Virtually all animals have beneficial symbioses with bacteria. The bacterial symbionts can have a major impact on their hosts, by influencing their development, defending them against potential natural enemies, or providing them with nutrition. Several symbioses are known to be highly specific, even in cases in which the symbionts are not transmitted directly to the offspring but are acquired from the environment. However, the molecular mechanisms for host-symbiont recognition are largely unknown.

One striking example of a highly specific nutritional symbiosis is found in *Bathymodiolus* mussels. These mussels are among the most successful fauna at cold seeps and hydrothermal vents of the deep sea, thanks to their association with intracellular sulfur-oxidizing bacteria (SOX), methane-oxidizing bacteria, or both. The SOX symbionts are acquired from the environment with every new generation of mussels, and also throughout the lifespan of the mussel, yet the symbiont species is always specific to the respective host species. This transmission mode gives the mussel the potential to acquire symbiont strains adapted to the local environment, with the associated risks of being infected by opportunistic strains that provide less nutrition, or of not encountering the symbiont at all. Three key questions arise from this mode of transmission: 1) How do symbionts and host recognize each other to maintain a long-term association? 2) Does the metabolism of newly acquired symbionts differ from the established population? 3) Is there strain variation in the symbiont population?

Host-Symbiont Interactions and Metabolism of Chemosynthetic Symbiosis

Lizbeth Sayavedra

Host-Symbiont Interactions and Metabolism of Chemosynthetic Symbiosis in Deep-Sea *Bathymodiolus* Mussels

Lizbeth Sayavedra

Host-symbiont interactions and metabolism of chemosynthetic symbiosis in deep-sea *Bathymodiolus* mussels

Dissertation
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
- Dr. rer. nat. -

dem Fachbereich Biologie/Chemie der
Universität Bremen
vorgelegt von

Lizbeth Sayavedra

Bremen

Mai 2016

Die vorliegende Arbeit wurde von Mai 2012 bis Mai 2016 in der Abteilung Symbiose am Max Planck Institut für marine Mikrobiologie in Bremen angefertigt.

1. Gutachterin: Dr. Jillian M. Petersen
2. Gutachter: Prof. Dr. Matthias Ullrich

Tag des Promotionskolloquiums: 5. Jul. 2016

Umschlag:

Bathymodiolus azoricus Muscheln an der hydrothermalen Quelle "Lucky Strike". BioBaz-Forschungsfahrt, 2013. Bild © Marum

"Pathogenicity […] usually results from inconclusive negotiations for symbiosis, an overstepping of the line by one side or the other, a biological misinterpretation of borders."

-Lewis Thomas-

"Life is a DNA software system."

-J Craig Venter-

Table of contents

Summary

Virtually all animals have beneficial symbioses with bacteria. The bacterial symbionts can have a dramatic impact on their hosts, by influencing their development, defending them against potential natural enemies, or providing them with nutrition. Several symbioses are known to be highly specific, even in cases where the symbionts are not transmitted directly to the offspring but are acquired from the environment. However, the molecular mechanisms for host-symbiont recognition are largely unknown.

One striking example of a highly specific nutritional symbiosis is found in mussels of the genus *Bathymodiolus*. These mussels are among the most successful fauna in cold seeps and hydrothermal vents of the deep sea, thanks to their association with intracellular sulfur-oxidizing bacteria (SOX), methane-oxidizing bacteria, or both. The SOX symbionts are acquired from the environment with every new generation of mussels, and also throughout the lifespan of the mussel, yet the symbiont species is always specific to the respective host species. This transmission mode gives the mussel the potential to acquire symbiont strains adapted to the local environment, with the associated risks of being infected by opportunistic strains that provide less nutrition, or of not encountering the symbiont at all. Three key questions arise from this mode of transmission: 1) How do symbionts and host recognize each other to maintain a long-term association? 2) Does the metabolism of newly acquired symbionts differ from the established population?, and 3) Is there strain variation in the symbiont population?

In the first part of this thesis, I compared the genomes of SOX symbionts from two *Bathymodiolus* species to their closest free-living and symbiotic relatives to find potential mechanisms for host-symbiont interaction. This study revealed that the SOX symbiont genomes encode an ‘arsenal’ of toxin-related genes (TRGs) that are unique to the *Bathymodiolus* symbiosis. These TRGs belong to three main classes: the YD-repeat, RTX, and MARTX protein families. The TRGs were expressed in both the transcriptome and proteome of the SOX symbionts, showing that they play an active role in the symbiosis. We hypothesized that these genes are being used for host-symbiont interactions and for defense against pathogens or parasites.

To understand the evolutionary context of these toxin-related genes, we sequenced the genomes of six additional SOX symbiont species to build a phylogenomic tree and mapped the occurrence of different TRG classes onto that tree. I found that the *Bathymodiolus*-SOX symbiosis had two independent lineages. Members of the MARTX family of TRGs were present in both symbiont lineages, suggesting that they are essential for host interaction. In contrast, YD repeats and RTX were unique to one symbiont lineage and their numbers are increasing, suggesting that they may not be essential for host recognition.

Bathymodiolus gills are continuously growing over the lifetime of the mussel. The budding zone is the site of active growth and is where the new tissue is colonized by bacterial symbionts from the environment. We, therefore, expected symbionts in the budding zone to be expressing a different set of genes compared to the established symbiont populations in older gill filaments, possibly including host-recognition factors. I sequenced and compared the metatranscriptomes of the budding zone vs. older gill filaments from five mussel individuals. In the nascent filaments, the SOX symbiont overexpressed 236 genes, including genes for central carbon metabolism, energy production, and protein repair. This expression pattern is similar to the adaptation phase

of bacterial cultures in laboratory models such as *Salmonella enterica*, which suggests that the SOX symbionts are adapting to a new environment in nascent filaments.

Finally, we showed that the SOX symbionts of mussels from the North Mid-Atlantic Ridge are more metabolically versatile than previously thought. Their genomes show that they have the potential to use methanol as an energy source to fuel chemosynthetic primary production. They may, therefore, be methylotrophs (using methanol) in addition to being chemolithoautotrophs (using reduced sulfur and hydrogen). Within a single mussel, the key gene for methanol oxidation, a methanol dehydrogenase, is present in only part of the SOX symbiont population, even though they all have identical 16S rRNA genes. The proportion of SOX symbionts with the gene is different at each venting site and might be influenced by individual site conditions.

These studies show that several mechanisms are at stake that promotes the highly specific, yet metabolically flexible association of the *Bathymodiolus* symbiosis. The transmission mode of the symbionts is likely playing a major role in the unique genetic traits of *Bathymodiolus* symbionts described in this thesis. The presence of different pools of SOX symbionts in the environment, as well as contrasting conditions between sites, might be driving the acquisition and evolution of these traits to ‘win’ the symbiosis establishment.

Zusammenfassung

Nahezu alle Tiere gehen nützliche Symbiosen mit Bakterien ein. Die bakteriellen Symbionten können einen bedeutenden Einfluss auf ihre Wirte haben, beispielsweise durch Beeinflussung ihrer Entwicklung, Verteidigung gegen potenzielle natürliche Feinde oder der Versorgung mit Nährstoffen. Es ist bekannt, dass einige symbiontische Assoziationen hochspezifisch sind, sogar in Fällen, in denen die Symbionten nicht direkt auf die Nachkommen übertragen, sondern aus der Umwelt aufgenommen werden. Trotzdem sind die molekularen Mechanismen für die gegenseitige Erkennung von Wirt und Symbiont unbekannt.

Ein bemerkenswertes Beispiel einer hochspezifischen nährstoffbasierten Symbiose ist für Muscheln der Gattung *Bathymodiolus* bekannt. Diese Muscheln zählen zu den erfolgreichsten Organismen an kalten Quellen und Hydrothermalquellen der Tiefsee, dank ihrer Assoziation mit intrazellulären schwefeloxidierenden Bakterien (SOX), methanoxidierenden Bakterien, oder beiden. Die SOX-Symbionten werden mit jeder neuen Muschelgeneration, sowie über die gesamte Lebensspanne der Muscheln, aus der Umwelt aufgenommen. Dabei ist die Symbiontenart stets spezifisch für eine Wirtsart. Diese Form der Symbiontenübertragung erlaubt es den Muscheln, die an die örtliche Umgebung angepassten Symbiontenstämme aufzunehmen, jedoch mit dem Risiko, von opportunistischen Stämmen infiziert zu werden, welche weniger Nahrung bereitstellen, oder aber, auf keinen Symbionten zu treffen. Drei Schlüsselfragen ergeben sich aus dieser Übertragungsart von Symbionten: 1) Wie erkennen sich Wirt und Symbiont, um eine langfristige Assoziation aufrechtzuerhalten? 2) Unterscheidet sich der Stoffwechsel der neu erworbenen Symbionten von dem der bereits bestehenden Population? 3) Gibt es unterschiedliche Bakterienstämme in der Symbiontenpopulation?

Im ersten Teil der vorliegenden Dissertation habe ich das Genom der SOX-Symbionten zweier *Bathymodiolus*-Arten mit nah verwandten frei- und in Symbiose lebenden Bakterien verglichen, um potenzielle Mechanismen der Wirts-Symbionten Interaktion zu finden. Diese Studie zeigte, dass in den Genomen der SOX-Symbionten ein ganzes "Arsenal" an toxin-verwandten Genen (engl. "toxin-related genes" = TRGs) vorhanden ist, welches einzigartig für die *Bathymodiolus*-Symbiose ist. Diese TRGs können in drei Hauptklassen eingeteilt werden: YD-repeat, RTX- und MARTX-Proteinfamilien. Die TRGs wurden sowohl im Transkriptom als auch im Proteom der SOX-Symbionten exprimiert, was impliziert, dass diese eine aktive Rolle in der Symbiose spielen. Wir postulierten, dass diese Gene der Interaktion zwischen Wirt und Symbiont sowie dem Schutz vor Krankheitserregern und Parasiten dienen.

Um den evolutionsbedingten Kontext der TRGs zu verstehen, sequenzierten wir das Genom zusätzlicher sechs SOX-Symbiontenarten, berechneten einen phylogenetischen Baum und zeichneten das Vorkommen der jeweiligen TRG-Klassen in diesem Baum ein. Für die SOX-*Bathymodiolus*-Symbiose konnte ich zwei voneinander unabhängige Abstammungslinien ausmachen. In beiden Abstammungslinien waren TRGs der MARTX-Familie zu finden, was darauf hinweist, dass diese von wesentlicher Bedeutung für die Wirtsinteraktion sind. Dagegen waren die YD-repeats und RTX auf eine einzige Symbionten-Abstammungslinie beschränkt, was darauf hindeutet, dass sie möglicherweise nicht entscheidend zur Wirterkennung beitragen.

Bathymodiolus-Kiemen wachsen kontinuierlich während der Lebenszeit der Muscheln. Die Knospenzone ist der Ort, an welchem das aktive Wachstum stattfindet und das neu gebildete Gewebe von bakteriellen Symbionten kolonisiert wird. Daher

erwarteten wir, dass die Symbionten in der Knospenzone einen, im Vergleich zu den Populationen in älterem Gewebe, unterschiedlichen Satz an Genen exprimieren würden, unter welchen sich Faktoren zur Wirtserkennung befinden könnten. Ich sequenzierte und verglich Metatranskriptome der Knospenzone und der älteren Kiemenfilamente fünf einzelner Muscheln. In den jüngeren Filamenten wurden 236 Gene von den SOX-Symbionten überexprimiert, einschließlich der Gene des zentralen Kohlenstoffmetabolismus, der Energieproduktion und Proteinreparatur. Dieses Expressionsmuster ähnelt dem Adaptationsphasenmuster von Bakterienkulturen in Labormodellen, wie z. B. bei *Salmonella enterica*, was darauf hindeutet, dass die SOX-Symbionten sich an eine neue Umgebung in den entstehenden Filamenten anpassen.

Schließlich konnten wir zeigen, dass die SOX-Symbionten von Muscheln des nördlichen Mittelatlantischen Rückens einen vielseitigeren Stoffwechsel haben als bisher angenommen. Ihr Genom zeigte, dass sie das Potenzial besitzen, Methanol als Energiequelle zu nutzen, um die chemosynthetische Primärproduktion anzutreiben. Sie könnten daher als autotrophe Methylotrophe (Nutzung von Methanol) angesehen werden, zusätzlich zu ihrem chemolithoautotrophen Stoffwechsel (Nutzung von reduzierten Schwefelverbindungen und Wasserstoff). Innerhalb einer einzelnen Muschel kommt das Schlüsselgen für die Methanoloxidation, eine Methanol-Dehydrogenase, nur in einem Teil der SOX-Symbionten-Population vor, und das obwohl die 16S rRNA-Gene bei allen Symbionten identisch sind. Das Verhältnis zwischen SOX-Symbionten mit und ohne Schlüsselgen unterscheidet sich an den Tiefseequellen und könnte von Umweltbedingungen an den jeweiligen Quellen beeinflusst werden.

Diese Untersuchungen zeigen, dass es bei der *Bathymodiolus*-Symbiose mehrere Mechanismen gibt, welche die hochspezifische, aber dennoch metabolisch flexible *Bathymodiolus*-Symbiose fördern. Die Symbiontenübertragungsart spielt sicherlich eine bedeutende Rolle bei der Ausprägung der einzigartigen genetischen Merkmale der in dieser Dissertation beschriebenen *Bathymodiolus*-Symbionten. Das Vorkommen verschiedener SOX-Symbiontenpopulationen in der Umwelt sowie die gegensätzlichen Bedingungen, die an verschiedenen Tiefseequellen herrschen, könnten die Aufnahme und Evolution dieser Merkmale antreiben, um das "Rennen" um die Etablierung der Symbiose zu "gewinnen".

Abbreviations

B.	*Bathymodiolus*
BazSym	Sulfur-oxidizing symbiont of *Bathymodiolus azoricus*
BputSym	Sulfur-oxidizing symbiont of *Bathymodiolus puteoserpentis*
BspSym	Sulfur-oxidizing symbiont of *Bathymodiolus* sp. from 9º SMAR
Ca.	*Candidatus*
***Ca.* R. magnifica**	*'Candidatus* Ruthia magnifica'
***Ca.* T. autotrophica**	'*Candidatus* Thioglobus autotrophica EF1'
***Ca.* T. singularis**	*'Candidatus* Thioglobus singularis PS1'
***Ca.* V. okutanii**	*'Candidatus* Vesicomyosocius okutanii'
CBB	Calvin-Benson-Bassham
DNA	Deoxyribonucleic acid
Dsr	Dissimilatory sulfite reduction complex
E. coli	*Escherichia coli*
EPR	East Pacific Rise
ET	Eiffel Tower
FDR	False discovery rate
FISH	Fluorescence *in situ* hybridization
FPKM	Fragments per kilobase mapped
kb	Kilobase pair
Log	Logatchev
LPS	Lipopolysaccharides
LS	Lucky Strike
MALDI-MS	Matrix-assisted laser desorption ionization mass spectrometry
MAR	Mid-Atlantic Ridge
MARTX	Multifunctional Autoprocessing RTX toxins
MDH	Methanol dehydrogenase
MG	Menez Gwen
MOX	Methane-oxidizing symbiont
mRNA	Messenger RNA
Ms	Montsegur

NMAR	Northern Mid-Atlantic Ridge
PCR	Polymerase chain reaction
PGN	Peptidoglycan
Rb	Rainbow
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
rTCA	Reverse tricarboxylic acid cycle
RTX	Repeat in toxins
RuBisCO	Ribulose 1,5-biphosphate carboxylase-oxygenase
RuMP	Ribulose monophosphate pathway
Sem	Semenov
SMAR	Southern Mid-Atlantic Ridge
SNP	Single nucleotide polymorphism
SOX	Sulfur-oxidizing symbiont
TEM	Transmission electron micrograph
TMM	Trimmed mean of M values
TRGs	Toxin-related genes
V. fischeri	*Vibrio fischeri*

Chapter I:

Introduction

1.1. Chemosynthesis

Before the 19th century, the sun was the only known energy source that could fuel life. Sergei Winogradsky then discovered in 1887 that some microbes could use inorganic substances as their sole energy source (Ackert Jr, 2006), a process that is now named chemosynthesis. Chemosynthesis was initially thought to be only a minor contributor to ecological carbon budgets, until the discovery of life at deep-sea hydrothermal vents in 1977 (Corliss *et al*., 1979; Weiss *et al*., 1977). The deep sea had been perceived as a food-deprived environment, where life could not thrive because of the constant cold temperatures, high pressures, and lack of light (Jannasch and Mottl, 1985; Van Dover, 2000). Researchers were therefore astonished to find that hydrothermal vents are colonized by dense communities of invertebrate animals, which seemed to be living of the inorganic chemicals coming from the sea floor (Cavanaugh *et al*., 1981). This discovery led to the start of a new and exciting research field which has revealed that the key to the success of these marine invertebrate communities is their symbiosis with chemosynthetic bacteria (Box 1) (DeChaine *et al*., 2006; Dubilier *et al*., 2008).

1.1.1. Diversity of chemosynthetic environments

Shortly after the discovery of hydrothermal vents, scientists discovered chemosynthetic symbioses in a wide diversity of marine environments. These include vents, cold seeps, mud volcanoes, continental margins, whale and wood falls, mangrove swamps, coral reefs and shallow water sediments (reviewed by Dubilier *et al*., 2008). These chemosynthetic marine environments occur worldwide. Although the deep-sea is harder to reach than shallow water systems, several sites have been studied in regard to their associated fauna (Figure 1-1**)** (German *et al*., 2011). These ecosystems are active for highly variable time frames and are often disturbed by geologic activity (German *et al*., 2011). The fauna colonizing vents and seeps is not very diverse and comprises many endemic specialist species (Van Dover, 2010, 2003). Depending on the length of time that

an ecosystem exists and the types of energy sources present, different fauna and microbial assemblages will succeed in the colonization (Vrijenhoek, 2010).

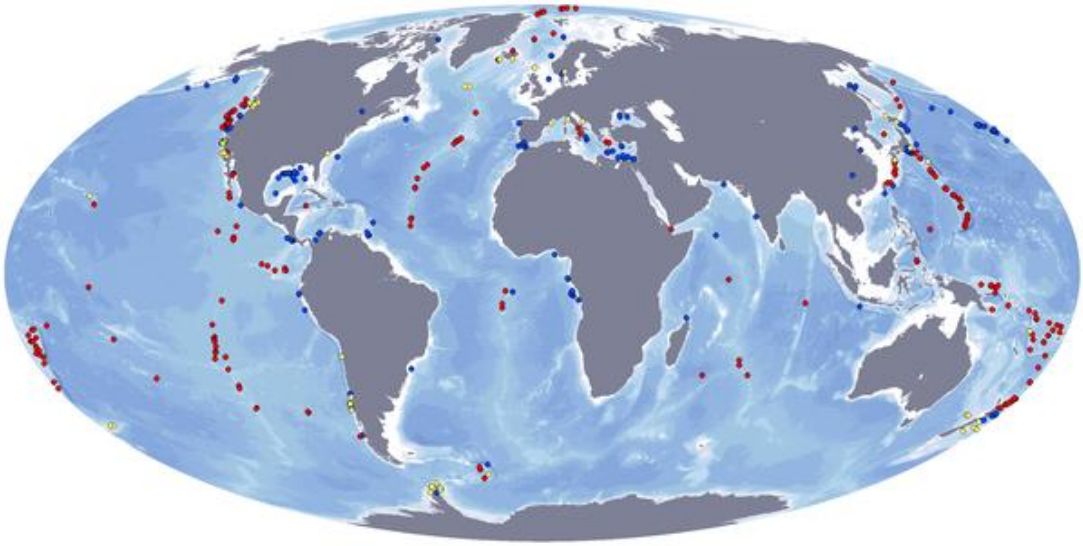


Figure 1-1. Global distribution of hydrothermal vents (red), cold seeps (blue), and organic falls (yellow) that have been studied with respect to their fauna. From German *et al.* (2011).

Hydrothermal vents are fissures on the sea floor from which heated water with reduced chemical substances such as sulfide, iron, manganese, methane, and hydrogen, pours out at sea-floor spreading zones (Van Dover, 2002). This chemically modified water is produced through the circulation of seawater through the ocean crust in three main steps (Humphris and McCollom, 1998):

1. The recharge zone, where seawater penetrates the surface of the permeable volcanic rocks of the ocean crust to an approximate depth of 300 meters. In this step, the oxygen content of the water decreases as the water oxidizes the crust and the reactions occur at a relatively low temperature. In deeper layers, less porous rocks restrict the water flow and the water gets heated as it approaches the heat sources. The water becomes acidic, and sulfate and calcium get removed as minerals, chlorite, and anhydrite precipitate.
2. The reaction zone, where water reacts with rocks at elevated temperatures of up to 400°C. This zone greatly influences the end-member fluids that will eventually come out. Due to the acidic nature of the water at this point, metals such as iron, copper, and sulfur get leached from the rocks.
3. The upflow zone, where the heated water quickly rises to the sea floor. The discharge can occur through the main chimney or through more laborious paths as “diffuse fluids”. The fluids react with the cold oxygenated seawater above the

sea floor. This causes black precipitates formed by metal sulfides creating the so-called "black smokers". If the volcanic pile is highly permeable, the metal sulfides will precipitate beneath the sea floor and will exudate water depleted of metal sulfides creating the "white-smokers" (Figure 1-2) (Humphris and McCollom, 1998; Lascelles, 2007; Tivey, 2007).

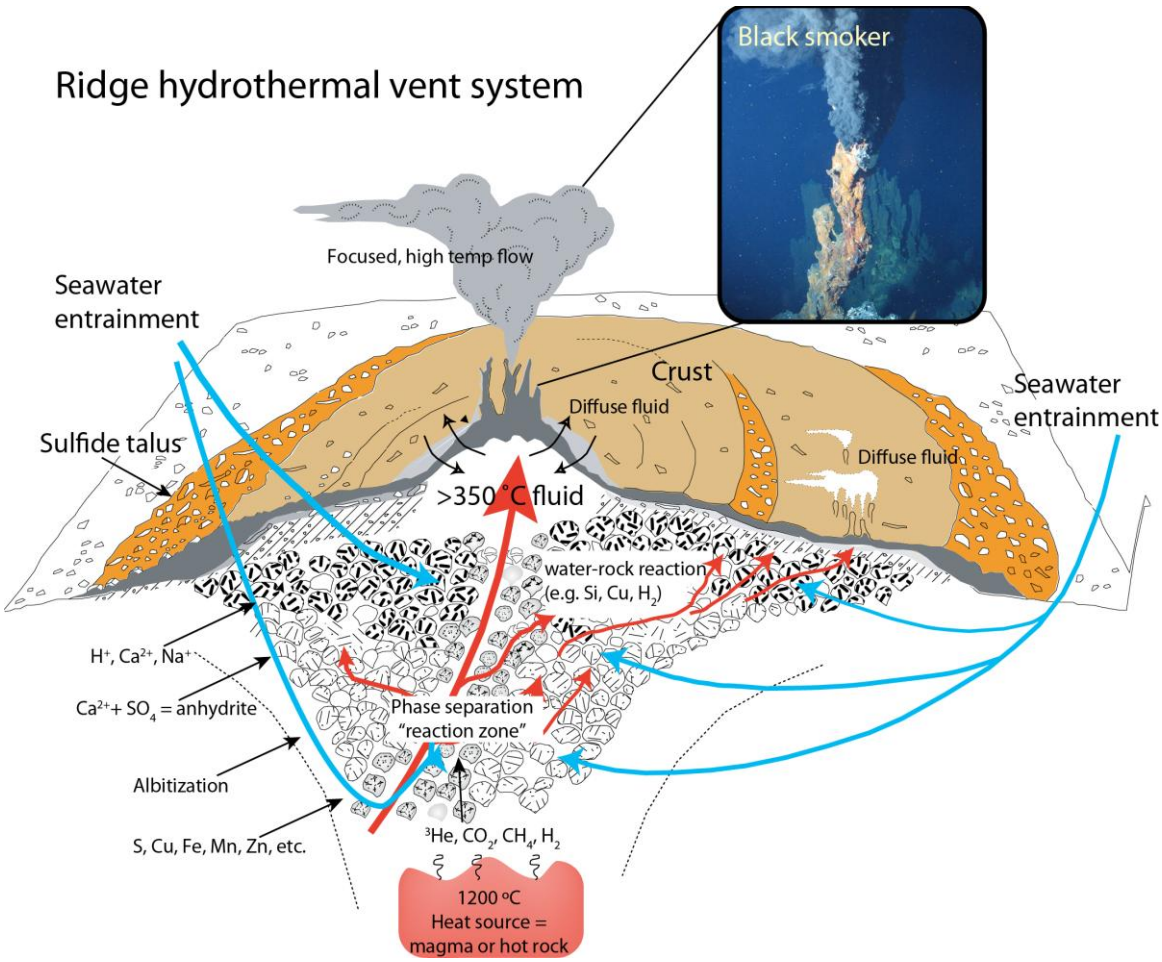


Figure 1-2. Typical hydrothermal vent setting of mid ocean ridges. Cold seawater penetrates into the crust and reacts with basaltic glass and other rocks at temperatures of 40-60°C. Fluids get heated as they penetrate deeper and Mg gets removed in exchange for Ca^{2+}, H^{+}, and Na^{+}. As the fluid moves deeper down it gets slightly acidic and anoxic which causes sulfur and metal to leach from the rocks into the fluid. Volatiles from magma, such as He, CO_2, CH_4, and H_2, are also dissolved into the fluid. Then the buoyant fluid rises rapidly with minor precipitation and dissolution of sulfide. The fluids can come out directly into the ocean or can be further modified if seawater enters other cracks from the vicinity. Adapted from Tivey (2007). Black smoker picture was taken from the Rainbow vent field during the BioBaz cruise (2013, ©Marum).

The fluid composition of different vents is highly heterogeneous and depends on i) the composition of the cold seawater before penetrating the crust, ii) the rock composition that will react with the fluids, and iii) the size, shape and depth of the heat source (Tivey, 2007). Most vent sites with ultrafast to intermediate spreading rates have

basaltic rocks, which are part of the earth's crust (Wetzel and Shock, 2000). These sites have usually high concentrations of sulfide, as high as 20 mM. However, concentrations can even reach 20-110 mM in the course of volcanic events (Lilley *et al*., 2003; Tivey, 2007). In contrast, some vents located at ultraslow and slow spreading ridges have ultramafic rocks such as peridotite and serpentinite, which originate from the earth's mantle (Tivey, 2007; Wetzel and Shock, 2000). Ultramafic vent sites are driven by serpentinization, a reaction that produces methane and hydrogen as byproducts (Blumenberg, 2010; Martin *et al*., 2008). Examples of these sites are Lost City, Rainbow, and Logatchev, located in the northern Mid-Atlantic Ridge (MAR) (Charlou *et al*., 2002; Douville *et al*., 2002; Kelley *et al*., 2005). Vents with ultramafic rocks are therefore supplying different energy sources that will eventually shape the microbial and faunal assemblages. The spreading rate of the tectonic plates at which the vent is located affects the topology and the stability of the vent. Depending on the spreading rate, the time between magma discharges ranges from <10 years to >50,000 years (Martin *et al*., 2008). Vents located along fast-spreading ridge axes are active only for some decades, while vents located along slow-spreading axes can exist for thousands of years (Lalou and Brichet, 1982; Vrijenhoek, 2010). The fauna that colonize these sites must be adapted for dispersal and colonization of the vent sites, which is often hampered by geographical barriers, such as transform faults, that limit larvae dispersal along the axis (Young *et al*., 2008).

Cold seeps exist at passive and active continental margins, where low-temperature fluids rich in hydrocarbons and chemically reduced substances emanate from the sea floor (Sibuet and Roy, 2002). In contrast to hydrothermal vents, most of the production of reduced substances is driven by microorganisms at cold seeps (Jørgensen and Boetius, 2007). Their distribution is patchy, even at short geographical scales (Barry *et al*., 1996). It is not well known how long cold seeps can last, as dating these ecosystems is challenging (MacDonald *et al*., 1990). Several seeps host large communities of fauna that grow slowly and live for a long time, which suggests that these ecosystems are not ephemeral (Bergquist *et al*., 2000). In contrast, **whale and wood falls** are short-term habitats. These organic falls provide an unusually large quantity of organic material, which cause a local depletion of oxygen and sulfide production that supports chemosynthetic communities (Smith and Baco, 2003). Their degradation speed depends on depth, oxygen concentration, and the microbial community associated with them (Smith and Baco, 2003). Wood falls are mainly found at terrestrial margins, while whale

carcasses are found at the whales feeding and breeding regions, and along their migration routes (Braby *et al*., 2007; Vrijenhoek, 2010). For all these chemosynthetic ecosystems, the frequency of the ecosystem extinction, sources of colonizing species, and dispersal rates will determine the number of faunal species present in the population (species richness) (Black *et al*., 1997).

1.1.2. Energy sources for chemosynthetic symbioses

In the ecosystems where there is no sunlight, chemosynthetic microorganisms are the primary producers. Marine invertebrates thrive in these ecosystems thanks to their association with chemoautotrophic and chemoorganotrophic bacteria. Chemoautotrophic bacteria use reduced substances such as sulfide and hydrogen as sources of energy and

Box 1 | Symbiosis. The term symbiosis (from the Greek *sym* 'with' and *bios* 'life') was introduced by Anton de Bary and Albert Bernhard Frank in the 1870's for the living together of two organisms (de Bary, 1879). This broad definition includes beneficial, neutral and detrimental associations. The role of the symbiotic interaction is not always easy to define, since the benefits can change depending on the environmental conditions. With the advent of molecular biology, it has become clear that sometimes the mechanisms used by pathogens for virulence are very similar to the ones necessary for interaction between symbionts (Walker and Crossman, 2007). The symbiosis definition provided by Moya *et al*. (2008) will be used in this thesis, stated as "*a long-term association between two or more species integrated at the behavioral, metabolic and genetic level*". The following terms related to symbiosis will be used:

- **Host:** considered as the bigger partner.
- **Symbiont:** the smaller-sized member of the association that uses the host's resources and does not necessarily provide a benefit in return (Ferrière *et al*., 2007; Starr, 1975).
- **Mutualism/beneficial:** Both partners get a benefit from the association.
- **Commensal**: The symbiont benefits from the association without any detrimental or beneficial effect to the host.
- **Pathogen/parasite**: The symbiont benefits from the host, but the fitness of the host is affected detrimentally.

electrons (reducing equivalents) that are released by abiotic and biotic processes, to fix inorganic carbon (Childress and Fisher, 1992; Van Dover, 2002). Methane is an organic compound, and thus bacteria that use methane as energy and carbon source are considered chemoorganotrophic organisms. These chemosynthetic symbioses occur ubiquitously at oxic-anoxic interfaces (Stewart *et al*., 2005). Although free-living bacteria and archaea are known to use a wide repertoire of reduced energy sources, so far only reduced sulfur compounds, methane, hydrogen and carbon monoxide have been shown to be used by chemosynthetic bacteria in symbiotic relationships with eukaryotic hosts (Dubilier *et al*., 2008; Kleiner *et al*., 2015, 2012b; Martin *et al*., 2008; Petersen *et al*., 2011; Stewart *et al*., 2008). Some of the favorable redox reactions that are used by free-living and symbiotic microorganisms at vents and other chemosynthetic environments are shown in Table 1-1.

Table 1-1. Some anaerobic and aerobic favorable redox reactions used by free-living and symbiotic microorganisms from chemosynthetic environments. Energy available ($\Delta G^{0'}$) is shown in kJ per mole per reaction. Modified from Martin *et al*. (2008).

Metabolism	Reaction	$\Delta G^{0'}$ (kJ per mole)	Examples of chemosynthetic microorganisms
Anaerobic			
Methanogenesis	$4\ H_2 + CO_2 \rightarrow CH_4 + 2\ H_2O$	-131	*Methanococcus* spp. common in magma-hosted vents
	$CH_3CO_2^- + H_2O \rightarrow CH_4 + HCO_3^-$	-36	Methanosarcinales at Lost City
	$4\ HCOO^- + H^+ \rightarrow 3\ HCO_3^- + CH_4$	-106	
S^0 reduction	$S^0 + H_2 \rightarrow H_2S$	-45	Lithotrophic and heterotrophic; hyperthermophilic archaea
Anaerobic CH_4 oxidation	$CH_4 + SO_4^{2-} \rightarrow HS^- + HCO^{3-} + H_2O$	-21	*Methanosarcina* spp. and Epsilonproteobacteria at mud volcanoes and methane seeps
Sulfate reduction	$SO_4^{2-} + H^+ + 4\ H_2 \rightarrow HS^- + 4\ H_20$	-170	Deltaproteobacteria
Fe reduction	$8\ Fe^{3+} + CH_3CO_2^- + 4\ H_2O \rightarrow 2\ HCO_3^- + 8\ Fe^{2+} + 9\ H^+$	Not calculated	Epsilonproteobacteria, thermophilic bacteria and hyperthermophilic Crenarchaeota
Fermentation	$C_6H_{12}O_6 \rightarrow 2\ C_2H_6O + 2\ CO_2$	-300	Many genera of bacteria and archaea
Sulfide oxidation/ nitrate reduction	$5\ HS^- + 8\ NO_3^- + 3\ H^+ \rightarrow 5\ SO_4^{2-} + 4\ N_2 + 4\ H_2O$	-3722	SOX symbionts of *Bathymodiolus* spp.
Aerobic			
Sulfide oxidation	$HS^- + 2\ O_2 \rightarrow SO_4^{2-} + H^+$	-750	Many genera of bacteria; Common vent animal symbionts (e.g. SOX symbionts of *Bathymodiolus* spp.)
CH_4 oxidation	$CH_4 + 2\ O_2 \rightarrow HCO_3^- + H^+ + H_2O$	-750	Common in hydrothermal systems; Vent animal symbionts (e.g. MOX symbiont of *Bathymodiolus spp*.)
Methanol oxidation	$CH_3OH + O_2 \rightarrow CO_2 + H_2O + 2e^-$	-661.6[#]	*Methylophaga sp.* SOX symbiont of *B. azoricus* and *B. puteoserpentis**
Carbon monoxide oxidation	$CO + 0.5\ O_2 \rightarrow CO_2$	-257	Gamma-3 symbiont of *Olavius algarvensis*[**]

Metabolism	Reaction	$\Delta G^{0'}$ (kJ per mole)	Examples of chemosynthetic microorganisms
H_2 oxidation	$H_2 + 0.5\ O_2 \rightarrow H_2O$	-230	Common in hydrothermal systems; Vent animal symbionts (e.g. SOX symbiont of *Bathymodiolus spp.*)
Fe oxidation	$Fe2+ + 0.5\ O_2 + H^+ \rightarrow Fe^{3+} + 0.5\ H_2O$	-65	Common in low-temperature vent fluids; rock-hosted microbial mats
Mn oxidation	$Mn^{2+} + 0.5\ O_2 + H_2O \rightarrow MnO_2 + 2\ H^+$	-50	Common in low-temperature vent fluids; rock-hosted microbial mats; hydrothermal plumes
Respiration	$C_6H_{12}O_6 + 6\ O_2 \rightarrow 6\ CO_2 + 6\ H_2O$	-2,870	Many genera of bacteria

[*] Sayavedra *et al*. – Chapter 5
[**] Kleiner *et al.* (2015)
[#] Pawlowska *et al*. (2014)

1.1.3. Diversity of chemosynthetic symbioses

Free-living chemosynthetic bacteria face the challenge of having to exploit spatially reduced redox gradients. A strategy that some free-living bacteria have used to be able to reach sulfidic and oxic zones is biofilm formation. For example, the sulfur-oxidizing *Beggiatoa* bacteria form filamentous mats that reach both oxygenated water at the sediment surface and reduced substrates in the bottom layers, and can track the chemocline via gliding (Fenchel, 2002). In contrast, chemosynthetic bacteria that live in symbiosis can bridge broader redox zones through the mobility or morphological adaptations of their hosts (Figure 1-3) (Stewart *et al*., 2005). Different behavioral, morphological and physical strategies are used by host animals to reach the oxic-anoxic interfaces in the environment. For example, the shrimp *Rimicaris exoculata* can rapidly move between hot reduced fluids and oxygenated water at vents to provide its epibiotic bacteria located within the gill chamber access to electron donors and acceptors (Jan *et al*., 2014; Petersen *et al*., 2010). The vesicomyid clam *Caplyptogena magnifica* uses its extensible foot to reach sulfidic water within rocks, while filtering oxygenated water with its siphon (Zal *et al*., 2000). These clams use their size to reach longer spatial gradients and can grow up to an impressive size of 30 cm (Cavanaugh *et al*., 2013). The tube worm *Riftia pachyptila* uses a feathery gill plume to absorb sulfide and oxygen from seawater and transport them via the circulatory system along its entire body (Stewart *et al*., 2005). To do so, *R. pachyptila* uses three extracellular haemoglobins, which can bind both oxygen and sulfide at different sites of the respective proteins (Zal *et al*., 1996).

Hosts and symbionts have evolved distinct forms of morphological integration to cooperate with each other. The symbiont can be i) attached to the host in specific regions or covering the entire surface (e.g. the shrimp *Rimicaris exoculata*, and the shallow water stilbonematine nematodes) ii) inside the host, but extracellular (e.g. gutless oligochaetes),

or iii) inside host cells (e.g. the tube worm *Riftia pachyptila*, and bathymodioline mussels) (Dubilier *et al*., 2008).

Associations between chemosynthetic bacteria and eukaryotes have independently evolved multiple times. So far, seven eukaryotic phyla are known to host chemosynthetic symbioses (Ciliophora, Mollusca, Nematoda, Annelida, Platyhelminthes, Arthropoda, and Porifera) (Cavanaugh *et al*., 2013; Dubilier *et al*., 2008). Chemoautotrophy is present in very diverse groups of bacteria including Actinobacteria, Aquificae, Bacili, Chloroflexi, Chlorobi, Spirochaeta, and Proteobacteria (Oren, 2009). It is therefore surprising that chemosynthetic bacteria with a symbiotic lifestyle are to date only known from the Proteobacteria, mostly Gammaproteobacteria, but also Epsilonproteobacteria, Alphaproteobacteria, and possibly Deltaproteobacteria and Zetaproteobacteria (Dubilier *et al*., 2008; Goffredi *et al*., 2007; Gruber-Vodicka *et al*., 2011; Jan *et al*., 2014; Kleiner *et al*., 2012b; Petersen *et al*., 2010; Woyke *et al*., 2006).

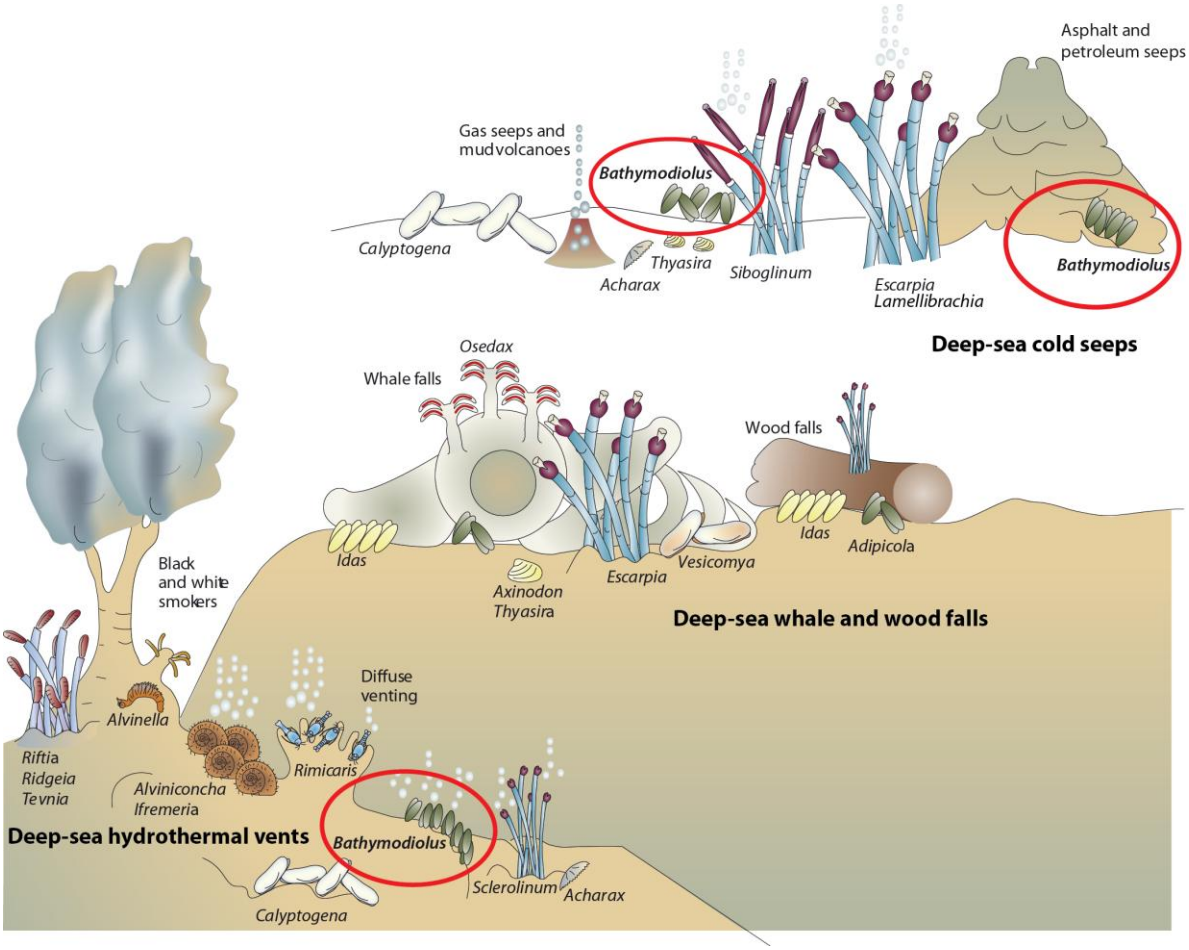


Figure 1-3. Chemosynthetic symbioses from deep-sea ecosystems. The habitats of *Bathymodiolus* mussels, the main chemosynthetic symbiosis studied in this study, are circled in red. Adapted from Dubilier *et al*. (2008).

As exploration of yet undiscovered vents, seeps, and organic falls continues, new symbioses will likely be discovered.

To date, most of the known chemosynthetic symbioses include sulfur-oxidizing (SOX) and methane-oxidizing (MOX) bacteria. Gammaproteobacterial SOX symbionts have been described in six host phyla, in contrast to only three host phyla that associate with MOX bacteria (Cavanaugh *et al*., 2013). Genome comparisons between SOX bacteria that evolved different morphological and behavioral mechanisms to adapt to their hosts could therefore reveal molecular mechanisms that are shared or unique to each symbiosis. In the next section, I will describe the general features of bathymodioline mussels, the host of the chemoautotrophic SOX symbiont that is the main focus of this thesis

1.2. The bathymodioline symbiosis

1.2.1. The host animals

Bathymodioline mussels are among the most successful fauna at hydrothermal vents and cold seeps (Van Dover, 2002; von Cosel, 2002). These mussels are specialists in deep-sea reducing environments, and besides the occasional presence in organic falls, their niche is limited (Duperron, 2010; Pailleret *et al*., 2007). They are a subfamily within the family Mytilidae, one of the six families of bivalves that have been reported to be associated with chemosynthetic bacterial symbionts (Figure 1-4) (Taylor and Glover, 2010). The taxonomy of the genera and species is still under debate, but detailed morphological studies indicate that there are three major groups within the bathymodiolines: the *Bathymodiolus thermophilus* group, the "*B*". *childressi* group, and the *B. aduloides* group (Figure 1-5). Other genera of mussels that are also found in chemosynthetic environments are *Gigantidas*, present at vents around New Zealand, and *Idas*, *Tamu*, and *Adipicola*, which are smaller in size and are present at seeps and organic falls (Figure 1-5) (Duperron, 2010).

Based on fossil records and molecular data, the contemporary seep and vent mussels appeared between the late Mesozoic and the early Cenozoic (Figure 1-5) (Kiel, 2006). Most of these mussels live attached to hard substrates by means of their byssal threads (Duperron, 2010). Thus, in contrast to the tubeworm *Riftia pachyptila* or vesicomyid clams, most bathymodioline mussels do not have access to reduced sediments. The mussels pump the energy-rich water through their gills and thereby supply the substrates to their symbionts via diffusion across the gill tissue (Duperron, 2010). The symbionts are located in the gills, within specialized host cells called bacteriocytes. The gut and feeding groove are functional for filter feeding; yet most of their nutrition is provided by their symbionts (Duperron, 2010; Riou *et al*., 2010).

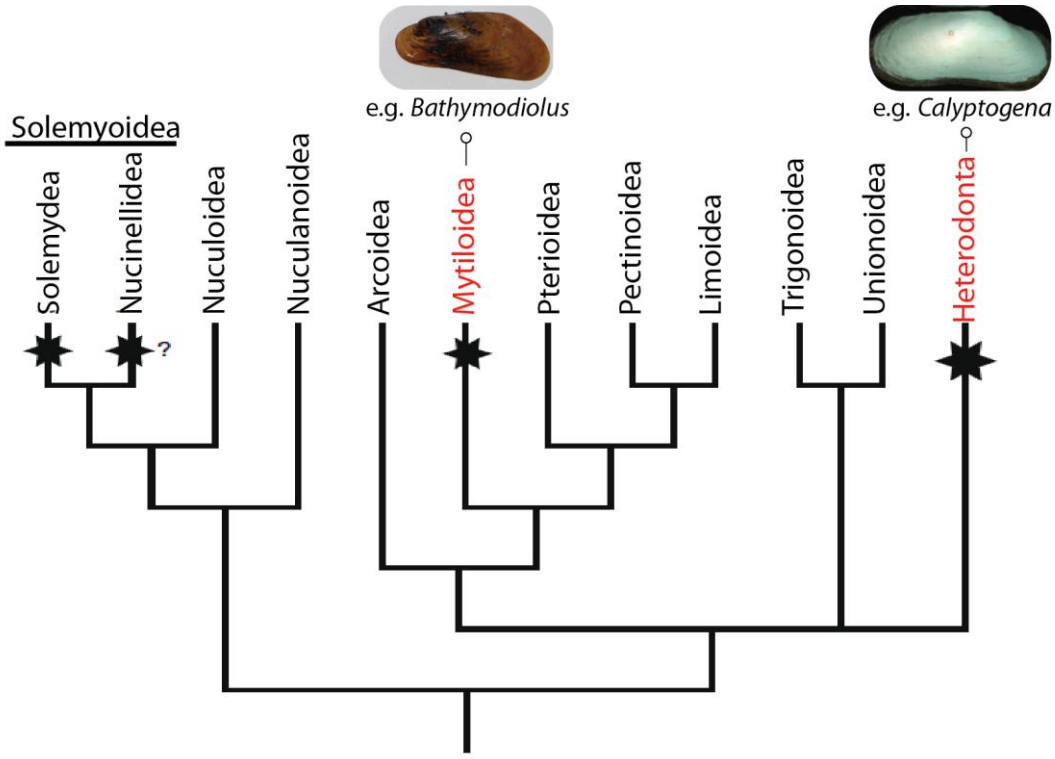


Figure 1-4. Bivalve phylogeny. Families that include symbiotic associations with chemosynthetic symbionts are shown with a star. The subject of this thesis, *Bathymodiolus*, belongs to the superfamily Mytiloidea. The SOX symbionts of *Calyptogena* are the closest relatives to the *Bathymodiolus* SOX symbionts, but this tree shows that their respective host animals are not close relatives. Modified from Taylor *et al*. (2010).

1.2.2. Single and dual symbioses with chemoautotrophic and methanotrophic bacteria

To thrive in the reducing environments, most bathymodioline mussels host a chemoautotrophic symbiont that uses CO_2 as a carbon source and reduced sulfur

compounds and hydrogen as energy source (referred as the SOX symbiont in this thesis) (Figure 1-6) (Dubilier *et al*., 2008; Duperron *et al*., 2006). To fix carbon, the SOX symbiont uses the Calvin-Benson-Bassham (CBB) cycle (Cavanaugh and Robinson, 1996; Fisher *et al*., 1987). Besides sulfide (H_2S), other reduced sulfur compounds such as thiosulfate ($S_2O_3^{2-}$) and sulfite (SO_3^{2-}) are present at vents (Gartman *et al*., 2011). So far, sulfide and thiosulfate have been shown to fuel inorganic carbon fixation by intracellular

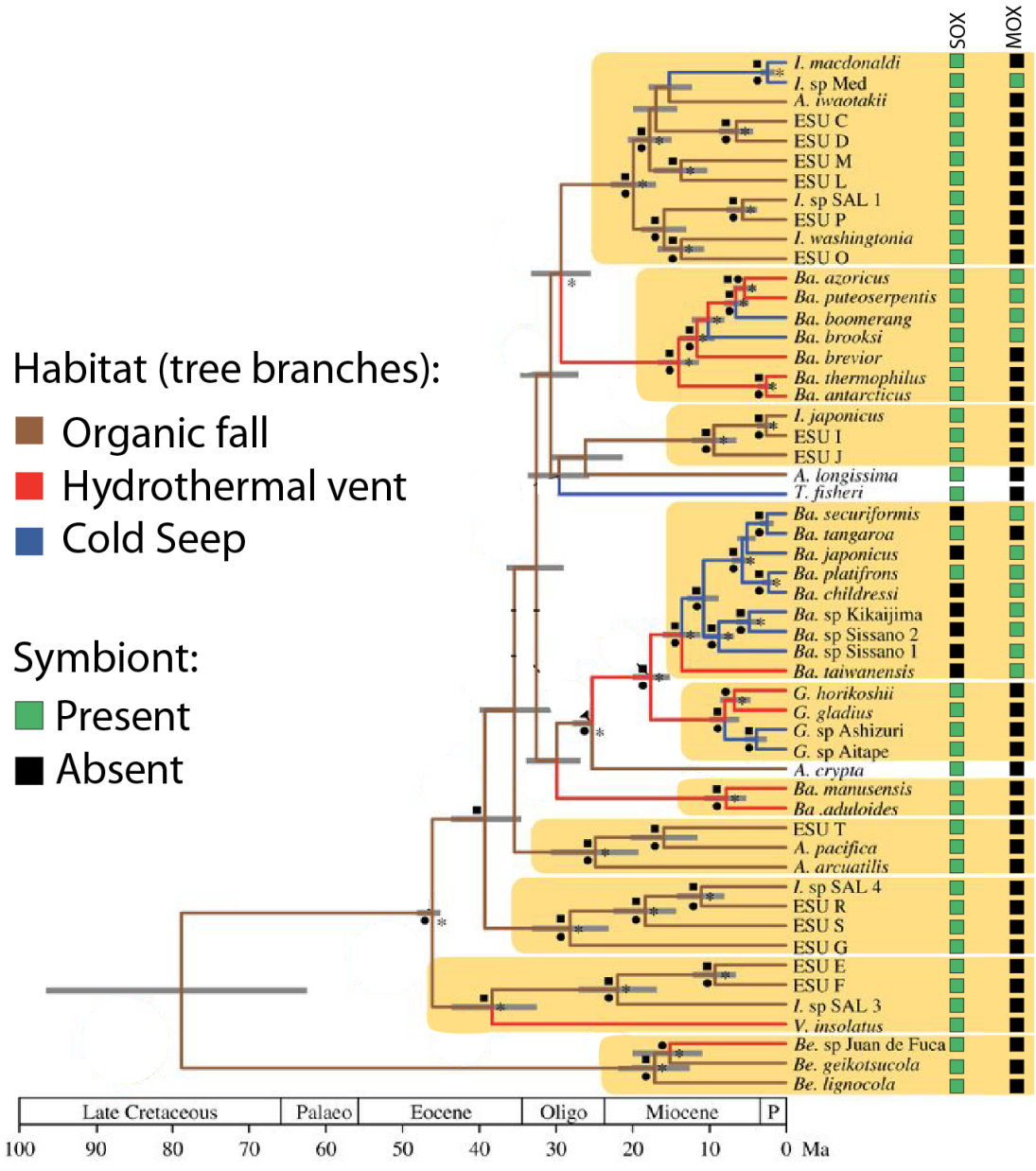


Figure 1-5. Phylogeny of deep-sea symbiotic mytilid mussels based on five genes (mitochondrial and nuclear). Grey bars indicate 95% highest posterior density (HPD) intervals of divergence time estimates. Black squares represent posterior probability >0.99, circles are bootstrap values >75%, asterisks represent bipartitions found with both nuclear and mitochondrial genes. The branches of the tree are color-coded depending on the habitat and it is indicated whether the host species has been reported to have the sulfur-oxidizing (SOX) symbiont and/or the methane-oxidizing (MOX) symbiont. *A.*, *Adipicola*; *Ba.*, *Bathymodiolus*; *Be.*; *Benthomodiolus*; *G.*, *Gigantidas*; *I.*, *Idas*. Modified from Lorion *et al*. (2013).

symbionts of vent bathymodioline mussels and vesicomyid clams (Beinart *et al*., 2015; Nelson *et al*., 1995). The genes required for the oxidation of reduced sulfur compounds are similar to those of free-living sulfur-oxidizing Gammaproteobacteria, even though some individual genes of the independent pathways are not present (Kleiner *et al*., 2012a).

Despite their name, the SOX symbiont may not be restricted to using only reduced sulfur compounds to fuel carbon fixation. Some SOX symbionts, such as those associated to the vent mussels *B. puteoserpentis*, *B. azoricus*, and *B. thermophilus*, have acquired the ability to use hydrogen as an additional energy source (Ikuta *et al*., 2015; Petersen *et al*., 2011). The potential to use hydrogen as an energy source by some SOX symbionts is a trait that has also been found in other symbioses, such as the clam *Solemya velum*, the tube worm *Riftia pachyptila*, and the shrimp *Rimicaris exoculata* (Dmytrenko *et al*., 2014; Petersen *et al*., 2011). This trait is however not present in the closest symbiotic SOX relatives, the vesicomyid clam symbionts (see section "Symbiotic close relatives") (Kuwahara *et al*., 2007; Newton *et al*., 2008).

At least ten species of bathymodioline mussels, such as *B. azoricus*, have acquired a second symbiont that can use methane as a carbon and energy source (referred in this thesis as the MOX symbiont), a rare feature for metazoans (Figure 1-6) (Dubilier *et al*.,

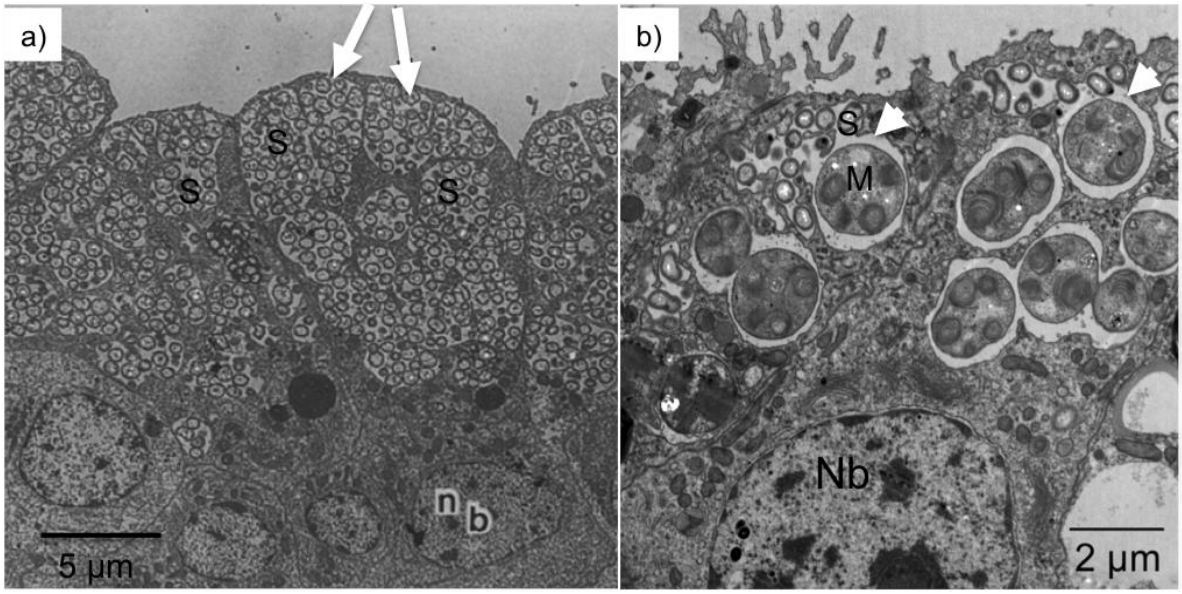


Figure 1-6. TEMs of gill cross-sections of two Bathymodiolus species hosting one or two symbionts, respectively. a) *B. thermophilus* from a vent on the Galapagos Rift has bacteriocytes containing endosymbiotic SOX symbionts (S) enclosed within vacuoles (white arrow). From Fisher *et al*. (1987). b) *B. brooksi* from a seep in the northern Gulf of Mexico host two distinct symbiont morphotypes. The small morphotype corresponds to the SOX symbiont (S) and the large morphotype with stacked membranes corresponds to the MOX symbiont (M). Some bacteriocytes contain both symbonts within a single vacuole (white arrow). Nb, Nucleus of bacteriocyte (courtesy of Niko Leisch).

2008; Duperron *et al*., 2006; Petersen and Dubilier, 2009). For at least three bathymodioline mussels the MOX symbiont appears to have replaced the SOX symbiont at some point of evolution (e.g. *"B." childressi*) (Figure 1-5) (Lorion *et al*., 2013). The activity of the MOX symbiont was confirmed by enzymatic assays for methanol dehydrogenase, which catalyzes the second step in methane oxidation (Fisher *et al*., 1987; Pimenov *et al*., 2002; Robinson *et al*., 1998). Genomic evidence suggests that the MOX symbiont can generate biomass from one-carbon compounds such as methane and methanol via the ribulose monophosphate (RuMP) pathway, which makes this symbiont a type I methanotroph (Antony CP *et al.*, in prep.). Acquiring a symbiont that can exploit an additional energy source has allowed mussels to colonize environments rich in sulfide, hydrogen, or methane (Figure 1-7) (Fisher *et al*., 1987; Petersen *et al*., 2011).

If a new symbiosis is established, or if one of the symbionts acquires the potential to use an additional energy source, a wider range of environmental niches could be

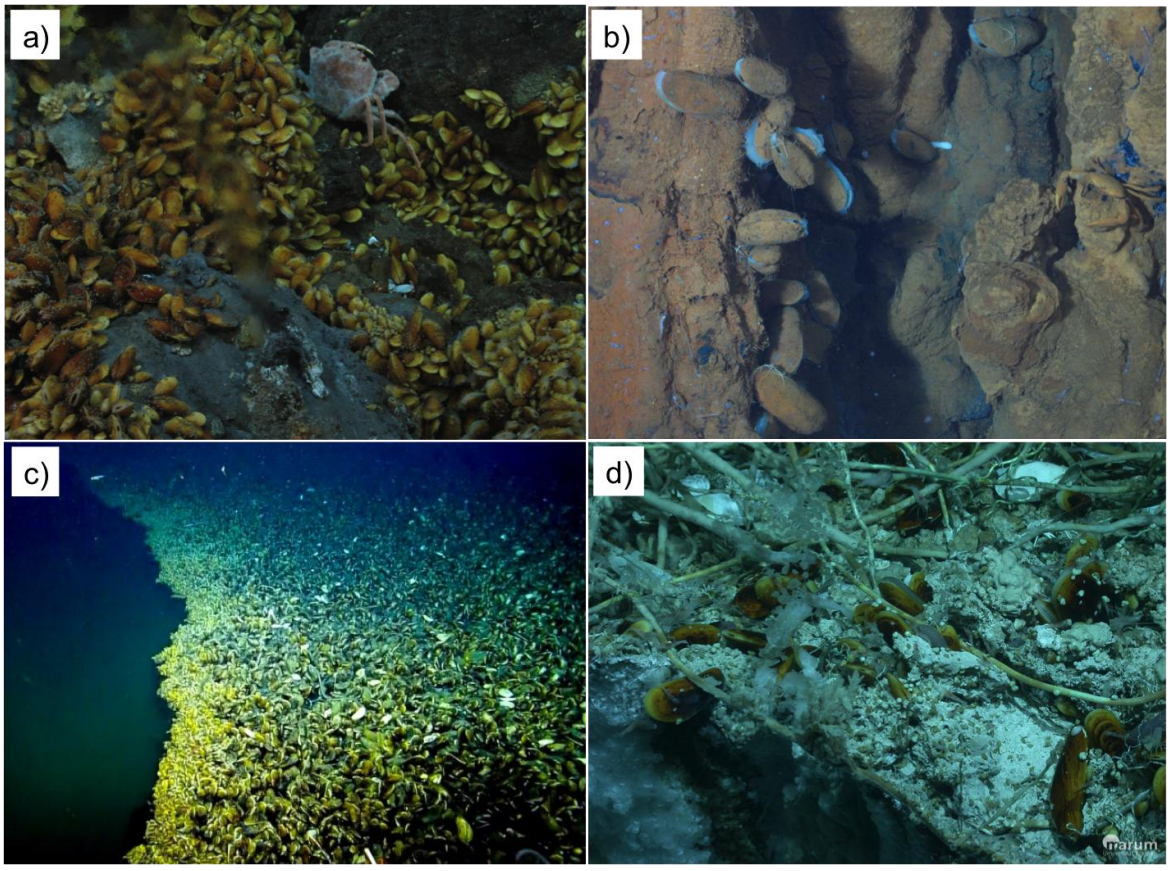


Figure 1-7. *Bathymodiolus* mussels colonize vents and seeps with different rock and fluid chemistry. a) A dense community of *B. azoricus* mussels from the Lucky Strike vent field located in the MAR. Note that fluids are coming out of the Earth´s crust. b) *B. azoricus* from the Rainbow vent field at NMAR. Iron oxides cover the fauna at this site, resulting in the orange coloration of the shells. (BioBaz cruise, 2013, Image © Marum). c) '*B'. childressi* form a dense community around a brine pool in the Northern Gulf of Mexico (NA044-45 cruise, 2013, Image © Ocean Exploration Trust). d) *Calyptogena* clams, tube worms, and mussels of the species *B. heckerae* and *B. brooksi* colonize a seep rich in gas hydrates located in Chapopote, Gulf of Mexico (M114-2 cruise, 2014, Image © Marum).

exploited by the host. Other free-living sulfur oxidizers have been shown to use other energy sources that are also readily available at vents. For example *Thiobacillus ferrooxidans* is an iron and sulfur oxidizer (Rawlings *et al*., 1999); *Paracoccus pantotrophus* and *Beggiatoa alba* can use reduced sulfur compounds and methanol as energy sources (Friedrich *et al*., 2001; Jewell *et al*., 2008). When key genes for energy use are known, genomics can be a powerful tool to predict further energy sources that could help the mussel's symbionts to exploit their niches.

1.2.3. Symbiont transmission of *Bathymodiolus* SOX symbionts

The SOX symbiont and its host are highly specific to each other, as suggested by the presence of only one SOX symbiont 16S rRNA phylotype[1] per host species (Dubilier *et al*., 2008). The exceptions are i) *B. azoricus* and *B. puteoserpentis*, two mussel species of the MAR that share the same SOX symbiont phylotype (Duperron *et al*., 2006), and ii) *B. heckerae* mussels that live in hydrocarbon-rich cold seeps in the Gulf of Mexico and host two different SOX symbiont phylotypes (Duperron *et al*., 2008; Raggi *et al*., 2013).

Based on morphological and phylogenetic evidence, *Bathymodiolus* symbionts are transmitted horizontally, i.e. from the environment to each generation (DeChaine *et al.*, 2006; Salerno *et al.*, 2005; Won *et al.*, 2003; Fontanez and Cavanaugh, 2014). Induced loss and reacquisition of the symbiont depending on the available energy source showed that the bacteria are taken up from the environment (Kadar *et al*., 2005). The symbiont infection process is highly specific, and begins early in the mussel's development. Juvenile mussels that are as small as seven millimeters already contain symbionts of their specific phylotype, and no other bacterial phylotypes (Wentrup *et al*., 2013). Symbiotic bacteria from the surrounding seawater continue to colonize the mussels across their entire life span, yet the symbiosis remains specific at the phylotype level (Wentrup *et al*., 2014).

Symbionts sharing exactly the same 16S rRNA sequence can however represent different strains with different gene contents and sequence variation. This has recently been shown for the SOX symbionts of *B. septemdierum* (Ikuta *et al*., 2015). Remarkably, the differences between SOX symbiont strains in *B. septemdierum* conferred metabolic

[1] A phylotype is defined as an identical 16S rRNA sequence.

flexibility. Some strains had the potential to use hydrogen as energy source in addition to reduced sulfur compounds, and some had the ability to use nitrate besides oxygen as electron acceptor.

1.2.4. Symbiotic close relatives

Based on the 16S rRNA gene phylogeny, the chemoautotrophic symbionts of *Bathymodiolus* mussels are most closely related to the chemoautotrophic symbionts of vesicomyid clams '*Ca*. Ruthia magnifica' and '*Ca*. Vesicomyosocious okutanii', which

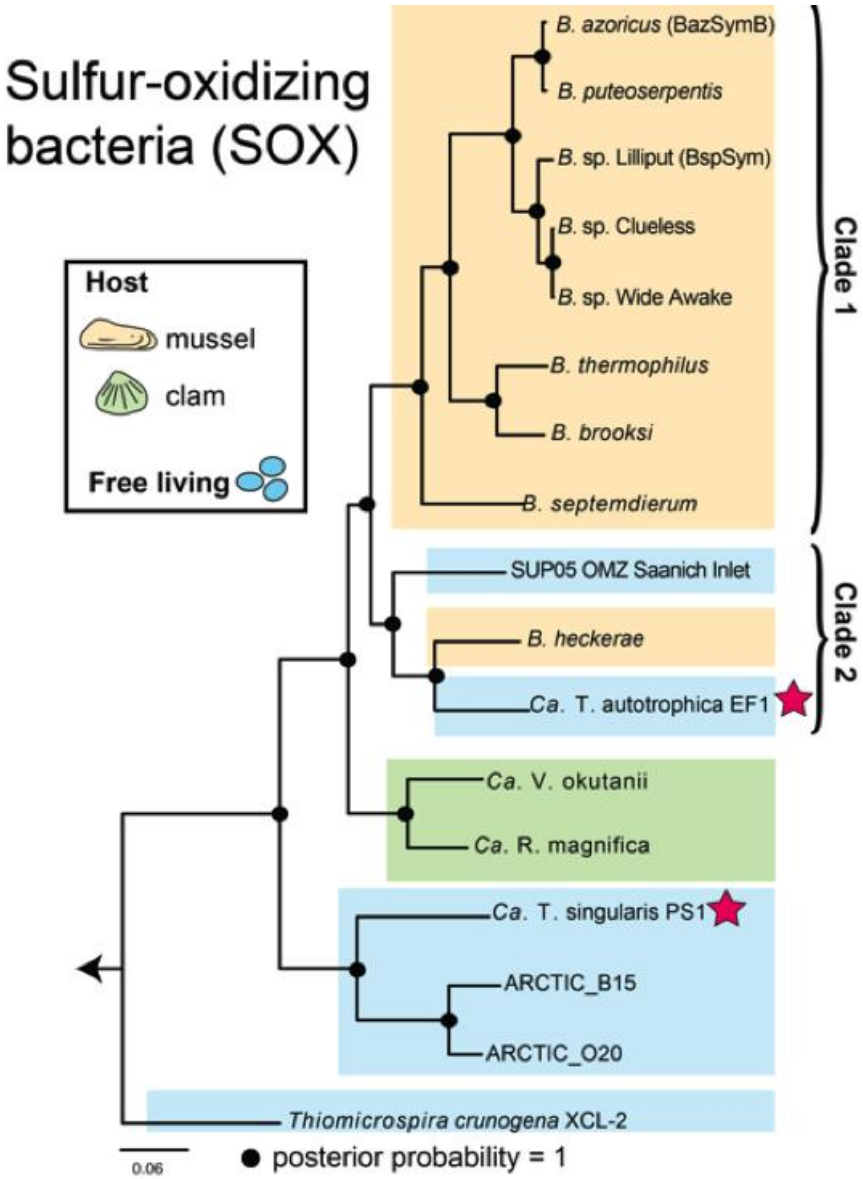


Figure 1-8. Species tree of the SOX symbionts and related sulfur-oxidizing bacteria that have a sequenced genome. Estimated from 19 orthologous genes and the 16S rRNA gene (modified from Chapter 3, Figure 1). Pink star represents free-living bacteria for which a culture and complete genome are now available (Marshall and Morris, 2013, 2015; Shah and Morris, 2015). Abbreviations: *Ca. Candidatus*; T, Thioglobus; V. Vesicomyosocius; R, Ruthia.

are also found at deep-sea hydrothermal vents and cold seeps (Figure 1-8) (Petersen *et al*., 2012; Sayavedra *et al*., 2015). Vesicomyid clams belong to the heterodonts, a different subclass from the bathymodiolines (Figure 1-4). Their gut and feeding groove are reduced and they depend almost entirely on their endosymbionts for nutrition (Figure 1-8) (Childress *et al*., 1991; Fiala-Médioni and Métivier, 1986). In contrast to the bathymodioline mussels, clam symbionts are vertically transmitted through the female germ line, with rare cases of mixed transmission mode (Box 2) (Cary and Giovannoni, 1993; Endow and Ohta, 1990; Stewart *et al*., 2008). Previous studies revealed that the symbiont genomes of the clam symbionts have a highly conserved gene content independent on whether the host was collected from a vent or seep ecosystem (Newton *et al*., 2008). As observed for other obligate intracellular symbionts, clam symbiont genomes appear to be in a state of ongoing genomic reduction and are AT biased due to the absence of the DNA repair gene *mutY* (see Box 2 – The effect of transmission mode in genome evolution) (Kuwahara *et al*., 2011, 2008, 2007; Rocha and Danchin, 2002).

Box 2 | The effect of transmission mode in genome evolution

One of the main questions in the symbiosis field is how the symbionts are transmitted from parent to offspring. This is especially important if the host depends on the symbiont for survival, as is the case for chemosynthetic nutritional symbioses. Symbionts are transmitted between host generations by two general routes: vertical transmission (e.g. symbionts transmitted though the germ line) and horizontal transmission (e.g. symbionts taken up from the environment with each new generation). The mode of transmission plays a crucial role in symbiont evolution, and recent whole-genome sequencing efforts have revealed some of the effects of transmission mode on symbiont genome evolution (Bright and Bulgheresi, 2010; Degnan *et al*., 2010; Gil *et al*., 2011).

Vertically transmitted symbionts suffer constant population bottlenecks, which reduces the effective size of the symbiont population (N_e). N_e is defined as "the size of the ideal population that would undergo the same amount of genetic drift as the actual population" (Soulé, 1987). As N_e shrinks, genetic drift has a higher effect in the population and mutations can get fixed despite being slightly deleterious (Ohta, 1987). Once genes become superfluous, they suffer an altered selective pressure, which often results in reduced genomes. Obligate symbionts may lose genes for pathways essential to free-living bacteria such as DNA repair, cytokinesis, and peptidoglycan synthesis (e.g. *Calyptogena* and some

Box 2 | … continued

insect symbionts have no *recA*, *mutZ*, and *ftsZ*) (Kuwahara *et al*., 2008, 2007; Moya *et al*., 2008). Still, in many cases they retain the genomic repertoire necessary for the production of metabolites, amino acids and vitamins necessary for the maintenance of their host's nutritional needs (Zientz *et al*., 2004). When the symbionts become dependent on the host, they often show similar evolutionary trajectories, a phenomenon called "cospeciation" (de Vienne *et al*., 2013). This often results in congruent phylogenies of the host and the symbiont, such as the *Buchnera* endosymbionts with their pea aphid hosts, as well as the vesicomyid endosymbionts with their clam hosts (reviewed by Moran *et al*., 2008; Peek *et al*., 1998). Symbiont transmission strategy is however not always clearly defined; occasional horizontal transmission can occur in vertical transmitted symbioses resulting in a **'mixed transmission mode'** (e.g. *Wolbachia*) (Bright and Bulgheresi, 2010).

Horizontally transmitted symbionts are taken up from the environment with each new generation and transmission is independent of the host reproduction. Phylogenies of host and symbiont often have an incongruent pattern (Bright and Bulgheresi, 2010). In contrast to the strict vertical transmission mode, symbionts acquired through horizontal transmission have often higher strain diversity and larger genomes with evidence of recombination (Baldo *et al*., 2006; Moran and Wernegreen, 2000; Wernegreen, 2005). The diversity of the symbiont strains that colonize a host will depend on the population present in the environment and in other hosts living close by (Vrijenhoek, 2010). The symbiont diversity within one host will also be affected by the time that a host remains "permissive" for infection. If the host can only be colonized during a brief period during development, a more limited number of symbiont strains will colonize the host. This is the case for vestimentiferan tubeworms and the squid-*Vibrio* symbiosis (Bright and Bulgheresi, 2010; Nyholm and McFall-Ngai, 2004). In contrast, hosts that can be colonized during their entire lifespans, such as bathymodioline mussels, will have a higher chance to have a more diverse symbiont population.

1.2.5. Free-living close relatives of the bathymodioline SOX symbionts

A number of free-living bacteria named 'SUP05' and 'ARCTIC' are closely related to bathymodioline symbionts, according to their 16S rRNA phylogeny (Chapter 2:

Figure 2- Figure Supplement 1) (Petersen *et al*., 2012; Sayavedra *et al*., 2015). SUP05 are abundant in oceanic oxygen minimum zones (OMZ) and hydrothermal plumes (Sunamura *et al*., 2004; Swan *et al*., 2011; Walsh *et al*., 2009). Like their symbiotic relatives, these bacteria are chemoautotrophic sulfur oxidizers, and some can use hydrogen as an energy source (Anantharaman *et al*., 2012; Walsh *et al*., 2009). ARCTIC are ubiquitous in the dark ocean, and they are a sister group to the symbiont/SUP05 cluster. ARCTIC is also a chemoautotrophic sulfur oxidizer that has a similar repertoire of pathways for oxidizing reduced sulfur compounds to SUP05, suggesting that similar electron donors are used in well-oxygenated waters (Swan *et al*., 2011). ARCTIC and SUP05 co-occur in oxygen minimum zones of the eastern South Pacific and eastern North Pacific. Besides their chemoautotrophic metabolism, alternative modes of energy metabolism have been suggested for ARCTIC and SUP05 bacteria, since they can be present in sulfidic and non-sulfidic environments (Walsh *et al*., 2009). Recently, bacterial strains falling within the SUP05 and ARCTIC clades have been isolated and sequenced, called *Ca*. 'Thioglobus autotrophica EF1' and '*Ca*. Thioglobus singularis PS1' (Figure 1-8) (Marshall and Morris, 2015; Shah and Morris, 2015). Previous metagenomic studies have shown that SUP05 and ARCTIC are obligate anaerobes that use nitrate as their sole electron acceptor (Walsh *et al*., 2009). However, both of the new isolates encode the genomic potential to use oxygen, and can grow both aerobically and anaerobically (Vega Shah, personal communication, 17 March 2016).

A major question for horizontally transmitted symbionts is how the host and symbiont recognize each other specifically in the environment and trigger the initiation of the symbiosis. The gills of bathymodioline mussels are exposed to all kinds of bacteria from the seawater throughout their lives, including pathogens, which potentially compete for the same niche than the symbionts. Because the SOX symbiont has close relatives with both free-living and symbiotic bacteria of which genomes are available, I aimed to look for genes that could mediate the symbiotic relationship in *Bathymodiolus* mussels. In this thesis, I compared the genomes of *Bathymodiolus* symbionts to the vertically transmitted clam symbionts and the free-living relatives SUP05 and ARCTIC. The available sequencing information within the time frame of each research study included in this thesis is described in detail in the corresponding chapters.

1.2.6. Molecular interactions of symbionts with their host

Microbial associations with eukaryotes are ubiquitous. The host is often dependent on their microbial symbionts for nutrition, defense or development (reviewed by Chaston and Goodrich-Blair, 2010; Gilbert *et al*., 2015; Oliver *et al*., 2014). Several associations have been shown to be highly specific, even when a large pool of different bacteria is in initial contact with the host (Bull and Rice, 1991; Duperron, 2010; Kaltenpoth *et al*., 2014; Kwong *et al*., 2014). To achieve the high specificity, host and symbiont need to communicate. The communication between bacteria and the host involves chemical signaling, a process called 'inter-kingdom signaling' (Hughes and Sperandio, 2008). The host usually responds to the symbiont by regulating the expression of its immune system, showing that the communication is not in one direction (Kendall and Sperandio, 2016). Invertebrates do not have an adaptive immune system nor an immunological memory, but the genes encoding the innate immune system are similar to those of vertebrates (Medzhitov and Janeway Jr, 1998). However, even if we would think of invertebrates as having only a simple immune system, it is nonetheless able to specifically recognize some pathogens (Lawniczak *et al*., 2007; Loker *et al*., 2004). Thus, in order for the symbiont to colonize the host, the symbiont must evade the host immune system, or the host immune system must recognize the symbiont without attacking it. Different molecular mechanisms have been described in insect symbioses. For example, during morphogenesis of the weevil insect *Sitophilus zeamais*, the host expresses peptidoglycan recognition proteins (PGRP) to prevent its intracellular symbiont from colonizing tissues other than the bacteriome (Anselme *et al*., 2006). In other cases, the insect host expresses an antimicrobial agent to which its symbiont is resistant (e.g. *Sodalis glossinidius* and its seconday symbiont) (Hao *et al*., 2001; Snyder and Rio, 2013). These examples show how flexible the mechanisms are that allow host-bacteria interactions.

Most of the well-characterized molecular mechanisms for host-bacteria interactions are from pathogens (Bhavsar *et al*., 2007). With the advent of genome sequencing, it is now recognized that beneficial symbionts and pathogens sometimes share common molecular mechanisms to recognize and colonize their hosts (reviewed by Goebel and Gross, 2001; Hentschel *et al*., 2000; Moya *et al*., 2008). The two most well recognized mechanisms shared between symbionts and pathogens are: i) quorum sensing, which is used by bacteria to communicate with each other and coordinate the behavior of

the population (reviewed by Whitehead *et al*., 2001). Both bacterial symbionts and pathogens can use quorum sensing to control the colonization of the host and the expression of virulence genes by hijacking the chemical signaling of the host (Sperandio *et al*., 2003); and ii) two-component regulatory systems, which are used to respond to changes in the environment. Two-component systems are controlled by a sensor kinase, which senses changes in pH, ion concentration, temperature or osmolarity, and a response regulator that converts the signal to a response from the cell (Beier and Gross, 2006; Hentschel *et al*., 2000).

Besides quorum sensing and two-component systems, other strategies shared by symbionts and pathogens involve effectors that control and corrupt the normal function of the cell. These effectors can be injected directly into the host with transporter apparatuses shaped like a needle, such as the type III, IV and VI secretion systems (Backert and Meyer, 2006; Bhavsar *et al*., 2007; Hueck, 1998). For example, the beneficial intracellular symbiont of *Sodalius glossinidius* uses a type III secretion system that has a very similar sequence to those used by pathogens from the Enterobacteriaceae (Goebel and Gross, 2001). *Buchnera,* the intracellular symbiont of the pea-aphid, encodes flagellar proteins, similar to those of *Yersinia enterocolitica*, which might be used for excreting proteins into the host (Shigenobu *et al*., 2000). The intracellular symbionts of *Amoeba proteus*, *Legionella*-like X bacteria, secrete a protein that reaches the nucleus and regulates the expression of amoeba's nuclear genes (Pak and Jeon, 1997). The traits required for colonization and maintenance of the symbiosis are prone to be transferred between different bacteria through horizontal gene transfer, mediated by phages, plasmids, and other mobile elements (reviewed by Gogarten and Townsend, 2005).

Often the effector 'virulence' genes are annotated as protein toxins, because by definition a microbial proteinaceous toxin is an agent that causes damage to a host tissue and that can disable the immune system (Olsnes *et al.*, 2013). Hentschel *et al*. (2000) has proposed the term 'symbiotic factor' to refer to the molecular mechanisms that overall contribute to beneficial symbioses. By this definition, toxins are considered symbiotic factors if they contribute to the fitness of the symbiosis in the long run. Indeed, beneficial symbionts that are acquired from the environment often cause an initial damage to their host during colonization but they are not detrimental in the long term. For example, during colonization of the beneficial symbiont *Vibrio fischeri,* the ducts and the crypt of the squid *Euprymna scolopes* are induced to produce the toxic gas nitric oxide (NO) (Nyholm and McFall-Ngai, 2004). This process is followed by apoptosis of epithelial

tissue induced by *V. fischeri*. A second example is the nitrogen-fixing bacteria *Rhizobium*, which during colonization of the root legumes can cause necrosis on the root nodules. These initially harmful mechanisms can eventually be compensated by the benefit of the symbiosis (Gourion *et al*., 2015).

Symbionts can also encode 'virulence genes', for which their target is not the host, but rather the host's enemies. Highly polymorphic toxins have been found in defensive symbionts such as the *Ca*. Hamiltonella defensa, which associates with the pea aphid *Acyrthosiphon pisum* and defends the host from parasitoid wasps, as well as the *Photorhabdus* and *Xenorhabdus* symbionts of soil nematodes (Goodrich-Blair and Clarke, 2007; Oliver *et al*., 2003; Rodou *et al*., 2010). In this thesis, I describe the presence of highly polymorphic 'toxins' in the nutritional *Bathymodiolus* symbionts that might be involved in the host-microbe crosstalk or in host defense (Sayavedra *et al*., 2015). I will however refer to the 'toxin' genes of the *Bathymodiolus* symbionts with the more neutral term 'toxin-related genes' as no detrimental effect has been shown to any eukaryote.

1.3. How to study chemosynthetic symbioses when the symbionts cannot be cultured

To date, only one chemosynthetic symbiont has been obtained in pure culture - the epsilonproteobacterial symbiont of *Alvinella pompejana* (Campbell *et al*., 2001). Methods that do not rely on cultivation are therefore invaluable for understanding these symbioses. Traditional methods for characterizing chemosynthetic symbioses include microscopy (light, confocal, scanning and transmission electron), enzymatic assays for key metabolic enzymes involved in chemoautotrophy and methanotrophy (e.g. RuBisCO, methanol dehydrogenase) and tracing the incorporation of isotopes labels (reviewed by Cavanaugh *et al*., 2013). These methods allow us to study morphology and physiology. Nonetheless, molecular methods are necessary to understand phylogenetic diversity and genetics, and can help us to uncover previously "hidden" metabolic pathways.

During the last two decades, a diverse array of molecular methods has been applied to symbiosis research. High throughput sequencing has become one of the most cost-effective methods to understand not only symbiotic, but also free-living communities (Iverson *et al*., 2012; Teeling and Glöckner, 2012). Genome and transcriptome sequencing is now being combined with proteomics, metabolomics, and contextual data,

which dramatically increases the complexity of the data to be analyzed (Teeling and Glöckner, 2012). Studying low-diversity systems such as highly specific symbiosis can therefore provide 'model systems' that are relatively easy to analyze and compare. Indeed, these so called 'omic' approaches have already proven to be highly valuable for understanding the genetic associations and metabolic interactions between different animals. For example, a syntrophic association between multiple bacterial symbionts of the gutless oligochaete *Olavius algarvensis* was analyzed using metagenomics and metaproteomics. *O. algarvensis* hosts two sulfur-oxidizing gammaproteobacterial symbionts and two sulfate-reducing deltaproteobacterial symbionts that cooperate to recycle the electron donors and acceptors present inside the worm (Kleiner *et al*., 2012b; Woyke *et al*., 2006). A second example is the tube worm *Riftia pachyptila* with its symbiont *Ca*. Endoriftia persephone, which showed that the symbiont from a single host expressed two different pathways for carbon fixation: the CBB cycle and the rTCA (tricarboxylic acid cycle) (Markert *et al*., 2011, 2007; Robidart *et al*., 2008).

Besides the metabolic traits, symbiont genome evolution has also been revealed through 'omic' approaches. For example, the genome evolution of the highly reduced genomes of *Buchnera*, as well as the Vesicomyd clam endosymbionts was revealed through sequencing approaches (Gil *et al*., 2002; Moran, 2003; Newton *et al*., 2008; Shigenobu *et al*., 2000). One of the most powerful tools at hand to unveil the host-microbe interactions is comparative genomic and transcriptomics with bacteria that are closely related and have different strategies for survival. These 'omic' approaches were therefore used in this thesis and are described in detail in the corresponding chapters.

1.4. Aims of this thesis

The broad aims of this thesis were to i) find potential molecular mechanisms that allow the SOX symbiont to achieve a very specific association with its host, and ii) to reconstruct the metabolism and energy sources of the SOX symbiont from different venting sites with contrasting environmental conditions. The specific questions examined in the next four chapters are the following:

1. **Which genes are unique to the horizontally transmitted SOX symbionts of *Bathymodiolus* and are potentially involved in host-symbiont interactions?**

The first aim was to investigate the effect of the transmission mode and the environment on genome evolution of the *Bathymodiolus* SOX symbionts and their closest

sequenced relatives. By comparing the genomes of *Bathymodiolus* SOX symbionts of two host species with the vertically transmitted clam symbionts and the free-living SUP05, I found a vast array of toxin-related genes (TRGs). The high abundance of TRGs was unprecedented for a nutritional symbiont (Chapter II). Because the presence of a gene does not necessarily imply that a gene is actually functioning in the organism, we aimed to examine which of the TRGs were expressed by using transcriptomics and proteomics. The results have been published and are therefore presented as a self-contained chapter (Chapter II).

2. At what point in evolution were the toxin-related genes (TRGs) of the SOX symbiont acquired?

This question was raised by our results from the previous question. I was able to sequence and assemble SOX symbiont genomes from *Bathymodiolus* species hosting single, dual, and multiple symbionts. A well-supported phylogenomic reconstruction showed that SOX symbionts have diverged in two independent lineages. One of the symbiont lineages grouped with the free-living relatives SUP05 and *Ca*. Thioglobus autotrophica (Figure 1-8 and Chapter III). Based on the results of Chapter II, I found that the abundant TRGs of the SOX symbionts had signatures of horizontal gene transfer. The unique phylogenetic relationship of the two SOX lineages gave me the opportunity to investigate the distribution of TRGs and to pinpoint the time in evolution where each class of TRG was acquired. The results are presented as a self-contained manuscript in Chapter III.

3. What genes do symbionts express to colonize their host?

Bathymodiolus mussels form new gill filaments and acquire their symbionts throughout their whole life span (Wentrup *et al*., 2014). The gill has a hot-spot for developing new filaments called the budding zone (Neumann and Kappes, 2003). This region was therefore likely to also represent a hot-spot for symbiont colonization. Because little is known about the behavior of chemosynthetic symbionts when they colonize their host, my aim was to investigate the differential gene expression between the budding zone and older gill filaments. As host and symbiont differentially expressed genes that are involved in metabolite production, we also aimed to confirm the transcriptomic results with metabolomics. The results are presented as a self-contained manuscript in Chapter IV.

4. What is the general metabolism of SOX symbionts?

I had access to the first sequenced SOX symbiont genomes of *Bathymodiolus* mussels. Thus, one of the aims of this thesis was to analyze the general metabolic pathways present in this symbiosis. The general metabolism and the energy sources of the SOX symbiont are described within the Appendix 1 of Chapter II, Chapter IV, and Chapter V.

5. When and how did the SOX symbiont acquire the ability to be a methylotroph? What roles do strain heterogeneity and the environment play in the acquisition of this trait?

Genome sequencing of the *B. puteoserpentis* SOX symbiont from the Logatchev vent field revealed the potential to use methanol as an additional energy source. Methanol is a byproduct of methane oxidation by the MOX symbiont, and might also be abundant in the vent fluids (Chapter V). The key gene for methanol oxidation, a methanol dehydrogenase (MDH), was not recovered in our previous draft genome 'bins' obtained from *B. azoricus* and *B.* sp 9° south (Chapter II). Genome coverage revealed that not all SOX symbiont strains encoded the gene, even though they had identical 16S rRNA. Thus, our major aim was to examine which *Bathymodiolus* species encoded the MDH and to know at what point the gene was acquired during the diversification of the SOX symbionts.

During the time of this thesis, new discoveries regarding the high strain heterogeneity of the SOX symbiont were made (e.g. Ikuta *et al.* (2015), Rebecca Ansorge *et al*. (in prep.), my own data). Thus, I also aimed to relate the venting conditions to the proportion of SOX symbiont strains that encoded and expressed the MDH gene. This is essential for understanding the effect of environmental conditions on symbiont activity and abundance. The results of these questions are presented in Chapter V.

List of publications and manuscripts with author's contributions:

1. **Abundant toxin-related genes in the genomes of beneficial symbionts from deep-sea hydrothermal vent mussels**

Lizbeth Sayavedra, Manuel Kleiner, Ruby Ponnudurai, Silke Wetzel, Eric Pelletier, Valerie Barbe, Nori Satoh, Eiichi Shoguchi, Dennis Fink, Corinna Breusing, Thorsten B.H. Reusch, Philip Rosenstiel, Markus B. Schilhabel, Dörte Becher, Thomas Schweder, Stephanie Markert, Nicole Dubilier, Jillian M. Petersen

Published in *eLife* (2015), doi: 10.7554/eLife.07966. Included in this thesis with publisher's permission.

2. **Comparative genomic insights into the roles of toxin-related genes in beneficial bacteria and their acquisition by horizontal gene transfer**

Lizbeth Sayavedra, Rebecca Ansorge, Nicole Dubilier, Jillian M. Petersen

Manuscript in preparation for *eLife* Research Advance.

3. **Understanding symbiont colonization in deep-sea mussels using differential gene expression**

Lizbeth Sayavedra, Rebecca Ansorge, Bruno Huettel, Manuel Liebeke, Nicole Dubilier, Jillian M. Petersen

Manuscript in preparation for ISME.

4. **Symbiont strain heterogeneity confers metabolic flexibility in deep-sea mussels: Methylotrophy in sulfur-oxidizing symbionts**

Lizbeth Sayavedra, Miguel Á. González-Porras, Chakkiath Paul Antony, Jimena Barrero-Canosa, Nicole Dubilier, Jillian M. Petersen

Manuscript in preparation.

Not included in this thesis:

5. **Metabolic and physiological interdependencies in the *Bathymodiolus azoricu*s symbiosis**

Ruby P. Ponnudurai, Manuel Kleiner, **Lizbeth Sayavedra**, Jillian M. Petersen, Martin Moche, Andreas Otto, Dörte Becher, Takeshi Takeuchi, Nori Satoh, Nicole Dubilier, Thomas Schweder, Stephanie Markert

Published in ISME, doi: 10.1038/ismej.2016.124.

6. **Biophysical and population genetic models predict the presence of "phantom" stepping stones connecting Mid-Atlantic Ridge vent ecosystems**

Corinna Breusing, Arne Biastoch, Annika Drews, Anna Metaxas, Didier Jollivet, Robert C. Vrijenhoek, Frank Melzner, **Lizbeth Sayavedra**, Jillian M. Petersen, Nicole Dubilier, Markus B. Schilhabel, Philip Rosenstiel, Thorsten B. H. Reusch

Published in Current Biology, doi: 10.1016/j.cub.2016.06.062.

7. **Short-chain alkanes fuel the metabolism of symbiotic *Cycloclasticus* from deep-sea gas and petroleum seeps**

Maxim Rubin Blum, Paul Antony Chakkiath, Christian Borowski, **Lizbeth Sayavedra**, Thomas Pape, Heiko Sahling, Gerhard Bohrmann, Nicole Dubilier

Manuscript under review in Nature Microbiology.

8. **Strain-level variation in symbiont populations of *Bathymodiolus* mussels along the Mid-Atlantic Ridge**

Rebecca Ansorge, **Lizbeth Sayavedra**, Jillian M. Petersen, Anne Kupczok, Tal Dagan, Halina Tegetmeyer, Nicole Dubilier

Manuscript in preparation.

9. **Metagenomic, metatranscriptomic and comparative genomics analyses of met hanotrophic endosymbionts of deep-sea mussels**

Chakkiath Paul Antony, **Lizbeth Sayavedra**, Harald Gruber-Vodicka, Maxim Rubin-Blum, Patricia L. Tavormina, Victoria J. Orphan, Nicole Dubilier

Manuscript in preparation.

10. **An alternative inorganic carbon fixation pathway in an epsilonproteobacterial epibiont of *Bathymodiolus* mussels**

Adrien Assié, **Lizbeth Sayavedra**, Harald Gruber-Vodicka, Samanta Joye, Matthew Saxton, Nicole Dubilier, Jillian M. Petersen

Manuscript in preparation.

Chapter II:

Abundant toxin-related genes in the genomes of beneficial symbionts from deep-sea hydrothermal vent mussels

Lizbeth Sayavedra[1], Manuel Kleiner[1,*], Ruby Ponnudurai[2,*], Silke Wetzel[1], Eric Pelletier[3,4,5], Valerie Barbe[3], Nori Satoh[6], Eiichi Shoguchi[6], Dennis Fink[1], Corinna Breusing[7], Thorsten B.H. Reusch[7], Philip Rosenstiel[8], Markus B. Schilhabel[8], Dörte Becher[9,10], Thomas Schweder[2;9], Stephanie Markert[2,9], Nicole Dubilier[1,11], Jillian M. Petersen[1#&]

[1]Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany

[2]Institute of Pharmacy, Ernst-Moritz-Arndt-University, Felix-Hausdorff-Strasse 3, 17487 Greifswald, Germany

[3]CEA - Genoscope, 2 rue Gaston Crémieux, 91000 Evry, France

[4]CNRS UMR8030, 2 rue Gaston Crémieux, 91000 Evry, France

[5]Université d'Evry Val d'Essonne, 2 rue Gaston Crémieux, 91000 Evry, France

[6]Marine Genomics Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495, Japan

[7]GEOMAR Helmholtz Centre for Ocean Research Kiel, Evolutionary Ecology, Düsternbrooker Weg 20, 24105 Kiel, Germany

[8]Institute of Clinical Molecular Biology (IKMB), Schittenhelmstr. 12 24105 Kiel, Germany

[9]Institute of Marine Biotechnology, Walther-Rathenau-Strasse 49a, 17489 Greifswald, Germany

[10]Institute of Microbiology, Ernst-Moritz-Arndt-University, Friedrich-Ludwig-Jahn-Strasse 15, 17487 Greifswald, Germany

[11]University of Bremen, 28359 Bremen, Germany

*Equal contribution

#Corresponding author – Jillian M. Petersen

Published in *eLife* (2015); 4:e07966

Author contributions

L.S., J.P. and N.D. designed research and conceived the project; **L.S.** analyzed genomic and transcriptomic data; M.K., R.P., D.B., T.S. and S.M. performed proteomic analyses and processed mass spectrometry data; S.W. developed symbiont extraction method; E.P. and V.B. sequenced and assembled samples sent to Genoscope; N.S. and E.S. sequenced and assembled samples in OIST; D.F. performed gradient centrifugation during the M82-3 cruise; C.B., T.R., P.R., and M.S. sequenced transcriptomes from *B.* sp.; **L.S**. and J.P. wrote the paper; All coauthors reviewed and revised the paper.

eLIFE elifesciences.org

RESEARCH ARTICLE

Abundant toxin-related genes in the genomes of beneficial symbionts from deep-sea hydrothermal vent mussels

Lizbeth Sayavedra[1], Manuel Kleiner[1†], Ruby Ponnudurai[2†], Silke Wetzel[1], Eric Pelletier[3,4,11], Valerie Barbe[3], Nori Satoh[5], Eiichi Shoguchi[5], Dennis Fink[1], Corinna Breusing[6], Thorsten BH Reusch[6], Philip Rosenstiel[7], Markus B Schilhabel[7], Dörte Becher[8,9], Thomas Schweder[2,8], Stephanie Markert[2,8], Nicole Dubilier[1,10], Jillian M Petersen[1*‡]

[1]Max Planck Institute for Marine Microbiology, Bremen, Germany; [2]Institute of Pharmacy, Ernst-Moritz-Arndt-University, Greifswald, Germany; [3]Genoscope - Centre National de Séquençage, Commissariat à l'énergie atomique et aux énergies alternatives, Evry, France; [4]Metabolic Genomics Group, Commissariat à l'énergie atomique et aux énergies alternatives, Evry, France; [5]Marine Genomics Unit, Okinawa Institute of Science and Technology, Onna, Japan; [6]Evolutionary Ecology, GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany; [7]Institute of Clinical Molecular Biology, Kiel, Germany; [8]Institute of Marine Biotechnology, Greifswald, Germany; [9]Institute of Microbiology, Ernst-Moritz-Arndt-University, Greifswald, Germany; [10]University of Bremen, Bremen, Germany; [11]University of Évry-Val d'Essonne, Evry, France

***For correspondence:** jmpeters@mpi-bremen.de

[†]These authors contributed equally to this work

Present address: [‡]Department of Microbiology and Ecosystem Science, Division of Microbial Ecology, Research Network Chemistry Meets Microbiology, University of Vienna, Vienna, Austria

Competing interests: The authors declare that no competing interests exist.

Funding: See page 22

Received: 09 April 2015
Accepted: 14 September 2015
Published: 15 September 2015

Reviewing editor: Axel A Brakhage, Friedrich Schiller University Jena and Hans-Knöll-Institut, Germany

Abstract *Bathymodiolus* mussels live in symbiosis with intracellular sulfur-oxidizing (SOX) bacteria that provide them with nutrition. We sequenced the SOX symbiont genomes from two *Bathymodiolus* species. Comparison of these symbiont genomes with those of their closest relatives revealed that the symbionts have undergone genome rearrangements, and up to 35% of their genes may have been acquired by horizontal gene transfer. Many of the genes specific to the symbionts were homologs of virulence genes. We discovered an abundant and diverse array of genes similar to insecticidal toxins of nematode and aphid symbionts, and toxins of pathogens such as *Yersinia* and *Vibrio*. Transcriptomics and proteomics revealed that the SOX symbionts express the toxin-related genes (TRGs) in their hosts. We hypothesize that the symbionts use these TRGs in beneficial interactions with their host, including protection against parasites. This would explain why a mutualistic symbiont would contain such a remarkable 'arsenal' of TRGs.
DOI: 10.7554/eLife.07966.001

Introduction

Mussels of the genus *Bathymodiolus* dominate deep-sea hydrothermal vents and cold seeps worldwide. The key to their ecological and evolutionary success is their symbiosis with chemosynthetic bacteria that provide them with nutrition (***von Cosel, 2002***; ***Van Dover et al., 2002***). *Bathymodiolus* mussels host their symbionts inside specialized gill epithelial cells called bacteriocytes (***Cavanaugh et al., 2006***; ***Petersen and Dubilier, 2009***).

Their filtering activity exposes *Bathymodiolus* mussels to a plethora of diverse microbes in their environment. Despite this, they are colonized by only one or a few specific types of chemosynthetic symbionts. Some mussel species associate exclusively with sulfur-oxidizing (SOX) symbionts that use

eLife digest Although bacteria are commonly associated with causing illness, many are actually beneficial to the organism they live in or on. The phenomenon of one species helping another to survive is known as symbiosis.

Animals thrive at hydrothermal vents in the deep sea because of their partnerships with symbiotic bacteria. The bacteria use the geochemical energy found at hydrothermal vents to convert carbon into sugars, thus providing their animal hosts with essential nutrients. Unlike the symbiotic communities that associate with humans and other mammals, in which thousands of bacterial species co-exist, deep-sea mussels associate with just one or two species of symbiotic bacteria. This relative simplicity is ideal for investigating how the intimate associations between animals and bacteria work.

Genes contain the instructions cells and organisms need to survive, and so one way that researchers investigate symbiosis is by studying the genes of the organisms involved. Such studies of beneficial bacteria are beginning to reveal that the molecular mechanisms involved in symbiosis are remarkably similar to those responsible for the harmful effects produced by some bacteria.

By performing genetic sequencing on the symbiotic bacteria from deep-sea mussels, Sayavedra et al. have discovered that the bacteria have an unusually large number of toxin-like genes, and that all of these genes are active in the bacteria when they are inside host mussels. This was unexpected, as the bacteria are known to benefit their mussel hosts. The toxin-like genes from the symbiotic bacteria are similar to toxins found in the bacteria that cause diseases such as cholera and the plague in humans and other animals.

Sayavedra et al. suggest that the symbiotic bacteria have 'tamed' these toxins to use them in beneficial interactions with their host. For example, some of the toxins could help the bacteria and mussels to recognize and interact with each other, and others could help to protect the mussel host from its natural enemies. The next step will be to test these ideas, which will be challenging as the mussels cannot be bred in the laboratory.

DOI: 10.7554/eLife.07966.002

reduced sulfur compounds and sometimes hydrogen as an energy source, and carbon dioxide as a carbon source. Some have only methane-oxidizing (MOX) symbionts that use methane as an energy source and carbon source. Some mussel species host both types in a dual symbiosis (**Fisher et al., 1993**; **Distel et al., 1995**; **Duperron et al., 2006**; **Dubilier et al., 2008**; **Petersen et al., 2011**). In all species except one, a single 16S rRNA phylotype for each type of symbiont (SOX or MOX) is found in the gills (**Dubilier et al., 2008**). There are more than 30 described *Bathymodiolus* species, and most associate with a characteristic symbiont phylotype, which is not found in other species (**Duperron et al., 2013**).

Although these associations are clearly very specific, the molecular mechanisms that underpin this specificity are still unknown. No chemosynthetic symbiont has ever been obtained in pure culture. Therefore, molecular methods for investigating uncultured microbes have been essential for understanding their biodiversity, function, and evolution (reviewed by **Dubilier et al., 2008**).

The *Bathymodiolus* symbionts are assumed to be horizontally transmitted, which means that each new host generation must take up their symbionts from the surrounding environment or co-occurring adults (**Won et al., 2003b**; **Kadar et al., 2005**; **DeChaine et al., 2006**; **Fontanez and Cavanaugh, 2014**; **Wentrup et al., 2014**). To initiate the symbiosis, hosts and symbionts must have evolved highly specific recognition and attachment mechanisms. Once they have been recognized, the symbionts need to enter host cells and avoid immediate digestion, just like other intracellular symbionts such as *Burkholderia rhizoxinica* and *Rhizobium leguminosarum*, or pathogens such as *Legionella*, *Listeria*, or *Yersinia* (**Hentschel et al., 2000**; **Moebius et al., 2014**). Indeed, like many intracellular pathogens, the *Bathymodiolus* symbionts seem to induce a loss of microvilli on the cells they colonize (**Cossart and Sansonetti, 2004**; **Bhavsar et al., 2007**; **Haglund and Welch, 2011**; **Wentrup et al., 2014**). Finally, the symbionts achieve dense populations inside the host cells (e.g., **Duperron et al., 2006**; **Halary et al., 2008**). Therefore, they must be able to avoid immediate digestion by their hosts. Although the mechanisms of host cell entry and immune evasion have been extensively studied in pathogens and plant–microbe associations such as the rhizobia-legume symbiosis, far less is known about the mechanisms beneficial symbionts use to enter and survive within animal host cells.

The symbiosis between *Vibrio fisheri* bacteria and *Euprymna scolopes* squid is one of the few beneficial host-microbe associations where the molecular mechanisms of host-symbiont interaction have been investigated. A number of factors are involved in initiating this symbiosis such as the symbiont-encoded 'TCT toxin', which is related to the tracheal cytotoxin of *Bordetella pertussis* (**McFall-Ngai et al., 2013**). A few studies of intracellular insect symbionts have shown that they use type III and type IV secretion systems to establish and maintain their association with their host (reviewed by **Dale and Moran, 2006**; **Snyder and Rio, 2013**). These secretion systems are commonly used by intracellular pathogens to hijack host cell processes, allowing their entry and survival within host cells (e.g., **Hueck, 1998**; **Steele-Mortimer et al., 2002**; **Backert and Meyer, 2006**). An example is the *Sodalis* symbionts of aphids and weevils, which use a type III secretion system for entry to the host cell and are thought to have evolved from pathogens (**Dale et al., 2001**; **Clayton et al., 2012**). The virulence determinants of their pathogenic ancestors might therefore have been co-opted for use in beneficial interactions with their insect hosts.

In contrast to the *Sodalis* symbionts of insects and the *Vibrio* symbionts of squid, the *Bathymodiolus* SOX symbionts are not closely related to any known pathogens. Moreover, because they fall interspersed between free-living SOX bacteria in 16S rRNA phylogenies, they are hypothesized to have evolved multiple times from free-living ancestors (**Figure 2—figure supplement 1**) (**Petersen et al., 2012**). Comparative genomics is a powerful tool for identifying the genomic basis of beneficial and pathogenic interactions, particularly if the symbionts or pathogens have close free-living relatives that do not associate with a host (e.g., **Galagan, 2014**; **Ogier et al., 2014**; **Zuleta et al., 2014**). Genomes of closely related free-living and symbiotic relatives of *Bathymodiolus* SOX symbionts were recently published. Their closest free-living relatives are marine SOX bacteria called SUP05, which are abundant in the world's oceans, particularly in oxygen minimum zones (OMZs) and hydrothermal plumes (**Sunamura et al., 2004**; **Lavik et al., 2009**; **Walsh et al., 2009**; **Anantharaman et al., 2012**; **Wright et al., 2012**; **Petersen and Dubilier, 2014**). The *Bathymodiolus* SOX symbionts and SUP05 bacteria form a monophyletic clade together with the SOX symbionts of vesicomyid clams based on 16S rRNA gene phylogenies (**Figure 2—figure supplement 1**) (**Distel et al., 1995**; **Petersen et al., 2012**). Closed genomes are available for the symbionts of two clam species (**Kuwahara et al., 2007**; **Newton et al., 2007**).

All members of this monophyletic group (the mussel and clam symbionts, and SUP05) share similar core metabolic features. They are all capable of autotrophic growth, and all use reduced sulfur compounds as an energy source (**Newton et al., 2008**; **Walsh et al., 2009**). They can differ in auxiliary metabolic capabilities such as hydrogen oxidation, nitrate reduction, or mixotrophy (**Newton et al., 2008**; **Petersen et al., 2011**; **Anantharaman et al., 2012**; **Murillo et al., 2014**). However, the major difference between these organisms is their lifestyle: SUP05 bacteria are exclusively free-living. The clam symbionts are exclusively host-associated, are vertically transmitted, and have reduced genomes. The *Bathymodiolus* symbionts appear to have adapted to both niches, as they have a host-associated stage and are assumed to also have a free-living stage.

The goal of this study was to identify the genomic basis of host-symbiont interactions in *Bathymodiolus* symbioses. We used high-throughput sequencing and binning techniques to assemble the first essentially complete draft genomes of the SOX symbionts from *Bathymodiolus* mussels. We used comparative genomics of the symbionts' genomes to those of their close free-living and obligate symbiotic relatives to reveal genes potentially involved in *Bathymodiolus* host-symbiont interactions. We used phylogenetics and bioinformatic prediction of horizontally acquired genes to investigate the origins of these genes. Finally, we used transcriptomics and proteomics to determine whether potential host-symbiont interaction genes are being expressed by the symbionts in their host.

Results

Draft genome sequences of *Bathymodiolus* symbionts

We sequenced the genomes of the SOX symbionts from three *Bathymodiolus* individuals: two were *Bathymodiolus azoricus* from the Menez Gwen vent field on the northern Mid-Atlantic Ridge (MAR) (**Figure 1**). We refer to these as BazSymA and BazSymB. The third mussel individual was an undescribed *Bathymodiolus* species (BspSym), from the Lilliput hydrothermal vent on the southern MAR (SMAR) (**Figure 1**). Symbiont draft genomes from each individual were almost complete (see 'Materials and methods'). Despite different sequencing and assembly strategies, the draft genomes

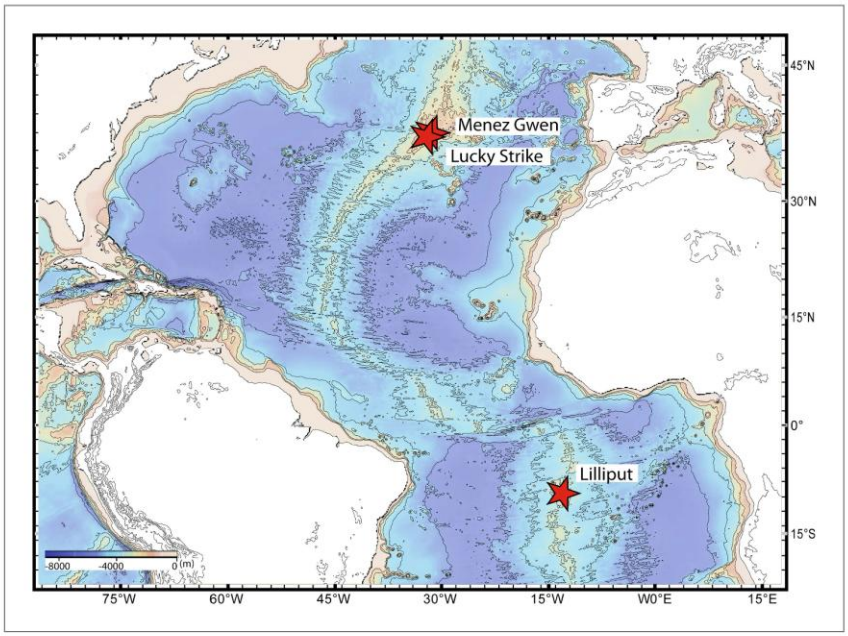


Figure 1. Sampling sites. Map showing the sampling sites of *Bathymodiolus* mussels at hydrothermal vents along the Mid-Atlantic Ridge (red stars). *B.* sp. is found at Lilliput (BspSym), *Bathymodiolus azoricus* at Menez Gwen (BazSymA and BazSymB) and Lucky Strike. The details of the sampling sites are described in ***Supplementary file 1E***. The map was produced with GeoMapApp 3.3.

DOI: 10.7554/eLife.07966.003

were 90.7–97.7% complete (***Table 1***). The total assembly sizes were between 1.7 and 2.3 Mbp, on 52 to 506 contigs (***Table 1***). Each draft genome contained one copy of the 16S rRNA gene. The BazSymB assembly only contained an 829 bp fragment of the 16S rRNA gene; however, we PCR amplified and sequenced this gene from the DNA used to generate the metagenome. The 16S rRNA genes from the two *B. azoricus* symbionts were 100% identical and were 99.3% identical to BspSym. The core metabolic potential of the *Bathymodiolus* SOX symbionts is described in Appendix 1-Symbiont metabolism. A detailed description of the genomes is beyond the scope of this article and will be published elsewhere.

General genome comparison

Bathymodiolus symbiont genomes were more similar to the free-living SUP05 than to the clam symbionts in terms of size and GC content (***Table 1***). Analysis of codon usage showed that all three *Bathymodiolus* SOX symbiont genomes had a greater proportion of genes that may have been acquired through recent horizontal gene transfer (HGT) compared to the clam symbionts *Candidatus* Vesicomyosocius okutanii, and *Candidatus* Ruthia magnifica (***Table 1***). BspSym, the genome that assembled into the fewest contigs, lacked synteny compared to SUP05 and the clam symbionts, as shown by whole genome alignment (***Figure 2—figure supplement 2***). Genome alignment of BazSymA and BazSymB was not attempted because the assemblies were highly fragmented. The possibility of incorrect genome assembly for BspSym was ruled out for four regions by PCR amplification of sequences spanning the regions without synteny. For confirmation, one PCR product was Sanger-sequenced and found to be identical to the draft genome assembly of BspSym. These regions without genome synteny therefore most likely represent true genome reshuffling in *Bathymodiolus* symbionts.

The *Bathymodiolus* symbiont genomes had more mobile elements compared to their closest relatives (***Supplementary file 1A***). *Bathymodiolus* symbionts had between 13 and 23 transposases and three to five integrases. SUP05 had 14 transposases and one integrase. We did not find any

Table 1. Overview of the genomes compared in this study: SOX symbiont of *B.* sp, two individual SOX symbionts of *B. azoricus*, SOX symbiont *Candidatus* Vesicomyosocious okutanii, SOX symbiont of *Calyptogena magnifica* (*Candidatus* Ruthia magnifica), and free-living SUP05

Genome	Collection site	Contigs	GC content (%)	Length/Span (Mbp)†	Number of CDSs	HGT	Estimated completeness‡	Coverage§	Separation method#	References
B. sp symbiont (BspSym)	Lilliput	52	38.23	1.8/2.3	2225	33%	95.39%	22X	Filtration	***Petersen et al., 2011***, this study
B. azoricus symbiont (BazSymB)	Menez Gwen	239	38.20	1.5/1.7	1802	30%	90.60%	8X	Gradient centrifugation/ binning	This study
B. azoricus symbiont (BazSymA)*	Menez Gwen	506	37.58	1.85/1.85	2008	35%	97.70%	59X	Binning	This study
Ca. V. okutanii	Sagami Bay	1	31.59	1.0/1.0	980	26%	93.58%	–	Whole genome assembly	***Kuwahara et al., 2007***
Ca. R. magnifica	East Pacific Rise, 9°N	1	34.03	1.2/1.2	1210	23%	94.84%	–	Whole genome assembly	***Newton et al., 2007***
SUP05	Saanich Inlet	97	39.29	1.4/2.5	1586	30%	85.76%	–	Binning	***Walsh et al., 2009***

SOX, sulfur-oxidizing.
*SOX symbiont sequences recovered from metagenome of adductor muscle.
HGT = Genes that potentially originated from horizontal gene transfer.
†Length is the total length of sequence information on contigs without Ns, and span is the entire length of scaffold assembly including Ns.
‡The completeness of the genome was estimated with CheckM using a set of lineage-specific genes for proteobacteria (***Parks et al., 2015***).
§Median coverage.
#Separation method indicates the experimental separation of symbionts from host tissue and co-occurring symbionts (filtration or gradient centrifugation), or the in silico separation of genomic information from hosts and co-occurring bacteria (binning).
DOI: 10.7554/eLife.07966.004

transposases or integrases in the clam symbiont genomes. The *Bathymodiolus* symbionts were highly enriched in restriction-modification system genes (between 10 and 22 genes), whereas SUP05 only had one, and the clam symbionts had none. This large difference raises the possibility that restriction-modification systems are involved in genome reshuffling in the *Bathymodiolus* symbionts.

Gene-based comparison reveals toxin-related genes specific to *Bathymodiolus* symbionts

Between 2.3 and 7.6% of the genes found only in the *Bathymodiolus* SOX symbionts but not in the clam symbionts and SUP05, genomes were annotated as toxin or virulence genes (***Figure 2***). Most were related to genes from one of three toxin classes: (1) the RTX (repeats in toxins) toxins, (2) MARTX (multifunctional autoprocessing RTX toxins), a sub-group of RTX toxins, and (3) YD repeat toxins (also called *rhs* genes as they were initially described as 'recombination *h*otspots'). Representatives from all three toxin-related genes (TRGs) classes were found in each of the three-draft genomes, except for RTX, which were not found in BazSymA. The number of genes from each class varied between the three genomes. We found the largest number of TRGs in the genome with the fewest contigs, BspSym, which had at least 33 YD repeat genes, eight RTX genes, and 19 MARTX-like genes. In the BazSymB genome, 14 YD repeat genes, two RTX genes, and up to 10 MARTX genes were found. BazSymA had 16 YD repeat genes, and one MARTX (***Figure 2***, ***Supplementary file 1B***). This indicates that these toxin-related classes are common to the SOX symbionts of both *B.* sp. and *B. azoricus*. In the BspSym genome, which assembled into the largest contigs, 22 out of 88 TRGs were found directly upstream or downstream of mobile elements.

Toxin genes are known to have unusually high substitution rates due to an 'evolutionary arms race' with their targets (***Jackson et al., 2009***; ***Linhartová et al., 2010***). Accordingly, many of the *Bathymodiolus* symbiont TRGs were highly variable between the symbionts of *B. azoricus* and *B.* sp., but also between the symbionts from the two *B. azoricus* individuals, and even between different copies within one genome (see 'Variability of TRGs within *Bathymodiolus* SOX symbiont populations'). We therefore searched for homologs of the TRGs in the clam symbionts and SUP05 with a lower BSR of up to 0.25, but even with this reduced stringency, no hits were found (see 'Materials and methods').

We ruled out the possibility that the TRGs were not found in SUP05 draft genome because of its incompleteness (~89% complete), by searching for homologs of the *Bathymodiolus* symbionts TRGs in unbinned metagenomes and metatranscriptomes from hydrothermal plumes and OMZs that are enriched in SUP05. If free-living SUP05 also encoded these TRGs, we would expect to find them regularly in SUP05-enriched metagenomes and metatranscriptomes. Instead, no hits were found in four out of these six data sets. In a metagenome from the Lost City hydrothermal vent, we found one weak hit to a YD repeat gene (31% similarity). In a metagenome from the Guaymas Basin hydrothermal vent, we found one weak hit to an RTX gene (34% similarity). However, both of these metagenomes were from sites colonized by either *Bathymodiolus* mussels (Lost City), or *Riftia pachyptila* tubeworms (Guaymas Basin), whose symbionts also encode a hemolysin gene of the RTX class (***Gardebrecht et al., 2012***). These rare hits might therefore come from contamination by symbionts in the environment (***Harmer et al., 2008***). Considering the almost complete absence of genes similar to the TRGs of *Bathymodiolus* in SUP05-enriched next-generation sequence data sets, we conclude that these genes are specific to the *Bathymodiolus* SOX symbionts and are not found in their close symbiotic or free-living relatives.

Relationships to other toxins

Most symbiont TRGs were so divergent that they could not be confidently aligned. One exception was the YD repeat genes, a few of which contained a conserved repeat region. We reconstructed the phylogeny of this conserved region. YD repeat sequences from the symbionts of both *Bathymodiolus* species formed a distinct cluster, distant from all other sequences in public databases (***Figure 3***, ***Figure 3—figure supplement 1***). The *Bathymodiolus* symbiont sequences did not cluster according to their host species, but instead were intermixed, suggesting gene duplication events prior to the divergence of the *B. azoricus* and *B.* sp. symbiont lineages. The *Bathymodiolus* symbiont sequences fell into a cluster that contained mostly pathogenic bacteria such as *Yersinia pestis* and *Burkholderia pseudomallei*. This cluster was well supported by Bayesian analysis (0.91 posterior probability). This cluster also contained a number of beneficial symbionts such as *B. rhizoxinica*, which is an intracellular

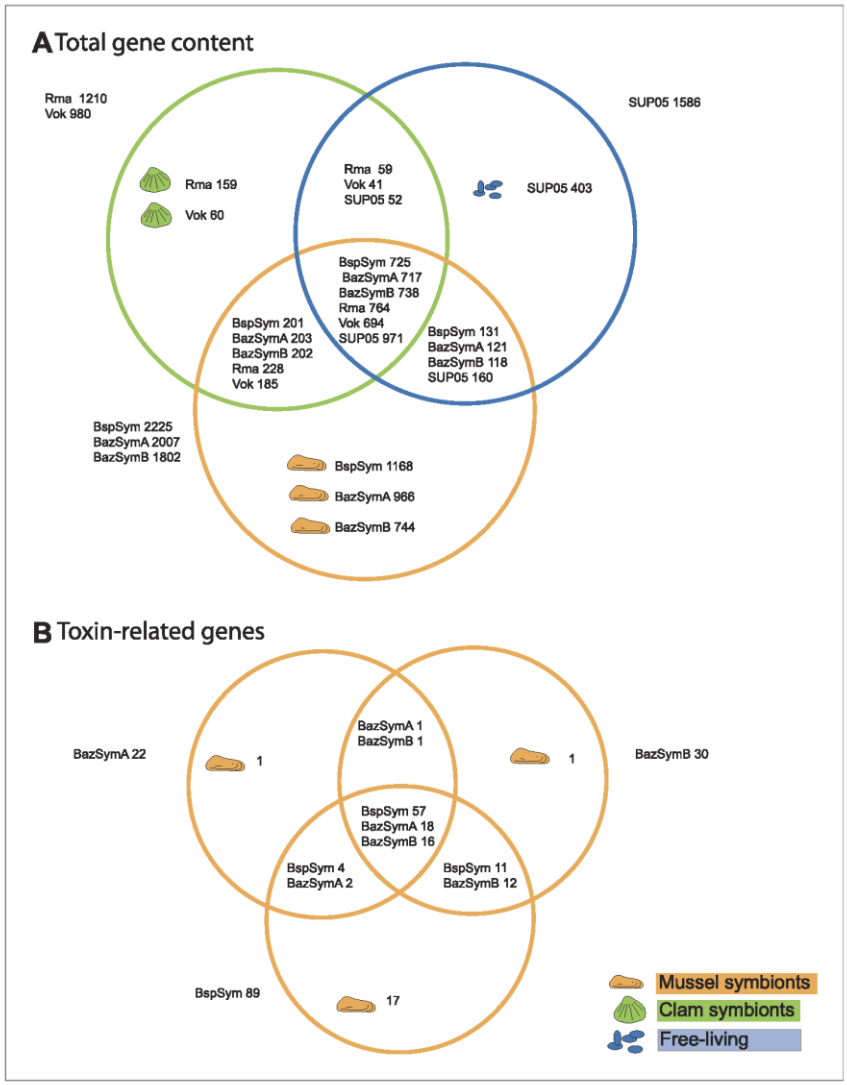


Figure 2. Genes shared between the *Bathymodiolus* and vesicomyid SOX symbionts and free-living SUP05. Protein-coding sequences from the *Bathymodiolus* sulfur-oxidizing (SOX) symbiont were compared to the clam symbiont genomes and to the SUP05 metagenome from ***Walsh et al. (2009)*** with BLAST score ratios (BSR). (**A**) Venn diagram of the shared and unique gene content in the clam symbionts, mussel symbionts, and SUP05 bacteria. Predicted protein sequences of each mussel symbiont were compared to a combined data set of the clam symbionts (Rma and Vok) and SUP05. Similarly, protein sequences of each clam symbiont were compared to a combined data set of mussel symbionts (BspSym, BazSymB, and BazSymA). Depending on the reference genome, the number of shared genes varies slightly and possibly reflects the presence of paralogous genes and redundant sequence information in these draft genomes. Abbreviations are explained in detail in ***Table 1***. The BLAST score ratio (BSR) threshold was 0.4. (**B**) Venn diagram of mussel symbiont toxin-related genes (TRGs), calculated with a BSR threshold of 0.2.
DOI: 10.7554/eLife.07966.005

The following figure supplements are available for figure 2:

Figure supplement 1. Maximum likelihood 16S rRNA phylogeny of the close relatives of the *Bathymodiolus* SOX symbionts.
DOI: 10.7554/eLife.07966.006

Figure supplement 2. Whole genome alignment.
DOI: 10.7554/eLife.07966.007

Figure 2. continued on next page

Figure 2. Continued

Figure supplement 3. Metabolic reconstruction of the Bathymodiolus symbiont.
DOI: 10.7554/eLife.07966.008

symbiont of the fungus *Rhizopus*, and the *Photorhabdus* and *Xenorhabdus* symbionts of soil nematodes (***Waterfield et al., 2001***; ***Goodrich-Blair and Clarke, 2007***; ***Moebius et al., 2014***).

To overcome the difficulties in aligning these highly divergent TRGs, we constructed gene sequence similarity networks based on BLAST to depict relationships among and between the symbiont TRGs, and those in public databases. This analysis revealed distinct sequence clusters that contained genes with >25% similarity over at least half of the length of the gene (***Figure 4***). If a cluster

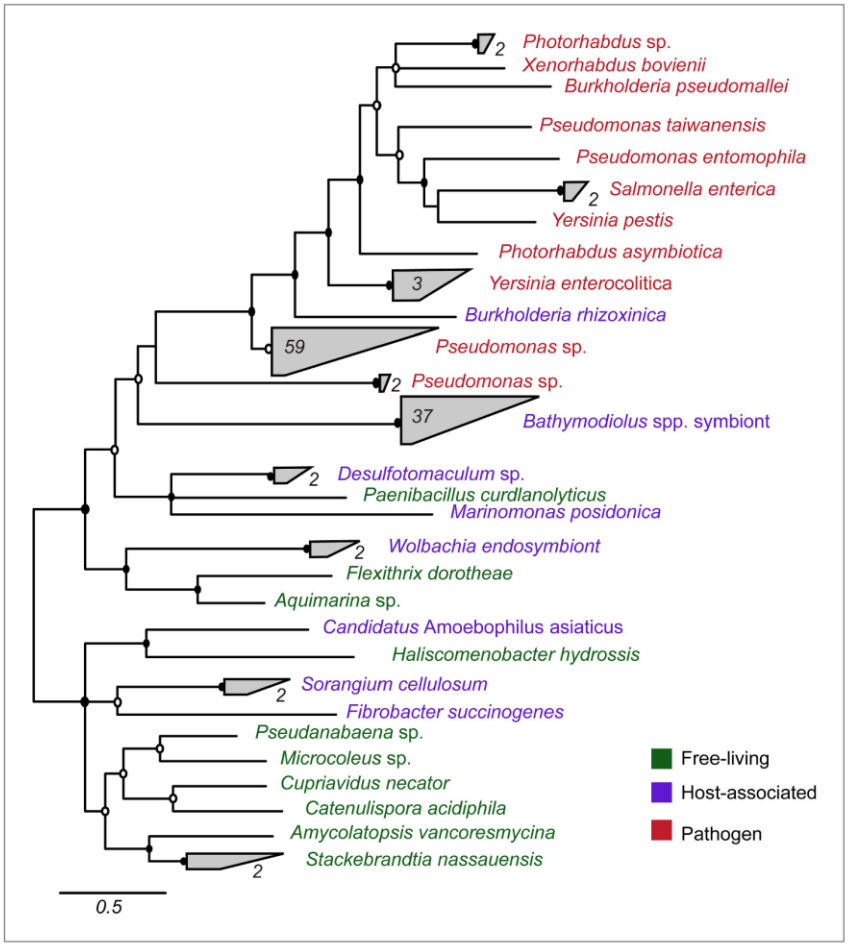


Figure 3. Phylogeny of YD repeat-containing proteins. The tree is a consensus of bayesian and maximum likelihood analyses, result of an alignment of 536 amino acids. Black circles represent branches with posterior probability >0.8 and bootstrap value >80. White circles represent branches with either posterior probability >0.8 or bootstrap value >80. The number of sequences per collapsed group is shown next to the gray bloks. Purple: organism found in intestinal microflora or in close association with another organism; green: free-living; red: pathogen.
DOI: 10.7554/eLife.07966.009

The following figure supplement is available for figure 3:

Figure supplement 1. Consensus of bayesian and maximum likelihood phylogeny of YD proteins with identifiers.
DOI: 10.7554/eLife.07966.010

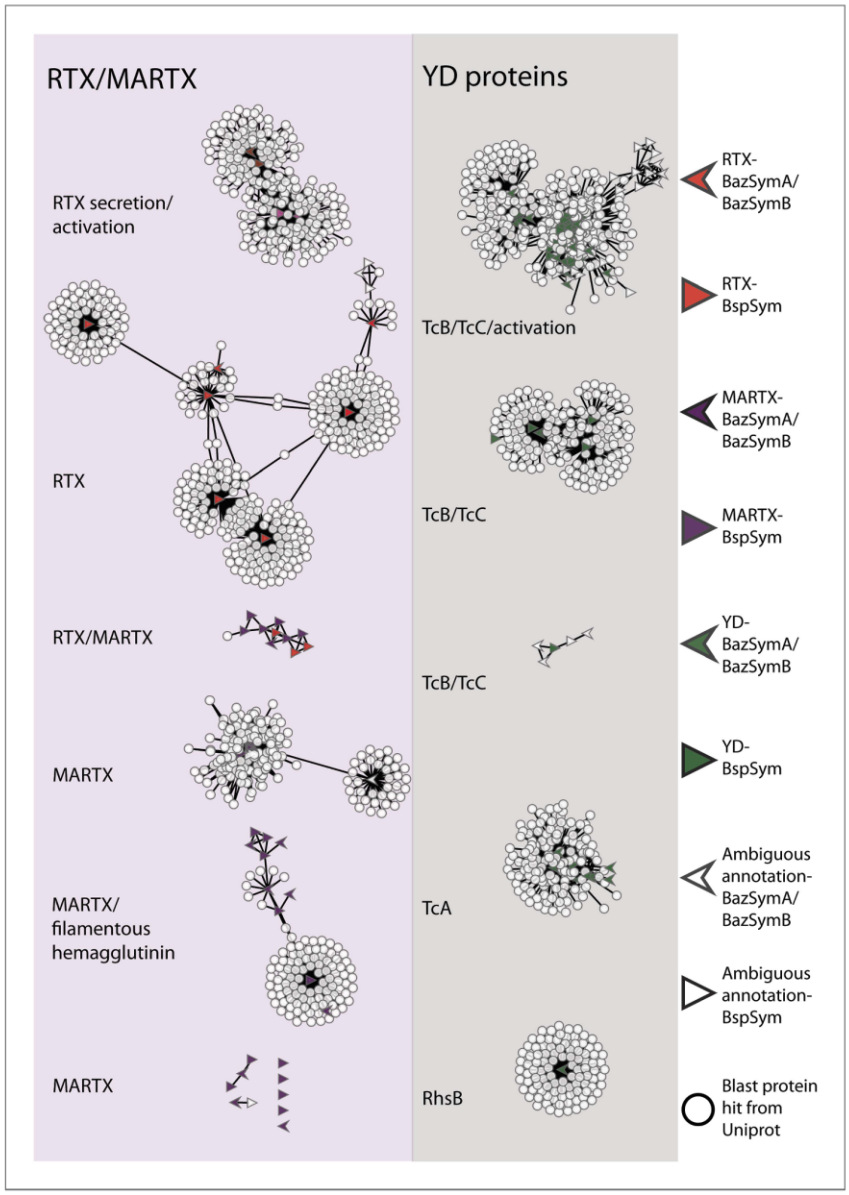


Figure 4. Protein similarity network of toxin-related proteins in the *Bathymodiolus* symbionts. Each node corresponds to a protein sequence and the links between nodes represent BLAST hits. The length of the edges is inversely proportional to the sequence similarity. Protein clusters containing RTX or multifunctional autoprocessing RTX (MARTX) proteins are shown in the red panel on the left, and sequence clusters containing YD repeats are shown in the gray panel on the right. Arrowheads are proteins from *B. azoricus* symbionts, and triangles are proteins from *B.* sp. symbionts. The symbols are colored in green if they were identified in the Bathymodiolus symbionts as YD repeat-containing genes, red if they were identified as RTX genes, and purple for MARTX genes. Some protein sequences were similar to the TRGs but not annotated as such as these are partial genes that did not have any conserved domain. If the clusters contained mostly genes with a particular annotation, we named the clusters after these annotations, for example, cluster 'TcB/TcC' contained proteins annotated as TcB or TcC.

DOI: 10.7554/eLife.07966.011

Figure 4. continued on next page

Figure 4. Continued

The following figure supplements are available for figure 4:

Figure supplement 1. Network of toxin-related proteins in the *Bathymodiolus* symbionts with BLAST hits from *Vibrio*, *Photorhabdus*, *Xenorhabdus*, and *Pseudomonas* highlighted.

DOI: 10.7554/eLife.07966.012

Figure supplement 2. Genomic architecture of MARTX regions.

DOI: 10.7554/eLife.07966.013

contained at least one gene that was similar to at least one other gene in another cluster (similarity cut-offs as above), then these clusters were joined to create a larger network. They were also joined if both clusters contained genes that had similarity to another gene in the database (i.e., if they could be joined by at most two steps). This allowed us to identify distinct sub-groups within the three toxin-related classes, and to identify toxin sequences from public databases that were most similar to the *Bathymodiolus* symbiont genes.

TRGs from the *Bathymodiolus* symbionts clustered together with toxin and TRGs from phylogenetically diverse organisms including characterized toxins of gammaproteobacterial *Vibrio* and *Pseudomonas*, and TRGs of the gammaproteobacteria *Shewanella*, the actinobacterial *Rhodococcus*, the cyanobacterium *Trichodesmium*, and the firmicute *Caldicellulosiruptor* (see e.g., ***Figure 4—figure supplement 1***). The RTX genes clustered into two separate networks, one that had similarity to RTX secretion and activation genes, and one that had similarity to RTX toxins. Five distinct networks contained MARTX genes. One of these included genes from the symbiont MARTX1 cluster, and genes from other organisms that were annotated as MARTX or filamentous hemagglutinin. One network contained some genes that we classified as RTX and some we classified as MARTX, reflecting their shared features such as the RTX repeats. Eight MARTX genes had no significant hits to any other gene in public databases.

The YD repeat genes formed five distinct networks. Sequences from the first three had structural similarity to TcB and TcC, two subunits of the ABC toxins of *Photorhabdus* and *Xenorhabdus*, the beneficial symbionts of entomopathogenic nematodes. The B and C subunits form a cage-like structure that encapsulates the toxic domain (an adenosine diphosphate (ADP) ribosylation domain, located at the C terminus of the C subunit). The A subunit forms a syringe-like structure, which delivers the toxin to the insect cell (***Meusch et al., 2014***). The genes in the fourth YD network had structural similarity to TcA genes that encode the syringe-like A subunit. The fifth YD network had similarity to genes annotated as RhsB, which was shown to play a role in bacteria–bacteria competition in *Escherichia coli* (***Poole et al., 2011***).

MARTX and YD repeat genes are enriched in the genomes of host-associated bacteria

The *Bathymodiolus* symbiont genomes encoded more YD repeat and MARTX genes than any other genome that we compared them to (***Figure 5***, ***Figure 5—figure supplements 1–3***). This is remarkable considering the relatively small size of their genomes, and the fact that they are still incomplete. A few published genomes encoded more RTX genes, but these were much larger (>5 Mbp) (***Figure 5—figure supplement 3***).

The vast majority of RTX, MARTX, and YD repeat proteins have not been functionally characterized. The few proteins whose function has been studied in detail are from bacteria that are known pathogens or cultured strains that can form biofilms. Because of this, it is generally assumed that RTX, MARTX, and YD repeat proteins function in host-microbe interactions, in microbe–microbe antagonism, or in biofilm formation, but this has not been extensively tested. To further investigate the functional role of the TRGs encoded by the *Bathymodiolus* symbionts, we tested whether similar genes are more likely to be found in bacteria that live in a particular niche (extracellular host-associated, intracellular host-associated, or free-living), or that express a particular phenotype (pathogenesis or biofilm formation).

First, we used the Kruskal–Wallis one-way analysis of variance to determine whether the distribution of the three TRGs classes differed significantly (I) between biofilm-forming vs non-biofilm-forming bacteria, (II) between pathogenic and non-pathogenic bacteria, and (III) between free-

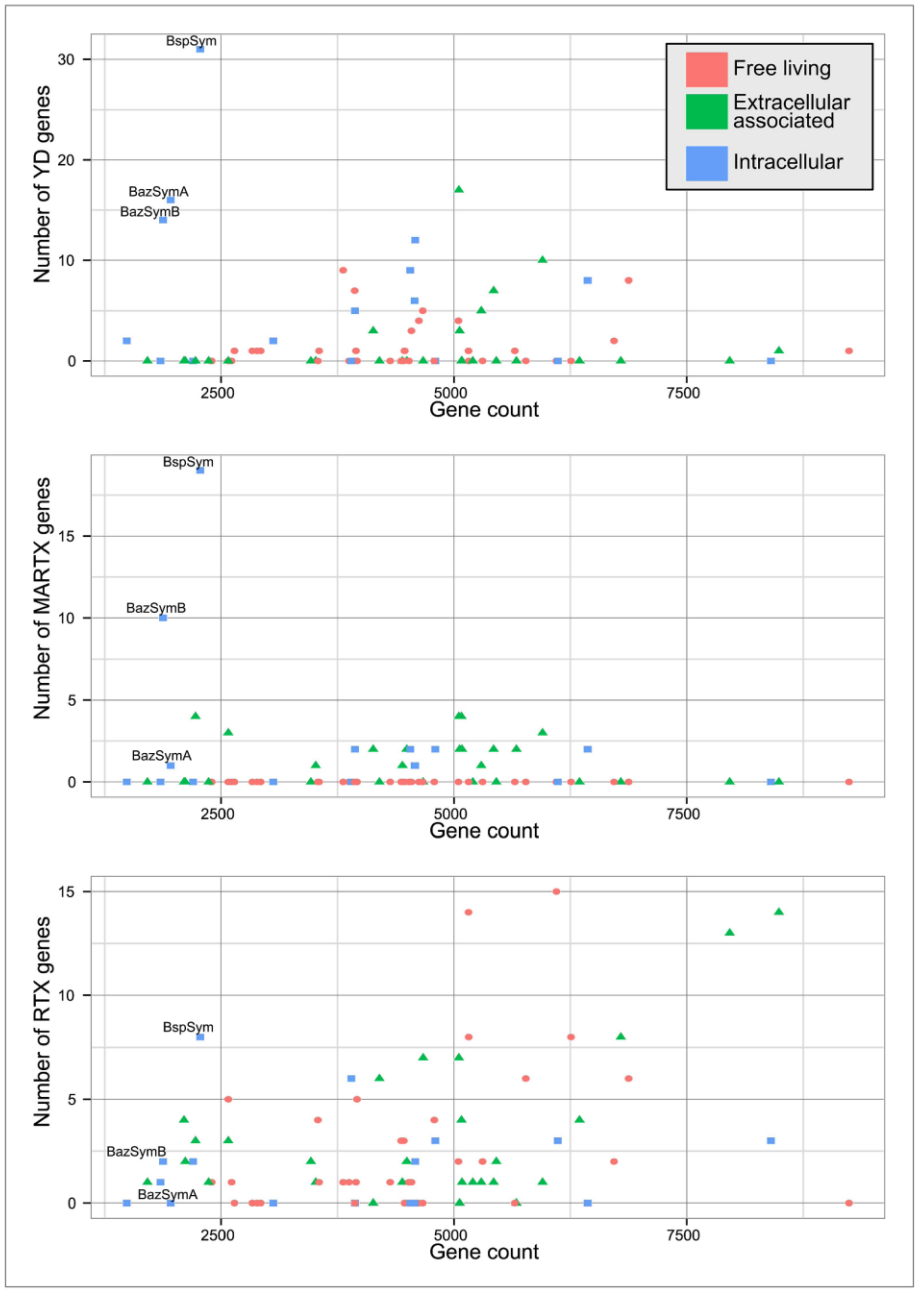


Figure 5. Distribution of the three major TRGs classes according to lifestyle. Each dot represents one sequenced genome. The sum of TRGs is on the Y axis, and the total number of genes predicted in each genome is on the X axis. Free-living bacteria are shown in red, host-associated bacteria that live outside of host cells are in green, and host-associated bacteria that can live inside host cells are shown in blue. The positions of the Bathymodiolus SOX symbionts are indicated. A detailed overview of all organisms that had similar TRGs to the SOX symbiont with the number of TRGs is shown in ***Supplementary file 1B***.

DOI: 10.7554/eLife.07966.014

The following figure supplements are available for figure 5:

Figure supplement 1. YD genes per genome, normalized to the total gene count.

DOI: 10.7554/eLife.07966.015

Figure 5. continued on next page

Figure 5. Continued

Figure supplement 2. MARTX genes per genome, normalized to the total gene count.
DOI: 10.7554/eLife.07966.016

Figure supplement 3. RTX genes per genome, normalized to the total gene count.
DOI: 10.7554/eLife.07966.017

living bacteria, host-associated intracellular bacteria, and host-associated extracellular bacteria, without considering whether the bacteria were pathogenic. There was no significant enrichment of any TRG category in bacteria known to form biofilms vs those that do not (***Table 2***). One class, MARTX was significantly enriched in the genomes of pathogenic vs non-pathogenic bacteria (p-value = 0.007, Kruskal–Wallis test). There was also a significant bias in the distribution of genes encoding MARTX and YD repeat genes in bacteria according to their lifestyle (extracellular host-associated, intracellular host-associated, or free-living).

When three categories are tested, such as in (III) above, the Kruskal–Wallis test does not identify which category the bias is associated with. To tease apart which of these three niche categories was most enriched in TRGs, we did Mann–Whitney–Wilcoxon tests (***Table 3***). These showed that both YD repeat and MARTX genes were enriched in the genomes of host-associated microbes (YD repeat: p-value = 0.026, MARTX: p-values = $2.125e^{-6}$, $1.618e^{-6}$, Mann–Whitney–Wilcoxon test). While MARTX genes were enriched in host-associated bacteria regardless of their location, YD repeat genes were only significantly enriched in intracellular bacteria. In contrast to YD repeat and MARTX genes, RTX did not show any enrichment in the three defined categories. RTX are therefore widely distributed among bacteria and are just as likely to be found in free-living and host-associated bacteria (Appendix 2).

Bacteria that are closely related often have similar genomic and physiological features. However, toxin genes are commonly gained through HGT, which may weaken the phylogenetic signal in their distribution patterns (reviewed by ***Dobrindt et al., 2004***; ***Gogarten and Townsend, 2005***). To tease apart the possible phylogenetic influence on the TRGs distribution, we used Permanova to test whether any of the three classes was enriched in particular phylogenetic groups at the class, order, and family levels. Only RTX genes were significantly enriched, and only at the order level (p-value = 0.0159) (***Supplementary file 1C***). Therefore, phylogeny is not the main driver in the toxin-related distribution of YD repeats genes and MARTX.

Variability of TRGs within *Bathymodiolus* SOX symbiont populations

Toxin genes often have unusually high substitution rates, making them highly variable compared to non-toxin genes (e.g., ***Ohno et al., 1997***; ***Davies et al., 2002***). We compared the substitution rates of all genes in the *Bathymodiolus* SOX symbiont genomes within the population of symbionts associated with each mussel species. This was done for each of the two species, *B. azoricus* and *B.* sp. by mapping transcriptome reads from three *Bathymodiolus* individuals to the draft genomes of their respective symbionts (see below for transcriptomes). We calculated the number of single nucleotide polymorphisms (SNPs) per kb per gene. The number of SNPs in most of the TRGs was not significantly higher than the genome-wide average (***Figure 6***). However, we found 22 TRGs that did have

Table 2. p-values obtained with Kruskal–Wallis rank sum test

	B/NB df = 1	P/NP df = 1	Ext/Int/FL df = 2
YD	0.097	0.5217	0.010*
RTX	0.715	0.793	0.308
MARTX	0.773	0.007*	$3.21e^{-06}$*

The three main lifestyle categories were tested against each toxin-related class. Number of TRGs per genome was normalized to the total gene count.
FL = free-living, Ext = extracellular host-associated, Int = intracellular host-associated, P = pathogen, NP = non-pathogen, B = found in biofilms, NB = not found in biofilms, df = degrees of freedom, TRG, toxin-related gene, MARTX, multifunctional autoprocessing RTX.
*p-value was considered to be significant ($p < 0.05$).
DOI: 10.7554/eLife.07966.018

Table 3. p-values obtained with Mann–Whitney–Wilcoxon test for enrichment of YD and MARTX genes similar to those from the SOX symbiont

	FL/Ext	FL/Int	Ext/Int
YD	0.129	0.026	0.006*
MARTX	$2.125e^{-06}$*	$1.618e^{-06}$*	0.751

FL = free-living, Ext = extracellular host-associated, Int = intracellular host-associated, MARTX, multifunctional autoprocessing RTX, SOX, sulfur-oxidizing.
*p-value was considered to be significant ($p < 0.05$).
DOI: 10.7554/eLife.07966.019

significantly more SNPs than most of the other genes in the genomes. Among the 22 highly variable genes, we found representatives of each TRG class: YD repeats, RTX, and MARTX (***Figure 6***).

Expression of TRGs

Transcriptome sequencing revealed that all predicted TRGs of the SOX symbionts in *B. azoricus* and *B.* sp gills were expressed. Reads mapping to TRGs accounted for 0.67–1.71% of mRNA in *B. azoricus* symbionts and 0.58–3.14% in *B.* sp. symbionts. All TRGs were found in the transcriptomes of at least one of the three individuals that we sequenced per species (***Supplementary file 1D***). The expression levels of some genes from the RTX, MARTX, and YD repeats classes were in some cases higher than the expression of the essential Calvin cycle gene ribulose bisphosphate carboxylase/oxidase (RuBisCO), which

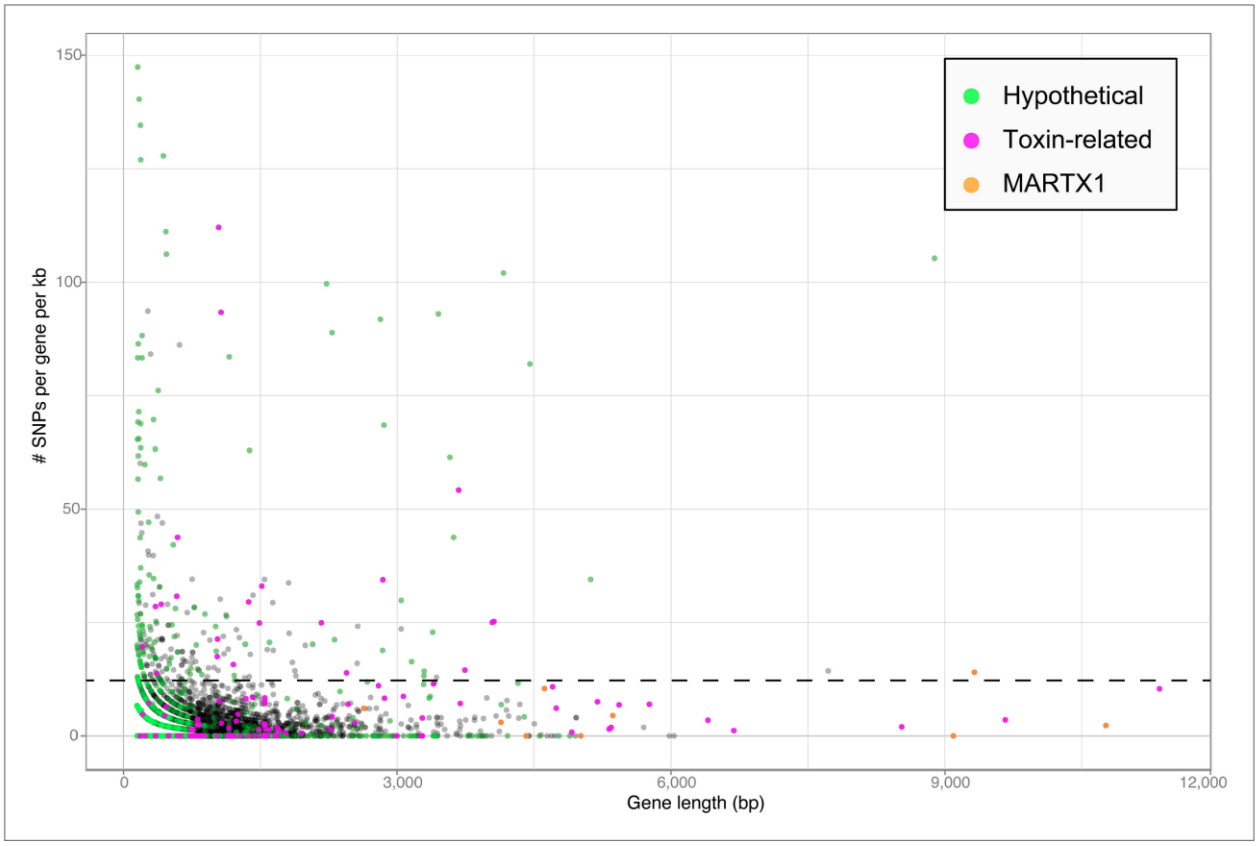


Figure 6. Single nucleotide polymorphisms per gene. The number of single nucleotide polymorphisms (SNPs) per gene was normalized according to the length minus regions of unknown sequence for genes containing N's. Genes smaller than 150 bp were not considered. The dotted line represents the median plus one standard deviation of the number of SNPs per gene per kb.
DOI: 10.7554/eLife.07966.020
The following source data is available for figure 6:

Source data 1. Variability in TRGs encoded by the *Bathymodiolus* SOX symbionts.
DOI: 10.7554/eLife.07966.021

accounted for 0.03–0.5% of mRNA in *B. azoricus* symbionts and 0.11–1.04% in *B.* sp. symbionts. We could also confirm expression of some TRGs at the protein level in *B. azoricus* mussels (samples of *B.* sp. were not available for proteomics). We analyzed (I) gradient centrifugation fractions that were enriched in the *Bathymodiolus* SOX symbiont and (II) whole gill tissue. Some of the characterized toxin proteins are associated with membranes. We therefore also analyzed a fraction enriched in membrane proteins. In these proteomes, we found 12 YD repeat proteins and one RTX. We also found a number of toxin-associated proteins such as one RTX activator and two RTX toxin transporters (***Table 4***). Nine of the 12 identified YD repeat proteins were present in the membrane proteome. Our method allowed us to identify symbiont-encoded proteins that were found in higher relative abundance in the whole gill fraction compared to the symbiont-enriched fraction (see Appendix 4). These proteins are potentially exported to the host tissue, indicating that they play a role in direct host-symbiont interactions. One RTX activator and one YD repeat protein were significantly enriched in the host tissue (***Table 4***).

Discussion

Origin of TRGs in *Bathymodiolus* symbionts

Two scenarios could explain the origin of the large complement of TRGs in the *Bathymodiolus* symbionts and not in their close relatives: firstly, the TRGs could have been in the genome of their last

Table 4. Toxin-related proteins found in the proteome of the SOX symbiont from *B. azoricus*

Proteome	Identifier	Category	Annotation	Molecular weight (kDa)	Max. number of unique peptides*
SM	Host_EST_000107	YD	IPR006530IYD repeat	43	11
SM	Host_EST_000115	YD	IPR006530IYD repeat	42	12
N	Host_EST_000248	YD	IPR006530IYD repeat	37	7
M	Host_EST_002123	YD	IPR006530IYD repeat	24	5
SM†	Thio_BAZ_1943_contig360420_0	RTX (activator)	Hemolysin-activating lysine-acyltransferase (Hemolysin C)	19	3
SM	Tox_BAZ_119_contig00027_0	YD	RHS repeat-associated core domain-containing protein	202	17
SM	Tox_BAZ_120_contig00027_1	YD	Virulence plasmid 28.1 kDa A protein	62	11
SM	Tox_BAZ_1734_contig02141_2	RTX (transporter)	Secretion protein HlyD family protein	43	10
SM	Tox_BAZ_2494_contig00030_0	YD	Virulence plasmid 28.1 kDa A protein	183	33
M	Tox_BAZ_3202_scaffold00038_7	RTX	Hemolysins and related proteins containing CBS domains	35	2
SM	Tox_BAZ_525_contig104979_0	YD	Virulence plasmid 28.1 kDa A protein	52	2
S†	ToxAzor_892893	YD	Rhs	114	2
SM	ToxSMAR_1260BAT01109	YD	[weak similarity to] Toxin complex/ plasmid virulence protein	321	8
N	ToxSMAR_2052BAT01788, Thio_BAZ_1733_contig02141_1 or Thio_BAZ_2580_scaffold00010_8	RTX (transporter)	Toxin secretion ATP-binding protein	79	5
S	ToxSMAR893-894	YD	Rhs family protein	103	1
SM	ToxAzor_890891	YD	Rhs family protein	67	6

S = soluble proteome, M = membrane-enriched proteome, SM = found in both proteomes, SOX, sulfur-oxidizing.
*The highest number of unique peptides detected in one sample.
†Proteins that are potentially exported by the symbiont to the host gill tissue.
DOI: 10.7554/eLife.07966.022

common ancestor but were all subsequently lost in both the clam symbionts and SUP05. Alternatively, the *Bathymodiolus* symbionts could have acquired these genes via HGT after their divergence from the clam symbionts and SUP05. Toxins are often found in 'genomic islands' that have been acquired through HGT (reviewed by **Lindsay et al., 1998**; **Ochman et al., 2000**; **Dobrindt et al., 2004**; **Soucy et al., 2015**). Several observations point towards HGT of the TRGs into the *Bathymodiolus* symbionts rather than their loss by the clam symbionts and SUP05. Firstly, 63–68% of the TRGs could be identified as potentially horizontally acquired based on codon usage analysis, in contrast to 30–35% predicted for all coding sequences (**Table 1**, **Figure 6—source data 1**). This means that their transfer was relatively recent, as the codon usage of these genes has not yet adapted to one typical of the symbiont genomes. Secondly, the content of the *Bathymodiolus* symbiont genomes attest to the major role that HGT has played in their evolution. They are enriched in mobile elements such as transposases and restriction-modification systems compared to their closest relatives (**Supplementary file 1A**). The lack of synteny we observed in the symbiont genomes is consistent with the presence of mobile elements and major HGT events (**Kobayashi, 2001**; **Rocha and Danchin, 2002**; **Achaz et al., 2003**). Thirdly, the TRGs from the *Bathymodiolus* symbionts are similar to genes from distantly related bacteria. Finally, mobile elements were regularly found directly upstream or downstream of the TRGs in the *Bathymodiolus* symbiont genomes. The linkage of mobile elements with some of these genes could explain the mechanism of their transfer into the *Bathymodiolus* symbiont genomes. This could also explain why each genome contained multiple copies of TRGs, as mobile elements are also prone to duplication (**Reams and Neidle, 2004**). Considering these observations, and the absence of these TRGs in their close relatives, we consider it most likely that they were acquired through HGT.

Are some of the TRGs used for direct beneficial interactions between hosts and symbionts?

Genes from the second major class of TRGs, MARTX, that were similar to those from *Bathymodiolus* symbiont, were significantly enriched in beneficial and pathogenic host-associated bacteria. One of the regions encoding MARTX genes from the *Bathymodiolus* SOX symbiont has a domain structure similar to the filamentous hemagglutinin FhaB from *B. pertussis* and *Bordetella bronchiseptica* (MARTX1, see Appendix 3). FhaB is involved in attachment of *Bordetella* to their human host and suppression of the immune response (reviewed by **Melvin et al., 2014**). *B. bronchiseptica* has two distinct phenotypic stages: an infective stage, where *fhaB* is upregulated, and a non-infective stage, where *fhaB* is downregulated. The non-infective stage is necessary for its survival in the environment outside of the host. The lifestyle of the *Bathymodiolus* symbiont has striking similarities to *B. bronchiseptica*. The symbionts must also survive in the environment to be transmitted from one host generation to the next. Our transcriptomes showed that these MARTX genes are expressed by *Bathymodiolus* symbionts within the host tissue. Unfortunately, we do not have samples to test whether they are downregulated in environmental symbiont stages. If so, it may have a similar function to its homologs in pathogenic *Bordetella*. MARTX-like genes also mediate cell–cell attachment in the symbiotic bacterial consortium '*Chlorochromatium aggregatum*' (**Vogl et al., 2008**; **Liu et al., 2013**).

Are some of the YD repeat proteins used for antagonistic bacteria–bacteria interactions?

Like the MARTX, YD repeat proteins were also significantly more enriched in host-associated compared to free-living bacteria. Some characterized YD repeat proteins function in competition between closely related bacterial strains (**Waterfield et al., 2001**; **Kung et al., 2012**; **Koskiniemi et al., 2013**). In its intracellular niche, why would the *Bathymodiolus* symbiont need to compete against other bacteria? Multiple strains of SOX symbionts can co-occur in *Bathymodiolus* mussels (**Won et al., 2003a**; **DeChaine et al., 2006**; **Duperron et al., 2008**). These may compete with each other for nutrients and energy, or for space within host cells. Bacteria that express toxins to inhibit their close relatives need an immunity protein to protect them from each toxin (**Zhang et al., 2012**; **Benz and Meinhart, 2014**). These immunity proteins are encoded immediately downstream or upstream of the toxin protein. The toxin-immunity pair is usually linked to genes encoding a type VI secretion system, which is the mechanism of toxin delivery for all so far described YD proteins involved in bacteria–bacteria competition (**Zhang et al., 2012**; **Benz and Meinhart, 2014**). None of the YD repeat proteins in the *Bathymodiolus* SOX symbiont genome was found in an operon with an

identifiable immunity protein, and no type VI secretion system gene was found in any of our draft genomes. The arrangement of YD repeat genes in the *Bathymodiolus* symbiont genomes and the lack of genes encoding type VI secretion systems are therefore inconsistent with a role in competition between closely related symbiont strains.

Bathymodiolus mussels are also infected by bacterial intranuclear pathogens called *Candidatus* Endonucleobacter bathymodioli, which are related to the genus *Endozoicomonas* (**Zielinski et al., 2009**). *Ca*. E. bathymodioli invades the mussel cell nuclei where it multiplies, eventually bursting the infected cell. Intranuclear bacteria are never found in the nuclei of symbiont-containing cells, which led **Zielinski et al. (2009)** to hypothesize that the symbionts can protect their host cells against infection. Consistent with this hypothesis, growth inhibition assays showed that *B. azoricus* gill tissue homogenates inhibit the growth of a broader spectrum of pathogens compared to the symbiont-free foot tissues (**Bettencourt et al., 2007**). The mechanisms of protection against cultured bacterial pathogens and *Ca*. E. bathymodioli are unknown. There is no evidence from their genomes that the SOX symbionts of *B.* sp. and *B. azoricus* produce antibiotics, but they do have genes for bacteriocin production that were expressed in the six transcriptomes analyzed in this study. Expression of some of the TRGs discovered here, for example, those related to toxins involved in bacteria–bacteria competition, or the production of bacteriocins by symbionts could explain the absence of intranuclear bacteria from symbiont-hosting cells.

Are some of the TRGs used for protection against eukaryotic parasites?

Symbionts of other marine invertebrate hosts are able to recognize, enter, and survive within host cells without a large number of TRGs. For example, the genomes of the SOX symbionts of hydrothermal vent *Riftia* tubeworms and the heterotrophic symbionts of whale-fall *Osedax* worms are virtually complete, but they only contain one or a few RTX genes, and no YD repeat genes (**Robidart et al., 2008**; **Gardebrecht et al., 2012**; **Goffredi et al., 2014**). There is overwhelming evidence that SOX symbionts are beneficial for their *Bathymodiolus* mussel hosts (Appendix 1). It is therefore highly unlikely that the *Bathymodiolus* SOX symbionts are pathogens that have been mistaken for beneficial symbionts. The remarkably large number of TRGs in the *Bathymodiolus* SOX symbionts bears striking similarity to the arsenal of toxins encoded by *Candidatus* Hamiltonella defensa, which is a facultative symbiont of aphids (**Oliver et al., 2003**, **2005**). Both symbionts encode multiple copies of RTX and YD repeat proteins (**Degnan and Moran, 2008**). *Ca*. Hamiltonella defensa is a defensive symbiont that protects its host from attack by parasitic wasps (**Oliver et al., 2003**). Its protective effect is linked to its complement of RTX and YD repeat toxins (**Degnan and Moran, 2008**; **Oliver et al., 2010**). Based on our phylogeny of YD repeats, those from the *Bathymodiolus* symbionts cluster together with sequences from both pathogens and beneficial symbionts. Of the beneficial symbionts in this cluster, all except the *Bathymodiolus* symbionts have been shown to produce exotoxins that damage the organisms their host parasitizes, either a plant in the case of *B. rhizoxinica* or an insect in the case of *Photorhabdus/Xenorhabdus* (**Partida-Martinez et al., 2007**). This raises the intriguing possibility that some of the TRGs in the SOX symbiont genomes might function in protecting the mussel hosts against eukaryotic parasites.

Compared to our knowledge of parasitism in shallow-water bivalves, little is known about parasitism in deep-sea *Bathymodiolus* mussels. This is surprising considering that these incredibly dense communities would be ideal habitats for parasites (**Moreira and López-García, 2003**). Two studies have investigated parasitism in *Bathymodiolus* mussels based on the microscopic identification of unusual 'inclusions' in mussel tissues (**Powell et al., 1999**; **Ward et al., 2004**). The most abundant parasites resembled *Bucephalus*-like trematodes of the phylum Platyhelminthes, which are common in shallow-water mussels (e.g., **Wardle, 1988**; **Lauenstein et al., 1993**; **da Silva et al., 2002**; **Minguez et al., 2012**). *Bucephalus* trematodes infect the gonads of their mussel hosts, which often results in sterilization (**Hopkins, 1957**; **Coustau et al., 1991**). Like their shallow-water relatives, the *Bucephalus*-like trematodes were abundant in the gonads of *Bathymodiolus childressi* from cold seeps in the Gulf of Mexico (**Powell et al., 1999**). **Powell et al. (1999)** estimated that due to this heavy infection, up to 40% of *B. childressi* populations are reproductively compromised.

The distribution of these trematode parasites has not yet been systematically investigated in *Bathymodiolus*. However, of the three species so far studied, only *B. childressi* was infected by trematodes (**Powell et al., 1999**; **Ward et al., 2004**). *B. childressi* is one of the few *Bathymodiolus*

species that only associates with MOX symbionts, but not with SOX symbionts. If many of the TRGs encoded by the *Bathymodiolus* SOX symbiont are being used to defend its host against parasites, as is hypothesized for *Ca*. H. defensa, then this could help to explain why *B. childressi* is so heavily infected by trematodes. The MOX symbionts of *B. azoricus* and *B. childressi* do not encode the abundant TRGs of the *Bathymodiolus* SOX symbiont (Antony CP, personal communication, May 2015).

Our SNP analysis provides further support for the hypothesis that some TRGs may be used for direct beneficial interactions, and some may be used for indirect interactions such as protection against parasites. Genes involved in direct host-symbiont interactions such as recognition and communication are expected to be conserved within the symbiont population of one host species (***Jiggins et al., 2002***; ***Bailly et al., 2006***). Consistent with this, eight out of nine genes in the MARTX1 region, which we hypothesize may be involved in attachment to the host, do not have a significantly larger number of SNPs per kb compared to the rest of the genome (***Figure 6—source data 1***). In the ninth gene, annotated as a hypothetical protein, SNPs per kb were slightly above average. In contrast to genes involved in direct host-symbiont interactions, those involved in indirect interactions such as defense against parasites are expected to be highly diverse (see Appendix 2). The large sequence variability in 22 of the TRGs is therefore consistent with a role for these genes in protecting the host against parasites.

Conclusions

The genomes of the uncultured *Bathymodiolus* SOX symbionts encode a unique arsenal of TRGs, unexpected for a beneficial, nutritional symbiont. We hypothesize that the *Bathymodiolus* SOX symbiont has 'tamed' these genes for use in beneficial interactions with their host. Some of the TRGs may benefit the symbiosis by protecting the symbionts and their hosts from their natural enemies. In most cases, symbionts are either nutritional, that is, their primary role is to provide their host with most or all of its nutrition, or they are defensive (***Douglas, 2014***). The *Bathymodiolus* symbiont is therefore unusual, as it may play an essential role in both nutrition and defense. The TRGs were most likely acquired by HGT, and this may be the mechanism by which its free-living ancestors acquired the ability to form an intimate relationship with marine animals.

Remarkably, the *Bathymodiolus* SOX symbionts encode a larger complement of these TRGs than any so far sequenced pathogen, suggesting that these 'toxins', although initially discovered in pathogens, may in fact belong to larger protein families that function in both beneficial and pathogenic host-microbe interactions. An alternative to the hypothesis that toxins may be tamed for use in beneficial interactions would be that 'symbiosis factors' may be commandeered for use in harmful interactions. Given our recent recognition of the ubiquity and vast natural diversity of mutualistic interactions between bacteria and eukaryotes (***McFall-Ngai et al., 2013***), it is possible that many of the genes that are currently annotated as toxins may have first evolved through beneficial host-microbe associations.

Materials and methods

Sampling and processing of *Bathymodiolus* mussels

We collected *Bathymodiolus* mussels in Lilliput, Menez Gwen, and Lucky Strike vent sites on the MAR. *Bathymodiolus* sp. from Lilliput on the SMAR were sampled and processed for genome sequencing as in ***Petersen et al. (2011)***. For transcriptomics, we sampled mussels from the SMAR at 09°32.85′S, 13°12.64′W. Specimens of *B. azoricus* were collected in three cruise expeditions to Menez Gwen at (i) 37°45.5777′N, 31°38.2611′W during the MOMARETO cruise, (ii) 37°50.68′N, 31°31.17′W during the RV Meteor cruise M82-3, and (iii) Lucky Strike at 37°16′58.5′′N, 32°16′32.2′′W during the Biobaz Cruise. The adductor muscle was dissected from samples of the MOMARETO cruise, while the gill tissue was dissected from mussels collected during the RV Meteor cruise. For samples from the RV Meteor cruise, we used a combination of differential and rate-zonal centrifugation to enrich *Bathymodiolus* SOX symbiont from gill tissue for genomic and proteomic analyses. Samples for transcriptomics were fixed on board in RNAlater (Sigma, Germany) according to the manufacturer instructions and stored at −80°C. An overview of the samples used in this study is shown in ***Supplementary file 1E***.

DNA extraction, sequencing, genome assembly and binning

DNA extraction and genome sequencing of the *B.* sp. SMAR SOX symbiont was described in ***Petersen et al. (2011)***. Briefly, the gill tissue of a single individual was ground in a glass tissue homogenizer and frozen until further processing. In the home laboratory, the homogenate was diluted in phosphate-buffered saline (PBS)1×, centrifuged at 400×*g* for one minute, and the supernatant filtered through a 12-µm GTTP filter (Millipore, Germany). Centrifugation and filtration was repeated 20 times. The filtrates were passed through GTTP filters of 8 µm, 5 µm, 3 µm, and 2 µm. Cells collected on the 0.2-µm filter were used for DNA extraction after ***Zhou et al. (1996)*** with an initial incubation overnight at 37˚C in extraction buffer and proteinase K. 6 kb mate-paired reads were sequenced with 454-Titanium and 36 bp Illumina reads. 454 reads were assembled with Newbler v2.3 (454 Life Sciences Corporation) and 569 pyrosequencing errors were corrected using the Illumina reads.

DNA from the *B. azoricus* SOX symbiont enrichments from three individuals (gradient pellets, see Appendix 4) was extracted according to ***Zhou et al. (1996)***. Genomic DNA was extracted from adductor muscle using a CTAB/PVP extraction procedure (2% CTAB, 1% PVP, 1.4 M NaCl, 0.2% beta-mercaptoethanol, 100 mM Tris HCl pH 8, 0.1 mg ml^{-1} proteinase K). After complete digestion of tissues (1 hr at 60˚C), the mixture was incubated with 1 µl of RNase for 30 min at 37˚C. An equal volume of chloroform-isoamyl alcohol (24:1) was added and tubes were slowly mixed by inversion for 3 min before a 10 min centrifugation at 14,000 rpm and 4˚C. The supernatant was collected in a fresh tube, and DNA was precipitated with 2/3 volume of cold isopropanol (1 hr at −20˚C). The DNA pellet was recovered by centrifugation (14,000 rpm at 4˚C for 20 min), washed with 75% cold ethanol, air-dried, and suspended in 100 µl of sterile water. 454 sequencing was done by Genoscope to sequence the gradient pellet from gill tissue, and by OIST to sequence the adductor muscle of *B. azoricus*. For the gradient pellet, a 3 kb insert 454 library was prepared according to manufacturer protocols for mate-pair sequencing. 630752 reads were generated on a Titanium FLX sequencing machine and assembled using Newbler software (version 08172012). 1310 contigs larger than 500 bp were obtained, forming 130 scaffolds of a total length of 1668565 bp. The assembly from the adductor muscle was done with Newbler v. 2.7 (454 Life Sciences Corporation) as described in ***Takeuchi et al. (2012)*** resulting in 644000 contigs of a total length of 510449434 bp. The adductor muscle and gradient pellet metagenomes of *B. azoricus* were binned to separate the SOX symbiont from the MOX symbiont and host genomes with Metawatt V. 1.7, which uses tetranucleotide frequencies, coverage, GC content, and taxonomic information for binning (***Strous et al., 2012***). Only sequences longer than 800 bp were considered for further analyses. Since we could only recover an 829 bp fragment of the 16S rRNA from BazSymB, the same DNA that was used for 454 sequencing was used as template for PCR amplification with the universal primers GM3f/GM4r (***Muyzer et al., 1995***). The PCR product was directly sequenced with Sanger and assembled using Geneious V7 (***Kearse et al., 2012***). The 16S rRNA of BazSymB can be found under the accession number (LN871183).

Genome annotation

We annotated the genomes of the *Bathymodiolus* symbionts (BspSym: PRJNA65421, BazSymA: PRJEB8263, and BazSymB: PRJEB8264), *Candidatus* V. okutanii (NC_009465), *Candidatus* R. magnifica (CP000488), and SUP05 metagenome (ACSG01000000; GQ351266 to GQ351269 and GQ369726) with the following workflow: we used Glimmer (***Delcher et al., 2007***) for open reading frame (ORF) prediction. Ribosomal RNA genes were detected with RNAmmer (***Lagesen et al., 2007***) and tRNAs with tRNAscan-SE (***Lowe and Eddy, 1997***). Annotation was done with GenDB 2.4 (***Meyer, 2003***) and supplemented by JCoast 1.7 (***Richter et al., 2008***) to integrate the results of BLASTp (cut-off e-value of 10.0) against sequence databases NCBI-nr (***Altschul et al., 1997***) SwissProt (***Boeckmann, 2003***), KEGG (***Kanehisa et al., 2011***), COG (***Tatusov, 2000***), Pfam (***Bateman et al., 2004***), and InterPro (***Hunter et al., 2009***). TMHMM (***Krogh et al., 2001***) was used for transmembrane helix analysis and SignalP (***Emanuelsson et al., 2007***) for signal peptide predictions. Sequences of the cytochrome c oxidase subunit 1 of the *Bathymodiolus* mussels were submitted to the European Nucleotide Archive when available (***Supplementary file 1E***).

PCR amplification of regions with lack of synteny

We designed primers to amplify four regions covering region with lack of synteny. The primer sequences and annealing temperatures are listed in ***Supplementary file 1F***. The PCR program

consisted of an initial denaturation step of 98˚C for 30 s, followed by 35 cycles at 98˚C for 10 s, specific annealing temperature for 30 s, 72˚C for 2 min, and a final extension at 72˚C for 10 min. We obtained a PCR product of the expected size based on our assembly of all four targeted regions. We sequenced one of these by Sanger sequencing using ABI BigDye v3.1 and the ABI PRISM 3100 genetic analyzer (Applied Biosystems, Foster City, CA).

Genome analysis and comparison to close relatives

We used CheckM to evaluate the completeness of our draft genomes with a set of single-copy marker genes that are specific to proteobacteria (lineage-specific marker set of CheckM p_Proteobacteria, UID3880) (***Parks et al., 2015***). We estimated the similarity of BazSymB and BazSymA draft genomes with the mean and standard deviation of genes with bi-directional best BLAST hits. The initial comparison of gene content between the clam symbionts, mussel symbionts draft genomes, and SUP05 metagenome was done by BLAST score ratio (BSR) with a BSR cut-off of 0.4 (***Rasko et al., 2005***). Since toxin genes are expected to have a higher mutation rate, we compared the toxin distribution among the three mussel symbiont draft genomes and their closest relatives with a BSR cut-off of 0.2.

Because *Ca*. R. magnifica is the largest vesicomyid genome available, a whole genome comparison with SUP05 was done using a Dotplot produced by Ugene (http://ugene.unipro.ru) with a minimum length of 50 bp and 90% similarity. Since a high gene synteny could be observed, *Ca*. R. magnifica was used as scaffold to reorder the contigs from SUP05 and the *B.* sp. symbiont using Mauve (***Rissman et al., 2009***). We did not do this analysis for the *B. azoricus* symbiont genomes because they were highly fragmented. To compare the influence of HGT on genome evolution, we used the method of ***Davis and Olsen (2010)*** to identify genes with different codon usage patterns, indicating that they may have been recently acquired via HGT.

Identification, classification, and structural analysis of TRGs

The initial genome annotation contained many genes annotated as toxins (see Appendix 4 for details of the automatic annotation). Most were related to toxins from three broad previously defined classes, the YD repeats, RTX, and MARTX toxins. Most of the YD repeat and RTX genes could be identified accurately by automatic annotation due to the presence of signature repeat regions. Some of the predicted YD genes appeared to be truncated. Alignment with the closest hits was used to extend the partial ORFs that were not correctly predicted. To curate the annotation of MARTX genes, we built Hidden Markov Model (HMM) profiles for the characterized MARTX domains in *Vibrio cholerae:* actin cross-linking domain, Rho GTPase inactivation, and cysteine protease domain (***Satchell, 2011***). The profiles were used to scan the SOX symbiont genomes with HMMER (***Finn et al., 2011***). We searched for functional domains in the proteins identified as TRGs with SMART (***Ponting et al., 1999***). To analyze the diversity and redundancy of the TRGs in the SOX genomes, we constructed a protein similarity network. All protein sequences of the SOX symbionts were searched against Uniprot with BLAST (coverage >50%, similarity >25%, e-value < e−5). We also used BLAST of all against all sequences of the *Bathymodiolus* SOX symbiont to recover partial TRGs that could not be identified by automatic annotation. The two searches were combined to produce a sequence similarity network based on transitivity clustering using the plugin Blast2SimilarityGraph in Cytoscape (***Srinivasan and Moon, 1999***; ***Shannon et al., 2003***; ***Wittkop et al., 2010***). Only the sub-networks that were connected in a maximum of two steps to the TRGs were considered. All TRGs of the SOX symbionts were submitted to Phyre2 (***Kelley and Sternberg, 2009***) to predict the secondary structure of the protein. We looked for clusters in the protein similarity network that could be associated with a subunit of a toxin complex or an active domain.

Search for TRGs in SUP05

None of the three toxin-related classes has characteristic patterns, profiles, or domains that can be searched with standard tools. Often, their only shared feature is a repeat region, such as the RTX repeat in RTX and MARTX, which is a calcium-binding site containing G and N residues. However, the number of G and N residues and the length of the repeat is highly variable, which makes profile searches impossible. Moreover, these repeat regions can be shared by other calcium-binding proteins such as integrins and fibronectin, which do not act as toxins. The sequence and length of the YD

repeat is similarly variable. To identify homologs of the *Bathymodiolus* symbiont TRGs in published metagenomes and metatranscriptomes, we therefore used BLASTp (coverage >50%, similarity >25%, e-value < 0.001) (***Supplementary file 1G***).

Best hits were blasted against the SOX symbiont proteins to search for signatures of the TRGs. We looked for bacterial genomes that had similar TRGs to the *Bathymodiolus* SOX symbiont, using the Integrated Microbial Genomes (IMG) database (***Markowitz et al., 2011***) with BLASTp (>50% coverage, > 25% similarity, e-value 0.001). 122 genomes were retrieved and manually curated for the potential of pathogenicity and biofilm formation, as well as the lifestyle categories extracellular host-associated, intracellular, and free-living bacteria. Kruskall–Wallis and Mann–Whitney–Wilcoxon tests were used to estimate if the increased number of a class of TR is biased in a certain lifestyle. All statistical analyses were done in R.

Phylogenetic analyses

The 16S rRNA gene sequences of the *Bathymodiolus* symbionts were imported into the Silva database (release Ref 119) and initially aligned with the SINA aligner (***Pruesse et al., 2012***). The final alignment was refined with MAFFT (***Katoh et al., 2002***). A maximum likelihood tree was estimated from the alignment of 1653 nucleotide positions using RaxML with 100 bootstrap replicates.

To construct the YD phylogeny, the data set from BspSym was used to obtain related sequences from GenBank and *B. azoricus*. Because two metagenomes of *B. azoricus* were analyzed, we used CD-hit to remove redundancy at 100% similarity (***Li and Godzik, 2006***). YD repeat proteins are very variable in the C and N terminus. Therefore, the selection criterion we used was the presence of the conserved *rhs* domain. Only this 'core' region was used for phylogeny (***Jackson et al., 2009***). YD proteins were aligned with MAFFT with BLOSUM30 (***Katoh et al., 2002***). Phylogenetic analyses were done in ARB with maximum likelihood and bayesian reconstructions using a filter of 10% similarity, which resulted in 536 amino acid positions (***Ronquist and Huelsenbeck, 2003***; ***Ludwig et al., 2004***; ***Stamatakis, 2006***). Bootstrap support was calculated with 100 replicates in maximum likelihood. MrBayes was run for 9 million generations and two independent runs of four heated chains. A consensus tree of both methods was constructed. Polytomies were introduced when both methods did not agree.

Statistical analyses of TRGs content in microbial genomes

Genomes with curated metadata are available through IMG (***Markowitz et al., 2011***). Nevertheless, most entries contained no information in categories we were interested in such as biofilm formation and intracellular lifestyle. IMG was used as starting point to search for curated genomes that had genes with similarity to the TRGs of the mussel symbionts (>10% similarity, e-value < e^{-5}). The protein sequences of these genomes were retrieved from NCBI based on the genome name (modified Python script from Sixing Huang, Max Planck Institute Bremen). BLASTp was used to retrieve protein sequences related to the SOX TRGs (>50% coverage, > 25% similarity, e-value of 0.001). 122 genomes had at least one gene that was similar to at least one TRG. We considered the genomes of different strains as independent events for statistical analysis, as even closely related strains can have different lifestyles. These genomes were manually curated based on a literature search and classified according to the following categories: (1) lifestyle: divided into three sub-categories (a) associated: if the organism is at any stage host-associated (based on IMG metadata), (b) intracellular: includes obligate and facultative symbionts (searched in Google scholar with the keywords 'intracellular bacteria' and references read for more details when the abstract was not sufficient), (c) free-living: bacteria that are not host-associated and not intracellular; (2) pathogen: bacteria that can produce disease (information obtained from IMG); (3) biofilm: bacteria found in biofilms (Google scholar keywords 'biofilm' and 'microbial mat' with literature analyses when not clear).

The data were formatted and merged with self-written Perl scripts and ***R Development Core Team (2011)***. Genomes with similar toxins or TRGs to the *Bathymodiolus* symbionts are shown in ***Supplementary file 1B***. The sum of the number of TRGs belonging to YD, RTX, and MARTX was normalized with the total gene count for each genome and multiplied by a factor of 1000. To compare the number of TRGs against the lifestyle categories, we used Kruskall–Wallis test. A post hoc analysis was carried out on significant p-values for the associated category with the Mann–Whitney–Wilcoxon test. Statistical analyses were done in R. We tested whether the bacteria were enriched in any of the

three toxin-related classes at the class, order, or family by using one-way Permanova with 9999 permutations and Euclidean distances. p-values were corrected with Bonferroni correction for multiple testing.

Transcriptomics

To extract the total RNA of three individuals of *B. azoricus*, the gill tissue was incubated overnight in RNAlater (Sigma) at 4˚C. A fragment of the gill was dissected and homogenized. RNA was extracted with RNeasy Plus MicroKit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. To remove cell debris and to improve RNA yield, we used QIAshreder Mini Spin Columns (Qiagen). The quality of the RNA was assessed with Agilent 2100 Bioanalyzer. The RNA was used for cDNA synthesis with the Ovation RNA-Seq System V2 (NuGEN, San Carlos, CA). To extract the total RNA of three individuals of *B.* sp., the gill tissue was placed separately on liquid nitrogen, homogenized, and stored overnight at 4˚C in self-made RNAlater (10 mM ethylenediaminetetraacetic acid (EDTA), 25 mM tri-sodiumcitrate-dihydrate, 5.3 M ammonium sulfate, adjusted to pH 5.2). After removal of RNAlater, samples were incubated in 600 µl RLT-β-mercaptoethanol buffer (1:100) for 10 min and homogenized on QIAshredder columns (Qiagen). Total RNA was extracted with RNeasy Mini Kit (Qiagen). We applied the DNA-free DNAse Treatment and Removal Kit according to the manufacturer's instructions (Invitrogen, Carlsbad, CA/Ambion, Austin, TX). RNA quality was checked on the Experion Automated Electrophoresis Station using the RNA StdSens Analysis protocol (BioRad, Hercules, CA).

Libraries of *B.* spp. were generated with the Illumina TruSeq RNA Sample Preparation Kit and sequenced 2 × 100 paired-end on an Illumina HiSeq 2000 platform at the Institute of Clinical Molecular Biology (Kiel). A total of 32.9, 38.2, and 38.4 million reads were sequenced per individual of *B. azoricus*. Libraries of *B. azoricus* were generated with DNA library prep kit for Illumina (BioLABS, Frankfurt am Main, Germany) and sequenced as single 100-bp reads on an Illumina HiSeq 2500 platform at the Max Planck Genome Centre (Cologne). A total of 4.3, 4.8, and 6.9 million reads were sequenced per individual of *B. azoricus*. Adaptor removal and quality trimming was done with Nesoni (http://www.vicbioinformatics.com/software.nesoni.shtml) using a quality threshold of 20. To remove ribosomal sequences from the data, we mapped the reads against the SILVA 115 SSU database with Bowtie2 and kept those reads that failed to align (***Langmead and Salzberg, 2012***; ***Quast et al., 2013***). The abundance of the transcripts per gene was estimated with Rockhopper that uses upper quartile normalization (***McClure et al., 2013***). The expression values of the TRGs were normalized to the expression of RubisCO (***Supplementary file 1D***). Transcriptome reads that mapped to the SOX symbionts with Bbmap (http://bbmap.sourceforge.net/) were deposited in the European Nucleotide Archive under the accession numbers PRJEB7941 for *B. azoricus* and PRJEB7943 for *B.* sp.

SNP analysis

We calculated SNPs to compare substitution rates of genes in the SOX symbiont genomes using three transcriptomes of *B.* sp and three of *B. azoricus* (***Supplementary file 1E***). Reads were normalized to a coverage of maximum 200 and minimum of five with BBNorm v33 (http://sourceforge.net/projects/bbmap/). Transcriptomes of *Bathymodiolus* sp. were mapped against the draft genome BspSym and transcriptomes of *B. azoricus* were mapped to the draft genomes of BazSymA and BazSymB with BBMap v33. We only considered those reads that mapped to a single position in the reference genome and that had higher than 90% identity alignments. SNPs were called independently for the draft genomes BspSym, BazSymA, and BazSymB with the Genome Analysis ToolKit as described by ***De Wit et al. (2012)*** with some modifications (***McKenna et al., 2010***). In summary, regions needing realignment were identified and realigned over intervals. SNPs and insertions or deletions (InDels) were called with the haplotype caller with a minimum confidence of 20. A filter was applied around InDels with a mask extension of 5. SNPs per gene were obtained with BEDTools (***Quinlan and Hall, 2010***). The number of SNPs per gene was normalized according to the gene length minus regions of unknown sequence for genes containing Ns. We did not consider genes shorter than 150 bp nucleotides or outlier genes at the ends of scaffolds that had an unusual number of SNPs.

Proteomics

Soluble proteins were extracted from SOX symbiont enrichments, host cytosolic fractions, and whole gill and foot tissue in biological duplicates (Appendix 4). All proteome samples were obtained from the M82-3 cruise. Membrane proteins were extracted from the SOX symbiont enrichments and the

whole gill tissue samples. We used 1D-PAGE followed by liquid chromatography (1D-PAGE-LC) to separate proteins and peptides as described previously with minor modifications (**Heinz et al., 2012**). MS/MS spectra were acquired with a LTQ Orbitrap Velos mass spectrometer (Thermo Fisher, Bremen, Germany) for soluble proteins and a LTQ Orbitrap Classic (Thermo Fisher Scientific Inc., Waltham, MA) for membrane proteins (Appendix 4). MS/MS data were searched against two databases using the SEQUEST algorithm (**Eng et al., 1994**). The first database, designated 'reduced' database, contained protein sequences from the SOX and MOX symbionts from *Bathymodiolus*, as well as from the host (see Appendix 4). The second database contained in addition sequences from host-related bivalves and symbiont-related bacteria (**Supplementary file 1H**). False discovery rates were determined with the target-decoy search strategy as described by Elias and Gygi (**Elias and Gygi, 2007**; **Kleiner et al., 2012a**).

Acknowledgements

We thank the captains and crews of the research vessels and ROVs involved in the sampling effort, including Francois Lallier and Arnaud Tanguy who provided samples for sequencing from the MoMARETO cruise (**Sarrazin et al., 2006**) with the N/O *Pourquoi Pas?* and ROV *Victor 6000*. This project was funded by the Max Planck Society, an ERC Advanced Grant and a Gordon and Betty Moore Foundation Marine Microbial Initiative Investigator Award to ND, the DFG Cluster of Excellence 'The Ocean in the Earth System' at MARUM, Bremen, a DAAD scholarship to LS, OIST funding to NS, and the Marie Curie Initial Training Network 'Symbiomics' (project no. 264774) for the proteomics. We thank Harald Gruber-Vodicka, Hanno Teeling, and Brandon Seah for useful discussions, Michael Richter for his help with GenDB, and Sixing Huang for providing a Python script. We thank the three reviewers whose excellent suggestions helped us to improve the manuscript.

Additional information

Funding

Funder	Grant reference	Author
Max-Planck-Gesellschaft (MPG)		Lizbeth Sayavedra, Manuel Kleiner, Silke Wetzel, Dennis Fink, Nicole Dubilier, Jillian M Petersen
Gordon and Betty Moore Foundation	MMI Investigator Grant 3811	Nicole Dubilier
European Research Council (ERC)	Bathybiome	Lizbeth Sayavedra, Manuel Kleiner, Silke Wetzel, Dennis Fink, Nicole Dubilier, Jillian M Petersen
Deutsche Forschungsgemeinschaft (DFG)		Lizbeth Sayavedra, Manuel Kleiner, Silke Wetzel, Nicole Dubilier, Jillian M Petersen
German Academic Exchange Service (DAAD)		Lizbeth Sayavedra
European Research Council (ERC)	Symbiomics ITN 264774	Ruby Ponnudurai, Thomas Schweder, Stephanie Markert
Okinawa Institute of Science and Technology Graduate University (OIST)		Nori Satoh

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

LS, Wrote the paper; analyzed genomic and transcriptomic data, Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; MK, RP, Performed proteomic analyses and processed mass spectrometry data, Drafting or revising the article; SW, Developed symbiont extraction method, Drafting or revising the article; EP, VB, Sequenced and assembled samples sent to Genoscope, Drafting or revising the article; NS, ES, Sequenced and

assembled samples in OIST, Drafting or revising the article; DF, DF performed gradient centrifugation during the M82-3 cruise, Drafting or revising the article; CB, TBHR, PR, MBS, Sequenced transcriptomes from B. sp., Drafting or revising the article; DB, TS, SM, Performed proteomic analyses and processed mass spectrometry data, Drafting or revising the article; ND, Conception and design, Drafting or revising the article; JMP, Wrote the paper, Conception and design, Drafting or revising the article

Additional files

Supplementary file

• Supplementary file 1. (**A**) Number of mobile elements in the genomes compared in this study. (**B**) Genomes with toxin-related genes (TRGs) similar to those of the sulfur-oxidizing (SOX) symbionts of *Bathymodiolus*. The number of genes per TRGs class is shown. (**C**) p-values obtained with one-way Permanova were corrected with Bonferroni correction for multiple testing. Number of TRGs per genome was normalized to the total gene count. (**D**) Transcriptome counts of three individuals from *B. azoricus* and three individuals from *B.* sp. were mapped to their respective reference genomes with Rockhopper. Expression values of TRGs were normalized to the expression of RubisCO. (**E**) Samples used in this study. (**F**) Primer sequences and annealing temperatures used to detect genome rearrangements. (**G**) Metagenomes and metatranscriptomes enriched in SUP05 from oxygen minimum zones (OMZ) or hydrothermal vents. (**H**) Amino acid sequences from the following genomes were used in the reference database for proteomic analysis (IncDB). The genomes belong to relatives of the SOX and methane-oxidizing (MOX) symbionts of *B. azoricus*, as well as the mussel host. (**I**) Details of expressed toxin-related proteins identified with proteomics. The values are given in % normalized spectral abundance factor (NSAF), which is a normalized spectral abundance factor that gives the relative abundance of a protein in a sample in %.

DOI: 10.7554/eLife.07966.023

Major datasets

The following datasets were generated:

Author(s)	Year	Dataset title	Dataset ID and/or URL	Database, license, and accessibility information
Sayavedra L, Kleiner M, Ponnudurai R, Wetzel S, Pelletier E, Barbe V, Satoh N, Shoguchi E, Fink D, Breusing C, Reusch TBH, Rosenstiel P, Schilhabel M, Becher D, Schweder T, Markert S, Dubilier N, Petersen JM	2015	Endosymbiont of Bathymodiolus sp.	http://www.ebi.ac.uk/ena/data/view/PRJNA65421	Publicly available at the EBI European Nucleotide Archive (Accession no: PRJNA65421).
Sayavedra L, Kleiner M, Ponnudurai R, Wetzel S, Pelletier E, Barbe V, Satoh N, Shoguchi E, Fink D, Breusing C, Reusch TBH, Rosenstiel P, Schilhabel M, Becher D, Schweder T, Markert S, Dubilier N, Petersen JM	2015	Genome of Sulfur-oxidizer endosymbiont of Bathymodiolus azoricus from Menez Gwen (BazSymA)	http://www.ebi.ac.uk/ena/data/view/PRJEB8263	Publicly available at the EBI European Nucleotide Archive (Accession no: PRJEB8263).
Sayavedra L, Kleiner M, Ponnudurai R, Wetzel S, Pelletier E, Barbe V, Satoh N, Shoguchi E, Fink D, Breusing C, Reusch TBH, Rosenstiel P, Schilhabel M, Becher D, Schweder T, Markert S, Dubilier N, Petersen JM	2015	Genome of Sulfur-oxidizer endosymbiont of Bathymodiolus azoricus from Menez Gwen (BazSymB)	http://www.ebi.ac.uk/ena/data/view/PRJEB8264	Publicly available at the EBI European Nucleotide Archive (Accession no: PRJEB8264).

Author(s)	Year	Dataset title	Dataset ID and/or URL	Database, license, and accessibility information
Sayavedra L, Kleiner M, Ponnudurai R, Wetzel S, Pelletier E, Barbe V, Satoh N, Shoguchi E, Fink D, Breusing C, Reusch TBH, Rosenstiel P, Schilhabel M, Becher D, Schweder T, Markert S, Dubilier N, Petersen JM	2015	Mitochondrial COI gene for cytochrome oxidase subunit 1	http://www.ebi.ac.uk/ena/data/view/LN833433-LN833440	Publicly available at the EBI European Nucleotide Archive (Accession no: LN833433-LN833440).
Sayavedra L, Kleiner M, Ponnudurai R, Wetzel S, Pelletier E, Barbe V, Satoh N, Shoguchi E, Fink D, Breusing C, Reusch TBH, Rosenstiel P, Schilhabel M, Becher D, Schweder T, Markert S, Dubilier N, Petersen JM	2015	Transcriptome of Sulfur-oxidizer endosymbiont of Bathymodiolus azoricus	http://www.ebi.ac.uk/ena/data/view/PRJEB7941	Publicly available at the EBI European Nucleotide Archive (Accession no: PRJEB7941).
Sayavedra L, Kleiner M, Ponnudurai R, Wetzel S, Pelletier E, Barbe V, Satoh N, Shoguchi E, Fink D, Breusing C, Reusch TBH, Rosenstiel P, Schilhabel M, Becher D, Schweder T, Markert S, Dubilier N, Petersen JM	2015	Transcriptome of Sulfur-oxidizer endosymbiont of Bathymodiolus sp. 9 South	http://www.ebi.ac.uk/ena/data/view/PRJEB7943	Publicly available at the EBI European Nucleotide Archive (Accession no: PRJEB7943).
Sayavedra L, Kleiner M, Ponnudurai R, Wetzel S, Pelletier E, Barbe V, Satoh N, Shoguchi E, Fink D, Breusing C, Reusch TBH, Rosenstiel P, Schilhabel M, Becher D, Schweder T, Markert S, Dubilier N, Petersen JM	2015	Bathymodiolus azoricus thioautotrophic gill symbiont partial 16S rRNA gene, isolate BazSymB	http://www.ebi.ac.uk/ena/data/view/LN871183	Publicly available at the EBI European Nucleotide Archive (Accession no: LN871183).

References

Achaz G, Coissac E, Netter P, Rocha EPC. 2003. Associations between inverted repeats and the structural evolution of bacterial genomes. *Genetics* **164**:1279–1289.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**:3389–3402. doi: 10.1093/nar/25.17.3389.

Anantharaman K, Breier JA, Sheik CS, Dick GJ. 2012. Evidence for hydrogen oxidation and metabolic plasticity in widespread deep-sea sulfur-oxidizing bacteria. *Proceedings of the National Academy of Sciences of USA* **110**:330–335. doi: 10.1073/pnas.1215340110.

Backert S, Meyer TF. 2006. Type IV secretion systems and their effectors in bacterial pathogenesis. *Current Opinion in Microbiology* **9**:207–217. doi: 10.1016/j.mib.2006.02.008.

Badger MR, Bek EJ. 2008. Multiple RuBisCO forms in proteobacteria: their functional significance in relation to CO_2 acquisition by the CBB cycle. *Journal of Experimental Botany* **59**:1525–1541. doi: 10.1093/jxb/erm297.

Bailly X, Olivieri I, De Mita S, Cleyet-Marel J-C, Béna G. 2006. Recombination and selection shape the molecular diversity pattern of nitrogen-fixing *Sinorhizobium* sp. associated to *Medicago*. *Molecular Ecology* **15**:2719–2734. doi: 10.1111/j.1365-294X.2006.02969.x.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. 2004. The Pfam protein families database. *Nucleic Acids Research* **32**:D138–D141. doi: 10.1093/nar/gkh121.

Benz J, Meinhart A. 2014. Antibacterial effector/immunity systems: it's just the tip of the iceberg. *Current Opinion in Microbiology* **17**:1–10. doi: 10.1016/j.mib.2013.11.002.

Bettencourt R, Roch P, Stefanni S, Rosa D, Colaço A, Serrão Santos R. 2007. Deep sea immunity: unveiling immune constituents from the hydrothermal vent mussel *Bathymodiolus azoricus*. *Marine Environmental Research* **64**:108–127. doi: 10.1016/j.marenvres.2006.12.010.

Bhavsar AP, Guttman JA, Finlay BB. 2007. Manipulation of host-cell pathways by bacterial pathogens. *Nature* **449**:827–834. doi: 10.1038/nature06247.

Boeckmann B. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* **31**:365–370. doi: 10.1093/nar/gkg095.

Bradford MM. 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry* **72**:248–254.

Brinkhoff T, Giebel H-A, Simon M. 2008. Diversity, ecology, and genomics of the *Roseobacter* clade: a short overview. *Archives of Microbiology* **189**:531–539. doi: 10.1007/s00203-008-0353-y.

Buchan A, González JM, Moran MA. 2005. Overview of the marine *roseobacter* lineage. *Applied and Environment Microbiology* **71**:5665–5677. doi: 10.1128/AEM.71.10.5665-5677.2005.

Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Subhraveti P, Weaver DS, Weerasinghe D, Zhang P, Karp PD. 2014. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research* **42**:D459–D471. doi: 10.1093/nar/gkt1103.

Cavanaugh CM, Robinson JJ. 1996. CO_2 fixation in chemoautotroph-invertebrate symbioses: expression of form I and form II RuBisCO. In: Lidstrom ME, Tabita FR, editors. *Microbial growth on C1 compounds:* Springer Netherlands. p. 285–292.

Cavanaugh CM, McKiness ZP, Newton ILG, Stewart FJ. 2006. Marine chemosynthetic symbioses. In: Dworkin M, Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E, editors. *The Prokaryotes:* Springer New York. p. 475–507.

Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak MY, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P. 2012. A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology* **30**:918–920. doi: 10.1038/nbt.2377.

Christie-Oleza JA, Piña-Villalonga JM, Bosch R, Nogales B, Armengaud J. 2012. Comparative proteogenomics of twelve *Roseobacter* exoproteomes reveals different adaptive strategies among these marine bacteria. *Molecular & Cellular Proteomics* **11**:M111.013110. doi: 10.1074/mcp.M111.013110.

Clayton AL, Oakeson KF, Gutin M, Pontes A, Dunn DM, von Niederhausern AC, Weiss RB, Fisher M, Dale C. 2012. A novel human-infection-derived bacterium provides insights into the evolutionary origins of mutualistic insect–bacterial symbioses. *PLOS Genetics* **8**:e1002990. doi: 10.1371/journal.pgen.1002990.

Cossart P, Sansonetti PJ. 2004. Bacterial invasion: the paradigms of enteroinvasive pathogens. *Science* **304**:242–248. doi: 10.1126/science.1090124.

Coustau C, Renaud F, Maillard C, Pasteur N, Delay B. 1991. Differential susceptibility to a trematode parasite among genotypes of the *Mytilus edulis/galloprovincialis* complex. *Genetics Research* **57**:207–212. doi: 10.1017/S0016672300029359.

da Silva PM, Magalhães AR, Barracco MA. 2002. Effects of *Bucephalus* sp. (Trematoda: Bucephalidae) on *Perna perna* mussels from a culture station in Ratones Grande Island, Brazil. *Journal of Invertebrate Pathology* **79**:154–162. doi: 10.1016/S0022-2011(02)00026-5.

Dale C, Moran NA. 2006. Molecular interactions between bacterial symbionts and their hosts. *Cell* **126**:453–465. doi: 10.1016/j.cell.2006.07.014.

Dale C, Young SA, Haydon DT, Welburn SC. 2001. The insect endosymbiont *Sodalis glossinidius* utilizes a type III secretion system for cell invasion. *Proceedings of the National Academy of Sciences of USA* **98**:1883–1888. doi: 10.1073/pnas.021450998.

Davies RL, Campbell S, Whittam TS. 2002. Mosaic structure and molecular evolution of the leukotoxin operon (lktCABD) in Mannheimia (Pasteurella) haemolytica, *Mannheimia glucosida*, and *Pasteurella trehalosi*. *Journal of Bacteriology* **184**:266–277. doi: 10.1128/JB.184.1.266-277.2002.

Davis JJ, Olsen GJ. 2010. Characterizing the native codon usages of a genome: an axis projection approach. *Molecular Biology and Evolution* **28**:211–221. doi: 10.1093/molbev/msq185.

De Wit P, Pespeni MH, Ladner JT, Barshis DJ, Seneca F, Jaris H, Therkildsen NO, Morikawa M, Palumbi SR. 2012. The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources* **12**:1058–1067. doi: 10.1111/1755-0998.12003.

DeChaine EG, Bates AE, Shank TM, Cavanaugh CM. 2006. Off-axis symbiosis found: characterization and biogeography of bacterial symbionts of *Bathymodiolus* mussels from Lost City hydrothermal vents. *Environmental Microbiology* **8**:1902–1912. doi: 10.1111/j.1462-2920.2005.01113.x.

Degnan PH, Moran NA. 2008. Diverse phage-encoded toxins in a protective insect endosymbiont. *Applied and Environmental Microbiology* **74**:6782–6791. doi: 10.1128/AEM.01285-08.

Delcher AL, Bratke KA, Powers EC, Salzberg SL. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**:673–679. doi: 10.1093/bioinformatics/btm009.

Distel DL, Lee HK, Cavanaugh CM. 1995. Intracellular coexistence of methano- and thioautotrophic bacteria in a hydrothermal vent mussel. *Proceedings of the National Academy of Sciences of USA* **92**:9598–9602. doi: 10.1073/pnas.92.21.9598.

Dmytrenko O, Russell SL, Loo WT, Fontanez KM, Liao L, Roeselers G, Sharma R, Stewart FJ, Newton IL, Woyke T, Wu D, Lang JM, Eisen JA, Cavanaugh CM. 2014. The genome of the intracellular bacterium of the

coastal bivalve, *Solemya velum*: a blueprint for thriving in and out of symbiosis. *BMC Genomics* **15**:924. doi: 10.1186/1471-2164-15-924.

Dobrindt U, Hochhut B, Hentschel U, Hacker J. 2004. Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews Microbiology* **2**:414–424. doi: 10.1038/nrmicro884.

Douglas AE. 2014. *The symbiotic habit*.

Dubilier N, Bergin C, Lott C. 2008. Symbiotic diversity in marine animals: the art of harnessing chemosynthesis. *Nature Reviews Microbiology* **6**:725–740. doi: 10.1038/nrmicro1992.

Duperron S, Bergin C, Zielinski F, Blazejak A, Pernthaler A, McKiness ZP, DeChaine E, Cavanaugh CM, Dubilier N. 2006. A dual symbiosis shared by two mussel species, *Bathymodiolus azoricus* and *Bathymodiolus puteoserpentis* (Bivalvia: Mytilidae), from hydrothermal vents along the northern Mid-Atlantic Ridge. *Environmental Microbiology* **8**:1441–1447.

Duperron S, Halary S, Lorion J, Sibuet M, Gaill F. 2008. Unexpected co-occurrence of six bacterial symbionts in the gills of the cold seep mussel *Idas* sp. (Bivalvia: Mytilidae). *Environmental Microbiology* **10**:433–445. doi: 10.1111/j.1462-2920.2007.01465.x.

Duperron S, Gaudron SM, Rodrigues CF, Cunha MR, Decker C, Olu K. 2013. An overview of chemosynthetic symbioses in bivalves from the North Atlantic and Mediterranean sea. *Biogeosciences* **10**:3241–3267. doi: 10.5194/bg-10-3241-2013.

Elias JE, Gygi SP. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4**:207–214. doi: 10.1038/nmeth1019.

Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols* **2**:953–971. doi: 10.1038/nprot.2007.131.

Eng JK, McCormack AL, Yates JR III. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **5**:976–989. doi: 10.1016/1044-0305(94)80016-2.

Eymann C, Dreisbach A, Albrecht D, Bernhardt J, Becher D, Gentner S, Tam le T, Büttner K, Buurman G, Scharf C, Venz S, Völker U, Hecker M. 2004. A comprehensive proteome map of growing *Bacillus subtilis* cells. *Proteomics* **4**:2849–2876. doi: 10.1002/pmic.200400907.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**:W29. doi: 10.1093/nar/gkr367.

Fisher CR, Brooks JM, Vodenichar JS, Zande JM, Childress JJ, Burke RA Jr. 1993. The co-occurrence of methanotrophic and chemoautotrophic sulfur-oxidizing bacterial symbionts in a deep-sea mussel. *Marine Ecology* **14**:277–289. doi: 10.1111/j.1439-0485.1993.tb00001.x.

Florens L, Carozza MJ, Swanson SK, Fournier M, Coleman MK, Workman JL, Washburn MP. 2006. Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* **40**:303–311. doi: 10.1016/j.ymeth.2006.07.028.

Fontanez KM, Cavanaugh CM. 2014. Evidence for horizontal transmission from multilocus phylogeny of deep-sea mussel (Mytilidae) symbionts: horizontal transmission of mussel symbionts. *Environmental Microbiology* **16**:3608–3621. doi: 10.1111/1462-2920.12379.

Galagan JE. 2014. Genomic insights into tuberculosis. *Nature Reviews Genetics* **15**:307–320. doi: 10.1038/nrg3664.

Gardebrecht A, Markert S, Sievert SM, Felbeck H, Thürmer A, Albrecht D, Wollherr A, Kabisch J, Le Bris N, Lehmann R, Daniel R, Liesegang H, Hecker M, Schweder T. 2012. Physiological homogeneity among the endosymbionts of *Riftia pachyptila* and *Tevnia jerichonana* revealed by proteogenomics. *The ISME Journal* **6**:766–776. doi: 10.1038/ismej.2011.137.

Gebruk AV, Chevaldonné P, Shank T, Lutz RA, Vrijenhoek RC. 2000. Deep-sea hydrothermal vent communities of the Logatchev area (14°45′N, Mid-Atlantic Ridge): diverse biotopes and high biomass. *Journal of the Marine Biological Association of the UK* **80**:383–393. doi: 10.1017/S0025315499002088.

Goffredi SK, Yi H, Zhang Q, Klann JE, Struve IA, Vrijenhoek RC, Brown CT. 2014. Genomic versatility and functional variation between two dominant heterotrophic symbionts of deep-sea *Osedax* worms. *The ISME Journal* **8**:908–924. doi: 10.1038/ismej.2013.201.

Gogarten JP, Townsend JP. 2005. Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology* **3**:679–687. doi: 10.1038/nrmicro1204.

Goodrich-Blair H, Clarke DJ. 2007. Mutualism and pathogenesis in *Xenorhabdus* and *Photorhabdus*: two roads to the same destination. *Molecular Microbiology* **64**:260–268. doi: 10.1111/j.1365-2958.2007.05671.x.

Haglund CM, Welch MD. 2011. Pathogens and polymers: microbe–host interactions illuminate the cytoskeleton. *The Journal of Cell Biology* **195**:7–17. doi: 10.1083/jcb.201103148.

Halary S, Riou V, Gaill F, Boudier T, Duperron S. 2008. 3D FISH for the quantification of methane- and sulphur-oxidizing endosymbionts in bacteriocytes of the hydrothermal vent mussel *Bathymodiolus azoricus*. *The ISME Journal* **2**:284–292. doi: 10.1038/ismej.2008.3.

Harada M, Yoshida T, Kuwahara H, Shimamura S, Takaki Y, Kato C, Miwa T, Miyake H, Maruyama T. 2009. Expression of genes for sulfur oxidation in the intracellular chemoautotrophic symbiont of the deep-sea bivalve *Calyptogena okutanii*. *Extremophiles* **13**:895–903. doi: 10.1007/s00792-009-0277-8.

Harmer TL, Rotjan RD, Nussbaumer AD, Bright M, Ng AW, DeChaine EG, Cavanaugh CM. 2008. Free-living tube worm endosymbionts found at deep-sea vents. *Applied and Environmental Microbiology* **74**:3895–3898. doi: 10.1128/AEM.02470-07.

Hart MM, Forsythe J, Oshowski B, Bücking H, Jansa J, Kiers ET. 2013. Hiding in a crowd—does diversity facilitate persistence of a low-quality fungal partner in the mycorrhizal symbiosis? *Symbiosis* **59**:47–56. doi: 10.1007/s13199-012-0197-8.

Heinz E, Williams TA, Nakjang S, Noël CJ, Swan DC, Goldberg AV, Harris SR, Weinmaier T, Markert S, Becher D, Bernhardt J, Dagan T, Hacker C, Lucocq JM, Schweder T, Rattei T, Hall N, Hirt RP, Embley TM. 2012. The genome of the obligate intracellular parasite *Trachipleistophora hominis*: new insights into microsporidian genome dynamics and reductive evolution. *PLOS Pathogens* **8**:e1002979. doi: 10.1371/journal.ppat.1002979.

Hentschel U, Steinert M, Hacker J. 2000. Common molecular mechanisms of symbiosis and pathogenesis. *Trends in Microbiology* **8**:226–231. doi: 10.1016/S0966-842X(00)01758-3.

Hopkins SH. 1957. B. Parasitism. *Geological society of america memoirs:* Geological Society of America. p. 413–428.

Hueck CJ. 1998. Type III protein secretion systems in bacterial pathogens of animals and plants. *Microbiology and Molecular Biology Reviews* **62**:379–433.

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C. 2009. InterPro: the integrative protein signature database. *Nucleic Acids Research* **37**:D211–D215. doi: 10.1093/nar/gkn785.

Jackson AP, Thomas GH, Parkhill J, Thomson NR. 2009. Evolutionary diversification of an ancient gene family (*rhs*) through C-terminal displacement. *BMC Genomics* **10**:584. doi: 10.1186/1471-2164-10-584.

Jiggins FM, Hurst GDD, Yang Z. 2002. Host-symbiont conflicts: positive selection on an outer membrane protein of parasitic but not mutualistic Rickettsiaceae. *Molecular Biology and Evolution* **19**:1341–1349.

Kadar 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.

Kaebernick M, Neilan BA, Börner T, Dittmann E. 2000. Light and the transcriptional response of the Microcystin biosynthesis gene cluster. *Applied and Environmental Microbiology* **66**:3387–3392. doi: 10.1128/AEM.66.8.3387-3392.2000.

Käll L, Storey JD, MacCoss MJ, Noble WS. 2008. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of Proteome Research* **7**:29–34. doi: 10.1021/pr700600n.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2011. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research* **40**:D109–D114. doi: 10.1093/nar/gkr988.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**:3059–3066. doi: 10.1093/nar/gkf436.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**:1647–1649. doi: 10.1093/bioinformatics/bts199.

Kelley LA, Sternberg MJE. 2009. Protein structure prediction on the Web: a case study using the Phyre server. *Nature Protocols* **4**:363–371. doi: 10.1038/nprot.2009.2.

Kleiner M, Petersen JM, Dubilier N. 2012a. Convergent and divergent evolution of metabolism in sulfur-oxidizing symbionts and the role of horizontal gene transfer. *Current Opinion in Microbiology* **15**:621–631. doi: 10.1016/j.mib.2012.09.003.

Kleiner M, Wentrup C, Lott C, Teeling H, Wetzel S, Young J, Chang YJ, Shah M, VerBerkmoes NC, Zarzycki J, Fuchs G, Markert S, Hempel K, Voigt B, Becher D, Liebeke M, Lalk M, Albrecht D, Hecker M, Schweder T, Dubilier N. 2012b. Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. *Proceedings of the National Academy of Sciences of USA* **109**:E1173–E1182. doi: 10.1073/pnas.1121198109.

Kobayashi I. 2001. Behavior of restriction–modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Research* **29**:3742–3756. doi: 10.1093/nar/29.18.3742.

Koskiniemi S, Lamoureux JG, Nikolakakis KC, t'Kint de Roodenbeke C, Kaplan MD, Low DA, Hayes CS. 2013. Rhs proteins from diverse bacteria mediate intercellular competition. *Proceedings of the National Academy of Sciences of USA* **110**:7032–7037. doi: 10.1073/pnas.1300627110.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes1. *Journal of Molecular Biology* **305**:567–580. doi: 10.1006/jmbi.2000.4315.

Kung VL, Khare S, Stehlik C, Bacon EM, Hughes AJ, Hauser AR. 2012. An *rhs* gene of *Pseudomonas aeruginosa* encodes a virulence protein that activates the inflammasome. *Proceedings of the National Academy of Sciences of USA* **109**:1275–1280. doi: 10.1073/pnas.1109285109.

Kuwahara H, Yoshida T, Takaki Y, Shimamura S, Nishi S, Harada M, Matsuyama K, Takishita K, Kawato M, Uematsu K, Fujiwara Y, Sato T, Kato C, Kitagawa M, Kato I, Maruyama T. 2007. Reduced genome of the thioautotrophic intracellular symbiont in a deep-sea clam, *Calyptogena okutanii*. *Current Biology* **17**:881–886. doi: 10.1016/j.cub.2007.04.039.

Kwak JS, Jeong H-G, Satchell KJF. 2011. *Vibrio vulnificus rtxA1* gene recombination generates toxin variants with altered potency during intestinal infection. *Proceedings of the National Academy of Sciences of USA* **108**:1645–1650. doi: 10.1073/pnas.1014339108.

Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Ussery DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* **35**:3100–3108. doi: 10.1093/nar/gkm160.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**:357–359. doi: 10.1038/nmeth.1923.

Lauenstein G, Harmon M, Gottholm B. 1993. *National status and trends program: monitoring site descriptions (1984–1990) for the national mussel watch and benthic surveillance projects:* National Oceanic and Atmospheric Administration. NOAA technical memorandum NOS ORCA-70. 1–358.
Lavik G, Stührmann T, Brüchert V, Van der Plas A, Mohrholz V, Lam P, Mußmann M, Fuchs BM, Amann R, Lass U, Kuypers MM. 2009. Detoxification of sulphidic African shelf waters by blooming chemolithotrophs. *Nature* **457**:581–584. doi: 10.1038/nature07588.
Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**:1658–1659. doi: 10.1093/bioinformatics/btl158.
Lindsay JA, Ruzin A, Ross HF, Kurepina N, Novick RP. 1998. The gene for toxic shock toxin is carried by a family of mobile pathogenicity islands in *Staphylococcus aureus*. *Molecular Microbiology* **29**:527–543. doi: 10.1046/j.1365-2958.1998.00947.x.
Linhartová I, Bumba L, Mašín J, Basler M, Osička R, Kamanová J, Procházková K, Adkins I, Hejnová-Holubová J, Sadílková L, Morová J, Sebo P. 2010. RTX proteins: a highly diverse family secreted by a common mechanism. *FEMS Microbiology Reviews* **34**:1076–1112. doi: 10.1111/j.1574-6976.2010.00231.x.
Liu X, Sheng J, Iii RC. 2011. Fatty acid production in genetically modified cyanobacteria. *Proceedings of the National Academy of Sciences of USA* **108**:6899–6904. doi: 10.1073/pnas.1103014108.
Liu Z, Müller J, Li T, Alvey RM, Vogl K, Frigaard NU, Rockwell NC, Boyd ES, Tomsho LP, Schuster SC, Henke P, Rohde M, Overmann J, Bryant DA. 2013. Genomic analysis reveals key aspects of prokaryotic symbiosis in the phototrophic consortium *"Chlorochromatium aggregatum"*. *Genome Biology* **14**:R127. doi: 10.1186/gb-2013-14-11-r127.
Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**:0955.
Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar , Buchner A, Lai T, Steppi S, Jobb G, Förster W, Brettske I, Gerber S, Ginhart AW, Gross O, Grumann S, Hermann S, Jost R, König A, Liss T, Lüssmann R, May M, Nonhoff B, Reichel B, Strehlow R, Stamatakis A, Stuckmann N, Vilbig A, Lenke M, Ludwig T, Bode A, Schleifer KH. 2004. ARB: a software environment for sequence data. *Nucleic Acids Research* **32**:1363–1371. doi: 10.1093/nar/gkh293.
Ma K, Vitek O, Nesvizhskii AI. 2012. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinformatics* **13**:S1. doi: 10.1186/1471-2105-13-S16-S1.
Markert S, Arndt C, Felbeck H, Becher D, Sievert SM, Hügler M, Albrecht D, Robidart J, Bench S, Feldman RA, Hecker M, Schweder T. 2007. Physiological proteomics of the uncultured endosymbiont of *Riftia pachyptila*. *Science* **315**:247–250. doi: 10.1126/science.1132913.
Markert S, Gardebrecht A, Felbeck H, Sievert SM, Klose J, Becher D, Albrecht D, Thürmer A, Daniel R, Kleiner M, Hecker M, Schweder T. 2011. Status quo in physiological proteomics of the uncultured *Riftia pachyptila* endosymbiont. *Proteomics* **11**:3106–3117. doi: 10.1002/pmic.201100059.
Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC. 2011. IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research* **40**:D115–D122. doi: 10.1093/nar/gkr1044.
McClure R, Balasubramanian D, Sun Y, Bobrovskyy M, Sumby P, Genco CA, Vanderpool CK, Tjaden B. 2013. Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Research* **41**:e140–e140. doi: 10.1093/nar/gkt444.
McFall-Ngai M, Hadfield MG, Bosch TC, Carey HV, Domazet-Lošo T, Douglas AE, Dubilier N, Eberl G, Fukami T, Gilbert SF, Hentschel U, King N, Kjelleberg S, Knoll AH, Kremer N, Mazmanian SK, Metcalf JL, Nealson K, Pierce NE, Rawls JF, Reid A, Ruby EG, Rumpho M, Sanders JG, Tautz D, Wernegreen JJ. 2013. Animals in a bacterial world, a new imperative for the life sciences. *Proceedings of the National Academy of Sciences of USA* **110**:3229–3236. doi: 10.1073/pnas.1218525110.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**:1297–1303. doi: 10.1101/gr.107524.110.
Meyer F. 2003. GenDB–an open source genome annotation system for prokaryote genomes. *Nucleic Acids Research* **31**:2187–2195. doi: 10.1093/nar/gkg312.
Melvin JA, Scheller EV, Miller JF, Cotter PA. 2014. *Bordetella pertussis* pathogenesis: current and future challenges. *Nature Reviews Microbiology* **12**:274–288. doi: 10.1038/nrmicro3235.
Meusch D, Gatsogiannis C, Efremov RG, Lang AE, Hofnagel O, Vetter IR, Aktories K, Raunser S. 2014. Mechanism of Tc toxin action revealed in molecular detail. *Nature* **508**:61–65. doi: 10.1038/nature13015.
Minguez L, Buronfosse T, Giambérini L. 2012. Different host exploitation strategies in two zebra mussel-trematode systems: Adjustments of host life history traits. *PLOS ONE* **7**:e34029. doi: 10.1371/journal.pone.0034029.
Moebius N, Üzüm Z, Dijksterhuis J, Lackner G, Hertweck C. 2014. Active invasion of bacteria into living fungal cells. *eLife* **3**:e03007. doi: 10.7554/eLife.03007.
Moreira D, López-García P. 2003. Are hydrothermal vents oases for parasitic protists? *Trends in Parasitology* **19**:556–558. doi: 10.1016/j.pt.2003.09.013.
Murillo AA, Ramírez-Flandes S, DeLong EF, Ulloa O. 2014. Enhanced metabolic versatility of planktonic sulfur-oxidizing γ-proteobacteria in an oxygen-deficient coastal ecosystem. *Aquatic Microbiology* **1**:18. doi: 10.3389/fmars.2014.00018.
Muyzer G, Teske A, Wirsen CO, Jannasch HW. 1995. Phylogenetic relationships of *Thiomicrospira* species and their identification in deep-sea hydrothermal vent samples by denaturing gradient gel electrophoresis of 16S rDNA fragments. *Archives of Microbiology* **164**:165–172. doi: 10.1007/BF02529967.
Newton I, Girguis P, Cavanaugh C. 2008. Comparative genomics of vesicomyid clam (Bivalvia: Mollusca) chemosynthetic symbionts. *BMC Genomics* **9**:585. doi: 10.1186/1471-2164-9-585.

Newton IL, Woyke T, Auchtung TA, Dilly GF, Dutton RJ, Fisher MC, Fontanez KM, Lau E, Stewart FJ, Richardson PM, Barry KW, Saunders E, Detter JC, Wu D, Eisen JA, Cavanaugh CM. 2007. The *Calyptogena magnifica* chemoautotrophic symbiont genome. *Science* **315**:998–1000. doi: 10.1126/science.1138438.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**:299–304. doi: 10.1038/35012500.

Ogier JC, Pagès S, Bisch G, Chiapello H, Médigue C, Rouy Z, Teyssier C, Vincent S, Tailliez P, Givaudan A, Gaudriault S. 2014. Attenuated virulence and genomic reductive evolution in the entomopathogenic bacterial symbiont species, *Xenorhabdus poinarii*. *Genome Biology and Evolution* **6**:1495–1513. doi: 10.1093/gbe/evu119.

Ohno M, Ménez R, Ogawa T, Danse JM, Shimohigashi Y, Fromen C, Ducancel F, Zinn-justin S, Le du MH, Boulain JC, Tamiya T, Ménez A. 1998. Molecular evolution of snake toxins: is the functional diversity of snake toxins associated with a mechanism of accelerated evolution? *Progress in Nucleic Acid Research and Molecular Biology* **59**:307–364.

Oliver KM, Russell JA, Moran NA, Hunter MS. 2003. Facultative bacterial symbionts in aphids confer resistance to parasitic wasps. *Proceedings of the National Academy of Sciences of USA* **100**:1803–1807. doi: 10.1073/pnas.0335320100.

Oliver KM, Moran NA, Hunter MS. 2005. Variation in resistance to parasitism in aphids is due to symbionts not host genotype. *Proceedings of the National Academy of Sciences of USA* **102**:12795–12800. doi: 10.1073/pnas.0506131102.

Oliver KM, Degnan PH, Burke GR, Moran NA. 2010. Facultative symbionts in aphids and the horizontal transfer of ecologically important traits. *Annual Review of Entomology* **55**:247–266. doi: 10.1146/annurev-ento-112408-085305.

Paerl H. 2008. Nutrient and other environmental controls of harmful cyanobacterial blooms along the freshwater–marine continuum. In: Hudnell HK, editor. *Cyanobacterial harmful algal blooms: state of the science and research needs*, Springer New York. p. 217–237.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **25**:1043–1055. doi: 10.1101/gr.186072.114.

Partida-Martinez LP, Groth I, Schmitt I, Richter W, Roth M, Hertweck C. 2007. *Burkholderia rhizoxinica* sp. nov. and *Burkholderia endofungorum* sp. nov., bacterial endosymbionts of the plant-pathogenic fungus *Rhizopus microsporus*. *International Journal of Systematic and Evolutionary Microbiology* **57**:2583–2590. doi: 10.1099/ijs.0.64660-0.

Pernthaler A, Pernthaler J, Amann R. 2002. Fluorescence *in situ* hybridization and catalyzed reporter deposition for the identification of marine bacteria. *Applied and Environmental Microbiology* **68**:3094–3101. doi: 10.1128/AEM.68.6.3094-3101.2002.

Petersen J, Dubilier N. 2014. Gene swapping in the dead zone. *eLife* **3**. doi: 10.7554/eLife.04600.

Petersen JM, Dubilier N. 2009. Methanotrophic symbioses in marine invertebrates: methanotrophic symbioses in marine invertebrates. *Environmental Microbiology Reports* **1**:319–335. doi: 10.1111/j.1758-2229.2009.00081.x.

Petersen JM, Zielinski FU, Pape T, Seifert R, Moraru C, Amann R, Hourdez S, Girguis PR, Wankel SD, Barbe V, Pelletier E, Fink D, Borowski C, Bach W, Dubilier N. 2011. Hydrogen is an energy source for hydrothermal vent symbioses. *Nature* **476**:176–180. doi: 10.1038/nature10325.

Petersen JM, Wentrup C, Verna C, Knittel K, Dubilier N. 2012. Origins and evolutionary flexibility of chemosynthetic symbionts from deep-sea animals. *The Biological Bulletin* **223**:123–137.

Ponting CP, Schultz J, Milpetz F, Bork P. 1999. SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Research* **27**:229–232. doi: 10.1093/nar/27.1.229.

Poole SJ, Diner EJ, Aoki SK, Braaten BA, t'Kint de Roodenbeke C, Low DA, Hayes CS. 2011. Identification of functional toxin/Immunity genes linked to contact-dependent growth inhibition (CDI) and rearrangement hotspot (Rhs) systems. *PLOS Genetics* **7**:e1002217. doi: 10.1371/journal.pgen.1002217.

Powell EN, Barber RD, Kennicutt MC II, Ford SE. 1999. Influence of parasitism in controlling the health, reproduction and PAH body burden of petroleum seep mussels. *Deep Sea Research Part I Oceanographic Research Papers* **46**:2053–2078. doi: 10.1016/S0967-0637(99)00035-7.

Pruesse E, Peplies J, Glöckner FO. 2012. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**:1823–1829. doi: 10.1093/bioinformatics/bts252.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**:D590–D596. doi: 10.1093/nar/gks1219.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841–842. doi: 10.1093/bioinformatics/btq033.

R Development Core Team. 2011. *R: a language and environment for statistical computing*.

Rasko DA, Myers GS, Ravel J. 2005. Visualization of comparative genomic analyses by BLAST score ratio. *BMC Bioinformatics* **6**:2. doi: 10.1186/1471-2105-6-2.

Reams AB, Neidle EL. 2004. Selection for gene clustering by tandem duplication. *Annual Review of Microbiology* **58**:119–142. doi: 10.1146/annurev.micro.58.030603.123806.

Richter M, Lombardot T, Kostadinov I, Kottmann R, Duhaime MB, Peplies J, Glöckner FO. 2008. JCoast—a biologist-centric software tool for data mining and comparison of prokaryotic (meta)genomes. *BMC Bioinformatics* **9**:177. doi: 10.1186/1471-2105-9-177.

Riou V, Duperron S, Halary S, Dehairs F, Bouillon S, Martins I, Colaço A, Serrão Santos R. 2010. Variation in physiological indicators in *Bathymodiolus azoricus* (Bivalvia: Mytilidae) at the Menez Gwen Mid-Atlantic Ridge deep-sea hydrothermal vent site within a year. *Marine Environmental Research* **70**:264–271. doi: 10.1016/j.marenvres.2010.05.008.

Rissman AI, Mau B, Biehl BS, Darling AE, Glasner JD, Perna NT. 2009. Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics* **25**:2071–2073. doi: 10.1093/bioinformatics/btp356.
Robidart JC, Bench SR, Feldman RA, Novoradovsky A, Podell SB, Gaasterland T, Allen EE, Felbeck H. 2008. Metabolic versatility of the *Riftia pachyptila* endosymbiont revealed through metagenomics. *Environmental Microbiology* **10**:727–737. doi: 10.1111/j.1462-2920.2007.01496.x.
Rocha EPC, Danchin A. 2002. Base composition bias might result from competition for metabolic resources. *Trends in Genetics* **18**:291–294. doi: 10.1016/S0168-9525(02)02690-2.
Roig FJ, Gonzalez-Candelas F, Amaro C. 2010. Domain organization and evolution of multifunctional autoprocessing repeats-in-toxin (MARTX) Toxin in *Vibrio vulnificus*. *Applied and Environmental Microbiology* **77**:657–668. doi: 10.1128/AEM.01806-10.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**:1572–1574. doi: 10.1093/bioinformatics/btg180.
Rouchet R, Vorburger C. 2014. Experimental evolution of parasitoid infectivity on symbiont-protected hosts leads to the emergence of genotype specificity. *Evolution* **68**:1607–1616. doi: 10.1111/evo.12377.
Russell JA, Weldon S, Smith AH, Kim KL, Hu Y, Łukasik P, Doll S, Anastopoulos I, Novin M, Oliver KM. 2013. Uncovering symbiont-driven genetic diversity across North American pea aphids. *Molecular Ecology* **22**:2045–2059. doi: 10.1111/mec.12211.
Sakiyama T, Ueno H, Homma H, Numata O, Kuwabara T. 2006. Purification and characterization of a hemolysin-like protein, Sll1951, a nontoxic member of the RTX protein family from the *Cyanobacterium Synechocystis* sp. Strain PCC 6803. *Journal of Bacteriology* **188**:3535–3542. doi: 10.1128/JB.188.10.3535-3542.2006.
Sakiyama T, Araie H, Suzuki I, Shiraiwa Y. 2011. Functions of a hemolysin-like protein in the cyanobacterium *Synechocystis* sp. PCC 6803. *Archives of Microbiology* **193**:565–571. doi: 10.1007/s00203-011-0700-2.
Sarrazin J, Sarradin P-M, Allais A-G, Momareto Cruise Participants X. 2006. MoMARETO: a cruise dedicated to the spatio-temporal dynamics and the adaptations of hydrothermal vent fauna on the Mid-Atlantic Ridge. *InterRidge News* **15**:24–33.
Satchell KJF. 2007. MARTX, multifunctional autoprocessing repeats-in-toxin toxins. *Infection and Immunity* **75**:5079–5084. doi: 10.1128/IAI.00525-07.
Satchell KJF. 2011. Structure and function of MARTX toxins and other large repetitive RTX proteins. *Annual Review of Microbiology* **65**:71–90. doi: 10.1146/annurev-micro-090110-102943.
Schmid M, Sieber R, Zimmermann Y-S, Vorburger C. 2012. Development, specificity and sublethal effects of symbiont-conferred resistance to parasitoids in aphids. *Functional Ecology* **26**:207–215. doi: 10.1111/j.1365-2435.2011.01904.x.
Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**:2498–2504. doi: 10.1101/gr.1239303.
Snyder AK, Rio RVM. 2013. Interwoven biology of the Tsetse holobiont. *Journal of Bacteriology* **195**:4322–4330. doi: 10.1128/JB.00487-13.
Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics* **16**:472–482. doi: 10.1038/nrg3962.
Srinivasan M, Moon YB. 1999. A comprehensive clustering algorithm for strategic analysis of supply chain networks. *Computers & Industrial Engineering* **36**:615–633. doi: 10.1016/S0360-8352(99)00155-2.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**:2688. doi: 10.1093/bioinformatics/btl446.
Steele-Mortimer O, Brumell JH, Knodler LA, Méresse S, Lopez A, Finlay BB. 2002. The invasion-associated type III secretion system of *Salmonella enterica* serovar Typhimurium is necessary for intracellular proliferation and vacuole biogenesis in epithelial cells. *Cellular Microbiology* **4**:43–54. doi: 10.1046/j.1462-5822.2002.00170.x.
Strous M, Kraft B, Bisdorf R, Tegetmeyer HE. 2012. The binning of metagenomic contigs for microbial physiology of mixed cultures. *Frontiers in Microbiology* **3**. doi: 10.3389/fmicb.2012.00410.
Sunamura M, Higashi Y, Miyako C, Ishibashi J, Maruyama A. 2004. Two bacteria phylotypes are predominant in the Suiyo Seamount hydrothermal plume. *Applied and Environmental Microbiology* **70**:1190–1198. doi: 10.1128/AEM.70.2.1190-1198.2004.
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. doi: 10.1093/dnares/dss005.
Tatusov RL. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* **28**:33–36. doi: 10.1093/nar/28.1.33.
Van Dover C. 2000. *The ecology of deep-sea hydrothermal vents*. Princeton, NJ: Princeton University Press.
Van Dover CL, German CR, Speer KG, Parson LM, Vrijenhoek RC. 2002. Evolution and biogeography of deep-sea vent and seep invertebrates. *Science* **295**:1253–1257. doi: 10.1126/science.1067361.
Vogl K, Wenter R, Dreßen M, Schlickenrieder M, Plöscher M, Eichacker L, Overmann J. 2008. Identification and analysis of four candidate symbiosis genes from "*Chlorochromatium aggregatum*", a highly developed bacterial symbiosis. *Environmental Microbiology* **10**:2842–2856. doi: 10.1111/j.1462-2920.2008.01709.x.
von Cosel R. 2002. A new species of bathymodioline mussel (Mollusca, Bivalvia, Mytilidae) from Mauritania (West Africa), with comments on the genus *Bathymodiolus* Kenk & Wilson, 1985. *Zoosystema* **24**:259–272.
Wagner-Döbler I, Biebl H. 2006. Environmental biology of the marine *Roseobacter* lineage. *Annual Review of Microbiology* **60**:255–280. doi: 10.1146/annurev.micro.60.080805.142115.

Walsh DA, Zaikova E, Howes CG, Song YC, Wright JJ, Tringe SG, Tortell PD, Hallam SJ. 2009. Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. *Science* **326**:578–582. doi: 10.1126/science.1175309.

Ward ME, Shields JD, Van Dover CL. 2004. Parasitism in species of *Bathymodiolus* (Bivalvia: Mytilidae) mussels from deep-sea seep and hydrothermal vents. *Diseases of Aquatic Organisms* **62**:1–16. doi: 10.3354/dao062001.

Wardle WJ. 1988. A Bucephalid Larva, *Cercaria pleuromerae* n. sp. (Trematoda: Digenea), parasitizing a deepwater Bivalve from the Gulf of Mexico. *The Journal of Parasitology* **74**:692–694. doi: 10.2307/3282191.

Waterfield NR, Bowen DJ, Fetherston JD, Perry RD, others. 2001. The *tc* genes of *Photorhabdus*: a growing family. *Trends in Microbiology* **9**:185–191. doi: 10.1016/S0966-842X(01)01978-3.

Wentrup C, Wendeberg A, Schimak M, Borowski C, Dubilier N. 2014. Forever competent: deep-sea bivalves are colonized by their chemosynthetic symbionts throughout their lifetime. *Environmental Microbiology* **16**:3699–3713. doi: 10.1111/1462-2920.12597.

West SA, Kiers ET, Pen I, Denison RF. 2002. Sanctions and mutualism stability: when should less beneficial mutualists be tolerated? *Journal of Evolutionary Biology* **15**:830–837. doi: 10.1046/j.1420-9101.2002.00441.x.

Wittkop T, Emig D, Lange S, Rahmann S, Albrecht M, Morris JH, Böcker S, Stoye J, Baumbach J. 2010. Partitioning biological data with transitivity clustering. *Nature Methods* **7**:419–420. doi: 10.1038/nmeth0610-419.

Won Y, Young C, Lutz R, Vrijenhoek R. 2003a. Dispersal barriers and isolation among deep-sea mussel populations (Mytilidae: *Bathymodiolus*) from eastern Pacific hydrothermal vents. *Molecular Ecology* **12**:169–184. doi: 10.1046/j.1365-294X.2003.01726.x.

Won YJ, Hallam SJ, O'Mullan GD, Pan IL, Buck KR, Vrijenhoek RC. 2003b. Environmental acquisition of thiotrophic endosymbionts by deep-sea mussels of the genus *Bathymodiolus*. *Applied and Environmental Microbiology* **69**:6785. doi: 10.1128/AEM.69.11.6785-6792.2003.

Wood AP, Aurikko JP, Kelly DP. 2004. A challenge for 21st century molecular biology and biochemistry: what are the causes of obligate autotrophy and methanotrophy? *FEMS Microbiology Reviews* **28**:335–352. doi: 10.1016/j.femsre.2003.12.001.

Wright JJ, Konwar KM, Hallam SJ. 2012. Microbial ecology of expanding oxygen minimum zones. *Nature Reviews Microbiology* **10**:381–394. doi: 10.1038/nrmicro2778.

Zhang D, de Souza RF, Anantharaman V, Iyer LM, Aravind L. 2012. Polymorphic toxin systems: comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. *Biology Direct* **7**:18. doi: 10.1186/1745-6150-7-18.

Zhou J, Bruns MA, Tiedje JM. 1996. DNA recovery from soils of diverse composition. *Applied and Environmental Microbiology* **62**:316–322.

Zielinski FU, Pernthaler A, Duperron S, Raggi L, Giere O, Borowski C, Dubilier N. 2009. Widespread occurrence of an intranuclear bacterial parasite in vent and seep bathymodiolin mussels. *Environmental Microbiology* **11**:1150–1167. doi: 10.1111/j.1462-2920.2008.01847.x.

Zuleta LF, Cunha Cde O, de Carvalho FM, Ciapina LP, Souza RC, Mercante FM, de Faria SM, Baldani JI, Straliotto R, Hungria M, de Vasconcelos AT. 2014. The complete genome of *Burkholderia phenoliruptrix* strain BR3459a, a symbiont of *Mimosa flocculosa*: highlighting the coexistence of symbiotic and pathogenic genes. *BMC Genomics* **15**:535. doi: 10.1186/1471-2164-15-535.

Appendix 1

The role of the symbionts in the *Bathymodiolus* association

In their complement of toxin-related genes (TRGs), the *Bathymodiolus* symbiont genomes closely resemble those of pathogenic bacteria. In fact, it has a larger arsenal of these particular types of TRGs than any so far sequenced pathogen. Considering this, is it possible that their role in the association with their mussel hosts has been misunderstood? Could they in fact be pathogens of *Bathymodiolus* rather than beneficial symbionts? Chemosynthetic symbionts are widely assumed to benefit their hosts by contributing a source of organic matter for their nutrition and so far this was also thought to be true for the *Bathymodiolus* symbioses (***Van Dover, 2000***; ***Cavanaugh et al., 2006***; ***Dubilier et al., 2008***). There is a variety of evidence supporting that the *Bathymodiolus* symbionts are in fact an effective source of nutrition, including: (1) *Bathymodiolus* mussels can grow to densities two to five times higher than their shallow-water relatives that rely exclusively on filter-feeding for their nutrition (***Gebruk et al., 2000***), (2) *Bathymodiolus* mussels have a reduced feeding groove and digestive system compared to their filter-feeding relatives (Gustafson et al., 1998), (3) carbon transfer from sulfur-oxidizing (SOX) symbionts to *B. azoricus* has been demonstrated experimentally (***Riou et al., 2010***), and (4) based on their genomes, the symbionts are autotrophs (see below). These observations are all consistent with the hypothesis that the SOX symbionts benefit their *Bathymodiolus* hosts by providing them with a source of nutrition.

Symbiont metabolism

The symbionts encode a modified version of the Calvin cycle that lacks the enzyme sedoheptulose-1,7-biphosphatase and fructose-1,6-bisphosphatase. Instead, the symbionts have a pyrophosphate-dependent phosphofructokinase (PPi-PFK) as described in other chemosynthetic symbioses (***Markert et al., 2011***; ***Kleiner et al., 2012b***; ***Dmytrenko et al., 2014***). In contrast to the clam symbionts and SUP05 that possess ribulose bisphosphate carboxylase/oxidase (RubisCO) form II, *Bathymodiolus* symbionts have RubisCO form I. RubiscCO form I is better adapted to higher O_2 concentrations and has a medium to low affinity to CO_2, indicating an adaptation to environments with O_2 present (***Cavanaugh and Robinson, 1996***; ***Badger and Bek, 2008***).

The genome did not provide sufficient evidence to determine whether the *Bathymodiolus* symbionts are obligate autotrophs or mixotrophs. None of the three essentially complete symbiont genomes encode the enzyme alpha-ketoglutarate dehydrogenase, which is often absent in the genomes of obligate autotrophs (***Wood et al., 2004***). Some other genes of the tricarboxylic acid (TCA) cycle were also missing, but this may be because the genomes are not yet closed. A few tripartite ATP-independent periplasmic (TRAP) transporters were present, and these could be involved in uptake of organic compounds.

To acquire the necessary energy to fuel carbon fixation, the *Bathymodiolus* symbionts can utilize reduced sulfur compounds and hydrogen (***Petersen et al., 2011***; ***Kleiner et al., 2012a***). The pathways for sulfur oxidation present in the genomes are similar to those of other SOX symbionts of *R. pachyptila* tubeworms, or vesicomyid and solemyid clam symbionts (***Figure 2—figure supplement 3***) (***Robidart et al., 2008***; ***Harada et al., 2009***; ***Kleiner et al., 2012a***; ***Dmytrenko et al., 2014***). We found the genes that encode SoxABXYZ and sulfide-quinone reductases (Sqr), which reduce cytochrome c and quinones using sulfide and/or thiosulfate as electron donor. The *Bathymodiolus* SOX symbionts might store elemental sulfur based on the absence of SoxCD sulfur dehydrogenase, although the symbionts lack the large intracellular inclusions typical of sulfur storage, and elemental sulfur has not yet been detected in mussel gills. Sulfide can be oxidized to sulfite with the reverse dissimilatory sulfite reduction pathway (rDsr), and sulfite can be further oxidized to sulfate with APS reductase and an ATP-generating ATP sulfurylase (Sat) (Dahl et al., 2008).

Besides the use of oxygen as electron acceptor, we found in the most complete genome of the symbiont from *B. azoricus* the genes for the use of nitrate as electron acceptor (nitrate reductase—*narGHIJ*); however, we did not find these genes in the other two symbiont genomes. The genes required to convert nitrite to nitrous oxide were found in all three-draft genomes (nitrite reductase—*nirK* and nitrite oxidoreductase—*norCB*). We might have not found the nitrate reductase in the draft genomes of BspSym and BazSymB due to the incompleteness (**Table 1**). The use of an alternative electron acceptor could reduce the competition for oxygen with their host.

Appendix 2

Other potential roles of the TRGs

RTX toxins as nutrient-scavenging proteins

RTX genes were not preferentially found in host-associated bacteria but were also regularly found in free-living bacteria. Bloom-forming aquatic cyanobacteria had some of the highest numbers of RTX genes (***Figure 5***, ***Figure 5—figure supplement 3***). Cyanobacteria are well-known producers of secondary metabolite toxins called cyanotoxins, but protein toxins of cyanobacteria have received far less attention (***Kaebernick et al., 2000***). A hemolysin-like protein from *Synechocystis* sp. PCC 6803 was shown to be located in the S-layer of *Synechocystis* cells, and its deletion causes cells to be more permeable and more sensitive to environmental toxins (***Liu et al., 2011***; ***Sakiyama et al., 2011***). This hemolysin-like protein therefore seems to have a protective rather than a toxic function, although it did have slight hemolytic activity when tested against sheep erythrocytes (***Sakiyama et al., 2006***). However, the function of most cyanobacterial RTX toxins has not yet been investigated. Intriguingly, cyanobacterial blooms are tightly coupled to nutrient availability (***Paerl, 2008***). If free-living cyanobacteria can use RTX proteins as toxins to obtain nutrients by lysing other organisms, analogous to pathogens lysing host cells, they could gain access to additional sources of nutrients.

Members of the marine *Roseobacter* clade also had an unusually large number of RTX genes (***Figure 5—figure supplement 3***). *Roseobacter* are some of the most abundant bacteria in marine habitats. They commonly associate with a range of eukaryotic organisms from single-celled phytoplankton to vertebrate and invertebrate animals (***Buchan et al., 2005***; ***Wagner-Döbler and Biebl, 2006***; ***Brinkhoff et al., 2008***). Previous studies have shown that the RTX genes of *Roseobacter* are expressed, and that they can make up to 90% of the excreted proteome (***Christie-Oleza et al., 2012***). However, their function in *Roseobacter* is still unknown.

Could strain variation of *Bathymodiolus* symbionts be promoted by a defensive role of TRGs?

A role for the *Bathymodiolus* SOX symbiont in host defense against pathogens could help to explain why the mussels seem to associate with multiple closely related SOX strains (***Duperron et al., 2006***). Associations with multiple very similar symbiont strains are considered evolutionarily unstable because they promote the emergence of 'cheaters', symbionts that benefit from the association but do not contribute to the cost (***West et al., 2002***; ***Hart et al., 2013***). For example, a 'cheater' could be a SOX symbiont that occupies space within the host, thus gaining access to the electron donors and acceptors it needs for its growth, but passes on less organic matter to the host than neighboring 'cooperators'. Cheaters have a distinct growth advantage, and could easily outcompete cooperators, ultimately leading to the breakdown of the association. Despite this, multiple strains of SOX bacteria can occur in individual *Bathymodiolus* mussels (***Won et al., 2003a***; ***DeChaine et al., 2006***). Intriguingly, the SOX symbiont strain variation identified in genetic markers such as the 16S-ITS-23S rRNA gene operon is similar to the strain variation seen in natural populations of defensive *Ca.* H. defensa symbionts (***Russell et al., 2013***). Moreover, we have shown that closely related SOX symbionts show extensive sequence variation in most TRGs, which is also the case for *Ca.* H. defensa (***Russell et al., 2013***). This large natural diversity is inconsistent with a role for the TRGs in host-symbiont recognition and communication, as genes mediating these processes should be conserved (***Jiggins et al., 2002***; ***Bailly et al., 2006***). However, this diversity could be explained by antagonistic co-evolution between protective symbionts and their host's parasites. Indeed, experimental evolution shows that natural enemies quickly acquire resistance to *Ca.* H. defensa (***Rouchet and Vorburger, 2014***). However, this resistance is highly specific to the strain of *Ca.* H. defensa, the parasite is

experimentally challenged with and is useless against other closely related strains (***Schmid et al., 2012***; ***Rouchet and Vorburger, 2014***). Associating with diverse symbiont strains, each with its own unique arsenal of TRGs, could therefore protect the host in the face of rapidly evolving parasite resistance.

Appendix 3

Structure of *Bathymodiolus* symbiont TRGs and genomic regions

In BspSym, the 19 multifunctional autoprocessing RTX (MARTX)-like genes were clustered in two genomic regions that spanned 58.5 kb (MARTX1) and 46.3 kb (MARTX2). Genes predicted in these regions encoded many of the unique features of MARTX genes and other large repetitive RTX proteins including RTX repeats, cysteine protease domains (CPDs), hemagglutinin, and adhesion domains (***Figure 4—figure supplement 2***). Moreover, some of these genes were very large, which is typical of MARTX and MARTX-like genes (***Satchell, 2007***, ***2011***). BazSymA and BazSymB also encoded genes with characteristic signatures of MARTX genes, but these two draft genomes were highly fragmented, and the MARTX genes were found on very small genome fragments.

Both of the BspSym MARTX regions contained genes encoding RTX repeats and RTX CPDs. The RTX CPD has so far only been found in MARTX genes. Its function is to cleave off effector domains, whose sequence and effects on eukaryotic hosts can vary substantially even in MARTX genes from different strains of the same bacterial species (***Roig et al., 2010***; ***Kwak et al., 2011***). Beyond these shared features, MARTX1 and MARTX2 of the BspSym had distinct differences. Genes in MARTX1 encoded six hemagglutinin domains, two adhesion domains, and a pectin lyase domain. Similar features are found in large repetitive RTX adhesins, except that the CPD domain is not usually found in this protein subfamily (***Satchell, 2011***). In fact, the combination of hemagglutinin, pectin lyase, and CPD domains has so far only been found in the filamentous hemagglutinin protein of the betaproteobacterial pathogens *B. pertussis* and *B. bronchiseptica*, which is a key determinant of epithelial cell attachment and immune response modulation in *Bordetella* infections (***Melvin et al., 2014***). In contrast to *Bordetella*, where these domains are all encoded in one very large ORF, these domains were encoded in neighboring ORFs in the *Bathymodiolus* symbiont. However, MARTX-encoding genes are notoriously difficult to predict due to their repetitive structure (***Satchell, 2007***, ***2011***). We therefore cannot exclude the possibility that these may in fact encode one large protein in the *Bathymodiolus* symbiont.

The second MARTX-like gene cluster in the BspSym genome, MARTX2, resembles a 'typical' hemolysin operon. It contains all the necessary genes for type I secretion, and a gene annotated as *rtxA*, which encodes a hemolysin toxin. However, in contrast to typical hemolysin operons, it is missing the hemolysin activator gene. Interestingly, MARTX1, but not MARTX2, contains a gene annotated as a hemolysin activator. Moreover, the MARTX2 region contains four genes that each has up to five CPD domains. The CPD domain is typical for MARTX genes but not for known hemolysins. This region therefore has features characteristic of both RTX and MARTX genes.

eLIFE Research article Genomics and evolutionary biology | Microbiology and infectious disease

Appendix 4

Supplementary materials and methods

Symbiont and host separation via density gradient centrifugation for genomics and proteomics from gill tissue

To separate the SOX symbionts from gill tissues and to obtain a fraction enriched in host cytosolic proteins from the gill, a combination of differential and rate-zonal centrifugation was applied. All steps were carried out at 4˚C. Around 1 g of gill tissue was removed from a mussel and transferred into a Duall homogenizer (Tissue grind pestle and tube SZ22, Kontes Glass Company, Vineland, New Jersey). The tissue was covered with 1× PBS and homogenized thoroughly with about 20 strokes. To remove host nuclei and large tissue fragments, the homogenate was filled into a 15 ml conical tube, 1× PBS added up to 15 ml, and centrifuged for 10 min at 700×*g* in a swing out rotor. The supernatant was carefully transferred into a new 15 ml tube and the pellet (first pellet) frozen at −80˚C. The supernatant was centrifuged again for 10 min at 700×*g* in a swing out rotor to remove any remaining nuclei and tissue fragments. The resulting supernatant was transferred to a new tube and centrifuged in a fixed angle rotor for 10 min at 15.000×*g* to pellet all symbionts, mitochondria, and small tissue fragments. The supernatant from this centrifugation step, which contained the cytosolic host proteins, was frozen at −80˚C. The pellet was resuspended in 1× PBS and the suspension was layered on top of a discontinuous density gradient made with HistoDenz (Sigma) dissolved in 1× PBS. The density gradient was set up in 5% steps from 5 to 25% (wt/vol) HistoDenz. The density gradient was centrifuged in a swing out rotor for 7 min at 3000×*g* and was then divided into equally sized fractions. The pellet in the gradient tube (gradient pellet) was frozen at −80˚C until processing. The gradient fractions were washed twice with 1× PBS to remove the HistoDenz and the resulting pellets were frozen at −80˚C. Subsamples for catalyzed reporter deposition-fluorescence in situ hybridization (CARD-FISH) analyses were taken from the homogenates, pellets, and gradient fractions prior to freezing for later analysis of sample composition.

Analysis of symbiont composition in density gradient fractions using CARD-FISH

To determine the abundance and composition of host material and the different symbiont species in the homogenates, supernatants, pellets, and gradient fractions, we used CARD-FISH. Subsamples of gradient pellets were fixed overnight in 1% formaldehyde at 4˚C, washed three times in 1× PBS, and stored at −20˚C in 50% 1× PBS and 50% ethanol. In the home laboratory, cells were poured through 0.22-µm filters. CARD-FISH was done as described previously (***Pernthaler et al., 2002***) with minor modifications. Endogenous peroxidases were inactivated with 0.01 M HCl for 10 min at room temperature. Double hybridization of rRNA was performed with the probes BMARt-193 for the SOX symbiont and BMARm-845 for the methane-oxidizing (MOX) symbionts (***Duperron et al., 2006***). Sections were hybridized at 46˚C for 2 hr, followed by washing at 48˚C for 20 min. Amplification was done for 10 min at 37˚C, followed by inactivation of horseradish peroxidase (HRP) using methanol. Prior to microscopic evaluation, the cells were counterstained with 1 µg ml^{-1} 4′,6-diamidino-2-phenylindole (DAPI). Cells with a symbiont-specific signal were counted against at least 500 DAPI signals per filter section.

Proteomic analysis

Extraction of cytosolic and membrane proteins

For the proteomic analyses, gradient pellets, host supernatant (described above), and frozen tissues from mussel gill and foot were processed in biological duplicates. For washing, the samples were resuspended in TE buffer (10 mM Tris pH 7.5, 10 mM EDTA pH 8.0, 1× complete Protease Inhibitor Cocktail [Roche Applied Science, Germany]) and centrifuged briefly (3 min, 4˚C and 21,500×*g*). The resultant pellets were homogenized using a Duall homogenizer and the

homogenate was transferred to low binding 1.7 ml reaction tubes (Sorenson BioScience Inc., Salt Lake City, UT). Cells were lysed on ice using a sonicator (Bandelin Sonopuls ultrasonic homogenizer, Germany); 2 × 25 s at 30% power and a cycle of 0.5 s, with a 30 s pause. The lysate was centrifuged to remove cell debris (10 min, 4˚C and 15,300×*g*). The resulting supernatants, that is, protein raw extracts, were further subjected to ultracentrifugation (100,000×*g* for 60 min at 4˚C), allowing for the enrichment of membranes and membrane-associated proteins in the pellets and accumulation of soluble proteins in the supernatants. Protein concentrations in the soluble protein extracts were determined using the method described by ***Bradford, 1976*** and aliquots for all four sample types (symbiont-enriched gradient pellet, host supernatant, gill tissue, and foot tissue) were stored at −80˚C until MS analysis. Additionally, for gill tissue and gradient pellet samples, membrane proteins were purified from the enriched membrane fraction (see above) according to the protocol of ***Eymann et al. (2004)*** as described by ***Markert et al. (2007)***. The purified membrane protein fraction (in 30 µl of 50 mM triethylammonium bicarbonate buffer, pH 7.8) was transferred to low binding 1.7 ml tubes and immediately used for MS analysis. Due to sample scarcity, the biological duplicates for these membrane protein samples were pooled together to obtain enough protein for MS analysis.

1D-PAGE-LC-MS/MS

Approximately, 20 µg of protein of the membrane and soluble protein extracts, respectively, was dissolved in 20 µl of sample loading buffer (100 mM Tris-HCl (pH 6.8), 10% sodium dodecyl sulfate (SDS), 20% glycerol, 5% β-mercaptoethanol, 0.1% bromophenol blue). Proteins were separated by one-dimensional (1D) SDS PAGE in pre-cast 10% polyacrylamide gels (BioRad), fixed and stained with Coomassie Brilliant Blue (G250, Sigma-Aldrich, Germany). The gel lanes were cut into 10 equal sized pieces, destained with washing buffer (200 mM ammonium hydrogen carbonate, 30% acetonitrile), dried and proteins were subjected to overnight in-gel digestion with trypsin (Promega, Germany) (***Heinz et al., 2012***). Prior to mass spectrometric analysis, the peptides were purified using Ziptips (P10, U-C_{18}, Millipore).

Tryptic digests of the cytosolic samples were subjected to liquid chromatography performed on an EASYnLC (Proxeon, Denmark) with self-packed columns (Luna 3µ C18(2) 100A, Phenomenex, Germany) in a one-column setup. Following loading and desalting at a flow of 700 nl/min at a maximum of 220 bar of water in 0.1% acetic acid, separation of the peptides was achieved by the application of a binary non-linear 70 min gradient from 5 to 50% acetonitrile in 0.1% acetic acid at a flow rate of 300 nl/min. The LC was coupled online to an LTQ Velos Orbitrap mass spectrometer (Thermo Fisher, Germany) at a spray voltage of 2.4 kV. After a survey scan in the Orbitrap (R = 30,000), MS/MS data were recorded for the twenty most intensive precursor ions in the linear ion trap. Singly charged ions were not taken into account for MS/MS analysis. The lock mass option was enabled throughout all analyses.

Tryptic digests of the membrane samples were analyzed the same way as the cytosolic samples with the modification that the LC was coupled online to an LTQ Orbitrap Classic mass spectrometer (Thermo Fisher, Germany) and that only the five most intensive precursor ions were chosen for MS/MS analysis in the linear ion trap.

After mass spectrometric measurement, MS data were converted into the mzXML format by msconvert (***Chambers et al., 2012***) and subsequently subjected to database searching via Sorcerer (SageN, Milptas, CA) using SEQUEST (Thermo Fisher Scientific, San Jose, CA; version 27, revision 11) without charge state deconvolution and deisotoping performed (details below).

Protein identification, validation, and quantitation

Two protein sequence databases were constructed to identify proteins. The first database—Reduced Database (RedDB)—contained a total of 52,546 protein sequences from *B. azoricus* and its symbionts. Of these, 7714 sequences were from the SOX symbiont, 5955 sequences from the MOX symbiont, and 38,877 from the host.

The second database—Incremented Database (IncDB)—contained a total of 242,947 protein sequences from *B. azoricus*, its symbionts, and from organisms phylogenetically related to the

symbionts and the host. 93,304 protein sequences were from relatives of the SOX symbiont, 26,622 from relatives of the MOX symbiont, and 123,021 from host-related bivalves. Redundant sequences were removed using the CD-hit-2D clustering algorithm (***Li and Godzik, 2006***) at 100% sequence clustering threshold. To determine false discovery rates (FDRs) of protein identification, both protein sequence databases were reversed and appended to the original databases as decoy sequences.

All MS/MS spectra from 1D-PAGE-LC-MS/MS experiments were searched against the two protein sequence databases using the SEQUEST (v.27, rev. 11) algorithm (***Eng et al., 1994***) with the following parameters: Parent Mass Tolerance, 0.0065 Da; Fragment Ion Tolerance, 1 Da; up to 2 missed cleavages allowed (internal lysine and arginine residues), fully tryptic peptides only, and oxidation of methionine as variable modification (+15.99 Da).

Scaffold (version 4.0.6.1, Proteome Software Inc., Portland, OR) was used for filtering and analysis of protein identifications. For the searches against the IncDB, protein identifications were filtered with the PeptideProphet and ProteinProphet algorithms implemented in Scaffold (***Ma et al., 2012***) using the following thresholds: 95% for peptides and 99% for proteins. For the searches against the RedDB, protein identifications were filtered at the peptide level using SEQUEST scores (XCorrs of at least 2.5 (charge +2) and 3.5 (charges >+2), DeltaCN >0.08) and by requiring that they contained at least two identified peptides.

Protein level FDRs for all database searches were determined according to the method of ***Käll et al. (2008)*** using the number of identified decoy sequences that passed the filtering thresholds. FDRs for searches against the IncDB were all <1.7% and for the RedDB <0.2%.

For relative quantitation of proteins, normalized spectral abundance factor values, which give the relative abundance of a protein in a sample in %, were calculated for each sample according to the method of ***Florens et al. (2006)*** (***Supplementary file 1I***).

Supplementary Figures

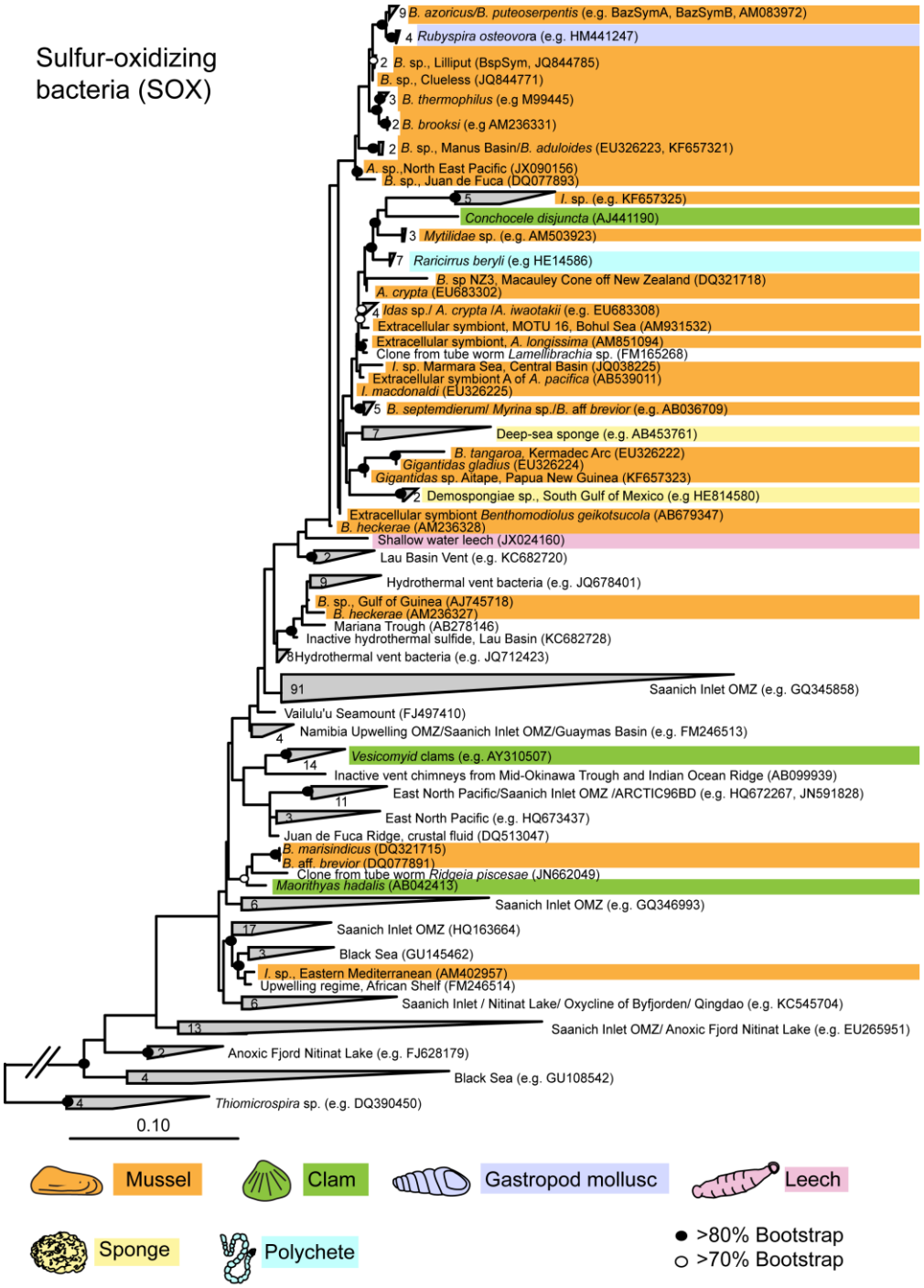


Figure 2—figure supplement 1. Maximum likelihood 16S rRNA phylogeny of the close relatives of the *Bathymodiolus* SOX symbionts. The tree was estimated from an alignment of 1653 nucleotide positions and was rooted with four sequences from *Thiomicrospira* species. The number of sequences per collapsed group is shown next to the gray blocks. Diagonal lines in the out-group branch indicate that the branch is not to scale. *B.* = *Bathymodiolus*; *A.* = *Adipicola; I.* = *Idas*.

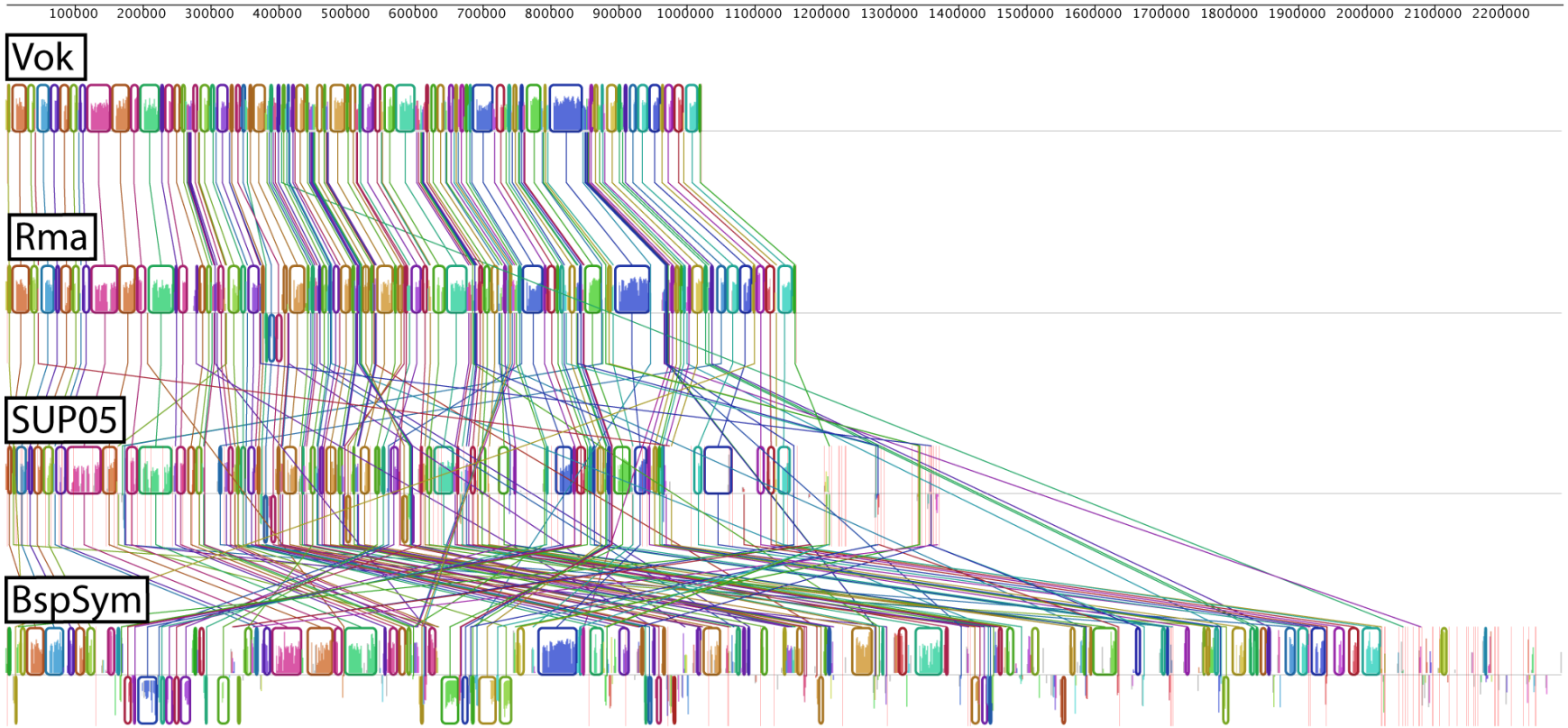


Figure 2—figure supplement 2. Whole genome alignment. Each colored block is a region of the genome that aligned to part of another genome because it is homologous and the genes are arranged in the same order. Lines crossing represent conflicting information when compared to other genomes. These are the sites where lack of synteny was observed. Red vertical lines represent contig boundaries. BspSym = SOX symbiont of *Bathymodiolus* sp., Vok = SOX symbiont *Candidatus Vesicomyosocious okutanii*, Rma = SOX symbiont of *Calyptogena magnifica* (*Ca.* Ruthia magnifica), SUP05 = free-living marine sulfur oxidizers

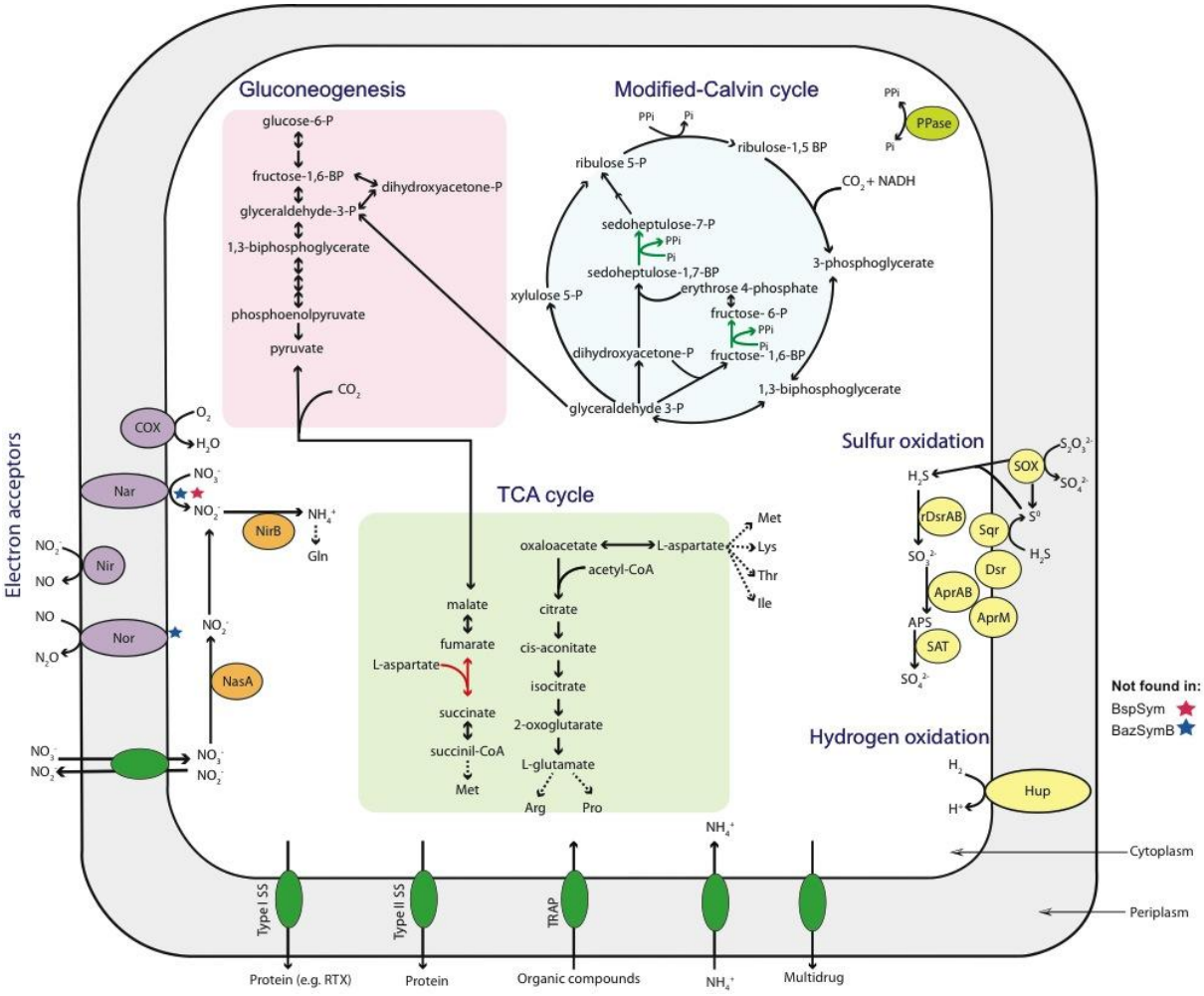


Figure 2—figure supplement 3. Metabolic reconstruction of the *Bathymodiolus* symbiont. Key metabolic pathways were inferred from genomic information using Pathway Tools (Caspi *et al*., 2014). Red stars indicate that the gene was not found in the *B.* sp. symbiont genome, and blue stars indicate that the gene was not found in BazSymB, but was found in BazSymA, both symbionts of *B. azoricus*. Red arrow indicates a missing enzyme that could be replaced with an alternative reaction. Green arrow indicates an inorganic pyrophosphate-dependent step in the modified version of the Calvin cycle. Nar = nitrate reductase; Nir = nitrite reductase; Nor = nitric oxide reductase; Hup = membrane-bound hydrogenase; SOX = sulfur oxidation; rDSR = reverse dissimilatory sulfite reductase; Sqr = sulfide-quinone reductase; Apr = adenylsulfate reductase; SAT = sulfate adenyltransferase; P = phosphate; BP = biphosphate; COX = cytochrome c oxidase; Gln = glutamine; Arg = arginine; Pro = proline; Met = methionine; Lys = lysine; Thr = threonine; Ile = isoleucine; PPi = inorganic pyrophosphate; PPase = soluble pyrophosphatase; SS = secretion system.

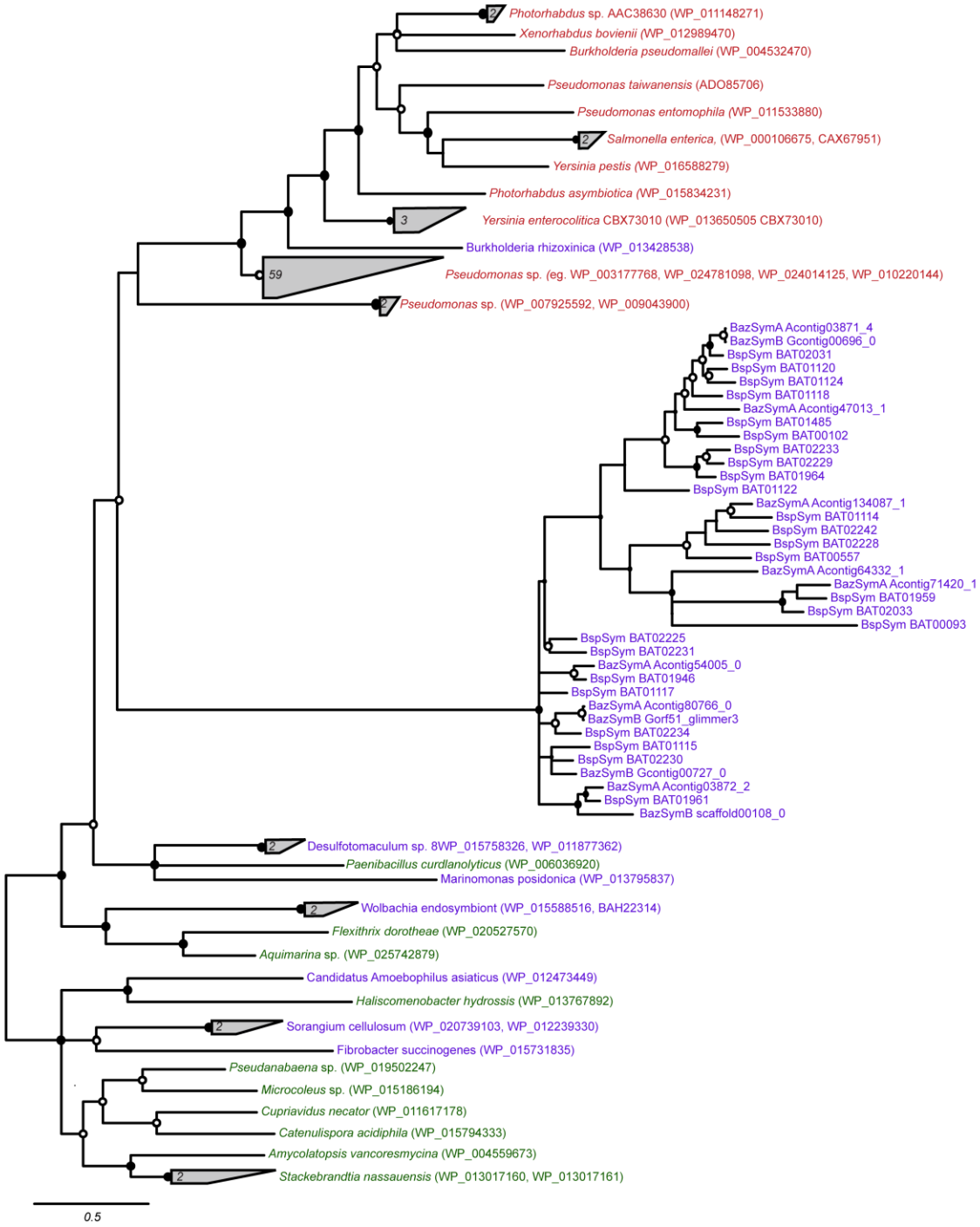


Figure 3—figure supplement 1. Consensus of bayesian and maximum likelihood phylogeny of YD proteins with identifiers. Trees were estimated from an alignment of 536 amino acids. Circles represent branches with posterior probability higher than 0.8 and bootstrap values higher than 80/100. If both reconstruction methods are significant, the circle is black, otherwise it is white. Purple: found in intestinal microflora or in close association with other organisms; green: free-living; red: pathogen.

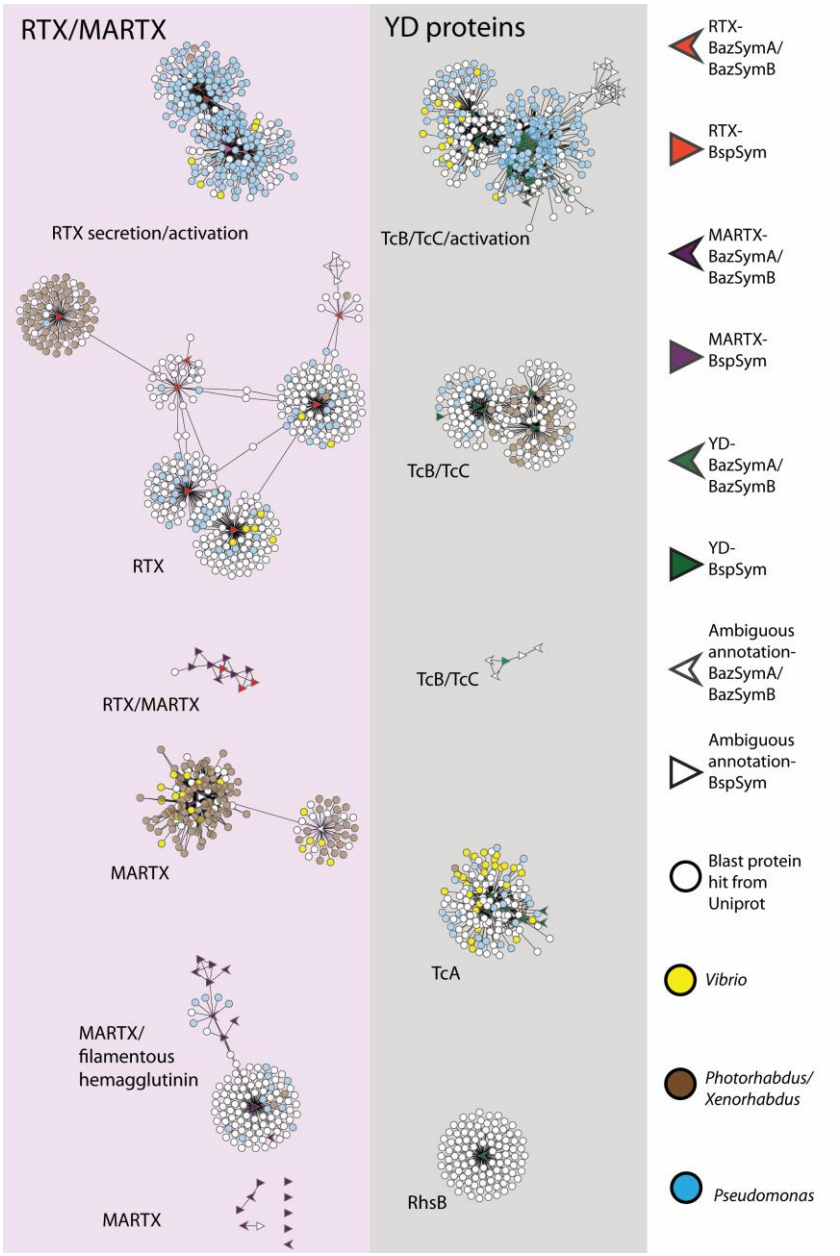


Figure 4—figure supplement 1. Network of toxin-related proteins in the *Bathymodiolus* symbionts with BLAST hits from *Vibrio*, *Photorhabdus*, *Xenorhabdus*, and *Pseudomonas* highlighted. Each node corresponds to a protein sequence and the links between nodes represent BLAST hits. The length of the link is proportional to the sequence similarity. Protein clusters containing RTX or MARTX are shown in the red panel on the left. Sequence clusters containing YD repeats are shown in the gray panel on the right. Arrowheads are proteins from *B. azoricus* symbionts, and triangles are proteins from *B.* sp. symbionts. The symbols are colored in green if they could be identified in the Bathymodiolus symbionts as YD repeat-containing proteins, red if they could be identified as RTX proteins, and purple for MARTX. If the clusters contained mostly proteins with a particular annotation, we named the clusters after these annotations, for example, cluster 'TcB/TcC' contained proteins annotated as TcB or TcC.

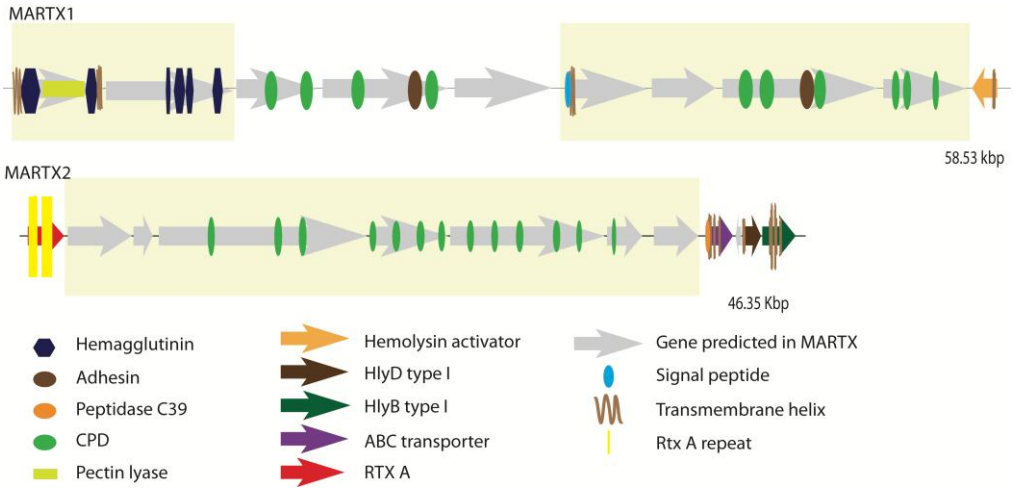


Figure 4—figure supplement 2. Genomic architecture of MARTX regions. The two MARTX regions in BspSym are shown. Operons identified by assembling transcriptome data are indicated in yellow boxes.

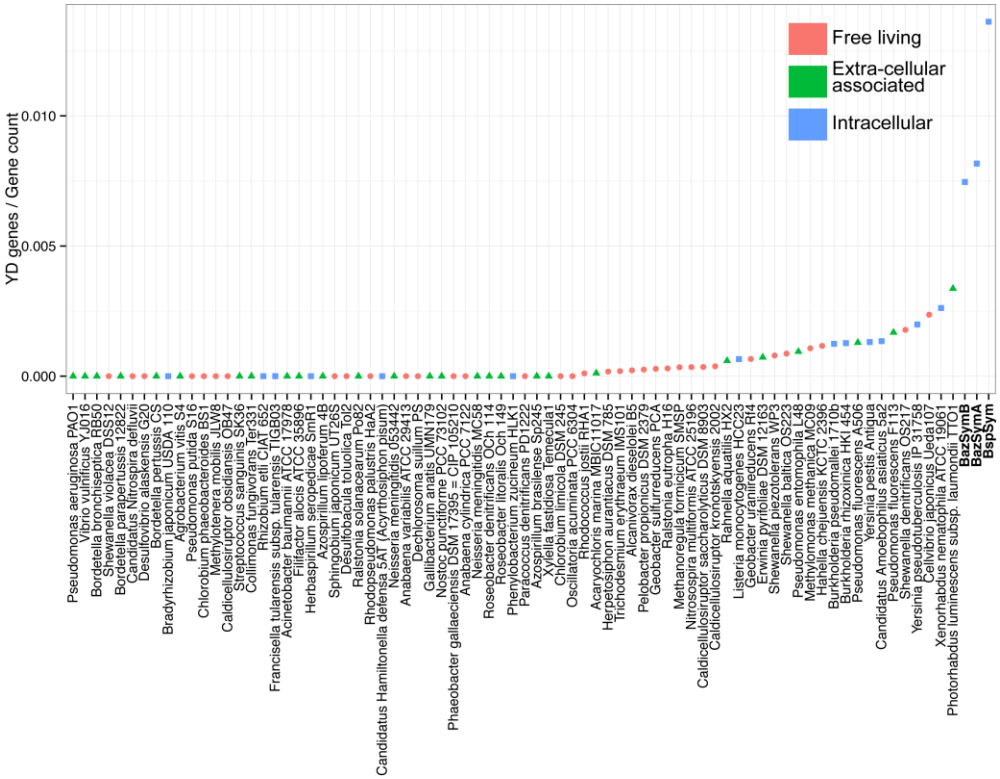


Figure 5—figure supplement 1. YD genes per genome, normalized to the total gene count. Each dot is colored by the category to which they belong. *Bathymodiolus* SOX symbionts are highlighted.

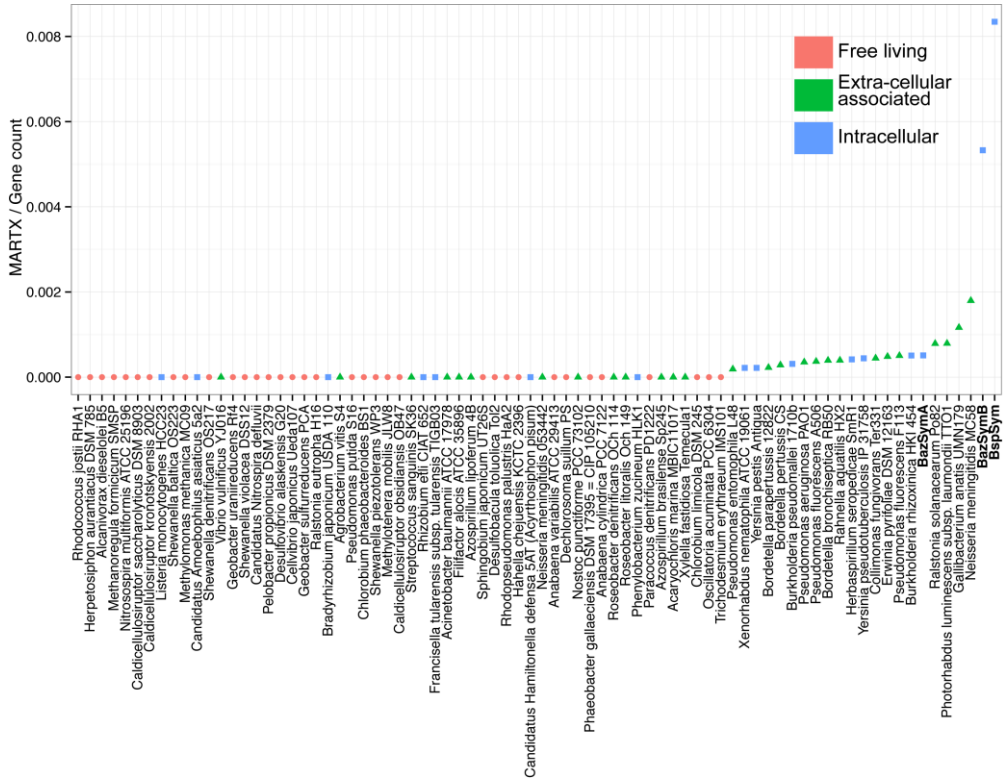


Figure 5—figure supplement 2. MARTX genes per genome, normalized to the total gene count. Each dot is colored by the category to which they belong. *Bathymodiolus* SOX symbionts are highlighted.

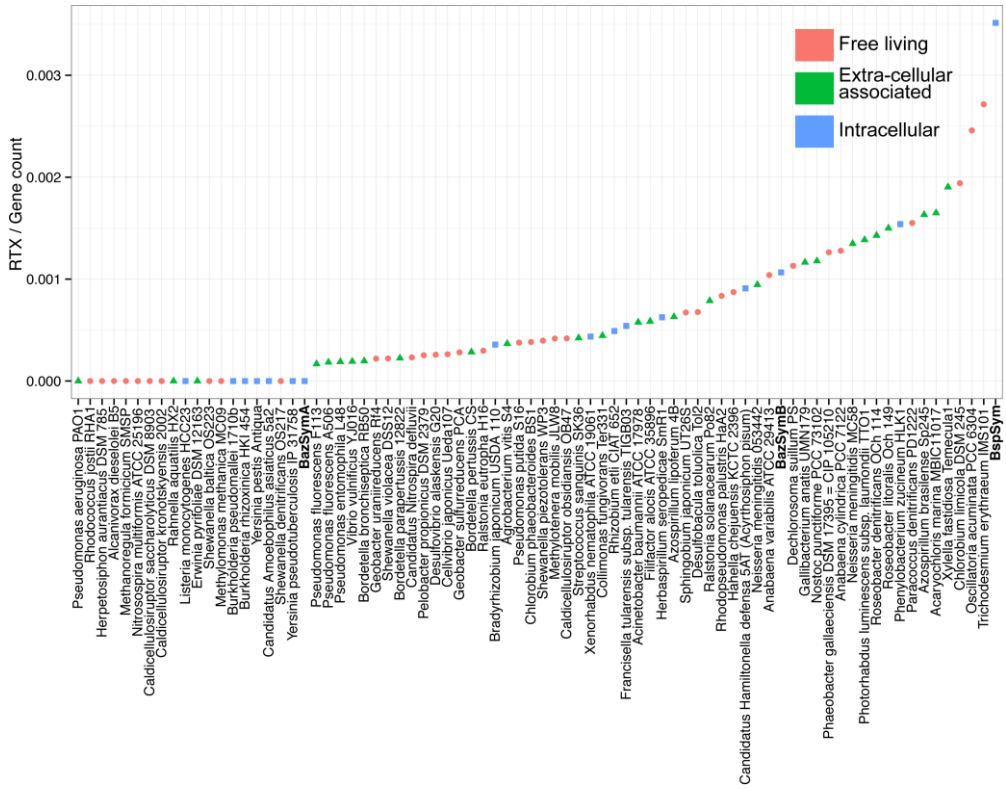


Figure 5—figure supplement 3. RTX genes per genome, normalized to the total gene count. Each dot is colored by the category to which they belong. *Bathymodiolus* SOX symbionts are highlighted.

Supplementary File 1

Supplementary File 1A. Number of mobile elements in the genomes compared in this study

Genome	# Transposases	# Integrases	# Restriction-modification systems
BspSym	17	5	18
BazSymA	23	4	22
BazSymB	13	3	10
SUP05[*]	14	1	1
***Candidatus* Ruthia magnifica**[#]	0	0	0
***Candidatus* Vesicomyosocius okutanii**[+]	0	0	0

[*](Walsh *et al*., 2009)
[#](Newton et al., 2007)
[+](Kuwahara et al., 2007)

Supplementary Table1B. Genomes with TRGs similar to those of the SOX symbionts of *Bathymodiolus*. The number of genes per TRGs class is shown.

Genome	Class	YD #	RTX#	MARTX #	NCBI Taxon ID	Gene count	Lifestyle -		
Acinetobacter baumannii ATCC 17978	Gammaproteobacteria	0	2	0	400667	3464	Ext	P	*
Agrobacterium vitis S4	Alphaproteobacteria	0	2	0	311402	5455	Ext	P	*
Anabaena cylindrica PCC 7122	Cyanophyceae	0	8	0	272123	6258	FL	NP	*
Anabaena variabilis ATCC 29413	Cyanophyceae	0	6	0	240292	5772	FL	NP	*
Azospirillum brasilense Sp245	Alphaproteobacteria	0	13	0	192	7962	Ext	NP	*
Azospirillum lipoferum 4B	Alphaproteobacteria	0	4	0	862719	6349	Ext	NP	*
Bordetella bronchiseptica RB50	Betaproteobacteria	0	1	2	257310	5086	Ext	P	*
Bordetella parapertussis 12822	Betaproteobacteria	0	1	1	257311	4447	Ext	P	*
Bordetella pertussis CS	Betaproteobacteria	0	1	1	1017264	3516	Ext	P	*
Bradyrhizobium japonicum USDA 110	Alphaproteobacteria	0	3	0	224911	8402	Int	NP	*
Caldicellulosiruptor obsidiansis OB47	Clostridia	0	1	0	608506	2389	FL	NP	*
Candidatus Hamiltonella defensa 5AT (Acyrthosiphon pisum)	Gammaproteobacteria	0	2	0	572265	2200	Int	P	
Candidatus Nitrospira defluvii	Nitrospira	0	1	0	330214	4317	FL	NP	
Chlorobium limicola DSM 245	Chlorobia	0	5	0	290315	2576	FL	NP	
Chlorobium phaeobacteroides BS1	Chlorobia	0	1	0	331678	2611	FL	NP	
Collimonas fungivorans Ter331	Betaproteobacteria	0	2	2	1005048	4493	Ext	NP	
Dechlorosoma suillum PS	Betaproteobacteria	0	4	0	640081	3539	FL	NP	
Desulfobacula toluolica Tol2	Deltaproteobacteria	0	3	0	651182	4435	FL	NP	
Desulfovibrio alaskensis G20	Deltaproteobacteria	0	1	0	207559	3874	FL	NP	
Filifactor alocis ATCC	Clostridia	0	1	0	546269	1709	Ext	NP	*

Genome	Class	YD #	RTX#	MARTX #	NCBI Taxon ID	Gene count	Lifestyle -		
35896									
Francisella tularensis subsp. tularensis TIGB03	Gammaproteobacteria	0	1	0	1001542	1850	Int	P	
Gallibacterium anatis UMN179	Gammaproteobacteria	0	3	3	1005058	2576	Ext	P	
Herbaspirillum seropedicae SmR1	Betaproteobacteria	0	3	2	757424	4799	Int	NP	*
Methylotenera mobilis JLW8	Betaproteobacteria	0	1	0	583345	2400	FL	NP	*
Neisseria meningitidis 053442	Betaproteobacteria	0	2	0	374833	2116	Ext	P	*
Neisseria meningitidis MC58	Betaproteobacteria	0	3	4	122586	2226	Ext	P	*
Nostoc punctiforme PCC 73102	unclassified	0	8	0	63737	6791	Ext	NP	*
Oscillatoria acuminata PCC 6304	unclassified	0	15	0	56110	6101	FL	NP	*
Paracoccus denitrificans PD1222	Alphaproteobacteria	0	8	0	318586	5158	FL	NP	*
Phaeobacter gallaeciensis DSM 17395 = CIP 105210	Alphaproteobacteria	0	5	0	391619	3960	FL	NP	*
Phenylobacterium zucineum HLK1	Alphaproteobacteria	0	6	0	450851	3899	Int	P	
Pseudomonas aeruginosa PAO1	Gammaproteobacteria	0	0	2	208964	5671	Ext	P	*
Pseudomonas putida S16	Gammaproteobacteria	0	2	0	1042876	5307	FL	NP	*
Ralstonia solanacearum Po82	Betaproteobacteria	0	4	4	1031711	5080	Ext	P	*
Rhizobium etli CIAT 652	Alphaproteobacteria	0	3	0	491916	6116	Int	NP	*
Rhodopseudomonas palustris HaA2	Alphaproteobacteria	0	4	0	316058	4788	FL	NP	*
Roseobacter denitrificans OCh 114	Alphaproteobacteria	0	6	0	375451	4201	Ext	NP	*
Roseobacter litoralis Och 149	Alphaproteobacteria	0	7	0	391595	4668	Ext	NP	*
Shewanella violacea DSS12	Gammaproteobacteria	0	1	0	637905	4515	FL	NP	
Sphingobium japonicum	Alphaproteobacteria	0	3	0	452662	4460	FL	NP	

Genome	Class	YD #	RTX#	MARTX #	NCBI Taxon ID	Gene count	Lifestyle -		
UT26S									
Streptococcus sanguinis SK36	Bacilli	0	1	0	388919	2367	Ext	P	*
Vibrio vulnificus YJ016	Gammaproteobacteria	0	1	0	196600	5202	Ext	P	*
Xylella fastidiosa Temecula1	Gammaproteobacteria	0	4	0	183190	2102	Ext	P	*
Rhodococcus jostii RHA1	Actinobacteria	1	0	0	101510	9242	FL	NP	
Acaryochloris marina MBIC11017	unclassified	1	14	0	329726	8488	Ext	NP	*
Herpetosiphon aurantiacus DSM 785	Chloroflexi	1	0	0	316274	5654	FL	NP	
Trichodesmium erythraeum IMS101	unclassified	1	14	0	203124	5156	FL	NP	*
Alcanivorax dieselolei B5	Gammaproteobacteria	1	0	0	930169	4470	FL	NP	*
Pelobacter propionicus DSM 2379	Deltaproteobacteria	1	1	0	338966	3949	FL	NP	*
Geobacter sulfurreducens PCA	Deltaproteobacteria	1	1	0	243231	3552	FL	NP	*
Ralstonia eutropha H16	Betaproteobacteria	2	2	0	381666	6718	FL	NP	
Methanoregula formicicum SMSP	Methanomicrobia	1	0	0	593750	2924	FL	NP	*
Nitrosospira multiformis ATCC 25196	Betaproteobacteria	1	0	0	323848	2885	FL	NP	*
Caldicellulosiruptor saccharolyticus DSM 8903	Clostridia	1	0	0	351627	2834	FL	NP	
Caldicellulosiruptor kronotskyensis 2002	Clostridia	1	0	0	632348	2642	FL	NP	
Rahnella aquatilis HX2	Gammaproteobacteria	3	0	2	1151116	5060	Ext	NP	
Listeria monocytogenes HCC23	Bacilli	2	0	0	552536	3059	Int	P	
Geobacter uraniireducens Rf4	Deltaproteobacteria	3	1	0	351605	4542	FL	NP	*
Erwinia pyrifoliae DSM 12163	Gammaproteobacteria	3	0	2	644651	4134	Ext	NP	*
Shewanella piezotolerans WP3	Gammaproteobacteria	4	2	0	225849	5047	FL	NP	
Shewanella baltica OS223	Gammaproteobacteria	4	0	0	407976	4622	FL	NP	*

Genome	Class	YD #	RTX#	MARTX #	NCBI Taxon ID	Gene count	Lifestyle -		
Pseudomonas entomophila L48	Gammaproteobacteria	5	1	1	384676	5293	Ext	P	*
Methylomonas methanica MC09	Gammaproteobacteria	5	0	0	857087	4664	FL	NP	*
Hahella chejuensis KCTC 2396	Gammaproteobacteria	8	6	0	349521	6875	FL	NP	
Burkholderia pseudomallei 1710b	Betaproteobacteria	8	0	2	320372	6436	Int	P	*
Burkholderia rhizoxinica HKI 454	Betaproteobacteria	5	0	2	882378	3938	Int	NP	*
Pseudomonas fluorescens A506	Gammaproteobacteria	7	1	2	1037911	5426	Ext	NP	*
Yersinia pestis Antiqua	Gammaproteobacteria	6	0	1	360102	4576	Int	P	*
Candidatus *Amoebophilus asiaticus* 5a2	unclassified	2	0	0	452471	1487	Int	P	
Pseudomonas fluorescens F113	Gammaproteobacteria	10	1	3	1114970	5952	Ext	NP	*
Shewanella denitrificans OS217	Gammaproteobacteria	7	0	0	318161	3935	FL	NP	
Yersinia pseudotuberculosis IP 31758	Gammaproteobacteria	9	0	2	349747	4529	Int	P	*
Cellvibrio japonicus Ueda107	Gammaproteobacteria	9	1	0	498211	3811	FL	NP	
Xenorhabdus nematophila ATCC 19061	Gammaproteobacteria	12	2	1	406817	4583	Int	P	*
Photorhabdus luminescens subsp. laumondii TTO1	Gammaproteobacteria	17	7	4	243265	5052	Ext	P	
BazSymB	Gammaproteobacteria	14	2	10	NA	1802	Int	NP	
BazSymA	Gammaproteobacteria	16	0	1	NA	2007	Int	NP	
BspSym	Gammaproteobacteria	33	8	19	174145	2225	Int	NP	

FL = free-living, Ext = extracellular host associated, Int = intracellular host associated, P = pathogen, NP = non-pathogen.

#The sum of the genes that had a significant BLAST hit against the YD, RTX and MARTX genes of the SOX symbionts.

*Found in biofilms.

Supplementary File 1C. P-values obtained with one-way Permanova were corrected with Bonferroni correction for multiple testing. Number of TRGs per genome was normalized to the total gene count

	Class	Order	Family
YD	0.3154	0.2731	0.0513
RTX	0.0812	0.0159*	0.214
MARTX	0.5479	0.3415	0.3093

Supplementary File 1D. Transcriptome counts of three individuals from B. azoricus and three individuals from B. sp. were mapped to their respective reference genomes with Rockhopper. Expression values of TRGs were normalized to the expression of RubisCO.

Genome	Identifier	Annotation	Class	Rep* 1	Rep* 2	Rep* 3
BazSymA	Acontig00027_1	YD repeat-containing protein	RHS	0.15	0.39	0.04
BazSymA	Acontig00030_0	Plasmid 28.1 kda A protein	RHS	0.07	0.11	0.03
BazSymA	Acontig03871_4	RHS repeat-associated core domain protein containing protein	RHS	0.19	0.71	0.05
BazSymA	Acontig03872_2	YD repeat-containing protein	RHS	0.47	1.69	0.19
BazSymA	Acontig104979_0	Virulence plasmid 28.1 kda A protein	RHS	0.00	0.03	0.00
BazSymA	Acontig134087_1	YD repeat-containing protein	RHS	0.14	0.74	0.05
BazSymA	Acontig21192_0	Peptidase C80 family [MARTX]	MARTX	0.61	1.81	0.17
BazSymA	Acontig47013_1	YD repeat-containing protein	RHS	0.53	1.83	0.14
BazSymA	Acontig64332_1	RHS repeat-associated core domain protein containing protein	RHS	0.32	1.27	0.08
BazSymA	Acontig71420_1	YD repeat-containing protein	RHS	0.28	0.97	0.07
BazSymA	Acontig80766_0	RHS repeat-associated core domain protein containing protein	RHS	0.25	0.87	0.05
BazSymA	Acontig88396_0	YD repeat-containing protein	RHS	0.41	1.75	0.07
BazSymA	Acontig211660_0	YD repeat-containing protein	RHS	0.44	1.55	0.15
BazSymA	Acontig134087_0	YD repeat-containing protein	RHS	0.35	1.68	0.10
BazSymA	Acontig54005_0	RHS repeat-associated core domain protein-containing protein	RHS	0.18	0.82	0.06
BazSymA	Acontig64332_0	Probable RHS	probable RHS	0.41	1.84	0.18
BazSymA	Acontig206100_1	RhsB protein	RHS	0.10	0.31	0.04
BazSymA	Acontig00027_2	Insecticidal toxin complex protein	RHS	0.07	0.14	0.04
BazSymA	Acontig96355_1	Probable RHS	probable RHS	0.24	1.10	0.04
BazSymA	Acontig32555_0	Hemolysin secretion/activation protein shlb family	RTX activator	0.11	0.50	0.04
BazSymB	Gcontig00696_0	YD repeat-containing protein	RHS	0.43	1.46	0.09
BazSymB	Gcontig00723_0	Probable RHS	probable RHS	1.20	3.78	0.26
BazSymB	Gcontig00727_0	Insecticidal toxin protein	RHS	1.37	4.39	0.25
BazSymB	scaffold00001_64	Conserved hypothetical protein [MARTX]	MARTX	0.57	1.55	0.15
BazSymB	scaffold00001_66	Filamentous hemagglutinin family N-terminal domain-containing protein	MARTX	1.03	2.80	0.24
BazSymB	scaffold00001_67	Hypothetical protein [MARTX]	MARTX	0.90	2.83	0.23
BazSymB	scaffold00001_68	Peptidase C80 family [MARTX]	MARTX	1.00	3.01	0.22
BazSymB	scaffold00001_69	Peptidase C80 family [MARTX]	MARTX	1.20	3.62	0.37

Genome	Identifier	Annotation	Class	Rep* 1	Rep* 2	Rep* 3
BazSymB	scaffold00001_70	Peptidase C80 family [MARTX]	MARTX	0.84	2.12	0.22
BazSymB	scaffold00007_28	Peptidase C80	MARTX	1.15	3.07	0.26
BazSymB	scaffold00007_30	Peptidase C80	MARTX	0.56	1.85	0.17
BazSymB	scaffold00033_1	YD repeat-containing protein	RHS	0.25	0.57	0.06
BazSymB	scaffold00033_4	Insecticidal toxin complex protein	RHS	0.27	0.33	0.14
BazSymB	scaffold00033_7	Virulence plasmid 28.1 kda A protein	RHS	0.19	0.29	0.06
BazSymB	scaffold00036_14	Hemolysin-type calcium-binding conserved site	RTX	0.23	0.70	0.04
BazSymB	scaffold00083_0	YD repeat-containing protein	RHS	1.33	2.34	0.15
BazSymB	scaffold00108_0	YD repeat-containing protein	RHS	0.64	2.04	0.15
BazSymB	scaffold00080_9	Peptidase C80	MARTX	0.17	0.43	0.05
BazSymB	scaffold00001_63	Hypothetical protein [MARTX]	MARTX	0.18	0.43	0.04
BazSymB	Gcontig00791_2	YD repeat-containing protein	RHS	0.77	2.78	0.25
BazSymB	Gorf51_glimmer3	YD repeat-containing protein	RHS	0.96	3.30	0.23
BazSymB	scaffold00033_6	Virulence plasmid 28.1 kda A protein	RHS	0.39	0.77	0.15
BazSymB	Gorf52_glimmer3	YD repeat-containing protein	RHS	0.48	1.84	0.15
BazSymB	scaffold00033_2	Insecticidal toxin complex protein	RHS	0.47	1.00	0.09
BazSymB	scaffold00033_3	Plasmid 28.1 kda A protein	RHS	0.16	0.26	0.06
BazSymB	scaffold00033_0	YD repeat-containing protein	RHS	0.37	1.01	0.05
BazSymB	Gcontig00849_0	Hemolysin type	RTX	1.25	4.18	0.42
BspSym	BAT00155	Hypothetical protein [MARTX]	MARTX	0.73	0.01	0.00
BspSym	BAT00157	Hypothetical protein [MARTX]	MARTX	0.58	0.01	0.00
BspSym	BAT00370	Hemolysin-type calcium-binding protein	RTX	0.50	0.02	0.02
BspSym	BAT00890-891	Rhs	RHS	0.78	0.06	0.04
BspSym	BAT00891	Probable RHS	Probable RHS	0.83	0.03	0.00
BspSym	BAT00892-894	Rhs	RHS	0.35	0.02	0.01
BspSym	BAT01108	Cell well associated rhsd protein precursor	RHS	0.47	0.09	0.11
BspSym	BAT01109	Toxin complex/plasmid virulence protein	RHS	0.51	0.07	0.13
BspSym	BAT01110	Insecticidal toxin complex protein	RHS	0.47	0.04	0.05
BspSym	BAT01111	Probable RHS	probable RHS	0.34	0.01	0.00
BspSym	BAT01113	Probable RHS	probable RHS	0.25	0.00	0.00
BspSym	BAT01114	Insecticidal toxin protein	RHS	0.41	0.00	0.00
BspSym	BAT01115	YD repeat-containing protein	RHS	0.65	0.01	0.00
BspSym	BAT01116	Hypothetical protein [RHS]	RHS	0.41	0.00	0.00

Genome	Identifier	Annotation	Class	Rep* 1	Rep* 2	Rep* 3
BspSym	BAT01117	YD repeat-containing protein	RHS	0.63	0.00	0.00
BspSym	BAT01118	YD repeat-containing protein	RHS	0.45	0.00	0.00
BspSym	BAT01120	YD repeat-containing protein	RHS	0.59	0.00	0.00
BspSym	BAT01122	Insecticidal toxin complex protein	RHS	0.35	0.00	0.00
BspSym	BAT01123	Probable RHS	probable RHS	0.62	0.00	0.00
BspSym	BAT01124	YD repeat-containing protein	RHS	0.59	0.01	0.01
BspSym	BAT01125	Probable RHS	probable RHS	1.48	0.01	0.00
BspSym	BAT01126	Probable RHS	probable RHS	0.56	0.01	0.00
BspSym	BAT01309	Probable RHS	probable RHS	0.44	0.01	0.00
BspSym	BAT01311	Probable RHS	probable RHS	0.35	0.01	0.01
BspSym	BAT01314	Probable RHS	probable RHS	0.38	0.01	0.00
BspSym	BAT01451	RTX toxin RtxA	MARTX	0.59	0.02	0.02
BspSym	BAT01452	Hypothetical protein [MARTX]	MARTX	0.49	0.01	0.00
BspSym	BAT01454	Peptidase C80	MARTX	0.17	0.01	0.01
BspSym	BAT01455	Peptidase C80	MARTX	0.28	0.01	0.01
BspSym	BAT01456	Peptidase C80	MARTX	0.32	0.01	0.01
BspSym	BAT01457	RTX toxin rtxa-like protein	MARTX	0.52	0.02	0.01
BspSym	BAT01458	Hypothetical protein	MARTX	1.29	0.07	0.01
BspSym	BAT01485	YD repeat-containing protein	RHS	0.35	0.00	0.00
BspSym	BAT01505	Rhs family protein	RHS	0.12	0.00	0.00
BspSym	BAT01658	Hemolysin-type calcium-binding region	RTX	0.69	0.04	0.05
BspSym	BAT01946	YD repeat-containing protein	RHS	0.47	0.01	0.00
BspSym	BAT01950	Probable RHS	probable RHS	0.36	0.00	0.01
BspSym	BAT01957	Probable RHS	probable RHS	0.35	0.00	0.00
BspSym	BAT01959	Tccc3	RHS	0.43	0.01	0.01
BspSym	BAT01964	Tccc4	RHS	0.46	0.00	0.00
BspSym	BAT02033	Insecticidal toxin complex-like protein	RHS	0.70	0.00	0.00
BspSym	BAT02222	Probable RHS	probable RHS	0.27	0.00	0.00
BspSym	BAT02223	Probable RHS	probable RHS	0.03	0.00	0.00

Genome	Identifier	Annotation	Class	Rep* 1	Rep* 2	Rep* 3
BspSym	BAT02225	Insecticidal toxin protein	RHS	0.20	0.00	0.00
BspSym	BAT02228	Insecticidal toxin protein	RHS	0.43	0.01	0.00
BspSym	BAT02229	Insecticidal toxin protein	RHS	0.28	0.00	0.00
BspSym	BAT02230	YD repeat-containing protein	RHS	0.68	0.01	0.00
BspSym	BAT02231	Insecticidal toxin protein	RHS	0.40	0.01	0.00
BspSym	BAT02232	Probable RHS	probable RHS	0.51	0.01	0.00
BspSym	BAT02232	YD repeat-containing protein	RHS	0.51	0.01	0.00
BspSym	BAT02233	YD repeat-containing protein	RHS	0.40	0.01	0.00
BspSym	BAT02234	Insecticidal toxin complex protein tccc (Toxin complex protein)	RHS	0.69	0.01	0.00
BspSym	BAT02235	Probable RHS	probable RHS	0.44	0.00	0.01
BspSym	BAT02238	Probable RHS	probable RHS	0.49	0.00	0.00
BspSym	BAT02239	Probable RHS	probable RHS	0.39	0.01	0.00
BspSym	BAT02240	Probable RHS	probable RHS	0.26	0.00	0.00
BspSym	BAT02242	Insecticidal toxin protein	RHS	0.52	0.00	0.00
BspSym	BAT02244	Probable RHS	probable RHS	0.26	0.00	0.00
BspSym	BAT00289-290	Rtxa	RTX	0.53	0.01	0.00
BspSym	BAT00052-EXT	RTX toxin	RTX	0.39	0.01	0.00
BspSym	BAT00290-end	Rtx	RTX	0.33	0.01	0.01
BspSym	BAT00101	Probable RHS	probable RHS	0.53	0.02	0.00
BspSym	BAT01112	Probable RHS	probable RHS	1.10	0.00	0.02
BspSym	BAT01459	RTX toxin RtxA	MARTX	0.48	0.02	0.03
BspSym	BAT01453	Hypothetical protein [MARTX]	MARTX	0.46	0.01	0.01
BspSym	BAT01127	Probable RHS	probable RHS	0.19	0.00	0.00
BspSym	BAT01128	Probable RHS	probable RHS	0.35	0.00	0.00
BspSym	BAT01129	Probable RHS	probable RHS	0.54	0.00	0.00
BspSym	BAT01130	Probable RHS	probable RHS	0.31	0.00	0.00
BspSym	BAT00154	Hypothetical protein [MARTX]	MARTX	0.51	0.01	0.00
BspSym	BAT00160	Hypothetical protein [MARTX]	MARTX	0.51	0.00	0.00

Genome	Identifier	Annotation	Class	Rep* 1	Rep* 2	Rep* 3
BspSym	BAT00158	Hypothetical protein [MARTX]	MARTX	0.43	0.00	0.00
BspSym	BAT00156	Hypothetical protein [MARTX]	MARTX	0.41	0.00	0.00
BspSym	BAT00152	Martx	MARTX	0.38	0.00	0.00
BspSym	BAT00153	Hemagglutination activity domain protein	MARTX	0.29	0.00	0.00
BspSym	BAT00159	Hypothetical protein [MARTX]	MARTX	0.27	0.01	0.00
BspSym	BAT01461	Hypothetical protein [MARTX]	MARTX	0.14	0.00	0.00
BspSym	BAT00161	Hemolysin activator protein	MARTX activator	0.25	0.00	0.00
BspSym	BAT02224	Probable RHS	probable RHS	0.22	0.00	0.00
BspSym	BAT01961	TccC4	RHS	0.65	0.01	0.00
BspSym	BAT00557	Insecticidal toxin protein	RHS	0.58	0.02	0.00
BspSym	BAT02236	Probable RHS	probable RHS	0.38	0.01	0.01
BspSym	BAT02243	Probable RHS	probable RHS	0.35	0.01	0.00
BspSym	BAT02031	Insecticidal toxin protein	RHS	0.42	0.00	0.00
BspSym	BAT00093	Insecticidal toxin complex-like protein	RHS	0.27	0.00	0.00
BspSym	BAT00102	Insecticidal toxin protein	RHS	0.13	0.00	0.00
BspSym	BAT00001	Hemolysin-type calcium binding protein	RTX	0.80	0.04	0.06
BspSym	BAT00371	Protein containing Type 1 secretion	RTX	0.45	0.02	0.03
BspSym	BAT00002	RTX toxin protein	RTX	0.22	0.02	0.02

#Rep = Replicate

Supplementary File 1E. Samples used in this study

Species	Cruise	Collection date	Site	Latitude	Longitude	Depth (m)	Sequencing center	Number of individuals	Purpose	Sequencing method	CO1 accession number
B. sp.	Marsued 5	2-May-09	Lilliput	09°32.82'S	13°12.56'W	1491	Genoscope (France)	1	Genomics*	454/ Illumina 36bp	-
B. sp.	M78-2	29-Apr-09	Lilliput	09°32.85'S	13°12.64'W	1489	Institute for Clinical Molecular Biology (Germany)	3	Transcriptomics	Illumina 100bp (PE[&])	LN833438 LN833439 LN833440
B. azoricus	RV Meteor cruise M82-3	10-Sep-10	Menez Gwen	37°50.68'N	31°31.17'W	835	Genoscope (France)	1	Genomics	454	LN833434
B. azoricus	RV Meteor cruise M82-3	10-Sep-10	Menez Gwen	37°50.68'N	31°31.17'W	835	Genoscope (France)	2	Proteomics	-	-
B. azoricus	MoMARETO	13-Aug-06	Menez Gwen	37°45.58N	31°38.26'W	840	OIST (Japan)	1	Genomics	454	LN833433
B. azoricus	Biobaz	9-Aug-13	Lucky Strike	37°16.98"N	32°16.54'W	1700	Max Planck Genome Centre (Germany)	3	Transcriptomics	Illumina 100bp	LN833435 LN833436 LN833437

*Longest scaffold published in Petersen *et al.* 2011

[#]Single host individuals were used for sequencing.

CO1 = cytochrome c oxidase subunit 1

[&]PE = Paired-end sequencing

Supplementary File 1F. Primer sequences and annealing temperatures used to detect genome rearrangements.

Primer pair	Forward	Reverse	Temp (°C)	Product size
BT1	TCATGCGCTGACTCGCCATT	TCGTTCTATGCAAGTGGCACGCT	62	1746*
BT4	TGCCCCAAATCCCAAGGTGC	TTGATGGATCCTCCTGCAATGGCT	62	1074*
BT5	TCCAGTCTGGCGCTCACAGG	TGCTGCTGACGCTTTTGGACAG	62	959
BT6	GGCCTAGCCCTATTTAGCAACCGGA	GCAAGTGCGTCAAAATCCACCGA	65	1199*

*The PCR product was of the expected size but not sequenced.

Supplementary File 1G. Metagenomes and metatranscriptomes enriched in SUP05 from oxygen minimum zones (OMZ) or hydrothermal vents.

Metagenome/ metatranscriptome	Accession number/Database	Class	Query	Best hit	E-value	% Identity	% Coverage
Juan de Fuca chimney (metagenome)	SRX012322	-	-	-	-	-	-
Lost City chimney biofilm (metagenome)	CAM_PROJ_Hydro-thermal Vent (CAMERA)	-	-	-	-	-	-
Lost City chimney (metagenome)	ACQI00000000	YD	BAT02242	ACQI01008516	7.00E-08	31	75
Guaymas basin plume (metagenome/metatranscriptome)	AJXC00000000	RTX	BAT00001	GB_4MN_MetaGALL_nosff_rep_c209590	6.85E-06	34	64
OMZ of East tropical South Pacific Chile (metagenome)	SRA025088	-	-	-	-	-	-
OMZ of East tropical South Pacific Chile (metagenome/metatranscriptome)	SRA023632	-	-	-	-	-	-

Supplementary File 1H. Amino acid sequences from the following genomes were used in the reference database for proteomic analysis (IncDB). The genomes belong to relatives of the sulfur-oxidizing (SOX) and methane-oxidizing (MOX) symbionts of *Bathymodiolus azoricus*, as well as the mussel host.

	Sequences of *Bathymodiolus*	**Sequences of phylogenetic relatives**	
		Symbiotic (*gamma-proteobacteria*)	***Free-living (*gamma-proteobacteria*)**
SOX	SOX sequences from available *Bathymodiolus azoricus* EST library (Bettencourt *et al*., 2010) SOX ORFs from draft genomes of *Bathymodiolus azoricus* (This study: BazSymA and BazSymB) SOX ORFs from draft genome of *Bathymodiolus sp.* from Lilliput, in the SouthMAR (This study: BspSym)	*Olavius algarvensis Gamma1* and *Gamma3 symbionts* (Kleiner *et al*., 2012) *Ca.* Endoriftia persephone and endosymbiont of *Tevnia jerichonana* (Gardebrecht *et al*., 2012) *Ca.* Ruthia magnifica str. Cm (Newton *et al*., 2007) *Ca.* Vesicomyosocius okutanii HA (Kuwahara *et al*., 2007)	*Acidithiobacillus ferrooxidans, Allochromatium vinosum DSM 180, Beggiatoa sp. SS, Beggiatoa sp. PS, Beggiatoa alba B18LD Caminibacter mediatlanticus TB-2, Halothiobacillus neapolitanus c2, ARCTIC020, ARCTICb15, SUP05, Magnetococcus marinus MC-1, Marichromatium purpuratum 984, Nitrosococcus halophilus Nc4, Nitrosococcus oceani ATCC 19707, Nitrosococcus watsonii C-113, Paracoccus denitrificans PD1222, Persephonella marina EX-H1, Starkeya novella* DSM 506, *Sulfurimonas gotlandica GD1, Sulfurimonas denitrificans* DSM 1251, *Sulfurimonas autotrophica* DSM 16294, *Thermovibrio ammonificans HB-1*, DSM 15698, *Thiobacillus denitrificans* ATCC 25259, *Thiocapsa marina* 5811, *Thiocystis violascens* DSM 198, *Thiomicrospira crunogena* XCL-2, *Thiorhodococcus drewsii* AZ1, *Thiorhodovibrio* sp. 970
MOX	MOX sequences from available *Bathymodiolus azoricus* EST library MOX ORFs from draft genome of *Bathymodiolus azoricus* (BazSymA) MOX ORFs from draft genome of *Bathymodiolus sp.* from SouthMAR (BspSym)	-	*Methylobacter tundripaludum SV96, Methylococcus capsulatus str. Bath, Methylomicrobium alcaliphilum, Methylomonas methanica* MC09, all amino acid sequences available in NCBI under phylogenetic group *Methylococcaceae*
Host	Host sequences from available *Bathymodiolus azoricus* EST library with bacterial sequences removed.	*Mytilus edulis, Mytilus galloprovincialis, Mizuhopecten yessoensis*, all amino acid sequences available in NCBI under phylogenetic group Bivalvia.	

*These sequences were downloaded from available NCBI genome BioProjects.

Supplementary File 1I. Details of expressed toxin-related proteins identified with proteomics. The values are given in % NSAF, which is a normalized spectral abundance factor that gives the relative abundance of a protein in a sample in %.

Identifier	Category	Max. coverage (%)[#]	Gill A	Gill B	Foot A	Foot B	Sup. A	Sup. B	GP A	GP B	MGP	MGill
Host_EST_000107	YD	33	0.0082	0.0234	0.0000	0.0000	0.0040	0.0055	0.0188	0.0628	0.0318	0.0115
Host_EST_000115	YD	34	0.0117	0.0319	0.0000	0.0000	0.0041	0.0113	0.0299	0.0814	0.0325	0.0265
Host_EST_000248	YD	33	0.0019	0.0204	0.0000	0.0000	0.0000	0.0064	0.0170	0.0389	0.0000	0.0000
Host_EST_002123	YD	24	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0142	0.0361
Thio_BAZ_1943_contig360420_0	RTX (activator)	21	0.0037	0.0044	0.0000	0.0000	0.0090	0.0125	0.0189	0.0000	0.0000	0.0130
Tox_BAZ_119_contig00027_0	YD	18	0.0054	0.0035	0.0000	0.0000	0.0005	0.0000	0.0210	0.0333	0.0000	0.0086
Tox_BAZ_120_contig00027_1	YD	27	0.0215	0.0284	0.0027	0.0000	0.0028	0.0019	0.0246	0.0653	0.0358	0.0420
Tox_BAZ_1734_contig02141_2	RTX (transporter)	46	0.0063	0.0024	0.0137	0.0000	0.0000	0.0000	0.0055	0.0287	0.0556	0.0317
Tox_BAZ_2494_contig00030_0	YD	29	0.0119	0.0116	0.0006	0.0000	0.0017	0.0007	0.0077	0.0257	0.0551	0.0582
Tox_BAZ_3202_scaffold00038_7	RTX	8.7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0098	0.0071
Tox_BAZ_525_contig104979_0	YD	6.8	0.0000	0.0016	0.0000	0.0000	0.0000	0.0000	0.0017	0.0069	0.0000	0.0024
ToxAzor_892893	YD	2.8	0.0009	0.0020	0.0000	0.0000	0.0000	0.0000	0.0000	0.0011	0.0000	0.0000
ToxSMAR_1260BAT01109	YD	3.4	0.0020	0.0069	0.0000	0.0000	0.0007	0.0000	0.0015	0.0115	0.0080	0.0081
ToxSMAR_2052BAT01788, Thio_BAZ_1733_contig02141_1 or Thio_BAZ_2580_scaffold00010_8	RTX (transporter)	8	0.0023	0.0038	0.0000	0.0000	0.0000	0.0000	0.0045	0.0184	0.0000	0.0000
ToxSMAR893-894	YD	1.8	0.0000	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000	0.0017	0.0000	0.0000
ToxAzor_890891	YD	9.9	0.0041	0.0076	0.0000	0.0000	0.0016	0.0000	0.0088	0.0284	0.0153	0.0148

Sup. = supernatant; GP = Gradient Pellet; MGP = Membrane proteome of the gradient pellet; MGill = Membrane proteome of the gill.

[#]The highest coverage of the protein sequence by identified peptides in one sample.

References of supplementary file 1

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

Gardebrecht, A., Markert, S., Sievert, S.M., Felbeck, H., Thürmer, A., Albrecht, D., Wollherr, A., Kabisch, J., Bris, N.L., Lehmann, R., *et al*. (2012). Physiological homogeneity among the endosymbionts of *Riftia pachyptila* and *Tevnia jerichonana* revealed by proteogenomics. *ISME* J. **6**, 766–776.

Kleiner, M., Petersen, J.M., and Dubilier, N. (2012). Convergent and divergent evolution of metabolism in sulfur-oxidizing symbionts and the role of horizontal gene transfer. *Curr. Opin. Microbiol*. **15**, 621–631.

Kuwahara, H., Yoshida, T., Takaki, Y., Shimamura, S., Nishi, S., Harada, M., Matsuyama, K., Takishita, K., Kawato, M., Uematsu, K., *et al*. (2007). Reduced genome of the thioautotrophic intracellular symbiont in a deep-sea clam, *Calyptogena okutanii*. *Curr. Biol*. **17**, 881–886.

Newton, I.L.G., Woyke, T., Auchtung, T.A., Dilly, G.F., Dutton, R.J., Fisher, M.C., Fontanez, K.M., Lau, E., Stewart, F.J., Richardson, P.M., *et al*. (2007). The *Calyptogena magnifica* chemoautotrophic symbiont genome. *Science* **315**, 998–1000.

Walsh, D.A., Zaikova, E., Howes, C.G., Song, Y.C., Wright, J.J., Tringe, S.G., Tortell, P.D., and Hallam, S.J. (2009). Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. *Science* **326**, 578–582.

Data source of figure 6

Figure 6-Source data file. Variability in toxin-related genes encoded by the *Bathymodiolus* SOX symbionts

ID	SNPs	Length	N's	SNPs per gene	Annotation	Foreign
BAT00001	4	1413	947	8.6	Hemolysin-type calcium binding protein	*
BAT00002	17	3741	2570	14.5	RTX toxin protein	*
BAT00052-EXT	54	2169	0	24.9	RTX toxin	*
BAT00093	18	1029	0	17.5	Insecticidal toxin complex-like protein	
BAT00101	2	453	0	4.4	Conserved hypothetical protein (RHS)	
BAT00102	12	5193	3599	7.5	Insecticidal toxin protein	*
BAT00152	11	2634	824	6.1	[MARTX1] protein, filamentous hemagglutinin family	
BAT00153	14	10764	4693	2.3	[MARTX1] hemagglutination activity domain protein	*
BAT00154	33	4614	1458	10.5	[MARTX1] hypothetical protein	*
BAT00155	0	9093	4373	0	[MARTX1] hypothetical protein	*
BAT00156	0	4413	2859	0	[MARTX1] hypothetical protein	*
BAT00157	0	5013	2894	0	[MARTX1] hypothetical protein	*
BAT00158	10	4137	787	3	[MARTX1] hypothetical protein	*
BAT00159	36	9324	6760	14	[MARTX1] hypothetical protein	*
BAT00160	24	5361	0	4.5	[MARTX1] hypothetical protein	*
BAT00161	2	1710	0	1.2	Hemolysin activator protein, hlyb family	*
BAT00289-290	4	1536	3	2.6	Rtxa	*
BAT00290-end	199	3672	0	54.2	Rtx	*
BAT00370	2	1158	0	1.7	Hemolysin-type calcium-binding protein	*
BAT00371	11	1341	0	8.2	Protein containing Type 1 secretion, C-terminal domain (RTX)	
BAT00557	100	1071	0	93.4	Insecticidal toxin protein	
BAT00890-891	39	1371	49	29.5	Rhs	*
BAT00891	5	123	0	40.7	Probable RHS	*
BAT00892-894	93	4059	365	25.2	Rhs	*
BAT01108	3	2274	0	1.3	Cell well associated rhsd protein precursor	*
BAT01109	17	8529	0	2	Toxin complex/plasmid virulence protein	*
BAT01110	10	5343	0	1.9	Insecticidal toxin complex protein	*
BAT01111	2	2463	2177	7	Probable RHS	*
BAT01112	1	135	0	7.4	Probable RHS	
BAT01113	1	3273	3020	4	Probable RHS	*
BAT01114	2	1596	0	1.3	Insecticidal toxin complex protein	*
BAT01115	0	726	76	0	YD repeat-containing protein	*
BAT01116	0	759	0	0	Hypothetical protein (RHS)	*
BAT01117	0	1530	0	0	YD repeat-containing protein	*

ID	SNPs	Length	N's	SNPs per gene	Annotation	Foreign
BAT01118	0	1443	0	0	YD repeat-containing protein	*
BAT01120	0	1542	0	0	YD repeat-containing protein	*
BAT01122	0	1158	359	0	Insecticidal toxin complex protein	*
BAT01123	0	591	0	0	Probable RHS	*
BAT01124	0	1533	0	0	YD repeat-containing protein	*
BAT01125	0	852	659	0	Probable RHS	*
BAT01126	0	756	0	0	Probable RHS	*
BAT01127	0	105	0	0	Probable RHS	
BAT01128	4	204	0	19.6	Probable RHS	
BAT01129	1	210	0	4.8	Probable RHS	
BAT01130	0	186	0	0	Probable RHS	
BAT01309	0	495	0	0	Probable RHS	*
BAT01311	0	627	0	0	Probable RHS	*
BAT01314	0	801	0	0	Probable RHS	*
BAT01451	43	2844	1594	34.4	[MARTX2] RTX toxin RtxA	*
BAT01452	39	3396	0	11.5	[MARTX2] Hypothetical protein	*
BAT01453	8	2862	1899	8.3	[MARTX2] Hypothetical protein	
BAT01454	39	11352	7602	10.4	[MARTX2] Peptidase C80, RTX self-cleaving toxin	*
BAT01455	51	4704	0	10.8	[MARTX2] Peptidase C80, RTX self-cleaving toxin	*
BAT01456	23	9663	3136	3.5	[MARTX2] Peptidase C80, RTX self-cleaving toxin	*
BAT01457	7	2286	646	4.3	[MARTX2] RTX toxin rtxa-like protein	*
BAT01458	5	363	0	13.8	[MARTX2] hypothetical protein	*
BAT01459	9	3069	2035	8.7	[MARTX2] RTX toxin rtxa	
BAT01461	0	129	0	0	[MARTX2] Hypothetical protein	
BAT01485	50	1515	0	33	YD repeat-containing protein	*
BAT01505	3	6690	4118	1.2	Rhs family protein	*
BAT01658	29	4743	0	6.1	Hemolysin-type calcium-binding region	*
BAT01946	37	1488	0	24.9	YD repeat-containing protein	*
BAT01950	2	819	0	2.4	Probable RHS	*
BAT01957	0	831	0	0	Probable RHS	*
BAT01959	0	1284	0	0	Tccc3	*
BAT01961	0	1140	0	0	Tccc4	
BAT01964	19	1206	0	15.8	Tccc4	*
BAT02031	22	1032	0	21.3	Insecticidal toxin complex protein	
BAT02033	117	1044	0	112.1	Insecticidal toxin complex-like protein	*
BAT02222	3	813	0	3.7	Probable RHS	*
BAT02223	4	4041	3881	25	Probable RHS	*
BAT02224	18	585	0	30.8	Probable RHS	

ID	SNPs	Length	N's	SNPs per gene	Annotation	Foreign
BAT02225	13	5763	3898	7	Insecticidal toxin complex protein	*
BAT02228	0	3273	2056	0	Insecticidal toxin complex protein	*
BAT02229	0	1215	0	0	Insecticidal toxin complex protein	*
BAT02230	0	1545	0	0	YD repeat-containing protein	*
BAT02231	0	3264	2097	0	Insecticidal toxin complex protein	*
BAT02232	0	1197	0	0	Insecticidal toxin complex protein TccC3	*
BAT02233	0	1521	0	0	YD repeat-containing protein	*
BAT02234	0	1548	0	0	Insecticidal toxin complex protein TccC (Toxin complex protein)	*
BAT02235	0	240	20	0	Probable RHS	*
BAT02236	0	351	0	0	Probable RHS	
BAT02238	0	825	0	0	Probable RHS	*
BAT02239	0	1578	1462	0	Probable RHS	*
BAT02240	0	3000	2266	0	Probable RHS	*
BAT02242	0	492	0	0	Insecticidal toxin complex protein	*
BAT02243	0	354	265	0	Probable RHS	
BAT02244	2	810	0	2.5	Probable RHS	
Gcontig00696_0	0	1023	0	0	YD repeat-containing protein	*
Gcontig00723_0	9	702	0	12.8	Sugar-binding protein (RHS)	*
Gcontig00727_0	27	660	0	40.9	Insecticidal toxin complex protein	*
Gcontig00791_2	3	465	0	6.5	YD repeat-containing protein partial	
Gcontig00849_0	10	351	0	28.5	Outer membrane channel lipoprotein (RTX)	
Gorf51_glimmer3	13	1548	0	8.4	YD repeat-containing protein	
Gorf52_glimmer3	6	1548	766	7.7	YD repeat-containing protein	
scaffold00001_63	3	2535	1494	2.9	Hypothetical protein [MARTX]	
scaffold00001_64	2	1485	0	1.3	Conserved hypothetical protein [MARTX]	*
scaffold00001_66	37	5430	20	6.8	Filamentous hemagglutinin family, N-terminal domain-containing protein [MARTX]	*
scaffold00001_67	1	744	20	1.4	Hypothetical protein [MARTX]	*
scaffold00001_68	29	2796	177	11.1	Peptidase C80 family [MARTX]	*
scaffold00001_69	24	3690	335	7.2	Peptidase C80 family [MARTX]	*
scaffold00001_70	22	6405	31	3.5	Peptidase C80 family [MARTX]	*
scaffold00007_28	14	2349	54	6.1	Peptidase C80 family [MARTX]	*
scaffold00007_30	14	3105	51	4.6	Peptidase C80 family [MARTX]	*
scaffold00033_0	0	543	231	0	YD repeat-containing protein	
scaffold00033_1	2	3645	0	0.5	YD repeat-containing protein	*
scaffold00033_2	5	843	0	5.9	Insecticidal toxin complex protein	
scaffold00033_3	0	891	0	0	Plasmid 28.1 kda A protein	
scaffold00033_4	1	1953	20	0.5	Insecticidal toxin complex protein	*

ID	SNPs	Length	N's	SNPs per gene	Annotation	Foreign
scaffold00033_6	2	276	0	7.2	Virulence plasmid 28.1 kda A protein	
scaffold00033_7	2	834	0	2.4	Virulence plasmid 28.1 kda A protein	*
scaffold00036_14	0	1680	0	0	Hemolysin-type calcium-binding conserved site	*
scaffold00080_9	7	2841	725	3.3	Hypothetical protein [MARTX]	
scaffold00083_0	35	4236	401	9.1	YD repeat-containing protein partial	*
scaffold00108_0	29	1932	1032	32.2	YD repeat-containing protein	*
Acontig00027_1	8	5322	0	1.5	YD repeat-containing protein	*
Acontig00027_2	0	1659	0	0	Insecticidal toxin complex protein	
Acontig00030_0	4	4914	0	0.8	Plasmid 28.1 kda A protein	*
Acontig03871_4	0	1023	0	0	RHS repeat-associated core domain protein containing protein	*
Acontig03872_2	8	1053	0	7.6	YD repeat-containing protein	*
Acontig104979_0	0	1371	0	0	Virulence plasmid 28.1 kda A protein	*
Acontig134087_0	0	183	0	0	YD repeat-containing protein	
Acontig134087_1	3	1074	0	2.8	YD repeat-containing protein partial	*
Acontig206100_1	12	414	0	29	Rhsb protein	
Acontig211660_0	26	594	0	43.8	YD repeat-containing protein	
Acontig21192_0	34	2445	0	13.9	Peptidase C80(MARTX)	*
Acontig32555_0	3	1686	0	1.8	Hemolysin secretion/activation protein shlb family	*
Acontig47013_1	11	1545	0	7.1	YD repeat-containing protein	*
Acontig54005_0	3	1554	0	1.9	RHS repeat-associated core domain protein-containing protein	
Acontig64332_0	0	120	0	0	Conserved hypothetical protein (RHS)	
Acontig64332_1	4	1551	0	2.6	RHS repeat-associated core domain protein containing protein	*
Acontig71420_1	6	1257	0	4.8	YD repeat-containing protein	*
Acontig80766_0	4	1239	0	3.2	RHS repeat-associated core domain protein containing protein	*
Acontig88396_0	1	765	0	1.3	YD repeat-containing protein	*
Acontig96355_1	0	129	0	0	Hypothetical protein (RHS)	

*Genes with a different codon usage, most likely acquired through horizontal gene transfer

Chapter III:

Comparative genomic insights into the roles of toxin-related genes in beneficial bacteria and their acquisition by horizontal gene transfer

Lizbeth Sayavedra[1], Rebecca Ansorge[1], Nicole Dubilier[1], Jillian M. Petersen[1,2]

[1]Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany
[2]University of Vienna, Althanstrasse 14, 1090 Vienna, Austria

This manuscript is in preparation and has not been reviewed by all authors.

Author contributions

LS, ND, and JP conceived the study; **LS** analyzed the data; RA provided assembly of *B.* sp. 5° SMAR (Clueless) and proofread the manuscript; **LS** and JP wrote the paper.

Abstract

Bathymodioline mussels associate with intracellular sulfur-oxidizing (SOX) bacteria that provide them with nutrition. In the genomes of two closely-related SOX symbionts, we recently discovered abundant toxin-related genes (TRGs) belonging to three classes: YD repeats, RTX and MARTX. We hypothesized that the TRGs were acquired by horizontal gene transfer, and could be used for beneficial interactions between hosts and symbionts. We have now sequenced the genomes of SOX symbionts from six additional host species. Based on a well-supported phylogenomic reconstruction, we found that the symbionts belong to two separate lineages that established symbioses with these mussels independently. MARTX were common to all SOX symbionts, suggesting that they are essential for colonization of bathymodioline hosts. The RTX and YD classes were found in only one of these two symbiont lineages, and the YD class has undergone gene family expansion. The second symbiont lineage is missing RTX and YD genes, but has acquired a Type IV secretion system. This raises the possibility that distinct but closely related free-living bacteria can acquire different molecular mechanisms for interacting with the same group of animal hosts in convergent evolution.

Introduction

Specific associations between animals and beneficial bacteria are virtually universal (McFall-Ngai *et al.*, 2013). Many of these bacteria are acquired from the environment, but the mechanisms that underlie host-bacteria recognition, invasion of host tissues or cells, and maintenance of the associations, are largely unknown (Cossart and Sansonetti, 2004; Pel and Pieterse, 2013). In contrast, the molecular mechanisms pathogens use to interact with their hosts have been intensively studied (Sansonetti, 2002; Kaufmann and Dorhoi, 2016; Di Genova and Tonelli, 2016; Kendall and Sperandio, 2016). A number of pathogen-encoded proteins that interfere with host cell activity have been described and characterized as toxins (Aktories, 2011; Lang *et al.*, 2010; Huber *et al.*, 2016). With the advent of large-scale bacterial genome sequencing, toxin-related genes have been found in the genomes of many beneficial bacteria suggesting that pathogens and beneficial bacteria may use similar molecular mechanisms to interact with their hosts (Pérez-Brocal *et al.*, 2011; Moya *et al.*, 2008).

We recently discovered diverse and abundant TRGs in the genomes of the beneficial symbionts of *Bathymodiolus* mussels, which we hypothesized could be

involved in beneficial host-microbe interactions including host-symbiont communication and defense against parasites (Sayavedra *et al.*, 2015). *Bathymodiolus* mussels are one of the most abundant animals at deep-sea hydrothermal vents and cold seeps worldwide (Dubilier *et al.*, 2008; Van Dover, 2000). To thrive in these harsh and nutrient-poor environments, the mussels associate with intracellular chemosynthetic bacteria that provide them with nutrition (Dubilier *et al.*, 2008; Duperron *et al.*, 2009; Fisher *et al.*, 1987; Fujiwara *et al.*, 2000; Raggi *et al.*, 2013; Ponnudurai *et al.*, 2016). The SOX symbionts are acquired from the environment by each new host generation (Won *et al.*, 2003; DeChaine *et al.*, 2006), but like most beneficial intracellular bacteria, virtually nothing is known about the mechanisms the symbionts use to invade and survive within host cells.

In this study, we investigated the distribution of toxin-related genes (TRGs) in distinct lineages of *Bathymodiolus*-associated symbionts. We hypothesized that TRGs specific to all bathymodioline symbionts may be involved in direct host-microbe interactions such as recognition and invasion of host cells, which are essential processes in all intracellular symbioses. We previously hypothesized that the symbionts acquired the TRGs by horizontal gene transfer (Sayavedra *et al.*, 2015). We therefore also aimed to pinpoint when the TRGs were most likely acquired by the symbionts during their evolution.

Results and Discussion

Phylogenomics reveals two well-supported independent symbiont lineages

Previously, genome sequences were available from the symbionts of three bathymodioline species (Sayavedra *et al.*, 2015; Ikuta *et al.*, 2015). We sequenced and assembled the draft genomes of SOX symbionts from six additional mussel species from hydrothermal vents and cold seeps around the world (Table 1). The draft genomes sequenced in this study were between 96.55 to 98.52% complete and were sequenced to depths between 24X and 3600X. Our estimates of the complete genome sizes ranged from 1.96 to 2.82 Mbp, which is larger than the closed 1.47 Mbp genome of the *B. septemdierum* SOX symbiont (Ikuta *et al.*, 2015).

Previous 16S rRNA gene phylogenies showed that the SOX symbionts of bathymodioline mussels do not form a single monophyletic clade but are interspersed with free-living sulfur-oxidizing bacteria called 'SUP05' or 'ARCTIC' (Petersen *et al.*, 2012; Sayavedra *et al.*, 2015). However, the highly conserved 16S rRNA gene could not reliably resolve the phylogenetic relationships between these organisms. We therefore reconstructed a species tree with 19 orthologous protein-coding genes and the 16S rRNA gene from the *Bathymodiolus* SOX symbionts and their close relatives. Consistent with the 16S rRNA phylogenies, most sulfur-oxidizing symbionts of *Bathymodiolus* mussels clustered in a single well-supported clade ('Clade 1', Figure 1). So far, this clade contains exclusively host-associated bacteria (Figure 1) (Petersen *et al.*, 2012; Sayavedra *et al.*, 2015). The *B. heckerae* endosymbionts (BheckSOX) clustered in a separate well-supported clade, together with the free-living bacteria '*Candidatus* Thioglobus autotrophica' EF1 and SUP05 ('Clade 2', Figure 1). This indicates that at least two lineages of free-living sulfur-oxidizing bacteria evolved symbiotic relationships with

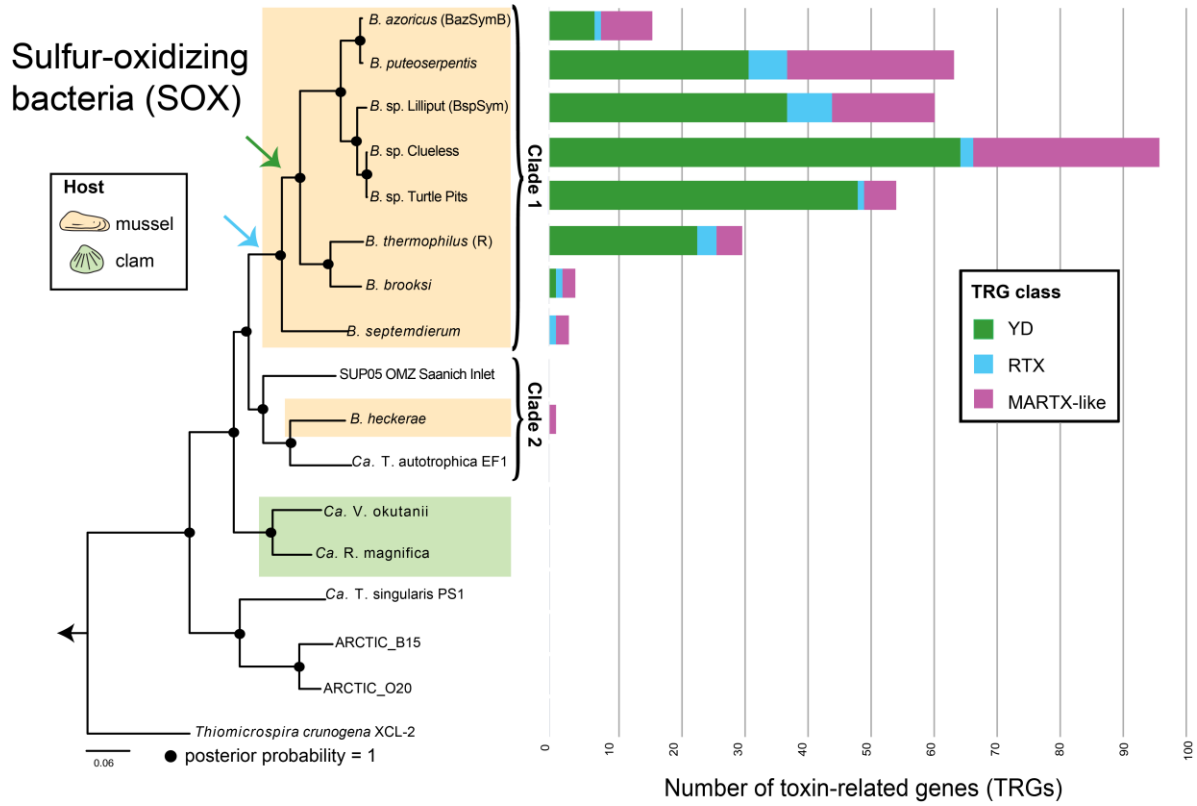


Figure 1. Species tree estimated with 19 orthologous genes and the 16S rRNA gene. Filled circles represent a posterior probability of one. The blue arrow indicates the point when RTX were acquired and the green arrow indicates when YD repeats were acquired. All genomes showed in this tree were searched for TRGs. T = Thioglobus; B = Symbiont of *Bathymodiolus*.

bathymodioline mussels independently. Having genome sequences of symbionts from both clades gave us the ideal opportunity to investigate how the genomic mechanisms underlying these evolutionary transitions from a free-living to a host-associated lifestyle might have evolved. Considering the hypothesized role of toxin-related genes in mediating host-microbe interactions, we focused on understanding their distribution among these closely-related free-living and symbiotic sulfur oxidizers.

Table 1. Overview of sulfur-oxidizing bacteria analyzed in this study. Completeness was estimated with CheckM (Parks *et al*., 2015).

Organism		Sampling site	Ecosystem	GC content (%)	Sequencing Depth	Completeness (%)	Genome Size (Mbp)	No. of scaffolds (>1000 bp)
Host	*B. azoricus*[1]	Menez Gwen, MAR	Vent	38.20	~8 X	90.60	1.66	239
	B. sp. 9° South, Lilliput[1]	9° South, Lilliput, MAR	Vent	38.23	~22 X	95.39	2.29	52
	B. thermophilus[+]	Crab-Spa, EPR	Vent	38.4	~24 X	97.86	2.25	149
	B. puteoserpentis[+]	Logatchev, MAR	Vent	37.67	~3600 X	97.7	2.19	77
	B. sp. 5° South, Clueless[+]	5° South, Clueless, MAR	Vent	37.81	~102 X	98.52	2.43	383
	B. sp. 5° South, Turtle Pits[+]	5° South, Turtle Pits	Vent	37.76	~226X	96.55	2.54	382
	B. heckerae[+]	Chapopote, GoM	Seep	37.41	~557 X	96.58	1.96	236
	B. brooksi[+]	Chapopote, GoM	Seep	36.63	~187 X	97.2	2.82	374
	B. septemdierum[2]	Myojin Knoll	Vent	38.74	~505X	98.68	1.47	1
	C. okutanni[3]	Sagami Bay	Seep	31.59	-	93.58	1.02	1
	C. magnifica[4]	9° N, EPR	Vent	34.03	-	94.84	1.16	1
Free-living	SUP05[5]	Saanich Inlet	OMZ	39.29	-	85.76	1.37	97
	Candidatus T. singularis PS1[6]	Puget Sound	Surface water (5 m)	37.44	-	98.68	1.71	1
	Candidatus T. autotrophica EF1[7]	Effingham Inlet	Redox gradient (60 m)	39.14	-	99.18	1.51	1
	Thiomicrospira crunogena XCL-2[8]	EPR	Vent	43.13	-	100	2.43	1

B., *Bathymodiolus*; *C.*, *Calyptogena*; T, Thioglobus; GoM, Gulf of Mexico; EPR, East Pacific Rise; MAR, Mid-Atlantic Ridge; OMZ, oxygen minimum zone.
[+]Sequenced in this study; [1](Sayavedra *et al*., 2015); [2](Ikuta *et al*., 2015); [3](Kuwahara *et al*., 2007); [4](Newton *et al*., 2007); [5](Walsh *et al*., 2009); [6](Marshall and Morris, 2015); [7](Shah and Morris, 2015); [8](Scott *et al*., 2006)

Horizontal acquisition and expansion of toxin-related gene families

In our previous study, we discovered TRGs from three toxin classes: 1) RTX, or 'repeat in toxin' proteins, 2) MARTX toxins, which are large proteins containing multiple repeat motifs and domains of diverse function, and 3) YD repeat toxins, named for their characteristic repeat sequence (Sayavedra *et al.*, 2015). We searched all of the genomes of interest for these toxin-related genes with BLAST against the TRGs previously described in Sayavedra *et al*. (2015). TRGs were classified into the three classes based on the presence of functional domains with the conserved domain database (CDD) (see Materials and methods). Consistent with our previous analysis, none of the genome sequences from free-living relatives or the obligate intracellular symbionts of clams encoded TRGs (Figure 1). Intriguingly, only one toxin class, MARTX, was found in all of the *Bathymodiolus* symbiont genomes, regardless of the Clade they belonged to. RTX (repeats in toxins) were found in all Clade 1 symbionts, but not in the genome of BheckSOX (See Materials and Methods). YD repeat proteins were found in all members of Clade 1 except the basal *B. septemdierum* symbiont, which is a closed genome (Figure 1). These observations support the following theories: 1) RTX and YD repeats are not essential for forming and maintaining an intracellular symbiotic association with *Bathymodiolus*, 2) RTX genes were acquired by the common ancestor of Clade 1, and 3) YD repeat genes were acquired by Clade 1 symbionts after divergence from the *B. septemdierum* symbiont, and are found in multiple copies, suggesting gene family expansion through duplication.

Considering that YD repeats do not seem essential for symbiosis with *Bathymodiolus* mussels, could their main function be in protecting the host against pathogens and parasites as we previously hypothesized (Sayavedra *et al*. 2015)? The need for symbiont protection would depend on multiple factors including host immunity and the abundance of parasites and pathogens in the environment. The apparent expansion of the YD repeat gene family in some Clade 1 lineages is consistent with a role in defense (Nuismer and Otto, 2004; Tellier *et al.*, 2014). If the YD repeat proteins are involved in defense, then symbiont lineages encoding these proteins may either experience more intense pressure from parasites and pathogens in the environment, or their hosts might have lost certain immune factors over evolution. For example, the marine invertebrates *Nematostella* and *Hydra* have lost part of the complement effector pathway (Miller *et al.*,

2007). Hosts with defensive symbionts can also become 'addicted' to symbiont defense, as selective pressure to maintain its own immunity is lower in the presence of the defensive symbiont (Martinez *et al.*, 2016). No complete genomes are available for *Bathymodiolus*, it is therefore currently unknown how complete their immune pathways are.

All symbionts from Clade 1 encoded RTX proteins but so far, the Clade 2 symbiont does not seem to encode proteins from this class. Initially, this seemed to support a role for these toxins in defense rather than essential host-symbiont crosstalk. However, there is an alternative explanation – RTX proteins might mediate host-symbiont interactions in Clade 1, and the Clade 2 symbiont, BheckSOX, independently evolved different molecular mechanisms to perform the functions fulfilled by the RTX proteins in Clade 1 symbionts. A 'communication channel' between host and symbionts is essential for the exchange of metabolites and other molecules such as protein effectors in an intracellular symbiosis. In *Bathymodiolus,* this would require transporters or pores in the vacuolar membranes that surround the symbionts (Fisher et al., 1987). RTX proteins have been shown to form pores in eukaryotic membranes (Lally et al., 1999; Linhartová et al., 2010; Los et al., 2013). If RTX proteins are used for establishing channels of communication with the host in Clade 1 symbionts through pore formation, how could this be achieved in the *B. heckerae* symbiosis? To answer this question, we searched for genes related to toxins or virulence factors that were specific to BheckSOX. Indeed, this symbiont encodes a type IV secretion system (T4SS) of the type VirB/D, which has not yet been found in any other bathymodioline mussel SOX symbionts (See Materials and Methods - Investigating the distribution of TRGs and T4SS in symbiont genomes). T4SSs of pathogens are known to be used for DNA transfer, host colonization, and sometimes the same T4SS can carry out both functions (Backert and Meyer, 2006; Dehio, 2008; Alvarez-Martinez and Christie, 2009). The BheckSOX VirB/D-T4SS shares a similar genomic architecture with systems used for conjugation (e.g. *Vibrio parahaemolyticus,* and the plasmids of *Edwardsiella tarda*, *Aeromonas punctata* and *Candidatus* Regiella insecticola), and for host cell invasion and persistence through secretion of toxic effectors (e.g. *Bartonella henselae* str. Houston-1 and *Salmonella enterica* sv. Typhimurium STm7) (Schmid *et al.*, 2004; Dehio, 2008; Seubert *et al.*, 2003; Gokulan *et al.,* 2013). The BheckSOX symbiont genome encodes a phage integrase upstream of the VirB/D T4SS gene cluster, raising the possibility that the symbiont acquired its T4SS from a bacteriophage. If the BheckSOX uses its T4SS to interact with

its host, this would be a remarkable example of convergent evolution in which free-living bacteria took unique evolutionary trajectories to become intracellular symbionts of animals, depending on the genes they acquired.

MARTX-like genes are enriched in the genomes of bacteria that associate with eukaryotes, and often have domains involved in attachment (Sayavedra *et al.*, 2015). We previously hypothesized that the MARTX-like proteins of the *Bathymodiolus* SOX symbiont could be used for host attachment or hijacking of host cell functions in a similar way to the characterized MARTX proteins of *Pseudomonas putida*, *P. fluorescens* and *Legionella pneumophila* (Satchell, 2011; Cirillo *et al.*, 2002). The specific presence of MARTX-like genes in all of the *Bathymodiolus* SOX symbionts, and not in their free-living or vertically-transmitted symbiotic relatives, is consistent with a role in essential host-symbiont interactions such as attachment and recognition.

The sequences and domain architecture of MARTX genes from *Bathymodiolus* symbionts were remarkably diverse. The repeated structure of the gene might help to create new domain architectures through homologous recombination (Satchell, 2011, 2007). Even though the domain architecture is highly variable, there were some common features. All *Bathymodiolus* symbionts had at least one MARTX-like gene with domains involved in attachment such as haemmagglutinin, VCBS repeat, FhaB, cadherin, and integrin, supporting their role in attachment to the host (Supplementary Table 1 and 2). The *Bathymodiolus* symbioses are clearly highly specific: one host species always associates with one or more symbiont 16S rRNA types, which are not found in any other host species (see Duperron *et al.*, 2006 for the only exception). This host specificity is strictly maintained even when two species co-occur, such as *B. brooksi* and *B. heckerae* at cold seeps in the Gulf of Mexico (Raggi *et al.*, 2013; Duperron *et al.*, 2007). If MARTX genes play a role in recognition, then the highly divergent sequence and domain architectures of the MARTX genes in different symbiont lineages might be one of the mechanisms that underlie this remarkable specificity.

To investigate whether the MARTX genes from Clades 1 and 2 had a common origin, or whether each lineage horizontally acquired MARTX genes independently from different sources, we compared the sequence and domain architecture of the SOX symbiont MARTX genes with those in public databases. The MARTX gene from BheckSOX was most similar to a MARTX gene from a Clade 1 *Bathymodiolus* SOX symbiont (51.5% similarity, 52% coverage), not to any other sequence available in public databases. Although we cannot rule out that a more similar MARTX gene is yet to be

sequenced, the similarity between Clade 1 and BheckSOX MARTX genes points toward a common origin for these genes. They might have been acquired from a common source in their environment, or conversely, this pattern may be the result of horizontal gene transfer between closely-related free-living and symbiotic sulfur-oxidizing bacteria. For example, the free-living ancestor of BheckSOX might have acquired a MARTX gene from a Clade 1 symbiont. Free-living SUP05 bacteria have been found in hydrothermal vent plumes in close proximity to *Bathymodiolus* mussels (Sunamura *et al.*, 2004; Anderson *et al.*, 2013; Anantharaman *et al.*, 2014; Sylvan *et al.*, 2012). Gene exchange between symbionts in a 'free-living' stage in the environment and obligate free-living SUP05 bacteria could lead to the evolution of novel symbiont lineages. Shared phages would be one mechanism for gene transfer, as both *Bathymodiolus* symbionts and SUP05 can be infected by phages (Roux *et al.*, 2014, our own unpublished data). Our results indicate that toxin-related genes might be acquired from distantly-related bacteria, but they may also be shuffled between symbiont lineages or between symbionts and close relatives in the environment.

Summary and Conclusions

Our comparative genomic analyses showed that MARTX toxin-like genes were common to all the *Bathymodiolus* SOX symbionts, but have not yet been found in any of the free-living or vertically transmitted symbiotic close relatives. These genes might therefore be essential 'symbiosis genes' used for recognition, attachment and symbiont uptake. Their remarkably variable sequence and domain structure is in stark contrast to the vast majority of proteins involved in core cellular functions, which are are highly conserved between symbionts. Thus, the highly variable MARTX are ideal candidate genes that might help to explain the highly specific nature of these symbioses. Out of the two lineages of *Bathymodiolus* SOX symbionts we identified, members of only one had TRGs belonging to the RTX and YD classes. We did not find RTX and YD repeat-encoding genes in the second lineage, but instead, we found a VirB/D-T4SS system, which in pathogens is responsible for secretion of toxic effectors. If the VirB/D-T4SS system of BheckSOX is indeed used to eject 'symbiotic effectors' to the host it may fulfill a function carried out by other TLG classes in other symbiont lineages. This would suggest that environmentally acquired symbionts can take independent evolutionary paths to be able to send molecular cues to the host. In another intriguing parallel between the

evolution of pathogens and beneficial symbionts, our study supports that just as for pathogens such as *P. aeruginosa* (Huber *et al.*, 2016), the genomes of beneficial bacteria require a patchwork of horizontally acquired genes to establish and maintain beneficial interactions with eukaryotic hosts.

Sample collection and DNA extraction

Sampling sites and coordinates of the mussels sequenced in this study are listed in Supplementary Table 1. The gills of *B.* sp from 5° SMAR (Turtle Pits) were dissected and homogenized in 1x PBS. The homogenate was centrifuged at 100 g for 10 min and the supernatant filtered through a series of GTTP filters: 12 µm, 5 µm, and 2 µm. The 2 µm cellulose acetate filter (Millipore, Germany) had an enriched fraction of SOX symbiont cells and was stored at -80°C. Gill tissues of *B. puteoserpentis, B.* sp from 5° SMAR (Clueless), *B. brooksi,* and *B. heckerae* were dissected and stored at -80°C. The gills of *Bathymodiolus thermophilus* were stored in 96% ethanol at -80°C. In the home laboratory, the gill tissue was transferred to innuSpeed Lysis tubes E (Analytic Jena, Germany) and homogenized in a FastPrep-24 Instrument (MP Biomedicals Europe) for 2 min at 4 m/s. DNA was extracted from the homogenates and the filter with cells as described by Zhou *et al*. (1996) or with the DNeasy blood and tissue kit (Qiagen, Germany) (Supplementary Table 1).

Sequencing, assembly and genome annotation

Genomic DNA libraries were generated with the Illumina TruSeq DNA Sample Prep Kit (BioLABS, Frankfurt am Main, Germany). Sequencing details are shown in Supplementary Table 1. Adaptors were removed and the reads were quality filtered (Q=2) with BBDuk V35 (Bushnell B. - sourceforge.net/projects/bbmap/). One initial assembly was done for each *Bathymodiolus* species with IDBA-UD (Peng *et al*., 2012). The assembly was binned with gbtools (Seah and Gruber-Vodicka, 2015) based on differential genome coverage, taxonomic affiliation and GC- content as described by Albertsen *et al*. (2013). The resulting genome bin was used as reference to map the reads from each individual with Bbmap using a percentage similarity of 98% with BBmap V35. The reads were used for re-assembly with Spades V. 3.6. The resulting genome was re-binned to remove possible contaminant contigs as described above. Contigs smaller than 800 bp were discarded from the final bins due to low confidence level (Bouck *et al.*, 1998; Luo *et*

al., 2012). The assembly of *B. puteoserpentis* SOX was gap-filled with GapFiller (Boetzer and Pirovano, 2012).

Investigating the distribution of TRGs and T4SS in symbiont genomes

Fragmented genomes obtained through binning methods naturally pose a challenge for comparative genomics and analyses of gene absence. To provide additional support for the observation that RTX and YD genes are absent in the SOX symbiont genome of *B. heckerae*, we scanned the entire metagenome assembly produced by IDBA-UD (Peng *et al.*, 2012) with Diamond (Buchfink *et al.*, 2014) using the collection of RTX and YD protein sequences of the SOX symbionts of Clade 1 as 'bait'. Because some contigs showed sequence similarity to the RTX and YD genes, we reassembled the metagenome with metaSpades V. 3.8, which outputs an assembly graph (Nurk *et al.*, 2016). This graph was visualized with Bandage (Wick *et al.*, 2015) to look for potential connections between the SOX symbiont genome and contigs with hits to YD and RTX. The assembly graph did not connect those contigs to the symbiont genome and the contigs had a different coverage compared to the main symbiont genome. Thus, we are confident that these contigs did not belong to the SOX symbiont of *B. heckerae* described in this study.

We investigated the Virb/D-T4SS as described above for the TRGs with the following modifications; as a query, we used the VirB/D-T4SS system from BheckSOX against the entire metagenome assemblies of the mussel samples that had symbionts from Clade 1. The VirB/D-T4SS gene cluster was curated with CONJdb (Guglielmini *et al.*, 2013). Similar genome architectures were searched with RAST and IMG (Aziz *et al.*, 2008; Markowitz *et al.*, 2011).

Phylogenetic analyses and identification of toxin-related genes

The draft genomes described in this study were annotated with RAST (Aziz *et al.*, 2008). The shared orthologous sequences were obtained with OrthoMCL (Li, 2003) implemented in Odose (Galaxy). The trimmed aligned sequences and the 16S rRNA gene were concatenated. Gene partitions were used to calculate a species tree with BEST (Liu, 2008), implemented in MrBayes version 3.2.6 (Ronquist and Huelsenbeck, 2003). We used two independent runs of 5,000,000 generations and two heated chains.

Candidate toxin-related genes were obtained with Blast searches of all protein sequences of the closest relatives against the TRGs described in the SOX symbiont genomes of *B. azoricus* and *B.* sp 9° South (Sayavedra *et al.*, 2015). We considered only those sequences that had a similarity higher than 25% and minimum query coverage of 25% or a minimum alignment of 200 amino acids. TRGs of the *Bathymodiolus* SOX symbionts are polymorphic systems – the domains that are present within the genes tend to vary in order and presence. The annotation of RTX and MARTX-like genes is therefore especially difficult. Thus, to avoid the difficulties of annotation, we designed a scoring script that predicts the super class of the TRG based on the domain combination present in the protein sequence and protein length. The protein sequences were submitted to the batch conserved domain database (CDD) website (Marchler-Bauer *et al.*, 2014). The resulting file was used as input for a self-written Perl script that assigned a score to the domains of each protein based on how often the domain was present in the classes YD, RTX or MARTX (see Supplementary Table 3 for the scores assigned to the domains observed in TRGs). If the class of the protein obtained based on the domain score system was different to the class of the best blast hit, then the annotation was checked manually. The presence of peptidase 80 or multiple functional domains was considered as a determinant factor for a protein to be part of the class MARTX-like. The domain scoring system has the drawback that genes that do not have any detectable domain will not be classified, but it allows a consistent and comparable method for identifying functional domains present in the TRGs. Only the proteins that had at least one functional domain were further considered (Figure 1).

To find proteins with similar domain content and structure to the MARTX-like protein of *B. heckerae* symbiont, we reconstructed a domain distance tree with the domains obtained with CDD using DoMosaics (Moore *et al.*, 2014). Overlapping domains were resolved based on e-value. Symbiont proteins that had similar domains to the MARTX-like protein of *B. heckerae* symbiont are shown in Supplementary Figure 1.

Acknowledgements

We thank the captains and crews of the sampling cruises AT16-23 and M114-2. We especially thank Maxim Rubin Blum, Christian Borowski, and Stephanie Markert for providing samples. We thank Brandon Seah for useful discussions. This work was funded by the Max Planck Society, the DAAD through a doctoral grant to LS, and the Vienna Science and Technology Fund (WWTF) through project VRG14-021 to JMP.

Statement of competing interests

The authors declare no competing interests.

References

Aktories K. (2011). Bacterial protein toxins that modify host regulatory GTPases. *Nat Rev Microbiol* **9**: 487–498.

Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**: 533–538.

Alvarez-Martinez CE, Christie PJ. (2009). Biological diversity of prokaryotic type IV secretion systems. *Microbiol Mol Biol Rev MMBR* **73**: 775–808.

Anantharaman K, Duhaime MB, Breier JA, Wendt KA, Toner BM, Dick GJ. (2014). Sulfur oxidation genes in diverse deep-sea viruses. *Science* **344**: 757–760.

Anderson RE, Beltrán MT, Hallam SJ, Baross JA. (2013). Microbial community structure across fluid gradients in the Juan de Fuca Ridge hydrothermal system. *FEMS Microbiol Ecol* **83**: 324–339.

Aziz R, Bartels D, Best A, DeJongh M, Disz T, Edwards R, *et al.* (2008). The RAST Server: Rapid annotations using subsystems technology. *BMC Genomics* **9**: 75.

Backert S, Meyer TF. (2006). Type IV secretion systems and their effectors in bacterial pathogenesis. *Curr Opin Microbiol* **9**: 207–217.

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

Bouck J, Miller W, Gorrell JH, Muzny D, Gibbs RA. (1998). Analysis of the quality and utility of random shotgun sequencing at low redundancies. *Genome Res* **8**: 1074–1084.

Buchfink B, Xie C, Huson DH. (2014). Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**: 59–60.

Cirillo SL, Yan L, Littman M, Samrakandi MM, Cirillo JD. (2002). Role of the *Legionella pneumophila rtxA* gene in amoebae. *Microbiology* **148**: 1667–1677.

Cossart P, Sansonetti PJ. (2004). Bacterial invasion: The paradigms of enteroinvasive pathogens. *Science* **304**: 242–248.

DeChaine EG, Bates AE, Shank TM, Cavanaugh CM. (2006). Off-axis symbiosis found: Characterization and biogeography of bacterial symbionts of *Bathymodiolus* mussels from Lost City hydrothermal vents. *Environ Microbiol* **8**: 1902–1912.

Dehio C. (2008). Infection-associated type IV secretion systems of *Bartonella* and their diverse roles in host cell interaction. *Cell Microbiol* **10**: 1591–1598.

Di Genova BM, Tonelli RR. (2016). Infection strategies of intestinal parasite pathogens and host cell responses. *Front Microbiol* **7**. e-pub ahead of print, doi: 10.3389/fmicb.2016.00256.

Dubilier N, Bergin C, Lott C. (2008). Symbiotic diversity in marine animals: The art of harnessing chemosynthesis. *Nat Rev Micro* **6**: 725–740.

Duperron S, Lorion J, Samadi S, Gros O, Gaill F. (2009). Symbioses between deep-sea mussels (Mytilidae: Bathymodiolinae) and chemosynthetic bacteria: Diversity, function and evolution. *C R Biol* **332**: 298–310.

Duperron S, Sibuet M, MacGregor BJ, Kuypers MMM, Fisher CR, Dubilier N. (2007). Diversity, relative abundance and metabolic potential of bacterial endosymbionts in three *Bathymodiolus* mussel species from cold seeps in the Gulf of Mexico. *Environ Microbiol* **9**: 1423–1438.

Fisher C, Childress J, Oremland R, Bidigare R. (1987). The importance of methane and thiosulfate in the metabolism of the bacterial symbionts of two deep-sea mussels. *Mar Biol* **96**: 59–71.

Fujiwara Y, Takai K, Uematsu K, Tsuchida S, Hunt J, Hashimoto J. (2000). Phylogenetic characterization of endosymbionts in three hydrothermal vent mussels: Influence on host distributions. *Mar Ecol Prog Ser* **208**: 147–155.

Gokulan K, Khare S, Rooney AW, Han J, Lynne AM, Foley SL. (2013). Impact of plasmids, including those encodingVirB4/D4 type IV secretion systems, on *Salmonella enterica* serovar Heidelberg virulence in macrophages and epithelial cells. *PLOS ONE* **8**: e77866.

Guglielmini J, de la Cruz F, Rocha EPC. (2013). Evolution of conjugation and type IV secretion systems. *Mol Biol Evol* **30**: 315–331.

Huber P, Basso P, Reboud E, Attrée I. (2016). *Pseudomonas aeruginosa* renews its virulence factors. *Environ Microbiol Rep*. e-pub ahead of print, doi: 10.1111/1758-2229.12443.

Ikuta T, Takaki Y, Nagai Y, Shimamura S, Tsuda M, Kawagucci S, *et al.* (2015). Heterogeneous composition of key metabolic gene clusters in a vent mussel symbiont population. *ISME J*. e-pub ahead of print, doi: 10.1038/ismej.2015.176.

Kaufmann SHE, Dorhoi A. (2016). Molecular determinants in phagocyte-bacteria interactions. *Immunity* **44**: 476–91.

Kendall MM, Sperandio V. (2016). What a dinner party! Mechanisms and functions of interkingdom signaling in host-pathogen associations. *mBio* **7**: e01748-15.

Lally ET, Hill RB, Kieba IR, Korostoff J. (1999). The interaction between RTX toxins and target cells. *Trends Microbiol* **7**: 356–361.

Lang AE, Schmidt G, Schlosser A, Hey TD, Larrinua IM, Sheets JJ, *et al.* (2010). *Photorhabdus luminescens* toxins ADP-Ribosylate actin and RhoA to force actin clustering. *Science* **327**: 1139–1142.

Li L. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**: 2178–2189.

Linhartová I, Bumba L, Mašín J, Basler M, Osička R, Kamanová J, *et al.* (2010). RTX proteins: A highly diverse family secreted by a common mechanism. *FEMS Microbiol Rev* **34**: 1076–1112.

Liu L. (2008). BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics* **24**: 2542–2543.

Los FCO, Randis TM, Aroian RV, Ratner AJ. (2013). Role of pore-forming toxins in bacterial infectious diseases. *Microbiol Mol Biol Rev* **77**: 173–207.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, *et al.* (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**: 18.

Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, *et al.* (2014). CDD: NCBI's conserved domain database. *Nucleic Acids Res* gku1221.

Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, *et al.* (2011). IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res* **40**: D115–D122.

Martinez J, Cogni R, Cao C, Smith S, Illingworth CJR, Jiggins FM. (2016). Addicted? Reduced host resistance in populations with defensive symbionts. *Proc R Soc B Biol Sci* **283**: 20160778.

McFall-Ngai M, Hadfield MG, Bosch TCG, Carey HV, Domazet-Lošo T, Douglas AE, *et al.* (2013). Animals in a bacterial world, a new imperative for the life sciences. *Proc Natl Acad Sci* **110**: 3229–3236.

Miller DJ, Hemmrich G, Ball EE, Hayward DC, Khalturin K, Funayama N, *et al.* (2007). The innate immune repertoire in Cnidaria-ancestral complexity and stochastic gene loss. *Genome Biol* **8**: R59.

Moore AD, Held A, Terrapon N, Weiner J, Bornberg-Bauer E. (2014). DoMosaics: Software for domain arrangement visualization and domain-centric analysis of proteins. *Bioinformatics* **30**: 282–283.

Moya A, Pereto J, Gil R, Latorre A. (2008). Learning how to live together: Genomic insights into prokaryote-animal symbioses. *Nat Rev Genet* **9**: 218–229.

Nuismer SL, Otto SP. (2004). Host–parasite interactions and the evolution of ploidy. *Proc Natl Acad Sci U S A* **101**: 11036–11039.

Nurk S, Meleshko D, Korobeynikov A, Pevzner P. (2016). metaSPAdes: A new versatile de novo metagenomics assembler. *ArXiv160403071 Q-Bio*. http://arxiv.org/abs/1604.03071 (Accessed October 17, 2016).

Pel MJC, Pieterse CMJ. (2013). Microbial recognition and evasion of host immunity. *J Exp Bot* **64**: 1237–1248.

Peng Y, Leung HCM, Yiu SM, Chin FYL. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**: 1420–1428.

Pérez-Brocal V, Latorre A, Moya A. (2011). Symbionts and pathogens: What is the difference? In: Dobrindt U, Hacker JH, Svanborg C (eds) Vol. 358. *Between Pathogenicity and Commensalism*. Springer Berlin Heidelberg: Berlin, Heidelberg, pp 215–243.

Petersen JM, Wentrup C, Verna C, Knittel K, Dubilier N. (2012). Origins and evolutionary flexibility of chemosynthetic symbionts from deep-sea animals. *Biol Bull* **223**: 123–137.

Ponnudurai R, Kleiner M, Sayavedra L, Petersen JM, Moche M, Otto A, *et al.* (2016). Metabolic and physiological interdependencies in the *Bathymodiolus azoricus* symbiosis. *ISME J*. e-pub ahead of print, doi: 10.1038/ismej.2016.124.

Raggi L, Schubotz F, Hinrichs K-U, Dubilier N, Petersen JM. (2013). Bacterial symbionts of *Bathymodiolus* mussels and *Escarpia* tubeworms from Chapopote, an asphalt seep in the southern Gulf of Mexico. *Environ Microbiol* **15**: 1969–1987.

Ronquist F, Huelsenbeck JP. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574.

Roux S, Hawley AK, Beltran MT, Scofield M, Schwientek P, Stepanauskas R, *et al.* (2014). Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta- genomics. *Elife* **3**. e-pub ahead of print, doi: 10.7554/eLife.03125.

Sansonetti P. (2002). Host–pathogen interactions: The seduction of molecular cross talk. *Gut* **50**: iii2-iii8.

Satchell KJF. (2011). Structure and function of MARTX toxins and other large repetitive RTX proteins. *Annu Rev Microbiol* **65**: 71–90.

Sayavedra L, Kleiner M, Ponnudurai R, Wetzel S, Pelletier E, Barbe V, *et al.* (2015). Abundant toxin-related genes in the genomes of beneficial symbionts from deep-sea hydrothermal vent mussels. *eLife* **4**. e-pub ahead of print, doi: 10.7554/eLife.07966.

Schmid MC, Schulein R, Dehio M, Denecker G, Carena I, Dehio C. (2004). The VirB type IV secretion system of *Bartonella henselae* mediates invasion, proinflammatory activation and antiapoptotic protection of endothelial cells. *Mol Microbiol* **52**: 81–92.

Seah BKB, Gruber-Vodicka HR. (2015). Gbtools: Interactive visualization of metagenome bins in R. *Microb Physiol Metab* 1451.

Seubert A, Hiestand R, De La Cruz F, Dehio C. (2003). A bacterial conjugation machinery recruited for pathogenesis. *Mol Microbiol* **49**: 1253–1266.

Sunamura M, Higashi Y, Miyako C, Ishibashi J, Maruyama A. (2004). Two bacteria phylotypes are predominant in the Suiyo Seamount hydrothermal plume. *Appl Environ Microbiol* **70**: 1190–1198.

Sylvan JB, Toner BM, Edwards KJ. (2012). Life and death of deep-sea vents: Bacterial diversity and ecosystem succession on inactive hydrothermal sulfides. *MBio* **3**: e00279-11.

Tellier A, Moreno-Gámez S, Stephan W. (2014). Speed of adaptation and genomic footprints of host–parasite coevolution under arms race and trench warfare dynamics. *Evolution* **68**: 2211–2224.

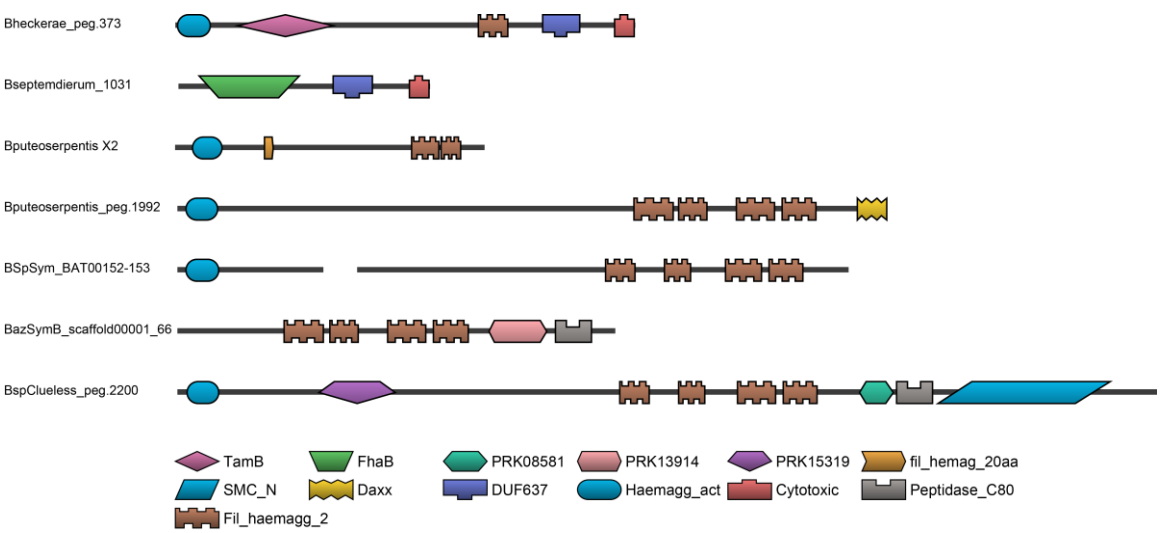
Van Dover C. (2000). The ecology of deep-sea hydrothermal vents. Princeton University Press: Princeton N.J.

Wick RR, Schultz MB, Zobel J, Holt KE. (2015). Bandage: Interactive visualization of *de novo* genome assemblies: Fig. 1. *Bioinformatics* **31**: 3350–3352.

Won YJ, Hallam SJ, O'Mullan GD, Pan IL, Buck KR, Vrijenhoek RC. (2003). Environmental acquisition of thiotrophic endosymbionts by deep-sea mussels of the genus *Bathymodiolus*. *Appl Environ Microbiol* **69**: 6785.

Zhou J, Bruns MA, Tiedje JM. (1996). DNA recovery from soils of diverse composition. *Appl Environ Microbiol* **62**: 316–322.

Supplementary Figures



Supplementary Figure 1. Domain structure of a MARTX-like gene of the *B. heckerae* symbiont and the similar domain architecture in the other *B.* symbionts. The description of the domains is listed in Supplementary Table 2.

Supplementary Tables

Supplementary Table 1. Samples sequenced in this study

Species	Cruise	Collection date	Site	Latitude	Longitude	Depth (m)	Library	No. of sequencing reads	Acc. number	Host CO1
B. thermophilus	AT26-23 Dive 4763-2014	08.11.2014	Crab-Spa, EPR	9°50.377N	104°17.533W	2512	1600R	$16.6^{+,1,6}$	X	KU597624
B. puteoserpentis	M64-2	17.05.05	Logatchev, MAR	14° 45.3N	44° 59.2662W	3009	GOSF2 -PE GOSN3-MP	$217.6^{*,1,3}$ $120.2^{*,1,4}$	X	KU597632
B. sp. 5° South, Clueless	M78-2 302 ROV15	22.04.2009	5° South, Clueless, MAR	4° 48.19599S	12° 22.308024W	2971.5	Clue	$15^{+,2,5}$	X	X
B. sp. 5° South, Turtle Pits	ATA 57 ROV 7/2	21.01.2008	5° South, Turtle Pits, MAR	4°48.558′S	12°22.463′W	2989	1501A, 1593A	$13.4^{+,1,7}$	X	X
B. heckerae	M114-2	14.03.2015	Chapopote, Gulf of Mexico	21° 54.003N	93° 26.124”W	-2925	1600D	$30.6^{+,2,6}$	X	X
B. brooksi	M114-2	14.03.2015	Chapopote, Gulf of Mexico	21° 54.003N	93° 26.124”W	-2925	1789A	$46.6^{+,2,5}$	X	X

Sequencing center = $^{+}$Max Planck Genome Centre (Germany); $^{\&}$University of Bielefeld (Germany); *Genoscope (France).

^{1}DNA extraction as described by Zhou *et al.* (1996); ^{2}DNA extraction with DNAsy Blood and Tissue kit; ^{3}Sequenced as 2 X 100 bp paired-end reads on an Illumina HiSeq 2000 instrument; ^{4}Sequenced as 2 x 100 bp mate-paired reads of a 5.6 kb library on an Illumina HiSeq 2000 instrument; ^{5}Sequenced as 2 x 150 bp paired-end reads on an Illumina HiSeq 2500 instrument; ^{6}Sequenced as 2 X 250 bp paired-end reads on an Illumina MiSeq instrument; and ^{7}Sequenced as 2 X 250 bp paired-end reads on an Illumina MiSeq instrument.

X = Accession numbers will be retrieved upon submission.

Supplementary Table 1. Description of the domains present in MARTX-like genes from *B. heckerae* symbiont and the most similar sequences in the *B.* symbionts.

Accession	Short name	Definition
TIGR01731	fil_hemag_20aa	Adhesin heca family 20-residue repeat. This model represents two copies of a 20-residue repeat found in *Bordetella pertussis* filamentous hemagglutinin family of adhesins
pfam03344	Daxx	Daxx Family; The Daxx protein (also known as the Fas-binding protein) is thought to play a role in apoptosis, but the precise role played by Daxx remains to be determined. Daxx forms a complex with Axin.
pfam05860	Haemagg_act	Haemagglutination activity domain; This domain is suggested to be a carbohydrate-dependent haemagglutination activity site. It is found in a range of haemagglutinins and haemolysins.
COG3210	FhaB	Large exoprotein involved in heme utilization or adhesion [Intracellular trafficking, secretion, and vesicular transport]; Large exoproteins involved in heme utilization or adhesion [Intracellular trafficking and secretion].
pfam09000	Cytotoxic	Cytotoxic; The cytotoxic domain confers cytotoxic activity to proteins, enabling the formation of nucleolytic breaks in 16S ribosomal RNA
PRK15319	PRK15319	AIDA autotransporter-like protein shda; Provisional
COG2911	TamB	Autotransporter translocation and assembly factor tamb [Intracellular trafficking, secretion, and vesicular transport].
pfam13332	Fil_haemagg_2	Haemagluttinin repeat.
pfam04830	DUF637	Possible hemagglutinin (DUF637); This family represents a conserved region found in a bacterial protein which may be a hemagglutinin or hemolysin.
PRK13914	PRK13914	Invasion associated secreted endopeptidase.
pfam11713	Peptidase_C80	Peptidase C80 family; This family belongs to cysteine peptidase family C80.
pfam02463	SMC_N	This domain is found at the N terminus of SMC (structural maintenance of chromosomes) proteins.
PRK08581	PRK08581	N-acetylmuramoyl-L-alanine amidase; Validated

Supplementary Table 2. Domain scores to classify the potential TRGs as MARTX, RTX or YD.

Accession	Short name	Definition	MARTX	RTX	YD	RTX activator	Class
TIGR03660	T1SS_rpt_143	T1SS-143 repeat domain	2	2	0	0	Type 1SS
cl11927	T1SS_rpt_143 superfamily	T1SS-143 repeat domain	2	2	0	0	Type 1SS
TIGR03661	T1SS_VCA0849	Type I secretion C-terminal target domain (VC_A0849 subclass)	2	2	0	0	Type 1SS
cl15337	T1SS_VCA0849 superfamily	Type I secretion C-terminal target domain, VC_A0849 subclass	2	2	0	0	Type 1SS
cl00296	Peptidase_C39_like superfamily	Peptidase family C39 mostly contains bacteriocin-processing endopeptidases from bacteria.	0	0.5	0	0	RTX?
PHA02584	34	Long tail fiber, proximal subunit	0	1	0	0	RTX (in combination)
cl21494	Esterase_lipase superfamily	Esterases and lipases (includes fungal lipases, cholinesterases, etc.)	0	1	0	0	RTX (in combination)
pfam13529	Peptidase_C39_2	Peptidase_C39 like family; Peptidase_C39 like family.	0	1	0	0	RTX (in combination)
COG5153	CVT17	Putative lipase essential for disintegration of autophagic bodies inside the vacuole	0	2	0	0	RTX
cl16721	DUF4329 superfamily	Domain of unknown function (DUF4329)	0	0	1	0	YD (in combination)
cl17169	RRM_SF superfamily	RNA recognition motif (RRM) superfamily	0	0	1	0	YD (in combination)
cl15554	SpvB superfamily	Salmonella virulence plasmid 65kda B protein	0	0	1	0	YD (in combination)
pfam13517	VCBS	Repeat domain in Vibrio, Colwellia, Bradyrhizobium and Shewanella. The large protein size and repeat copy numbers, species distribution, and suggested activities of several member proteins suggests a role for this domain in adhesion (TIGR)	1	0	1	0	YD (in combination)
PRK15244	PRK15244	Virulence protein spvb	0	0	2	0	YD
TIGR03696	Rhs_assc_core	RHS repeat-associated core domain. This class includes secreted bacterial insecticidal toxins and intercellular signalling proteins such as the teneurins in animals	0	0	2	0	YD
cl14012	Rhs_assc_core superfamily	RHS repeat-associated core domain	0	0	2	0	YD
pfam05593	RHS_repeat	Rhs repeat	0	0	2	0	YD
cl11982	RHS_repeat superfamily	Rhs repeat	0	0	2	0	YD

Accession	Short name	Definition	MARTX	RTX	YD	RTX activator	Class
COG3209	RhsA	Uncharacterized conserved protein rhas	0	0	2	0	YD
pfam03534	SpvB	Salmonella virulence plasmid 65kda B protein	0	0	2	0	YD
pfam12255	TcdB_toxin_midC	Insecticide toxin tcdb middle/C-terminal region	0	0	2	0	YD
cl13662	TcdB_toxin_midC superfamily	Insecticide toxin tcdb middle/C-terminal region	0	0	2	0	YD
pfam12256	TcdB_toxin_midN	Insecticide toxin tcdb middle/N-terminal region	0	0	2	0	YD
cl13663	TcdB_toxin_midN superfamily	Insecticide toxin tcdb middle/N-terminal region	0	0	2	0	YD
cl21563	VCBS superfamily	Repeat domain in Vibrio, Colwellia, Bradyrhizobium and Shewanella. The large protein size and repeat copy numbers, species distribution, and suggested activities of several member proteins suggests a role for this domain in adhesion (TIGR)	2	0	2	0	YD
pfam03538	VRP1	Salmonella virulence plasmid 28.1kda A protein	0	0	2	0	YD
cl21676	VRP1 superfamily	Salmonella virulence plasmid 28.1kda A protein	0	0	2	0	YD
TIGR01643	YD_repeat_2x	YD repeat (two copies)	0	0	2	0	YD
cl00125	RHOD superfamily	Rhodanese homology domain (rhod)	-10	-10	-10	0	Rhodonase
TIGR02172	Fb_sc_TIGR02172	Fibrobacter succinogenes paralogous family TIGR02172	0.01	0.01	0	0	MARTX/RTX, not specific
pfam11713	Peptidase_C80	Peptidase C80 family	1.01	0.5	0	0	MARTX/RTX, more common in MARTX
cl21453	PKc_like superfamily	Protein Kinases, catalytic domain	0.05	0.05	0	0	MARTX/RTX (in combination)
pfam03160	Calx-beta	Calx-beta domain; Calx-beta domain.	1	1	0	0	MARTX/RTX
COG2931	COG2931	Ca2+-binding protein, RTX toxin-related [Secondary metabolites biosynthesis, transport and catabolism]	2	2	0	0	MARTX/RTX
pfam05557	MAD	Mitotic checkpoint protein	1	1	0	0	MARTX/RTX
cl13207	Peptidase_C80 superfamily	Peptidase C80 family; This family belongs to cysteine peptidase family C80.	1	1	0	0	MARTX/RTX
pfam08548	Peptidase_M10_C	Peptidase M10 serralysin C terminal	0.6	0.5	0	0	MARTX/RTX

Accession	Short name	Definition	MARTX	RTX	YD	RTX activator	Class
pfam05701	WEMBL	Weak chloroplast movement under blue light	1	0	1	0	MARTX/RHS
COG2831	FhaC	Hemolysin activation/secretion	0.01	0	0	1	MARTX activator
pfam08479	POTRA_2	POTRA domain, shlb-type. Shlb is important in the secretion and activation of the haemolysin shla	0.01	0.01	0	1	MARTX activator
PHA02562	46	Endonuclease subunit	0.01	0	0	0	MARTX (In combination)
TIGR01612	235kDa-fam	Reticulocyte binding/rhoptry protein	0.01	0	0	0	MARTX (In combination)
pfam13166	AAA_13	AAA domain; This family of domains contain a P-loop motif that is characteristic of the AAA superfamily. Many of the proteins in this family are conjugative transfer proteins	0.01	0	0	0	MARTX (In combination)
pfam13514	AAA_27	AAA domain; This domain is found in a number of double-strand DNA break proteins	0.01	0	0	0	MARTX (In combination)
TIGR01901	adhes_NPXG	Filamentous hemagglutinin family N-terminal domain. A number of the members of this family have been designated adhesins, filamentous haemagglutinins, heme/hemopexin-binding protein, etc.	0.01	0	0	0	MARTX (In combination)
pfam04111	APG6	Autophagy protein Apg6; In yeast, 15 Apg proteins coordinate the formation of autophagosomes	0.01	0	0	0	MARTX (In combination)
pfam13754	Big_3_4	Bacterial Ig-like domain (group 3)	0.01	0	0	0	MARTX (in combination)
cl16375	Big_3_4 superfamily	Bacterial Ig-like domain (group 3)	0.01	0	0	0	MARTX (in combination)
pfam13205	Big_5	Bacterial Ig-like domain	0.01	0	0	0	MARTX (in combination)
pfam05262	Borrelia_P83	Borrelia P83/100 protein	0.01	0	0	0	MARTX (In combination)
cd11304	Cadherin_repeat	Cadherins are glycoproteins involved in Ca2+-mediated cell-cell adhesion. They play numerous roles in cell fate, signalling, proliferation, differentiation, and migration	0.01	0	0	0	MARTX (in combination)
pfam11600	CAF-1_p150	Chromatin assembly factor 1 complex p150 subunit, N-terminal	0.01	0	0	0	MARTX (In combination)
pfam07888	CALCOCO1	Calcium binding and coiled-coil	0.01	0	0	0	MARTX (In

Accession	Short name	Definition	MARTX	RTX	YD	RTX activator	Class
		domain (CALCOCO1) like					combination)
pfam10174	Cast	RIM-binding protein of the cytomatrix active zone; This is a family of proteins that form part of the CAZ (cytomatrix at the active zone) complex which is involved in determining the site of synaptic vesicle fusion	0.01	0	0	0	MARTX (In combination)
TIGR03346	chaperone_ClpB	ATP-dependent chaperone clpb. This molecular chaperone does not act as a protease, but rather serves to disaggregate misfolded and aggregated proteins	0.01	0	0	0	MARTX (In combination)
COG4372	COG4372	Uncharacterized conserved protein, contains DUF3084 domain	0.01	0	0	0	MARTX (In combination)
COG4913	COG4913	Uncharacterized protein	0.01	0	0	0	MARTX (In combination)
COG5281	COG5281	Phage-related minor tail protein	0.01	0	0	0	MARTX (In combination)
pfam03344	Daxx	Daxx Family; The Daxx protein (also known as the Fas-binding protein) is thought to play a role in apoptosis	0.01	0	0	0	MARTX (In combination)
pfam09756	DDRGK	DDRGK domain	0.01	0	0	0	MARTX (In combination)
pfam07263	DMP1	Dentin matrix protein 1 (DMP1); This family consists of several mammalian dentin matrix protein 1 (DMP1) sequences	0.01	0	0	0	MARTX (In combination)
pfam07423	DUF1510	Protein of unknown function (DUF1510)	0.01	0	0	0	MARTX (In combination)
cl22487	DUF1510 superfamily	Protein of unknown function (DUF1510)	0.01	0	0	0	MARTX (In combination)
pfam12128	DUF3584	Protein of unknown function (DUF3584)	0.01	0	0	0	MARTX (In combination)
pfam04094	DUF390	Protein of unknown function (DUF390)	0.01	0	0	0	MARTX (In combination)
pfam05667	DUF812	Protein of unknown function (DUF812)	0.01	0	0	0	MARTX (In combination)
pfam05917	DUF874	Helicobacter pylori protein of unknown function (DUF874)	0.01	0	0	0	MARTX (In combination)
COG4942	EnvC	Septal ring factor envc, activator of murein hydrolases amia and amib	0.01	0	0	0	MARTX (In combination)
COG4477	EzrA	Septation ring formation regulator ezra	0.01	0	0	0	MARTX (In combination)

Accession	Short name	Definition	MARTX	RTX	YD	RTX activator	Class
pfam13332	Fil_haemagg_2	Haemagluttinin repeat	0.01	0	0	0	MARTX (In combination)
cl16241	Fil_haemagg_2 superfamily	Haemagluttinin repeat	0.01	0	0	0	MARTX (In combination)
pfam05860	Haemagg_act	Haemagglutination activity domain	0.01	0	0	0	MARTX (In combination)
smart00912	Haemagg_act	Haemagglutination activity	0.01	0	0	0	MARTX (In combination)
cl05436	Haemagg_act superfamily	Haemagglutination activity domain	0.01	0	0	0	MARTX (In combination)
pfam07111	HCR	Alpha helical coiled-coil rod protein (HCR). The function of HCR is unknown but it has been implicated in psoriasis in humans and is thought to affect keratinocyte proliferation.	0.01	0	0	0	MARTX (In combination)
PRK11448	hsdR	Type I restriction enzyme ecoki subunit R	0.01	0	0	0	MARTX (In combination)
smart00283	MA	Methyl-accepting chemotaxis-like domains (chemotaxis sensory transducer)	0.01	0	0	0	MARTX (In combination)
pfam09726	Macoilin	Transmembrane protein	0.01	0	0	0	MARTX (In combination)
pfam05672	MAP7	MAP7 (E-MAP-115) family; The organisation of microtubules varies with the cell type and is presumably controlled by tissue-specific microtubule-associated proteins (maps). The 115-kda epithelial MAP (E-MAP-115/MAP7) has been identified as a microtubule-stabilizing protein predominantly expressed in cell lines of epithelial origin	0.01	0	0	0	MARTX (In combination)
COG5271	MDN1	Midasin, AAA atpase with vwa domain, involved in ribosome	0.01	0	0	0	MARTX (In combination)
TIGR04523	Mplasa_alph_rch	Helix-rich Mycoplasma protein	0.01	0	0	0	MARTX (In combination)
COG3264	MscK	Small-conductance mechanosensitive channel	0.01	0	0	0	MARTX (In combination)
COG3096	MukB	Chromosome condensin mukbef, atpase and DNA-binding subunit mukb Uncharacterized protein involved in chromosome partitioning	0.01	0	0	0	MARTX (In combination)
PRK04863	mukB	Cell division protein mukb	0.01	0	0	0	MARTX (In combination)

Accession	Short name	Definition	MARTX	RTX	YD	RTX activator	Class
pfam05616	Neisseria_TspB	Neisseria meningitidis tspb protein	0.01	0	0	0	MARTX (In combination)
pfam10168	Nup88	Nuclear pore component; Nup88 is overexpressed in tumor cells.	0.01	0	0	0	MARTX (In combination)
PRK00247	PRK00247	Putative inner membrane protein translocase component yidc.	0.01	0	0	0	MARTX (In combination)
PRK00409	PRK00409	Recombination and DNA strand exchange inhibitor protein.	0.01	0	0	0	MARTX (In combination)
PRK02224	PRK02224	Chromosome segregation protein	0.01	0	0	0	MARTX (In combination)
PRK03918	PRK03918	Chromosome segregation protein	0.01	0	0	0	MARTX (In combination)
PRK05035	PRK05035	Electron transport complex protein rnfc	0.01	0	0	0	MARTX (In combination)
PRK05771	PRK05771	V-type ATP synthase subunit I	0.01	0	0	0	MARTX (In combination)
PRK06975	PRK06975	Bifunctional uroporphyrinogen-III synthetase/uroporphyrin-III C-methyltransferase	0.01	0	0	0	MARTX (In combination)
PRK08581	PRK08581	N-acetylmuramoyl-L-alanine amidase	0.01	0	0	0	MARTX (In combination)
PRK09418	PRK09418	Bifunctional 2',3'-cyclic nucleotide 2'-phosphodiesterase/3'-nucleotidase precursor protein	0.01	0	0	0	MARTX (In combination)
PRK11281	PRK11281	Hypothetical protein	0.01	0	0	0	MARTX (In combination)
PRK12704	PRK12704	Phosphodiesterase	0.01	0	0	0	MARTX (In combination)
PRK12705	PRK12705	Hypothetical protein	0.01	0	0	0	MARTX (In combination)
PRK13914	PRK13914	Invasion associated secreted endopeptidase	0.01	0	0	0	MARTX (In combination)
PTZ00108	PTZ00108	DNA topoisomerase 2-like protein	0.01	0	0	0	MARTX (In combination)
PTZ00121	PTZ00121	Maebl	0.01	0	0	0	MARTX (In combination)
TIGR00606	rad50	Rad50; All proteins in this family for which functions are known are involvedin recombination, recombinational repair, and/or non-homologous end joining.	0.01	0	0	0	MARTX (In combination)
COG0419	SbcC	DNA repair exonuclease sbccd atpase subunit	0.01	0	0	0	MARTX (In combination)
TIGR00618	sbcc	Exonuclease sbcc	0.01	0	0	0	MARTX (In combination)
pfam05483	SCP-1	Synaptonemal complex protein 1	0.01	0	0	0	MARTX (In

Accession	Short name	Definition	MARTX	RTX	YD	RTX activator	Class
		(SCP-1)					combination)
TIGR04211	SH3_and_anchor	SH3 domain protein; Members of this protein family have a signal peptide, a strongly conserved SH3 domain, a variable region, and then a C-terminal hydrophobic transmembrane alpha helix region.	0.01	0	0	0	MARTX (In combination)
pfam03865	ShlB	Haemolysin secretion/activation protein shlb/fhac/hecb	0.01	0.01	0	1	MARTX (In combination)
COG1196	Smc	Chromosome segregation atpase [Cell cycle control, cell division, chromosome partitioning]; Chromosome segregation atpases	0.01	0	0	0	MARTX (In combination)
pfam02463	SMC_N	Recf/recn/SMC N terminal domain; This domain is found at the N terminus of SMC proteins. The SMC (structural maintenance of chromosomes)	0.01	0	0	0	MARTX (In combination)
TIGR02169	SMC_prok_A	Chromosome segregation protein SMC	0.01	0	0	0	MARTX (In combination)
TIGR02168	SMC_prok_B	Chromosome segregation protein SMC	0.01	0	0	0	MARTX (In combination)
pfam08317	Spc7	Spc7 kinetochore protein; This domain is found in cell division proteins which are required for kinetochore-spindle association.	0.01	0	0	0	MARTX (In combination)
TIGR04320	Surf_Exclu_PgrA	SEC10/pgra surface exclusion domain. The tetracycline resistance plasmid pcf10 in Enterococcus faecalis promotes conjugal plasmid transfer in response to sex pheromones, but pgra/Sec10 encoded by that plasmid, a member of this family, specifically inhibits the ability of cells to receive homologous plasmids. The phenomenon is called surface exclusion.	0.01	0	0	0	MARTX (In combination)
TIGR02680	TIGR02680	TIGR02680 family protein	0.01	0	0	0	MARTX (In combination)
COG3064	TolA	Membrane protein involved in colicin uptake [Cell wall/membrane/envelope biogenesis]; Membrane protein involved in colicin uptake [Cell envelope biogenesis, outer membrane].	0.01	0	0	0	MARTX (In combination)

Accession	Short name	Definition	MARTX	RTX	YD	RTX activator	Class
PRK09510	tolA	Cell envelope integrity inner membrane protein tola; Provisional	0.01	0	0	0	MARTX (In combination)
TIGR02794	tolA_full	Tola protein. The Tol-Pal complex is required for maintaining outer membrane integrity. Also involved in transport (uptake) of colicins and filamentous DNA, and implicated in pathogenesis.	0.01	0	0	0	MARTX (In combination)
pfam13868	Trichoplein	Trichoplein or mitostatin, was first defined as a meiosis-specific nuclear structural protein. It has since been linked with mitochondrial movement	0.01	0	0	0	MARTX (In combination)
COG4717	YhaN	Uncharacterized protein yhan, contains AAA domain	0.01	0	0	0	MARTX (In combination)
COG2268	YqiK	Uncharacterized membrane protein yqik	0.01	0	0	0	MARTX (In combination)
smart00191	Int_alpha	Integrin alpha (beta-propellor repeats); Integrins are cell adhesion molecules that mediate cell-extracellular matrix and cell-cell interactions.	-10	-10	-10	0	Integrin
cl05885	HCBP_related superfamily	Haemolysin-type calcium binding protein related domain	0	1	0	0	Hemolysin cabind

Chapter IV:

Understanding symbiont colonization of deep-sea mussels using differential gene expression

Lizbeth Sayavedra[1#], Rebecca Ansorge[1], Bruno Huettel[2], Manuel Liebeke[1], Nicole Dubilier[1], Jillian M. Petersen[1,3#]

[1]Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany

[2]Max Planck Institute for Plant Breeding Research, Carl-von-Linee-Weg 10, 50829 Cologne, Germany

[3]University of Vienna, Althanstrasse 14, 1090 Vienna, Austria

[#]Corresponding authors

This manuscript is in preparation and has not been reviewed by all authors.

Author contributions

LS conceived the study, analyzed the data, and wrote the manuscript; RA did the SNP analysis; BH sequenced the samples; ML supervised metabolomics analysis; ND helped to conceive the study; JP helped to conceive the study and revised the manuscript.

Abstract

Deep-sea *Bathymodiolus* mussels host intracellular sulfur-oxidizing (SOX) and methane-oxidizing (MOX) symbionts in their gills that provide them with nutrition. New gill tissues are formed throughout the lifetime of the mussels. These are initially symbiont-free and are continuously colonized by the symbiotic bacteria as they grow throughout the mussel's lifetime. To better understand the molecular processes involved in symbiont colonization we compared metatranscriptomes of young, newly colonized gill tissues with older, mature regions of gill tissues from five mussels. The mussel host overexpressed genes for cell proliferation and migration in young tissues, confirming that they are actively growing. The SOX symbiont overexpressed 236 genes in young tissues that included protein repair, central carbon metabolism, and energy production genes. These expression patterns are reminiscent of the lag phase of bacterial growth during adaptation to new environmental conditions. The SOX symbiont also overexpressed fatty acid biosynthesis in young tissues and metabolomics confirmed that fatty acids were more abundant in these than in mature tissues. This raises the intriguing possibility that the SOX symbiont provides its host with fatty acids to help power the energy-demanding gill growth. In contrast to the SOX symbiont, the MOX symbiont overexpressed only four genes in young tissues compared with older tissues. Our results suggest that SOX symbionts in young gill tissues are in an adaptation phase after their acquisition from a dormant environmental population. In contrast, the MOX symbiont may be acquired from an actively growing population, either from older mature tissues through ‘self-infection’ or from an active environmental population.

Introduction

Mussels from the genus *Bathymodiolus* dominate deep-sea hydrothermal vent and cold seep ecosystems worldwide (Cavanaugh *et al.*, 2013; Dubilier *et al.*, 2008; Van Dover, 2002). The mussels, as well as other fauna in these habitats, depend on an intimate association with chemosynthetic bacteria that harness the abundant reduced compounds to obtain energy, fix carbon and provide the host with nutrition (Cavanaugh *et al.*, 2013). *Bathymodiolus* mussels host their symbionts in specialized gill cells called bacteriocytes. They can host a sulfur oxidizing-symbiont (SOX), a methane-oxidizing symbiont (MOX), or both simultaneously (Duperron *et al.*, 2008; Dubilier *et al.*, 2008; Petersen *et al.*, 2011).

Although *Bathymodiolus* mussels are exposed to countless bacterial species in their environment, the symbiosis is highly specific. In most species, only one 16S rRNA phylotype of each symbiont type (SOX and MOX) is found (DeChaine *et al.*, 2006; Dubilier *et al.*, 2008; Duperron *et al.*, 2006). Nevertheless, the genomic content of the SOX symbiont can vary between distinct co-occurring strains (Ikuta *et al.*, 2015). Both SOX and MOX symbionts are most likely horizontally transmitted, which means that the host needs to take up its symbionts from the environment with every generation anew (Won *et al.*, 2003; DeChaine *et al.*, 2006; Fontanez and Cavanaugh, 2014). New gill filaments are grown during the host's life from a region known as the budding zone. The budding zone is initially symbiont-free and therefore needs to be continuously colonized as the gill grows (Neumann and Kappes, 2003; Wentrup *et al.*, 2014). To colonize the budding zone, *Bathymodiolus* mussels could acquire new symbionts from the environment, or alternatively, symbionts from older filaments could re-infect the budding zone (Wentrup *et al.*, 2014). A combination of both strategies is also possible. The molecular mechanisms that allow hosts and symbionts to recognize one another and to establish their intimate symbiosis are currently unknown.

The goal of this study was to identify potential molecular mechanisms involved in symbiosis establishment. Like most host-associated microorganisms, the symbionts of *Bathymodiolus* mussels have not been successfully cultivated, making cultivation-independent techniques essential for understanding molecular host-symbiont interactions. Thus, we used metatranscriptome sequencing of five individuals to compare the activity of symbionts and hosts in the budding zone vs older gill filaments, where the symbiosis is well-established. In this study, the term 'overexpression' will be used to mean that the transcripts of a gene are more abundant in a gill region vs. another. Moreover, we assessed the symbiont genome variation between budding zone vs established tissue with metagenome sequencing and single nucleotide polymorphism (SNP) analyses of the same samples used for transcriptome sequencing.

Results

Genome and transcriptome *de novo* assemblies

We obtained essentially complete genomes of the sulfur-oxidizing (SOX) and methane-oxidizing (MOX) symbionts from the metagenomes of single *Bathymodiolus*

individuals. The best assembly of each symbiont was used as a reference for differential gene expression in all mussel individuals (assembly metrics are described in Table 1). The relative abundance of symbionts and host in each sample was determined based on the median genome coverage. This showed that both symbionts were more abundant in established tissue compared to the budding zone (Figure 1a). The change in symbiont abundance affects the proportion of symbiont mRNA present in each metatranscriptome, as the mRNA will reflect the abundance and activity per cell of host and symbionts. To account for the symbiont abundance variation, we used the SOX and MOX symbiont genome assemblies as a reference to map the mRNA reads. Analysis of differential gene expression was then carried out per symbiont, only considering the reads mapped to the reference genome for normalization. Genes were therefore considered overexpressed if they made up a larger proportion of the total reads mapping to a symbiont in one sample vs another. For the *Bathymodiolus* mussel host, we assembled the transcriptome *de novo* and removed most bacterial reads using the symbiont reference genomes (Metrics described in Supplementary Table 1) (see Methods – *de novo* transcriptome assembly and differential expression of the host).

Table 1. Metrics of symbiont genome assemblies. Completeness was estimated with CheckM (Parks *et al.*, 2015).

Symbiont	Completeness	No. of contigs	Largest contig	Total length	N50
SOX	98.4 % (p_Proteobacteria, UID4274)	482	31,807 bp	2,294,222	7,981
MOX	98.6 % (c_Gammaproteobacteria, UID4274)	512	62,793 bp	3,084,637	12,681

We compared the percentage of mRNA or rRNA reads from each symbiont in the five mussel individuals. The genome of the SOX symbiont clearly outnumbered the MOX symbiont in both, new and old gill filaments, suggesting that the SOX symbiont is more abundant than the MOX symbiont along the whole gill (Figure 1a). If the symbiont cells are at a similar growth stage with similar environmental conditions in new and old filaments, the transcriptional activity of the symbionts should reflect their abundance (Blazewicz *et al.*, 2013). The relative proportion of SOX symbiont rRNA was higher than for the MOX symbiont in all ten samples, reflecting the abundance of SOX and MOX symbiont reads in the metagenomes (Figure 1b). Intriguingly, the percentage of mRNA

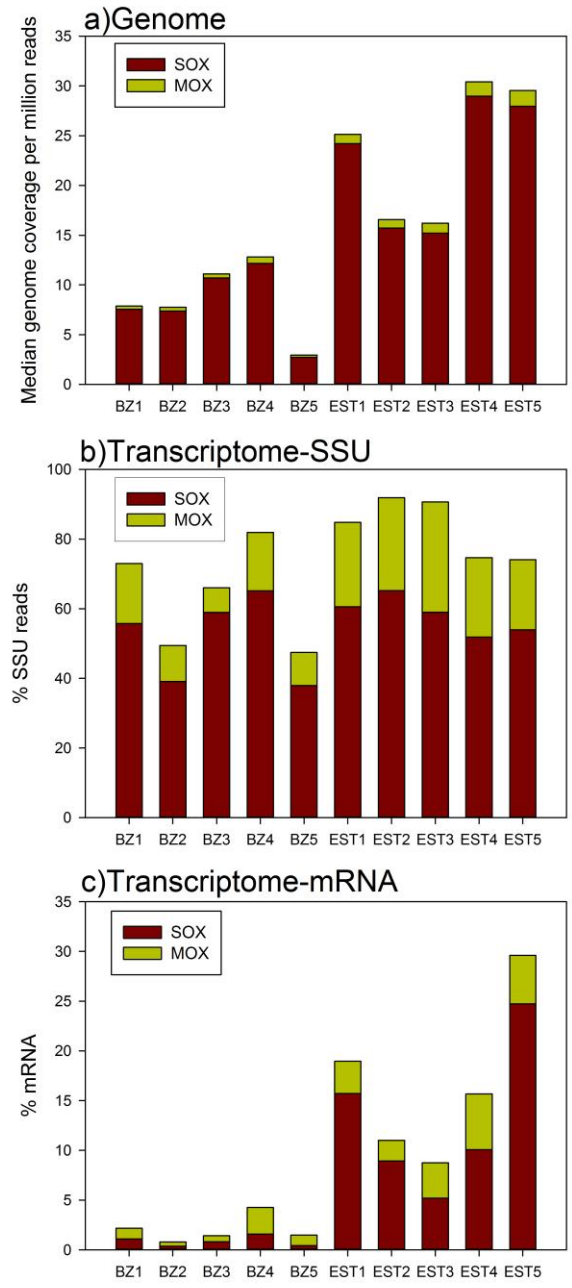


Figure 1. a) Median genome coverage of sulfur-oxidizing (SOX) and methane-oxidizing (MOX) symbionts normalized to sequencing depth of the library. b) Percentage of transcriptome reads mapping to the small subunit (SSU) of the ribosomal RNA c) Percentage of transcriptome reads mapping unambiguously to the SOX or MOX symbiont genomes. The reads were first depleted of mitochondrial reads and rRNA of symbionts and host. BZ = budding zone; EST = established tissue.

for each symbiont did not follow the same trend: although in older gill filaments the percentage of mRNA of the SOX was higher compared to the MOX symbiont, both symbionts accounted for a similar percentage of mRNA reads in the budding zone (Figure 1c).

The high abundance of 16S rRNA SOX reads in the budding zone might suggest that they are actively using the translation machinery. In agreement, 18 ribosomal subunits, as well as chaperones such as GroEL were overexpressed in the budding zone. These results suggest that the SOX symbiont had a higher rate of protein synthesis in the budding zone compared to the established tissue.

The *Bathymodiolus* host overexpressed genes related to growth in the budding zone

The host overexpressed 19 gene isoforms in the budding zone and 38 in the established tissue (Figure 2). In the budding zone, the host overexpressed gene isoforms related to the regulation of cell proliferation and migration, as expected for a region of active growth. Among these isoforms, we found i) Four tensin-like isoforms. Tensins anchor cells to the extracellular matrix by coupling the contractile actin cytoskeleton to integrins. Integrins then bind

signaling molecules for regulatory pathways that play a critical role in migration, proliferation and apoptosis (Chen *et al.*, 2003; Lo, 2004); ii) Two microtubule-binding proteins. Microtubule-based binding and microtubule-associated proteins help in the organization of the cytoskeleton, a process that helps to move internal cell structures during cell division (Mollinedo and Gajate, 2003); iii) Two homeobox domain-containing isoforms. Homeobox domains are found in *hox* genes, which are transcriptional regulators involved in the control of gene expression along the anterior and posterior regions of animals and have a crucial role in cell division, migration, and apoptosis (Alonso, 2002); iv) A retinoid receptor. Retinoid receptors induce epigenetic modifications that are involved in cell differentiation (Gudas, 2013; Urvalek *et al.*, 2014); and v) A transcriptional regulator of cell proliferation. These results support the growth of the gill tissue from the budding zone at the time of sampling.

Only a few of the host genes overexpressed in the established tissue were similar to characterized proteins. These included methyltransferases and histone-like proteins, which are known to be involved in chromatin remodeling (reviewed in Jaenisch and Bird, 2003). Chromatin structure influences gene expression and can play a crucial role in the control of cell division and cell differentiation (Urvalek *et al.*, 2014). The overexpression of these genes could be a response of the cell to avoid massive cell proliferation. Indeed, we found an isoform similar to the recombinase Rad51 overexpressed, which can have a role in cell growth control (Scully *et al.*, 1997). The overexpression of these genes could therefore help to restrict cell proliferation in older gill filaments.

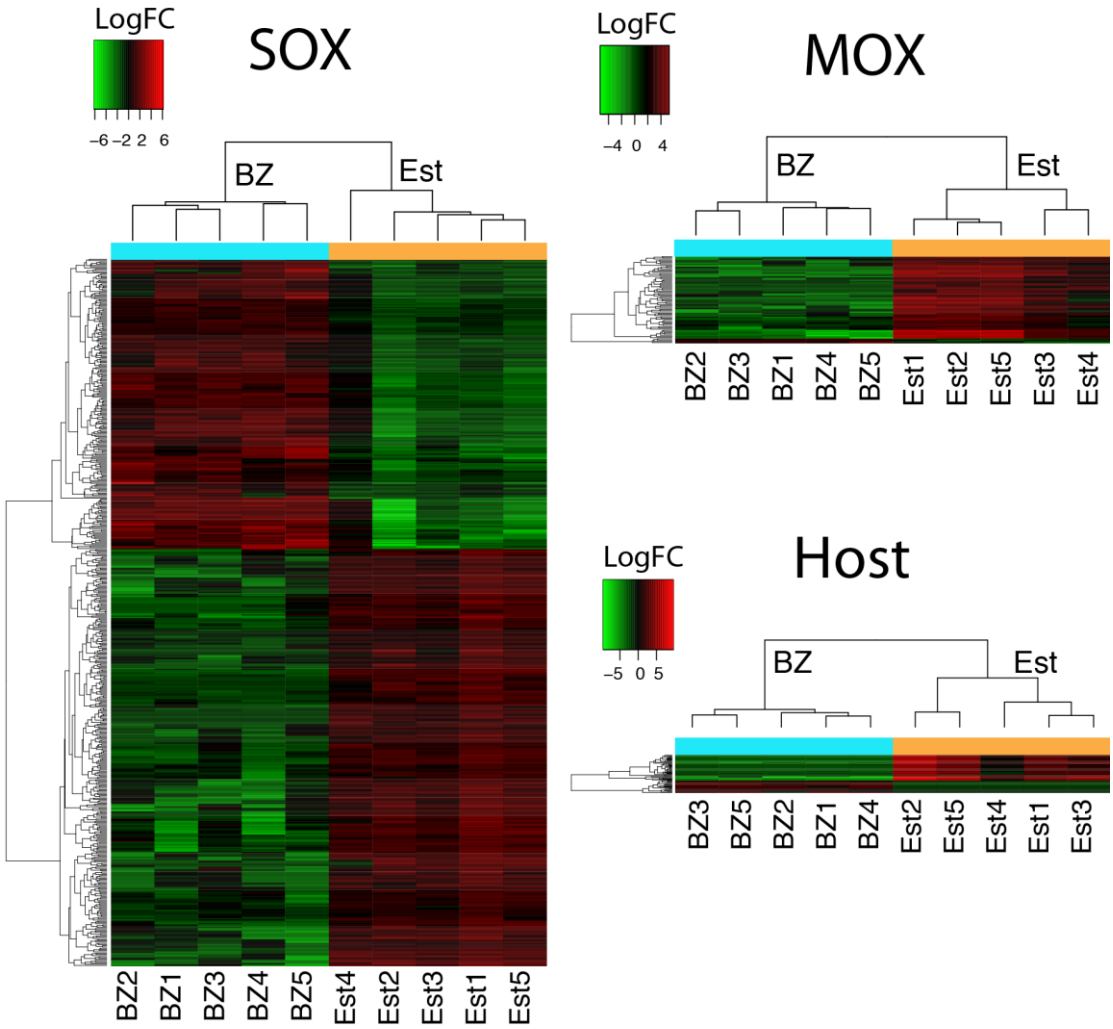


Figure 2. Differential gene expression of the SOX symbiont, MOX symbiont, and the *Bathymodiolus* host. The minimum fold change (LogFC) in expression considered was three ($p = 1e^{-3}$). BZ, budding zone; Est, established tissue.

Single nucleotide polymorphism (SNP) analysis

Since *Bathymodiolus* mussels can be colonized by their symbionts during their entire life span, they have a high chance to 'select' their symbionts from a diverse pool of symbiont strains present in the surrounding environment or their own established filaments (Vrijenhoek, 2010; Wentrup *et al.*, 2014). A key question arises from this mode of transmission: What is the main source of the symbiont population in the newly formed gill filaments? We would expect the heterogeneity of the budding zone to be the same or smaller than the established tissue if i) there are many more symbiont strains present in the environment than those that are taken up by new gill filaments, and if ii) symbionts

from older gill filaments can migrate to the budding zone to inoculate the nascent gill filaments. On the contrary, if the symbiont population of the budding zone is taken up from the environment, we would expect all three possibilities in genomic variation in the budding zone compared to the established tissue: lower, the same or higher, which would depend on i) the number of strains present in the environment and ii) the number of strains that can colonize every new forming gill cell. Thus, we aimed to estimate the genomic variation in the symbiont population between the two gill regions of each mussel by mapping the genome reads to the reference genomes and calculating the number of single nucleotide polymorphisms (SNPs) per kb. The number of SNPs per kb of the MOX symbiont was not significantly different between the two gill regions, suggesting a similar genomic variation between the two gill regions (budding zone = 1.22±0.6; established tissue = 0.98±0.07; p-value 0.5, Kruskal-Wallis test). In contrast, the SOX symbiont had significantly more SNPs per kb in the budding zone compared to the established tissue in all five individuals (budding zone = 5.84±0.38; established tissue = 4.34±0.68; p-value 0.008611, Kruskal-Wallis test). This indicates that the genomic variation of the SOX symbiont was higher in the budding zone.

The higher variation in the genomes of SOX symbionts located in the budding zone raised the next question: Are the strains of the budding zone more similar to those of the established tissue of the same mussel? We used the genome variation of the SOX symbionts to reconstruct a principal component analysis (PCA) with the frequency of the nucleotide variations against our reference SOX symbiont genome per gill region and per individual. If the mussels would use old filaments to self-infect the new filaments, we would expect a cluster in the PCA plot of the budding zone and the established tissue per individual. Instead, we did not observe a grouping per individual or gill region, suggesting that the budding zone does not acquire most of the population through self-infection (Supplementary Figure 1). Therefore, we hypothesize that the SOX symbionts in the budding zone are taken up from the environment or co-occurring hosts.

The SOX symbiont overexpressed energy, carbon metabolism, and fatty acid biosynthesis in the budding zone

The SOX symbiont overexpressed 236 genes in the budding zone and 339 genes in the established tissue. Genes of pathways involved in energy production and electron transport were overexpressed in the budding zone. These pathways included hydrogen

oxidation, sulfur oxidation, nitrate respiration, and oxygen respiration, suggesting that the SOX symbionts require more energy in the budding zone (Figure 3 and Supplementary Figure 2). Indeed, pathways from the central carbon metabolism that could use the energy produced in the form of ATP and NAD(P)H were overexpressed in the budding zone: a modified version of the Calvin cycle for CO_2 fixation, the incomplete tricarboxylic acid (TCA) cycle, and glycolysis/gluconeogenesis (Figure 3 and Supplementary Figure 2). The pathway for the synthesis of the fatty acid palmitate was also overexpressed in the budding zone (Figure 3 and 4). Palmitate is the most common short chain fatty acid in bacteria (Gunina *et al.*, 2014). Fatty acids are essential for membrane components and for metabolic energy (Fujita *et al.*, 2007). Thus, the overexpression of the biosynthesis of palmitate might indicate that the SOX symbiont is storing energy and carbon substrates for later use, or that is preparing the building blocks for the membrane of future daughter cells.

In the established tissue, the SOX symbiont overexpressed a TRAP transporter that can be used for the import of organic acids such as succinate, fumarate, and malate. The SOX symbiont overexpressed in this same gill region a gene encoding a malic acid enzyme that can convert malate to pyruvate (Supplementary Figure 2). Pyruvate can be converted to phosphoenolpyruvate (PEP), which in turn can be used to synthesize several glycolytic intermediates and precursors of biosynthetic pathways (Supplementary Figure 2) (Munoz-Elias and McKinney, 2006). The SOX symbiont could be using this anaplerotic reaction to feed biosynthetic building blocks into the synthesis of the amino acids isoleucine, tyrosine, serine and tryptophan, whose biosynthesis was overexpressed in the established tissue (Table 2).

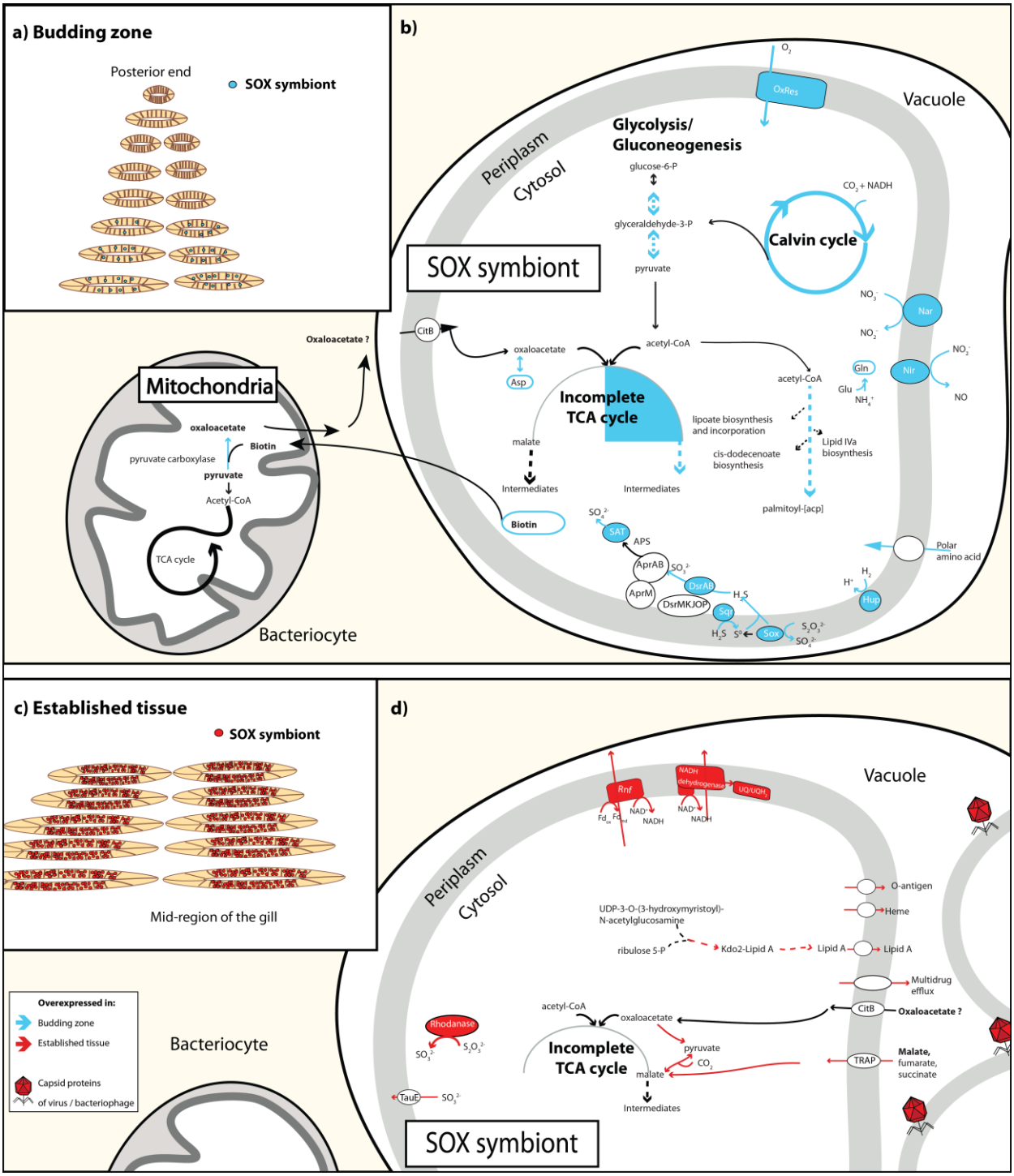


Figure 3. Schematic representation of the gill filaments, which host only a low number of symbionts (a); and the established tissue, which host a dense symbiont population (c). An overview of the pathways overexpressed by the SOX symbiont in the budding zone (b) and the established tissue (d) is shown. b) The metabolic profile of the SOX symbiont in the budding zone suggests a metabolic coupling between the SOX symbiont and the host. The SOX symbiont needs external oxaloacetate to fulfill its metabolic needs, which could be provided by the host. In exchange, the SOX may provide the cofactor biotin, which is required for the conversion of pyruvate to oxaloacetate by the host. Pathways for obtaining energy, the Calvin cycle, gluconeogenesis, and TCA cycle are overexpressed in the budding zone. d) In the established tissue, multiple capsid proteins from viruses were found overexpressed.

Figure 3. *continued on next page.*

Figure 3. *continued*
Arows with reactions overexpressed in the budding zone are colored blue, reactions overexpressed in the established tissue are colored red, and reactions encoded in the genome but not differentially expressed are colored black. Cyt, Cytochrome; COX, Cytochrome c oxidase; Nar; membrane bound respiratory nitrate reductase; Dsr, dissimilatory sulfite reductase gene cluster; Sox, sulfur oxidation gene cluster; AprAB, adenosine-5'-phosphosulfate (APS) reductase; AprM, APS reductase membrane anchor; SQR, sulfide-quinone reductase; SAT, sulfate adenylyltransferase; CitB, citrate transporter.

Table 2. Amino acids, vitamins, and cofactors synthesized by the SOX and MOX symbionts. The gill region where the biosynthesis of the amino acid or cofactor is overexpressed is indicated.

	Compound	SOX – overexpressed in:	MOX – overexpressed in:
Amino acid	isoleucine	Est (only partial)	+
	lysine	Est (only partial)	+
	arginine	Est (only partial, on the way to ornithine)	+
	tryptophan	Est (one before the last step)	+
	tyrosine	Est (one before the last step)	+
	serine	Est (last step)	+
	cysteine	BZ (partial)	+
	glycine	BZ (last step)	+
	aspartate	BZ (last step)	+
	glutamine	BZ (last step)	+
	threonine	BZ (last step)	+
	alanine	+	+
	asparagine	+	+
	glutamate	+	+
	histidine	+	+
	leucine	+	+
	methionine	+	+
	proline	+	+
	selenocysteine	+	+
	valine	+	+
	phenylalanine	-	-
	pyrrolysine	-	-
Vitamins and cofactors	Cobalamin (B_{12})	Only synthesis of coenzyme B_{12} from cobalamin	Est
	Heme (Porphyrin)	+	+
	Ubiquinone	Est (partial)	+
	Nicotinate and nicotinamide	+	+
	Folate	Est (partial)	+
	Lipoate	+	+
	Riboflavin (B_2)	+	+
	Pantothenate (B_5)	Est (2 genes)	+
	Pyridoxine (B_6)	Est (partial)	+
	Thiamine (B_1)	BZ (partial)	+
	Biotin	+	+
	Coenzyme A	Est (last step)	+

BZ, Budding zone; Est, Established tissue
+ Encoded in the genome
-Not encoded in the genome

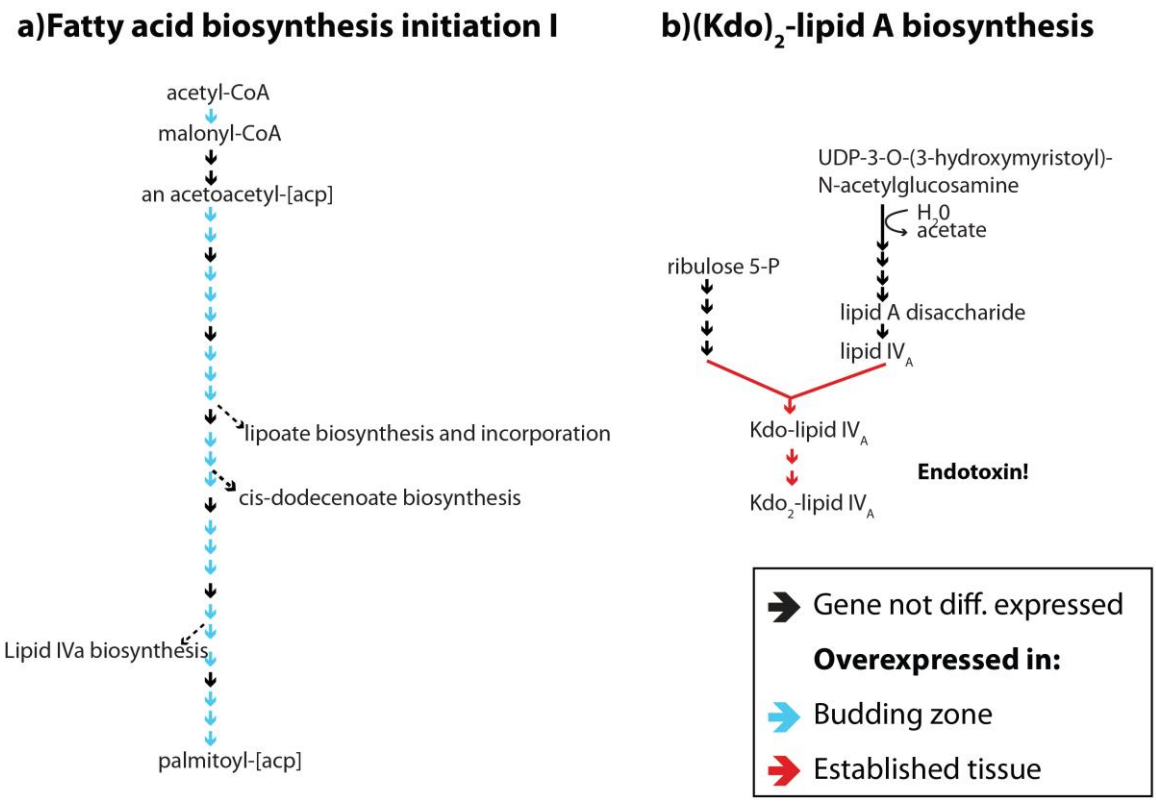


Figure 4. a) Fatty acid biosynthesis is overexpressed in the budding zone, while lipid A biosynthesis is overexpressed in the established tissue.

Establishing the winner: possible mechanisms for host-symbiont communication and for reducing strain diversity

If a significant proportion of the SOX symbiont population is acquired from the environment in the budding zone, genes related to symbiont-host recognition are expected to be overexpressed in this region. Indeed, bacterial genes that could mediate the direct interaction of adjacent cells through adhesion molecules were overexpressed: six integrins, a cadherin, and nine genes with discoidin domains (F5/8 type C). These genes were originally described in eukaryotes as mediators of cell-to-cell communication (Clark and Brugge, 1995; Seveau *et al.*, 2004; Springer *et al.*, 1984). The signaling proteins for host-bacteria interactions are often similar to those used for cell-to-cell communication in eukaryotes (Foster *et al.*, 2000; Hughes and Sperandio, 2008). Thus, the overexpression of eukaryotic-like genes in the budding zone might reflect the activity of the SOX symbionts to actively colonize and maintain the symbiosis.

To invade and communicate with their host, bacteria can produce different microbial-associated molecular patterns (MAMPs). Lipopolysaccharides (LPS) and peptidoglycan (PGN) are among the molecules that are recognized to play a major role in modulating the host immune response (Koropatnick, 2004; Nürnberger and Brunner, 2002; Pel and Pieterse, 2013). Two pathways involved in LPS biosynthesis were overexpressed in older gill filaments by the SOX symbiont: i) 3-deoxy-d-*manno*-octulosonic acid $(Kdo)_2$-lipid A biosynthesis (Figure 3 and 4b) and ii) synthesis of the monosaccharide dTDP-N-acetylthomosamine. In the same gill region, two enzymes of the PGN biosynthesis pathway were overexpressed: i) UDP-N-acetylmuramoylalanyl-D-glutamate-2,6,-diaminopimelate ligase (E:C: 6.3.2.13) and ii) UDP-N-acetylmuramyl pentapeptide phosphotransferase/ phospho-N-acetylmuramoyl-pentapeptide-transferase (E.C. 2.7.8.13). LPS and PGN can induce extensive morphological changes in animals. For example, the LPS and PGN molecules of beneficial symbionts and pathogens induce the loss of microvilli and ciliated appendages to prevent further recognition, attachment and invasion by other bacteria (Gilbert *et al.*, 2015; Nyholm and McFall-Ngai, 2004). This phenomenon has been observed in several host-microbe associations: intracellular pathogens such as *Yersinia* and *Listeria* induce microvilli loss after invasion; the beneficial *Vibrio fischeri* induces gradual loss of cilia from the bobtail squid once inside the light organ crypts; and microbial symbionts of tubeworms induces the loss of ciliated structures in the mouth region after colonization (Bhavsar *et al.*, 2007; Cossart and Sansonetti, 2004; Nyholm and McFall-Ngai, 2004; Southward, 1988). *Bathymodiolus* mussels also lose the microvilli in bacteriocytes that are fully colonized (Wentrup *et al.*, 2014). In a similar way to other intracellular pathogens and symbionts, LPS and PGN overexpression from the SOX could be triggering the loss of microvilli in the bacteriocytes that are fully colonized to avoid host recognition and invasion by other bacteria. This mechanism could prevent the infection of potential pathogens. For example, the intranuclear *Candidatus* Endonucleobacter bathymodioli sometimes infect *Bathymodiolus* mussels (Zielinski *et al.*, 2009). This pathogen invades the cell nuclei of the mussel, multiplies, and eventually bursts the host cell, but is never found in bacteriocytes. To first recognize the host, *Ca*. E. bathymodioli might use the host cells microvilli. Thus, the loss of microvilli induced by the SOX symbiont would be safeguarding the host bacteriocytes from the pathogen infection.

To respond to the environment provided inside the host bacteriocytes, bacteria need to receive, process and respond to the chemical signals. In the established tissue, two

genes involved in cyclic AMP (cAMP) signal transduction were overexpressed: i) 3',5'-cyclic AMP phosphodiesterase CpdA, which modulates cAMP internal concentration, and ii) a gene containing a cAMP binding domain. cAMP is a second messenger that can trigger the coordinated activity of different intracellular signaling pathways (Camilli and Bassler, 2006; Kalia *et al.*, 2012). Bacteria use cAMP levels to control virulence gene expression, manipulate the host immunity, form biofilms, and regulate catabolic operons for the use of alternative carbon sources (reviewed by Kalia *et al.*, 2012; McDonough and Rodriguez, 2012). For example, cAMP signaling is used by *Vibrio cholera* to respond to carbon availability, biofilm formation, quorum sensing, virulence expression and phage resistance (McDonough and Rodriguez, 2012). Thus, the change in cAMP concentration by the SOX symbiont in the established tissue could be triggering the activation and repression of several pathways to respond to the environment provided by the host.

In fact, the SOX symbiont overexpressed two classes of toxin-related genes in the established tissue whose expression could be responding to the change in cAMP concentration: two multifunctional autoprocessing repeats-in-toxin genes (MARTX) and 12 YD repeat genes. In agreement, the expression of cAMP has been linked to the activity of MARTX from *Vibrio vulnificus* (Ziolo *et al.*, 2014). Moreover, genes with the domain spvB, often found in YD repeat genes of the SOX symbiont, are known to respond to the expression of cAMP (O'Byrne and Dorman, 1994). Although the role of these toxin-related genes (TRGs) in the symbiosis is not fully understood, it has been hypothesized that the symbionts can use YD genes for bacteria-bacteria competition or parasite defense, and MARTX may have a role in attachment (Sayavedra *et al.*, 2015). These two classes of TRGs are preferentially found in genomes of host-associated bacteria (Sayavedra *et al.*, 2015). Thus, the overexpression of these TRGs could be a population response of the SOX symbiont to coordinate their behavior when the symbiont population increases in the established tissue and to distribute the resources provided by the host inside the bacteriocytes.

The MOX symbiont has a subtle change of expression

The MOX overexpressed four genes in the budding zone and 67 in the established tissue. The four genes overexpressed in the budding zone were involved in replication, transcription, translation and fatty acid biosynthesis. Nonetheless, more genes from all four of these categories were overexpressed in the established tissue.

In the established tissue, three enzymes related to the production of cobalamin or vitamin B_{12} were overexpressed: an enzyme for the synthesis of cob(II)yrinate a,c-diamide from cobyrinate, adenosylcobyric acid synthase, and cobalamin biosynthesis protein. The SOX symbiont does not have the full pathway to synthesize coenzyme B_{12}, but it can import cobalamin and convert it to coenzyme B_{12} (Table 2). Thus, the overexpression of the pathway by the MOX symbiont could be a response to the relative higher number of SOX cells in the established tissue in order to provide them with cobalamin (Table 2). The MOX needs to interact with the SOX symbiont, but also with its host. The host also requires cobalamin, as so far no eukaryote has been described to produce it (McDowell, 2012; Roth *et al.,* 1996). The MOX symbiont could, therefore, be also synthesizing the cobalamin for both the SOX symbiont and the host.

To communicate with the host, bacteria commonly export protein effectors (Bhavsar *et al.*, 2007). The MOX symbionts overexpressed a type II secretion system (T2SS) in the established tissue, which is used for protein secretion from the cytoplasm to the extracellular space and/or the host (Cianciotto, 2005). T2SS can promote bacterial growth depending on the environmental niche. For example, the pathogen *Legionella pneumophila* needs T2SS for intracellular growth (Soderberg *et al.*, 2004). The proteins exported by the MOX symbiont T2SS might be important to maintain the symbiosis and control their growth inside the bacteriocyte. As the SOX and MOX populations increase their numbers in the established tissue, the space will become a limiting resource for which they will need to compete and coordinate their behavior. Thus, the effectors exported by the MOX with the help of the T2SS in the established tissue might help to regulate their numbers.

Metabolite signature of budding zone and established tissue

Based on the transcriptome results, several metabolites should be in different abundances between the budding zone and the established tissue. Thus, we used GC-MS-based metabolomics to compare the abundance of metabolites between budding zone and the established tissue. Out of 65 detected compounds, 44 were significantly more abundant in the budding zone (P-values of Mann-Whitney test are shown Supplementary Table 3). Oxaloacetate was significantly more abundant in the budding zone (Supplementary Figure 3), which is in agreement with the transcriptome results from the host displaying pyruvate carboxylase upregulated. We were also able to detect 11 fatty

acids more abundant in the budding zone, from which four were derivatives of palmitate: palmitoleic acid, cis-9-hexadecanoic acid, n-hexadecanoic acid and palmitelaidic acid, which matches the overexpression of fatty acids from the SOX symbiont in the budding zone. The high abundance of fatty acids and oxaloacetate might indicate that the budding zone has a high-energy demand.

Highly expressed mobile elements and viral sequences in established tissue

Mobile elements were among the most highly overexpressed genes in both symbionts. The SOX symbiont overexpressed three restriction-modification systems, ten transposases, and five integrases in the established tissue, compared to one restriction-modification system, six transposases, two integrases and one phage-associated protein in the budding zone. The MOX overexpressed nine transposases and one integrase in the established tissue. Mobile elements are involved in gene deletion, gene duplications, genome rearrangements, and horizontal gene transfer (Frost *et al.*, 2005). Kleiner *et al*. (2013) hypothesized that the high expression of mobile elements can lead to an expansion of mobile elements in the genome and eventually could help in the gene acquisition to adapt to the host environment. This raises the intriguing possibility that the symbionts reshuffle their genomes and exchange genetic material at a high rate even when they are in a confined space inside the bacteriocytes.

Viruses are a potential source of new genetic material to the host and to the symbiont populations (Suttle, 2007). We found 67 genes in the metatranscriptome assembly that had the best taxonomic affiliation to viruses, of which 23 were overexpressed in the established tissue (Figure 3). Most of the sequences were annotated as capsid proteins, except for two replication initiation proteins and a hypothetical protein. The expression of capsid genes might indicate that the viruses are moving from a lysogenic to a lytic stage (reviewed by Feiner *et al.*, 2015; Molineux and Panja, 2013). Viruses could i) help in the gene transfer between species (bacteria-host or bacteria-bacteria), and ii) could shape the composition and abundance of the symbiont microbial population at the strain and species level (Suttle, 2007; Weinbauer and Rassoulzadegan, 2004). If viruses reach the lytic stage, they might carry eukaryotic or bacterial genes that could be transferred to other symbiont cells. When bacteria of the same kind are very abundant, viruses that are host-specific can be major players in the control of the

microbial composition (Weinbauer and Rassoulzadegan, 2004). According to the model 'killing the winner', the most abundant bacteria are more prone to viral infections, which would be the case for a fully colonized gill tissue. As for SUP05 -the free-living relatives of the SOX symbionts-, viruses could even provide new metabolic traits that could eventually provide a selective advantage to some SOX symbiont strains (Anantharaman *et al.*, 2014). Viruses could, therefore, have a major role in the control of the symbiont population.

Discussion

Is the SOX symbiont adjusting to a new environment in the budding zone, while the MOX is not?

Our goal was to identify potential molecular mechanisms involved in symbiosis establishment. To understand the symbiont response to the host in the new filaments, we first aimed to find whether the symbionts in the budding zone are acquired through self-infection or from the environment. The SOX symbiont genome variation is different in the budding zone compared to the established tissue of the same individual (Supplementary Figure 1). This suggests that the mussel does not use self-infection as the main source of the symbiont population in the budding zone. If self-infection is not the main source of symbiont acquisition, then the budding zone must be taking up most of the SOX symbionts from the surrounding fluids.

Our results showed that the SOX symbiont had a distinct profile of gene expression between the budding zone and the established tissue. The SOX symbiont is thought to depend on some external organic acids because it lacks the enzymes malate dehydrogenase and succinate dehydrogenase of the TCA cycle, which would allow the symbiont to produce essential metabolic intermediates (see below- Does the SOX symbiont take advantage of host metabolites?) (Ponnudurai *et al*., 2016). In the free-living stage of the symbiont, the metabolism of the symbiont is probably slower, as the symbiont will be in a state of nutrient limitation. When the free-living symbionts finally have access to the host, they might have a population behavior similar to the delay before exponential growth of bacteria when they enter a new niche, also known as the lag phase.

Indeed, genes that are typically overexpressed during the lag phase were overexpressed by the SOX symbiont in the budding zone. During the lag phase, bacteria need to prepare

the machinery for cell division. In this stage, bacteria overexpress ribosomal proteins (Martin, 1932). For example, *Salmonella enterica* serovar Typhimurium overexpresses ribosomal proteins during the lag phase when nutrients become available (Rolfe *et al.*, 2012). During the stationary phase, bacteria accumulate different kinds of protein damage, such as oxidation of amino acid residues, protein carbonylation, and aberrant disulfide bond formation (Dukan and Nyström, 1998). This protein deterioration is repaired by bacteria during the lag phase (Rolfe *et al.*, 2012). The aberrant disulfide bonds of cytosolic proteins can be repaired with the help of disulfide reductases like thioredoxins and glutaredoxin (Rolfe *et al.*, 2012). Moreover, RNA polymerase core subunits peak their expression during lag-phase (Rolfe *et al.*, 2012). The SOX symbiont overexpressed these genes in the budding zone, which suggests that they must adapt to a new nutrient-rich niche within the host. This supports our hypothesis that the budding zone acquires the SOX symbiont population mainly from the environment.

In contrast to the SOX symbiont, the MOX did not show a difference in the genomic variation between the budding zone and the established tissue. These results can be explained under the following scenarios: i) The MOX symbiont may preferentially colonize the host at an early ontogenetic stage, and subsequently take up symbionts from the older filaments by self-infection, or ii) The genetic pool of strains available from the environment is low and therefore we are not able to detect a significant genomic variation between the budding zone and the established tissue. The transcriptome profile of the MOX symbiont showed only a few genes differentially expressed between the budding zone and the established tissue, which suggests that the MOX symbionts acquired in the budding zone were active and might have experienced similar conditions before the colonization. Thus, self-infection (scenario i) is more likely since the MOX symbionts present in the budding zone and in the established tissue have a similar genetic variation and do not have a drastic change in expression.

The role of carbon in the SOX-host symbiosis

The analysis of differentially expressed genes between budding zone and established tissue showed that the SOX symbiont overexpressed the carbon metabolism in the budding zone. As shown with our SNP analyses, the genome heterogeneity of the SOX symbiont is higher in the budding zone, which suggests that a higher number of SOX strains are present. The space competition between different SOX strains in the

newly formed filaments would not leave a place for social cheaters, as otherwise the competitiveness of the whole SOX population would decrease (Hibbing *et al.*, 2010; Travisano and Velicer, 2004). To maintain the cooperative behavior of the symbionts and get rid of potential cheaters, the host might reward the beneficial SOX strains while it might sanction the cheaters. This evolutionary model is named 'partner choice' and is thought to play a major role in maintaining the cooperative behavior of horizontally transmitted symbionts (e.g. squid-vibrio and legume-rhizobia symbioses) (Foster and Wenseleers, 2006; Kiers *et al.*, 2003; Koch *et al.*, 2014; Sachs *et al.*, 2011). The major role of chemoautotrophic symbionts is to fix inorganic carbon into organic compounds that the host uses as a source of nutrition (Cavanaugh *et al.*, 2013). The host could be selecting for those SOX symbionts that are actively fixing carbon and producing 'public goods' for the holobiont. Considering that part of the fixed carbon could act as 'public goods', this raises the possibility that the SOX symbiont cooperator-strains are overexpressing the carbon metabolism as a way to show that they are worth keeping in symbiosis. This mechanism might be triggered at a higher rate in the budding zone due to the presence of a higher number of SOX strains contained within the bacteriocytes.

To understand the role of the carbon metabolism in infection, we need to compare the *Bathymodiolus* symbiosis with the better-studied pathogens and their mechanisms for host interaction. Bacterial pathogens drastically change their carbon metabolism when they are in contact with their host compared to free-living stages (reviewed in Eisenreich *et al.*, 2010; Munoz-Elias and McKinney, 2006). In contrast to the *Bathymodiolus* symbionts, most pathogens are heterotrophs. In *Mycobacterium tuberculosis*, carbon metabolism determines the ability of replication and persistence within the host (Marrero *et al.*, 2010). *M. tuberculosis* and *Salmonella enterica* preferentially use fatty acids as a substrate for *in vivo* growth (Munoz-Elias and McKinney, 2006). *Vibrio cholera* induces the expression of fatty acid degradation, glycerol transport, and metabolism at a higher rate when it is host-associated compared to when it is in culture (Mandlik *et al.*, 2011). This behavior is most likely a response of *V. cholera* to utilize host-derived lipids. In the case of the SOX symbiont, rather than using fatty acids from the host for catabolism, it seems like they do the opposite and synthesize the fatty acids at a higher rate in the budding zone, which could be taken up by the host. Thus, the high abundance of fatty acids detected by GC-MS in the budding zone might be explained by the activity of the SOX symbiont. This would be one major difference to pathogens: instead of using fatty acids from the host, the symbionts are synthesizing their own. Indeed, the ß-oxidation

cycle, which is the dominant route for oxidative degradation of fatty acids in bacteria and eukaryotes, is not present in the *Bathymodiolus* symbionts or the vesicomyid clam symbionts. In contrast, the closest free-living relative of the SOX symbiont, SUP05, does have the ß-oxidation cycle that would allow the free-living bacteria to use fatty acids from the environment (Walsh *et al.*, 2009). Even if the ß-oxidation cycle is not present, fatty acids are still required for the generation of the building blocks of the membrane, which are essential to produce new cells (Fujita *et al.*, 2007). This raises the next question: do SOX symbiont cells divide at a higher rate in the budding zone or the established tissue?

The rRNA content of a cell does not have a linear correlation to the growth rate in all bacteria, and thus, cannot be used to estimate growth rates (Blazewicz *et al.*, 2013). A second alternative is to assess where genes for cell division are overexpressed. Like the vesicomyid SOX symbionts, the *Bathymodiolus* SOX symbionts do not have the key gene for cell division, *ftsZ*. In most bacteria, FtsZ polymerizes to form a ring-like structure, which attracts other cell division proteins like FtsA (Madigan *et al.*, 2006). *FtsA* is encoded in the genome and was found overexpressed in the established tissue, suggesting that the SOX symbiont divides at a higher rate in the established tissue. If the SOX symbionts are indeed dividing more in the established tissue, the overexpression of fatty acids in the budding zone might reflect the cell preparation for cell division. At the same time, the host could be exploiting some of the SOX symbionts, not only with fixed carbon in the form of short organic acids but also with long-chain fatty acids during the expensive process of growth in the budding zone.

Does the SOX symbiont take advantage of host metabolites?

The symbiosis with the SOX symbiont does not only benefit the host, but may also benefit the symbiont. The genomes of the SOX symbionts of *B. azoricus* mussels lack the enzyme malate dehydrogenase (E.C. 1.1.1.37), which catalyzes the reversible reaction of malate to oxaloacetate (Sayavedra *et al.*, 2015, Ponnudurai *et al*., 2016). To close the TCA cycle for generation of redox equivalents and biomass precursors, the SOX symbiont needs to produce or import oxaloacetate. In order to produce oxaloacetate, the enzyme aspartate transaminase (E.C. 2.6.1.1), which catalyzes the reversible conversion of oxaloacetate to aspartate, is needed. The enzyme is present but as this reaction is not favored in the direction of oxaloacetate, the symbionts might also take up TCA

intermediates such as oxaloacetate from its host when available. In the budding zone, the host overexpressed pyruvate carboxylase, which mediates the carboxylation of pyruvate to oxaloacetate. Metabolite analyses confirmed that oxaloacetate was more abundant in the budding zone. The host might be overproducing oxaloacetate at a higher rate in the budding zone to provide the SOX symbiont when the symbiont expresses more its incomplete TCA cycle (e.g. the budding zone). The required oxaloacetate needs to be imported. The SOX symbiont genome does not encode a dedicated oxaloacetate transporter. Alternatively, the SOX symbionts might use a citrate transporter as was suggested for the vesicomyid SOX symbionts and for the promiscuous citrate transporters of other organisms (Groeneveld *et al.*, 2010; Newton *et al.*, 2008; Pudlik and Lolkema, 2011). Interestingly, the SOX symbionts overexpressed in the budding zone the synthesis of biotin. Biotin is a cofactor for carboxylase enzymes, involved in several pathways like fatty acids biosynthesis and gluconeogenesis, which would explain why the SOX symbiont overexpressed biotin production (Waldrop *et al.*, 2012). Biotin is also a cofactor for the host's pyruvate carboxylase. Biotin biosynthesis has only been found in yeast and plants, but no animal has been found that can do so (McDowell, 2012; Streit and Entcheva, 2002). Thus, the overexpression of biotin by the SOX symbionts might induce the host to produce more oxaloacetate that in turn the SOX symbiont could import to close its TCA cycle. This would imply a remarkably intimate metabolic coupling between the SOX symbiont and its host (Figure 3).

Conclusions

Our combined analyses using metagenomics, metatranscriptomics and metabolomics allowed us to identify the role that the host and the two symbionts are playing in the budding zone compared to the established tissue. Our analyses showed that the budding zone does not acquire most of the SOX symbiont population through self-infection but rather from a metabolically less active population present in the environment. Once the symbiosis is established, the host and the SOX symbiont express in synchrony pathways that produce metabolites that could benefit both partners. Further comparative transcriptomic analyses are necessary to investigate whether there is a gradual adaptation in the expression profile from the budding zone to the oldest filaments.

Outlook

When do the SOX symbionts divide?

One of the most basic questions to answer regarding the host-symbiont colonization is “When do the symbionts divide?” To further investigate where the SOX symbionts divide the most, we plan to analyze the morphology of the symbiont cells, as well as to use immunocytochemistry with an antibody targeting FtsA as described by Leisch *et al*. (2012). These techniques will be done in collaboration with N. Leisch.

Methods

Sampling and processing of Bathymodiolus mussels

Bathymodiolus mussels from the Lucky Strike vent site on the MAR were sampled from two very close sites: Eiffel Tower (37° 16' 58.512"N, 32° 16' 32.196"W) and Montsegur (37° 17' 17.016"N, 32° 16' 32.052"W) during the Biobaz Cruise. The gills of the mussels were dissected and either frozen immediately at -80°C or fixed on board in RNAlater (Sigma®, Germany) according to the manufacturer instructions and stored at -80°C. An overview of the samples used in this study is shown in Table Supplementary Table 2.

Genome and transcriptome sequencing

To extract the total RNA and DNA, we dissected the budding zone, and a piece in the mid-region of five mussel’s gills. The tissue was stored overnight in RNAlater (Sigma®) at 4°C. RNA and DNA were extracted with AllPrep DNA/RNAMini Kit (Qiagen, Germany) according to the manufacturer’s instructions. To get rid of the cell debris and to improve RNA yield, we used QIAshreder Mini Spin Columns (Qiagen). The quality of the RNA was assessed with Agilent 2100 Bioanalyzer. The RNA was used for cDNA synthesis with the Ovation RNA-Seq System V2 (NuGEN, San Carlos, CA, USA). Libraries of the cDNA and genomic DNA were generated with DNA library prep kit for Illumina (Biolabs, Germany). We sequenced the cDNA libraries as paired-end reads of 150 bp, and the genomic DNA as paired-end reads of 100 bp on an Illumina HiSeq 2500 platform in the Max Planck Genome Centre (Cologne).

Genome assembly of the sulfur-oxidizing (SOX) and methane-oxidizing (MOX) symbionts

An initial assembly of each *Bathymodiolus* individual was done with IDBA_UD (Peng *et al.*, 2012). The symbiont genomes were binned based on the genome coverage, GC content, and taxonomic affiliation with GBtools (Albertsen *et al.*, 2013; Seah and Gruber-Vodicka, 2015; Wang and Wu, 2013). We reassembled the MOX and SOX symbiont genomes from the individuals for which the coverage of the SOX or the MOX was the highest. For reassembly, reads were mapped to the bins and assembled with Spades v.3.5 (Bankevich *et al.*, 2012). The quality metrics and completeness of the resulting assemblies were estimated with Quast and CheckM (Marker lineage for SOX symbiont: p_Proteobacteria, UID4274; MOX c_Gammaproteobacteria, UID4274) (Gurevich *et al.*, 2013; Parks *et al.*, 2015). Genomes were annotated with the IMG pipeline (Markowitz *et al.*, 2011). The SOX and MOX symbiont genomes will be submitted to the ENA repository upon acceptance of the manuscript.

Genome coverage of the MOX and SOX symbionts was estimated with bbmap v33 (BBMap tools, Bushnell B., sourceforge.net/projects/bbmap) and a custom R script to calculate median coverage for multiple files. The median coverage was normalized to sequencing depth with the following formula:

$$Median\ cov.\ x\ million\ reads = \left(\frac{Median\ cov.\ of\ all\ scaffolds}{Number\ of\ reads}\right) * 1{,}000{,}000$$

Single nucleotide polymorphisms of the symbionts

Sequencing reads (Phred quality score > 20) were mapped with bbmap V33 to the corresponding genome assembly for each host individual and tissue type separately. PCR duplicates were removed using samtools (Li *et al.*, 2009). The reads were realigned around InDels using GATK IndelRealigner (DePristo *et al.*, 2011; McKenna *et al.*, 2010). Subsequently, the read coverage was adjusted for all samples to 45x (SOX symbiont) and 8x (MOX symbiont) with samtools to ensure comparability among samples (Li *et al.*, 2009). Thresholds were chosen according to the samples with the lowest read coverage per symbiont type. SNPs and InDels were called with GATK HaplotypeCaller with a ploidy setting 10 and subsequent quality filtering of the variants (DePristo *et al.*, 2011; McKenna *et al.*, 2010). SNPs were annotated according to their effects on coding genes

using snpEff and GATK VariantAnnotator (Cingolani *et al.*, 2012). To compare the number of SNPs between the two gill regions for the SOX and MOX symbionts, we used Kruskal-Wallis test implemented in PAST (Hammer *et al.*, 2001).

To know whether the SOX symbiont population in the budding zone was more alike within than between individuals, we used Varan (Malde, 2014). Varan outputs how many times a specific base is mapping to a position of the genome for every sample. Using a self-written Perl script, we created a matrix that has in each row the frequency of a nucleotide base, without considering the positions for which all samples had the same nucleotide base. This matrix was used to reconstruct a principal component analysis with the R packages vegan and RDA (Oksanen *et al.*, 2011).

Differential expression of the SOX and MOX symbionts

We removed adaptors, ribosomal genes of the symbionts and the host, as well as the mitochondria from transcriptome reads with bbduk v33. The proportion of reads mapping to the small ribosomal subunits was estimated with Phyloflash 2.0 (https://github.com/HRGV/phyloFlash). Statistics about the number of mRNA reads mapping to the SOX and MOX symbionts were estimated with bbsplit v33. Differential gene expression of the SOX and MOX symbionts was done using a genome-based approach, for which we used as reference the aforementioned genome assemblies. To estimate the amount of mRNA of each symbiont, transcriptome reads were mapped to the symbiont reference genomes per sample with bbmap v33 using a minimum identity of 98%. Transcriptome reads were downsampled to have a similar number of transcriptome reads in all ten samples depending on the SOX or MOX symbiont mRNA abundance and re-mapped to the reference genomes with reformat and bbmap v33 (sourceforge.net/projects/bbmap). The number of transcripts per gene was estimated with featureCounts (Liao *et al.*, 2014). Differential gene expression analysis was done using edgeR with trimmed mean of M values (TMM) normalization (Robinson and Oshlack, 2010; Robinson *et al.*, 2010).

De novo transcriptome assembly and differential expression of the host

We *de novo* assembled the transcriptome to analyze genes differentially expressed by the *Bathymodiolus* host. Reads were normalized to a maximum of 500X coverage and error corrected with bbnorm v33. The transcriptome was assembled with Trinity

r20140413p1 (Grabherr *et al.*, 2011). We removed poorly supported transcripts with an isoform percentage value less than five (Grabherr *et al.*, 2011) and duplicate sequences at a 98% threshold with dedupe v33. To remove bacterial symbiont transcripts, we used blast against the symbiont draft genomes (e-value = 1 x e^{-5}). Metrics of the resulting transcriptome are shown in Supplementary Table 1. All transcript were searched against the nr database with DIAMOND (Buchfink *et al.*, 2014). Transcripts were taxonomically classified with MEGAN5 based on the diamond blast results (Albertsen *et al.*, 2013; Huson *et al.*, 2011). Differential gene expression analysis was done with those transcripts that did not have a significant blast hit against the symbiont reference sequences as described by Haas *et al*. (2013), with at least a 3-fold change in expression. Transcripts were annotated and gene ontology terms assigned with Blast2GO (Conesa *et al.*, 2005, 2).

GC-MS

Five *B. azoricus* individuals were collected from Lucky Strike, Montsegur at 37°2836'N, 32°16.534W. The gills were dissected and stored at -80°C until further processing. To dissect the budding zone tissue, the gill was placed in chilled methanol on top of dry ice. Tissue was transferred to 2 ml tubes containing 1 g of 1 mm zirconia beads to enhance tissue disruption and to ensure a standardized mixing procedure. The pre-cooled extraction solvent mixture acetonitrile, methanol, water (2:1:1 vol./vol.) was then added directly to the tube. All samples were extracted twice sequentially, with the first extraction having 1 ml AMW added and bead beating in a FastPrep (MP Biomedicals) at 4M*s^{-1} for 40 s. Samples were centrifuged at 10000 g for 1 min and the resulting supernatant was transferred into a 2 ml reaction tube. The extraction was repeated by adding 750 µl solvent mixture to the pellet and repetition of the bead beating and centrifugation step. Supernatants of each extraction step were combined and dried under vacuum (Concentrator Plus, Eppendorf) at 30^{o}C for 3 h. The dried sample was derivatized first with 80 µl methoxyamine in pyridine (20mg*ml^{-1}) for 90 min at 37^{o}C and constant agitation at 1250 rpm and second with 100 µl N,O-Bis(trimethylsilyl)trifluoroacetamide for 30 min at 37^{o}C and 1250 rpm. Samples were centrifuged for 1 min and the supernatant transferred to GC-MS vials and analyzed with an Agilent 7890B Gas-chromatograph coupled with a 7693 autosampler and a 5944A mass selective detector. Separation was done on a DB-5MS+DG column with a length of 30 m, a diameter of 0.25 mm and a film size of 0.25 µm. The injector temperature was set at 290^{o}C. The initial

temperature of the oven was 60°C without hold, followed by a ramp of 20°C*min^{-1} to 325°C and hold for 2 min. The solvent delay was set to 7.25 min. Helium was used as a carrier gas with a fixed flow rate of 1ml*min^{-1}. The quadrupole mass spectrometer was operated in electron ionization mode at 70 eV and the scanning range was set to 50-600 m/z. Metabolites were assigned using the NIST 2.0 Library and verified by injection of pure compounds under same conditions. We used the Mann-Whitney-Wilcoxon test to compare whether the compounds were more abundant in a specific gill region. Statistical analyses were done in R.

Acknowledgements

We thank the captain and crew of the BioBaz Cruise (2013) involved in the sampling effort. We especially thank Adrien Assié and Christian Borowski for sample collection. We thank Erik Puskas for measuring the samples on the GC-MS, Cecilia Wentrup for her help with mussel dissection, and Brandon Seah and Harald Gruber-Vodicka for helpful discussions. We thank the Max Planck Society for funding the project and the DAAD for funding to L.S.

Statement of competing interests

The authors declare no competing interests.

References

Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**: 533–538.

Alonso CR. (2002). Hox proteins: Sculpting body parts by activating localized cell death. *Curr Biol* **12**: R776–R778.

Anantharaman K, Duhaime MB, Breier JA, Wendt KA, Toner BM, Dick GJ. (2014). Sulfur oxidation genes in diverse deep-sea viruses. *Science* **344**: 757–760.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, *et al.* (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**: 455–477.

Bhavsar AP, Guttman JA, Finlay BB. (2007). Manipulation of host-cell pathways by bacterial pathogens. *Nature* **449**: 827–834.

Blazewicz SJ, Barnard RL, Daly RA, Firestone MK. (2013). Evaluating rRNA as an indicator of microbial activity in environmental communities: Limitations and uses. *ISME J* **7**: 2061–2068.

Buchfink B, Xie C, Huson DH. (2014). Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**: 59–60.

Camilli A, Bassler BL. (2006). Bacterial small-molecule signaling pathways. *Science* **311**: 1113–1116.

Cavanaugh CM, McKiness ZP, Newton ILG, Stewart FJ. (2013). Marine chemosynthetic symbioses. In: Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F (eds). *The Prokaryotes*. Springer Berlin Heidelberg, pp 579–607.

Chen CS, Alonso JL, Ostuni E, Whitesides GM, Ingber DE. (2003). Cell shape provides global control of focal adhesion assembly. *Biochem Biophys Res Commun* **307**: 355–361.

Cianciotto NP. (2005). Type II secretion: A protein secretion system for all seasons. *Trends Microbiol* **13**: 581–588.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, *et al*. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; iso-2; iso-3. *Fly (Austin)* **6**: 80–92.

Clark EA, Brugge JS. (1995). Integrins and signal transduction pathways: The road taken. *Science* **268**: 233.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**: 3674–3676.

Cossart P, Sansonetti PJ. (2004). Bacterial invasion: The paradigms of enteroinvasive pathogens. *Science* **304**: 242–248.

DeChaine EG, Bates AE, Shank TM, Cavanaugh CM. (2006). Off-axis symbiosis found: Characterization and biogeography of bacterial symbionts of *Bathymodiolus* mussels from Lost City hydrothermal vents. *Environ Microbiol* **8**: 1902–1912.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, *et al*. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491–498.

Dubilier N, Bergin C, Lott C. (2008). Symbiotic diversity in marine animals: The art of harnessing chemosynthesis. *Nat Rev Micro* **6**: 725–740.

Dukan S, Nyström T. (1998). Bacterial senescence: Stasis results in increased and differential oxidation of cytoplasmic proteins leading to developmental induction of the heat shock regulon. *Genes Dev* **12**: 3431–3441.

Duperron S, Bergin C, Zielinski F, Blazejak A, Pernthaler A, McKiness ZP, *et al*. (2006). A dual symbiosis shared by two mussel species, *Bathymodiolus azoricus* and *Bathymodiolus puteoserpentis* (Bivalvia: Mytilidae), from hydrothermal vents along the northern Mid-Atlantic Ridge. *Environ Microbiol* **8**: 1441–1447.

Duperron S, Halary S, Lorion J, Sibuet M, Gaill F. (2008). Unexpected co-occurrence of six bacterial symbionts in the gills of the cold seep mussel *Idas* sp. (Bivalvia: Mytilidae). *Environ Microbiol* **10**: 433–445.

Eisenreich W, Dandekar T, Heesemann J, Goebel W. (2010). Carbon metabolism of intracellular bacterial pathogens and possible links to virulence. *Nat Rev Micro* **8**: 401–412.

Feiner R, Argov T, Rabinovich L, Sigal N, Borovok I, Herskovits AA. (2015). A new perspective on lysogeny: Prophages as active regulatory switches of bacteria. *Nat Rev Microbiol* **13**: 641–650.

Fontanez KM, Cavanaugh CM. (2014). Evidence for horizontal transmission from multilocus phylogeny of deep-sea mussel (Mytilidae) symbionts: Horizontal transmission of mussel symbionts. *Environ Microbiol* 3608–3621.

Foster JS, Apicella MA, McFall-Ngai MJ. (2000). *Vibrio fischeri* lipopolysaccharide induces developmental apoptosis, but not complete morphogenesis, of the *Euprymna scolopes* symbiotic light organ. *Dev Biol* **226**: 242–254.

Foster KR, Wenseleers T. (2006). A general model for the evolution of mutualisms. *J Evol Biol* **19**: 1283–1293.

Frost LS, Leplae R, Summers AO, Toussaint A. (2005). Mobile genetic elements: The agents of open source evolution. *Nat Rev Microbiol* **3**: 722–732.

Fujita Y, Matsuoka H, Hirooka K. (2007). Regulation of fatty acid metabolism in bacteria. *Mol Microbiol* **66**: 829–839.

Gilbert SF, Bosch TCG, Ledón-Rettig C. (2015). Eco-Evo-Devo: Developmental symbiosis and developmental plasticity as evolutionary agents. *Nat Rev Genet* **16**: 611–622.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, *et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**: 644–652.

Groeneveld M, Weme RGJDO, Duurkens RH, Slotboom DJ. (2010). Biochemical characterization of the C_4-dicarboxylate transporter DctA from *Bacillus subtilis*. *J Bacteriol* **192**: 2900–2907.

Gudas LJ. (2013). Retinoids induce stem cell differentiation via epigenetic changes. *Semin Cell Dev Biol* **24**: 701–705.

Gunina A, Dippold MA, Glaser B, Kuzyakov Y. (2014). Fate of low molecular weight organic substances in an arable soil: From microbial uptake to utilisation and stabilisation. *Soil Biol Biochem* **77**: 304–313.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**: 1072–1075.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, *et al.* (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**: 1494–1512.

Hammer Ø, Harper DAT, Ryan PD. (2001). PAST: Paleontological statistics software package for education and data analysis. Palaeontolia Electronica 4.

Hibbing ME, Fuqua C, Parsek MR, Peterson SB. (2010). Bacterial competition: Surviving and thriving in the microbial jungle. *Nat Rev Microbiol* **8**: 15–25.

Hughes DT, Sperandio V. (2008). Inter-kingdom signalling: Communication between bacteria and their hosts. *Nat Rev Microbiol* **6**: 111–120.

Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Res* **21**: 1552–1560.

Ikuta T, Takaki Y, Nagai Y, Shimamura S, Tsuda M, Kawagucci S, *et al.* (2015). Heterogeneous composition of key metabolic gene clusters in a vent mussel symbiont population. *ISME J*. e-pub ahead of print, doi: 10.1038/ismej.2015.176.

Jaenisch R, Bird A. (2003). Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nat Genet* **33**: 245–254.

Kalia D, Merey G, Nakayama S, Zheng Y, Zhou J, Luo Y, *et al.* (2012). Nucleotide, c-di-GMP, c-di-AMP, cGMP, cAMP, (p)ppGpp signaling in bacteria and implications in pathogenesis. *Chem Soc Rev* **42**: 305–341.

Kiers ET, Rousseau RA, West SA, Denison RF. (2003). Host sanctions and the legume–rhizobium mutualism. *Nature* **425**: 78–81.

Kleiner M, Young JC, Shah M, VerBerkmoes NC, Dubilier N. (2013). Metaproteomics reveals abundant transposase expression in mutualistic endosymbionts. *mBio* **4**: e00223-13-e00223-13.

Koch EJ, Miyashiro T, McFall-Ngai MJ, Ruby EG. (2014). Features governing symbiont persistence in the squid–vibrio association. *Mol Ecol* **23**: 1624–1634.

Koropatnick TA. (2004). Microbial factor-mediated development in a host-bacterial mutualism. *Science* **306**: 1186–1188.

Leisch N, Verheul J, Heindl NR, Gruber-Vodicka HR, Pende N, den Blaauwen T, *et al.* (2012). Growth in width and FtsZ ring longitudinal positioning in a gammaproteobacterial symbiont. *Curr Biol* **22**: R831–R832.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Liao Y, Smyth GK, Shi W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923–930.

Lo SH. (2004). Tensin. *Int J Biochem Cell Biol* **36**: 31–34.

Madigan MT, Martinko JM, Brock TD. (2006). Brock biology of microorganisms. Pearson Prentice Hall: Upper Saddle River, NJ.

Malde K. (2014). Estimating the information value of polymorphic sites using pooled sequences. *BMC Genomics* **15**: S20.

Mandlik A, Livny J, Robins WP, Ritchie JM, Mekalanos JJ, Waldor MK. (2011). RNA-seq-based monitoring of infection-linked changes in *Vibrio cholerae* gene expression. *Cell Host Microbe* **10**: 165–174.

Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, *et al.* (2011). IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res* **40**: D115–D122.

Marrero J, Rhee KY, Schnappinger D, Pethe K, Ehrt S. (2010). Gluconeogenic carbon flow of tricarboxylic acid cycle intermediates is critical for *Mycobacterium tuberculosis* to establish and maintain infection. *Proc Natl Acad Sci* **107**: 9819–9824.

Martin DS. (1932). The oxygen consumption of *Escherichia coli* during the lag and logarithmic phases of growth. *J Gen Physiol* **15**: 691–708.

McDonough KA, Rodriguez A. (2012). The myriad roles of cyclic AMP in microbial pathogens: From signal to sword. *Nat Rev Microbiol* **10**: 27–38.

McDowell LR. (2012). Vitamins in animal nutrition: Comparative aspects to human nutrition. Elsevier.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, *et al.* (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–1303.

Molineux IJ, Panja D. (2013). Popping the cork: Mechanisms of phage genome ejection. *Nat Rev Microbiol* **11**: 194–204.

Mollinedo F, Gajate C. (2003). Microtubules, microtubule-interfering agents and apoptosis. *Apoptosis* **8**: 413–450.

Munoz-Elias EJ, McKinney JD. (2006). Carbon metabolism of intracellular bacteria. *Cell Microbiol* **8**: 10–22.

Neumann D, Kappes H. (2003). On the growth of bivalve gills initiated from a lobule-producing budding zone. *Biol Bull* **205**: 73–82.

Newton I, Girguis P, Cavanaugh C. (2008). Comparative genomics of vesicomyid clam (Bivalvia: Mollusca) chemosynthetic symbionts. *BMC Genomics* **9**: 585.

Nürnberger T, Brunner F. (2002). Innate immunity in plants and animals: Emerging parallels between the recognition of general elicitors and pathogen-associated molecular patterns. *Curr Opin Plant Biol* **5**: 318–324.

Nyholm SV, McFall-Ngai M. (2004). The winnowing: Establishing the squid-vibrio symbiosis. *Nat Rev Microbiol* **2**: 632–642.

O'Byrne CP, Dorman CJ. (1994). The spv virulence operon of *Salmonella typhimurium* LT2 is regulated negatively by the cyclic AMP (cAMP)-cAMP receptor protein system. *J Bacteriol* **176**: 905–912.

Oksanen J, Guillaume Blanchet F, Kindt R, Legendre P, Minchin PR, O'Hara RB, *et al.* (2011). Vegan: Community ecology package. *R Package Version 20-0*. http://cran.r-project.org/web/packages/vegan/index.html.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**: 1043–1055.

Pel MJC, Pieterse CMJ. (2013). Microbial recognition and evasion of host immunity. *J Exp Bot* **64**: 1237–1248.

Peng Y, Leung HCM, Yiu SM, Chin FYL. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**: 1420–1428.

Petersen JM, Zielinski FU, Pape T, Seifert R, Moraru C, Amann R, *et al.* (2011). Hydrogen is an energy source for hydrothermal vent symbioses. *Nature* **476**: 176–180.

Pudlik AM, Lolkema JS. (2011). Mechanism of citrate metabolism by an oxaloacetate decarboxylase-deficient mutant of *Lactococcus lactis* IL1403. *J Bacteriol* **193**: 4049–4056.

Robinson MD, McCarthy DJ, Smyth GK. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140.

Robinson MD, Oshlack A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**: R25.

Rolfe MD, Rice CJ, Lucchini S, Pin C, Thompson A, Cameron ADS, *et al.* (2012). Lag phase is a distinct growth phase that prepares bacteria for exponential growth and involves transient metal accumulation. *J Bacteriol* **194**: 686–701.

Roth J, Lawrence J, Bobik T. (1996). Cobalamin (coenzyme B_{12}): Synthesis and Biological Significance. *Annu Rev Microbiol* **50**: 137–181.

Sachs JL, Skophammer RG, Regus JU. (2011). Evolutionary transitions in bacterial symbiosis. *Proc Natl Acad Sci* **108**: 10800–10807.

Sayavedra L, Kleiner M, Ponnudurai R, Wetzel S, Pelletier E, Barbe V, *et al.* (2015). Abundant toxin-related genes in the genomes of beneficial symbionts from deep-sea hydrothermal vent mussels. *eLife* **4**. e-pub ahead of print, doi: 10.7554/eLife.07966.

Scully R, Chen J, Plug A, Xiao Y, Weaver D, Feunteun J, *et al.* (1997). Association of BRCA1 with Rad51 in mitotic and meiotic cells. *Cell* **88**: 265–275.

Seah BKB, Gruber-Vodicka HR. (2015). Gbtools: Interactive visualization of metagenome bins in R. *Microb Physiol Metab* 1451.

Seveau S, Bierne H, Giroux S, Prévost M-C, Cossart P. (2004). Role of lipid rafts in E-cadherin–and HGF-R/Met–mediated entry of *Listeria monocytogenes* into host cells. *J Cell Biol* **166**: 743–753.

Soderberg MA, Rossier O, Cianciotto NP. (2004). The type II protein secretion system of *Legionella pneumophila* promotes growth at low temperatures. *J Bacteriol* **186**: 3712–3720.

Southward EC. (1988). Development of the gut and segmentation of newly settled stages of *Ridgeia* (Vestimentifera): Implications for relationship between Vestimentifera and Pogonophora. *J Mar Biol Assoc U K* **68**: 465.

Springer WR, Cooper DNW, Barondes SH. (1984). Discoidin I is implicated in cell-substratum attachment and ordered cell migration of dictyostelium discoideum and resembles fibronectin. *Cell* **39**: 557–564.

Streit WR, Entcheva P. (2002). Biotin in microbes, the genes involved in its biosynthesis, its biochemical role and perspectives for biotechnological production. *Appl Microbiol Biotechnol* **61**: 21–31.

Suttle CA. (2007). Marine viruses - major players in the global ecosystem. *Nat Rev Microbiol* **5**: 801–812.

Travisano M, Velicer GJ. (2004). Strategies of microbial cheater control. *Trends Microbiol* **12**: 72–78.

Urvalek A, Laursen KB, Gudas LJ. (2014). The roles of retinoic acid and retinoic acid receptors in inducing epigenetic changes. In: Asson-Batres MA, Rochette-Egly C (eds) Vol. 70. *The Biochemistry of retinoic acid receptors I: Structure, activation, and function at the molecular level*. Springer Netherlands: Dordrecht, pp 129–149.

Van Dover CL. (2002). Evolution and biogeography of deep-sea vent and seep invertebrates. *Science* **295**: 1253–1257.

Vrijenhoek RC. (2010). Genetics and evolution of deep-sea chemosynthetic bacteria and their invertebrate hosts. In: Kiel S (ed) Vol. 33. *The vent and seep biota*. Springer Netherlands: Dordrecht, pp 15–49.

Waldrop GL, Holden HM, Maurice MS. (2012). The enzymes of biotin dependent CO_2 metabolism: What structures reveal about their reaction mechanisms. *Protein Sci Publ Protein Soc* **21**: 1597–1619.

Walsh DA, Zaikova E, Howes CG, Song YC, Wright JJ, Tringe SG, *et al.* (2009). Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. *Science* **326**: 578–582.

Wang Z, Wu M. (2013). A phylum-level bacterial phylogenetic marker database. *Mol Biol Evol* **30**: 1258–1262.

Weinbauer MG, Rassoulzadegan F. (2004). Are viruses driving microbial diversification and diversity? *Environ Microbiol* **6**: 1–11.

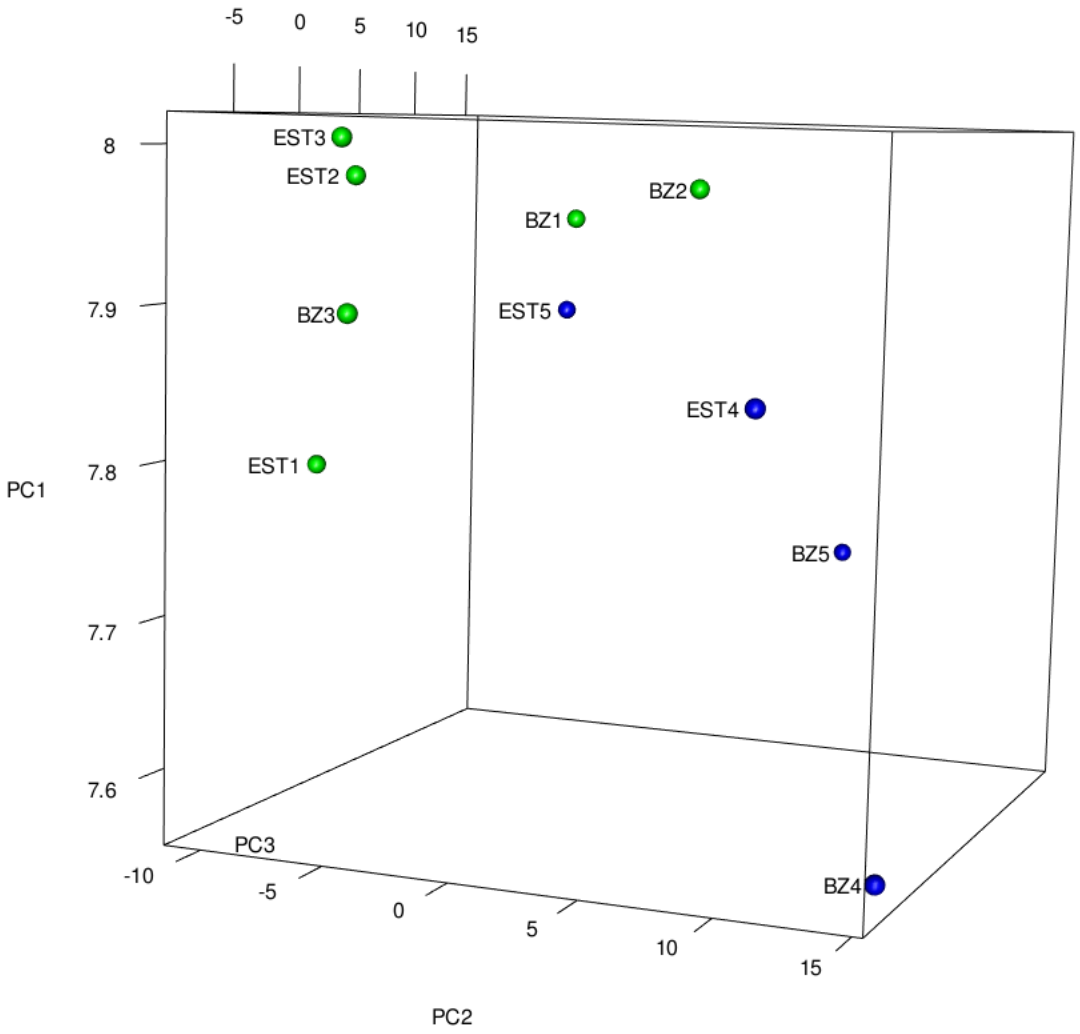
Wentrup C, Wendeberg A, Schimak M, Borowski C, Dubilier N. (2014). Forever competent: Deep-sea bivalves are colonized by their chemosynthetic symbionts throughout their lifetime. *Environ Microbiol* 3699–3713.

Won YJ, Hallam SJ, O'Mullan GD, Pan IL, Buck KR, Vrijenhoek RC. (2003). Environmental acquisition of thiotrophic endosymbionts by deep-sea mussels of the genus *Bathymodiolus*. *Appl Environ Microbiol* **69**: 6785.

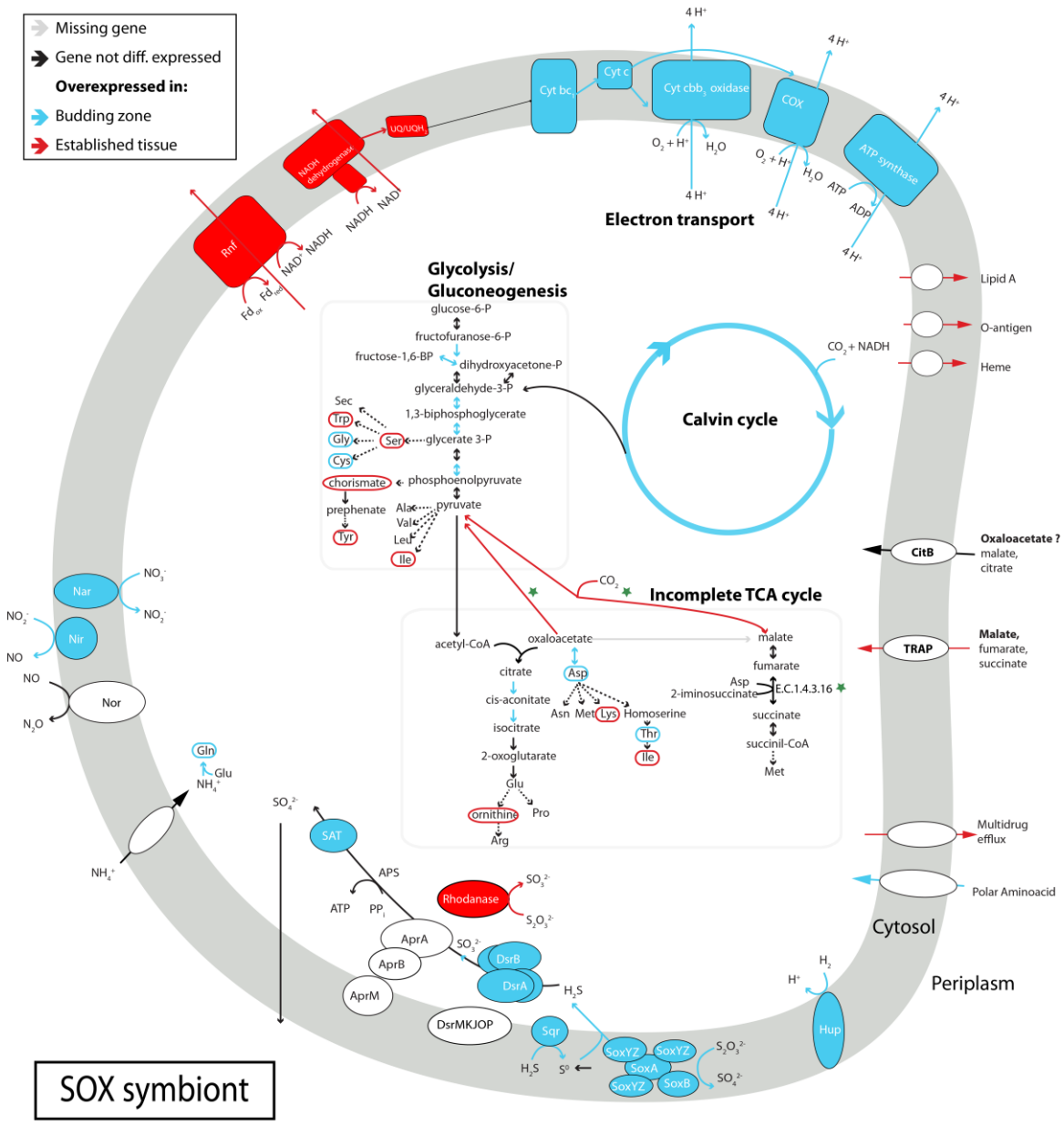
Zielinski FU, Pernthaler A, Duperron S, Raggi L, Giere O, Borowski C, *et al.* (2009). Widespread occurrence of an intranuclear bacterial parasite in vent and seep bathymodiolin mussels. *Environ Microbiol* **11**: 1150–1167.

Ziolo KJ, Jeong H-G, Kwak JS, Yang S, Lavker RM, Satchell KJF. (2014). *Vibrio vulnificus* biotype 3 multifunctional autoprocessing RTX toxin is an adenylate cyclase toxin essential for virulence in mice. *Infect Immun* **82**: 2148–2157.

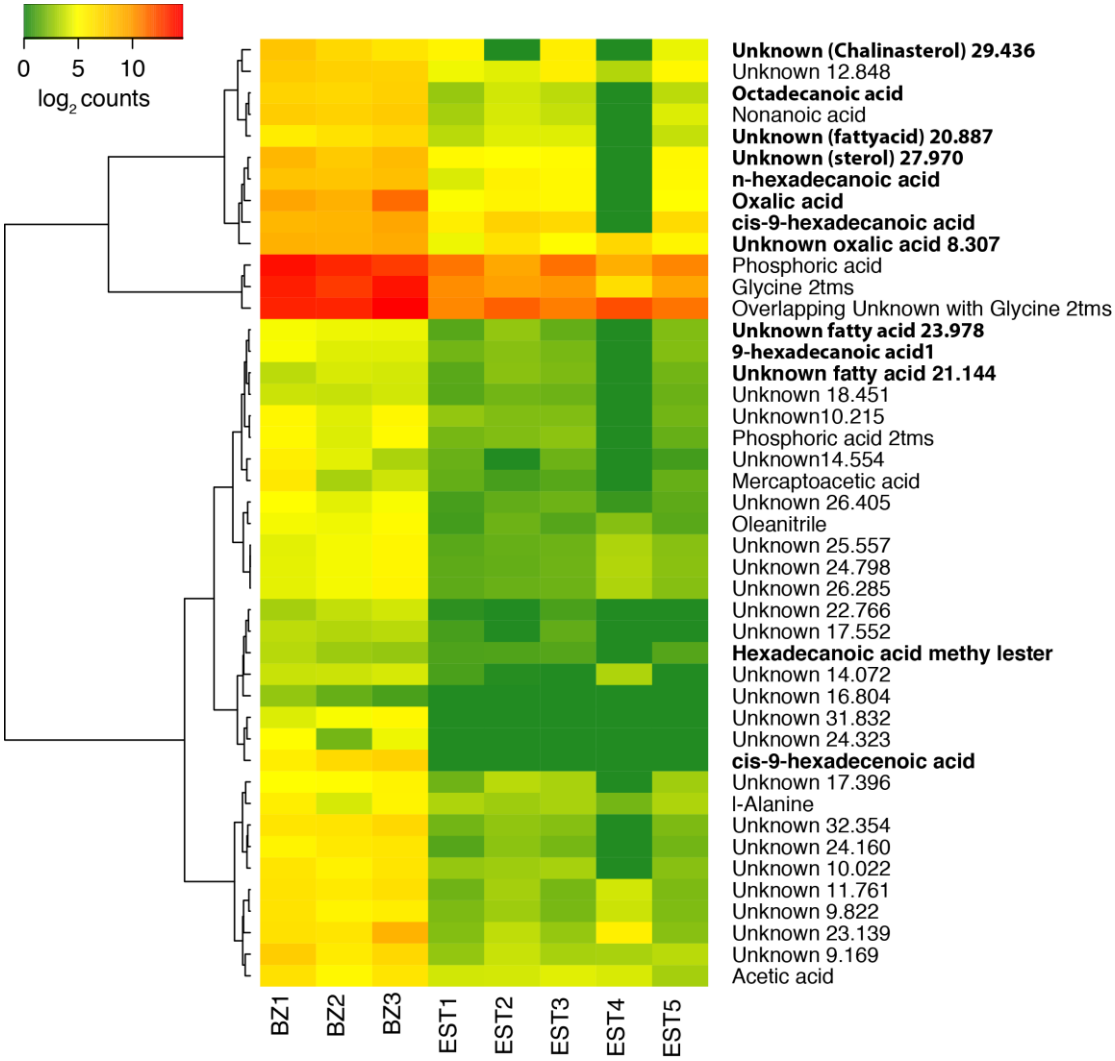
Supplementary Figures



Supplementary Figure 1. Principal component analysis (PCA) plot of the variation in the SOX symbiont population. Blue= Mussels from Eiffel Tower. Green= Mussels from Montsegur. EST= established tissue, BZ= budding zone. Each number corresponds to an individual. The samples did not cluster per individual or gill region.



Supplementary Figure 2. Metabolic reconstruction of the SOX symbiont with the detailed reactions of the TCA cycle, glycolysis pathway, and synthesized amino acids that are overexpressed in the budding zone (blue) or the established tissue (red). Reactions encoded in the genome but not differentially expressed are represented as black arrows; missing genes are represented as grey arrows. Alternative reactions that could complete the cycles are indicated with green stars. Cyt, Cytochrome; COX, Cytochrome c oxidase; Nar; membrane bound respiratory nitrate reductase; Dsr, dissimilatory sulfite reductase gene cluster; Sox, sulfur oxidation gene cluster; AprAB, adenosine-5'-phosphosulfate (APS) reductase; AprM, APS reductase membrane anchor; SQR, sulfide-quinone reductase; SAT, sulfate adenylyltransferase; CitB, citrate transporter.



Supplementary Figure 3. Metabolite abundance comparison between budding zone (BZ) and established tissue (EST). Only the metabolites that had a significant abundance difference between BZ and EST are shown. Fatty acids and oxaloacetate are shown in bold.

Supplementary Tables

Supplementary Table 1. Metrics of *de novo* transcriptome assembly depleted of symbiont sequences using as reference the genome assemblies of the SOX and MOX symbiont from this study.

	Statistics based on transcripts depleted of symbiont sequences
Total transcripts	83,923
Percent GC	35.91
Contig N10	8,951
Contigs N30	4,034
Contig N50	2,137
Median	556
Total assembled bases	95,153,614

Supplementary Table 2. Overview of samples used for genome and transcriptome sequencing

#	**CO1 accession number**	**Site**	**Collection date**	**Fixation**	**Gill region**	**No. of transcriptome reads***	**No. of genome reads***
1	LN833435[$]	LS-Montsegur	9.Aug.2013	Frozen	BZ	27.36	25.19
					Est	20.09	33.57
2	LN833436[$]	LS-Montsegur	9.Aug.2013	Frozen	BZ	39.04	30.60
					Est	9.91	20.20
3	LN833437[$]	LS-Montsegur	9.Aug.2013	Frozen	BZ	31.58	29.00
					Est	8.84	23.64
4	XX	LS-Eiffel Tower	17.Aug.2013	RNA later	BZ	23.35	25.96
					Est	10.12	16.17
5	XX	LS-Eiffel Tower	17.Aug.2013	RNA later	BZ	40.64	23.45
					Est	9.26	16.66

LS = Lucky Strike
XX = accession numbers will be obtained prior to submission
[#]Individual
*Million reads, paired-end sequencing
BZ= Budding zone
Est = Mid-region of the gill, where the symbiosis is fully established
[$]Sayavedra *et al*. 2015

Supplementary Table 3. Relative abundances of metabolites detected with GC-MS from budding zone (BZ) and established tissue (EST).

Compound	Retention time (min)	BZ (µ±sd of rel. abundance)	EST (µ±sd of rel. abundance)	P-value (Mann-Whitney)
Dimethylglycine	7.389876667	1062.71±358.51	482.46±244.01	0.071
Lactic acid	8.199326667	8.84±8.89	2.51±2.03	0.250
Unknown (oxalic acid) 8,307	8.2869	575.1±96.59	61.42±43.88	0.036
Acetic acid	8.4045	67.1±27.06	13.49±4.17	0.036
L-Alanine	8.76356	37.51±20.29	6.18±2.12	0.036
Overlapping Unknown with Glycine 2 tms	8.988756667	16434.51±7736.57	2876.34±1447.43	0.036
Glycine 2tms	9.028786667	12736.53±4774.26	863.54±481.12	0.036
Unknown 9,169	9.142636667	123.15±62.98	7.9±2.68	0.036
Oxalic acid	9.22646	1377.83±1211.95	27.01±16.25	0.036
Unknown 9,306		0±0	68.14±43.12	0.081
Phosphoric acid 2tms	9.743156667	28.7±10.62	2.43±1.6	0.036
Unknown 9,822	9.815716667	60.44±20.69	5.42±3.98	0.036
Unknown 10,022	10.01088667	66.41±17.17	4.33±2.65	0.036
Unknown 10,215	10.18604	31.39±12.37	2.77±1.76	0.036
Valine	10.28154444	100.45±107.98	8±8.71	0.143
Mercaptoacetic? Acid	10.89039333	28.7±32.62	1.18±0.8	0.036
Serine 2tms	10.86537778	12.2±21.13	7.25±9.89	1.000
Ethanolamine	10.97484583	41.61±17.27	21.65±13.02	0.143
Phosphoric acid	11.04427667	12352.48±5405.49	1522.6±851.08	0.036
Isoleucine	11.33036111	37.73±41.02	3.71±3.01	0.071
Threonine 1	11.44296111	29.35±50.36	17.16±25.12	1.000
Proline	11.42002222	39.52±52.32	4.91±5.6	0.250
Glycine 3tms	11.52219333	4326.44±2075.19	1803.06±1079.1	0.250
Succinic acid	11.63634583	836.66±920.61	150.82±133.09	0.250
Glyceric acid (overlap)	11.70987917	6.31±10.93	0.25±0.36	1.000
Unknown 11,761	11.78491667	79.33±15.76	5.69±4.78	0.036
Serine 3tms		0±0	6.89±9.45	0.329
Fumaric acid	12.13834583	594.99±547.55	19.62±27.26	0.359
Nonanoic acid	12.25616111	181.62±25.49	9.78±6.71	0.036
Unknown 12,848	12.85208667	175.77±23.61	27.52±17.18	0.036
Unknown 13,055	13.046925	22.73±39.38	8.91±12.33	1.000
B-Alanine	13.1424	19.84±34.37	65.2±102.26	0.733
Unknown 13,531	13.52395	17.71±30.67	6.36±8.74	1.000
Malic acid		0±0	2.47±5.51	0.606
Unknown 14,072	14.05062	12.96±1.42	1.7±3.27	0.036
Unknown 14,231	14.22872778	94.26±163.11	2.31±5.15	0.306
Unknown 14,554	14.5644	25.37±22.43	0.98±1.07	0.036
Unknown 14,785	14.78750667	2.21±2.11	0±0	0.079
Unknown 16,804	16.80800556	2.52±1.95	0±0	0.017
Unknown 17,396	17.3819375	35.08±8.3	4.94±3.73	0.036
Unknown 17,552	17.54925833	9.06±0.82	0.51±0.77	0.032
Unknown 18,451	18.45317222	12.44±1.07	1.62±1	0.036
9-hexadecanoic acid 1	18.65804167	20.26±4.95	2.53±1.5	0.036
Hexadecanoic acid methyl ester	18.86602333	6.47±2.15	0.89±0.5	0.036
Cis-9-hexadecanoic acid	19.22383333	564.73±204.87	88.56±62.48	0.036
N-hexadecanoic acid	19.46904	277.25±28.41	26.71±18.69	0.036

Compound	Retention time (min)	BZ (µ±sd of rel. abundance)	EST (µ±sd of rel. abundance)	P-value (Mann-Whitney)
Cis-9-hexadecenoic acid	19.93725833	106.99±50.52	0±0	0.017
Palmitelaidic acid	19.93725833	54.71±31.27	20.36±15.51	0.143
Oleanitrile	20.46990333	27.16±5.52	1.83±1.2	0.036
Unknown (fatty acid) 20,887	20.8886	89.02±34.73	10.95±7.12	0.036
Unknown (fatty acid) 21,144	21.1382875	13.11±3.17	2.18±1.56	0.036
Octadecanoic acid	21.3369	149.2±19.49	7.54±5.15	0.036
Unknown 22,766	22.77147222	10.22±3.61	0.24±0.41	0.032
Unknown 23,139	23.15972	224.05±246.17	14.32±19.43	0.036
Umnown (fatty acid) 23,978	24.00170556	22.75±2.22	2.2±1.81	0.036
Unknown 24,160	24.16226111	60.14±15.55	2.11±1.58	0.036
Unknown 24,323	24.32072333	18.32±14.25	0±0	0.017
Unknown 24,798	24.8149	28.06±11.03	3.57±2.76	0.036
Unknown 25,557	25.5693	27.94±11.04	3.39±2.53	0.036
Unknown 26,285	26.29993	29.06±12.94	3.39±2.48	0.036
Unknown 26,405	26.40001667	24.82±5.94	1.33±0.66	0.036
Unknown (sterol) 27,970	27.98169167	333.39±111.05	26.94±15.26	0.036
Unknown (chalinasterol) 29,436	29.44889	152.71±95.16	23.25±24.22	0.036
Unknown 31,832	31.89756667	26.66±10.42	0±0	0.017
Unknown 32,354	32.3338875	94.28±27.38	2.79±1.74	0.036

Chapter V:

Symbiont strain heterogeneity confers metabolic flexibility in deep-sea mussels: Methylotrophy in sulfur-oxidizing symbionts

Lizbeth Sayavedra[1], Miguel Á. González-Porras[1], Chakkiath Paul Antony[1], Jimena Barrero-Canosa[1], Nicole Dubilier[1], Jillian M. Petersen[1,2]

[1]Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany
[2]University of Vienna, Althanstrasse 14, 1090 Vienna, Austria

This manuscript contains preliminary results and has not been reviewed by all authors.

Author contributions

LS conceived the project, analyzed genome and transcriptome sequencing, found that not all SOX symbionts encode the MDH, designed probes for gene-FISH, and wrote the manuscript; MGP designed gene-FISH probes, produce preliminary results of gene-FISH hybridizations; CPA helped to develop the concept and revised the manuscript; JBC showed the gene-FISH method and assisted with troubleshooting; ND helped to conceive the project; JP found the MDH in the genome and conceived the project.

Abstract

Deep-sea *Bathymodiolus* mussels are among the most successful invertebrates colonizing hydrothermal vents. They can host intracellular methane-oxidizing symbionts (MOX) that use methane as carbon and energy source, and sulfur-oxidizing symbionts (SOX) that use energy from the oxidation of reduced sulfur compounds and/or hydrogen to assimilate inorganic carbon (CO_2). Using metagenome sequencing, we discovered the genomic potential for methanol oxidation in the SOX symbionts of *Bathymodiolus* mussels from vents of the northern Mid-Atlantic Ridge (MAR). The cluster of genes involved in methanol oxidation, which included the methanol dehydrogenase (MDH) key gene of the XoxF type, was only present and expressed in SOX symbionts of the NMAR. These symbionts, therefore, exhibit unexpected versatility in their capacity to perform multiple metabolic strategies: chemolithoautotrophy and methylotrophy. Genome coverage revealed that within a single mussel only some symbiont strains encode the MDH. Preliminary results with single-gene fluorescence *in situ* hybridization confirmed the link between the MDH gene cluster and the SOX symbiont, as well as a mosaic-like distribution of the SOX symbionts encoding the MDH in *B. azoricus*. Transcriptome sequencing revealed that the MDH has an expression pattern that reflects the strain heterogeneity and the environmental conditions at the hydrothermal vents. The MDH was most likely acquired through horizontal gene transfer from Alphaproteobacteria based on phylogeny and codon usage analysis. We hypothesize that the successful symbiotic establishment of the SOX symbiont strains with the capacity to use methanol is determined by the environmental conditions.

Introduction

Deep-sea hydrothermal vents are hot spots where fluids rich in reduced chemicals sustain chemosynthetic oases of life (Van Dover, 2002). Since the discovery of vents in 1977, reduced sulfur compounds, hydrogen and methane have been found to fuel chemosynthesis at deep-sea hydrothermal vents (Dubilier *et al.*, 2008; Petersen *et al.*, 2011). At hydrothermal vents, steep chemical gradients occur temporally and spatially, even at the millimeter scale (Schrenk *et al.*, 2003; Zielinski *et al.*, 2011). Associations with symbionts that can exploit different energy sources can allow invertebrate animals to respond to their quickly changing environment (Ikuta *et al.*, 2015; Beinart *et al.*, 2012; Distel *et al.*, 1995).

Bathymodiolus mussels are among the dominant fauna at hydrothermal vents and cold seeps worldwide. They obtain their nutrition through the association with intracellular sulfur-oxidizing (SOX) symbionts, or methane-oxidizing (MOX) symbionts, or both. One exceptional mussel species from the Gulf of Mexico, *B. heckerae*, can associate with up to six symbiont phylotypes that include two sulfur-oxidizing symbionts, a methanotroph and a methylotroph – a bacterium capable of using C1 carbon compounds such as methanol as an energy and carbon source (Raggi *et al.*, 2013; Duperron *et al.*, 2007).

Methylotrophs oxidize methanol to formaldehyde using a periplasmic pyrroloquinoline quinone (PQQ)-linked methanol dehydrogenase enzyme (MDH). The MDH enzyme in most methylotrophs can be a calcium-dependent MxaFI and/or the lanthanide-dependent XoxF (Keltjens *et al.*, 2014). XoxF is present in all characterized gram-negative methylotrophs (Chistoserdova, 2011) and clusters into five distinct phylogenetic clades (XoxF1-5) (Keltjens *et al.*, 2014). The *xoxF* gene and its product are known to be widespread in marine environments (Sowell *et al.*, 2011; Taubert *et al.*, 2015). Strain HTCC2181, an abundant marine methylotroph, relies on XoxF to grow on methanol (Giovannoni *et al.*, 2008). Formaldehyde, formed from methanol oxidation in methylotrophs, is i) oxidized all the way to CO_2 to generate energy and ii) assimilated into biomass via the ribulose monophosphate (RuMP) pathway or the serine cycle (at the level of methylene-H4F and CO_2) or the Calvin-Benson-Bassham (CBB) cycle (at the level of CO_2). Methylotrophs that use the CBB cycle for C1 assimilation are considered methylotrophic autotrophs and have mostly been restricted to a few representatives from the Alphaproteobacteria, Verrucomicrobia and NC10 phyla (Chistoserdova, 2011).

In this study, we show that the SOX symbiont can use not only reduced sulfur compounds and hydrogen as an energy source (Petersen *et al.*, 2011; Dubilier *et al.*, 2008), but also has the potential to use methanol. Methanol is the second most abundant organic compound in the atmosphere (Dixon *et al.*, 2011). In hydrothermal vent fluids, C1 and other low molecular weight carbon compounds such as methanethiol and formate have been identified, but methanol is a challenging compound to measure in dissolved organic carbon (Reeves *et al.*, 2014; McDermott *et al.*, 2015; Rossel *et al.*, 2015). The most obvious source of methanol for the SOX symbiont would be the MOX symbiont, that produces methanol as the primary product of methane oxidation (Petersen and Dubilier, 2009). As is the case with most symbiotic bacteria, cultivation attempts have been unsuccessful, so molecular methods are the best tools at hand to explore the genetic

potential of these communities. We sequenced metagenomes and metatranscriptomes of mussels from vents with different geological settings that determine whether hydrogen, methane (and possibly methanol) are in higher concentrations. Genome coverage of the MDH suggested that the SOX symbionts have high strain heterogeneity in the potential to use methanol, although the SOX symbiont population has only one 16S rRNA gene. Similar strain heterogeneity was recently reported for the potential to use hydrogen in SOX symbionts of *B. thermophilus* (Ikuta *et al.*, 2015). The goals of this study were i) to examine the distribution of the SOX symbiont strains encoding the MDH among *Bathymodiolus* mussels around the world, ii) to determine the strain variation in the potential to use methanol both within vent regions and between different vent regions, and iii) to correlate the expression of the MDH to the venting sites and the proportion of strains that encode the gene.

Results

Some sulfur-oxidizing symbionts of *Bathymodiolus puteoserpentis* and *B. azoricus* acquired the potential for methylotrophy

The sulfur-oxidizing (SOX) symbiont genomes from *B. puteoserpentis* and *B. azoricus* encoded a PQQ-methanol dehydrogenase (XoxF), a key enzyme for methanol oxidation (genomes to be released in Sayavedra *et al.*, in prep - Chapter 3 and 4). The SOX symbiont did not encode any of the known pathways for C1 carbon assimilation from formaldehyde (i.e, RuMP or serine pathway) and instead, only possessed the CBB cycle for carbon assimilation at the level of CO_2. The SOX symbiont potentially oxidizes methanol to formaldehyde using the XoxF type of MDH and subsequently, oxidizes the formaldehyde to formate via mainly the glutathione-linked formaldehyde dehydrogenase (GSH) pathway. Thus, the SOX symbiont is potentially an autotrophic methylotroph (Chistoserdova, 2011). The gene neighborhood of *xoxF* included four genes upstream of the sequence involved in formaldehyde oxidation and the interaction of XoxF with cytochrome c (MxaD), as well as 11 genes downstream involved in the synthesis and transport of the PQQ cofactor of XoxF.

Previous SOX symbiont draft genome 'bins' obtained from metagenome assemblies of *B. azoricus* from Menez Gwen did not encode the MDH-associated gene cluster (Sayavedra *et al.*, 2015), but the gene was present in genome 'bins' obtained from *B. azoricus* individuals of the Lucky Strike (LS) vent field and *B. puteoserpentis* from

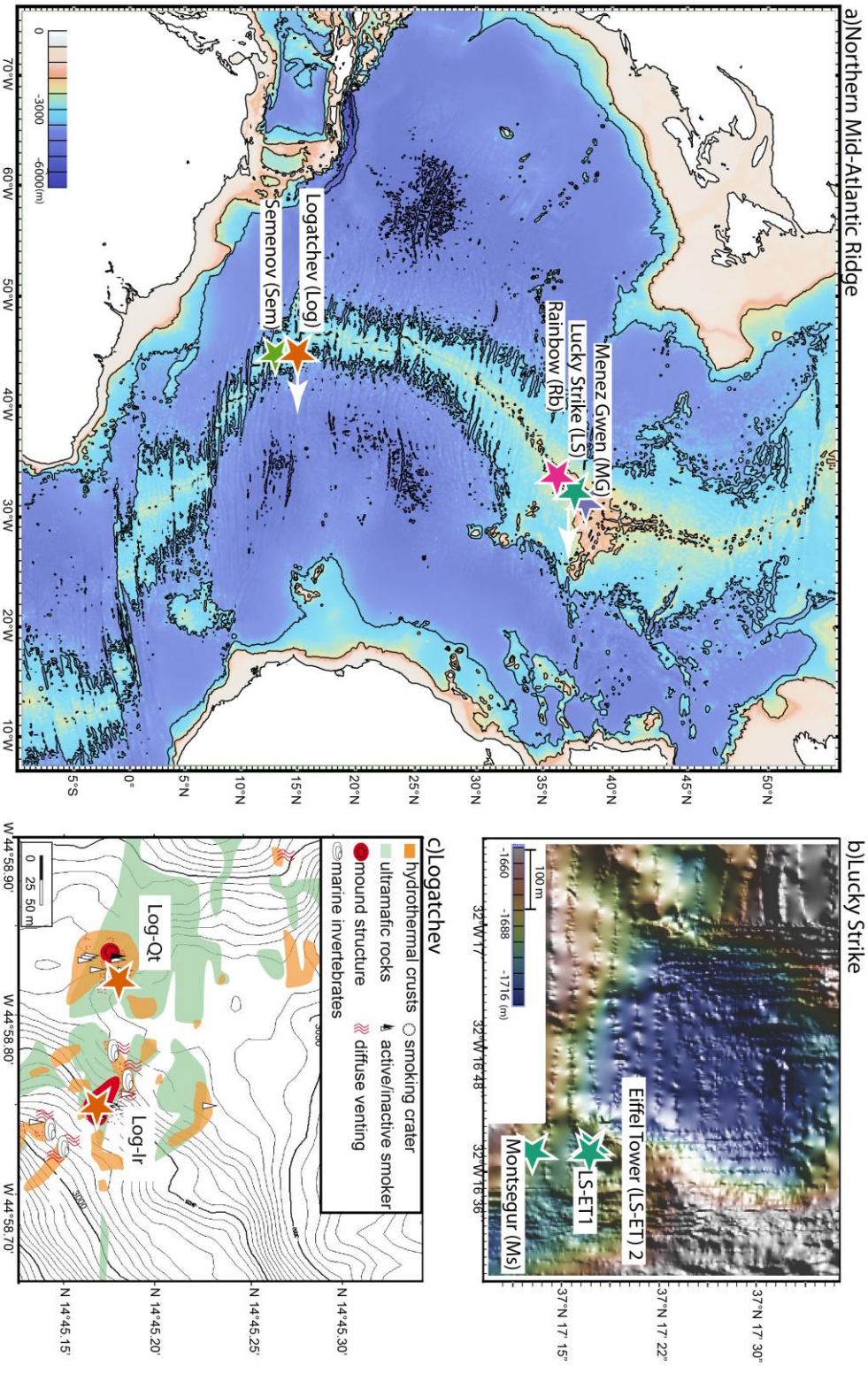


Figure 1. a) Overview of sampling sites where the SOX symbiont encodes the MDH gene. b) Zoom in of the three sampling locations within the Lucky Strike vent field - Eiffel Tower (LS-ET) and Montsegur (LS-Ms). c) Zoom in of the two sampling sites within Logatchev - Quest (Log-Qt) and Irina II (Log-Ir). Bathymetric data from Humphris *et al.* (2002) was used to produce the map from b) with GeoMapApp. Map from c) was modified from S. Petersen *et al.* (2009).

Logatchev-Irina II (Log-Ir) (Figure 1 - Vent Sites) (Sayavedra *et al.*, in prep - Chapter 3 and 4). To know the distribution of the MDH in *B. azoricus* and *B. puteoserpentis* mussels, we sequenced metagenomes of vents along the northern MAR in vents with contrasting environmental conditions where either *B. azoricus* or *B. puteoserpentis* live: Menez Gwen (MG) (n=5), Lucky Strike (Montsegur (LS-Ms) (n=5) and Eiffel Tower (LS-ET) (n=10 divided among two spots separated by three meters)), Rainbow (Rb) (n=5), Semenov (Sem) (n=3), and Logatchev – Irina II (n=1) (Figure 1). We also searched for the MDH gene of the SOX symbiont in metagenome reads of *B. heckerae* and *B. brooksi* mussels (Gulf of Mexico), *B. thermophilus* (Crab Spa, East Pacific Rise), *B.* sp from 9° SMAR (Lilliput), *B.* sp 5° SMAR (Turtle Pits), *B.* sp 5° SMAR (Clueless), and the closed genome of the SOX symbiont of *B. septemdierum* (Myojin knoll). A detailed list with the samples used in this study is described in Supplementary Table 1. All *B. azoricus* and *B. puteoserpentis* metagenomes from this study had reads mapping to the MDH gene cluster of the SOX, and none of the other *Bathymodiolus* species analyzed in this study had reads mapping to the SOX symbiont MDH cluster. This suggests that the potential to use methanol by sulfur-oxidizing symbionts is exclusive to the *B. azoricus* and *B. puteoserpentis* symbiosis.

Single genome assemblies of SOX symbionts of *B. azoricus* and *B. puteoserpentis* were done per individual mussel (Supplementary Table 2). The best genome assembly based on the completeness, the number of scaffolds, and the N50, belonged to the SOX symbiont of *B. puteoserpentis* from Logatchev. This assembly was therefore used as the reference genome for further analysis.

Are the SOX symbionts of *B. azoricus* and *B. puteoserpentis* the same species?

Based on the 16S rRNA gene, the SOX symbiont of *B. azoricus* (BazSym) and *B. puteoserpentis* (BputSym) are the same phylotype (Duperron *et al.*, 2006). The two host-mussel species are distributed along the northern Mid-Atlantic Ridge, and even a hybrid zone of intermixing between these two populations has been described (Won *et al.*, 2003a). A fundamental question, now that we have essentially complete symbiont genomes available, is whether their SOX symbionts are indeed the same species. SOX symbiont bins from different individuals of the same host species at the same site had an average nucleotide identity (ANI) ≥ 99.68%. SOX symbiont genomes from the most

remote geographical locations that we have available for BazSym and BputSym - Semenov and Menez Gwen - had an ANI value of 98.5% between them (Supplementary Figure 1). The symbionts of these two species would, therefore, be considered one species based on widely accepted cut-offs (Rodriguez-R and Konstantinidis, 2010).

Relationships to other MDHs

Since the MDH cluster was absent in the SOX symbiont genomes from most mussel species we investigated, this gene region might have been present in the common ancestor and subsequently lost, or the *B. puteoserpentis* and *B. azoricus* SOX symbionts acquired the gene region by horizontal gene transfer (HGT). To investigate the possible origin of the *xoxF* methanol dehydrogenase of the SOX symbiont, we reconstructed a MDH phylogeny that included sequences from the SOX, MOX, methylotrophic symbionts and other free-living bacteria. All three types of symbionts encoded a MDH from the XoxF5 sub-clade (Figure 2). Proteobacterial methanotrophs, in general, encode both versions of the MDH – the canonical MxaFI and the XoxF. Intriguingly, from the six species of *Bathymodiolus* mussels containing MOX symbionts analyzed in this study, only *B. azoricus* and *B. puteoserpentis* had a MOX symbiont containing both forms while the rest encoded exclusively *xoxF* (Figure 2) (Antony CP *et al. in prep*). *mxaF* and *xoxF* of the MOX symbiont grouped with MDH sequences from bacteria of the order Methylococcales while *xoxF* of the methylotroph grouped with sequences of the order Thiotrichales, in agreement with the taxonomic affiliation of the MOX and methylotroph symbionts based on the 16S rRNA gene (Raggi *et al.*, 2013; Petersen and Dubilier, 2009). In contrast, *xoxF* of the SOX symbiont (Gammaproteobacteria) was more similar to sequences of the order Rhodospirillales (Alphaproteobacteria). The discrepancy in the taxonomic affiliation of the SOX symbiont's *xoxF* gene suggests that the MDH region was acquired through HGT.

To further investigate if the genes for methanol oxidation have been recently acquired through HGT, we used analysis of codon usage with the genome of *B. puteoserpentis* SOX (BputSym) (Davis and Olsen, 2010). The *xoxF* figured among the list of genes that were predicted to have been acquired through HGT (36.5% of total genes from the genome). Four genes in the *xoxF* gene neighborhood, involved in formate oxidation and the synthesis of the pyrroloquinoline (PQQ) cofactor, were also predicted

to be the product of recent HGT events (Figure 3). These results support the recent acquisition of the MDH cluster through HGT, possibly from an alphaproteobacterium.

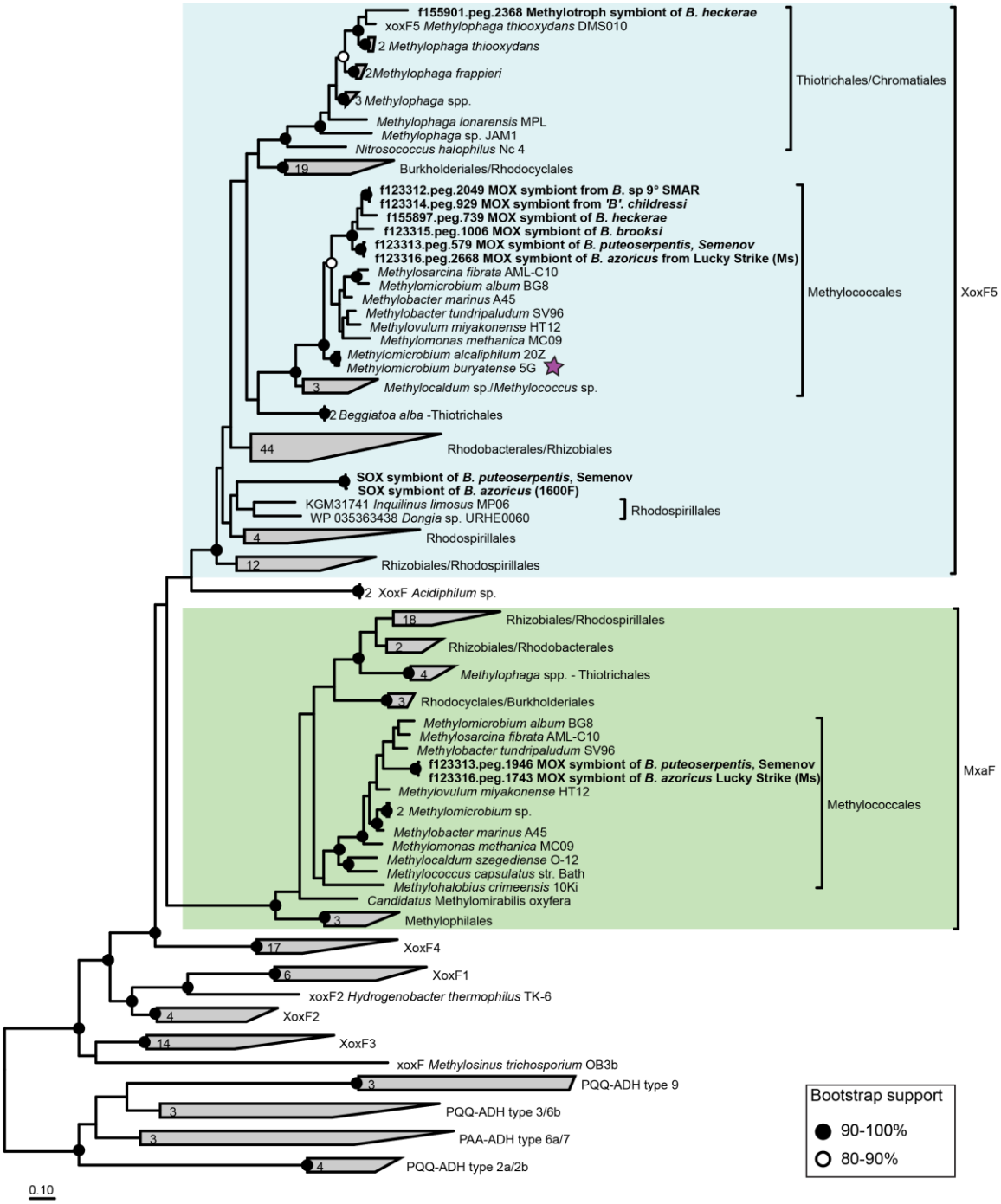


Figure 2. Maximum likelihood phylogeny of the different clades of MDH: XoxF , MxaF and PQQ-dependent dehydrogenases. Purple star = the enzymatic activity of the XoxF5 of *Methylomicrobium buryatense* was experimentally confirmed (Chu and Lidstrom, 2016).

Not all SOX symbionts have the potential to use methanol

Although all *B. azoricus* and *B. puteoserpentis* metagenomes had reads mapping to the MDH gene cluster, the coverage of the *xoxF* and the gene neighborhood varied from site to site (the mapping profile of one individual per site is shown in Figure 3). Variations in gene coverage could indicate that only some SOX symbiont strains can utilize methanol as an additional energy source. The ability to use methanol by the SOX symbiont may not be critical for its survival, as it can use other energy sources (Petersen *et al.*, 2011). Vents have contrasting fluid composition that, depending on which reduced compound is in higher concentrations in the vent, could determine the success of a specific strain of SOX symbiont to colonize and persist within the mussel. We tested whether the proportion of SOX strains with the potential to use methanol was higher in venting sites with higher concentrations of methane. The coverage of each coding sequence (CDS) was estimated by mapping the metagenome reads to the BputSym reference genome per single individual mussel. As a proxy for the number of strains that can utilize methanol, we used the ratio of 16 genes within the MDH gene cluster to the single copy gene gyrase A (*gyrA*) (Supplementary Figure 2). A ratio of one would suggest that all SOX strains have the potential to use methanol while a ratio of zero would suggest that no SOX symbiont encodes the gene cluster.

We used the Kruskal-Wallis one-way analysis of variance to determine whether the ratio of the 16 genes in the MDH gene cluster against *gyrA* was significantly different between the two mussel host species or vent sites. BazSym from all sites had a significantly lower coverage ratio of the MDH gene cluster compared to BputSym (BazSym = 0.16±0.19 vs. BputSym = 0.75±0.22; p-value < 2e-16, Kruskal-Wallis test), and there was a significant difference in the ratios between different vent regions (p-value < 2.2e-16, Kruskal-Wallis test). The SOX symbiont of Log-Ir had the highest ratio of the MDH gene cluster (0.8±0.11) followed by Sem (0.72±0.24), Rb (0.47±0.2), LS-ET (0.10±0.05), LS-MS (0.07±0.04), and MG (0.05±0.05). To tease apart which of the sites are different, we used Dunn's test. This test compares pairs of vent samples. All pairwise comparisons of the ratios by venting sites were significantly different except for LS-Ms versus MG and the two sites that host BputSym (Sem and Log) (Table 1). This is surprising because even within the vent Lucky Strike there was a significant difference between two sites that are ~150 meters apart (LS-ET vs LS-Ms, p-value=0.0004, Dunn's test) (Figure 1b). This analysis indicates that i) there are significant differences between

the symbiont populations of these two different mussel species, with one species (*B. puteoserpentis*) consistently hosting a larger proportion of symbionts capable of using methanol) and ii) there are significant differences in the proportion of symbionts encoding MDH genes within a species, and these differences are vent site-specific.

Table 1. Comparison of gene coverage of the MDH gene cluster between vents. The counts were normalized to the gene coverage of the single copy gene *gyrA*. P-values were obtained with Dunn's pairwise comparison test with 16 single genes in the neighborhood of the *xoxF* (Figure 3 and Supplementary Figure 2). Mean gene coverage and standard deviation of the 16 single genes is shown. P-values were corrected for multiple testing with the false-discovery rate.

Site	MDH gene cluster coverage/*gyrA*	P-values				
		Sem	RB	LS-ET	LS-Ms	MG
Log-Ir	0.84±0.11	0.201	0.021*	4.10E-11*	1.10E-15*	< 2e-16*
Sem	0.72±0.24		0.037*	< 2e-16*	< 2e-16*	< 2e-16*
RB	0.47±0.2			< 2e-16*	< 2e-16*	< 2e-16*
LS-ET	0.1±0.05				0.0004*	3.70E-07*
LS-Ms	0.07±0.04					0.090
MG	0.05±0.05					

*P-value was considered to be significant when lower than 0.05.

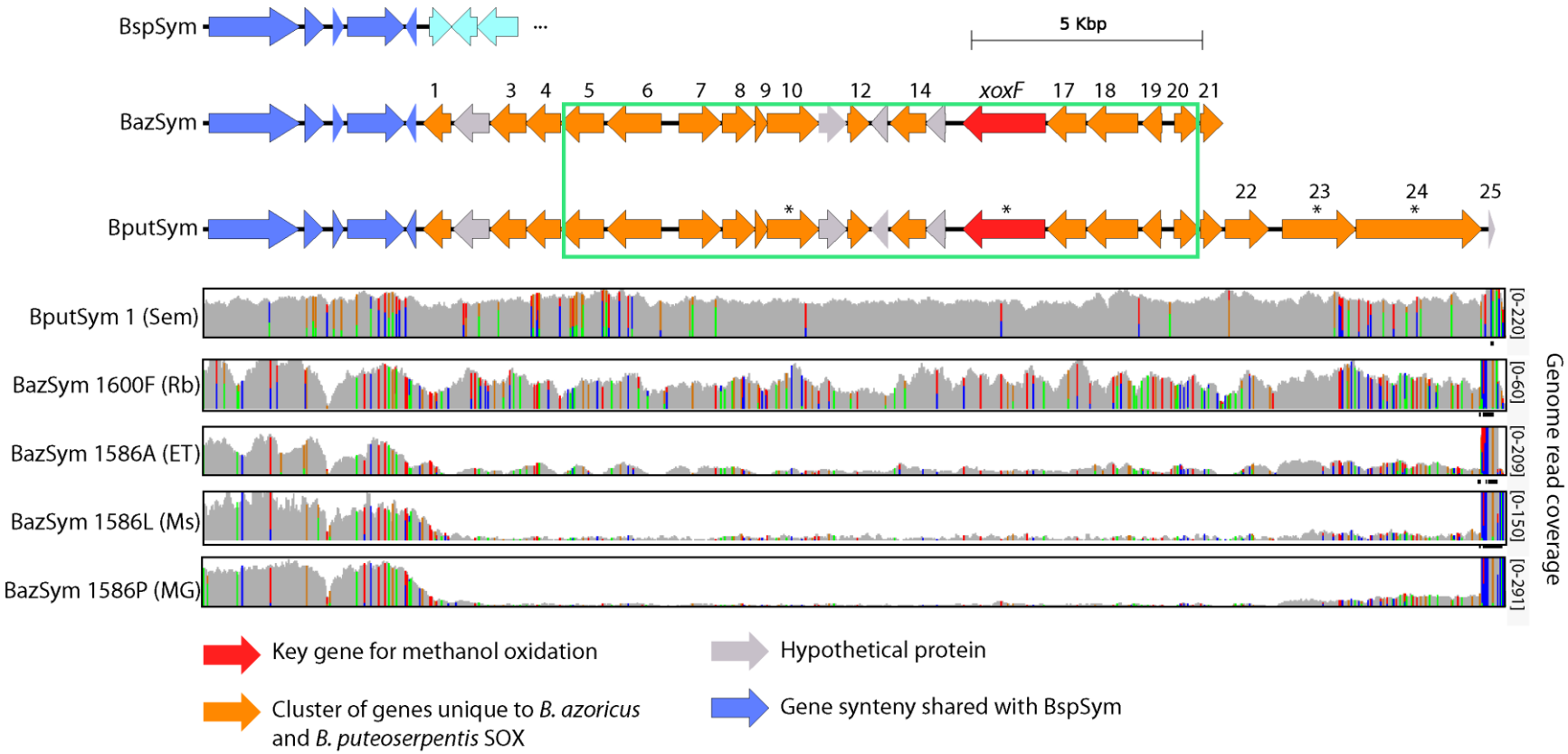


Figure 3. Gene synteny of MDH gene cluster and read mapping profile. Upper panel: Gene region containing the MDH (*xoxF*) of the SOX symbiont of *B. azoricus* (BazSym) and *B. puteoserpentis* (BputSym) is not present in the SOX symbiont of *B.* sp. 9° SMAR (BspSym). The green box contains the genes that were used for comparing read mapping coverage between sites (Table 1). Genes that were predicted to be acquired through HGT are marked with an asterisk. Lower panel: Comparison of genome read mapping profiles of single individuals from Semenov (Sem), Rainbow (Rb), Eiffel Tower (LS-ET), Montsegur (LS-Ms), and Menez Gwen (MG) compared to the reference genome (BputSym). Only one individual per site is shown. The scale of the genome coverage was chosen based on the read mapping coverage of the single copy gene recA. Nucleotide variations from the reference sequence (BputSym) are shown as color bars with the proportion of each base (A = green, C = blue, G = yellow, T = red). Figure 3. *continued on next page*

Figure 3. *continued*

Gene 1 = DNA-binding response regulator; 3 = ABC-2 type transport system permease protein; 4 = ABC-2 type transport system ATP-binding protein; 5 = PQQ-dependent catabolism-associated beta-propeller; 6 = ABC transporter, PQQ-dependent alcohol dehydrogenase system; 7 = Coenzyme PQQ synthesis protein B; 8 = pyrroloquinoline-quinone synthase; 9 = Coenzyme PQQ synthesis protein D; 10 = Coenzyme PQQ synthesis protein E; 12 = protein of unknown function DUF302; 14 = quinoprotein dehydrogenase-associated probable ABC transporter substrate-binding protein; 17 = S-formylglutathione hydrolase; 18 = S-(hydroxymethyl)glutathione dehydrogenase; 19 = S-(hydroxymethyl)glutathione synthase; 20 = MxaD; 21 = Protein of unknown function (DUF2380); 22 = 40-residue YVTN family beta-propeller repeat-containing protein; 23 = formate dehydrogenase beta subunit; 24 = formate dehydrogenase alpha subunit.

From genes to activity: comparing the potential to use methanol to the relative expression levels of genes involved in methanol oxidation

The SOX symbiont encodes the potential to use at least three main different energy sources: reduced sulfur compounds, hydrogen, and methanol. Thus, we investigated whether the genes for methanol use were indeed expressed. We sequenced the transcriptomes of three individuals per site of the same mussels that were used for the metagenome sequencing. Because we did not have the material for transcriptome sequencing of the same individual from Log-Ir, we sequenced different individuals than those used for metagenome sequencing. We sequenced transcriptomes from two sites within Logatchev, Quest (Log-Qt) and Irina II (Log-Ir), with three individuals per site (detail of samples used is shown in Supplementary Table 1). These two sites were chosen because the fluids from these two spots had contrasting availability of energy sources (Table 4) and the mussels from Irina II seemed to be in a poor condition: they had smaller and paler gills (Christian Borowski, personal communication).

The expression values of the key genes for methanol oxidation were normalized to the expression of *gyrA*, a gene that is constitutively expressed and is commonly used in qPCR normalization (Rocha *et al.*, 2015). Sem had the highest expression of *xoxF* compared to *gyrA* (4.99±0.77x *gyrA*), followed by Log-Qt (2.67±1.3x *gyrA*), Rb (1.17±1.04x *gyrA*), LS-ET (0.80±0.56x *gyrA*), MG (0.65±0.17x *gyrA*), LS-Ms (0.53±0.17x *gyrA*), and Log-Ir (0.20±0.18x *gyrA*) (Table 2). We also compared the expression pattern of the four genes upstream of *xoxF*, involved in formaldehyde oxidation and XoxF activation (Figure 4). These four genes upstream had the same trend of expression as *xoxF* at all sites except for

MG and LS-Ms. Sem had the highest expression (27.68±39.22X *gyrA*), followed by Log-Qt (3.76±4.96 x *gyrA*), Rb (1.10±1.3X *gyrA*), LS-ET (0.28±0.29X *gyrA*), LS-Ms (0.19±0.39X *gyrA*), MG (0.08±0.08X *gyrA*), and Log-Ir (0.0±0.0X *gyrA*). We could not detect the expression of the genes involved in synthesis and processing of the PQQ cofactor for most samples (Table 2). The subunit B of the PQQ cofactor, which had the highest expression of the PQQ subunits, followed a similar trend of expression than *xoxF:* the highest expression was found in Sem (1.16±1.4X gyrA), followed by Rb (0.21±0.12X gyrA), Log-Qt (0.25±0.44X gyrA) and LS-ET (0.05±0.09). The expression of the cofactor is therefore likely required only at very low levels for the activity of *xoxF*. For this reason, we used only the expression values of *xoxF* and the four genes upstream to evaluate whether there was a significant difference in expression between sites (Figure 4).

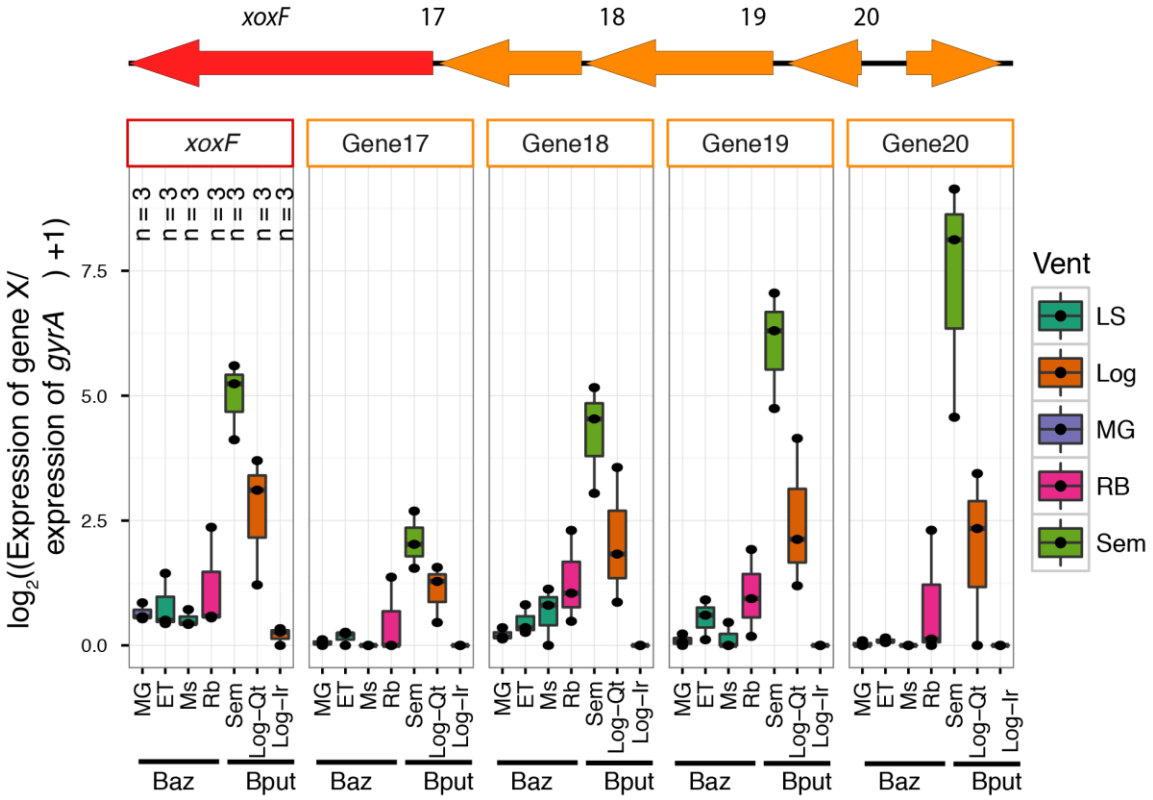


Figure 4. Gene expression of the five genes of the MDH gene cluster that were expressed in most samples. The red arrow corresponds to the key gene for methanol oxidation *xoxF*, orange arrows are genes of the MDH cluster with a functional annotation. Gene coverage was first normalized to fragments per kb mapped (FPKM) to account for differences in gene length. The ratio of expression between the gene of interest and the constitutively expressed gene *gyrA* was then calculated (normalized FPKM counts). This ratio of expression is shown as the logarithm in base 2 plus one. Gene 17 = S-formylglutathione hydrolase; 18 = S-(hydroxymethyl)glutathione dehydrogenase; 19 = S-(hydroxymethyl)glutathione synthase; 20 = MxaD protein.

Table 2. Expression of the single genes within the MDH gene cluster and P-values of expression comparisons between vents. The counts were normalized to the gene expression of the constitutively expressed gene *gyrA.* P-values were obtained with Dunn's pairwise comparison test with five genes: MDH-xoxF, S-formylglutathione hydrolase, S-(hydroxymethyl)glutathione dehydrogenase, S-(hydroxymethyl)glutathione synthase, and *mxaD* (Figure 3). We considered only these five genes because the gene expression in the other 11 genes of the MDH gene cluster was very low or zero. P-values were corrected for multiple testing with the false-discovery rate.

	Expression of gene X/expression of *gyrA* - Mean±SD						
	Sem	**Log-Qt**	**Rb**	**LS-ET**	**MG**	**LS-Ms**	**Log-Ir**
PQQ-dependent catabolism-associated beta-propeller protein	0.25±0.43	0.6±0.53	0.07±0.07	0±0	0.13±0.13	0±0	0±0
ABC transporter, substrate binding protein, PQQ-dependent alcohol dehydrogenase system	0.33±0.56	0±0	0±0	0±0	0.02±0.03	0±0	0±0
Coenzyme PQQ synthesis protein B	1.16±1.4	0.21±0.12	0.25±0.44	0.05±0.09	0±0	0±0	0±0
pyrroloquinoline-quinone synthase	0.47±0.46	0.1±0.18	0.08±0.13	0.1±0.1	0±0	0±0	0±0
Coenzyme PQQ synthesis protein D	0±0	0.13±0.23	0±0	0±0	0±0	0±0	0±0
Coenzyme PQQ synthesis protein E	0.57±0.99	0±0	0±0	0±0	0±0	0±0	0±0
protein of unknown function DUF302	0.62±1.08	0±0	0±0	0±0	0.03±0.04	0±0	0±0
quinoprotein dehydrogenase-associated	0.26±0.46	0.09±0.16	0.08±0.14	0±0	0.02±0.04	0±0	0±0
hypothetical protein	0.94±1.63	0.72±0.63	0.42±0.37	0±0	0.03±0.05	0±0	0±0
Methanol dehydrogenase XoxF	4.99±0.77	2.67±1.3	1.17±1.04	0.8±0.56	0.65±0.17	0.53±0.17	0.2±0.18
S-formylglutathione hydrolase	2.09±0.58	1.1±0.57	0.46±0.79	0.17±0.14	0.05±0.06	0±0	0±0
S-(hydroxymethyl)glutathione dehydrogenase	4.25±1.09	2.09±1.37	1.28±0.93	0.48±0.3	0.22±0.12	0.64±0.58	0±0
S-(hydroxymethyl)glutathione synthase	6.03±1.18	2.49±1.51	1.01±0.87	0.54±0.4	0.1±0.12	0.15±0.27	0±0
mxaD protein	3±0.46	1.3±1.14	0.63±0.95	0.13±0.06	0.04±0.07	0±0	0±0
	P-values - Dunn's Test						
	Sem	**Log-Qt**	**Rb**	**LS-ET**	**MG**	**LS-Ms**	**Log-Ir**
Log-Qt	0.1323						
Rb	0.00157*	0.117					
LS-ET	0.00012*	0.022*	0.474				
MG	9.80E-07*	0.00097*	0.108	0.309			
LS-Ms	2.10E-07*	0.00028*	0.0512	0.190	0.729		
Log-Ir	1.80E-10*	1.30E-06*	0.00157*	0.017*	0.154	0.259	

*P-value was considered to be significant when lower than 0.05.

We detected a significant difference in expression between the vent sites (p-value = 1.19e-12, Kruskal Wallis test). Dunn's test of pairwise comparisons of expression values for vent regions revealed that i) Semenov had a significantly higher expression than any site with *B. azoricus* mussels (p-values are shown in Table 2); ii) BputSym of Log-Qt had significantly higher expression than any site that hosted *B. azoricus* mussels. The comparison with Rb was not significantly different but the expression was higher in Log-Qt; and iii) Although all SOX symbionts of mussels from Log-Ir likely had the MDH gene cluster based on our coverage analysis, its expression

was significantly lower than the two other sites with *B. puteoserpentis* (Sem and Log-Qt) and two of the sites where *B. azoricus* lives (Rb and LS-ET). This expression pattern probably reflects the poor conditions of the mussels resulting from the low availability of methane and other reduced compounds (Table 3 and 4).

Table 3. Mineral and gas concentrations at end-member fluids of the sites used in this study. The data was obtained from (Charlou *et al.*, 2002; Douville *et al.*, 2002; Von Damm *et al.*, 1998; Charlou *et al.*, 2000). End-member fluid composition of Semenov is not available.

	MG	LS	Rb	Log
Depth (m)	850	1700	2300	3000
Temperature (˚C)	275/284	170/364	365	347/352
pH	4.2/4.3	3.5/3.7	2.8	3.3
SO_4^{2-} (mM)	0	0	0	0
Gases				
H_2S (mM)	< 1.5	2.5/3.0	1.2	0.5/0.8
CO_2 (mM)	17/20	13/28	16	10.1
CH_4 (mM)	1.35/2.63	0.50/0.97	2.5	2.1
CO (mM)	–	–	5000	–
N_2 (mM)	0.60/1.90	0.61/0.97	1.8	3
H_2 (mM)	0.024/0.048	0.02/0.73	16	12

MG = Menez Gwen; LS = Lucky Strike; Rb = Rainbow; Log = Logatchev

Table 4. Fluid *in situ* data of the two mussel beds from Logatchev (Log) and the mussel bed from Menez Gwen (White Flames) (MG). Logatchev fluid composition was obtained from the Meteor Cruise Report M64-2. Menez Gwen data was obtained from Meier *et al.* (2016).

	Log-Irina II (2A)	Log-Quest	MG
CH_4 (µM)	1.63	63.68	1.2
H_2 (µM)	1.36	4.20	0.26
H_2S (µM)	max. 0.16	70	8.1

Discussion

Methylotrophy– a novel metabolic ability acquired through HGT

Our findings show that the SOX symbiont has the potential to use methanol, a new energy source for a symbiont that was previously only known to exhibit chemolithoautotrophy via the usage of hydrogen and reduced sulfur compounds (Petersen *et al.*, 2011; Dubilier *et al.*, 2008). The methylotrophic mode of metabolism observed in

the SOX symbiont strains suggests the interesting possibility of a growth strategy involving the simultaneous use of methanol, reduced sulfur compounds and hydrogen to fix inorganic carbon. Sulfur-oxidizing Gammaproteobacteria that are also capable of methylotrophy are not common. So far, only some strains from *Beggiatoa* and *Methylophaga* have been described to do so (Zwart *et al.*, 1996; Jewell *et al.*, 2008; Boden *et al.*, 2010).

Based on the phylogeny of the MDH genes and the different codon usage to the rest of the genome, the potential to use methanol was most likely acquired through horizontal gene transfer from Alphaproteobacteria. This is in agreement with the shared metabolic features of the SOX with alphaproteobacterial methylotrophs: based on the transcriptome, the GSH pathway for formaldehyde oxidation of the SOX symbiont was the main route for formaldehyde oxidation (Figure 5). This pathway is also the major route for energy generation in methylotrophic autotrophs from the Alphaproteobacteria (e.g., *Paracoccus*, *Rhodobacter*) (Chistoserdova, 2011). All Alphaproteobacteria sequenced to date carry gene transfer agents, which are phage-like entities that can be linked to complete array of functional genes (Lang *et al.*, 2012). Indeed, Alphaproteobacteria are abundant in the proximity of hydrothermal vent fluids (Meier *et al*, 2016). Before colonizing its host, the SOX symbiont presumably has a free-living stage, during which shares the niche with the bacteria in the proximity to the vents (Won *et al.*, 2003b; DeChaine *et al.*, 2006; Fontanez and Cavanaugh, 2014). The SOX symbiont could have acquired the potential to use methanol from an alphaproteobacterium of the Rhodospirillales order while in its free-living stage.

What factors could help the SOX symbiont with the potential to use methanol to 'win' over other SOX symbiont strains?

Not all SOX symbiont strains of *B. azoricus* encoded the gene for methanol oxidation. We used mussels from vent sites that had contrasting methane concentrations in the end-member fluids (Table 3), a factor that is known to shape the chemosynthetic free-living bacterial community (Sheik *et al.*, 2015). Methanol could also be present in the vent fluids, but unfortunately, no data is available regarding its concentration at vents. Rossel *et al*. (2015) speculated that methanol could be present in fluids from MG, but could not accurately characterize the non-extractable fraction of dissolved organic carbon, where methanol would be expected. The production of methanol is favored under the

same conditions that favor the production of methane through abiotic methanogenesis (Shock and Canovas, 2010; Voglesonger *et al.*, 2001). An additional source of methanol could come from the microbial activity of methanotrophs and methanogens, which consume or produce methane, respectively (Nercessian *et al.*, 2005; Martin *et al.*, 2008; Reeburgh, 2007). Thus, even though the methanol concentrations in the environment are not known, the methanol available for the mussel symbionts might correlate with how much methane is available.

The most obvious source of methanol for the symbiont is the one provided

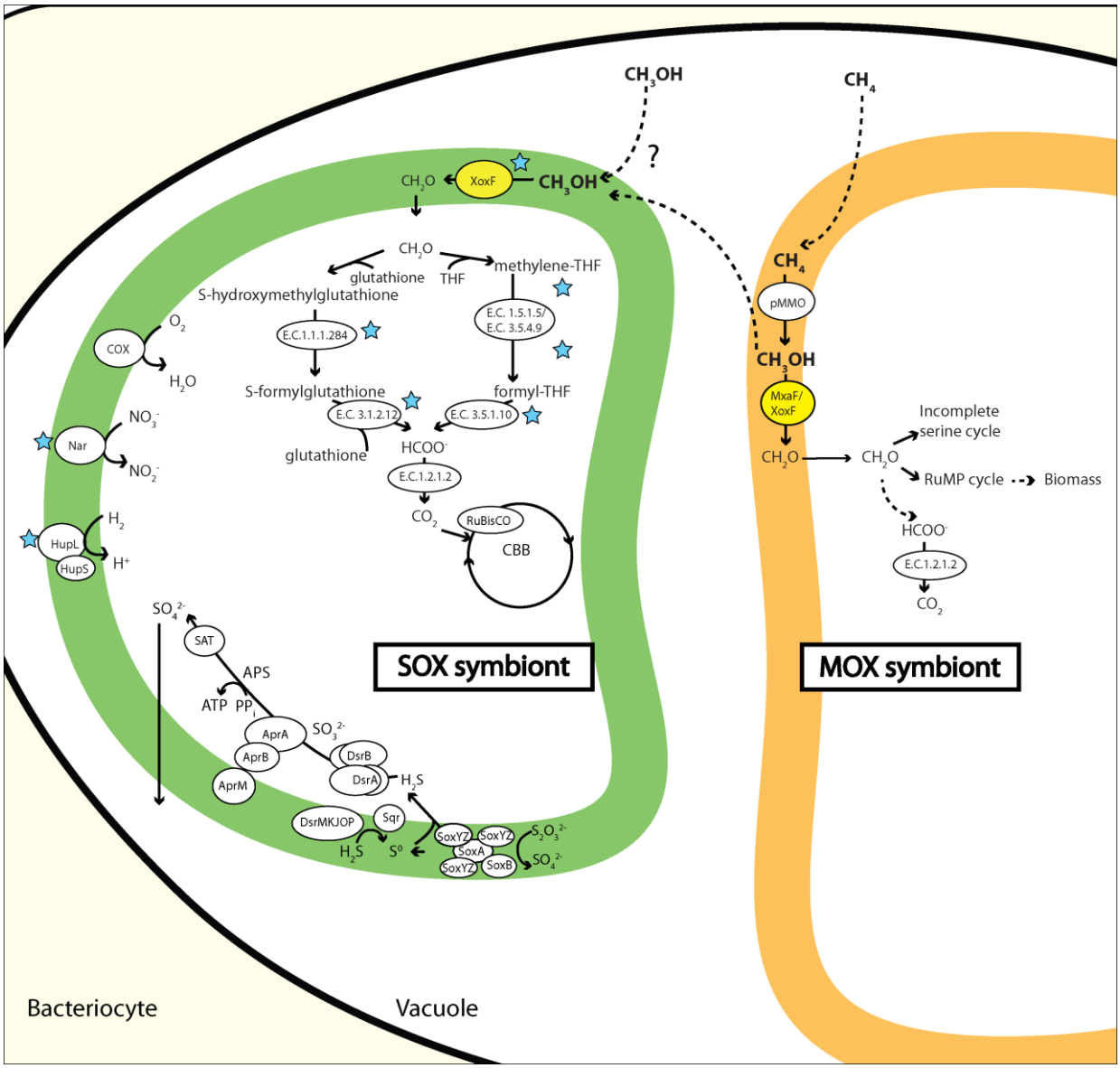


Figure 5. Model of the symbiotic relationship between the sulfur oxidizing (SOX) symbiont and the methanotroph symbiont (MOX). Methane (CH_4) is oxidized to methanol (CH_3OH) by the MOX symbiont with the membrane-bound methane monooxygenase (pMMO). The SOX symbiont could use methanol that is made available by the MOX symbiont or from the environment. Two formaldehyde oxidation routes are present, but only the glutathione-dependent is expressed (expression data not shown). The methanol dehydrogenase enzymes are shown in yellow. CH_2O = formaldehyde; $HCOO^-$ = formate; THF = tetrahydrofolate; Blue star = not all SOX strains encode the enzyme.

through methane oxidation by the co-occurring MOX symbiont. Most methanotrophs are also known to grow on methanol in addition to methane (Kelly and Wood, 2010). If the SOX symbiont use methanol, that would imply that both symbionts, the MOX and SOX, could use the same energy source (a model is shown in Figure 5). This could give place to a cooperative or competitive behavior of the two symbionts depending on the environmental conditions. The competitive behavior could occur in an environment where methane is in low abundance but sulfide is in high concentration. These conditions typically occur in vents that host basaltic rocks, such as the LS vent field (Langmuir *et al.*, 1997; Charlou *et al.*, 2000). Under these conditions, the SOX that can use methanol and the methanotroph symbiont would be competing for the same energy source. The SOX symbiont that can use methanol would be negatively selected, and the SOX symbiont strains that use as energy source reduced sulfur compounds would be positively selected.

A contrasting behavior of the population would occur when methane is in high concentrations and sulfide is in low concentrations in the environment. This setting is typical of vents that host ultramafic rocks that originate from the earth's mantle, as methane and hydrogen are produced through serpentinization reactions (e.g. Rainbow) (Blumenberg, 2010; Martin *et al.*, 2008; Tivey, 2007). Under these conditions, the SOX symbiont strains that use as energy source reduced sulfur compounds would have a disadvantage and those strains that can use methanol would be positively selected. A cooperative behavior between the SOX with the potential to use methanol and the MOX symbiont could occur. Concentrations above 300 µM CH_4 can inhibit methane oxidation by *Bathymodiolus* symbionts probably caused by the accumulation of the toxic intermediates methanol, formaldehyde, and formate (Kochevar *et al.*, 1992). Under high methane concentrations in the environment, the MOX symbiont might make the excess methanol available in the extracellular environment. Methanol is a polar molecule that could cross the membrane by passive diffusion (Petersen and Dubilier, 2009). Indeed, preliminary experiments with the closest free-living relative of the MOX symbiont, *Methyloprofundus sedimenti* strain WF1 (Tavormina *et al.*, 2015), suggest the presence of methanol in the medium of log-phase cells that were grown under high methane concentrations (Patricia L. Tavormina, personal communication with Antony CP). The dual symbiosis with a SOX strain that can make use of the excess methanol might help the MOX to still be active when methane concentrations are present at inhibitory levels.

Depending on the environmental conditions, methanol concentrations in the gill could be playing a role shaping the community of SOX strains present.

The potential to use methanol by SOX symbionts was restricted to mussels that inhabit vents along the NMAR. The mussels acquire their symbionts with every new generation from the surrounding seawater (Won *et al.*, 2003b). Two major factors must influence whether a specific SOX symbiont strain colonizes the mussel – it has to be present in the environment and it must be selected from the plethora of free-living bacteria that exist in the surrounding seawater. The symbionts could come from mussels in the proximity, or from remote distances, carried by the water currents. Geographical barriers could slow down or stop the passive migration of the bacteria carried by the water flow (Van Dover, 2002). Two large geographical barriers exist between the NMAR and SMAR, namely the Romanche and Chain Transform Fault (Mercier and Speer, 1998; Heezen *et al.*, 1964). The distribution of the vent fauna suggests that these geographical barriers did not stop the dispersal of animals along the MAR (van der Heijden *et al.*, 2012; Gebruk *et al.*, 2000), and yet such geographical barriers might have stopped the transfer of the SOX strain with the MDH gene cluster to SMAR communities until now. Alternatively, the SOX strains with the MDH can reach the SMAR but they are outcompeted by local SOX strains under the local conditions.

Based on the genome coverage, the MDH gene cluster was present in all symbiont strains of *B. puteoserpentis*, but not all symbionts of *B. azoricus* had it. From all *B. azoricus* mussels, Rb had the highest number of strains with MDH. Our codon usage analyses suggest that the acquisition of *xoxF* and other genes in the cluster is a recent event in evolution as there has been not enough time to adapt to the codon usage of the rest of the genome. Thus, the potential to use methanol was not likely in a common ancestor that depending on the local conditions lost the genes. Instead, the SOX symbiont strain with the MDH cluster might have spread or is under current spreading among the NMAR and depending on the environmental conditions the trait might be positively selected. Two scenarios could explain the spread of the SOX strains with the MDH gene cluster among the NMAR: spreading from i) north to south, or ii) from south to north. Under scenario i) BazSym with the MDH gene cluster from northern sites of Rb traveled south. In Rb, they would have proliferated due to the environmental conditions, as this site hosts ultramafic rocks characterized by high methane and hydrogen concentrations (Charlou *et al.*, 2002; Douville *et al.*, 2002). Methane concentrations influence the activity and abundance of the MOX symbiont, which in turns determine how much

methane is oxidized to methanol (Boutet *et al.*, 2011). With the presence of higher methanol concentrations, SOX symbiont strains that could exploit this additional carbon and energy source would have an inherent benefit and are therefore likely to be positively selected. Under scenario ii) SOX symbiont strains with the MDH gene cluster traveled from the sites with *B. puteosepentis* –Log and Sem - to Rb, and further norther vents, where the gene cluster is being integrated into the population. As in i) the SOX strain carrying the MDH gene cluster would have had a higher chance of survival and proliferation at Rb thanks to the higher methane concentrations. Scenario ii) is more likely because the MDH-associated gene cluster appears to be already in all SOX symbiont strains of *B. puteoserpentis*. We might be seeing the process of establishment of a new symbiont population along the NMAR. Migration analysis of the SOX symbiont population could provide further evidence of whether most SOX bacteria are carried from south to north along the NMAR. Migration population studies have been applied mainly to eukaryotes, but they can be also applied to bacterial populations as described by Beerli and Felsenstein (2001).

Correlating the environment to the expression of the methanol dehydrogenase

We observed that only some SOX symbiont strains of *B. azoricus* encoded the MDH cluster, which suggests a high strain heterogeneity that confers different metabolic traits. This heterogeneity was observed between sites within the same vent field at a distance of less than 150 meters in LS. Within the LS vent field, the symbiont populations seem to be the most heterogeneous compared to the mussels sequenced from any other site. This was confirmed with the assembly metrics, as relatively worse quality assemblies were obtained for SOX symbionts of LS, a phenomenon that is known to result from strain heterogeneity (Supplementary Table 2) (Albertsen *et al.*, 2013).

Three factors will influence the gene expression: i) the number of SOX symbiont strains that encode the gene, ii) methane and methanol concentrations in the environment, as the expression of *xoxF* is likely changing depending on whether the substrate is present or not, as has been observed for other methylotrophs depending on the *in situ* conditions (Vorobev *et al.*, 2013), and iii) the time of sample fixation, as mussels that live in shallower chemosynthetic communities will take less time to come on board for processing compared to mussels that live in higher depths. The proportion of strains

encoding the MDH cluster in *B. azoricus* mussels followed a similar trend to the expression of *xoxF* and the four genes in the neighborhood (Figure 4). In Rb, the high proportion of strains encoding the MDH cluster, as well as the high methane concentrations (2.5 mM in end-member fluids) can account for the high expression of the MDH cluster. In agreement with the higher proportion of strains that encode the MDH cluster in LS-ET compared to LS-Ms, the expression of the MDH cluster was higher in LS-ET compared to LS-Ms. Even though Rb has a higher proportion of strains with the MDH compared to LS-ET, the expression of the MDH was not significantly different between Rb and LS-ET (p-value 0.474, Dunn's test). Thus, the symbionts in LS-ET must express the MDH at a higher rate compared to Rb to compensate for the lower number of strains with the MDH gene cluster. The methane concentrations in end-member fluids in LS range from 0.5 to 0.97 mM, the lowest for the sites we used for our study (Table 3). With a low concentration of methane, local gradients might have a higher impact on the expression and selection of a strain that can use a reduced compound that is probably not so abundant. Indeed, methane measurements of water column samples from LS showed that methane could be detected in LS-ET but not in LS-Ms (Cecile Konn and Cecile Cathalot, personal communication with Christian Borowski). The large variability of the environmental conditions in LS might have favored the presence of multiple strains that would allow the mussel to respond to the changing environmental conditions. Future studies should investigate whether the availability of methane and methanol within the LS vent field has a more drastic change at short geographical scales compared to other venting sites along the NMAR.

The expression data of BputSym give us the perfect study case to observe the change in gene expression without the influence of the strain heterogeneity or a large effect of variability in fixation time as the sites have relatively similar depths (2432-3045 m depth). Even though our data suggests that all BputSym encode the MDH gene cluster, we observed a high variability in expression between Sem, Log-Qt, and Log-Ir. The venting sites of Sem and Log have a similar depth, yet they have contrasting expression profiles. Semenov had the highest expression of the methanol dehydrogenase, followed by Log-Qt. Semenov has massive sulfide deposits and hosts ultramafic and basaltic rocks, which makes possible that the fluid composition is typical for any of these rock types (high or low methane) (Beltenev *et al.*, 2008). Unfortunately, the end-member fluid composition of this site is not yet available. Based on our transcriptome results, Semenov must have a high abundance of methanol, either produced by methane oxidation activity

of the MOX symbiont or present in the vent fluids. Logatchev is a methane-rich site (2.1 mM in end-member fluids) (Douville *et al.*, 2002; Charlou *et al.*, 2010). In Log-Qt, the methane concentration on the mussel beds was much higher than Log-Ir (63.68 vs 1.63 µM) (Table 4). The expression of the MDH from the SOX symbionts of BputSym is therefore likely influenced by the availability of methanol.

Conclusions

The genomes of the *Bathymodiolus* SOX symbionts from the northern MAR encode the potential to use methanol. The genes for the use of methanol were likely acquired recently through horizontal gene transfer. This is the first chemosynthetic symbiont known to use reduced sulfur compounds, hydrogen, and methanol as an energy source. We hypothesize that having symbionts with the potential to use methanol as an energy source might help the host to cope with the fluctuating environmental conditions in vents. The results of this study emphasize how important it is to acquire the environmental parameters along with the samples collected, as even bacteria that have the same 16S rRNA genes can have a high metabolic flexibility. The potential to use methanol by the SOX symbiont poses a new challenge while measuring the activity of methanotrophic symbionts. Previous studies have relied on the activity of the MDH enzyme as a proxy for methane oxidation, since measuring the direct activity of methane monooxygenases is difficult (Petersen and Dubilier, 2009). As we now know that two types of symbionts have this activity, the results from MDH-based enzyme assays on gill tissues should be interpreted with care.

Preliminary results and outlook

Confirming the link of the MDH gene region to the 16S rRNA with gene fluorescence *in situ* hybridization (gene-FISH)

To further confirm the link between the MDH gene cluster and the SOX symbiont, we designed 12 gene probes that specifically target the MDH gene cluster of the SOX symbiont. The probes were synthesized by PCR amplification and fluorescently labeled with Alexa594 according to Barrero-Canosa *et al.*, (2016). For the visual

identification of the SOX symbiont, we used the BMARt_193 rRNA probe (Duperron *et al.*, 2006) four times labeled with ATTO488. Preliminary results showed that i) the gene signal of the MDH overlaps with the signal of bacteria labeled with the specific 16S rRNA of the SOX symbiont and that ii) the distribution of the SOX symbiont with the potential to use methanol in *B. azoricus* mussels from LS-ET has a patchy distribution (Figure 6). These results confirmed our genome analysis and showed that the gene was not the result of miss-assembly or spurious binning. Further experiments to confirm differences between mussels of the venting regions used for sequencing are currently being developed.

Do all SOX symbionts of Log-Qt and Log-Ir encode the MDH?

We found that the SOX symbiont from Log-Ir had a significantly lower expression of the genes for methanol oxidation. Our current genome data suggests that all SOX symbiont strains of *B. puteoserpentis* encode the MDH. The strain heterogeneity also seems to be lower considering that the SOX symbiont genomes obtained from *B. puteoserpentis* mussels were among the best assemblies considering the N50 metric. To confirm that the differences between Log-Ir and Log-Qt are not the bias of yet unseen heterogeneity, we will sequence metagenomes of the same gill homogenates of the individuals that were used for transcriptome sequencing.

Are the SOX symbionts true methylotrophs?

During the Meteor Cruise 2016, we plan to test whether the sulfur-oxidizing symbiont can oxidize methanol. A crucial part of this experiment will be to separate the SOX from the MOX symbionts before incubation. Preliminary experiments by J. Petersen and R. Ansorge have shown that the SOX can be separated from the MOX symbiont through gill homogenization and filtration, as the SOX symbiont is smaller than the MOX. The SOX symbiont will be separated from the MOX with this technique, as well as differential centrifugation. The enriched SOX fraction, MOX fraction, and *M. sediment* str. WF1 will be incubated with increasing methanol concentrations (0, 10 µM, 100 µM, and 12 mM) for up to 96 hrs. We expect to see formate production over time by using a colorimetric assay. Preliminary results have shown that formate is produced in almost pure fractions of SOX symbionts fed with methanol.

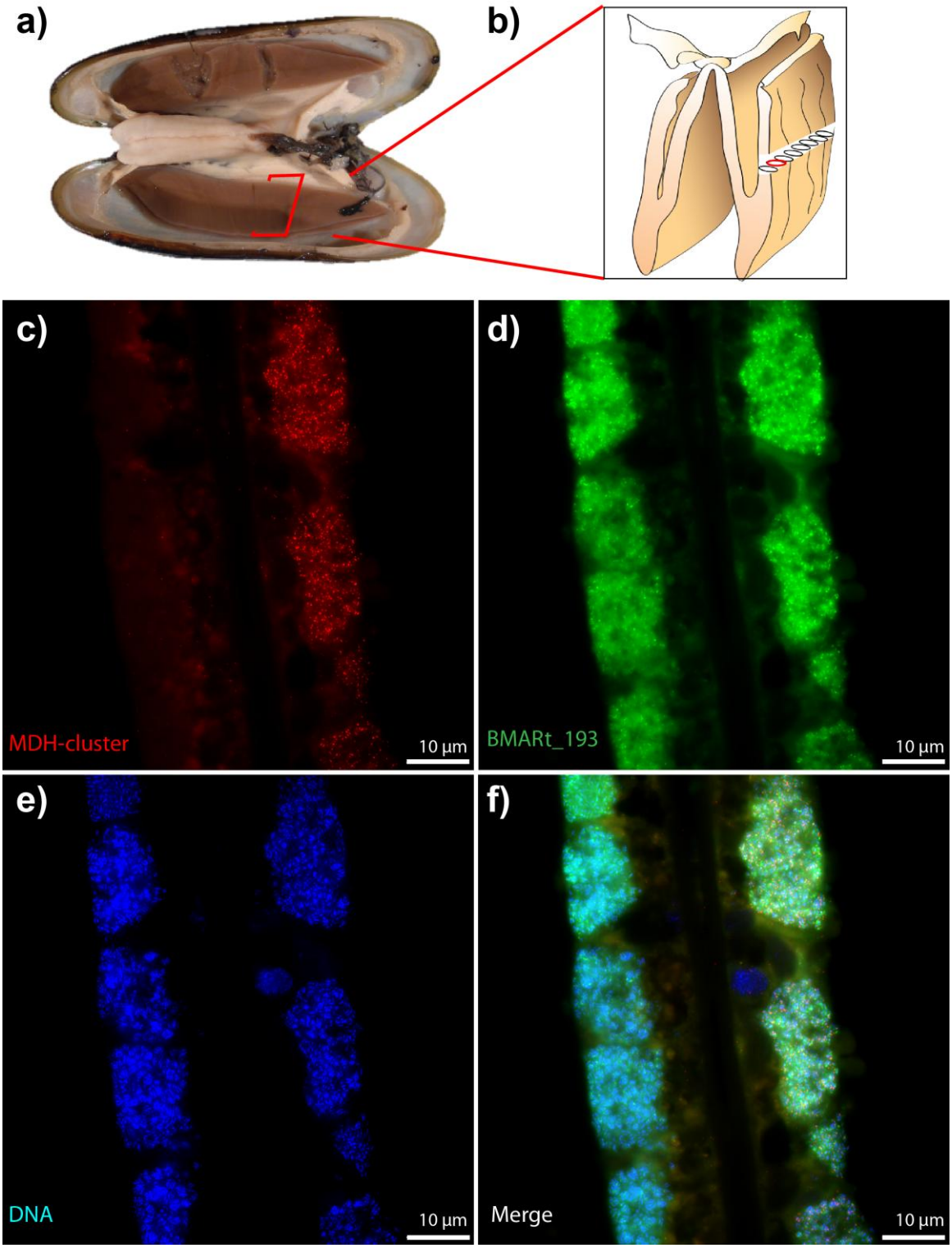


Figure 6. *Bathymodiolus* mussels have oversized gills (a) composed of multiple filaments (b). A cross-section of a single gill filament of *B. azoricus* from Lucky Strike-Eiffel Tower hybridized with gene-FISH probes targeting the MDH cluster (c) and the SOX symbiont (d), counterstained with DAPI (e) shows that the distribution of the SOX-symbiont strains encoding the MDH gene cluster has a mosaic-like distribution. Section (f) shows an overlay of c-e. b) was produced by A. Assié.

Materials and methods

Sample collection and sequencing

Sampling sites of the *Bathymodiolus* mussels used for sequencing in this study are described in Supplementary Table 1. Gill homogenates were used for DNA extractions to minimize the effects of symbiont strain variability across the gill, when material was available, as described by Zhou *et al*. (1996) or with the DNeasy blood and tissue kit (Qiagen, Germany). Genomic DNA libraries were generated with the Illumina TruSeq DNA Sample Prep Kit (BioLABS, Frankfurt am Main, Germany). Sequencing was done as either 100, 150, or 250 bp paired-end reads on an Illumina HiSeq 2500 machine at the Max Planck Genome Center (Supplementary Table 1).

Phylogenetic analysis

The nucleotide alignment from Taubert *et al*. (2015) was used as a comprehensive database of methanol dehydrogenase sequences. Sequences were translated to protein sequences and aligned with MAFFT (Katoh *et al.*, 2002). A maximum likelihood tree was reconstructed using RAxML (Stamatakis, 2006) with 100 bootstrap replicates.

Estimation of differentially represented genes in the metagenome and metatranscriptome

The average nucleotide identity (ANI) was estimated with enve-omics.ce.gatech.edu/ani. The metagenome and metatranscriptome reads of each individual were mapped against the *B. puteoserpentis* SOX from Log-Ir with a minimum identity of 95%. This threshold was selected to avoid spurious mapping and because most of the genome had a higher similarity (Supplementary Figure 1). We removed adapters, quality trimmed (Q2), and error corrected the metagenome and metatranscriptome reads with BBduk and BBnorm V35 (Bushnell B. - sourceforge.net/projects/bbmap/). Reads were mapped to the *B. puteoserpentis* SOX symbiont with a minimum similarity of 95% with Bbmap, and only the reads that mapped to the reference genome were further considered for normalization. The number of reads mapping to each coding sequence (CDS) was estimated with FeatureCounts (Liao *et al.*, 2014). To compare metagenome

and metatranscriptome sequencing data among individuals and between sites, a normalization factor was estimated with trimmed mean of M-values (TMM) (calcNormFactors, edgeR) (Robinson and Oshlack, 2010). The counts were converted to reads per kilobase per million (RPKM). To determine whether the venting areas or the mussel species had significantly different ratios of genome or transcriptome coverage for the gene regions of interest, we used a second normalization step. We used the ratio of the normalized RPKM counts of the MDH gene cluster to the single copy gene *gyrA*. This strategy was used for both, the genome and transcriptome data. We chose *gyrA* because this single copy gene was expressed in all samples and was among the top 52 genes with the least variation in expression among our samples. *gyrA* is also commonly used in qPCR studies for gene normalization (Rocha *et al.*, 2015). The ratios of gene coverage or gene expression were compared between sites and between species with Kruskal-Wallis test in R (R Development Core Team, 2011). When comparing more than two categories, Kruskal Wallis cannot tell which category is different. Thus, we used Dunn's test as a *post hoc* analysis to compare which of the venting sites had significantly different ratios of expression or gene coverage with the FSA package implemented in R (Ogle, 2016). Because multiple pairwise comparisons were calculated, p-values were corrected with the false-discovery rate.

GeneFISH on *Bathymodiolus* spp. gill tissue

The potential metabolic differences between SOX strains of five mussels from different locations along the Mid-Atlantic Ridge were visualized with direct-geneFISH (Barrero-Canosa *et al.*, 2016). DNA was extracted from a *B. azoricus* gill homogenate with the DNeasy blood and tissue kit according to the manufacturer instructions (Qiagen, Germany). Part of the MDH gene cluster was amplified by PCR with the Phusion® DNA Polymerase (Thermo Fisher Scientific, Waltham, MA) with the PCR primers described in Supplementary Table 3. The following conditions were used: initial denaturation at 98°C for 30 s, 35 cycles at 98°C for 10 s, 64.6°C for 30 s and 72°C for 4 min, and a final elongation at 72°C for 10 min. The PCR product was cloned with the TOPO® XL cloning kit (Invitrogen, Carlsbad, CA, USA) according to the manufacturer descriptions. A total of 12 polynucleotide probes were designed to target the gene cluster of the MDH with Geneious V.9. (Kearse *et al.*, 2012). The PCR primers used to amplify the probes, as well as the target gene of each amplified probe, are shown in Supplementary Table 3. The

DNA of the plasmid was used as a template to synthesize the polynucleotides probes by PCR with the Taq DNA Polymerase (5 PRIME, Germany) using the following conditions: initial denaturation at 95°C for 5 min, 35 cycles at 95°C for 1 min, 55°C for 1 min 30 s and 72°C for 2 min, and a final elongation step at 72°C for 10 min. As a negative control, we used the non-sense polynucleotide probe NonPolyPr350 described by Moraru *et al*., (2010). PCR products were purified with QiaQuick PCR purification kit (Qiagen, Germany) and chemically labeled with Alexa594 dye with ULYSIS® Nucleic Acid Labeling Kit (Molecular Probes, Oregon, USA).

Cross sections (3 µm) of formalin-fixed paraffin-embedded gills were immobilized on poly-L-lysine-coated glass (Sigma-Aldrich, San Louis, Missouri, USA). The gill sections were dewaxed with decreasing ethanol series of 96, 80, 70 and 50% (v/v) for 15, 3, 3 and 3 min, respectively. For permeabilization, the sections were incubated with lysozyme solution (0.5 mg ml^{-1} lysozyme (Sigma-Aldrich, Germany) in 1X PBS, 0.05 M EDTA and 0.1 M Tris-HCl pH 7.8) for 1 h on ice. Tissue sections were washed with milliQ water followed by 1 min in ethanol 96% (v/v) and air-dried. The gene- and rRNA-targeted probes were hybridized simultaneously in a hybridization buffer containing 35% formamide (5X SSC buffer, 20% dextran sulfate (v/v), 20 mM EDTA, 35% formamide (v/v), 0.1% SDS (v/v), sheared salmon sperm 0.25 mg ml^{-1} (Invitrogen, MA, USA), yeast RNA 0.25 mg ml^{-1} (Invitrogen, MA, USA), 1% blocking reagent for nucleic acids (Roche, Switzerland)), 31 pg μl^{-1} per individual Alexa594 labeled MMI-SOX polynucleotide probes and 5 ng μl^{-1} of the SOX 16S rRNA probe BMARt_193 (Duperron *et al.*, 2006). As a negative control, we used 372 pg μl^{-1} of Alexa594 labeled NonPoly350 polynucleotide probe (Moraru *et al.*, 2010). For the denaturation of the gene target, the samples were incubated for 40 min at 85°C. After denaturation, the samples were hybridized at 46°C for 4 h. Following the hybridization, the sections were rinsed in pre-warmed 48°C washing buffer (0.07 M NaCl, 0.02 M Tris-HCl pH 7.8, 5 mM EDTA pH 8, and 0.01% SDS (v/v)) and transferred to fresh pre-warmed washing buffer for 15 min followed by 20 min in 1X PBS, 1 min in milliQ water, 1 min 96% ethanol and air-dried. Sections were counter-stained with 1 µg μl^{-1} 4', 6-Diamino-2-Phenylindole, Dihydrochloride (DAPI) for 10 min at room temperature and consecutively washed with milliQ water and 96% ethanol (v/v). After drying, the sections were mounted in a mixture of Citifluor (Agar Scientific, Unite Kingdom) and VECTASHIELD Mounting Medium (Vector Laboratories, CA, USA) (4:1). Sections were stored at -20°C until observation with super resolution – structure illumination

microscopy SR-SIM with the ELYRA PS.1 LSM 780 microscope (Carl Zeiss, Jena, Germany)

Acknowledgements

We thank the captains and crews of the M64-2, ODEMAR and BioBaz cruises involved in the sampling effort. We especially thank Adrien Assié, Christian Borowski and Corinna Breusing for providing samples. We thank Niko Leisch for useful discussions regarding the FISH. The planning for the incubations on board of the Meteor cruise 126 was done with the help of Manuel Liebeke, Maxim Rubin Blum, and Judith Zimmermann. We thank Carolin Peter for testing the formate assay kit with seawater and Dolma Michellod for her willingness to perform the methanol incubation experiments during the Meteor Cruise. We thank the Max Planck Society for funding the project, and the DAAD for funding to L.S.

Statement of competing interests

The authors declare no competing interests.

References

Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**: 533–538.

Barrero-Canosa J., Moraru C., Zeugner L., Fuchs B.M., Amann R. Direct-geneFISH: A simplified protocol for the simultaneous detection and quantification of genes and rRNA in microorganisms. *Environ. Microbiol*. e-pub ahead of print, doi: 10.1111/1462-2920.13432.

Beerli P, Felsenstein J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc Natl Acad Sci* **98**: 4563–4568.

Beinart RA, Sanders JG, Faure B, Sylva SP, Lee RW, Becker EL, *et al.* (2012). Evidence for the role of endosymbionts in regional-scale habitat partitioning by hydrothermal vent symbioses. *Proc Natl Acad Sci* **109**: E3241–E3250.

Beltenev V, Ivanov V, Rozhdestvenskaya I, Cherkashov G, Stepanova T, Shilov V, *et al.* (2008). New data about hydrothermal fields on the Mid-Atlantic Ridge between 11-14 N: 32nd Cruise of R/V Professor Logatchev. *Geochem Geophys Geosyst* **10**: 1029.

Blumenberg M. (2010). Microbial chemofossils in specific marine hydrothermal and methane cold seep settings. In: Kiel S (ed) Vol. 33. *The vent and seep biota*. Springer Netherlands: Dordrecht, pp 73–106.

Boden R, Kelly DP, Murrell JC, Schäfer H. (2010). Oxidation of dimethylsulfide to tetrathionate by *Methylophaga thiooxidans* sp. nov.: A new link in the sulfur cycle. *Environ Microbiol* **12**: 2688–2699.

Boutet I, Ripp R, Lecompte O, Dossat C, Corre E, Tanguy A, *et al*. (2011). Conjugating effects of symbionts and environmental factors on gene expression in deep-sea hydrothermal vent mussels. *BMC Genomics* **12**: 530.

Charlou JL, Donval JP, Douville E, Jean-Baptiste P, Radford-Knoery J, Fouquet Y, *et al.* (2000). Compared geochemical signatures and the evolution of Menez Gwen (37°50′N) and Lucky Strike (37°17′N) hydrothermal fluids, south of the Azores Triple Junction on the Mid-Atlantic Ridge. *Chem Geol* **171**: 49–75.

Charlou JL, Donval JP, Fouquet Y, Jean-Baptiste P, Holm N. (2002). Geochemistry of high H_2 and CH_4 vent fluids issuing from ultramafic rocks at the Rainbow hydrothermal field (36°14′N, MAR). *Chem Geol* **191**: 345–359.

Charlou JL, Donval JP, Konn C, OndréAs H, Fouquet Y, Jean-Baptiste P, *et al*. (2010). High production and fluxes of H_2 and CH_4 and evidence of abiotic hydrocarbon synthesis by serpentinization in ultramafic-hosted hydrothermal systems on the Mid-Atlantic Ridge. In: Rona PA, Devey CW, Dyment J, Murton BJ (eds). *Diversity Of Hydrothermal Systems On Slow Spreading Ocean Ridges*. American Geophysical Union, pp 265–296.

Chistoserdova L. (2011). Modularity of methylotrophy, revisited. *Environ Microbiol* **13**: 2603–2622.

Chu F, Lidstrom ME. (2016). XoxF acts as the predominant methanol dehydrogenase in the Type I methanotroph *Methylomicrobium buryatense*. *J Bacteriol* **198**: 1317–1325.

Davis JJ, Olsen GJ. (2010). Characterizing the native codon usages of a genome: An axis projection approach. *Mol Biol Evol* **28**: 211–221.

DeChaine EG, Bates AE, Shank TM, Cavanaugh CM. (2006). Off-axis symbiosis found: Characterization and biogeography of bacterial symbionts of *Bathymodiolus* mussels from Lost City hydrothermal vents. *Environ Microbiol* **8**: 1902–1912.

Distel DL, Lee HK, Cavanaugh CM. (1995). Intracellular coexistence of methano- and thioautotrophic bacteria in a hydrothermal vent mussel. *Proc Natl Acad Sci* **92**: 9598–9602.

Dixon JL, Beale R, Nightingale PD. (2011). Microbial methanol uptake in northeast Atlantic waters. *ISME J* **5**: 704–716.

Douville E, Charlou JL, Oelkers EH, Bienvenu P, Jove Colon CF, Donval JP, *et al*. (2002). The Rainbow vent fluids (36°14′N, MAR): The influence of ultramafic rocks and phase separation on trace metal content in Mid-Atlantic Ridge hydrothermal fluids. *Chem Geol* **184**: 37–48.

Dubilier N, Bergin C, Lott C. (2008). Symbiotic diversity in marine animals: The art of harnessing chemosynthesis. *Nat Rev Micro* **6**: 725–740.

Duperron S, Bergin C, Zielinski F, Blazejak A, Pernthaler A, McKiness ZP, *et al*. (2006). A dual symbiosis shared by two mussel species, *Bathymodiolus azoricus* and *Bathymodiolus puteoserpentis* (Bivalvia: Mytilidae), from hydrothermal vents along the northern Mid-Atlantic Ridge. *Environ Microbiol* **8**: 1441–1447.

Duperron S, Sibuet M, MacGregor BJ, Kuypers MMM, Fisher CR, Dubilier N. (2007). Diversity, relative abundance and metabolic potential of bacterial endosymbionts in three *Bathymodiolus* mussel species from cold seeps in the Gulf of Mexico. *Environ Microbiol* **9**: 1423–1438.

Fontanez KM, Cavanaugh CM. (2014). Evidence for horizontal transmission from multilocus phylogeny of deep-sea mussel (Mytilidae) symbionts: Horizontal transmission of mussel symbionts. *Environ Microbiol* 3608–3621.

Gebruk AV, Chevaldonné P, Shank T, Lutz RA, Vrijenhoek RC. (2000). Deep-sea hydrothermal vent communities of the Logatchev area (14˚ 45′ N, Mid-Atlantic Ridge): Diverse biotopes and high biomass. *J Mar Biol Assoc UK* **80**: 383–393.

Giovannoni SJ, Hayakawa DH, Tripp HJ, Stingl U, Givan SA, Cho J-C, *et al.* (2008). The small genome of an abundant coastal ocean methylotroph. *Environ Microbiol* **10**: 1771–1782.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**: 1072–1075.

Heezen BC, Bunce ET, Hersey JB, Tharp M. (1964). Chain and romanche fracture zones. *Deep Sea Res Oceanogr Abstr* **11**: 11–33.

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.

Humphris SE, Fornari DJ, Scheirer DS, German CR, Parson LM. (2002). Geotectonic setting of hydrothermal activity on the summit of Lucky Strike Seamount (37°17′N, Mid-Atlantic Ridge). *Geochem Geophys Geosystems* **3**: 1–25.

Ikuta T, Takaki Y, Nagai Y, Shimamura S, Tsuda M, Kawagucci S, *et al.* (2015). Heterogeneous composition of key metabolic gene clusters in a vent mussel symbiont population. *ISME J*. e-pub ahead of print, doi: 10.1038/ismej.2015.176.

Jewell T, Huston SL, Nelson DC. (2008). Methylotrophy in freshwater *Beggiatoa alba* strains. *Appl Environ Microbiol* **74**: 5575–5578.

Katoh K, Misawa K, Kuma K, Miyata T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**: 3059–3066.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, *et al.* (2012). Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647–1649.

Kelly DP, Wood AP. (2010). Isolation and characterization of methanotrophs and methylotrophs: Diversity of methylotrophic organisms and of one-carbon substrates. In: Timmis KN (ed). *Handbook of Hydrocarbon and Lipid Microbiology*. Springer Berlin Heidelberg, pp 3827–3845.

Keltjens JT, Pol A, Reimann J, Op den Camp HJM. (2014). PQQ-dependent methanol dehydrogenases: Rare-earth elements make a difference. *Appl Microbiol Biotechnol* **98**: 6163–6183.

Kochevar RE, Childress JJ, Fisher CR, Minnich E. (1992). The methane mussel: Roles of symbiont and host in the metabolic utilization of methane. *Mar Biol* **112**: 389–401.

Lang AS, Zhaxybayeva O, Beatty JT. (2012). Gene transfer agents: Phage-like elements of genetic exchange. *Nat Rev Microbiol* **10**: 472–482.

Langmuir C, Humphris S, Fornari D, Van Dover C, Von Damm K, Tivey MK, *et al.* (1997). Hydrothermal vents near a mantle hot spot: The Lucky Strike vent field at 37°N on the Mid-Atlantic Ridge. *Earth Planet Sci Lett* **148**: 69–91.

Liao Y, Smyth GK, Shi W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923–930.

Martin W, Baross J, Kelley D, Russell MJ. (2008). Hydrothermal vents and the origin of life. *Nat Rev Microbiol* **6**: 805–814.

McDermott JM, Seewald JS, German CR, Sylva SP. (2015). Pathways for abiotic organic synthesis at submarine hydrothermal fields. *Proc Natl Acad Sci* **112**: 7668–7672.

Meier D, Bach W, Girguis PR, Gruber-Vodicka H, Reeves EP, Richter M, *et al.* (2016). Heterotrophic *Proteobacteria* in the vicinity of diffuse hydrothermal venting. *Environ Microbiol*. e-pub ahead of print, doi: 10.1111/1462-2920.13304.

Mercier H, Speer KG. (1998). Transport of bottom water in the Romanche Fracture Zone and the Chain Fracture Zone. *J Phys Oceanogr* **28**: 779–790.

Moraru C, Lam P, Fuchs BM, Kuypers MMM, Amann R. (2010). GeneFISH – an *in situ* technique for linking gene presence and cell identity in environmental microorganisms. *Environ Microbiol* **12**: 3057–3073.

Nercessian O, Bienvenu N, Moreira D, Prieur D, Jeanthon C. (2005). Diversity of functional genes of methanogens, methanotrophs and sulfate reducers in deep-sea hydrothermal environments. *Environ Microbiol* **7**: 118–132.

Ogle DH. (2016). FSA: Fisheries stock analysis.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**: 1043–1055.

Petersen JM, Dubilier N. (2009). Methanotrophic symbioses in marine invertebrates: Methanotrophic symbioses in marine invertebrates. *Environ Microbiol Rep* **1**: 319–335.

Petersen JM, Zielinski FU, Pape T, Seifert R, Moraru C, Amann R, *et al.* (2011). Hydrogen is an energy source for hydrothermal vent symbioses. *Nature* **476**: 176–180.

Petersen S, Kuhn K, Kuhn T, Augustin N, Hékinian R, Franz L, *et al.* (2009). The geological setting of the ultramafic-hosted Logatchev hydrothermal field (14°45′N, Mid-Atlantic Ridge) and its influence on massive sulfide formation. *Lithos* **112**: 40–56.

R Development Core Team. (2011). R: A language and environment for statistical computing. *R Found Stat Comput*. http://www.r-project.org/ (Accessed February 1, 2012).

Raggi L, Schubotz F, Hinrichs K-U, Dubilier N, Petersen JM. (2013). Bacterial symbionts of *Bathymodiolus* mussels and *Escarpia* tubeworms from Chapopote, an asphalt seep in the southern Gulf of Mexico. *Environ Microbiol* **15**: 1969–1987.

Reeburgh WS. (2007). Oceanic methane biogeochemistry. *Chem Rev* **107**: 486–513.

Reeves EP, McDermott JM, Seewald JS. (2014). The origin of methanethiol in midocean ridge hydrothermal fluids. *Proc Natl Acad Sci* **111**: 5474–5479.

Robinson MD, Oshlack A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**: R25.

Rocha DJP, Santos CS, Pacheco LGC. (2015). Bacterial reference genes for gene expression studies by RT-qPCR: survey and analysis. *Antonie Van Leeuwenhoek* **108**: 685–693.

Rodriguez-R LM, Konstantinidis KT. (2010). Bypassing cultivation to identify bacterial species. *Issues*.

Rossel PE, Stubbins A, Hach PF, Dittmar T. (2015). Bioavailability and molecular composition of dissolved organic matter from a diffuse hydrothermal system. *Mar Chem* **177**, **Part 2**: 257–266.

Sayavedra L, Kleiner M, Ponnudurai R, Wetzel S, Pelletier E, Barbe V, *et al.* (2015). Abundant toxin-related genes in the genomes of beneficial symbionts from deep-sea hydrothermal vent mussels. *eLife* **4**. e-pub ahead of print, doi: 10.7554/eLife.07966.

Schrenk MO, Kelley DS, Delaney JR, Baross JA. (2003). Incidence and diversity of microorganisms within the walls of an active deep-sea sulfide chimney. *Appl Environ Microbiol* **69**: 3580–3592.

Sheik CS, Anantharaman K, Breier JA, Sylvan JB, Edwards KJ, Dick GJ. (2015). Spatially resolved sampling reveals dynamic microbial communities in rising hydrothermal plumes across a back-arc basin. *ISME J* **9**: 1434–1445.

Shock E, Canovas P. (2010). The potential for abiotic organic synthesis and biosynthesis at seafloor hydrothermal systems. *Geofluids* **10**: 161–192.

Sowell SM, Abraham PE, Shah M, Verberkmoes NC, Smith DP, Barofsky DF, *et al.* (2011). Environmental proteomics of microbial plankton in a highly productive coastal upwelling system. *ISME J* **5**: 856–865.

Stamatakis A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688.

Taubert M, Grob C, Howat AM, Burns OJ, Dixon JL, Chen Y, *et al.* (2015). *xoxF* encoding an alternative methanol dehydrogenase is widespread in coastal marine environments. *Environ Microbiol* **17**: 3937–3948.

Tavormina PL, Hatzenpichler R, McGlynn S, Chadwick G, Dawson KS, Connon SA, *et al.* (2015). *Methyloprofundus sedimenti* gen. nov., sp. nov., an obligate methanotroph from ocean sediment belonging to the 'deep sea-1' clade of marine methanotrophs. *Int J Syst Evol Microbiol* **65**: 251–259.

Tivey M. (2007). Generation of seafloor hydrothermal vent fluids and associated mineral deposits. *Oceanography* **20**: 50–65.

Van Dover CL. (2002). Evolution and biogeography of deep-sea vent and seep invertebrates. *Science* **295**: 1253–1257.

Voglesonger KM, Holloway JR, Dunn EE, Dalla-Betta PJ, O'Day PA. (2001). Experimental abiotic synthesis of methanol in seafloor hydrothermal systems during diking events. *Chem Geol* **180**: 129–139.

Von Damm KL, Bray AM, Buttermore LG, Oosting SE. (1998). The geochemical controls on vent fluids from the Lucky Strike vent field, Mid-Atlantic Ridge. *Earth Planet Sci Lett* **160**: 521–536.

Vorobev A, Beck DAC, Kalyuzhnaya MG, Lidstrom ME, Chistoserdova L. (2013). Comparative transcriptomics in three *Methylophilaceae* species uncover different strategies for environmental adaptation. *PeerJ* **1**: e115.

Won Y, Hallam SJ, O'Mullan GD, Vrijenhoek RC. (2003a). Cytonuclear disequilibrium in a hybrid zone involving deep-sea hydrothermal vent mussels of the genus *Bathymodiolus*. *Mol Ecol* **12**: 3185–3190.

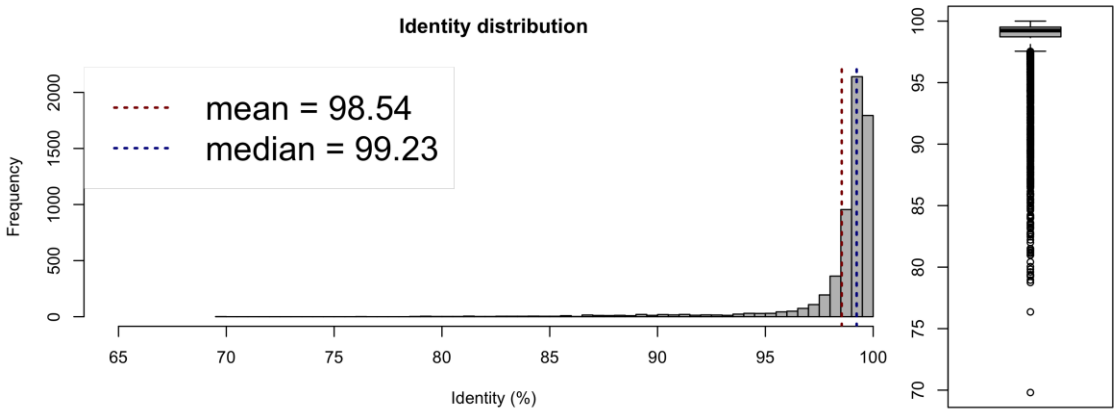
Won YJ, Hallam SJ, O'Mullan GD, Pan IL, Buck KR, Vrijenhoek RC. (2003b). Environmental acquisition of thiotrophic endosymbionts by deep-sea mussels of the genus *Bathymodiolus*. *Appl Environ Microbiol* **69**: 6785.

Zhou J, Bruns MA, Tiedje JM. (1996). DNA recovery from soils of diverse composition. *Appl Environ Microbiol* **62**: 316–322.

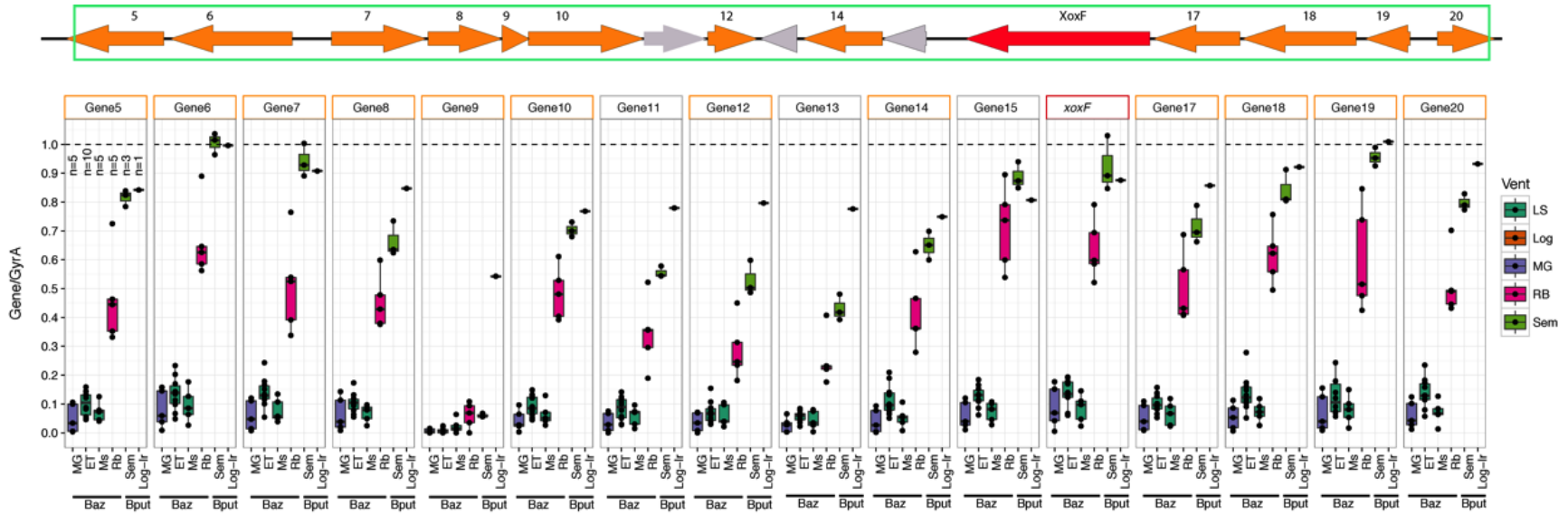
Zielinski FU, Gennerich H-H, Borowski C, Wenzhöfer F, Dubilier N. (2011). *In situ* measurements of hydrogen sulfide, oxygen, and temperature in diffuse fluids of an ultramafic-hosted hydrothermal vent field (Logatchev, 14° 45′ N, Mid-Atlantic Ridge): Implications for chemosymbiotic bathymodiolin mussels. *Geochem Geophys Geosystems* **12**.

Zwart JMM de, Nelisse PN, Kuenen JG. (1996). Isolation and characterization of *Methylophaga sulfidovorans* sp. nov.: An obligately methylotrophic, aerobic, dimethylsulfide oxidizing bacterium from a microbial mat. *FEMS Microbiol Ecol* **20**: 261–270.

Supplementary Figures



Supplementary Figure 1. Average nucleotide identity (ANI) distribution between the SOX symbionts of *B. azoricus* and *B. puteoserpentis*. The ANI was estimated with the web server enve-omics.ce.gatech.edu/ani



Supplementary Figure 2. Read genome coverage variation of the16 genes of the MDH region. Gene coverage is normalized to the coverage of the single copy gene *gyrA*. The red arrow corresponds to the key gene for methanol oxidation *xoxF*, orange arrows are genes of the MDH region with a functional annotation, and gray arrows are hypothetical proteins. Baz = Genome information from *B. azoricus* reads. Bput = Genome information from *B. puteoserpentis* reads. Gene 5 = PQQ-dependent catabolism-associated beta-propeller protein; 6 = ABC transporter, substrate binding protein, PQQ-dependent alcohol dehydrogenase system; 7 = Coenzyme PQQ synthesis protein B; 8 = pyrroloquinoline-quinone synthase; 9 = Coenzyme PQQ synthesis protein D; 10 = Coenzyme PQQ synthesis protein E; 12 = protein of unknown function DUF302; 14 = quinoprotein dehydrogenase-associated probable ABC transporter substrate-binding protein; 17 = S-formylglutathione hydrolase; 18 = S-(hydroxymethyl)glutathione dehydrogenase; 19 = S-(hydroxymethyl)glutathione synthase; 20 = MxaD.

Supplementary Tables

Supplementary Table 1. Sequencing data analyzed in this study.

Cruise	Ecosystem	Site	Species	DNA lib	RNA lib	Sampling date	Latitude	Longitude	Depth (m)	DNA seq	RNA seq
M78-2***	Vent	5° SMAR, Clueless	*B. sp.*$^{\$\$}$	Clue	-		-4.803267	-12.3718	-2971.5	-	-
ATA 57 ROV 7/2*	Vent	5° SMAR, Turtle Pits	*B. sp.* $^{\$\$}$	1501A, 1593A	-	21-Jan-08	-4.809167	-12.374383	-2989	13.4$^{\&\&}$	-
M78-2 319***	Vent	9° SMAR, Lilliput	*B. sp.* $^{\$\$}$	Lil	-	29-Apr-09	-9.5475	-13.2088	-1490	15$^{\&}$	-
M114-2	Seep	Ch-GoM	*B. brooksi*$^{\$\$}$	1600C	-	20-Mar-15	21.90002	-93.43525	-2925	17.5$^{\#}$	-
M114-2	Seep	Ch-GoM	*B. heckerae*$^{\$\$\$}$	1600A	-	20-Mar-15	21.90002	-93.43525	-2925	3.6$^{\#}$	-
M114-2	Seep	Ch-GoM	*B. heckerae*$^{\$\$\$}$	1600B	-	20-Mar-15	21.90002	-93.43525	-2925	20.4$^{\#}$	-
M114-2	Seep	Ch-GoM	*B. heckerae*$^{\$\$\$}$	1600D	-	14-Mar-15	21.90005	-93.43540	-2925	30.6$^{\#}$	-
M114-2	Seep	Ch-GoM	*B. heckerae*$^{\$\$\$}$	1600E	-	14-Mar-15	21.90005	-93.43540	-2925	12.6$^{\#}$	-
M114-2	Seep	Ch-GoM	*B. heckerae*$^{\$\$\$}$	1600O	-	10-Mar-15	21.89999205	-93.43524	-2925	17.7$^{\#}$	-
M114-2	Seep	Ch-GoM	*B. heckerae*$^{\$\$\$}$	1600P	-	10-Mar-15	21.89999205	-93.43524	-2925	33.0$^{\#}$	-
M114-2	Seep	Ch-GoM	*B. heckerae*$^{\$\$\$}$	1600Q	-	10-Mar-15	21.89999205	-93.43524	-2925	10.4$^{\#}$	-
AT26-23 Dive 4763*	Vent	Crab Spa-EPR	*B. thermophilus*$^{\$}$	1600R	-	8-Nov-14	9.839628	-104.292217	-2512	16.6$^{\#}$	-
AT26-23 Dive 4763	Vent	Crab Spa-EPR	*B. thermophilus*$^{\$}$	1600S	-	8-Nov-14	9.839628	-104.292217	-2512	18.0$^{\#}$	-
AT26-23 Dive 4763	Vent	Crab Spa-EPR	*B. thermophilus*$^{\$}$	1600T	-	8-Nov-14	9.839628	-104.292217	-2512	17.2$^{\#}$	-
R/V Atlantis AT26-13^{+}	Seep	GC600-GoM	*B. brooksi*$^{\$\$}$	M3	-	11-Apr-14	28.384564	-88.580279	-1225	60$^{\&}$	-
R/V Atlantis AT26-13^{+}	Seep	GC600-GoM	*B. brooksi*$^{\$\$}$	M4	-	11-Apr-14	28.384564	-88.580279	-1225	60$^{\&}$	-
Logatchev 2005 M64-2 Dive 244-9-1 *	Vent	Log-Ir	*Bput*$^{\$\$}$	Log	-	17-May-05	14.753	-44.97933	-3009	217.6$^{\#}$	-
Logatchev 2005 M64-2 232 ROV5-	Vent	Log-Ir	*Bput*$^{\$\$}$	2065D	1848P	26-May-05	14.753	-44.978667	-3042	-	17.8$^{\&}$
Logatchev 2005 M64-2 232 ROV5-	Vent	Log-Ir	*Bput*$^{\$\$}$	2065E	1848Q	26-May-05	14.753	-44.978667	-3042	-	17.9$^{\&}$
Logatchev 2005 M64-2 232 ROV5-	Vent	Log-Ir	*Bput*$^{\$\$}$	2065F	1848R	26-May-05	14.753	-44.978667	-3042	-	18.1$^{\&}$
Logatchev 2005 M64-2 281 ROV-3	Vent	Log-Qt	*Bput*$^{\$\$}$	2065A	1848M	30-May-05	14.753167	-44.978833	-3045	-	17.8$^{\&}$
Logatchev 2005 M64-2 281 ROV-3	Vent	Log-Qt	*Bput*$^{\$\$}$	2065B	1848N	30-May-05	14.753167	-44.978833	-3045	-	16.5&

Cruise	Ecosystem	Site	Species	DNA lib	RNA lib	Sampling date	Latitude	Longitude	Depth (m)	DNA seq	RNA seq
Logatchev 2005 M64-2 281 ROV-3	Vent	Log-Qt	*Bput*[ss]	2065C	1848 O	30-May-05	14.753167	-44.978833	-3045	-	17.6[&]
BioBaz 2013	Vent	LS-ET	*B. azoricus*[ss]	1586A	1848 D	10-Aug-13	37.28911833	-32.27549	-1690	14.1[#]	17.9[&]
BioBaz 2013	Vent	LS-ET	*B. azoricus*[ss]	1586E	1848 E	10-Aug-13	37.28911833	-32.27549	-1690	13.7[#]	16.3[&]
BioBaz 2013	Vent	LS-ET	*B. azoricus*[ss]	1586C	1848 F	10-Aug-13	37.28911833	-32.27549	-1690	16.8[#]	15.0[&]
BioBaz 2013	Vent	LS-ET	*B. azoricus*[ss]	1586B	-	10-Aug-13	37.28911833	-32.27549	-1690	13.6[#]	-
BioBaz 2013	Vent	LS-ET	*B. azoricus*[ss]	1586D	-	10-Aug-13	37.28911833	-32.27549	-1690	13.6[#]	-
BioBaz 2013	Vent	LS-ET2	*B. azoricus*[ss]	1586F	-	17-Aug-13	37.28922	-32.27561	-1689	14.3[#]	-
BioBaz 2013	Vent	LS-ET2	*B. azoricus*[ss]	1586G	-	17-Aug-13	37.28922	-32.27561	-1689	13.8[#]	-
BioBaz 2013	Vent	LS-ET2	*B. azoricus*[ss]	1586H	-	17-Aug-13	37.28922	-32.27561	-1689	12.7[#]	-
BioBaz 2013	Vent	LS-ET2	*B. azoricus*[ss]	1586I	-	17-Aug-13	37.28922	-32.27561	-1689	14.9[#]	-
BioBaz 2013	Vent	LS-ET2	*B. azoricus*[ss]	1586J	-	17-Aug-13	37.28922	-32.27561	-1689	14.3[#]	-
BioBaz 2013	Vent	LS-Ms	*B. azoricus*[ss]	1586K	1848 A	9-Aug-13	37.28806167	-32.27557	-1700	12.8#	17.7[&]
BioBaz 2013	Vent	LS-Ms	*B. azoricus*[ss]	1586L	1848 B	9-Aug-13	37.28806167	-32.27557	-1700	15.7#	11.8[&]
BioBaz 2013	Vent	LS-Ms	*B. azoricus*[ss]	1586 M	1848 C	9-Aug-13	37.28806167	-32.27557	-1700	13.2[#]	18.2[&]
BioBaz 2013	Vent	LS-Ms	*B. azoricus*[ss]	1586N	-	9-Aug-13	37.28806167	-32.27557	-1700	13.1[#]	-
BioBaz 2013	Vent	LS-Ms	*B. azoricus*[ss]	1586O	-	9-Aug-13	37.28806167	-32.27557	-1700	12.3[#]	-
M82-3	Vent	MG	*B. azoricus*[ss]	1586P	1848 J	13-Sep-10	37.8444492	-31.51890	-836	24.2[#]	15.3[&]
M82-3	Vent	MG	*B. azoricus*[ss]	1586Q	1848 K	13-Sep-10	37.8444492	-31.51890	-836	12.2[#]	17.5[&]
M82-3	Vent	MG	*B. azoricus*[ss]	1586T	1848 L	13-Sep-10	37.8444492	-31.51890	-836	16.2[#]	17.5[&]
M82-3	Vent	MG	*B. azoricus*[ss]	1586R	-	13-Sep-10	37.8444492	-31.51890	-836	13.6[#]	-
M82-3	Vent	MG	*B. azoricus*[ss]	1586S	-	13-Sep-10	37.8444492	-31.51890	-836	11.9[#]	-
NT10-08[**]	Vent	Myojin Knoll	*B. septemdierum*[$]	BSEPE	-	14-May-10	32°6.234′N	139°52.152′E	-1228	-	-
BioBaz 2013	Vent	Rb	*B. azoricus*[ss]	1600G	1848 G	13-Aug-13	36.22942667	-33.90196	-2273	11.2[#]	17.9[&]
BioBaz 2013	Vent	Rb	*B. azoricus*[ss]	1600H	1848 H	13-Aug-13	36.22942667	-33.90196	-2273	11.0[#]	18.0[&]
BioBaz 2013	Vent	Rb	*B. azoricus*[ss]	1600I	1848 I	13-Aug-13	36.22942667	-33.90196	-2273	12.5[#]	13.2[&]
BioBaz 2013	Vent	Rb	*B. azoricus*[ss]	1600F	-	13-Aug-13	36.22942667	-33.90196	-2273	13.6[#]	-

Cruise	Ecosystem	Site	Species	DNA lib	RNA lib	Sampling date	Latitude	Longitude	Depth (m)	DNA seq	RNA seq
BioBaz 2013	Vent	Rb	*B. azoricus*[$$]	1600J	-	13-Aug-13	36.22942667	-33.90196	-2273	10.4[#]	-
Odemar	Vent	Sem	*Bput*[$$]	SemA-M2-5	1844 J	22-Nov-14	13.513467	-44.963133	-2432	7.3[&]	13.4[&]
Odemar	Vent	Sem	*Bput*[$$]	SemB-M3-12	1844 K	22-Nov-14	13.513467	-44.963133	-2432	7.8[&]	13.4[&]
Odemar	Vent	Sem	*Bput*[$$]	SemC-M4-16	1844 L	22-Nov-14	13.513467	-44.963133	-2432	9.5[&]	13.4[&]

MG = Menez Gwen (White Flames); LS-ET = Lucky Strike-Eiffel Tower; LS-Ms = Lucky Strike-Montsegur; Rb = Rainbow; Log = Logatchev; Sem = Semenov-2, East Flank, chimney "Ash Lighthouse"; Chapopote-Gulf of Mexico = Ch-GoM; GC600-Gulf of Mexico; East Pacific Rise = EPR.

[$]Symbiosis with only SOX symbiont.

[$$]Symbiosis with SOX and MOX symbionts.

[$$$]Symbiosis with up to six phylotypes of bacteria, including two SOX, two MOX, and a methylotroph symbiont.

[#]Sequenced as 2 X 100 bp paired-end reads.

[&]Sequenced as 2 X 150 bp paired-end reads.

[&&]Sequenced as 2 X 250 bp paired-end reads.

[*]Sequencing described in Sayavedra *et al*. *in prep* - Chapter 3.

[**]Sequencing and assembly of SOX symbiont from one individual by Ikuta *et al*. (2015).

[***]Sequencing of five individuals will be described by Ansorge *et al*. (*in prep*).

[+]Sequencing will be described by A. Assié *et al*. (*in prep*).

Supplementary Table 2. Metrics of SOX symbiont genome assemblies per individual. Completeness was estimated with CheckM using the p_Proteobacteria (UID3880) gene marker set (Parks *et al.*, 2015). Assembly statistics were estimated with Quast (Gurevich *et al.*, 2013). Bput = *B. puteoserpentis*; Baz = *B. azoricus*; Bther = *B. thermophilus*

Site	Species	DNA library	Completeness	Contamination	# contigs	Largest contig	Total length	N50	# N's per 100 kbp
Log-Ir	Bput[$$]	Log	97.7	2.63	81	317,428	2.19	123,055	8389.87
Crab Spa-EPR	Bther[$]	1600S	97.2	1.32	365	347,385	2.78	81,021	3.41
Crab Spa-EPR	Bther[$]	1600R	97.86	1.32	253	144,199	2.39	41,860	0
Crab Spa-EPR	Bther[$]	1600T	97.86	4.41	273	126,494	2.49	36,571	4.01
Sem	Bput[$$]	bputeoB	98.36	3.29	248	113,340	2.32	28,100	11.42
Sem	Bput[$$]	bputeoA	97.7	4.61	320	108,651	2.44	27,693	4.1
Sem	Bput[$$]	bputeoC	98.36	2.14	232	101,558	2.21	26,002	0
Rb	Baz[$$]	1600F	97.7	0.99	355	62,242	2.14	12,345	0
Rb	Baz[$$]	1600J	98.36	0.99	469	36,168	2.38	11,912	53.09
Rb	Baz[$$]	1600I	98.36	4.82	597	48,528	2.51	9,020	74.03
MG	Baz[$$]	1586P	98.36	1.97	563	50,747	2.44	8,646	10.33
MG	Baz[$$]	1586S	98.36	0.66	601	51,470	2.48	8,532	21.36
MG	Baz[$$]	1586Q	98.36	1.32	599	47,708	2.50	8,341	55
MG	Baz[$$]	1586T	98.36	1.32	661	47,928	2.55	8,264	73.31
LS-Ms	Baz[$$]	1586O	98.36	4.56	636	51,549	2.78	7,951	47.16
Rb	Baz[$$]	1600G	98.36	1.37	551	55,000	2.42	7,869	3.35
MG	Baz[$$]	1586R	98.36	1.32	643	51,128	2.52	7,624	26.05
LS-ET2	Baz[$$]	1586J	98.36	3.29	606	37,186	2.75	7,553	46.38
LS-ET	Baz[$$]	1586D	98.36	3.62	638	39,746	2.80	7,426	29.15
LS-ET	Baz[$$]	1586E	98.36	3.62	670	48,453	2.86	7,362	46.4
LS-Ms	Baz[$$]	1586N	98.36	3.62	723	35,307	2.83	6,746	27.57
LS-ET2	Baz[$$]	1586H	98.36	2.96	804	48,914	2.90	6,481	44.48
LS-ET2	Baz[$$]	1586F	98.36	3.29	844	31,728	2.95	6,380	35.18
LS-ET2	Baz[$$]	1586G	98.36	2.3	864	32,506	3.09	6,219	78.18
LS-Ms	Baz[$$]	1586L	98.36	1.97	874	31,405	2.49	5,666	32.98
Rb	Baz[$$]	1600H	97.65	0.66	791	26,700	2.54	5,475	132.68

Site	Species	DNA library	Completeness	Contamination	# contigs	Largest contig	Total length	N50	# N's per 100 kbp
LS-Ms	Baz[$$]	1586K	98.36	3.62	1,092	31,659	3.25	5,237	11.04
LS-ET	Baz[$$]	1586C	98.36	4.28	1,169	31,704	3.15	4,923	39.2
LS-ET2	Baz[$$]	1586I	98.36	2.52	1,578	31,676	3.67	3,493	14.67
LS-ET	Baz[$$]	1586A	98.36	6.25	2,211	31,541	4.11	1,922	39.35
LS-Ms	Baz[$$]	1586M	97.37	13.14	2,591	24,058	4.52	1,783	103.99
LS-ET	Baz[$$]	1586B	98.36	6.3	2,752	31,113	4.87	1,762	65.06

[$]Symbiosis with only SOX symbiont
[$$]Symbiosis with SOX and MOX symbionts

Supplementary Table 3. PCR primers and annealing temperatures used for the amplification of the MDH gene cluster and the probe synthesis.

Primer	Target gene	Forward (5'-3')	Reverse (5'-3')	Product size	Temp (°C)
MDH_cluster	MDH gene cluster	TACGCCCCTTTCCAAGTGAC	ATTTTGGTGGCTGTCGTTGC	8334	64.6
PP2	Conserved protein	GCTGCGAGGAAAATCCAGTAC	GCCCTAAATCAACACCACGC	299	55
PP3	S-(hydroxymethyl) glutathione dehydrogenase	GCGGTTGGGAAGCAACTTCT	CCTTCGTGACCTAACACACA	299	55
PP4		TGTGGTCCAGGCGTTAAAGA	ACCACAACCCAGCAAGCATA	299	55
PP5		CGGGTCAACGGTTGCAGTAT	ACTGTAGAAACACCCCAGCC	299	55
PP6	S-formylglutathione hydrolase	GGGCGTACAGAGTTACCAGG	TTCTGTGCTTGAGGTGGTAGG	299	55
PP7		ATGTTAGAACCAAGGCAGGCA	GGCACAAATCAACGCACCAT	298	55
PP8		CTCCGATTTGCAATCCTACACA	CCATATGTTCACCTATAAACGATGCA	299	55
MDHPP1	*xoxF*	TTGTCCAGCAGCGTTAGGTT	ATAGTGTCTTCACCGCCTGC	268	55
MDHPP2		GGTTTTGCCTTGCTACTATTGTTG	AAAGCACCACTCCAAACGGA	257	55
MDHPP3		GGGAATATAAACCAATACAAGACCCG	TGTTAGGGAATGGCGTGTGG	260	55
MDHPP4		ACGGTTACTCACAAGGTCCAG	ACCATACTCACCGCCAGAAAT	268	55
MDHPP5		ATCAAATGACCCCGCACGAT	TCGTGCCCAAATCGTCATAGA	270	55

Chapter VI:

Conclusions, preliminary results, and future directions

This thesis contributes substantially to several topics within microbiology and symbiosis research by i) describing highly abundant toxin-related genes (TRGs) that are unique to the horizontally transmitted SOX symbionts and proposing that the symbiont uses these genes for host-symbiont interactions (Chapter II), ii) investigating the evolutionary history of the TRGs and showing that some TRGs are present in only one of the two lineages of SOX symbionts, probably acquired by horizontal gene transfer (Chapter III), iii) revealing which genes are more expressed during the active colonization of the host by the SOX symbiont (Chapter IV), iv) describing how the potential to use methanol, a new energy source for the SOX symbiont, is present only in some strains with the same 16S rRNA (Chapter V), and v) revealing that the metabolic flexibility of the SOX symbiont might be influenced by the contrasting vent conditions (Chapter V). In the following sections, the results of this thesis will be briefly summarized and some preliminary results will be presented as a basis for future research studies.

6.1. The role of toxin-related genes (TRGs) in the *Bathymodiolus* symbiosis

Before I began this thesis, virtually nothing was known about how a horizontally transmitted chemosynthetic symbiont could interact with its host. By using comparative genomics of free-living bacteria against vertically and horizontally transmitted symbionts, I found an abundant number of TRGs in the genomes of beneficial SOX symbionts of two *Bathymodiolus* species. The TRGs belonged to three main classes: YD repeats, RTX and MARTX. We hypothesized that the role of these TRGs were: i) for direct host-symbiont interactions, to recognize, infect or escape the host immune defense; or ii) for indirect host-symbiont interactions, to protect the host against pathogens or eukaryotic parasites.

These two possible functions – interacting with the host, and defending against external pathogens – are expected to have different genetic signatures. According to the

"honest signaling" model of host-microbe communication (Dawkins and Guilford, 1991; Hillman and Goodrich-Blair, 2016), microbes might express signals to the host in order to display their willingness to cooperate, even if it initially causes some cost to the microbe (Hillman and Goodrich-Blair, 2016). Toxin-related genes may serve as such "honest signals" that the host can recognize, which would imply that the host had to evolve mechanisms to recognize the polymorphic TRGs. In this case there would be selective pressure for slower mutational rates. However, if TRGs are involved in a defensive role, having antagonistic interactions with external pathogens and parasites, we would expect them to have accelerated mutation rates (Oliver *et al*., 2014). What we actually observe is that some of the TRGs of the *Bathymodiolus* symbionts have a high number of single nucleotide polymorphisms, while others are more conserved (Chapter II). Thus those which are conserved might be playing a role in "honest signaling" to interact with the host, while the highly variable ones might be playing a role in host defense.

Proving that the TRGs have any of these roles will be challenging. In model organisms that are well-established in the laboratory, the function of a gene is usually demonstrated by targeted knockout mutations. To show that TRGs of the symbiont are used for direct host-symbiont interactions, we would have to knock out different TRGs and show that the SOX bacterium is no longer able to colonize or maintain the symbiosis (Lipsett, 1987). Then, it would also be necessary to reintroduce the TRGs and show that the lost function is restored. To confirm that the SOX symbiont can defend its host from pathogens or parasites, the first step would be to identify which parasites and pathogens are present in the vent and seep ecosystems. Then, we would need to prove that a host with no symbionts, or at least a very reduced number of symbionts, has a reduced fitness when exposed to the parasite/pathogen compared to a host with a "normal" symbiont population (Lopanik, 2014). Such experiments are still not feasible, as the primary requirements are to have the host living under optimal conditions in aquaria and obtaining a pure culture of the symbiont. Most symbiont cultivation attempts have been, however, unsuccessful in the long term (e.g. A. Kreutzmann and CP Antony, Symbiosis Department), and *Bathymodiolus* mussels tend to lose their symbionts when kept in aquaria (Kadar *et al*., 2005).

Even though we cannot yet directly prove the role of the TRGs, indirect culture-independent techniques can help us in understanding their role in symbiosis. For example, if the 'free-living' stage of the SOX symbiont is found in the proximity of mussel beds, transcriptome sequencing could reveal whether the TRGs are being expressed or not

outside the host. We would expect that if TRGs were used for host-symbiont interactions, then they would not be expressed in the free-living stage of the SOX symbiont. A second approach that could help us understanding their role is the combination of matrix-assisted laser desorption ionization as a mass spectrometry imaging technique (MALDI-IMS) and proteomics. This technique is currently under development for *Bathymodiolus* mussels and will allow us to see the protein and metabolite distribution across the gills and other soft tissues (B. Geier and M. Liebeke, Symbiosis Department). Screening for parasites in a significant number of individuals might reveal that a specific metabolite or protein is expressed in the tissue that is in the proximity of a mussel parasite. A third approach would be to express different TRG proteins heterologously, purify them, and test their activity. Purified protein extracts could then be tested for antimicrobial, hemolytic, and antagonistic activity against different eukaryotes.

An alternative role that has been suggested for some TRGs of the YD repeat class is bacterial competition. Some characterized YD repeat proteins have been shown to have antagonistic bacterial interactions, although they would normally have an immunity gene in the neighborhood that would protect them (Koskiniemi *et al*., 2013; Sayavedra *et al*., 2015). An *Escherichia coli* clone heterologously expressing candidate proteins could be used to test for antagonistic bacterial interactions with the cultures of the close free-living relatives of the SOX symbiont i.e., '*Candidatus* Thioglobus autotrophica' and '*Ca*. T. singularis'. This experiment could potentially show whether SOX symbionts use a TRG to achieve its high specificity to the host, even when very close relatives are present in the seawater. In fact, during the time of my PhD, I successfully created an *E. coli* expression clone of one TRG of the YD repeat class. The recombinant protein could be potentially used for the aforementioned experiment.

6.2. Convergent acquisition of symbiotic factors and the role of horizontal gene transfer in the acquisition of TRGs

By sequencing genomes from the two distinct lineages of SOX symbionts, I was able to pinpoint when the YD-repeat and RTX classes were acquired during the evolution of the SOX symbiont by horizontal gene transfer (Chapter III). These two classes are present in only one of the two lineages, so we concluded that they are not essential to maintain the symbiosis with bathymodioline mussels, because we would otherwise see them in all symbiont genomes. The MARTX class of TRGs, on the other hand, is always

present in SOX symbionts, although their domain structure is very flexible. Thus, MARTX could be among the genes that help in promoting the highly specific symbiosis. However, this result cannot rule out that the two lineages have converged on different mechanisms for host-symbiont interaction, e.g. a different symbiotic factor like the type IV secretion system found in the SOX symbiont genome without MARTX.

Sequencing additional SOX symbionts from other mytilid mussel hosts might help us to test our hypotheses for which genes are necessary or not for host interaction. We have restricted our sequencing efforts to vent and seep bathymodioline mussels, mainly due to the lack of available sample material from other types of mussels (Chapter II, III, IV, and V). The 16S rRNA phylogeny of the SOX symbiont shows potentially interesting genome comparisons for future studies (Chapter II – Figure 2-figure supplement 1). For example, the symbionts of *Adipicola pacifica*, *A. longissima* and MOTU 16, which are found in organic falls, are very closely related to the clade of the SOX symbionts of *B. azoricus* (Duperron *et al*., 2008; Fujiwara *et al*., 2010; Lorion *et al*., 2009). In contrast to their intracellular counterparts, these symbionts are extracellular, showing that different morphological adaptations to the host have evolved in mytilid mussels. By sequencing the symbionts of these mytilid mussels, we could have the ideal case scenario to study the evolutionary transition from free-living bacteria to extracellular symbiosis, to intracellular symbiosis. This kind of comparison is relevant so as to understand the evolution of the symbiont lineage with their host.

Moving beyond the SOX symbionts of mytilid mussels, we are now in a position to begin exploring whether chemosynthetic symbionts from other groups of animal hosts, such as nematodes, sponges, and tube worms, also share similar mechanisms for host-microbe interaction. Such analyses could reveal mechanisms acquired through convergent evolution or horizontal gene transfer. For example, the highly specific ectosymbiont *Ca*. Thiosymbion ectosymbiont of the *Leptonemella* nematode also encodes genes from the YD and MARTX classes (J. Zimmermann, personal communication). The transmission mode of this symbiosis has not been fully resolved, but in this symbiosis, each host species maintains a specific symbiont phylotype (Zimmermann *et al*., 2016). The presence of the same classes of TRGs might contribute to the highly specific association between the host and the symbiont. The symbiont genomes might have acquired through convergent evolution molecular mechanisms similar to that of the *Bathymodiolus* symbionts. Future studies should identify how widespread the TRGs are in highly specific

symbioses. Patterns might start to emerge for symbionts that have a direct contact with the environment in, at least, one stage of their life cycle.

6.3. Understanding the symbiont colonization process

6.3.1. The origin of the symbiont population

The SOX symbiont

The colonization process of most horizontally transmitted symbionts in marine invertebrates is poorly understood, with the exception of the *Vibrio*-squid and *Ca*. Endoriftia persephone-*Riftia pachyptila* symbioses (Chun *et al*., 2008; McFall-Ngai *et al*., 2012; Nyholm and McFall-Ngai, 2004). One of the aims of this thesis was to understand the symbiont colonization process of *Bathymodiolus* mussels. Wentrup *et al.* (2014) showed that the newly formed gill filaments of the budding zone are colonized throughout the mussel's lifetime. However, it was not clear whether most of the symbiont population came from older gill filaments or from the environment.

Based on our single nucleotide polymorphism (SNP) analyses, most of the SOX symbiont population is likely coming from the environment (Chapter IV). We still do not know, however, whether i) the SOX symbionts colonize only newly-developing gill filaments, so that the populations are stable after establishment, or ii) older gill filaments are still susceptible to colonization, and patches of "new" symbionts can occur among established populations. Furthermore, if the free-living SOX bacteria in the surrounding water are composed of different strains, and the ambient strain composition changes over time, we may be able to track the history of the SOX at a given site by identifying strain variants in filaments of different ages, in analogy to the growth rings of a tree trunk. Thus, future studies should look at how the SOX symbiont population varies along the gill.

The MOX symbiont

The MOX symbiont is believed to be horizontally transmitted, because the MOX symbiont phylogeny is incongruent with the phylogeny of the host mussels, but so far no further evidence is available to support this mode of transmission (Petersen and Dubilier, 2009). SNP analysis (Chapter IV) shows that the MOX symbiont population in the budding zone is not genetically distinct from older gill filaments, unlike the SOX where there is more symbiont variability in the budding zone. If MOX transmission is

horizontal, like the SOX, how can we explain this difference? The MOX symbiont may preferentially colonize the host at an early ontogenetic stage, and subsequently take up symbionts from the older filaments by self-infection. If this is the case, the question of "at which stage the initial MOX symbiont population colonizes the mussel?" remains open.

Comparative genomic analyses suggest that the MOX symbiont genomes are smaller, have a reduced GC-content, and lack key genes for osmoregulation that are present in the available genomes of free-living marine methanotrophs (Antony CP, personal communication, April 2016). If the MOX symbiont cannot survive outside the host, a vertical transmission mode might actually be occurring in *Bathymodiolus* mussels. This would explain the poorly differentiated MOX symbiont population between the two gill regions. To have further evidence for the transmission mode of the MOX symbiont, future studies should inspect whether very young larvae already host MOX symbionts.

6.3.2. Is the SOX symbiont response similar in other *Bathymodiolus* host species?

The SOX symbiont has a distinct expression profile in the budding zone compared to the established tissue in *B. azoricus* mussels of the Lucky Strike vent field (Chapter IV). The expression profile of the symbiont showed that they overexpressed genes for the central carbon metabolism and the use of energy sources in the budding zone. However, we now know that the genomic potential of the SOX symbiont population between vents can hugely differ, and in some vents the population can be more heterogeneous than others (R. Ansorge *et al*., in prep., and Chapter V). This raises the following questions: Does the SOX symbiont express a similar set of genes to recognize and maintain the symbiosis under different environmental conditions? Is the SOX symbiont response similar when a larger pool of strains from the environment can colonize the host? Is the competition between different strains sets triggering a different host-symbiont response?

Based on the poor quality of the assembly metrics of SOX symbionts from the Lucky Strike vent filed, the SOX symbiont genomes of this site are one of the most heterogeneous (R. Ansorge, in prep., and Chapter V). With so many different SOX strains present, the emergence of cheaters that benefit from the association without contributing themselves is prone to happen (Hibbing *et al*., 2010). Thus, comparing the expression profile of *B. azoricus* mussels to a species that hosts less strain variation might reveal i) the molecular interaction mechanisms conserved for host colonization, and ii) how the

symbiont populations behave under the presence of a different collection of SOX strains with possibly different metabolic potential. A good candidate species for such comparison is *B. thermophilus* from the East Pacific Rise because it only hosts the SOX symbiont and the strain heterogeneity must be lower (based on the N50 metric of the genome assemblies) (Chapter III and V).

6.3.3. Do SOX symbionts have a slower metabolism outside the host?

The differential expression analysis revealed that the SOX symbiont has an expression profile reminiscent of the lag-phase of a bacterial culture in nascent gill filaments (Chapter IV). This suggests that the symbiont has a slower metabolism outside the host, which is expected, as host commodities provided to the symbiont, such as a more stable niche, and metabolites produced by the host are not present. Indeed, the SOX symbiont might depend on the host's oxaloacetate to complete the TCA cycle, as it does not encode the enzyme malate dehydrogenase (Sayavedra *et al*., 2015, Ponnudurai *et al*., 2016). The degree of the symbiont dependence on oxaloacetate from the host needs to be however verified. Oxaloacetate could theoretically also be generated with the enzyme aspartate transaminase if there is enough aspartate to drive the enzyme in reverse. Ponnudurai *et al.* (2016) argued that aspartate is heavily used for the production of lysine, threonine, and methionine based on proteomic data. However, the recently sequenced free-living relative '*Ca*. T. autotrophica' does not encode the enzyme and yet is able to live outside the host (Figure 6-1). One possibility is that '*Ca*. T. autotrophica' depends on the oxaloacetate present in the medium culture. Future studies could investigate whether this bacterium takes up ^{13}C labeled oxaloacetate from the medium for further incorporation into amino acids and other metabolites. If the free-living close relative indeed takes the label oxaloacetate, a similar mechanism for the uptake might be present in *Bathymodiolus* symbionts.

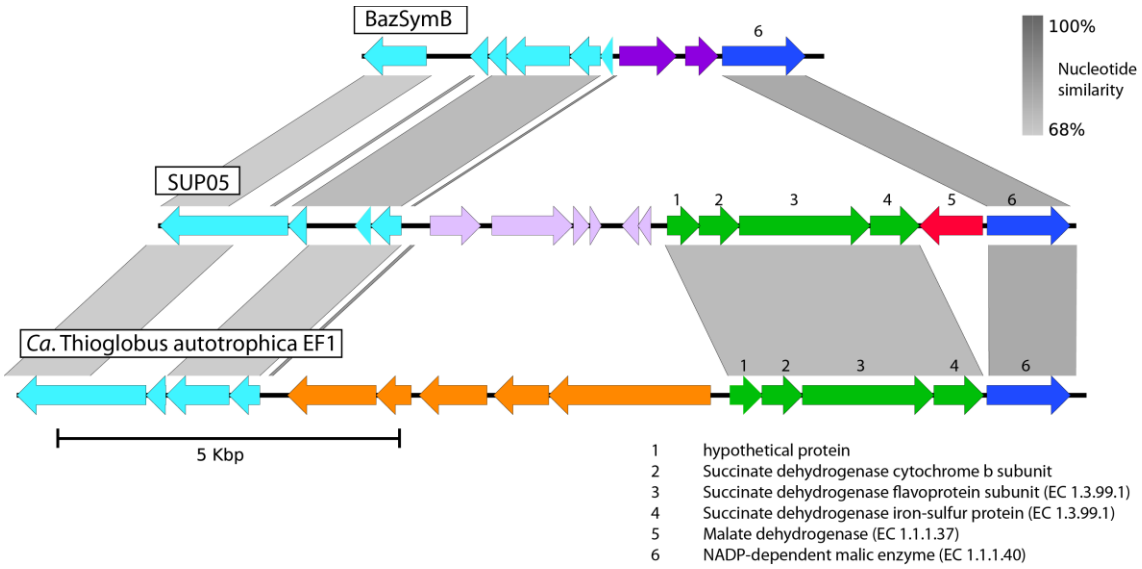


Figure 6-1. Multiple gene alignment of the SOX symbiont of *B. azoricus* from Menez Gwen (BazSymB), the free-living SUP05 obtained from Saanich Inlet (Walsh *et al*., 2009), and the now-cultivated "Ca. T. autotrophica EF1" (Shah and Morris, 2015). Genes that are similar between genomes have the same color code. The enzyme malate dehydrogenase is missing from Ca. T. autotrophica and BazSymB. The genome comparison was done with Mauve and Easyfig (Rissman *et al*., 2009; Sullivan *et al*., 2011). The analysis and the figure were prepared for Ponnudurai *et al*. (2016).

6.4. Lack of *ftsZ* in *Bathymodiolus* sulfur-oxidizing symbionts and their close relatives

One of the basic question in the symbiosis establishment of an environmentally acquired symbiont is: When and what triggers cells division? This is essential to understand the symbiont population behavior when is actively colonizing the gills compared to when the symbiont population is established (Chapter IV). A common gene marker for cell division in transcriptomes and proteomes is *ftsZ* (Sharma *et al*., 2006). *ftzZ* encodes a tubulin homolog that is involved in bacterial cell division, and is widely distributed and conserved in bacterial genomes (Rothfield *et al*., 1999). However, the *Bathymodiolus* SOX symbiont genome does not encode *ftsZ* (Chapter IV), even though it is not reduced as compared to free-living relatives (Chapter II-V). Conserved "essential genes" like *ftsZ* have been lost by genome erosion in many vertically-transmitted obligate symbionts and in intracellular pathogens (Chapter I-Box2) (McCutcheon and Moran, 2012; Moran, 2002), such as the insect endosymbionts '*Ca*. Sulcia muelleri str. GWSS', '*Ca*. Zinderia insecticola', and '*Ca*. Carsonella ruddii'; the intracellular pathogen *Chlamydia;* the pathogens *Ureplasma urealiticum* and *Mycoplasma mobile;* and the clam

endosymbionts '*Ca*. Ruthia magnifica' and '*Ca*. Vesicomyosocius okutanii' (reviewed by Erickson and Osawa, 2010).

Given that SOX symbiont genomes do not appear to have undergone genome reduction, how can the lack of *ftsZ* be explained? It has been hypothesized that the host's genome could have integrated the genes required for the division of the intracellular symbionts (Bernander and Ettema, 2010). If this were so, the lack of *ftsZ* would be correlated with the symbiotic life-style. However, when I analyzed all the available genomes of free-living close relatives of *Bathymodiolus* SOX symbionts, namely SUP05 from Saanich Inlet, '*Ca.* T. autotrophica', and '*Ca.* T. singularis', none of them turned out to encode *ftsZ* (Marshall and Morris, 2015; Shah and Morris, 2015; Walsh *et al*., 2009). Thus, a new cell division mode independent of FtsZ might have evolved in this group of Gammaproteobacteria.

Three fundamental cell division systems have been recognized so far: i) FtsZ-dependent division, which is the most widespread in Eubacteria and the phylum Euryarchaeota; ii) actin-myosin mechanism, that is widespread in eukaryotes, with the exception of mitochondria and chloroplast organelles that kept an FtsZ-dependent machinery derived from their prokaryotic ancestors (Bernander and Ettema, 2010); and iii) Cdv-dependent division, which was recently found in the phylum Crenarchaeota, and constricts the cells in the middle like FtsZ (Lindås *et al*., 2008). Apart from these three recognized division systems, different mechanisms have been proposed for organisms that lack *ftsZ*. For example, mitochondrial genomes have lost the *ftsZ* mode of division several times during evolution and rely mainly on dynamin-related proteins for cell division (Cerveny *et al*., 2007); Planctomycetes use budding reproduction and might depend on a peptidoglycan ring for division (Bernander and Ettema, 2010; Fuerst, 2005); *Mycoplasma* uses the force generated by its cell-motility machinery to divide (Lluch-Senar *et al*., 2010); *Chlamydia* uses MreB as a central coordinator at the division site to compensate the lack of *ftsZ* (Ouellette *et al*., 2012). The mode of division for most intracellular symbionts and pathogens with reduced genomes is however not well understood.

One of the main difficulties in characterizing the cell division modes of those intracellular symbionts that lack *ftsZ* is that there are no cultures available. Thus, the availability of the cultures of '*Ca*. T. autotrophica EF1' and '*Ca*. T. singularis PS1' opens new possibilities for investigating the cell division mode of the Gammaproteobacteria related to the *Bathymodiolus* symbiont. In collaboration with N. Leisch and S. Bulgheresi,

we plan to first use antibodies against MreB and FtsA to characterize the localization of these proteins in different cell division stages. In most bacteria, FtsA recruits the FtsZ polymers to the membrane (Loose and Mitchison, 2014). As this protein is encoded in all the close relatives of the symbionts, we might get an insight of the alternative mechanism of cell division by knowing which proteins bind FtsA. This could be achieved by creating a recombinant protein of FtsA with a polyhistidine-tag. Because histidine binds to several types of metals like nickel, the His tag is commonly used for purification and protein binding assays. Thus, FtsA could be purified and the proteins that bind it could be characterized by using mass spectrometry. Once we understand the cell division mechanism of this clade of symbionts and free-living bacteria, we might get further insights on how the symbiont population behaves during host colonization.

6.5. Symbiont strain heterogeneity and the use of alternative electron donors and acceptors

As more sequencing data become available, it is increasingly clearer that the SOX symbiont population in a single mussel host animal can be quite genetically heterogeneous. These may thus represent multiple strains despite having the same 16S rRNA gene sequence, which has been the standard marker for phylogenetic identity. Strain variation can be at the level of single-nucleotide polymorphisms, or even the presence/absence of entire genes or operons.

In Chapter V, I describe how only some SOX symbiont strains have acquired the potential to use methanol as energy source. The SOX symbiont's metabolic flexibility in using methanol as a carbon and energy source might help *Bathymodiolus* mussels to colonize environmental gradients with high methane and methanol concentrations, therefore improving the host fitness. When methanol is not available, the use of other energy sources might be playing an important role. Indeed, the metabolic flexibility is not only restricted to the use of methanol. My preliminary results also show metabolic flexibility in the use of hydrogen and thiosulfate as energy sources.

6.5.1. Hydrogen

As described in Chapter V, we sequenced metagenomes of vents along the northern Mid-Atlantic Ridge in vents with contrasting environmental conditions where either *B. azoricus* or *B. puteoserpentis* live: Menez Gwen (MG) (n=5), Lucky Strike

(Montsegur (LS-Ms) (n=5) and Eiffel Tower (LS-ET) (n=10)), Rainbow (Rb) (n=5), Semenov (Sem) (n=3), and Logatchev – Irina II (Log-Ir) (n=1) (Chapter V-Figure 1). The gene coverage variation of the hydrogenase operon against the single copy marker gene gyrase A (*gyrA*) was used as a proxy to the proportion of the SOX strains that could make use of hydrogen, as described in Chapter V for the MDH gene cluster. The venting sites For the key gene *hupL* alone, expression was highest in the SOX symbionts at Sem (34.6±44.46 × *gyrA*), followed by Log-Qt (9.52±8.67), Rb (2.78±0.65), LS-Ms (2.38±0.9), Log-Ir (1.73±2.41), MG (0.59±0.26), and LS-ET (0.1±0.09) (Figure 6-2). For the hydrogenase operon as a whole, the highest ratio of the hydrogenase operon to *gyrA* was found at Log-Ir (0.78±0.098) followed by Sem (0.55±0.259), Rb (0.47±0.224), MG (0.38±0.16), LS-MS (0.22±0.103), and LS-ET (0.11±0.067) (Figure 6-2). As for the MDH, the coverage ratio for the hydrogenase operon was significantly different between sites (p-value < 2.2e-16, Kruskal-Wallis test). Teasing apart which of the pairwise comparison between vent sites are different revealed that there was a significant difference in the ratio of the hydrogenase between all venting sites with *B. azoricus* mussels (p-values with Dunn's test are shown in Table 1). These results suggest that i) as for the use of methanol, only some strains within a population of symbionts with identical 16S rRNA have the potential to use hydrogen; and ii) in some vents, the strains that use hydrogen dominate the population more than others (Figure 6-2).

Table 6-1. Comparison of the mean gene coverage ratio of the 19 genes of the hydrogenase operon against *gyrA*. P-values obtained with Dunn's test were corrected for multiple testing with the Bonferroni correction. Bput, *B. puteoserpentis*; Baz, *B. azoricus*.

Host species	Site	Hydrogenase operon coverage/*gyrA*	P-values				
			Sem	RB	LS-ET	LS-Ms	MG
Bput	**Log-Ir**	0.78±0.098	0.195	0.001*	< 2e-16*	4e-11*	3e-5*
Bput	**Sem**	0.55±0.259	-	0.261	< 2e-16*	1.2e-9*	0.205
Baz	**RB**	0.47±0.224	-	-	< 2e-16*	8.6e-9*	0.130
Baz	**LS-ET**	0.11±0.067	-	-	-	< 2e-16*	2e-16*
Baz	**LS-Ms**	0.22±0.103	-	-	-	-	2.4e-5*
Baz	**MG**	0.38±0.16	-	-	-	-	

We hypothesize that the expression levels could be correlated with hydrogen concentrations at the different sites. Unfortunately, the environmental data from Sem is not available yet, although this site has the highest expression profile for the use of

hydrogen. The sites with the next highest expression of *hupL*, Log and Rb, are sites that host ultramafic rocks, which result in high methane and hydrogen concentrations (Chapter V – Table 4) (Charlou *et al*., 2002; Douville *et al*., 2002; Marbler *et al*., 2010; Zielinski *et al*., 2011). Even though these two sites have high hydrogen concentrations, the differences in expression are probably influenced by the contrasting proportion of strains that encode the hydrogenase in Log compared to Rb. LS and MG host basaltic rocks, which results in lower hydrogen concentrations (Charlou *et al*., 2000; Ondréas *et al*., 2009; Riou *et al*., 2010; Von Damm *et al*., 1998). Thus, the lower expression of the

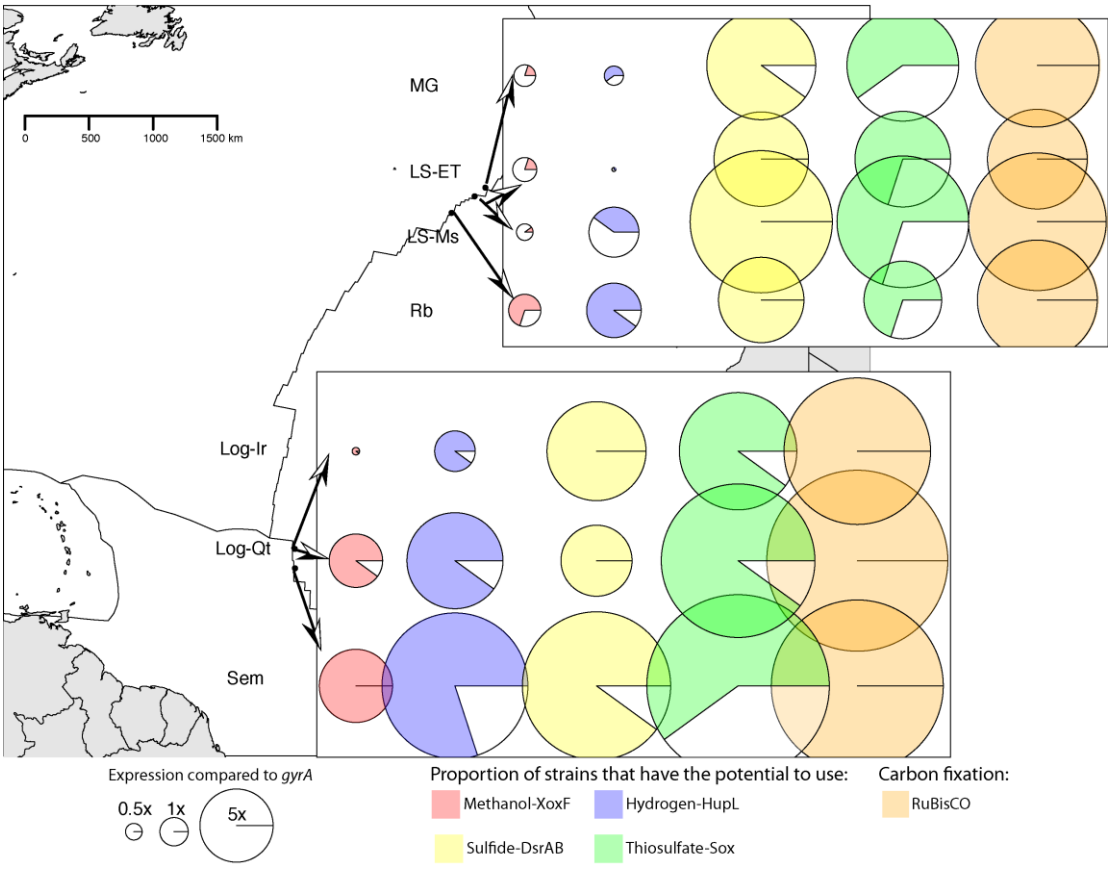


Figure 6-2. Comparison of the genes encoded and expressed for the use of different energy sources in SOX symbionts of *B. azoricus* (upper panel) and *B puteoserpentis* (lower panel). Pie chart sectors represent proportions of SOX strains with the genetic potential (estimated from metagenome) to use respective substrates (colors), whereas area of each entire disk is proportional to the expression level of respective key gene (estimated from metatranscriptome). Proportion of strains was calculated with the average ratio of the gene coverage between the target gene regions and *gyrA* plus one standard deviation. Gene expression of the key genes was normalized with FPKM, TMM and *gyrA* as described in Chapter V. This expression value was further normalized with the $\log_2$ scale. The following key genes were used: methanol, *xoxF* ; hydrogen, *hupL*; sulfide, *dsrAB* complex; thiosulfate *soxAZYX* complex; and for carbon fixation *cbbL* (as comparison). Locality abbreviations: MG, Menez Gwen; LS-ET, Lucky Strike-Eiffel Tower; LS-Ms, Lucky Strike-Montsegur; Rb, Rainbow; Log-Ir, Logatchev-Irina II; Log-Qt, Logatchev-Quest; amd Sem, Semenov.

hydrogenase relates to both the lack of the substrate present and the lower abundance of strains encoding the gene operon.

6.5.2. Reduced sulfur compounds

Bathymodiolus SOX symbionts encode the reverse *dsr*, *sox*, *apr* and *sat* systems for the oxidation of reduced sulfur compounds (Figure 6-3b). These pathways are similar to their close relatives *Ca*. R. magnifica and *Ca*. V. okutanii (Figure 6-3a) (Kuwahara *et al*., 2008; Newton *et al*., 2008, 2007). The reverse-*dsr* system is involved in sulfide oxidation, while the *sox* system is involved in thiosulfate oxidation. Thiosulfate and sulfide both stimulate carbon fixation in *Bathymodiolus* mussels, but they are consumed at different rates (Beinart *et al*., 2012; Nelson *et al*., 1995; Riou *et al*., 2010). Originally, it was thought that the difference in the consumption of reduced sulfur compounds would be due to differences in the absence or presence of genes encoding the respective pathways. Later, Kleiner *et al*. (2012a) showed that the pathways for sulfur oxidation are conserved among Gammaproteobacteria SOX symbionts belonging to different clades. The consumption rates, however, could be affected by strain heterogeneity in terms of the absence or presence of genes encoding the *dsr*, *sox*, *apr* and *sat* systems. Because the SOX symbiont displays high genomic heterogeneity with respect to the use of hydrogen and methanol, it is worth looking in detail at the genome coverage variation between sites. The following results are from preliminary analysis with respect to the use of sulfide and thiosulfate.

The coverage ratio of the *dsrAB*, *dsrMKLJOPN* genes were significantly lower in Sem (0.74±0.15) and MG (0.76±0.13) as compared to the other venting sites (0.83-0.92) (p-values are shown in Table 6-2). The expression of *dsr* for the use of sulfide was however not affected by the slight variation in the number of strains that encoded the *dsr* system (see bellow) (Figure 6-2). As a proxy for the use of thiosulfate, I considered the ratio of the genes *soxBAZYX* to *gyrA*. The only site with a ratio close to one was observed in Log-Ir (0.9±0.07). In contrast, all other venting sites had a significantly lower ratio (ratio = 0.57-0.64, p-values of pairwise comparison between sites are shown in Table 6-3). The deviation in coverage of the *sox* genes suggests that not all SOX symbionts can utilize thiosulfate.

As a proxy for the expression of the *dsr* and *sox* systems, I considered the expression of the genes *dsrAB* and *soxAZYX*, respectively. The highest expression of

dsrAB was in Sem (36.88±46.29x *gyrA*), followed by LS-Ms (28.98±33.45x *gyrA*), MG (14.66±10.83x *gyrA*), LS-ET (10.26±6.95x *gyrA*), Log-Ir (9.76±5.37x *gyrA*), Rb (6.98±2.6x *gyrA*), and Log-Qt (5.38±2.14x *gyrA*). The expression of *soxAZYX* was higher in Sem (86.22±97.49x *gyrA*), followed by Log-Qt (41.99±9.96x *gyrA*), LS-Ms (24.14±4.39x *gyrA*), Log-Ir (16.73±15.32x *gyrA*), MG (14.43±15.12x *gyrA*), LS-ET (9.43±2.05x *gyrA*), and Rb (5.7±1.4x *gyrA*). The ratio of *dsrAB* to *soxAZYX* can be used to form an approximate idea of whether sulfide or thiosulfate plays a more important role in the symbiosis (Kleiner *et al*., 2012b). This ratio suggests that sulfide and thiosulfate play a similar role in Ms and Rb (dsr:sox = 1.2), LS-Et (dsr:sox =1.09), and MG (dsr:sox = 1.02). In contrast, the expression of the sox system is higher compared to dsr in Log-Ir (dsr:sox = 0.58), Sem (dsr:sox = 0.43) and Log-Qt (dsr:sox = 0.13). Therefore i) symbionts of *B. azoricus* mussels may use sulfide and thiosulfate at similar rates (MG, LS, and Rb venting sites), and ii) symbionts of *B. puteoserpentis* likely use more thiosulfate than sulfide.

Table 6-2. Comparison of the gene coverage ratio of the *dsrAB* and *dsrMKLJOPN* genes against *gyrA*. P-values were obtained with Dunn's test using nine genes of the *dsr* system. P-values were corrected for multiple testing with the FDR correction.

	P-values - Dunn's Test for genes encoding *dsrAB* and *dsrMKLJOPN* /*gyrA*					
	MG	**LS-ET**	**LS-Ms**	**Rb**	**Sem**	**Log-Ir**
LS-ET	0.0000015*	-	-	-	-	-
LS-Ms	0.000089*	0.71195	-	-	-	-
Rb	0.00793*	0.07437	0.21494	-	-	-
Sem	0.71195	0.0000095*	0.00017*	0.00793*	-	-
Log-Ir	0.00402*	0.70266	0.62221	0.20281	0.00376*	-
	Gene coverage/*gyrA* coverage					
	MG	**ET**	**MS**	**RB**	**Sem**	**Log-Ir**
dsrA	0.91±0.05	0.96±0.15	1.06±0.11	0.95±0.08	0.88±0.05	0.99±NA
dsrB	0.79±0.07	0.91±0.17	0.9±0.05	0.87±0.08	0.78±0.05	0.91±NA
dsrM	0.77±0.1	0.98±0.09	1.02±0.12	0.98±0.13	0.87±0.08	1.03±NA
dsrK	0.81±0.03	0.9±0.06	0.94±0.08	0.9±0.05	0.78±0.03	0.87±NA
dsrL	0.84±0.04	0.96±0.07	0.88±0.03	0.85±0.05	0.79±0.06	0.9±NA
dsrJ	0.49±0.1	0.58±0.07	0.61±0.22	0.47±0.04	0.38±0.03	0.79±NA
dsrO	0.69±0.09	0.82±0.18	0.77±0.11	0.73±0.05	0.64±0.06	0.88±NA
dsrP	0.79±0.09	1±0.1	0.96±0.1	0.91±0.08	0.76±0.04	1±NA
dsrN	0.77±0.06	0.92±0.09	0.84±0.1	0.86±0.06	0.8±0.03	0.91±NA

Incubation experiments on different *Bathymodiolus* species have shown contrasting preferences for the use of sulfide and thiosulfate as electron donors. For example, *B. thermophilus* showed a greater preference for thiosulfate over sulfide (Nelson

et al., 1995). In contrast *B. brevior* did not show carbon incorporation in the presence of thiosulfate (Beinart *et al*., 2015). The differences in consumption rates for sulfide and thiosulfate might be relevant for those *Bathymodiolus* species, which have substantial strain heterogeneity among their SOX symbionts. Even if not all strains of the *Bathymodiolus* symbionts encode the *sox* system, these genes seem to have a major role in their overall activity (Figure 6-2). In *B. azoricus* mussels, the expression of the *dsr* and *sox* systems has a similar level of expression (Figure 6-2). This is in stark contrast to the chemosynthetic symbionts of *Olavius algarvensis* and *Riftia* that express the *dsr* system ~2-5 times more than the *sox* system (Kleiner *et al*., 2012b, Ponnudurai *et al*., 2016). It is surprising therefore, that although not all SOX symbiont strains seem to encode the *sox* sytem, this system is actively expressed in the population, at a level similar or higher than the *dsr*.

The use of thiosulfate as an energy source might confer an ecological benefit to the *Bathymodiolus* symbionts. *Bathymodiolus* mussels do not possess any special proteins to retain the sulfide present in the environment, which makes the use of sulfide dependent on the supply by the chemosynthetic environment (Nelson and Fisher, 1995). Sulfide can be abiotically oxidized spontaneously to elemental sulfur under oxic conditions, making the compound unavailable for the *Bathymodiolus* symbionts. In contrast, thiosulfate is a more stable compound that is therefore more abundant in certain environments (Beinart *et al*., 2015; Hensen *et al*., 2006; Zielinski *et al*., 2011). The metabolic flexibility in the use of sulfide and thiosulfate, as well as the aforementioned energy sources might help the host to respond to their local ephemeral environments.

Table 6-3. Comparison of the gene coverage ratio of the *soxB* and *soxAZYX* genes against *gyrA*. P-values were obtained with Dunn's test using four genes of the sox system. P-values were corrected for multiple testing with the FDR correction. NA, not applicable.

	P-values - Dunn's Test for genes encoding *soxB, soxAZYX*					
	MG	**LS-ET**	**LS-Ms**	**Rb**	**Sem**	**Log-Ir**
LS-ET	0.546	-	-	-	-	-
LS-Ms	0.388	0.762	-	-	-	-
Rb	0.689	0.912	0.762	-	-	-
Sem	0.546	0.841	0.912	0.841	-	-
Log-Ir	0.032*	0.046*	0.079	0.046*	0.079	-
	Gene coverage /*gyrA* coverage					
	MG	**ET**	**MS**	**RB**	**Sem**	**Log-Ir**
soxB	0.85±0.06	1.02±0.1	0.93±0.06	1.01±0.13	0.83±0.04	1.01±NA
soxA	0.63±0.06	0.59±0.09	0.73±0.07	0.65±0.19	0.72±0.05	0.85±NA
soxZ	0.39±0.12	0.32±0.12	0.45±0.06	0.37±0.16	0.39±0.05	0.87±NA
soxY	0.41±0.11	0.53±0.08	0.54±0.04	0.49±0.09	0.54±0.04	0.85±NA
soxX	0.6±0.09	0.74±0.15	0.67±0.11	0.66±0.09	0.76±0.04	0.95±NA

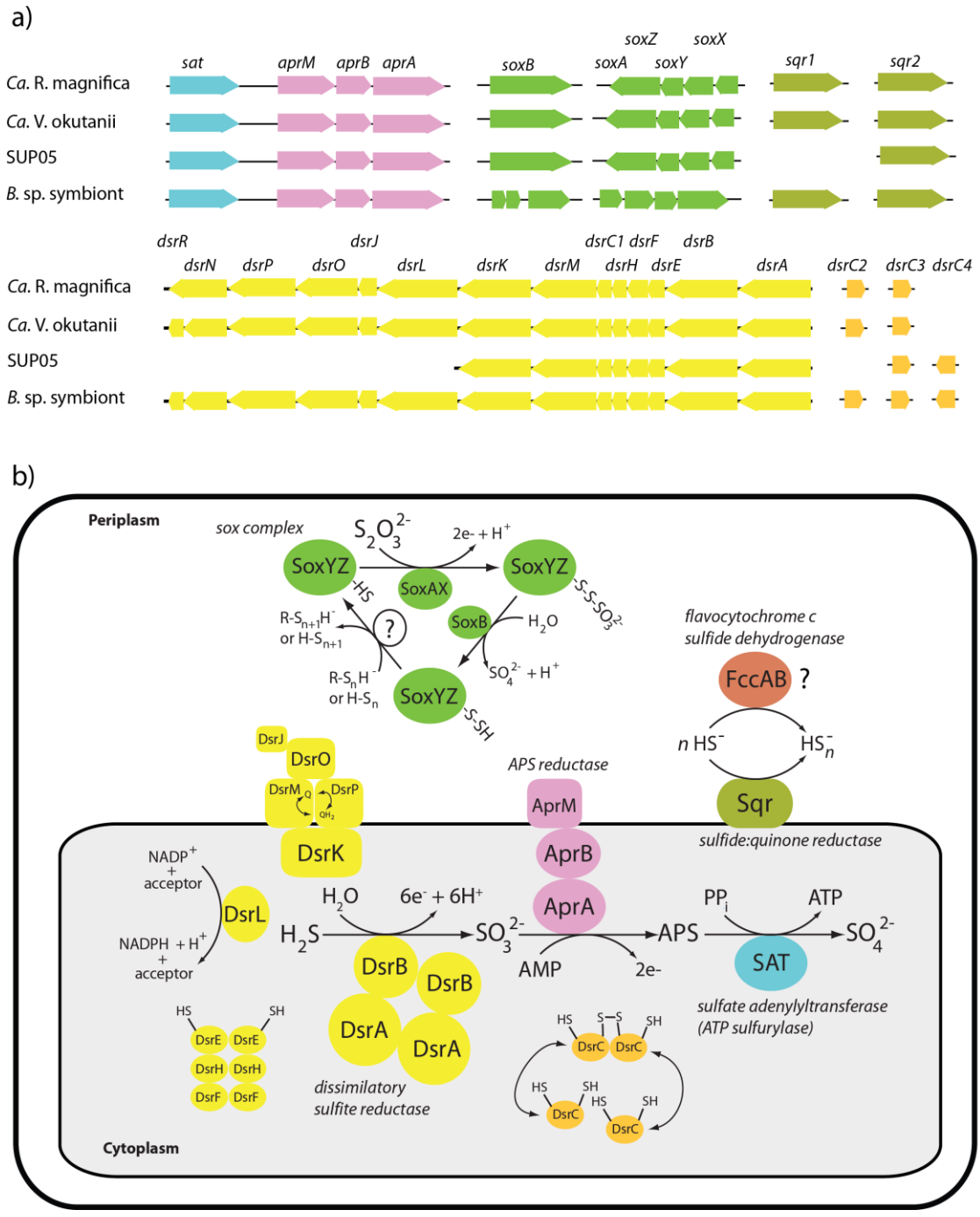


Figure 6-3. a) The order of the genes used for oxidation of reduced sulfur compounds is similar between the *Bathymodiolus* SOX symbionts, the free-living SUP05 and the clam symbionts *Ca*. R.magnifica and *Ca*. V. okutanii (Genomes from Kuwahara *et al*., 2007; Newton *et al*., 2007; Walsh *et al*., 2009). b) Pathways for the use of reduced sulfur compounds in the SOX symbionts. *dsr*, dissimilatory sulfite reductase gene cluster; *sox*, sulfur oxidation gene cluster; *aprAB*, adenosine-5'-phosphosulfate (APS) reductase; *aprM*, APS reductase membrane anchor; SQR, sulfide-quinone reductase; SAT, sulfate adenylyltransferase. Modified from Stewart *et al*. (2011) and Dahl *et al*. (2008).

6.5.3. Concluding remarks on metabolic flexibility

The structure of the SOX symbiont population that can use a particular energy source is likely influenced by the vent conditions during colonization. The results presented in this thesis highlight the importance of obtaining physical and chemical data along with the biological samples, although making these measurements is not straightforward (Zielinski *et al*., 2011). The analysis of the metabolic flexibility should be extended to the presence and expression of the electron acceptors nitrate and oxygen. Ideally, to prove that there is variability in the presence or absence of those genes, a visual technique such as gene-FISH should be used (Moraru *et al*., 2010). This technique would have the advantage that it might allow us to find patterns in the distribution of different symbiont populations, similar to what we expect to see for the MDH region (Chapter V). The variation in the metabolic potential across symbiont populations should be considered when estimating the flux dynamics of carbon fixation under the presence of different energy sources.

The analyses regarding the symbiont variation were restricted in this thesis to a big pool of SOX strains per individual, as sequencing was done with gill homogenates. The mussels are however dynamic organisms that experience different conditions within the gill depending on the water flux. Also, when mussels close their shells (e.g. in response to disturbance) they can experience temporary anoxia, which can trigger the use of alternative electron acceptors such as nitrate. Future studies should investigate the existence of micro-niches within a mussel individual, as these factors might play a major role towards the selection of the symbiont strains within specific regions of the gill.

The next steps will be to i) mine further the wealth of genomic and transcriptomic data that we have now at hand from different *Bathymodiolus* species, and to ii) test the various hypotheses that arise from these analyses. The techniques used could include protein visualization and metabolite measurements (e.g. MALDI-MS, GC-MS), as well as established methods such as enzymatic assays. The results of this thesis set the ground for a vast array of future studies.

References for Chapter I and VI

Ackert LT. (2006). The role of microbes in agriculture: Sergei Vinogradskii's discovery and investigation of chemosynthesis, 1880–1910. J Hist Biol 39: 373–406.

Anantharaman K, Breier JA, Sheik CS, Dick GJ. (2012). Evidence for hydrogen oxidation and metabolic plasticity in widespread deep-sea sulfur-oxidizing bacteria. *Proc Natl Acad Sci* **110**: 330–335.

Anselme C, Vallier A, Balmand S, Fauvarque M-O, Heddi A. (2006). Host PGRP gene expression and bacterial release in endosymbiosis of the weevil *Sitophilus zeamais*. *Appl Environ Microbiol* **72**: 6766–6772.

Backert S, Meyer TF. (2006). Type IV secretion systems and their effectors in bacterial pathogenesis. *Curr Opin Microbiol* **9**: 207–217.

Baldo L, Bordenstein S, Wernegreen JJ, Werren JH. (2006). Widespread recombination throughout *Wolbachia* genomes. *Mol Biol Evol* **23**: 437–449.

Barry JP, Gary Greene H, Orange DL, Baxter CH, Robison BH, Kochevar RE, *et al.* (1996). Biologic and geologic characteristics of cold seeps in Monterey Bay, California. *Deep Sea Res Part Oceanogr Res Pap* **43**: 1739–1762.

de Bary A. (1879). Die Erscheinung der Symbiose: Vortrag gehalten auf der Versammlung Deutscher Naturforscher und Aerzte zu Cassel. Trübner.

Beier D, Gross R. (2006). Regulation of bacterial virulence by two-component systems. *Curr Opin Microbiol* **9**: 143–152.

Beinart RA, Gartman A, Sanders JG, Luther GW, Girguis PR. (2015). The uptake and excretion of partially oxidized sulfur expands the repertoire of energy resources metabolized by hydrothermal vent symbioses. *Proc R Soc B* **282**: 20142811.

Beinart RA, Sanders JG, Faure B, Sylva SP, Lee RW, Becker EL, *et al.* (2012). Evidence for the role of endosymbionts in regional-scale habitat partitioning by hydrothermal vent symbioses. *Proc Natl Acad Sci* **109**: E3241–E3250.

Bergquist DC, Williams FM, Fisher CR. (2000). Longevity record for deep-sea invertebrate. *Nature* **403**: 499–500.

Bernander R, Ettema TJ. (2010). FtsZ-less cell division in archaea and bacteria. *Curr Opin Microbiol* **13**: 747–752.

Bhavsar AP, Guttman JA, Finlay BB. (2007). Manipulation of host-cell pathways by bacterial pathogens. *Nature* **449**: 827–834.

Black MB, Halanych KM, Maas P a. Y, Hoeh WR, Hashimoto J, Desbruyères D, *et al.* (1997). Molecular systematics of vestimentiferan tubeworms from hydrothermal vents and cold-water seeps. *Mar Biol* **130**: 141–149.

Blumenberg M. (2010). Microbial chemofossils in specific marine hydrothermal and methane cold seep settings. In: Kiel S (ed) Vol. 33. *The vent and seep biota*. Springer Netherlands: Dordrecht, pp 73–106.

Braby CE, Rouse GW, Johnson SB, Jones WJ, Vrijenhoek RC. (2007). Bathymetric and temporal variation among *Osedax* boneworms and associated megafauna on whale-falls in Monterey Bay, California. *Deep Sea Res Part Oceanogr Res Pap* **54**: 1773–1791.

Bright M, Bulgheresi S. (2010). A complex journey: Transmission of microbial symbionts. *Nat Rev Microbiol* **8**: 218.

Bull J, Rice W. (1991). Distinguishing mechanisms for the evolution of co-operation. *J Theor Biol* **149**: 63–74.

Campbell BJ, Jeanthon C, Kostka JE, Luther GW, Cary SC. (2001). Growth and phylogenetic properties of novel bacteria belonging to the epsilon subdivision of the Proteobacteria enriched from *Alvinella pompejana* and deep-sea hydrothermal vents. *Appl Environ Microbiol* **67**: 4566–4572.

Cary SC, Giovannoni SJ. (1993). Transovarial inheritance of endosymbiotic bacteria in clams inhabiting deep-sea hydrothermal vents and cold seeps. *Proc Natl Acad Sci* **90**: 5695.

Cavanaugh CM, Gardiner SL, Jones ML, Jannasch HW, Waterbury JB. (1981). Prokaryotic cells in the hydrothermal vent tube worm *Riftia pachyptila* Jones: Possible chemoautotrophic symbionts. *Science* **213**: 340–342.

Cavanaugh CM, McKiness ZP, Newton ILG, Stewart FJ. (2013). Marine chemosynthetic symbioses. In: Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F (eds). *The Prokaryotes*. Springer Berlin Heidelberg, pp 579–607.

Cavanaugh CM, Robinson JJ. (1996). CO_2 fixation in chemoautotroph-invertebrate symbioses: Expression of form I and form II RuBisCO. In: Lidstrom ME, Tabita FR (eds). *Microbial Growth on C1 Compounds*. Springer Netherlands, pp 285–292.

Cerveny KL, Tamura Y, Zhang Z, Jensen RE, Sesaki H. (2007). Regulation of mitochondrial fusion and division. *Trends Cell Biol* **17**: 563–569.

Charlou JL, Donval JP, Douville E, Jean-Baptiste P, Radford-Knoery J, Fouquet Y, *et al.* (2000). Compared geochemical signatures and the evolution of Menez Gwen (37°50′N) and Lucky Strike (37°17′N) hydrothermal fluids, south of the Azores Triple Junction on the Mid-Atlantic Ridge. *Chem Geol* **171**: 49–75.

Charlou JL, Donval JP, Fouquet Y, Jean-Baptiste P, Holm N. (2002). Geochemistry of high H_2 and CH_4 vent fluids issuing from ultramafic rocks at the Rainbow hydrothermal field (36°14′N, MAR). *Chem Geol* **191**: 345–359.

Chaston J, Goodrich-Blair H. (2010). Common trends in mutualism revealed by model associations between invertebrates and bacteria. *FEMS Microbiol Rev* **34**: 41–58.

Childress JJ, Fisher CR. (1992). The biology of hydrothermal vent animals: Physiology, biochemistry, and autotrophic symbioses.

Childress JJ, Fisher CR, Favuzzi JA, Sanders NK. (1991). Sulfide and carbon dioxide uptake by the hydrothermal vent clam, *Calyptogena magnifica*, and its chemoautotrophic symbionts. *Physiol Zool* **64**: 1444–1470.

Chun CK, Troll JV, Koroleva I, Brown B, Manzella L, Snir E, *et al.* (2008). Effects of colonization, luminescence, and autoinducer on host transcription during development of the squid-vibrio association. *Proc Natl Acad Sci* **105**: 11323–11328.

Corliss JB, Dymond J, Gordon LI, Edmond JM. (1979). On the Galapagos Rift. *Science* **203**: 16.

von Cosel R. (2002). A new species of bathymodioline mussel (Mollusca, Bivalvia, Mytilidae) from Mauritania (West Africa), with comments on the genus *Bathymodiolus* Kenk & Wilson, 1985. *Zoosystema* **24**: 259–272.

Dahl C, Friedrich C, Kletzin A. (2008). Sulfur oxidation in prokaryotes. In: John Wiley & Sons, Ltd (ed). *Encyclopedia of Life Sciences*. John Wiley & Sons, Ltd: Chichester, UK. http://doi.wiley.com/10.1002/9780470015902.a0021155 (Accessed September 27, 2011).

Dawkins MS, Guilford T. (1991). The corruption of honest signalling. *Anim Behav* **41**: 865–873.

DeChaine EG, Bates AE, Shank TM, Cavanaugh CM. (2006). Off-axis symbiosis found: Characterization and biogeography of bacterial symbionts of *Bathymodiolus* mussels from Lost City hydrothermal vents. *Environ Microbiol* **8**: 1902–1912.

Degnan PH, Leonardo TE, Cass BN, Hurwitz B, Stern D, Gibbs RA, *et al*. (2010). Dynamics of genome evolution in facultative symbionts of aphids. *Environ Microbiol* **12**: 2060–2069.

Dmytrenko O, Russell SL, Loo WT, Fontanez KM, Liao L, Roeselers G, *et al*. (2014). The genome of the intracellular bacterium of the coastal bivalve, *Solemya velum*: A blueprint for thriving in and out of symbiosis. *BMC Genomics* **15**: 924.

Douville E, Charlou JL, Oelkers EH, Bienvenu P, Jove Colon CF, Donval JP, *et al*. (2002). The Rainbow vent fluids (36°14′N, MAR): The influence of ultramafic rocks and phase separation on trace metal content in Mid-Atlantic Ridge hydrothermal fluids. *Chem Geol* **184**: 37–48.

Dubilier N, Bergin C, Lott C. (2008). Symbiotic diversity in marine animals: The art of harnessing chemosynthesis. *Nat Rev Micro* **6**: 725–740.

Duperron S. (2010). The diversity of deep-sea mussels and their bacterial symbioses. In: Kiel S (ed) Vol. 33. *The vent and seep biota*. Springer Netherlands: Dordrecht, pp 137–167.

Duperron S, Bergin C, Zielinski F, Blazejak A, Pernthaler A, McKiness ZP, *et al*. (2006). A dual symbiosis shared by two mussel species, *Bathymodiolus azoricus* and *Bathymodiolus puteoserpentis* (Bivalvia: Mytilidae), from hydrothermal vents along the northern Mid-Atlantic Ridge. *Environ Microbiol* **8**: 1441–1447.

Duperron S, Halary S, Lorion J, Sibuet M, Gaill F. (2008a). Unexpected co-occurrence of six bacterial symbionts in the gills of the cold seep mussel *Idas* sp. (Bivalvia: Mytilidae). *Environ Microbiol* **10**: 433–445.

Duperron S, Laurent MC, Gaill F, Gros O. (2008b). Sulphur-oxidizing extracellular bacteria in the gills of Mytilidae associated with wood falls. *FEMS Microbiol Ecol* **63**: 338–349.

Endow K, Ohta S. (1990). Occurrence of bacteria in the primary oocytes of vesicomyid clam *Calyptogena soyoae*. *Mar Ecol Prog Ser Oldendorf* **64**: 309–311.

Erickson HP, Osawa M. (2010). Cell division without FtsZ - a variety of redundant mechanisms. *Mol Microbiol* **78**: 267–270.

Fenchel T. (2002). Microbial behavior in a heterogeneous world. *Science* **296**: 1068–1071.

Ferrière R, Gauduchon M, Bronstein JL. (2007). Evolution and persistence of obligate mutualists and exploiters: competition for partners and evolutionary immunization. *Ecol Lett* **10**: 115–126.

Fiala-Médioni A, Métivier C. (1986). Ultrastructure of the gill of the hydrothermal vent bivalve *Calyptogena magnifica*, with a discussion of its nutrition. *Mar Biol* **90**: 215–222.

Fisher C, Childress J, Oremland R, Bidigare R. (1987). The importance of methane and thiosulfate in the metabolism of the bacterial symbionts of two deep-sea mussels. *Mar Biol* **96**: 59–71.

Fontanez KM, Cavanaugh CM. (2014). Evidence for horizontal transmission from multilocus phylogeny of deep-sea mussel (Mytilidae) symbionts: Horizontal transmission of mussel symbionts. *Environ Microbiol* 3608–3621.

Friedrich CG, Rother D, Bardischewsky F, Quentmeier A, Fischer J. (2001). Oxidation of reduced inorganic sulfur compounds by bacteria: Emergence of a common mechanism? *Appl Environ Microbiol* **67**: 2873–2882.

Fuerst JA. (2005). Intracellular compartmentation in planctomycetes. *Annu Rev Microbiol* **59**: 299–328.

Fujiwara Y, Kawato M, Noda C, Kinoshita G, Yamanaka T, Fujita Y, *et al*. (2010). Extracellular and mixotrophic symbiosis in the whale-fall mussel *Adipicola pacifica*: A trend in evolution from extra- to intracellular symbiosis. *PLoS ONE* **5**. e-pub ahead of print, doi: 10.1371/journal.pone.0011808.

Gartman A, Yücel M, Madison AS, Chu DW, Ma S, Janzen CP, *et al.* (2011). Sulfide oxidation across diffuse flow zones of hydrothermal vents. *Aquat Geochem* **17**: 583–601.

German CR, Ramirez-Llodra E, Baker MC, Tyler PA, and the ChEss Scientific Steering Committee. (2011). Deep-water chemosynthetic ecosystem research during the census of marine life decade and beyond: A proposed deep-ocean road map. *PLoS ONE* **6**: e23259.

Gil R, Latorre A, Moya A. (2011). Evolution of prokaryote-animal symbiosis from a genomics perspective. *Endo Symbiotic Methanogenic Archaea* 207–233.

Gil R, Sabater-Muñoz B, Latorre A, Silva FJ, Moya A. (2002). Extreme genome reduction in *Buchnera* spp.: Toward the minimal genome needed for symbiotic life. *Proc Natl Acad Sci* **99**: 4454–4458.

Gilbert SF, Bosch TCG, Ledón-Rettig C. (2015). Eco-Evo-Devo: Developmental symbiosis and developmental plasticity as evolutionary agents. *Nat Rev Genet* **16**: 611–622.

Goebel W, Gross R. (2001). Intracellular survival strategies of mutualistic and parasitic prokaryotes. *Trends Microbiol* **9**: 267–273.

Goffredi SK, Johnson SB, Vrijenhoek RC. (2007). Genetic diversity and potential function of microbial symbionts associated with newly discovered species of *Osedax* polychaete worms. *Appl Environ Microbiol* **73**: 2314–2323.

Gogarten JP, Townsend JP. (2005). Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* **3**: 679–687.

Goodrich-Blair H, Clarke DJ. (2007). Mutualism and pathogenesis in *Xenorhabdus* and *Photorhabdus*: Two roads to the same destination. *Mol Microbiol* **64**: 260–268.

Gourion B, Berrabah F, Ratet P, Stacey G. (2015). Rhizobium–legume symbioses: The crucial role of plant immunity. *Trends Plant Sci* **20**: 186–194.

Gruber-Vodicka HR, Dirks U, Leisch N, Baranyi C, Stoecker K, Bulgheresi S, *et al.* (2011). *Paracatenula*, an ancient symbiosis between thiotrophic Alphaproteobacteria and catenulid flatworms. *Proc Natl Acad Sci* **108**: 12078–12083.

Hao Z, Kasumba I, Lehane MJ, Gibson WC, Kwon J, Aksoy S. (2001). Tsetse immune responses and trypanosome transmission: Implications for the development of tsetse-based strategies to reduce trypanosomiasis. *Proc Natl Acad Sci* **98**: 12648–12653.

Hensen D, Sperling D, Trüper HG, Brune DC, Dahl C. (2006). Thiosulphate oxidation in the phototrophic sulphur bacterium *Allochromatium vinosum*. *Mol Microbiol* **62**: 794–810.

Hentschel U, Steinert M, Hacker J. (2000). Common molecular mechanisms of symbiosis and pathogenesis. *Trends Microbiol* **8**: 226–231.

Hibbing ME, Fuqua C, Parsek MR, Peterson SB. (2010). Bacterial competition: Surviving and thriving in the microbial jungle. *Nat Rev Microbiol* **8**: 15–25.

Hillman K, Goodrich-Blair H. (2016). Are you my symbiont? Microbial polymorphic toxins and antimicrobial compounds as honest signals of beneficial symbiotic defensive traits. *Curr Opin Microbiol* **31**: 184–190.

Hueck CJ. (1998). Type III protein secretion systems in bacterial pathogens of animals and plants. *Microbiol Mol Biol Rev* **62**: 379–433.

Hughes DT, Sperandio V. (2008). Inter-kingdom signalling: Communication between bacteria and their hosts. *Nat Rev Microbiol* **6**: 111–120.

Humphris SE, McCollom T. (1998). The cauldron beneath the seafloor. *Oceanus* **41**: 18.

Ikuta T, Takaki Y, Nagai Y, Shimamura S, Tsuda M, Kawagucci S, *et al.* (2015). Heterogeneous composition of key metabolic gene clusters in a vent mussel symbiont population. *ISME J*. e-pub ahead of print, doi: 10.1038/ismej.2015.176.

Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV. (2012). Untangling genomes from metagenomes: Revealing an uncultured class of marine Euryarchaeota. *Science* **335**: 587–590.

Jan C, Petersen JM, Werner J, Teeling H, Huang S, Glöckner FO, *et al*. (2014). The gill chamber epibiosis of deep-sea shrimp *Rimicaris exoculata*: An in-depth metagenomic investigation and discovery of Zetaproteobacteria. *Environ Microbiol* **16**: 2723–2738.

Jannasch HW, Mottl MJ. (1985). Geomicrobiology of deep-sea hydrothermal vents. *Science* **229**: 717–725.

Jewell T, Huston SL, Nelson DC. (2008). Methylotrophy in freshwater *Beggiatoa alba* strains. *Appl Environ Microbiol* **74**: 5575–5578.

Jørgensen BB, Boetius A. (2007). Feast and famine - microbial life in the deep-sea bed. *Nat Rev Microbiol* **5**: 770–781.

Kadar 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*. *J Exp Mar Biol Ecol* **318**: 99–110.

Kaltenpoth M, Roeser-Mueller K, Koehler S, Peterson A, Nechitaylo TY, Stubblefield JW, *et al.* (2014). Partner choice and fidelity stabilize coevolution in a Cretaceous-age defensive symbiosis. *Proc Natl Acad Sci* **111**: 6359–6364.

Kelley DS, Karson JA, Früh-Green GL, Yoerger DR, Shank TM, Butterfield DA, *et al.* (2005). A serpentinite-hosted ecosystem: The Lost City hydrothermal field. *Science* **307**: 1428–1434.

Kendall MM, Sperandio V. (2016). What a dinner party! Mechanisms and functions of interkingdom signaling in host-pathogen associations. *mBio* **7**: e01748-15.

Kiel S. (2006). Cold-seep mollusks are older than the general marine mollusk fauna. *Science* **313**: 1429–1431.

Kleiner M, Petersen JM, Dubilier N. (2012a). Convergent and divergent evolution of metabolism in sulfur-oxidizing symbionts and the role of horizontal gene transfer. *Curr Opin Microbiol* **15**: 621–631.

Kleiner M, Wentrup C, Holler T, Lavik G, Harder J, Lott C, *et al.* (2015). Use of carbon monoxide and hydrogen by a bacteria–animal symbiosis from seagrass sediments. *Environ Microbiol* **17**: 5023–5035.

Kleiner M, Wentrup C, Lott C, Teeling H, Wetzel S, Young J, *et al.* (2012b). Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. *Proc Natl Acad Sci U S A* **109**: E1173-1182.

Koskiniemi S, Lamoureux JG, Nikolakakis KC, Roodenbeke C t'Kint de, Kaplan MD, Low DA, *et al.* (2013). Rhs proteins from diverse bacteria mediate intercellular competition. *Proc Natl Acad Sci* **110**: 7032–7037.

Kuwahara H, Takaki Y, Shimamura S, Yoshida T, Maeda T, Kunieda T, *et al.* (2011). Loss of genes for DNA recombination and repair in the reductive genome evolution of thioautotrophic symbionts of *Calyptogena* clams. *BMC Evol Biol* **11**: 285.

Kuwahara H, Takaki Y, Yoshida T, Shimamura S, Takishita K, Reimer JD, *et al.* (2008). Reductive genome evolution in chemoautotrophic intracellular symbionts of deep-sea *Calyptogena* clams. *Extremophiles* **12**: 365–374.

Kuwahara H, Yoshida T, Takaki Y, Shimamura S, Nishi S, Harada M, *et al.* (2007). Reduced genome of the thioautotrophic intracellular symbiont in a deep-sea clam, *Calyptogena okutanii*. *Curr Biol* **17**: 881–886.

Kwong WK, Engel P, Koch H, Moran NA. (2014). Genomics and host specialization of honey bee and bumble bee gut symbionts. *Proc Natl Acad Sci* **111**: 11509–11514.

Lalou C, Brichet E. (1982). Ages and implications of East Pacific Rise sulphide deposits at 21 °N. *Nature* **300**: 169–171.

Lascelles DF. (2007). Black smokers and density currents: A uniformitarian model for the genesis of banded iron-formations. *Ore Geol Rev* **32**: 381–411.

Lawniczak M, Barnes A, Linklater J, Boone J, Wigby S, Chapman T. (2007). Mating and immunity in invertebrates. *Trends Ecol Evol* **22**: 48–55.

Lilley MD, Butterfield DA, Lupton JE, Olson EJ. (2003). Magmatic events can produce rapid changes in hydrothermal vent chemistry. *Nature* **422**: 878–881.

Lindås A-C, Karlsson EA, Lindgren MT, Ettema TJG, Bernander R. (2008). A unique cell division machinery in the Archaea. *Proc Natl Acad Sci* **105**: 18942–18946.

Lipsett M. (1987). Causation in toxic tort litigation: The role of epidemiologic evidence. *J Hazard Mater* **15**: 185–203.

Lluch-Senar M, Querol E, Piñol J. (2010). Cell division in a minimal bacterium in the absence of *ftsZ*. *Mol Microbiol* **78**: 278–289.

Loker ES, Adema CM, Zhang S-M, Kepler TB. (2004). Invertebrate immune systems – not homogeneous, not simple, not well understood. *Immunol Rev* **198**: 10–24.

Loose M, Mitchison TJ. (2014). The bacterial cell division proteins FtsA and FtsZ self-organize into dynamic cytoskeletal patterns. *Nat Cell Biol* **16**: 38–46.

Lopanik NB. (2014). Chemical defensive symbioses in the marine environment. *Funct Ecol* **28**: 328–340.

Lorion J, Duperron S, Gros O, Cruaud C, Samadi S. (2009). Several deep-sea mussels and their associated symbionts are able to live both on wood and on whale falls. *Proc R Soc Lond B Biol Sci* **276**: 177–185.

Lorion J, Kiel S, Faure B, Kawato M, Ho SYW, Marshall B, *et al.* (2013). Adaptive radiation of chemosymbiotic deep-sea mussels. *Proc R Soc Lond B Biol Sci* **280**: 20131243.

MacDonald IR, Jr NLG, Reilly JF, Brooks JM, Callender WR, Gabrielle SG. (1990). Gulf of Mexico hydrocarbon seep communities: VI. Patterns in community structure and habitat. *Geo-Mar Lett* **10**: 244–252.

Marbler H, Koschinsky A, Pape T, Seifert R, Weber S, Baker ET, *et al.* (2010). Geochemical and physical structure of the hydrothermal plume at the ultramafic-hosted Logatchev hydrothermal field at 14°45′N on the Mid-Atlantic Ridge. *Mar Geol* **271**: 187–197.

Markert S, Arndt C, Felbeck H, Becher D, Sievert SM, Hügler M, *et al.* (2007). Physiological Proteomics of the Uncultured Endosymbiont of *Riftia pachyptila*. *Science* **315**: 247–250.

Markert S, Gardebrecht A, Felbeck H, Sievert SM, Klose J, Becher D, *et al.* (2011). Status quo in physiological proteomics of the uncultured *Riftia pachyptila* endosymbiont. *Proteomics* **11**: 3106–3117.

Marshall KT, Morris RM. (2015). Genome sequence of '*Candidatus* Thioglobus singularis' Strain PS1, a mixotroph from the SUP05 clade of marine Gammaproteobacteria. *Genome Announc* **3**: e01155-15.

Martin W, Baross J, Kelley D, Russell MJ. (2008). Hydrothermal vents and the origin of life. *Nat Rev Microbiol* **6**: 805–814.

McCutcheon JP, Moran NA. (2012). Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* **10**: 13–26.

McFall-Ngai M, Heath-Heckman EA, Gillette AA, Peyer SM, Harvie EA. (2012). The secret languages of coevolved symbioses: Insights from the *Euprymna scolopes–Vibrio fischeri* symbiosis. In: Vol. 24. Elsevier, pp 3–8.

Medzhitov R, Janeway Jr CA. (1998). Innate immune recognition and control of adaptive immune responses. *Semin Immunol* **10**: 351–353.

Moran NA. (2002). Microbial minimalism: Genome reduction in bacterial pathogens. *Cell* **108**: 583–586.

Moran NA. (2003). Tracing the evolution of gene loss in obligate bacterial symbionts. *Curr Opin Microbiol* **6**: 512–518.

Moran NA, McCutcheon JP, Nakabachi A. (2008). Genomics and evolution of heritable bacterial symbionts. *Annu Rev Genet* **42**: 165–190.

Moran NA, Wernegreen JJ. (2000). Lifestyle evolution in symbiotic bacteria: Insights from genomics. *Trends Ecol Evol* **15**: 321–326.

Moraru C, Lam P, Fuchs BM, Kuypers MMM, Amann R. (2010). GeneFISH – an *in situ* technique for linking gene presence and cell identity in environmental microorganisms. *Environ Microbiol* **12**: 3057–3073.

Moya A, Pereto J, Gil R, Latorre A. (2008). Learning how to live together: Genomic insights into prokaryote-animal symbioses. *Nat Rev Genet* **9**: 218–229.

Nelson D, Fisher C. (1995). Chemoautotrophic and methanotrophic endosymbiotic bacteria at deep-sea vents and seeps. *Microbiol Deep-Sea Hydrothermal Vents* 125–167.

Nelson DC, Hagen KD, Edwards DB. (1995). The gill symbiont of the hydrothermal vent mussel *Bathymodiolus thermophilus* is a psychrophilic, chemoautotrophic, sulfur bacterium. *Mar Biol* **121**: 487–495.

Neumann D, Kappes H. (2003). On the growth of bivalve gills initiated from a lobule-producing budding zone. *Biol Bull* **205**: 73–82.

Newton I, Girguis P, Cavanaugh C. (2008). Comparative genomics of vesicomyid clam (Bivalvia: Mollusca) chemosynthetic symbionts. *BMC Genomics* **9**: 585.

Newton ILG, Woyke T, Auchtung TA, Dilly GF, Dutton RJ, Fisher MC, *et al*. (2007). The *Calyptogena magnifica* chemoautotrophic symbiont genome. *Science* **315**: 998–1000.

Nyholm SV, McFall-Ngai M. (2004). The winnowing: Establishing the squid-vibrio symbiosis. *Nat Rev Microbiol* **2**: 632–642.

Ohta T. (1987). Simulating evolution by gene duplication. *Genetics* **115**: 207–213.

Oliver KM, Russell JA, Moran NA, Hunter MS. (2003). Facultative bacterial symbionts in aphids confer resistance to parasitic wasps. *Proc Natl Acad Sci* **100**: 1803–1807.

Oliver KM, Smith AH, Russell JA. (2014). Defensive symbiosis in the real world – advancing ecological studies of heritable, protective bacteria in aphids and beyond. *Funct Ecol* **28**: 341–355.

Olsnes S, Wesche J, Falnes P. (2013). Uptake of protein toxins acting inside cells. In: Aktories K, Just I (eds) Vol. 145. *Bacterial protein toxins*. Springer Science & Business Media.

Ondréas H, Cannat M, Fouquet Y, Normand A, Sarradin PM, Sarrazin J. (2009). Recent volcanic events and the distribution of hydrothermal venting at the Lucky Strike hydrothermal field, Mid-Atlantic Ridge. *Geochem Geophys Geosystems* **10**: Q02006.

Oren A. (2009). Chemolithotrophy. *Encycl Life Sci*. e-pub ahead of print, doi: 10.1002/9780470015902.a0021153.

Ouellette SP, Karimova G, Subtil A, Ladant D. (2012). *Chlamydia* co-opts the rod shape-determining proteins MreB and Pbp2 for cell division. *Mol Microbiol* **85**: 164–178.

Pailleret M, Haga T, Petit P, Privé-Gill C, Saedlou N, Gaill F, *et al.* (2007). Sunken wood from the Vanuatu Islands: Identification of wood substrates and preliminary description of associated fauna. *Mar Ecol* **28**: 233–241.

Pak JW, Jeon KW. (1997). A symbiont-produced protein and bacterial symbiosis in *Amoeba proteus*. *J Eukaryot Microbiol* **44**: 614–619.

Pawlowska M. (2014). Mitigation of landfill gas emissions. CRC Press: Balkema, London, UK.

Peek AS, Feldman RA, Lutz RA, Vrijenhoek RC. (1998). Cospeciation of chemoautotrophic bacteria and deep sea clams. *Proc Natl Acad Sci* **95**: 9962.

Petersen JM, Dubilier N. (2009). Methanotrophic symbioses in marine invertebrates: Methanotrophic symbioses in marine invertebrates. *Environ Microbiol Rep* **1**: 319–335.

Petersen JM, Ramette A, Lott C, Cambon-Bonavita M-A, Zbinden M, Dubilier N. (2010). Dual symbiosis of the vent shrimp *Rimicaris exoculata* with filamentous gamma- and epsilonproteobacteria at four Mid-Atlantic Ridge hydrothermal vent fields. *Environ Microbiol* **12**: 2204–2218.

Petersen JM, Wentrup C, Verna C, Knittel K, Dubilier N. (2012). Origins and evolutionary flexibility of chemosynthetic symbionts from deep-sea animals. *Biol Bull* **223**: 123–137.

Petersen JM, Zielinski FU, Pape T, Seifert R, Moraru C, Amann R, *et al.* (2011). Hydrogen is an energy source for hydrothermal vent symbioses. *Nature* **476**: 176–180.

Pimenov NV, Kalyuzhnaya MG, Khmelenina VN, Mityushina LL, Trotsenko YA. (2002). Utilization of methane and carbon dioxide by symbiotrophic bacteria in gills of Mytilidae (*Bathymodiolus*) from the Rainbow and Logachev hydrothermal fields on the Mid-Atlantic Ridge. *Microbiology* **71**: 587–594.

Ponnudurai R, Kleiner M, Sayavedra L, Petersen JM, Moche M, Otto A, et al. (2016). Metabolic and physiological interdependencies in the *Bathymodiolus azoricus* symbiosis. ISME J. e-pub ahead of print, doi: 10.1038/ismej.2016.124.

Raggi L, Schubotz F, Hinrichs K-U, Dubilier N, Petersen JM. (2013). Bacterial symbionts of *Bathymodiolus* mussels and *Escarpia* tubeworms from Chapopote, an asphalt seep in the southern Gulf of Mexico. *Environ Microbiol* **15**: 1969–1987.

Rawlings D, Tributsch H, Hansford G. (1999). Reasons why '*Leptospirillum*'-like species rather than *Thiobacillus ferrooxidans* are the dominant iron-oxidizing bacteria in many commercial processes for the biooxidation of pyrite and related ores. *Microbiology* **145**: 5–13.

Riou V, Duperron S, Halary S, Dehairs F, Bouillon S, Martins I, *et al.* (2010). Variation in physiological indicators in *Bathymodiolus azoricus* (Bivalvia: Mytilidae) at the Menez Gwen Mid-Atlantic Ridge deep-sea hydrothermal vent site within a year. *Mar Environ Res* **70**: 264–271.

Rissman AI, Mau B, Biehl BS, Darling AE, Glasner JD, Perna NT. (2009). Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics* **25**: 2071–2073.

Robidart JC, Bench SR, Feldman RA, Novoradovsky A, Podell SB, Gaasterland T, *et al.* (2008). Metabolic versatility of the *Riftia pachyptila* endosymbiont revealed through metagenomics. *Environ Microbiol* **10**: 727–737.

Robinson J, Polz M, Fiala-Medioni A, Cavanaugh C. (1998). Physiological and immunological evidence for two distinct C1-utilizing pathways in *Bathymodiolus puteoserpentis* (Bivalvia: Mytilidae), a dual endosymbiotic mussel from the Mid-Atlantic Ridge. *Mar Biol* **132**: 625–633.

Rocha EPC, Danchin A. (2002). Base composition bias might result from competition for metabolic resources. *Trends Genet* **18**: 291–294.

Rodou A, Ankrah DO, Stathopoulos C. (2010). Toxins and secretion systems of *Photorhabdus luminescens*. *Toxins* **2**: 1250–1264.

Rothfield L, Justice S, García-Lara J. (1999). Bacterial cell division. *Annu Rev Genet* **33**: 423–448.

Salerno JL, Macko SA, Hallam SJ, Bright M, Won Y-J, McKiness Z, *et al.* (2005). Characterization of symbiont populations in life-history stages of mussels from chemosynthetic environments. *Biol Bull* **208**: 145–155.

Sayavedra L, Kleiner M, Ponnudurai R, Wetzel S, Pelletier E, Barbe V, *et al.* (2015). Abundant toxin-related genes in the genomes of beneficial symbionts from deep-sea hydrothermal vent mussels. *eLife* **4**. e-pub ahead of print, doi: 10.7554/eLife.07966.

Shah V, Morris RM. (2015). Genome Sequence of '*Candidatus* Thioglobus autotrophica' strain EF1, a chemoautotroph from the SUP05 clade of marine Gammaproteobacteria. *Genome Announc* **3**: e01156-15.

Sharma D, Bose A, Shakila H, Das TK, Tyagi JS, Ramanathan VD. (2006). Expression of mycobacterial cell division protein, FtsZ, and dormancy proteins, DevR and Acr, within lung granulomas throughout guinea pig infection. *FEMS Immunol Med Microbiol* **48**: 329–336.

Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* **407**: 81–86.

Sibuet M, Roy KO-L. (2002). Cold seep communities on continental margins: Structure and quantitative distribution relative to geological and fluid venting patterns. In: Wefer PDG, Billett DD, Hebbeln DD, Jørgensen PDBB, Schlüter PDM, Weering DTCE Van (eds). *Ocean margin systems*. Springer Berlin Heidelberg, pp 235–251.

Smith CR, Baco AR. (2003). Ecology of whale falls at the deep-sea floor. https://books.google.de/books?hl=en&lr=&id=64crGFXWn5gC&oi=fnd&pg=PA311&ots=LYHALsFf--&sig=e78H4iSK9T41pRMMqVw4vP5szMk (Accessed April 12, 2016).

Snyder AK, Rio RVM. (2013). Interwoven biology of the Tsetse holobiont. *J Bacteriol* **195**: 4322–4330.

Soulé ME. (1987). Viable populations for conservation. Cambridge University Press.

Sperandio V, Torres AG, Jarvis B, Nataro JP, Kaper JB. (2003). Bacteria–host communication: The language of hormones. *Proc Natl Acad Sci* **100**: 8951–8956.

Starr M. (1975). A generalized scheme for classifying organismic associations. In: Vol. 29. *Symposia of the Society for Experimental Biology*. p 1.

Stewart FJ, Dmytrenko O, DeLong EF, Cavanaugh CM. (2011). Metatranscriptomic analysis of sulfur oxidation genes in the endosymbiont of *Solemya Velum*. *Front Microbiol* **2**. e-pub ahead of print, doi: 10.3389/fmicb.2011.00134.

Stewart FJ, Newton ILG, Cavanaugh CM. (2005). Chemosynthetic endosymbioses: Adaptations to oxic–anoxic interfaces. *Trends Microbiol* **13**: 439–448.

Stewart FJ, Young CR, Cavanaugh CM. (2008). Lateral symbiont acquisition in a maternally transmitted chemosynthetic clam endosymbiosis. *Mol Biol Evol* **25**: 673–687.

Sullivan MJ, Petty NK, Beatson SA. (2011). Easyfig: A genome comparison visualizer. *Bioinformatics* **27**: 1009–1010.

Sunamura M, Higashi Y, Miyako C, Ishibashi J, Maruyama A. (2004). Two bacteria phylotypes are predominant in the Suiyo Seamount hydrothermal plume. *Appl Environ Microbiol* **70**: 1190–1198.

Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D, *et al.* (2011). Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* **333**: 1296–1300.

Taylor JD, Glover EA. (2010). Chemosymbiotic bivalves. In: Kiel S (ed) Vol. 33. *The vent and seep biota*. Springer Netherlands: Dordrecht, pp 107–135.

Teeling H, Glöckner FO. (2012). Current opportunities and challenges in microbial metagenome analysis—a bioinformatic perspective. *Brief Bioinform* bbs039.

Tivey M. (2007). Generation of seafloor hydrothermal vent fluids and associated mineral deposits. *Oceanography* **20**: 50–65.

Van Dover C. (2000). The ecology of deep-sea hydrothermal vents. Princeton University Press: Princeton N.J.

Van Dover CL. (2002). Evolution and biogeography of deep-sea vent and seep invertebrates. *Science* **295**: 1253–1257.

Van Dover CL. (2010). Mining seafloor massive sulphides and biodiversity: What is at risk? *ICES J Mar Sci J Cons* fsq086.

Van Dover CL. (2003). Variation in community structure within hydrothermal vent mussel beds of the East Pacific Rise. *Mar Ecol Prog Ser* **253**: 55–66.

de Vienne D, G Refrégier, López‐Villavicencio M, Tellier A, Hood M, Giraud T. (2013). Cospeciation vs host–shift speciation: Methods for testing, evidence from natural associations and relation to coevolution. *New Phytol* **198**: 347–385.

Von Damm KL, Bray AM, Buttermore LG, Oosting SE. (1998). The geochemical controls on vent fluids from the Lucky Strike vent field, Mid-Atlantic Ridge. *Earth Planet Sci Lett* **160**: 521–536.

Vrijenhoek RC. (2010). Genetics and evolution of deep-sea chemosynthetic bacteria and their invertebrate hosts. In: Kiel S (ed) Vol. 33. *The vent and seep biota*. Springer Netherlands: Dordrecht, pp 15–49.

Walker A, Crossman LC. (2007). This place is big enough for both of us. *Nat Rev Microbiol* **5**: 90–92.

Walsh DA, Zaikova E, Howes CG, Song YC, Wright JJ, Tringe SG, *et al.* (2009). Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. *Science* **326**: 578–582.

Weiss RF, Lonsdale P, Lupton JE, Bainbridge AE, Craig H. (1977). Hydrothermal plumes in the Galapagos Rift. *Nature* **267**: 600–603.

Wentrup C, Wendeberg A, Huang JY, Borowski C, Dubilier N. (2013). Shift from widespread symbiont infection of host tissues to specific colonization of gills in juvenile deep-sea mussels. *ISME J* **7**: 1244–1247.

Wentrup C, Wendeberg A, Schimak M, Borowski C, Dubilier N. (2014). Forever competent: Deep-sea bivalves are colonized by their chemosynthetic symbionts throughout their lifetime. *Environ Microbiol* 3699–3713.

Wernegreen JJ. (2005). For better or worse: Genomic consequences of intracellular mutualism and parasitism. *Curr Opin Genet Dev* **15**: 572–583.

Wetzel LR, Shock EL. (2000). Distinguishing ultramafic-from basalt-hosted submarine hydrothermal systems by comparing calculated vent fluid compositions. *J Geophys Res Solid Earth* **105**: 8319–8340.

Whitehead NA, Barnard AML, Slater H, Simpson NJL, Salmond GPC. (2001). Quorum-sensing in Gram-negative bacteria. *FEMS Microbiol Rev* **25**: 365–404.

Won YJ, Hallam SJ, O'Mullan GD, Pan IL, Buck KR, Vrijenhoek RC. (2003). Environmental acquisition of thiotrophic endosymbionts by deep-sea mussels of the genus *Bathymodiolus*. *Appl Environ Microbiol* **69**: 6785.

Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, *et al.* (2006). Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443**: 950–955.

Young C, Fujio S, Vrijenhoek R. (2008). Directional dispersal between mid - ocean ridges: Deep - ocean circulation and gene flow in *Ridgeia piscesae*. *Mol Ecol* **17**: 1718–1731.

Zal F, Lallier FH, Wall JS, Vinogradov SN, Toulmond A. (1996). The multi-hemoglobin system of the hydrothermal vent tube worm *Riftia pachyptila*. *J Biol Chem* **271**: 8869–8874.

Zal F, Leize E, Oros DR, Hourdez S, Van Dorsselaer A, Childress JJ. (2000). Haemoglobin structure and biochemical characteristics of the sulphide-binding component from the deep-sea clam *Calyptogena magnifica*. *Cah Biol Mar* **41**: 413–424.

Zielinski FU, Gennerich H-H, Borowski C, Wenzhöfer F, Dubilier N. (2011). *In situ* measurements of hydrogen sulfide, oxygen, and temperature in diffuse fluids of an ultramafic - hosted hydrothermal vent field (Logatchev, 14° 45′ N, Mid - Atlantic Ridge): Implications for chemosymbiotic bathymodiolin mussels. *Geochem Geophys Geosystems* **12**.

Zientz E, Dandekar T, Gross R. (2004). Metabolic interdependence of obligate intracellular bacteria and their insect hosts. *Microbiol Mol Biol Rev* **68**: 745–770.

Zimmermann J, Wentrup C, Sadowski M, Blazejak A, Gruber-Vodicka H, Kleiner M, *et al.* (2016). Closely coupled evolutionary history of ecto- and endosymbionts from two distantly-related animal phyla. *Mol Ecol*. e-pub ahead of print, doi: 10.1111/mec.13554.

Acknowledgments

This thesis would not have been possible without the help of many people, which gave me scientific and moral support during the years I spent in Germany. I am especially grateful to:

The **German Academic Exchange Service (DAAD)** and the **Max Planck Society**, for financial support during the time of my PhD studies.

Prof. Dr. Nicole Dubilier, for accepting me in your group - now department -, for showing me my strengths and weakness, and for your advice of the scientific world.

Dr. Jill Petersen, for taking the time to review this thesis, great supervision and for life advice. I take with me even the tip of "righty tighty" that you gave me during the Nautilus cruise. If I stay in science, I hope I can transmit your knowledge to future generations.

Dr. Matthias Ullrich, for taking the time to review this thesis.

Prof. Dr. Michael Friedrich and **Tina Enders**, for taking the time to be on the examination board.

Brandon Seah, for all the scientific and non-scientific discussions, for your friendship, for proofreading parts of this thesis, and for agreeing to be a member of the examination board.

Rebecca Ansorge, for all the great times as student, colleague, and friend. Also for proofreading parts of this thesis.

Miguel A. Glez. Porras, for being such an enthusiastic student with a strong laugh.

Adrien Assié, for all the *Bathy* discussions and for facilitating so many samples.

Miriam Sadowsky and **Silke Wetzel** for her technical assistance.

Christian Borowsky, for all the help looking for samples in the freezers, and providing information about multiple cruises.

Oli Jäckle, for proofreading my German summary and being such a great guy.

Niko Leisch, for kindly providing a TEM picture of the Bathy symbionts.

The former and current members of the **Symbiosis Department**, for being an extraordinary team.

All the people acknowledged in the corresponding chapters.

My former and current office-mates, Chia-I, Manuel K., Harald, Rebecca, Max, Dolma, for spontaneous discussions, and a great atmosphere.

Dr. Hanno Teeling, for all the advice.

Antony CP, for your friendship and proofreading parts of this thesis.

Judith Zimmermann, for your friendship and all the moral support, for listening to me in the good and bad times, and for proofreading parts of this thesis.

To my friends, those who already left and those who are still here. Without you, life in Germany would have never been the same: **Jime**, **Betty**, **Rodrigo**, **Saar**, **Laura Gallego**, **Antonio Ramos**, **Gerd**, **Max** and **Jenny**.

To **my family in Germany**, for all the new years, Christmas, and other celebrations, and for all the help during the adaptation process.

Mi familia en México, tan lejos y tan cerca al mismo tiempo. **Mamá**, gracias por tu apoyo en mis decisiones de vida y por alegrarme cuando fue necesario. Gracias a mi abuelito **Rubén** y mi abuelita **Mercedes** que me brindaron su apoyo incondicional y dejaron éste mundo recientemente.

My husband **Dani**, for being the sunshine of the cold and rainy days of Germany. Thanks for your love and support! I also thank you for proofreading parts of this thesis and for translating my summary.