nature ecology & evolution

SUPPLEMENTARY INFORMATION

VOLUME: 1 | ARTICLE NUMBER: 0121

In the format provided by the authors and unedited.

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

Jin Sun[1,2], Yu Zhang[3], Ting Xu[2], Yang Zhang[4], Huawei Mu[2], Yanjie Zhang[2], Yi Lan[1], Christopher J. Fields[5], Jerome Ho Lam Hui[6], Weipeng Zhang[1], Runsheng Li[2], Wenyan Nong[6], Fiona Ka Man Cheung[6], Jian-Wen Qiu[2*], Pei-Yuan Qian[1*]

[1] Division of Life Science, Hong Kong University of Science and Technology, Hong Kong, China
[2] Department of Biology, Hong Kong Baptist University, Hong Kong, China
[3] Shenzhen Key Laboratory of Marine Bioresource and Eco-Environmental Science, College of Life Science, Shenzhen University, Shenzhen, China
[4] Key Laboratory of Tropical Marine Bio-resources and Ecology, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou 510301, China
[5] High Performance Computing in Biology, Roy J. Carver Biotechnology Centre, University of Illinois Urbana-Champaign, USA
[6] Simon 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

Correspondence and requests for materials should be addressed to P.Y.Q* (boqianpy@ust.hk) or J.W.Q*. (qiujw@hkbu.edu.hk).

This PDF file includes:
Supplementary Notes
Figs. S1 to S7
Tables S1 to S7, S9, and S16

Mussel Genome Supplementary Notes

1. Background information

The discoveries of hydrothermal vents in the late 1970s (Corliss et al. 1979) and cold seeps in the early 1980s (Paull et al. 1984) have revolutionised our view of life on earth (Van Dover 2000). It is now widely known that these ecosystems, which depend on simple organic molecules such as hydrogen sulfide and methane as the energy source, can support dense populations of macrobenthos, with productivities rivaling those of photosynthesis-based tropical forest ecosystems (Tunnicliffe 1991; Van Dover 2000; Levin 2005; Cordes et al. 2007). Besides the common features of high pressure, darkness and low temperature in the deep-sea, hydrothermal vents and methane seeps are often characterised by hypoxia and high concentrations of toxic substances such as hydrogen sulfide and heavy metals.

Another drastic difference between the fauna of deep-sea chemosynthetic ecosystems and their shallow-water counterparts is their dependence on chemotrophic (including and methanotrophic) microbes (Van Dover 2000; Dubilier et al. 2008; Moya et al. 2008). These microbes can utilise simple organic molecules such as hydrogen sulfide and methane as a source of energy to fuel the ecosystem in the absence of solar energy. Many of the dominant macrobenthos in these chemosynthesis-based ecosystems harbour symbiotic bacteria that also provide nutrients to the host. For instance, the giant tubeworm *Riftia pachyptila* harbours the endosymbiotic bacterium *Candidatus Endoriftia persephone* which is a sulfide-oxidizing gamma proteobacterium in a special structure called trophosome (Robidart et al. 2008). The giant clams *Calyptogena* spp. harbour the sulfur-oxidizing endosymbiotic bacterium *Candidatus Ruthia magnifica* in their gill epithelial cells (Newton et al. 2007). The Yeti crab *Kiwa puravida* farms epibiotic bacteria on the fine hairs of its chelipeds (claws) (Thurber et al. 2011). Characterisation of these symbiont-host relationships, including how the hosts recruit the bacteria and control their population, and how the host and the symbiont benefit from each other, is the key to understand the success of these macrobenthos in the special chemosynthesis ecosystems.

Deep-sea mussels (Mytilidae, Bathymodiolinae) are often dominant species in hydrothermal vents and cold seeps around the world, playing a key role in the food-web of these ecosystems. Due to their wide distribution and remarkable biological characteristics including an ability to survive for an extended period under atmospheric pressure (Kádár et al. 2005), these mussels have been recognised as an emerging model in studies of symbiosis (Dubilier et al. 2008; Minic 2009), immunity (Dubilier et al. 2008; Bettencourt et al. 2010; Detree et al. 2016), adaptation to abiotic stress (Boutet et al. 2009), ecotoxicology (Tanguy et al. 2008; Company et al. 2011, 2012), biogeography and evolution (Van Dover et al. 2001; Won et al. 2003b; Vrijenhoek 2010; van der Heijden et al. 2012; Johnson et al. 2013; Lorion et al. 2013; Miyazaki et al. 2013; Breusing et al. 2015; Xu et al. 2016), and discovery of molecules for biotechnological application (Thornburg et al. 2010).

Since the fossil record of cold seep mytilid mussels dates back to only about 150 million years ago (MYA) in the late Jurassic, it is clear that the modern vent and seep bathymodiolin mussels are not living fossils from the Palaeozoic (Little and Vrijenhoek 2003). However, due to the conserved morphology in mytilids, it is uncertain whether the Jurassic seep mytilids are close relatives of the modern bathymodiolins that have become widely distributed in fossil records since about 45 MYA in the middle Eocene (Kiel and Amano 2013). Molecular data from the COI gene indicate that modern vent and seep bathymodiolin mussels arose more recently (Distel et al. 2010), but the estimated onset time of bathymodiolin diversification and world-wide dispersal varied substantially from the early Miocene (ca. 21 MYA, Miyazaki et al. 2010) to early Eocene–late Cretaceous (ca. 56–94 MYA by Little and Vrijenhoek 2003; 48.2–74.3 MYA by Jones et al. 2006; and 48.6 MYA by Lorion et al. 2013).

Different from their shallow-water relatives of *Modiolus* which are symbiont free,

deep-sea mussels include members that harbour epibiotic bacteria on their gill filaments (i.e., *Benthomodiolus geokotsucola* and *Adipicola pacifica* living on sunken woods and whale bones) and that habour endosymbiotic bacteria in their gill epithelial cells (i.e., *Apicola crypta* living on whale bones, and various species of *Bathymodiolus* and *Gigantidas* inhabiting vents and seeps) (Miyazaki et al. 2010). This difference in mussel-symbiont relationship has led to the hypothesis that shallow-water mussels invaded the deep-sea sunken woods and whale carcasses before colonising cold seeps and hydrothermal vents around the world ocean (Distel et al. 2000; Little and Vrijenhoek 2003; Olu-Le Roy et al. 2007; Miyazaki et al. 2010; Lorion et al. 2013). Having whole genome sequence would allow a better estimation of the time of origin for modern bathymodiolin mussels, and a better understanding of the contribution of symbiosis to the adaption of various vent and seep animals such as giant tubeworms and giant clams. In addition, sequencing the mussel genome could facilitate biological studies of Mollusca, a large phylum with more than 85,000 extant species but only five sequenced genomes [i.e., the California sea hare *Aplysia californica*, the owl limpet *Lottia gigantea* (Simakov et al. 2013), the Pacific oyster *Crassostrea gigas* (Zhang et al. 2012a), the pearl oyster *Pinctada fucata* (Takeuchi et al. 2012; Takeuchi et al. 2016) and the octopus *Octopus bimaculoides* (Albertin et al. 2015)].

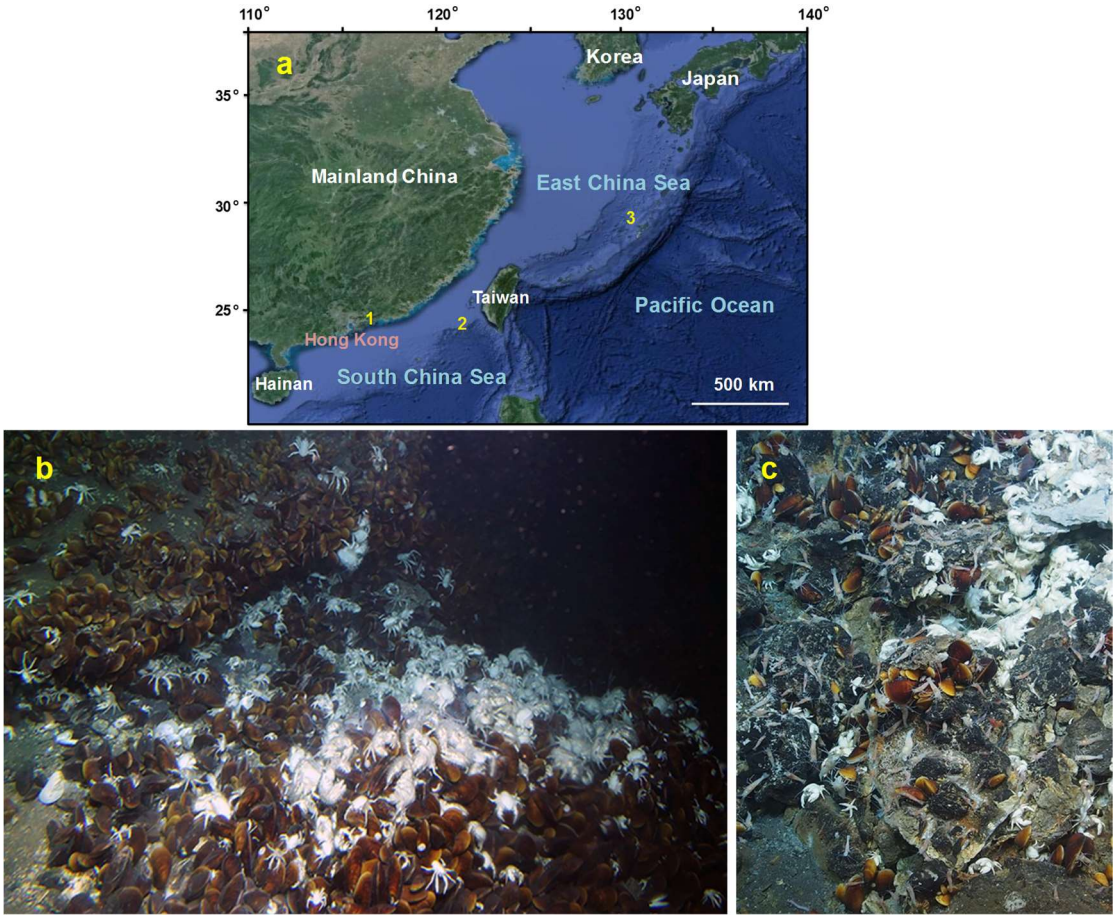


Fig. S1. Sampling sites for mussels used in the present study. *Modiolus philippinarum* was collected from an intertidal flat in Ting Kok (22° 46.989′ N, 114° 21.389′ E), Hong Kong (1). *Bathymodiolus platifrons* used for genome and transcriptome sequencing were collected from in a methane seep (22° 06.921′ N, 119° 17.131′E, 1122 m deep) in the northern South China Sea (2, b). This species is also common in hydrothermal vents in the Okinawa Trough (3, c). The conspicuous white squat lobster *Shinkaia crosnieri* was associated with *Bathymodiolus platifrons* in both (b) and (c). Photo courtesy: (b) Jian-Wen Qiu; (c) ROV Kaiko Mk IV crew.

2. Sequencing the genome of *Bathymodiolus platifrons*

Individuals of *B. platifrons* were collected using the manned submersible Jiaolong from a methane seep (22° 06.921′ N, 119° 17.131′E, 1122 m deep; Fig. S1a and b) in the South China Sea in July 2013 (Feng et al. 2015). The adductor muscle of an individual was dissected, cut into small pieces and fixed in RNAlater immediately when the submersible was brought on board the research vessel. The sample was transported to Hong Kong Baptist University (HKBU) on dry ice and stored at -80 °C until use. DNA was extracted using the CTAB method (Murray and Thompson 1980)

and DNA quality and quantity were checked with standard agarose gel electrophoresis and a Qubit Fluorometer, respectively.

To increase the DNA library diversity as well as to reduce the PCR-bias from a single library preparation, nine libraries were constructed with different insert sizes (180 bp to 16 kb) and different preparation methods (see Table S1a for details). The library named "180bp" was constructed using a PCR-free method by the Beijing Genomics Institute (BGI). The libraries with the prefix "Nx" were constructed using Illumina's Nextera Mate Pair Sample Preparation Kit. The libraries "500bp", "800bp" and "Nx16kb" were sequenced on the Illumina HiSeq 2500 platform with the rest sequenced on the Illumina HiSeq 2000 platform.

A total of 17 lanes were sequenced, including five lanes by HiSeq 2500 and 12 lanes by HiSeq 2000. The raw reads from each library were checked using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and low quality reads and sequencing adaptor contaminated reads were trimmed by Trimmomatic v0.33 (Bolger et al. 2014) with the settings of LEADING = 5, TRAILING = 5, SLIDINGWINDOW = 4:15, and MINLEN = 40. Reads from short insert-size (i.e., < 1 kb) libraries were error corrected by using Musket v1.1 with the kmer size of 21 (Liu et al. 2013).

Due to the potential chimeric sequence contamination and high PCR duplication level in mate-pair libraries prepared using the Illumina's Nextera protocol, sequencing reads were de-duplicated using FastUniq (Xu et al. 2012), and subsequently examined and sorted using NxTrim v0.3.1 which can recognize and trim the artificial Nextera mate-pair circulation adapters. Reads from category "mp" (exclusively true mate-pair reads) and "unknown" (mostly large insert size mate-pair reads) were concatenated and used for downstream scaffolding. Reads from categories "pe" (paired-end reads) and "se" (single-end reads) were discarded. The "2kb" and "5kb" mate-pair libraries

were also de-duplicated using FastUniq (Xu et al. 2012). Collectively, there were 4.15 billion clean reads from all of the libraries (i.e., short insert size libraries and mate-pair libraries) resulting in approximate 554 gigabases (Gb) of raw reads.

3. Sequencing the genome of *Modiolus philippinarum*

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, Fig. S1a) during a 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 morphological description (Lee and Morton 1985). DNA from the adductor muscle was extracted using DNeasy's Blood & Tissue Kit (Qiagen, Hilden Germany).

A total of six DNA libraries with insert sizes ranging from 180 bp to 5 kb were constructed and sequenced on 13 lanes, including five lanes by a Hiseq2500 and seven lanes by a Hiseq2000 (see Table S1b for details).

The raw reads were checked using FastQC and trimmed using Trimmomatic v0.33 (Bolger et al. 2014) with the same settings described in section 1. Reads from short insert-size (i.e., < 1 Kb) libraries were error corrected using Musket v1.1 with the kmer size of 21 (Liu et al. 2013). Reads from the 3 Kb and 5 Kb mate-pair libraries were de-duplicated with FastUniq (Xu et al. 2012), resulting in 4.52 billion clean reads with a total of 510 Gb nucleotides.

Table S1. Summary of data generated in the genome sequencing projects.

a. *Bathymodiolus platifrons*

Library	Insert size (bp)	Standard deviation (bp)	Platform	Reads No. (million)	Read length	Coverage (X)	Physical coverage (X)
180bp	170	15	Hiseq2000	2168.1	100	129.6	112.6
500bp	466	22	Hiseq2500	465.9	250	67.2	66.3
800bp	783	51	Hiseq2500	417.8	250	58.0	100.0
2kb	2258	238	Hiseq2000	481.1	100	28.7	331.9
Nx3kb	2684	359	Hiseq2000	121.3	101	6.4	99.5
Nx5kb	4580	600	Hiseq2000	102.7	101	5.6	143.7
5kb	4881	658	Hiseq2000	195.7	100	11.1	291.9
Nx10kb	9565	1873	Hiseq2000	87.0	101	4.7	254.1
Nx16kb	15784	2922	Hiseq2500	109.6	160	7.9	528.6
		Total				319.0	1928.7

b. *Modiolus philippinarum*

Library	Insert size (bp)	Standard deviation (bp)	Platform	Reads No. (million)	Read length	Coverage (X)	Physical coverage (X)
180bp	164	28	Hiseq2000	1312.4	100	54.0	45.2
500bp_1	403	115	Hiseq2000	656.3	100	26.9	55.5
500bp_2	458	73	Hiseq2500	1464.2	125	75.5	140.7
800bp	783	55	Hiseq2500	846.4	125	43.3	139.1
3kb	3042	680	Hiseq2000	133.8	100	5.5	85.4
5kb	4897	801	Hiseq2000	108.2	100	4.3	111.2
		Total				209.5	577.1

4. Characteristics of the two mussel genomes

The genome size, heterozygosity and repeat content of the two mussel genomes were determined by analysing the kmer histograms generated from the short insert-size libraries using Platanus v1.21 (Kajitani et al. 2014) and SGA-preqc (Simpson 2014). The 17-mer histograms (Fig. S2) showed that both genomes were highly heterozygous and repetitive. Dividing the total number of effective kmer (total number of kmer – total number of errorous kmer) by the number of homo-peak showed that the genome size of *B. platifrons* and *M. philippinarum* was 1.63 Gb and 2.21 Gb, respectively. Repeating the calculations using a longer kmer size of 32 (which is also used in the following contig assemblies) led to a similar result of 1.66 Gb for the *B. platifrons*

genome and 2.32 Gb for the *M. philippinarum* genome. Applying SGA-preqc (Simpson 2014), based on a kmer size of 31, produced a genome size of 1.64 Gb for *B. platifrons* and 2.38 Gb for *M. philippinarum*, which are similar to the corresponding values obtained using Platanus v1.21. The genome sizes generated using SGA-preqc (Simpson 2014) were used for characterisation hereafter as this method has incorporated sequencing error correction.

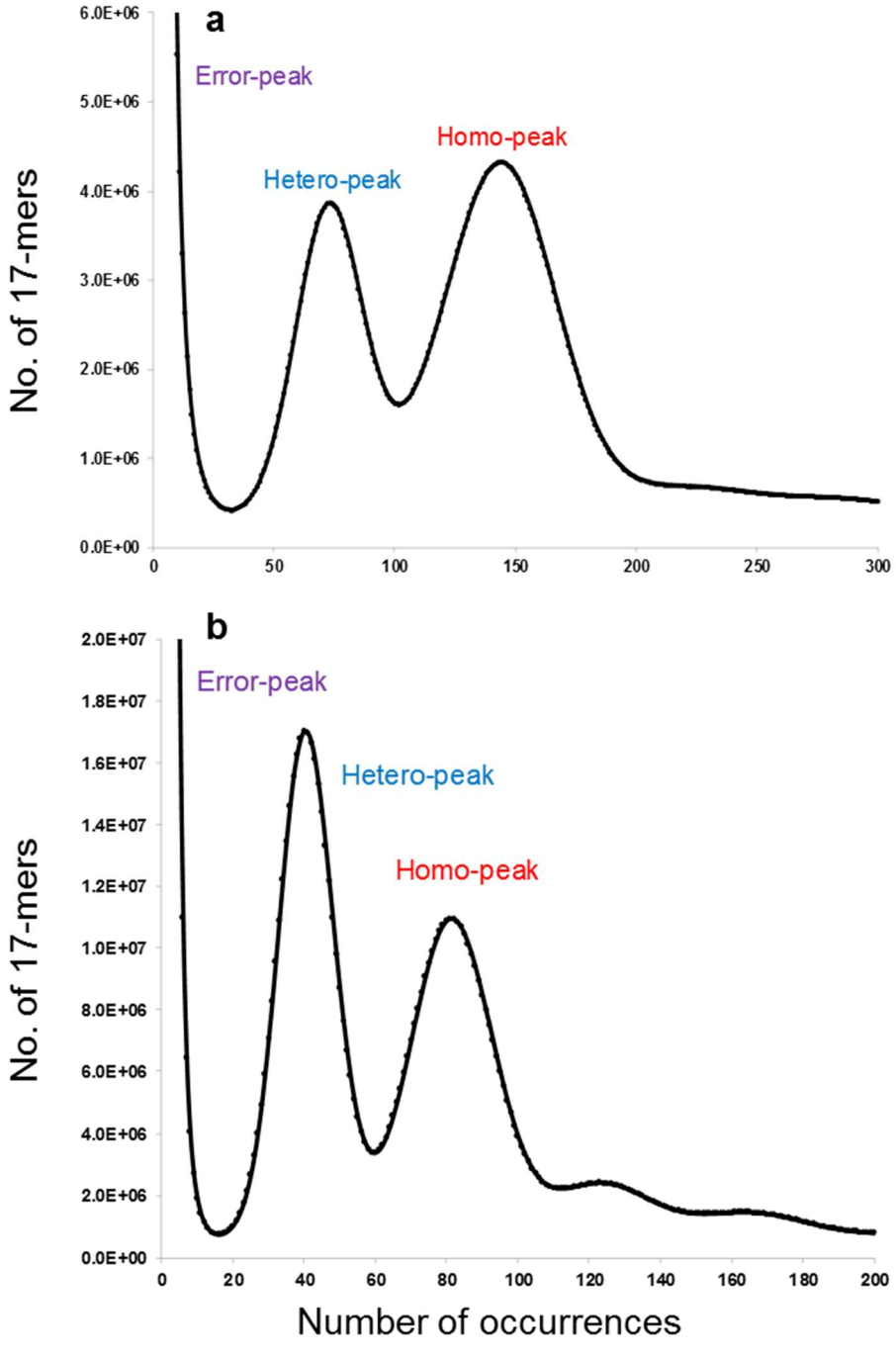


Fig. S2. Kmer histogram for *B. platifrons* (a), and *M. philippinarum* (b).

In Fig. S2, the kmer histogram for both species has a hetero-peak and a homo-peak (Kajitani et al. 2014), which conform to negative binomial distribution. Genome-wide heterozygosity rate was estimated as *(hetero_kmers/2k) / (hetero_kmers/2 + homo_kmers)*, where k is the kmer size (Vij et al. 2016).

Calculations using different kmer sizes (i.e., 17 and 32) showed that the genome-wide heterozygosity rate was 1.24% for *B. platifrons* and 2.02% for *M. philippinarum*. Comparison with other bivalve genomes showed that the heterozygosity of the Pacific oyster *C. gigas* (Zhang et al. 2012a) was 1.95%, and that of the Zhikong scallop *Chlamys farreri* (Jiao et al. 2013) was 1.77%. Using the SGA-preqc (Simpson 2014) method produced similar heterozygosity values, with *B. platifrons* having the lowest heterozygosity while *M. philippinarum* having the highest heterozygosity among the four bivalves with substantial genomic resources (Fig. S3). The lowest heterozygosity rate in *B. platifrons* could be due to recurrent population bottlenecks as a result of population extinction and re-colonisation, as suggested for *Bathymodiolus* mussels distributed at seeps in the Gulf of Mexico (Faure et al. 2015). Nevertheless, these four bivalve genomes are much more heterozygous than the human genome.

Applying SGA-preqc (Simpson 2014) with different kmer sizes showed that the repeat content was high in both *B. platifrons* and *M. philippinarum*. Between the two species, *M. philippinarum* had higher repeat content than *B. platifrons*, which is consistent with the bigger genome size of the former species (Fig. S3). More details of the contribution of repeats to the genome size can be found in section 10.

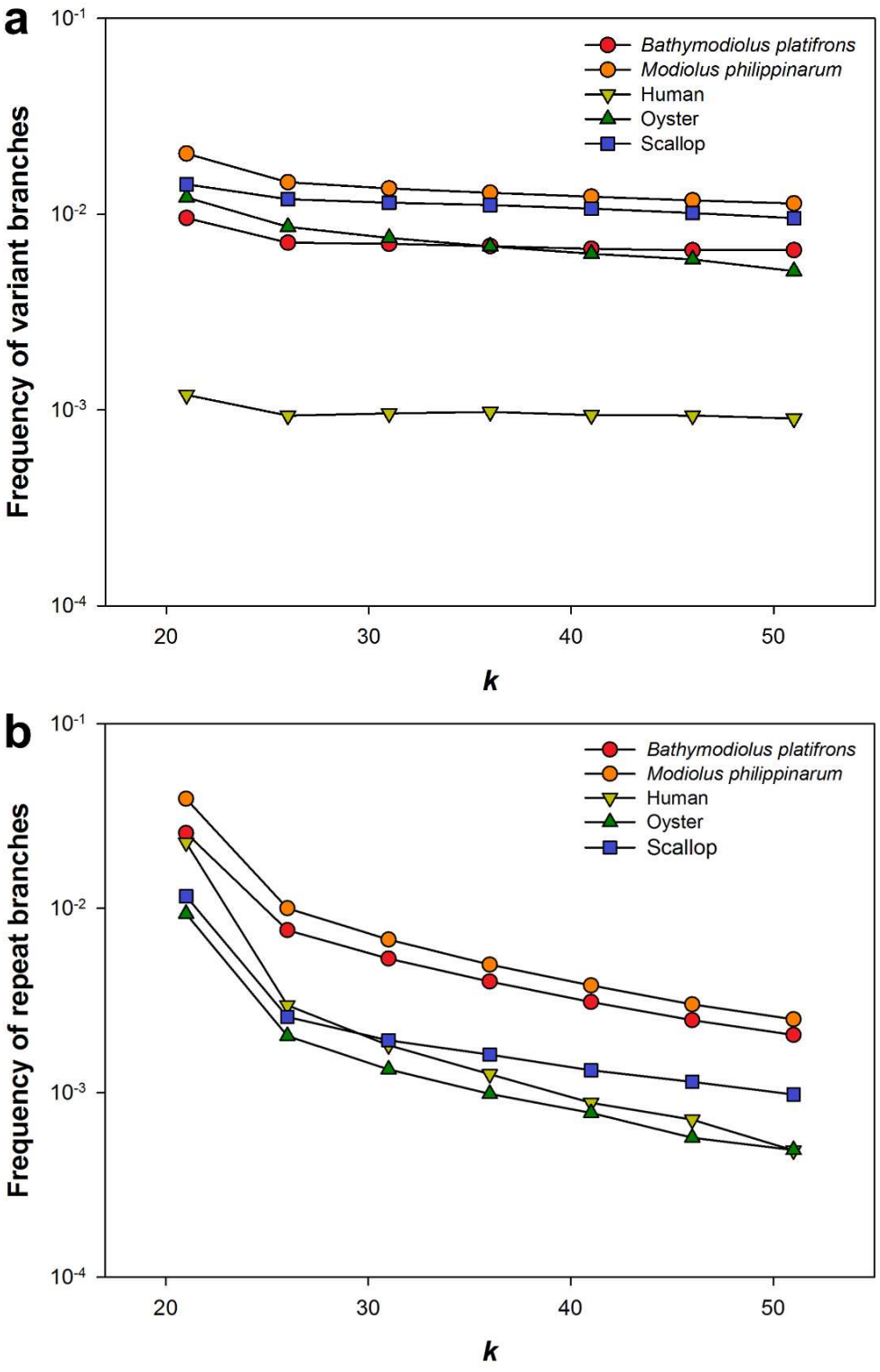


Fig. S3. Comparison of genome repeat contents among the two mussels, oyster, scallop, and human. Estimated variation branch rate (a) and repeat branch rate (b) for each genome as a function of kmer size between 21–51 in increments of 5. Human: *Homo sapiens*; Oyster: *Crassostrea gigas*; Scallop: *Chlamys farreri*.

5. Transcriptome sequencing

Transcriptome sequencing (RNA-Seq) was conducted for the two species of mussels to aim genome assembly (i.e., scaffolding contigs) and to annotate the genomes (i.e., gene prediction). For *B. platifrons*, both published (Wong et al. 2015) and new transcripts generated from RNA-Seq in this study were used. The mantle, foot and gill transcriptomes have been reported (Wong et al. 2015). The adductor muscle, visceral mass and ovary tissues were dissected on board, the tissue samples were fixed in RNAlater, the RNA was extracted using Trizol, and the samples were sent to BGI, Shenzhen for sequencing using a eukaryotic mRNA enrichment and library preparation protocol. For *M. philippinarum*, RNA samples were extracted from mantle, foot, gill, adductor muscle and visceral mass using Trizol and sent to BGI, Shenzhen for sequencing. The RNA-Seq data are summarised in Table S5.

Raw reads were trimmed using Trimmomatic v0.33 (Bolger et al. 2014) and assembled using Trinity (Grabherr et al. 2011). Due to the high redundancy in the Trinity output, all clean reads were mapped to the assembled contigs and the most abundant isoforms were retained according to the Trinity manual (Grabherr et al. 2011). CDHIT-EST (Li and Godzik 2006) was then applied to further absorb those short contigs which showed > 95% sequence similarities to the remaining contigs.

Table S2. A list of RNA-seq samples used for the transcriptome sequencing. Adapter contamination and low quality raw Illumina reads were trimmed to produce clean reads.

Abbreviation	Description	Clean read pairs	Read length (bp)
Bathymodiolus platifrons			
AM	Adductor muscle	15,627,245	150
F	Foot	25,625,920	90
O	Ovary	13,798,156	150
G	Gill	25,967,758	90
M	Mantle	26,901,035	90
VM	Visceral mass	15,640,503	150
Modiolus philippinarum			
AM	Adductor muscle	15,063,341	100
F	Foot	19,023,474	100
G	Gill	18,632,229	100
M	Mantle	18,933,368	100
VM	Visceral mass	30,054,038	100

6. Genome assembly

A strategy involving high sequencing coverage and constructing genomic libraries with different insert sizes was employed to assemble the mussel genomes with high heterozygosity and high repeat contents. In total, the sequences *B. platifrons* and *M. philippinarum* covered 319 X and 209.5 X of the genome size, respectively. The genomes were assembled using Platanus v1.21, which is appropriate for assembly of highly heterozygous eukaryote genomes (Kajitani et al. 2014). A Dell server with Intel(R) Xeon(R) CPU E7-4860 v2 2.60 GHz processors and 1.5 TB RAM was used for genome assembly.

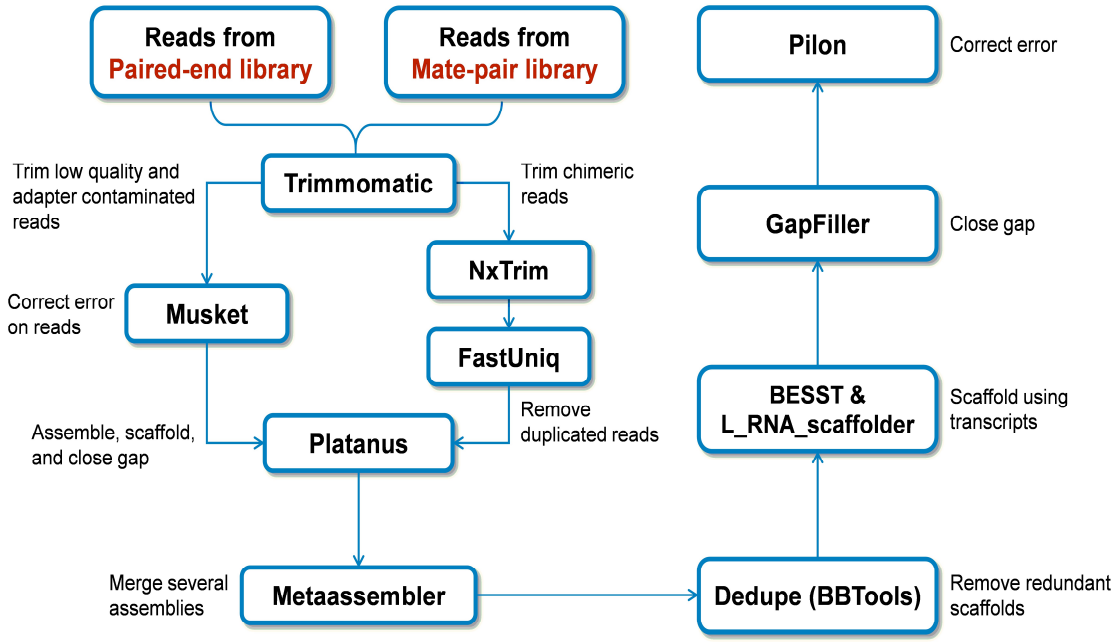


Fig. S4. Analytical pipeline for the genome assembly.

The genome assembly pipeline involved assembly of DNA sequences using different parameters, merging results from different assemblies, removing redundant scaffolds, further scaffolding using transcripts, gap filling and error correcting (Fig. S4). Briefly, the clean and error corrected short insert-size library reads were assembled using Platanus v1.21 (Kajitani et al. 2014) with the following settings: -k 32, -t 29, -s 10, -d 0.3 and -u 0.2 for *B. platifrons* and -u 0.3 for *M. philippinarum* to account for their different heterozygosity rates as suggested by the user manual. The insert sizes of all of the libraries were estimated during the first scaffolding step. Then, the scaffolding was re-run by changing the "minimum insert size" and "SD of insert size" for all of the mate-pair library reads according to the estimated values. The "gap_close" function was applied to fill the gaps between contigs with all of the short insert-size library reads and "mp" reads from the Nextera mate-pair libraries.

For the *B. platifrons* genome, Metassembler (Wences and Schatz 2015) was applied to obtain consensus scaffolds after merging three versions of the former Platanus v1.21 (Kajitani et al. 2014) assemblies with three different settings. According to the manual of Metassembler, the "2kb" mate-pair library was used as an internal competition

322 reference to help select the best assembly. Then, the genome scaffold and contig N50
323 reached 332 kb and 9.6 kb, respectively; and the total scaffold size was 1.69 Gb.
324 BUSCO (Benchmarking Universal Single-Copy Orthologs) was applied to assess the
325 assembly completeness using 843 universal metazoan single copy orthologs (Simão et
326 al. 2015), which showed that there were 87% complete single copy gene models (C),
327 2.3% complete duplicate gene models (D), 8.4% fragmented gene models (F) and 3.9%
328 missing gene models (M).
329
330 For the *M. philippinarum* genome, there was only one assembly from Platanus v1.21
331 (Kajitani et al. 2014), with a scaffold of N50 62.4 kb and a contig N50 of 9.6 kb. Its
332 total scaffold size was 3.27 Gb. Applying BUSCO (Simão et al. 2015) revealed 59%
333 C, 3.5% D, 26% F and 13% M.
334
335 The assembled scaffolds of these two genomes were larger than their genome sizes
336 predicted by kmer analysis, especially for the more heterozygous *M. philippinarum*
337 genome. Therefore BBTools dedupe (https://sourceforge.net/projects/bbmap/) was
338 applied to further remove the redundant scaffolds due to heterozygosity, and to
339 explore a series of settings and monitor the resultant scaffolds by BUSCO (Simão et
340 al. 2015) to determine the appropriate parameters. A similar strategy was applied in
341 the post-assembly of another highly heterozygous genome (Cong et al. 2015).
342
343 The scaffolds were extended by incorporating the RNA-Seq data using BESST
344 (Sahlin et al. 2014) and L_RNA_scaffolder (Xue et al. 2013) following the authors'
345 instructions. The scaffolds were then gap filled using GapFiller (Boetzer and Pirovano
346 2012). Afterwards, all of the short insert-size reads and mate-pair reads with a fixed
347 insert size range determined by Platanus v1.21 (Kajitani et al. 2014) were mapped to
348 the scaffolds using Bowtie 2 (Langmead and Salzberg 2012), and Pilon (Walker et al.
349 2014) was applied to analyse the mapping results and correct the erroneous bases, fix
350 the small indels and correct the mis-assemblies. The resultant scaffold parameters are

listed in Table S2. These post-assembly steps have led to significant improvement in BUSCO (Simão et al. 2015) results (Table S4), especially for the *M. philippinarum* genome assembly. Comparison with other available lophotrochozoan genomes showed that the two assembled mussel genomes are of good quality.

Table S3. Comparison between the assembled genomes of *B. platifrons* and *M. philippinarum* and other six lophotrochozoan genomes sequenced by next generation sequencing technology. Abbreviations: Bpl, *Bathymodiolus platifrons*; Mph, *Modiolus philippinarum*; Pfu, *Pinctada fucata*; Cgi, *Crassostrea gigas*; Bgl, *Biomphalaria glabrata*; Aca, *Aplysia californica*; Obi, *Octopus bimaculoides*; and Lan, *Lingula anatina*.

	Bpl	**Mph**	**Cgi**	**Pfu**	**Aca**	**Bgl**	**Obi**	**Lan**
Genome size	1.64G	2.38G	545Mb	1.14G	1.8G	931M	2.7G	463M
GC content	34.3%	34.0%	33.4%	35.1%	40.3%	36.0%	36.0%	36.4%
Total scaffolds assembled	1.66G	2.63G	558M	815M	927M	916M	2.37G	425M
No. of scaffolds	65,664	74,575	7658	29,306	4331	331,539	379,696	3830
No. of scaffolds > 10Kb	9,998	39,195	3,167	8,502	2,368	11,102	9,987	2,854
Longest scaffold	2.79M	715.4Kb	1.96M	1.16M	6.10M	2.18M	4.06M	1.77M
Unknown sequences in scaffold	11.77%	4.84%	11.81%	6.71%	20.44%	1.91%	14.98%	4.11%
Scaffold N50/NG50	343.4Kb / 350.0Kb	100.2Kb / 110.3Kb	401.7Kb / 412.2Kb*	167Kb / 87.9Kb	917.5Kb / NA	48.1Kb / 46.0Kb	466.1Kb / 381.1Kb	294.4K/ 266.8Kb
Scaffold L50/LG50	1,261/1,228	7,755/6,581	400 / 384*	1,346 / 2,674	291 / NA	3,094 / 3,249	1,369 / 1,761	419 / 487
Contig N50/NG50	13.2Kb / 11.3Kb	19.7Kb / 20.8Kb	32.6Kb / 28.9Kb*	21.9Kb / 10.3Kb	9.6Kb / NA	16.7Kb / 15.2Kb	6.5Kb / 4.1Kb	55.1Kb / 46.4Kb

Table S4. BUSCO results for 11 available lophotrochozoan genomes. Abbreviations: C, number of complete single copy gene models in BUSCOs. D, number of complete duplicated gene models; a high value of D may indicate an erroneous assembly (Simão et al. 2015). F, number of fragmented gene models; M, number of missing gene models.

Species	C	D	F	M	Reference/Source
Capitella teleta	799 (94%)	65 (7.7%)	30 (3.5%)	14 (1.6%)	Simakov et al. 2013
Lottia gigantea	785 (93%)	32 (3.7%)	35 (4.1%)	23 (2.7%)	Simakov et al. 2013
Lingula anatina	768 (91%)	181 (21%)	43 (5.1%)	32 (3.7%)	Luo et al. 2015
Crassostrea gigas	765 (90%)	60 (7.1%)	49 (5.8%)	29 (3.4%)	Zhang et al. 2012a
Aplysia californica	761 (90%)	21 (2.4%)	36 (4.2%)	46 (5.4%)	GenBank accession no. GCA_000002075
Bathymodiolus platifrons	**751 (89%)**	**21 (2.4%)**	**60 (7.1%)**	**32 (3.7%)**	**The present study**
Octopus bimaculoides	711 (84%)	16 (1.8%)	62 (7.3%)	70 (8.3%)	Albertin et al. 2015
Pinctada fucata	709 (84%)	38 (4.5%)	78 (9.2%)	56 6.6%)	Takeuchi, et al. 2016
Modiolus philippinarum	**663 (78%)**	**35 (4.1%)**	**118 (13%)**	**62 (7.3%)**	**The present study**
Biomphalaria glabrata	620 (73%)	20 (2.3%)	106 (12%)	117 (13%)	GenBank accession no. GCA_000457365
Hellobdella robusta	618 (73%)	24 (2.9%)	95 (11%)	130 (15%)	Simakov et al. 2013

7. Checking the quality of genome assembly

Our assembled transcripts (over 600 bp in length) were mapped to the assembled genomic sequences using BLASTn with a stringent criterion of 1e-50 to further examine the completeness of the genomes. For *B. platifrons*, 99.2% (31,840 out of 32,109) of the transcriptome contigs could match to the assembled scaffolds with an average sequence similarity of 99.0%. Although the transcriptome contained 255 transcripts from symbiotic methane-oxidizing bacteria belonging to the family Methylococcaceae (Wong et al. 2015), none of these transcripts matched to our genomic scaffolds. For *M. philippinarum*, the overall matching rate was 96.1% (57,274 out of 59,598) with an average sequence similarity of 97.6%. Thus, both genome assemblies had an excellent coverage of the transcripts. Comparing our genomic scaffolds with NCBI's nt database using BLASTn with an E-value of 1e-5 led to the identification of 3 scaffolds in *B. platifrons* and 5 scaffolds in *M.*

philippinarum that showed similarity to bacterial sequences. These 8 scaffolds were all quite short (376 bp to 3400 bp only) and did not match sequences from any known deep-sea sulfur-oxidizing bacteria (SOB) or methane oxidizing bacteria (MOB), and therefore were considered contaminants and removed. Next, all of the short insert-size paired-end reads were mapped to the scaffolds, and the scaffold average GC content and reads coverage were calculated using Samtools (Li et al. 2009) (Fig. S5). In both *B. platifrons* and *M. philippinarum*, the scaffolds were each packed into a cluster. Should there be substantial bacteria contamination, there would be a cluster of reads with different GC contents (Takeuchi et al. 2012). This result also implies very little possibility of horizontal gene transfer from bacteria.

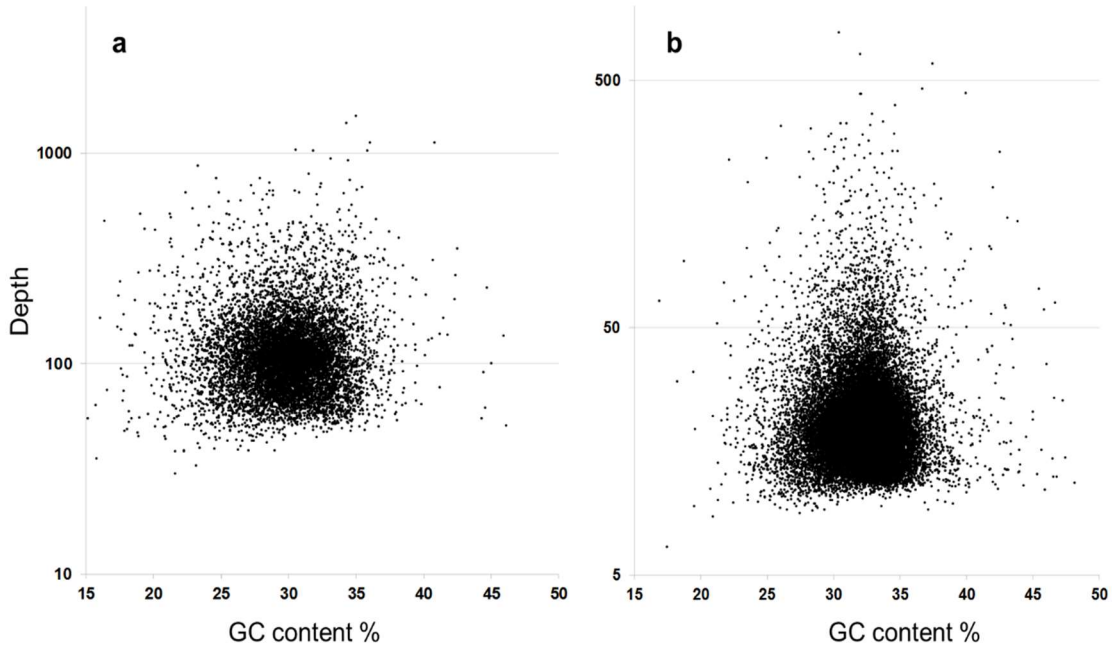


Fig. S5. Overall GC content vs. sequencing depth (times of genome size) in the genome of *Bathymodiolus platifrons* (a), and *Modiolus philippinarum* (b).

Paired-end reads from three different short insert-size libraries (i.e., 200 bp, 500 bp and 900 bp) generated from another individual of *B. platifrons* (Xu et al. 2016) were also mapped to the present new *B. platifrons* assembly using Bowtie 2 (Langmead and Salzberg 2012). The overall mapping rate was 94.54%, 95.01% and 95.04% for the

three libraries, respectively. Among the mapped reads, 88.0%, 91.4% and 94.9% read pairs respectively were mapped to the same scaffold and oriented towards each other with a reasonable insert size. Based on the mapping position in the same scaffolds (Fig. S6), the three libraries generated three peaks of insert size (i.e., 174 bp, 462 bp and 775 bp), closely corresponding to three nominal insert size of the three libraries. This result further indicated very small possibility of genome mis-assembly, therefore the assembled genomes were used in the downstream analysis.

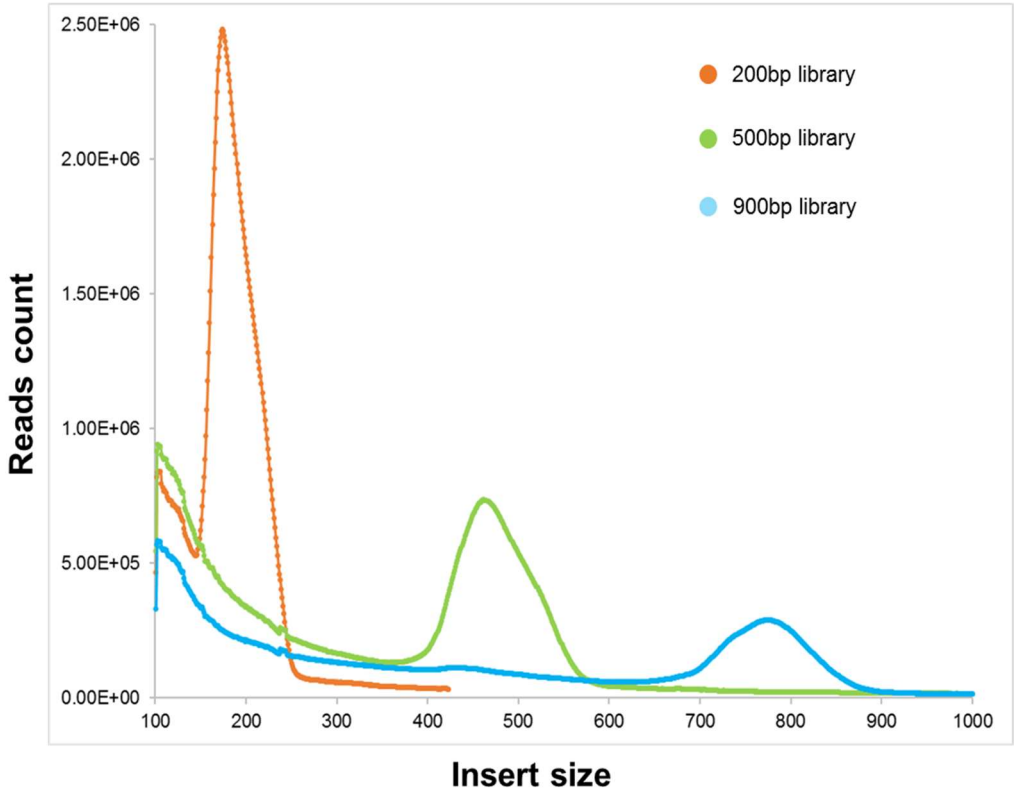


Fig. S6. Insert size distribution based on the mapping distance using paired-end reads from three libraries of another individual compared with the *B. platifrons* genome scaffolds.

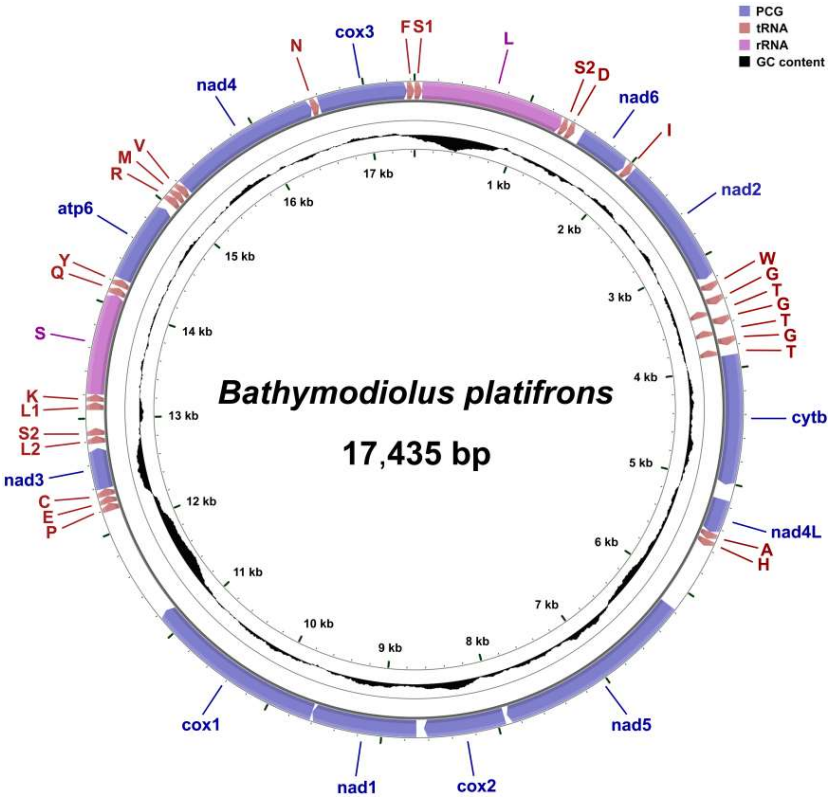
8. Mitochondrial genome assembly and characterisation

As a by-product of the nuclear genome assembly, mitochondrial geneomes (mitogenomes) of the two mussels were obtained for the first time for use in phylogenetic and evolutionary studies of Mytilidae. Currently most reported mitigenomes in the family Mytilidae have come from shallow water marine and

freshwater genera, without any from deep-sea genera (Robicheau 2016). Due to the high copy number of mitochondrial DNA and the similarity between mitochondrial sequences and partial nuclear sequences, it is not feasible to retrieve mitochondrial genomes (mitogenomes) directly from whole genome assembly (Hahn et al. 2013). Mitogenomes of *B. platifrons* and *M. philippinarum* were therefore assembled using CLC Genomics Workbench v7.5 with default settings (http://www.clcbio.com) and one lane of Hiseq reads only for each species. The sequences were blasted against several published Mytilidae mitogenomes with an *E*-value of 1e-5 using tBLASTn to extract the mitogenomes and then annotated as in Xu et al. (2015). In brief, protein-coding genes (PCGs), ribosomal RNA (rRNA) and transfer RNA (tRNA) genes using BLAST implemented in MITOS with an *E*-value of 1e-5 (Bernt et al. 2005). The boundary between each pair of protein-coding genes (PCGs) was determined and manually checked by alignment with several published mytilid mitogenomes. The tRNA genes were further verified by their putative secondary structures obtained using ARWEN v1.2. The boundary between two rRNAs was deduced based on the assumption that rRNA occupies all the space between their adjacent tRNAs. A mitogenome gene map was constructed for each species using the CGView server (Grant and Stothard 2008).

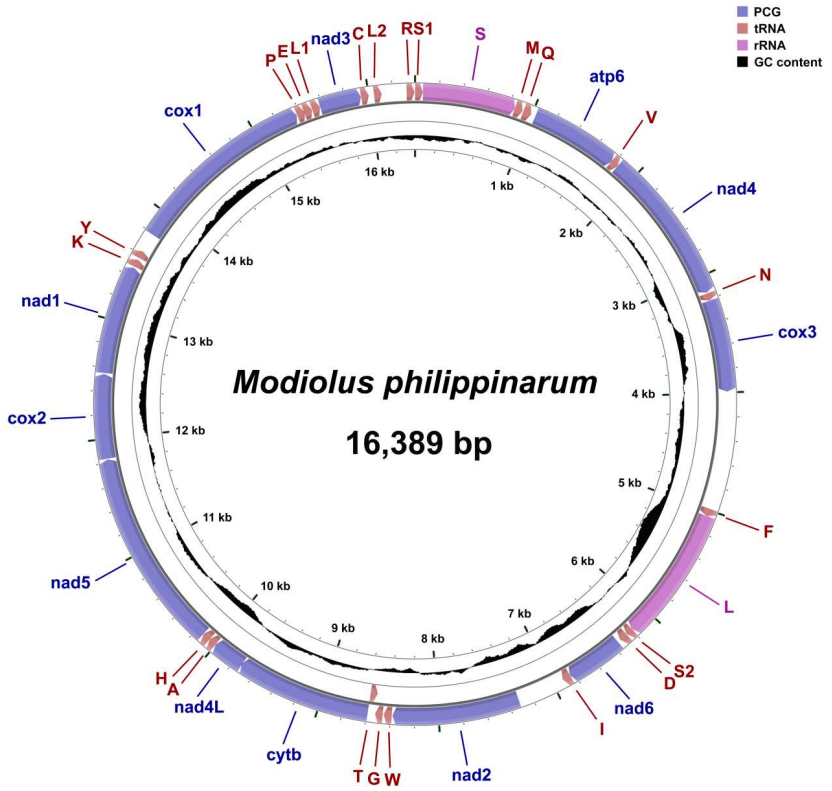
443

444 a.



445

446 b.



447

448

Fig. S7. Mitogenome gene map for *B. platifrons* (a) and *M. philippinarum* (b). Purple, red and pink arrows represent the locations of PCGs, tRNAs and rRNAs, respectively. The arrows show the direction of transcription. The tRNA genes are identified by the one-letter codes for their corresponding amino acids. Genes encoded on the heavy strand and light strand are displayed on the outer and middle circle respectively. The inner circle shows the average GC content of the corresponding genomic regions.

The mitogenome of *B. platifrons* was 17,435 bp in length, which contained 12 PCGs, 27 tRNAs and 2 rRNAs (Fig. S7a, Table S5a). In comparison with the 37 most common genes in metazoan mitogenomes (Boore 1999), the *B. platifrons* mitogenome had an additional *trnS2* but no *atp8*. This deviation in mitochondrial gene content has been known in other mussels, such as *Unio japanensis* [Genbank accession numbers AB055624 (M-type) and AB055625 (F-type)] and *Mytilus edulis* (Boore et al. 2004). Besides, the *B. platifrons* mitogenome had two more *trnG* genes as well as two more *trnT* genes. Five start codons (i.e., ATA, ATG, ATT, GTG and TTG) and two stop codons (i.e., TAA and TAG) were used for the PCGs. Most genes were encoded on the heavy strand, except for three *trnT*s which all were encoded on the light strand. In the *B. platifrons* mitogenome, three gene borders had 1 bp overlap (i.e., *nad4L* and *trnA*; *nad4* and *trnN*; *cox3* and *trnF*). There were and two large non-coding regions (NRs), one located between *trnH* and *nad5* (599 bp) with 68.4% A+T, and another between *cox1* and *trnP* (994 bp) with 73.8% A+T. The remaining 29 NRs were small, ranging from 1 to 156 bp in length.

The mitogenome of *M. philippinarum* was 16,389 in length, consisting of 12 PCGs, 22 tRNAs and 2 rRNAs (Fig.S7b, Table S5b). The mitogenome of *M. philippinarum* had no *atp8* either. ATA, ATG and ATT were used as the start codons, and TAA and TAG were used as the common stop codons. However, *nad4L* used T as the stop codon. The TAA stop codon is known to be further completed by adding 3′ A residues to the mRNA during translation (Mizi et al. 2005). All genes, except for *trnT*, were encoded on the heavy strand. In the *M. philippinarum* mitogenome, three were three pairs of genes with overlapped regions, including 2 bp overlap between *atp6* and *trnV*,

4 bp overlap between *cytb* and *nad4L*, and 3 bp overlap between *trnP* and *trnE*. There were and two large non-coding regions (NRs), one located between *trnI* and *nad2* (405bp) with 71.9% A+T, and another between *cox3* and *trnF* (996 bp) with 69.0% A+T. The remaining 23 NRs were small, ranging from 1 to 209 bp in length.

In most animal taxa, the relative positions of the larger PCGs and rRNA genes are more conserved than the shorter tRNA genes (Boore 1999). However, the mtgenomes of some animal taxa, such as Brachiopoda (Endo et al. 2005) and Mollusca (Boore et al. 2004), exhibit substantial gene order rearrangements, and such information can be used in phylogentic analysis. The gene arrangements of rRNA and tRNA genes of the deep-sea mussel *B. platifrons* and the shallow water mussel *M. philippinarum* were quite different, but their order of the 12 PCGs was conserved (i.e., *cox1-nad3-atp6-nad4-cox3-nad6-nad2-cytb-nad4L-nad5-cox2-nad1*). Their order of PCGs was identical to that of the F-type *Modiolus modiolus* mitogenome, although *M. modiolus* had an extra *atp8* gene (Robicheau 2016). However, the PCGs arrangements in *Bathymodiolus* and *Modiolus* were quite different from those of other 10 mytilid mitogenomes (Uliano-Silva et al. 2016). For instance, five species of *Mytilus* have an identical PCGs order (Uliano-Silva et al. 2016), but they only share three small synteny blocks with *B. platifrons* and *M. philippinarum*: *nad4-cox3*, *nad4L-nad5* and *cox2-nad1*. The freshwater golden mussel *Limnoperna fortune* (Uliano-Silva et al. 2016) shares the closest synteny pattern with *B. platifrons* and the two species of *Modiolus*, with two gene blocks: *cytb-nad4L-nad5-cox2* and *nad3-atp6-nad4*. Together, these results showed that the order of PCGs had not changed during the divergence between *Modiolus* and *Bathymodiolus*, but their gene order was quite different from other genera/species of mytilids. However, more sampling of other genera of deep-sea mussels is required to understand the evolution of mitogenomes.

Doubly uniparental inheritance (DUI) is a special type of mtDNA inheritance found in some bivalves including Mytilidae (Mizi et al. 2005). In DUI, paternally (M-type

genome) and maternally (F-type genome) transmitted mtDNAs produce two different sex-associated mtDNA lineages. Both M- and F-type mitogenomes have been reported in *Modiolus modiolus* (Robicheau 2016). The gene content and the protein-coding gene order of the *M. philippinarum* mitogenome were very similar to those of the *M. modiolus* F-type mitogenome. However, it should be noted that 1) A previous study revealed no heteroplasmy, a sign of DUI, in the mitochondrial sequences of *Bathymodiolus* spp. (Jones et al. (2006); 2), the M-type mitogenome in female individuals is usually lost during development in many mytilids (Robicheau 2016); and 3), in our study, only one individual of both species was used in genome sequencing. Therefore, whether the two species used in our study harbour DUI requires further investigation. Nevertheless, the mitogenomes of *B. platifrons* and *M. philippinarum* will be useful in phylogenetic studies and will also contribute to a better knowledge of the evolution of DUI in mytilids.

525 Table S5a. Characteristics of *Bathymodiolus platifrons* mitogenome

Seq name	Type	Start	End	Length (bp)	Intergenic nucleotides	Start/Stop codon	Strand
trnS1(gct)	tRNA	1	66	66	0		+
rrnL	rRNA	67	1319	1253	0		+
trnS2(tga)	tRNA	1320	1380	61	5		+
trnD(gtc)	tRNA	1386	1450	65	65		+
nad6	PCG	1516	1974	459	10	TTG/TAA	+
trnI(gat)	tRNA	1985	2050	66	3		+
nad2	PCG	2054	3187	1134	44	ATG/TAA	+
trnW(tca)	tRNA	3232	3297	66	58		+
trnG(tcc)	tRNA	3356	3426	71	7		+
trnT(tgt)	tRNA	3434	3497	64	42		-
trnG(tcc)	tRNA	3540	3610	71	7		+
trnT(tgt)	tRNA	3618	3681	64	42		-
trnG(tcc)	tRNA	3724	3795	72	8		+
trnT(tgt)	tRNA	3804	3867	64	13		-
cytb	PCG	3881	5017	1137	131	ATT/TAG	+
nad4L	PCG	5149	5451	303	-1	ATA/TAG	+
trnA(tgc)	tRNA	5451	5515	65	3		+
trnH(gtg)	tRNA	5519	5581	63	599		+
nad5	PCG	6181	7905	1725	27	ATG/TAA	+
cox2	PCG	7933	8637	705	66	ATG/TAG	+
nad1	PCG	8704	9621	918	2	ATG/TAG	+
cox1	PCG	9624	11189	1566	994	GTG/TAG	+
trnP(tgg)	tRNA	12184	12251	68	0		+
trnE(ttc)	tRNA	12252	12315	64	1		+
trnC(gca)	tRNA	12317	12383	67	1		+
nad3	PCG	12385	12741	357	41	GTG/TAA	+
trnL2(taa)	tRNA	12783	12848	66	4		+
trnS2(tga)	tRNA	12853	12917	65	156		+
trnL1(tag)	tRNA	13074	13142	69	4		+
trnK(ttt)	tRNA	13147	13209	63	0		+
rrnS	rRNA	13210	14095	886	0		+
trnQ(ttg)	tRNA	14096	14162	67	11		+
trnY(gta)	tRNA	14174	14238	65	5		+
atp6	PCG	14244	14960	717	60	GTG/TAA	+
trnR(tcg)	tRNA	15021	15088	68	0		+
trnM(cat)	tRNA	15089	15151	63	0		+
trnV(tac)	tRNA	15152	15217	66	1		+
nad4	PCG	15219	16523	1305	-1	GTG/TAG	+
trnN(gtt)	tRNA	16523	16590	68	3		+
cox3	PCG	16594	17370	777	-1	ATG/TAA	+

trnF(gaa)	tRNA	17370	17434	65	1		+

526

527 Table S5b Characteristics of *Modiolus philippinarum* mitogenome

Seq name	Type	Start	End	Length (bp)	Intergenic nucleotides	Start/Stop codon	Strand
trnS1(gct)	tRNA	1	66	66	0		+
rrnS	rRNA	67	848	782	1		+
trnM(cat)	tRNA	850	915	66	8		+
trnQ(ttg)	tRNA	924	993	70	38		+
atp6	PCG	1032	1775	744	-2	ATG/TAG	+
trnV(tac)	tRNA	1774	1838	65	0		+
nad4	PCG	1839	3143	1305	7	ATT/TAA	+
trnN(gtt)	tRNA	3151	3214	64	1		+
cox3	PCG	3216	3992	777	996	ATT/TAA	+
trnF(gaa)	tRNA	4989	5053	65	0		+
rrnL	rRNA	5054	6195	1142	0		+
trnS2(tga)	tRNA	6196	6255	60	4		+
trnD(gtc)	tRNA	6260	6326	67	27		+
nad6	PCG	6354	6842	489	0	ATG/TAG	+
trnI(gat)	tRNA	6843	6911	69	405		+
nad2	PCG	7317	8384	1068	6	ATA/TAG	+
trnW(tca)	tRNA	8391	8458	68	8		+
trnG(tcc)	tRNA	8467	8530	64	0		+
trnT(tgt)	tRNA	8531	8591	61	1		-
cytb	PCG	8593	9729	1137	-4	ATT/TAG	+
nad4L	PCG	9726	10002	277	0	ATA/T	+
trnA(tgc)	tRNA	10003	10067	65	0		+
trnH(gtg)	tRNA	10068	10135	68	1		+
nad5	PCG	10137	11828	1692	6	ATG/TAA	+
cox2	PCG	11835	12545	711	8	ATG/TAA	+
nad1	PCG	12554	13471	918	10	ATG/TAA	+
trnK(ttt)	tRNA	13482	13546	65	17		+
trnY(gta)	tRNA	13564	13627	64	161		+
cox1	PCG	13789	15375	1587	5	ATA/TAA	+
trnP(tgg)	tRNA	15381	15445	65	-3		+
trnE(ttc)	tRNA	15443	15509	67	5		+
trnL1(tag)	tRNA	15515	15580	66	3		+
nad3	PCG	15584	15934	351	3	ATG/TAG	+
trnC(gca)	tRNA	15938	16000	63	45		+
trnL2(taa)	tRNA	16046	16110	65	209		+
trnR(tcg)	tRNA	16320	16384	65	5		+

528

9. Identification of repeats and transposable elements

Repeats and transposable elements were identified using the RepeatModeler pipeline v1.0.8 (Smit and Hubley 2008) that includes the three repeat identifiers RECON, RepeatScout and TRF (Benson 1999). RMblast was used to search and construct a species-specific repeats library of the genome; and RepeatMasker (http://www.repeatmasker.org/) was used to identify, classify and mask the repeats with the species-specific repeats library and repeats in the RepBase (Kapitonov and Jurka 2008). Repeats were counted as one if they were split into fragments by an indel.

The genomes of both mussels were highly repetitive, with 47.9% of the *B. platifrons* genome and 62.0% of the *M. philippinarum* genome being repeats (Table S6). Among the repeats identified, a large proportion was unknown (i.e., 20.8% in *B. platifrons* and 27.1% *M. philippinarum*). The DNA transposons, retrotransposons and SINEs accounted for 10.4%, 14.5% and 1.41% of the *B. platifrons* genome, and 22.1%, 10.9% and 0.34% of the *M. philippinarum* genome, respectively. The larger repeats content in *M. philippinarum* had contributed to its larger genome size, which was 340.7Mb bigger than that of *B. platifrons*. In particular, the Helitron DNA transposons showed significant difference between the two mussels, with the abundance in *B. platifrons* being 340.7Mb smaller than that in *M. philippinarum*. The Helitron transposable elements and unclassified repeats together contributed 707 Mb of the total of 740 Mb difference between the two genomes.

555 Table S6. Composition of repeats in the two mussel genomes. Classification of
556 transposable elements is at the superfamily level. The category of "others" includes
557 repeats match to upper class classification only and also repeats match to other minor
558 superfamilies.
559

Repeats	*Bathymodiolus platifrons*		*Modiolus philippinarum*	
	Length occupied	Percentage	Length occupied	Percentage
DNA transposons	172,923,785	10.4216	580,270,180	22.0664
Helitron	69,608,151	4.1951	410,351,674	15.6048
TcMar	26,011,893	1.5677	8,255,577	0.3139
hAT	13,693,094	0.8252	13,942,880	0.5302
Crypton	13,342,201	0.8041	9,056,060	0.3444
PIF	9,456,778	0.5699	14,849,500	0.5647
Academ	5,418,635	0.3266	3,462,527	0.1317
Maverick	5,352,629	0.3226	43,057,071	1.6374
Zator	5,172,781	0.3117	4,111,179	0.1563
MULE	2,283,021	0.1376	358,642	0.0136
Ginger	1,945,688	0.1173	1,616,868	0.0615
PiggyBac	1,371,292	0.0826	936,329	0.0356
CMC	945,214	0.0570	10,560,123	0.4016
Kolobok	760,474	0.0458	374,960	0.0143
IS3EU	341,970	0.0206	791,263	0.0301
Sola	284,118	0.0171	392,107	0.0149
P	215,409	0.0130	18,629	0.0007
Harbinger	142,338	0.0086	6,699	0.0003
Merlin	16,333	0.0010	162,394	0.0062
Dada	13,348	0.0008	134,313	0.0051
Others	16,548,418	0.9973	57,831,385	2.1992
Retrotransposons	240,519,207	14.4954	286,704,504	10.9028
LINEs				
Penelope	60,773,515	3.6626	174,633,460	6.6409
RTE	22,816,358	1.3751	15,142,383	0.5758
CR1	21,974,828	1.3244	14,200,531	0.5400
L1	17,863,938	1.0766	11,218,910	0.4266
L2	9,036,239	0.5446	5,708,943	0.2171
I	2,937,548	0.1770	8,853,603	0.3367
Jockey	458,228	0.0276	68,989	0.0026
R1	31,878	0.0019	121,523	0.0046
Tad1	11,845	0.0007	1,268,630	0.0482
R2	11,749	0.0007	10,279	0.0004
Rex	10,186	0.0006	7,627	0.0003
Others	763,732	0.0460	1,033,310	0.0393
LTRs				
Ngaro	66,024,948	3.9791	1,186,538	0.0451

Gypsy	9,080,454	0.5473	26,927,627	1.0240
DIRS	1,878,112	0.1132	1,490,442	0.0567
Copia	696,450	0.0420	1,844,492	0.0701
ERV	592,883	0.0357	5,942,845	0.2260
Pao	552,452	0.0333	4,951,746	0.1883
Others	400,379	0.0241	1,007,907	0.0383
SINEs	23,439,374	1.4126	9,043,502	0.3439
Low complexity	1,875,194	0.1130	3,253,236	0.1237
rRNA, tRNA & snRNA	67,157	0.0040	64,271	0.0024
Satellite	25,817,905	1.5560	28,209,369	1.0727
Simple repeat	10,129,150	0.6105	23,708,164	0.9016
Unknown	345,161,363	20.8018	711,525,733	27.0578
Total	**795,329,650**	**47.9321**	**1,631,694,240**	**62.0499**

10. Genome annotation

Trinity was used to generate another version of transcripts using the genome-based mode. The transcriptome data from both the *de novo* and genome-based approaches were further assembled using the PASA pipeline v2.0.2 (Haas et al. 2003) with BLAT as the aligner.

The two mussel genomes were annotated using MAKER v3.0 (Cantarel et al. 2008) which recently integrated EVidenceModeler (EVM) to enhance the predication sensitivity. MAKER v3.0 was run twice as follows in order to train the gene modelers and increase the gene predication accuracy.

Round 1

This initial run used scaffolds that were over 200 kb in length. RepeatMasker (http://www.repeatmasker.org/) was used to “hard-mask” the scaffolds by converting the repetitive sequences into unknown sequences. The integrated transcriptome contigs generated from PASA were then searched against these scaffolds using BLAST + Exonerate in MAKER v3.0 with the setting of “est2genome = 1”. Metazoan protein sequences from the Uniprot Swissprot database and bivalve proteins from NCBI, with the exclusion of mitochondrial sequences, ribosome sequences and histone sequences that are commonly used in phylogenetic and population genetic

studies, corresponding to the setting of "protein2genome = 1", were used as homology-based evidence. The former BUSCO's Augustus v3.1 (Stanke et al. 2008) gene models were used in this run for *de novo* gene model prediction. The result of this run served as future gene modeler training. The resultant gff format files were converted to fasta format files and the gene models were filtered to remove genes with single exon, genes without full coding sequences, and any gene pairs with an inter gene region less than 3 kb in length. The "bona-fide" training models were optimized with Augustus' "optimize_augustus.pl" script.

Round 2

RepeatMasker (http://www.repeatmasker.org/) was applied to "soft-mask" the two mussel genomes with the repeat libraries of all model organisms in the RepBase (Kapitonov and Jurka 2008) plus species-specific repeat libraries. Augustus was re-run using the training model generated in Round 1. EVM was applied with different weights assigned to different predication methods: "evmtrans:est2genome = 10, evmprot:protein2genome = 10, and evmab:augustus = 5". These settings have been used in annotating another genome (Cong et al. 2015). Only scaffolds longer than 3 kb and a protein length at least 50 amino acids were retained for gene predication.

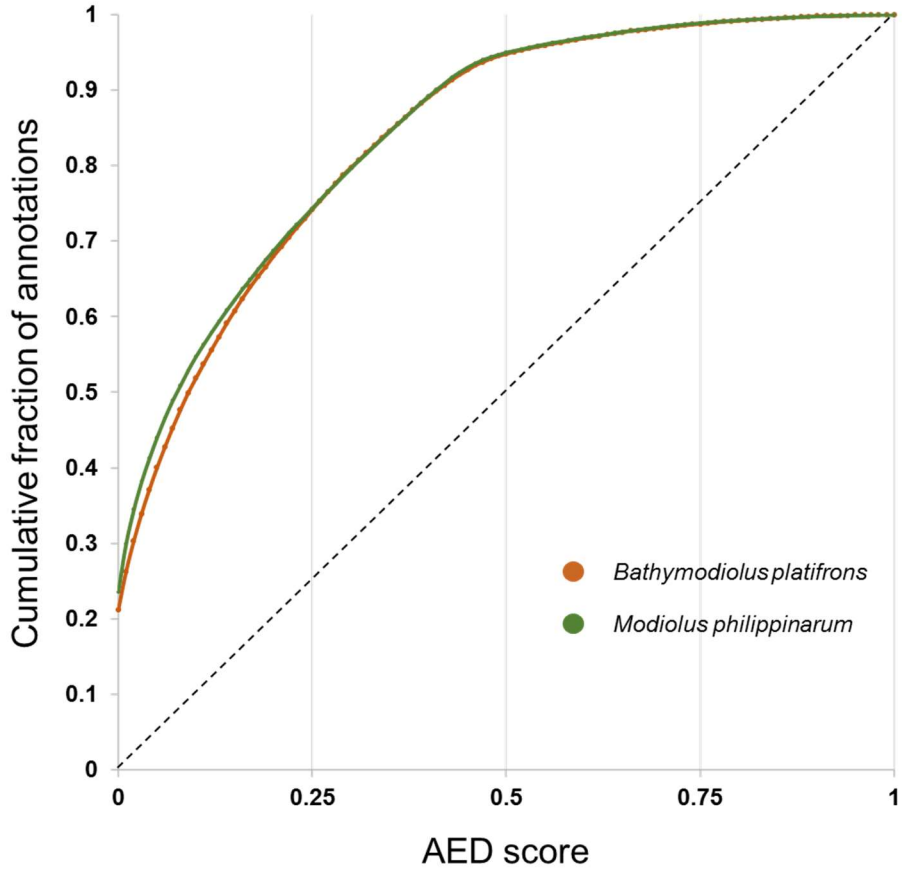


Fig. S8. Distribution of the overall annotation edit distance (AED) score. The dashed line indicates a theoretically reasonable annotation according to Maker's user manual.

Genome annotation resulted in 39,639 and 46,251 predicted genes in *B. platifrons* and *M. philippinarum*, respectively. These values were larger than the total gene numbers in most sequenced metazoans. Thus, the gene models were further filtered by running OrthoMCL (Li et al. 2003) with the protein models from six species of mollusks (i.e., *Crassostrea gigas*, *Pinctada fucata*, *Lottia gigantea*, *Aplysia californica*, *Biomphalaria glabrata* and *Octopus bimaculoides*) and retaining the sequences that could be clustered with these mollusks, as done in a recent genome assembly project (Takeuchi et al. 2016). The transcriptome reads were then mapped to the remaining "orphan" genes and only those supported by at least 10 ppm reads were retained. This filtering reduced the gene models to 33,584 in *B. platifrons*, among them 28,962 (86.2%) with transcriptome support; and 36,549 in *M. philippinarum*, among them

31,429 (86.0%) with transcriptome support.

Intron and exon sizes were calculated based on the gff gene annotation files. The mean intron and exon sizes were 2061 and 281 bp respectively for *B. platifrons*; and 2766 and 323 bp respectively for *M. philippinarum* (Table S7).

Table S7. Gene composition in the two mussel genomes.

	Bathymodiolus platifrons	*Modiolus philippinarum*
No. of protein-coding genes	33,584	36,549
No. of single-exon genes (%)	4507 (13.4%)	7603 (20.8%)
Mean exon length	281 bp	323 bp
Mean exon length of single-exon genes	770 bp	811 bp
Mean exon length of multi-exon genes	269 bp	301 bp
Total size of exons (%)	52.3 Mb (3.2%)	56.3 Mb (2.1%)
Mean No. of exons per gene	5.5	4.8
Mean intron length	2061 bp	2766 bp
Total size of introns (%)	314.2 Mb (18.9%)	380.6 Mb (14.5%)

11. Functional annotation

BLASTp was used to search the predicted mussel protein models against the whole NCBI non-redundant database with an E-value of 1e-5. The xml format files were uploaded to BLAST2GO to obtain the GO and InterprotScan annotation. The online KAAS server (Moriya et al. 2007) was used to search against the default common metazoan genomes appended with several invertebrate genomes [*Helobdella robusta* (leech), *Lottia gigantea* (limpet), *Crassostrea gigas* (oyster), *Schistosoma mansoni* (flatworm), *Nematostella vectensis* (sea anemone) and *Hydra vulgaris* (hydra)] to obtain the KEGG (Kyoto Encyclopedia of Genes and Genomes) annotation, with the BBH (bi-directional best hit) method applied in the search to increase the sensitivity. Pfam annotation was performed by searching the 16,295 Pfam entries using a hidden Markov model (HMM) (Finn et al. 2014).

A WEGO (Ye et al. 2006) diagram was plotted based on the GO annotation on these two species (Fig. S9). In the GO categories of biological adhesion, biological regulation, death, developmental process, immune system process, locomotion, multicellular organismal process and pigmentation, *B. platifrons* had significantly more genes than *M. philippinarum* (Chi-square test, $P < 0.05$), indicating possible gene family expansion in these categories in *B. platifrons*. However, in no GO categories *B. platifrons* had significantly less genes than *M. philippinarum*.

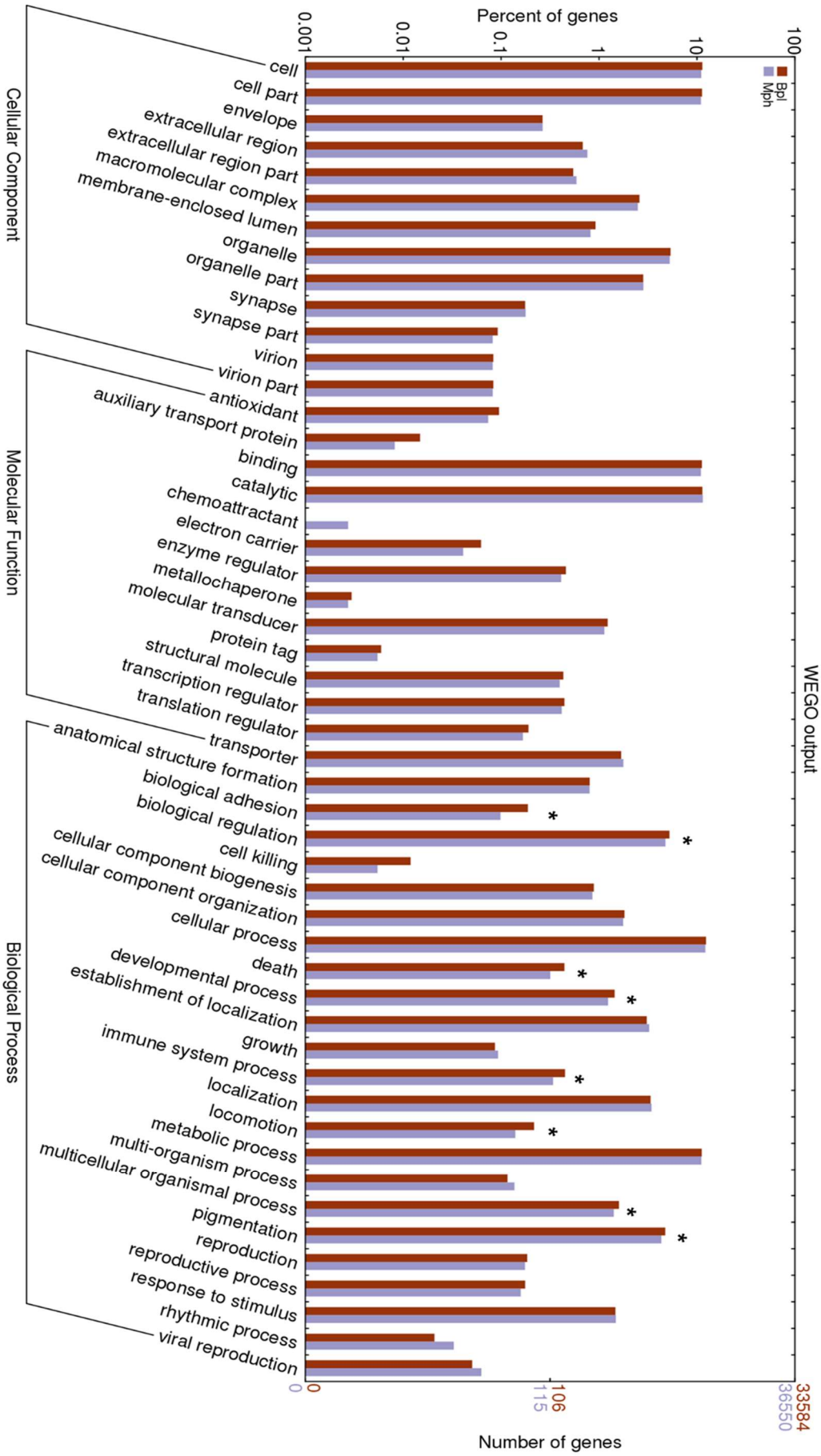

Percent of genes
0.001
0.01
0.1
1
10
100
Bpl
Mph
WEGO output
Cellular Component
Molecular Function
Biological Process
cell
cell part
envelope
extracellular region
extracellular region part
macromolecular complex
membrane-enclosed lumen
organelle
organelle part
synapse
synapse part
virion
virion part
antioxidant
auxiliary transport protein
binding
catalytic
chemoattractant
electron carrier
enzyme regulator
metallochaperone
molecular transducer
protein tag
structural molecule
transcription regulator
translation regulator
transporter
anatomical structure formation
biological adhesion
biological regulation
cell killing
cellular component biogenesis
cellular component organization
cellular process
death
developmental process
establishment of localization
growth
immune system process
localization
locomotion
metabolic process
multi-organism process
multicellular organismal process
pigmentation
reproduction
reproductive process
response to stimulus
rhythmic process
viral reproduction
0
0
106
115
33584
36550
Number of genes

Fig. S9. A WEGO diagram showing the GO category distribution in *B. platifrons* (Bpl) and *M. philippinarum* (Mph). GO category with significantly different gene number ($P < 0.05$) between the two species is labeled with an asterisk.

12. Phylogenetics

Protein sequences of the nine available lophotrochozoan genomes (Table S3) were downloaded and used for phylogenetic analysis with the two mussels. OrthoMCL (Li et al. 2003) was used to determine and cluster gene families. The bidirectional best hits (BBH) approach in BLASTp was applied with a threshold value of 1e-5. This analysis resulted in 35,057 gene families among the 11 species. Only single copy orthologues in each gene cluster were used in phylogenetic analysis. Protein sequences were firstly aligned by MUSCLE (Edgar 2004), and gaps and uncertainty alignments were further trimmed by Gblocks with a stringent threshold (Talavera and Castresana 2007). The resultant bona fide alignments were concatenated. Maximum likelihood trees were built using RAxML (Stamatakis et al. 2005) by applying the GTR + Γ model and slow bootstrapping method.

Divergent time among the mollusks was estimated using MCMCTree, which applied the Bayesian estimation method for species divergence (dos Reis & Yang 2011). 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 (Bond et al. 2014). The tree was calibrated with the following time frames to constrain the age of nodes: minimum = 305.5 MYA and soft maximum = 581 MYA for *C. teleta* and *H. robusta* (Benton et al., 2009); minimum = 168.6 MYA and soft maximum = 473.4 MYA for *A. californica* and *B. glabrata* (Benton et al., 2009); minimum = 470.2 MYA and soft maximum = 531.5 MYA for *A. californica* (or *B. glabrata*) and *L. gigantea* (Benton et al., 2009); minimum = 532 MYA and soft maximum = 549 MYA for the first appearance of the mollusks (Benton et al., 2015); and minimum = 550.25 MYA and soft maximum = 636.1 MYA for the first appearance of Lophotrochozoa (Benton et al., 2015).

To better understand the phylogenetic positions of these two mussels in the family Mytilidae, transcriptome sequences from six additional mytilid species were downloaded from NCBI, with details shown below:

Bathymodiolus azoricus transcriptome

A gill transcriptome of *B. azoricus* collected from an Atlantic hydrothermal vent was sequenced by 454 pyrosequencing (Bettencourt et al. 2010). Since the original assembly was redundant, we re-assembled the raw reads (NCBI SRA number: SRR067874) using GS *De Novo* Assembler v2.6 (Roche, CT, USA), and only the longest isotigs were retained for each isogroup. The resultant assembled transcriptome had 11,879 sequences with a N50 of 948 bp. The former assembly contained 75,407 contigs, with an average contig length of 509 bp, thus our assembly had significantly improved the assembly and reduced the redundancy.

Limnoperna fortunei transcriptome

Transcriptome sequencing from five main tissues of the freshwater golden mussel *L. fortunei* was conducted by 454 pyrosequencing (NCBI SRA number: SRR942484) (Uliano-Silva et al. 2014). The assembled contigs were deposited in NCBI TSA database. The assembly contained 64,494 sequences with a N50 of 566 bp.

Perna viridis transcriptome

Transcriptome sequencing of the tropical green mussel *P. viridis* was performed on an Illumina Hiseq2000 (Leung et al. 2014). Since the assembled contigs were not published, we re-assembled the *P. viridis* transcriptome using our transcriptome assembly pipeline described above. Raw Illumina reads (NCBI BioProject accession number of SRP043984) were downloaded and trimmed with Trimmomatic v0.33 (Bolger et al. 2014) and the sequences were assembled using Trinity (Grabherr et al. 2011). Redundancy was reduced by removing low abundance isoforms for each “gene”

and further clustered by CDHIT-EST (Li and Godzik 2006) to reach at least 95% similarity to each other. The resultant assembly contained 64,903 sequences with a N50 of 1788 bp.

Mytilus transcriptomes

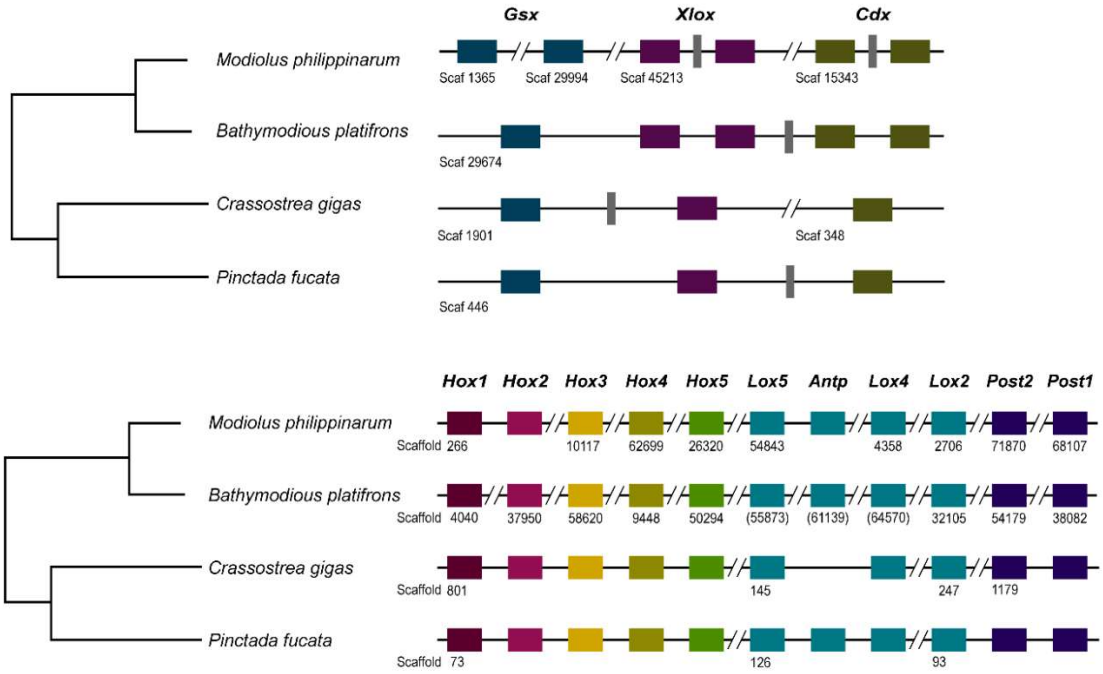
The blue mussels *Mytilus* spp. are distributed all over the world ocean and have been intensively used in aquaculture. Although many *Mytilus* transcriptome datasets are available, only transcriptomes from three species sequenced on the Illumina platforms and with sufficient coverage were chosen. For *Mytilus edulis*, there were 32.8 Gb raw reads (NCBI SRA number: SRR1424831). For *Mytilus galloprovincialis*, there were 5.4 Gb raw reads (NCBI SRA number: SRP011280.2) (Gerdol et al. 2014). For *Mytilus coruscus*, there were 4.4 Gb raw reads (NCBI SRA number: SRR2895132) (Xu et al. 2016). Using our previously described transcriptome analytical pipeline, raw reads were cleaned, trimmed, *de novo* assembled and redundant reads removed. The N50 value of the newly assembled *M. edulis*, *M. galloprovincialis* and *M. coruscus* transcriptome was 1248 bp, 1278 bp and 1577 bp, respectively.

The assembled contigs were translated into proteins using TransDecoder, a module in Trinity package (Haas et al. 2013), with only the longest protein sequence from each transcript being used as the representative. Protein sequences from each species were grouped into gene families using OrthoMCL (Li et al. 2003) as described above. Only the one-to-one orthologs were aligned, trimmed and concatenated. Maximum likelihood trees were built using RaxML (Stamatakis et al. 2005) by applying the GTR + Γ model and slow bootstrapping method.

Both the DNA and RNA based trees supported that *Modiolus* is the closest taxon to *Bathymodiolus* with high bootstrap value support (Fig. 1c and d). The divergent time between *B. platifrons* and *M. philippinarum* was estimated as ca. 110.4 MYA (the early/late Cretaceous). In previous studies using COI, modern bathymodiolin mussels

738 which harbour symbionts were estimated to arise over a wide time period from 21
739 MYA to 94 MYA (Little and Vrijenhoek 2003; Jones et al. 2006; Miyazaki et al. 2010;
740 Lorion et al. 2013). Our calculation using genome-wide data resulted in an estimation
741 that was slightly older than the upper range of the previously estimated divergence
742 age between shallow and deep-sea mussels. This result was remarkable because
743 deep-sea mussels started to radiate and colonise various vents and seeps only about 45
744 MYA (Little and Vrijenhoek 2003; Kiel and Amano 2013), indicating there was a long
745 period of time (ca. 59 MYA) from the origin to radiation. This result lends support to
746 the hypothesis that the ancestor of modern deep-sea mussel may have experienced an
747 extinction event during the global anoxia period in association with a thermal
748 maximum at ca. 57 MYA (the Palaeocene/Eocene) (Jones et al. 2006; Olu-Le Roy et
749 al. 2007; Lorion et al. 2013).
750
751 **13. Hox and ParaHox gene characterisation**
752 Hox and ParaHox genes are conserved metazoan genes involved in the embryonic
753 development by patterning the anterior-posterior body axis. To determine distribution
754 patterns of Hox and ParaHox genes in the mussel genomes, a tBLASTn (Altschul et al.
755 1990) was conducted using Hox and ParaHox sequences of *C. teleta* along with
756 homeodomain sequences downloaded from the HomeoDB (Zhong and Holland 2011).
757 Identity of each putatively identified gene was confirmed by reciprocal comparison
758 against the NCBI nr database using BLASTx. Homeodomain sequences were
759 translated into amino acid sequences and aligned using MUSCLE (Edgar 2004) to fly
760 *Drosophila melanogaster*, human *Homo sapiens* and amphioxus *Branchiostoma*
761 *floridae* sequences of known homology. Aligned regions were gap-removed using
762 MEGA (Kumar et al. 2008) and Maximum-likelihood phylogenetic tree was
763 constructed using the LG model plus gamma distribution and invariant site (LG + Γ
764 +I) with 4 gamma categories implemented in MEGA-CC (Kumar et al. 2012), and the
765 robustness of the tree was assessed by 1000 bootstraps.
766
767 Similar to other bilaterians, the two mussel genomes investigated here contained all

768 the 11 expected Hox and ParaHox genes as reported in other bivalve genomes (Fig.
769 S10). In *B. platifrons*, these 11 Hox genes were scattered on different scaffolds; while
770 these 11 Hox genes were located on 9 scaffolds in *M. philippinarum* with *Hox1* and
771 *Hox2* connected and *Lox5* and *Antp* connected. Although the scaffold sizes ranged
772 from 60.0 kb to 2.40 Mb in length (average length 218.5 kb) in *B. platifrons*, we could
773 not identify any protein-coding genes other than the Hox genes in respective scaffolds
774 in both genomes, and thus it remains unclear whether the Hox genes are arranged in
775 an intact cluster. On the other hand, the ParaHox genes were arranged in a cluster in *B.*
776 *platifrons* with intervening non-homeobox genes located between the *Xlox* and *Cdx*,
777 which mirrored the ParaHox cluster genes organisation in other lophotrochozoans.
778
779 Although in many metazoans Hox and ParaHox genes form only one to few synteny
780 blocks (Garcia-Fernandez 2005), in some mollusks they are located in many clusters
781 (Zhang et al. 2012a; Albertin et al. 2015; Takeuchi et al. 2016). Albertin et al. (2015)
782 suggested that this broken synteny pattern could be the result of either whole-genome
783 duplication followed by extensive gene loss, or increase in chromosome number by
784 chromosome fragmentation. They considered that the former scenario was unlikely
785 because of the great extent of gene loss required. Our results from the two mussels
786 appeared to support the latter hypothesis.
787



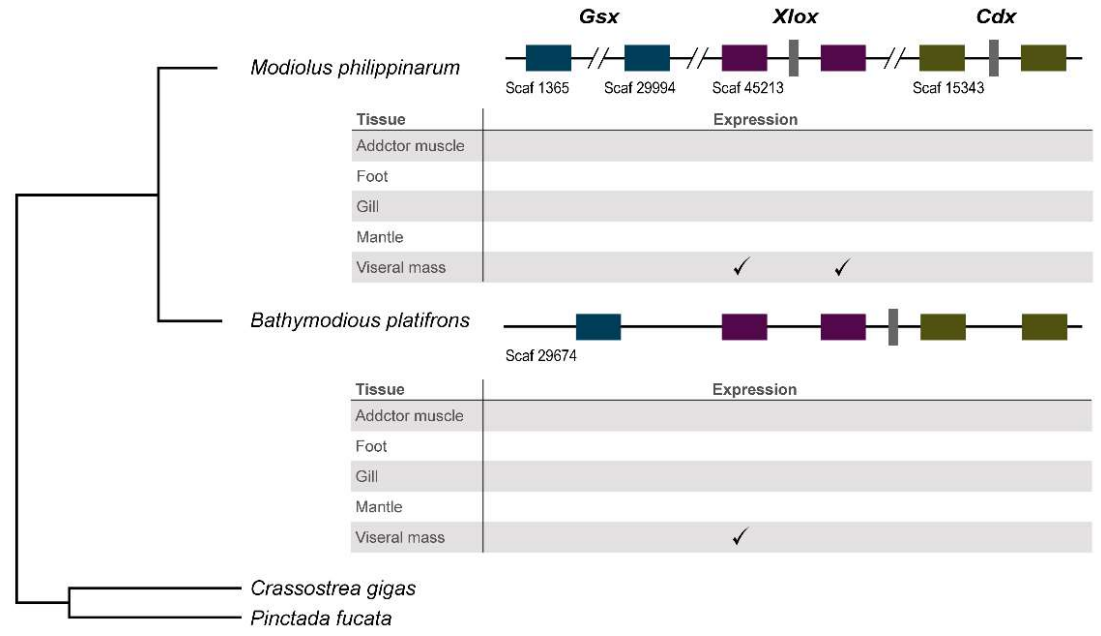
788

789

790 Fig. S10. Organisation of Hox and ParaHox cluster genes in four bivalve genomes

791

792 The expression levels of Hox and ParaHox genes in these two species of mussels were
793 also examined in the transcriptomic data (Fig. S11). In the examined tissues, only the
794 *Xlox* gene in the ParaHox genes was expressed in the visceral mass of both species.
795 Meanwhile, the Hox cluster genes were expressed in the different examined tissues
796 (Fig. S11).



797

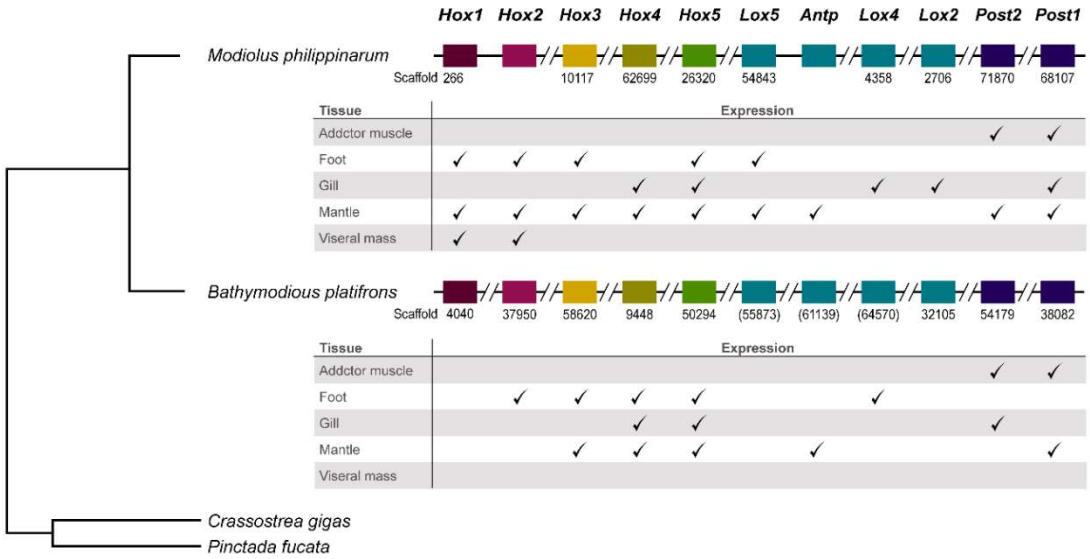


Fig. S11. Expression of Hox and Hox and Parahox genes

14. microRNA characterisation

A total of 602 unique mature microRNA sequences were downloaded from different lophotrochozoan genomes deposited in miRBase release 21 (Griffiths-Jones et al. 2008). Systematic searches were performed using MapMi version v150b01 with maximum mature mismatch set to 1 (Guerra-Assunção and Enright 2010). MapMi used a combination of Bowtie (Langmead and Salzberg 2012) and RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) to map the mature miRNA sequences to the genomes.

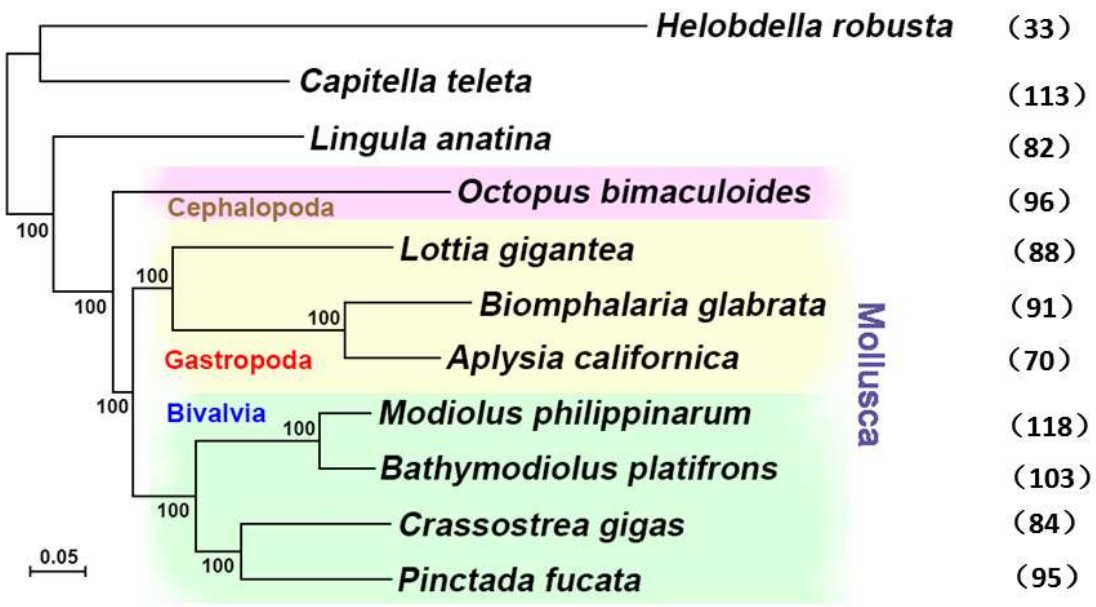


Fig. S12. Number of microRNAs predicted in different lophotrochozoan genomes. The phylogenetic tree (same as Figure 1a) was used to show the relationship among lophotrochozoans.

MicroRNAs are conserved short non-coding RNAs that act as post-transcriptional regulators in animals. Comparison of the numbers of microRNAs in the two mussel genomes to other lophotrochozoan genomes suggested that they contained similar numbers of microRNAs without having undergone extensive duplications nor losses (Fig. S12 and Table S8).

15. Gene expression analysis

Identifying highly expressed genes can help us understand the main functions of the tissue. In both species of mussels, the clean RNA-Seq reads in various tissues were quantified using kallisto (Bray et al. 2016) which applies a pseudo-alignment method to rapidly determine the compatibility of reads with targets. Gene expression level was quantified based on transcripts per million (TPM). A tissue specifically expressed gene was defined as a gene that had over 75% of the total transcripts in a particular tissue based on TPM (Albertin et al. 2015). GO enrichment analysis was performed based on the GO annotation using GO-EAST with the Hypergeometric test, and the result was further corrected by the Yekutieli method (Zheng and Wang 2008). The

gene expression data were $\log_2$(TPM+1) transformed, and principal component analysis (PCA) analysis was performed using PAST v3.12 (Hammer et al. 2001) to determine the similarity in gene expression among the tissues and in the two species. The PCA result can be represented by a 2-dimensional plot, with the two components capturing a total of 49% variance (Fig. S13). The first component (PC1) separated *B. platifrons* (Bpl) and *M. philippinarum* (Mph) and explained 29% of variance, indicating the species factor was more important that the tissue type in influencing gene expression.

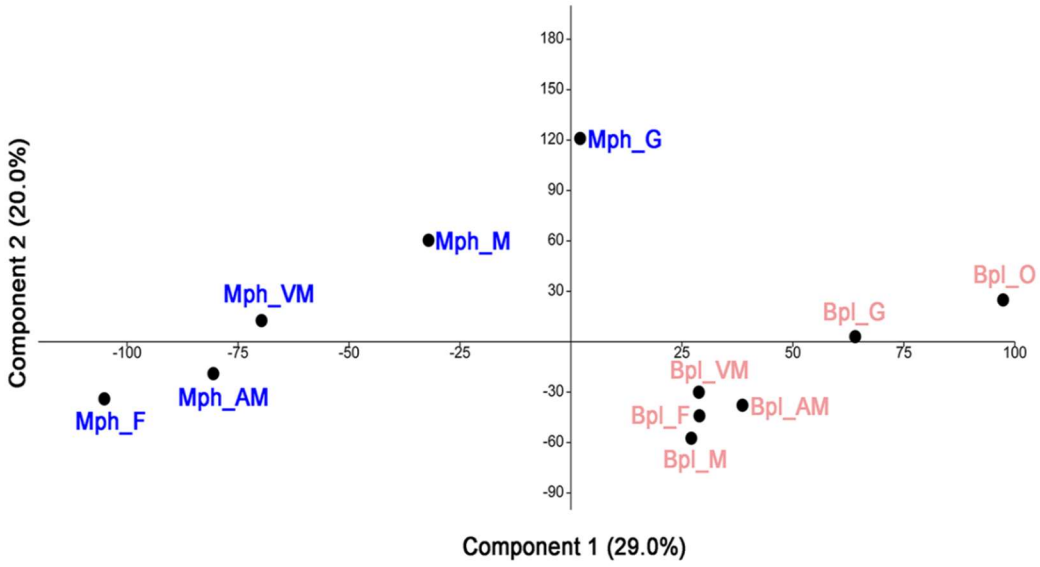


Fig. S13. Principle component analysis (PCA) of gene expression in different tissues of the two species of mussels. Only 1 to 1 orthologs were used. Abbreviations: AM, adductor muscle; F, foot; O, ovary; G, gill; M, mantle; and VM, visceral mass.

The enriched GO categories of genes specifically expressed in the gills of the two species exhibited remarkable differences (Table S9). For *M. philippinarum*, the gill specific genes were mainly related to cytoskeleton. For *B. platifrons*, the enriched gene categories were mainly related to ammonia transportation, low-density lipoprotein transport, toxin transporter activity, response to sterol and WNT signalling pathway (to be further discussed in section 19). Abundant toxin-related genes have been identified in two *Bathymodiolus* species that harbour sulfur-oxidizing endosymbionts (Sayavedra et al. 2015). The toxin-related genes were suggested to be

involved in defense against pathogens (Sayavedra et al. 2015). The highly expressed toxin transporter genes in the gill of *B. platifrons* might be related to the need to coordinate with the symbiont for toxin secretion. This hypothesis can be further tested when the endosymbiont genome of *B. platifrons* is sequenced.

Table S9. GO enrichment of gill specifically expressed genes.

GOID	Term	Level	log_odds_ratio	*P*-value
Bathymodiolus platifrons				
GO:0008519	ammonium transmembrane transporter activity	2	3.910	0.000681
GO:0015696	ammonium transport	3	3.910	8.84E-06
GO:0072488	ammonium transmembrane transport	6	4.173	0.000349
GO:0060485	mesenchyme development	4	2.531	0.00148
GO:0009953	dorsal/ventral pattern formation	4	2.822	0.00205
GO:0005041	low-density lipoprotein receptor activity	1	2.910	0.00645
GO:0030228	lipoprotein particle receptor activity	1	2.911	0.00645
GO:0014029	neural crest formation	8	2.911	0.00645
GO:0019534	toxin transporter activity	1	2.911	0.00645
GO:0034391	regulation of smooth muscle cell apoptotic process	1	2.911	0.00645
GO:0034392	negative regulation of smooth muscle cell apoptotic process	5	2.911	0.00645
GO:0009798	axis specification	4	2.450	0.00552
GO:0009950	dorsal/ventral axis specification	5	2.794	0.00817
GO:0044332	Wnt signaling pathway involved in dorsal/ventral axis specification	6	2.911	0.00645
GO:0036314	response to sterol	4	2.911	0.00645
GO:0036315	cellular response to sterol	7	2.911	0.00645
GO:0070723	response to cholesterol	4	2.911	0.00645
GO:0071397	cellular response to cholesterol	8	2.911	0.00645
GO:0015026	coreceptor activity	2	2.911	0.00645
GO:0071936	coreceptor activity involved in Wnt signaling pathway	2	2.911	0.00645
GO:0030177	positive regulation of Wnt signaling pathway	5	2.386	0.00645
GO:2000053	regulation of Wnt signaling pathway involved in dorsal/ventral axis specification	3	2.911	0.00645
GO:2000055	positive regulation of Wnt signaling pathway involved in dorsal/ventral axis specification	5	2.911	0.00645

GOID	Term	Level	log_odds_ratio	*P*-value
Modiolus philippinarum				
GO:0005874	microtubule	7	1.194	0.00221
GO:0015630	microtubule cytoskeleton	4	0.751	0.00499
GO:0005200	structural constituent of cytoskeleton	1	1.726	0.00128
GO:0006928	movement of cell or subcellular component	2	1.263	8.39E-05
GO:0007017	microtubule-based process	2	1.380	4.21E-07
GO:0043623	cellular protein complex assembly	6	1.430	0.00352
GO:0051258	protein polymerization	6	1.828	0.00075
GO:0005875	microtubule associated complex	7	1.352	0.00022
GO:0005929	cilium	3	1.360	0.00505
GO:0003774	motor activity	1	1.326	0.00013
GO:0003777	microtubule motor activity	1	1.693	4.79E-06
GO:0007018	microtubule-based movement	3	1.569	1.07E-05
GO:0046912	transferase activity, transferring acyl groups, acyl groups converted into alkyl on transfer	1	3.278	0.00215
GO:0006633	fatty acid biosynthetic process	9	2.089	0.00504
GO:0030286	dynein complex	8	1.968	4.07E-06

The enrichment of genes related to ammonia transporters in *B. platifrons* was consistent with the result of a previous study showing *Bathymodiolus* sp. could actively absorb ammonium from the ambient environment (Lee et al. 1992). Ammonia transporters (Amts) have been widely identified in prokaryotes and eukaryotes, playing essential roles of uptaking reduced nitrogen and of energy metabolism (Andrade and Einsle 2007). More evidence of the coordination between the host and symbiont in nitrogen metabolism can be found in our meta-proteomics analysis of the *B. platifrons* gill (section 21).

16. Gene family evolution

Because gene family classification based on BLASTp homology search may split a big gene family (e.g. TLR and *wnt*) into several small gene sub-families, we applied the HMM scanning method on known Pfam functional protein domains to classify the gene families (Albertin et al. 2015). Annotation of Pfam domains were performed on all the 11 lophotrochozoan species after excluding protein domains of all kinds of

transposable elements (i.e., DDE_Tnp, transposase, helitron, helicase and transposon). A domain with multiple copies in a protein was counted only once. Gene family expansion/contraction events in *B. platifrons* were determined using a two-tailed Fisher's exact test in R, with the Pfam domain counts in *B. platifrons* compared with the background average domain counts of the 9 non-mussel mollusks (i.e., species other than *B. platifrons* and *M. philippinarum*). The *P*-values were sequentially corrected using the Benjamini and Hochberg method (Benjamini and Hochberg 1995), and only gene families with corrected *P*-values less than 0.05 and the number higher/lower than at least half of the outgroup species were considered as expanded/contracted.

The events of Pfam domain expansion and contraction in each lineage are shown in Fig. 2b and Table S10. Among all of the lineages examined, the lineage of *B. platifrons* had the least number of contracted functional domains– only 39 which is the smallest number among all the species examined. Likewise, comparison of the major GO gene categories between *B. platifrons* and *M. philippinarum* using WEGO did not find any gene category in which *B. platifrons* contains significantly fewer genes than *M. philippinarum*. Collectively, these results indicated that the adaptation of *B. platifrons* to deep-sea chemosynthetic environment was mainly mediated by gene family/domain expansion, and the deep-sea mussel had retained most genes of their shallow-water ancestors.

Besides analysing the protein families using the Pfam domain method, we also applied OrthoMCL (Li et al. 2003) to cluster gene families and a Bayesian method using Computational Analysis of Gene Family Evolution (CAFÉ v 3.1) to determine the significance of gene family expansion/contraction (Han, et al. 2013). Because this approach assumes that there is at least one ancestral gene in the Most Recent Common Ancestor (MRCA), only gene families that were shared between at least one of the outgroups (non-mollusks, i.e., *Capitella teleta*, *Hellobdella robusta* and *Lingula*

914 *anatina*) and at least one of the molluscan species were included in the analysis. This
915 filtering reduced the number of gene families to 11,261 for analysis with CAFÉ. For
916 each gene family, CAFÉ generated a family-wide *P*-value, with a significant *P*-value
917 indicating a possible gene family expansion or contraction event. Also, in each gene
918 family, a branch-specific *P*-value was generated for each branch/node using the
919 Viterbi method. In this study, family-wide *P*-value less than 0.01 and a branch/node
920 'Viterbi *P*-value' less than 0.001 was considered as having expansion/contraction for a
921 specific gene family and specific species, respectively, as recommended by the
922 authors of CAFÉ v 3.1.
923
924 The results showed that 81 gene families have expanded and 11 families have
925 contracted (Supplementary table S11). The CAFÉ result is consistent with the domain
926 analysis result. For example, the gene families fibrinogen, cadherin, Toll-like receptor
927 (which contains a TIR domain), and methyltransferase 24 were also found to be
928 expanded. However, the protein domain analysis tends to cluster proteins into bigger
929 gene families, whereas the CAFÉ analysis tends to provide a finer clustering of
930 smaller gene families. For instance, the domain analysis showed that there are 214
931 proteins with the TIR domain in the deep-sea mussel, whereas the CAFÉ analysis
932 showed that several Toll-like receptors (TLRs) which are characterised by having the
933 TIR domain. Specifically, TLR4 and TLR13 were both expanded in the deep-sea
934 mussel.
935
936 Animals in vents and seeps heavily rely on chemotrophic and methanotrophic
937 microbes that can utilise simple organic molecules such as hydrogen sulfide and
938 methane as a source of energy to fuel the ecosystem in the absence of solar energy
939 (Van Dover 2000). Thus, deep-sea mussels may have lost some digestive enzymes
940 that are used to digest planktonic food captured by filter-feeding. To test this
941 hypothesis, we examined the expansion/contraction of glycosyl hydrolases – proteins
942 involved in hydrolysis of glycosidic bonds in complex sugar and suggested to be the

crucial digestive enzymes in shallow-water mussels (Xu 2002). The *B. platifrons* genome had 53 glycosyl hydrolase gene families in the Pfam clan of CL0058. None of these gene families have undergone extensive expansion or contraction in *B. platifrons* when compared with other 10 lophotrochozoan species including *M. philippinarum* (Table S12). Our result was thus consistent with the finding that *Bathymodiolus* mussels still retained the ability to filter-feed (Lee et al. 1992), as they retained all these enzymes for digestion. In fact, the flexible feeding behavior of *Bathymodiolus* has been considered to be a key trait contributing to their survival when most macrobenthos have died during the vent or seep extinction process (Van Dover 2000). Additionally, since the larvae of *Bathymodiolus* will swim up to the surface to feed on phytoplankton (Arellano and Young 2009; Arellano et al. 2014), retaining a repertoire of glycosyle hydrolase coding genes in the *B. platifrons* genome can be important for its larvae before settling in a chemosymthesis-based environment.

There were 111 expanded protein domains in *B. platifrons* (Fig. 2c and Table S10). These expanded protein domains were related to protein folding (e.g. HSP70), detoxification (e.g. ABC transporter), immune response (e.g. immunoglobiulin, C1q, fibrinogen, and TIR etc.), endocytosis (e.g. Ras, syndecan, and protocadherin), apoptosis (e.g. DEATH, TNF, CARD, caspase, and IAP) and protein modification (e.g. dual specificity phosphatase and galactosyltransferase) (Fig. 2c). The expansion and expression of selected gene families will be further discussed in section 18.

17. Positively selected genes by using branch-site model

A total of 5,039 single-copy orthologues were identified among *B. platifrons*, *M. philippinarum*, *C. gigas*, and *P. fucata* using OrthoMCL (Li et al. 2003). Sequences from each single-copy orthologue group were aligned using MUSCLE (Edgar 2004), trimmed using Gblocks with a stringent threshold (Talavera and Castresana 2007), and concatenated. Phylogenetic tree was constructed using RAxML (Stamatakis et al.

2005) with the GTR + Γ model and rapid bootstrapping on 1,337,786 amino acids.

For each single-copy gene family, ParaAT, a programme in the MUSCLE package (Zhang et al. 2012b), was used to align the protein sequences and back-translate them to their corresponding cDNA sequences for use in downstream analysis. The POTION pipeline (Hongo et al. 2015) was applied to further refine the alignment and change the format required for the downstream analyses.

Positive selection of genes in a specific branch can be detected by branch-site models implemented in the codeml module of the phylogenetic analysis by maximum likelihood (PAML) package v4.8 (Yang 2007). Application of a branch-site model requires a lineage to be designated as the "foreground" (i.e., *B. platifrons*) and the rest as the "background" (i.e., *M. philippinarum*, *C. gigas* and *P. fucata*). Alternative model (model = 2, NSsite = 2 and fix_omega = 0) and null hypothesis (model = 2, NSsite = 2, fix_omega = 1 and omega = 1) likelihood values were calculated using Chi-square distribution, and the resultant significance level was further corrected by the Benjamini and Hochberg method (Benjamini and Hochberg 1995). Only genes with > 90% Bayesian Empirical Bayes (BEB) sites and having a corrected *P*-value less than 0.1 were considered as positively selected. A KEGG pathway enrichment was performed with a Chi-square test (if any of the component were over 5) or Fisher's exact test (otherwise).

This analysis showed that 162 genes in the *B. platifrons* lineage had a clear signal of positive selection (Supplementary Table S13). These positively selected genes function in many essential biological processes. However, only one KEGG pathway (i.e., endocytosis) with six positively selected genes was enriched ($P = 0.00526$) (Fig. S14). Since *Bathymodiolus* mussels recruit symbiotic bacteria into the gill epithelial cells through endocytosis (Won et al. 2003a), the enrichment of the endocytosis pathway indicated that the gill of *B. platifrons* has been remodeled in adaptation to the

need to conduct phagocytosis of symbionts in the chemosynthetic environment.

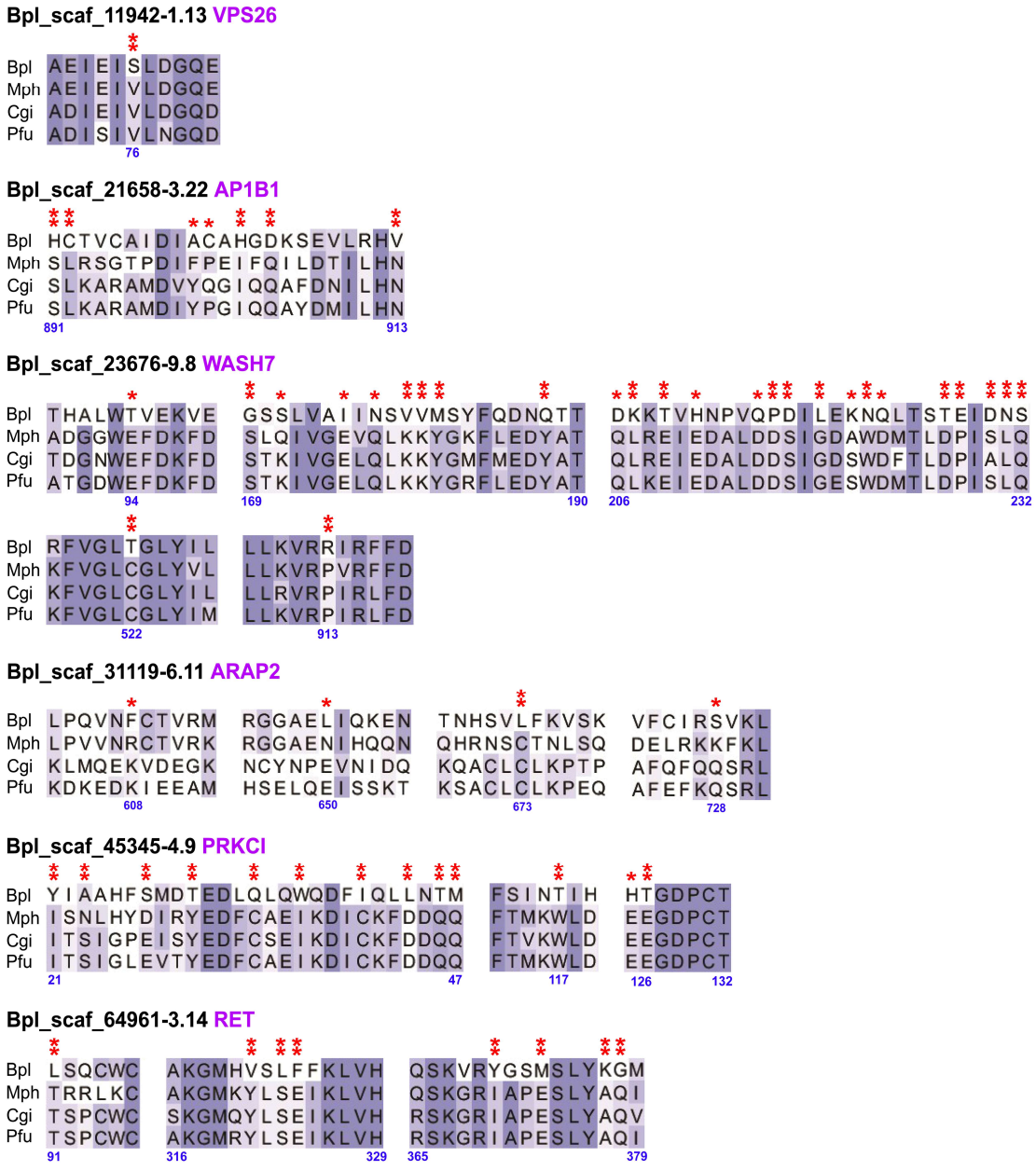


Fig. S14. Alignment of six endocytosis related genes in four bivalves. The asterisks indicate the significant level of positive selection calculated by the PAML branch-site model, where '*' indicates $P > 0.9$ and '**' indicates $P > 0.95$. The blue numbers under the sequences indicate the positions of the amino acids in the protein sequence. Abbreviation: Bpl, *B. platifrons*; Mph, *Modiolus philippinarum*, Pfu, *Pinctada fucata*; Cgi, *Crassostrea gigas*; VPS, vacuolar protein-sorting proteins; AP-1, adaptor protein complexes 1; WASH, Wiskott-Aldrich syndrome protein and SCAR homologue; ARAP, arf-GAP with Rho-GAP domain, ANK repeat and PH domain-containing protein; PRKCI, protein kinase C iota type; and RET, proto-oncogene tyrosine- kinase receptor Ret.

18. Specific gene family analyses

Selected *B. platifrons* gene families that are putatively involved in adaptation to the deep-sea environment were analysed for details of lineage-specific expansion and/or high expression in the gill. For each gene family, sequences in the assembled genome were identified using BLASTp and aligned using the hmmalign programme of the HMMER package with their corresponding reference sequence in the Pfam-A database (Mount 2009). The FastTree programme (Price et al. 2010), which applied an "approximately-maximum-likelihood" and was designed for large gene family analysis, was used to construct a phylogenetic tree using the model of WAG + Γ with pseudo-counts. When appropriate, a heat-map was also used to plot the tissue-specific expression pattern of these genes.

18.1 Heat shock protein 70

Molecular chaperones, including the heat shock protein 70 (HSP70) family, are important in maintaining protein function as they assist protein covalent folding, unfolding, assembly and disassembly (Daugaard et al. 2007; Verghese et al. 2012). Expansion of the HSP70 gene family in the Pacific oyster *Crassotrea gigas* (80 compared to 17 in human and 39 in sea urchins) was suggested to be an adaptation to the intertidal zone with fluctuating environmental factors such as desiccation and temperature during tidal cycles (Zhang et al. 2012a). A high number of HSP70 genes (i.e., 120) was also found in the pearl oyster *Pinctada fucata* (Table S10). Sequencing the genome of two mussels in this study revealed high numbers of HSP70 genes in both *M. philippinarum* (99) *and B. platifrons* (179), indicating expansion of this family is common to bivalves (Table S10).

Nevertheless, comparing the shallow water and deep-sea mussel showed that there was expansion of this gene family (i.e., 80 more genes) in association of the invasion of mussels to the deep-sea. Considering the harsh hydrothermal vent and methane seep environment (e.g. high concentrations of hydrogen sulfide and heavy metal,

hypoxia, variable temperatures), having more chaperone proteins may help deep-sea chemosynthesis-based fauna maintain proper protein functions. In fact, frequent DNA damage has been reported in the vent mussel *Bathymodiolus azoricus*, and the level of DNA damage is positively correlated with the expression level of HSP70 genes (Pruski and Dixon 2007), indicating the active involvement of HSP70 in maintaining proper protein functions of other proteins in deep-sea mussels. The number of HSP70 genes in *B. platifrons* was the highest among all of the bivalve species examined. Among them, 76 were validated by mass spectrometry in this study. The expansion of HSP70 genes was further evaluated by phylogenetic analysis using the four bivalve genomes with a sequenced genome, which showed that at least one cluster of HSP70 genes was dominated by *B. platifrons* genes, providing further evidence of gene expansion in this particular lineage (Fig. S15a). At least two scaffolds each contained a tandem cluster of 6 and 7 HSP70 genes respectively (Fig. S15c), indicating a recent and specific duplication of this gene family in the deep-sea mussel. Examining the gene expression profile showed that most of the HSP70 domain containing genes were highly expressed in the gill (Fig. S15b), again indicating the active involvements of this gene family in coping with the ambient environment because the gill is the organ responsible for respiration and filtration.

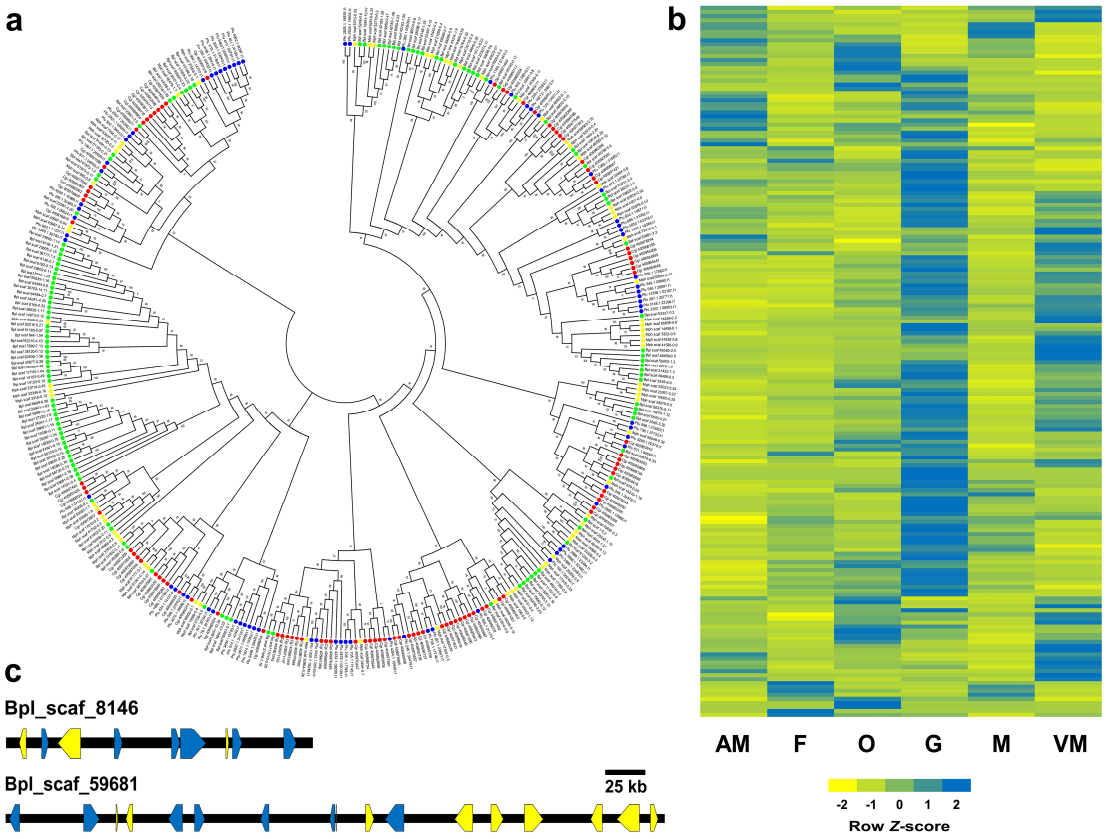


Fig. S15 Expansion of the HSP70 family in *Bathymodiolus platifrons* (Bpl). (a) an unrooted gene genealogy of HSP70 genes in Bpl (green), Mph (yellow), Cgi (red) and Pfu (blue); (b) A heat map of expression of HSP70 genes in six tissues; (c) genomic arrangement of two HSP70 gene clusters. HSP70 genes are shown in blue, whereas other genes are labelled in yellow. Scaffold Bpl_scaf_8146 contains 6 HSP70 genes that are transcribed in the same direction, whereas scaffold Bpl_scaf_59681 contains 7 HSP70 genes which are transcribed in opposite directions. Abbreviations: Mph, *Modiolus philippinarum*, Pfu, *Pinctada fucata*; Cgi, *Crassostrea gigas*; AM, adductor muscle; F, foot; O, ovary; G, gill; M, mantle; and VM, visceral mass.

18.2 ATP-binding cassette transporters

ATP-binding cassette (ABC) transporters, a large gene family with transmembrane domain, are involved in trans-membrane trafficking and biochemical defense through removing toxicants (Sturm et al. 2009). At least one ABC transporter was shown to remove cadmium from the cell and thus confer the animal with heavy metal resistance (Kim et al. 2007). In mussels, ABC transporters were suggested to form an environment-tissue barrier (Luckenbach and Epel 2008). Millimolar level thiosulfate, a product of sulfide oxidation that can be toxic to the cell, was detected in the blood of the vent mussel *Bathymodiolus thermophilus* (Childress et al. 1992). In addition,

striking high levels of trace metals (e.g., iron, chromium, copper, zinc, molybdenum, silver, antimony, barium, mercury, magnesium, and uranium) were documented in the clam *Calyptogena magnifica* (Roesijadi et al. 1985), the mussel *B. thermophiles* (Childress et al. 1992), the polychaete *Alvinella pompejana* (Gaill et al. 1984), and the limpet *Neomphalus fretterae* (Childress et al. 1992) inhabiting hydrothermal vents. The accumulated level of some trace metals (e.g. 148 mg kg^{-1} of copper and 2152 mg kg^{-1} of zinc) in these vent species could be lethal if applied to their shallow-water relatives (Childress et al. 1992). There were 393 ABC transporter containing genes in *B. platifrons*, among them 111 had proteomics support. Among the 11 lophotrochozoan species examined, *B. platifrons* had the highest number of ABC transporter genes (Table S10a), and the expression level of most these genes were highest in the gill (Fig. S16). Collectively, the high number and high expansion of ABC transporter genes in the deep-sea mussel indicated their role of removing toxic substances from the deep-sea mussel.

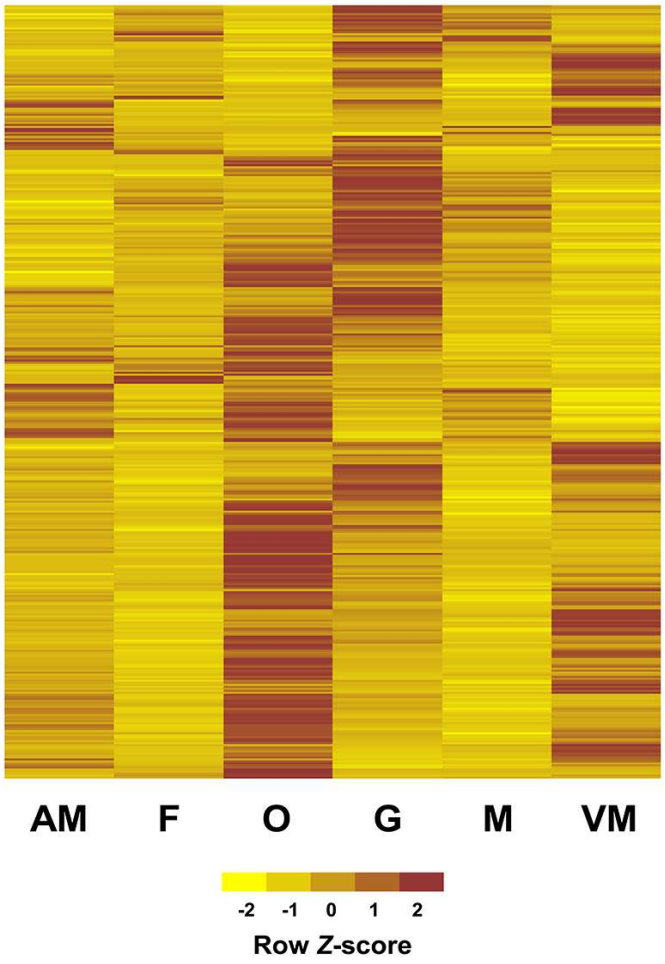


Fig. S16. Heat map showing the expression pattern of ABC transporter genes among six tissues of *B. platifrons*. Abbreviations: AM, adductor muscle; F, foot; O, ovary; G, gill; M, mantle; and VM, visceral mass.

18.3 Immune recognition

As the first step to establish the host-symbiont relationship, pattern recognition is required to distinguish the specific symbiotic bacteria from other microbes in the environment. This process involves a variety of host pattern recognition receptors (PRRs) which can detect the signature molecule on the surface of the microbe, named microbe-associated molecular patterns (MAMPs) (Hoffmann et al. 1999). Protein domains that are most significantly expanded in the deep-sea mussel were various types of immunoglobin domains (i.e., I-set, V-set, C2-set, C1-set, and Izumo-like) (Table S10a), which are believed to be involved in the protein to protein/ligand

interactions (Williams and Barclay 1988; Zhang et al. 2004). Expansion of immune related genes was found in other host-symbiont systems, such as the coral *Acropora digitifera* (Shinzato et al. 2011) and the sea anemone *Aiptasia* (Baumgarten et al. 2015). Likewise, there was great expansion in genes containing immune-related domains (i.e., immunoglobin, C1q, TIR and fibrinogen) in the deep-sea mussel, which might be important to establish the host-symbiont relationship, and defend the host from pathogen attack. Below are two detailed examples of such gene expansions.

18.3.1 Peptidoglycan recognition proteins

Peptidoglycan recognition proteins (PGRPs) are evolutionary conserved innate immunity molecules present widely in metazoans. They are similar to bacterial type 2 amidases in structure and thus can bind to bacterial cell wall peptidoglycan (Dziarski and Gupta 2006). Analysis of the endosymbiosis relationship in the weevil *Nardonella* spp. indicated that the presence of endosymbiotic bacteria could enhance PGRPs expression, and the high expression of PGRPs could reduce the weevils' defense system in order to maintain a long term symbiotic relationship (Moya et al. 2008). In the squid *Euprymna scolopes* - *Vibrio* symbiosis, PGRPs had the highest expression levels in epithelial cells that harbour the bacteria, indicating the host's response to the peptidoglycan of the symbiont (Nyholm and Graf 2012). A study of *Bathymodiolus azoricus* collected from Atlantic vent sites indicated that PGRPs might be involved in establishing the immune response in the juveniles when they incorporated symbiotic bacteria into the body (Bettencourt et al. 2014). The present study showed that *B. platifrons* had the highest number of PGRPs among the four bivalves studied, i.e., 23 in *B. platifrons*, 9 in *M. philippinarum*, 10 in *C. gigas* and 10 in *P. fucata* (Fig. S17a). There was a tandem duplication with five PGRP genes in scaffold Bpl_scaf_21183, indicating a recent duplication event in *B. platifrons* (Fig. S17c). Phylogenetic analysis revealed two distinct clusters of PGRPs in the bivalves, and their expression levels were high in both the gill and visceral mass (Fig. S17b), indicating that various PGRPs might serve different functions in different issues, perhaps mainly symbiont

recognition in the gill and pathogen recognition in the visceral mass.

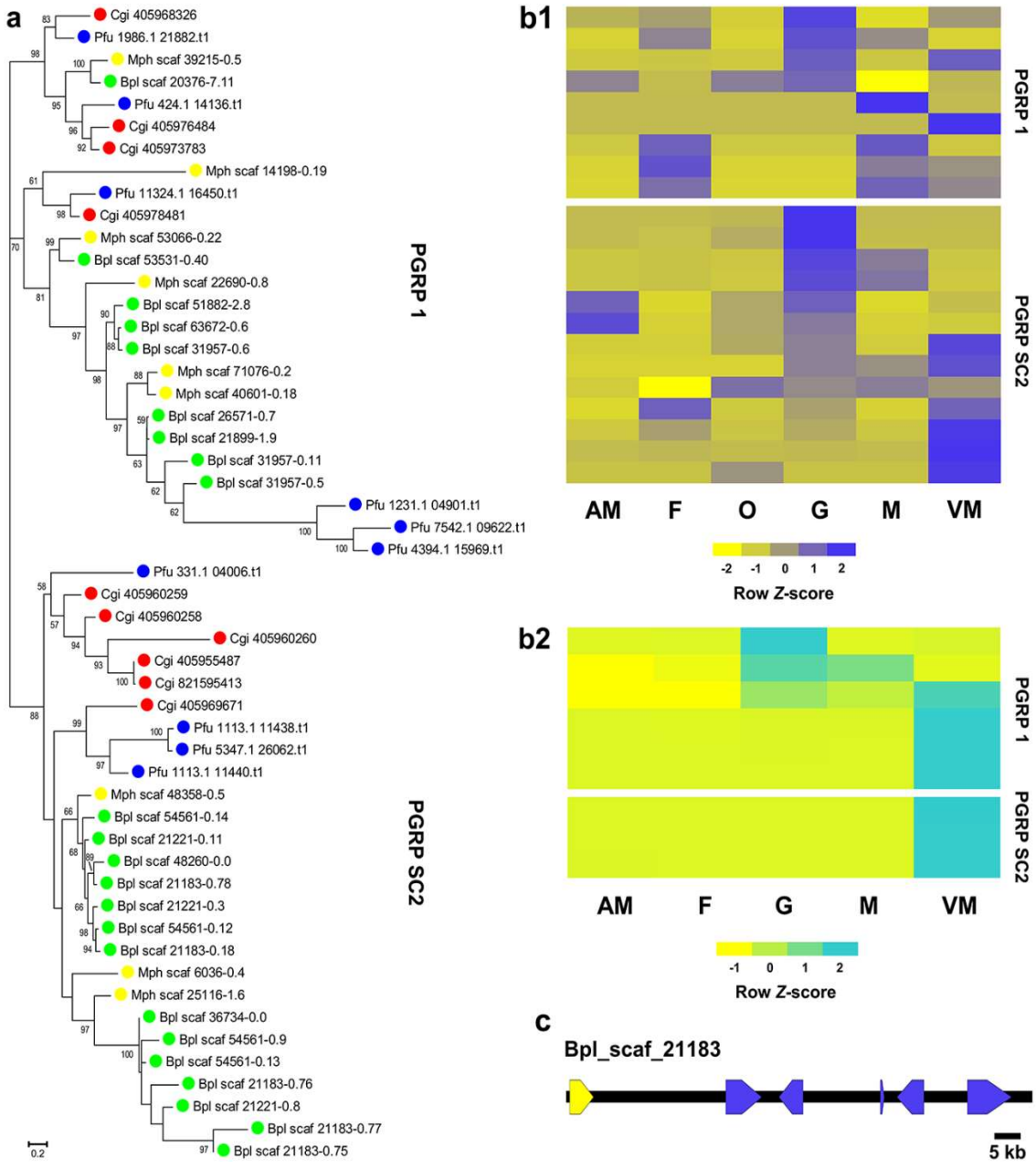


Fig. S17. PGRPs genes in *B. platifrons*. (a) Gene genealogy of PGRPs using sequences from four bivalves. (b1 and b2) Expression of two clusters of PGRPs. (c) genomic arrangement of a PGRP gene cluster indicating possible tandem duplication. PGRP genes are shown in blue, whereas another gene is shown in yellow. Abbreviations: AM, adductor muscle; F, foot; O, ovary; G, gill; M, mantle; and VM, visceral mass.

18.3.2 Fibrinogens

Although not well studied, invertebrates do possess some anticipatory and memory-like molecules to distinguish self and non-self molecules, and fibrinogens are among these molecules (Nyholm and Graf 2012). Fibrinogens are glycoproteins critical for immune process in animals (Mosesson 2005). They bind to parasites and bacteria, and their expression can be triggered by pathogen attack (Zhang et al. 2004). In the genome of *Aiptasia* which harbours symbiotic dinoflagellate algae, the expansion of the fibrinogen gene family was suggested as an example of the plasticity in the repertoire of pattern-recognition genes in invertebrates (Baumgarten et al. 2015). The *B. platifrons* genome contained 206 fibrinogen genes, the highest number of this family of genes among the 11 species of lophotrochozoans examined, and many of these genes were highly expressed in the gill (Fig. S18a). At least five scaffolds each contained a tandem duplication of 5, 5, 6, 6 and 6 fibrinogen genes (Fig. S18b). Due to their inherent role of pattern recognition, fibrinogens could be involved in the symbiotic bacteria recognition.

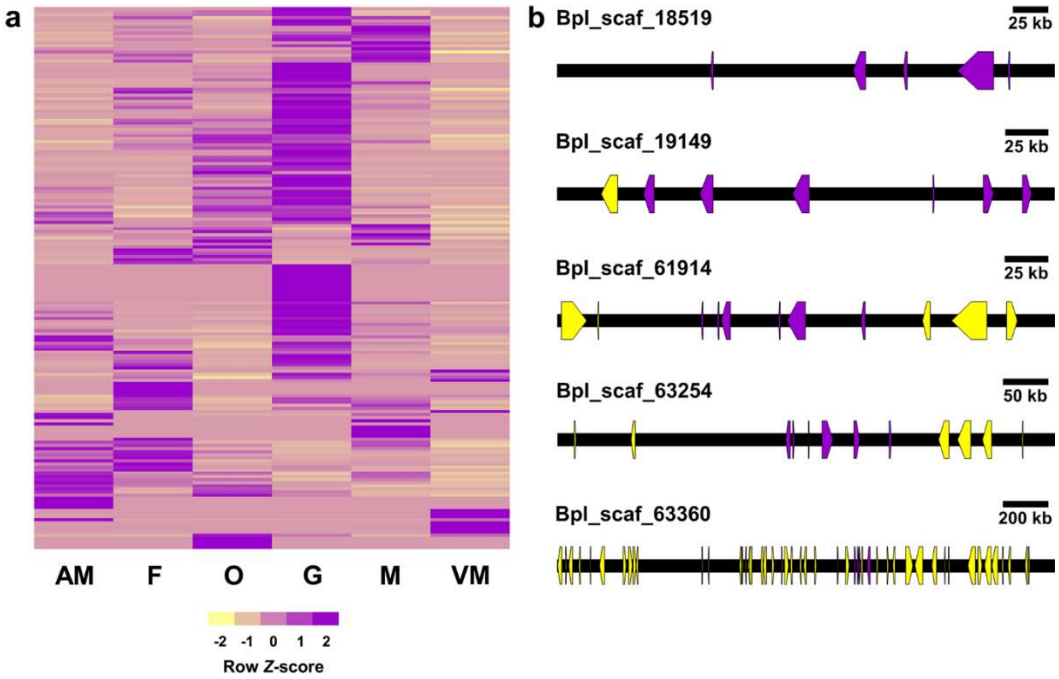


Fig. S18. Fibrinogen genes in *B. platifrons*. (a) Heat map of their expression in six tissues. (b) genomic arrangement of five gene clusters indicating possible tandem duplication. Fibrinogen genes are shown in purple, whereas other genes are shown in yellow. Abbreviations: AM, adductor muscle; F, foot; O, ovary; G, gill; M, mantle; and VM, visceral mass.

18.4 Endocytosis related genes/gene families

18.4.1 Wnt gene family

As a conserved signaling pathway in the development of metazoans, Wnt signalling pathway transfers signals into a cell through cell surface receptors (Nusse and Varmus 1992). In *B. platifrons*, 16 Wnt genes were identified, which can be grouped into Wnt1, Wnt2b, Wnt4, Wnt5b, Wnt6, Wnt7b, Wnt8a, Wnt9a, Wnt10a, Wnt11b and Wnt16 (Fig. S19a). Analysis of the gene expression indicated that Wnt4, Wnt7b, Wnt8a, Wnt10a and Wnt16 were highly expressed in the gill. The synteny of *wnt9-wnt1-wnt6-wnt10* in *B. platifrons* was the same as that in *Lottia gigantean* and *Pinctada fucata* but different from the synteny in *Daphnia* and *Drosophila* (Janssen et al. 2010; Takeuchi et al. 2016) (Fig. S19c). This result indicated that the Wnt gene cluster is conserved in the phylum Mollusca.

Wnt ligands could form a complex with Frizzled and low-density lipoprotein receptor related protein (LRP). This complex could further link to vacuolar ATPase (vATPase) which acidifies the vesicle, a process critical for the vesicle traffic. The protein complex could further trigger the endocytosis pathway with the help of endosomal sorting complex required for transport (ESCRT) (Niehrs 2012). The essential role of the Wnt signalling pathway in host defense was demonstrated in a study where *Drosophila* S2 cells were challenged by white spot syndrome virus (Niehrs 2012). The virus was engulfed by the cell but not digested. Further gene expression analysis showed differential expression of genes in the Wnt pathway (Zhu and Zhang 2013). Analysis of protein expression upon exposure to *Symbiodinium* in the sea anemone *Aiptasia* revealed the up-regulation of proteins involved in endocytosis (protein AP-1 and Rho GTPase) and low-density lipoprotein receptor (LPR) (Oakley et al. 2016), suggesting LRP and endocytosis were involved in the symbiont interaction. The high expression of genes in the Wnt signalling pathway and LRP receptor in *B. platifrons* implies active endocytosis, which may be related to the engulfment of MOB cells (to

be further discussed in section 20).

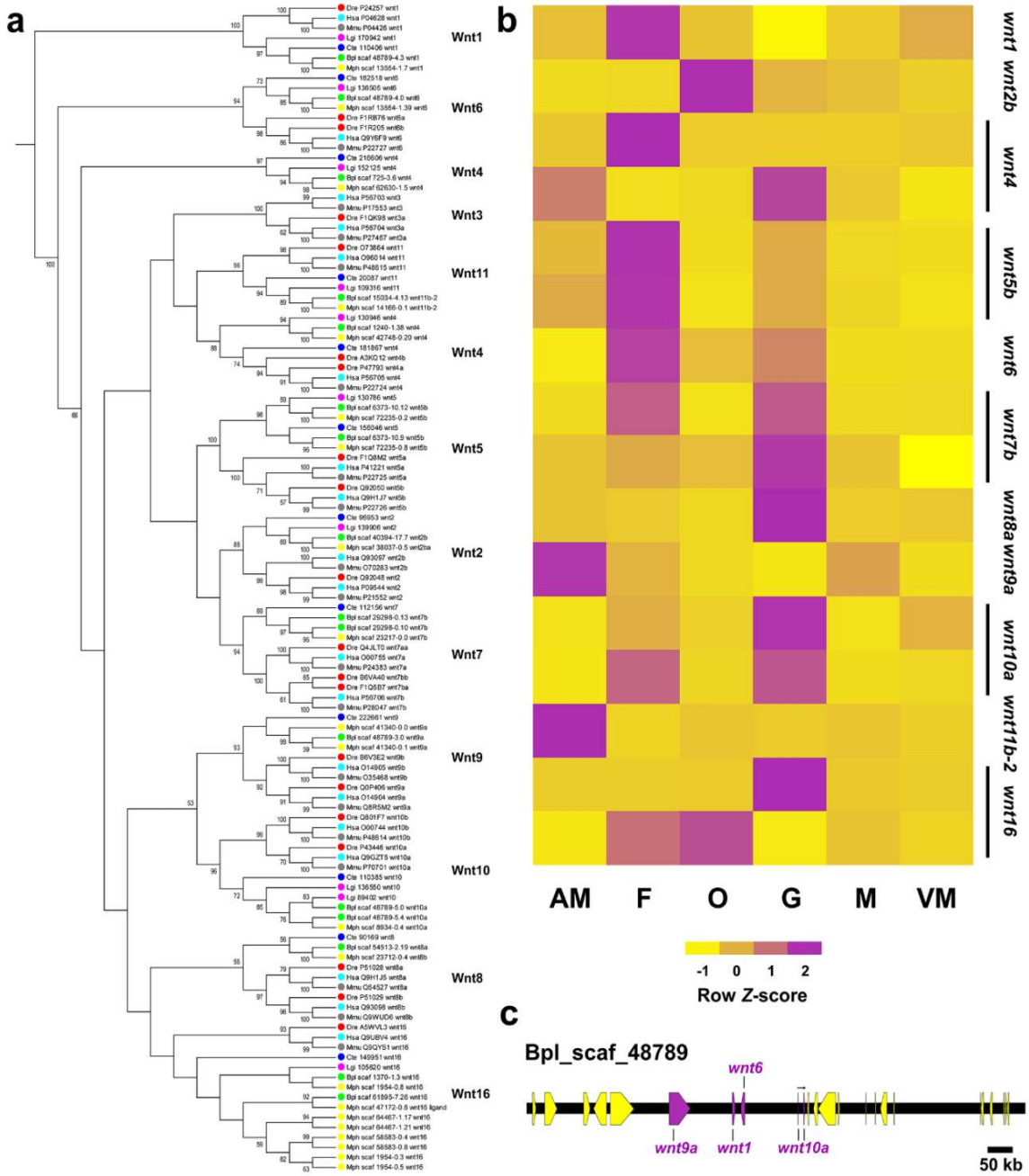


Fig. S19. Genes in the Wnt signalling pathway in *B. platifrons*. (a) Gene genealogy of *wnt* genes with typical *wnt* genes from Uniprot database. (b) gene expression profile of all of the 16 *wnt* genes in *B. platifrons*. (c) Genomic structure of *wnt9-wnt1-wnt6-wnt10*. Abbreviations: AM, adductor muscle; F, foot; O, ovary; G, gill; M, mantle; and VM, visceral mass.

18.4.2 Toll-like receptor 13

Toll-like receptors (TLRs) are are a group of critical innate immune recognition genes. Among them are TLR13, endosomal receptors in various groups of animals that can sense microbes only after their internalization by phagocytosis (Biondo et al., 2012). We found that TLR 13 gene families including 59 genes were significantly expanded in *B. platifrons*. Analyzing these TLR13 genes revealed that they exclusively contained the canonical Toll/Interleukin receptor (TIR) domain and the leucine-rich repeat (LRR) domain. The TLR 13 receptors were not only greatly expanded in the genome of *B. platifrons*, but also highly expressed in the gill (Fig. S20), suggesting their involvement in immune responses to symbiotic bacteria in the gill

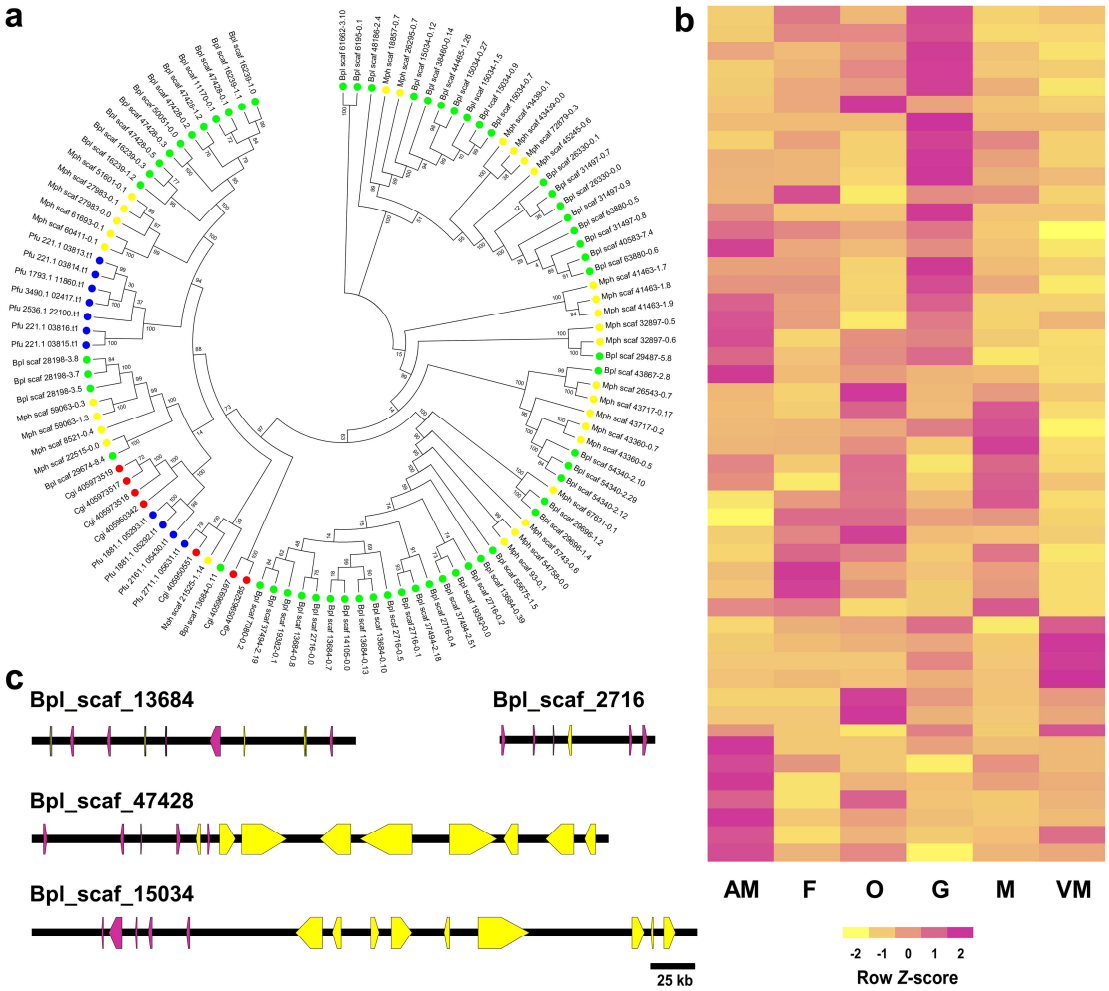


Fig. S20. Toll-like receptor (TLR) 13 gene family. (a) unrooted gene genealogy of TLR13 genes in Bpl (green), Mph (yellow), Cgi (red) and Pfu (blue) showing the

expansion of this gene family in the deep-sea mussel. (b) gene expression profile of all of the TLR13 genes in *B. platifrons*. (c) Genomic structure of a TLR13 gene containing scaffold, indicating tandem duplication of these genes. The TLR13 genes are shown in red, and other genes are shown in yellow. Abbreviations: AM, adductor muscle; F, foot; O, ovary; G, gill; M, mantle; and VM, visceral mass.

18.4.3 Syndecan and Protocadherin

Syndecan, a heparan sulfate proteoglycan protein, was shown to initiate raft-dependent endocytosis upon clustering (Chen and Williams 2013). A total of 143 syndecan genes were identified in the *B. platifrons* genome, the largest number of this family of genes among the 11 lophotrochozoan species examined. Many of them were highly expressed in the gill (Fig. S21), suggesting their active involvement in endocytosis of the symbiont.

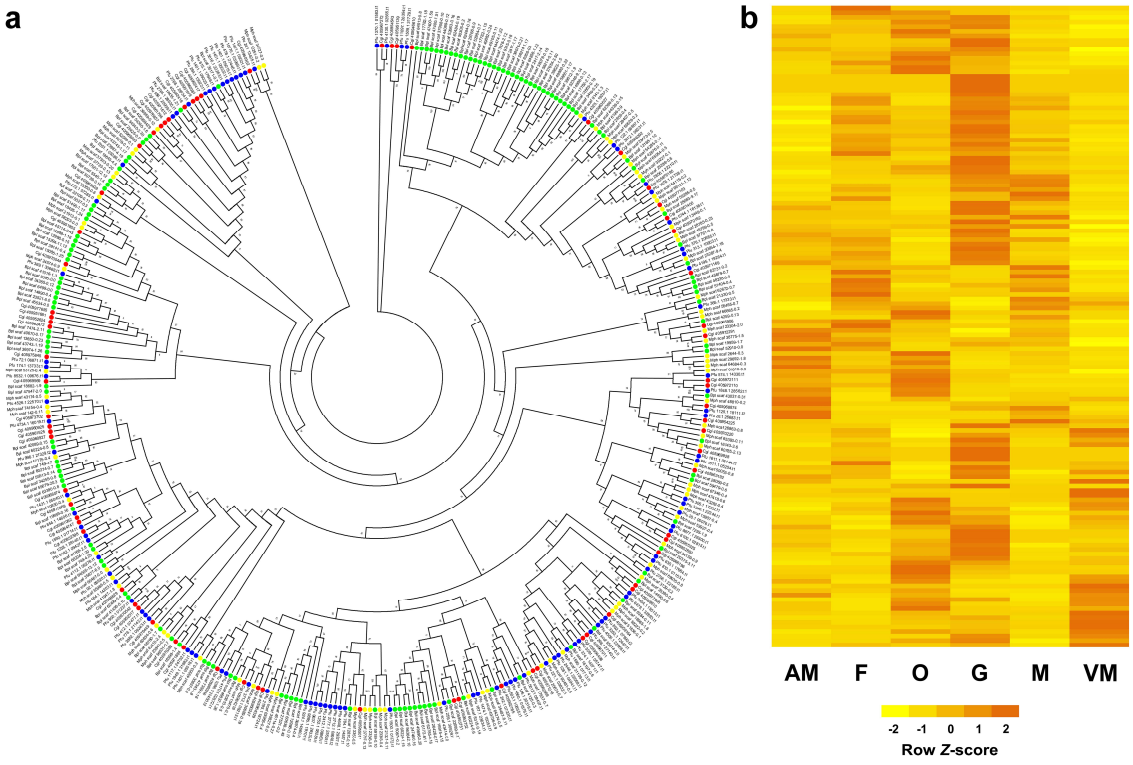


Fig. S21. Expansion of the Syndecan family in *Bathymodiolus platifrons* (Bpl). (a) an unrooted gene genealogy of Syndecan genes in Bpl (green), Mph (yellow), Cgi (red) and Pfu (blue); (b) A heat map of expression of Syndecan genes in six tissues.

Abbreviations: Mph, *Modiolus philippinarum*, Pfu, *Pinctada fucata*; Cgi, *Crassostrea gigas*; AM, adductor muscle; F, foot; O, ovary; G, gill; M, mantle; and VM, visceral mass.

Cadherin plays important roles in intercellular adhension and cell-cell communication, and protocadherins are a subgroup of cadherins (Buchanan et al. 2010). The expression of cadherins had positive correlation with endocytosis intensity in mammalian cells, indicating the coupling between this family of proteins and endocytosis intensity (de Beco et al. 2009). Also, cadherins could interact with ß-catenins, which is another way of the involvement of the Wnt signalling pathway in cell-cell interaction (Akiyama et al. 2004). A total of 102 protocadherins (the largest number in 11 species examined) were identified in the *B. platifrons* genome and their expressions were high in the gill (Fig. S22), suggesting their involvement of the symbiont endocytosis regulation.

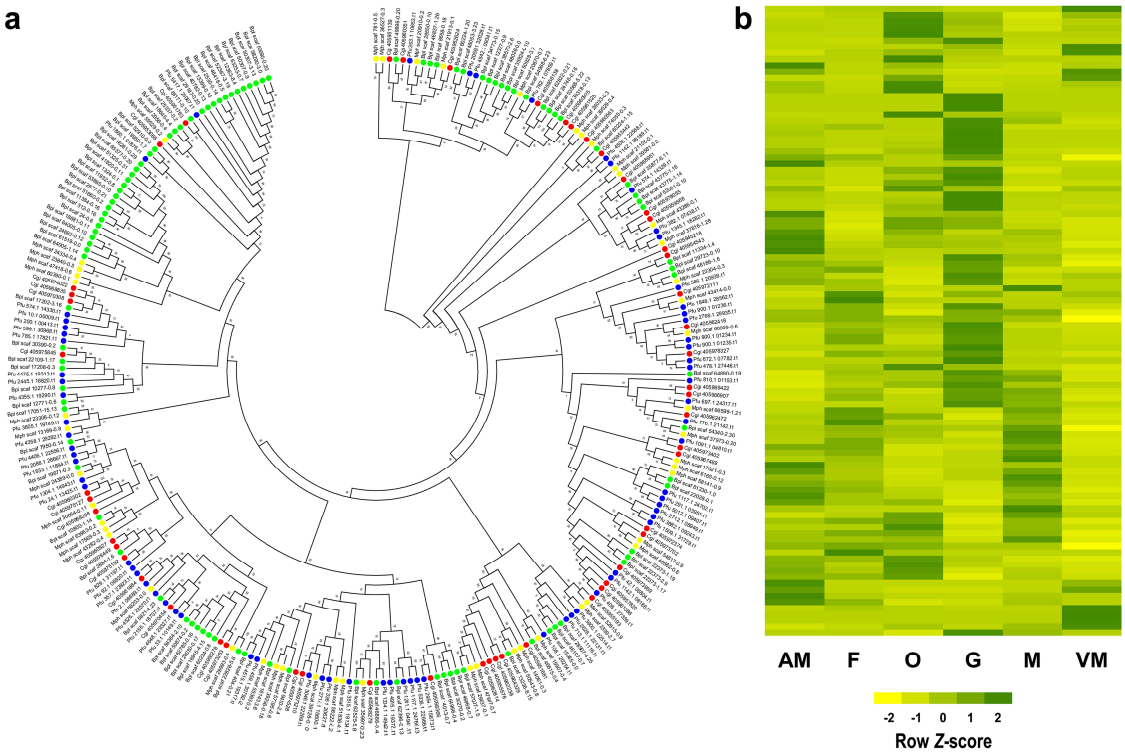


Fig. S22. Expansion of the Protcadherin family in *Bathymodious platifrons* (Bpl). (a) an unrooted gene genealogy of Protcadherin genes in Bpl (green), Mph (yellow), Cgi (red) and Pfu (blue); (b) A heat map of expression of Protcadherin genes in six tissues. Abbreviations: Mph, *Modiolus philippinarum*, Pfu, *Pinctada fucata*; Cgi, *Crassostrea*

gigas; AM, adductor muscle; F, foot; O, ovary; G, gill; M, mantle; and VM, visceral mass.

18.4.4 Ras gene family.

The Ras family of small guanosine triphosphatases (GTPases) is evolutionary conserved and functions in a variety of cell signalling pathways such as apoptosis, cell growth and cell differentiation. This gene family can be divided into five major subfamilies based on sequence variation and function: Ras, Rho, Rab, Ran and Arf (Wennerberg et al. 2005). The Ras gene family can regulate membrane complex formation (Rocks et al. 2006). Proteins in the subfamily Rab have recently shown as a key component of endocytosis and exocytosis. For example, Rab5, Rab7 and Rab11 are known to control intracellular vesicle trafficking, thereby preventing phagosomal maturation and eventually maintaining a stable relationship between cnidarian host and its symbiotic dinoflagellate (Chen et al. 2003; Chen, et al. 2004; Chen et al. 2005). There were 418 Ras genes in *B. platifrons*, the highest number of Ras among the 11 species of lophotrochozoans examined. Among them, 102 Ras proteins had mass spectrometry support. Many of the Ras proteins were highly expressed in the gill, suggesting their active roles in the interaction with the symbiont (Fig. S23a). Also, at least five scaffolds had obvious tandem duplications, with each scaffold containing 5 to 10 Ras genes (Fig. S23b).

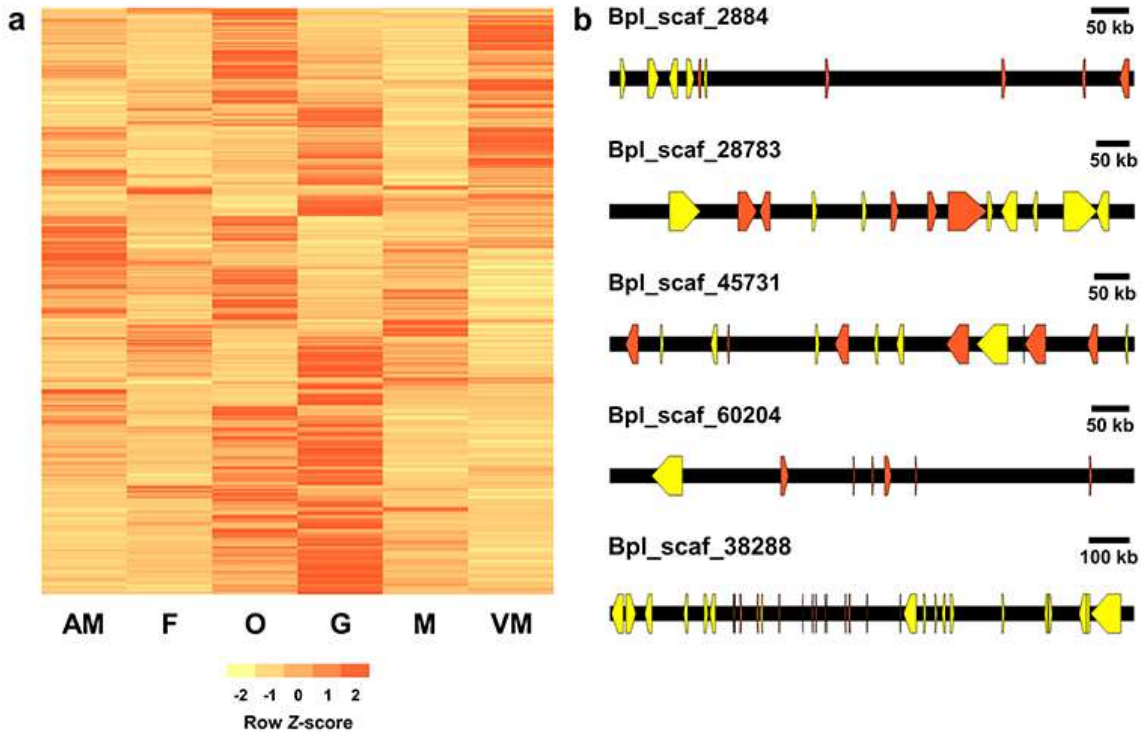


Fig. S23. Ras gene family in *B. platifrons*. (a) Expression profile of 418 Ras genes in six tissues. (b) Genomic structure of five Ras containing scaffolds, indicating a tandem duplication of Ras genes. The Ras genes are shown in red, and other genes are shown in yellow. Abbreviations: AM, adductor muscle; F, foot; O, ovary; G, gill; M, mantle; and VM, visceral mass.

18.5 Apoptosis as a means to control symbiont population

Apoptosis plays critical roles of eliminating pathogens and controlling cell survival. In coral-dinoflagellate symbiosis, the host apoptosis pathway was triggered when uptaking the undesirable algal type (Dunn and Weis 2009). Apoptosis was also documented in the juvenile stage of deep-sea vestimentiferan tubeworms, when the symbiotic bacteria colonising the host (Nussbaumer et al. 2006). In the squid-*vibrio* symbiosis system, apoptosis of epithelial cells was induced by the introduction of *V. fischeri* lipopolysaccharide (Nyholm and McFall-Ngai 2004). Searching our assembled mussel genomes revealed many apoptosis related functional domains. The number of genes in most of these domains of *B. platifrons* were higher than those in other 11 lophotrochozoan species examined, and five of these domains were significantly different in abundance (i.e., caspase, IAP repeat, TNF, DEATH and CARD) (see Table 1 for more details). Collectively, these results emphasised the

importance of apoptosis in initiating the host-symbiont relationship and controlling the endosymbiont load.

There were significant apoptosis related gene family expansions in *B. platifrons*, including 155 Death genes, 75 caspase genes and 130 IAP genes (Fig. S24 and Fig. S25). Most of them were highly expressed in the gill (Fig. 4). Among them, caspase 8, a central controller of programmed cell death, is activated through the extrinsic pathway involving members of the TNF superfamily and their DEATH receptor to initiate the apoptotic pathway (Romero et al. 2015). TUNEL assay showed the presence of numerous apoptotic cells in *Bathymodiolus thermophiles*, a vent species that harbours symbiotic sulfur-oxidizing bacteria (SOB) (Guezi et al. 2013). Therefore, the two species of *Bathymodiolus* harbouring different types of bacteria appeared to control their symbionts through an identical apoptosis pathway. In addition, two ceramide derivatives (i.e., bathymodiolamides A and B) isolated from the gill of vent mussel *Bathymodiolus thermophilus* exhibited strong apoptosis induction activity (Andrianasolo et al. 2011), indicating they were candidates for initiating apoptosis process in the gills. Also, at least one scaffold had obvious tandem duplications, containing 6 genes (Fig. S25b).

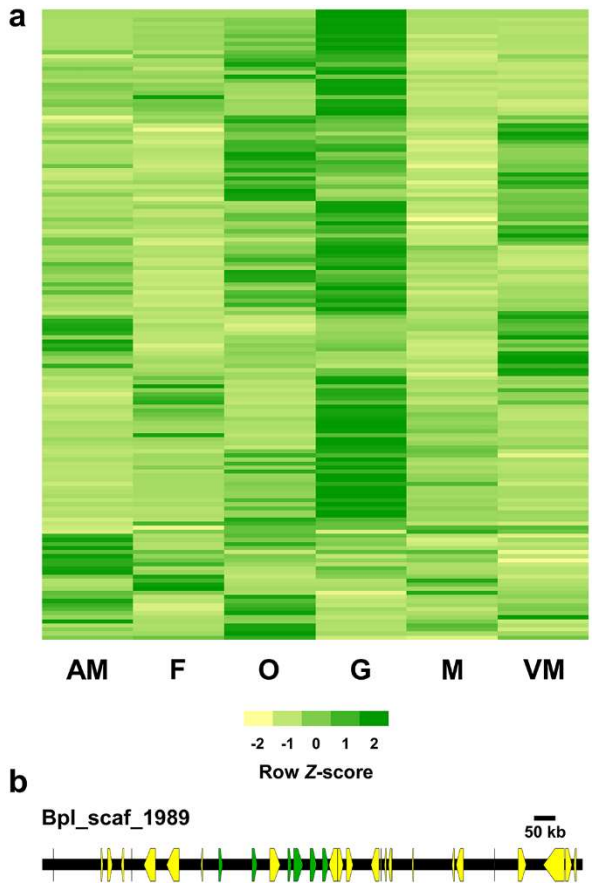


Fig. S24. DEATH domain containing proteins in *B. platifrons*. (a) A heat map showing the expression level of DEATH genes in six tissues. (b) Genomic structure of a DEATH containing scaffold, indicating tandem duplication of these genes. The DEATH genes are shown in green, and other genes are shown in yellow. Abbreviations: AM, adductor muscle; F, foot; O, ovary; G, gill; M, mantle; and VM, visceral mass.

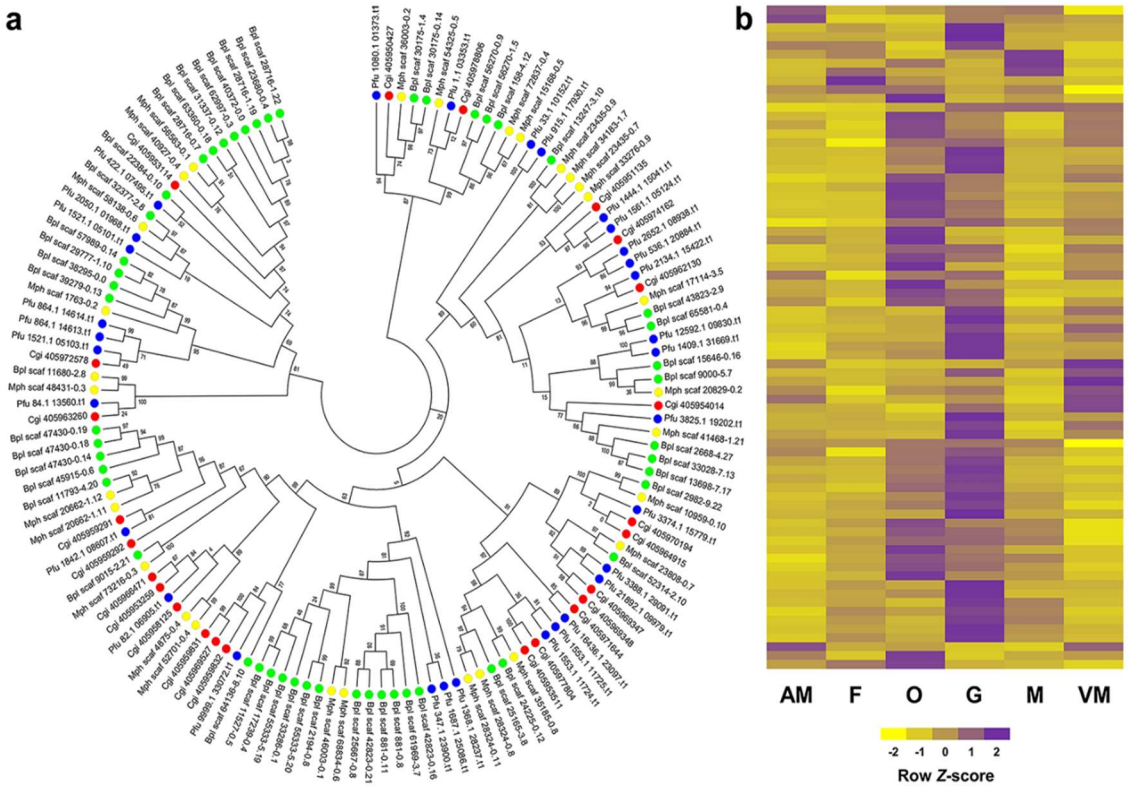


Fig. S25. The caspase family in *B. platifrons* (Bpl). (a) unrooted gene genealogy of protcadherin genes in Bpl (green), Mph (yellow), Cgi (red) and Pfu (blue) showing the expansion of this gene family in the deep-sea mussel. (b) A heat map of expression of caspase genes in six tissues. Abbreviations: Mph, *Modiolus philippinarum*; Cgi, *Crassostrea gigas*; Pfu, *Pinctada fucata*; AM, adductor muscle; F, foot; O, ovary; G, gill; M, mantle; and VM, visceral mass.

18.6 Lysozymes

Lysozymes are glycoside hydrolases that damage bacterial cell walls. They are used extensively in host defense (Xue et al. 2010) and regulating symbiont population (Moya et al. 2008). In the deep-sea mussel *Bathymodiolus azoricus*, the increase in activity and gene expression of three lysozymes corresponded to the decline in symbiont population in a transplant experiment in the field, clearly indicating the function of lysozymes in controlling the symbiont population (Detree et al. 2016). The present study showed that *B. platifrons* had six lysozyme genes, and in three of them the expression levels were highest in the gill (Fig. S26), suggesting their involvement in symbiont population control.

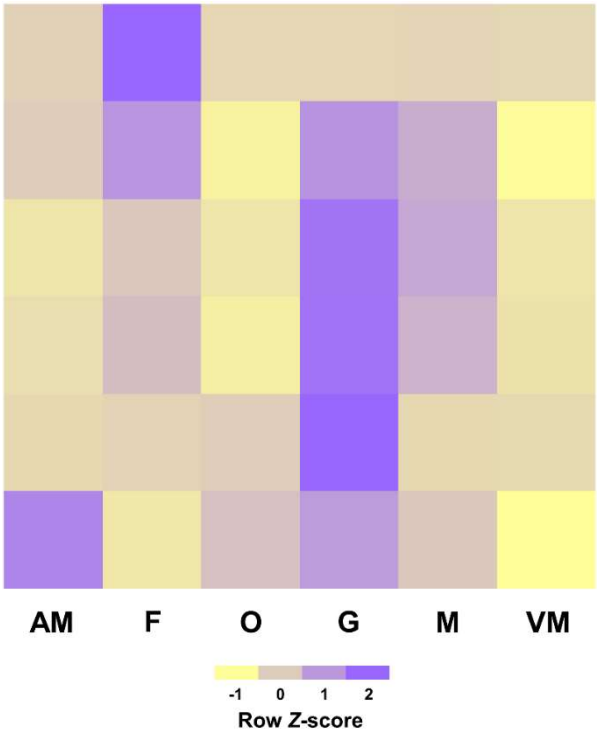


Fig. S26. Expression profile of six lysozyme genes in *B. platifrons*. Abbreviations: AM, adductor muscle; F, foot; O, ovary; G, gill; M, mantle; and VM, visceral mass.

19. Taxonomically restricted genes (TRG)

TRGs have been suggested to be crucial in adaptive evolution (Kaessmann 2010). The protein sequences of *B. platifrons* without any homolog to both NCBI nr database and Swiss-prot database were searched against *M. philippinarum* protein sequences using BLASTp with a threshold value of 1e-5. This resulted in 3067 TRGs in *B. platifrons* (Table S14). Among them, 2193 genes were supported by transcriptomic evidence, and 48 were supported by proteomic evidence. Analysis of the InterPro Scan result showed that 261 of the TRGs had signal peptides, indicating these proteins were potentially secreted proteins. Meanwhile, 776 of them had a transmembrane domain, indicating membrane proteins potentially engaged novel gene formations Observation of the gene expression profiles of these 2193 TRGs showed that 250 (11.4%), 159 (7.3%), 482 (22.0%), 846 (38.6%), 192 (8.8%) and 264 (12.0%) of them had the highest expression level in adductor muscle, foot, ovary, gill, mantle and visceral mass, respectively. The gill had the highest number of highly expressed TRGs, which

indicated that the gill might have contributed to adaptation to the deep-sea environment, consistent with the functions of the gill in interacting with the environment and hosting symbiotic bacteria.

20. Proteome characterisation

Gill samples of three individuals of *B. platifrons* fixed separately in RNAlater were used in proteomic analysis following our published method (Sun et al. 2012). In brief, the samples were homogenized in a lysis buffer containing 8M urea and 40 mM HEPES (pH = 7.5), sonicated to break up cells and high molecular weight DNA and centrifuged under 4°C at 15,000 g for 15 min. The supernatant containing proteins was collected, purified with a methanol/chloroform protein precipitation method (Wessel and Flügge 1984) and re-constituted in lysis buffer. Protein quality and quantity were determined by SDS-PAGE and RC-DC protein assay, respectively (Sun et al. 2012). The three samples were then pooled in equal amount, 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 on a Waters 2695 HPLC with the following gradient: 100% buffer A (10 mM KH_2PO_4 and 20% ACN, pH = 3.0) for the first 5 min, 0-30% buffer B (10 mM KH_2PO_4, 0.5 M KCl and 20% ACN, pH = 3.0) for the next 28 min, 30−100% buffer B for another 5 min, 100% buffer B for 5 min and a final 100% Buffer A for 7 min (Mu et al. 2015). The fractions were collected every 1 min and some of them containing low abundance of peptides were pooled. The resultant 30 fractions were desalted and dried.

Mass spectrometry was conducted by following a method used in human proteome map analysis to maximize the number of identified proteins (Wilhelm et al. 2014). Each SCX fraction was re-dissolved with 0.1% formic acid. Approximate 500 ng peptides were analyzed on a LTQ-Orbitrap Elite mass spectrometer coupled with an Easy-nLC (Thermo Fisher, Bremen, Germany) equipped with an Acclaim PepMap

RSLC C18 column (50 μm X 25 cm, with 2 μm particle sizes and 100 Å pore sizes) (Thermo Fisher Scientific, Sunnyvale, CA USA). The LC gradient was as follows: 100% Solution A (0.1% formic acid) for 5 min, 4−32% (linear changing gradient) Solution B (0.1% formic acid in ACN) for 105 min, 32%−98% Solution B for 10 min, 98% Solution B for 5 min and a final 5 min of 100% Solution A to equilibrate the column. The flow rate was set as 250 nL/min. The mass spectrometry scan range was 360−1300 m/z with a resolution of 30,000 in the positive charge mode. Top 15 high abundant multiple-charge precursors were chosen for fragmentation using high-energy collision-induced dissociation (HCD). The MS/MS scan settings were as follows: normalized collision energy = 30%; automatic gain control target value = 2×10^4; resolution = 15,000; max injection time = 100 ms; dynamic exclusion = 30 s; and precursor ion isolation width = 2.0 Th.

Raw MS/MS data were converted to *mgf* format files using Proteome Discoverer v1.3.0.339. All of the *.mgf* files were merged and submitted to Mascot v2.3.2 (Matrix Sciences, London, U.K.) against the deduced *B. platifrons* protein sequences, with their reversed sequences used to calculate the false discovery rate (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. Peptides with a score lower than 27, equivalent to the 5% significance threshold, and peptides with less than 8 amino acids were removed. The FDR was dynamically controlled at 0.01. Protein abundance was quantified using the semi-quantitative emPAI method (Ishihama et al. 2005).

At a false positive rate of 1%, we identified 5527 proteins with an average number of significant peptide match of 10.7 and an average coverage of 13.9% (Table S15). Analysis of the abundance of these 5527 protein abundance and their corresponding gene expression revealed high positive correlation ($P < 1e\text{-}5$) (Fig. S27).

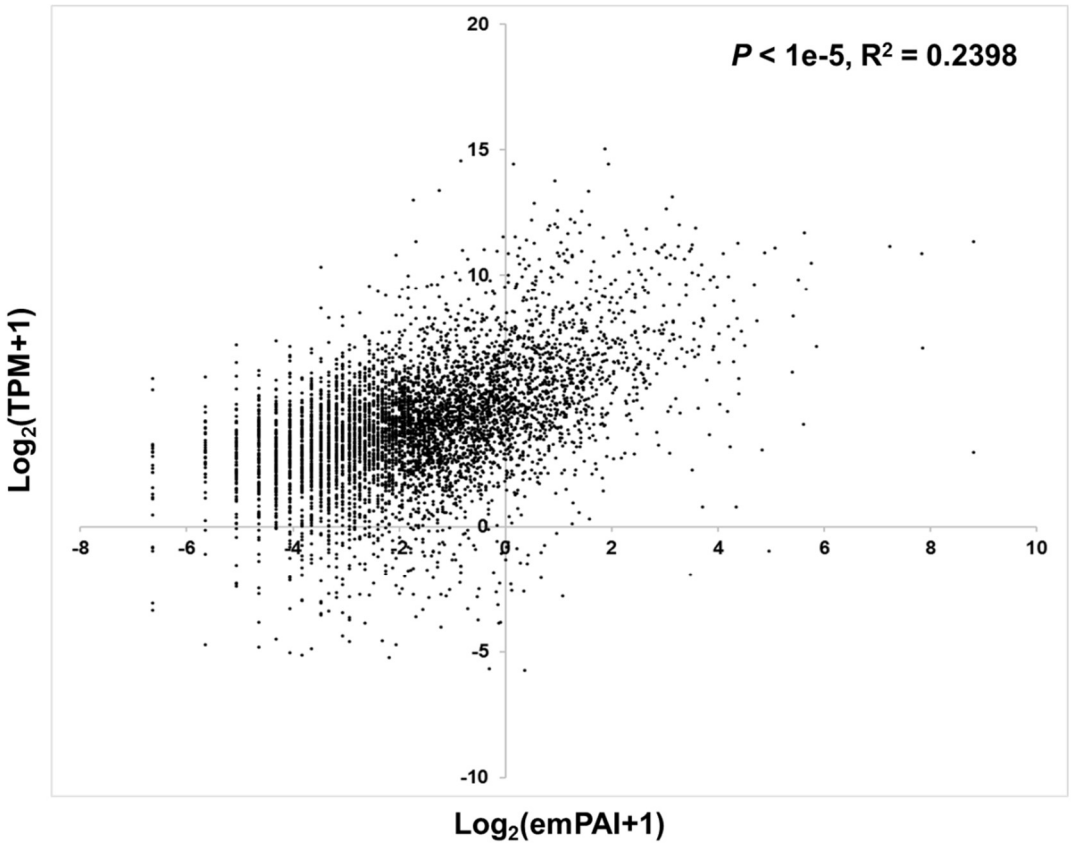


Fig. S27. Correlation analysis of protein expression and RNA-Seq gene expression. Protein abundance was quantified based on the emPAI peptide count method, while gene expression was quantified using TPM.

This set of proteomic data further indicated the high quality of our good gene model prediction. The proteins identified here could be used for target proteomic analysis in the future as well as analysis of the functional composition of symbiont proteins in the gill described below.

Among the top 50 highly expressed proteins identified (Table S16) were several functional groups that are involved in adaptation to the deep-sea environment. The high expression of several heat shock proteins (heat shock cognate 70, heat shock 20 and heat shock cognate 70), disulfide isomerase, 14-3-3 epsilon and peroxiredoxin indicated their involvement in protein structure stabilisation and detoxification, which should be advantageous in the deep-sea environment. The high expression of several immune related proteins (fibrinogen, C-type lectin 6 and C-type lectin 6) might be

1455 related to their active involvement in defense against pathogenic bacteria, as well as
1456 control of endosymbionts.
1457
1458
1459

Table S16. Fifty most abundant proteins in the gill of *B. platifrons*. Abbreviations: Prot_cov., percentage of protein sequences covered by matched peptides; G, gill; TPM, transcripts per million.

Protein	Description	Score	No. of Peptides matched	Prot_ cov.	emPAI	G_TPM
Bpl_scaf_58211-0.2	histone H4-like	8789	308	53.4	449.58	7.69377
Bpl_scaf_20231-3.3	heat shock cognate 70, partial	1414	48	90.2	449.58	2586.1
Bpl_scaf_34414-0.3	THY domain-containing -1, partial	2042	72	66.5	232.26	134.753
Bpl_scaf_38338-12.2 2	Novel protein	4024	95	69.2	229.16	1887.61
Bpl_scaf_33596-7.3	carbonic anhydrase 1-like	10276	325	60.5	150.78	2282.97
Bpl_scaf_18893-4.53	tubulin beta-2C chain	1540	54	75.5	57.74	140.263
Bpl_scaf_21209-0.8	actin, cytoplasmic	13753	577	58.2	54.06	1452.43
Bpl_scaf_57259-5.3	cytoplasmic intermediate filament A	7939	232	54.5	50.69	693.939
Bpl_scaf_36595-6.57	actin	13708	582	58.2	49.53	3344.42
Bpl_scaf_41685-0.4	neurocan core	1473	55	44.2	48.94	16.6537
Bpl_scaf_36595-6.55	actin	13073	568	58.2	45.86	907.305
Bpl_scaf_31966-3.37	actin	12606	554	58.2	42.88	325.932
Bpl_scaf_25704-1.3	fibrinogen 1	1192	37	62.2	42.35	69.8604
Bpl_scaf_48274-0.3	carbonic anhydrase 1-like	8581	243	60.3	33.7	2178.87
Bpl_scaf_36595-7.55	actin	12158	545	56.4	31.14	0.510106
Bpl_scaf_27238-0.4	peroxiredoxin 6	729	28	59.6	29.27	1918.99
Bpl_scaf_11635-1.9	heat shock 20	830	35	65.9	28.57	8.21971
Bpl_scaf_24343-1.41	small nuclear ribonucleo Sm D2	942	28	47.9	26.72	285.212
Bpl_scaf_18893-5.11	tubulin beta-4B chain	6302	198	46	25.59	793.445
Bpl_scaf_8188-0.18	disulfide isomerase	6692	218	64.2	22.84	145.022
Bpl_scaf_56904-9.39	Barrier-to-autointegration factor	707	24	47.3	21.66	216.168
Bpl_scaf_62441-3.7	persulfide dioxygenase ETHE1, mitochondrial-like isoform X1	1967	62	66.7	21.07	58.0492
Bpl_scaf_43573-4.9	calumenin-like isoform X2	2172	76	53.8	20.88	38.3499
Bpl_scaf_44465-0.53	cytoplasmic actin	11798	528	56.5	20.78	2477.98
Bpl_scaf_30989-2.24	Calmodulin	981	28	69.3	20.52	308.754
Bpl_scaf_38133-0.8	C-type lectin 6	1056	40	36.5	20.34	1.71048
Bpl_scaf_10139-14.6	cytochrome c oxidase subunit 5B, mitochondrial-like	1115	36	63.8	19.73	501.912
Bpl_scaf_63657-2.1	14-3-3 epsilon	3044	96	48.8	19.64	173.468
Bpl_scaf_1759-1.0	hypothetical protein LOTGIDRAFT_165383	814	37	66.2	19.59	93.8496
Bpl_scaf_32287-1.5	galactose-specific lectin nattectin-like	1197	44	38.7	18.84	9.076
Bpl_scaf_5538-4.10	Tropomyosin	2776	84	46.8	18.39	245.259

Bpl_scaf_28622-0.8	heat shock cognate 70	5265	168	43.6	18.28	993.506
Bpl_scaf_1776-0.0	histone H2B	3054	101	32.7	18.06	23.4883
Bpl_scaf_5285-5.10	dynein light chain 2, cytoplasmic	622	20	58.4	17.05	1878.39
Bpl_scaf_43487-0.0	C-type lectin 9	851	40	39.4	16.42	29.2738
Bpl_scaf_25667-1.22	Myosin regulatory light chain sqh	882	32	52.1	16.22	90.8487
Bpl_scaf_54882-2.12	14-3-3 zeta isoform X1	2352	83	40.3	16.15	201.54
Bpl_scaf_21636-3.22	60S ribosomal L30	739	21	64	15.93	736.359
Bpl_scaf_38703-10.8	Tubulin beta chain	4701	159	49.3	15.25	436.893
Bpl_scaf_46480-0.7	peptidyl-prolyl cis-trans isomerase A	1062	34	63.4	15.2	996.432
Bpl_scaf_33270-1.25	myosin heavy chain, non-muscle isoform X11	14353	378	48.6	15.11	141.235
Bpl_scaf_34427-9.6	retinal dehydrogenase 2-like	3588	98	53.9	15	128.685
Bpl_scaf_5516-0.0	Novel protein	857	33	47	14.47	901.32
Bpl_scaf_19132-0.0	histone H2B	2846	84	55.2	14.25	12.4868
Bpl_scaf_51518-1.0	EF-hand domain-containing D2-like	532	19	58.3	13.91	56.667
Bpl_scaf_13985-0.24	ependymin-related 1-like	2031	65	40.2	13.29	691.073
Bpl_scaf_44507-2.13	transgelin -3	1083	35	62.2	13.28	284.658
Bpl_scaf_20376-10.6	Sarcoplasmic calcium-binding	1220	49	50	13.1	1.72614
Bpl_scaf_32538-11.39	c-binding -like isoform X2	978	40	53.6	13.08	308.687
Bpl_scaf_64340-0.2	calreticulin	3808	128	53.7	12.89	1360.96

Since the gills of *B. platifrons* harbour methane-oxidizing bacteria (MOB) (Barry et al. 2002; Lorion et al. 2013) belonging to the family Methylococcaceae, we also identified and quantified the bacterial proteins in the gill samples. Protein sequences from 24 published bacterial genomes within Methylococcaceae were downloaded from NCBI and appended with 1402 bacterial proteins from our previous meta-transcriptome analysis (Wong et al. 2015). This created a database of 86,023 bacterial proteins. The database was searched using criteria 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 1e-07 (Kanehisa et al. 2016).

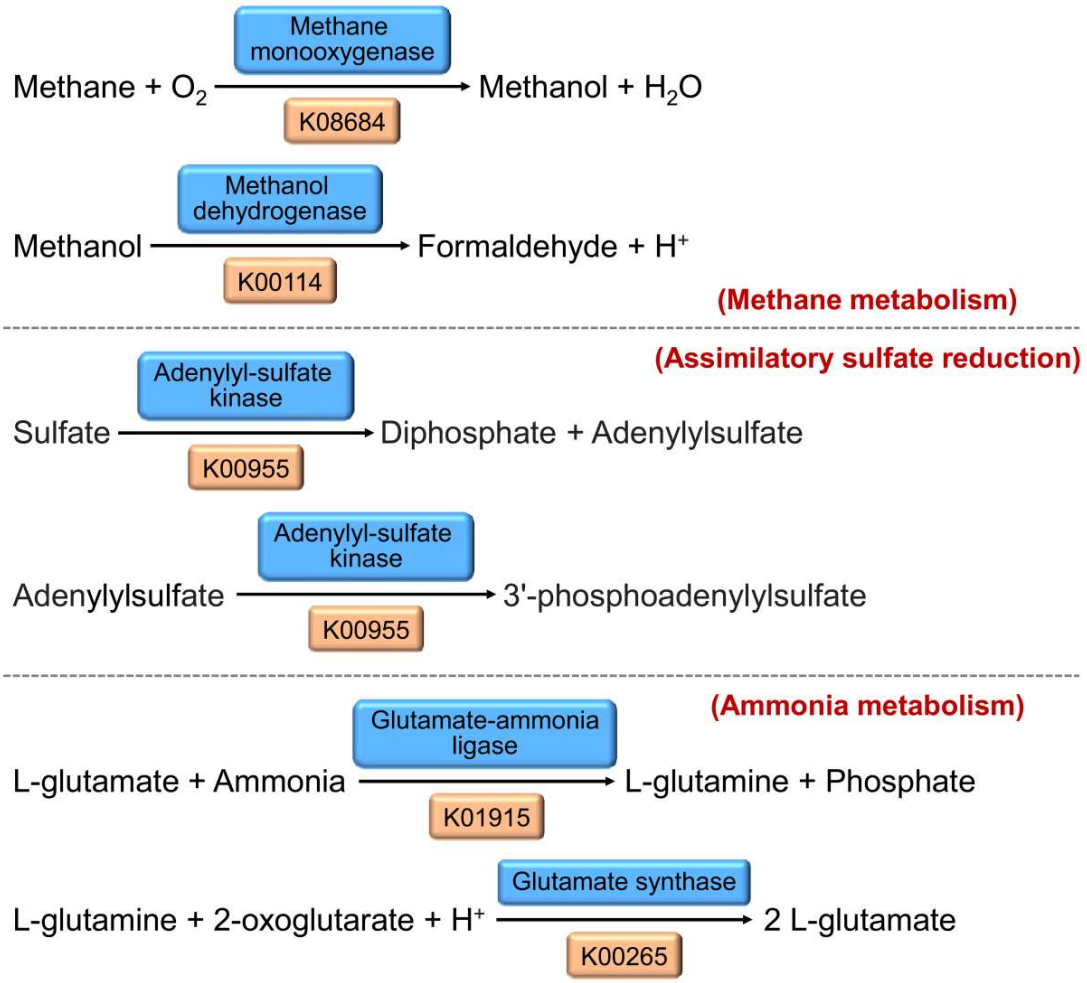


Fig. S28. Key reactions in the biochemical synthesis pathways of MOB in the gill of *B. platifrons*. The number in highlighted in yellow background is the KO number annotated by searching the KEGG database. These reactions point to three important metabolic pathways (i.e., methane metabolism, assimilatory sulfate reduction and ammonia metabolism).

Among the 5527 proteins identified in the gill samples of *B. platifrons* in our meta-proteomic analysis, 220 were putative methane oxidizing bacteria (MOB) proteins (Table S17). Functional annotation of these proteins revealed three major metabolic pathways in the symbiont (Fig. 3 and Fig. S28). The methanotrophic pathway, identified by the high expression of two signature enzymes of methane oxidation (i.e., methane monooxygenase and methanol dehydrogenase), provides fuel to drive the host-symbiont relationship (DeChaine and Cavanaugh 2006). The assimilatory sulfate reduction pathway, identified by several adenylyl-sulfate kinases,

produces sulfur-containing amino acids that are needed for synthesis of proteins (Canfield and Des Marais 1991). Previous studies have indicated the potential involvement of the glutamine synthetase/glutamate dehydrogenase pathway in nitrogen assimilation by the MOX symbiont (Dechaine and Cavanaugh 2006). We identified two key enzymes of this pathway (i.e., glutamate-ammonia ligase and glutamate synthase), providing direct evidence to support the involvement of this pathway in nitrogen metabolism in *Bathymodiolus*. This pathway could have two benefits to the symbiont and host: 1) removal of ammonia, which could reach a level that is toxic to the cell in vent and seep fluids; and 2) obtaining nitrogenous nutrient that is scarce in the deep-sea environment for synthesis of various large molecules. Overall, the identification of enzymes in these critical metabolic pathways in the symbiont sheds light on their metabolic benefits to the host that live in an environment with very little, if any, photosynthesis-derived nutrients.

References

Akiyama H, Lyons JP, Mori-Akiyama Y, Yang X, Zhang R, Zhang Z, Deng JM, Taketo MM, Nakamura T, Behringer RR, McCrea PD, de Crombrugghe B. 2004. Interactions between Sox9 and β-catenin control chondrocyte differentiation. *Genes & Development* **18**:1072–1087.

Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, Brenner S, Ragsdale CW, Rokhsar DS. 2015. The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature* **524**:220–224.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**:403–410.

Andrade SLA, Einsle O. 2007. The Amt/Mep/Rh family of ammonium transport proteins. *Molecular Membrane Biology* **24**:357–365.

Andrianasolo EH, Haramaty L, McPhail KL, White E, Vetriani C, Falkowski P, Lutz R. 2011 Bathymodiolamides A and B, ceramide derivatives from a deep-sea hydrothermal vent invertebrate mussel, *Bathymodiolus thermophilus*. *Journal of Natural Products* **74**:842–846.

Arellano SM, Van Gaest AL, Johnson SB, Vrijenhoek RC, Young CM. 2014. Larvae from deep-sea methane seeps disperse in surface waters. *Proceedings of the Royal Society B: Biological Sciences* **281**:20133276.

Arellano SM, Young CM. 2009. Spawning, development, and the duration of larval life in a deep-sea cold-seep mussel. *The Biological Bulletin* **216**:149–162.

Barry JP, Buck KR, Kochevar RK, Nelson DC, Fujiwara Y, Goffredi SK, Hashimoto J. 2002. Methane-based symbiosis in a mussel, *Bathymodiolus platifrons*, from cold seeps in Sagami Bay, Japan. *Invertebrate Biology* **121**:47–54.

Baumgarten S, Simakov O, Esherick LY, Liew YJ, Lehnert EM, Michell CT, Li Y, Hambleton EA, Guse A, Oates ME, et al. 2015. The genome of *Aiptasia*, a sea anemone model for coral symbiosis. *Proceedings of the National Academy of Sciences of the United States of America* **112**:11893–11898.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**:289–300.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**:573–580.

Benton MJ, Donoghue PCJ, Asher RJ. In The Timetrees of Life: Calibrating and constraining molecular clocks. (eds Hedges SB & Kumar S). 35–86 (Oxford Univ Press, New York, 2009).

Benton MJ, Donoghue PCJ, Asher RJ, Friedman M, Near TJ, Vinther J. 2015. Constraints on the timescale of animal evolutionary history. *Palaeontologia Electronica* **18.1.1FC**:1–106.

Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, Pütz J, Middendorf M, Stadler PF. 2013. MITOS: Improved *de novo* metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution* **69**:313–319.

Bettencourt R, Pinheiro M, Egas C, Gomes P, Afonso M, Shank T, Santos R. 2010. High-throughput sequencing and analysis of the gill tissue transcriptome from the deep-sea hydrothermal vent mussel *Bathymodiolus azoricus*. *BMC Genomics* **11**:559.

Bettencourt R, Rodrigues M, Barros I, Cerqueira T, Freitas C, Costa V, Pinheiro M, Egas C, Santos RS. 2014. Site-related differences in gene expression and bacterial densities in the mussel *Bathymodiolus azoricus* from the Menez Gwen and Lucky Strike deep-sea hydrothermal vent sites. *Fish & Shellfish Immunology* **39**:343–353.

Biondo C, Mancuso G, Beninati C, Iaria C, Romeo O, Cascio A, Teti G. 2012. The role of endosomal Toll-like receptors in bacterial recognition. *European Review for Medical and Pharmacological Sciences* **16**:1506–1512.

Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. *Genome Biology* **13**:R56.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**:2114–2120.

Bond JE, Garrison NL, Hamilton CA, Godwin RL, Hedin M, Agnarsson I. 2014. Phylogenomics resolves a spider backbone phylogeny and rejects a prevailing paradigm for Orb Web Evolution. *Current Biology* **24**:1765–1771.

Boore JL. 1999. Animal mitochondrial genomes. *Nucleic Acids Research*, **27**:1767–1780.

Boore JL, Medina M, Rosenberg LA. 2004. Complete sequences of the highly rearranged molluscan mitochondrial genomes of the Scaphopod *Graptacme eborea* and the bivalve *Mytilus edulis*. *Molecular Biology and Evolution*, **21**:1492–1503.

Boutet I, Jollivet D, Shillito B, Moraga D, Tanguy A. 2009. Molecular identification of differentially regulated genes in the hydrothermal-vent species *Bathymodiolus thermophilus* and *Paralvinella pandorae* in response to temperature. *BMC Genomics* **10**:222.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**:525–527.

Breusing C, Johnson S, Tunnicliffe V, Vrijenhoek R. 2015. Population structure and connectivity in Indo-Pacific deep-sea mussels of the *Bathymodiolus septemdierum* complex. *Conservation Genetics* **16**:1415–1430.

Buchanan SM, Schalm SS, Maniatis T. 2010. Proteolytic processing of protocadherin proteins requires endocytosis. *Proceedings of the National Academy of Sciences of United States of America* **107**:17774–17779.

Canfield D, Des Marais D. 1991. Aerobic sulfate reduction in microbial mats. *Science* **251**:1471–1473.

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M. 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research* **18**:188–196.

Chen K, Williams KJ. 2013. Molecular mediators for raft-dependent endocytosis of syndecan-1, a highly conserved, multifunctional receptor. *Journal of Biological*

Chemistry **288**:13988–13999.
Chen M-C, Cheng Y-M, Hong M-C, Fang L-S. 2004. Molecular cloning of Rab5 (ApRab5) in *Aiptasia pulchella* and its retention in phagosomes harboring live zooxanthellae. *Biochemical and Biophysical Research Communications* **324**:1024–1033.
Chen M-C, Cheng Y-M, Sung P-J, Kuo C-E, Fang L-S. 2003. Molecular identification of Rab7 (ApRab7) in *Aiptasia pulchella* and its exclusion from phagosomes harboring zooxanthellae. *Biochemical and Biophysical Research Communications* **308**:586–595.
Chen M-C, Hong M-C, Huang Y-S, Liu M-C, Cheng Y-M, Fang L-S. 2005. ApRab11, a cnidarian homologue of the recycling regulatory protein Rab11, is involved in the establishment and maintenance of the *Aiptasia–Symbiodinium* endosymbiosis. *Biochemical and Biophysical Research Communications* **338**:1607–1616.
Childress, J. J, Fisher, R. C. 1992. The biology of hydrothermal vent animals: physiology, biochemistry, and autotrophic symbioses. Boca Raton, FL, ETATS-UNIS: CRC Press.
Company R, Antúnez O, Bebianno MJ, Cajaraville MP, Torreblanca A. 2011. 2-D difference gel electrophoresis approach to assess protein expression profiles in *Bathymodiolus azoricus* from mid-Atlantic ridge hydrothermal vents. *Journal of Proteomics* **74**:2909–2919.
Company R, Torreblanca A, Cajaraville M, Bebianno MJ, Sheehan D. 2012. Comparison of thiol subproteome of the vent mussel *Bathymodiolus azoricus* from different Mid-Atlantic Ridge vent sites. *Science of the Total Environment* **437**:413–421.
Cong Q, Borek D, Otwinowski Z, Grishin Nick V. 2015. Tiger swallowtail genome reveals mechanisms for speciation and caterpillar chemical defense. *Cell Reports* **10**:910–919.
Cordes EE, Carney SL, Hourdez S, Carney RS, Brooks JM, Fisher CR. 2007. Cold seeps of the deep Gulf of Mexico: community structure and biogeographic comparisons to Atlantic equatorial belt seep communities. *Deep Sea Research Part I: Oceanographic Research Papers* **54**:637–53.
Corliss JB, Dymond J, Gordon LI, Edmond JM, von Herzen RP, Ballard RD, Green K, Williams D, Bainbridge A, Crane K, van Andel TH. 1979. Submarine thermal springs on the Galápagos Rift. *Science* **203** (4385):1073–1083.
Daugaard M, Rohde M, Jäättelä M. 2007. The heat shock protein 70 family: Highly homologous proteins with overlapping and distinct functions. *FEBS Letters* **581**:3702–3710.
de Beco S, Gueudry C, Amblard F, Coscoy S. 2009. Endocytosis is required for E-cadherin redistribution at mature adherens junctions. *Proceedings of the National Academy of Sciences of the United States of America* **106**:7010–7015.
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).
Detree C, Chabenat A, Lallier FH, Satoh N, Shoguchi E, Tanguy A, Mary J. 2016.

Multiple I-type lysozymes in the hydrothermal vent mussel *Bathymodiolus azoricus* and their role in symbiotic plasticity. *PLoS ONE* **11**:e0148988.Distel DL, Baco AR, Chuang E, Morrill W, Cavanaugh C, Smith CR. 2000. Marine ecology: Do mussels take wooden steps to deep-sea vents? *Nature* **403**:725–726.

dos Reis M, Yang Z. 2011. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Molecular Biology and Evolution* **28**:2161–2172.

Dubilier N, Bergin C, Lott C. 2008. Symbiotic diversity in marine animals: the art of harnessing chemosynthesis. *Nature Reviews Microbiology* **6**:725–740.

Dunn SR, Weis VM. 2009. Apoptosis as a post-phagocytic winnowing mechanism in a coral–dinoflagellate mutualism. *Environmental Microbiology* **11**:268–276.

Dziarski R, Gupta D. 2006. The peptidoglycan recognition proteins (PGRPs). *Genome Biology* **7**:232.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**:1792–1797.

Endo K, Noguchi Y, Ueshima R, Jacobs HT. 2005. Novel repetitive structures, deviant protein-encoding sequences and unidentified ORFs in the mitochondrial genome of the brachiopod *Lingula anatina*. *Journal of Molecular Evolution*, **61**:36–53.

Faure B, Schaeffer SW, Fisher CR. 2015. Species distribution and population connectivity of deep-sea mussels at hydrocarbon seeps in the Gulf of Mexico. *PLoS ONE* **10**:e0118460.

Feng D, Cheng M, Kiel S, Qiu JW, Yang Q, Zhou H, Peng Y, Chen DF. 2015. Using *Bathymodiolus* tissue stable carbon, nitrogen and sulfur isotopes to infer biogeochemical process at a cold seep in the South China Sea. *Deep-Sea Research Part I: Oceanographic Research Papers* **104**:52–59.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M. 2014. Pfam: the protein families database. *Nucleic Acids Research* **42**:D222–D230.

Gaill F, Halpern S, Quintana C, Desbruyeres D. 1984. Intracellular presence of arsenic and zinc associated with sulfur in a hydrothermal vents polychaetous (*Alvinella-pompejana*). *Comptes Rendus De L Academie Des Sciences Serie Iii-Sciences De La Vie-Life Sciences* **298**:331.

Garcia-Fernàndez J. 2005. Hox, ParaHox, ProtoHox: facts and guesses. *Heredity* **94**:145–152.

Gerdol M, De Moro G, Manfrin C, Milandri A, Riccardi E, Beran A, Venier P, Pallavicini A. 2014. RNA sequencing and *de novo* assembly of the digestive gland transcriptome in *Mytilus galloprovincialis* fed with toxinogenic and non-toxic strains of *Alexandrium minutum*. *BMC Research Notes* **7**:722.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**:644–652.

Grant JR, Stothard P. 2008. The CGView Server: a comparative genomics tool for

circular genomes. *Nucleic Acids Research* **36**:181–184.
Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Research* **36**:154–158.
Guerra-Assunção JA, Enright AJ. 2010. MapMi: automated mapping of microRNA loci. *BMC Bioinformatics* **11**:133.
Guezi H, Boutet I, Tanguy A, Lallier FH. 2013. The potential implication of apoptosis in the control of chemosynthetic symbionts in *Bathymodiolus thermophilus*. *Fish & Shellfish Immunology* **34**:1709.
Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O. 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**:5654–5666.
Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8**:1494–1512.
Hahn C, Bachmann L, Chevreux B. 2013. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads - a baiting and iterative mapping approach. *Nucleic Acids Research* **41**:e129
Hammer Ø, Harper D, Ryan P. 2001. PAST: Paleontological statistics software package for education and data analysis. *Palaeontologia Electronica* **4**:9.
Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular Biology and Evolution* **30**:1987–1997.
Hoffmann JA, Kafatos FC, Janeway CA, Ezekowitz RAB. 1999. Phylogenetic perspectives in innate immunity. *Science* **284**:1313–1318.
Hongo JA, de Castro GM, Cintra LC, Zerlotini A, Lobo FP. 2015. POTION: an end-to-end pipeline for positive Darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes. *BMC Genomics* **16**:567.
Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, Mann M. 2005. Exponentially Modified Protein Abundance Index (emPAI) for Estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Molecular & Cellular Proteomics* **4**:1265–1272.
Janssen R, Le Gouar M, Pechmann M, Poulin F, Bolognesi R, Schwager EE, Hopfen C, Colbourne JK, Budd GE, Brown SJ, Prpic NM, Kosiol C, Vervoort M, Damen WG, Balavoine G, McGregor AP.2010. Conservation, loss, and redeployment of Wnt ligands in protostomes: implications for understanding the evolution of segment formation. *BMC Evolutionary Biology* **10**:1–21.
Jiao W. Fu X, Dou J, Li H, Su H, Mao J, Yu Q, Zhang L, Hu X, Huang X, Wang Y, Wang S, Bao Z. 2013. High-resolution linkage and quantitative trait locus mapping aided by genome survey sequencing: building up an integrative genomic

framework for a bivalve mollusc. *DNA Research* 1–17.
Jones WJ, Won Y-J, Maas PAY, Smith PJ, Lutz RA, Vrijenhoek RC. 2006. Evolution of habitat use by deep-sea mussels. *Marine Biology* **148**:841–851.
Johnson S, Won Y-J, Harvey J, Vrijenhoek R. 2013. A hybrid zone between *Bathymodiolus* mussel lineages from Eastern Pacific hydrothermal vents. *BMC Evolutionary Biology* **13**:21.
Kádár E, Bettencourt R, Costa V, Santos RS, Lobo-da-Cunha A, Dando P. 2005. Experimentally induced endosymbiont loss and re-acquirement in the hydrothermal vent bivalve *Bathymodiolus azoricus*. *Journal of Experimental Marine Biology and Ecology* **318**:99–110.
Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Research* **20**:1313–1326.
Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, Kohara Y, Fujiyama A, Hayashi T, Itoh T. 2014. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research* **24**:1384–1395.
Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**:457–462.
Kapitonov VV, Jurka J. 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Review Genetics* **9**:411–412.
Kiel S, Amano K. 2013. The earliest bathymodiolin mussels: an evaluation of Eocene and Oligocene taxa from deep-sea methane seep deposits in western Washington State, USA. *Journal of Paleontology* **87**:589–602.
Kim D-Y, Bovet L, Maeshima M, Martinoia E, Lee Y. 2007. The ABC transporter AtPDR8 is a cadmium extrusion pump conferring heavy metal resistance. *The Plant Journal* **50**:207–218.
Kumar S, Nei M, Dudley J, Tamura K. 2008. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in Bioinformatics* **9**:299–306.
Kumar S, Stecher G, Peterson D, Tamura K. 2012. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics* **28**:2685–2686.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**:357–359.
Lee RW, Thuesen EV, Childress JJ, Fisher CR. 1992. Ammonium and free amino acid uptake by a deep-sea mussel (*Bathymodiolus* sp., undescribed) containing methanotrophic bacterial symbionts. *Marine Biology* **113**:99–106.
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 University Press, 1985).
Leung P, Ip J, Mak S, Qiu JW, Lam P, Wong C, Chan L, Leung K. 2014. *De novo* transcriptome analysis of *Perna viridis* highlights tissue-specific patterns for environmental studies. *BMC Genomics* **15**:804.

1771 Levin LA. 2005. Ecology of cold seep sediments: Interactions of fauna with flow,
1772 chemistry and microbes. *Oceanography and marine biology* **43**:1–46.
1773 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
1774 Durbin R, Subgroup GPDP. 2009. The sequence alignment/map format and
1775 SAMtools. *Bioinformatics* **25**:2078–2079.
1776 Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of ortholog groups for
1777 eukaryotic genomes. *Genome Research* **13**:2178–2189.
1778 Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets
1779 of protein or nucleotide sequences. *Bioinformatics* **22**:1658–1659.
1780 Little CTS, Vrijenhoek RC. 2003. Are hydrothermal vent animals living fossils?
1781 *Trends in Ecology & Evolution* **18**:582–588.
1782 Liu Y, Schröder J, Schmidt B. 2013. Musket: a multistage *k*-mer spectrum-based error
1783 corrector for Illumina sequence data. *Bioinformatics* **29**:308–315.
1784 Lorion J, Kiel S, Faure B, Kawato M, Ho SYW, Marshall B, Tsuchida S, Miyazaki J-I,
1785 Fujiwara Y. 2013. Adaptive radiation of chemosymbiotic deep-sea mussels.
1786 *Proceedings of the Royal Society B: Biological Sciences* **281**:20132972.
1787 Luckenbach T, Epel D. 2008. ABCB- and ABCC-type transporters confer
1788 multixenobiotic resistance and form an environment-tissue barrier in bivalve gills.
1789 *American Journal of Physiology-Regulatory, Integrative and Comparative*
1790 *Physiology* **294**:1919–1929.
1791 Luo Y-J, Takeuchi T, Koyanagi R, Yamada L, Kanda M, Khalturina M, Fujie M,
1792 Yamasaki S, Endo K, Satoh N. 2015. The *Lingula* genome provides insights into
1793 brachiopod evolution and the origin of phosphate biomineralization. *Nature*
1794 *Communication* **6**:8301.
1795 Minic Z. 2009. Organisms of deep sea hydrothermal vents as a source for studying
1796 adaptation and evolution. *Symbiosis* **47**:121–132.
1797 Miyazaki J-I, Beppu S, Kajio S, Dobashi A, Kawato M, Fujiwara Y, Hirayama H.
1798 2013. Dispersal ability and environmental adaptability of deep-sea mussels
1799 *Bathymodiolus* (Mytilidae: Bathymodiolinae). *Open Journal of Marine Science*
1800 **3**:31–39.
1801 Miyazaki J-I, Martins LdO, Fujita Y, Matsumoto H, Fujiwara Y. 2010. Evolutionary
1802 process of deep-sea *Bathymodiolus* mussels. *PLoS ONE* **5**:e10363.
1803 Mizi A, Zouros E, Moschonas N, Rodakis GC. 2005. The complete maternal and
1804 paternal mitochondrial genomes of the Mediterranean mussel *Mytilus*
1805 *galloprovincialis*: implications for the doubly uniparental inheritance mode of
1806 mtDNA. *Molecular Biology Evolution*, **22**:952–967.
1807 Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic
1808 genome annotation and pathway reconstruction server. *Nucleic Acids Research*
1809 **35**:W182–W185.
1810 Mosesson MW. 2005. Fibrinogen and fibrin structure and functions. *Journal of*
1811 *Thrombosis and Haemostasis* **3**:1894–1904.
1812 Mount DW. 2009. Using hidden Markov models to align multiple sequences. *Cold*
1813 *Spring Harbor Protocols* **2009**:pdb.top41.
1814 Moya A, Pereto J, Gil R, Latorre A. 2008. Learning how to live together: genomic

1815 insights into prokaryote-animal symbioses. *Nature Reviews Genetics* **9**:218–229.
1816 Mu H, Sun J, Fang L, Luan T, Williams GA, Cheung SG, Wong CKC, Qiu J-W. 2015.
1817 Genetic basis of differential heat resistance between two species of congeneric
1818 freshwater snails: insights from quantitative proteomics and base substitution rate
1819 analysis. *Journal of Proteome Research* **14**:4296–4308.
1820 Murray MG, Thompson WF. 1980. Rapid isolation of high molecular weight plant
1821 DNA. *Nucleic Acids Research* **8**:4321–4326.
1822 Newton ILG, Woyke T, Auchtung TA, Dilly GF, Dutton RJ, Fisher MC, Fontanez KM,
1823 Lau E, Stewart FJ, Richardson PM, Barry KW, Saunders E, Detter JC, Wu D,
1824 Eisen JA, Cavanaugh CM. 2007. The *Calyptogena magnifica* chemoautotrophic
1825 symbiont genome. *Science* **315**:998–1000.
1826 Niehrs C. 2012. The complex world of WNT receptor signalling. *Nature Reviews*
1827 *Molecular Cell Biology* **13**:767–779.
1828 Nussbaumer AD, Fisher CR, Bright M. 2006. Horizontal endosymbiont transmission
1829 in hydrothermal vent tubeworms. *Nature* **441**:345–348.
1830 Nusse R, Varmus HE. 1992. *Wnt* genes. *Cell* **69**:1073–1087.
1831 Nyholm SV, Graf J. 2012. Knowing your friends: invertebrate innate immunity fosters
1832 beneficial bacterial symbioses. *Nature Reviews Microbiology* **10**:815–827.
1833 Nyholm SV, McFall-Ngai M. 2004. The winnowing: establishing the squid-*vibrio*
1834 symbiosis. *Nature Reviews Microbiology* **2**:632–642.
1835 Oakley CA, Ameismeier MF, Peng L, Weis VM, Grossman AR, Davy SK. 2016.
1836 Symbiosis induces widespread changes in the proteome of the model cnidarian
1837 *Aiptasia*. *Cellular Microbiology* **18**:1009–1023.
1838 Olu-Le Roy K, von Cosel R, Hourdez S, Carney SL, Jollivet D. 2007. Amphi-Atlantic
1839 cold-seep *Bathymodiolus* species complexes across the equatorial belt. *Deep-Sea*
1840 *Research Part I: Oceanographic Research Papers* **54**:1890–1911.
1841 Paull CK, Hecker B, Commeau R, Freeman-Lynde RP, Neumann C, Corso WP,
1842 Golubic S, Hook JE, Sikes E, Curray J. 1984. Biological communities at the
1843 Florida Escarpment resemble hydrothermal vent taxa. *Science* **226**:964–967.
1844 Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately
1845 maximum-likelihood trees for large alignments. *PLoS ONE* **5**:e9490.
1846 Pruski AM, Dixon DR. 2007. Heat shock protein expression pattern (HSP70) in the
1847 hydrothermal vent mussel *Bathymodiolus azoricus*. *Marine Environmental*
1848 *Research* **64**:209–224.
1849 Robidart JC, Bench SR, Feldman RA, Novoradovsky A, Podell SB, Gaasterland T,
1850 Allen EE, Felbeck H. 2008. Metabolic versatility of the *Riftia pachyptila*
1851 endosymbiont revealed through metagenomics. *Environmental Microbiology*
1852 **10**:727–737.
1853 Robicheau BM. 2016. Doubly uniparental inheritance of mitochondrial DNA in the
1854 marine horse mussel (*Modiolus modiolus*): a comparison of female- and
1855 male-associated mitochondrial genomes. *PhD thesis*, Acadia University Wolfville,
1856 Canada.
1857 Rocks O, Peyker A, Bastiaens PIH. 2006. Spatio-temporal segregation of Ras signals:
1858 one ship, three anchors, many harbors. *Current Opinion in Cell Biology* **18**:351–

357.
Roesijadi G, Young J, Crecelius E, Thomas L. 1985. Distribution of trace metals in the hydrothermal vent clam, *Calyptogena magnifica*. *Bulletin of the Biological Society of Washington* **6**:311–324.
Romero A, Novoa B, Figueras A. 2015. The complexity of apoptotic cell death in mollusks: An update. *Fish & Shellfish Immunology* **46**:79–87.
Sahlin K, Vezzi F, Nystedt B, Lundeberg J, Arvestad L. 2014. BESST-efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics* **15**:281.
Sayavedra L, Kleiner M, Ponnudurai R, Wetzel S, Pelletier E, Barbe V, Satoh N, Shoguchi E, Fink D, Breusing C, Reusch TB, Rosenstiel P, Schilhabel MB, Becher D, Schweder T, Markert S, Dubilier N, Petersen JM. 2015. Abundant toxin-related genes in the genomes of beneficial symbionts from deep-sea hydrothermal vent mussels. *eLife* **4**:e07966.
Shinzato C, Shoguchi E, Kawashima T, Hamada M, Hisata K, Tanaka M, Fujie M, Fujiwara M, Koyanagi R, Ikuta T, Fujiyama A, Miller DJ, Satoh N. 2011. Using the *Acropora digitifera* genome to understand coral responses to environmental change. *Nature* **476**:320–323.
Simakov O, Marletaz F, Cho SJ, Edsinger-Gonzales E, Havlak P, Hellsten U, Kuo DH, Larsson T, Lv J, Arendt D, Savage R, Osoegawa K, de Jong P, Grimwood J, Chapman JA, Shapiro H, Aerts A, Otillar RP, Terry AY, Boore JL, Grigoriev IV, Lindberg DR, Seaver EC, Weisblat DA, Putnam NH, Rokhsar DS. 2013. Insights into bilaterian evolution from three spiralian genomes. *Nature* **493**:526–531.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**:3210–3212.
Simpson JT. 2014. Exploring genome characteristics and sequence quality without a reference. *Bioinformatics* **30**:1228–1235.
Smit A, Hubley, R. 2008. RepeatModeler Open-1.0. Available at http://www.repeatmasker.org
Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21**:456–463.
Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* **24**:637–644.
Sturm A, Cunningham P, Dean M. 2009. The ABC transporter gene family of *Daphnia pulex*. *BMC Genomics* **10**:170.
Sun J, Zhang H, Wang H, Heras H, Dreon MS, Ituarte S, Ravasi T, Qian P-Y, Qiu J-W. 2012. First proteome of the egg perivitelline fluid of a freshwater gastropod with aerial oviposition. *Journal of Proteome Research* **11**:4240–4248.
Takeuchi T, Kawashima T, Koyanagi R, Gyoja F, Tanaka M, Ikuta T, Shoguchi E, Fujiwara M, Shinzato C, Hisata K, Fujie M, Usami T, Nagai K, Maeyama K, Okamoto K, Aoki H, Ishikawa T, Masaoka T, Fujiwara A, Endo K, Endo H, Nagasawa H, Kinoshita S, Asakawa S, Watabe S, Satoh N. 2012. Draft genome of

the pearl oyster *Pinctada fucata*: a platform for understanding bivalve biology. *DNA Research* **19**:117–130.
Takeuchi T, Koyanagi R, Gyoja F, Kanda M, Hisata K, Fujie M, Goto H, Yamasaki S, Nagai K, Morino Y, Miyamoto H, Endo K, Endo H, Nagasawa H, Kinoshita S, Asakawa S, Watabe S, Satoh N, Kawashima T. 2016. Bivalve-specific gene expansion in the pearl oyster genome: implications of adaptation to a sessile lifestyle. *Zoological Letters* **2**:3.
Talavera G, Castresana J. 2007. Improvement of Phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* **56**:564–577.
Tanguy A, Bierne N, Saavedra C, Pina B, Bachère E, Kube M, Bazin E, Bonhomme F, Boudry P, Boulo V, Boutet I, Cancela L, Dossat C, Favrel P, Huvet A, Jarque S, Jollivet D, Klages S, Lapègue S, Leite R, Moal J, Moraga D, Reinhardt R, Samain JF, Zouros E, Canario A. 2008. Increasing genomic information in bivalves through new EST collections in four species: development of new genetic markers for environmental studies and genome evolution. *Gene* **408**:27–36.
Thornburg CC, Zabriskie TM, McPhail KL. 2010. Deep-sea hydrothermal vents: potential hot spots for natural products discovery? *Journal of Natural Products* **73**:489–499.
Thurber AR, Jones WJ, Schnabel K. 2011. Dancing for food in the deep sea: bacterial farming by a new species of Yeti crab. *PLoS ONE* **6**:e26243.
Tunnicliffe V. 1991. The biology of hydrothermal vents: ecology and evolution. *Oceanography and Marine Biology: An Annual Review* 29: 319–408.
Uliano-Silva M, Americo JA, Brindeiro R, Dondero F, Prosdocimi F, de Freitas Rebelo M. 2014. Gene discovery through transcriptome sequencing for the invasive mussel *Limnoperna fortunei*. *PLoS ONE* **9**:e102973.
Uliano-Silva M, Americo JA, Costa I, Schomaker-Bastos A, de Freitas Rebelo M, Prosdocimi F. 2016. The complete mitochondrial genome of the golden mussel *Limnoperna fortunei* and comparative mitogenomics of Mytilidae. *Gene* **577**:202–208.
van der Heijden K, Petersen JM, Dubilier N, Borowski C. 2012. Genetic connectivity between North and South Mid-Atlantic ridge chemosynthetic bivalves and their symbionts. *PLoS ONE* **7**:e39994.
Van Dover CL. 2000. The ecology of deep-sea hydrothermal vents. Princeton University Press, New Jersey.
Van Dover CL, Humphris SE, Fornari D, Cavanaugh CM, Collier R, Goffredi SK, Hashimoto J, Lilley MD, Reysenbach AL, Shank TM, Von Damm KL, Banta A, Gallant RM, Gotz D, Green D, Hall J, Harmer TL, Hurtado LA, Johnson P, McKiness ZP, Meredith C, Olson E, Pan IL, Turnipseed M, Won Y, Young CR, Vrijenhoek RC. 2001. Biogeography and ecological setting of Indian Ocean hydrothermal vents. *Science* **294**:818–823.
Verghese J, Abrams J, Wang Y, Morano KA. 2012. Biology of the heat shock response and protein chaperones: budding yeast (*Saccharomyces cerevisiae*) as a model system. *Microbiology and Molecular Biology Reviews* **76**:115–158.

Vij S, Kuhl H, Kuznetsova IS, Komissarov A, Yurchenko AA, Van Heusden P, Singh S, Thevasagayam NM, Prakki SRS, Purushothaman K, Saju JM, Jiang J, Mbandi SK, Jonas M, Hin Yan Tong A, Mwangi S, Lau D, Ngoh SY, Liew WC, Shen X, Hon LS, Drake JP, Boitano M, Hall R, Chin CS, Lachumanan R, Korlach J, Trifonov V, Kabilov M, Tupikin A, Green D, Moxon S, Garvin T, Sedlazeck FJ, Vurture GW, Gopalapillai G, Kumar Katneni V, Noble TH, Scaria V, Sivasubbu S, Jerry DR, O'Brien SJ, Schatz MC, Dalmay T, Turner SW, Lok S, Christoffels A, Orbán L. 2016. Chromosomal-level assembly of the Asian seabass genome using long sequence reads and multi-layered scaffolding. *PLoS Genetics* **12**:e1005954.

Vrijenhoek RC. 2010. Genetic diversity and connectivity of deep-sea hydrothermal vent metapopulations. *Molecular Ecology* **19**:4391–4411.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**:e112963.

Wences A, Schatz M. 2015. Metassembler: merging and optimizing *de novo* genome assemblies. *Genome Biology* **16**:207.

Wennerberg K, Rossman KL, Der CJ. 2005. The Ras superfamily at a glance. *Journal of Cell Science* **118**:843–846.

Wessel D, Flügge UI. 1984. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Analytical Biochemistry* **138**:141–143.

Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, Mathieson T, Lemeer S, Schnatbaum K, Reimer U, Wenschuh H, Mollenhauer M, Slotta-Huspenina J, Boese JH, Bantscheff M, Gerstmair A, Faerber F, Kuster B. 2014. Mass-spectrometry-based draft of the human proteome. *Nature* **509**:582–587.

Williams AF, Barclay AN. 1988. The immunoglobulin superfamily—domains for cell surface recognition. *Annual Review of Immunology* **6**:381–405.

Won Y-J, Hallam SJ, O'Mullan GD, Pan IL, Buck KR, Vrijenhoek RC. 2003a. Environmental acquisition of thiotrophic endosymbionts by deep-sea mussels of the genus *Bathymodiolus*. *Applied and Environmental Microbiology* **69**:6785–6792.

Won Y, Young CR, Lutz RA, Vrijenhoek RC. 2003b. Dispersal barriers and isolation among deep-sea mussel populations (Mytilidae: *Bathymodiolus*) from Eastern Pacific hydrothermal vents. *Molecular Ecology* **12**:169–184.

Wong YH, Sun J, He LS, Chen LG, Qiu J-W, Qian P-Y. 2015. High-throughput transcriptome sequencing of the cold seep mussel *Bathymodiolus platifrons*. *Scientific Reports* **5**:16597.

Xu B. 2002. Endoglucanase and mannanase from blue mussel, *Mytilus edulis*: purification, characterization, gene and three dimensional structure. *Ph.D. Thesis*, *Uppsala University*, Uppsala, Sweden.

Xu H, Luo X, Qian J, Pang X, Song J, Qian G, Chen J, Chen S. 2012. FastUniq: A fast *de novo* duplicates removal tool for paired short reads. *PLoS ONE* **7**:e52249.

Xu M, Jiang L, Shen K-N, Wu C, He G, Hsiao C-D. 2016. Transcriptome response to copper heavy metal stress in hard-shelled mussel (*Mytilus coruscus*). *Genomics Data* **7**:152–154.

Xu T, Sun J, Chen C, Qian P-Y, Qiu J-W. 2015. The mitochondrial genome of the deep-sea snail *Provanna* sp. (Gastropoda: Provannidae). *Mitochondrial DNA* DOI:10.3109/19401736.2014.1003827.

Xu T, Sun J, Lv J, Watanabe H, Li T, Zou W, Rouse GW, Wang S, Qian P-Y, Bao Z, Qiu JW. 2016. Genome-wide discovery of single nucleotide polymorphisms (SNPs) and single nucleotide variants (SNVs) in deep-sea mussels: Potential use in population genomics and cross-species application. *Deep Sea Research Part II: Topical Studies in Oceanography* doi:10.1016/j.dsr2.2016.03.011.

Xue Q, Hellberg ME, Schey KL, Itoh N, Eytan RI, Cooper RK, La Peyre JF. 2010. A new lysozyme from the eastern oyster, *Crassostrea virginica*, and a possible evolutionary pathway for i-type lysozymes in bivalves from host defense to digestion. *BMC Evolutionary Biology* **10**:213.

Xue W, Li J-T, Zhu Y-P, Hou G-Y, Kong X-F, Kuang Y-Y, Sun X-W. 2013. L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC Genomics* **14**:604.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**:1586–1591.

Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, Wang J, Li S, Li R, Bolund L, Wang J. 2006. WEGO: a web tool for plotting GO annotations. *Nucleic Acids Research* **34**:293–297.

Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, Yang P, Zhang L, Wang X, Qi H, Xiong Z, Que H, Xie Y, Holland PW, Paps J, Zhu Y, Wu F, Chen Y, Wang J, Peng C, Meng J, Yang L, Liu J, Wen B, Zhang N, Huang Z, Zhu Q, Feng Y, Mount A, Hedgecock D, Xu Z, Liu Y, Domazet-Lošo T, Du Y, Sun X, Zhang S, Liu B, Cheng P, Jiang X, Li J, Fan D, Wang W, Fu W, Wang T, Wang B, Zhang J, Peng Z, Li Y, Li N, Wang J, Chen M, He Y, Tan F, Song X, Zheng Q, Huang R, Yang H, Du X, Chen L, Yang M, Gaffney PM, Wang S, Luo L, She Z, Ming Y, Huang W, Zhang S, Huang B, Zhang Y, Qu T, Ni P, Miao G, Wang J, Wang Q, Steinberg CE, Wang H, Li N, Qian L, Zhang G, Li Y, Yang H, Liu X, Wang J, Yin Y, Wang J. 2012a. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **490**:49–54.

Zhang S-M, Adema CM, Kepler TB, Loker ES. 2004. Diversification of Ig superfamily genes in an Invertebrate. *Science* **305**:251–254.

Zhang Z, Xiao J, Wu J, Zhang H, Liu G, Wang X, Dai L. 2012b. ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments. *Biochemical and Biophysical Research Communications* **419**:779–781.

Zheng Q, Wang X-J. 2008. GOEAST: a web-based software toolkit for gene ontology enrichment analysis. *Nucleic Acids Research* **36**:358–363.

Zhong Y-F, Holland PWH. 2011. HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evolution & Development* **13**:567–568.

Zhu F, Zhang X. 2013. The Wnt signaling pathway is involved in the regulation of

2035 phagocytosis of virus in *Drosophila*. *Scientific Reports* **3**:2069.
2036