sequencing the hydrothermal vent mussel *Bathymodiolus marisindicus* and the endosymbiont genomes combined with a transcriptomic and proteomic approach that explore the mechanisms of symbiosis.

**Results:** Here, we provide the first coupled mussel-endosymbiont genome assembly. Comparative genome analysis revealed that both *Bathymodiolus marisindicus* and its endosymbiont reshape their genomes through the expansion of gene families, likely due to chemosymbiotic adaptation. Functional differentiation of host immune-related genes and attributes of symbiont self-protection that likely facilitate the establishment of endosymbiosis. Hologenomic analyses offer new evidence that metabolic complementarity between the host and endosymbionts enables the host to compensate for its inability to synthesize some essential nutrients, and two pathways (digestion of symbionts and molecular leakage of symbionts) that can supply the host with symbiont-derived nutrients. Results also showed that bacteriocin and abundant toxins of symbiont may contribute to the defense of the *B. marisindicus* holobiont. Moreover, an exceptionally large number of anti-virus systems were identified in the *B. marisindicus* symbiont, which likely work synergistically to efficiently protect their hosts from phage infection, indicating virus-bacteria interactions in intracellular environments of a deep-sea vent mussel.

**Conclusions:** Our study provides novel insights into the mechanisms of symbiosis enabling deep-sea mussels to successfully colonize the special hydrothermal vent habitats.

**Key words*:*** Deep-sea**,** Mussel holobionts, Chemosynthetic symbiosis, Hydrothermal vents, Hologenome, Bacterial endosymbiont, Adaptation, Holobiont defense.

## Background

Symbioses with microorganisms are widespread in eukaryotic organisms, which have shaped the ecology and evolution of both the hosts and symbionts [1, 2]. Such a mutual beneficial relationship has enabled a variety of eukaryotic organisms to colonize some "extreme" habitats on Earth, such as hydrothermal vents and cold seeps in the deep oceans. In deep-sea vent and seep ecosystems, many macrobenthos, such as clams, tubeworms, snails, and mussels, harbor endosymbiotic bacteria within their specialized host cells, termed bacteriocytes, and access energy and nutrients through the oxidation of reducing substances, including methane, hydrogen sulfide, thiosulfate, and hydrogen[3]. Genomic tools are used in exploring the molecular mechanisms of such chemosynthetic symbioses in deep-sea animals[4, 5], but the lack of high-quality hologenomes limited the resolution in most of these studies. Compared with the sequences of the symbiont genomes of deep-sea animals, currently available sequences of host genomes are fewer (e.g., the tubeworms *Lamellibrachia luymesi* [6] and *Paraescarpia echinospica* [7], the clam *Archivesica marissinica* [8], and the snails *Gigantopelta aegis* [9] and *Chrysomallon squamiferum* [10].

Deep-sea mussels (Mytilidae, Bathymodiolinae) that host chemosynthetic bacterial symbionts have been found to flourish in diverse marine habitats including vent fields, seep areas, whale carcasses, and sunken wood [11, 12]. Apart from habitat diversity, deep-sea mussels are known to associate with an exceptional range of symbiotic types, such as intracellular and extracellular symbionts, and they can host methanotrophic or sulfur-oxidizing symbionts or both. Deep-sea mussels often form dense populations, which can serve as an important habitat for many other animals, such as polychaetas, snails, and limpets [13]. Owing to their remarkable ecological and biological features, deep-sea mussels are regarded as a feasible holobiont model for studying adaptation

and symbiosis [14]. Therefore, the complete genomes of a deep-sea mussel and its symbiont will enable studies that seek to understand the chemosynthetic symbiosis. Among bathymodulines, only the genome of the cold-seep mussel *Gigantidas plantifrons* has been sequenced; however, the assembly is quite fragmented, with a contig N50 value of 13.2 kb only, because a second generation sequencing technology alone was used [15]. Furthermore, since this mussel harbors methane-oxidizing bacteria (MOBs), and thus some of the mechanisms of host-symbiont association discovered in that study may not be applicable to deep-sea mussels harboring sulfur-oxidizing bacteria (SOBs), or both MOBs and SOBs [16].

Although eukaryote-controlled mechanisms are critical for host protection and this has been a focus of previous research, there is a growing understanding that symbiotic micro-organisms may also play a role in defense against natural enemies [17]. Diverse symbiotic bacterial species that protect insects against parasites, parasitoids, predators, and pathogens have been found [18]. As for deep-sea mussels, earlier studies showed that multiple strains of bacteria can coexist in the bacteriocytes of *Bathymodiolus* mussels gills[19, 20], but no bacteria have ever been found in the nuclei of these host cells [21]. These findings suggested that the symbionts can protect their mussel host cells from infection [22]. However, information on the roles of endosymbiotic bacteria in defending the deep-sea animal holobionts is scarce. Moreover, the endosymbiotic bacteria are likely to be protected against phage infection because the intracellular environment is isolated from the external environment. Phages, nevertheless, have been discovered in multiple arthropods endosymbiotic systems such as the flour moths [23] , mosquitos [24], and wasps [25]. Antivirus-related genes (i.e., *Cas* genes) are present in endosymbiont genomes of deep-sea mussels [4] and tubeworms [5], indicating the potential interactions between viruses and symbiotic bacteria. Interestingly, a recent study revealed that phage-bacteria interplay was likely present in deep-sea vent snail holobionts, which might contribute to regulate the population size of endosymbiotic bacteria [26]. To sum up, the diversity of molecular mechanisms governing host-

symbiont interactions is still understudied.

The deep-sea mussel *Bathymodiolins marisindicus* is a dominant epifaunal species in the hydrothermal vent fields in the Indian Ocean (Fig. 1A). This mussel harbors SOBs in its gills [27]. The goal of this study was to extend the knowledge of chemosynthetic endosymbiosis in deep-sea mussels. We generated a high-quality hologenome of *B. marisindicus*, and produced transcriptomic and proteomic data for the quantitative analyses of gene and protein expression that are pertinent to the symbiosis of this holobiont.

## Results and discussion

### Characteristics of the hologenome

Using PacBio long-read sequencing (~110-fold coverage), and Illumina short-read sequencing (~200-fold coverage), we de novo assembled the genome of *B. marisindicus*. This genome assembly was ~1.04Gb in total length, with a contig N50 of 301.96 kb. BUSCO assessment showed that 96.6% (92.6% complete and 4% fragmented) of the conserved metazoan genes were represented in the assembly (Table S1), indicating that the genome is of high completeness compared with the other sequenced lophotrochozoans. The genome of *B. marisindicus* is smaller than that of the cold-seep mussel *G. plantifrons* (~1.64 Gb) mainly due to its fewer repeats [15]. In addition, 27,190 protein-coding genes (PCGs; Table S2) were annotated in the *B. marisindicus* genome, of which 784 genes were highly or exclusively expressed in the symbiont-hosting gill (Table S3), indicating a symbiosis-specific function. A phylogenetic tree reconstructed from 388 shared single-copy genes in 19 lophotrochozoan species revealed the divergence between *B. marisindicus* and *G. platifrons* approximately 34.1 million years ago (Ma; Fig. 1B and Fig. S3), which is consistent with the result of our previous estimation based on the entire mitogenomes [28]. Gene family analyses revealed the expansion of 17 pfam domains in *B. marisindicus* (Fig. 1C), and many of them are likely involved in the chemosynthetic symbiosis (see below). The assembled genome of the SOB symbiont was 2.1 Mb in length and encodes 2,164 genes (Table

S7). CheckM analysis showed that the symbiont genome has a high completeness of 97.88% and low potential contamination (3.98%) (Table S8). Gene family analysis showed that 17 pfam domains have undergone expansion in the *B. marisindicus* symbiont compared with the SOBs of other *Bathymodiolus* mussels (Fig. 1D), and many of these expanded domains are related to the chemosynthetic symbiosis (see below). Metaproteomic analyses of the gill tissues revealed 6,379 host proteins (Table S9) and 1,020 SOB proteins (Table S10), providing additional protein-based evidence for tracing the metabolism of the *B. marisindicus* holobiont.

Transposable elements (TEs) may influence the function of gene. It is helpful in obtaining novel genetic material and disseminating regulatory elements, which induce the formation of stress-inducible regulatory networks [29]. We found bursts of TE insertion activities in the bathymodiolin mussel lineage approximately 160 Ma (Fig. S4), which was close to the upper age limit of chemoautotrophic symbiont-hosting Bathymodiolinae mussels (160.2 Ma) estimated herein (Fig. S3). Moreover, multiple pfam protein domains involved in gene fusion were expanded in the *B. marisindicus* genome (Fig. 1C) including DNA transposases (such as DDE superfamily endonuclease), and retrotransposons (such as reverse transcriptase, RNA-dependent DNA polymerase) [8]. Interestingly, the TEs have also been expanded in the SOBs (Fig. 1D). The numerous transposase genes may have facilitated the SOBs to acquire "foreign" DNA (i.e., toxin-related genes) and obtain new functions (i.e., new metabolic properties, detoxification, pathogenicity, virulence, and colonization of host intracellular environment). These results indicated that the enrichment of TEs might have enabled the *B. marisindicus* holobiont to acquire beneficial genetic materials and thereby adapt to an endosymbiotic lifestyle. TEs were also expanded in other deep-sea endosymbiotic animals, such as the clam *A. marissinica* [8] and the snail *G. aegis* [9], indicating that the expansion of TEs might be a convergent feature in deep-sea animals that host chemoautotrophic symbionts.

In eukaryotes, the creation of pseudogenes (i.e., nonfunctional DNA sequences that mimic functional genes) can be induced by transposable elements undergoing insertion or retrotransposition in the coding region [30]. Although often presumed to lack function, pseudogenes may play important biological roles [30], particularly in the regulation of symbiosis [8]. In the present study, 6,026 pseudogenes were identified in the *B. marisindicus* genome, significantly lower than those reported in *A. marissinica* (10,211), although the latter has a somewhat higher number of PCGs (28,949) [8]. This larger number of pseudogenes in *A. marissinica* may be related to the expansion of TEs [8]. Comparative transcriptome analysis showed that 411 of the *B. marisindicus* pseudogenes exhibited higher expression in the gill than in other tissues (Table S11 and S12). A functional classification showed that these highly expressed pseudogenes were enriched in 20 COG functional categories (Fig. S7), and many of them are associated with the host's defense, genomic DNA integrity, nutrient production, metabolism, and transport (Fig. S7), showing potential involvement in the regulation of gene functions [30] in this endosymbiont-hosting organ.

**Genetic regulation related to the establishment of symbiosis**

Animal hosts need to remodel their immune system to accommodate their obligate endosymbiont [9]. In the present study, comparative genomic analyses showed that many immunity-related gene families were expanded in *B. marisindicus,* including Leucine-rich repeat (LRR), C1q domain (C1qD) and immunoglobulin domain (Ig; Fig. 1C), indicating their potential roles in host-symbiont interactions. LRR exhibits high binding affinity with lipopolysaccharides (LPSs), and this binding plays crucial role in the recognition of symbiotic bacteria [6]. Remarkably, the Ig family and C1qD were expanded in the cold-seep mussel *G. platifrons* [15]. This finding highlights their importance to deep-sea mussel symbiosis. These gene families are essential because they enable hosts to recognize the surface patterns of various symbiotes with high specificity [31].

The immune recognition of symbionts mediated by pattern recognition receptors (PRRs)

could be the first step of symbiosis [32]. Our data showed that some PRRs were enriched or exclusively expressed in the gill in contrast to those in other organs (Fig. S8 and Table S3), including LPS, peptidoglycan recognition proteins (PGRPs), toll-like receptors (TLRs), LRRs, fibrinogen-related proteins (FBGs), galectin (Gal) protein, and C1q proteins, highlighting their importance in the establishment and maintenance of symbiosis. We found that multiple PRRs (LPS, PGRP, TLR, Gal and LRR) possibly play important roles in symbiont recognition in deep-sea animals, such as *L. luymesi* [6], *P. echinospica* [5], *A. marissinica* [8], *B. azoricus* [33], *G. platifrons*[15, 32], and *C. squamiferum* [9]. Interestingly, our results showed that the expression levels of many PRRs, such as Gals, PGRPS, TLRs, intergrin and C1q proteins, were down-regulated in the gill compared with those in other tissues (Fig. S8). The gene expression profiles indicated that these immunity-related genes potentially have different functional roles apart from the establishment of symbiosis. Symbionts in some invertebrates [8], including deep-sea mussels [33] suppress host immune response. Therefore, down-regulation of these immune genes suggested that endosymbionts may evolved mechanisms that do not activate host immune system. Collectively, our data implied that the up-regulated PRRs are involved in symbiont recruitment and the down-regulated PRRs are related to pathogen invasion. These findings also highlighted that *B. marisindicus* may restructure its innate immunity and thereby acquire endosymbionts and tolerant pathogens.

The endocytosis of exogenous bacteria is also a crucial step in the establishment of the host-symbiont relationship. In *G. platifrons*, multiple gene families associated with endocytosis, such as TLR13, syndecan, and protocadherin are expanded [15]. By contrast, *G. platifrons* gene families responsible for endocytosis were not expanded in *B. marisindicus*. Nevertheless, some genes that function in endocytosis including TLR13, vacuolar sorting proteins (VSP), low-density lipoprotein receptor-related proteins (LDLRs), and protocadherin, were notably more highly or even exclusively expressed in gill (Fig. S8). TLR13 and LDLRs function as an endosomal receptors in

various groups of animals, and can recognize bacteria [15], and VSP and protocadherin are involved in endocytosis regulation in other species [34]. Syndecan, adaptor protein complexes (APs), and Wiskott–Aldrich syndrome protein and SCAR homologue (WASH) were detected in the host proteome (Tables S2 and S9), mediating endocytosis and vesicular trafficking [35]. These results showed the diverse mechanisms of endosymbiont endocytosis by deep-sea mussels.

A recent study showed that the endosymbionts of the deep-sea snail *G. aegis* possess pathogen-associated molecular patterns (PAMPs), such as peptidoglycan associated lipoprotein (Pal), porins, OmpA family proteins, and OmpH family proteins, which facilitate the invasion of hosts [9]. In the SOBs of *B. marisindicus*, some PAMPs, including multifunctional autoprocessing RTX toxins (MARTX; Table 1), LRRs, fucolectins (FUCLs), OmpA family proteins and cadherins, were among the highly expressed proteins (Table S14), possibly enabling SOBs recognize and invade their mussel hosts. MARTX mediated host recognition and specificity in deep-sea mussels [36]. Cadherin was expanded in these SOBs compared with other *Bathymodiolus* symbionts (Fig. 1D). In sponge symbioses, symbiont LRRs and cadherins play an essential roles in host recognition [37]. PAMPs on the surfaces of symbiotic bacteria likely bind to the mussel host PRRs that induce symbiont recognition (Fig. 5). For instance, the OmpA family proteins can interact with host TLR and enable a symbiont to invade a host's intracellular environment [38], and FUCLs can serve as immune-recognition molecules that bind directly to a mussel's cell surface glycans [39].

To survive in the hosts' intracellular environments, endosymbionts must possess mechanisms for resisting the hosts' defenses. Symbiont bacteria might specifically mimic hosts' immune functions for the purpose of immune evasion [40]. In the present study, we found that the proteins of the immune-recognition gene FUCLs (pfam09603) are abundant in the hosts and endosymbionts (Tables S9 and S10), suggesting that this SOB mimics host immune function and thereby avoid its immune recognition.

Furthermore, the secretion system (SS) is essential for bacteria to survive inside host cells [41], as many symbionts (e.g., siboglinid symbionts, and rhizobia) use the SS to evade phagocytosis and facilitate infection [42]. In the *B. marisindicus* SOBs, we found the protein products of components of the Type II secretion system (T2SS) (Fig. 4A and 4B). In some pathogens, T2SS possesses a dual function of virulence and mediating environmental survival. For example, it can facilitate the intracellular replication of bacteria and secrete substrates (e.g., peptidases, lipases, nucleases etc.) to exploit nutrient and energy sources. Remarkably, the peptidases M4 and M48, which are involved in the degradation of the structural barriers of hosts by pathogens [5], were significantly expanded in this SOB (Fig. 1D). In addition, various proteolytic enzymes, such as the peptidases S41, S11, S26A, M41, M16, M23, and M20, were identified in the symbiont proteome (Table S10). These proteases were secreted through T2SS, which possibly facilitates the bacterial digestion and use of host nutrients. Additionally, a bacterial surface antigen of the SOB was highly expressed at the protein level (27th most abundant in proteome; Table S10). This protein plays a role in bacterial interactions with the environment, such as evasion of host defenses and induction of toxicity [43]. These symbiont attributes likely contributed to the establishment of an endosymbiotic lifestyle.

**The source and acquisition of nutrients in the host**

In a reduced digestive system, bathymodiolines obtain most nutrition from their gill endosymbionts to meet their metabolic requirements [44]. Our genomic analysis indicated that *B. marisindicus* might not synthesize many amino acids or digest phytoplankton-derived organic particles. The *B. marisindicus* genome included only 61 genes related to amino acid biosynthesis (Table S15). The absence of these amino acid biosynthesis genes indicates that either (a) it was lost over evolution, or (b) *B. marisindicus* genome was not complete. The high quality of the *B. marisindicus* genome assembly (96.6% completeness) as well as the dispersed distribution of these genes in the genome indicated that absence of these genes in amino acid biosynthesis

was not due to sequencing bias or assembly error. The SOB genome encodes an essentially complete gene set (110 genes) for the biosynthesis of all 20 essential proteinogenic amino acids and 11 vitamins or cofactors (Fig. 2 and Table S15). Genes required in the biosynthesis of 13 amino acids were discovered in the SOB genome but not in *B. marisindicus* (Fig. 2), indicating that *B. marisindicus* relies on its endosymbionts to compensate for this nutritional deficiency (see below). Moreover, a number of glycosyl hydrolase families (GHF) that can catalyze the hydrolysis of complex polysaccharides particularly cellulose [15], including GHF5, GHF15, and GHF27, are missing in the *B. marisindicus* genome. By contrast, an average of 5.2, 5, and 3.6 genes in these three families, respectively, are present in other bivalve genomes (Fig. S7 and Table S16). None of these GHF families is contracted in *G. platifrons* which lives in shallower waters (~642 m to 1,684 m); although heavily depending on its endosymbionts for nutrition, it retains the capacity to digest organic particles originally produced in surface water [15]. Notably, *B. marisindicus* lives in a much deeper water depth (2,757 m) and thus has access to less sinking biomass for filter-feeding in contrast to *G. platifrons*. These results indicated that the contraction of GHF families and the absence of multiple key amino acid synthesis-related genes in the *B. marisindicus* genome are adaptation to a greater dependence on its symbionts for nutrition.

The direct digestion of symbionts is an important nutrient acquisition strategy for deep-sea bathymodiolines, and lysozymes are responsible for symbiont digestion [4]. However, owing to the reduced peptidoglycan biosynthesis pathways, lysozymes in molluscs are suggested to be used in defense against pathogens, instead of symbiont digestion, and a host may utilize other mechanisms to obtain their nutrient from the symbionts [45], including the use of cathepsins for symbiont digestion [8]. Our transcriptome analyses showed that multiple cathepsins (cathepsin A,B,C,D,F,L,X,C1A) and one lysozyme were more highly or exclusively expressed in the gills of *B. marisindicus* (Fig. 3c, Tables S3 and S17) , and the corresponding proteins were found

in our proteomic analysis. Our speculation that cathepsins and lysozyme are used in digesting the symbionts in *B. marisindicus* is consistent with the findings of a previous study on *Bathymodiolus azoricus* [4]. Moreover, our results revealed that "milking" (translocation of nutrients) can be an overlooked strategy used by bathymodiolines to gain nutrients from the endosymbionts. All the genes associated with T2SS, the general secretory (Sec) pathway and Twin-arginine translocation (Tat) pathway were identified in the *B. marisindicus* symbiont genome (Fig. 3A and Table S18). Many of these genes were evidently expressed in the gills at the protein level (Fig. 3B). Nutrients, such as sugars and amino acids, can be transported across the inner membrane and into the periplasm of an endosymbionts via the Sec or the Tat pathway, and then they are secreted out of the cell through T2SS. In addition, the host solute carrier (SLC) family (437 genes; Table S2), which can rely on ion gradients to transport small molecular metabolites across the cell membrane, was significantly expanded in the *B. marisindicus* genome. Remarkably, this gene family is also expanded in deep-sea clam *A. marissinica* with 180 genes, suggesting its involved in "milking" nutrients from the symbionts [8]. Moreover, SLCs that transport small molecular metabolites (i.e., glucose, folate, glutamate and neutral amino acid), were all expressed at the protein level (Fig. S10 and Table S19). The above results indicated the involvement of both direct digestion and "milking" as a host's strategies to obtain nutrients from the endosymbionts (Fig. 5).

**Metabolic support between host and endosymbiont**

Considering the host's dependency on endosymbionts, we described the key metabolic pathways of the symbionts and their interactions with hosts. The habitat of *B. marisindicus* contains reduced components (hydrogen and sulfur compounds) that can provide a steady energy source to the holobionts [9]. Many deep-sea animals, such as the tubeworm *L. luymesi* [6] and clam *A. marissinica* [8], rely on hemoglobins for both oxygen and hydrogen sulfide transport. In *Bathymodiolus* mussels, no dedicated host proteins for sulfide transport have been identified. Instead, two host cytoglobin genes,

which function in the transportation of sulfide and oxygen [46], exhibited higher expression levels in the gills than in other tissues herein (Tables S3 and S20), implying their involvement in oxygen and hydrogen sulfide transport in *B. marisindicus.*

The genes involved for all of the key metabolic pathways for energy generation were found in the *B. marisindicus* symbiont genome (Fig. 4A). The abundance of symbiont SOB proteins involved in central metabolism are summarized in Fig. 4B and Table S21. Similar to the SOBs of *B. azoricus* [4], proteins required for sulfur oxidation in the *B. marisindicus* SOBs, including Sox genes, *dsrAB*, *sqr*, *aprAB,* and *sat*, were abundant in the gills (Fig. 4B), demonstrating their active involvement in detoxifying sulfide for the holobiont and in oxidizing sulfide, thiosulfate, and sulfite in the gills. Moreover, the identification of *hupL* and *hups* in *B. marisindicus* symbiont genome indicated that the SOBs of *B. marisindicus* can utilize hydrogen for energy production. Proteins related to hydrogen oxidation and sulfur oxidation were both highly expressed (Fig. 4B), indicating hydrogen oxidation is as important as thiosulfate oxidation in the SOB of *B. marisindicus*. This result was inconsistent with the findings in the SOB of *B. azoricus*, in which the energy-generating thiosulfate oxidation process is more prominent than the hydrogen oxidation process [4]. In contrast to *B. azoricus*, which also hosts MOB [4], *B. marisindicus* holobiont cannot use methane as an energy source, indicating *B. marisindicus* holobiont maybe more dependent on hydrogen oxidation than the *B. azoricus* holobiont.

The *B. marisindicus* symbiont genome lacks the key reverse TCA (rTCA) cycle genes (*oor*, *por*, and *acl*) but encode all the genes necessary for carbon fixation via the Calvin-Benson-Bassham (CBB) cycle (Fig. 4A). These results supported the hypothesis that deep-sea mussel symbionts switched from the rTCA cycle to a fully functional CBB cycle during evolution [47]. In the proteomes of this SOB, the CBB pathways proteins were abundant, indicating their significance in energy production. Similar to the thiotrophic *B. azoricus* symbiont, the SOB relies on the CBB cycle to fix $CO_2$ by

utilizing a form I ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO) (*cbbL* and *cbbS*). However, the SOBs of *G. aegis* from the same vent field use form II RuBisCO (*cbbM*) for carbon fixation[9], showing the diversity of carbon fixation machineries in different animal symbionts. In addition, the SOB genome also encodes a complete set of TCA cycle genes, and clear expressional evidence of these genes has been found at the protein level (Fig. 4A). This result was different from the result in vent mussel *B. azoricus* SOBs, which seem to be unable to replenish the essential carbon metabolism intermediates oxaloacetate and succinate because of the lack of several key TCA cycle enzymes, including 2-oxoglutarate dehydrogenase (Odh) malate dehydrogenase (Mdh), and succinate dehydrogenase (Sdh) [4]. However, the MOBs of *B. azoricus* possesses a complete TCA cycle [4], indicating the missing metabolic intermediates of the SOBs can be provided by the hosts and the MOBs.

As mentioned above, *B. marisindicus* does not encode genes for synthesizing several essential amino acids. However, we found evidence that *B. marisindicus* provides its symbiont with some metabolic intermediates and receive amino acids from its symbiont. In a host, a 3-mercaptopyruvate sulfurtransferase (MPST), two sulfide:quinone reductases (Sqr), two thiosulfate sulfurtransferase (Tst) and one sulfur dioxygenase (Sdo) were identified. MPST is known to be involved in hydrogen sulfide generation [48] whereas Sqr, Tst, and Sdo are associated with the mitochondrial oxidation of sulfide to thiosulfate. Both enzymes showed significantly elevated expression levels in the gills compared with other tissues (Tables S3 and S20), suggesting the abundance of thiosulfates in the gill tissues. These data supported the idea that in *Bathymodiolus* gills mitochondrial sulfide oxidation may create a pool of thiosulfate as a stable energy source for the thiotrophic symbionts [16]. Moreover, numerous copies of the host enzyme carbonic anhydrase (CA) were discovered (one of the CAs ranked first in the proteome and transcriptome) and significantly higher abundances in the gills than in other tissues (Tables S3 and S9), implying that these enzymes are involved in symbiotic processes. Similar to Cas in other deep-sea invertebrates [16, 49], CA in the gill tissue

may convert $CO_2$ to $HCO_3^-$, thus immobilizing and concentrating it for efficient fixation by the SOBs. In the genome and proteome, L-amino acid ABC transporters (AapJQAMP) were found, which may facilitate the transport of amino acids from endosymbionts to the host [16] (Fig. 4A). This mechanism may enable the *B. marisindicus* host to compensate for its inability to synthesize many essential amino acids. These results are consistent with previous findings in the vent mussel *Bathymodiolus thermophilus* hosting a SOB and *B. azoricus* hosting a SOB and a MOB [16], implying this might be a common feature of deep-sea mussel symbiosis.

**Endosymbionts also play a role in *B. marisindicus* holobiont defense**

*Bathymodiolus* mussels are infected by bacterial intranuclear pathogens called *Candidatus Endonucleobacter* bathymodiolin [21]. Additionally, the absence of intranuclear bacteria in the nuclei of symbiont-containing cells and growth inhibition assays indicated that *B. azoricus* gill tissue homogenates inhibit the growth of a wide spectrum of pathogens; this feature led to the hypothesis that the symbionts can protect their host cells from infection [22]. Our results showed that the symbionts of *B. marisindicus* possess a gene cluster involved in bacteriocin production (Fig. S11), and most of these genes were expressed at the protein level (Table S22). Bacteriocins are effector proteins that bacteria release into the environment. They are the most well-studied antibacterial effector proteins [50]. The bacteriocin is likely to be secreted into the bacteriocyte cytosol of the mussel host via the T2SS and Sec pathway (Fig. 3A), and complements host defense against other bacteria. Moreover, the symbiotic bacteria of deep-sea mussels have been suggested to "tame" some toxins such as YD repeats, and use them in beneficial interactions, and provide mussel hosts protection against natural enemies [36]. In the present study, we found that the toxin-related gene YD repeats was expanded in this SOB (Fig. 1D and Table S23). The YD repeat genes of the SOB in *Bathymodiolus* mussels provide protection against parasites and are involved in competition between closely related bacterial strains [36, 51]. Furthermore, the proteins of many YD repeat toxins (two of them are ranked the 1st and the 9th in

proteome, respectively) showed remarkably high expression levels (Table 1 and Table S23). Bacteria can use their type VI secretion system (T6SS) to inject antibacterial toxin into competing bacterial cells [51]. Intriguingly, two highly conserved genes of T6SS, *vgrG* and *Hcp,* were identified near the toxin-related genes (Table S18). Hcp can form hexameric rings that stack upon each other to form a membrane spanning nanotube, and the trimeric VgrG complex forms a closed cap on the Hcp nanotube [52], which enables SOB to deliver their YD repeat proteins to competing bacterial cells and exert its toxicity.

**Virus-endosymbiont interactions**

The intracellular space is general thought to be a closed environment that guards symbiotic bacteria against phage infection. The identification of bacteriophages infecting endosymbiont Wolbachia bacteria in insects indicates that this assumption may not be true in many invertebrates [53]. In this study, we found 569 unique viral genome sequences in the metagenome data of the *B. marisindicus* gills (Table S24), and 18 out of 21 viral genome sequences with hallmark genes were classified as belonging to the dsDNA phages, which can infect bacteria (Table S25). The mussel-associated phages might enter the gill cells through horizontal transfer processes, such as transcytosis, phagocytosis, active bacterial infection, or activation of a bacterial carrier [54]. After successfully entering mussel gill cells, phages may invade its bacterial host. Phages might primarily use a lysogenic infection strategy [55]. Our analysis of mussel SOB genome indicates the lysogenic lifestyles, as indicated by the identification of prophages based on virus-specific genes (phage integrase; Table S7). To withstand infection, bacteria have evolved numerous antiviral defense mechanisms that provide protection against phage predation [56]. In the *B. marisindicus* endosymbiont, we found over 150 genes related to 13 defense systems against phage infection and lysis (Table S26). We found the components of type I-F and type II CRISPR-Cas systems (Fig. 6A), which were regularly and compactly distributed in the symbiont genome. CRISPR-Cas systems are adaptive immunity systems that protect bacteria from their bacteriophages

and may respond to new threats by acquiring new spacers from invading nucleic acids. The Type I-F CRISPR-Cas system has 67 spacers, whereas the type II system possesses 48 spacers. Only one spacer matched the set of phage genome sequences possibly because of the rapid mutations for escaping CRISPR [57]. Furthermore, complete gene sets for all the four categories of RM systems, including type I, II, III, and IV (Fig. 6A), were encoded in the genome of the *B. marisindicus* symbiont. CRISPR-Cas and RM systems are functionally coupled. They can target specific sequences on the invading phages [58]. Moreover, the type II TA system was found by a toxic protein and its cognate antitoxin protein (Fig. 6A). As non-DNA-targeting systems, TA provides another line of defense. When phages successfully inject their DNA and start replication, TA systems induce the dormancy of infected cells by inhibiting gene expression [59]. Notably, many genes involved in these antiphage systems were expressed at the protein level (Fig.6B), indicating that they function together to establish complementary defense lines and may work synergistically to efficiently protect their hosts from phage infection [60]. The establishment of abundant defense systems in this SOB might be the result of the long-term co-evolution of the endosymbionts and phages. Many CRISPR-Cas and RM proteins were detected in *B. azoricus* symbionts (SOBs and MOBs) [4]. Furthermore, phage-bacteria interactions were found in deep-sea vent snail holobionts [26]. These observations indicate that this interaction is likely widespread in deep-sea animal symbionts.

Viruses and their bacterial hosts have a density-dependent association, which can result in the selective death of numerically dominant, and highly competitive taxa (termed the "killing the winner") [61]. Virus-mediated cell lysis is a major cause of bacterial death in deep-sea sediments, resulting in the release of cellular components to the environment and the microbial community changes [62]. Given that the SOB was the dominant strain in the *B. marisindicus* gill, we speculate that phages might regulate endosymbiont population through lysis, and allow the mussel hosts to have an additional pathway to obtain nutrients from their endosymbionts (Fig. 5).

## Conclusions

We have assembled the first deep-sea mussel genome that harbors SOB and the SOB genome. Through integrated multi-omic analyses, we have discovered a variety of specific evolutionary innovations that should help to elucidate their adaptation to endosymbiotic lifestyle. Our data revealed that expansion and functional differentiation of immunity-related gene families are key adaptive strategies of the deep-sea mussel, and the lack of many genes essential to amino acid biosynthesis and the contraction of GHF families highlight the dependence of a host on its endosymbionts. Furthermore, hologenomic analyses revealed that metabolic complementarity between the host and endosymbionts. Analyzing symbiont genome and proteome uncovered the potential role of endosymbiotic bacteria in host recognition and defense of the *B. marisindicus* holobiont, and its possible adaptions to the endosymbiotic lifestyle. Moreover, we have discovered an extensive antivirus system of endosymbiont and possible phage-endosymbiont interactions. Overall, this study has enriched our knowledge the mechanisms of symbiosis that has allowed these mussels to flourish in deep-sea hydrothermal vent ecosystems, and provide resources for understanding the evolution of deep-sea mussels and their symbionts.

## Material and methods

### Sample collection

*Bathymodiolus marisindicus* were collected in April 2019 from the Longqi hydrothermal vent field (49.65° E, 37.78° S; 2,757 m depth) situated on the Southwest Indian Ridge with a remotely operated vehicle (ROV) Hailong III on board the research vessel (R/V) *Dayang Yihao* during cruise 52III. Once the mussels were brought onboard the research vessel, the gill, mantle, adductor muscle, foot and visceral mass were dissected from one individual, fixed separately in RNAlater and then stored at −80 °C.

### DNA and RNA extraction

High-molecular-weight genomic DNA was extracted from the foot and gill separately with a the MagAttract High-Molecular-Weight DNA Kit (QIAGEN, Hilden,

Netherlands) according to the manufacturer's protocol, for sequencing the genomes of the host and the symbionts. Genomic DNA Clean & ConcentratorTM -10 kit (ZYMO Research, Irvine, CA, USA) was used for purifying the extracted DNA. TRIzol (Thermo Fisher Scientific, United States) was used for extracting total RNA extraction from the five dissected tissues. The quantity and quality of both DNA and RNA were examined using 1% agarose gel electrophoresis and NanoDrop 2000 (Thermo Fisher Scientific, United States), respectively. The DNA concentration was assessed using a QubitTM 3 Fluorometer (Thermo Fisher Scientific, Singapore).

**Library preparation and sequencing**

The host genome was sequenced from the foot tissue of the same individual with Oxford Nanopore Technology, PacBio sequel sequencing and Illumina platforms. The long-read DNA was used in constructing an 8-10 kb Nanopore DNA library with a ligation sequencing Kit (SQK-LSK109) according to the manufacturer's protocol and sequenced with the FLO-MIN106 R9.4 flow cell coupled to the MinIONTM platform (Oxford Nanopore Technologies, Oxford, UK) at the Hong Kong University of Science and Technology. The raw reads were processed by adopting high-accuracy base calling mode by Oxford Nanopore basecaller Guppy version 2.1.3 according to the manufacturer's protocol. Other purified DNA was used in constructing a 20k PacBio single-molecule real-time (SMRT) library (Pacific Biosciences, USA) and sequenced in SMRT Cell by Bioinformatics Technology Co., Ltd., Beijing, China (www.novogene.cn). Illumina DNA sequencing of the foot was performed on an Illumina HiSeq ™ X-Ten platform for the generation of 150 bp paired-end reads with the 500 bp short-insert DNA library at Novogene (Beijing, China). The symbiont genome was sequenced with both the Oxford Nanopore Technology and Illumina platforms from the gill of the same individual. A long-read DNA library of the gill was constructed and sequenced as mentioned above at Novogene with the Guppy version 3.2.10 for basecalling. Illumina short-reads DNA libraries with an insert size of 350 bp for the gill were constructed at Novogene and sequenced on an Illumina HiSeq 2500

platform. The Illumina RNA libraries of the five disserted tissues were constructed individually and sequenced on an Illumina HiSeq 2500 platform (PE150) at Novogene.

**Assembly and scaffolding of the host genome**

Trimmomatic version0.39 [63] was used in removing the adaptors and low-quality reads (quality score < 20, length < 40 bp) of the Illumina data. PacBio subreads over 5 kb were corrected and trimmed using Canu version 1.7.1 [64] with the following settings: genome Size = 1.15 Gb, corMhapSensitivity = normal, corMinCoverage = 4, corOutCoverage=200, correctedErrorRate = 0.105, then wtdbg2 version 2.1 and wtpoa-cns [65] were applied to assemble the genome under default settings. to improve the accuracy of the draft genome assembly, we conducted two rounds of error correction using PacBio subreads by Racon version1.441 [66] and polished twice with Illumina reads using Pilon version1.13 [67]. Bacterial contamination in the host genome assembly was further filtered using MaxBin version 2.2.5 [68]. The quality of the host genome assembly was assessed using BUSCO version5.1.3 [69] and the Metazoa database.

**Symbiont genome assembly**

The Illumina raw data were filtered using Trimmomatic version 0.39 [63] to remove adapters and low-quality bases. Clean reads were first assembled using SPAdes version 3.13.0 [70] with the --meta setting and k-mer sizes of 55, 77, 99 and 127 bp. Genome binning of symbiont genome followed previous study [71]. In brief, clean reads were mapped to the initial assemble result by Bowtie version 2.3.5 [72] and the coverage of each contig was calculated by SAMtools version 1.9 [73]. The GC and tetranucleotide content were calculated by calc.gc.pl and calc.kmerfreq.pl [71]. Conserved marker proteins were identified step by step using Prodigal version 2.6.3 (Hyatt et al. 2010), HMMER version 3.2.1 [74], BLASTp version 2.9.0 [75] and imported to MEGAN version 6.2.1 [76] to cluster their taxonomic affiliation. The results were analyzed in RStudio following the metagenome.workflow.modified.R script [71] to extract the

symbiont genome based on the sequencing coverage and the GC content. Further scaffolding using the Single Molecular Integrative Scaffolding (SMIS) pipeline (https://github.com/fg6/smis) by adding filtered Nanopore long reads, which classified against the NCBI RefSeq database of bacterial using Kraken2 [77]. The gap-closing software TGS-GapCloser [78] added above-mentioned Nanopore long reads to fill the gaps and enhance genome assembly. CheckM version 1.1.3 [79] was utilized to evaluate the completeness and the contamination of the final assembly.

**Gene prediction and functional annotation**

The prediction of protein coding sequences and proteins of symbiont genome was performed using Prodigal version 2.6.3 [80] with default parameters. RepeatProteinMask in RepeatMasker [81] was used in identifying the repetitive sequences in *B. marisindicus* genome, and then RepeatModeler [81] and LTR FINDER.x86 64-1.0.6 [82] were used in constructing a de novo repeat library. Repetitive elements were predicted using Tandem Repeat Finder (version 4.07b) [83]. The protein-coding genes of *B. marisindicus* genome were predicted using a combination of *ab initio*, homology-based, and transcriptome-based methods. A*b initio* prediction was performed using AUGUSTUS version 3.2.1 [84]. Transcriptome-based annotation was conducted using RNA-seq data from five *B. marisindicus* tissues (gill, foot, adductor muscle, visceral mass, and mantle). For the homology-based gene prediction, homologous proteins of several reported mollusk species (*Archivesica marissinica*, *Crassostrea gigas*, *Mizuhopecten yessoensis*, *Modiolus philippinarum*; *Bathymodiolus platifrons*) were downloaded from NCBI and aligned to *B. marisindicus* genome using tBLASTn version 2.4.0+ with e-value ≤ 1e–5. Subsequently, all the achieved alignments were analyzed using Genewise version 2.2.0 software [85] to search for precise gene structures, and these prematurely, terminated frame-shifted, or short (less than 200 bp) genes were removed. The gene structures obtained using these three approaches were integrated with MAKER version 2.31.10 [86] to yield a nonredundant gene set. For achieving a functional annotation, predicted protein

sequences were aligned against public databases including Swiss-Prot [87], NCBI non-redundant (NR) database, Clusters of Orthologous Groups (COG) [88], Gene Ontology (GO) [89], InterPro [90], and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway [91].

**Phylogenomic analysis**

The orthologous groups shared between the predicted proteins of *B. marisindicus* and those of other 19 selected molluscan genomes were identified utilizing OrthoMCL version 1.1 [92]. All possible matches among the retained protein sequences were identified through All-vs-All Blast. An e-value of 1e-7 was used in the search for potential matches among the retained protein sequences. Lastly, OrthoMCL with an inflation index of 1.5 were used in grouping the alignments into gene families. MAFFT version 7.237 [93] was used to align the amino acid sequences of each remaining single-copy gene. All of the aligned sequences were also concatenated and then served as the concatenated dataset. Phylogenetic analysis was carried out utilizing IQ-TREE version 1.6.10 [94] with settings of 1000 ultrafast bootstraps. By calibrating the phylogenetic tree with seven fossil records and geographic events, the software MCMCtree [95] was utilized to yield the time-calibrated tree (Fig. S3). To investigate phylogenetic relationship of the symbionts, we examined the genomes of bacterial symbionts belonging to Gammaproteobacteria from deep-sea invertebrates (Fig. S6). The same pipelines applied to *B. marisindicus* were used in the symbiont analyses for the generation of the symbiotic orthologue clusters. MAFFT version 7.237 [93] was used for protein alignments. The alignments of single-copy orthologues were concatenated for subsequent phylogeny analysis. The IQ-TREE version 1.6.10 [94] was used in performing the phylogeny analysis of the symbionts.

**Host gene family analyses**

Gene families shared by *B. marisindicus* and six bivalve genomes (i.e., *Pinctada fucata*,

*Crassostrea gigas*, *Mizuhopecten yessoensis*, *Modiolus philippinarum*; *Ruditapes philippinarum* and *Argopecten purpuratus*) were used in the gene family analyses. The expansion and contraction of these gene families in *B. marisindicus* were detected using CAFE version 4.2.1 [96] and Fisher's exact test. The p-values were corrected using the false discovery rate with an adjusted *p*-value of < 0.05. The IQ-TREE with 1000 bootstrap replicates was used for phylogenetic analyses of selected genes. The expanded domains in *B. marisindicus* genome were annotated using Pfam with an e-value of <1e-5.

**Host gene expression analysis**

The adaptors and low-quality reads (> 10% Ns, Phred value Q ≤ 20; < 40 bp in length) in RNA-seq data were removed using Trimmomatic version 0.39 [63]. Gene expression levels were normalized as transcripts per million (TPM) using Salmon version 0.9.1 [97] under default settings. The highly expressed genes in the host gill were determined by differential expression analysis versus foot, visceral mass, adductor muscle and mantle (n = 5) using edgeR [98] based on the reads counts. Only genes with >2-fold expressional difference and a significant FDR *p*-value of < 0.05 were considered as highly expressed genes.

**Pseudogenes**

The repeat regions and genes of the host genome were masked, and a tBLASTn version 2.4.0+ with *e*-value of < 1e-20 and the SEG low-complexity filter was used in homologous search for pseudogene candidates in the intergenic regions. Candidate pseudogenes were identified utilizing the Pseudogene Pipeline (https://github.com/ShiuLab/PseudogenePipeline) with the following settings: identity > 60%, match length > 50 amino acids, and query coverage > 70% of the query sequence [8]. Putative processed pseudogenes were classified by scanning for insertion of retrotransposons on their 2 kb flanking regions. RNAseq data were mapped to the genome assembly utilizing histat2 version 2.1.0 [99] with default parameters for

assessment the expression of candidate pseudogenes. SAMtools version 1.9 with default settings was used to sort and index aligned reads (with mapping quality ≥10). The read counts in each tissue were produced by running the multicov program in BEDTools version 2.24.0 [100] under default parameters. Pseudogenes with read counts of > 5 were considered as expressed.

**Metaproteomics**

From the gills of three *B. marisindicus* individuals, proteins were extracted utilizing the methanol-chloroform method [101]. To separate different sizes of proteins ranging from 10 kDa to 150 kDa, SDS-PAGE gel was run for ~30 μg of the extracted protein from each sample and stained by colloidal Coomassie blue. The peptide for LC-MS/MS was acquired by alkylation and digestion, protein reduction, drying, and peptide extraction. Dionex UltiMate 3000 RSLCnano coupled with an Orbitrap Fusion Lumos Mass Spectrometer (Thermo Fisher Scientific, Bremen, Germany) was used for analyzing each protein fraction. The search database includes the protein sequences predicted from the genome and the corresponding reversed sequences (decoy) of both *B. marisindicus* and its endosymbiont. Proteome Discoverer software version 2.4 (Thermo Fisher Scientific, Bremen, Germany) was used in the quantification and identification of proteins based on the raw mass spectrometry data. Proteins were identified with the assigned peptides' identification confidence level of over 0.95 and false discovery rate of 2.5%.

**Identification of bacteriocin gene cluster and virus genomes**

The identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in the symbiont genome were conducted using antiSMASH version 6.0 [102] with default parameters. Besides, VirSorter2 [103] with default parameters was used to predict and classify viral sequences from the initial assemblies that was generated from SPAdes version 3.13.0 [70]. In the following, sequences longer than 3kb with maxscore > 0.5 were identified as putative viral sequences. CheckV [104] with default settings was

used to estimate the completeness and contamination of the putative viral sequences and to identify proviruses among the viral sequences.

**Identification of defense system genes of the symbiont**

To explore the variety of defence systems, BLASTp in the DIAMOND program [75] was used in searching the genes against the PADS Arsenal database [105] with custom settings as below: more sensitive mode, identity ≥50%, *e*-value $<10^{-10}$. Bacterial genes mapped to the PADS database were examined to confirm that the discovered genes contained conserved domains engaged in the prokaryotic defense against phages through the use of HMMScan in the HMMER version 3.3 tool suite [106] against PFAM version 32.0 [107] (*e*-value $<10^{-3}$, bit score ≥30) from a past research [56], and the pfam accessions of the conserved domains was manually retrieved. To forecast the completeness of the defense systems, the gene components of a system were identified in a contig sequence as reported earlier [108]. A system was deemed complete if it included all the genes necessary for that system to operate. Besides, MetaCRT [109] was used in predicting the CRISPR spacers, and spacers >6 bp in length were matched to phage genome sequences with fuzznuc [110].

**Abbreviations**

MOB: Methane-oxidising bacteria; SOB: Sulphur-oxidising bacteria; PRRs: Pattern recognition receptors (PRRs); PAMPs: Pathogen-associated molecular patterns; T2SS: Type II secretion system; T6SS: type VI secretion system; CRISPR: Clustered regularly interspaced short palindromic repeat; DISARM: Defence island system associated with restriction-modification; RM: Restriction–modification system; TA: Toxin–antitoxin system;

**Supplementary Information**

Additional file 1: Supplementary Figures, and Tables S1, S4 and S8.

Additional file 2: Supplementary Tables S2, S5-S7, S9-S26.

**Acknowledgements**

We thank the captain and crew of the R/V Dayang Yihao as well as the operation team of the ROV Sea Dragon III during the third leg of the China Ocean Mineral Resources Research and Development Association DY52nd cruise, and Dr. Yanan Sun from Hong Kong Baptist University for her help with sample collection.

**Authors' contributions**

P.-Y.Q. conceived the project. K.Z., Y.X., J.S., J.-W.Q. and P.-Y.Q. designed the experiments. J.S. and T.X. collected the *Bathymodiolus marisindicus*. K.Z. and J.S. performed host genome assembly. Y.H.K. and K.Z. performed the proteome analysis. Y.X. and K.Z. performed the DNA extraction, RNA extraction, ONT sequencing, symbiont genome assembly, and gene expression analysis. K.Z. conducted other data analyses. K.Z. and Y.X. drafted the manuscript. All authors contributed to improvement of the manuscript and approved it for submission and publication.

**Funding**

This work was supported by grants from the Major Project of Basic and Applied Basic Research of Guangdong Province (2019B030302004), Key Special Project for Introduced Talents Team of Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou) (GML2019ZD0404, GML2019ZD0409), the Hong Kong Branch of Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou) (SMSEGL20SC01, SMSEGL20SC02), and China Ocean Mineral Resources Research and Development Association (DY135-E2-1-03).

**Availability of data and materials**

All sequencing data and assembly data of *B. marisindicus* and its symbiont were deposited to the National Centre for Biotechnology Information (NCBI) database under BioProject PRJNA772587.

**Declarations**

**Ethics approval and consent to participate**

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

The authors declare no competing interests

## Author details

[1] Department of Ocean Science and Hong Kong Branch of the Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), The Hong Kong University of Science and Technology, Hong Kong, China;

[2] Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Guangzhou 511458, China;

[3] Institute of Evolution & Marine Biodiversity, Ocean University of China, Qingdao, 266003, China

[4] Department of Biology, Hong Kong Baptist University, Hong Kong, China

## References


1. Engelstädter J, Hurst GDD. The ecology and evolution of microbes that manipulate host reproduction. Annu Rev Ecol Evol Syst. 2009;40:127–49.

2. Foster KR, Schluter J, Coyte KZ, Rakoff-Nahoum S. The evolution of the host microbiome as an ecosystem on a leash. Nature. 2017;548(7665):43–51.

3. Dubilier N, Bergin C, Lott C. Symbiotic diversity in marine animals: The art of harnessing chemosynthesis. Nat Rev Microbiol. 2008;6(10):725–40.

4. Ponnudurai R, Kleiner M, Sayavedra L, Petersen JM, Moche M, Otto A, et al. Metabolic and physiological interdependencies in the *Bathymodiolus azoricus* symbiosis. ISME J. 2017;11(2):463–77.

5. Yang Y, Sun J, Sun Y, Kwan YH, Wong WC, Zhang Y, et al. Genomic, transcriptomic, and proteomic insights into the symbiosis of deep-sea tubeworm holobionts. ISME J. 2020;14(1):135–50.

6. Li Y, Tassia MG, Waits DS, Bogantes VE, David KT, Halanych KM. Genomic adaptations to chemosymbiosis in the deep-sea seep-dwelling tubeworm *Lamellibrachia luymesi*. BMC Biol. 2019;17(1):1–14.

7. Sun Y, Sun J, Yang Y, Lan Y, Ip JC-H, Wong WC, et al. Genomic signatures supporting the symbiosis and formation of chitinous tube in the deep-sea tubeworm *Paraescarpia echinospica*. Mol Biol Evol. 2021;38(10): 4116-4134.

8. Ip JCH, Xu T, Sun J, Li R, Chen C, Lan Y, et al. Host-Endosymbiont Genome Integration in a Deep-Sea Chemosymbiotic Clam. Mol Biol Evol. 2021;38(2):502–18.

9. Lan Y, Sun J, Chen C, Sun Y, Zhou Y, Yang Y, et al. Hologenome analysis reveals dual symbiosis in the deep-sea hydrothermal vent snail *Gigantopelta aegis*. Nat Commun. 2021;12(1):1–15.

10. Sun J, Chen C, Miyamoto N, Li R, Sigwart JD, Xu T, et al. The Scaly-foot Snail genome and implications for the origins of biomineralised armour. Nat Commun. 2020;11(1):1–12.

11. Distel DL, Lee HKW, Cavanaugh CM. Intracellular coexistence of methano- and thioautotrophic bacteria in a hydrothermal vent mussel. Proc Natl Acad Sci U S A.

1995;92(21):9598–602.
12. Miyazaki JI, de Oliveira Martins L, Fujita Y, Matsumoto H, Fujiwara Y. Evolutionary process of deep-sea *Bathymodiolus* mussels. PLoS One. 2010;5(4):1–11.
13. Lorion J, Kiel S, Faure B, Kawato M, Ho SYW, Marshall B, et al. Adaptive radiation of chemosymbiotic deep-sea mussels. Proc R Soc B Biol Sci. The Royal Society; 2013;280(1770):20131243.
14. Govenar B. Shaping vent and seep communities: habitat provision and modification by foundation species. Vent seep biota. Springer; 2010. p. 403–432.
15. Sun J, Zhang Y, Xu T, Zhang Y, Mu H, Zhang Y, et al. Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. Nat Ecol Evol. 2017;1(5):1–7.
16. Ponnudurai R, Heiden SE, Sayavedra L, Hinzke T, Kleiner M, Hentschker C, et al. Comparative proteomics of related symbiotic mussel species reveals high variability of host–symbiont interactions. ISME J. 2020;14:649–56.
17. Oliver KM, Smith AH, Russell JA. Defensive symbiosis in the real world - advancing ecological studies of heritable, protective bacteria in aphids and beyond. Funct Ecol. 2014;28(2):341–55.
18. Schmid M, Sieber R, Zimmermann Y, Vorburger C. Development, specificity and sublethal effects of symbiont-conferred resistance to parasitoids in aphids. Funct Ecol. Wiley Online Library; 2012;26:207–15.
19. DeChaine EG, Bates AE, Shank TM, Cavanaugh CM. Off-axis symbiosis found: Characterization and biogeography of bacterial symbionts of *Bathymodiolus* mussels from Lost City hydrothermal vents. Environ Microbiol. 2006;8(11):1902–12.
20. Duperron S, Halary S, Lorion J, Sibuet M, Gaill F. Unexpected co-occurrence of six bacterial symbionts in the gills of the cold seep mussel *Idas* sp. (Bivalvia: Mytilidae). Environ Microbiol. 2008;10(2):433–45.
21. Zielinski FU, Pernthaler A, Duperron S, Raggi L, Giere O, Borowski C, et al. Widespread occurrence of an intranuclear bacterial parasite in vent and seep bathymodiolin mussels. Environ Microbiol. 2009;11(5):1150–67.

22. Bettencourt R, Roch P, Stefanni S, Rosa D, Colaço A, Serrão Santos R. Deep sea immunity: Unveiling immune constituents from the hydrothermal vent mussel *Bathymodiolus azoricus*. Mar Environ Res. 2007;64(4):108–27.
23. Fujii Y, Kubo T, Ishikawa H, Sasaki T. Isolation and characterization of the bacteriophage WO from Wolbachia, an arthropod endosymbiont. Biochem Biophys Res Commun. Elsevier; 2004;317(4):1183–8.
24. Chauvatcharin N, Ahantarig A, Baimai V, Kittayapong P. Bacteriophage WO-B and Wolbachia in natural mosquito hosts: infection incidence, transmission mode and relative density. Mol Ecol. 2006;15(9):2451–61.
25. Bordenstein SR, Marshall ML, Fry AJ, Kim U, Wernegreen JJ. The tripartite associations between bacteriophage, *Wolbachia*, and arthropods. PLoS Pathog. 2006;2(5):e43.
26. Zhou K, Xu Y, Zhang R, Qian PY. Arms race in a cell: genomic, transcriptomic, and proteomic insights into intracellular phage–bacteria interplay in deep-sea snail holobionts. Microbiome. Microbiome; 2021;9(1):1–13.
27. Yamanaka T, Mizota C, Fujiwara Y, Chiba H, Hashimoto J, Gamo T, et al. Sulphur-isotopic composition of the deep-sea mussel Bathymodiolus marisindicus from currently active hydrothermal vents in the Indian Ocean. J Mar Biol Assoc United Kingdom. 2003;83(4):841–8.
28. Zhang K, Sun J, Xu T, Qiu JW, Qian PY. Phylogenetic relationships and adaptation in deep-sea mussels: Insights from mitochondrial genomes. Int J Mol Sci. 2021;22(4):1–13.
29. Casacuberta E, González J. The impact of transposable elements in environmental adaptation. Mol Ecol. 2013;22(6):1503–17.
30. Cheetham SW, Faulkner GJ, Dinger ME. Overcoming challenges and dogmas to understand the functions of pseudogenes. Nat Rev Genet. 2020;21(3):191–201.
31. Baumgarten S, Simakov O, Esherick LY, Liew YJ, Lehnert EM, Michell CT, et al. The genome of *Aiptasia*, a sea anemone model for coral symbiosis. Proc Natl Acad Sci U S A. 2015;112(38):11893–8.

32. Li M, Chen H, Wang M, Zhong Z, Zhou L, Li C. Identification and characterization of endosymbiosis-related immune genes in deep-sea mussels *Gigantidas platifrons*. J Oceanol Limnol. 2020;38(4):1292–303.
33. Détrée C, Haddad I, Demey-Thomas E, Vinh J, Lallier FH, Tanguy A, et al. Global host molecular perturbations upon in situ loss of bacterial endosymbionts in the deep-sea mussel *Bathymodiolus azoricus* assessed using proteomics and transcriptomics. BMC Genomics. BMC Genomics; 2019;20(1):1–14.
34. de Beco S, Gueudry C, Amblard F, Coscoy S. Endocytosis is required for E-cadherin redistribution at mature adherens junctions. Proc Natl Acad Sci. 2009;106(17):7010–5.
35. Derivery E, Sousa C, Gautier JJ, Lombard B, Loew D, Gautreau A. The Arp2/3 activator WASH controls the fission of endosomes through a large multiprotein complex. Dev Cell. Elsevier; 2009;17(5):712–23.
36. Sayavedra L, Kleiner M, Ponnudurai R, Wetzel S, Pelletier E, Barbe V, et al. Abundant toxin-related genes in the genomes of beneficial symbionts from deep-sea hydrothermal vent mussels. Elife. 2015;4:1–39.
37. Hentschel U, Piel J, Degnan SM, Taylor MW. Genomic insights into the marine sponge microbiome. Nat Rev Microbiol. 2012;10(9):641–54.
38. Jeannin P, Bottazzi B, Sironi M, Doni A, Rusnati M, Presta M, et al. Complexity and complementarity of outer membrane protein A recognition by cellular and humoral innate immunity receptors. Immunity. 2005;22(5):551–60.
39. Hirabayashi J. Lectin Purification and Analysis. Springer; 2020.
40. Maculins T, Fiskin E, Bhogaraju S, Dikic I. Bacteria-host relationship: Ubiquitin ligases as weapons of invasion. Cell Res. 2016;26(4):499–510.
41. Sayavedra L, Ansorge R, Rubin-Blum M, Leisch N, Dubilier N, Petersen J. Horizontal acquisition followed by expansion and diversification of toxin-related genes in deep-sea bivalve symbionts. bioRxiv. 2019;605386.
42. Nivaskumar M, Francetic O. Type II secretion system: A magic beanstalk or a protein escalator. Biochim Biophys Acta - Mol Cell Res. 2014;1843(8):1568–77.

43. Hu YF, Zhao D, Yu XL, Hu YL, Li RC, Ge M, et al. Identification of bacterial surface antigens by screening peptide phage libraries using whole bacteria cell-purified antisera. Front Microbiol. 2017;8:1–9.
44. Le Pennec M, Beninger PG, Herry A. Feeding and digestive adaptations of bivalve molluscs to sulphide-rich habitats. Comp Biochem Physiol -- Part A Physiol. 1995;111(2):183–9.
45. Conway N. Occurrence of lysozyme in the common cockle Cerastoderma edule and the effect of the tidal cycle on lysozyme activity. Mar Biol. 1987;95(2):231–5.
46. Tsujino H, Yamashita T, Nose A, Kukino K, Sawai H, Shiro Y, et al. Disulfide bonds regulate binding of exogenous ligand to human cytoglobin. J Inorg Biochem. 2014;135:20–7.
47. Assié A, Leisch N, Meier D V., Gruber-Vodicka H, Tegetmeyer HE, Meyerdierks A, et al. Horizontal acquisition of a patchwork Calvin cycle by symbiotic and free-living Campylobacterota (formerly Epsilonproteobacteria). ISME J. 2020;14(1):104–22.
48. Pedre B, Dick TP. 3-Mercaptopyruvate sulfurtransferase: An enzyme at the crossroads of sulfane sulfur trafficking. Biol Chem. 2021;402(3):223–37.
49. Hongo Y, Nakamura Y, Shimamura S, Takaki Y, Uematsu K, Toyofuku T, et al. Exclusive localization of carbonic anhydrase in bacteriocytes of the deep-Sea clam *calyptogena okutanii* with thioautotrophic symbiotic bacteria. J Exp Biol. 2013;216(23):4403–14.
50. Ishibashi N, Himeno K, Masuda Y, Perez RH, Iwatani S, Zendo T, et al. Gene cluster responsible for secretion of and immunity to multiple bacteriocins, the NKR-5-3 enterocins. Appl Environ Microbiol. 2014;80(21):6647–55.
51. Koskiniemi S, Lamoureux JG, Nikolakakis KC, De Roodenbeke CTK, Kaplan MD, Low DA, et al. Rhs proteins from diverse bacteria mediate intercellular competition. Proc Natl Acad Sci U S A. 2013;110(17):7032–7.
52. Benz J, Meinhart A. Antibacterial effector/immunity systems: it's just the tip of the iceberg. Curr Opin Microbiol. Elsevier; 2014;17:1–10.

906 53. Kent BN, Bordenstein SR. Phage WO of Wolbachia: lambda of the endosymbiont
907 world. Trends Microbiol. 2010;18(4):173–81.
908 54. Abad FX, Pinto RM, Gajardo R, Bosch A. Viruses in mussels: public health
909 implications and depuration. J Food Prot. 1997;60(6):677–81.
910 55. Knowles B, Silveira CB, Bailey BA, Barott K, Cantu VA, Cobián-Güemes AG, et
911 al. Lytic to temperate switching of viral communities. Nature. 2016;531(7595):466–
912 70.
913 56. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, et al. Systematic
914 discovery of antiphage defense systems in the microbial pangenome. Science.
915 2018;359(6379):0–12.
916 57. Landsberger M, Gandon S, Meaden S, Rollie C, Chevallereau A, Chabas H, et al.
917 Anti-CRISPR Phages Cooperate to Overcome CRISPR-Cas Immunity. Cell.
918 2018;174(4):908-916.e12.
919 58. Isaev AB, Musharova OS, Severinov K V. Microbial Arsenal of Antiviral
920 Defenses – Part I. Biochem. 2021;86(3):319–37.
921 59. Heaton BE, Herrou J, Blackwell AE, Wysocki VH, Crosson S. Molecular
922 structure and function of the novel BrnT/BrnA toxin-antitoxin system of Brucella
923 abortus. J Biol Chem. 2012;287(15):12098–110.
924 60. Dupuis M-È, Villion M, Magadán AH, Moineau S. CRISPR-Cas and restriction–
925 modification systems are compatible and increase phage resistance. Nat Commun.
926 2013;4(1):1–7.
927 61. Winter C, Bouvier T, Weinbauer MG, Thingstad TF. Trade-Offs between
928 competition and defense specialists among unicellular planktonic organisms: the
929 “killing the winner” hypothesis revisited. Microbiol Mol Biol Rev. 2010;74(1):42–57.
930 62. Heinrichs ME, Tebbe DA, Wemheuer B, Niggemann J, Engelen B. Impact of viral
931 lysis on the composition of bacterial communities and dissolved organic matter in
932 deep-sea sediments. Viruses. 2020;12(9):22.
933 63. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina
934 sequence data. Bioinformatics. 2014;30(15):2114–20.

64. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Res. 2017;27(5):722–36.
65. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2020;17(2):155–8.
66. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27(5):737–46.
67. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014; 9(11): e112963.
68. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 2016;32(4):605–7.
69. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.
70. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.
71. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol. 2013;31(6):533–8.
72. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
73. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
74. Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. 2009;23:205–11.
75. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using

964 DIAMOND. Nat Methods. Nature Publishing Group; 2015;12(1):59–60.
965 76. Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC. Integrative analysis
966 of environmental sequences using MEGAN4. Genome Res. 2011;21(9):1552–60.
967 77. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2.
968 Genome Biology; 2019;20(1):1–13.
969 78. Xu M, Guo L, Gu S, Wang O, Zhang R, Peters BA, et al. TGS-GapCloser: a fast
970 and accurate gap closer for large genomes with low coverage of error-prone long
971 reads. Gigascience. 2020;9(9):1–11.
972 79. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM:
973 assessing the quality of microbial genomes recovered from isolates, single cells, and
974 metagenomes. Genome Res. 2015;25(7):1043–55.
975 80. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal:
976 prokaryotic gene recognition and translation initiation site identification. BMC
977 Bioinformatics. BioMed Central; 2010;11(1):1–11.
978 81. Chen N. Using RepeatMasker to identify repetitive elements in genomic
979 sequences. Curr Protoc Bioinforma. 2004;5(1):4–10.
980 82. Xu Z, Wang H. LTR-FINDER: an efficient tool for the prediction of full-length
981 LTR retrotransposons. Nucleic Acids Res. 2007;35(suppl_2):265–8.
982 83. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic
983 Acids Res. 1999;27(2):573–80.
984 84. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in
985 eukaryotes that allows user-defined constraints. Nucleic Acids Res.
986 2005;33(suppl_2):465–7.
987 85. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res.
988 2004;14(5):988–95.
989 86. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an
990 easy-to-use annotation pipeline designed for emerging model organism genomes.
991 Genome Res. 2008;18(1):188–96.
992 87. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et

993 al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.
994 Nucleic Acids Res. 2003;31(1):365–70.
995 88. Tatusov RL, Galperin MY, Natale DA, Koonin E V. The COG database: a tool for
996 genome-scale analysis of protein functions and evolution. Nucleic Acids Res.
997 2000;28(1):33–6.
998 89. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, et al. The Gene
999 Oncology (GO) database and informatics resource. Nucleic Acids Res. 2004;32
1000 (suppl_1):258–61.
1001 90. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al.
1002 InterPro: the integrative protein signature database. Nucleic Acids Res.
1003 2009;37(suppl_1):211–5.
1004 91. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at
1005 GenomeNet. Nucleic Acids Res. 2002;30(1):42–6.
1006 92. Li L, Stoeckert CJJ, Roos DS. OrthoMCL: identification of ortholog groups for
1007 eukaryotic genomes. Genome Res. 2003;13(9):2178–89.
1008 93. Rozewicki J, Li S, Amada KM, Standley DM, Katoh K. MAFFT-DASH:
1009 integrated protein sequence and structural alignment. Nucleic Acids Res.
1010 2019;47(W1):W5–10.
1011 94. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and
1012 effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol
1013 Biol Evol. 2015;32(1):268–74.
1014 95. Reis M Dos, Yang Z. Approximate likelihood calculation on a phylogeny for
1015 bayesian estimation of divergence times. Mol Biol Evol. 2011;28(7):2161–72.
1016 96. De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for
1017 the study of gene family evolution. Bioinformatics. 2006;22(10):1269–71.
1018 97. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and
1019 bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417–9.
1020 98. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for
1021 differential expression analysis of digital gene expression data. Bioinformatics.

2009;26(1):139–40.
99. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.
100. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
101. Wessel D, Flügge UI. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. Anal Biochem. 1984;138(1):141–3.
102. Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Res. 2021;1:0–7.
103. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome. 2021;9(1):1–13.
104. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol. 2021;39(5):578–85.
105. Zhang Y, Zhang Z, Zhang H, Zhao Y, Zhang Z, Xiao J. PADS Arsenal: a database of prokaryotic defense systems related genes. Nucleic Acids Res. 2020;48(D1):D590–8.
106. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 2013;41(12):e121–e121.
107. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47(D1):D427–32.
108. Bernheim A, Sorek R. The pan-immune system of bacteria: antiviral defence as a community resource. Nat Rev Microbiol. 2020;18(2):113–9.
109. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics. 2007;8(1):1–8.

110. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet. 2000;16(6):276–7.

## Figures

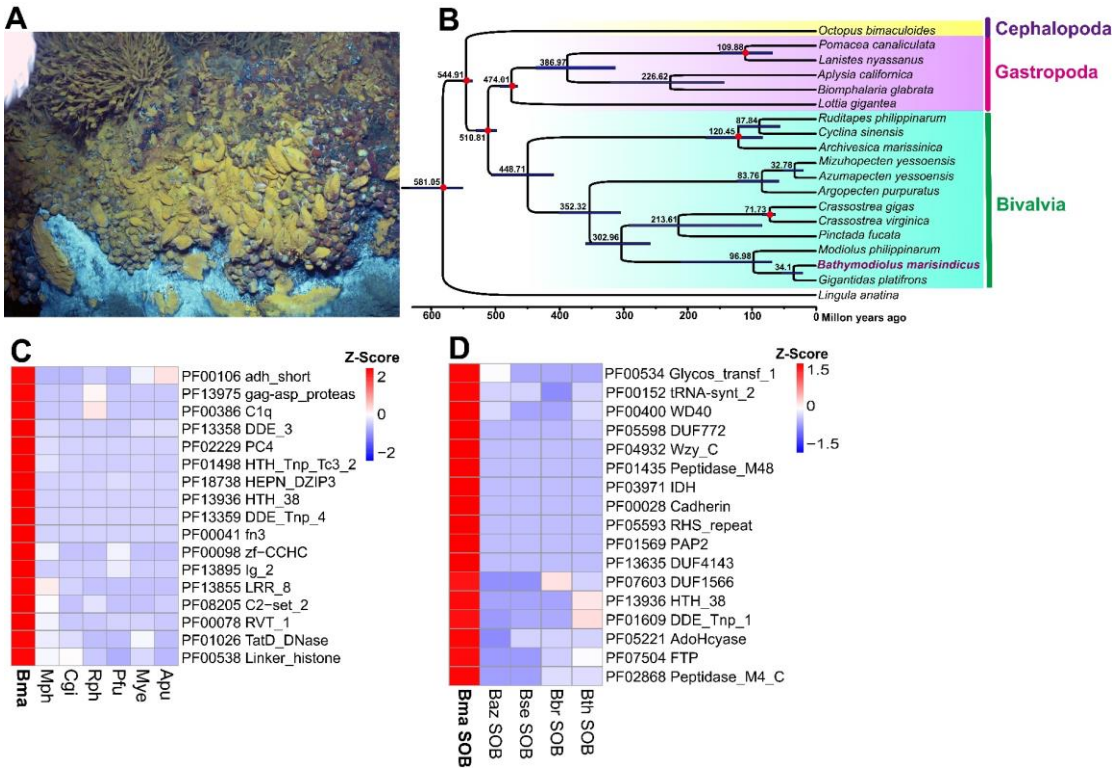


**Fig. 1** Phylogenetic position and gene family analysis of *Bathymodiolus marisindicus*. **A** A dense population of *B. marisindicus* on a chimney in the Longqi vent field; the surfaces of most mussels were covered with sulfide. **B** Maximum likelihood phylogenetic relationships among 19 molluscs with a brachiopod as the outgroup. The tree was calibrated at seven nodes (indicated by red dots) using fossils and geological events (Fig. S3). **C** and **D** Heat maps of the representative pfam domains that are expanded in *B. marisindicus* and its endosymbionts, with multiple domains in a given gene being counted as one. Abbreviations: Apu, *Argopecten purpuratus*; Bma, *Bathymodiolus marisindicus*; Cgi, *Crassostrea virginica*; Mph, *Modiolus philippinarum*; Mye, *Mizuhopecten yessoensis*; Pfu, Pinctada fucata; Rph, *Ruditapes philippinarum*. SOB, sulfur-oxidizing bacteria; Baz, *Bathymodiolus azoricus*; Bbr, *Bathymodiolus brooksi*; Bse, *Bathymodiolus septemdierum*; Bth, *Bathymodiolus thermophilus.*

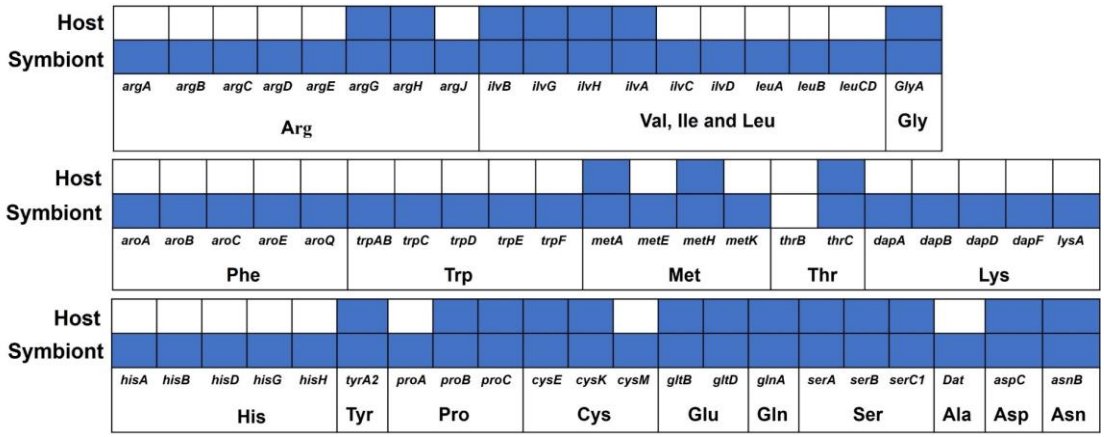


**Fig. 2** *Bathymodiolus marisindicus* lacks some essential amino acid biosynthesis genes. The presence (blue boxes) or absence (white boxes) of key genes are related to amino acid biosynthesis in the genomes of *B. marisindicus* and its symbionts.

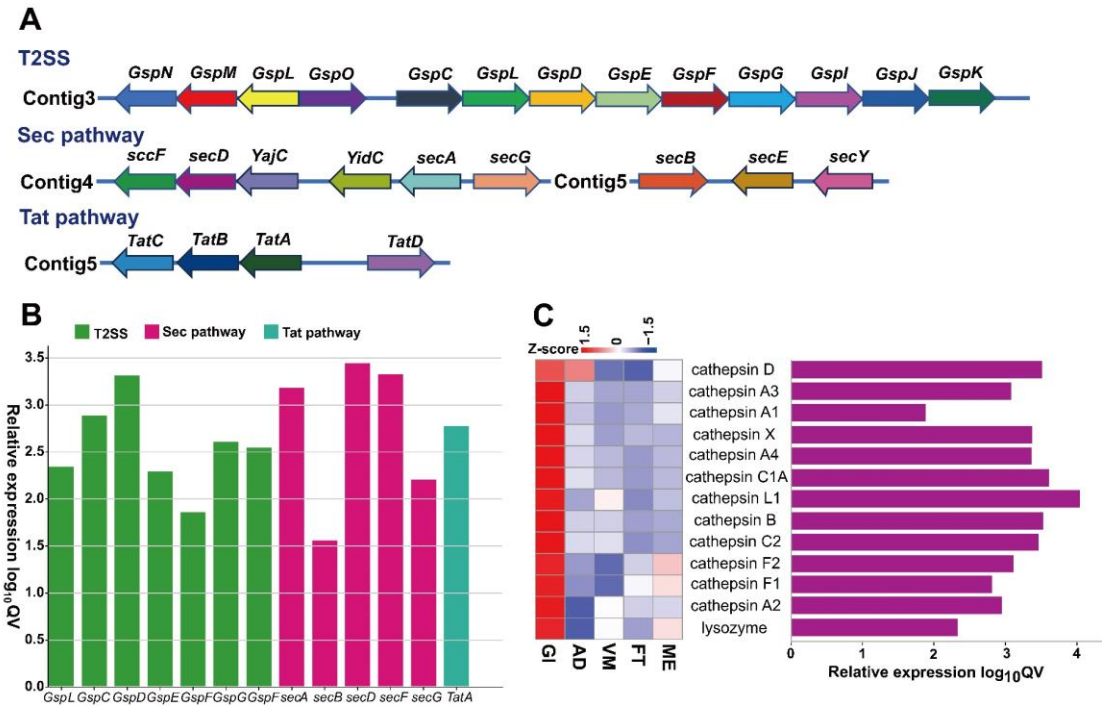


**Fig. 3** Secretion systems in the endosymbionts and high expression of digestive enzymes in the gills of *Bathymodiolus marisindicus*. **A** Schematic representation of the Type II secretion system (T2SS), the general secretory (Sec) pathway and Twin-arginine translocation (Tat) pathway in the SOBs. **B** The relative protein expression levels ($log_{10}$ QV) of genes associated with T2SS, Sec pathway and Tat pathway. **C** Left shows the tissue-specific expression (i.e., GI, gill; Ad, adductor muscle; VM, Visceral mass; FT, foot; ME, mantle) of cathepsins and lysozyme; the right panel shows the protein abundances in the gills. QV, Quantitative value.

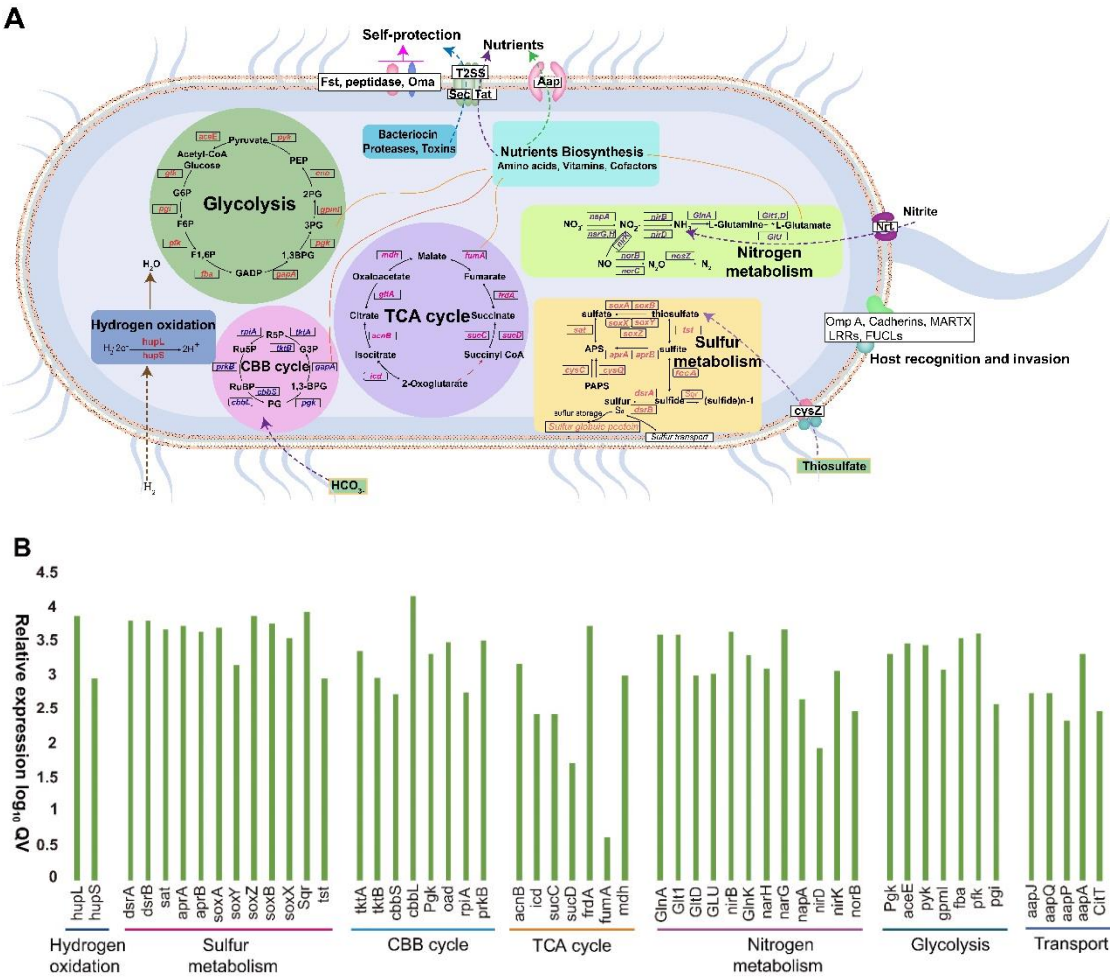


**Fig. 4** Central metabolism of the *Bathymodiolus marisindicus* symbiont. **A** A diagram showing the central metabolic pathways of the sulfur-oxidizing endosymbiont. **B** The relative gene expression levels (log10 QV) of key enzymes in the central metabolism. QV, Quantitative value. Abbreviations are provided in Supplementary Table S21.

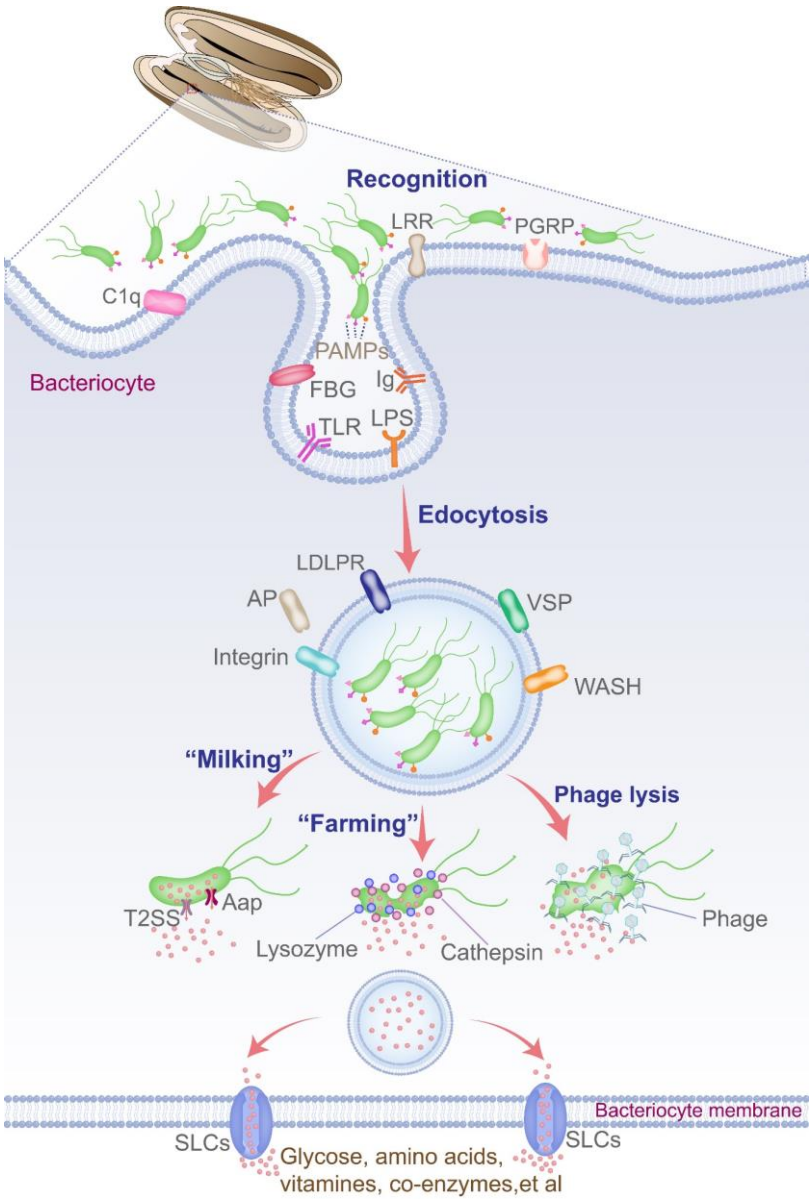


**Fig. 5** Model of symbiosis between *Bathymodiolus marisindicus* and its sulfur-oxidizing symbionts. The pathogen-associated molecular patterns (PAMPs) (i.e., MARTX, OmpA, LRR, FUCL, and cadherin) on the surface of SOB likely interact with the host pattern recognition receptors (PRRs; i.e., TLRs, LRRs, PGRPs, FBGs, C1q, Ig, LPS, LDLPRs, and VSP) that induce symbiont recognition and endocytosis. The nutrients synthesized by the SOBs are released through "farming" (direct digestion of symbionts), "milking" (molecular leakage of symbionts) and phage-mediated endosymbiont lysis, then the SLCs of the host can transport the nutrients across the cell membrane. Moreover, the SOB population can be regulated through symbiont digestion and phage-mediated endosymbiont lysis. AP, adaptor protein complex; C1qD, C1q domain; FBG, fibrinogen-related protein; FUCL, fucolectins; LDLR, low-density lipoprotein receptor-related protein; LRR, Leucine-

rich repeat; LPS, lipopolysaccharide; MARTX, multifunctional autoprocessing RTX toxins; Pal, peptidoglycan associated lipoprotein; PGRP, peptidoglycan recognition proteins; TLR, toll-like receptor; SLC, solute carrier; T2SS, type II secretion system; VSP, vacuolar sorting proteins; WASH, Wiskott-Aldrich syndrome protein and SCAR homologue.

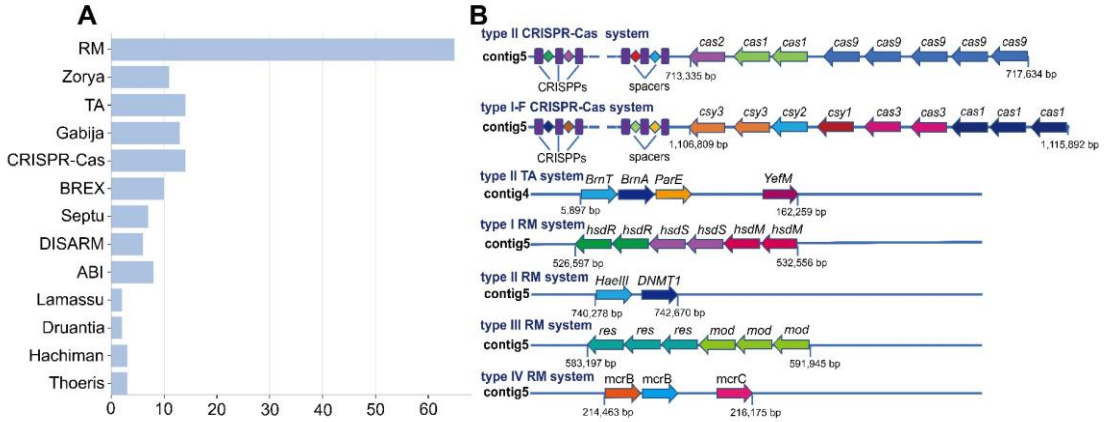


**Fig. 6** Antiviral defense systems in the sulfur-oxidizing symbionts of *Bathymodiolus marisindicus*. **A** Several genes detected in each defense system in the SOB. **B** Representative sequences of defense systems showing a complete set of required gene components.

## Tables

**Table 1** Toxin-related proteins found in the proteome of the SOB from *B. marisindicus*

| Identifier | Annotation | Category | Quantitative value |
|---|---|---|---|
| contig3_137 | RHS repeat-associated core domain-containing protein | YD | 26531.1 |
| contig3_138 | YD repeat-containing protein | YD | 10559.3 |
| contig3_140 | RHS repeat-associated core domain-containing protein | YD | 5304.4 |
| contig3_141 | RHS repeat-associated core domain-containing protein | YD | 111.4 |
| contig3_172 | RHS repeat-associated core domain-containing protein | YD | 1330.3 |
| contig3_175 | YD repeat-containing protein | YD | 322.2 |
| contig3_189 | RHS repeat-associated core domain-containing protein | YD | 157.7 |
| contig3_96 | RHS repeat-associated core domain-containing protein | YD | 19.2 |
| contig3_98 | insecticidal toxin complex protein | YD | 39 |
| contig3_116 | RHS repeat-associated core domain-containing protein | YD | 22.5 |
| contig3_360 | RHS repeat-associated core domain-containing protein | YD | 15 |
| contig16_2 | RHS repeat-associated core domain-containing protein | YD | 124.9 |
| contig4_51 | RHS family protein | YD | 268.2 |
| contig8_11 | RHS family protein | YD | 708.6 |
| contig8_12 | RHS family protein | YD | 548.2 |
| contig3_124 | Outer membrane adhesin-like protein | MARTX | 645.8 |
| contig3_126 | Cadherin repeat domain-containing protein | MARTX | 1065.2 |
| contig3_361 | Cadherin repeat domain-containing protein | MARTX | 939.6 |
| contig4_187 | Cadherin repeat domain-containing protein | MARTX | 846.7 |
| contig3_238 | RTX toxins and related $Ca^{2+}$-binding proteins | MARTX | 3405.7 |
| contig5_131 | RTX toxins and related $Ca^{2+}$-binding proteins | MARTX | 502.6 |
| contig5_158 | $Ca^{2+}$-binding protein, RTX toxin-related | MARTX | 9965.6 |
| contig5_1315 | $Ca^{2+}$-binding protein, RTX toxin-related | MARTX | 3080.1 |
| contig4_185 | $Ca^{2+}$-binding protein, RTX toxin-related | MARTX | 3138.3 |
| contig3_258 | $Ca^{2+}$-binding protein, RTX toxin-related | MARTX | 1084.9 |
| contig5_838 | $Ca^{2+}$-binding protein, RTX toxin-related | MARTX | 1932.5 |
| contig3_126 | $Ca^{2+}$-binding protein, RTX toxin-related | MARTX | 1075.1 |
| contig8_4 | $Ca^{2+}$-binding protein, RTX toxin-related | MARTX | 411 |
| contig5_179 | $Ca^{2+}$-binding protein, RTX toxin-related | MARTX | 194.8 |
| contig5_1312 | $Ca^{2+}$-binding protein, RTX toxin-related | MARTX | 79.8 |
| contig3_101 | Cadherin repeat domain-containing protein | MARTX | 148.3 |
| contig4_184 | Cadherin repeat domain-containing protein | MARTX | 359.8 |

## Supplementary Files

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

- Additionalfile1.pdf
- Additionalfile2.xlsx