Strain diversity and evolution in endosymbionts of *Bathymodiolus* mussels

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

dem Fachbereich Biologie/Chemie der
Universität Bremen
vorgelegt von

Rebecca Ansorge

Bremen
Februar 2019

Die Untersuchungen zur vorliegenden Doktorarbeit wurden in der Abteilung Symbiose am Max-Planck-Institut für Marine Mikrobiologie unter der Leitung von Prof. Dr. Nicole Dubilier durchgeführt.

Gutachter

1. Prof. Dr. Nicole Dubilier
2. Prof. Dr. Matthias Horn

Prüfer

2. Dr. Bernhard Fuchs
3. Prof. Dr. Marko Rohlfs

Tag des Promotionskolloquiums

26. März 2019

To Life.

“It is not the mountain we conquer but ourselves”

- Edmund Hillary

“The important thing is not to stop questioning. Curiosity has its own reason for existence. One cannot help but be in awe when he contemplates the mysteries of eternity, of life, of the marvelous structure of reality. It is enough if one tries merely to comprehend a little of this mystery each day.”

—"Old Man's Advice to Youth: 'Never Lose a Holy Curiosity.'" *LIFE Magazine* (2 May 1955) p. 64”, Albert Einstein

Summary

Bacterial populations in the environment are often complex and characterized by high genetic diversity. The factors that determine microbial community composition, spatial organization, co-existence and genome evolution are still not well understood. Accumulating evidence shows that even closely related strains of the same species can differ strongly in their functions. However, it is often unknown whether these closely related lineages can co-exist. Especially in intimate symbioses between bacteria and animal hosts, the impact of strain-level diversity is largely unexplored. Evolutionary theory predicts that mutualistic symbioses, based on reciprocal exchange of costly goods, should be destabilized by genetic diversity among closely-related symbiont strains. Yet, to date it is unknown if this prediction holds true in environmental symbioses. This is because the majority of bacteria has not been cultured under standard laboratory conditions, and due to the methodological challenges of teasing apart highly similar genomes from a complex bacterial community. In this thesis I therefore aimed to deepen our current understanding of mutualistic symbioses, and the potential role that symbiont strain diversity and evolution may play in it. To do so, I preformed high-resolution genomic analyses of the endosymbionts of *Bathymodiolus* mussels. These mussels dominate hydrothermal vents and cold seeps in the deep sea, and form a chemosynthetic symbiosis with gammaproteobacterial sulfur-oxidizing (SOX) or methane-oxidizing (MOX) symbionts, or both. The symbionts reside intracellularly in gill epithelial cells of the host and are horizontally transmitted between host generations.

The *Bathymodiolus* symbiosis is an ideal system to study intra-specific symbiont diversity, because of the small number of symbiont phylotypes that colonize each host individual. Their low community diversity allowed the detection of strain diversity within the symbiont populations, yet, standard binning approaches currently cannot resolve highly similar strain genomes from metagenomes. I therefore developed a custom workflow to resolve strain-level diversity in natural symbiont populations from high-resolution metagenomic sequencing data. My analyses revealed an extensive nucleotide diversity in the SOX symbionts, showing up to 11 single-nucleotide polymorphisms (SNPs) per kilobase pair (kbp; chapter II). Using this polymorphism data, up to 16 strains could be detected within single host individuals. To assess whether these strains also differed in their functional repertoires, I used read coverage information to identify regions of low coverage, representing differences in gene content among symbiont strains. Surprisingly, this analysis unveiled an extensive genomic plasticity among co-existing strains, affecting the energy metabolism, phage defense, lipopolysaccharide synthesis, phosphate uptake and regulation. Most of these functions were transcribed, indicating that the genotypic diversity also affected the symbiont phenotypes. The study included different *Bathymodiolus* species from a range of geographic locations, which made us conclude that high strain diversity of the SOX endosymbiont is pervasive among *Bathymodiolus* mussels.

The SOX symbionts are members of the SUP05 clade, a widespread bacterial group that impacts marine geochemical cycles of sulfur and nitrogen. Lineages of this clade have been reported to occur both as free-living cells, and as symbionts of mussels, clams and sponges. Whether the different lifestyles are associated with specific sets of genes and whether *Bathymodiolus* symbionts thus share unique genetic features distinct from the other lifestyles and lineages in the SUP05 clade remained unclear. An extensive comparative genomic analysis (chapter III) revealed that the link between lifestyle and gene content similarity was weak, indicating that convergent evolution has led to different genetic solutions with the same outcome (e.g. the colonization of *Bathymodiolus* mussels). Further, our analyses revealed that the SUP05 clade displays extensive gene content variation and metabolic plasticity on all phylogenetic levels, from genus to single strains, possibly driven by horizontal gene transfer. This suggests that evolvability, referring to the potential of a population to evolve adaptive solutions to unknown future conditions, might be a trait that is selected for in the SUP05 clade, contributing to the global success of this bacterial group.

The impact of natural selection on symbiont genome evolution is strongly affected by the type of association and transmission mode. In vertically transmitted symbionts the effective population size (N_e) is small, which diminishes selection pressure and results in symbiont genome evolution predominantly driven by genetic drift. In horizontally transmitted symbionts, such as the SOX symbionts, N_e is often unknown and the impact of genetic drift and natural selection is less obvious. Therefore, I used allele frequencies derived from polymorphism data to detect signatures of selection in the symbiont genomes (chapter IV). In order to gain a comprehensive picture of diversifying selection in the SOX symbiont, this study encompasses 88 metagenomes, spanning 15 chemosynthetic sampling sites from all over the world, and including 9 distinct host species. This analysis allowed the identification of genes affected by diversifying selection between symbiont populations, which included core traits, such as genes involved in sulfur oxidation. Our results therefore indicate that natural selection is effective in the SOX symbionts, and that environmental conditions and the interaction with the host may be important drivers in diversifying selection.

Analyses of allele frequencies in co-occurring host individuals further revealed a large symbiont population overlap (chapter II and IV), suggesting a continuous or extended re-sampling of symbionts either derived from a free-living population or from symbiont release by other host individuals. Intriguingly, the population overlap and degree of heterogeneity between the host species *B. brooksi* and *B. thermophilus* was smaller, possibly resulting from a reduced rate of symbiont exchange compared to the other investigated host species (chapter IV). Factors affecting symbiont exchange between hosts may be the age and density of host individuals.

The depth and high-resolution of the sequencing approach in this thesis together with the extensive sampling effort allowed me to uncover previously hidden strain diversity in the SOX symbionts of *Bathymodiolus*. My findings challenge and extent current evolutionary theories and point out the value of in-depth analysis of environmental bacterial communities to deepen our understanding of evolution, microbial interaction and symbiosis.

Zusammenfassung

Natürliche Bakterienpopulationen sind oft komplex und zeichnen sich durch große genetische Diversität aus. Die zugrunde liegenden Prozesse, die die Zusammensetzung mikrobieller Gemeinschaften bestimmen, sowie ihre räumliche Strukturierung, Koexistenz und Genomevolution, sind größtenteils unbekannt. Mittlerweile häufen sich Studien, die aufzeigen, dass sich auch nah verwandte Bakterienstämme stark in ihren Funktionen unterscheiden können. Inwiefern diese nah verwandten Arten dennoch koexistieren können, ist allerdings oft nicht bekannt. Vor allem in engen Symbiosen zwischen Bakterien und ihren Tierwirten sind die Auswirkungen von hoher Stammdiversität bisher kaum untersucht. Laut evolutionären Theorien ist genetische Diversität zwischen nah verwandten Symbiontenstämmen nachteilig, da sie innerhalb von mutualistischen Symbiosen, in der teure Güter zwischen den Partnern ausgetauscht werden, zu einer Destabilisierung der Gemeinschaft führen kann. Ob diese Theorie tatsächlich für natürliche Symbiosen zutrifft, ist allerdings noch nicht bestätigt. Das liegt vor allem daran, dass die meisten Bakterien bis heute nicht kultiviert wurden, und zudem an den methodischen Schwierigkeiten, die damit verbunden sind sehr ähnliche Genome aus einer komplexen Bakterienpopulation zu extrahieren. Das Ziel meiner Arbeit war es daher, ein tiefergehendes Verständnis für die Funktionsweise mutualistischer Symbiosen zu erlangen, und die potentielle Rolle, die die Diversität und Evolution der Symbiontenstämme darin spielt, zu verstehen. Alle Ergebnisse dieser Arbeit basieren auf hochauflösenden genomischen Analysen von Endosymbionten der *Bathymodiolus* Muscheln. Diese Muscheln leben in großen Zahlen an Hydrothermalquellen und kalten Quellen in der Tiefsee, wo sie chemosynthetische Symbiosen mit gammaproteobakteriellen Schwefeloxidierern (SOX) oder Methanoxidierern (MOX) bilden, oder beiden. Die Symbionten leben intrazellulär in den Kiemenepithelien der Wirte und werden horizontal von einer auf die nächste Wirtsgeneration übertragen.

Die *Bathymodiolus*-Symbiose ist besonders gut für die Untersuchung von Symbiontendiversität geeignet, da jeder Wirt nur eine kleine Anzahl an Symbionten-Phylotypen beherbergt. Die geringe Diversität ermöglichte daher, die Stammdiversität innerhalb der Symbiontenpopulationen mit hoher Auflösung zu untersuchen. Allerdings sind die erhältlichen Standardmethoden zur Extrahierung von Bakteriengenomen aus komplexen Datensätzen momentan noch nicht in der Lage zwischen sehr ähnlichen Stämmen zu unterscheiden. Daher habe ich im Rahmen dieser Arbeit einen methodischen Workflow entwickelt, der sich die hohe Auflösung von metagenomischer Sequenzierung zunutze macht, um die Stammdiversität natürlicher Symbiontenpopulationen aufzudecken. Mit Hilfe dieser Analyse wurde die starke Polymorphie der Nukleotidzusammensetzung innerhalb der SOX Symbiontenpopulationen deutlich, mit Werten von bis zu 11 Einzelnukleotid-Polymorphismen (SNP, engl. Single Nucleotide Polymorphism) pro Kilobasenpaar (kbp; Kapitel II). Anhand dieser Polymorphie konnten bis zu 16 Symbiontenstämme innerhalb eines einzelnen Wirtes identifiziert werden. Um zu verstehen, ob diese Stämme sich auch funktionell voneinander unterscheiden, habe ich in den Symbiontengenomen nach Regionen gesucht, die durch die Sequenzierung unterschiedlich häufig abgedeckt wurden und sich somit in der Sequenziertiefe unterschieden. Diese Regionen geben Hinweise auf Unterschiede in der Genzusammensetzung der verschiedenen Stämme. Erstaunlicherweise hat diese Analyse zu der Entdeckung von weitreichender Plastizität in den Genomen koexistierender Symbionenstämme geführt, in Bereichen, die für Energiestoffwechsel, Virusabwehr, Lipopolysaccharidsynthese, Phosphataufnahme und -regulation kodieren. Die meisten dieser Funktionen waren ebenfalls transkribiert, was darauf hindeutet, dass die genetische Diversität sich auch auf den symbiontischen Phenotyp auswirkt. Da diese Studie verschiedene *Bathymodiolus*-Arten umfasste, die an unterschiedlichen geographischen Standorten beprobt wurden, nehmen wir an, dass die hohe Stammdiversität der SOX Endosymbionten innerhalb der *Bathymodiolus*-Symbiose weitverbreitet ist.

Die SOX Symbionten sind Mitglieder der SUP05 Klade, einer weitverbreiteten Gruppe von Bakterien, die die geochemischen Stoffkreisläufe von Schwefel, Stickstoff und Sauerstoff

im Meer beeinflussen. Aufgrund früherer Studien ist bekannt, dass diese Klade sowohl freilebende Bakterienarten beinhaltet, als auch Symbionten von Muscheln und Schwämmen. Ob sich diese verschiedenen Lebensweisen auch in der genetischen Zusammensetzung der Arten widerspiegeln, und dementsprechend alle *Bathymodiolus* Symbionten einzigartige genetische Merkmale teilen, die in den anderen Lebensweisen und Arten der SUP05 Klade fehlen, blieb bisher unklar. Innerhalb einer umfassenden vergleichenden genetischen Untersuchung (Kapitel III) konnte ich in dieser Arbeit zeigen, dass eine ähnliche Lebensweise nur bedingt mit Ähnlichkeit in der Zusammensetzung von Genen in Zusammenhang stand. Dies deutet darauf hin, dass konvergente Evolution innerhalb der SUP05 Klade zu unterschiedlichen genetischen Lösungen bei gleichem Ausgang geführt hat (z. B. die Kolonisierung von *Bathymodiolus* Muscheln). Weiterhin haben die Ergebnisse unserer Analyse eine große Variation in der Genzusammensetzung und metabolischen Plastizität der SUP05 Klade aufgezeigt, die auf allen phylogenetischen Ebenen, von Genus bis zu einzelnen Stämmen, deutlich wurde, und potenziell auf lateralem Gentransfer beruht. Unsere Ergebnisse legen daher nahe, dass Evolvierbarkeit, also das Potential einer Population angepasste Lösungen für eine unbekannte Zukunft entwickeln zu können, ein Merkmal sein könnte, für welches in der SUP05 Klade selektiert wird, und was somit zum weltweiten Erfolg dieser Gruppe beiträgt. Dies ist fundiert auf den Ergebnissen unserer Analyse, welche eine weitreichende Variation in der Genzusammensetzung und metabolischen Plastizität auf allen phylogenetischen Ebenen, von Genus zu einzelnen Stämmen, aufgezeigt hat, potenziell getrieben durch lateralen Gentransfer.

Der Einfluss von natürlicher Selektion auf Genomevolution in Symbionten wird stark durch die Art der symbiotischen Gemeinschaft und die Art der Symbiontenübertragung beeinflusst. In vertikal übertragenen Symbionten ist die effektive Populationsgröße (N_e) in der Regel klein, was dazu führt, dass vor allem genetische Drift die Genomevolution der Symbionten antreibt. In horizontal übertragenen Symbionten dagegen ist N_e oft unbekannt und der Einfluss von genetischer Drift und natürlicher Selektion weniger offensichtlich. Um Signaturen von Selektion in den Symbiontengenomen dennoch erkennen zu können, habe ich Allelfrequenzen der Polymorphiedaten untersucht (Kapitel IV). Diese Studie umfasst Daten aus 88 Metagenomen, deren Proben an 15 chemosynthetische Standorten auf der ganzen Welt gesammelt wurden und 9 verschiedene Wirtstypen abdeckten, und ermöglichte somit einen umfangreichen Einblick in die diversifizierende Selektion in SOX Symbionten. Diese Analyse erlaubte die Identifizierung von Genen, die diversifizierender Selektion zwischen den Symbiontenpopulationen unterliegen, und unter anderem Gene für Kernfunktionen wie die Schwefeloxidation beinhalteten. Unsere Ergebnisse machen deutlich, dass natürliche Selektion effektiv ist, und dass Umweltbedingungen und die Interaktion mit dem Wirt wichtige Antriebskräfte für diversifizierende Selektion in SOX Symbionten zu sein scheinen.

Die Analyse von Allelfrequenzen in an einem Standort gemeinsam vorkommenden Wirten hat ergeben, dass sich ihre Symbiontenpopulationen weitgehend überschneiden (Kapitel II und IV). Diese Ergebnisse deuten daher darauf hin, dass Symbionten entweder kontinuierlich oder über einen langen Zeitraum immer wieder neu aufgenommen werden können, wobei unklar bleibt ob sie von einer freilebenden Population oder von freigesetzten Symbionten anderer Wirte stammen. Faszinierenderweise, haben wir geringe Symbiontenüberschneidung und niedrige Stammdiversität in den Wirtsarten *B. brooksi* und *B. thermophilus* entdeckt. Dies deutet auf eine niedrigere Möglichkeit des Symbiontenaustausches im Vergleich zu den anderen Wirtsarten hin (Kapitel IV), was möglicherweise im Alter oder der Dichte der Wirtsindividuen begründet ist.

Innerhalb dieser Arbeit habe ich eine bisher unerkannte Stammdiversität in den SOX Symbionten von *Bathymodiolus* Muscheln enthüllt. Dies war nur mit Hilfe eines umfangreichen Datensatzes tief-sequenzierter Metagenome möglich. Die präsentierten Ergebnisse hinterfragen und erweitern anerkannte evolutionäre Theorien und zeigen den Wert von tiefgehenden Analysen der natürlichen Bakteriengemeinschaften auf, um unser Verständnis von Evolution, mikrobieller Interaktion und Symbiose zu vertiefen.

Chapter I | Introduction

1.1 Symbiosis

“*Das fortwährende und innige Zusammenleben ungleichnamiger Organismen*”. This was how de Bary first defined the term ‘symbiosis’ in 1879 (de Bary, 1879). It translates to ‘the continuous and intimate living together of not-alike organisms’ and was introduced to describe lichens. Lichens were previously thought of as plants, but instead represent an intimate association between fungi and algae, a symbiosis. The word symbiosis is composed of two parts derived from Greek – *syn* which translates to ‘together’ and *bios* which translates to ‘life’. This first definition of ‘living together’ by de Bary is still valid today and forms the foundation for all symbiosis research.

Symbiosis describes three kinds of interactions. Two associated organisms can be either mutually beneficial (mutualism), neutral (commensalism) or one partner can negatively impact the other (parasitism). Although de Bary did not limit the concept of symbiosis to any of these types, symbiosis is often used as a synonym for mutualism (Martin and Schwab, 2012; Douglas, 2010). Nevertheless, the broader definition of symbiosis is important to describe intimate associations where the type of association is unknown or where it changes under different circumstances. One example where the type of assication can change is the human-associated bacterium *Neisseria meningitidis*. While normally, this bacterium resides in the human body as commensal it can become parasitic under specific conditions, such as in dense human populations, causing severe disease in the human brain (Soriani, 2017). Under the broad definition of symbiosis that includes mutualists, commensalists and parasites, *N. meningitidis* would thus be considered a human symbiont. Instead,

under the restricted definition of symbiosis that equals mutualism, one has to differentiate between the different lifeforms. In addition to the challenge that the type of association (beneficial, neutral or parasitic) it often not known, it is also not clearly defined how long two organisms have to be associated to form a symbiosis. The currently accepted description is that two organisms have to be associated for a significant proportion of their lifetime, which is vague in itself (Douglas, 2010). It is thus evident, that the continuum of associations among species in terms of time, contact, and reciprocal impact observed in nature still poses a challenge for scientists to grasp and define symbiotic associations.

Despite the controversy in definition, symbiosis has undoubtedly shaped life on earth as it is today. The entire eukaryotic domain of life would not exist today if it was not for symbiosis. It is through endosymbiosis that the eukaryotic life evolved, as mitochondria and plastids originated from bacteria (Margulis and Fester, 1991; Margulis, 1970). This has marked the starting point for an explosion in the evolution of multicellular lifeforms resulting in todays extensive diversity of eukaryotes. Eukaryotic organisms continue to evolve symbioses with microbial partners and today it is believed that the vast majority of multicellular eukaryotes, if not all, are symbiotic (Little, 2010). Symbioses between eukaryotes and microbes provide a valuable source of evolutionary innovation (Margulis and Fester, 1991; Moran and Telang, 1998). For example, through the association with a symbiotic partner, new traits can be gained, such as the access to novel metabolic capabilities, protection from antagonists or extended dispersal and mobility (Douglas, 2010). Symbiosis can also allow organisms to invade and adapt to new niches, which would not be possible without a symbiont (Little, 2010; Moran and Telang, 1998). One famous example is

the discovery, only 40 years ago, of chemosynthetic symbioses, a mutualistic association in the absence of light and photosynthetic primary production which will be discussed in detail as this is the subject of this thesis. A symbiosis between animal hosts and bacteria can have different complexities. For example, it can involve just two partners, such as the intimate association between *Euprymna scolopes* squids and the bioluminescent bacterium *Vibrio fischeri* (Visick and McFall-Ngai, 2000), few partners such as the gutless oligochaete *Olavius algarvensis* associated with less than ten bacterial species (Dubilier et al., 2008; Ruehland et al., 2008), or it can involve hundreds of bacterial partners leading to very complex microbiomes (e.g. Turnbaugh et al., 2007). The human microbiome has experienced a drastic increase in attention over the past years where the scientific focus on merely parasitic symbionts was extended to commensal and mutualistic ones. Throughout my study I will refer to the bacterial partner as the symbiont and to the animal partner as the host.

1.2 Chemosynthetic symbiosis

"*Isn't the deep ocean supposed to be like a desert?* [...] *Well there's all these animals down here.*" These are famous words from the geologist Jack Corliss on board of the submersible "Alvin" when hydrothermal vents were discovered in 1977. At that time there was no explanation as to how these extensive communities of animals could exist in complete darkness and far away from photosynthetic primary production in surface waters (Lonsdale, 1977). Without knowing it, the team in the Alvin, Jack Corliss, Tjeerd van Andel and Jack Donnelly were the first people that ever laid eyes on chemosynthetic symbioses – or so it seemed. In reality, chemosynthetic symbioses

in shallow-water sediments have been encountered unrecognized by scientists a long time before (Stewart and Cavanaugh, 2006; Cavanaugh, 1983; Reid and Bernard, 1980; Owen, 1961). However, it needed an observation such as the one in the deep sea, that lacked any explanation as to how animals can thrive there, to discover the beneficial association between chemosynthetic bacteria and animals (Cavanaugh, 1983; Cavanaugh et al., 1981). Chemosynthesis is a process in which chemical energy from e.g. reduced sulfur compounds, methane or hydrogen, is used to fix inorganic carbon (or methane) into biomass (reviewed in Dubilier et al., 2008; Jannasch and Mottl, 1985; Jannasch, 1985). So far only bacteria are described to be able to perform chemosynthesis which made the discovery of dense animal communities in the deep sea all the more surprising. Before the discovery of the hydrothermal vent ecosystem most life in the ocean was assumed to be sustained by photosynthetic primary production fueled by sun light. The deep ocean was considered a desert as very little organic material (~ 1%, reviewed in Jannasch, 1985) reaches the bottom to feed the deep-sea fauna. There is a wide diversity of chemosynthetic environments in the world’s oceans, most of which are now known to harbor chemosynthetic communities and symbioses. These communities are almost entirely based on bacterial chemosynthesis (**Box 1**).

Box 1 | Photosynthesis and chemosynthesis

Photosynthesis: The fixation of inorganic carbon into organic carbon with the use of sun light as energy source.

Chemosynthesis: The fixation of inorganic carbon (or methane) into organic carbon with the use of reduced chemicals (inorganic or organic) as energy source.

	Inorganic electron donor (e.g. $S_2O_3^{2-}$, H_2S, H_2, CO, etc.)	Organic electron donor (e.g. CH_4, sugars, etc.)
Inorganic carbon source (HCO_3^-, CO_2)	*Chemolithoautotroph*	*Chemoorganoautotroph*
Organic carbon source (e.g. CH_4, sugars, etc.)	*Chemolithoheterotroph*	*Chemoorganoheterotroph*

1.2.1 Diversity of chemosynthetic environments

Chemosynthetic habitats can be diverse, but share characteristic features (Smith, 2012). These include the presence of a chemical energy source of reduced compounds, such as sulfide, methane and hydrogen (Jannasch and Mottl, 1985). For chemosynthesis to be possible, also oxidants have to be present, such as oxygen and nitrate (Jannasch and Mottl, 1985). If a chemosynthetic environment is to sustain large animal communities, oxygen is required for the animals to respire, although the oxygen requirement for distinct animal groups can differ (Zhang and Cui, 2016; Sperling et al., 2015). Therefore, chemosynthetic communities are not entirely independent from sun light, as photosynthesis produces the oxygen needed in those habitats (Smith, 2012). There is a variety of different habitat types that were described to fuel chemosynthetic communities, such as hydrothermal vents, cold seeps, whale and wood falls, mud volcanoes, mangrove swamps and shallow-water sediments (**Fig. 1**, Dubilier et al., 2008). In my thesis I focused on chemosynthetic

symbioses that predominantly occur at hydrothermal vents and cold seeps, which I will describe below in some more detail.

Whale and wood falls represent temporal and local hot spots with reduced chemical compounds allowing the formation of chemosynthetic communities. When large whale carcasses sink to the ocean sea-floor they provide a rich resource for microbial degradation of the organic material. This produces a sulfide-rich environment that attracts diverse communities of organisms, including chemosynthetic and other symbioses, such as *Osedax* tubeworms (a heterotrophic symbiosis), vestimentiferan tubeworms, vesicomyid clams and polychaetes (Smith et al., 2015; Vrijenhoek, 2010). Similarly, wood falls and shipwrecks represent temporally restricted chemosynthetic environments, where the decomposing organic material produces sulfidic environments to fuel chemosynthesis, illustrating how human activities can impact the evolution of deep-sea symbioses. Such temporally restricted chemosynthetic environments have been considered ancient ‘stepping-stones’ in the evolution of chemosynthetic symbioses into deep-sea habitats (Distel et al., 2000). This is supported by recent findings of a giant mud-boring teredinid bivalve that evolved by replacing a heterotrophic gill symbiont with a chemoautotrophic symbiont (Distel et al., 2017).

As mentioned above, chemosynthetic symbioses have also been identified in shallow-water habitats. More specifically, the biological degradation of organic material in shallow-water sediments leads to oxygen depletion and sulfide enrichment. *Solemya* clams, *Codakia* clams, *Olavius* oligochaetes, *Paracatenula* flatworms and meiofaunal

Kentrophoros ciliates are just a few examples of the diversity of chemosynthetic symbioses found in shallow-water sediments (Dubilier et al., 2008).

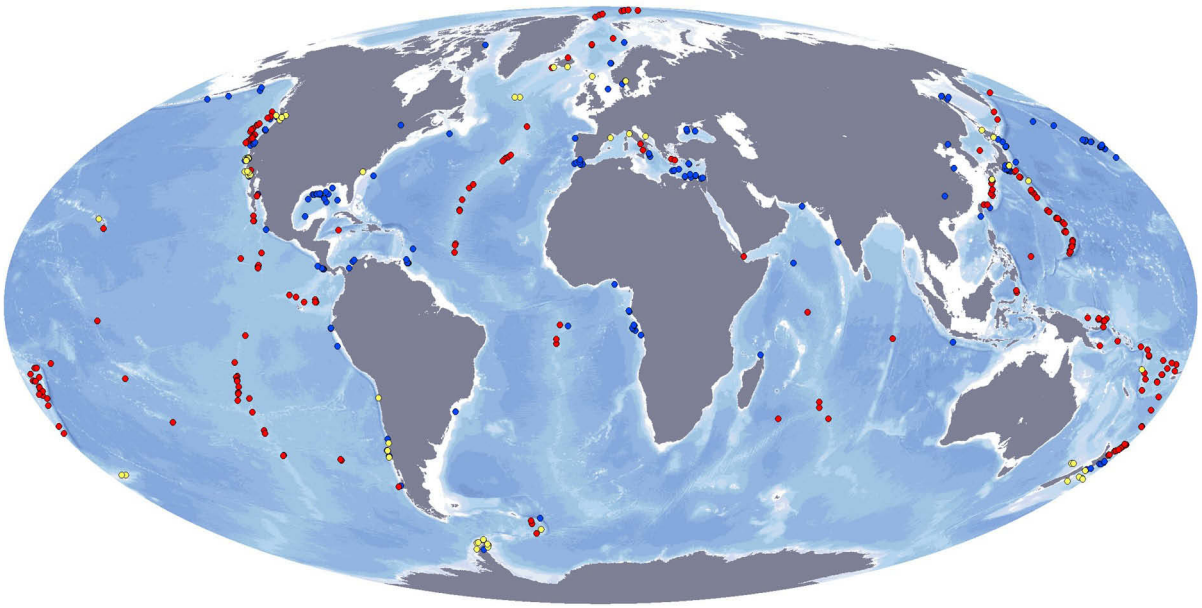


Figure 1 | "Global distribution of hydrothermal vent (red), cold seep (blue) and whale fall (yellow) sites that have been studied with respect to their fauna", adapted from *German et al., 2011*. This illustration only depicts sites that have been discovered and studied so far. The ongoing discovery of these habitats and their associated fauna might fill some of the gaps in the future.

Hydrothermal vents and cold seeps

Hydrothermal vents occur worldwide at spreading zones along the edges of the Earth's tectonic plates (**Fig. 1**). Seawater enters the porous ocean sea-floor and is heated up as it approaches magma chambers of molten rock below the sea floor. During its way through the crust, the seawater becomes enriched in dissolved gases (e.g. CO_2, H_2S, H_2, CH_4) and metals (e.g. Fe and Mn) from the volcanic ocean crust (Martin et al., 2008). These gases serve as energy sources fueling chemosynthetic life in those environments (Jannasch and Mottl, 1985). However, the concentration of dissolved compounds depends on a variety of factors such as the origin and composition of the oceanic crust, the original composition of the entering fluid and

the heat source itself (Tivey, 2007). Hot fluids can reach temperatures of more than 400 °C (Perner et al., 2014). When these fluids re-emerge from the sea-floor, they mix with the cold surrounding seawater. This mixing leads to mineral precipitation forming the typical ‘chimney’ structures of hot vents. In addition, the sudden precipitation of metal sulfides can cause the impression of black smoke being expelled by those chimneys, which is the reason that some of these vents are referred to as ‘black smokers’. The rate of continental spreading differs extremely with ultra-slow (< 20 mm $year^{-1}$), slow, intermediate, fast and ultra-fast (up to 150 mm $year^{-1}$) velocities (Snow and Edmonds, 2007; Charlou and Donval, 1993). The speed of continental spreading influences the geological origin of a venting system. Fast spreading zones are mostly influenced by volcanic activities (Kelley et al., 2002). Slow spreading instead is characterized by tectonic activities and can be more stable over time (Snow and Edmonds, 2007; Kelley et al., 2002). Ultra-slow spreading ridges and off-axis venting systems are fundamentally different from the other spreading systems in that they can be characterized by ultramafic rocks deriving from the Earth’s mantle, instead of basalts from the Earth’s crust (Jørgensen and Boetius, 2007; Snow and Edmonds, 2007). The geological composition of venting systems strongly influences the concentration of pH and reduced gases that are available for the vent fauna (Martin et al., 2008). For example vent fluids from basalt-hosted vents are typically enriched in sulfide, whereas those from ultramafic rocks are enriched in hydrogen (Amend et al., 2011; McCollom, 2007; Charlou et al., 2002).

In addition to hydrothermal vents, a second type of geologically-derived chemosynthetic environment was discovered in the Gulf of Mexico (Kennicutt et al.,

1985). Here, fluids with dissolved gases and energy sources to sustain chemosynthetic metabolism ‘seep’ upwards out of the ocean sediment. These ‘cold seeps’ are widespread on active and passive continental margins. In contrast to hydrothermal venting, the fluids are not characterized by hot temperatures. However, seeping pore-water is usually enriched in hydrocarbons, gases (e.g. CH_4, H_2S) that can fuel the energy demands of the diverse chemosynthetic communities in these habitats and also contains valuable nutrients (e.g. phosphate, ammonia) (Suess, 2010; Levin, 2005). There are different forces that pressure the waters to seep out of the sea-floor such as tectonic activities, disintegration of gas hydrates and sub-surface migration of salt (Cordes et al., 2010).

Vent and seep environments can be considered oases in the deep sea. Both habitats are often dominated by chemosynthetic symbioses such as vestimentiferan tubeworms, snails, *Rimicaris* shrimps, vesicomyid clams and bathymodiolin mussels (**Fig. 2**). These animals occur in high numbers and large biomasses, a productivity in the deep sea that would not be possible without living in chemosynthetic symbioses (Dubilier et al., 2008). For example, *Bathymodiolus azoricus* has been shown to reach up to 31 630 individuals and more than 3 kg of dry weight per m^2 at hydrothermal vents on the Mid Atlantic Ridge (Husson et al., 2017). *Riftia* tubeworms represent another example, which have been described as the fastest growing invertebrate known to date (Gaill et al., 1997; Lutz et al., 1994).

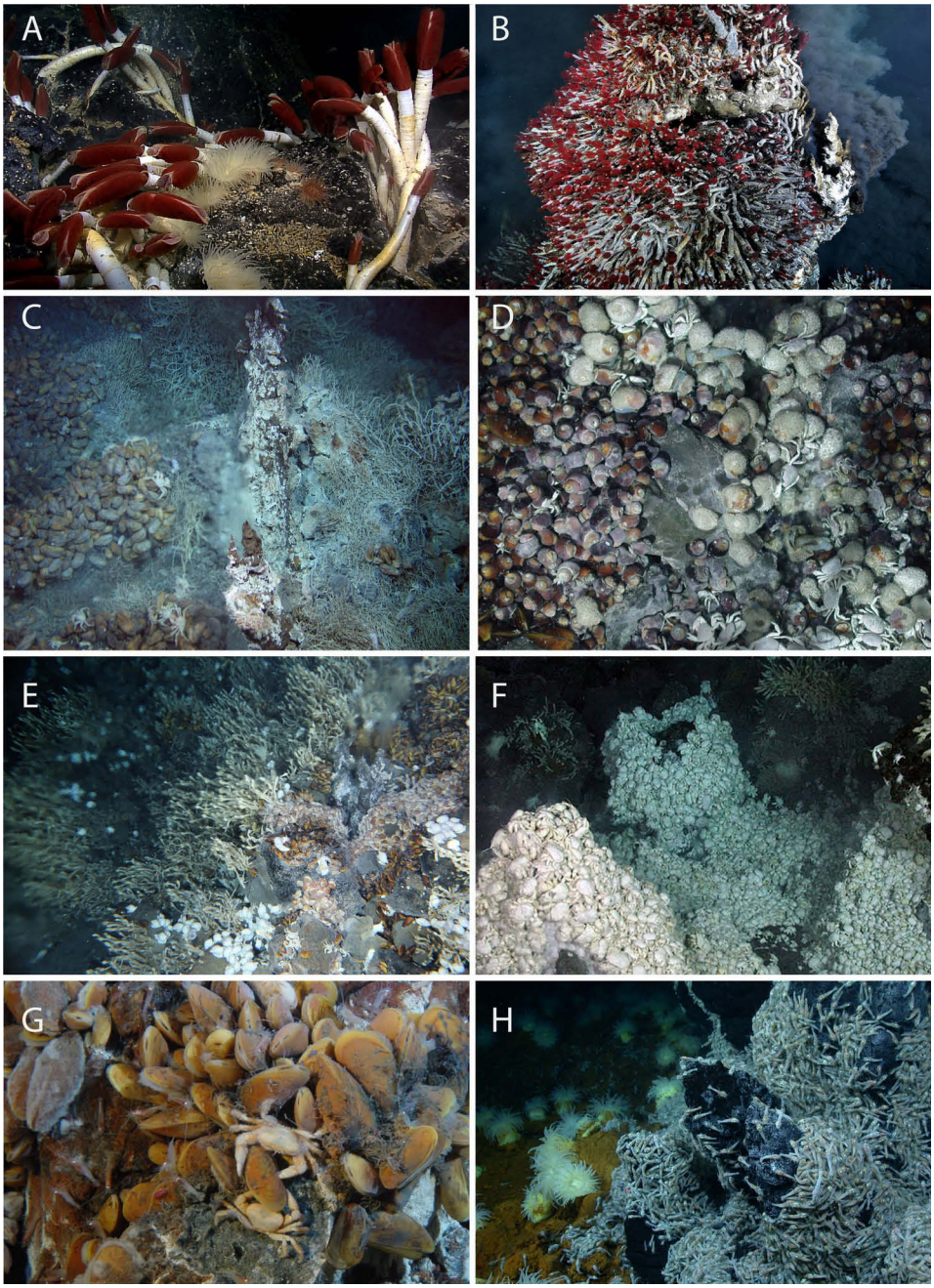


Figure 2 | Variation in the dominant, symbiont-hosting invertebrates at active hydrothermal vents. Eastern Pacific Ocean: (A) Tubeworms (*Riftia pachyptila*) with limpets and anemones; courtesy Wikipedia; B) Tubeworms (*Ridgeia piscesae*) with alvinellid polychaetes; courtesy Ocean Networks Canada. Western Pacific Ocean: (C) Mussels (*Bathymodiolus septemdierum*) and tubeworms (*Lamellibrachia* sp) with lithodid crabs; courtesy ROV Kiel 6000, GEOMAR; D) Hairy (*Alviniconcha* spp.) and black (*Ifremeria nautilei*) snails with bythograeid crabs (*Austinograea alaysae*); courtesy Woods Hole Oceanographic Institution. Indian Ocean: (E) Lepadid barnacles, scaly-foot snails (*Chrysomallon squamiferum*), mussels (*Bathymodiolus aff. brevior*); courtesy JAMSTEC. Southern Ocean: (F) Yeti squat lobster (*Kiwa tyleri*); courtesy NERC ChEsSo Consortium. Atlantic Ocean: (G) Mussels (*Bathymodiolus azoricus*) with bythograeid crabs; courtesy IFREMER. Caribbean Sea: (H) Swarming shrimp (*Rimicaris hybisae*); courtesy Woods Hole Oceanographic Institution. *Figure adapted from Van Dover et al., 2018.*

1.2.2 Diversity of chemosynthetic symbioses

Since the discovery of hydrothermal vents and chemosynthetic symbiosis in the 1970s, there has been almost half a century of continuous discoveries of new species. Just recently, a new chemosynthetic symbiosis was discovered in the giant teredinid bivalve (Distel et al., 2017). The association with bacteria that are able to fix inorganic carbon by using chemical energy has emerged multiple times over the course of evolution in animals as well as ciliates (Dubilier et al., 2008). Most chemosynthetic symbionts belong to the *Gammaproteobacteria*, but also *Epsilonproteobacteria* (Assié et al., 2016) and *Alphaproteobacteria* (Gruber-Vodicka et al., 2011) have been discovered in chemosynthetic symbioses. The first described chemosynthetic symbionts were sulfur oxidizers (SOX) of deep-sea tubeworms of the species *Riftia pachyptila* (Cavanaugh et al., 1981) and only shortly after, methane oxidizers were discovered in bathymodiolin mussels (Childress et al., 1986). In addition to sulfide and methane, hydrogen and carbon monoxide have been described to serve as energy substrate in chemosynthetic symbionts (Kleiner et al., 2015, 2012; Petersen et al., 2011).

Chemosynthetic symbioses are highly diverse in the types of host adaptations to the symbionts, the locations of bacterial symbionts within the hosts, and in the types of transmission modes how bacteria are transferred between host generations. The association with chemosynthetic symbionts has often caused drastic adaptations in host morphology. For example vestimentiferan tubeworms (e.g. *Riftia pachyptila*) don’t have a mouth or gut and instead harbor their sulfur-oxidizing symbionts in a specialized organ called the trophosome (Cavanaugh, 1983; Cavanaugh et al., 1981;

Felbeck et al., 1981). Despite the reduced gut, these animals can grow 2 m long and are among the fastest growing invertebrates known so far (Bright et al., 2013; Gaill et al., 1997). Bathymodiolin mussels represent another example as they have enlarged gills compared to their non-symbiotic relatives from mytilid bivalves (von Cosel and Olu, 1998). In addition, the hosts have evolved a variety of strategies to provide their chemosynthetic symbionts with both the energy substrate and oxidants required to perform chemosynthesis (Cavanaugh et al., 2013). For the symbionts this is assumed to represent an advantage of symbiotic life in comparison to free-living bacteria (Vrijenhoek, 2010). For a free-living chemoautotroph that uses oxygen as electron acceptor it can be challenging to get access to both substrates: reduced energy sources and oxygen, because these rarely co-occur in nature (Cavanaugh et al., 2013; Zhang and Millero, 1993). Usually, free-living bacteria such as filamentous *Beggiatoa* form biofilms in locations were reduced gases flow out of the ground and the overlaying seawater still has enough dissolved oxygen (Vrijenhoek, 2010). Instead a symbiont that is associated with an animal host may profit from its size, mobility and other adaptations to span the redox gradient (Stewart et al., 2005). For example, vesicomyid clams can span chemical gradients by their large size of up to 30 cm. In addition, these bivalves can extend their foot to reach sulfide-rich conditions, but they can simultaneously filter oxygen-rich seawater, thus providing their symbionts with both substrates (Vrijenhoek, 2010). Mobility between reduced and oxic conditions has been suggested to be a strategy in meiofaunal organisms with chemosynthetic symbionts (Giere et al., 1991; Ott, 1989). There are also adaptations in host proteins, as described for *Riftia* tubeworms. In these hosts, hemoglobins that bind oxygen and sulfide are transported via the blood circulation to

the symbiont-harboring organ, the trophosome (Arp et al., 1987; Arp and Childress, 1983).

Chemosynthetic symbionts can be associated to their host in different ways: i) extracellular on the host surface, also referred to as ectosymbionts or epibionts (e.g. in *Rimicaris* shrimps, *Leptonemella* nematodes), ii) extracellular but inside the host (e.g. *Olavius* worms) iii) intracellular in host cells (e.g. *Bathymodiolus* mussels, *Riftia* tubeworms, *Vesicomyosocious* clams). Both, ii) and iii) are referred to as endosymbionts, due to their location within the host. It is important to note that an endosymbiotic relationship is not necessarily more specific than an ectosymbiotic association, as it has been shown for chemoautotrophic ectosymbionts colonizing the cuticle of *Leptonemella* nematodes that can be highly specific (e.g. Zimmermann et al., 2016).

1.3 Symbiont transmission

There are two routes of symbiont transmission from one host generation to the next: vertical and horizontal transmission. During vertical transmission, symbionts are directly passed from the parent to the offspring. In most described cases this happens through the female germline via the egg. But symbionts can also be vertically transmitted in other ways, for example via feeding or oral smearing (Bright and Bulgheresi, 2010). A vertical transmission mode through the germline has been described for vesicomyid clams (Cary and Giovannoni, 1993; Endow and Ohta, 1990). The host benefits from vertically transmitted symbionts, as it provides a secure way to equip all offspring with a symbiont that is needed to survive e.g. in

chemosynthetic environments. On the other hand, the transmitted strain might not be the best adapted to new environmental conditions and therefore can hamper the hosts ability to disperse and colonize new locations or habitats (Vrijenhoek, 2010). Horizontal transmission, instead, can happen either laterally from other hosts (not only parental individuals) or through a free-living stage of the symbiont (Bright and Bulgheresi, 2010). The latter has been described for vestimentiferan tubeworms (Nussbaumer et al., 2006). The successful association between the symbiont and its host is determined by a variety of factors that include recognition mechanisms and the right timing. Some hosts only have a short developmental stage or time window during which they are permissive to symbiont colonization. This is for example the case in squid juveniles and vestimentiferan tubeworms where larvae can only be infected during a short time after settlement (Nussbaumer et al., 2006). Horizontal transmission allows the host to acquire the best-adapted symbiont in the local conditions (Won et al., 2003). However, it also poses a risk to the association in finding its beneficial partner and in the possibility of being invaded by parasitic strains (Vrijenhoek, 2010). In addition, genotype heterogeneity resulting from horizontal transmission might lead to the emergence of ‘cheater’ strains (see section **1.5.1**).

There are also mixed modes of vertical and horizontal transmission. Here, predominantly vertically transmitted symbiont populations can be supplemented by symbionts that are taken up from the environment or from other hosts. Evidence for a mixed transmission mode has been recently shown for solemyid and vesicomyid clams that were previously thought of as vertically transmitted (Ozawa et al., 2017; Russell and Cavanaugh, 2017; Russell et al., 2017). In fact, strict vertical

transmission with the complete absence of horizontal transmission events may be rare (Bright and Bulgheresi, 2010).

1.3.1 Impact of transmission mode on symbiont genome evolution

The transmission mode of symbionts can have manifold influences on the genome evolution in symbionts (Wernegreen, 2015; Bright and Bulgheresi, 2010; Moran, 1996). Vertical transmission in long-term relationships imposes a physical bottleneck that allows a limited number of symbiont cells to be transmitted from one host generation to the next (Mira and Moran, 2002). This has profound effects on the so-called effective symbiont population size (N_e). N_e can be defined as the "[...] *size of a population evolving in the absence of selection that would generate as much neutral diversity as is actually observed.*" (Fraser et al., 2009). Albeit it is extremely difficult to estimate N_e for bacteria, it is clear that bottlenecking drastically reduces it (Bobay and Ochman, 2018). Small N_e decreases the effect of natural selection and instead makes a population more susceptible to genetic drift. This leads to the accumulation of slightly deleterious mutations that are not purged by purifying selection, a process that has also been referred to as Muller's ratchet (Moran, 1996). Evolution under genetic drift can result in a reduction of genome size due the loss of gene functions as has been shown for many obligate insect symbionts (Moran and Bennett, 2014). This process has to be distinguished from *adaptive* genome reduction that can also lead to the evolution of smaller genome sizes in bacterial populations with large effective genome sizes (Wernegreen, 2015). In vertically transmitted endosymbionts such genome reduction can be so extreme that it ultimately leads to the loss of essential symbiont functions, a process that has been described as the 'evolutionary

rabbit hole' (Bennett and Moran, 2015). In addition to N_e, vertically transmitted endosymbionts are genetically isolated from other bacterial populations and thus from any source of "new" genetic material that would allow recombination. In contrast, horizontally transmitted symbionts have potentially much larger N_e. If N_e is large, natural selection can be effective. This results in removal of slightly deleterious mutations via purifying selection (Wernegreen, 2015). Adaptive alleles can be selected for and reach fixation in the affected population. However, as described above, also in horizontally transmitted symbionts there can be population bottleneck effects. N_e is largely influenced by how much the symbiont population in co-occurring hosts and the environmental population contributes to the colonization of the next host generation (Vrijenhoek, 2010). In addition, the opportunities for horizontal gene transfer and recombination with environmental populations depend on whether the environmental symbiont stage is active or dormant towards these processes (Vrijenhoek, 2010). Therefore, our abilities to predict the impact of selection and genetic drift on horizontally transmitted symbionts are limited.

The impact of transmission modes on the evolution of organisms often becomes visible in the phylogeny of host and symbiont. The phylogenies of strictly vertically transmitted symbionts show congruent phylogenies with those of their hosts. For example, such coupling was shown in chemosynthetic vesicomyid clams from deep-sea environments and *Paracatenula* flatworms from shallow-water sediments (Gruber-Vodicka et al., 2011; Goffredi et al., 2003; Hurtado et al., 2003). Points of incongruence in otherwise congruent phylogenies can potentially reveal occasional horizontal transmission, as well as host switching events over the course of evolution (Ozawa et al., 2017; Stewart et al., 2008). In contrast, there is typically only little

congruence between host and symbiont phylogenies in horizontally transmitted symbionts, such as between *Bathymodiolus* mussels and their symbionts (Bright and Bulgheresi, 2010; Won et al., 2008).

1.3.2 Impact of transmission mode on symbiont heterogeneity

The mode of symbiont transmission greatly influences the genetic heterogeneity of symbionts within single hosts. Vertical transmission is expected to lead to homogeneity within hosts, due to the strong bottleneck effects and reduced opportunities for gene exchange and recombination. In horizontal transmission, the heterogeneity of symbionts within a host depends on the heterogeneity in the potentially infectious population and on the number of cells that can colonize a host. The latter is influenced by the time of the permissive period. For example, squids and vestimentiferan tubeworms only have a short time period where symbionts can colonize, and in line with this, they show limited symbiont diversity (Nussbaumer et al., 2006; Nyholm and McFall-Ngai, 2004). In contrast, a host that can continuously acquire symbionts throughout an extended period or its entire lifetime, potentially a multitude of infectious strains could colonize these hosts, presuming that they occur in the environment. This has been hypothesized for bathymodiolin mussels (Wentrup et al., 2014).

1.4 Bathymodiolin symbiosis

Symbiotic deep-sea mussels in the family Mytilidae form a monophyletic subfamily called Bathymodiolinae (**Fig. 3, 4**). These mussels dominate many hydrothermal vents and cold seeps and have also been observed at sunken wood and whale falls (**Fig. 4**; Distel et al., 2000; Duperron et al., 2013; Van Dover et al., 2002). These deep-sea mussels are thought to have evolved approx. 89 million years ago from their shallow-water relatives, and invaded the deep-sea via stepping-stones such as wood and whale falls (Liu et al., 2018; Lorion et al., 2013; Samadi et al., 2007; Distel et al., 2000). Compared to the other lineages within the Mytilidae family, Bathymodiolinae form a young clade and have undergone rapid adaptive divergence (Liu et al., 2018; Lorion et al., 2013). Although the phylogeny within this group is an ongoing debate, most recent analyses have suggested nine genera within the Bathymodiolinae, namely *Bathymodiolus*, *Benthomodiolus*, *Vulcanidas*, *Lignomodiolus*, *Idas*, *Terua*, *"Bathymdiolus"*, *Gigantidas* and *Nypamodiolus* (Liu et al., 2018).

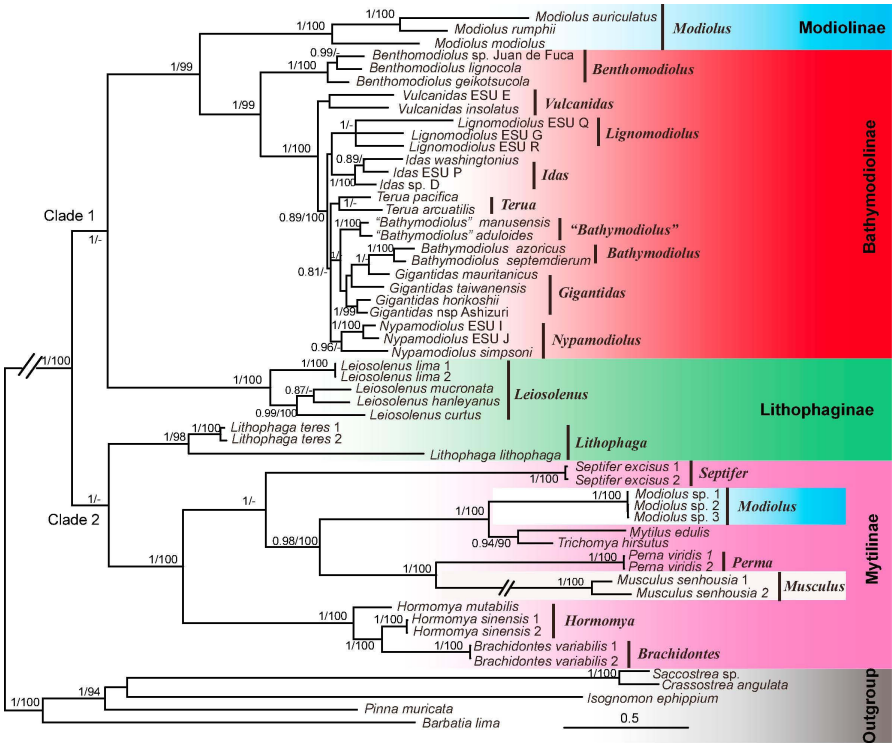


Figure 3 | "Phylogenetic tree of marine mussels based on the Bayesian analysis of a combined dataset of COI, 16S rRNA, 18S rRNA, 28S rRNA and H3 genes. Values of Bayesian posterior probability (greater than 0.8)/maximum likelihood bootstrap (greater than 90%) are shown above the branches." *Taken from Liu et al., 2018.*

Despite a functioning digestive system, deep-sea bathymodiolin mussels depend on their chemosynthetic bacterial symbionts for nutrition (Duperron, 2010; DeChaine and Cavanaugh, 2005; Stewart et al., 2005). *Bathymodiolus* mussels were shown to be associated with a SOX, MOX or both symbiont types (**Fig. 4**). SOX symbionts are associated with most species and only absent in a few mussel species from the *"B." childressi* clade (Duperron, 2010; Lorion et al., 2013). Both symbionts have been suggested to be horizontally transmitted from one host generation to the next (Wentrup et al., 2014; Won et al., 2003). Most of the symbionts reside intracellularly in gill epithelial cells called bacteriocytes but symbionts have also been described to occur extracellularly in these mussels (Duperron, 2010). Multiple studies have shown

the transfer of fixed or acquired carbon from the symbionts to the host (Riou et al., 2008; Nelson et al., 1995; Fisher and Childress, 1992; Childress et al., 1986). However, the details in the mode of this transfer are not fully understood yet. Most evidence suggests that intracellular digestion, as well as ‘leaking’ from intact symbiont cells to the host both play a role (Kádár et al., 2008; Fiala-Médioni et al., 2002; Streams et al., 1997). In addition to nutrition, the symbionts may be beneficial for the host in sulfide detoxification (Powell and Somero, 1986), supply with amino acids (Ponnudurai et al., 2017) or potential defense against parasites (Sayavedra et al., 2015).

Apart from the primary gammaproteobacterial SOX and MOX symbionts, other symbiont types have been described in bathymodiolin hosts. For example, a gammaproteobacterial *Cycloclasticus* symbiont (Rubin-Blum et al., 2017; Raggi et al., 2013) and extracellular epsilonproteobacterial symbionts (Assié et al., 2016) were described for some host species. Also parasitic symbiont types occur in bathymodiolin mussels, invading host nuclei of gill epithelial cells, that are not colonized by beneficial symbionts (Zielinski et al., 2009). All mussels species that are investigated in this thesis belong to the *Bathymodiolus* genus and have either both, SOX and MOX or only SOX as primary symbionts (**Fig. 4**).

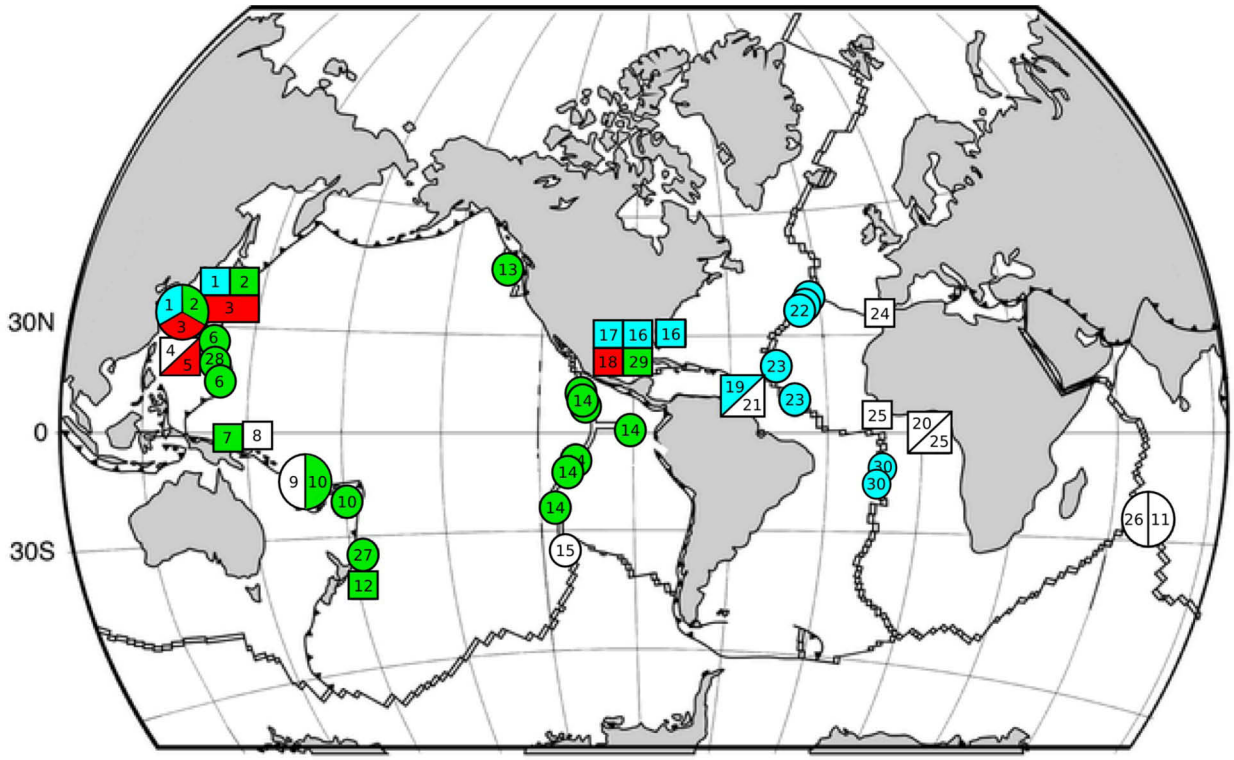


Figure 4 | Distribution of described bathymodiolin species and their association with only SOX (green), only MOX (red) or both symbionts (blue). Species included in this thesis are marked with a star in the legend. *B. = Bathymodiolus, "B." = "Bathymodiolus", I. = Idas, T. = Tamu. Modified after Génio et al., 2008.*

1.4.1 Metabolism of SOX and MOX symbionts

Both types of symbionts in *Bathymodiolus* mussels are housed in the gills where they are in close proximity to access seawater that is enriched in energy substrates and oxidants (Distel et al., 1995). The common energy sources of all bathymodiolin SOX symbionts are reduced sulfur compounds (Kleiner et al., 2012). The oxidation of sulfide and thiosulfate is coupled to the fixation of inorganic carbon as has been shown experimentally in *B. azoricus* and *B. thermophilus* (**Fig. 5**; Riou et al., 2008;

Nelson et al., 1995). In SOX symbionts the autotrophic fixation of carbon dioxide into biomass occurs through the Calvin-Benson-Bassham cycle (CBB) (Cavanaugh and Robinson, 1996). Besides sulfur other energy sources have been shown to serve as electron donors for the SOX symbionts. The use of hydrogen as an electron donor has been observed experimentally and the symbiont's potential for the oxidation is encoded in the genomes of SOX symbionts of *Bathymodiolus* species from hydrothermal vent sites (Ikuta et al., 2016; Petersen et al., 2011). In addition, Sayavedra (2016) discovered genes encoding the capability to use methanol as an alternative energy source in the genomes of the SOX symbionts and hypothesized that the methanol could originate from the co-occurring MOX symbiont. This may result in a syntrophic relationship between the two symbionts which has been suggested previously for other pathways (**Fig. 5**; Ponnudurai et al., 2017). The SOX symbionts do not encode a complete tricarboxylic acid (TCA) cycle as its genome does not contain the genes encoding a malate dehydrogenase and a succinate dehydrogenase. An incomplete TCA cycle led to the hypothesis of a tight coupling between host and symbiont metabolism as the SOX symbiont presumably relies on the host to provide key intermediates it cannot produce (Ponnudurai et al., 2017). However, also in *Ca.* Thioglobus autotrophicus, a free-living relative of the SOX symbiont (see section **1.4.2**), the malate dehydrogenase is not encoded in its genome, thus, it remains uncertain how metabolically dependent the SOX symbiont is on the host (Ponnudurai et al., 2017).

This thesis focuses on the SOX symbiont of *Bathymodiolus* mussels and the key metabolic characteristics of the MOX symbiont are only briefly outlined below. The association of metazoans in chemosynthetic environments with a MOX symbiont is by

far not as common as the association with a SOX symbiont (Duperron, 2010). The presence of a MOX symbiont has been detected in at least ten described *Bathymodiolus* species from vent and seep sites, most of which also have SOX symbionts (Lorion et al., 2013). Experimental and genomic evidence suggests that the MOX symbiont can aerobically oxidize C1 compounds such as methane and methanol (Pimenov et al., 2002; Robinson et al., 1998; Fisher et al., 1987). In addition, the MOX symbiont encodes and expresses the complete ribulose monophosphate (RuMP) pathway for the assimilation of C1 compounds, such as formaldehyde (Ponnudurai et al., 2017). In contrast to the SOX symbiont, the MOX symbiont encodes a complete TCA cycle (Ponnudurai et al., 2017).

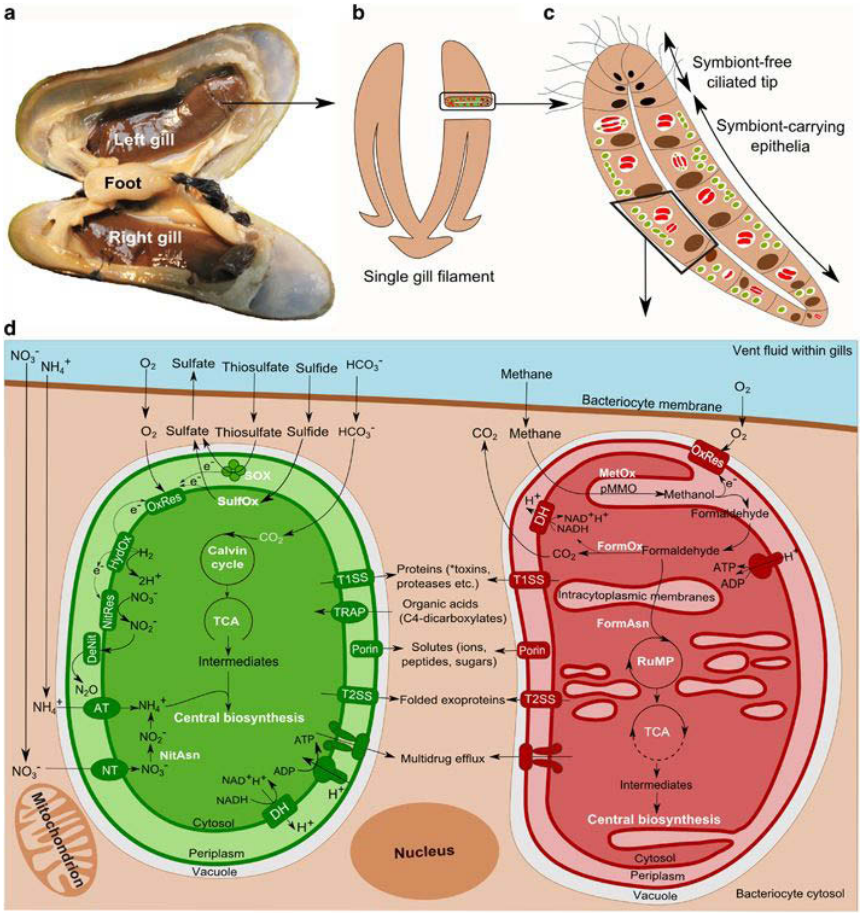


Figure 5 | "Schema of a *B. azoricus* gill bacteriocyte. (a) Dissected *B. azoricus* specimen showing gills and gill filaments. (b) Schema of single gill filament. (c) Schematic cross-section of a gill filament showing bacteriocytes. (d) A single bacteriocyte showing the central pathways of the symbionts. Symbionts (SOX: green, MOX: red) are located inside vacuoles (white) surrounded by the bacteriocyte cytosol (brown) with gases and substrates exchanged between vent fluids (blue) flushing the gills. An overview of the basic metabolic processes occurring in the symbionts is shown. AT: ammonium transporter, DH: dehydrogenase, DeNit: denitrification, FormAsn: formaldehyde assimilation, FormOx: formaldehyde oxidation to formate and CO2, HydOx: hydrogen oxidation, MetOx: methane oxidation by pMMO (particulate methane-monooxygenase enzyme complex) to methanol and then to formaldehyde, NitAsn: nitrogen assimilation, NitRes: nitrate respiration, NT: nitrate transporter, OxRes: Oxidative phosphorylation with oxygen as terminal electron acceptor, SOX: thiosulfate oxidation, SulfOx: sulfide oxidation via the rDSR-APS-Sat pathway, TCA: tricarboxylic acid cycle, TRAP: tripartite ATP-independent periplasmic transporter, T1SS: Type I secretion system, T2SS: Type II secretion system. *toxins are known to be secreted by the thiotrophs (Sayavedra et al., 2015)." *Adapted from Ponnudurai et al. 2017, the legend was modified to fit names in main text.*

1.4.2 The SOX symbiont relatives from the SUP05 clade

This thesis mainly focuses on the SOX symbiont of bathymodiolin mussels. This symbiont falls into the so-called SUP05 clade of *Gammaproteobacteria*, which, together with its sister clade Arctic96BD, was recently reclassified by the Genome Taxonomy Database as belonging to the *Thioglobaceae* family within the order of Thiomicrospirales (Parks et al., 2018). In addition to bathymodiolin symbionts, this bacterial family encompasses other symbiotic and free-living lineages of gammaproteobacterial sulfur oxidizers (GSO). The phylogeny of free-living and symbiotic lineages is interspersed which suggests that within this family the lifestyle has been switched multiple times over the course of evolution either from free-living to symbiotic, or from symbiotic to free-living, or both (Sayavedra, 2016; Petersen et al., 2012).

Free-living relatives of Bathymodiolus SOX symbionts

The free-living lineages within the *Thioglobaceae* have been found in a broad range of marine habitats, particularly oxygen minimum zones (OMZs), anoxic marine zones (AMZs) and hydrothermal vents (Callbeck et al., 2018; Meier et al., 2017; Swan et al., 2011; Walsh et al., 2009; Sunamura et al., 2004). Evidence of high cell abundances and metabolic activity shaped the view that free-living SUP05/ArcticBD96 lineages have an important impact on biogeochemical cycles of sulfur and nitrogen in the ocean (Callbeck et al., 2018; Murillo et al., 2014; Stewart et al., 2012; Ulloa et al., 2012). Unlike the symbiotic lineages, two strains of the *Thioglobaceae* have been successfully isolated and their genomes were sequenced;

Ca. Thioglobus autotrophicus EF1 from the SUP05 clade and *Ca*. Thioglobus singularis PS1 from the Arctic96BD clade (Shah et al., 2017; Marshall and Morris, 2015; Shah and Morris, 2015; Marshall and Morris, 2013). Another SUP05-lineage *Ca.* Thioglobus perditus was recently described, though not cultivated (Callbeck et al., 2018). In addition, recent efforts in metagenomic and single-cell sequencing of natural bacterial populations have revealed a stunning metabolic versatility among the free-living SUP05/Arctic96BD lineages (Meier et al., 2017; Murillo et al., 2014; Roux et al., 2014; Anantharaman et al., 2013) which is briefly summarized below.

All known lineages of the family *Thioglobaceae* have the potential to oxidize reduced sulfur compounds. Beyond this commonality, there is a great plasticity in metabolic potential among the lineages. Some lineages of the SUP05 clade were suggested to be obligate chemolithoautotrophs (Shah et al., 2017), whereas other SUP05/Arctic96BD lineages might be heterotrophs or mixotrophs, capable of both a autotrophic and heterotrophic metabolism (Marshall and Morris, 2013; Murillo et al., 2014; Swan et al., 2011). Although it has been first postulated that SUP05-lineages were obligate anaerobes (Walsh et al., 2009), experimental and genomic evidence on SUP05/Arctic96BD lineages suggested a facultative lifestyle with both nitrate and oxygen as potential electron acceptors (Shah et al., 2017; Anantharaman et al., 2013). SUP05/Arctic96BD bacteria also differ substantially in their genetic potential to reduce nitrate nitrogen gas (N_2) (Anantharaman et al., 2013). The genetic set-up for sulfur oxidation was more consistent within each clade but clearly different between the SUP05 and Arctic96BD clades (Murillo et al., 2014). Interestingly, Arctic96BD lack the two common sulfur oxidation pathways, *sox* for the oxidation of thiosulfate, and the reverse *dsr* for the oxidation of sulfite (Murillo et al., 2014).

However, the cultivated Arctic96BD lineage showed enhanced heterotrophic growth in the presence of thiosulfate (Marshall and Morris, 2013). There is also metabolic plasticity among the free-living lineages in their possible energy sources. Like bathymodiolin SOX symbionts, also free-living SUP05 bacteria from hydrothermal vents encoded the genetic potential for the oxidation of hydrogen (Anantharaman et al., 2013). Generally, the bacterial family *Thioglobaceae* appears to be characterized by a metabolic plasticity affecting the sulfur, nitrogen, and carbon metabolism and the cycling of these in the environment.

Symbiotic relatives of Bathymodiolus SOX symbionts

In addition to bathymodiolin symbionts, the SUP05 clade also includes symbionts of other host families (**Fig. 6**). In fact, the first known genomes from the SUP05 clade were endosymbionts of deep-sea vesicomyid clams, *Ca*. Ruthia magnifica and *Ca*. Vesicomyosocious okutanii (Kuwahara et al., 2008, 2007; Newton et al., 2007). In contrast to the phylogeny of their symbionts, vesicomyid clams are phylogenetically very distant from bathymodiolin mussels (Taylor and Glover, 2010). Similar to bathymodiolin symbionts, vesicomyid symbionts are housed within gill epithelial cells. However, the symbiont transmission mode is vertical with occasional horizontal transmission events and thus different from horizontally transmitted bathymodiolin symbionts (Stewart et al., 2008; Cary and Giovannoni, 1993; Endow and Ohta, 1990). In line with this, the symbiont genomes show signatures of ongoing genome reduction as a result of genetic drift (Kuwahara et al., 2011, 2008; and see **1.3**). SUP05-lineages also have been identified as symbionts of deep-sea sponges (Rubin-

Blum et al., 2019; Nishijima et al., 2010). However, the transmission mode and cellular location in the sponge tissue remain to be shown.

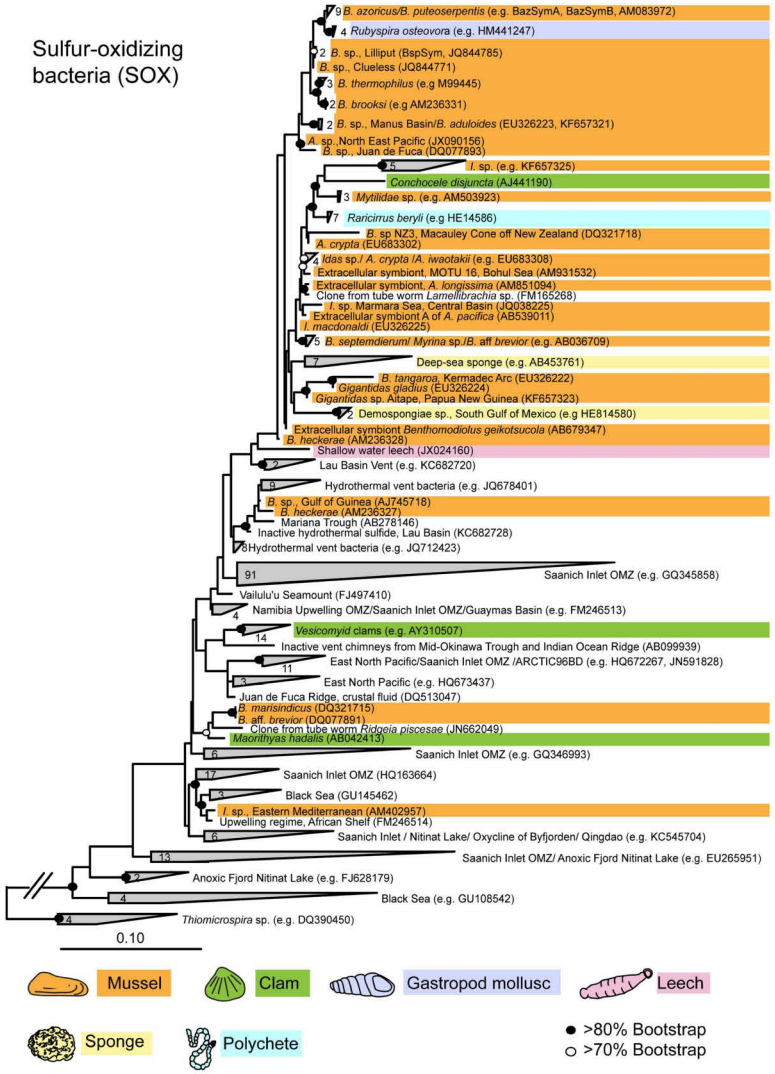


Figure 6 | "Maximum likelihood 16S rRNA phylogeny of the close relatives of the *Bathymodiolus* SOX symbionts. The tree was estimated from an alignment of 1653 nucleotide positions and was rooted with four sequences from *Thiomicrospira* species. The number of sequences per collapsed group is shown next 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*." *Taken from Sayavedra et al., (2015).*

1.4.3 Microdiversity in *Bathymodiolus* SOX symbionts

Usually, each bathymodiolin host species is associated with one 16S rRNA phylotype of SOX or MOX symbiont, or both (Duperron, 2010). Exceptions occur for example in the two host species *B. azoricus* and *B. puteoserpentis* from hydrothermal vents at the Mid Atlantic Ridge which share an identical SOX and MOX symbiont phylotype (Duperron et al., 2006). On the other hand, a single bathymodiolin species can harbor two distinct SOX phylotypes, as shown in *B. heckerae* and *I.* sp. MED (Duperron et al., 2008, 2007). How these two SOX phyloypes co-exist and whether there is competition between them is not known. Whereas these phylotypes appeared to be physically separated in *B. heckerae*, mutual exclusion was not observed between the phylotypes in *I.* sp. MED (Duperron et al., 2008, 2007). This raises the question as to how diverse are the symbionts within and between host species and individuals. Investigations of the 16S rRNA gene and the internal transcribed spacer region (ITS), that has a higher resolution than the 16S rRNA gene, revealed the co-occurrence of distinct 16S and ITS types in the SOX symbiont (Duperron et al., 2008, 2007; Won et al., 2003). This indicated that there is a level of microdiversity that is not captured with the common marker gene encoding the 16S rRNA. Applying recent sequencing technologies, Ikuta et al. (2016) confirmed this, and were able to reveal metabolic diversity among co-existing symbiont strains of a single SOX symbiont species in *B. septemdierum*.

The presence of multiple endosymbiont species and closely related strains of the same species within the same host individual poses the question of how these can stably co-exist. The role of competition or, perhaps, cooperation is often not known in

natural communities where the symbiotic partners cannot be cultured. In my thesis I aim to shed light on the strain diversity in the bathymodiolin endosymbiosis and to understand possible implications for the symbiotic association.

1.5 Strain diversity and endosymbiosis

1.5.1 Diversity in mutualism

Theoretical models state that the evolution of mutualism is influenced by three key factors: i) high benefit to cost ratio, (ii) high within-species relatedness and (iii) high between-species fidelity (Foster and Wenseleers, 2006). Instead, the persistence of mutualistic associations over evolutionary time has long challenged evolutionary theory (Frederickson, 2013). The diversity of symbionts, more specifically intra-specific diversity among closely related strains of the same bacterial species can strongly impact the stability of mutualistic associations as described below. Here I refer to ‘strain’, when bacterial organisms belong to the same species.

Most mutualistic associations are based on the exchange of benefits and the associated costs (Bronstein, 2015). This can cause conflicts and lead to the emergence of so-called cheaters. These are predicted to destabilize symbiotic relationships that involve reciprocal exchange of costly resources. A ‘cheater’ strain belongs to the same species as the cooperative strain, but provides no or less resources to their symbiotic partners while still receiving the full benefit (Douglas, 2008). In comparison to the cooperating strain, the cheater has a fitness advantage

and therefore is expected to outcompete the other. This leads to a decrease in host fitness and thus to the destabilization of the mutualistic relationship (Douglas, 2008). Evolutionary theory predicts that mutualists should favor the association with high-quality partners, and consequently select for a reduced symbiont diversity within hosts (Poisot et al., 2011; Thrall et al., 2007; Thompson, 2005; Frank, 1996; Bull and Rice, 1991). There are different mechanisms that can ensure the association with only few and/or high-quality partners. First, there is *partner fidelity feedback*. This means that the future symbiont fitness depends directly on current investment (Sachs et al., 2004), such as in the vertical transmission mode (Yamamura, 1996). On the other hand, there is *partner control*, where one partner can discriminate between multiple other partners. Here, partner choice or a screening of the partner may lead to filter out the cooperating symbiont type (Archetti et al., 2011; Sachs et al., 2004; Bull and Rice, 1991). One example of partner choice can be found in the *Vibrio*-squid symbiosis that has evolved a highly selective mechanism for the beneficial symbiont strain (Nyholm and McFall-Ngai, 2004; Visick and McFall-Ngai, 2000). In addition, partner control can also be exercised after the establishment of the symbiosis by sanctioning. This has been described for the *Rhizobium*-legume symbiosis where non-cooperating strains are punished by e.g. lowering the amount of rewards (Westhoek et al., 2017). The incentive to cheat is highest when the costs associated with the symbionts are high. However, not all associations, where both partners benefit, invoke a cost. For example, in the *Blattabacterium*-cockroach symbiosis, the toxic waste-products of the animal serve as a nitrogen source for the bacterial symbiont. Therefore, both partners benefit from the nitrogen transfer but do not have to pay a cost, although there might be undetected costs e.g. in the evolution of adaptations to the symbiosis (Douglas, 2008; Cochran, 1975). Such

association has also been referred to as ‘byproduct-mutualism’ (Connor, 1995). The type of exchanged commodities and associated costs in a symbiotic association often remain elusive, which limits our ability to make informed predictions how common these byproduct mutualisms are in nature (Cushman and Beattie, 1991).

Despite the theoretical expectation that selection for specialized, one-to-one associations favors the evolutionary stability of mutualism, there is a lot of evidence for one-to-many interactions (Batstone et al., 2018). One of the most popular systems is the human gut microbiome encompassing hundreds of bacterial species. There have been recent attempts to explain such apparent permissiveness to symbiont diversity. The cost in harboring a diverse symbiont community can be reduced by a number of factors including: equal partner quality, non-obligate associations, complementary functions, niche-partitioning, to name a few (Batstone et al., 2018). More importantly, the environment has an essential impact on the evolution of permissiveness to diversity. For example, temporal or spatial variability in the availability of high-quality partners can select against partner-specificity (Batstone et al., 2017). In addition, diversity within a symbiotic community can be beneficial if it leads to a higher resilience to variable environmental conditions and to the protection from potentially invading parasites. The latter has been observed in the skin microbiota of frogs, where a higher species richness led to a better protection against a parasitic lineage (Piovia-Scott et al., 2017). Taken together, and as pointed out by Batstone et al. (2018): “*There is a pressing need for future work that quantifies partner breadth across a wide range of systems and scales in order for us to fully appreciate the resiliency of mutualistic interactions in the face of environmental change.*”.

In addition to conflicts between both symbiotic partners, one also has to consider the ecological concepts of bacterial co-existence. If two members of the symbiont community share some or all resources, they will compete, which consequently prevents a stable co-existence of the two (Russel et al., 2017). Competition will either i) drive one partner to extinction, ii) lead to a physical separation between both partners, or iii) cause a separation between the metabolic niches of both partners (Ghoul and Mitri, 2016). The more closely related two organisms are, the more likely they have overlapping resource requirements (Russel et al., 2017). This implies that closely related strains that belong to the same species are less likely to co-exist than two divergent species. Thompson (2005) stated that in mutualism, similar functions among symbionts should be selected against my minimizing genetic diversity within species, whereas complementary symbionts may be favored. To understand the impact of symbiont co-existence in nature, I explored the level of heterogeneity among closely related strains of the SOX symbiont in bathymodiolin mussels.

1.5.2 Intra-specific diversity of symbionts in nature

Most studies on strain diversity within symbiont species has been done on a few symbiotic model systems, such as the *Rhizobium*-legume and *Vibrio*-squid symbioses. These systems can be cultivated under controlled conditions which allows for testing of theoretical hypotheses. In addition, these provide excellent systems to understand processes such as the emergence of cheaters and host control, e.g. partner choice and sanctioning (Kiers and Denison, 2008). Despite these advantages, there are two limitations to this approach. First, our inability to culture a wide range of symbiotic

systems which means these theories cannot be broadly tested. And second, the unsuitability of this approach to assess true natural conditions. However, understanding the complexity of symbiotic associations in nature is equally important to determine the validity of the above-mentioned theoretical predictions. Recently, the study of strain diversity in natural bacterial populations has become more feasible due to the advances in sequencing technologies (e.g. Delmont and Eren, 2018; Quince et al., 2017). Metagenomic and metatranscriptomic sequencing are currently revealing the extent of strain heterogeneity in natural microbial populations, including symbiotic ones (Ellegaard and Engel, 2019). In my thesis I make use of this resolution to assess strain-level heterogeneity in bathymodiolin symbionts.

As the field of metagenomic sequencing is relatively young, there are not many high-resolution studies of strain diversity in endosymbioses that thrive in chemosynthetic habitats. In the following, I summarize a few examples that have been described. The endosymbionts of *Solemya* clams have assumed for a long time to be vertically transmitted because symbionts were found in host ovaries and embryos (Krueger et al., 1996; Cary and Giovannoni, 1993). Surprisingly, high-resolution sequencing has revealed a low level of micro-diversity within hosts that suggested occasional horizontal symbiont transmission events (Russell and Cavanaugh, 2017; Russell et al., 2017). In another clam host belonging to the Lucinidae family, studies on single copy marker genes have revealed symbiont strain diversity between the lucinid species (Lim et al., 2018; Brissac et al., 2016). However, intra-specific diversity and implications on metabolic functions within single hosts remain to be shown. In the deep sea, two different tubeworm species, *Ridgeia piscesae* and *Riftia pachyptila*

have been shown to harbor symbionts with intra-specific heterogeneity within single individuals (Polzin et al., 2019; Russell and Cavanaugh, 2017). Despite this heterogeneity, within each host individual there was always one dominant strain whereas the other strains were low in abundance. Other non-autotrophic symbioses in chemosynthetic environments have been investigated using single-copy marker genes. Here, bone-eating *Osedax* worms (heterotrophic symbionts) as well as wood-boring *Lyrodus* bivalves have been shown to harbor up to nine 16S ribotypes within single hosts (Verna et al., 2010; Goffredi et al., 2007; Luyten et al., 2006). These may represent intra-specific and/or inter-specific differences and the level of genomic diversity remains to be shown. For *Osedax*, genomic studies have revealed two dominating genotypes within hosts but this was cultivation-based and true strain diversity may be higher (Goffredi et al., 2014).

Aims of this thesis

When I started my doctoral studies, little was known about the intra-specific diversity in symbionts of bathymodiolin mussels. At the same time, evidence was accumulating that diversity beyond species boundaries can be pervasive in bacterial communities and can have profound effects on their functioning. Small genomic changes can strongly impact the lifestyle of a bacterial strain. For example, a single mutation in a regulatory gene can change the host range of a symbiont (Pankey et al., 2017). As outlined above, there are a lot of open questions on how strain diversity in symbioses emerges, persists and how it impacts mutualistic associations. Understanding these processes in environmental symbioses is challenging, as the symbionts cannot be cultivated and capturing natural symbiont diversity requires high-resolution methods. The *Bathymodiolus* symbioses represents an ideal system to study the presence and the impact of strain diversity on symbioses, because it associates only with a few bacterial species (Duperron, 2010; Duperron et al., 2008, 2005). This low level of diversity allows for an in-depth analysis that can resolve intra-specific heterogeneity in the symbionts. Using cultivation-independent methods I wanted to understand the role of strain diversity in the SOX symbiont of bathymodiolin mussels. The specific questions that I aim to address in the following chapters are as follows.

What is the extent of genomic diversity in the SOX symbiont of Bathymodiolus mussels?

Previous studies suggested that the SOX symbionts in *Bathymodiolus* have a level of unexplored microdiversity (Duperron et al., 2008, 2007; Won et al., 2003). For *B. septemdierum* symbionts, a recent study revealed heterogeneity in genes encoding proteins of the energy metabolism (Ikuta et al., 2016). However, the extent of strain diversity within and between host individuals of different *Bathymodiolus* species remained unknown. Determining the level of strain diversity in mutualistic endosymbioses is highly relevant because it is often thought to destabilize the association. *Bathymodiolus* mussels associate only with few endosymbiotic species, making it an ideal system to explore questions as to how many strains can co-exist, how the strains differ in their encoded functions and how this connects to theoretical predictions. Low complexity of the symbiont community allows the application of high-resolution and high-coverage metagenomics and to resolve symbiont diversity in their natural context.

To answer the above-mentioned question, I developed a metagenomics pipeline that uses metagenomic sequencing to tease apart strain-level differences that cannot be distinguished by standard binning methods (Ansorge et al., 2019) (chapter II). The developed pipeline allowed me to reveal a high level of nucleotide diversity among co-existing strains within host individuals. Using the information of read coverage, my studies showed substantial functional heterogeneity among these strains, which raised the question as to how these symbionts can co-exist in a single organism. Together with my co-authors I could show that the metabolic plasticity among co-

existing strains affected protein functions that were detected in the transcriptomes. Therefore, the gene content variation is likely linked to phenotypic differences as well.

The manuscript in chapter II has been deposited as a preprint at *bioRxiv* (Ansorge et al., 2019) and is currently under revision at *Nature Microbiology*.

How did metabolic diversity evolve in the SOX symbiont?

The bacterial pangenome consists of two components, the ‘core’ genome referring to all the genes that are shared among the members of a phylogenetic group (e.g. species) and the ‘accessory’ genome that represents dispensable functions not found in each genome (McInerney et al., 2017; Young et al., 2006). Pangenome sizes differ enormously among bacterial species but underlying evolutionary forces driving these differences are unclear and subject to ongoing debates (McInerney et al., 2017; Vos and Eyre-Walker, 2017). Based on our observations of extensive gene content variation in *Bathymodiolus* symbiont species and strains (chapter II), the question emerged whether this is a general feature of this bacterial clade and how it evolved. Lifestyle is thought to a have strong influence on the evolution of pangenomes. The SOX symbionts belong to the widespread SUP05 clade that includes a diversity of lifestyles, namely free-living bacteria, clam endosymbionts (vertically transmitted), *Bathymodiolus* endosymbionts (horizontally transmitted) and sponge symbionts (unknown transmission mode). This provided me with the opportunity to analyze the impact of lifestyle and genetic relatedness on pangenome evolution. In addition, previous studies of the free-living lineages revealed metabolic plasticity among them. I performed a pangenome analysis which suggested that gene content similarity is

mostly explained by phylogenetic relationships and not by lifestyle. This brought me to hypothesize that convergent evolution led to the emergence of similar mechanisms that favor particular lifestyles (e.g. association with *Bathymodiolus*) but with distinct genomic solutions. A possible explanation for the global success of the SUP05 clade could thus be the selection for evolvability as a trait (chapter III).

Do Bathymodiolus mussels continuously sample symbiont strains from the environment?

Despite common agreement that *Bathymodiolus* symbionts are horizontally transmitted, this question is still not answered (Wentrup et al., 2014). Together with my collaborators I discovered high degrees of similarity in the symbiont populations between host individuals from the same location. This was an indication for frequent exchange of symbionts between host individuals. However, there were differences between host species *B. brooksi* and the other host species, raising the possibility of different transmission dynamics in this species. In chapter IV, I provide some possible explanations for these observations, but in order to fully explain the differences in symbiont transmission in *B. brooksi*, a broader sampling effort would be required. Therefore, this remains an intriguing topic for future exploration.

What drives genome evolution in Bathymodiolus SOX symbionts?

Genome evolution is a field that is heavily explored in symbiosis research, especially symbioses with strong bottlenecking, to try to tease apart the effects of genetic drift and natural selection. Nevertheless, the processes of genome evolution are often difficult to determine especially in natural populations. In horizontally transmitted symbionts, the sizes of potential bottlenecks usually are unclear and therefore the assessment of genomic signatures of evolution can help to understand symbiont transmission dynamics (Russell et al., 2017) as well as evolutionary pressures. Also in *Bathymodiolus* SOX symbionts, the details of transmission are not fully understood (see above). This also hampers our ability to make predictions about the impact of genetic drift and natural selection on the symbiont genomes. Therefore, I targeted the above-mentioned question by analyzing allele frequencies among populations of the SOX symbiont (chapter IV). The analysis included 88 metagenomes of single host individuals including seven distinct symbiont species from 15 different locations, which allows me to detect intra-specific signatures of genome evolution. My results indicated that the SOX symbiont experiences divergent selection in traits that were possibly driven by host-symbiont interaction and environmental conditions. This study was a first step towards elucidating the signatures of genome evolution in *Bathymodiolus* endosymbionts in their natural habitat.

List of publications

1. Diversity matters: Deep-sea mussels harbor multiple symbiont strains

Rebecca Ansorge, Stefano Romano, Lizbeth Sayavedra, Anne Kupczok, Halina E. Tegetmeyer, Nicole Dubilier, Jillian Petersen

manuscript under revision at *Nature Microbiology;* preprint available at *bioRxiv* https://doi.org/10.1101/531459

2. Genome structure reflects phylogeny rather than lifestyle in a widespread group of free-living and symbiotic marine bacteria from the SUP05 clade

Rebecca Ansorge, Stefano Romano, Lizbeth Sayavedra, Maxim Rubin-Blum, Nicole Dubilier, Jillian Petersen

manuscript *in preparation*

3. Evolutionary signatures in genomes of horizontally transmitted endosymbionts of *Bathymodiolus* mussels

Rebecca Ansorge, Stefano Romano, Nicole Dubilier, Jillian Petersen

manuscript *in preparation*

Contributed works not included in this thesis

4. Horizontally transmitted symbiont populations in deep-sea mussels are genetically isolated

Devani Romero Picazo, Tal Dagan, **Rebecca Ansorge**, Jillian Petersen, Nicole Dubilier, Anne Kupczok

manuscript submitted to *ISME;* preprint available at *bioRxiv* https://doi.org/10.1101/536854

5. Horizontal acquisition followed by expansion and diversification of toxin-related genes in deep-sea bivalve symbionts

Lizbeth Sayavedra, **Rebecca Ansorge**, Maxim Rubin-Blum, Nikolaus Leisch, Nicole Dubilier, Jillian Petersen

manuscript *in preparation*

6. Understanding symbiont colonization of deep-sea mussels using differential gene expressions

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

manuscript *in preparation*

References of introduction

Amend, J.P., McCollom, T.M., Hentscher, M., and Bach, W. (2011). Catabolic and anabolic energy for chemolithoautotrophs in deep-sea hydrothermal systems hosted in different rock types. Geochim. Cosmochim. Acta *75*, 5736–5748.

Anantharaman, K., Breier, J.A., Sheik, C.S., and Dick, G.J. (2013). Evidence for hydrogen oxidation and metabolic plasticity in widespread deep-sea sulfur-oxidizing bacteria. Proc. Natl. Acad. Sci. *110*, 330–335.

Ansorge, R., Romano, S., Sayavedra, L., Kupczok, A., Tegetmeyer, H.E., Dubilier, N., and Petersen, J. (2019). Diversity matters: Deep-sea mussels harbor multiple symbiont strains. BioRxiv 531459.

Archetti, M., Scheuring, I., Hoffman, M., Frederickson, M.E., Pierce, N.E., and Yu, D.W. (2011). Economic game theory for mutualism and cooperation. Ecol. Lett. *14*, 1300–1312.

Arp, A.J., and Childress, J.J. (1983). Sulfide binding by the blood of the hydrothermal vent tube worm *Riftia pachyptila*. Science *219*, 295–297.

Arp, A.J., Childress, J.J., and Vetter, R.D. (1987). The sulphide-binding protein in the blood of the vestimentiferan tube-worm, *Riftia Pachyptila*, is the extracellular haemoglobin. J. Exp. Biol. *128*, 139–158.

Assié, A., Borowski, C., Heijden, K. van der, Raggi, L., Geier, B., Leisch, N., Schimak, M.P., Dubilier, N., and Petersen, J.M. (2016). A specific and widespread association between deep-sea *Bathymodiolus* mussels and a novel family of Epsilonproteobacteria. Environ. Microbiol. Rep. *8*, 805–813.

de Bary, A. (1879). Die Erscheinung der Symbiose: Vortrag, gehalten auf der Versammlung deutscher Naturforscher und Aerzte zu Cassel.

Batstone, R.T., Dutton, E.M., Wang, D., Yang, M., and Frederickson, M.E. (2017). The evolution of symbiont preference traits in the model legume *Medicago truncatula*. New Phytol. *213*, 1850–1861.

Batstone, R.T., Carscadden, K.A., Afkhami, M.E., and Frederickson, M.E. (2018). Using niche breadth theory to explain generalization in mutualisms. Ecology.

Bennett, G.M., and Moran, N.A. (2015). Heritable symbiosis: The advantages and perils of an evolutionary rabbit hole. Proc. Natl. Acad. Sci. *112*, 10169–10176.

Bobay, L.-M., and Ochman, H. (2018). Factors driving effective population size and pan-genome evolution in bacteria. BMC Evol. Biol. *18*.

Bright, M., and Bulgheresi, S. (2010). A complex journey: transmission of microbial symbionts. Nat. Rev. Microbiol. *8*, 218–230.

Bright, M., Klose, J., and Nussbaumer, A.D. (2013). Giant tubeworms. Curr. Biol. *23*, R224–R225.

Brissac, T., Higuet, D., Gros, O., and Merçot, H. (2016). Unexpected structured intraspecific diversity of thioautotrophic bacterial gill endosymbionts within the Lucinidae (Mollusca: Bivalvia). Mar. Biol. *163*, 176.

Bronstein, J.L. (2015). Mutualism (Oxford University Press).

Bull, J.J., and Rice, W.R. (1991). Distinguishing mechanisms for the evolution of co-operation. J. Theor. Biol. *149*, 63–74.

Callbeck, C.M., Lavik, G., Ferdelman, T.G., Fuchs, B., Gruber-Vodicka, H.R., Hach, P.F., Littmann, S., Schoffelen, N.J., Kalvelage, T., Thomsen, S., et al. (2018). Oxygen minimum zone cryptic sulfur cycling sustained by offshore transport of key sulfur oxidizing bacteria. Nat. Commun. *9*, 1729.

Cary, S.C., and Giovannoni, S.J. (1993). Transovarial inheritance of endosymbiotic bacteria in clams inhabiting deep-sea hydrothermal vents and cold seeps. Proc. Natl. Acad. Sci. *90*, 5695–5699.

Cavanaugh, C.M. (1983). Symbiotic chemoautotrophic bacteria in marine invertebrates from sulphide-rich habitats. Nature *302*, 58.

Cavanaugh, C.M., and Robinson, J.J. (1996). CO2 Fixation in chemoautotroph-invertebrate symbioses: expression of form I and form II RubisCO. In *Microbial growth on C1 compounds: Proceedings of the 8th International Symposium on Microbial Growth on C1 Compounds*, M.E. Lidstrom, and F.R. Tabita, eds. (Dordrecht: Springer Netherlands), pp. 285–292.

Cavanaugh, C.M., Gardiner, S.L., Jones, M.L., Jannasch, H.W., and Waterbury, J.B. (1981). Prokaryotic cells in the hydrothermal vent tube worm *Riftia pachyptila* Jones: possible chemoautotrophic symbionts. Science *213*, 340–342.

Cavanaugh, C.M., McKiness, Z.P., Newton, I.L.G., and Stewart, F.J. (2013). Marine chemosynthetic symbioses. In *The Prokaryotes: Prokaryotic Biology and Symbiotic Associations*, E. Rosenberg, E.F. DeLong, S. Lory, E. Stackebrandt, and F. Thompson, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 579–607.

Charlou, J.-L., and Donval, J.-P. (1993). Hydrothermal methane venting between 12°N and 26°N along the Mid-Atlantic Ridge. J. Geophys. Res. Solid Earth *98*, 9625–9642.

Charlou, J.L., Donval, J.P., Fouquet, Y., Jean-Baptiste, P., and Holm, N. (2002). Geochemistry of high H2 and CH4 vent fluids issuing from ultramafic rocks at the Rainbow hydrothermal field (36°14′N, MAR). Chem. Geol. *191*, 345–359.

Childress, J.J., Fisher, C.R., Brooks, J.M., Kennicutt, M.C., Bidigare, R., and Anderson, A.E. (1986). A methanotrophic marine molluscan (Bivalvia, Mytilidae) symbiosis: mussels fueled by gas. Science *233*, 1306–1308.

Cochran, D.G. (1975). Excretion in insects. In *Insect Biochemistry and Function*, D.J. Candy, and B.A. Kilby, eds. (Boston, MA: Springer US), pp. 177–281.

Connor, R.C. (1995). The benefits of mutualism: a conceptual framework. Biol. Rev. *70*, 427–457.

Cordes, E.E., Cunha, M.R., Galéron, J., Mora, C., Olu-Le Roy, K., Sibuet, M., Van Gaever, S., Vanreusel, A., and Levin, L.A. (2010). The influence of geological, geochemical, and biogenic habitat heterogeneity on seep biodiversity: Seep habitat heterogeneity. Mar. Ecol. *31*, 51–65.

von Cosel, R., and Olu, K. (1998). Gigantism in Mytilidae. A new *Bathymodiolus* from cold seep areas on the Barbados accretionary Prism. Comptes Rendus Académie Sci. - Ser. III - Sci. Vie *321*, 655–663.

Cushman, J.H., and Beattie, A.J. (1991). Mutualisms: assessing the benefits to hosts and visitors. Trends Ecol. Evol. *6*, 193–195.

DeChaine, E.G., and Cavanaugh, C.M. (2005). Symbioses of methanotrophs and deep-sea mussels (Mytilidae: Bathymodiolinae). In *Molecular Basis of Symbiosis*, P.D.J. Overmann, ed. (Springer Berlin Heidelberg), pp. 227–249.

Delmont, T.O., and Eren, A.M. (2018). Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. PeerJ *6*, e4320.

Distel, D.L., Lee, H.K., and Cavanaugh, C.M. (1995). Intracellular coexistence of methano- and thioautotrophic bacteria in a hydrothermal vent mussel. Proc. Natl. Acad. Sci. U. S. A. *92*, 9598–9602.

Distel, D.L., Baco, A.R., Chuang, E., Morrill, W., Cavanaugh, C., and Smith, C.R. (2000). Marine ecology: Do mussels take wooden steps to deep-sea vents? Nature *403*, 725–726.

Distel, D.L., Altamia, M.A., Lin, Z., Shipway, J.R., Han, A., Forteza, I., Antemano, R., Limbaco, M.G.J.P., Tebo, A.G., Dechavez, R., et al. (2017). Discovery of chemoautotrophic symbiosis in the giant shipworm *Kuphus polythalamia* (Bivalvia: Teredinidae) extends wooden-steps theory. Proc. Natl. Acad. Sci. *114*, E3652–E3658.

Douglas, A.E. (2008). Conflict, cheats and the persistence of symbioses. New Phytol. *177*, 849–858.

Douglas, A.E. (2010). The symbiotic habit (Princeton University Press).

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

Duperron, S. (2010). The diversity of deep-sea mussels and their bacterial symbioses. In *The vent and seep biota*, S. Kiel, ed. (Springer Netherlands), pp. 137–167.

Duperron, S., Nadalig, T., Caprais, J.-C., Sibuet, M., Fiala-Médioni, A., Amann, R., and Dubilier, N. (2005). Dual Symbiosis in a *Bathymodiolus* sp. Mussel from a methane seep on the Gabon Continental Margin (Southeast Atlantic): 16S rRNA phylogeny and distribution of the symbionts in gills. Appl. Environ. Microbiol. *71*, 1694–1700.

Duperron, S., Bergin, C., Zielinski, F., Blazejak, A., Pernthaler, A., McKiness, Z.P., DeChaine, E., Cavanaugh, C.M., and 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. Environ. Microbiol. *8*, 1441–1447.

Duperron, S., Sibuet, M., MacGregor, B.J., Kuypers, M.M.M., Fisher, C.R., and 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.

Duperron, S., Halary, S., Lorion, J., Sibuet, M., and 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.

Duperron, S., Gaudron, S.M., Rodrigues, C.F., Cunha, M.R., Decker, C., and Olu, K. (2013). An overview of chemosynthetic symbioses in bivalves from the North Atlantic and Mediterranean Sea. Biogeosciences *10*, 3241–3267.

Ellegaard, K.M., and Engel, P. (2019). Genomic diversity landscape of the honey bee gut microbiota. Nat. Commun. *10*, 446.

Endow, K., and Ohta, S. (1990). Occurrence of bacteria in the primary oocytes of vesicomyid clam *Calyptogena soyoae*. Mar. Ecol. Prog. Ser. *64*, 309–311.

Felbeck, H., Childress, J.J., and Somero, G.N. (1981). Calvin-Benson cycle and sulphide oxidation enzymes in animals from sulphide-rich habitats. Nature *293*, 291.

Fiala-Médioni, A., McKiness, Z., Dando, P., Boulegue, J., Mariotti, A., Alayse-Danet, A., Robinson, J., and Cavanaugh, C. (2002). Ultrastructural, biochemical, and immunological characterization of two populations of the mytilid mussel *Bathymodiolus azoricus* from the Mid-Atlantic Ridge: evidence for a dual symbiosis. Mar. Biol. *141*, 1035–1043.

Fisher, C.R., and Childress, J.J. (1992). Organic carbon transfer from methanotrophic symbionts to the host hydrocarbon-seep mussel. Symbiosis *12*, 221-235

Fisher, C.R., Childress, J.J., Oremland, R.S., and Bidigare, R.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.

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

Frank, S.A. (1996). Host–symbiont conflict over the mixing of symbiotic lineages. Proc R Soc Lond B *263*, 339–344.

Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G., and Hanage, W.P. (2009). The bacterial species challenge: making sense of genetic and ecological diversity. Science *323*, 741–746.

Frederickson, M.E. (2013). Rethinking mutualism stability: cheaters and the evolution of sanctions. Q. Rev. Biol. *88*, 269–295.

Gaill, F., B, S., F, M., G, G., and Jj, C. (1997). Rate and process of tube production by the deep-sea hydrothermal vent tubeworm *Riftia pachyptila*. Mar. Ecol. Prog. Ser. *148*, 135–143.

Génio, L., Johnson, S.B., Vrijenhoek, R.C., Cunha, M.R., Tyler, P.A., Kiel, S., and Little, C.T.S. (2008). New record of *"Bathymodiolus"* Mauritanicus Cosel 2002 from the Gulf of Cadiz (NE Atlantic) Mud Volcanoes. J. Shellfish Res. *27*, 53–61.

German, C.R., Ramirez-Llodra, E., Baker, M.C., Tyler, P.A., and committee, and the C.S.S. (2011). Deep-water chemosynthetic ecosystem research during the census of marinelife decade and beyond: A proposed deep-ocean road map. PLOS ONE *6*, e23259.

Ghoul, M., and Mitri, S. (2016). The ecology and evolution of microbial competition. Trends Microbiol. *24*, 833–845.

Giere, O., Conway, N.M., Gastrock, G., and Schmidt, C. (1991). "Regulation" of gutless annelid ecology by endosymbiotic bacteria. Mar. Ecol. Prog. Ser. *68*, 287–299.

Goffredi, S., Hurtado, L., Hallam, S., and Vrijenhoek, R. (2003). Evolutionary relationships of deep-sea vent and cold seep clams (Mollusca: Vesicomyidae) of the “pacifica/lepta” species complex. Mar. Biol. *142*, 311–320.

Goffredi, S.K., Johnson, S.B., and Vrijenhoek, R.C. (2007). Genetic diversity and potential function of microbial symbionts associated with newly discovered species of *Osedax* polychaete worms. Appl Env. Microbiol *73*, 2314–2323.

Goffredi, S.K., Yi, H., Zhang, Q., Klann, J.E., Struve, I.A., Vrijenhoek, R.C., and Brown, C.T. (2014). Genomic versatility and functional variation between two dominant heterotrophic symbionts of deep-sea *Osedax* worms. ISME J. *8*, 908–924.

Gruber-Vodicka, H.R., Dirks, U., Leisch, N., Baranyi, C., Stoecker, K., Bulgheresi, S., Heindl, N.R., Horn, M., Lott, C., Loy, A., et al. (2011). Paracatenula, an ancient symbiosis between thiotrophic Alphaproteobacteria and catenulid flatworms. Proc. Natl. Acad. Sci. *108*, 12078–12083.

Hurtado, L.A., Mateos, M., Lutz, R.A., and Vrijenhoek, R.C. (2003). Coupling of bacterial endosymbiont and host mitochondrial genomes in the hydrothermal vent clam *Calyptogena magnifica*. Appl Env. Microbiol *69*, 2058–2064.

Husson, B., Sarradin, P.-M., Zeppilli, D., and Sarrazin, J. (2017). Picturing thermal niches and biomass of hydrothermal vent species. Deep Sea Res. Part II Top. Stud. Oceanogr. *137*, 6–25.

Ikuta, T., Takaki, Y., Nagai, Y., Shimamura, S., Tsuda, M., Kawagucci, S., Aoki, Y., Inoue, K., Teruya, M., Satou, K., et al. (2016). Heterogeneous composition of key metabolic gene clusters in a vent mussel symbiont population. ISME J. *10*, 990-1001.

Jannasch, H.W. (1985). Review Lecture - The chemosynthetic support of life and the microbial diversity at deep-sea hydrothermal vents. Proc. R. Soc. Lond. B Biol. Sci. *225*, 277–297.

Jannasch, H.W., and Mottl, M.J. (1985). Geomicrobiology of deep-sea hydrothermal vents. Science *229*, 717–725.

Jørgensen, B.B., and Boetius, A. (2007). Feast and famine — microbial life in the deep-sea bed. Nat. Rev. Microbiol. *5*, 770–781.

Kádár, E., Davis, S.A., and Lobo-da-Cunha, A. (2008). Cytoenzymatic investigation of intracellular digestion in the symbiont-bearing hydrothermal bivalve *Bathymodiolus azoricus*. Mar. Biol. *153*, 995–1004.

Kelley, D.S., Baross, J.A., and Delaney, J.R. (2002). Volcanoes, fluids, and life at Mid-Ocean Ridge spreading centers. Annu. Rev. Earth Planet. Sci. *30*, 385–491.

Kennicutt, M.C., Brooks, J.M., Bidigare, R.R., Fay, R.R., Wade, T.L., and McDonald, T.J. (1985). Vent-type taxa in a hydrocarbon seep region on the Louisiana slope. Nature *317*, 351.

Kiers, E.T., and Denison, R.F. (2008). Sanctions, cooperation, and the stability of plant-rhizosphere mutualisms. Annu. Rev. Ecol. Evol. Syst. *39*, 215–236.

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.

Kleiner, M., Hooper, L.V., and Duerkop, B.A. (2015). Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. BMC Genomics *16*, 7.

Krueger, D.M., Gustafson, R.G., and Cavanaugh, C.M. (1996). Vertical transmission of chemoautotrophic symbionts in the bivalve *Solemya velum* (Bivalvia: Protobranchia). Biol. Bull. *190*, 195–202.

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.

Kuwahara, H., Takaki, Y., Yoshida, T., Shimamura, S., Takishita, K., Reimer, J.D., Kato, C., and Maruyama, T. (2008). Reductive genome evolution in chemoautotrophic intracellular symbionts of deep-sea *Calyptogena* clams. Extrem. Life Extreme Cond. *12*, 365–374.

Kuwahara, H., Takaki, Y., Shimamura, S., Yoshida, T., Maeda, T., Kunieda, T., and Maruyama, T. (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.

Levin (2005). Oceanography and marine biology - an Annual Review, Vol. 43 (Boca Raton: Crc Press-Taylor & Francis Group).

Lim, S.J., Davis, B.G., Gill, D.E., Walton, J., Nachman, E., Engel, A.S., Anderson, L.C., and Campbell, B.J. (2018). Taxonomic and functional heterogeneity of the gill microbiome in a symbiotic coastal mangrove lucinid species. ISME J. 1.

Little, A.E.F. (2010). Parasitism is a strong force shaping the fungus-growing ant–microbe symbiosis. In *Symbioses and stress: joint ventures in biology*, J. Seckbach, and M. Grube, eds. (Dordrecht: Springer Netherlands), pp. 245–264.

Liu, J., Liu, H., and Zhang, H. (2018). Phylogeny and evolutionary radiation of the marine mussels (Bivalvia: Mytilidae) based on mitochondrial and nuclear genes. Mol. Phylogenet. Evol. *126*, 233–240.

Lonsdale, P. (1977). Clustering of suspension-feeding macrobenthos near abyssal hydrothermal vents at oceanic spreading centers. Deep Sea Res. *24*, 857–863.

Lorion, J., Kiel Steffen, Faure Baptiste, Kawato Masaru, Ho Simon Y. W., Marshall Bruce, Tsuchida Shinji, Miyazaki Jun-Ichi, and Fujiwara Yoshihiro (2013). Adaptive radiation of chemosymbiotic deep-sea mussels. Proc. R. Soc. B Biol. Sci. *280*, 20131243.

Lutz, R.A., Shank, T.M., Fornari, D.J., Haymon, R.M., Lilley, M.D., Von Damm, K.L., and Desbruyeres, D. (1994). Rapid growth at deep-sea vents. Nature *371*, 663–664.

Luyten, Y.A., Thompson, J.R., Morrill, W., Polz, M.F., and Distel, D.L. (2006). Extensive variation in intracellular symbiont community composition among members of a single population of the wood-boring bivalve *Lyrodus pedicellatus* (Bivalvia: Teredinidae). Appl Env. Microbiol *72*, 412–417.

Margulis, L. (1970). Origin of eukaryotic cells: evidence and research implications for a theory of the origin and evolution of microbial, plant, and animal cells on the precambrian Earth (Yale University Press).

Margulis, L., and Fester, R. (1991). Bellagio conference and book. Symbiosis as source of evolutionary innovation: speciation and morphogenesis. Symbiosis Phila. Pa *11*, 93–101.

Marshall, K.T., and Morris, R.M. (2013). Isolation of an aerobic sulfur oxidizer from the SUP05/Arctic96BD-19 clade. ISME J. *7*, 452–455.

Marshall, K.T., and Morris, R.M. (2015). Genome sequence of "*Candidatus* Thioglobus singularis" strain PS1, a mixotroph from the SUP05 clade of marine Gammaproteobacteria. Genome Announc. *3*.

Martin, B., and Schwab, E. (2012). Current usage of symbiosis and associated terminology. Int. J. Biol. *5*, 32.

Martin, W., Baross, J., Kelley, D., and Russell, M.J. (2008). Hydrothermal vents and the origin of life. Nat. Rev. Microbiol. *6*, 805–814.

McCollom, T.M. (2007). Geochemical constraints on sources of metabolic energy for chemolithoautotrophy in ultramafic-hosted deep-sea hydrothermal systems. Astrobiology *7*, 933–950.

McInerney, J.O., McNally, A., and O'Connell, M.J. (2017). Why prokaryotes have pangenomes. Nat. Microbiol. *2*, 17040.

Meier, D.V., Pjevac, P., Bach, W., Hourdez, S., Girguis, P.R., Vidoudez, C., Amann, R., and Meyerdierks, A. (2017). Niche partitioning of diverse sulfur-oxidizing bacteria at hydrothermal vents. ISME J. *11*, 1545–1558.

Mira, A., and Moran, N.A. (2002). Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria. Microb. Ecol. *44*, 137–143.

Moran, N.A. (1996). Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. *93*, 2873–2878.

Moran, N.A., and Bennett, G.M. (2014). The tiniest tiny genomes. Annu. Rev. Microbiol. *68*, 195–215.

Moran, N.A., and Telang, A. (1998). Bacteriocyte-associated symbionts of insects. BioScience *48*, 295–304.

Murillo, A.A., Ramírez-Flandes, S., DeLong, E.F., and Ulloa, O. (2014). Enhanced metabolic versatility of planktonic sulfur-oxidizing γ-proteobacteria in an oxygen-deficient coastal ecosystem. Front. Mar. Sci. *1*.

Nelson, D.C., Hagen, K.D., and Edwards, D.B. (1995). The gill symbiont of the hydrothermal vent mussel *Bathymodiolus thermophilus* is a psychrophilic, chemoautotrophic, sulfur bacterium. Mar. Biol. *121*, 487–495.

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.

Nishijima, M., Lindsay, D.J., Hata, J., Nakamura, A., Kasai, H., Ise, Y., Fisher, C.R., Fujiwara, Y., Kawato, M., and Maruyama, T. (2010). Association of thioautotrophic bacteria with deep-sea sponges. Mar. Biotechnol. N. Y. N *12*, 253–260.

Nussbaumer, A.D., Fisher, C.R., and Bright, M. (2006). Horizontal endosymbiont transmission in hydrothermal vent tubeworms. Nature *441*, 345–348.

Nyholm, S.V., and McFall-Ngai, M. (2004). The winnowing: establishing the squid–*Vibrio* symbiosis. Nat. Rev. Microbiol. *2*, 632–642.

Ott, J.A. (1989). Living at an interface: Meiofauna at the oxygen/sulfide boundary of marine sediments Jdrg A. Ott & R. Novak Institute of Zoology, University of Vienna. In *Reproduction, genetics and distributions of marine organisms: 23rd european*

marine biology symposium, School of Biological Sciences, University of Wales, Swansea, (Olsen & Olsen), p. 415.

Owen, G. (1961). A Note on the Habits and Nutrition of *Solemya parkinsoni* (Protobranchia: Bivalvia). J. Cell Sci. *s3-102*, 15–21.

Ozawa, G., Shimamura, S., Takaki, Y., Takishita, K., Ikuta, T., Barry, J.P., Maruyama, T., Fujikura, K., and Yoshida, T. (2017). Ancient occasional host switching of maternally transmitted bacterial symbionts of chemosynthetic vesicomyid clams. Genome Biol. Evol. *9*, 2226–2236.

Pankey, M.S., Foxall, R.L., Ster, I.M., Perry, L.A., Schuster, B.M., Donner, R.A., Coyle, M., Cooper, V.S., and Whistler, C.A. (2017). Host-selected mutations converging on a global regulator drive an adaptive leap towards symbiosis in bacteria. eLife *6*, e24414.

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. *36*, 996–1004.

Perner, M., Gonnella, G., Kurtz, S., and LaRoche, J. (2014). Handling temperature bursts reaching 464°C: different microbial strategies in the sisters peak hydrothermal chimney. Appl Env. Microbiol *80*, 4585–4598.

Petersen, J.M., Zielinski, F.U., Pape, T., Seifert, R., Moraru, C., Amann, R., Hourdez, S., Girguis, P.R., Wankel, S.D., Barbe, V., et al. (2011). Hydrogen is an energy source for hydrothermal vent symbioses. Nature *476*, 176–180.

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

Pimenov, N.V., Kalyuzhnaya, M.G., Khmelenina, V.N., Mityushina, L.L., and Trotsenko, Y.A. (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.

Piovia-Scott, J., Rejmanek, D., Woodhams, D.C., Worth, S.J., Kenny, H., McKenzie, V., Lawler, S.P., and Foley, J.E. (2017). Greater species richness of bacterial skin symbionts better suppresses the amphibian fungal pathogen *Batrachochytrium dendrobatidis*. Microb. Ecol. *74*, 217–226.

Poisot, T., Bever, J.D., Nemri, A., Thrall, P.H., and Hochberg, M.E. (2011). A conceptual framework for the evolution of ecological specialisation. Ecol. Lett. *14*, 841–851.

Polzin, J., Arevalo P., Nussbaumer T., Polz M.F., and Bright M. (2019). Polyclonal symbiont populations in hydrothermal vent tubeworms and the environment. Proc. R. Soc. B Biol. Sci. *286*, 20181281.

Ponnudurai, R., Kleiner, M., Sayavedra, L., Petersen, J.M., Moche, M., Otto, A., Becher, D., Takeuchi, T., Satoh, N., Dubilier, N., et al. (2017). Metabolic and physiological interdependencies in the *Bathymodiolus azoricus* symbiosis. ISME J. *11*, 463–477.

Powell, M.A., and Somero, G.N. (1986). Adaptations to sulfide by hydrothermal vent animals: sites and mechanisms of detoxification and metabolism. Biol. Bull. *171*, 274–290.

Quince, C., Delmont, T.O., Raguideau, S., Alneberg, J., Darling, A.E., Collins, G., and Eren, A.M. (2017). DESMAN: a new tool for de novo extraction of strains from metagenomes. Genome Biol. *18*, 181.

Raggi, L., Schubotz, F., Hinrichs, K.-U., Dubilier, N., and Petersen, J.M. (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.

Reid, R.G.B., and Bernard, F.R. (1980). Gutless bivalves. Science *208*, 609–610.

Riou, V., Halary, S., Duperron, S., Bouillon, S., Elskens, M., Bettencourt, R., Santos, R., Dehairs, F., and Colaço, A. (2008). Influence of CH4 and H2S availability on symbiont distribution, carbon assimilation and transfer in the dual symbiotic vent mussel *Bathymodiolus azoricus*. Biogeosciences *5*, 1681–1691.

Robinson, J.J., Polz, M.F., Fiala-Medioni, A., and Cavanaugh, C.M. (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.

Roux, S., Hawley, A.K., Beltran, M.T., Scofield, M., Schwientek, P., Stepanauskas, R., Woyke, T., Hallam, S.J., and Sullivan, M.B. (2014). Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. eLife *3*, e03125.

Rubin-Blum, M., Antony, C.P., Borowski, C., Sayavedra, L., Pape, T., Sahling, H., Bohrmann, G., Kleiner, M., Redmond, M.C., Valentine, D.L., et al. (2017). Short-chain alkanes fuel mussel and sponge *Cycloclasticus* symbionts from deep-sea gas and oil seeps. Nat. Microbiol. *2*, 17093.

Rubin-Blum, M., Antony, C.P., Sayavedra, L., Martínez-Pérez, C., Birgel, D., Peckmann, J., Wu, Y.-C., Cardenas, P., MacDonald, I., Marcon, Y., et al. (2019).

Fueled by methane: deep-sea sponges from asphalt seeps gain their nutrition from methane-oxidizing symbionts. ISME J. 1.

Ruehland, C., Blazejak, A., Lott, C., Loy, A., Erséus, C., and Dubilier, N. (2008). Multiple bacterial symbionts in two species of co-occurring gutless oligochaete worms from Mediterranean sea grass sediments. Environ. Microbiol. *10*, 3404–3416.

Russel, J., Røder, H.L., Madsen, J.S., Burmølle, M., and Sørensen, S.J. (2017). Antagonism correlates with metabolic similarity in diverse bacteria. Proc. Natl. Acad. Sci. *114*, 10684–10688.

Russell, S.L., and Cavanaugh, C.M. (2017). Intrahost genetic diversity of bacterial symbionts exhibits evidence of mixed infections and recombinant haplotypes. Mol. Biol. Evol. *34*, 2747–2761.

Russell, S.L., Corbett-Detig, R.B., and Cavanaugh, C.M. (2017). Mixed transmission modes and dynamic genome evolution in an obligate animal–bacterial symbiosis. ISME J. *11*, 1359.

Sachs, J.L., Mueller, U.G., Wilcox, T.P., and Bull, J.J. (2004). The Evolution of cooperation. Q. Rev. Biol. *79*, 135–160.

Samadi, S., Quéméré, E., Lorion, J., Tillier, A., von Cosel, R., Lopez, P., Cruaud, C., Couloux, A., and Boisselier-Dubayle, M.-C. (2007). Molecular phylogeny in mytilids supports the wooden steps to deep-sea vents hypothesis. C. R. Biol. *330*, 446–456.

Sayavedra, L. (2016). Host-symbiont interactions and metabolism of chemosynthetic symbiosis in deep-sea Bathymodiolus mussels. PhD Thesis. University of Bremen.

Sayavedra, L., Kleiner, M., Ponnudurai, R., Wetzel, S., Pelletier, E., Barbe, V., Satoh, N., Shoguchi, E., Fink, D., Breusing, C., et al. (2015). Abundant toxin-related genes in the genomes of beneficial symbionts from deep-sea hydrothermal vent mussels. eLife e07966.

Shah, V., and Morris, R.M. (2015). Genome Sequence of "*Candidatus* Thioglobus autotrophica" Strain EF1, a Chemoautotroph from the SUP05 Clade of Marine Gammaproteobacteria. Genome Announc. *3*.

Shah, V., Chang, B.X., and Morris, R.M. (2017). Cultivation of a chemoautotroph from the SUP05 clade of marine bacteria that produces nitrite and consumes ammonium. ISME J. *11*, 263–271.

Smith, C. (2012). Chemosynthesis in the deep-sea: life without the sun. Biogeosciences Discuss. *9*, 17037–17052.

Smith, C.R., Glover, A.G., Treude, T., Higgs, N.D., and Amon, D.J. (2015). Whale-fall ecosystems: recent insights into ecology, paleoecology, and evolution. Annu. Rev. Mar. Sci. *7*, 571–596.

Snow, J.E., and Edmonds, H.N. (2007). Ultraslow-spreading ridges rapid paradigm changes. Oceanography *20*, 90–101.

Soriani, M. (2017). Unraveling *Neisseria meningitidis* pathogenesis: from functional genomics to experimental models. F1000Research *6*.

Sperling, E.A., Knoll, A.H., and Girguis, P.R. (2015). The ecological physiology of Earth's second oxygen revolution. Annu. Rev. Ecol. Evol. Syst. *46*, 215–235.

Stewart, F.J., and Cavanaugh, C.M. (2006). Bacterial endosymbioses in *Solemya* (Mollusca: Bivalvia)—model systems for studies of symbiont–host adaptation. Antonie Van Leeuwenhoek *90*, 343–360.

Stewart, F.J., Newton, I.L.G., and Cavanaugh, C.M. (2005). Chemosynthetic endosymbioses: adaptations to oxic–anoxic interfaces. Trends Microbiol. *13*, 439–448.

Stewart, F.J., Young, C.R., and Cavanaugh, C.M. (2008). Lateral symbiont acquisition in a maternally transmitted chemosynthetic clam endosymbiosis. Mol. Biol. Evol. *25*, 673–687.

Stewart, F.J., Ulloa, O., and DeLong, E.F. (2012). Microbial metatranscriptomics in a permanent marine oxygen minimum zone. Environ. Microbiol. *14*, 23–40.

Streams, M.E., Fisher, C.R., and Fiala-Médioni, A. (1997). Methanotrophic symbiont location and fate of carbon incorporated from methane in a hydrocarbon seep mussel. Mar. Biol. *129*, 465–476.

Suess, E. (2010). Marine cold seeps. In *Handbook of hydrocarbon and lipid microbiology*, K.N. Timmis, ed. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 185–203.

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

Swan, B.K., Martinez-Garcia, M., Preston, C.M., Sczyrba, A., Woyke, T., Lamy, D., Reinthaler, T., Poulton, N.J., Masland, E.D.P., Gomez, M.L., et al. (2011). Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. Science *333*, 1296–1300.

Taylor, J.D., and Glover, E.A. (2010). Chemosymbiotic bivalves. In *The vent and seep biota: aspects from microbes to ecosystems*, S. Kiel, ed. (Dordrecht: Springer Netherlands), pp. 107–135.

Thompson, J.N. (2005). Coevolution: The geographic mosaic of coevolutionary arms races. Curr. Biol. *15*, R992–R994.

Thrall, P.H., Hochberg, M.E., Burdon, J.J., and Bever, J.D. (2007). Coevolution of symbiotic mutualists and parasites in a community context. Trends Ecol. Evol. *22*, 120–126.

Tivey, M.K. (2007). Generation of seafloor hydrothermal vent fluids and associated mineral deposits. Oceanography *20*, 50–65.

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., and Gordon, J.I. (2007). The human microbiome project. Nature *449*, 804–810.

Ulloa, O., Canfield, D.E., DeLong, E.F., Letelier, R.M., and Stewart, F.J. (2012). Microbial oceanography of anoxic oxygen minimum zones. Proc. Natl. Acad. Sci. *109*, 15996–16003.

Van Dover, C.L., German, C.R., Speer, K.G., Parson, L.M., and Vrijenhoek, R.C. (2002). Evolution and biogeography of deep-sea vent and seep invertebrates. Science *295*, 1253–1257.

Van Dover, C.L., Arnaud-Haond, S., Gianni, M., Helmreich, S., Huber, J.A., Jaeckel, A.L., Metaxas, A., Pendleton, L.H., Petersen, S., Ramirez-Llodra, E., et al. (2018). Scientific rationale and international obligations for protection of active hydrothermal vent ecosystems from deep-sea mining. Mar. Policy *90*, 20–28.

Verna, C., Ramette, A., Wiklund, H., Dahlgren, T.G., Glover, A.G., Gaill, F., and Dubilier, N. (2010). High symbiont diversity in the bone-eating worm *Osedax* mucofloris from shallow whale-falls in the North Atlantic. Environ. Microbiol. *12*, 2355–2370.

Visick, K.L., and McFall-Ngai, M.J. (2000). An exclusive contract: specificity in the *Vibrio fischeri-Euprymna scolopes* Partnership. J. Bacteriol. *182*, 1779–1787.

Vos, M., and Eyre-Walker, A. (2017). Are pangenomes adaptive or not? Nat. Microbiol. *2*, 1576.

Vrijenhoek, R.C. (2010). Genetics and evolution of deep-sea chemosynthetic bacteria and their invertebrate hosts. In *The vent and Seep biota*, S. Kiel, ed. (Springer Netherlands), pp. 15–49.

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.

Wentrup, C., Wendeberg, A., Schimak, M., Borowski, C., and Dubilier, N. (2014). Forever competent: deep-sea bivalves are colonized by their chemosynthetic symbionts throughout their lifetime. Environ. Microbiol. *16*, 3699–3713.

Wernegreen, J.J. (2015). Endosymbiont evolution: Predictions from theory and surprises from genomes. Ann. N. Y. Acad. Sci. *1360*, 16–35.

Westhoek, A., Field, E., Rehling, F., Mulley, G., Webb, I., Poole, P.S., and Turnbull, L.A. (2017). Policing the legume-*Rhizobium* symbiosis: a critical test of partner choice. Sci. Rep. *7*, 1419.

Won, Y.-J., Hallam, S.J., O'Mullan, G.D., Pan, I.L., Buck, K.R., and Vrijenhoek, R.C. (2003). Environmental acquisition of thiotrophic endosymbionts by deep-sea mussels of the genus *Bathymodiolus*. Appl. Environ. Microbiol. *69*, 6785–6792.

Won, Y.-J., Jones, W.J., and Vrijenhoek, R.C. (2008). Absence of cospeciation between deep-sea mytilids and their thiotrophic endosymbionts. J. Shellfish Res. *27*, 129–138.

Yamamura, N. (1996). Evolution of mutualistic symbiosis: a differential equation model. Res. Popul. Ecol. *38*, 211–218.

Young, J.P.W., Crossman, L.C., Johnston, A.W., Thomson, N.R., Ghazoui, Z.F., Hull, K.H., Wexler, M., Curson, A.R., Todd, J.D., Poole, P.S., et al. (2006). The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. Genome Biol. *7*, R34.

Zhang, J.-Z., and Millero, F.J. (1993). The products from the oxidation of H2S in seawater. Geochim. Cosmochim. Acta *57*, 1705–1718.

Zhang, X., and Cui, L. (2016). Oxygen requirements for the Cambrian explosion. J. Earth Sci. *27*, 187–195.

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

Zimmermann, J., Wentrup, C., Sadowski, M., Blazejak, A., Gruber-Vodicka, H.R., Kleiner, M., Ott, J.A., Cronholm, B., De Wit, P., Erséus, C., et al. (2016). Closely coupled evolutionary history of ecto- and endosymbionts from two distantly related animal phyla. Mol. Ecol. *25*, 3203–3223.

Chapter II | Diversity matters

Diversity matters: Deep-sea mussels harbor multiple symbiont strains

Rebecca Ansorge[1,2], Stefano Romano[2#], Lizbeth Sayavedra[1#], Anne Kupczok[3], Halina E. Tegetmeyer[1,5], Nicole Dubilier[1,4]*, Jillian Petersen[1,2]*

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

[2]Department for Microbiology and Ecosystem Science, University of Vienna, Austria

[3]Christian-Albrechts-University Kiel, Germany

[4]MARUM, University of Bremen, Bremen, Germany

[5]Center for Biotechnology, Bielefeld University, Bielefeld, Germany

[#]These authors contributed equally

*Corresponding authors

*An earlier version of this manuscript has been deposited as preprint in BioRxiv under https://doi.org/10.1101/531459. The manuscript is **currently under review at Nature Microbiology** and the present version contains adjustments according to reviewer's comments.*

This chapter contains extended data and supplementary material in line with the journal format

Author contributions can be found in the end of manuscript, in line with the journal format.

Abstract

Genetic diversity of closely-related free-living microbes is widespread and underpins ecosystem functioning, but most evolutionary theories predict that it destabilizes intimate mutualisms. Indeed, symbiont strain diversity is assumed to be restricted in intracellular bacteria associated with animals. Here, we sequenced the metagenomes and metatranscriptomes of 18 *Bathymodiolus* mussel individuals from four species, covering their known distribution range at deep-sea hydrothermal vents in the Atlantic. We show that as many as 16 strains of intracellular, sulfur-oxidizing symbionts coexist in individual *Bathymodiolus* mussels. Co-occurring symbiont strains differed extensively in key metabolic functions, such as the use of energy and nutrient sources, electron acceptors and viral defense mechanisms. Most strain-specific genes were expressed, highlighting their adaptive potential. We show that fine-scale diversity is pervasive in *Bathymodiolus* symbionts, and hypothesize that it may be widespread in low-cost symbioses where the environment, rather than the host, feeds the symbionts.

Introduction

Within-species variability is ubiquitous in natural bacterial populations and occurs at many levels, from single nucleotide polymorphisms (SNPs) to differences in gene content and regulation[1,2]. These fine-scale differences can have major functional consequences and thus define microbial lifestyles. For example, a single regulatory gene or a mutation can dramatically alter the host range of bacterial symbionts and human pathogens[3,4]. In the human gut microbiome, gene copy number variation

among different strains of the same bacterial species is linked to host disease[5]. However, many of these functional differences are invisible at the level of marker genes commonly used in microbiome studies, such as the gene encoding 16S rRNA.

In free-living microbial communities, diversity underpins ecosystem functioning and resilience[6,7]. However, in symbiotic associations, intra-specific genetic diversity of microbes within host individuals can destabilize relationships between hosts and their symbionts, because if more than one symbiont strain co-exist, those that contribute less to the symbiosis, and more to their own proliferation would necessarily out-compete the more cooperative partner that directs more of its resources towards the host[8]. These inherent evolutionary conflicts can be prevented by restricting symbiont strain diversity through mechanisms such as vertical transmission, partner choice and sanctioning, and discriminating against 'low quality' partners[9–11]. Consideration of these conflicts is a central theme of evolutionary theory of mutualisms, as they explain why intra-specific diversity of symbionts is so limited in a range of associations including the well-known aphids with their *Buchnera* endosymbionts and legume nodules that contain only a single strain of rhizobial symbiont.

Until recently, most studies of intra-specific variation relied on sequencing one or a few marker genes. Metagenomics has revolutionized our ability to detect intra-specific variation across entire genomes in uncultured microbes, revealing that this variation has been underestimated[12]. So far, few studies have considered intra-specific variation of symbiont populations within individual hosts sampled from the wild. In contrast to the well-known intracellular symbioses such as *Buchnera*, symbiont strain diversity is common in the human gut microbiome, where the symbionts occur extracellularly and thus come into frequent contact with external

DNA, other microbes and viruses. Symbiont strain variability within hosts may be influenced by factors such as the diversity of the colonizing population, transmission mode, effective population size, and intra- or extracellular location of the symbionts[13,12,1,14]. But does such within-species symbiont diversity incur a cost to the host? Evolutionary theories predict that higher diversity can lead to conflicts among symbionts residing in a single host, however, it could be beneficial if these fine-scale genetic differences are reflected in functional differences between even very closely-related symbiont strains that would prevent them from competing for the same niche[15,16]. It is still poorly understood under which conditions intra-specific symbiont diversity is beneficial to hosts, and efforts to understand the evolutionary implications of complex host-associated communities are in their infancy[17,18].

Metagenomes are essential for understanding natural within-species diversity, how such diversity evolves, and how it affects function, particularly in uncultivable organisms. However, teasing apart highly similar strain genomes in metagenomes remains a major challenge[19–21]. Deep-sea *Bathymodiolus* mussels are ideal for investigating the functional and evolutionary implications of symbiont strain diversity, as they host only two bacterial symbiont species: One sulfur-oxidizing (SOX), and one methane-oxidizing (MOX) symbiont[22–24]. These symbionts co-occur inside specialized gill epithelial cells called bacteriocytes and use reduced compounds from hydrothermal fluids as energy sources for carbon fixation. The symbionts thus provide their hosts with nutrition in the nutrient-poor deep sea, allowing these mussels to dominate hydrothermal vent and cold seep communities worldwide[22–24]. *Bathymodiolus* juveniles acquire their symbionts horizontally[25–29]. However, it is unclear whether *Bathymodiolus* symbionts are taken up throughout the

mussel's lifetime or only during a permissive window early in the host development, imposing a bottleneck for the symbiont[30].

The SOX symbionts of *Bathymodiolus* are very closely related to a ubiquitous group of free-living bacteria called SUP05, and their symbioses with deep-sea mussels have likely evolved multiple times from within the SUP05 clade[31]. With few exceptions, each *Bathymodiolus* host harbors a single 16S SOX symbiont phylotype[26,32]. However, studies of the more variable ribosomal internal transcribed spacer indicated that more than one symbiont strain may colonize individual mussels[28,29]. Metagenomics of one *Bathymodiolus* species recently showed that 'subpopulations' of SOX symbionts differed in key functions such as hydrogen oxidation and nitrate respiration[33]. These observations raise a number of questions: How widespread is strain diversity, how many strains co-exist in a host individual, how is such fine-scale diversity stably maintained in symbiosis over evolutionary time[34], and is strain variation linked to functional variation? To address these questions, we performed high-resolution metagenomic and metatranscriptomic analyses of the symbiont populations of 18 host individuals from four *Bathymodiolus* species that were collected from four geochemically distinct, hydrothermal vents along the Mid-Atlantic Ridge.

Results and Discussion

Genome-wide symbiont heterogeneity

We sequenced metagenomes of five mussel individuals from the hydrothermal vent fields Lucky Strike, Lilliput and Clueless, three individuals from the vent field

Semenov and one individual from the vent field Wideawake. We assembled Illumina metagenomes and used differential coverage and contig connectivity data to retrieve a consensus reference genome of the *Bathymodiolus* SOX symbiont for each vent field and host species (only one host species was found at each vent field, see Methods) (**Fig. 1**). The symbiont reference genomes ranged from 2 to 3 Mbp and were ≥ 94% complete (**Tab. S1**). In 12 out of 18 host individuals we did not detect any SNPs in the symbiont 16S rRNA genes. In the other six, we detected low-frequency SNPs, present in 8-16% of the symbiont population and some SNPs appeared in more than one individual (**Extended Data Tab. 1**). This finding supports a previous study detecting low-abundance SOX 16S rRNA phylotypes in some host individuals that are closely related to the known *Bathymodiolus* symbionts (> 98.8% similarity)[24,29].

Heterogeneity in symbiont populations of individual mussels was 1 to 3 SNPs/kbp in the core genome, defined as the set of genes shared among the symbionts from all vent fields, and 5 to 11 SNPs/kbp in entire genome bins (**Fig. S1**, **Extended Data Fig. 1**). Heterogeneity was remarkably consistent in symbiont populations of different mussel individuals from the same vent field, but differed considerably between fields.

Compared to previous studies[e.g. 35,36] the high intra-specific variability we observed is unexpected in intracellular symbionts. Other sulfur-oxidizing intracellular symbionts, from *Solemya* clams and *Ridgeia* tubeworms, which were also sequenced with Illumina, had polymorphism rates one order of magnitude lower than in *Bathymodiolus* (**Extended Data Tab. 2**). The *Bathymodiolus* SOX symbionts had polymorphism rates more similar to those of human gut bacteria, which are 7-18 SNPs/kbp in individual microbial species within single host individuals[14]. This

similarity is unexpected as, in contrast to the intracellular SOX symbiont, most human gut microbes are extracellular, have a heterotrophic metabolism and frequently come into contact with a myriad of diverse microorganisms and bacteriophages within the gut, allowing rampant gene exchange[37,38]. The polymorphism rates in the SOX symbionts were also of the same order of magnitude as those observed in subpopulations of *Prochlorococcus*, the most abundant free-living bacterium in the ocean[1,39].

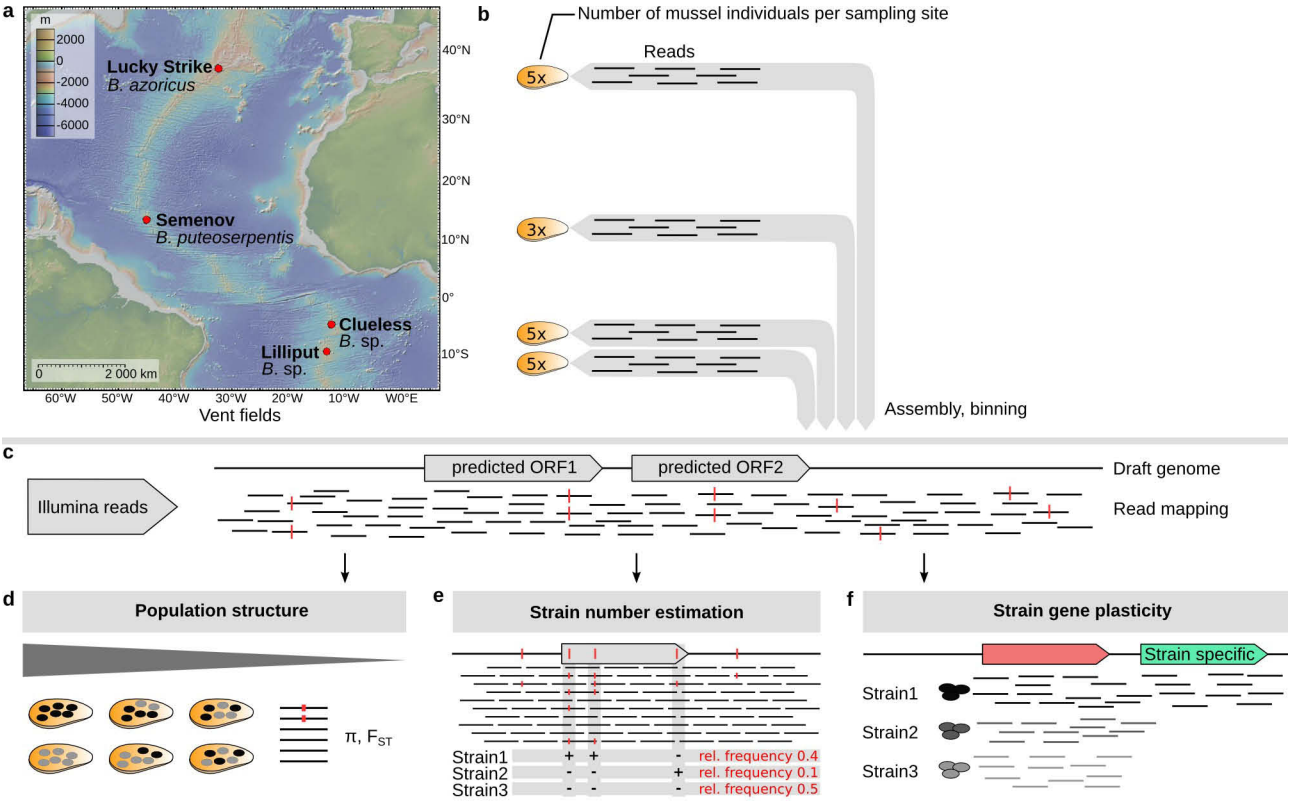


Figure 1 | Overview of the workflow developed for this study. (**a**) *Bathymodiolus* mussels were sampled at four vent fields along the Mid-Atlantic Ridge (MAR), (**b**) metagenomes of the sulfur-oxidizing symbiont (SOX) were assembled as a consensus for each site, binned and annotated, (**c**) for each sample, reads were mapped to per-site consensus draft genomes. We analyzed three aspects of symbiont strain diversity: (**d**) symbiont population structure by single nucleotide polymorphism (SNP) calling and the population genetic measures nucleotide diversity π and population differentiation F_{ST}, (e) estimation of strain numbers by gene version reconstruction based on SNP clusters partitioned according to their linkage by read overlap and relative frequency of local haplotypes (tool ViQuaS[84], see methods for details), and (**f**) differences in gene content among symbiont strains using read coverage information.

Population genomic insights into transmission and infection

Symbiont transmission mode influences heterogeneity, with vertically transmitted symbionts often displaying less heterogeneity than symbionts that are acquired horizontally[40]. Consistent with our findings of extensive SNP heterogeneity, symbiont nucleotide diversity π was 10 to 100 times higher in single *Bathymodiolus* mussels compared to *Solemya* clams[41]. Unlike *Solemya* symbionts that are predominantly vertically transmitted, *Bathymodiolus* juveniles acquire their symbionts horizontally[25–29]. Even in horizontally-transmitted symbioses, symbiont heterogeneity may be driven by bottleneck effects if symbionts are acquired only during a short period during the host's development, such as in *Ridgeia* tubeworms[42]. In fact, *Ridgeia* symbiont populations had little heterogeneity, similar to *Solemya* symbionts. Assuming genetic heterogeneity in the free-living stage of symbionts, symbiont populations within individual hosts would be genetically isolated from each other, reminiscent of population dynamics in vertically transmitted symbionts (**Extended Data Fig. 2**). To test if this is the case, we compared the nucleotide diversity of the core genome within host individuals (π_{within}) to that between hosts (pairwise, $\pi_{between}$). Principal component analysis (PCA) and a PERMANOVA test on pairwise Bray-Curtis dissimilarities comparing π_{within} to $\pi_{between}$ revealed that there was no significant difference between π values of hosts from the same vent field, whereas π_{within} differed significantly between vent fields (**Fig. 2**, **Tab. S2**, **S3**, **S4**, **Extended Data Fig. 3**). This indicates intermixed symbiont populations among co-occurring hosts, rather than genetic isolation of the symbiont populations within individual hosts (**Fig. 2**, **Extended Data Fig. 2**). Moreover, the fixation index (F_{ST}), a measure of population differentiation[43,44] expressed as values between 0 (no differentiation) and 1 (complete differentiation), was mostly low in pairwise comparisons of individuals sampled from

one vent field (0.04-0.24) (**Fig. 2**, **Extended Data Fig. 4**). The low F_{ST}-values between symbiont populations from different host individuals deviate substantially from values observed in bacterial populations characterized by clear differentiation. In fact, in the human gut microbiome inter-individual F_{ST} values were 0.4 among most similar and 0.8 among all individuals[14] and in the marine, free-living cyanobacterium *Prochlorococcus* F_{ST} values were ~ 0.7 between subpopulations[1]. This indicates that, despite its intracellular lifestyle, *Bathymodiolus* symbionts exhibit a higher degree of intermixing than is reported for other well studied systems with extracellular or free-living microbial populations of similar heterogeneity levels. Low genetic heterogeneity across symbiont populations from the same vent field supports a model of intermixed symbiont populations. Low F_{ST} between mussels further implies that mussels sample the entire environmentally available strain diversity, as stochastic subsampling, as would for example be expected if symbionts are taken up during a short developmental window, would increase F_{ST} between individuals. Together, our nucleotide diversity analyses thus indicate that in addition to self-infection of new gill tissues, *Bathymodiolus* symbionts may be repeatedly or continuously acquired from the environment throughout the host's lifetime, confirming an earlier study based on morphological observations of continuous symbiont uptake in *Bathymodiolus*[30]. Constant filtering of surrounding seawater through the symbiont-bearing gills would transport abundant symbiont cells released by dense surrounding mussel beds and lead to the colonization of young aposymbiotic tissue.

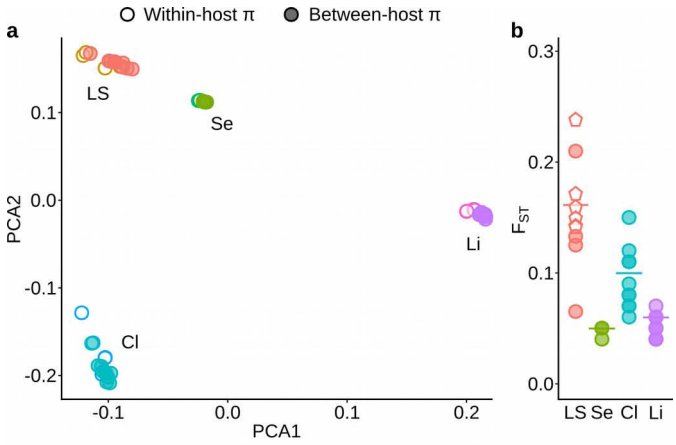


Figure 2 | Population genetic measures π and F_{ST} show that mussels from the same site host similar symbiont populations. (**a**) Principle component analysis (PCA) of π-values (nucleotide diversity) within and in pairwise comparison between individuals for core genes of the SOX symbiont in *B.* spp from the vent fields Lucky Strike (LS), Semenov (Se), Clueless (Cl) and Lilliput (Li). Filled circles represent pairwise π-values between two hosts; empty circles represent within-host π-values. π-values cluster according to vent field but no sub-clusters appear to separate within- and between-host π-values. This is confirmed by a PERMANOVA analysis on pairwise Bray-Curtis dissimilarities: no significant difference (Pseudo-F < 1.5, P > 0.2) between within-host and pairwise between-host π; significant difference between within-host π among fields (Pseudo-F > 85, Pr < 0.001) (see Tab. S3, S4). (**b**) F_{ST}-values: pairwise (symbols) and mean (line) across all host individuals per site. For vent site LS, circles represent host pairs from the same vent field, open pentagonal symbols represent host pairs from two different sampling sites that are separated by approx. 150 m. At the LS vent field, a few mussel pairs showed elevated F_{ST}-values, which could be explained by environmental differences between the two collection sites (discussed in Supplement).

Distinct symbiont strains co-exist in single host individuals

Understanding the true level of strain diversity in natural populations is a fundamental challenge in microbial ecology. To quantify strains, SNPs must be linked across genes or, if possible, entire genomes. The most sensitive ‘marker gene’ for resolving strain variability is the one that evolves most rapidly, but this is unlikely to be the same gene in all natural populations[45]. Therefore, we consider each distinct sequence of any coding gene to represent a different strain. We used more than 200 gammaproteobacterial single-copy marker genes to determine the maximum number of versions of each of these 200 genes, in each metagenome. Furthermore, we also

analyzed all genes that had coverages similar to those of these single-copy marker genes, and were therefore likely present in all strains within the population. We considered a single well-supported SNP sufficient to distinguish different strains (see Methods and Supplement section 1.5).

Both approaches produced similar results, detecting up to 16 versions of the most variable symbiont genes within single mussel individuals (**Fig. 3**, **Extended Data Fig. 5**). To investigate whether sequencing depth influenced estimated strain numbers, we repeated our analyses after down-sampling the reads to the lowest coverage found in our libraries (100x; **Tab. S1**). This reduced the estimated strain numbers to 4-9 per host individual, showing that read coverage influenced our results (**Fig. 3**).

We validated our approach for estimating strain numbers by analyzing a test dataset with simulated reads from 10 published *Escherichia coli* strains with 1% genetic heterogeneity, similar to that of the *Bathymodiolus* symbionts (**Tab. S5**). In this test dataset, read coverage also affected estimated strain numbers: these were underestimated at 100x coverage but were closest to accurate numbers at 300x coverage (see Supplement section 2.2; **Fig. S3**). Our estimate of 16 co-occurring SOX strains, from a library with 373x coverage, may therefore be realistic. We could further confirm the accuracy of our approach with long PacBio reads of a *B.* sp. individual sampled at the vent field Wideawake. We detected a maximum number of 11 distinct contigs containing the same single-copy gene, which was similar to the 12 strains we estimated using Illumina reads from the same individual (**Fig. 4**). Taken together, these analyses support our conclusion that, as few as 4 to 9, but as many as 16 symbiont strains co-occured within single *Bathymodiolus* individuals. Increasing

the sequencing depth may reveal an even larger number of strains (**Fig. S2**). These results are surprising, as a very low level of symbiont diversity was previously assumed to be typical for these hosts based on commonly used marker genes[23,24]. Nucleotide polymorphisms in the different gene versions may not all lead to different phenotypes for this particular gene, however, these polymporphisms are likely linked to other genomic changes including gene content variation[1] which likely result in phenotypic differences among the strains.

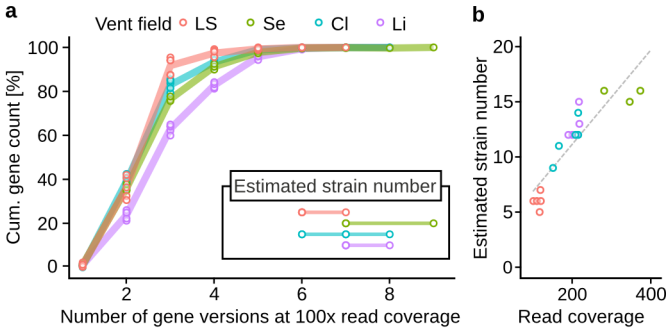


Figure 3 | Gene version reconstruction reveals up to 16 co-occurring SOX symbiont strains in individual *Bathymodiolus* mussels. (**a**) Cumulative count shows how many genes resulted in a specific number of reconstructed gene versions. This was performed for a set of 584 to 941 genes that had a read coverage within the coverage range of gammaproteobacterial marker genes, indicating that each strain in the population encoded these. Each line represents the average cumulative gene counts across all individuals from a site and each circle represents the gene count of a single individual. These plots reveal the spectrum of variability in SOX symbiont genomes – for each gene, there were between 1 and 9 different versions in populations of single host individuals at a read coverage of 100x. The gene with the most variation, and therefore the most versions, gives the most sensitive estimate for the number of strains that may co-exist in one mussel individual. Estimates of the number of co-existing strains are shown in the inset – these are ranges of estimates, derived from the maximum number of gene versions for each host individual. (**b**) Strain numbers were estimated with full read coverage ranging from 100 to 370x, revealing that up to 16 strains can co-exist in a single host individual. The sensitivity of strain detection correlates with read coverage (Spearman correlation: rs = 96, p = $4x10^{-10}$). LS: Lucky Strike (*B. azoricus*), Se: Semenov (*B. puteoserpentis*), Cl: Clueless (*B.* sp.), Li: Lilliput (*B.* sp.).

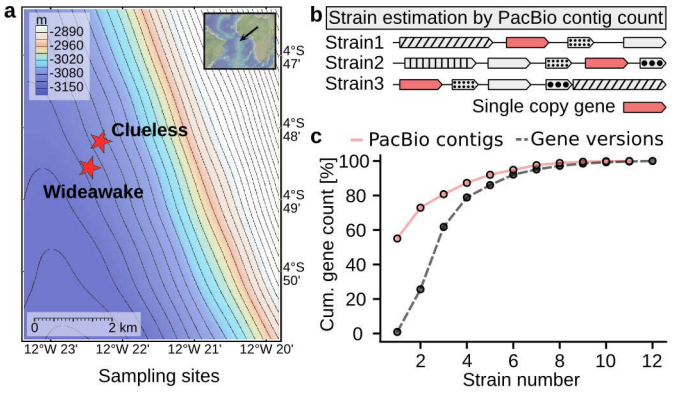


Figure 4 | Strain number estimates from PacBio sequencing confirms strain estimation approach from gene version reconstruction of Illumina sequences. (**a**) To verify our strain number estimation workflow, we obtained long read PacBio sequences and Illumina sequences from a single *B.* sp. individual from the Wideawake vent field (730 m from Clueless). (**b**) PacBio sequences revealed genome rearrangements around phylogenetic marker genes. (**c**) PacBio contigs and Illumina gene version reconstructions result in similar estimates of 11 and 12 strains, respectively. Continuous red line: number of PacBio contigs containing the same single-copy genes, dashed line: number of gene versions based on Illumina sequences and plotted as in Fig. 3a.

From the pangenome to the environment: Habitat chemistry drives symbiont genome heterogeneity

Understanding the geochemical environment experienced by deep-sea organisms is challenging. In addition, the relative availability of potential energy sources can be more important than absolute availability in determining which microbial energy-generating processes are most favorable[46]. We compared symbionts from vent fields with different environmental conditions, an ideal natural experiment for investigating potential links between strain diversity and the environment. We developed a bioinformatic pipeline that used metagenomic read coverage to identify differences in gene content among co-occurring strains in our dataset of four host species from geographically and geochemically distinct vent fields. Due to uneven DNA replication rates across the entire genome, even single-copy genes encoded by all strains have a range of coverages in metagenomes[47]. To define this range, we calculated the coverage of known, single-copy gammaproteobacterial genes in each metagenome (**Fig. S4**). Genes with coverage values below this range were likely only encoded by a subset of the population, and were thus considered strain-specific.

Between 30 and 50% of all genes in symbiont populations from individual mussels were potentially strain-specific, indicating massive differences in the gene contents of co-occurring strains (**Extended Data Tab. 3**). The functions of proteins encoded by the strain-specific genes differed markedly between the four vent fields, but within a field, these were mostly consistent among host individuals (**Tab. S6**, **Fig. 5**). With few exceptions, all strain-specific genes with annotated functions could also be detected in metatranscriptomes, suggesting that differences in gene content between

different strains resulted in functional differences that likely influence the fitness of symbionts and host (see Supplement section 2.3 for details, **Tab. S6**).

More than 80% of the strain-specific genes encoded hypothetical proteins with unknown functions. The strain-specific genes that could be annotated, encode proteins involved in functions such as synthesis of cell-surface components, environmental phosphate (P_i) sensing and acquisition, cell-cell interactions and phage defense (**Fig. 5**, **Extended Data Fig. 6**, **Extended Data Fig. 7**, **Tab. S6**). Hydrogen oxidation and nitrate reduction genes were also strain-specific in Mid-Atlantic Ridge populations, as shown previously in *B. septemdierum* from the West Pacific (**Fig. 5**)[33]. Some of the strain-specific symbiont genes may provide a selective advantage depending on the vent environment. For example, all mussels from vent fields with the highest hydrogen concentrations had a larger proportion of strains encoding hydrogenases, and those from fields with the lowest concentrations had the smallest proportion of strains encoding these enzymes (see Supplement 2.3). Ikuta *et al*.[33] also found differences in the relative proportions of strains that could oxidize hydrogen in a single *Bathymodiolus* species sampled from two vents. However, as most individuals sampled from one field were small juveniles, and most collected from the second field were adults, it was unclear whether this reflected site-specific differences in hydrogen availability, or changes during host development.

Genes involved in phosphate metabolism were another example of strain-specific variability that could provide a selective advantage depending on vent conditions. These genes were in a single cluster and encoded the high-affinity phosphate transport system PstSCAB, the regulatory protein PhoU and the two-component regulatory system PhoR-PhoB[48]. In addition to phosphorous metabolism, PhoR-B can

also affect other functions such as secondary metabolite production and virulence[49–51]. Considering the key role of these genes in cellular metabolism, it is surprising that this gene cluster was only encoded by the entire population of symbiont strains in mussels from two vent fields (Fig. 5). These genes were not present in any of the symbiont strains from Lilliput mussels, and only in some symbiont strains of Lucky Strike mussels, as confirmed by read mapping against symbiont bins (**Fig. 5**). At most hydrothermal vents P_i concentrations are unknown. However, soluble P_i depends on iron concentrations, which are reported to vary substantially between vent fields, raising the possibility that environmental P_i availability drives the loss or gain of P_i-related genes in symbiont populations (see Supplement section 2.3). Genes involved in P_i uptake and regulation were also strain-specific in *Prochlorococcus,* and their presence was linked to environmental P_i concentrations[52,53]. The SOX symbionts of vesicomyid clams and free-living relatives *Thioglobus* spp. from the SUP05 clade appear to also lack the PstSCAB genes based on our analyses of their published genomes (accession numbers of symbionts: JARW01000002, DDCF01000009, NC_009465, NC_008610; SUP05: CP010552, CP006911, CP008725, GG729964). However, to our knowledge no other bacteria have been described to miss both, the PhoR-B and PstSCAB systems. The symbiotic and free-living SOX bacteria that lack PstSCAB might encode unknown proteins that replace these missing functions, or use a low-affinity P_i-transporter to acquire P_i, as genes for these transporters were encoded in all of the analyzed genomes.

Oxygen concentrations fluctuate at vents due to dynamic mixing of anoxic hydrothermal fluids and oxygen-rich deep-sea seawater, and accordingly, mussel symbionts can use alternative electron acceptors such as nitrate[54–57]. Complete reduction of nitrate to dinitrogen gas (N_2) requires four enzymes: respiratory nitrate

reductase (Nar), nitrite reductase (Nir), nitric oxide reductase (Nor) and nitrous oxide reductase (Nos)[58]. In contrast to the genes needed for oxygen respiration, which were present in all symbiont populations, the prevalence of genes encoding all four steps of nitrate reduction to N_2 was highly variable among vent fields, among mussels from the same field, and even within symbiont populations of single mussels (**Fig. 5**, **Extended Data Fig. 6**). For example, in Lucky Strike individuals, the enzyme for the reduction of nitrate to nitrous oxide (N_2O) was encoded by 30 to 100% of the population, whereas the ability to perform the last step from N_2O to N_2 was not encoded at all. These variable abundances within symbiont populations suggest that each of the three steps of nitrate reduction to N_2O might be performed by a different subset of strains within a single host (**Fig. 5**, **Extended Data Fig. 6)**. The remarkable modularity of nitrate respiration genes in *Bathymodiolus* symbionts, as well as in other symbiotic and free-living bacteria[59,60], suggests that these genes are particularly prone to loss and gain. This raises the intriguing possibility that intricate interactions between microbes exchanging N intermediates are widespread in natural populations. Such a ‘division of labor’ may be beneficial as it could increase community productivity and avoid accumulation of intermediates[61,62].

Our results revealed that *Bathymodiolus* SOX symbionts have pangenomes with considerable functional diversity among co-existing strains, and this metabolic diversity may be linked to vent geochemistry. Together with our hypothesis of continuous uptake of symbionts throughout the mussel's lifetime, these results suggest constant symbiont strain shuffling between the environment and host as well as among co-occurring hosts. Given that the *Bathymodiolus* SOX symbiosis has evolved multiple times in convergent evolution from within the SUP05 clade[31], it is possible that some of this strain reshuffling might also involve ongoing evolutionary

emergence of new symbiont lineages from free-living SUP05 bacteria. This hypothesis would allow *Bathymodiolus* to associate with those strains that are best adapted to the vent environment. Rapid reshuffling of microbes has also been observed in other systems such as the human gut microbiota, where food intake has a direct and immediate effect on the microbial community[63]. At hydrothermal vents, exchange of symbiont strains would result in rapid adaptation of the metaorganism to local conditions within the lifetime of individual mussel hosts. Such genomic flexibility of the symbionts may underpin the productivity and global success of *Bathymodiolus* mussels in these ecosystems.

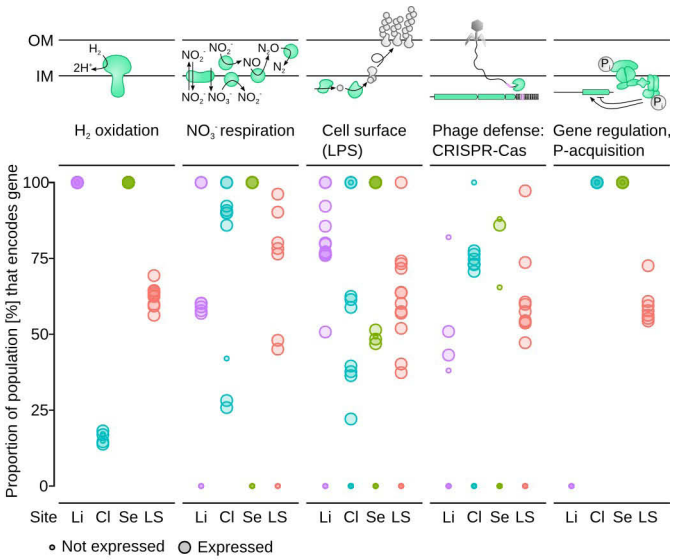


Figure 5 | Strain-specific genes encode potential key functions in SOX symbionts, including energy production and interactions with hosts and phages. The proportion of strains encoding these functions was different at each hydrothermal vent field. Large dots represent single genes that were detected in the transcriptomes and small dots represent genes that were not detectable in the transcriptomes. If the capability to perform a particular function is encoded by multiple genes, then it will have multiple points, e.g. hydrogen oxidation was encoded in a cluster containing multiple genes (see Tab. S6 for more details). The proportion of a population encoding each function was calculated as the average mean coverage (3 host individuals from Semenov and 5 host individuals for each of the other vent fields) compared to the mean coverage of genes encoded by the entire population (see Materials and Methods for details). If a gene had a coverage of 0, the gene was not encoded in the symbiont genome. Colors correspond to the four different sampling sites. Li: Lilliput, Cl: Clueless, Se: Semenov, LS: Lucky Strike.

A new model of evolutionary stability for one-to-many symbioses

Ecological theory predicts that if two different organisms share a limited resource, one will out-compete the other, unless mechanisms such as niche partitioning allow their stable co-existence[64,65]. Can these theories explain our results of co-existing strains, and how strain diversity and competition impact symbiosis stability? If competing symbionts differ in the net mutualistic benefit they provide, hosts can benefit by evolving mechanisms to differentially distribute costly resources to their partners. This can drive the evolution of specialized structures, such as compartments, with low symbiont diversity. In these cases, discrimination by hosts is important because a costly resource, such as photosynthate in the legume-*Rhizobia* symbioses, is provided[66]. But what if the symbiosis has low costs to the host?

Knowledge of costs and benefits of symbiotic associations is central to understanding their evolutionary trajectories. Beyond nutritional benefits gained through symbiont digestion[67], the benefits as well as the costs for *Bathymodiolus* mussels have not been extensively investigated. Possible costs include maintaining host-symbiont recognition mechanisms, transporting symbiont substrates into bacteriocyte vacuoles and waste products out, and dealing with toxic reactive oxygen species produced by symbiont metabolism.

In contrast to many well-characterized symbioses (e.g. see[66]), there is one substantial cost that *Bathymodiolus* does not have to bear - the cost of ‘feeding’ its symbionts. This is because the symbionts’ major energy sources come from the vent environment and not from the host itself. The SOX symbionts encode enzymes for the synthesis of

all 20 amino acids and 11 vitamins and cofactors, although they may require some metabolic intermediates from the host[68]. *Bathymodiolus* symbioses therefore more closely resemble byproduct mutualisms, where symbiotic partners benefit from metabolic byproducts that impose little or no cost to the producer[69]. Such low costs for the host, compared to associations where the symbionts rely entirely on the host for carbon or energy sources, would shift the balance between costs and benefits so that a greater range of symbiont strains, for example, even those that cannot make use of additional energy sources such as hydrogen, could still provide a net benefit to the host. Moreover, strain diversity has additional ecological and evolutionary benefits such as protection against bacteriophage attacks, and adaptation of the metaorganism to new and changing environments[70].

An increased benefit-to-cost ratio for the host can also decrease the incentive for bacterial partners to 'cheat'[71]. 'Cheating' is defined as using services provided by the host, and providing fewer or no services in return[69]. In the case of *Bathymodiolus*, the host would appear fully in control of the transfer of benefits from symbionts to the host. Regardless of whether symbionts share the products of carbon fixation immediately with their hosts through 'leaking' of small compounds, or whether these are primarily directed towards symbiont cell biosynthesis, intracellular digestion of symbiont cells ensures that all the products of symbiont primary production are eventually transferred to the host. Blocking intracellular digestion could be one way for the symbionts to cheat, but this may in turn be controlled by the host by sloughing off these 'defect' bacteriocytes, as is common for epithelia, and transporting them towards the mouth for digestion.

Finally, because the symbionts gain the bulk of their energy from the environment, instead of destabilizing the association as described by current evolutionary models, competition between different symbiont types could be beneficial for the host, if it results in the dominance of strains that more effectively transform geochemical energy in the vent environment into biomass and thus into host nutrition[15].

Conclusion

Our view of microbial diversity has long been shaped by a limited ability to accurately assess the enormous diversity of natural communities[72]. Metagenomics is rapidly changing this view, revealing that strain diversity has been vastly underestimated. Our study shows that strain diversity is pervasive in the sulfur-oxidizing symbionts of *Bathymodiolus* mussels. This diversity, invisible at the level of marker genes, is linked with functional diversity and likely has adaptive potential. Symbioses between corals and their intracellular photosynthetic algae are another prominent example where strain diversity may be common, although it is still unclear how much of this diversity is due to different gene copies within a single eukaryotic genome[16,73,74]. High symbiont diversity was also recently identified in the photosynthetic symbionts of marine protists[75].

This diversity has wide-ranging implications for the function and evolution of host-microbe associations. Despite this, it is currently not considered by most evolutionary theories, because these theories have been shaped by decades of study focused on models of symbiosis in which the host bears the enormous cost of ‘feeding’ the symbionts, and symbiont genetic diversity is highly restricted. We provide a new

theoretical framework that could explain the prevalence and evolutionary stability of strain diversity in beneficial host-microbe associations, where the environment provides for the symbionts' nutrition. This is the case for a diverse range of host-microbe associations from marine chemosynthetic and photosynthetic symbioses to the human digestive tract. Considering the substantial evidence that biodiversity underpins ecosystem stability, productivity, and resistance to invasion and parasitism[4], we predict that strain variation should be widespread in 'low-cost' associations such as these. Clearly, these hypotheses need testing and new concepts that extend evolutionary theories that were developed based on earlier studies of beneficial associations to a more united framework that can explain the wide range of host – microbe associations recent research is unveiling.

Methods

Sample collection

Four *Bathymodiolus* species from four vent fields were collected during three research cruises at hydrothermal vents along the Mid-Atlantic Ridge (MAR). Mussels from the same vent field belonged to the same host species based on their mitochondrial cytochrome c oxidase subunit I sequences. Symbiont-containing gill tissues were dissected from five mussel individuals from each of the following vent fields: Lucky Strike (site 'Montsegur' 37°17'19.1760"N, 32°16'32.0520"W; site 'Eiffel Tower' 37°17'20.8320"N, 32°16'31.7640"W), Clueless (4°48'11.7594"S, 12°22'18.4814"W) and Lilliput (9°32'47.6412"S, 13°12'35.0388"W). From these fields, samples were always dissected from the middle of each gill. From the

Semenov-2 field, gill pieces were dissected from the gill edges of three individuals (location 'Ash Lighthouse' 13°30'48.4812"N, 44°57'47.2788"W). One additional mussel individual was sampled at Wideawake (4°48'37.5599"S, 12°22'20.5201"W, 730 m from Clueless). From this individual, the whole gill was homogenized in a Dounce tissue grinder (Sigma, Germany) and a subsample used for DNA sequencing. For an overview of these locations and samples, see the map in **Fig. 1** and **Tab. S7**. Gill tissue pieces were either frozen directly at -80 °C or fixed in RNAlater according to the manufacturer's instructions (Sigma, Germany) and subsequently frozen at -80 °C.

Nucleic acid extraction and metagenome sequencing

DNA was extracted from gill pieces with commercially available kits (**Tab. S8**). RNA was extracted using the AllPrep kit (**Tab. S9**, Qiagen, Germany). From the symbiont homogenate from Wideawake, DNA was extracted according to Zhou *et al*.[76]. For each vent field, one reference SOX symbiont bin was produced from co-assemblies of metagenomes from multiple individuals as follows (see **Tab. S1** for reference genome statistics). Metagenomes were sequenced with Illumina or PacBio technology (see Supplement section 1.1 for details). Metagenomes were assembled from Illumina reads using IDBA-ud (v 1.1.1)[77] and SPAdes (v 3.2.2)[78], and genome bins were produced using a custom combination of differential coverage analysis with GBtools (v 2.4.5)[79] and contig connectivity analysis[80], and annotated with RASTtk[81] (see Supplement section 1.2 for details).

Transcriptome sequencing and analysis

Transcriptome reads were mapped to reference genomes with BBMap (v 36.x, Bushnell B. - BBMap - sourceforge.net/projects/bbmap/). The number of transcripts per gene was estimated with featureCounts[82]. Transcripts were normalized for different sequencing depths across libraries and for the gene length using edgeR with trimmed mean of M values (TMM) normalization[83,84].

SNP calling and population structure analysis

SNPs were called from reads of each individual sample mapped to the consensus symbiont bin and filtered, both performed with the Genome Analysis Toolkit (*GATK* v3.3.0; see Supplement section 1.3 for details)[85]. Rather than using the default settings for diploid genomes, we chose a ploidy setting of 10, as this better reflects a mixture of coexisting bacterial strains (**Tab. S10**). The symbiont population structure within and between host individuals was investigated by calculating nucleotide diversity π and the fixation index F_{ST} based on SNP frequencies and code is available on the github repository https://github.com/deropi/BathyBrooksiSymbionts/tree/master/Population_structure_analyses (see Supplement section 1.6 for details)[14].

Core genome calculation and detection of strain-specific genes

We developed a bioinformatic pipeline to identify strain-specific genes in metagenomes based on relative read coverage (see Supplement section 1.4 for details). Briefly, we defined the coverage range of genes that are encoded by each

strain in the population, based on single-copy gammaproteobacterial marker genes[86], and regarded all genes with coverage below this range as potentially strain specific. For some of these, multiple gene copies (coding sequences with the same annotation) were present in one metagenome. We excluded these genes from further analyses because it is possible that all strains encoded these, but that rearrangements led to different gene neighborhoods, causing these genes to fall on different contigs in the genome assemblies.

Strain number estimation and test simulation

We estimated the number of strains by using the number of gene sequence versions that could be reconstructed with the tool ViQuaS[87] as a proxy. *De novo* assembly of mapped reads partitions linked SNPs into clusters, frequencies of these local haplotypes resolve falsely clustered SNPs as well as connects SNPs that are too far apart for direct linkage (see supplementary material in Jayasundra et al., (2015)[87], for details). These distinct sequence versions were reconstructed for gammaproteobacterial marker genes from PhylaAmphora[86], as well as for all the genes encoded by each strain in the symbiont population in a single mussel (identified by read coverage, see above) using the tool ViQuaS (v 1.3)[87]. We created a test dataset with parameters that were similar to the sequencing data used in this study, by simulating Illumina reads from 10 publicly available *E. coli* genomes with ART (v 2.5.8)[88] (**Tab. S5**). Reads were pooled in even and uneven ratios to simulate different abundance patterns of strains in the population. Both datasets were analyzed with our strain estimation pipeline for two coverage depths 100x and 300x (see Supplement section 1.5 for details).

Code and data availability

Custom code is available on the github repository https://github.com/rbcan/MARsym_paper for detailed information of the computing steps. All sequencing reads and symbiont bins used in this study can be found at ENA under the accession number PRJEB28154.

Acknowledgements

We thank the captains, crews and ROV teams on the cruises BioBaz (2013), ODEMAR (2014), M78-2 (2009), Atalante Cruise Leg – 2 (2008) on board of the research vessels Pourquoi Pas?, FS Meteor and L'Atalante and the chief scientists François Lallier, Javie Excartin and Muriel Andreani, Richard Seifert and Colin Devey. Thank you to Adrien Assié, Christian Borowski, Corinna Breusing and Karina van der Heijden for sample treatment and fixation on board, and Målin Tietjen for the extraction of RNA from the samples of vent fields Semenov, Clueless and Lilliput. We also thank Christian Quast and Hanno Teeling for technical support. We thank Tal Dagan for the discussions and input during the project and on the written manuscript.

This study was funded by the Max Planck Society, the MARUM DFG-Research Center / Excellence Cluster “The Ocean in the Earth System” at the University of Bremen, the DFG CRC 1182 "Origin and Function of Metaorganisms", the German Research Foundation (RV Meteor M78-2 cruise), an ERC Advanced Grant (BathyBiome,

340535), and a Gordon and Betty Moore Foundation Marine Microbial Initiative Investigator Award to ND (Grant GBMF3811).

Contributions

R.A., J.P., L.S. and N.D conceived the study. R.A. and J.P. wrote the manuscript, with support from N.D., and contributions and revisions from all other co-authors. R.A. developed the metagenomic workflow for polymorphism detection, strain reconstruction, identification of strain-specific genes and analyzed the data with the exceptions described in the following. S.R. conducted the core-genome calculation, read simulation analyses, provided support for the statistical analyses and drafted respective manuscript sections. L.S. extracted nucleic acids for samples from Lucky Strike, Semenov and Wideawake, and conducted and evaluated the PacBio assembly. A.K. developed and provided an R-script for the calculation of π and F_{ST}. H.T. sequenced metagenomes from vent fields Clueless and Lilliput.

Extended Data

Extended Data Table 1 | Number of SNPs in the symbiont 16S rRNA gene within individual hosts at the four vent fields. The position on the 16S rRNA gene, SNP frequencies and nucleotide changes are indicated in light grey boxes. The following host species correspond to listed vent fields: Lucky Strike – *B. azoricus*, Semenov – *B. puteoserpentis*, Clueless – *B.* sp., Lilliput – *B.* sp.

	Lucky Strike		Semenov	Clueless		Lilliput	
Host individual 1	0		0	0		2*	1021 (8%) T>A
							1022 (8%) T>A
Host individual 2	1’	1018 (11%) G>A	0	0		2*	1021 (8%) T>A
							1022 (8%) T>A
Host individual 3	1’	1018 (16%) G>A	0	0		0	
Host individual 4	2’	1018 (13%) G>A	-	1	1036 (10%) T>C	0	
		930 (9%) C>T					
Host individual 5	0		-	0		0	

‘ 1 SNP is the same in all 3 individuals
* 2 SNPs are the same in both individuals

Extended Data Table 3 | Counts and percentages of low-coverage genes in within-host symbiont populations. Subsets represent counts that excluded all genes annotated as “hypothetical protein” and are further defined as strain-specific genes (one-copy genes with lower coverage in most or all symbiont populations from that site) and low-coverage genes with further copies in the genome. The following host species correspond to listed vent fields: Lucky Strike – *B. azoricus*, Semenov – *B. puteoserpentis*, Clueless – *B.* sp., Lilliput – *B.* sp.

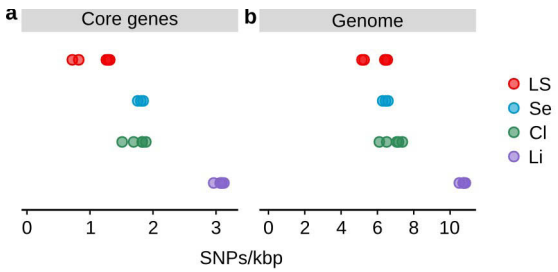
Vent site	Total low-coverage genes	w/o hypotheticals		
		Total low-coverage genes w/o hypotheticals	Strain-specific genes*	Genes with other copies in genome
Lucky Strike	1534-1685 (45-50%)	199 (5.8%)	101 (3.0%)	98 (2.8%)
Semenov	835-918 (33-36%)	60 (2.4%)	30 (1.2%)	30 (1.2%)
Clueless	929-1111 (30-36%)	132 (4.3%)	58 (1.9%)	74 (2.4%)
Lilliput	1442-1553 (42-45%)	122 (3.6%)	60 (1.8%)	62 (1.8%)

* detailed information in Tab S.6

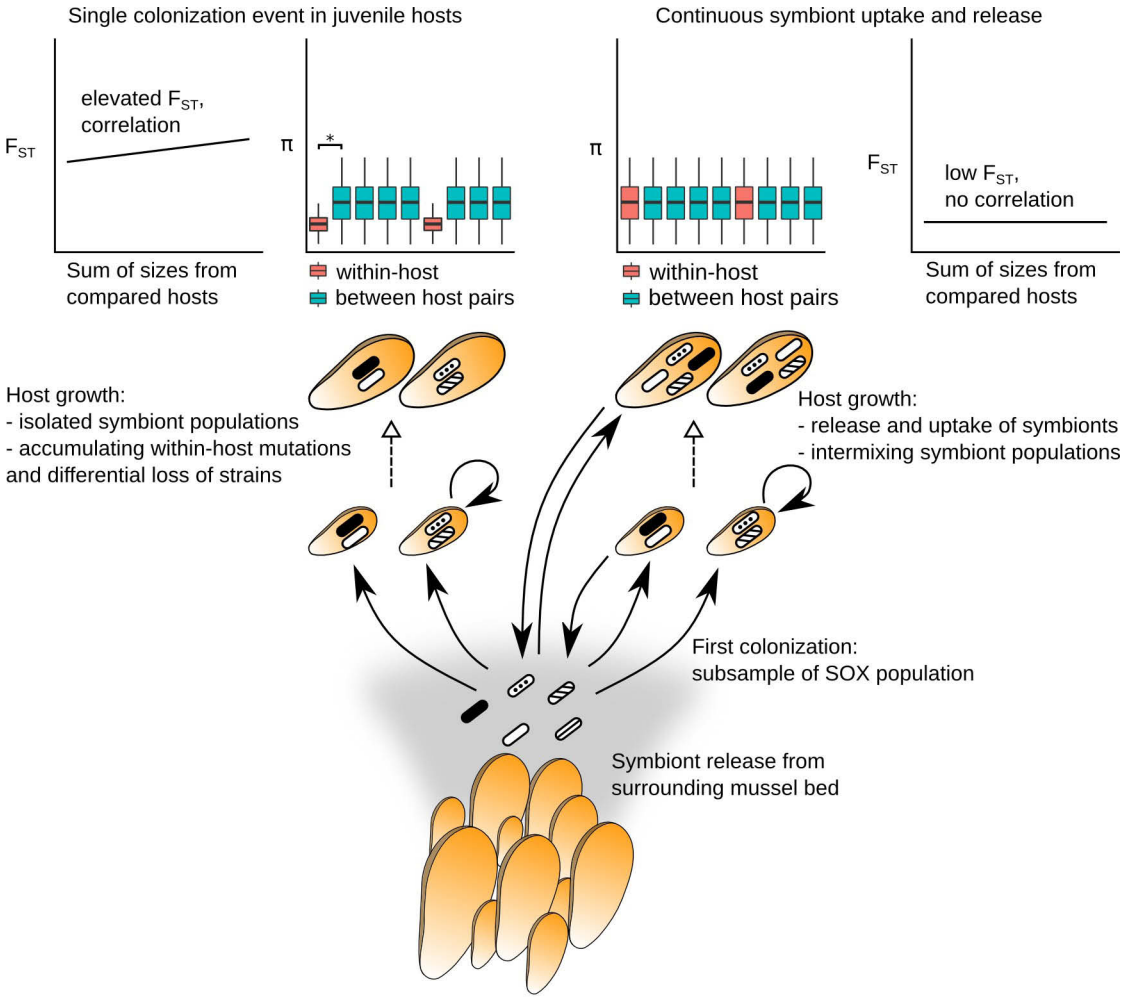
Extended Data Table 2 | Comparison of reported SNP densities from published studies to our present dataset. Data from *Bathymodiolus* SOX symbionts of this study are depicted in bold. Many of reported values were previously summed up in Wilmes *et al.* (2009). *NA* = information was missing or could not be retrieved.

Organism	SNPs/kb	Tool	Environment	Technology	Coverage	Reference
Ca. Accumulibacter phosphatis	0.01 (US) 0.02 (OZ)	Consed	Sludge bioreactor	*NA*	9.2-17.5 x (US) 5.36-7.68 x (OZ)	(Kunin et al., 2008)
Leptospirillum group II type UBA	0.04	*NA*	Acid mine drainage	Sanger	25 x	(Lo et al., 2007)
Kuenenia stuttgartiensis	0.07	*NA*	Anammox bioreactor	Sanger	22 x	(Strous et al., 2006)
Solemya velum endosymbionts	0.1-1*	Custom, GATK	Shallow water clam	Illumina	52-1115 x	(Russell and Cavanaugh, 2017)
Endoriftia symbionts in *Ridgeia piscesae*	0.3* 0.9*	VarScan GATK	Hydrothermal vent tubeworm	Illumina	175 x	(Perez and Juniper, 2017)
Leptospirillum group II type 5-way CG	0.9	Consed, Strainer	Acid mine drainage	Sanger	20 x	(Simmons, 2008)
Buchnera aphidicola endosymbiont	1.8*ω	Consed	Insects	Sanger	9 x (?)	(Ham et al., 2003)
'Iplasma'	2.7	*NA*	Acid mine drainage	*NA*	20 x	(Wilmes et al., 2009)
Endoriftia symbionts in *Riftia pachyptila*	2.9#	Custom	Hydrothermal vent tubeworm	Sanger	18.6 x	(Robidart et al., 2008)
'Eplasma'&	5.3	*NA*	Acid mine drainage	*NA*	10 x	(Wilmes et al., 2009)
***Bathymodiolus spp.* SOX symbiont**	**5 - 11***	**GATK**	**Hydrothermal vent mussel**	**Illumina**	**100 x**	**This study**
Prochlorococcus subpopulations	12	Custom	Free-living, marine	Illumina	*NA*	(Kashtan et al., 2014)
Gut microbiome	7-18*°	Custom	Human gut	Illumina	10-32 400 x	(Schloissnig et al., 2013)
Ferroplasma type II	22	Blastn	Acid mine drainage	Sanger	10 x	(Tyson et al., 2004)
Ferroplasma type I	30	Blastn	Acid mine drainage	*NA*	4.5 x	(Allen et al., 2007)
Archaeal virus contig from metagenome	70	*NA*	Yellowstone hot springs	Sanger	11 x	(Schoenfeld et al., 2008)
Archaeal virus AMDV2	270	Consed,	Acid mine drainage	*NA*	17.5 x	(Andersson and Banfield,

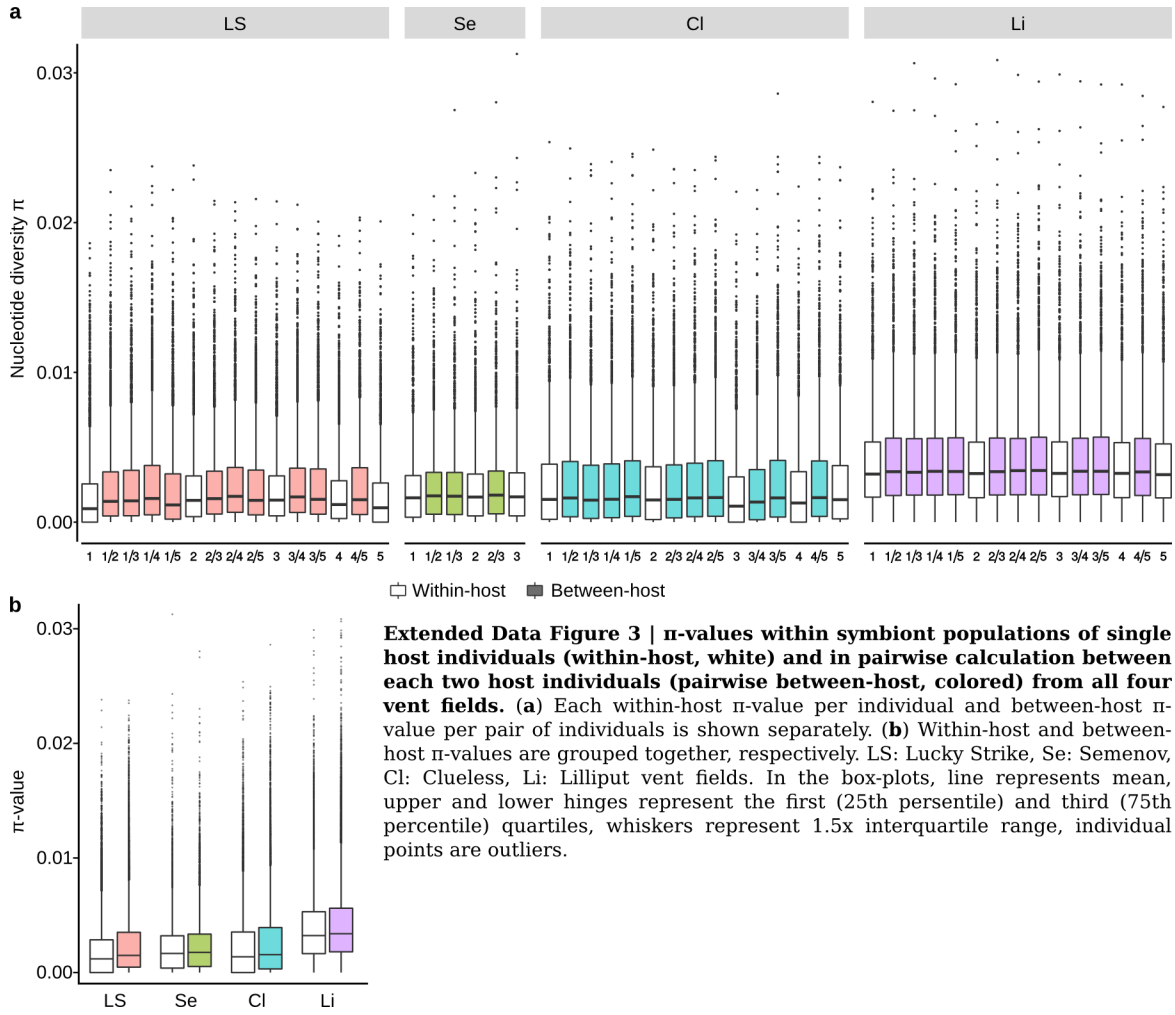
* in single host individuals
# gene subset
ω pooled host individuals
& partial assembly
° total polymorphism percentage of all sampled species per microbiome



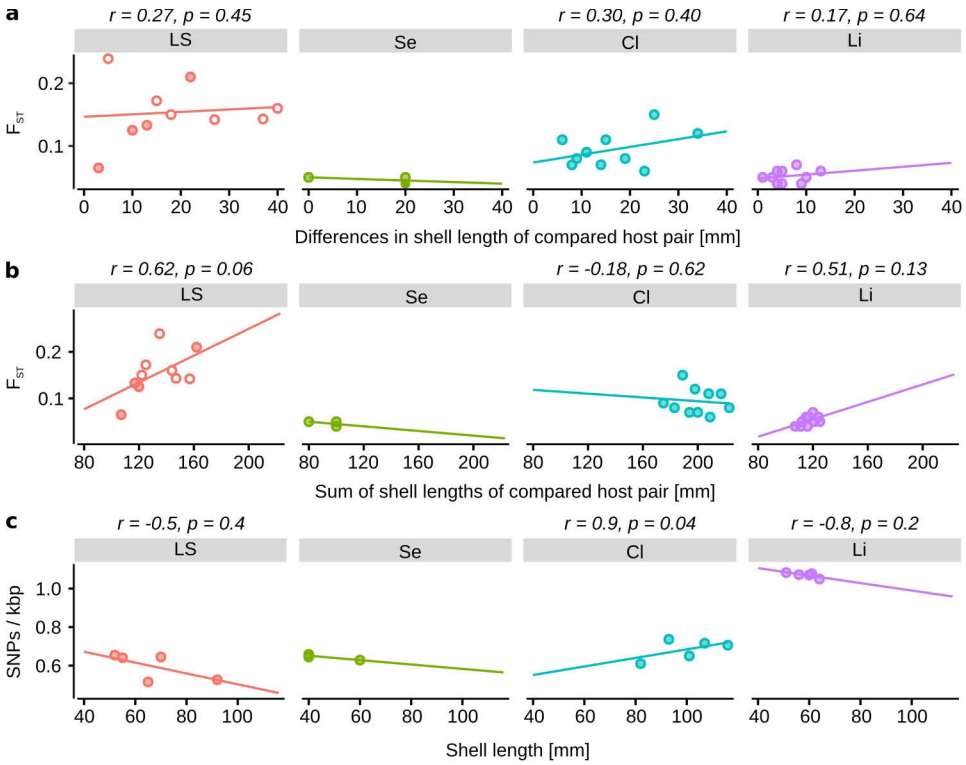
Extended Data Figure 1 | Single nucleotide polymorphisms (SNPs) of within-host symbiont populations in (a) core genes and (b) whole genome including non-coding regions. Vent fields (host species) are LS: Lucky Strike (*B. azoricus*), Se: Semenov (*B. puteoserpentis*), Cl: Clueless (*B.* sp.), Li: Lilliput (*B.* sp.).



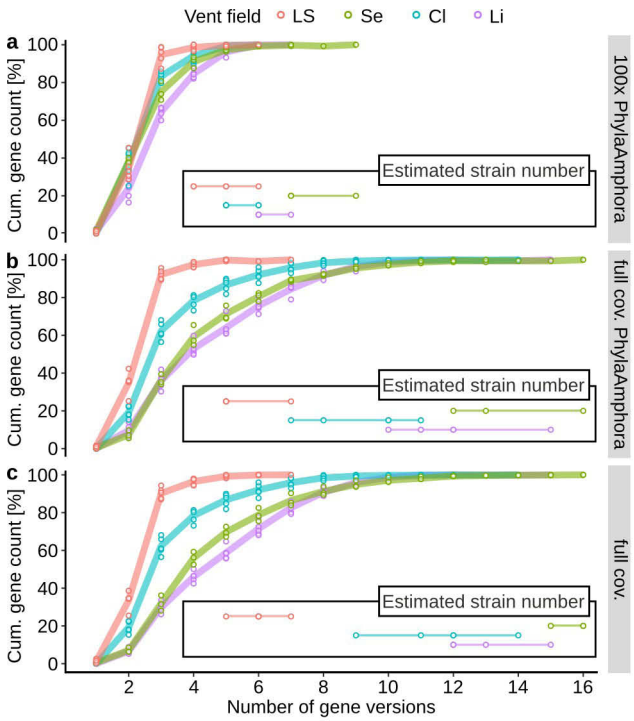
Extended Data Figure 2 | Theoretical model predicting the influence of symbiont transmission on population genomic signatures (π, F_{ST}). On the left a scenario is depicted in which the symbionts are acquired only once by juvenile mussels during a restricted time window, followed exclusively by repeated self-infection. On the right a scenario is depicted in which the symbionts are continuously released and taken up by host individuals throughout their lifetime.



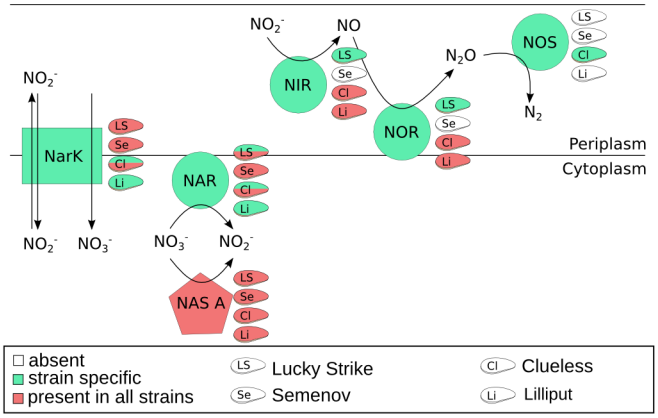
Extended Data Figure 3 | π-values within symbiont populations of single host individuals (within-host, white) and in pairwise calculation between each two host individuals (pairwise between-host, colored) from all four vent fields. (**a**) Each within-host π-value per individual and between-host π-value per pair of individuals is shown separately. (**b**) Within-host and between-host π-values are grouped together, respectively. LS: Lucky Strike, Se: Semenov, Cl: Clueless, Li: Lilliput vent fields. In the box-plots, line represents mean, upper and lower hinges represent the first (25th persentile) and third (75th percentile) quartiles, whiskers represent 1.5x interquartile range, individual points are outliers.



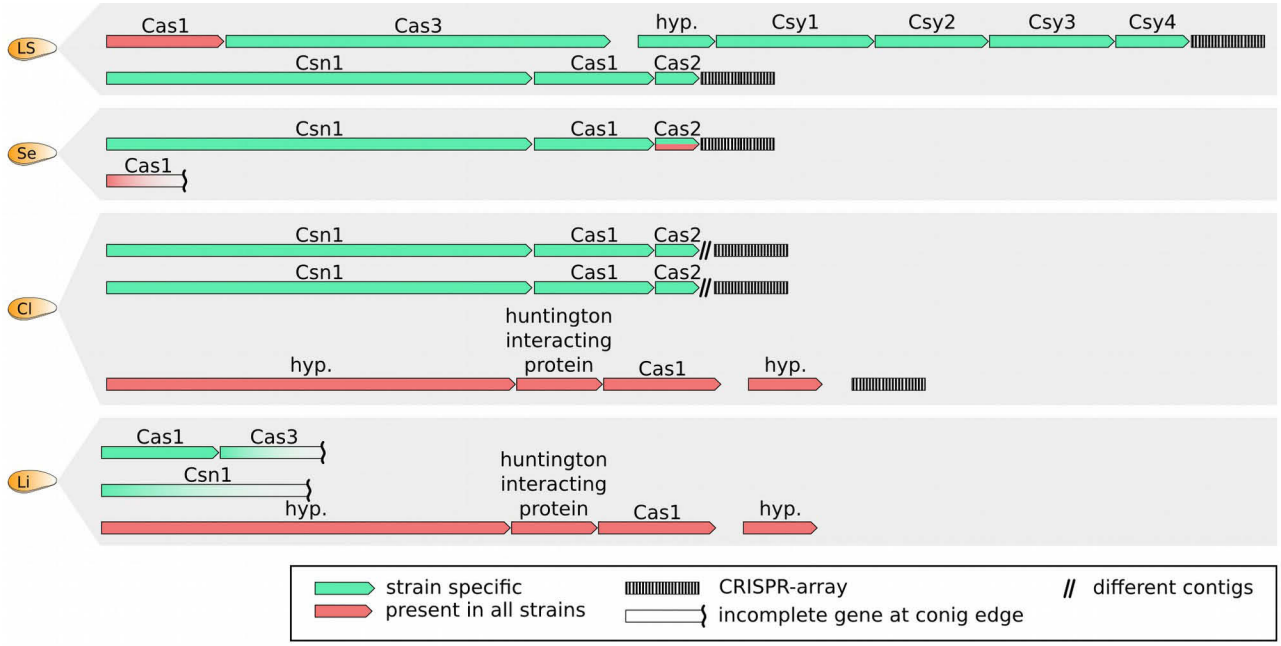
Extended Data Figure 4 | Spearman correlation between the difference (a) or sum (b) in shell lengths of two compared hosts with the pairwise F_{ST}; and correlation of shell length with intra-host SNP density (c). r = spearmans correlation coefficient rho, p = p-value; for vent field Lucky Strike white symbols = host pairs from different sites Eiffeltower and Montsegur, red symbols = host pairs from the same site; LS: Lucky Strike, Se: Semenov, Cl: Clueless, Li: Lilliput.



Extended Data Figure 5 | Cumulative gene counts of distinct numbers of gene versions on gammaproteobacterial marker genes from PhylaAmphora and the extended set of genes that had a read coverage within the coverage range of gammaproteobacterial marker genes, indicating that each strain in the population encoded these. Strain numbers were estimated for the marker gene set with 100x read coverage (**a**) and full read coverage (**b**) and for the entire gene set with full read coverage per host individual. Full read coverage was 100–120x for LS, 280–370x for Se, 150–215x for Cl and 190–218x for Li mussels. LS: Lucky Strike, Se: Semenov, Cl: Clueless, Li: Lilliput.



Extended Data Figure 6 | Representation of denitrification genes among strains of the SOX symbiont. A gene is absent (white mussel symbols), strain specific (green mussel symbols), or present in all strains (red mussel symbols) in a single host individual. When some host individuals from the same vent site had symbiont populations where a gene is strain specific and others where the entire population encoded that gene, mussel symbol is split into red and green color. LS: Lucky Strike (*B. azoricus*), Se: Semenov (*B. puteoserpentis*), Cl: Clueless (*B.* sp.), Li: Lilliput (*B.* sp.), NAR: respiratory nitrate reductase, NIR: nitrite reductase, NOR: nitric oxide reductase, NOS: nitrous oxide reductase, NarK: nitrate transporter, NAS A: assimilatory nitrate reductase.



Extended Data Figure 7 | Representation of CRISPR-Cas gene clusters in the SOX symbiont strains showing strain-specific genes (green), genes present in all strains (red), and CRISPR-arrays (striped boxes). LS: Lucky Strike, Se: Semenov, Cl: Clueless, Li: Lilliput.

References of chapter 2

1. Kashtan, N. *et al.* Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* **344**, 416–420 (2014).

2. Ackermann, M. Microbial individuality in the natural environment. *ISME J.* **7**, 465–467 (2013).

3. Pankey, M. S. *et al.* Host-selected mutations converging on a global regulator drive an adaptive leap towards symbiosis in bacteria. *eLife* **6**, e24414 (2017).

4. Viana, D. *et al.* A single natural nucleotide mutation alters bacterial pathogen host-tropism. *Nat. Genet.* **47**, 361–366 (2015).

5. Greenblum, S., Carr, R. & Borenstein, E. Extensive strain-level copy-number variation across human gut microbiome species. *Cell* **160**, 583–594 (2015).

6. Tilman, D., Isbell, F. & Cowles, J. M. Biodiversity and ecosystem functioning. *Annu. Rev. Ecol. Evol. Syst.* **45**, 471–493 (2014).

7. Hooper D. U. *et al.* Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. *Ecol. Monogr.* **75**, 3–35 (2005).

8. Frank, S. A. Host–symbiont conflict over the mixing of symbiotic lineages. *Proc R Soc Lond B* **263**, 339–344 (1996).

9. Sachs, J. L. *et al.* Host control over infection and proliferation of a cheater symbiont. *J. Evol. Biol.* **23**, 1919–1927 (2010).

10. Bulgheresi, S. *et al.* A New C-Type Lectin similar to the human immunoreceptor DC-SIGN mediates symbiont acquisition by a marine nematode. *Appl. Environ. Microbiol.* **72**, 2950–2956 (2006).

11. Nyholm, S. V. & McFall-Ngai, M. The winnowing: establishing the squid–*vibrio* symbiosis. *Nat. Rev. Microbiol.* **2**, 632–642 (2004).

12. Engel, P., Stepanauskas, R. & Moran, N. A. Hidden diversity in honey bee gut symbionts detected by single-cell genomics. *PLoS Genet.* **10**, e1004596 (2014).

13. Delmont, T. O. & Eren, A. M. Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. *PeerJ* **6**, e4320 (2018).

14. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013).

15. Batstone, R. T., Carscadden, K. A., Afkhami, M. E. & Frederickson, M. E. Using niche breadth theory to explain generalization in mutualisms. *Ecology* (2018).

16. Rowan, R. & Knowlton, N. Intraspecific diversity and ecological zonation in coral-algal symbiosis. *Proc. Natl. Acad. Sci.* **92**, 2850–2853 (1995).

17. Foster, K. R., Schluter, J., Coyte, K. Z. & Rakoff-Nahoum, S. The evolution of the host microbiome as an ecosystem on a leash. *Nature* **548**, 43–51 (2017).

18. Bongrand, C. *et al.* A genomic comparison of 13 symbiotic *Vibrio fischeri* isolates from the perspective of their host source and colonization behavior. *ISME J.* **10**, 2907–2917 (2016).

19. Quince, C. *et al.* DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol.* **18**, 181 (2017).

20. Cleary, B. *et al.* Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat. Biotechnol.* **33**, 1053–1060 (2015).

21. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).

22. Petersen, J. M. & Dubilier, N. Methanotrophic symbioses in marine invertebrates. *Environ. Microbiol. Rep.* **1**, 319–335 (2009).

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

24. Duperron, S. *et al.* 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 (2006).

25. Laming, S. R., Duperron, S., Cunha, M. R. & Gaudron, S. M. Settled, symbiotic, then sexually mature: adaptive developmental anatomy in the deep-sea, chemosymbiotic mussel Idas modiolaeformis. *Mar. Biol.* **161**, 1319–1333 (2014).

26. Duperron, S. The diversity of deep-sea mussels and their bacterial symbioses. in *The vent and seep biota* (ed. Kiel, S.) **33**, 137–167 (Springer Netherlands, 2010).

27. Won, Y.-J., Jones, W. J. & Vrijenhoek, R. C. Absence of cospeciation between deep-sea mytilids and their thiotrophic endosymbionts. *J. Shellfish Res.* **27**, 129–138 (2008).

28. DeChaine, E. G. & Cavanaugh, C. M. Symbioses of methanotrophs and deep-sea mussels (Mytilidae: Bathymodiolinae). in *Molecular basis of symbiosis* (ed. Overmann, P. D. J.) 227–249 (Springer Berlin Heidelberg, 2005).

29. Won, Y.-J. *et al.* Environmental acquisition of thiotrophic endosymbionts by deep-sea mussels of the genus *Bathymodiolus*. *Appl. Environ. Microbiol.* **69**, 6785–6792 (2003).

30. Wentrup, C., Wendeberg, A., Schimak, M., Borowski, C. & Dubilier, N. Forever competent: deep-sea bivalves are colonized by their chemosynthetic symbionts throughout their lifetime. *Environ. Microbiol.* **16**, 3699–3713 (2014).

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

32. Duperron, S. *et al.* 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 (2007).

33. Ikuta, T. *et al.* Heterogeneous composition of key metabolic gene clusters in a vent mussel symbiont population. *ISME J.* **10**, 990-1001 (2016).

34. Heath, K. D. & Stinchcombe, J. R. Explaining mutualism variation: a new evolutionary paradox? *Evolution* **68**, 309–317 (2014).

35. Perez, M. & Juniper, S. K. Is the trophosome of *Ridgeia piscesae* monoclonal? *Symbiosis* 1–11 (2017).

36. Russell, S. L. & Cavanaugh, C. M. Intrahost genetic diversity of bacterial symbionts exhibits evidence of mixed infections and recombinant haplotypes. *Mol. Biol. Evol.* **34**, 2747–2761 (2017).

37. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241–244 (2011).

38. Hehemann, J.-H. *et al.* Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**, 908–912 (2010).

39. McInerney, J. O., McNally, A. & O'Connell, M. J. Why prokaryotes have pangenomes. *Nat. Microbiol.* **2**, 17040 (2017).

40. Douglas, A. E. The ecology of symbiotic micro-organisms. in *Advances in ecological research* (eds. Begon, M. & Fitter, A. H.) **26**, 69–103 (Academic Press, 1995).

41. Russell, S. L., Corbett-Detig, R. B. & Cavanaugh, C. M. Mixed transmission modes and dynamic genome evolution in an obligate animal–bacterial symbiosis. *ISME J.* **11**, 1359 (2017).

42. Chaston, J. & Goodrich-Blair, H. Common trends in mutualism revealed by model associations between invertebrates and bacteria. *FEMS Microbiol. Rev.* **34**, 41–58 (2010).

43. Wright, S. *Evolution and the Genetics of Populations. The theory of gene frequencies.* **Vol. 2**, (University of Chicago Press, 1969).

44. Wright, S. Isolation by Distance. *Genetics* **28**, 114–138 (1943).

45. Lan, Y., Rosen, G. & Hershberg, R. Marker genes that are less conserved in their sequences are useful for predicting genome-wide similarity levels between closely related prokaryotic strains. *Microbiome* **4**, (2016).

46. Perner, M. *et al.* Linking geology, fluid chemistry, and microbial activity of basalt- and ultramafic-hosted deep-sea hydrothermal vent environments. *Geobiology* **11**, 340–355 (2013).

47. Brown, C. T., Olm, M. R., Thomas, B. C. & Banfield, J. F. Measurement of bacterial replication rates in microbial communities. *Nat. Biotechnol.* **34**, 1256 (2016).

48. Hsieh, Y.-J. & Wanner, B. L. Global Regulation by the seven-component Pi signaling system. *Curr. Opin. Microbiol.* **13**, 198–203 (2010).

49. Romano, S., Schulz-Vogt, H. N., González, J. M. & Bondarev, V. Phosphate limitation induces drastic physiological changes, virulence-related gene expression, and secondary metabolite production in *Pseudovibrio* sp. strain FO-BEG1. *Appl. Environ. Microbiol.* **81**, 3518–3528 (2015).

50. Santos-Beneit, F. The Pho regulon: a huge regulatory network in bacteria. *Front. Microbiol.* **6**, (2015).

51. Lamarche, M. G., Wanner, B. L., Crépin, S. & Harel, J. The phosphate regulon and bacterial virulence: a regulatory network connecting phosphate homeostasis and pathogenesis. *FEMS Microbiol. Rev.* **32**, 461–473 (2008).

52. Martiny, A. C., Huang, Y. & Li, W. Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions. *Environ. Microbiol.* **11**, 1340–1347 (2009).

53. Martiny, A. C., Coleman, M. L. & Chisholm, S. W. Phosphate acquisition genes in *Prochlorococcus* ecotypes: Evidence for genome-wide adaptation. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 12552–12557 (2006).

54. Zielinski, F. U., Gennerich, H.-H., Borowski, C., Wenzhöfer, F. & Dubilier, N. 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**, Q0AE04 (2011).

55. Kuwahara, H. *et al.* Reduced genome of the thioautotrophic intracellular symbiont in a deep-sea clam, *Calyptogena okutanii*. *Curr. Biol.* **17**, 881–886 (2007).

56. Hentschel, U., Hand, S. & Felbeck, H. The contribution of nitrate respiration to the energy budget of the symbiont-containing clam *Lucinoma aequizonata*: a calorimetric study. *J. Exp. Biol.* **199**, 427–433 (1996).

57. Hentschel, U., Cary, S. C. & Felbeck, H. Nitrate respiration in chemoautotrophic symbionts of the bivalve *Lucinoma aequizonata*. *Mar. Ecol. Prog. Ser.* **94**, 35–41 (1993).

58. Kraft, B., Strous, M. & Tegetmeyer, H. E. Microbial nitrate respiration – Genes, enzymes and environmental distribution. *J. Biotechnol.* **155**, 104–117 (2011).

59. Shah, V., Chang, B. X. & Morris, R. M. Cultivation of a chemoautotroph from the SUP05 clade of marine bacteria that produces nitrite and consumes ammonium. *ISME J.* **11**, 263–271 (2017).

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

61. Savage, V. M., Webb, C. T. & Norberg, J. A general multi-trait-based framework for studying the effects of biodiversity on ecosystem functioning. *J. Theor. Biol.* **247**, 213–229 (2007).

62. Lindemann, S. R. *et al.* Engineering microbial consortia for controllable outputs. *ISME J.* **10**, 2077–2084 (2016).

63. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).

64. Ghoul, M. & Mitri, S. The ecology and evolution of microbial competition. *Trends Microbiol.* **24**, 833–845 (2016).

65. Hardin, G. The Competitive Exclusion Principle. *Science* **131**, 1292–1297 (1960).

66. Udvardi, M. & Poole, P. S. Transport and metabolism in legume-*Rhizobia* symbioses. *Annu. Rev. Plant Biol.* **64**, 781–805 (2013).

67. Zheng, P. *et al.* Insights into deep-sea adaptations and host-symbiont interactions: a comparative transcriptome study on *Bathymodiolus* mussels and their coastal relatives. *Mol. Ecol.* **26**, 5133–5148 (2017).

68. Ponnudurai, R. *et al.* Metabolic and physiological interdependencies in the *Bathymodiolus azoricus* symbiosis. *ISME J.* **11**, 463-477 (2017).

69. Douglas, A. E. Conflict, cheats and the persistence of symbioses. *New Phytol.* **177**, 849–858 (2008).

70. Palmer, T. M. *et al.* Synergy of multiple partners, including freeloaders, increases host fitness in a multispecies mutualism. *Proc. Natl. Acad. Sci.* **107**, 17234–17239 (2010).

71. Foster, K. R. & Wenseleers, T. A general model for the evolution of mutualisms. *J. Evol. Biol.* **19**, 1283–1293

72. McLaren, M. R. & Callahan, B. J. In nature, there is only diversity. *mBio* **9**, e02149-17 (2018).

73. Wooldridge Scott A. Is the coral-algae symbiosis really ‘mutually beneficial’ for the partners? *BioEssays* **32**, 615–625 (2010).

74. Oppen, M. J. H. van, Palstra, F. P., Piquet, A. M.-T. & Miller, D. J. Patterns of coral–dinoflagellate associations in *Acropora*: significance of local availability and physiology of *Symbiodinium* strains and host–symbiont selectivity. *Proc. R. Soc. Lond. B Biol. Sci.* **268**, 1759–1767 (2001).

75. Brisbin, M. M., Mesrop, L. Y., Grossmann, M. M. & Mitarai, S. Intra-host symbiont diversity and extended symbiont maintenance in photosymbiotic *Acantharea* (clade F). *bioRxiv* 299495 (2018).

76. Zhou, J., Bruns, M. A. & Tiedje, J. M. DNA recovery from soils of diverse composition. *Appl. Environ. Microbiol.* **62**, 316–322 (1996).

77. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).

78. Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

79. Seah, B. K. B. & Gruber-Vodicka, H. R. gbtools: interactive visualization of metagenome bins in R. *Microb. Physiol. Metab.* **6**, 1451 (2015).

80. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 (2013).

81. Brettin, T. *et al.* RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci. Rep.* **5**, 8365 (2015).

82. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinforma. Oxf. Engl.* **30**, 923–930 (2014).

83. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

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

85. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

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

87. Jayasundara, D. *et al.* ViQuaS: an improved reconstruction pipeline for viral quasispecies spectra generated by next-generation sequencing. *Bioinforma. Oxf. Engl.* **31**, 886–896 (2015).

88. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).

Supplementary Information

1. Methods

Table S7 | Sample overview.

Host species	Vent fields	Mussel sizes [mm]	Year (Cruise)	Depth [m]
B. azoricus	Lucky Strike (2 sites)		2013 (BioBaz)	
	Montsegur (MS, 3 mussels)	65, 55, 52		ET 1690
	Eiffeltower (ET, 2 mussels)	70, 92		MS 1700
B. puteoserpentis	Semenov-2	60, 40, 40	2014 (ODEMAR)	2432
	Ash lighthouse			
B. sp.	Clueless	93, 116, 82,101, 107	2009 (M78-2)	2972
B. sp.	Lilliput (2 dives)	60, 56, 51, 61, 64	2009 (M78-2)	1490
B. sp.	Wideawake	-	2008 (ATA57)	2989

1.1 DNA and RNA extraction

Nucleic acids (DNA and RNA) were extracted from the same gill of *B. azoricus* from Lucky Strike with the AllPrep DNA/RNAMini Kit (Qiagen, Germany) according to manufacturer's instructions[1]. Cell debris was removed with QIAshreder Mini Spin Columns (Qiagen, Hilden, Germany). RNA was transcribed into cDNA, using the Ovation RNA-Seq System V2 (NuGEN, USA) and libraries generated with DNA library prep kit for Illumina (Biolabs, Germany). From Semenov, Clueless and Lilliput samples DNA was extracted with the Blood and Tissue Kit (Qiagen, Germany). RNA from Semenov, Clueless and Lilliput samples was extracted from a separate RNAlater-fixed gill tissue sample deriving from the same gill used for DNA extraction with the AllPrep DNA/RNAMini Kit according to manufacturer's instructions (Qiagen, Germany) (Tab. S8, S9).

Table S8 | Metagenomic sequencing details.

DNA	# Individuals	DNA extraction	Seq. Technology	Read length	Sequencing facility
B. azoricus	5	AllPrep	HiSeq2500	2 x 150bp	MPI Cologne
B. puteoserpentis	3	BloodTissue	MiSeq	2 x 250bp	MPI Cologne
B. sp. (Clueless)	5	BloodTissue	HiSeq2500	2 x 150bp	CeBiTec Bielefeld
B. sp. (Lilliput)	5	BloodTissue	HiSeq2500	2 x 150bp	CeBiTec Bielefeld
B. sp. (Wideawake)	1	Zhou et al., (1996)	MiSeq Pacific Biosciences RS II	2 x 250bp 6 SMRT cells	MPI Cologne

Table S9 | Metatranscriptomics sequencing details.

RNA	# Individuals	RNA extraction	Seq. Technology	Read length	Sequencing facility
B. azoricus	5	AllPrep	HiSeq2500	2 x 150bp	MPI Cologne
B. puteoserpentis	3	AllPrep	HiSeq2500	2 x 150bp	MPI Cologne
B. sp. (Clueless)	4	AllPrep	HiSeq2500	2 x 150bp	MPI Cologne
B. sp. (Lilliput)	5	AllPrep	HiSeq2500	2 x 150bp	MPI Cologne

Symbionts from B. sp (Wideawake) gill tissue were enriched by homogenization of the entire gill, centrifugation at 100 xg for 10 min, and serial filtration through 12µm, 5µm, 2µm filters[1]. DNA was extracted according to Zhou et al.[2].

1.2 Sequencing, assembly and annotation

TruSeq library preparation and sequencing was conducted by the Max Planck Genome Centre in Cologne, Germany, and the Centrum für Biotechnologie (CeBiTec) in Bielefeld, Germany, using Illumina HiSeq2500 with a read length of 150 bp, paired-end, for Lucky Strike, Clueless and Lilliput, and Illumina MiSeq sequencing with read length 250 bp for Semenov and Wideawake (Tab. S8, S9). Before assembly, Illumina TruSeq adapters and the internal Illumina standard for sequencing, phiX, were removed from the raw reads using BBMap (v36.x,

Bushnell B. - BBMap – sourceforge.net/projects/bbmap/). Read quality was trimmed to PHRED score 2 and the first 10 bp of all reads were removed.

Metagenomes were assembled according to the following workflow. For each vent field an initial consensus assembly was produced with IDBA-ud (v 1.1.1)[3] with the pooled reads of all mussel individuals of that field. Subsequently, clean reads of each mussel individual were mapped separately to the consensus IDBA assembly to perform differential coverage binning of the SOX symbiont of each mussel individual as implemented in GBtools (v 2.4.5)[4] and the bins further improved with contig connectivity analysis[5]. Clean reads from each mussel individual were then mapped separately to the symbiont bin and reassembled with SPAdes (v 3.1.1)[6] independently to obtain the best bin per host individual. Best individual bins were produced to extract the completest and cleanest read set for the symbiont from each metagenome before a best consensus assembly was produced. When assembly statistics of each individual symbiont bin could not be improved anymore with binning and reassembly, all reads mapping to this optimal bin were pooled and a consensus draft genome per vent field was reassembled using SPAdes[6]. Having a consensus draft genome per field as a reference was essential to be able to compare SNPs across host individuals. The final draft genome statistics, including completeness estimates based on CheckM (v 1.0.7)[7] using gammaproteobacterial 280 marker genes are summarized in Tab. S1. Briefly, symbiont genome sizes were between 2.22 and 2.83 Mb, GC content was around 37.5% and genome completeness was > 94%. Long-read sequencing was done with Pacific Biosciences RS II (PacBio) for the sample from Wideawake (Tab. S7). We sequenced six SMRT cells of extracted DNA from a symbiont enriched fraction. De novo assembly was done using the

SMRT software version 2.3.0 with the RS_HGAP_Assembly 2 workflow. We used RAST[8] to annotate the draft genomes we assembled from Illumina and PacBio sequences.

Table S1 | Details of the SOX symbiont bins.

Host species	# Individuals	Completeness* [%]	# Contigs	# CDS	GC [%]	Genome size [Mb]	Read coverage [x]
B. azoricus	5	94.5	607	3403	37.2	2.80	101-120
B. puteoserpentis	3	94.5	250	2521	37.7	2.22	281-373
B. sp. (Clueless)	5	94.5	452	3103	37.7	2.59	151-215
B. sp. (Lilliput)	5	94.6	578	3430	37.5	2.83	190-218
B. sp (Wideawake)#	1	94.5	379	3178	37.6	2.68	162

*based on gammabroteobacterial marker genes (Parks *et al.*, 2015)
#from MiSeq symbiont bin

1.3 SNP calling

For single nucleotide polymorphisms (SNPs) calling clean reads from each sample were further trimmed to a quality PHRED score of 20 and subsequently mapped with minimum identity of 95% to the consensus draft genome from that respective vent field using BBMap. Subsequently, PCR duplicates were removed using samtools (v 1.3.1)[9], reads realigned around INDELs with the Genome Analysis Toolkit (*GATK* v3.3.0)[10] and down-sampled to an average read coverage of 100x for each sample with samtools (per-sample coverages ranged from 100x to 370x).

SNPs were called with GATK HaplotypeCaller, ploidy 10, to allow for polyploidy that accounts for a mixture of multiple coexisting bacterial strains. At ploidy settings above 10 the SNP numbers did not increase anymore (Tab. S10). Subsequently, unreliable SNPs were filtered with GATK VariantFiltration

(settings: QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -20, ReadPosRankSum < -8).

Comparisons of SNP densities with other ploidy settings for all individuals of vent field Clueless showed that lower ploidy settings (2 and 5) reported fewer SNPs whereas a higher ploidy setting of 20 did not increase the detection of SNPs further (Tab. S10). We calculated the SNPs/kbp from the total number of polymorphic sites divided by the consensus genome size per vent field. For calculation of SNPs/kbp in only the core genes, SNPs were extracted for only the ORFs of that gene subset per vent field and divided by the number of nucleotides in the core genes.

To test whether using a per-site consensus reference can introduce artifacts or influence our estimates of symbiont heterogeneity we included an additional control. Using each individuals own symbiont genome bin as reference, we calculated the SNPs/kbp for all individuals from Clueless. All other parameters in the pipeline were kept the same. This revealed that SNP densities with the consensus draft bin as reference made up 93-98% of the single-bin SNP numbers (Tab. S10). This showed that SNP counts were similar, or slightly underestimated, when using the per-site consensus reference. Bash scripts for mapping and SNP calling are provided in https://github.com/rbcan/MARsym_paper.

Table S10 | SNP densities for the within-host SOX symbiont population of each of 5 host inidivuals from vent field Clueless. Densities were calculated, using a site-consensus genome assembly (GATK ploidy settings 2, 5, 10 and 20) and each individual-specific symbiont genome bin as reference (GATK ploidy setting 10), respectively to evaluate whether using the consensus bin introduces artifacts. In the present study ploidy setting 10 was used with the consensus reference sequence (marked in grey).

Host individuals	SNPs / kbp				
	Consensus reference				Per-individual reference
	Ploidy 2	Ploidy 5	Ploidy 10	Ploidy 20	Ploidy 10
C2	6.1	7.1	7.4	7.4	7.9
C3	5.8	6.8	7.1	7.1	7.6
C4	4.9	5.9	6.1	6.2	6.4
C5	5.4	6.3	6.5	6.6	6.7
C6	5.8	6.9	7.2	7.2	7.6

1.4 Core genome calculation and low coverage gene analysis for the detection of strain-specific genes

The core genome of all SOX symbionts from all four vents, using the consensus draft genome, was estimated with GET_HOMOLOGUES (v1.0)[11], using the OMCL algorithm and a sequence identity setting of 0.5 (Fig. S1). The core genome contained 1283 gene families, representing 40-53% of the predicted ORFs per reference genome.

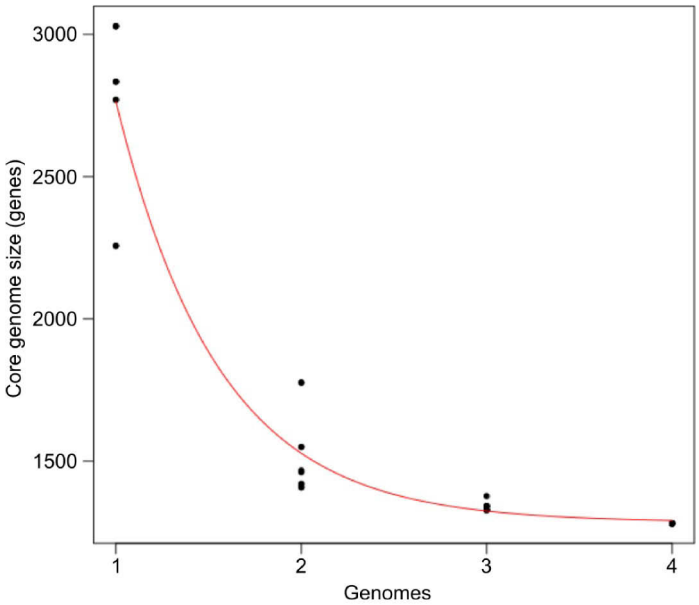


Figure S1 | Core genome plot of the SOX symbiont along MAR vent fields. Core genome contains 1283 gene clusters. Gene duplications can lead to different numbers of genes in these gene clusters and we detected 1344 (Lucky Strike), 1350 (Semenov) and 1395 (Clueless and Lilliput) genes, representing 1283 clusters. The red line represents the fitted function of the model explaining the decrease in genes that belong to the core genome

In metagenomes, strain-specific genes not encoded in the genomes of all co-occurring symbionts have a lower coverage compared to single-copy genes which are found in all strains[12]. To systematically identify strain-specific genes, we calculated the per-gene read coverage from our metagenomes. To calculate the per-gene coverage, we used the read mapping files (minimum identity 95%) that were created for the SNP calling (see above). Per-base coverage was calculated with bedtools (v2.16.2)[13], and per-gene coverage for all ORFs was calculated as the mean of the per-base coverages for each ORF with a custom R-script (R version 3.2.2)[14]. To determine the expected coverage of genes that are encoded once by each cell in the population, we extracted more than 200 gammaproteobacterial marker genes (from PhylaAmphora[15]) from the symbiont bins and calculated the coverage for each gene. We plotted the coverage

distribution of each gene and of the marker gene set in R (Fig. S4). Some genes had coverage values exceeding those of the single-copy markers, likely resembling unresolved paralogs and repeat elements. All genes that were below the lowest coverage of the marker genes were considered ‘low coverage' and potentially strain-specific.

Low-coverage gene sequences that reached into the last 200 bp of contig edges were not considered because low coverage, as well as high coverage in case of repeat elements, is typical for contig ends and may not reflect strain-specific differences. For functional analysis we excluded all genes that were annotated as “hypothetical proteins”.

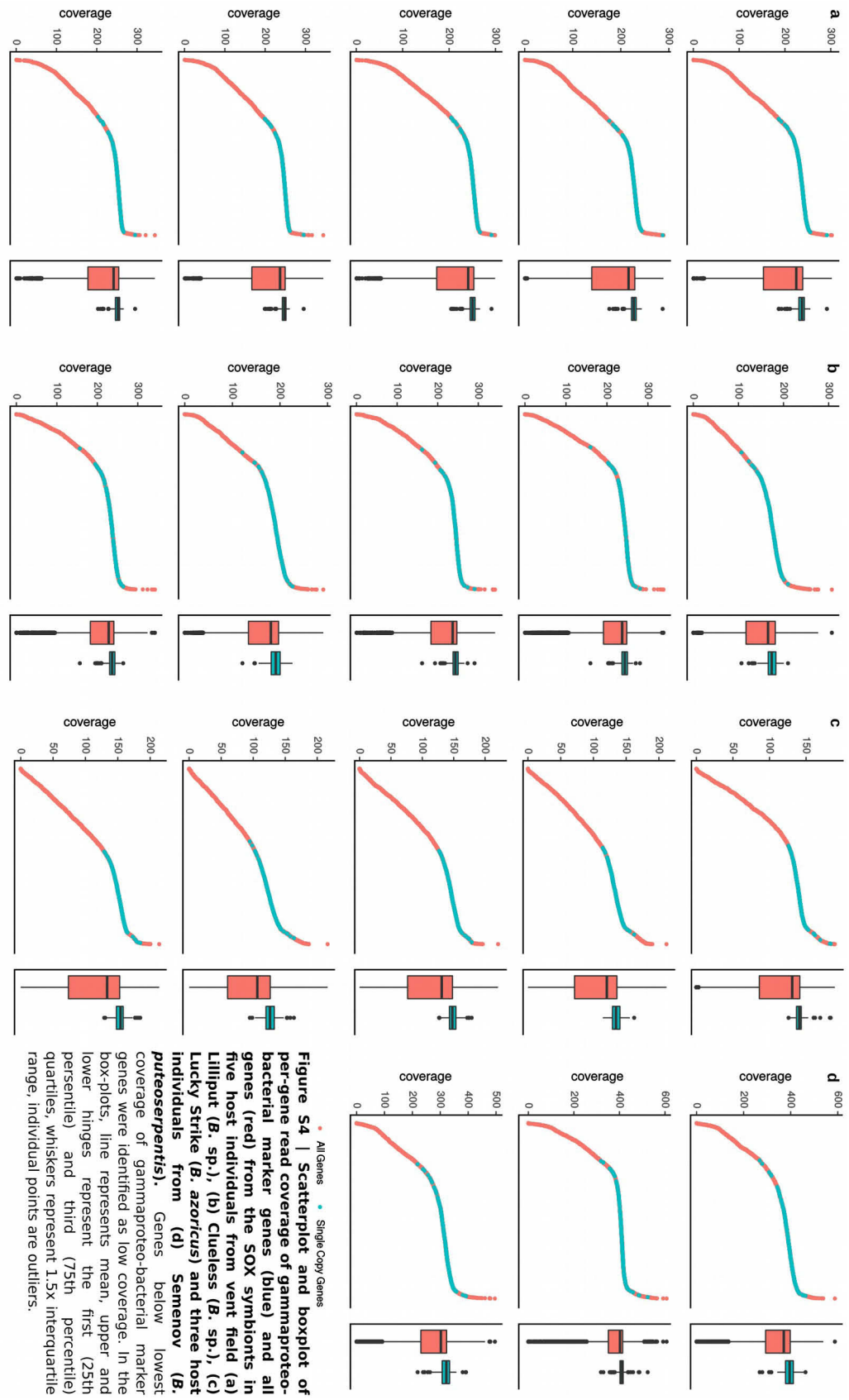


Figure S4 | Scatterplot and boxplot of per-gene read coverage of gammaproteobacterial marker genes (blue) and all genes (red) from the SOX symbionts in five host individuals from vent field (a) Lilliput (*B.* sp.), (b) Clueless (*B.* sp.), (c) Lucky Strike (*B. azoricus*) and three host individuals from (d) Semenov (*B. puteoserpentis*). Genes below lowest coverage of gammaproteo-bacterial marker genes were identified as low coverage. In the box-plots, line represents mean, upper and lower hinges represent the first (25th persentile) and third (75th percentile) quartiles, whiskers represent 1.5x interquartile range, individual points are outliers.

1.5 Strain number estimation and test simulation

Sequencing read coverage of gammaproteobacterial phylogenetic marker genes, as defined by the PhylaAmphora database, was used to detect genes that are present in each strain of the symbiont population as they fell into the same read coverage range (see Method section *Low coverage gene analysis)*. Gene versions were reconstructed for PhylaAmphora markers, as well as all genes encoded by each strain in the population using the tool ViQuaS (v 1.3)[16]. We applied the recommended tool settings for bacterial genomes and discarded all reconstructed sequences below the relative strain frequency above which the haplotype reconstruction is reliable ($f_{\min}$) and which was calculated for each gene (Bash scripts in https://github.com/rbcan/MARsym_paper). The highest number of versions for a single gene served as estimation of strain numbers in that particular sample.

To evaluate this approach, a test dataset was created from 10 *E. coli* genomes which showed a heterogeneity of 1% when pooled. Illumina sequencing reads were simulated with ART (v. 2.5.8)[17] for each of the genomes according to Tab. S5. All parameters were chosen to be similar to the features in the sequencing data used in this study. The reads of all 10 genomes were pooled in two different ways, even abundances and uneven abundances of contributing strains, but we consider uneven abundances to be more similar to natural populations.

Table S5 | Test dataset, created from the listed 10 *E. coli* strain genomes. Reads were simulated with ART[#] with default settings unless specified otherwise. Two conditions of uneven and even strain abundances were created (relative abundances shown in table).

	Strain	BioSample°	Accession°	Settings in ART[#]
Escherichia coli	P12b	SAMN02603904	CP002291.1	Seq. system: HiSeq2500 Error: Illumina / default Read length: 150bp Reads: paired-end Insert size: 215bp Min/max quality: 26/40
	ATCC 8739	SAMN02598405	CP000946.1	
	K-12 substr. DH10B	SAMN02604262	CP000948.	
	HS	SAMN02604037	CP000802.1	
	K-12 substr. MG1655	SAMN02604091	U00096.3	
	DH1	SAMN02598470	CP001637.1	
	REL606	SAMN02603421	CP000819.1	
	K-12 substr. BW2952	SAMN02603900	CP001396.1	
	BL21-Gold(DE3)pLysS AG	SAMN00002656	CP001665.1	
	BL21(DE3)	SAMN02603478	CP001509.3	
Heterogeneity (pooled)	1%[*]			
Uneven strain abundances	1, 0.5, 0.25, 0.125, 0.063, 0.031, 0.016, 0.008, 0.004, 0.002			
Even strain abundances	0.2 all strains			

[*]calculated with Parsnp v. 1.1.2 (Treangen *et al.*, 2014)
°NCBI https://www.ncbi.nlm.nih.gov/
[#]tool ART (Huang *et al.*, 2012)

Both samples were assembled with SPAdes (v 3.1.1)[6], mapped with BBMap at 95% identity and subjected to the strain estimation pipeline as described above. The latter was applied to two read coverage depths, 100x and 300x on average, to reach best comparability to our sample data. The strain number estimation from the PacBio assembly was conducted by counting the number of contigs per gene that had full read coverage (defined as above) and were present in one copy only in the Illumina draft genome annotation from the same sample.

1.6 Population structure

SNP frequencies for each sample were extracted from the vcf-files created by our SNP calling analysis.

The F_{ST} calculation is based on Schloissnig *et al.*[17]. Briefly, nucleotide diversity (π) within single host individuals (referred to as π_{within}) was calculated per gene according to

$$\pi_{s,g} = \frac{1}{|g|} \sum_{i=1}^{|g|} \sum_{n_1 \in \{A,C,T,G\}} \sum_{n_2 \in \{A,C,T,G\} \setminus n_1} \frac{x_{i,n_1,s}\, x_{i,n_2,s}}{c_{i,s}(c_{i,s}-1)}$$

where s is the sample, g is the gene, $|g|$ is the length of the gene, $c_{i,s}$ is the coverage at position i in sample s and $x_{i,nj,s}$ is the number of nucleotides n_j at position i in sample s.

Correspondingly, the pairwise between hosts diversity (referred to as $\pi_{between}$) is calculated as

$$\pi_{s_1,s_2,g} = \frac{1}{|g|} \sum_{i=1}^{|g|} \sum_{n_1 \in \{A,C,T,G\}} \sum_{n_2 \in \{A,C,T,G\} \setminus n_1} \frac{x_{i,n_1,s_1}\, x_{i,n_2,s_2}}{c_{i,s_1} c_{i,s_2}}$$

Then the fixation index F_{ST} for all samples S is defined by the ratio of the average sample diversity and the average between-sample diversity[18]:

$$F_{ST}(S,g) = 1 - \frac{\frac{1}{|S|} \sum_{s \in S} \pi_{s,g}}{\frac{2}{|S|(|S|-1)} \sum_{s_1 \in S} \sum_{s_2 \in S \setminus s_1} \pi_{s_1,s_2,g}}$$

where $|S|$ is the number of samples.

We calculated pairwise per gene F_{ST}-values between symbiont populations of each two host individuals per vent field and mean per-gene F_{ST}-values of all host individuals per field. We further calculated average F_{ST} over all genes for each field.

PCA plots of per-gene π_{within} and $\pi_{between}$, considering all core genes among the vent fields, were created with R (v.3.2.2)[14] using the *prcomp* function in the *stats* (v 3.2.3) package and *autoplot* function in the *ggplot2* (v 3.0.0) package. We considered our dataset as multivariate with per-gene π representing the variables. We tested if significant differences exist between π_{within} and $\pi_{between}$ at a single vent field, as well as whether significant differences in π_{within} exist among vent fields. Finally, we performed a pairwise comparison of π_{within} between fields to examine which of the fields differed significantly. These tests were performed using permutational multivariate analyses of variance (PERMANOVA, a*donis* function of the *vegan*, v 2.5.2, package) on a Bray-Curtis dissimilarity matrix (*vegdist* function of the *vegan*, package) calculated for the π-values. Our null hypothesis was that there is no difference among the compared groups, which we defined as follows: For the comparison between vent fields, the π_{within} of all the mussels at a single field are considered as one group resulting in a total of four groups; for the analysis at each vent field we consider all π_{within} as one group and the $\pi_{between}$, as the second group, with the single mussels or pairs as replicates.

To understand whether there is a correlation between (a) the difference in host shell length of each compared two host individuals and pairwise F_{ST}-values, (b) the sum of host shell lengths of each compared two host individuals and pairwise F_{ST}-values and (c) shell length and within-host SNP density, we used Spearman correlation (*rcorr* package). Spearman‘s correlation coefficient rho, referred to as *r*, and p-values are shown in Extended Data Fig. 4.

Table S2 | Average nucleotide diversity π among individuals of the same vent field. Standard deviation is indicated in the brackets.

Vent field	Mean π	Within-host π	Pairwise between-host π
Lucky Strike	0.0024 (± 0.00022)	0.0022 (± 0.00022)	0.0026 (± 0.00012)
Semenov	0.0024 (± 0.00007)	0.0023 (± 0.00006)	0.0025 (± 0.00005)
Clueless	0.0026 (± 0.00022)	0.0024 (± 0.00022)	0.0027 (± 0.00013)
Lilliput	0.0040 (± 0.00011)	0.0039 (± 0.00004)	0.0041 (± 0.00003)

Tab. S3 | PERMANOVA results comparing within-host π-values of symbiont populations between vent fields. Analysis performed on values of core genes; 3 degrees of freedom (df) and 14 residual (total df = 17) for the comparison of all 4 vent fields; 1 df and 6 residual (total df = 7) for comparisons of Semenov samples with the other fields and 8 residual (total df = 9) for comparisons between Lucky Strike, Clueless and Lilliput; Pseudo-F = F-value by permutation; P = P-value by 5000 permutations.

	Lucky Strike		Semenov		Clueless		All	
	Pseudo-F	P	Pseudo-F	P	Pseudo-F	P	Pseudo-F	P
Semenov	19.18	0.018*						
Clueless	30.538	0.008**	55.542	0.019*				
Lilliput	38.634	0.008**	109.7	0.018**	72.931	0.008**		
							85.822	**< 0.001***

Tab. S4 | PERMANOVA results comparing within-host π-values of symbiont populations against pairwise between-host π- values of the populations. Analysis performed on values of all encoded genes with 3609 for Lucky Strike, 2730 for Semenov, 3304 for Clueless and 3640 for Lilliput as well as for core genes; 1 degree of freedom (df) for all; 4 residual (total df = 5) for Semenov and 13 residual (total df = 14) for Lucky Strike, Clueless and Lilliput; Pseudo-F = F-value by permutation; P = P-value by 5000 permutations for Lucky Strike, Clueless and Lilliput and 719 permutations (maximum) for Semenov.

π_{within} / $\pi_{between}$	All genes		Core genes	
	Pseudo-F	P	Pseudo-F	P
Lucky Strike	1.311	0.253	0.982	0.399
Semenov	0.321	1	0.084	1
Clueless	0.848	0.529	0.721	0.602
Lilliput	0.749	0.628	0.354	0.864

2. Results and Discussion

There is some debate about the definition of a microbial ‘strain’ and even species[19]. However, most agree that genotypic differences are required and sufficient for microorganisms to be classified as different strains, without information about phenotypic differences, which can also be found within genetically identical microbial populations[20]. In this study, we define a strain by a distinct sequence version of any coding gene that is encoded by all cells in the symbiont population.

2.1 Population structure of the SOX symbiont

The population genetic measure F_{ST} can determine how differentiated populations are compared to each other. Our extended sampling effort allowed us to investigate data from a range of mussel sizes, which reflect differences in age as shown for other species of bathymodiolin mussels[23]. As shown previously, the gills in *Bathymodiolus* mussels are continuously growing and newly formed gill tissue needs to be colonized by the symbionts throughout the hosts life[24]. However, it was unclear whether this happens exclusively through repeated self-infection within one host or through the uptake of symbionts from the environment, or both. If there is no exchange of symbionts among host individuals, as during repeated self-infection, the symbiont populations of two co-occurring mussels would evolve isolated from each other after the first colonization of juvenile hosts. This will result in a gradual differentiation of the two bacterial populations over time possibly leading to an increase of F_{ST}.

Accumulation of within-host mutations and differential loss of symbiont strains, decreasing the similarity among populations over time, could result in a positive correlation between the sum of shell lengths of the compared host pair, representing the duration of population isolation, with F_{ST} (Extended Data Fig. 2). We did not detect significant correlations for any vent field, however Lucky Strike samples showed low p-values ($p = 0.06$; Extended Data Fig. 4). Additionally, large and small hosts would have been colonized at very different time points, possibly leading to a correlation between F_{ST} and the difference in host shell lenghts (Extended Data Fig. 1). However, there was no positive correlation of difference in shell length with F_{ST}, which could have resulted from two compositionally different seeding populations that colonized the hosts at different times (Extended Data Fig. 4). Instead mussels may continue to take up symbionts from the environment throughout their life. In the case of such continuous symbiont uptake, we do not know whether the symbionts originate from an active free-living population or the release from the surrounding mussel bed. If the symbionts are acquired exclusively from an active free-living population, older individuals would possibly have had contact to other strains that young individuals never encountered. This possibly results in a correlation between heterogeneity and shell length. In the other case where symbionts are acquired from released cells from co-occuring individuals, we would not expect a positive correlation as the symbiont populations would continuously intermix among hosts. We did not detect any significant correlation of heterogeneity with shell length, except for Clueless. The limited number of samples (five) calls for caution when interpreting these results (Extended Data Fig. 4). The fact that we did not detect correlations of F_{ST} with difference in shell length, sum in shell

length or between heterogeneity and shell length supports our hypothesis of a continuous uptake mode with intermixing SOX symbiont populations. Note that the factors leading to such correlations may be minimal in the lifetime of sampled mussels and therefore below the detection limit (Extended Data Fig. 4, Extended Data Fig. 2).

Interestingly, at Lucky Strike we observed the least similar symbiont populations among hosts from the same vent field and potentially a correlation between the sum of host shell lengths and F_{ST} values (p-value = 0.06, Extended Data Fig. 4). However, these mussels were sampled from two distinct patches 150 m apart, which might be influencing population dynamics. Typical patchiness of biogeochemical conditions even at small distances at hydrothermal vent fields might favor the higher abundance or presence of certain bacterial strains with advantageous genomic potentials. A recent study by Ho et al.[25], showed high divergence between SOX symbionts of different host individuals, possibly caused by patchiness as the authors detected little divergence when mussels originated from the same patch. This divergence even at a small distance between patches is further support for the continuous exchange of symbionts among mussels that are in close proximity to each other. It also indicates that two different patches might select for different strain compositions.

2.2 Number of strains – test data

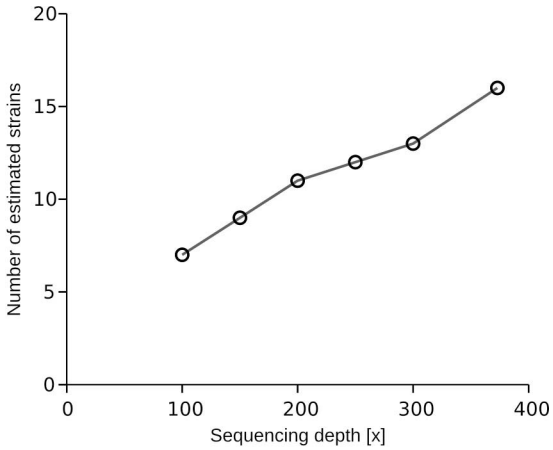


Figure S2 | Strain estimation for different sequencing depths of one *B. puteoserpentis* individual from site Semenov. Detection of strains increases with sequencing depth raising the possibility that more than 16 strains co-occur in *Bathymodiolus* individuals.

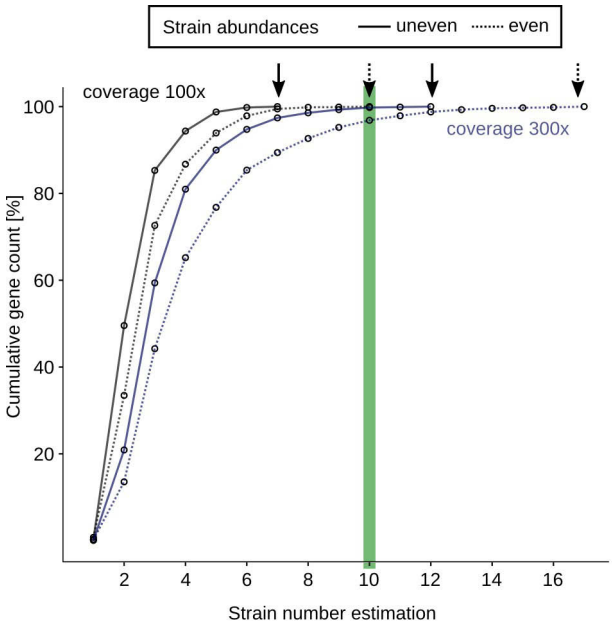


Figure S3 | Cumulative gene count of numbers of reconstructed gene versions for *E. coli* simulation dataset. Even (dashed line) and uneven (continuous line) strain abundances are each displayed for 100x (black) and 300x (blue) read coverage, respectively.

We created a test dataset with simulated reads from 10 *Escherichia coli* strains with 1% nucleotide polymorphisms among them when pooled. We included two scenarios of evenly abundant and differently abundant strains in the population and assembled the consensus genomes for both setups. We assumed that in a natural population, the abundances of different strains are not all even but rather similar to the test with uneven abundances. Following the same gene version construction approach as for the symbiont data we estimated 7 (different abundances) and 10 (even abundances) *E. coli* strains for a coverage of 100x (Fig. S3). Increasing the coverage to 300x, our analysis estimated 12 (different abundances) and 17 (equal abundances) strains.This analysis showed that at a coverage of 300x with even abundances we overestimated the number

of strains, but are closest to accurate numbers with uneven strain abundances. For 100x the estimated strain numbers were accurate with even abundances and underestimated with uneven strain abundances.

2.3 Accessory genome

Table S6 (Excel File) | List of strain-specific genes (all hypothetical proteins were excluded). List contains gene ID, orthologous family ID, predicted functions, indications of strain-specificity, coverage information [% of single-copy gene coverage], transcript counts [% of single-copy gene mean; TMM normalized]. Purple - Lilliput (*B*. sp. Individuals L102, L104, L105, L51, L54); Green – Clueless (*B*. sp. Individuals C2, C3, C4, C5, C6); Blue – Semenov (*B. puteoserpentis* individuals BputSemA, BputSemB, BputSemC); Red – Lucky Strike (*B. azoricus* individuals BazF, BazG, BazH, BazI, BazJ). In Summary Column: () - other lower covered versions, ‘ - not all lower covered, * - other fully covered version, a – all strains encode it, l – lower covered (strain-specific), x – gene not present; darker color when all (except max. one) individual express it, lighter color when some but not all express it, no shade when none express it (for Lucky Strike samples: all below 10 considered not/low expressed). Expression and gene coverage values are only reported for “l” low-coverage genes. Upper part of the list contains all genes with functional annotation that were “truly” strain-specific at at least one hydrothermal vent field, as no further copy of these genes were identified in the genomes. The lower part of the list reports some additional genes that were detected to have low coverage but had mostly additional low-coverage of full-coverage copies. Here, low coverage might be caused by the fact that different gene neighborhoods (due to genome rearrangements) result in multiple contigs with the same gene in the assembly. For these genes we can not be certain whether every strain carries this gene or not, and considered these genes not “truly” strain specific. (**file can be found on CD**)

In the following we report details about the distribution of the most prominent categories of strain-specific genes.

1) Cell-surface

Multiple genes whose functions were assigned to be involved in the synthesis of cell-surface components, such as O-antigens, and cell-cell interaction mechanisms were strain-specific and expressed at all vent fields. Many of these genes encoded proteins involved in the dTDP-L-Rhamnose biosynthesis pathway. Also, genes involved in biosynthesis of the capsular polysaccharide GDP-L-fucose were strain-specific in Clueless, only lower covered in a few individuals from Lilliput or they had a second low-coverage copy of the same gene at Lucky Strike. These genes encode proteins involved in the biosynthesis of the O-antigen sugar residues. Consistently, previous studies showed that the combination and number of sugar residues in O-antigens are highly variable surface structures and have been reported to be strain-specific[26]. The variability is important as these structures can determine whether a phage can attach to the bacterial cell or whether a cell is being recognized by a particular host. In the SOX symbionts, these genes were often found in clusters and in some cases had no occurrences of SNPs which is further support for their presence in single strains. The second step in O-antigen production is the linkage of the diverse sugars, a process commonly performed by highly specific glycosyl transferases[26]. Consistently, we found glycosyl transferases, possibly involved in these functions, to be strain-specific. Due to their high strain-specificity, O-antigens are commonly used for serotyping to distinguish between medically relevant strains (e.g for *Salmonella* and *Klebsiella*[27–29]). A serotyping approach might provide a good starting point for visualization of the symbiont strain distribution in *Bathymodiolus* gill tissue.

2) Hydrogen oxidation

A gene cluster encoding the complete enzymatic machinery for the oxidation of hydrogen, was present at all vent fields. The hydrogenase cluster appeared to be strain-specific at fields Lucky Strike and Clueless with different abundances of the hydrogenase-carrying strain, whereas at the other two fields this cluster was encoded in the entire symbiont population. The membrane-bound group 1 NiFe hydrogenase consists of a large and small subunit of the uptake hydrogenase which are encoded in one gene cluster together with its maturation factors in the SOX symbiont[30,31]. In Clueless mussels, the hydrogenase cluster was encoded by 7-25% of the symbiont population and the small subunit was detected expressed. The other genes in the hydrogenase cluster were barely or not detected in the transcriptome. Conversely, in Lucky Strike symbionts, all genes in the cluster were expressed and based on read coverage we inferred that 40-83% of the symbiont population carried the cluster.

We investigated the link between hydrogen concentrations in the mussel habitats and presence of the hydrogenase gene cluster in symbiont populations. At Semenov, an ultramafic-hosted vent field with hydrogen concentrations in the mM range, most likely all symbiont strains could use hydrogen (unpublished data from Meteor expedition M126 in 2016) and in fact all of them harbored the hydrogenase gene cluster. In contrast, at Clueless, where hydrogen concentrations were below 1 nM[32], only one in ten symbionts had a hydrogenase. Low hydrogen availability, resulting in competition for this limited

resource among symbiont strains, would favor the loss of hydrogenase genes, as the cost of maintaining the gene cluster in the genome may outweigh the benefit of additional energy. The link between hydrogen availability and hydrogenase prevalence was not as clear at the other two sampling sites. Lucky Strike had 122 µM hydrogen, and 60% of symbiont strains encoded hydrogenases. Lilliput had 0.87 µM hydrogen, and hydrogenases were detected in all strains. However, hydrogen concentrations alone do not always indicate which energy source is most favorable for free-living microbes, as this may depend on the availability of other energy sources such as sulfide[33]. Additionally, measuring the concentrations of symbiont energy sources such as hydrogen at hydrothermal vents is not trivial, as they can change over time and space even within one mussel bed down to a scale of milimeters[34]. Our data showed that differences in gene content may be tighlty linked to the environment and resource availability, however our ability to predict the precise conditions experienced by the symbiont populations is still influenced by our limitations to accurately determine all environmental parameters.

3) Denitrification

The reduction of nitrate to nitrogen gas (N_2) is performed in four steps that require the following enzymes: respiratory nitrate reductase (Nar), nitrite reductase (Nir), nitric oxide reductase (Nor) and nitrous oxide reductase (Nos). We detected a striking variability in the distribution of these genes among coexisting strains (Extended Data Fig. 6). At Lucky Strike, the Nar subunits had variable read coverage representing 60% to 100% of the symbiont population.

In contrast, at the same site the genes encoding the enzymes Nir and Nor were strain-specific in all host individuals with a representation of 30 – 75% in the symbiont population. Surprisingly, no strain was capable to perform the last step in nitrate reduction to N_2, as the Nos was missing in all strains at Lucky Strike. The entire Semenov symbiont population encoded only the Nar subunits for nitrate reduction to nitrite, but no strain could reduce nitrite to dinitrogen. In Clueless and Lilliput symbionts the Nar was strain-specific (in two out of five host individuals for Clueless and all individuals for Lilliput), while the genes encoding the enzymes Nir and Nor were present in all strains. Clueless was the only vent where the genes for Nos, which performs the last step in nitrate reduction and its two maturation proteins were present. This gene was also strain-specific and encoded in 15 – 50% of the population. Except for the Nos maturation proteins, all genes encoding enzymes involved in the nitrate reduction were expressed at the different vent fields.

It might seem counterintuitive why some strains lack the Nar as it likely provides an advantage during fluctuating oxygen concentrations. However, all strains additionally encode the assimilatory nitrate reductase (NasA) which performs the same reaction and both enzymes can possibly complement each other. A redundancy between NasA and Nar was hypothesized for vesicomyid symbionts where *Ca.* Ruthia magnifica and *Ca.* Vesicomyosocious okutanii both encoded only one of the two enzymes Nar and Nas[35]. Intriguingly, the recently cultivated, free-living relative of the SOX symbiont, *Ca.* T. autotrophicus, was also shown to encode only the Nar and Nor but lacks the enzymes to perform intermediate reactions of nitrate reduction[36]. Unlike the hydrogenase genes,

where either the entire gene cluster was present or absent, nitrate reduction genes were much more flexible at the level of single genes in *Bathymodiolus* symbionts. Such a high variation in presence and absence of nitrate reduction genes, in both free-living bacteria and symbionts, suggests that these genes seem very prone to be aquired and lost from genomes. Moreover, this suggests that each of the genes can provide a selective advantage or disadvantage on its own, without necessarily being dependent on the presence of all genes encoding the whole denitrification pathway.

4) Methanol oxidation

A gene cluster containing three genes for the oxidation of methanol to formate and four genes for coenzyme pyrroloquinoline (PQQ) synthesis was identified as strain-specific in Lucky Strike symbionts, while present in all strains in Semenov individuals and absent from Clueless and Lilliput symbionts.

5) P_i-dependent gene regulation and P_i uptake

The high-affinity phosphate transport system PstSCAB and the two-component regulatory system PhoR-PhoB was strain-specific at vent field Lucky Strike and completely absent from vent field Lilliput, whereas it was encoded by all strains at vent fields Semenov and Clueless. Phosphate strongly adsorbs to iron oxyhydroxides. At hydrothermal vents, dissolved phosphate is scavenged by iron minerals released in great quantities by the hydrothermal plume[37,38]. Since iron concentrations differ substantially between vent fields, we would expect that

phosphate concentrations in the mussel beds also differ, depending on the iron concentration in the vent plumes. In habitats where phosphate is plentiful, the advantage of a phosphate-dependent regulatory system may not outweigh the cost of maintaining it, thus, it is more likely to be lost. As far as we are aware, phosphate has only been measured at one of our four sampling locations, Lucky Strike. Here, dissolved phosphate concentrations were relatively high when compared to surface seawater (0.6-2.6µM)[39,40]. Hydrothermal fluids at this vent field have relatively low iron concentrations of 60-770 µmol/kg, which is consistent with the inverse correlation existing between iron and phosphate availability[41]. Although no phosphate concentrations have been reported for Lilliput, it is characterized by diffuse venting of low-temperature fluids with low iron concentrations (0.1 – 43 µM)[42]. Considering the interaction between iron and phosphate, this data suggests that less phosphate might be removed by the co-precipitation, resulting in higher phosphate concentrations. The Pho gene cluster provides yet another example of the remarkable flexibility in gene content of the SOX symbiont even among co-existing strains. Moreover, although we cannot draw final conclusions, the examples of hydrogenase and the P_i-related genes strongly suggest that the environmental conditions are key drivers of strain diversity and population dynamics.

6) Phage defense: CRISPR-Cas and RM-systems

CRISPR-Cas are described as the adaptive immune response of bacteria and archaea towards virus infections. The palindromic repeats are separated by spacers that represent fragments of previously invading virus or plasmid DNA

and therefore serve as a memory for a targeted defense during the next infection[43–45]. In Lucky Strike symbionts, two gene clusters contained *cas* genes of the types Cas, Csy and Csn, followed by the CRISPR array of integrated spacer sequences and repeats. All of the *cas* genes were strain-specific with abundances of 30-75%, except for one *cas1* that was present in the entire population (Extended Data Fig. 7). In Semenov symbionts, we identified one CRIPSR-associated gene cluster with Cas1, Cas2 and Csn1 present in 60-91% of the population. No expression of the CRISPR-associated genes could be detected in the transcriptome. At Clueless, two contigs with identical CRISPR-associated genes Cas1, Cas2 and Csn1 were found in 30-90% of the population. These may not be strain-specific but potentially caused by rearrangements around the *cas* genes resulting in different gene neighbourhoods around these genes and thus different contigs during the assembly. The *cas* genes were expressed in some individuals and the Csn1 in all. One additional contig carrying Cas1 was also identified in the entire symbiont population (Extended Data Fig. 7). In Lilliput symbionts no strain specificity could be identified with certainty in the presence of *cas* genes because the genes were detected in contig extremities. Only *cas1* was shown to be present in 41-67% of the population but with a second copy encoded by the entire population (Extended Data Fig. 7).

Further evidence of strain-specific phage pressure is the vast number of restriction-modification (RM) systems we identified in our dataset. RM are phage defense mechanisms, as well, consisting of a restriction enzyme to degrade invading DNA and a methylation unit to protect self-DNA. RM systems

of type I, II and III, as well as other genes involved in DNA metabolism appeared in vast numbers in the variable strain genomes with strong indications of genome rearrangements around them (as indicated by mobile elements, transposases and short contig lengths). Furthermore a large fraction of these was expressed, sometimes with variation according to host individual. Besides the function of RM systems in phage defense they were suggested to be involved in other processes such as increase of genomic variation, stabilization of genomic islands, control of HGT among closely related strains and regulation of transcription via methylation causing e.g. altered surface structures[46]. Therefore, all of these may be important factors in the host-symbiont, symbiont-symbiont and symbiont-phage interactions.

2.4 How does symbiont diversity emerge and how is it maintained?

Phages are key drivers of strain diversity in many bacterial populations[47]. They can also strongly influence pangenome evolution. There are several examples of horizontal acquisition of ecologically important gene clusters through phage infection, such as photosynthesis genes in *Prochlorococcus*[48]. Recently, sulfur-oxidation genes were discovered in phages that infect free-living SUP05 bacteria, close relatives of the SOX symbiont[49,50]. Our study revealed a number of genes that suggest widespread phage infection in the SOX symbionts. These include surface structure genes, restriction-modification (RM) systems and CRISPR-Cas and all of these were regularly found in the strain-specific pangenome (Extended Data Fig. 7). Strain diversity in surface lipopolysaccharides and *cas* genes possibly represent diverging phage defense

mechanisms among symbiont strains in a typical arms race[51]. Most of the strain-specific CRISPR-Cas systems were expressed, which could indicate that despite their intracellular lifestyle, the symbionts still face phage infection. While the mechanisms are yet to be investigated, recent findings showed that phage particles can enter mammalian epithelial cells to infect the intracellular pathogen *Staphylococcus aureus*[52]. Alternatively, CRISPR-Cas can also play a role in host-bacteria interactions through modification of surface structures by targeting self-mRNA for degradation[53,54]. Besides phages, other mechanisms can produce and maintain diversity in the populations, such as genome rearrangements through RM systems and exchange of genetic material via outer membrane vesicles[55,56]. While we have indications for a number of potential processes, the mechanisms of how diversity is maintained in the symbiont populations are yet to be determined and will be important to the understanding of the evolution of these symbioses.

References for supplementary material

1. Sayavedra, L. Host-symbiont interactions and metabolism of chemosynthetic symbiosis in deep-sea *Bathymodiolus* mussels. PhD dissertation, *Univ. Bremen* (2016).

2. Zhou, J., Bruns, M. A. & Tiedje, J. M. DNA recovery from soils of diverse composition. *Appl. Environ. Microbiol.* **62,** 316–322 (1996).

3. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28,** 1420–1428 (2012).

4. Seah, B. K. B. & Gruber-Vodicka, H. R. gbtools: Interactive visualization of metagenome bins in R. *Front. Microbiol.* **6,** (2015).

5. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31,** 533–538 (2013).

6. Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).

7. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25,** 1043–1055 (2015).

8. Brettin, T. *et al.* RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci. Rep.* **5,** 8365 (2015).

9. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

10. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

11. Contreras-Moreira, B. & Vinuesa, P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl. Environ. Microbiol.* **79,** 7696–7701 (2013).

12. Ikuta, T. *et al.* Heterogeneous composition of key metabolic gene clusters in a vent mussel symbiont population. *ISME J.* **10**, 990-1001 (2016).

13. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

14. R Core Team. *A language and environment for statistical computing. R Foundation for Statistical Computing.* (2016).

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

16. Jayasundara, D. *et al.* ViQuaS: an improved reconstruction pipeline for viral quasispecies spectra generated by next-generation sequencing. *Bioinforma. Oxf. Engl.* **31,** 886–896 (2015).

17. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28,** 593–594 (2012).

18. Hudson, R. R., Slatkin, M. & Maddison, W. P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132,** 583–589 (1992).

19. Achtman, M. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu. Rev. Microbiol.* **62,** 53–70 (2008).

20. Ackermann, M. Microbial individuality in the natural environment. *ISME J.* **7,** 465–467 (2013).

21. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493,** 45–50 (2013).

22. Kashtan, N. *et al.* Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* **344,** 416–420 (2014).

23. Smith, E. B., Scott, K. M., Nix, E. R., Korte, C. & Fisher, C. R. Growth and condition of seep mussels (*Bathymodiolus childressi*) at a Gulf of Mexico brine pool. *Ecology* **81,** 2392–2403 (2000).

24. Wentrup, C., Wendeberg, A., Schimak, M., Borowski, C. & Dubilier, N. Forever competent: deep-sea bivalves are colonized by their chemosynthetic symbionts throughout their lifetime. *Environ. Microbiol.* **16,** 3699–3713 (2014).

25. Ho, P.-T. *et al.* Geographical structure of endosymbiotic bacteria hosted by *Bathymodiolus* mussels at eastern Pacific hydrothermal vents. *BMC Evol. Biol.* **17,** 121 (2017).

26. Samuel, G. & Reeves, P. Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. *Carbohydr. Res.* **338,** 2503–2519 (2003).

27. Blixt, O., Hoffmann, J., Svenson, S. & Norberg, T. Pathogen specific carbohydrate antigen microarrays: a chip for detection of *Salmonella* O-antigen specific antibodies. *Glycoconj. J.* **25,** 27–36 (2008).

28. Erridge, C., Bennett-Guerrero, E. & Poxton, I. R. Structure and function of lipopolysaccharides. *Microbes Infect.* **4,** 837–851 (2002).

29. Hansen, D. S. *et al. Klebsiella pneumoniae* lipopolysaccharide O typing: revision of prototype strains and O-group distribution among clinical isolates from different sources and countries. *J. Clin. Microbiol.* **37,** 56–62 (1999).

30. Petersen, J. M. *et al.* Hydrogen is an energy source for hydrothermal vent symbioses. *Nature* **476,** 176–180 (2011).

31. Vignais, P. M. & Billoud, B. Occurrence, classification, and biological function of hydrogenases: an overview. *Chem. Rev.* **107,** 4206–4272 (2007).

32. Perner, M. *et al.* Short-term microbial and physico-chemical variability in low-temperature hydrothermal fluids near 5°S on the Mid-Atlantic Ridge. *Environ. Microbiol.* **11,** 2526–2541 (2009).

33. Perner, M. *et al.* Linking geology, fluid chemistry, and microbial activity of basalt- and ultramafic-hosted deep-sea hydrothermal vent environments. *Geobiology* **11,** 340–355 (2013).

34. Zielinski, F. U., Gennerich, H.-H., Borowski, C., Wenzhöfer, F. & Dubilier, N. 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,** Q0AE04 (2011).

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

36. Shah, V., Chang, B. X. & Morris, R. M. Cultivation of a chemoautotroph from the SUP05 clade of marine bacteria that produces nitrite and consumes ammonium. *ISME J.* **11,** 263–271 (2017).

37. Feely, R. A., Trefry, J. H., Lebon, G. T. & German, C. R. The relationship between P/Fe and V/Fe ratios in hydrothermal precipitates and dissolved phosphate in seawater. *Geophys. Res. Lett.* **25,** 2253–2256 (1998).

38. Feely, R. A., Trefry, J. H., Massoth, G. J. & Metz, S. A comparison of the scavenging of phosphorus and arsenic from seawater by hydrothermal iron oxyhydroxides in the Atlantic and Pacific Oceans. *Deep Sea Res. Part Oceanogr. Res. Pap.* **38,** 617–623 (1991).

39. Martiny, A. C., Coleman, M. L. & Chisholm, S. W. Phosphate acquisition genes in *Prochlorococcus* ecotypes: evidence for genome-wide adaptation. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 12552–12557 (2006).

40. Sarradin, P.-M., Caprais, J.-C., Riso, R., Kerouel, R. & Aminot, A. Chemical environment of the hydrothermal mussel communities in the Lucky Strike and Menez Gwen vent fields, Mid Atlantic ridge. *Cah. Biol. Mar.* **40,** 93–104 (1999).

41. Von Damm, K. L., Bray, A. M., Buttermore, L. G. & Oosting, S. E. The geochemical controls on vent fluids from the Lucky Strike vent field, Mid-Atlantic Ridge. *Earth Planet. Sci. Lett.* **160,** 521–536 (1998).

42. Haase, K. M. *et al.* Diking, young volcanism and diffuse hydrothermal activity on the southern Mid-Atlantic Ridge: the Lilliput field at 9°33′S. *Mar. Geol.* **266,** 52–64 (2009).

43. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315,** 1709–1712 (2007).

44. Makarova, K. S. *et al.* Evolution and classification of the CRISPR–Cas systems. *Nat. Rev. Microbiol.* **9,** 467–477 (2011).

45. van der Oost, J., Westra, E. R., Jackson, R. N. & Wiedenheft, B. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **12,** 479–492 (2014).

46. Vasu, K. & Nagaraja, V. Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol. Mol. Biol. Rev. MMBR* **77,** 53–72 (2013).

47. Abeles, S. R. *et al.* Microbial diversity in individuals and their household contacts following typical antibiotic courses. *Microbiome* **4,** 39 (2016).

48. Lindell, D. *et al.* Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 11013–11018 (2004).

49. Anantharaman, K. *et al.* Sulfur oxidation genes in diverse deep-sea viruses. *Science* **344,** 757–760 (2014).

50. Roux, S. *et al.* Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *eLife* **3,** e03125 (2014).

51. Bertozzi Silva, J., Storms, Z. & Sauvageau, D. Host receptors for bacteriophage adsorption. *FEMS Microbiol. Lett.* **363,** (2016).

52. Zhang, L. *et al.* Intracellular *Staphylococcus aureus* control by virulent bacteriophages within MAC-T bovine mammary epithelial cells. *Antimicrob. Agents Chemother.* **61,** e01990-16 (2017).

53. García-Gutiérrez, E., Almendros, C., Mojica, F. J. M., Guzmán, N. M. & García-Martínez, J. CRISPR content correlates with the pathogenic potential of *Escherichia coli*. *PLOS ONE* **10,** e0131935 (2015).

54. Sampson, T. R. & Weiss, D. S. Alternative roles for CRISPR/Cas systems in bacterial pathogenesis. *PLoS Pathog.* **9,** (2013).

55. Biller, S. J. *et al.* Bacterial vesicles in marine ecosystems. *Science* **343,** 183–186 (2014).

56. Batstone, R. T., Carscadden, K. A., Afkhami, M. E. & Frederickson, M. E. Using niche breadth theory to explain generalization in mutualisms. *Ecology* **99**, 1039–1050 (2018).

Chapter III | Genome structure in the SUP05 clade

Genome structure reflects phylogeny rather than lifestyle in a widespread group of free-living and symbiotic marine bacteria from the SUP05 clade

Rebecca Ansorge[1], Stefano Romano[2], Lizbeth Sayavedra[3], Maxim Rubin-Blum[4], Nicole Dubilier[1], Jillian Petersen[2]*

*Author order is not fixed

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

[2]Division for Microbiology and Ecosystem Science, University of Vienna, Austria

[3]Quadram Institute Bioscience, Norwich Research Park, Norwich, United Kingdom

[4]Israel Limnology and Oceanography Research, Tel Shikmona, Haifa, Israel

The manuscript is in preparation and has not been revised by all authors.

Author contributions

RA conceived the study, analyzed the data, prepared the figures/tables and wrote the manuscript. SR performed the COG/KEGG enrichment analysis and contributed to the conceptual design and data interpretation of the study. LS and MR obtained samples, extracted DNA and performed genome binning of some of the included draft genomes, both authors revised the manuscript. JP and ND helped to conceive the study, JP revised the manuscript.

Abstract

Sulfur-oxidizing bacteria of the SUP05 clade are widespread in diverse marine habitats from hydrothermal vents, to oxygen minimum zones, where they impact global biogeochemical cycles and have adopted a diverse repertoire of host-associated and free-living lifestyles. The evolutionary driving forces that led to the versatility and global success of this clade, including the adoption of multiple lifestyles, have been unknown so far. To understand the underlying evolutionary signatures among SUP05 lineages of different lifestyles, we investigated the genomic diversity within the SUP05 clade on a range of phylogenetic levels from genus to strain. Our study revealed that genes of the core genome were partitioned into distinct strains. This suggested that the encoded proteins are essential to a species, however they may be dispensable in single symbiont cells of the community. Thus, by forming a multi strain, or multi-cellular consortium, functions that appear essential to the populations or a host are divided among the strains. We further discovered that differences in gene content among SUP05 lineages are mostly explained by phylogenetic relationships and not by lifestyle. This was unexpected, because specific genomic traits are likely to be selected by specific lifestyles, such as the intracellular colonization of the particular animal host. We concluded that convergent evolution led to similar mechanisms that favor particular lifestyles but with distinct genomic solutions. Our results suggest that the free-living and host-associated SUP05 bacteria own their global success to evolvability, which is the cryptic potential of a population to evolve adaptive solutions to unforeseeable conditions or environments.

Introduction

Until recently, our understanding of microbial diversity and population dynamics, such as the co-existence and competition of bacterial strains, has been shaped by studies of cultivated organisms in the laboratory. However, improvements in next-generation sequencing technologies have resulted in the large-scale generation of vast amounts of metagenomic data, which allows us to resolve natural microbial diversity at unprecedented resolution[1]. Alongside laboratory experiments, metagenomic data is emerging as an essential source of information to understand processes such as microbial community responses to environmental change, antibiotic resistance in pathogenic bacteria, microbial community dynamics in the human microbiome and evolutionary history in species differentiation. Microbial diversity permeates all phylogenetic levels and it fundamentally affects microbial interactions, community dynamics and evolution[2–5].

Comparative genomic studies of bacterial lineages have vastly increased over the past years, revealing a complex landscape of extensive genomic variation even among members of single bacterial species[6]. Genomes of bacterial populations are often partitioned into a ‘core genome’, defined as the set of genes encoded by all members of a group (e.g. species), and a ‘pangenome’ which is the entire set of genes known to be encoded by members of the same group. The term ‘accessory genome’ refers to the variable part that is not encoded by all members of the group. The mechanisms that lead to the emergence and evolution of accessory genomes and their role in the biology of species is subject to ongoing debates[7]. Such debate is mainly based on cultivated, mainly pathogenic bacteria[8,6], whereas studies on natural free-living and beneficial host-associated bacteria are scarce. In fact, it is rarely known what

proportion of accessory genomes is present within natural populations of bacteria, information that would contribute to a more representative and realistic view of natural microbes and help to develop more generalizable theories[9]. The accessory genome plays a key role in the ecology and evolution of bacterial populations[10], underlining the need for more systematic analyses of genomic diversity in natural populations. Accumulating evidence suggests that strains of the same bacterial species can occupy different ecological niches[11,12]. Recent studies using metagenomic analyses suggest that coexistence of multiple highly related strains that differ in nucleotide diversity and gene content may be the norm rather than the exception in natural bacterial populations[10,13] (Chapter II).

In this study we investigated the genomic diversity in the SUP05 clade, a highly successful group of gammaproteobacterial sulfur oxidizers (GSOs) widespread in the world's oceans[14]. Based on the Genome Taxonomy Database[15], two well-known GSO sister clades SUP05 and Arctic96BD, have recently been classified as belonging to the *Thioglobaceae* family within the Thiomicrospirales order. Particularly in oxygen minimum zones (OMZs), anoxic marine zones (AMZs) and hydrothermal vent habitats, *Thioglobaceae* have been found to be abundant, active and are recognized to have an important impact on biogeochemical cycles of sulfur and nitrogen[16–20]. To date, only two strains of free-living bacteria that belong to the SUP05 and its sister clade Arctic96BD, *Ca*. Thioglobus autotrophicus EF1 and *Ca*. Thioglobus singularis PS1, have successfully been cultivated and sequenced[21–24]. However, additional metagenomic and single-cell sequencing efforts of natural microbial communities have shed light on the extensive metabolic versatility among free-living *Thioglobaceae* lineages, affecting sulfur, nitrogen, oxygen and carbon metabolism[17,18,25,26]. Alongside free-living lineages, this clade has evolved intimate

chemosynthetic symbioses with deep-sea invertebrates such as vesicomyid clams, sponges and bathymodiolin mussels[27,28]. The intracellular association with bathymodiolin mussel has been suggested to have evolved more than once within the SUP05 clade[27]. Whether the last common ancestor of the SUP05 clade was free-living or in fact host-associated remains to be shown[29,30]. Underlining their versatility, four different types of lifestyles have been observed in SUP05-related lineages: vertically transmitted endosymbionts (i.e. in vesicomyid clams), horizontally transmitted endosymbionts (i.e. in *Bathymodiolus* mussels), symbionts of unknown transmission mode (i.e. in poecilosclerid sponges) and free-living bacteria (i.e. *Ca*. Thioglobus spp.). Specific genes or other genomic signatures such as broad metabolic potential or capability of frequent horizontal acquisition of novel genes could allow bacteria to adopt a particular lifestyle. Yet, it is unclear which genetic features determine the lifestyle or ecology of these distinct, but closely related lineages within the SUP05 clade. Hunt *et al*.[4] revealed that in marine *Vibrio* environmental specialization occurs across a range of phylogenetic levels, showing that ecological selection can lead to genetic differentiation. Indeed, some *Bathymodiolus* SOX symbionts show genomic differences, such as uneven distributions of genes encoding proteins involved in the energy metabolism, between ecologically distinct sites that have different concentrations of available substrates[10], but also among closely related strains in populations within single mussels[10,31] (Chapter II). These differences appear to reflect the environmental conditions and suggest resource-driven partitioning among co-existing SUP05 strains within animal hosts.

We hypothesized that the different lifestyles adopted by SUP05 bacteria could be characterized by specific genomic signatures, which would allow these microorganisms to, for example, successfully colonize and thrive within a specific

host animal (e.g. mussels of the *Bathymodiolus* genus[32,33]). To investigate our hypothesis, we conducted a pangenome analysis on natural communities belonging to the *Thioglobaceae* family. If a lifestyle or particular environment required specific genes and pathways, we would expect these genes to be shared among all members with the same lifestyle, whereas these genes would not be required and hence absent in the members with a different lifestyle. For example, if genetic repertoires reflect the lifestyle of association with *Bathymodiolus* hosts, we hypothesize that all SUP05 lineages symbiotic to *Bathymodiolus* mussels share specific genes that allow the colonization of this host type. Alternatively, the genetic repertoires might reflect the phylogenetic relationship between bacterial lineages which would imply that a similar lifestyle developed through convergent evolution with different genetic solutions. We determined the extent of genetic variation within and among *Thioglobaceae* sub-clades of different relatedness in order to tease apart whether we can detect genetic repertoires that reflect ecological factors, such as the association with a particular host or geographic location.

Methods

Sample acquisition and DNA extraction

Samples and collection sites are listed in supplementary **Tab. S2**. An overview of the sampling locations is plotted in a geographic map using GeoMapApp (v 3.6.4, http://www.geomapapp.org, **Fig. 1**). The mussels sequenced for this study, *Bathymodiolus brooksi*, were collected in the southern Gulf of Mexico, dissected on board and gill tissue was frozen and stored at -80°C (**Tab. S2**). Sample acquisition and DNA extraction from mussel and sponge samples deriving from other studies are

described in the respective publications[10,34–36,30] (Chapter II) and summarized in **Tab. S2**. Briefly, the tissue samples were either frozen at -80°C, preserved in RNA later (Sigma, Germany) and frozen at -80°C or first fixed in 96% ethanol and subsequently frozen at -80°C. For *Bathymodiolus brooksi* individuals from site MC853, whole gills were homogenized after collection. The homogenate was preserved in RNA later and frozen at -80°C[34]. For *Bathymodiolus* sp. from site Wideawake, symbionts were enriched from a mussel homogenate on board after collection[29,30]. Water samples of free-living *Ca*. 'Thioglobus' perditus were filtered and stored on filters between -20°C and -80°C[16]. DNA was extracted either with the AllPrep DNA/RNA MiniKit (Qiagen, Germany), DNeasy blood and tissue Kit (Qiagen, Germany) or according to Zhou *et al.*[37]. Accession numbers of published genome sequences included in this study (*Bathymodiolus septemdierum* symbionts[31], vesicomyid symbionts[38–40], *Ca*. 'Thioglobus' autotrophicus EF1[22] and *Ca*. 'Thioglobus' singularis[23] strains) are listed in **Tab. S1**.

Metagenomics and genome binning

Libraries of genomic DNA were generated for each sample with the Illumina TruSeq DNA Sample Prep Kit (BioLABS, Germany). Details of the sequencing are shown in **Tab. S2**. Raw sequences were processed, assembled and binned either as described by the respective publication or as described in the following. For the metagenomes in this study, read adapters were removed and read quality was filtered to a minimum of two (Q2), using BBDuk (v 38.34, Bushnell B. - sourceforge.net/projects/bbmap/). Metagenomes were assembled for each sample individually with metaSPAdes using the default parameters and kmer sizes of 33, 55, 77, 99 and 127[41,42]. Binning of the sulfur-oxidizing and methane-oxidizing (MOX) symbiont, that often co-occurs with the

SOX symbiont in *Bathymodiolus* hosts, was performed using Bandage[43], differential coverage analysis combined with taxonomy and GC content[44] using gbtools[45], or Metabat2[46]. This produced high-quality genome bins with the completeness of > 90%, according to gammaproteobacterial marker genes in CheckM[47]. Genome statistics of all genome bins used in this study are listed in **Tab. S3**. Contigs smaller than 500 bp were excluded from all final bins.

Pangenome analysis

Each draft genome created in this or previous studies was annotated with Rast-tk[48] at the patricbrc.org online resource[49]. Average nucleotide identity (ANI) and average amino acid identity (AAI) were calculated with the enveomics collection[50], and clusters of *Thioglobaceae* species were defined according to ANI/AAI cutoff > 95%. Amino acid sequences of all coding sequences for each genome bin and for the published genomes were used to calculate core, accessory and unique genes among all bins within a single species cluster with the tool BPGA (v 1.3.0)[51] using default settings of 0.5 sequence identity for clustering with USEARCH[52].

To obtain a representative gene set encompassing all co-occurring SUP05 strains, we produced a single representative MAG per site and SUP05 species, which included one sequence for each encoded protein. For the calculation of representative SUP05 MAGs, we determined the pangenomes for all genome bins from the same site. The entire set of genes that were encoded by SUP05 strains of a single site were combined and one representative sequence for each gene, was subsequently used as input to compare representative MAGs and genomes to each other in a second pangenome analysis with BPGA (v 1.3.0). The resulting presence-absence matrix of

gene clusters was formatted and shared genes among *Thioglobaceae* groups of interest were visualized (**Fig. 3**), using the package UpsetR (v 1.3.3) in R (v.3.2.2)[53].

COG clustering

We assigned the genes of each genome bin to clusters of orthologous groups (COG)[54] and the kyoto encyclopedia of genes and genomes (KEGG)[55–57] using BPGA (v 1.3.0) in order to assess and compare their functional genetic potentials. For the COG categories we performed a principle component analysis on the presence and absence of broad (COG letters) and fine (COG numbers) categories using the prcomp function in the stats (v 3.2.3) package and autoplot function in the ggplot2 (v 3.0.0) package in R (v.3.2.2). The loadings for principle component (PC) 1 and PC2 were retrieved and the categories and annotation for the most extreme values of < -0.1 or > 0.1 are listed in **Tab. S6**. In addition, we used the enricher function in the clusterProfiler (v 3.10.1) package in R to identify COG and KEGG categories that were enriched in any of the genomes (**Tab. S4, 5**).

Phylogenetic tree calculation

Two phylogenetic trees were calculated for this study. First, we calculated the core genome for all genome bins in this study and two additional genomes, *Thiomicrospira arctica* (GCA_000381085.1) and *Thiomicrospira crunogena* (GCA_000012605.1), which were used as an outgroup. This resulted in a set of 171 of core genes. The protein sequences were aligned with muscle (v3.8.31)[58], concatenated and used to calculate a maximum likelihood tree using IQ-TREE (v1.6.9)[59] (**Fig. 2**). Secondly, we calculated a tree using just one representative genome bin per site (**Fig. 3**). For this,

we used the same set of core genes, alignment and tree building approach as for the first tree which resulted in the same phylogenetic groups and clusters.

Protein domain prediction

We used Interproscan 5[60] to identify protein domains from the Pfam database[61,62] in all amino acid sequences from the representative MAGs (see Pangenome analysis section). All sequences that had one or more Pfam domains identified were compared between datasets and two proteins were considered 'identical' in their domain composition when the exact same Pfam but no other domains were detected. The order of domains in the protein was not considered. Proteins with identical domain composition were clustered and visualized using the package UpsetR (v 1.3.3) in R (v.3.2.2)[53].

Strain-specific gene analysis

Sequencing reads were trimmed to a quality of 20 using BBDuk (v 38.34) and mapped to each genome bin with a minimum identity of 0.95 using BBMap (v 38.34). For the *Bathymodiolus* SUP05 datasets the average read coverage of all samples was reduced to 100x using samtools (v 1.3.1)[63], and the bins with lower coverage were excluded from the analysis. The maximum read coverage was kept for the sponge genome bins (library 1600L: 56x, library 1600N: 55x, library 1600M: 191x) and free-living genome bins (library C2484: 84x; library C2488: 73x). For the MOX the read coverage was reduced to 70x for each genome bin and all datasets with lower coverage were excluded.

Identification of strain-specific genes was performed as described in Ansorge *et al*.[10] (Chapter II), for each genome bin. Briefly, all genes that had lower coverage than gammaproteobacterial marker genes were classified as strain-specific. Subsequently, all genes that had an overlap with the contig proximities (100 bp on both edges of a contig were regarded as contig proximities) were filtered out.

We plotted the percentage of strain-specific genes of all coding sequences and the percentages of strain-specific genes that fall into the core, accessory or unique genome of a *Thioglobaceae* species using the ggplot2 (v 3.0.0) package in R (**Fig. 6**). Additionally, the fraction of strain-specific genes in each COG category was calculated and plotted with the ggplot2 (v 3.0.0) package (**Fig. 7**).

Results

Diversity and phylogenetic relationships within the Thioglobaceae family

In order to determine the genomic diversity among members of the *Thioglobaceae* family we generated 14 new environmental sequencing datasets and included 80 previously generated datasets from *Bathymodiolus* mussels from hydrothermal vents and cold seeps, poecilosclerid sponges from a cold seep, and free-living *Thioglobaceae* (**Fig. 1, Tab. S2**). We only used datasets that resulted in draft genomes of > 90% completeness. We also included genome sequences from three vertically transmitted vesicomyid symbionts, which had lower completeness due to their known genome reduction[64,65], two free-living SUP05 bacteria (*Ca.* Thioglobus autotrophicus and *Ca.* Thioglobus perditus), and three free-living bacteria of the sister clade Arctic96BD (*Ca.* Thioglobus singularis) (**Tab. S1**). Single-cell and

metagenomic sequences of other studies have not been included in our comparative analyses, because they did not fulfill the requirement of 90% completeness[17,18,25,26].

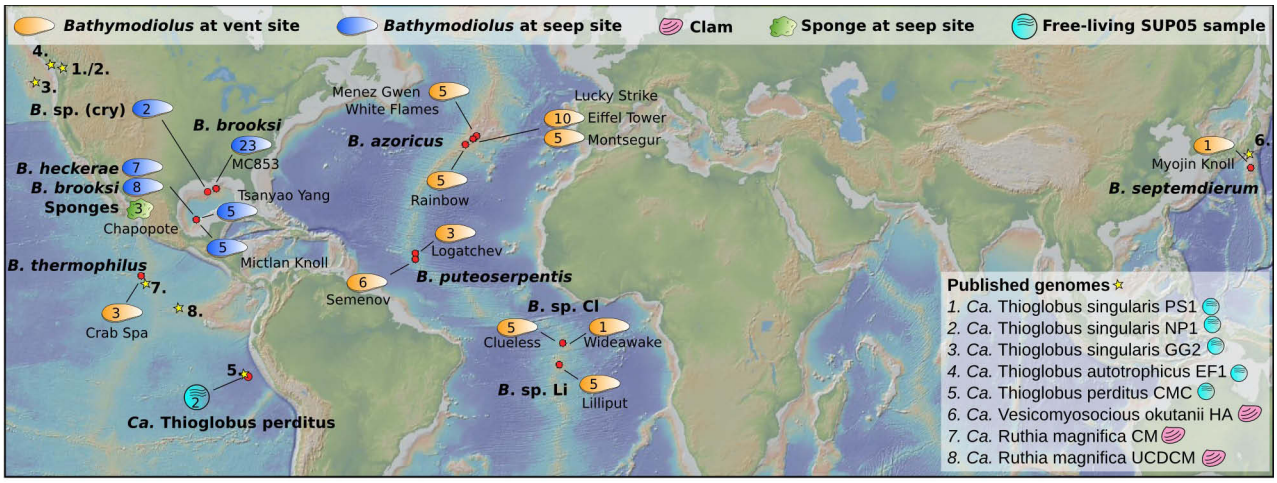


Fig. 1 | Environmental samples used in this study. Numbers indicate how many samples were available.Bold names refer to host species or type, except for free-living *Ca.* Thioglobus perditus. Sampling sites are written in regular font. Published genomes included in this studies are listed in the grey inset and their orginating location is shown as star symbols on map (accession numbers in Tab. S1). The map was created with GeoMapApp http://www.geomapapp.org.

In studies that assess microbial diversity, it is helpful to determine the number of species and the degree of phylogenetic relatedness to ensure comparability between studies and to define the level of genomic resolution that is covered in a study. The bacterial species concept is an ongoing debate since the discovery of bacteria and, despite the significance of strain-level diversity, is recognized to be essential to ensure comparability among different studies of biological communities including those limited to 16S rRNA sequencing[66]. We classified the phylogenetic relationships of all (draft) genomes in our dataset, using average amino acid identity (AAI), average nucleotide identity (ANI), gene content dissimilarity and phylogenetic analysis based on the core protein sequences. We grouped all genomes that showed ANI of >95% as belonging to the same species[66,67] (**Fig. S1, S2**). According to these widely accepted cut-offs, we estimated that in our dataset, we covered 14 species from the SUP05 clade and two species in the Arctic96BD clade (**Fig. 2**). Half of the SUP05 species

were symbionts of *Bathymodiolus* mussels, sampled from 15 different sites and were termed ‘NMAR’ (from host species *B. azoricus* and *B. puteoserpentis*), ‘SMAR’ (from host species *B.* sp. from the southern Mid-Atlantic-Ridge), ‘Bbro’ (from host species *B. brooksi*), ‘Bh_sup05’ (from host species *B. heckerae*), Bcr_sup05 (from host species B. sp., site DC673), ‘Bthe’ (from host species *B. thermophilus*), ‘Bsep’ (from host species *B. septemdierum*), ‘sox2’ (from host species *B. heckerae* and *B.* sp. cryptic species, site DC673).

A clear genomic similarity cut-off at the genus level has been not been identified yet[68]. AAI values around 70% have been considered as a possible genus boundary cutoff[67], but other studies based on the gene content within a group found that, the genus level appears to cover a broad range of gene content dissimilarity mostly below 0.4[68]. All members of the SUP05 clade had gene content dissimilarities below 0.3 when orthologous groups of proteins were compared and AAI values > 68% (Fig. S2). Although we cannot clearly determine whether these lineages all belong to the same genus, a well-supported clade, AAI and gene content dissimilarity values indicate that these lineages belong to a coherent phylogenetic group[68,67]. The published *Ca.* Thioglobus genomes included three members from the sister clade Arctic96BD and two species from the SUP05 clade. Despite sharing the genus name *Ca*. ‘Thioglobus’, the AAI < 65% and gene content dissimilarities between 0.3 to 0.45 between genomes that have been assigned to this genus indicate that these lineages belong to separate genera (**Fig. 2, S1, S2**). We decided to include all *Ca. ‘*Thioglobus’ strains and refer to the combined clades SUP05 and Arctic96BD as the *Thioglobaceae* family as it was has recently been classified [15] (see **Fig. 2**).

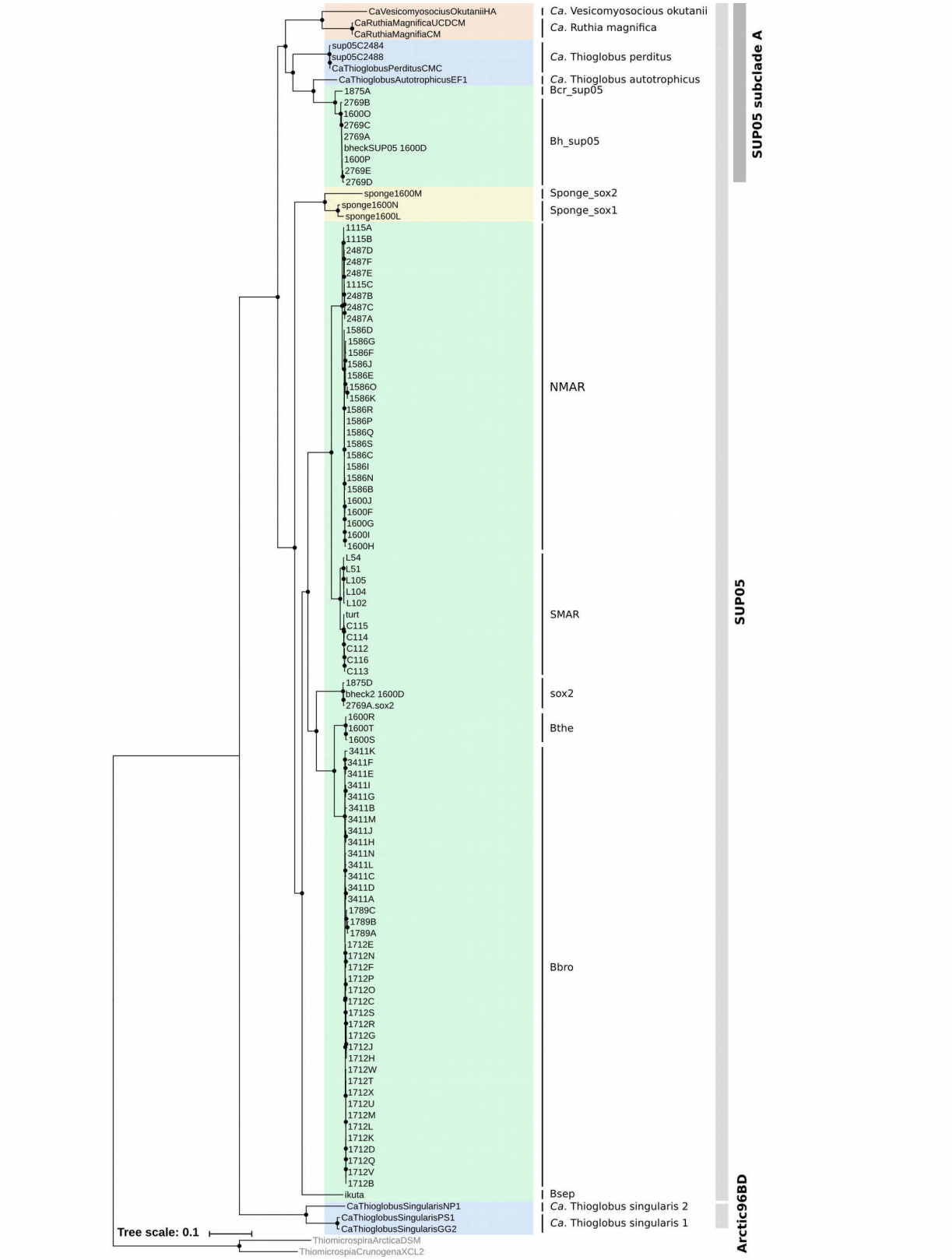


Figure 2 | Phylogenetic maximum likelihood tree based on a protein alignment of the core genes of all included lineages from the *Thioglobaceae* family (171). Colors in the tree correspond to lifestyle: blue: free-living, green: horizontally transmitted intracellular symbiont (*Bathymodiolus* associated), yellow: sponge symbiont of unknown transmission mode, red: vertically transmitted intracellular symbiont (clam-associated), grey: outgroup. Next to the tree, species clusers based on ANI/AAI > 95% are indicated. On the right side, clade names used in the main text are shown. Nodes with bootstrap > 0.9 are shown as black circles.

Species pangenomes and genetic signatures in the Thioglobaceae

We calculated the accessory and core genomes for each of the identified species groups containing three or more draft genomes (i.e. ‘NMAR’, ‘SMAR’, ‘Bbro’, ‘Bh_sup05’ and ‘Bthe’ - all associated with *Bathymodiolus*). Per species, the core genome was formed by 33 to 77% of the genes encoded in each genome (**Fig. 3**). These numbers fall among those of well-known medically important human-associated bacteria (**Fig. 4**). For SUP05 species ‘NMAR’ and ‘Bbro’ we had a large number of draft genomes available (29 and 36, respectively), therefore the fraction of accessory genes can be considered to be within known ranges. The SUP05 species ‘SMAR’, ‘Bh_sup05’ and ‘Bthe’ had only a few samples and their accessory genomes may still increase with additional genomes.

Our phylogenetic analysis revealed that a polyphyletic group of lineages within the *Thioglobaceae* exhibits a free-living lifestyle when the three genomes of *Ca.* ‘Thioglobus’ singularis from the Arctic96BD clade are included (**Fig. 2**). However, even considering only the SUP05 clade, the free-living lifestyle was paraphyletic, as the subclade with *Ca.* ‘Thioglobus’ autotrophicus included other SUP05 lineages that are symbiotic in *Bathymodiolus* mussels. The seven horizontally transmitted intracellular symbiont species that colonize *Bathymodiolus* mussels form a polyphyletic group within the SUP05 clade. The fact that lineages of these lifestyles do not form monophyletic clades allowed us to distinguish between differentiation that reflects phylogeny or ecology (e.g. lifestyle), respectively.

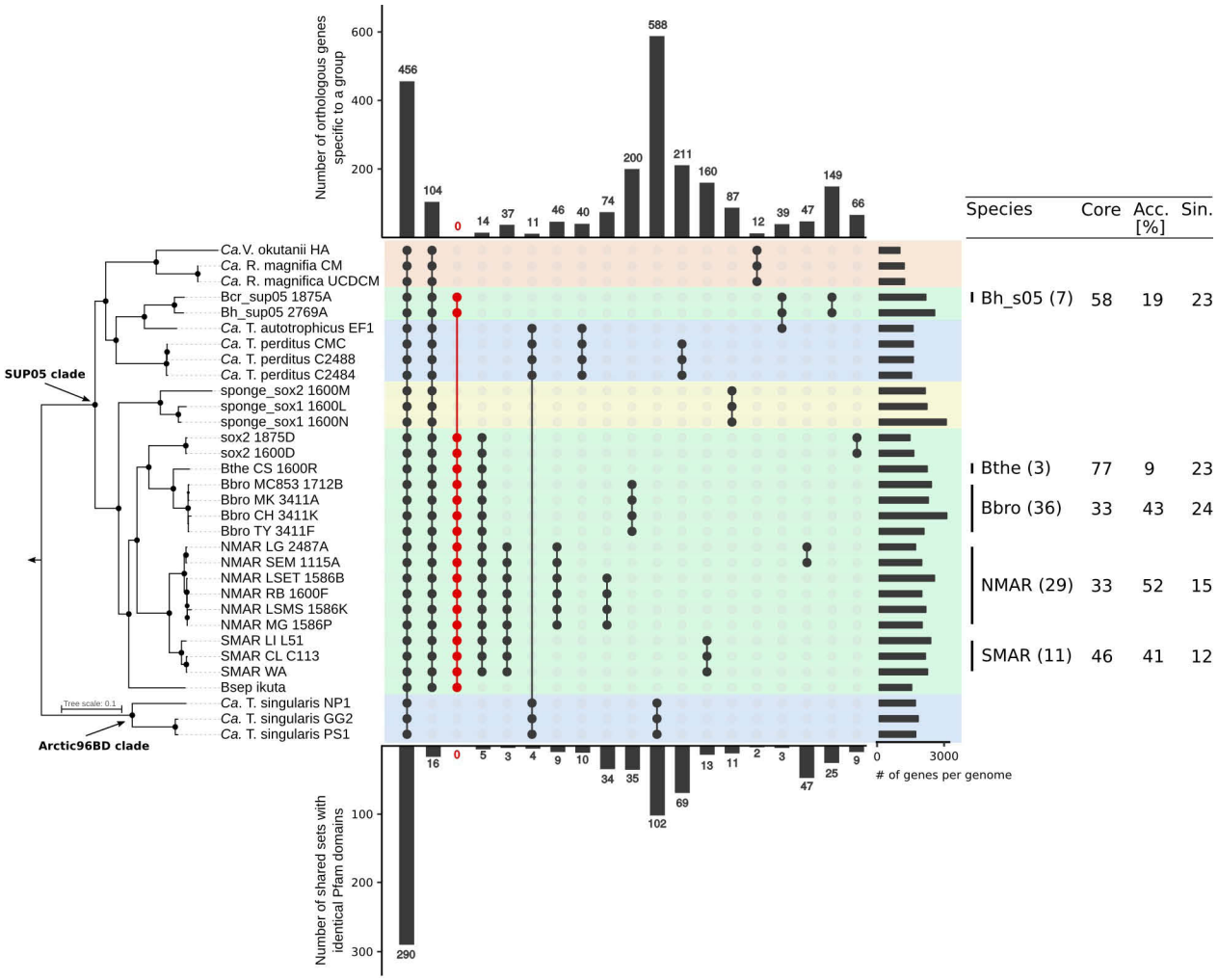


Figure 3 | Gene content similarity among SOX bacteria of the *Thioglobaceae* lineages of various lifestyles.
Left: Phylogenetic (ML) tree based on protein alignments of core genes shared by all representative metagenome assembled genomes (MAGs). The tree was calculated using the same gene set as used for constructing the phylogenetic tree in **Fig. 2**. The representative MAGs were constructed for each sampling site based on all genes encoded in the SOX symbionts at this site.
Top: We performed a clustering of the orthologous genes among all of these MAGS to investigate lifestyle-specific differences. The colors correspond to four different lifestyles: red: vertically transmitted, intracellular symbiont; green: horizontally transmitted, intracellular symbionts (*Bathymodiolus* symbionts); yellow: sponge symbionts of unknown transmission mode; blue: free-living bacteria.
Bottom: Annotation and analysis of Pfam domains was performed to detect lifestyle-specific differences based in function instead of sequence similarity. Symbiont lineages did not share features unique to endosymbiosis with *Bathymodiolus* mussels (marked in red).
Right: Species-specific core, accessory genomes (Acc.) and singleton genes (Sin.) of all *Thioglobaceae* species (according to average nucleotide identity > 95%) with three or more genomes are displayed in percentages next to the figure.

One common mechanism in the adaptation to new lifestyles and environments is the acquisition of new genes by horizontal gene transfer as well as loss of genes encoding proteins involved in functions that are not needed anymore[69]. To tease apart the role of lifestyle and environment in the *Thioglobaceae* family, we performed a two-step

analysis. First, as the draft metagenome assembled genomes (MAGs) represent a mixture of closely related strains, we produced a single representative MAG per site and SUP05 species, that included all genes encoded. Thus, we ensured that all strains and genomic potentials per SUP05 species were covered for each site. Second, we performed clustering of orthologous protein sequences of these representative MAGs to determine which protein sequences are unique to specific lifestyles (**Fig. 3**).

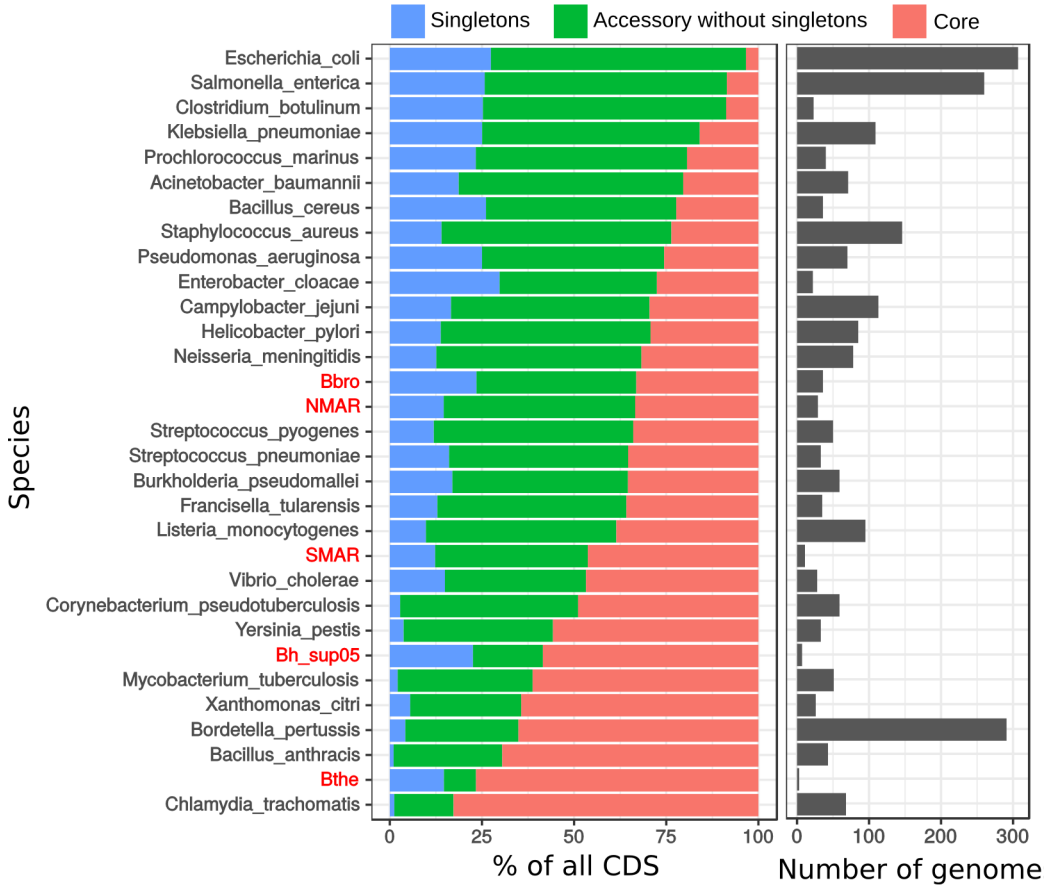


Figure 4 | Number of core, accessory and unique genes in five SUP05 species (in red) in comparison with well-described bacterial species. Only SUP05 species ('Bbro', 'NMAR', 'SMAR', 'Bh_sup05', 'Bthe' - all associated with *Bathymodiolus*) with three or more genomes were included. The other genomes retrieved from Ding *et al.*, 2018, https://academic.oup.com/view-large/107886354.

The core genome shared by all representative MAGs consisted of 456 genes, whereas 104 additional genes were exclusively shared among all the members of the SUP05 clade. Sulfur oxidation is considered the core metabolic feature shared by all members of this group. Interestingly, all compared species had the potential to

oxidize sulfur but the pathways differed substantially between the SUP05 and Arctic96BD-19 (*Ca.* 'Thioglobus' singularis) clades. In fact, *Ca.* 'Thioglobus' singularis genomes lacked the *Sox* sulfur oxidation enzyme system and the *Dsr* dissimilatory sulfite reductase system, the most common mechanisms for the oxidation of reduced sulfur compounds, as well as intermediate steps of the 2-thiouridine sulfur relay system. In cultivation experiments, sulfur oxidation and the formation of sulfur globules has been observed in *Ca.* 'Thioglobus' singularis and the presence of thiosulfate increased heterotrophic growth in this species isolate[24]. However, only the *Apr* system, which facilitates the oxidation of sulfite to sulfate[18], is encoded in the genome of *Ca.* 'Thioglobus' singularis. Instead, *Ca.* 'Thioglobus' singularis lineages included in this study encoded the proteins that are needed to use the extracellular amino acid taurine, which could potentially be used as a sulfur source[70]. In line with this, we found amino acid transport and metabolism mechanisms enriched in KEGG and COG categories of *Ca.* 'Thioglobus' singularis genomes (**Tab. S4, S5**).

We hypothesized that convergent evolution of a particular lifestyle, such as the colonization of an animal host, is based on similar mechanisms and therefore reflected in the presence of orthologous genes that may have been horizontally acquired. Therefore, we tested whether we can identify shared orthologous protein-coding genes that are specific to some of the lifestyles in the *Thioglobaceae*. First, we compared the host-associated and free-living lifestyles and did not detect any orthologous gene sets specific to a symbiotic lifestyle within the SUP05 clade. In contrast, we could detect 11 genes that were unique to all free-living *Thioglobaceae* lineages. These included a malate synthase A and isocitrate dehydrogenase phosphatase/kinase, both part of the glyoxylate cycle[71] for the use of alternative carbon-sources. Additionally, some of these genes encoded proteins involved in

antimicrobial defense. 40 additional genes unique to the free-living lineages within the SUP05 clade included proteins involved in propionate catabolism, reductive synthesis of deoxyribonucleotides from ribonucleotides, nitric oxide reduction, glycosyltransferases and a putative extracellular protein. These proteins represent potential candidates for gene content specific to the free-living lifestyle that may have been convergently lost in the host-associated lineages.

We also considered the different symbiotic lifestyles: intracellular and vertically transmitted, intracellular and horizontally transmitted, and symbionts of unknown transmission mode and cellular location. We could identify 12 genes that were exclusively shared by the vertically transmitted clam symbionts and 87 genes unique to the sponge symbionts. However, these two clades form monophyletic groups associated with specific host groups, hence it is not possible to distinguish between phylogenetic differentiation and lifestyle-specific gene signatures. As stated above, the polyphyletic clade of horizontally transmitted *Bathymodiolus* symbionts allows to tease apart phylogeny and ecology. Intriguingly, we could not identify gene signatures unique to the *Bathymodiolus*-associated lifestyle. In contrast, when we considered all intracellular symbionts independent of the host animal group and transmission mode, we could detect two genes unique to this group. These were annotated as *LemA*-like protein, a transmembrane protein with unknown function and a heat-shock protein *HtpX* with possible protease activities. In addition to lifestyle (i.e. association with a host, free-living), the environment could select for genetic signatures in bacterial lineages. One sampling site in our dataset, Chapopote, harbors four distinct host-associated SUP05 species, but there were no genes specific to all SUP05 bacteria from this site that were missing from all others (**Fig. 3**).

We expected that genetic features specific to horizontally transmitted *Bathymodiolus* symbionts might not be detectable by the orthologous clustering approach, as this is mainly based on sequence similarity. We have shown that the *Bathymodiolus* symbionts form a polyphyletic group, thus theoretically, the same function could be carried out by genes that are too dissimilar at the sequence level to be detected by this method. Therefore, we performed an additional clustering based on the functional similarity of proteins as inferred by their Pfam domain composition. In this analysis, the protein clusters are not defined based on sequence similarity, but on their identical domain composition, which we would expect if the proteins have similar functions. Only proteins that shared all their predicted domains were grouped together. Consistent with the orthologous clustering, this analysis did not reveal proteins with identical domain composition unique to the *Bathymodiolus* SOX symbionts (**Fig. 3**).

As a second approach, we aimed to determine whether functional profiles based on genetic repertoires are more strongly influenced by the lifestyle or the environment. Therefore, we assigned each gene to clusters of orthologous groups (COGs) of broad- (COG letters) and fine-scale (COG numbers) categories and used these to infer genome similarities based on their functional profiles **(Fig. 5**). The clustering based on functional profiles revealed that the genomes grouped mainly according to phylogenetic relatedness rather than lifestyle, even when fine-scale COG numbers were used. This finding aligns with the lack of genes unique to horizontally transmitted symbionts of *Bathymodiolus* mussels. Especially members of the SUP05 subclade A showed high similarity in their functional profiles. Subclade A consists of intracellular symbionts with both horizontal and vertical transmission, and free-living lineages, together comprising a highly-supported subclade within the SUP05 at genus

level (**Fig. 2, 5**). We subsequently analyzed which functional categories (COG) mainly defined two principal components, and detected toxin-related genes such as *RTX* and *Rhs*, mobile elements and restriction-modification systems, which can be considered drivers of the functional separation between clusters (**Tab. S6**). The evolutionary history of different toxin classes has recently been elucidated in the *Bathymodiolus* SOX symbiosis, showing extensive variation in the presence and number of toxin-related genes according to lineage or clade[29,36,30]. This is in line with our clustering analysis and indicates that phylogenetic factors, rather than metabolic functions drive differences between the clades. Sayavedra *et al.*[36] also identified an enrichment of mobile elements, such as restriction-modification (RM) systems, transposases and integrases in *Bathymodiolus* symbionts compared to related lineages within the SUP05-clade and hypothesized that these mechanisms play a role in mediating possibly lineage-specific genome rearrangements. In addition to enrichment in mobile elements such as transposases and RM-systems (**Tab. S4, S5**), we identified CRISPR-Cas systems within the genomes of symbiotic SUP05-lineages (**Fig. S4**). CRISPR-Cas systems are a prokaryotic defense mechanism against bacteriophages and foreign DNA[72]. Our results show that except for a single species, 'Bh_sup05', all symbiont species associated with *Bathymodiolus* mussels encoded *Cas* genes, up to 14 per genome, with high variability between individuals and sites. The sponge-associated SUP05 lineages also encoded *Cas* genes, two to nine per genome. In contrast, neither the free-living nor the vertically transmitted lineages in the SUP05 group *sensu-lato* encoded any *Cas* gene. *Bathymodiolus* and sponge symbionts therefore clearly encode a much higher number of mobile elements than SUP05-lineages with other lifestyles.

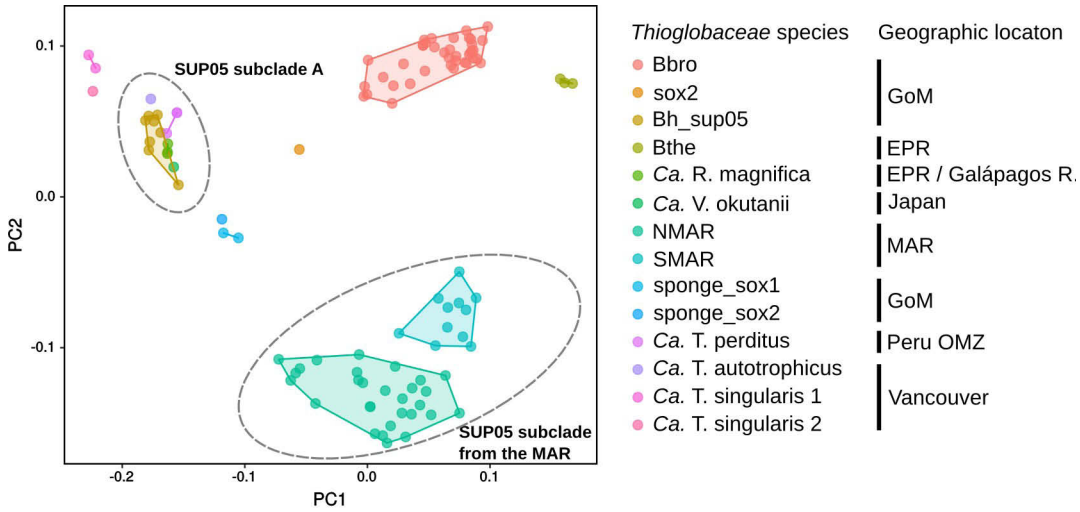


Figure 5 | Clusters of orthologous groups (COGs, fine categories) among SOX bacteria of the *Thioglobaceae* family. Functional profiles generally corresponded more to phylogenetic relatedness than to lifestyle. Particularly the cluster SUP05 subclade A (also see **Fig. 2**) appears very similar. The largest difference can be observed in principle component 2 (PC2) between *Bathymodiolus* SOX symbionts from the Mid Atlantic Ridge (MAR) compared to those from other ocean basins. However, these contain different host and symbiont species and thus the driving factors for the separation are unclear. Geographic locations are indicated and can also be found in **Fig. 1**. GoM: Gulf of Mexico, EPR: East Pacific Rise, Galápagos R.: Galápagos Ridge, MAR: Mid Atlantic Ridge, OMZ: Oxygen Minimum Zone.

Gene content variation among strains within SUP05 populations

Extensive gene content variation among highly related strains within single *Bathymodiolus* mussels has been described recently[10] (Chapter II). Such strain diversity is usually overlooked with standard genome binning approaches. These are not sensitive enough to tease apart highly similar strain genomes resulting in consensus genomes of potentially many different strains. Here, we built on our previous analyses and expanded our approach to the entire SUP05 clade. Our aim was to test whether within-population variation in gene content is a general feature of host-associated and free-living SUP05 bacteria across different bacterial species, sampling sites, and animal hosts for those with a symbiotic lifestyle (**Fig. 6**). Herein, we define a strain-specific gene as any gene that had a lower read coverage than defined single-copy phylogenetic marker genes from the PhylaAmphora database[73] that are assumed to be encoded by all strains. The strain-specific genes in co-

occurring SOX populations ranged from 2-33% of the protein-coding genes. The percentage of genes that were not encoded by all symbiont strains associated with one host individual was generally similar in individuals sampled from the same site with the exception of individuals from site MC853 in the northern Gulf of Mexico (**Fig. 1, 6)**. The major fraction of strain-specific genes was annotated as 'hypothetical proteins' with unknown functions (**Fig. 6**). The distribution of genes that could be assigned to COG categories revealed that in *Bathymodiolus* and sponge symbionts, strain-specific genes covered all 20 categories (**Fig. 7**). In the free-living SUP05 bacteria, strain-specific genes could be found in all except the four COG categories D (Cell cycle control, cell division, chromosome partitioning), F (Nucleotide transport and metabolism), P (Inorganic ion transport and metabolism) and Q (Secondary metabolites biosynthesis, transport and catabolism).

Most strain-specific genes in *Bathymodiolus* SOX were part of the accessory genomes which was expected as these are not encoded in all members of a species and hence are part of the variable genome. Surprisingly, we detected that a prominent number of strain-specific genes, as detected by low read coverage within SUP05 populations (e.g. within a single mussel individual), could also be assigned to the core genome of single SUP05 species (**Fig. 6**). This revealed that there were genes that appeared to be essential to all investigated populations, but not to each bacterial cell in a SUP05 species. Thus by definition, these proteins should not be regarded as part of the core genome. All COG categories were found among these genes. The strain-specific genes of the species core genome included annotations involved in the following processes that could shift the niche regimes of the different strains: denitrification (e.g. *Nar*, *NarK*), phosphorous acquisition (e.g. *PhnA*, polyphosphate kinase 2), iron-related transporters and amino acid synthesis.

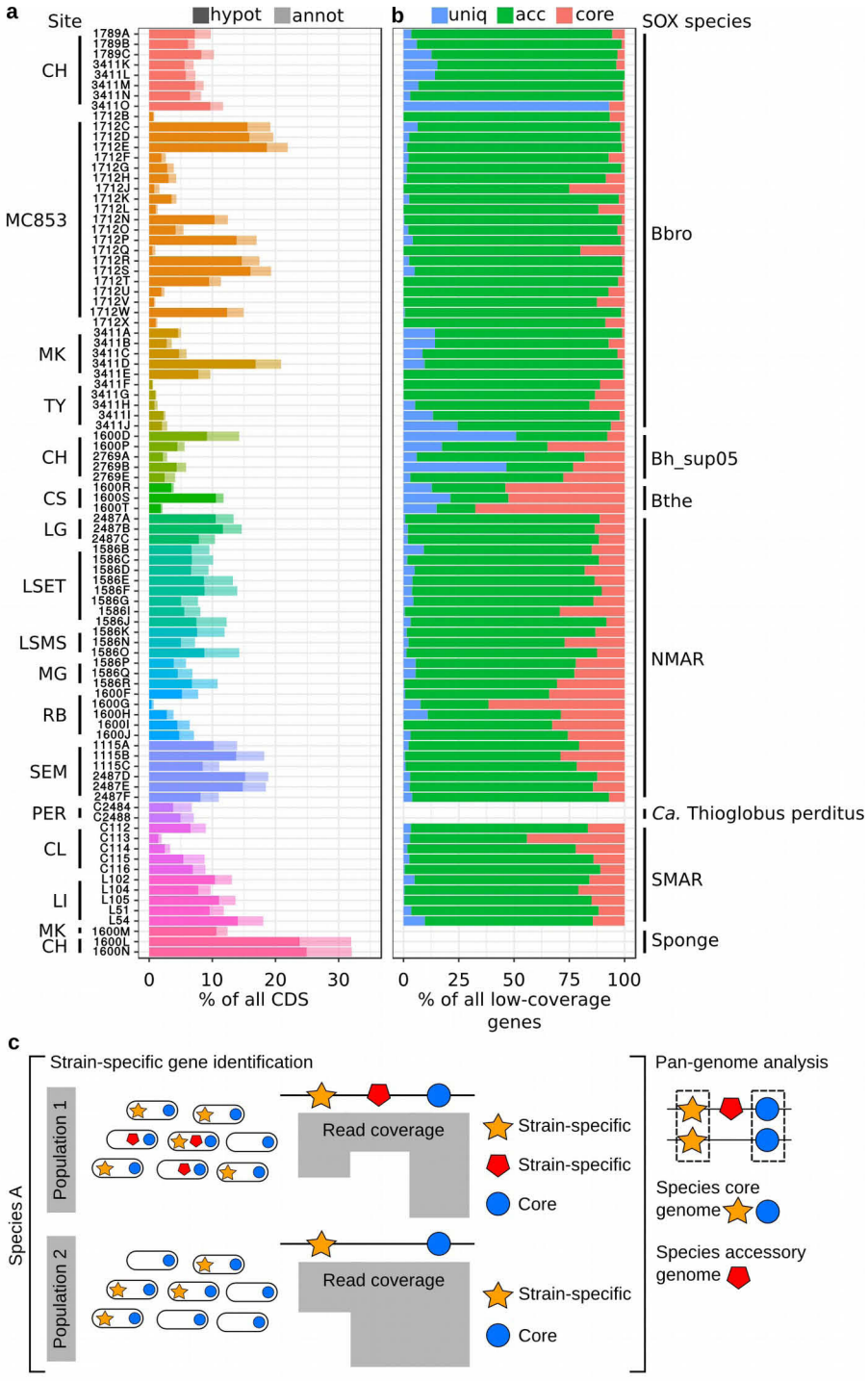


Figure 6 | Strain-specific genes of SOX bacteria of the SUP05 clade. (a) The percentage of genes that were classified as strain-specific in the SOX bacteria. Colors correspond to SUP05 species and sampling site. For most sites the amount of strain-specific genes is similar in co-occurring host individuals. Site MC853 is the exception where the number of strain-specific genes highly depends on the individual host. **(b)** Strain-specific genes in the SOX symbionts were assigned to core, accessory and unique genes for each species. Up to 50% of strain-specific genes are part of the species core genome. **(c)** Concept of the analysis: A strain-specific gene within a population is detected by low coverage of the sequencing reads mapped to the SUP05 draft genome of a sample (e.g. pentagon symbol). The unique, accessory and core genomes within each SUP05 species were identified in our pan-genome analysis using the draft genome per sample. These draft genomes cannot indicate whether a gene was strain-specific in the population or not. This results in the identification of strain-specific genes that are part of the species core genome (e.g. star symbol).

We extended the gene content variation analysis to the MOX symbiont of *Bathymodiolus,* which has recently been shown to have less nucleotide diversity, and possibly less strain variability, within single hosts compared to the SOX[34] (see Chapter V). Consistent with this, we discovered less gene content variation in the MOX symbiont compared to many SOX populations, with a proportion of strain-specific genes ranging from 2 to 12% of their genes within single populations (**Fig. 6, S3**).

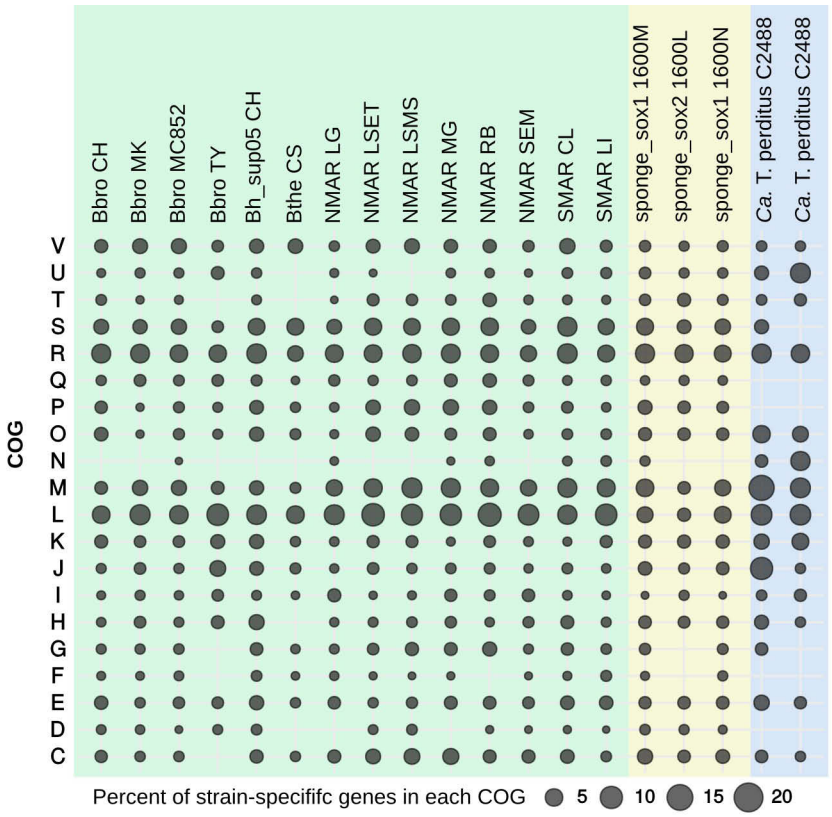


Figure 7 | Relative proportions of strain-specific genes at each site assigned to COG categories. Species and sites are shown above the panel. Proportions are relative to the total count of strain-specific genes within each group. The colors correspond to three different lifestyles: green: horizontally transmitted, intracellular symbionts (*Bathymodiolus* symbionts); yellow: sponge symbionts of unknown transmission mode; blue: free-living bacteria. Genes that could not be assigned to a COG category are not shown. COG categories are represented as follows: [V] Defense mechanisms, [U] Intracellular trafficking, secretion & vesicular transport, [T] Signal transduction mechanisms, [S] Function unknown, [R] General function prediction only, [Q] Secondary metabolites biosynthesis, transport & catabolism, [P] Inorganic ion transport & metabolism, [O] Post-translational modification, protein turnover & chaperones, [N] Cell motility, [M] Cell wall/membrane/envelope biogenesis, [L] Replication, recombination & repair, [K] Transcription, [J] Translation, ribosomal structure & biogenesis, [I] Lipid transport & metabolism, [H] Coenzyme transport & metabolism, [G] Carbohydrate transport & metabolism, [F] Nucleotide transport & metabolism, [E] Amino acid transport & metabolism, [D] Cell cycle control, cell division, chromosome partitioning, [C] Energy production & conversion. A large fraction of strain-specific genes (34 to 77%) could not be assigned to any COG category.

Discussion

Phylogeny, rather than lifestyle drives genetic composition in the Thioglobaceae family

Our results reveal the genetic diversity within and between 16 species of mostly environmental samples of bacteria belonging to the *Thioglobaceae* family. Bacteria from this group are widespread in the world's oceans and have evolved distinct lifestyles, including the association with various invertebrate hosts. We hypothesized that lifestyle-specific gene content has evolved within the *Thioglobaceae* to allow them to thrive in different habitats. Indeed, we detected proteins specific to either the symbionts of clams, or the symbionts of sponges. However, these symbionts that associate with the same group of animal hosts, sharing lifestyle traits such as transmission mode and location within or outside host cells, belonged to two distinct monophyletic clades. Therefore, these traits could potentially be linked to a number of factors including transmission mode, the host animal group, or the phylogenetic history. For the polyphyletic group of free-living *Ca.* 'Thioglobus' singularis strains and SUP05 species, we identified a number of proteins that are potentially specific to a free-living lifestyle. Some of the functional annotations (e.g. proteins involved in the glyoxylate bypass) suggested importance when environmental resources, such as carbon, vary. In contrast to free-living bacteria that are subjected to currents and differences in the availability of resources, symbiotic lineages may experience more uniform availabilities of these resources by residing in a host that keeps them in favorable conditions. Therefore, some of these functions may not be necessary or beneficial in the symbiotic lineages and have potentially been lost throughout evolution[74].

Similar to the free-living lifestyle, we expected shared genetic traits that are specific to lineages associated with mussel hosts from the genus *Bathymodiolus*. We showed that the symbionts colonizing *Bathymodiolus* form a polyphyletic group, a feature that could enable us to tease apart whether gene content overlaps are explained by phylogeny or lifestyle. To our surprise, none of the analyses on orthologous proteins, protein domains, and functional COG profiles revealed genes common to all *Bathymodiolus* symbionts. This suggests that *Bathymodiolus* symbionts have evolved different mechanisms to occupy the same niche, which is living as intracellular symbionts of very closely related host animals from the same genus. Such signatures could result from convergent evolution which has been suggested to affect other divergent bacterial lineages, such as ruminant pathogens and sponge symbionts, leading to the emergence of similar functions with different genetic solutions[75,76]. Similarly, genes encoding the key enzyme in carbon fixation RubisCO have been suggested to have emerged by convergent evolution in SOX symbionts, including the *Bathymodiolus* symbionts[77]. Intriguingly, the SUP05 lineages associated with sponges included in this study are the only SUP05 bacteria known to encode a type 1C RubisCO, whereas other lineages encode either type 1A or type II, including another *Thioglobaceae* lineage from deep-sea sponges[78] (Maxim Rubin-Blum, *personal communication*). This supports our hypothesis that phylogeny, rather than lifestyle drives genetic composition in the SUP05 clade. Previous findings revealed massive variation in the distribution of genes potentially involved in host-symbiont interaction among lineages of *Bathymodiolus* symbionts[29,30]. Expansion and loss of toxin-related genes, often replaced by secretion systems, support our hypothesis of convergent evolution of distinct host-microbe interaction mechanisms in *Bathymodiolus* symbioses.

In addition to the host and the type of symbiotic association, environmental conditions could select for shared genetic signatures among sympatric bacterial lineages inhabiting a particular environment. However, just as we found no genomic signatures shared by all symbiotic or *Bathymodiolus*-associated lineages, the four different SUP05 species sampled at the same seep site did not share any unique gene sets. Instead, with the exception of the free-living lineages, we could only detect gene content specific to closely related lineages within sub-clades. Thus, our results indicate that genetic repertoires mostly resemble phylogenetic relatedness rather than ecological differentiation, encompassing both lifestyle and environmental conditions. One key feature uniting these bacteria is their sulfur metabolism linked to the fixation of inorganic carbon. Considering the variety of lifestyles within the clade, this suggests that the core genetic backbone in SUP05 bacteria provides a repertoire that is adaptable to these different lifestyles and can support an exclusively free-living lifestyle, or provides benefits to an animal host. This is supported by the fact that chemosynthetic symbioses with different animal hosts, such as *Bathymodiolus* mussels, vesicomyid clams and sponges have evolved multiple times over the course of evolution, possibly from free-living lineages[27,30]. Our data show that rather than one optimal solution, *Bathymodiolus* symbionts may have evolved distinct genetic solutions that allowed the establishment, interaction and persistence of these intimate associations through convergent evolution.

Genetic flexibility is key

For the first time, we could estimate core and accessory genomes for five different species of *Bathymodiolus* symbionts in order to determine their flexible gene pool in nature. This revealed that the extent of genomic flexibility in *Bathymodiolus*

symbionts is similar to medically relevant human-associated bacteria. Effective population size, adaptation to new niches, and lifestyle have been suggested to be major factors driving accessory genome size[6]. For example, a reduction in effective population size due to strong physical bottlenecks during transmission of the human pathogen *Chlamydia trachomatis* may be an explanation for their very small accessory genomes (see **Fig. 4**). In contrast, bacterial species with lots of different lifestyles such as *Escherichia coli*, or *Prochlorococcus marinus*, both with potentially enormous effective population sizes[3], had very large accessory genomes (**Fig. 4**). In *Bathymodiolus* symbionts, effective population size could be large, considering their horizontal transmission and the high density of host colonies. However, the transmission and colonization process is not fully understood and the effective population size is still unknown[79]. In fact, effective population sizes are extremely difficult to quantify in most natural bacterial populations. Migration and adaptation to new niches may be essential for the success of SUP05 symbionts in the ephemeral and dynamic habitats, like hydrothermal vents and cold seeps, in which the environmental conditions change with location and time over scales from seconds to thousands of years, and from millimeters to thousands of kilometers[80]. Even within a single animal host individual, distinct niches might be occupied by different symbiont strains[10,31]. The necessity to adapt to new niches may be one of the defining features of life at vents and seeps, and thus one of the main factors driving gene content variation in *Bathymodiolus* symbionts and free-living SUP05.

In theoretical models, the more phylogenetically related two co-occurring strains are, the more likely they have the same growth requirements, which would likely result in competition[81,82]. Increased genetic variation among highly related strains can allow these strains to occupy different niches and therefore co-exist[10] (Chapter II).

Populations with strains that have distinct genetic repertoires allow flexibility of the consortium to react to changing conditions through shifts in the strain abundances. Some strain-specific genes may be subject to negative frequency-dependent selection; hence these genes only provide a benefit when rare in the population[83,84]. For example this could apply to denitrification genes that have already been shown to be variable in populations of SUP05 symbionts in *Bathymodiolus* mussels of the MAR[10] (Chapter II) and *Bathymodiolus septemdierum*[31] which was confirmed for other SUP05 lineages and hosts in this study. Lower abundance of some of these genes in the population could be selected for by costs of carrying that gene, avoidance of accumulation of metabolic intermediates and possibly the exchange of intermediates between bacterial cells. This hypothesis is supported by our findings that a substantial fraction of variable genes within single populations are part of the species core genome spanning multiple sites and even host species (e.g. NMAR-SOX). These genes are present in each host individual sharing the same SUP05 species. However, they are not encoded by each strain in the community. We hypothesize that the encoded proteins are potentially involved in functions that are essential for the population or host while they are not essential to every single strain in the community. Therefore, these could be considered conserved core functions only at the population level.

Also evolvability itself, referred to as the potential of a population to evolve adaptive solutions to unknown future conditions (www.nature.com/subjects/evolvability *last accessed 1rst Feb 2019*) could be favored by selection. This could be the case if (acquired) variable genes are neutral or slightly deleterious and the cost of keeping a gene can be balanced by the benefit of adaptability to a potential environmental change[85]. A wider repertoire of mobile genetic elements could increase the frequency

of gene loss and acquisition, increasing genetic variation on which selection can act upon. This may result in persistence of within-population heterogeneity predominantly in non-constant environments[85].

The key to success in the SUP05 clade may be the ability to keep their genomic set up flexible, first on evolutionary time scales in terms of plasticity in their gene content as has been suggested for the free-living lineages[18], and second on very short time scales with mosaics of strains occupying distinct micro-niches. SUP05 bacteria partition these capabilities into distinct strains with small genomes rather than one versatile strain that carries the potential to deal with many different conditions. One factor that can increase genomic plasticity, and therefore lead to convergent evolution of similar lifestyles with different genetic set ups, is horizontal gene transfer (HGT). Metabolic genes, such as the hydrogenase, have been shown to be affected by HGT in symbiotic SUP05 bacteria[77]. This, as well as the duplication and loss of genes, can lead to variation in gene content and increase the accessory genome of bacterial species.

Phage-mediated exchange of genetic material

Evolvability of the SUP05 clade could be selected for by a higher number of mobile elements increasing or regulating gene acquisition and loss. Indeed, multiple RM and CRISPR-Cas systems were identified in *Bathymodiolus* SUP05 symbionts[10,36] (Chapter II). RM systems have been described to be involved in phage defense, increase of genomic variation, control of HGT and stabilization of genomic islands[46]. In addition, naturally competent bacteria, such as *Helicobacter pylori*, have been characterized by increased numbers of RM systems[86]. *Bathymodiolus* SUP05 symbionts have been

shown to be highly enriched in RM systems compared to their free-living and vertically transmitted relatives[36] which may be one factor in increasing genetic diversity within these populations.

Our results show that *Bathymodiolus* and sponge symbionts have a variety of CRISPR-Cas systems, as well as CRISPR-spacer sequences, representing records of past phage infections (**Fig. S4**)[87,72]. We can therefore assume that *Bathymodiolus* and sponge symbionts have been infected by bacteriophages, which are considered important mediators of HGT[88,89]. Although we could not identify CRISPR-Cas systems in the free-living SUP05 members, massive phage infection has been shown for bacteria of this group with sulfur-oxidation genes in the phage gene pool[90,25]. Therefore phages likely are one of the driving forces of genetic diversity in the SUP05 clade, possibly leading to an increase in evolvability[88].

Assuming that HGT plays a role for most members of the SUP05 clade, it was puzzling to discover that symbionts of sponges and *Bathymodiolus* were so heavily enriched in CRISPR-Cas and RM systems compared to the free-living relatives, particularly because their location within a host may offer a sheltered environment from phage predation (**Tab S4, S5, Fig. S4**). However, phages have been shown to be able to even reach bacteria that reside intracellularly in eukaryotic cells[91]. One factor that could potentially increase the selective pressure towards increased rates of HGT may be the association with a host, adding the ‘meta-organism’ as unit of selection. For a host at hydrothermal vents and seeps, it might be more beneficial to harbor a flexible and genomically adaptable symbiont population as has been suggested for this and other systems before[10,92] (Chapter II). Compared to a free-living bacterial population, this could increase selective pressure towards

mechanisms that increase evolvability within the symbiont population through frequent exchange of genetic material. Alternatively, phage predation could be enhanced in the low-diversity host-associated microbial communities, as opposed to the more diverse free-living communities (e.g. kill-the-winner hypothesis). Thus, phage defense mechanisms may have been selected for in symbionts to avoid extinction[93]. RM- and CRISPR-Cas systems are both variable among co-occurring strains[10,94], and RM systems have been shown to be differentially regulated via phasing in bacterial population[95]. This leads to different susceptibility of co-occurring strains to phage infection and HGT in general.

Conclusion

Genetic diversity, co-existence and genomic plasticity are important factors for microbial community dynamics, particularly in host-associated microbial species. In humans, even small differences in the genetic set up may determine whether a microbial partner is beneficial or harmful to our health[12]. Our study sheds light on the diversity of bacteria belonging to the widespread SUP05 clade, revealing a 'hidden' within-population diversity which has previously been overlooked. We show that a substantial fraction of up to 33% of all genes can be variable in single populations (e.g. within single host individuals), some of which appear to be core functions present in all populations. Distinct co-existing strains could therefore be considered 'jigsaw pieces' to fulfill a composition of functions as a consortium instead of a single versatile strain.

Except for the free-living lineages, we could not detect genetic signatures that can be confidentially considered as features of a particular lifestyle and concluded that in

the SUP05 clade, the orthologous gene repertoire is mainly reflected by phylogeny rather than ecology. Instead we suggest that convergent evolution may have led to the emergence of similar lifestyles (e.g. the association with *Bathymodiolus* mussels) with different genomic solutions in divergent bacterial lineages. We further hypothesize that selection for increased evolvability could create conditions for multiple distinct lifestyles to emerge and explain persistence of within-population heterogeneity in the SUP05 clade. As postulated previously by Hunt *et al.*[4], *"[…] a large (flexible) gene pool […], if shared by horizontal gene transfer, gives rise to large numbers of ecologically adaptive phenotypes."*. This aligns with our results which suggest that genomic plasticity in closely related lineages of the SUP05 clade appears to facilitate adaptation to various different lifestyles and habitats. Integrating laboratory studies on strain competition, co-existence and genetic exchange with high-resolution studies on natural communities, such as this one, is key to develop testable hypotheses. These can help us understand the dynamics in natural microbial communities in a changing environment.

Acknowledgements

We thank the captains, crews and ROV teams on the cruises BioBaz (2013), ODEMAR (2014), M126 (2016), M78-2 (2009), NA58 (2015), NA43 (2014), M114-2 (2015), AT26-23 (2014), ATA57 (2008) and M93 (2013) on board of the research vessels Pourquoi Pas?, FS Meteor, E/V Nautilus, and L'Atalante and the chief scientists on these research expeditions. This study was funded by the Max Planck Society, the MARUM DFG-Research Center / Excellence Cluster "The Ocean in the Earth System" at the University of Bremen, the German Research Foundation, an ERC Advanced

Grant (BathyBiome, 340535), and a Gordon and Betty Moore Foundation Marine Microbial Initiative Investigator Award to ND (Grant GBMF3811).

References for chapter 3

1. Delmont, T. O. & Eren, A. M. Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. *PeerJ* **6**, e4320 (2018).

2. Kashtan, N. *et al.* Fundamental differences in diversity and genomic population structure between Atlantic and Pacific *Prochlorococcus*. *ISME J.* **11**, 1997 (2017).

3. Kashtan, N. *et al.* Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* **344**, 416–420 (2014).

4. Hunt, D. E. *et al.* Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **320**, 1081–1085 (2008).

5. Konstantinidis, K. T. & Tiedje, J. M. Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 2567–2572 (2005).

6. McInerney, J. O., McNally, A. & O'Connell, M. J. Why prokaryotes have pangenomes. *Nat. Microbiol.* **2**, 17040 (2017).

7. Vos, M. & Eyre-Walker, A. Are pangenomes adaptive or not? *Nat. Microbiol.* **2**, 1576 (2017).

8. Ding, W., Baumdicker, F. & Neher, R. A. panX: pan-genome analysis and exploration. *Nucleic Acids Res.* **46**, e5–e5 (2018).

9. Ellegaard, K. M. & Engel, P. Genomic diversity landscape of the honey bee gut microbiota. *Nat. Commun.* **10**, 446 (2019).

10. Ansorge, R. *et al.* Diversity matters: Deep-sea mussels harbor multiple symbiont strains. *bioRxiv* 531459 (2019).

11. Pankey, M. S. *et al.* Host-selected mutations converging on a global regulator drive an adaptive leap towards symbiosis in bacteria. *eLife* **6**, e24414 (2017).

12. Greenblum, S., Carr, R. & Borenstein, E. Extensive strain-level copy-number variation across human gut microbiome species. *Cell* **160**, 583–594 (2015).

13. Vatanen, T. *et al.* Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. *Nat. Microbiol.* **4**, 470–479 (2018).

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

15. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).

16. Callbeck, C. M. *et al.* Oxygen minimum zone cryptic sulfur cycling sustained by offshore transport of key sulfur oxidizing bacteria. *Nat. Commun.* **9**, 1729 (2018).

17. Meier, D. V. *et al.* Niche partitioning of diverse sulfur-oxidizing bacteria at hydrothermal vents. *ISME J.* **11**, 1545–1558 (2017).

18. Murillo, A. A., Ramírez-Flandes, S., DeLong, E. F. & Ulloa, O. Enhanced metabolic versatility of planktonic sulfur-oxidizing γ-proteobacteria in an oxygen-deficient coastal ecosystem. *Front. Mar. Sci.* **1**, (2014).

19. Glaubitz, S., Kießlich, K., Meeske, C., Labrenz, M. & Jürgens, K. SUP05 Dominates the gammaproteobacterial Sulfur oxidizer assemblages in pelagic redoxclines of the central baltic and black seas. *Appl. Environ. Microbiol.* **79**, 2767–2776 (2013).

20. Ulloa, O., Canfield, D. E., DeLong, E. F., Letelier, R. M. & Stewart, F. J. Microbial oceanography of anoxic oxygen minimum zones. *Proc. Natl. Acad. Sci.* **109**, 15996–16003 (2012).

21. Shah, V., Chang, B. X. & Morris, R. M. Cultivation of a chemoautotroph from the SUP05 clade of marine bacteria that produces nitrite and consumes ammonium. *ISME J.* **11**, 263–271 (2017).

22. Shah, V. & Morris, R. M. Genome sequence of "*Candidatus* Thioglobus autotrophica" strain EF1, a chemoautotroph from the SUP05 clade of marine Gammaproteobacteria. *Genome Announc.* **3**, (2015).

23. Marshall, K. T. & Morris, R. M. Genome sequence of "*Candidatus* Thioglobus singularis" strain PS1, a mixotroph from the SUP05 clade of marine Gammaproteobacteria. *Genome Announc.* **3**, (2015).

24. Marshall, K. T. & Morris, R. M. Isolation of an aerobic sulfur oxidizer from the SUP05/Arctic96BD-19 clade. *ISME J.* **7**, 452–455 (2013).

25. Roux, S. *et al.* Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *eLife* **3**, e03125 (2014).

26. Anantharaman, K., Breier, J. A., Sheik, C. S. & Dick, G. J. Evidence for hydrogen oxidation and metabolic plasticity in widespread deep-sea sulfur-oxidizing bacteria. *Proc. Natl. Acad. Sci.* **110**, 330–335 (2013).

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

28. Nishijima, M. *et al.* Association of thioautotrophic bacteria with deep-sea sponges. *Mar. Biotechnol. N. Y. N* **12**, 253–260 (2010).

29. Sayavedra, L. Host-symbiont interactions and metabolism of chemosynthetic symbiosis in deep-sea *Bathymodiolus* mussels. (University of Bremen, 2016).

30. Sayavedra, L. *et al.* Horizontal acquisition followed by expansion and diversification of toxin-related genes in deep-sea bivalve symbionts. *in prep*

31. Ikuta, T. *et al.* Heterogeneous composition of key metabolic gene clusters in a vent mussel symbiont population. *ISME J.,* **10**, 990-1001 (2016).

32. Liu, J., Liu, H. & Zhang, H. Phylogeny and evolutionary radiation of the marine mussels (Bivalvia: Mytilidae) based on mitochondrial and nuclear genes. *Mol. Phylogenet. Evol.* **126**, 233–240 (2018).

33. Lorion, J. *et al.* Adaptive radiation of chemosymbiotic deep-sea mussels. *Proc. R. Soc. B Biol. Sci.* **280**, 20131243 (2013).

34. Picazo, D. R. *et al.* Horizontally transmitted symbiont populations in deep-sea mussels are genetically isolated. *bioRxiv* 536854 (2019).

35. Rubin-Blum, M. *et al.* Short-chain alkanes fuel mussel and sponge *Cycloclasticus* symbionts from deep-sea gas and oil seeps. *Nat. Microbiol.* **2**, 17093 (2017).

36. Sayavedra, L. *et al.* Abundant toxin-related genes in the genomes of beneficial symbionts from deep-sea hydrothermal vent mussels. *eLife* e07966 (2015).

37. Zhou, J., Bruns, M. A. & Tiedje, J. M. DNA recovery from soils of diverse composition. *Appl. Environ. Microbiol.* **62**, 316–322 (1996).

38. Lee, R. D., Jospin, G., Coil, D. A. & Eisen, J. A. Draft genome sequence of the endosymbiont '*Candidatus* Ruthia magnifica' UCD-CM (Phylum Proteobacteria). *Genome Announc.* **2**, (2014).

39. Kuwahara, H. *et al.* Reduced genome of the thioautotrophic intracellular symbiont in a deep-sea clam, *Calyptogena okutanii*. *Curr. Biol.* **17**, 881–886 (2007).

40. Newton, I. L. G. *et al.* The *Calyptogena magnifica* chemoautotrophic symbiont genome. *Science* **315**, 998–1000 (2007).

41. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

42. Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

43. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).

44. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 (2013).

45. Seah, B. K. B. & Gruber-Vodicka, H. R. gbtools: Interactive visualization of metagenome bins in R. *Front. Microbiol.* **6**, (2015).

46. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).

47. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

48. Brettin, T. *et al.* RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci. Rep.* **5**, 8365 (2015).

49. Wattam, A. R. *et al.* Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res.* **45**, D535–D542 (2017).

50. Rodriguez-R, L. M. & Konstantinidis, K. T. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Inc.* (2016).

51. Chaudhari, N. M., Gupta, V. K. & Dutta, C. BPGA- an ultra-fast pan-genome analysis pipeline. *Sci. Rep.* **6**, 24373 (2016).

52. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

53. R Core Team. *A language and environment for statistical computing. R Foundation for Statistical Computing.* (2016).

54. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).

55. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–D595 (2019).

56. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).

57. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

58. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

59. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

60. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

61. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).

62. Sonnhammer, E. L., Eddy, S. R. & Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405–420 (1997).

63. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

64. Kuwahara, H. *et al.* Loss of genes for DNA recombination and repair in the reductive genome evolution of thioautotrophic symbionts of *Calyptogena* clams. *BMC Evol. Biol.* **11**, 285 (2011).

65. Kuwahara, H. *et al.* Reductive genome evolution in chemoautotrophic intracellular symbionts of deep-sea *Calyptogena* clams. *Extrem. Life Extreme Cond.* **12**, 365–374 (2008).

66. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).

67. Konstantinidis, K. T. & Tiedje, J. M. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr. Opin. Microbiol.* **10**, 504–509 (2007).

68. Tu, Q. & Lin, L. Gene content dissimilarity for subclassification of highly similar microbial strains. *BMC Genomics* **17**, 647 (2016).

69. Rubin-Blum, M. *et al.* Fueled by methane: deep-sea sponges from asphalt seeps gain their nutrition from methane-oxidizing symbionts. *ISME J.* **1**, (2019).

70. Chien, C.-C., Leadbetter, E. R. & Godchaux, W. *Rhodococcus* spp. utilize taurine (2-aminoethanesulfonate) as sole source of carbon, energy, nitrogen and sulfur for aerobic respiratory growth. *FEMS Microbiol. Lett.* **176**, 333–337 (1999).

71. Laporte, D. C., Stueland, C. S. & Ikeda, T. P. Isocitrate dehydrogenase kinase/phosphatase. *Biochimie* **71**, 1051–1057 (1989).

72. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).

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

74. Ponnudurai, R. *et al.* Metabolic and physiological interdependencies in the *Bathymodiolus azoricus* symbiosis. *ISME J.* **11**, 463-477 (2017).

75. Lo, W.-S., Gasparich, G. E. & Kuo, C.-H. Convergent evolution among ruminant-pathogenic *Mycoplasma* involved extensive gene content changes. *Genome Biol. Evol.* **10**, 2130–2139 (2018).

76. Fan, L. *et al.* Functional equivalence and evolutionary convergence in complex communities of microbial sponge symbionts. *Proc. Natl. Acad. Sci.* **109**, E1878–E1887 (2012).

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

78. Tian, R.-M. *et al.* Genome reduction and microbe-host interactions drive adaptation of a sulfur-oxidizing bacterium associated with a cold seep sponge. *mSystems* **2**, e00184-16 (2017).

79. Wentrup, C., Wendeberg, A., Schimak, M., Borowski, C. & Dubilier, N. Forever competent: deep-sea bivalves are colonized by their chemosynthetic symbionts throughout their lifetime. *Environ. Microbiol.* **16**, 3699–3713 (2014).

80. Zielinski, F. U., Gennerich, H.-H., Borowski, C., Wenzhöfer, F. & Dubilier, N. 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) *Geochem. Geophys. Geosystems* **12**, (2011).

81. Russel, J., Røder, H. L., Madsen, J. S., Burmølle, M. & Sørensen, S. J. Antagonism correlates with metabolic similarity in diverse bacteria. *Proc. Natl. Acad. Sci.* **114**, 10684–10688 (2017).

82. Zelezniak, A. *et al.* Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proc. Natl. Acad. Sci.* **112**, (2015).

83. Brisson, D. Negative frequency-dependent selection is frequently confounding. *Front. Ecol. Evol.* **6**, (2018).

84. Levin, B. R., Antonovics, J. & Sharma, H. Frequency-dependent selection in bacterial populations [and discussion]. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **319**, 459–472 (1988).

85. Heuer, H., Abdo, Z. & Smalla, K. Patchy distribution of flexible genetic elements in bacterial populations mediates robustness to environmental uncertainty. *FEMS Microbiol. Ecol.* **65**, 361–371 (2008).

86. Vasu, K. & Nagaraja, V. Diverse functions of Restriction-Modification systems in addition to cellular defense. *Microbiol. Mol. Biol. Rev. MMBR* **77**, 53–72 (2013).

87. van der Oost, J., Westra, E. R., Jackson, R. N. & Wiedenheft, B. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **12**, 479–492 (2014).

88. Braga, L. P. P., Soucy, S. M., Amgarten, D. E., da Silva, A. M. & Setubal, J. C. Bacterial diversification in the light of the interactions with phages: the genetic symbionts and their role in ecological speciation. *Front. Ecol. Evol.* **6**, (2018).

89. Lindell, D. *et al.* Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 11013–11018 (2004).

90. Anantharaman, K. *et al.* Sulfur oxidation genes in diverse deep-sea viruses. *Science* **344**, 757–760 (2014).

91. Zhang, L. *et al.* Intracellular *Staphylococcus aureus* control by virulent vacteriophages within MAC-T bovine mammary epithelial cells. *Antimicrob. Agents Chemother.* **61**, e01990-16 (2017).

92. LaJeunesse, T. C. *et al.* Host–symbiont recombination versus natural selection in the response of coral–dinoflagellate symbioses to environmental disturbance. *Proc. R. Soc. B Biol. Sci.* **277**, 2925–2934 (2010).

93. Rodriguez-Valera, F. *et al.* Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol.* **7**, 828–836 (2009).

94. Crawley, A. B., Henriksen, E. D., Stout, E., Brandt, K. & Barrangou, R. Characterizing the activity of abundant, diverse and active CRISPR-Cas systems in lactobacilli. *Sci. Rep.* **8**, 11544 (2018).

95. Hoskisson, P. A. & Smith, M. C. Hypervariation and phase variation in the bacteriophage 'resistome'. *Curr. Opin. Microbiol.* **10**, 396–400 (2007).

Supplementary figures and tables

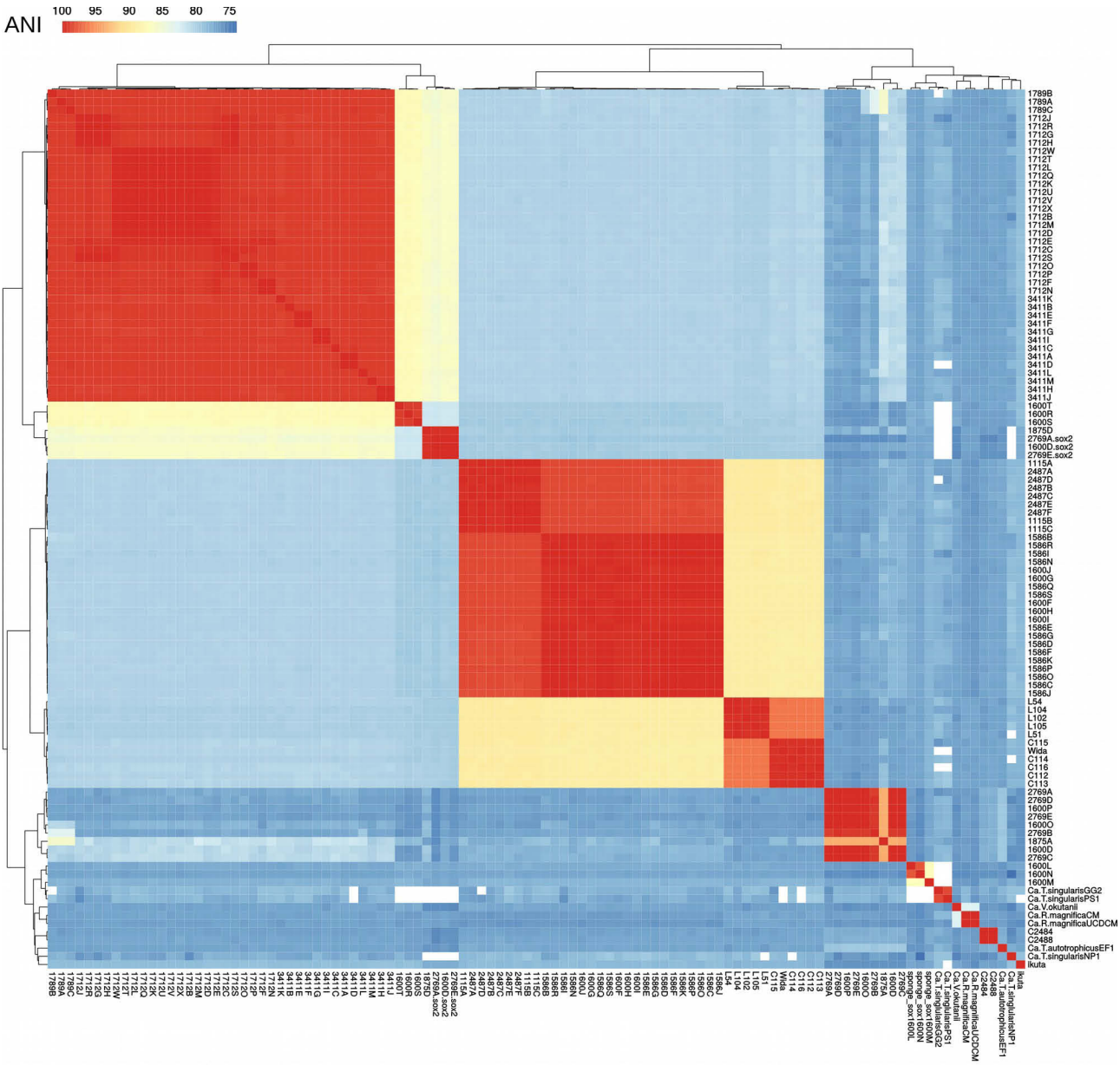


Figure S1 | Average nucleotide identity (ANI) between genomes / draft genomes in this study. A list of samples can be found in Tab. S2. All genomes with ANI > 95% were considered the same species.

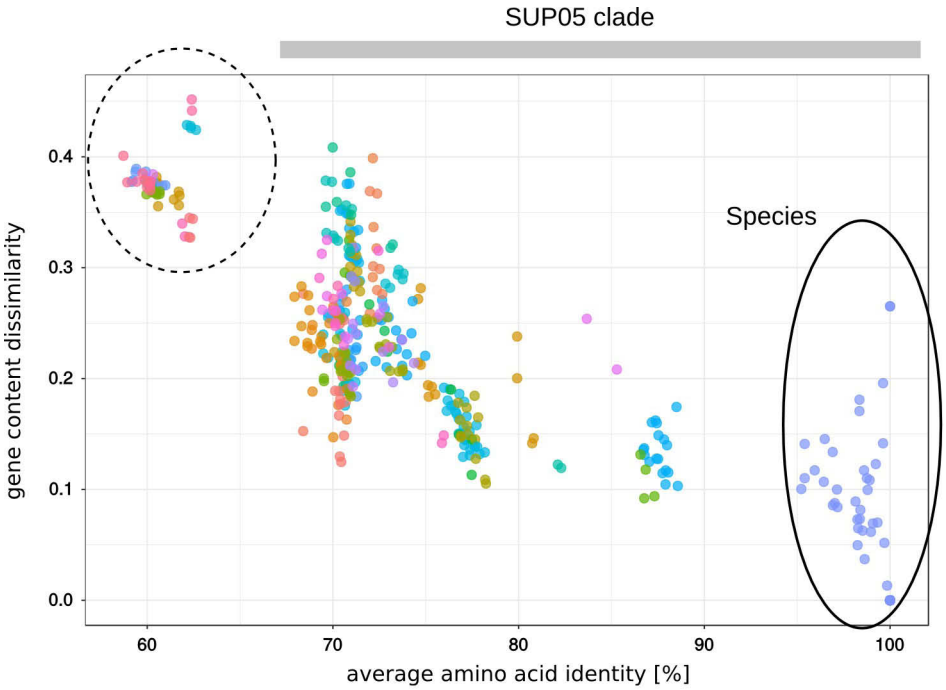


Figure S2 | Pairwise average amino acid identity (AAI) and gene content dissimilarity (Bray-Curtis dissimilarity of presense-absence matrix of orthologous genes). AAI values of > 95% also showed ANI > 95% and were therefore considered the same species (circle with continuous line). The SUP05 clade was considered a single genus in this study, with AAI > 68% and gene content dissimilarity < 0.4. Pairwise comparisons between lineages of the SUP05 clade and Arctic96BD clade are circled by broken line.

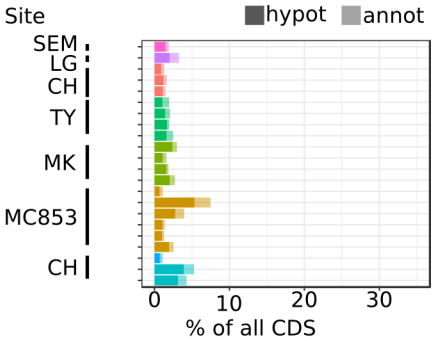


Figure S3 | Percentage of genes that were classified as strain-specific in the *Bathymodiolus* MOX symbionts. Colors correspond to MOX species and sampling site. For all MOX populations, the number of strain-specific genes was similar and did not reach as high percentages at the SOX symbionts. Abbreviations for the sampling sites: SEM: Semenov, LG: Logatchev, CH: Chapopote, MK: Mictlan Knoll, MC853: site MC853.

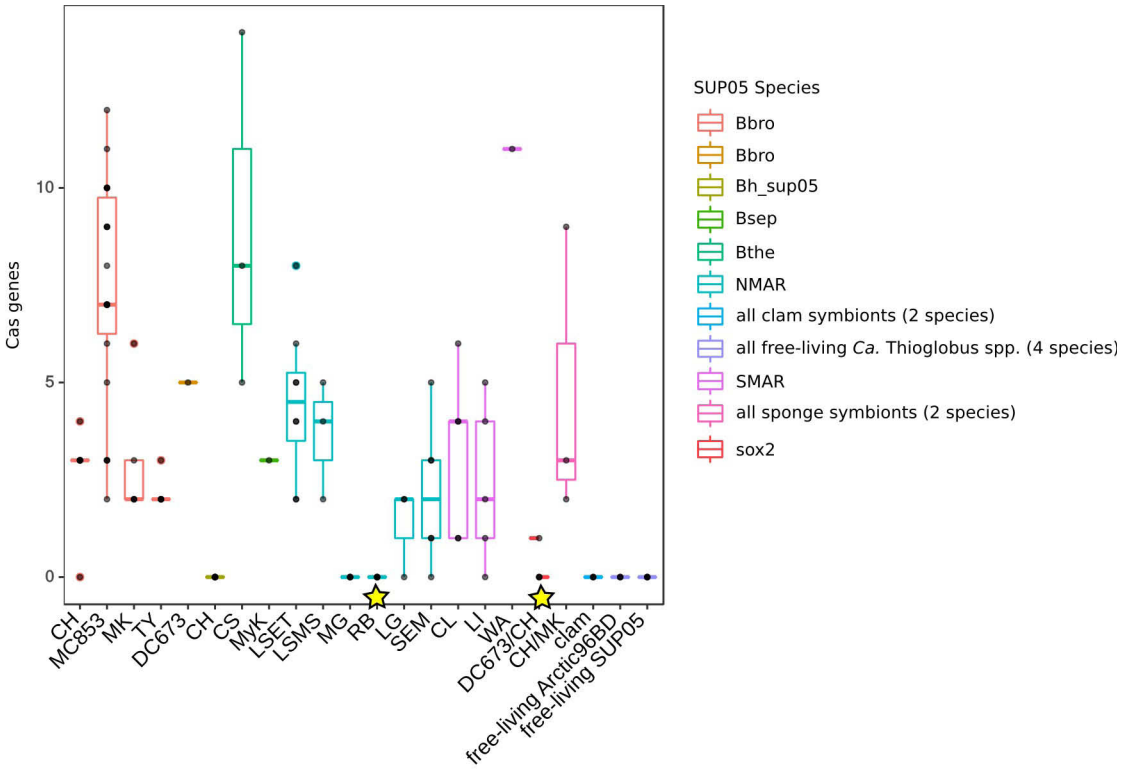


Figure S4 | Numer of Cas genes per draft genome for each sampling site. Yellow star indicates genomes where no Cas genes have been found but only CRISPR-spacers.

Table S6 | Loadings, COG category and annotation for principle component 1 and 2 in Fig. 5. Only highest loadings with a value < -0.1 and > 0.1 are displayed.

PC1	COG	Annotation	PC2	COG	Annotation
-0.10	COG0438_M	Glycosyltransferase involved in cell wall bisynthesis	-0.31	COG3209_M	Rhs family protein
			-0.21	COG2826_L	Transposase and inactivated derivatives, IS30 family
			-0.14	COG3039_L	Transposase and inactivated derivatives, IS5 family
			-0.13	COG3328_L	Transposase (or an inactivated derivative)
			-0.13	COG2801_L	Transposase InsO and inactivated derivatives
			-0.11	COG0532_J	GTPase
			-0.11	COG2865_K	Predicted transcriptional regulator, contains HTH domain
			-0.10	COG1192_D	ATPases involved in chromosome partitioning
0.72	COG2931_Q	RTX toxin	0.23	COG0859_M	ADP-heptose:LPS heptosyltransferase
0.20	COG3209_M	Rhs family protein	0.16	COG2931_Q	RTX toxin
0.16	COG0270_L	DNA-cytosine methylase	0.13	COG0526_OC	Thiol-disulfide isomerase or thioredoxin
0.16	COG0457_R	Tetratricopeptide (TPR) repeat	0.12	COG0790_R	TPR repeat
0.14	COG0286_V	system, DNA methylase subunit	0.11	COG0583_K	A-binding transcriptional regulator, LysR family
0.13	COG3177_S	Fic family protein	0.11	COG1280_E	Threonine/homoserine/homoserine lactone efflux protein
0.13	COG0732_V	Restriction endonuclease S subunit			
0.11	COG2801_L	Transposase InsO and inactivated derivatives			

Table S1 | Accession numbers and references for published genomes used in this study.

Species	Bioproject	Biosample	Accession	Reference
Ca. Thioglobus singularis GG2	PRJNA252014	SAMN02848474	CP008725.1	none
Ca. Thioglobus singularis NP1	PRJNA414010	SAMN07775767	CP023860.1	none
Ca. Thioglobus singularis PS1	PRJNA229178	SAMN04054243	CP006911.1	Marshall and Morris, 2015
Ca. Thioglobus autotrophicus EF1	PRJNA224116	SAMN03280981	NZ_CP010552.1	Shah and Morris, 2015
Ca. Thioglobus perditus CMC	PRJNA421259	SAMN08136125	PNQY00000000.1	Callbeck et al., 2018
Ca. Ruthia magnifica CM	PRJNA16841	SAMN02598375	CP000488	Newton et al., 2007
Ca. Ruthia magnifica UDC-CM	PRJNA236124	SAMN02641591	JARW00000000.1	Lee et al., 2014
Ca. Vesicomyosocious okutanii HA	PRJDA18267	SAMD00060908	AP009247	Kuwahara et al., 2007
B. septemdierum thiotrophic symbiont	PRJDB949	SAMD00024725	AP013042.1	Ikuta et al., 2016

Table S2 | Details of samples from environmental metagenomes. Samples marked in grey were processed in this study. For all draft genomes derived from other studies, the references to the respective publications are shown. RB: Rainbow, LSET: Lucky Strike – Eiffeltower, LSMS: Lucky Strike – Montsegur, MG: Menes Gwen – White Flames, SEM: Semenov, LG: Logatchev – Quest, WA: Wide Awake, CL: Clueless, LI: Lilliput, MC853: site MC853, CH: Chapopote, DC673: site DC673, MK: Mictlan, Knoll 2201, TY: Tsanyao Yang, Knoll 2223, CS: Crab Spa, OMZ: Peru upwelling oxygen minimum zone.

Cruise	Ecosyste	Host species	Site		Latitude	Longitude	pth [m]	Year	DNA Librar	DNA extraction	Seq	Ref
BioBaz 2013	Vent	*B. azoricus*	RB	MAR	36.22943	-33.90196	2273	2013	1600F	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
BioBaz 2013	Vent	*B. azoricus*	RB	MAR	36.22943	-33.90196	2273	2013	1600G	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
BioBaz 2013	Vent	*B. azoricus*	RB	MAR	36.22943	-33.90196	2273	2013	1600H	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
BioBaz 2013	Vent	*B. azoricus*	RB	MAR	36.22943	-33.90196	2273	2013	1600I	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
BioBaz 2013	Vent	*B. azoricus*	RB	MAR	36.22943	-33.90196	2273	2013	1600J	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
BioBaz 2013	Vent	*B. azoricus*	LSET	MAR	37.28912	-32.27549	1690	2013	1586B	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
BioBaz 2013	Vent	*B. azoricus*	LSET	MAR	37.28912	-32.27549	1690	2013	1586C	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
BioBaz 2013	Vent	*B. azoricus*	LSET	MAR	37.28912	-32.27549	1690	2013	1586D	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
BioBaz 2013	Vent	*B. azoricus*	LSET	MAR	37.28912	-32.27549	1690	2013	1586E	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
BioBaz 2013	Vent	*B. azoricus*	LSET	MAR	37.28292	-32.27561	1689	2013	1586F	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
BioBaz 2013	Vent	*B. azoricus*	LSET	MAR	37.28292	-32.27561	1689	2013	1586G	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
BioBaz 2013	Vent	*B. azoricus*	LSET	MAR	37.28292	-32.27561	1689	2013	1586I	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
BioBaz 2013	Vent	*B. azoricus*	LSET	MAR	37.28292	-32.27561	1689	2013	1586J	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
BioBaz 2013	Vent	*B. azoricus*	LSMS	MAR	37.28806	-32.27557	1700	2013	1586K	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
BioBaz 2013	Vent	*B. azoricus*	LSMS	MAR	37.28806	-32.27557	1700	2013	1586N	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
BioBaz 2013	Vent	*B. azoricus*	LSMS	MAR	37.28806	-32.27557	1700	2013	1586O	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
BioBaz 2013	Vent	*B. azoricus*	MG	MAR	37.84445	-31.5189	836	2010	1586P	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
BioBaz 2013	Vent	*B. azoricus*	MG	MAR	37.84445	-31.5189	836	2010	1586Q	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
BioBaz 2013	Vent	*B. azoricus*	MG	MAR	37.84445	-31.5189	836	2010	1586R	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
BioBaz 2013	Vent	*B. azoricus*	MG	MAR	37.84445	-31.5189	836	2010	1586S	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Sayavedra, 2016
ODEMAR	Vent	*B. puteoserpentis*	SEM	MAR	13.51347	-44.963133	2432	2014	1115A	DNeasy blood & tissue	2x250bp	Sayavedra, 2016
ODEMAR	Vent	*B. puteoserpentis*	SEM	MAR	13.51347	-44.963133	2432	2014	1115B	DNeasy blood & tissue	2x250bp	Sayavedra, 2016
ODEMAR	Vent	*B. puteoserpentis*	SEM	MAR	13.51347	-44.963133	2432	2014	1115C	DNeasy blood & tissue	2x250bp	Sayavedra, 2016
M126	Vent	*B. puteoserpentis*	LG	MAR	14.75325	-44.98014	3047.3	2016	2487A	DNeasy blood & tissue or Zhou et al., 1996	2x150bp	Sayavedra, 2016
M126	Vent	*B. puteoserpentis*	LG	MAR	14.75325	-44.98014	3047.3	2016	2487B	DNeasy blood & tissue or Zhou et al., 1996	2x150bp	Sayavedra, 2016
M126	Vent	*B. puteoserpentis*	LG	MAR	14.75325	-44.98014	3047.3	2016	2487C	DNeasy blood & tissue or Zhou et al., 1996	2x150bp	Sayavedra, 2016
M126	Vent	*B. puteoserpentis*	SEM	MAR	13.51353	-44.963047	2446.6	2016	2487D	DNeasy blood & tissue or Zhou et al., 1996	2x150bp	Sayavedra, 2016
M126	Vent	*B. puteoserpentis*	SEM	MAR	13.51353	-44.963047	2446.6	2016	2487E	DNeasy blood & tissue or Zhou et al., 1996	2x150bp	Sayavedra, 2016
M126	Vent	*B. puteoserpentis*	SEM	MAR	13.51353	-44.963047	2446.6	2016	2487F	DNeasy blood & tissue or Zhou et al., 1996	2x150bp	Sayavedra, 2016
ATA57	Vent	*B.* sp.	WA	MAR	-4.81043	-12.37237	2989	2008	1501A, 159	Zhou et al., 1996	2x150bp	Ansorge et al., 2019
M78-2	Vent	*B.* sp.	CL	MAR	-4.80327	-12.3718	2971	2009	C11-2	DNeasy blood & tissue	2x150bp	Ansorge et al., 2019
M78-2	Vent	*B.* sp.	CL	MAR	-4.80327	-12.3718	2971	2009	C11-3	DNeasy blood & tissue	2x150bp	Ansorge et al., 2019
M78-2	Vent	*B.* sp.	CL	MAR	-4.80327	-12.3718	2971	2009	C11-4	DNeasy blood & tissue	2x150bp	Ansorge et al., 2019
M78-2	Vent	*B.* sp.	CL	MAR	-4.80327	-12.3718	2971	2009	C11-5	DNeasy blood & tissue	2x150bp	Ansorge et al., 2019
M78-2	Vent	*B.* sp.	CL	MAR	-4.80327	-12.3718	2971	2009	C11-6	DNeasy blood & tissue	2x150bp	Ansorge et al., 2019
M78-2	Vent	*B.* sp.	LI	MAR	-9.54657	-13.209733	1490	2009	L10-2	DNeasy blood & tissue	2x150bp	Ansorge et al., 2019
M78-2	Vent	*B.* sp.	LI	MAR	-9.54657	-13.209733	1490	2009	L10-4	DNeasy blood & tissue	2x150bp	Ansorge et al., 2019
M78-2	Vent	*B.* sp.	LI	MAR	-9.54657	-13.209733	1490	2009	L10-5	DNeasy blood & tissue	2x150bp	Ansorge et al., 2019
M78-2	Vent	*B.* sp.	LI	MAR	-9.54657	-13.209733	1490	2009	L5-1	DNeasy blood & tissue	2x150bp	Ansorge et al., 2019
M78-2	Vent	*B.* sp.	LI	MAR	-9.54657	-13.209733	1490	2009	L5-4	DNeasy blood & tissue	2x150bp	Ansorge et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1237	-89.14040	1073	2015	1712B	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1237	-89.14040	1073	2015	1712C	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1237	-89.14040	1073	2015	1712D	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1237	-89.14040	1073	2015	1712E	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1237	-89.14040	1073	2015	1712F	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1241	-89.14010	1073	2015	1712G	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1241	-89.14010	1073	2015	1712H	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1241	-89.14010	1073	2015	1712I	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1241	-89.14010	1073	2015	1712J	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1241	-89.14010	1073	2015	1712K	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1247	-89.14110	1078	2015	1712L	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1247	-89.14110	1078	2015	1712M	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1237	-89.14040	1073	2015	1712N	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1237	-89.14040	1073	2015	1712O	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1237	-89.14040	1073	2015	1712P	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1237	-89.14040	1073	2015	1712Q	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1237	-89.14040	1073	2015	1712R	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1247	-89.14110	1073	2015	1712S	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1247	-89.14110	1078	2015	1712T	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1247	-89.14110	1078	2015	1712U	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1237	-89.14040	1073	2015	1712V	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1247	-89.14110	1078	2015	1712W	Zhou et al., 1996	2x250bp	Picazo et al., 2019
NA58	Seep	*B. brooksi*	MC853	GoM	28.1241	-89.14010	1073	2015	1712X	Zhou et al., 1996	2x250bp	Picazo et al., 2019
M114-2	Seep	*B. heckerae*	CH	GoM	21.90005	-93.4354	2925	2015	1600D	DNeasy blood & tissue	2x100 bp	Rubin-Blum et al., 2017
M114-2	Seep	*B. heckerae*	CH	GoM	21.90005	-93.4354	2925	2015	2769A	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Rubin-Blum et al., 2017
M114-2	Seep	*B. heckerae*	CH	GoM	21.90005	-93.4354	2925	2015	1600D	DNeasy blood & tissue	2x100 bp	Rubin-Blum et al., 2017
M114-2	Seep	*B. heckerae*	CH	GoM	21.90005	-93.4354	2925	2015	1600O	DNeasy blood & tissue	2x100 bp	Rubin-Blum et al., 2017
M114-2	Seep	*B. heckerae*	CH	GoM	21.90005	-93.4354	2925	2015	1600P	DNeasy blood & tissue	2x100 bp	Rubin-Blum et al., 2017
M114-2	Seep	*B. heckerae*	CH	GoM	21.90005	-93.4354	2925	2015	2769A	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Rubin-Blum et al., 2017
M114-2	Seep	*B. heckerae*	CH	GoM	21.90005	-93.4354	2925	2015	2769B	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Rubin-Blum et al., 2017
M114-2	Seep	*B. heckerae*	CH	GoM	21.90005	-93.4354	2925	2015	2769C	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Rubin-Blum et al., 2017
M114-2	Seep	*B. heckerae*	CH	GoM	21.90005	-93.4354	2925	2015	2769D	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Rubin-Blum et al., 2017
M114-2	Seep	*B. heckerae*	CH	GoM	21.90005	-93.4354	2925	2015	2769E	DNeasy blood & tissue or Zhou et al., 1996	2x100 bp	Rubin-Blum et al., 2017
NA043	Seep	*B.* sp. (Cryptic)	DC673	GoM	28.30933	-87.31085	2604	2014	1875A	DNeasy blood & tissue	2x150bp	Sayavedra et al, in prep
NA043	Seep	*B.* sp. (Cryptic)	DC673	GoM	28.30933	-87.31085	2604	2014	1875D	DNeasy blood & tissue	2x150bp	Sayavedra et al, in prep
M114-2	Seep	*B. brooksi*	CH	GoM	21.90002	-93.43525	2925	2015	1789A	DNeasy blood & tissue	2x150bp	Rubin-Blum et al., 2017
M114-2	Seep	*B. brooksi*	CH	GoM	21.90002	-93.43525	2925	2015	1789B	DNeasy blood & tissue	2x150bp	Rubin-Blum et al., 2017
M114-2	Seep	*B. brooksi*	CH	GoM	21.90002	-93.43525	2925	2015	1789C	DNeasy blood & tissue	2x150bp	Rubin-Blum et al., 2017

M114-2	**Seep**	***B. brooksi***	**CH**	**GoM**	**21.90002**	**-93.43525**	**2925**	**2015**	**3411K**	**AllPrep DNA/RNA Mini Kit**	**2x150bp**	**this study**
M114-2	**Seep**	***B. brooksi***	**CH**	**GoM**	**21.90002**	**-93.43525**	**2925**	**2015**	**3411L**	**AllPrep DNA/RNA Mini Kit**	**2x150bp**	**this study**
M114-2	**Seep**	***B. brooksi***	**CH**	**GoM**	**21.90002**	**-93.43525**	**2925**	**2015**	**3411M**	**AllPrep DNA/RNA Mini Kit**	**2x150bp**	**this study**
M114-2	**Seep**	***B. brooksi***	**CH**	**GoM**	**21.90002**	**-93.43525**	**2925**	**2015**	**3411N**	**AllPrep DNA/RNA Mini Kit**	**2x150bp**	**this study**
M114-2	**Seep**	***B. brooksi***	**MK**	**GoM**	**22.02238**	**-93.24648**	**3106**	**2015**	**3411A**	**AllPrep DNA/RNA Mini Kit**	**2x150bp**	**this study**
M114-2	**Seep**	***B. brooksi***	**MK**	**GoM**	**22.02238**	**-93.24648**	**3106**	**2015**	**3411B**	**AllPrep DNA/RNA Mini Kit**	**2x150bp**	**this study**
M114-2	**Seep**	***B. brooksi***	**MK**	**GoM**	**22.02238**	**-93.24648**	**3106**	**2015**	**3411C**	**AllPrep DNA/RNA Mini Kit**	**2x150bp**	**this study**
M114-2	**Seep**	***B. brooksi***	**MK**	**GoM**	**22.02238**	**-93.24648**	**3106**	**2015**	**3411D**	**AllPrep DNA/RNA Mini Kit**	**2x150bp**	**this study**
M114-2	**Seep**	***B. brooksi***	**MK**	**GoM**	**22.02238**	**-93.24648**	**3106**	**2015**	**3411E**	**AllPrep DNA/RNA Mini Kit**	**2x150bp**	**this study**
M114-2	**Seep**	***B. brooksi***	**TY**	**GoM**	**22.39300**	**-93.40512**	**3365**	**2015**	**3411F**	**AllPrep DNA/RNA Mini Kit**	**2x150bp**	**this study**
M114-2	**Seep**	***B. brooksi***	**TY**	**GoM**	**22.39300**	**-93.40512**	**3365**	**2015**	**3411G**	**AllPrep DNA/RNA Mini Kit**	**2x150bp**	**this study**
M114-2	**Seep**	***B. brooksi***	**TY**	**GoM**	**22.39300**	**-93.40512**	**3365**	**2015**	**3411H**	**AllPrep DNA/RNA Mini Kit**	**2x150bp**	**this study**
M114-2	**Seep**	***B. brooksi***	**TY**	**GoM**	**22.39300**	**-93.40512**	**3365**	**2015**	**3411I**	**AllPrep DNA/RNA Mini Kit**	**2x150bp**	**this study**
M114-2	**Seep**	***B. brooksi***	**TY**	**GoM**	**22.39300**	**-93.40512**	**3365**	**2015**	**3411J**	**AllPrep DNA/RNA Mini Kit**	**2x150bp**	**this study**
AT26-23	Vent	*B. thermophilus*	CS	EPR	9.939167	-105.6561	2512.0	2014	1600R	Zhou et al., 1996	2x150bp	Sayavedra et al, in prep
AT26-23	Vent	*B. thermophilus*	CS	EPR	9.939167	-105.6561	2512.0	2014	1600S	Zhou et al., 1996	2x150bp	Sayavedra et al, in prep
AT26-23	Vent	*B. thermophilus*	CS	EPR	9.939167	-105.6561	2512.0	2014	1600T	Zhou et al., 1996	2x150bp	Sayavedra et al, in prep
M93	pelagic	-	OMZ	Peru	-12.23	-77.18	-	2013	C2484	AllPrep DNA/RNA Mini Kit	-	Callbeck et al., 2018
M93	pelagic	-	OMZ	Peru	-12.23	-77.18	-	2013	C2488	AllPrep DNA/RNA Mini Kit	-	Callbeck et al., 2018
M114-2	Seep	Poecilosclerid bra. sp	MK	GoM	22.02238	-93.24648	3106	2015	1600M	AllPrep DNA/RNA Mini Kit	2x150bp	Rubin-Blum et al., 2017
M114-2	Seep	Poecilosclerid encr. s	CH	GoM	21.90002	-93.43525	2925	2015	1600L	AllPrep DNA/RNA Mini Kit	2x150bp	Rubin-Blum et al., 2017
M114-2	Seep	Poecilosclerid encr. s	CH	GoM	21.90002	-93.43525	2925	2015	1600N	AllPrep DNA/RNA Mini Kit	2x150bp	Rubin-Blum et al., 2017

Table S3 | Completeness, contamination and strain heterogeneity (of the contamination) for SUP05 draft genomes used in this study. Samples marked in grey were assembled and binned in this study. Other draft genomes were obtained from previous studies as listed in **Tab. S2.**

Bin	Completeness [%]	Contam. [%]	Strain het. [%]
1115A	94.53	3.09	75
1115B	93.41	3.09	85.71
1115C	94.53	0.56	100
1586B	93.41	0.84	100
1586C	92.85	0.84	50
1586D	93.97	0	0
1586E	92.85	0	0
1586F	92.85	1.4	33.33
1586G	93.97	0	0
1586I	94.53	0	0
1586J	93.97	0.56	0
1586K	92.28	0.56	0
1586N	94.53	0	0
1586O	93.6	1.12	50
1586P	91.72	0	0
1586Q	92.28	0	0
1586R	93.97	0	0
1586S	93.97	0	0
1600D	92.66	4.31	63.64
1600D.sox2	93.97	0	0
1600F	93.97	0	0
1600G	93.97	0	0
1600H	93.93	0	0
1600I	91.72	0	0
1600J	93.97	0.14	100
1600O	91.82	1.69	0
1600P	90.41	1.12	50
1600R	94.25	0.56	100
1600S	93.97	0.56	100
1600T	94.25	0.84	100
1712B	94.25	0.56	0
1712C	93.55	7.07	80
1712D	93.97	6.41	82.35
1712E	94.25	1.69	75
1712F	94.81	0	0
1712G	94.25	0.56	0
1712H	94.81	0.56	0
1712J	93.69	0	0
1712K	94.25	0.56	0
1712L	94.25	0.56	0
1712M	94.25	0.56	0
1712N	94.81	1.12	75
1712O	93.69	0.56	0
1712P	93.41	3.09	87.5
1712Q	94.25	0.56	0
1712R	94.25	1.52	100
1712S	93.27	7.4	95.65
1712T	94.25	0.56	0
1712U	93.69	0.56	0
1712V	94.25	0.56	0
1712W	93.97	5.99	95.24
1712X	94.25	0.56	0
1789A	93.27	0.56	0
1789B	93.83	0.6	0
1789C	93.83	0.56	0
1875A	93.97	1.69	66.67
1875Dcryp.sox2	87.13	1.12	0
2487A	94.16	1.4	57.14
2487B	94.72	1.4	57.14
2487C	94.72	1.4	83.33
2487D	94.53	1.12	66.67
2487E	94.72	1.12	66.67
2487F	94.53	1.97	83.33
2769A	93.5	0.84	50
2769A.sox2	93.13	0.41	100
2769B	92.47	2.53	40
2769C	93.22	3.65	80
2769D	91.96	4.12	91.67
2769E	90.13	1.69	33.33
3411A	**94.25**	**1.12**	**0**
3411B	**93.69**	**1.12**	**0**
3411C	**94.25**	**0.56**	**100**
3411D	**93.13**	**2.53**	**55.56**
3411E	**94.81**	**0.56**	**0**
3411F	**94.81**	**0.56**	**0**
3411G	**94.81**	**0**	**0**
3411H	**94.81**	**0.56**	**0**
3411I	**94.25**	**0**	**0**
3411J	**94.81**	**0.56**	**0**
3411K	**94.53**	**1.9**	**40**
3411L	**94.25**	**0.56**	**100**
3411M	**94.53**	**0.84**	**50**
3411N	**89.61**	**0.56**	**0**
3411O	**97.47**	**0.56**	**0**
Bsep	94.83	0.56	100
C112	94.53	1.31	75
C113	94.53	1.4	66.67
C114	94.53	1.4	66.67
C115	94.25	1.4	100
C116	93.69	1.03	100
L102	93.33	1.12	100
L104	94.64	0.56	100
L105	92.96	0.56	100
L51	94.08	1.69	100
L54	94.36	0.84	100
turt	94.53	2.81	87.5
sponge_1600L	91.5	12.02	81.82
sponge_1600M	91.27	16.57	0
sponge_1600N	91.46	15.16	73.68
C2484_SUP05_PeruOMZ	92.28	3.14	63.64
C2488_SUP05_PeruOMZ	95.09	3.37	66.67
CaRuthiaMagnificaCM	86.67	0	0
CaRuthiaMagnificaUCDCM	87.23	0	0
CaThioglobusAutotrophicusEF1	94.64	0	0
CaThioglobusPerditusCMC	95.09	3.37	66.67
CaThioglobusSingularisGG2	91.86	0	0
CaThioglobusSingularisNP1	93.88	0	0
CaThioglobusSingularisPS1	94.36	0	0
CaVesicomyosociousOkutanii	85.69	0	0

Table S4 | Enrichment of COG categories in SUP05 draft genomes. Colors mark the corresponding lifestyle: green: horizontally transmitted intracellular symbiont (*Bathymodiolus*-associated), yellow: sponge symbiont of unknown transmission mode, blue: free-living.

ID	COG	Description	P-value	Count
1875D_SOX2	COG0726_G	Peptidoglycan/xylan/chitin deacetylase, PgdA/CDA1 family	9.51222E-08	12
1875D_SOX2	COG1401_V	5-methylcytosine-specific restriction endonuclease McrBC, GTP-binding regulatory subunit McrB	6.56747E-05	6
1875D_SOX2	COG2801_L	Transposase InsO and inactivated derivatives	0.00010272	19
1875D_SOX2	COG2826_L	Transposase and inactivated derivatives, IS30 family	5.89854E-07	19
CaThioglobusAutotrophicusEF1	COG0859_M	ADP-heptose:LPS heptosyltransferase	7.18684E-06	11
CaThioglobusSingularisGG2	COG0010_E	Arginase family enzyme	0.00020666	4
CaThioglobusSingularisGG2	COG0318_IQ	Acyl-CoA synthetase (AMP-forming)/AMP-acid ligase II	0.00063849	6
CaThioglobusSingularisGG2	COG0395_G	ABC-type glycerol-3-phosphate transport system, permease component	4.37965E-05	6
CaThioglobusSingularisGG2	COG0518_F	GMP synthase - Glutamine amidotransferase domain	0.00065022	4
CaThioglobusSingularisGG2	COG0665_E	Glycine/D-amino acid oxidase (deaminating)	5.09949E-09	16
CaThioglobusSingularisGG2	COG0673_R	Predicted dehydrogenase	4.37965E-05	6
CaThioglobusSingularisGG2	COG0687_E	Spermidine/putrescine-binding periplasmic protein	0.00029864	5
CaThioglobusSingularisGG2	COG0697_GER	Permease of the drug/metabolite transporter (DMT) superfamily	1.47929E-06	16
CaThioglobusSingularisGG2	COG1012_C	Acyl-CoA reductase or other NAD-dependent aldehyde dehydrogenase	5.86116E-06	11
CaThioglobusSingularisGG2	COG1175_G	ABC-type sugar transport system, permease component	5.8549E-05	6
CaThioglobusSingularisGG2	COG1176_E	ABC-type spermidine/putrescine transport system, permease component I	0.00088672	4
CaThioglobusSingularisGG2	COG1177_E	ABC-type spermidine/putrescine transport system, permease component II	0.00088672	4
CaThioglobusSingularisGG2	COG1522_K	DNA-binding transcriptional regulator, Lrp family	0.00075397	6
CaThioglobusSingularisGG2	COG1653_G	ABC-type glycerol-3-phosphate transport system, periplasmic component	0.00016786	5
CaThioglobusSingularisGG2	COG2303_E	Choline dehydrogenase or related flavoprotein	8.60467E-05	5
CaThioglobusSingularisGG2	COG3839_G	ABC-type sugar transport system, ATPase component	4.37965E-05	6
CaThioglobusSingularisGG2	COG4175_E	ABC-type proline/glycine betaine transport system, ATPase component	0.00046214	4
CaThioglobusSingularisGG2	COG4176_E	ABC-type proline/glycine betaine transport system, permease component	0.00046214	4
CaThioglobusSingularisGG2	COG4638_PR	Phenylpropionate dioxygenase or related ring-hydroxylating dioxygenase, large terminal subunit	0.00065022	4
CaThioglobusSingularisGG2	COG5598_H	Trimethylamine:corrinoid methyltransferase	0.00012178	5
CaThioglobusSingularisNP1	COG0395_G	ABC-type glycerol-3-phosphate transport system, permease component	0.00046603	5
CaThioglobusSingularisNP1	COG0665_E	Glycine/D-amino acid oxidase (deaminating)	4.23656E-09	16
CaThioglobusSingularisNP1	COG0673_R	Predicted dehydrogenase	7.88277E-09	9
CaThioglobusSingularisNP1	COG0697_GER	Permease of the drug/metabolite transporter (DMT) superfamily	6.00288E-06	15
CaThioglobusSingularisNP1	COG1012_C	Acyl-CoA reductase or other NAD-dependent aldehyde dehydrogenase	0.00018705	9
CaThioglobusSingularisNP1	COG1175_G	ABC-type sugar transport system, permease component	0.00058709	5
CaThioglobusSingularisNP1	COG1744_R	Basic membrane lipoprotein Med, periplasmic binding protein (PBP1-ABC) superfamily	0.00019647	4
CaThioglobusSingularisNP1	COG2303_E	Choline dehydrogenase or related flavoprotein	8.086E-05	5
CaThioglobusSingularisNP1	COG2849_S	Antitoxin component YwqK of the YwqJK toxin-antitoxin module	3.88906E-05	7
CaThioglobusSingularisNP1	COG3839_G	ABC-type sugar transport system, ATPase component	0.00046603	5
CaThioglobusSingularisNP1	COG4638_PR	Phenylpropionate dioxygenase or related ring-hydroxylating dioxygenase, large terminal subunit	3.65823E-05	5
CaThioglobusSingularisPS1	COG0010_E	Arginase family enzyme	0.00021609	4
CaThioglobusSingularisPS1	COG0395_G	ABC-type glycerol-3-phosphate transport system, permease component	4.6728E-05	6
CaThioglobusSingularisPS1	COG0665_E	Glycine/D-amino acid oxidase (deaminating)	4.46245E-08	15
CaThioglobusSingularisPS1	COG0673_R	Predicted dehydrogenase	4.6728E-05	6
CaThioglobusSingularisPS1	COG0687_E	Spermidine/putrescine-binding periplasmic protein	0.00031509	5
CaThioglobusSingularisPS1	COG0697_GER	Permease of the drug/metabolite transporter (DMT) superfamily	1.71813E-06	16
CaThioglobusSingularisPS1	COG1012_C	Acyl-CoA reductase or other NAD-dependent aldehyde dehydrogenase	6.5401E-06	11
CaThioglobusSingularisPS1	COG1175_G	ABC-type sugar transport system, permease component	6.24472E-05	6
CaThioglobusSingularisPS1	COG1176_E	ABC-type spermidine/putrescine transport system, permease component I	0.00092605	4
CaThioglobusSingularisPS1	COG1177_E	ABC-type spermidine/putrescine transport system, permease component II	0.00092605	4
CaThioglobusSingularisPS1	COG1522_K	DNA-binding transcriptional regulator, Lrp family	0.000801	6
CaThioglobusSingularisPS1	COG1653_G	ABC-type glycerol-3-phosphate transport system, periplasmic component	0.00017722	5
CaThioglobusSingularisPS1	COG2303_E	Choline dehydrogenase or related flavoprotein	9.09028E-05	5
CaThioglobusSingularisPS1	COG3839_G	ABC-type sugar transport system, ATPase component	4.6728E-05	6
CaThioglobusSingularisPS1	COG4175_E	ABC-type proline/glycine betaine transport system, ATPase component	0.00048293	4
CaThioglobusSingularisPS1	COG4176_E	ABC-type proline/glycine betaine transport system, permease component	0.00048293	4
CaThioglobusSingularisPS1	COG4638_PR	Phenylpropionate dioxygenase or related ring-hydroxylating dioxygenase, large terminal subunit	0.00067927	4
CaThioglobusSingularisPS1	COG5598_H	Trimethylamine:corrinoid methyltransferase	0.00012861	5
Bsep	COG3328_L	Transposase (or an inactivated derivative)	1.87529E-06	12
site_rep.bbro_mc853	COG0732_V	Restriction endonuclease S subunit	1.4578E-05	14
site_rep.bbro_mc853	COG2931_Q	Ca2+-binding protein, RTX toxin-related	2.82367E-06	35
site_rep.bbro_MK	COG0286_V	Type I restriction-modification system, DNA methylase subunit	4.14333E-05	14
site_rep.bheck_CH	COG1192_D	Cellulose biosynthesis protein BcsQ	4.63965E-05	10
site_rep.bheck_CH	COG3436_L	Transposase	5.39614E-07	7
site_rep.bheck_CH	COG3609_K	Transcriptional regulator, contains Arc/MetJ-type RHH (ribbon-helix-helix) DNA-binding domain	2.48667E-06	6
site_rep.bthe_CS	COG2931_Q	Ca2+-binding protein, RTX toxin-related	1.16792E-10	43
site_rep.nmar_MG	COG3385_L	IS4 transposase	3.14078E-05	5
sponge1600L	COG0340_H	Biotin-(acetyl-CoA carboxylase) ligase	6.37414E-07	10
sponge1600L	COG0666_R	Ankyrin repeat	8.8776E-16	12
sponge1600L	COG0790_R	TPR repeat	8.703E-07	25
sponge1600M	COG2026_JD	mRNA-degrading endonuclease RelE, toxin component of the RelBE toxin-antitoxin system	8.3135E-06	9
sponge1600N	COG0845_M	Multidrug efflux pump subunit AcrA (membrane-fusion protein)	4.93844E-05	17
sponge1600N	COG1598_S	Predicted nuclease of the RNAse H fold, HicB family	2.96466E-05	5
sponge1600N	COG2026_JD	mRNA-degrading endonuclease RelE, toxin component of the RelBE toxin-antitoxin system	2.97915E-09	13
sponge1600N	COG2826_L	Transposase and inactivated derivatives, IS30 family	0.0001276	20
sponge1600N	COG3666_L	Transposase	2.54479E-07	11
sponge1600N	COG3668_R	Plasmid stabilization system protein ParE	0.00017389	7
sup05C2484	COG0458_EF	Carbamoylphosphate synthase large subunit	4.72633E-07	10
Wida	COG0494_LR	8-oxo-dGTP pyrophosphatase MutT and related house-cleaning NTP pyrophosphohydrolases, NUDIX family	4.69197E-14	26
Wida	COG0823_U	Periplasmic component of the Tol biopolymer transport system	3.2909E-08	13
Wida	COG2801_L	Transposase InsO and inactivated derivatives	4.53736E-08	33
Wida	COG2931_Q	Ca2+-binding protein, RTX toxin-related	2.18318E-15	63
Wida	COG3039_L	Transposase and inactivated derivatives, IS5 family	1.74401E-24	47
Wida	COG3209_M	Uncharacterized conserved protein RhaS, contains 28 RHS repeats	9.28341E-58	93
Wida	COG3210_U	Large exoprotein involved in heme utilization or adhesion	0.00026943	14
Wida	COG3335_L	Transposase	7.15873E-18	25
Wida	COG3415_L	Transposase	2.31739E-07	9
Wida	COG3676_L	Transposase and inactivated derivatives	1.32832E-09	18
Wida	COG4637_R	Predicted ATPase	0.00055587	10
Wida	COG5492_N	Uncharacterized conserved protein YjdB, contains Ig-like domain	1.22207E-05	8

Table S5 | Enrichment of KEGG categories in SUP05 draft genomes. Colors mark the corresponding lifestyle: green: horizontally transmitted intracellular symbiont (*Bathymodiolus*-associated), yellow: sponge symbiont of unknown transmission mode, blue: free-living.

ID	KEGG	Description	P-value	Count
1875A_SUP05SOX	K13735	adhesin/invasin	4.410406163705E-10	10
1875D_SOX2	K07482	transposase, IS30 family	2.4462009100588E-07	19
1875D_SOX2	K07452	mcrB; 5-methylcytosine-specific restriction enzyme B	5.4143449355543E-05	6
1875D_SOX2	K07481	transposase, IS5 family	7.8001290333206E-05	10
CaThioglobusPerditusCMC	K02841	waaC, rfaC; heptosyltransferase I	4.3482080073431E-05	7
CaThioglobusSingularisGG2	K00108	betA, CHDH; choline dehydrogenase	0.000113331187464	5
CaThioglobusSingularisGG2	K00315	DMGDH; dimethylglycine dehydrogenase	0.000113331187464	5
CaThioglobusSingularisNP1	K02057	ABC.SS.P; simple sugar transport system permease protein	3.670481626657E-06	6
CaThioglobusSingularisNP1	K00108	betA, CHDH; choline dehydrogenase	0.000102847861325	5
CaThioglobusSingularisNP1	K00315	DMGDH; dimethylglycine dehydrogenase	0.000102847861325	5
CaThioglobusSingularisPS1	K00108	betA, CHDH; choline dehydrogenase	0.000118873374627	5
CaThioglobusSingularisPS1	K00315	DMGDH; dimethylglycine dehydrogenase	0.000118873374627	5
site_rep.bheck_CH	K03496	parA, soj; chromosome partitioning protein	3.0441871056657E-05	10
sponge1600L	K03524	birA; BirA family transcriptional regulator, biotin operon repressor / biotin---[acetyl-CoA-carboxylase] lig	4.0865042334278E-07	10
sponge1600L	K07126	uncharacterized protein	1.124585224546E-06	24
sup05C2484	K01955	carB, CPA2; carbamoyl-phosphate synthase large subunit	5.5806963389653E-07	10
sup05C2488	K02841	waaC, rfaC; heptosyltransferase I	4.3738672276035E-05	7
Wida	K07494	putative transposase	2.2535574867654E-24	25
Wida	K08312	nudE; ADP-ribose diphosphatase	9.8342899641883E-21	25
Wida	K07488	transposase	1.0195523971613E-10	18
Wida	K03641	tolB; TolB protein	8.6072330249283E-09	13
Wida	K07481	transposase, IS5 family	1.6119260593315E-05	13
Wida	K07497	putative transposase	3.5901368924705E-05	18
Wida	K07482	transposase, IS30 family	0.000127669316091	18

Chapter IV | Symbiont evolution

Evolutionary signatures in genomes of horizontally transmitted endosymbionts of *Bathymodiolus* mussels

Rebecca Ansorge[1], Stefano Romano[2], Nicole Dubilier[1], Jillian Petersen[2]*

*Author order is not fixed

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

[2]Division for Microbiology and Ecosystem Science, University of Vienna, Austria

The manuscript is in preparation and has not been revised by all authors.

Author contributions

RA conceived the study, analyzed the data and wrote the manuscript, SR contributed to the conceptual design and interpretation of the data and revised the manuscript, JR and ND helped to conceive the study, JR revised the manuscript.

Abstract

The studies that shape our understanding of genome evolution in endosymbionts have long been biased towards vertically transmitted, obligate endosymbionts. In contrast, the driving forces shaping genome evolution in horizontally-transmitted endosymbionts are not well understood. In our study we shed light on the impact of nucleotide and amino acid variation and natural selection on horizontally transmitted endosymbionts of deep-sea mussels in their natural context. Investigation of seven symbiont species from 15 geographic locations revealed that amino acid variation among co-existing strains and positive selection in symbiont traits appeared to be driven by host-symbiont interaction and environmental conditions. Previous analyses have only focused on gene content variation, but our findings reveal that divergent evolution among symbiont populations also affects core functions, such as sulfur oxidation. This trait is encoded in the entire genus and could therefore be considered conserved. The sensitivity of our analysis allowed us to detect loci under positive selection possibly reflecting adaptation of the symbiont to local environmental conditions. Heterogeneity and allele frequencies further hold signatures that allow us to understand the extent of symbiont exchange among hosts. Population genomic analyses on these signatures suggested that symbiont exchange among host individuals may be less prominent in host species *Bathymodiolus brooksi* than in other host species – a characteristic that appears to be influenced by multiple factors such as host age and density. Our study shows that high-resolution genomic analyses aid our understanding of the intimate association of *Bathymodiolus* hosts with specific symbiont strains in their natural context, which is essential for deciphering the signatures of genome evolution in *Bathymodiolus* endosymbionts.

Introduction

Genome evolution in symbiotic bacteria is massively influenced by the interaction with their host and the mode of symbiont transmission between host generations[1,2]. Endosymbionts have mostly been described to be shaped by genome reduction, genetic drift and non-adaptive processes[2,3]. However, these observations and theories usually refer to obligate, vertically transmitted symbionts, while the evolutionary drivers in horizontally transmitted endosymbionts are not well understood. The concept of effective population size N_e, the size of a theoretical population that would explain the observed genetic variation in the investigated population, is considered a key factor influencing the strength of selection and genetic drift in bacterial species[4,5]. Most vertically transmitted symbionts undergo a population bottleneck with each host generation, which decreases N_e and therefore shifts the balance away from purifying selection and towards genetic drift[3]. In addition, isolation from the external environment of vertically transmitted endosymbionts restricts genetic exchange with other bacteria through horizontal gene transfer. These factors, increasing the impact of genetic drift, also reduce the strength of purifying selection. Consequently, slightly deleterious mutations can become fixed in bacterial populations under genetic drift[4]. The obligate lifestyle within animal hosts and the relaxed purifying selection due to reduced N_e often lead to reductive genome evolution in vertically transmitted endosymbionts. For example, in some vertically transmitted insect symbionts these phenomena have led to extreme genome reduction that resulted in the loss of symbiont function. This so-called ‘evolutionary rabbit hole’ can be detrimental to a host that depends on its symbiont[6]. Replacement of the primary ‘defective’ symbiont with secondary symbionts that fulfill these lost functions may represent an escape route for the host from that rabbit hole[7–9]. Instead,

for the symbiont with its reduced genome, there seems to be no escape from that obligate association. Intriguingly, genome reduction has also been described for some free-living bacteria, such as *Prochlorococcus*[10]. In contrast to vertically transmitted endosymbionts, however, this process is considered adaptive and favored by the selection for smaller and 'streamlined' genomes. *Prochlorococcus*, as most free-living bacteria, is characterized by large N_e, which implies strong effective selection[4]. The direction of selection can be purifying, eliminating deleterious mutations from a population; or adaptive which can lead to the fixation of a mutation in the population. Selection can also be disruptive between different populations of a bacterial species, leading to the diversification in particular genes that are subject to different selection pressures (e.g. imposed by environmental conditions). Such diversifying selection among bacterial populations can be detected by significant frequency shifts of the alleles[11,12].

Besides vertically transmitted endosymbionts, with typically small N_e, and free-living bacteria with potentially enormous N_e, there are horizontally transmitted endosymbionts. These endosymbionts are acquired by each host generation from the environment, imposing a challenge for host and symbiont in maintaining specific recognition[1]. The mechanisms that ensure successful association with each host generation and the time window during which the host is permissive to symbiont colonization are highly diverse and which reflects the impact on symbiont genome evolution. Horizontally transmitted symbionts are thought to be subject to evolutionary processes similar to free-living bacteria, which means their genomes should be similar in size to their free-living relatives and undergo frequent recombination while deleterious mutations are selected against[1]. Yet, these factors are strongly influenced by N_e which can vary greatly in horizontally transmitted

endosymbionts. For example, horizontally transmitted endosymbionts can undergo a strong population bottleneck with each host generation similar to vertically transmitted symbionts (e.g. *Endoriftia*[13], *Vibrio fischerii*[14]), whereas other endosymbionts continue to colonize their host over a longer period, avoiding population bottlenecks. The latter results in a larger N_e and has been hypothesized for symbionts of *Bathymodiolus* deep-sea mussels[15]. A key characteristic of horizontally transmitted endosymbionts is that they have to retain the ability to survive both inside and outside their hosts, occupying two vastly different niches. Furthermore, the physiological state of the free-living population, active or 'dormant', will strongly influence the frequency of genetic exchange with microbes from the environment, and thus genome evolution. For most symbioses, N_e and details in symbiont transmission are often unknown, which makes prediction of the evolutionary driving forces difficult. Investigation of genome signatures in horizontally transmitted endosymbiont populations can therefore help to unravel the impact of natural selection and genetic drift in these communities.

We investigated genomic signatures of selection in 7 different species of horizontally transmitted endosymbionts associated with *Bathymodiolus* mussels. This group of bivalves dominates hydrothermal vents and cold seeps in the deep sea, forming dense mussel beds in these extreme habitats[16]. Their nutrition depends on intracellular sulfur-oxidizing (SOX) endosymbionts that use chemical energy for primary production in a process termed chemosynthesis[17]. This intimate association is thought to have evolved multiple times independently from distinct lineages of free-living sulfur-oxidizing bacteria[18]. The symbionts are horizontally transmitted, potentially frequently throughout the hosts lifetime[15,19]. Thus, despite being housed intracellularly in their hosts' gill epithelia, the symbionts come into contact with the

external environment and have access to genetic material from free-living bacteria. In contrast to vertically transmitted endosymbionts, *Bathymodiolus* endosymbionts likely do not undergo strong population bottlenecks and may therefore evolve under completely different selection regimes. Indeed, previous studies revealed that the core genome of SOX symbionts in *Bathymodiolus brooksi* evolves under purifying selection rather than the neutral processes typical for vertically-transmitted symbionts[20]. In addition, although *Bathymodiolus* symbionts have relatively small genome sizes, 1.4 to 2.8 mbp[15,21,22], similar genome sizes (1.5 to 1.7 mbp) have been described for their closest free-living relatives *Ca*. Thioglobus spp.[23,24] (see Chapter III). This indicates that the symbiotic lifestyle *per se* in *Bathymodiolus* symbionts did not lead to genome reduction. In our study we investigated environmental metagenomes of 9 *Bathymodiolus* species from 15 different sampling sites (Tab. 1). The SOX symbionts belonged to 7 different species as defined in Chapter III.

Variation that leads to functional differences among members within a population is the foundation for natural selection and evolution. The populations of *Bathymodiolus* SOX symbionts have been shown to be heterogeneous at the level of single nucleotides and their gene content[15,20,22]. However, it is unknown whether the degree of nucleotide variation is consistent across *Bathymodiolus* SOX symbiont species. Indeed, two previous studies revealed different results in symbiont heterogeneity and population structure that warrant explanation[15,20]. In this study we compared genome signatures such as nucleotide variation, functional implications of these polymorphisms and population structure in different symbiont species, host species and sampling sites. We further investigated signatures of diversifying genome evolution between symbiont populations within each species, providing a snapshot of the impact of selection in horizontally transmitted endosymbionts.

Methods

Sample acquisition and processing

Sample acquisition, DNA extraction and metagenomic processing have been described and summarized in the method section and Tab. S1, S2, S3 of Chapter III. All datasets included in this study are summarized in **Tab. 1**. We investigated 88 metagenomes belonging to seven symbiont species (after ANI cutoff > 95% as defined in Chapter III) and originating from 15 distinct sampling locations. Each metagenome originated from gill tissue of a single host individual.

Table 1 | Mussel metagenome samples included in this study.

Sampling site	Site type	# of individuals	Symbiont species	Host species	Sample library[#]
Menez Gwen (MG)	Vent	4	NMAR	*B. azoricus*	1586P,Q,R,S
Lucky Strike ET (LSET)	Vent	8	NMAR	*B. azoricus*	1586B,C,D,E,F,G,I,J
Lucky Strike MS (LSMS)	Vent	3	NMAR*	*B. azoricus*	1586K,N,O
Rainbow (RB)	Vent	5	NMAR	*B. azoricus*	1600F,G,H,I,J
Logatchev (LG)	Vent	3	NMAR	*B. puteoserpentis*	2487A,B,C
Semenov (SEM)	Vent	6	NMAR	*B. puteoserpentis*	2487D,E,F, 1115A,B,C
Clueless (CL)	Vent	5	SMAR	*B.* sp. 1	C112,C113,C114,C115,C116
Lilliput (LI)	Vent	5	SMAR*	*B.* sp. 2	L102,L104,L105,L51,L54
DC673	Seep	1	Bcr_sup05	*B.* sp. cryptic	1875A
MC853	Seep	21	Bbro	*B. brooksi*	1712B-H,J,K,L,N-X
Chapopote (CH)	Seep	8	Bbro*	*B. brooksi*	3411K-N,1789A-C
		5	Bh_sup05*	*B. heckerae*	2769A,B,E, 1600D,P
Tsanyao Yang (TY)	Seep	5	Bbro	*B. brooksi*	3411F-J
Mictlan Knoll (MK)	Seep	5	Bbro	*B. brooksi*	3411A-E
Crab Spa (CS)	Vent	3	Bthe*	*B. thermophilus*	1600R,S,T
Myonin Knoll (MyK)	Vent	1	Bsep	*B. septemdierum*	ikuta

*indicates site from which species reference genome for F_{ST} analysis originated
[#]Details of library treatment listed in Chapter III Tab. S2, S3

SNP calling and F_{ST} analysis

In order to determine the heterogeneity within symbiont populations of single hosts, we performed a SNP calling analyses. First, we trimmed all Illumina sequencing reads to a minimum quality of 20 using BBDuk (v 38.34, Bushnell B. -

sourceforge.net/projects/bbmap/) and mapped these with a minimum identity of 95% to the symbiont genome bin of each sample separately using BBMap (v 38.34, Bushnell B. - sourceforge.net/projects/bbmap/). Mapping files were all sampled to the same read depth of 100x and SNP calling on each library was performed with Genome Analysis Toolkit (v 3.8.0) (GATK)[25] and with parameters described previously[15] (Chapter II). SNPs were annotated with snpEff (v 4.3) [26] and nonsynonymous (N) and synonymous (S) counts per gene were retrieved from the output file.

For the F_{ST} analysis we mapped all quality trimmed reads to a reference genome per symbiont species with minimum identity of 95% with BBMap (v 38.34) and sampled all mapping files to the same read depth of 100x. Both was necessary to compare polymorphic loci between samples. Allele frequencies from all polymorphic site were retrieved from the variant call format (vcf) file obtained from SNP calling we performed with GATK (v 3.8.0)[25] according to the parameters described in previously[15]. Calculation of pairwise F_{ST} for each locus was performed according to custom scripts that were used in Ansorge et al.[15] (deposited in https://github.com/deropi/BathyBrooksiSymbionts) and are based on Schloissnig et al.[27]. Pairwise F_{ST} was plotted per site and between sites for each symbiont species using the ggplot2 (v 3.0.0) package in R (v.3.2.2)[28].

Outlier identification

We plotted the number of S, N and the N/S ratio per gene within each sample in a box plot using the ggplot2 (v 3.0.0) package in R (v 3.2.2)[28]. Since the data were not normally distributed and highly skewed we applied robust statistics with the

adjboxStats function in the robustbase (v 0.92-7) package in R (v 3.2.2)[28] to identify outliers for the number of S, N and N/S ratio per gene (**Fig. S1**).

To identify candidate loci under diversifying selection we retrieved F_{ST} outliers with BayeScan (v. 2.1)[29]. We used a setting of prior odds 100 and regarded all genes with positive values of alpha, indicative of diversifying selection, and a q-value of 0.05 as candidate loci under diversifying selection (**Fig. S2**). All genes that had one or more diversifying loci were regarded as potentially diversifying.

Orthologous clustering

To compare shared sets of genes that were identified as N, S, N/S or F_{ST} outliers we clustered orthologous genes according to sequence similarity of 0.5 with USEARCH as implemented in BPGA (v 1.3.0)[30]. Shared clusters of orthologous genes were visualized with the UpsetR (v 1.3.3) package in R (v 3.2.2). Each gene identified as outlier was furthermore assigned to Clusters of Orthologous Groups (COGs)[31] using BPGA (v 1.3.0) to assess to which functional categories the outliers belonged to.

Results

Symbiont nucleotide variation is predominantly nonsynonymous

Single nucleotide polymorphisms (SNPs) are powerful signatures to elucidate population dynamics and evolutionary processes in bacterial genomes. To understand whether different species of SOX associated with mussels of the genus *Bathymodiolus* show similar evolutionary signatures in their genomes, we analyzed 88 metagenomes from 7 different symbiont species (NMAR, SMAR, Bbro, Bthe,

Bh_sup05, Bcry_sup05 and Bsep; as determined in chapter III) from 15 different sampling locations (**Tab. 1**). We calculated SNPs per kbp (SNPs/kbp) for each symbiont population within single host individuals to determine how heterogeneous these populations are. This analysis revealed that a SOX symbiont population within single hosts can have a broad range of population nucleotide diversity between < 1 and 13 SNPs/kbp. Interestingly, two symbiont species Bbro and Bthe showed a broad range of heterogeneity values at each of the five sampling locations, whereas the other three species from 9 distinct sites showed a rather narrow range of heterogeneity per site (**Fig. 1**).

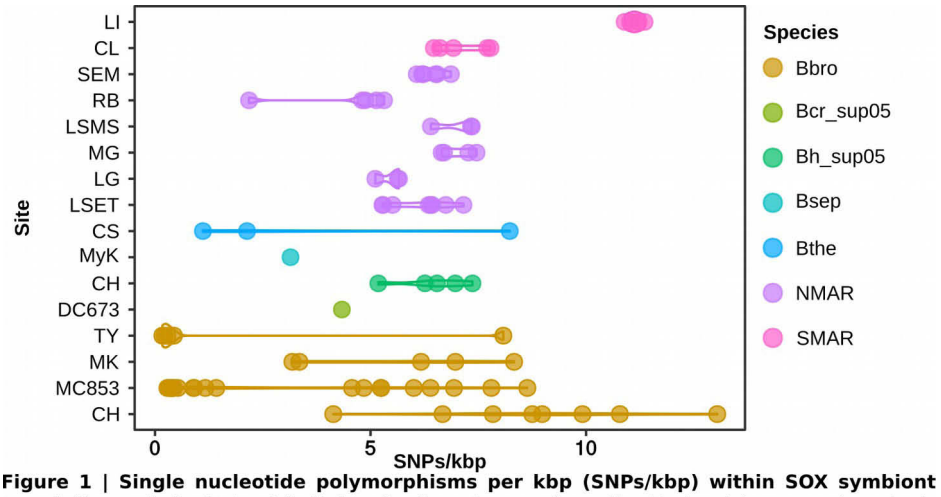


Figure 1 | Single nucleotide polymorphisms per kbp (SNPs/kbp) within SOX symbiont populations of single host individuals at each sampling site. Each point represents a single metagenome, originating from one mussel individual. Colors correspond to symbiont species (species were defined according to average nucleotide identity > 95%, see chapter III). The read coverage of all shown dataset was sampled the 100x. Sampling sites are listed in **Tab. 1**.

Two types of SNPs can be detected within a population: synonymous or silent substitutions (S), where the nucleotide change does not result in an amino acid change of the protein, and nonsynonymous substitutions (N) where the nucleotide change also leads to a change in the amino acid. Generally, within symbiont populations in single host individuals approx. 60% of all detected SNPs, and 69 to 75% of all SNPs in coding sequences, were nonsynonymous (**Fig. 2**). This means that

most polymorphic sites in the *Bathymodiolus* symbionts lead to different protein sequences. Intriguingly, a substantial proportion – 7 to 15% of all N substitutions affected the length of the protein sequence by either causing a premature stop codon or the loss of a stop codon (data not shown). To determine which proteins were enriched in amino acid substitutions, and potentially divergent between the co-existing symbiont strains, we identified outlier genes that had a high N/S ratio as a proxy to detect genes particularly enriched in N over S. (**Fig. S1**).

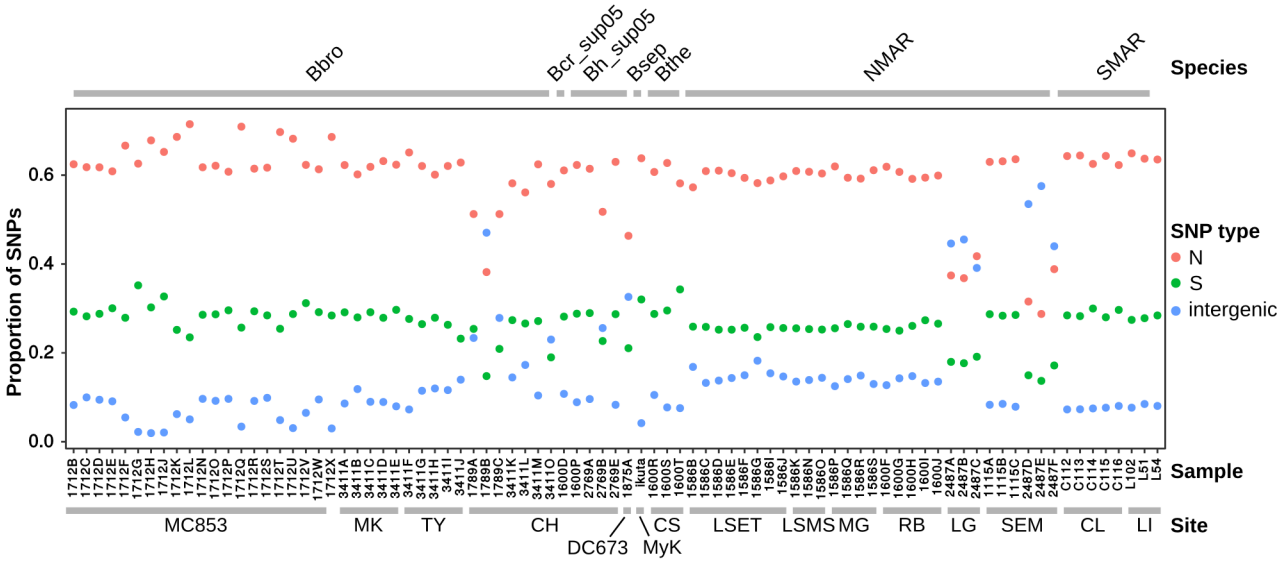


Figure 2 | Proportion of SNP type within symbiont populations of single host individuals (sample). Symbiont species are shown above the diagram. Sampling sites are indicated below the figure and are listed in **Tab. 1**. N: nonsynonymous SNP, S: synonymous SNP, inter: intergenic SNP. Proportions are in relation to the total number of SNPs per sample.

To verify whether symbiont populations of different host individuals shared genes with particularly large N/S ratio, we clustered orthologous genes according to sequence similarity (**Fig. 3**). This revealed an extreme variability of the genes showing a high N/S ratio, as every population had a unique characteristic profile and such genes were rarely shared between populations. Furthermore, very few genes were identified as outliers (< 8) for the SOX symbiont species Bbro and Bthe, whereas the other three species, NMAR, SMAR and Bh_sup05, had much more outliers (> 8, Fig. 3).

In addition, we investigated whether the outlier genes were associated with similar functional categories and assigned the outliers of high S, N, N/S values to Clusters of Orthologous Groups (COGs) (**Fig. 4**). For the N/S ratio we could identify two COG categories, Q and L, that appeared characteristic for members of the species NMAR, from some vent fields. Genes that belong to these two categories are involved in secondary metabolite synthesis and transport, and DNA replication, recombination and repair, respectively. All samples of symbiont species Bh_sup05 and the closely related lineage Bcr_sup05, as well as the divergent lineage Bsep, had N/S outliers in COG category C, which includes genes coding for proteins involved in energy production and conservation. Finally, we did not detect common patterns in N/S outliers for species SMAR and Bbro. However, both species showed N outliers in COG category L (DNA replication, recombination and repair). These were not detected as enriched in S and might have been absent from the N/S plot if there were no S in these genes. In summary, COG categories Q, L and C were most likely to have amino acid substitutions among co-existing strains. However, confirming the lack of shared patterns, the functional categories enriched in N or N/S outliers seem to be dependent on the symbiont species and even within a species, potentially also on sampling location.

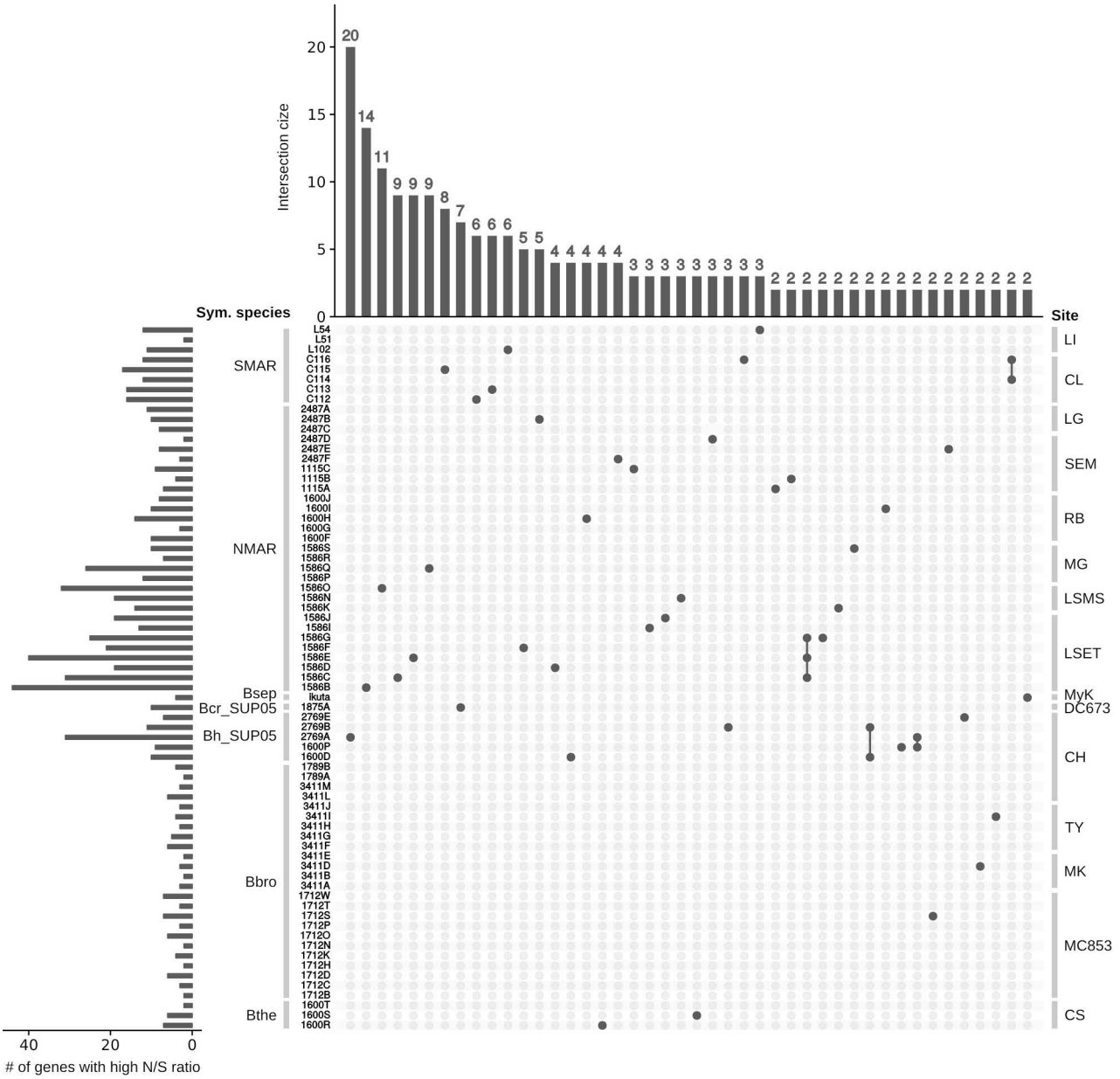


Figure 3 | Number of orthologous gene sets that were identified as outliers with a high N/S ratio within symbiont populations of single host individuals. Symbiont species (left of the panel) and sampling sites (right of the panel) are shown, and are also listed in **Tab. 1.** Only shared sets with more than two genes are displayed (singletons were excluded). There is little overlap of genes that were identified as N/S outliers between the samples.

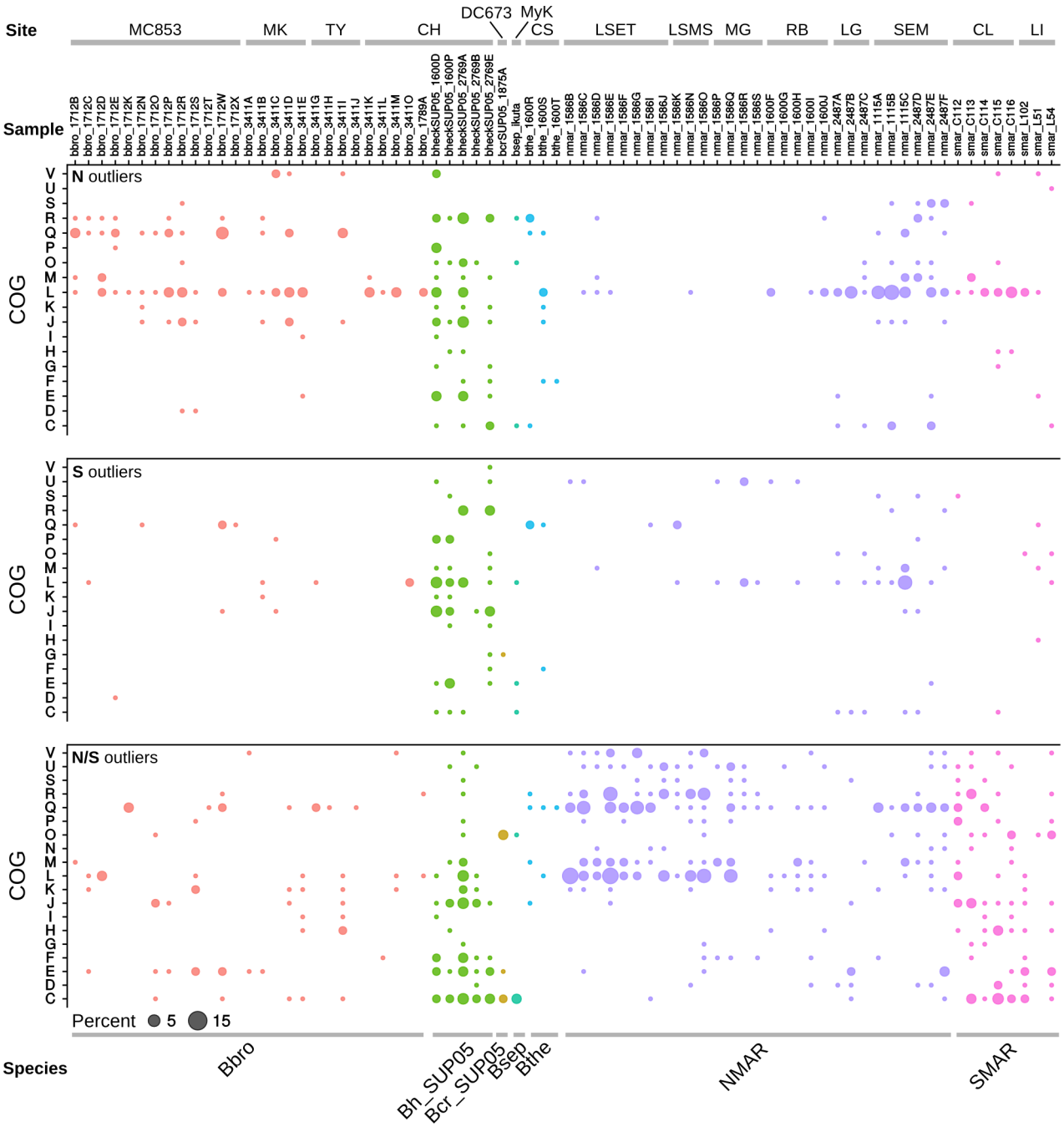


Figure 4 | Proportions of COG categories among genes that were identified as N, S or N/S outliers. Colors correspond to symbiont species and species affiliations are shown below the diagrams. Sampling sites are shown above the diagrams. Genes that could not be assigned to a COG category are not shown. COG categories are represented as follows: [V] Defense mechanisms, [U] Intracellular trafficking, secretion & vesicular transport, [T] Signal transduction mechanisms, [S] Function unknown, [R] General function prediction only, [Q] Secondary metabolites biosynthesis, transport & catabolism, [P] Inorganic ion transport & metabolism, [O] Post-translational modification, protein turnover & chaperones, [N] Cell motility, [M] Cell wall/ membrane/envelope biogenesis, [L] Replication, recombination & repair, [K] Transcription, [J] Translation, ribosomal structure & biogenesis, [I] Lipid transport & metabolism, [H] Coenzyme transport & metabolism, [G] Carbohydrate transport & metabolism, [F] Nucleotide transport & metabolism, [E] Amino acid transport & metabolism, [D] Cell cycle control, cell division, chromosome partitioning, [C] Energy production & conversion.

Diversifying selection acts upon symbiont populations

Most tests for adaptive selection, including the commonly used dN/dS[32,33], require sufficient divergence, or knowledge of distinct strain genome sequences. In natural populations, mixtures of closely related strains make these tests unsuitable. Yet, population genetic differentiation can be indicated by the frequency of SNPs at different loci in a reference genome (F_{ST}). F_{ST} outliers among populations of highly related bacteria can therefore identify candidate loci that are potentially evolving under diversifying selection[29]. We calculated F_{ST} outliers among all symbiont populations per species to identify genes that carry loci affected by diversifying selection (**Fig. S2**).

In contrast to the heterogeneity *within* single bacterial populations (shown above) the F_{ST} outlier test identifies significant differences in allele frequencies *between* populations of the same species and therefore can indicate whether loci are evolving under selection. The direction of selection can be either purifying, which means that deleterious mutations are selected against, or disruptive, selecting for allele frequencies that diversify among the compared populations. To identify traits that are possibly adaptive, we focused on the latter and identified genes with one or more loci that were affected by diversifying selection[29]. Excluding all hypothetical proteins from the analysis, the number of genes that contained diversifying loci differed greatly between symbiont species: 848 genes in Bbro, 272 in Bh_sup05, 146 genes in Bthe, 1171 genes in NMAR, and 126 genes in SMAR. It was not surprising to identify more diversifying loci in symbiont species Bbro and NMAR, as the populations were collected from many different sites that differ in biogeochemical conditions, which may lead to differences in the selection pressure. Surprisingly though, orthologous

clustering revealed that almost half of the genes, identified as diversifying were shared between species Bbro and NMAR (**Fig. 5**). This was confirmed when we analyzed the COG categories of the diversifying candidate genes and found highest similarities, according to abundance of COG categories, between these two species (**Fig. 5**).

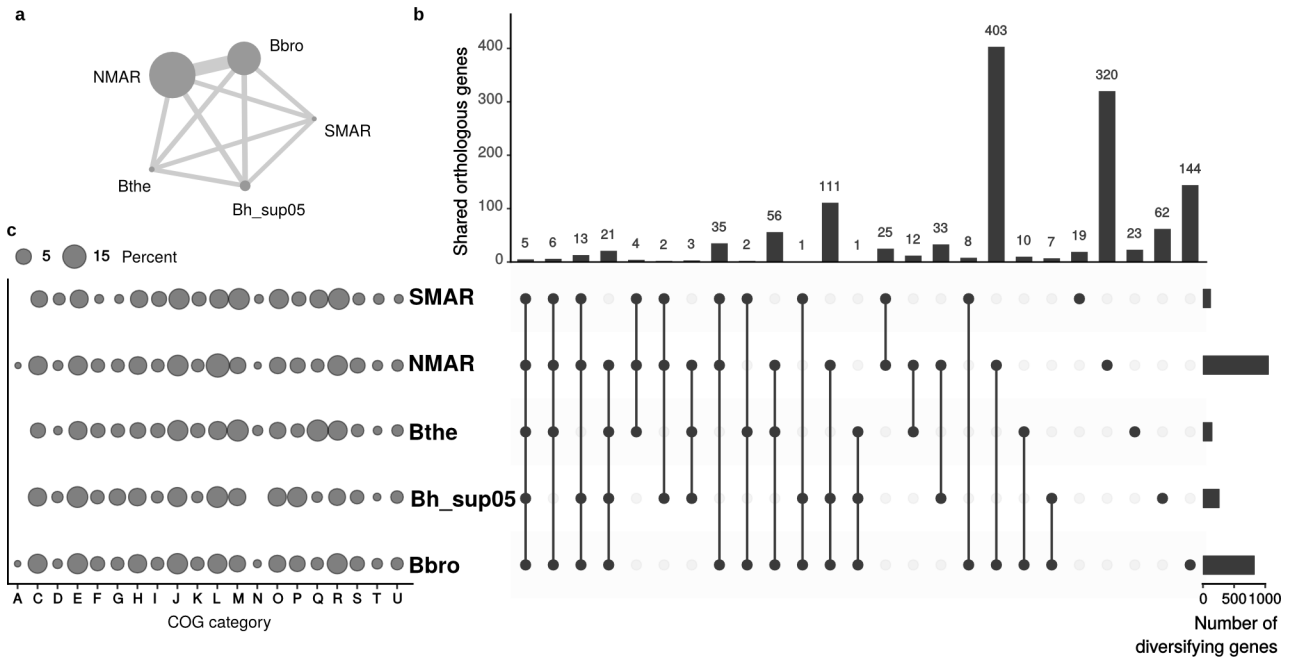


Figure 5 | Candidate genes under diversifying selection within symbiont species as inferred from F_{ST} outliers (**a**, **b**, **c**). (**a**) Similarities (based on Bray-Curtis dissimilarities) of orthologous gene clusters among different symbiont species (NMAR, SMAR, Bbro, Bh_sup05 and Bthe) are displayed in a network. Thickness of edge indicates similarity and size of the node corresponds to the number of diversifying candidate genes per species. (**b**) Absolute numbers of diversifying annotated genes (including hypothetical proteins) were 848 (1658) genes in Bbro, 272 (403) in Bh_sup05, 146 (221) genes in Bthe, 1171 (2066) genes in NMAR, and 126 (315) genes in SMAR. For the comparison of shared orthologs between species, hypothetical proteins were exluded. (**c**) Diversifying genes were assigned to COG categories and their relative representation is shown for each symbiont species. COG categories are represented as follows: [V] Defense mechanisms, [U] Intracellular trafficking, secretion & vesicular transport, [T] Signal transduction mechanisms, [S] Function unknown, [R] General function prediction only, [Q] Secondary metabolites biosynthesis, transport & catabolism, [P] Inorganic ion transport & metabolism, [O] Post-translational modification, protein turnover & chaperones, [N] Cell motility, [M] Cell wall/membrane/envelope biogenesis, [L] Replication, recombination & repair, [K] Transcription, [J] Translation, ribosomal structure & biogenesis, [I] Lipid transport & metabolism, [H] Coenzyme transport & metabolism, [G] Carbohydrate transport & metabolism, [F] Nucleotide transport & metabolism, [E] Amino acid transport & metabolism, [D] Cell cycle control, cell division, chromosome partitioning, [C] Energy production & conversion.

Genes with diversifying loci could be found in most COG categories (**Tab. S1 only digital on CD**). Gene categories that had most diversifying loci included toxin-related genes, such as RTX, Rhs or insecticidal toxins, secretion systems or mobile elements. However, we also detected diversifying loci in genes encoding important

metabolic processes, such as energy generation. For example, we identified three genes, HypE, HypD and HypB with diversifying loci in the SMAR symbiont species that included samples from two distinct vent field (Clueless and Lilliput). These genes are part of the hydrogenase operon that encodes the capability to use hydrogen as an additional energy source and has been shown to be part of the accessory genome among co-existing symbiont strains[15,22]. For three species, Bh_sup05, Bbro and NMAR, we detected the sulfite reduction-associated complex DsrMKJOP as diversifying. For NMAR, we additionally identified diversifying loci in the SoxABXYZ genes. The proteins encoded by these two gene sets are essential to the sulfur oxidation pathways, a key component of the energy metabolism in the symbionts. Thus, our results suggest that diversifying selection acts upon both accessory and core functions.

Symbiont population structure differs between B. brooksi and other host species

In addition to the identification of diversifying loci, F_{ST} can be used to determine population structure. Especially in natural symbiont populations that cannot be cultured in the laboratory this is a powerful approach to understand transmission and symbiont population dynamics between host individuals[15,20]. We used F_{ST} to determine how similar the different symbiont populations were within a species. Hence, pairwise F_{ST} per gene was calculated among all samples per symbiont species (**Fig. 6**). Large F_{ST} values indicate differentiation and low values indicate high similarity between populations. We detected that for symbiont species NMAR, SMAR and Bh_sup05, the individuals from the same site have very similar symbiont populations (relatively small F_{ST}), whereas those from different sites tend to be more differentiated. Exceptions were vent fields Lucky Strike (site Montsegur and

Eiffeltower) and Menez Gwen, which also had rather similar symbiont populations despite the distance separating these sampling sites (**Fig. 6**). Intriguingly, this pattern was completely different for symbiont species Bbro and Bthe, both of which showed highly differentiated signals also among individuals from the same site. This is consistent with and linked to the SNP densities that differed strongly according to host individual (**Fig. 1**).

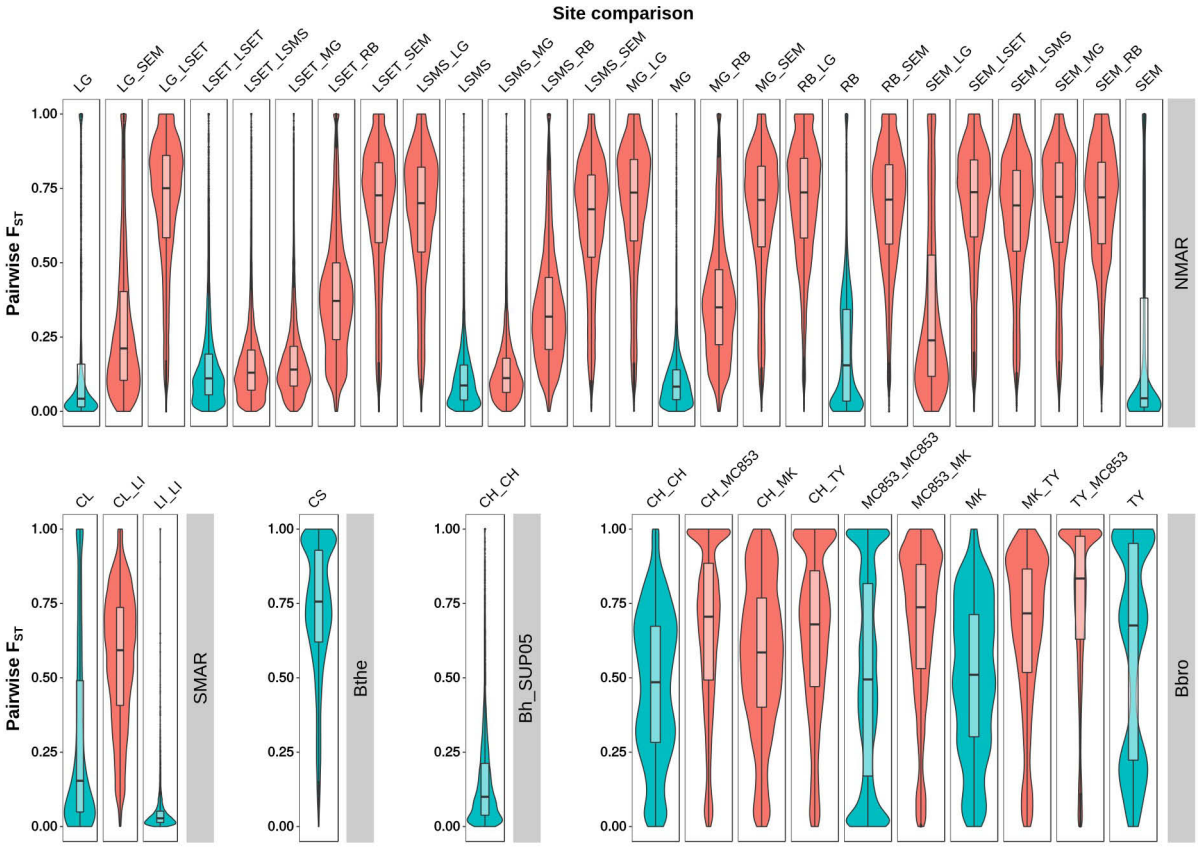


Figure 6 | Distribution of pairwise F_{ST} between individuals of the same site (blue) and between different sites (red) within one species (grey bars). Low F_{ST} indicates low differentiation, hence the populations are more similar. Large F_{ST} indicates strong differentiation, hence the populations are more diverse. The sampling site comparisons are shown above each panel.

We investigated if host age could be a factor that is influencing the broad range of SNP densities and differences in F_{ST} between individuals of the same site in the symbiont species Bbro and Bthe. We plotted SNP counts against mussel shell length which is an indicator of host age[34] and showed that regardless of shell length, some

individuals had low symbiont diversity (for Bthe hosts shell lengths were unknown). This analysis also revealed that at a shell length above 17 cm there seems to be a drop in heterogeneity in most host individuals, and this is mainly driving the negative correlation between shell length and SNPs/kbp (**Fig. 7**).

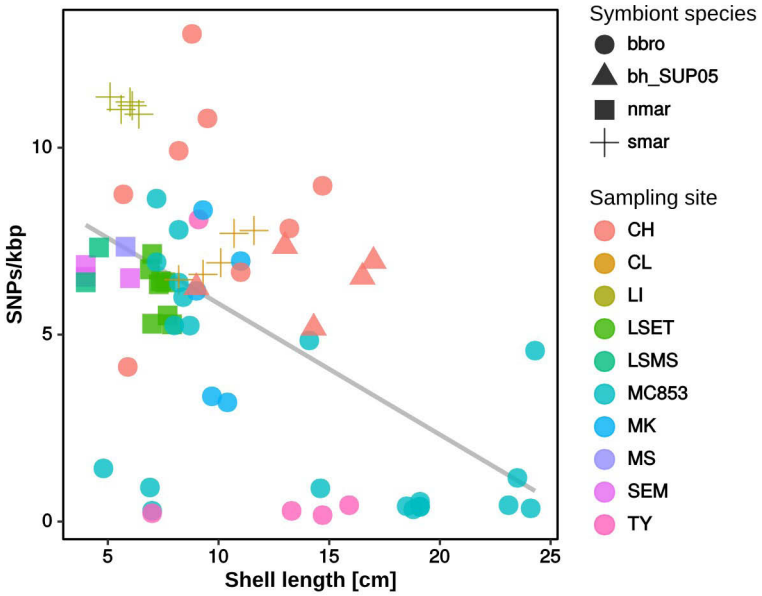


Figure 7 | Single nucleotide polymorphisms per kbp (SNPs/kbp) within symbiont populations against host shell length. Symbiont species Bthe, Bcr_sup05 and Bsep were not included as no shell length data were avalable. Colors correspond to sampling site and symbols to symbiont species (see **Tab. 1**). Spearman correlation coefficient (rho) was -0.39 with a p-value of 0.0009, comparing SNPs/kbp agains shell length.

Discussion

Non-neutral symbiont heterogeneity reflects increased evolvability

Our study provides the foundation for elucidating the genome signatures of allelic variation and selective pressure in horizontally transmitted endosymbiont populations in*Bathymodiolus* hosts. The high number of N SNPs in intra-host populations suggest functionally meaningful variation among co-existing strains. In fact, N substitutions can potentially have strong impact on bacterial lifestyle, including host specificity as

shown for pathogenic *Salmonella enterica*[35]. In the SOX symbionts up to 75% of detected SNPs were N substitutions, a value that is consistent with previous findings in *S. enterica* strains, where N SNPs represented 79% of all coding SNPs. However, in contrast to the *S. enterica* strains that infected multiple different animal hosts, and thus likely experienced starkly varying selective pressures, *Bathymodiolus* symbiont strains that co-exist in a single host individual showed similar N polymorphism numbers, and thus possibly functional heterogeneity. It is worth noting that not all amino acid substitutions are equally effective. Whether, and how much the function of a protein is altered by a substitution depends on the position in the protein where the amino acid was replaced, and the chemical properties of the amino acid[36]. For example, if an amino acid with similar chemical properties as the original one was replaced in a position that is not important for the protein task, protein function may not be affected. In contrast, if the substitution is in the active site of an enzyme, or affects protein folding, a single replacement may have large impact on function. On the other hand, premature or lost stop codons, which in our dataset accounted for up to 15% of all Ns, very likely affects protein function. Overall, our data indicate that the genetic variability in the SOX endosymbiont is not “neutral”, but rather the product of natural selection.

Within symbiont populations, three COG categories were particularly affected by amino acid substitutions: energy metabolism, secondary metabolism, DNA recombination and repair. This suggests that co-existing strains show variability in different functional categories, which is in line with previous findings of functional heterogeneity in the genomic content of co-existing strains in *Bathymodiolus* symbionts[15,22]. In our analyses we obtained a snapshot of sequence variation within single bacterial populations at a single time point. The direction of selection at N loci

within symbiont populations remains unclear, as the observed polymorphisms may be transient and have not yet been fixed or purged by selection. However, such functionally different polymorphisms may represent signatures of parallel evolution of co-existing strains within the host[37], as has been shown for bacterial pathogens. This study showed that even if a trait is adaptive, it does not necessarily sweep to fixation in the population leading to a decrease in heterogeneity in favor of this allele. Instead different alleles can persist in parallel over years[37], increasing the population heterogeneity. In fact we hypothesized in Chapter III that evolvability itself, the potential of a population to adapt to unknown future conditions, could be a trait selected for. A 'standing stock' of functionally different loci, as observed in the present study, supports this hypothesis. While slightly deleterious mutations are expected to be purged by selection in populations with large N_e[38], heterogeneity in adaptive traits could persist and increase evolvability.

Divergent selection affects core functions in Bathymodiolus endosymbionts

As for most horizontally transmitted endosymbionts, N_e is unknown in *Bathymodiolus* symbionts. If N_e is reduced, e.g. by transmission bottlenecks such as those observed in *Ca*. Endoriftia[13], selection is expected to be weak and evolution mostly driven by (neutral) genetic drift[2,4]. If there is no such bottleneck and symbionts continue to colonize their host over an extended period of time, N_e is expected to be larger, and genome evolution driven by both, selection and genetic drift. Previously, Picazo et al.[20] showed that the SOX symbiont species Bbro at site MC853 evolves mainly under a purifying selection regime, showing that selection is effective in this species. This result was based on analyses over the entire core genome of the symbiont strains from one sampling site. In our study we compared F_{ST} among symbiont populations to

identify candidate genes that have been evolving under selection within five different symbiont species. Our analysis included core and accessory genes, different sites with distinct biochemical conditions (e.g. site MC853, CH, MK and TY for symbiont species Bbro) and different host species (e.g. *B. azoricus* and *B. puteoserpentis* for symbiont species NMAR, **Tab. 1**). We detected candidate genes under diversifying selection by significant shifts in allele frequencies. While this approach detects divergent selection between the populations within one species, it is not possible to determine which type of selection is affecting a single population. Many of these diversifying loci identified in toxin-related genes that have been hypothesized to be involved in host-symbiont interaction[21]. Diversifying selection in the interaction mechanisms can help to define host specificity as has been suggested for *S. enterica,* in which N in interaction-related proteins influence the type of animal host which could be infected[35]. Potential interaction mechanisms in *Bathymodiolus* symbionts could thus be selected for by the interaction with the host, ensuring specific recognition and successful colonization that is essential in horizontally transmitted symbionts[1]. For example, this could be the case if *Bathymodiolus* hosts that are isolated (e.g. geographically at different sites). These hosts may (co-)evolve different recognition mechanisms with their local symbiont population. Another example can be observed in the NMAR symbiont species. This symbiont species colonizes two different host species, *B. azoricus* and *B. puteoserpentis*, and thus host-symbiont interaction mechanisms may be different. Alternatively, toxin-related genes may also be involved in other processes such as symbiont-mediated host defense against pathogens[21]. Diversifying selection therefore could indicate differences in pathogen abundance or types.

In addition to the host, external environmental conditions can apply selective pressure on the symbionts, particularly considering that they obtain their energy from the environment[17,39]. We previously identified differences in gene abundances within and among *Bathymodiolus* symbiont populations and hypothesized that this is an adaptive process driven by environmental conditions[15]. Here, we tested whether we can find further evidence for such selective processes, based on allele frequencies. We identified several diversifying loci in genes encoding proteins for energy generation pathways, such as hydrogen and sulfur oxidation. Such data can be interpreted in the light of the concentration of resources in the environment. In fact, between the two sampling sites Clueless and Lilliput hydrogen concentrations differ substantially[15,40,41], which could explain our finding of diversifying selection in hydrogenase-related genes among the symbiont populations at these two sites. The substrate concentration could therefore affect the direction of selection on the metabolizing enzymes. Unlike hydrogen oxidation, sulfur oxidation genes are part of the core genome in *Bathymodiolus* symbionts (Chapter III). Yet, we also detected diversifying selection on these genes in three out of five symbiont species. Therefore, our study supports the previous hypothesis of adaptive processes driven by environmental conditions in *Bathymodiolus* symbionts and extends this observation to core genes, which, although not subject to variation in presence or absence in co-occurring strains, nevertheless showed signatures of diversifying selection between populations.

The detection of diversifying loci among populations of the same symbiont species suggests that N_e is large enough to allow selection to be effective, and to minimize the effect of genetic drift. Therefore, the within-host heterogeneity, which included an enrichment of N substitutions in energy generation pathways, is not neutral. These

polymorphisms may instead serve as diverse repertoire of traits that can be selected for or against in diverse environmental settings. Despite effective selection, heterogeneity nevertheless persists in symbiont populations, thus contributing to the evolvability of the symbiont populations. Our results align well with previous analyses, and altogether strengthen the idea that functional heterogeneity based on both gene content variation and nucleotide and amino acid polymorphisms may be beneficial, and evolvability may be selected for[15] (see Chapter III).

Continuous symbiont exchange might be linked to host age and abundance

The comparison of the F_{ST} values within and between sites revealed that for most *Bathymodiolus* species, *B. azoricus*, *B. puteoserpentis*, *B.* sp. 1 Clueless, *B.* sp. 2 Lilliput and *B. heckerae*, mussel individuals from the same sampling site harbor similar symbiont populations. This was in line with previous findings[15] and supports the hypothesis of symbiont exchange among co-occurring host individuals. Symbiont exchange would lead to increased heterogeneity within symbiont populations and low F_{ST} values between populations of the same site. However, for host species *B. brooksi* and *B. thermophilus* we observed extensive differences in the symbiont population structure within and among host individuals, which possibly reflects differences in the symbiont acquisition. For *B. thermophilus,* the limited number of individuals (3), sampling sites (1) and unknown shell lengths hinder our ability to explain the observed differences between individuals. Our observations for *B. brooksi* are in line with the results of Picazo et al.[20], who analyzed samples from a single site (MC853), which were also part of our analysis. In addition, we included three more sampling sites for the same host and symbiont species and showed that this observation of

variation in symbiont population heterogeneity and F_{ST} between host individuals still holds true. In order to fully understand transmission processes in the symbiont population it is essential to tease apart the reason for this discrepancy between *B. brooksi* and the other host species we analyzed. Our results indicate a possible connection with shell length indicative of host age, since *B. brooksi* samples included specimen larger than 17 cm which were not available for any of the other species. Of the species sampled, only *B. brooksi* grows to such large sizes. We see a clear drop in heterogeneity at shell lengths above 17 cm (also shown by Picazo et al.[20]). Growth rates and maximum sizes differ between species of *Bathymodiolus*[42,43]. Growth rates consistently slow down with age in the few species so far investigated, eventually reaching saturation, however, the length at which growth ceases differs from species to species. An individual that grows very slowly has less growing gill tissue that needs to be colonized by symbionts. Symbiont diversity may decrease as the host ages through stochastic loss of symbiont strains. In actively growing individuals these might be constantly replenished by strains from the environment. These observations could explain the drop of heterogeneity in older hosts, and partially explain the differences in F_{ST} between *B. brooksi* and the other species because only in *B. brooksi* we had hosts larger than 17 cm. However, the patterns we see (Fig. 7) cannot be fully explained by host age, as some small *B. brooksi* host individuals also show relatively low symbiont diversity.

Another possible explanation for the similarity of symbiont populations in co-occuring host individuals is the number of symbiont cells in the water column, e.g. the environmental symbiont “titer”. The symbiont species along the Mid Atlantic Ridge originate from mussel beds that are extremely dense. In contrast, *Bathymodiolus brooksi* from the sampling sites MC853, MK, TY and CH occurred in small isolated

patches with CH being the site with most mussels per patch (*personal communication with Christian Borowski and Maxim Rubin-Blum and images from field work*)[20]. This more scattered distribution of mussels may decrease the symbiont titer in the water column, which could make the uptake from the water column less efficient and the re-infection within the animals therefore the more dominant colonization process. Loss of symbiont strains may thus not be replenished in all individuals, leading to a decrease in heterogeneity in the *B. brooksi* mussels which consequently leads to an increase in F_{ST} between individuals.

Altogether, these data indicate that the colonization dynamics over the lifetime of *Bathymodiolus* hosts are still not fully understood. This variation observed in *B. brooksi* and *B. thermophilus* might be due to a variety of factors including host species, symbiont titer, host age and symbiont loss. We suggest a systematic approach with different size ranges of not only adult mussels but also juveniles of different host species to investigate, how symbiont transmission happens throughout the lifetime of *Bathymodiolus*.

Conclusion and outlook

Our study revealed the potential of high-resolution metagenomics datasets for understanding both genome evolution in natural symbiotic bacteria and symbiont transmission in deep-sea animal hosts. A strong bias towards the study of genome evolution in vertically transmitted endosymbionts has left a lot of ‘unknowns’ in horizontally transmitted symbionts. The selection regimes under which bacteria evolve is strongly affected by N_e, which can vary substantially between different horizontally transmitted endosymbionts. Our observations of high heterogeneity,

effective natural selection and high similarity of symbiont populations between co-occurring hosts all support the hypothesis of continuous symbiont uptake and exchange in *Bathymodiolus*. *B. brooksi* is an exception that can possibly be explained by factors such as age and abundance of free-living symbionts. The high density of amino acid substitutions within symbiont populations and diversifying genes between populations are in line with previous findings of functional heterogeneity within and between symbiont populations in some *Bathymodiolus* symbionts[15] (see Chapter III). Our findings show that natural selection affects not only accessory but also core functions and strongly supports the idea that evolvability is selected for in these symbioses, and both the environment and the host can drive such selection.

In addition to the environment, other factors such as phage predation, interaction among symbiont strains or with other symbiont species, and differences in lifestyles can also act as selective forces. A more targeted analysis based on multi-factorial combinations of e.g. environments, hosts species, hosts age, would help to tease apart some of the factors driving diversifying selection in the symbionts. The analyses of allele frequencies and F_{ST} outliers is a sensitive method that captures ongoing evolutionary processes implying that the genomic signals we described were of relatively recent selection within symbiont species[11]. Additional analyses could be performed to detect more ancient selection based on substitution rates of fixed mutations between the symbiont species (e.g. dN/dS)[33]. In order to perform a dN/dS analyses, the genomes have to be divergent. For this analysis representative genomes for each species is required. The symbiont lineages in this study are closely related to free-living bacteria and symbionts of other host types (e.g. sponges). A dN/dS analysis among all of these would allow us to detect signatures of selection between these lifestyles. Finally, by investigating the genes under strong purifying or

balancing selection, within species through F_{ST} outliers or between species through dN/dS, we could start to understand which functions are fundamental in the symbiont lineages, hence need to be conserved.

The population-wide genome signatures we detected in *Bathymodiolus* symbionts can be explained by a large N_e. This deviates substantially from most studies on the genome evolution of endosymbionts, as these considered usually vertically transmitted symbionts with small N_e. The high functional heterogeneity and the effective selection we identified are not observed in the 'typical' endosymbiont, but are reminiscent of a 'typical' free-living bacterium. Compared to the latter, a horizontally transmitted endosymbiont has a host-associated and a free-living stage, and therefore is characterized by the necessity to survive under both conditions. The consequences of this multidimensional lifestyle on genome evolution are of strong interest because the latter can help use to understand fundamental processes underpinning symbiotic associations. For example, genomic signatures in *Solemya velum* symbionts have revealed that instead of strict vertical transmission, the symbionts experience occasional horizontal transmission events resulting in a mixed transmission mode[44,45]. Horizontal transmission of endosymbionts can result in small and large N_e and studies such as ours show the potential of elucidating the evolutionary dynamics affecting these communities. We encourage to continue and deepen the investigation of genomic evolutionary signatures in natural endosymbionts undergoing horizontal transmission, which may be a bacterial 'solution' to avoid going down the evolutionary rabbit hole.

Acknowledgements

We thank the captains, crews and ROV teams on the cruises BioBaz (2013), ODEMAR (2014), M126 (2016), M78-2 (2009), NA58 (2015), NA43 (2014), M114-2 (2015), AT26-23 (2014), ATA57 (2008) on board of the research vessels Pourquoi Pas?, FS Meteor, E/V Nautilus, and L'Atalante and the chief scientists on these research expeditions. This study was funded by the Max Planck Society, the MARUM DFG-Research Center / Excellence Cluster “The Ocean in the Earth System” at the University of Bremen, the German Research Foundation, an ERC Advanced Grant (BathyBiome, 340535), and a Gordon and Betty Moore Foundation Marine Microbial Initiative Investigator Award to ND (Grant GBMF3811).

References for chapter 4

1. Bright, M. & Bulgheresi, S. A complex journey: transmission of microbial symbionts. *Nat. Rev. Microbiol.* **8**, 218–230 (2010).

2. Moran, N. A., McCutcheon, J. P. & Nakabachi, A. Genomics and evolution of heritable bacterial symbionts. *Annu. Rev. Genet.* **42**, 165–190 (2008).

3. Moran, N. A. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc. Natl. Acad. Sci.* **93**, 2873–2878 (1996).

4. Wernegreen, J. J. Endosymbiont evolution: predictions from theory and surprises from genomes. *Ann. N. Y. Acad. Sci.* **1360**, 16–35 (2015).

5. Fraser, C., Alm, E. J., Polz, M. F., Spratt, B. G. & Hanage, W. P. The bacterial species challenge: making sense of genetic and ecological diversity. *Science* **323**, 741–746 (2009).

6. Bennett, G. M. & Moran, N. A. Heritable symbiosis: The advantages and perils of an evolutionary rabbit hole. *Proc. Natl. Acad. Sci.* **112**, 10169–10176 (2015).

7. Husnik, F. & McCutcheon, J. P. Repeated replacement of an intrabacterial symbiont in the tripartite nested mealybug symbiosis. *Proc. Natl. Acad. Sci.* **113**, E5416–E5424 (2016).

8. Koga, R. & Moran, N. A. Swapping symbionts in spittlebugs: evolutionary replacement of a reduced genome symbiont. *ISME J.* **8**, 1237–1246 (2014).

9. Smith, W. A. *et al.* Phylogenetic analysis of symbionts in feather-feeding lice of the genus *Columbicola*: evidence for repeated symbiont replacements. *BMC Evol. Biol.* **13**, 109 (2013).

10. Sun, Z. & Blanchard, J. L. Strong Genome-Wide Selection Early in the Evolution of *Prochlorococcus* Resulted in a Reduced Genome through the loss of a large number of small effect genes. *PLOS ONE* **9**, e88837 (2014).

11. Hohenlohe, P. A., Phillips, P. C. & Cresko, W. A. Using population genomics to detect selection in natural populations: key concepts and methodological considerations. *Int. J. Plant Sci.* **171**, 1059–1071 (2010).

12. Beaumont, M. A. Adaptation and speciation: what can Fst tell us? *Trends Ecol. Evol.* **20**, 435–440 (2005).

13. Nussbaumer, A. D., Fisher, C. R. & Bright, M. Horizontal endosymbiont transmission in hydrothermal vent tubeworms. *Nature* **441**, 345–348 (2006).

14. Sycuro, L. K., Ruby, E. G. & McFall-Ngai, M. Confocal microscopy of the light organ crypts in juvenile *Euprymna scolopes* reveals their morphological complexity and dynamic function in symbiosis. *J. Morphol.* **267**, 555–568 (2006).

15. Ansorge, R. *et al.* Diversity matters: Deep-sea mussels harbor multiple symbiont strains. *bioRxiv* 531459 (2019).

16. Van Dover, C. L. V. *The Ecology of Deep-Sea Hydrothermal Vents*. (Princeton University Press, 2000).

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

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

19. Wentrup, C., Wendeberg, A., Schimak, M., Borowski, C. & Dubilier, N. Forever competent: deep-sea bivalves are colonized by their chemosynthetic symbionts throughout their lifetime. *Environ. Microbiol.* **16**, 3699–3713 (2014).

20. Picazo, D. R. *et al.* Horizontally transmitted symbiont populations in deep-sea mussels are genetically isolated. *bioRxiv* 536854 (2019).

21. Sayavedra, L. *et al.* Abundant toxin-related genes in the genomes of beneficial symbionts from deep-sea hydrothermal vent mussels. *eLife* e07966 (2015).

22. Ikuta, T. *et al.* Heterogeneous composition of key metabolic gene clusters in a vent mussel symbiont population. *ISME J.* **10**, 990-1001 (2015).

23. Shah, V. & Morris, R. M. Genome sequence of "*Candidatus* Thioglobus autotrophica" strain EF1, a chemoautotroph from the SUP05 clade of marine Gammaproteobacteria. *Genome Announc.* **3**, (2015).

24. Marshall, K. T. & Morris, R. M. Genome sequence of "*Candidatus* Thioglobus singularis" strain PS1, a mixotroph from the SUP05 clade of marine Gammaproteobacteria. *Genome Announc.* **3**, (2015).

25. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

26. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118, *Fly (Austin)* **6**, 80–92 (2012).

27. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013).

28. R Core Team. *A language and environment for statistical computing. R Foundation for Statistical Computing.* (2016).

29. Foll, M. & Gaggiotti, O. A Genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a bayesian perspective. *Genetics* **180**, 977–993 (2008).

30. Chaudhari, N. M., Gupta, V. K. & Dutta, C. BPGA- an ultra-fast pan-genome analysis pipeline. *Sci. Rep.* **6**, 24373 (2016).

31. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).

32. Kryazhimskiy, S. & Plotkin, J. B. The population genetics of dN/dS. *PLoS Genet* **4**, e1000304 (2008)

33. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43 (2000).

34. Rhoads, D. C., Lutz, R. A., Revelas, E. C. & Cerrato, R. M. Growth of Bivalves at Deep-Sea Hydrothermal Vents Along the Galápagos Rift. *Science* **214**, 911–913 (1981).

35. Yue, M. *et al.* Allelic variation contributes to bacterial host specificity. *Nat. Commun.* **6**, (2015).

36. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* **7**, (2012).

37. Lieberman, T. D. *et al.* Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nat. Genet.* **46**, 82–87 (2014).

38. Hughes, A. L., Friedman, R., Rivailler, P. & French, J. O. Synonymous and nonsynonymous polymorphisms versus divergences in bacterial genomes. *Mol. Biol. Evol.* **25**, 2199–2209 (2008).

39. Petersen, J. M. *et al.* Hydrogen is an energy source for hydrothermal vent symbioses. *Nature* **476**, 176–180 (2011).

40. Perner, M. *et al.* Driving forces behind the biotope structures in two low-temperature hydrothermal venting sites on the southern Mid-Atlantic Ridge. *Environ. Microbiol. Rep.* **3**, 727–737 (2011).

41. Perner, M. *et al.* Short-term microbial and physico-chemical variability in low-temperature hydrothermal fluids near 5°S on the Mid-Atlantic Ridge. *Environ. Microbiol.* **11**, 2526–2541 (2009).

42. Nedoncelle, K., Lartaud, F., de Rafelis, M., Boulila, S. & Le Bris, N. A new method for high-resolution bivalve growth rate studies in hydrothermal environments. *Mar. Biol.* **160**, 1427–1439 (2013).

43. Schöne, B. R. & Giere, O. Growth increments and stable isotope variation in shells of the deep-sea hydrothermal vent bivalve mollusk *Bathymodiolus brevior* from the North Fiji Basin, Pacific Ocean. *Deep Sea Res. Part Oceanogr. Res. Pap.* **52**, 1896–1910 (2005).

44. Russell, S. L., Corbett-Detig, R. B. & Cavanaugh, C. M. Mixed transmission modes and dynamic genome evolution in an obligate animal–bacterial symbiosis. *ISME J.* **11**, 1359 (2017).

45. Russell, S. L. & Cavanaugh, C. M. Intrahost genetic diversity of bacterial symbionts exhibits evidence of mixed infections and recombinant haplotypes. *Mol. Biol. Evol.* **34**, 2747–2761 (2017).

Supplementary figures

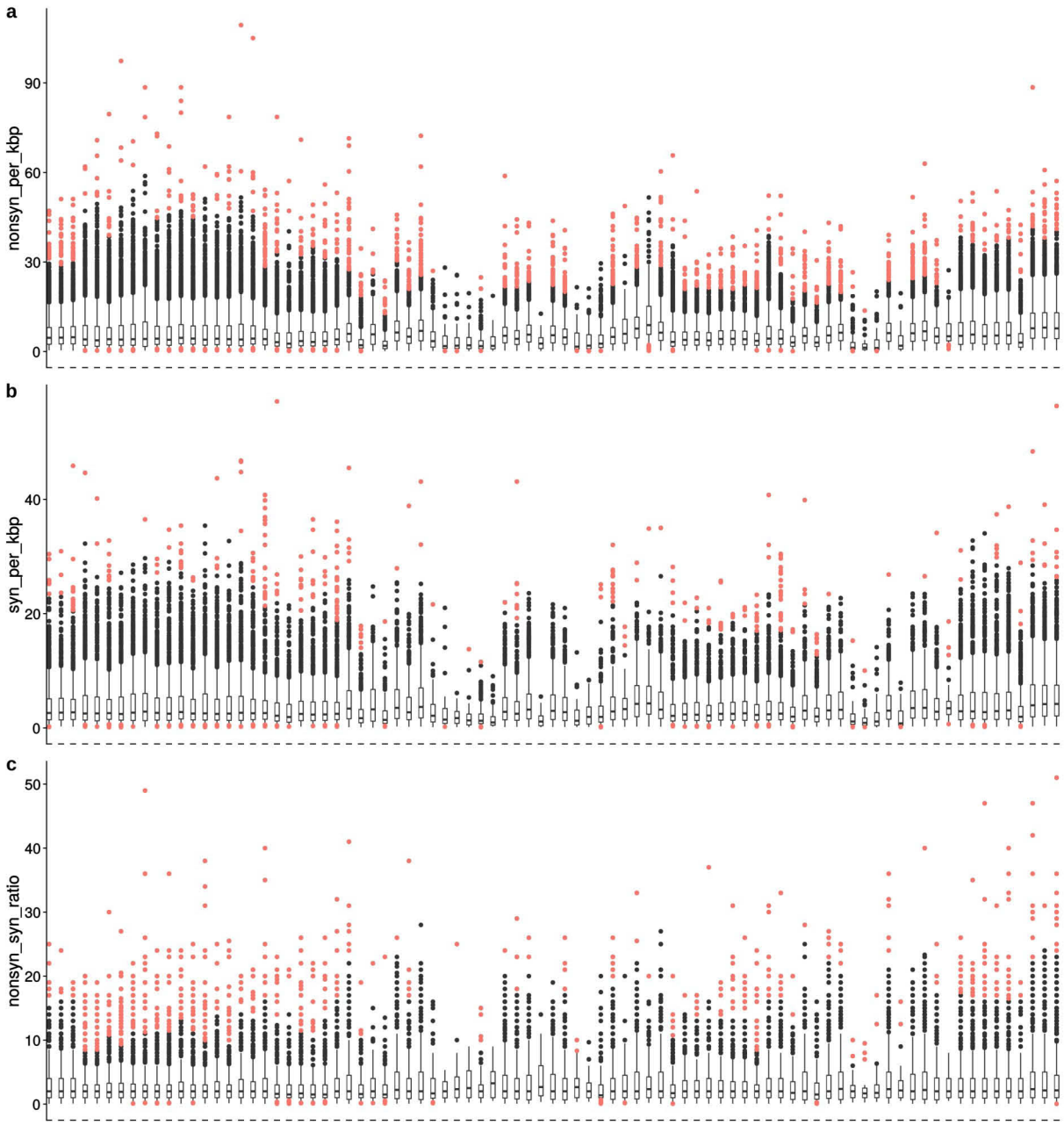


Figure S1 | Distributions of (a) nonsynonymous (N), (b) synonymous (S) and (c) N/S ratio per gene and metagenome SOX bin. Outliers that were identified with the adjboxStats function in the robustbase package in R are marked in red.

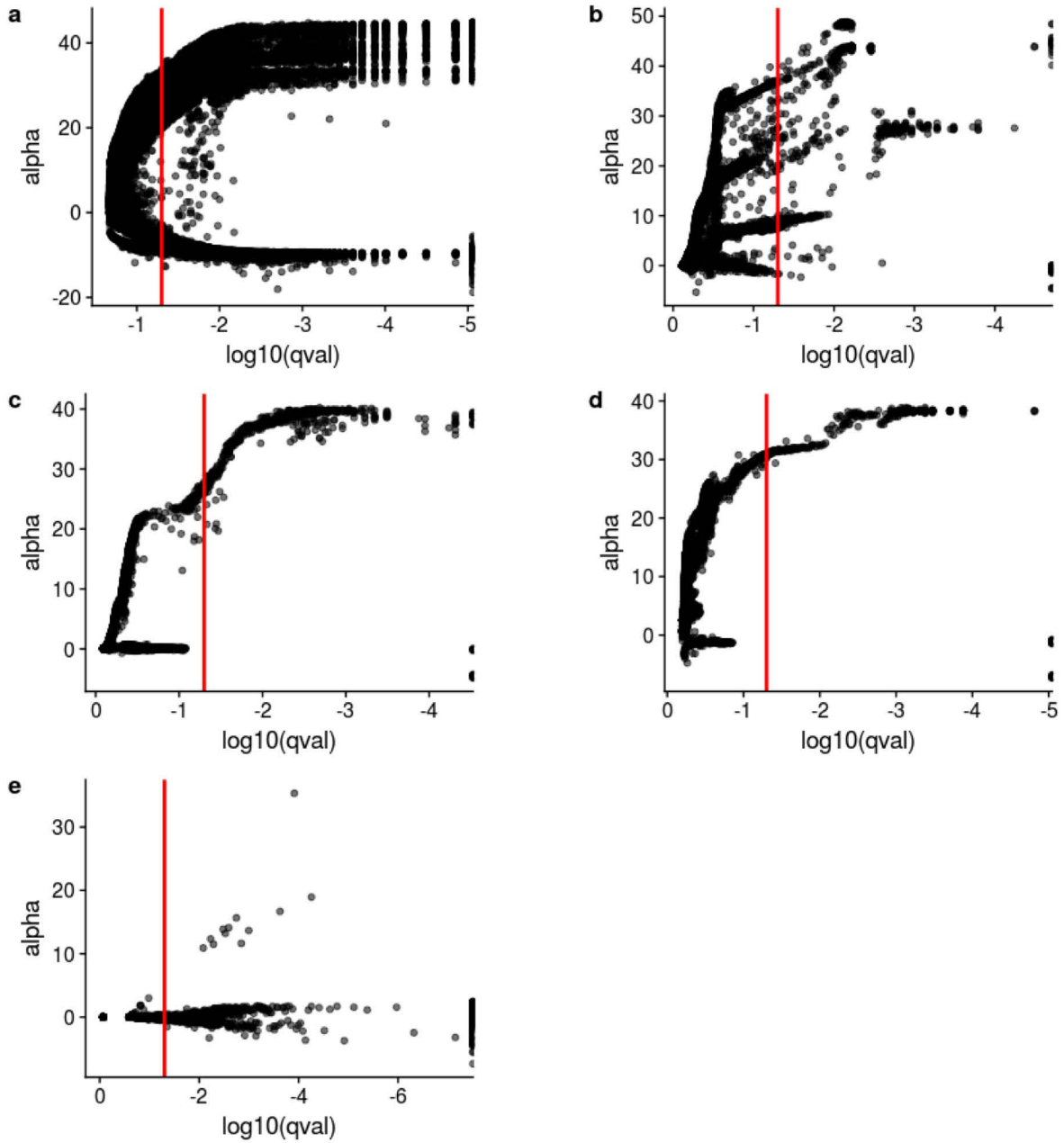


Figure S2 | Results from the F_{ST} outlier analysis using BayeScan. Each data point corresponds to one gene. Positive alpha values are indicative of diversifying selection and negative alpha values are indicative of purifying selection. The red line marks a q-value cut-off of 0.05 above which the result can be considered significant. Each panel corresponds to one symbiont species: (**a**) NMAR, (**b**) SMAR, (**c**) Bh_sup05, (**d**) Bthe, (**e**) Bbro.

Chapter V | Preliminary results, concluding remarks and future directions

Life at hydrothermal vents and cold seeps is almost exclusively supported by chemosynthesis (Jannasch H. W., 1985). Most of the fauna in these habitats owe their success to their association with bacterial, chemosynthetic symbionts that provide their nutrition (Dubilier et al., 2008). *Bathymodiolus* mussels are one of the most dominating animal groups and extremely widespread in chemosynthetic habitats all around the world, including hydrothermal vents, cold seeps, whale and wood falls (Duperron, 2010). The diversity of symbionts in the *Bathymodiolus* symbiosis has been subject of study since their discovery. In 1995 it was shown that some *Bathymodiolus* species have evolved a dual symbiosis with two divergent 16S phylotypes – a SOX and a MOX symbiont – that coexist in single bacteriocytes (Distel et al., 1995). Since then up to 6 different symbiont species have been detected to associate with single *Bathymodiolus* species and individuals (Duperron et al., 2008). A few studies based on 16S sequences and the intergenic spacer region discovered more closely related phylotypes in single mussels of *B. puteoserpentis* and *B. azoricus* in both SOX and MOX symbionts (DeChaine and Cavanaugh, 2005; Won et al., 2003). This raised the suspicion that there is more hidden diversity in *Bathymodiolus* communities than previously thought. For a long time it was unknown whether and to what extent these phylotypes differed functionally from each other. And this is a key question, because this has potentially major consequences for the host and for the interaction among strains within the symbiont population. In fact, during the time of my studies Ikuta et al. (2016) discovered hidden diversity in the SOX symbiont in *B. septemdierum* hosts by using recent methodologies and sequencing technologies. However, in this study a single host individual was

sequenced, leaving the extend of intra-specific diversity in other *Bathymodiolus* species and potential implications for the symbiotic association elusive. The combination of high-resolution metagenomic and metatranscriptomic sequencing, as well as the development of computational workflows has a great potential of resolving such microdiversity. During my studies I developed a metagenomic approach to study intra-specific strain diversity in natural populations of *Bathymodiolus*, which can potentially also be applied to other systems. My results have revealed extensive intra-specific functional heterogeneity in the SOX symbiont and its relatives from the widespread SUP05 clade. Using polymophism data and nucleotide frequencies, we obtained insights into symbiont population structure that can help us understand key processes such as transmission, which are currently impossible to observe in nature or in the laboratory. My results have challenged current evolutionary theories and offer explanations how these can be extended to account for our observations of pervasive strain diversity in environmental deep-sea symbioses.

5.1 Microdiversity in *Bathymodiolus* symbionts

Investigating intra-specific diversity within bacterial populations poses an immense challenge when the bacteria cannot be cultured in the laboratory. This is also the case for all *Bathymodiolus* symbionts. But even for bacteria that have been cultured, it is unclear how much these cultures represent the diversity of populations in the environment. Both factors stress the need for cultivation-independent methods to investigate intra-specific diversity. Metagenomic sequencing has the potential to detect even single nucleotide changes within bacterial species. The major remaining challenge is to tease apart highly similar strains from metagenomes, to link single

nucleotides that stem from the same cell. For example, the high sequence similarity among strains of the same species often cannot be resolved during the assembly. This leads to the production of contigs that represent a consensus sequence of a strain mixture, thus masking the underlying sequence heterogeneity among single strains. Detecting this heterogeneity requires sophisticated approaches and workflows as those I developed for this study (chapter II). In addition, I took advantage of the low species diversity in *Bathymodiolus* symbionts with a maximum of six, but usually only one or two species of symbionts in a single animal. This low diversity allowed me to obtain the high sequencing coverage within symbiont species (e.g. > 100x in the SOX symbiont) that is needed to investigate strain-level diversity. The workflow I developed detected extensive genome-wide heterogeneity in the SOX symbiont, which was unexpected based on their homogeneity at the level of single marker genes (Duperron, 2010; Duperron et al., 2008, 2006). This heterogeneity was manifested in nucleotide polymorphisms and gene content variation. Using this approach I was able to detect up to 16 different SOX strains in single *Bathymodiolus* hosts and revealed functional heterogeneity among these (chapter II). Such a high number was also surprising as endosymbioses are usually expected to have low strain heterogeneity. This is due to genetic conflicts, that can arise when two strains compete for the same resource (Russel et al., 2017; Frank, 1996). Recent studies that used high-resolution sequencing technology have revealed some strain-level diversity in chemosynthetic endosymbionts of *Ridgeia piscesae* (Perez and Juniper, 2017) and *Riftia pachyptila* (Polzin et al., 2019; Robidart et al., 2008) tubeworms and *Solemya velum* clams (Russell and Cavanaugh, 2017). However, these numbers were one order of magnitude lower than those in *Bathymodiolus* mussels. Former studies, based on the 16S rRNA gene have shown the co-existence of up to nine ribotypes in Osedax worms (Verna et al., 2010; Goffredi et al., 2007) and up to 7 ribotypes in shipworm

symbionts (Luyten et al., 2006). However, the extent for genomic variation within these host species has not been shown. Thus, the number of 16 strains in *Bathymodiolus* is the highest reported so far to co-exist intracellularly within single animal hosts.

There are different possible explanations why no study has so far observed such a high number of co-existing strains. First, the limitations in sequencing technologies. Only in the recent years these have given us the resolution to tease apart strain-level diversity. Ellegaard and Engel (2019) have postulated that nucleotide diversity increased up to a sequencing depth of 1000x, a number that so far is rarely reached in metagenomic studies. Also in our study we observed that when increasing the sequencing coverage, we can detect a higher number of strains. Whereas at a sequencing depth of 100x we detected 9 co-existing symbiont strains in *B. puteoserpentis*, this number increased to 16 when we reached a sequencing coverage of 370x. Therefore, other endosymbiont populations may have a hidden micro-diversity that will be discovered as high-resolution studies increase. Second, there may be biological reasons. In endosymbioses the co-existence of closely-related strains can have drastic consequences. For example, competition between these strains for host resources can lead to the destabilization or potentially even breakdown of the symbiotic association (Frank, 1996). This can be caused by the emergence of cheater strains which get all the resources provided by the host but gives nothing or less in return. A cheater strain therefore has a competitive advantage over the mutualistic strain, as can be observed in rhizobial symbionts of legumes (Sachs et al., 2010a, 2010b). We offered one possible explanation under which circumstances high endosymbiont strain diversity may be tolerated (chapter II). Specifically, we suggested that when the symbionts' energy source comes from

the environment and not directly from the host there is less competition for host resources and therefore less opportunity to cheat (chapter II, see **5.4**). This would limit the endosymbioses with high strain diversity to those systems with external energy sources (e.g. chemosynthetic and photosynthetic symbioses).

5.1.1 How diverse is the MOX symbiont?

The extent of strain diversity was unknown for the SOX symbiont until recently. Similarly, it was unclear how heterogeneous the MOX symbiont is and how it compares to the SOX symbiont – i.e. whether it is characterized by a high or low intra-specific diversity. In chapter III we have shown, that within most MOX populations less than 5% of all protein-coding sequences were strain-specific. These results indicate that the MOX population is not clonal, and strains within a single host individual can differ in their gene content. However, the degree of this variation was vastly less that what we observed for the SOX symbiont in *Bathymodiolus* mussels (chapter II and III). Therefore, our results suggested a lower degree of overall variability within the MOX symbiont. Consistently, Picazo et al. (2019, contributed works) found lower nucleotide diversity in the MOX symbiont when it was compared to the SOX. This analysis was performed on a single host species, *B. brooksi*, from a single site, MC853, in the Gulf of Mexico (GoM). Here, I extended this analysis to three host species from seven sites and the preliminary results confirmed the low degree of heterogeneity (**Fig. 1**). Within MOX populations of single host individuals, genome-wide heterogeneity was mostly below 3 SNPs/kbp (for comparison, SOX symbionts had up to 13 SNPs/kbp). However, there were a few exceptions with > 5 SNPs/kbp in *B. brooksi* individuals from site MC853. Such a striking difference in the intra-specific diversity of MOX and SOX symbionts is remarkable and needs to be

interpreted in the light of multiple factors that define this symbiotic consortium: i) heterogeneity in the free-living symbiont population, ii) selective symbiont uptake and iii) restriction of symbiont uptake to short time window (physical bottleneck).

The first and most obvious reason that can explain a difference in the symbiont heterogeneity is the heterogeneity of the potentially colonizing symbiont population. Both SOX and MOX symbionts have been suggested to be horizontally transmitted (Fontanez and Cavanaugh, 2014; Wentrup et al., 2014; Won et al., 2003). This implies, that there has to be a free-living stage that colonizes juvenile host individuals. If the pool of free-living MOX bacteria is less heterogeneous than the SOX population, this would lead to a lower heterogeneity of MOX symbionts within the host as well. To determine this we need to have an estimate of the strain composition in the environmental MOX population. However, despite many years of research investigating the *Bathymodiolus* symbiosis, this is still unclear. Symbiont sequences have been found in environmental samples from water and biofilms but their genomic diversity remains to be shown (Fontanez and Cavanaugh, 2014; Crépeau et al., 2011).

Secondly, the host might be discriminating between strains that can colonize the gill tissue and those that cannot. Such processes have been described for example in *Euprymna scolopes* squids that are highly selective towards strains that can colonize their light organs (Nyholm and McFall-Ngai, 2003; Visick and McFall-Ngai, 2000). It is therefore possible that the host is selective towards just a few MOX strains that can successfully colonize the host. Instead, such discrimination might be more relaxed or absent towards different SOX strains, explaining the differences in heterogeneity.

Lastly, there could be a difference in the transmission mode between SOX and MOX symbionts. We have hypothesized that the SOX symbiont is acquired continuously from the environment (discussed further down). However, this time frame where host tissue can be colonized from an external population could be restricted in the MOX symbiont. A short time window for colonization from the external population would impose a physical bottleneck that can reduce population heterogeneity within the host. An example for such a bottleneck was shown in *Riftia* tubeworms where only a few symbiont cells can colonize host juveniles (Nussbaumer et al., 2006). If chance was to dictate which strain can colonize the juvenile host, this could also cause a difference in the strains between host individuals.

Interestingly, preliminary analyses of nucleotide polymorphisms in the MOX revealed a wider span in the degree of heterogeneity for *B. brooksi* from the GoM, compared to other host species (**Fig. 1**). This was similar to the SOX where I also observed a wide span in the degree of heterogeneity in host species *B. brooksi*, whereas the degree of heterogeneity was always similar among co-occurring hosts of other species (chapter II., III and IV).

In addition to measures of heterogeneity in the symbiont populations of individual hosts, the degree of similarity between symbiont populations of two host individuals can help to untangle symbiont transmission, which otherwise is virtually impossible to observe in nature (chapter II and IV). F_{ST} is a population genetic measure indicating how similar two populations are. Therefore I analyzed the F_{ST} between MOX populations of different host individuals from the same site. This analysis revealed that the MOX populations between individuals from a single site are very similar for host species *B. puteoserpentis* (vent site SEM, LG) and *B. heckerae* (seep

site CH) (**Fig. 1**). This is in line with our observations in the SOX symbiont that also showed a high similarity between the symbiont populations of individual mussels from the same site (chapter IV). Intriguingly, for *B. brooksi* (only site MC853 investigated for the MOX) the symbiont populations between host individuals from the same site differed substantially for both symbiont types (**Fig. 1** and chapter IV). This observation suggests that both symbionts are subject to similar population dynamics. These preliminary results also strengthen our observation for the SOX symbiont, indicating that *B. brooksi* is fundamentally different from the other host species, although the reasons for this difference are still unclear.

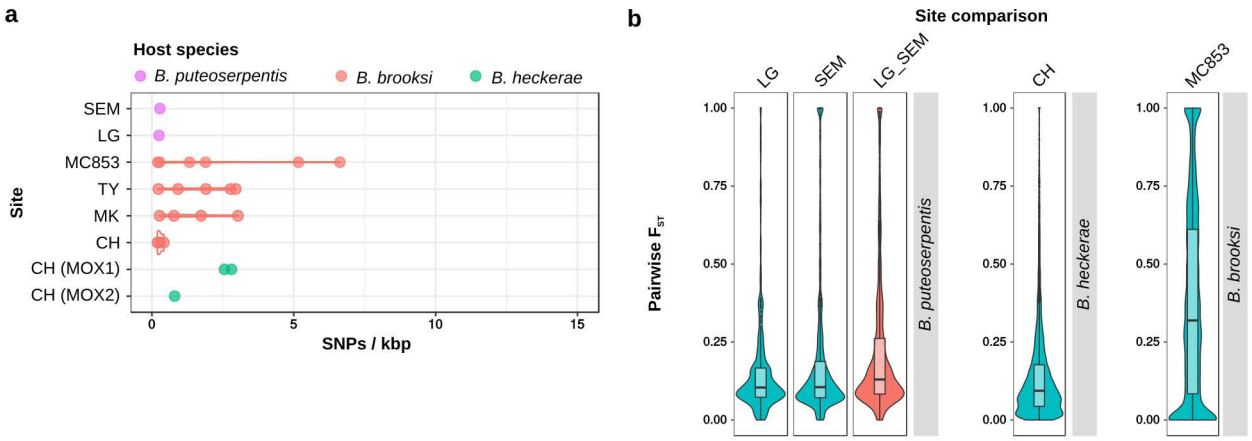


Figure 1 | Nucleotide variability in the MOX symbiont. a) Single nucleotide polymorphisms (SNPs) in MOX populations of single host individuals. MOX1 and MOX2 in *B. heckerae* are two different MOX phylotypes. Each data point corresponds to one host individual **b**) Pairwise F_{ST} per gene in the symbiont populations between host individuals from the same site (blue) or different sites (red) for the same host species (grey bars). The sites are indicated above the panels. The geographic locations of the sites are shown in **Fig. 2**. LG: Logatchev, SEM: Semenov, CH: Chapopote, MC853: site MC853.

5.1.2 What makes *B. brooksi* unusual?

In Chapter IV we discovered that *B. brooksi* had a much wider range of strain diversity compared to other *Bathymodiolus* species. Some explanations for such discrepancy have been outlined in Chapter IV. Here I expand on the possible reasons that were discussed and speculate about additional scenarios that could be considered. *B. brooksi* individuals differ from those of other host species in two ways:

i) The symbiont population heterogeneity in *B. brooksi* is extremely low for some host individuals. Instead, other individuals harbor a heterogeneous SOX population. This leads to a wide range of values in the degree of symbiont heterogeneity. ii) The similarity of symbiont populations is very low between some co-occurring hosts (chapter IV). Both is in stark contrast to what we observe for the other host species *B. azoricus*, *B. puteoserpentis*, *B.* sp. 1 from Clueless, *B.* sp. 2 from Lilliput and *B. heckerae* (**Fig. 2**). For these host species, the degree of symbiont heterogeneity is similar among individuals from the same site. In addition, the symbiont populations are very similar between the individuals, as shown by low F_{ST} values. This indicates that generally co-occurring host individuals share most of their SOX symbiont strains.

There are three factors that can possibly explain the difference between *B. brooksi* and the other host species: i) host age, ii) symbiont abundance in the environment and iii) host species. First, the extremely low symbiont diversity in some host individuals could by influenced by host age. As discussed in chapter IV and in Picazo et al. (2019, contributed works), all very old individuals with shell lengths above 17 cm investigated so far have low symbiont diversity (**Fig. 2**). *B. brooksi* individuals can grow very large and the largest individuals we collected were up to 25 cm in shell length. As pointed out in chapter IV, the growth rates for *B. brooksi* are unknown. However, *B. brevior* and *B. thermophilus* mussels were shown to slow down their growth when they reached a certain shell length (Nedoncelle et al., 2013; Schöne and Giere, 2005). Interestingly, for *B. thermophilus,* a reduction in growth rate starts at a shell length of around 16 cm. This host species reaches shell lengths of up to 18.4 cm (Nedoncelle et al., 2013; Rhoads et al., 1981) which is larger than *B. brevior* individuals but still smaller than maximum shell lengths that we observed in *B. brooksi*. Assuming that new gill tissue is continuously colonized by a mixture of self-

infection and environmental symbionts, this balance of symbiont source might become shifted when tissue growth slows down. For example, this could be due to the velocity of tissue growth. Internal self-infection might not be fast or efficient enough to colonize fast growing gill tissue. Instead, when SOX strains from the environment come in contact with young symbiont-free tissue, they can possibly colonize it. If gill tissue is produced at a much slower rates in old individuals, the chance of self-infection by the internal symbiont population could be higher than the colonization by a (possibly rare) strain from the environment. Slow growth therefore could lead to a decrease in the chance of environmental symbionts to 'meet' symbiont-free tissue. Such a process might also be influenced by the loss of microvili on bacteriocytes after being colonized by symbionts, possibly indicating that there is only a window of opportunity for the colonization (Wentrup et al., 2014). This would explain why there are no new strains coming in. However, this does not explain why symbiont heterogeneity drops in old individuals.

A decrease in symbiont diversity could be explained by the loss of symbiont strains. This could happen through internal competition among the bacteria. If strains from the environment can colonize new gill cells, symbiont strain loss through competition may be balanced by replenishment with external strains. A connection between changes in microbiota diversity and host age has been observed in other symbioses. For example, the gut microbiota of honey bees becomes unstable in old bees compared to young and mid-aged individuals (Ellegaard and Engel, 2019). Similarly, instability and reduced genetic diversity are characteristic for the gut microbiome in elderly humans (Nagpal et al., 2018; Salazar et al., 2017; Claesson et al., 2011). Therefore, I hypothesize that old age induces a change in the symbiont population, possibly due to inefficient acquisition of external strains in *B. brooksi*. Unfortunately,

the lack of other host species with shell lengths above 17 cm hampers our ability to tease apart the impact of host age on symbiont strain heterogeneity from other factors, such as the environment or host species.

A second factor that can possibly cause the difference between *B. brooksi* and other host species is the density of symbionts in the surrounding seawater. There is no evidence so far for a proliferating free-living population of the SOX symbiont (Fontanez and Cavanaugh, 2014). In fact, Ponnudurai et al. (2017) suggested that the metabolic capabilities including an incomplete TCA in the SOX genomes indicates that these symbionts depend on the host for survival. Thus the free-living stage of the SOX symbiont might be only a transition stage where the cells can survive but not proliferate. If this is the case, the uptake of symbionts highly depends on the density of mussel individuals at a vent or seep site as these would 'inoculate' the environment with symbiont strains. Following this line of thought, it might be possible that the mussel density explains the diversity patterns we see in *B. brooksi*. The host species *B. brooksi* occurs at many cold seep sites in the GoM. Our samples originated from one site in the northern GoM (MC853) and three sites in the southern GoM (CH: Chapopote, MK: Mictlan Knoll 2201 and TY: Tsanyao Yang Knoll 2223). A general feature that is common to the seep sites in this study is the low density and patchy distribution of host individuals (**Fig. 2**). This was in contrast to the hydrothermal vent sites at the Mid Atlantic Ridge (MAR) which typically had extensive mussel beds with many thousands of individuals (**Fig. 2**). We observed extremely low heterogeneity values of approx. 1 SNP/kbp only at two out of four seep sites – in 4 out of 5 individuals at site TY, and in 11 out of 21 individuals at site MC853. If there are only few host individuals that release symbionts into the environment this might shift the balance between self-infection and colonization of

young gill tissue from external strains towards the former. Instead, stochastic factors may dictate whether strains get exchanged among host individuals. This hypothesis is supported by the fact that the similarity of symbiont populations between *B. brooksi* individuals was lower than in all other species examined, indicating that not all strains get efficiently exchanged (chapter IV). In the case of extremely low strain heterogeneity in *B. brooksi* individuals at TY and MC853, it is unclear whether this is due to symbiont loss or a reduced colonization by environmental strains. Symbiont loss could potentially happen through competition among strains within a host or due to preferential digestion by the host. The latter would require the host to differentiate among strains or alternatively certain strains may have found more efficient ways to escape host digestion.

Finally, the difference in *B. brooksi* compared to other host species could be the result of differences in the biology of the host species that affect how they interact with symbionts. It is intriguing that in addition to having unusual symbiont population structure, phylogenetically, *B. brooksi* is quite divergent to the other host species included in this study (Lorion et al., 2013). We have hypothesized convergent evolution of different mechanisms that allowed the establishment of a symbiosis for different symbiont (and host) species (Chapter III, Sayavedra, 2016). Thus, a difference in host-symbiont interaction mechanisms could lead to a less efficient colonization of *B. brooksi* hosts by external symbionts. *B. brooksi* co-occurs with host species *B. heckerae* at site CH (**Fig. 2**). Unlike *B. brooksi*, *B. heckerae* had a narrow range in the degree of symbiont heterogeneity and a high symbiont population overlap between individuals. This supports the hypothesis that the host species can influence the processes of symbiont uptake and exchange.

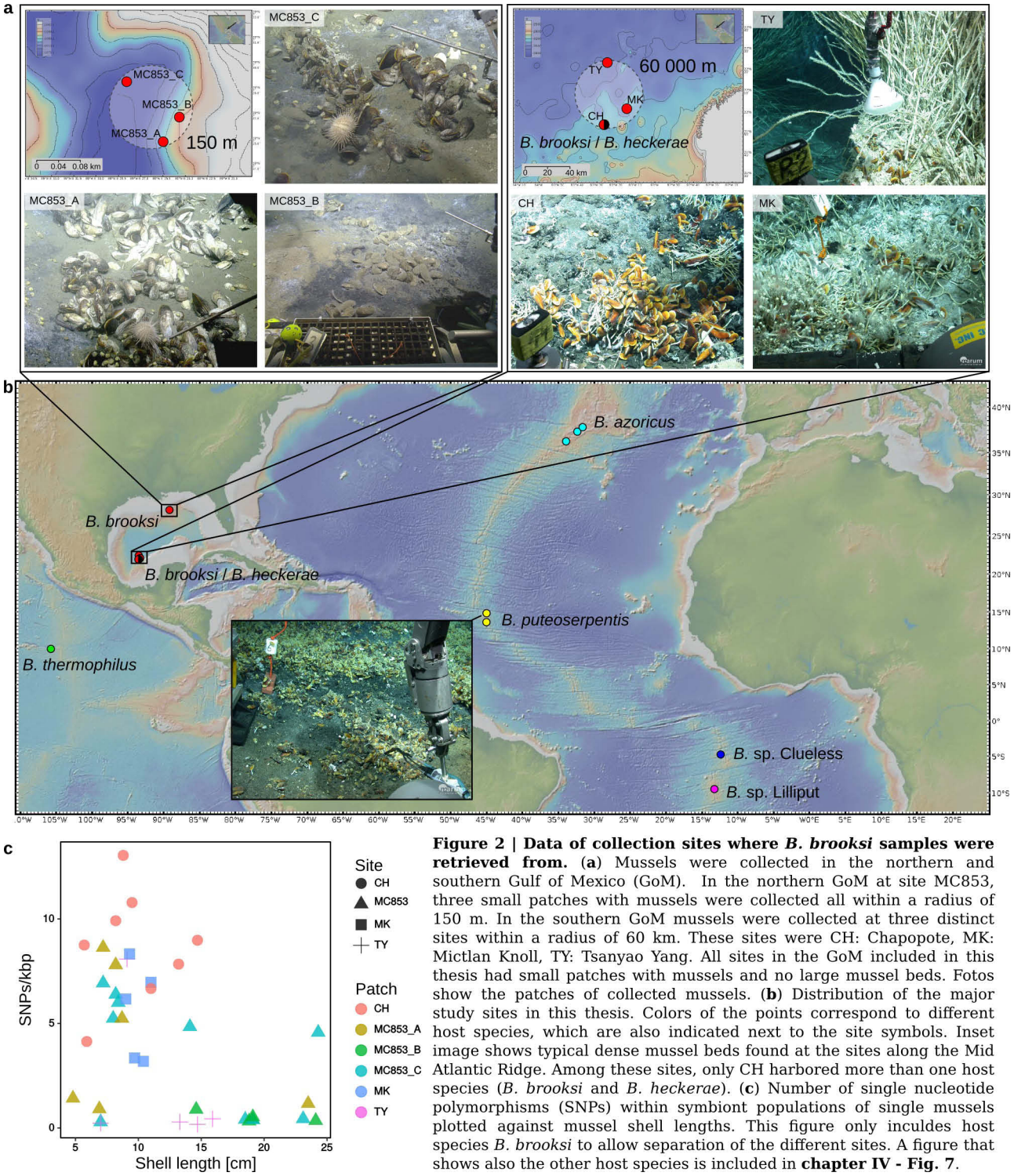


Figure 2 | Data of collection sites where *B. brooksi* samples were retrieved from. (**a**) Mussels were collected in the northern and southern Gulf of Mexico (GoM). In the northern GoM at site MC853, three small patches with mussels were collected all within a radius of 150 m. In the southern GoM mussels were collected at three distinct sites within a radius of 60 km. These sites were CH: Chapopote, MK: Mictlan Knoll, TY: Tsanyao Yang. All sites in the GoM included in this thesis had small patches with mussels and no large mussel beds. Fotos show the patches of collected mussels. (**b**) Distribution of the major study sites in this thesis. Colors of the points correspond to different host species, which are also indicated next to the site symbols. Inset image shows typical dense mussel beds found at the sites along the Mid Atlantic Ridge. Among these sites, only CH harbored more than one host species (*B. brooksi* and *B. heckerae*). (**c**) Number of single nucleotide polymorphisms (SNPs) within symbiont populations of single mussels plotted against mussel shell lengths. This figure only inculdes host species *B. brooksi* to allow separation of the different sites. A figure that shows also the other host species is included in **chapter IV - Fig. 7**.

In order to tease apart the above-mentioned possibilities I suggest to perform a systematic sequencing effort across a wide range of age cohorts in the host species

(including very young juveniles and very old individuals) and from sites with different densities in the mussel beds. Ideally this was performed for two or more host species, including *B. brooksi*. This implies a huge sampling effort, which might be difficult, considering the mussel's deep-sea habitat.

5.2. Functional heterogeneity in regulatory mechanisms

In chapter II we observed fundamental gene content differences in the high-affinity phosphate transport system PstSCAB and the two-component phosphate-dependent regulatory system PhoR-PhoB (Pho regulon, Santos-Beneit, 2015) between *Bathymodiolus* SOX symbiont species but also within populations of the same species and even within single host individuals (Ansorge et al., 2019, chapter II). We screened additional datasets, most of which were briefly discussed in chapter II (**Tab. 1**). We observed that these Pst-system and PhoR-PhoB were absent at three hyrothermal vent sites (Lilliput, Myonin Knoll, Crab Spa), corresponding to three different symbiont species (one at each vent site). In addition these systems were strain-specific at two vent sites at least. In contrast, Pst and PhoR-PhoB were absent from the clam symbionts and most free-living SOX lineages (**Tab. 1**). As pointed out in chapter II we could not find any other description of a bacterial species lacking both the Pst and PhoR-PhoB system. Our observations suggest that the SOX symbionts of *Bathymodiolus* vary substantially not only in their phosphorus acquisition mechanisms, but also in phosphate-dependent regulation of gene expression. In other bacteria this system has been described to be involved in regulating mechanisms that are potentially important in symbiosis, such as the production of secondary metabolites and virulence factors (Romano et al., 2015; Santos-Beneit, 2015; Lamarche et al., 2008).

Table 1 | Presence and absence of Pst genes, PhoR-PhoB, PhoU and a low-affinity transporter for phosphate in the SOX symbionts and other relatives from the *Thioglobaceae* family. Light grey indicates that these genes were strain-specific within symbiont populations, medium grey indicates that this gene was present and dark grey indicates that more than one copy of the gene were present.

Host or free-living	Symbiosis	Species	Site/ strain	PstS	PstC	PstA	PstB	PhoB	PhoR	Pho U	PhoU homolog	Low-affinity P-transporter	Library / Accession
B. azoricus	mussel	NMAR	LSET	1	1	1	1	1	1	1	1	present	1586B
B. azoricus	mussel	NMAR	LSMS	1	1	1	1	1	1	1	1	present	1586K
B. azoricus	mussel	NMAR	RB	1	1	1	1	1	1	1	1	present	1600G
B. azoricus	mussel	NMAR	MG	0	1	1	1	1	1	1	1	present	1586Q
B. puteoserpentis	mussel	NMAR	SEM	1	1	1	1	1	1	1	1	present	1115C
B. puteoserpentis	mussel	NMAR	LG	1	1	1	1	1	1	1	1	present	2487A
B. sp. Clueless	mussel	SMAR	CL	1	1	1	1	1	1	1	1	present	C112
B. sp. Lilliput	mussel	SMAR	LI	0	0	0	0	0	0	0	1	present	L102
B. sp. Wideawake	mussel	SMAR	WA	1	1	1	1	1	1	1	1	present	WA
B. brooksi	mussel	Bbro	TY	1	1	1	1	1	1	1	1	present	3411F
B. brooksi	mussel	Bbro	CH	1	1	1	1	1	1	1	1	present	3411K
B. brooksi	mussel	Bbro	MC853	1	1	1	1	1	1	1	1	present	1712M
B. brooksi	mussel	Bbro	MK	1	1	1	1	1	1	1	1	present	3411A
B. heckerae	mussel	Bh_sup05	CH	1	1	1	1	2	1	1	1	present	1600D_SUP05SOX
B. heckerae	mussel	Bh_sox2	CH	1	1	1	1	1	1	1	1	present	1600D_SOX2
*B. septemdierum**	mussel	Bsep	MyK	0	1	0	0	0	0	0	1	present	AP013042.1
B. sp. Cryptic GoM	mussel	Bcry_sox2	DC673	1	1	1	1	1	1	1	1	present	1875A
B. sp. Cryptic GoM	mussel	Bcry_sup05	DC673	1	1	1	1	1	1	1	1	present	1875D
B. thermophilus	mussel	Bthe	CS	0	0	0	0	0	0	0	1	present	1600T
*C. magnifica**	clam	Rmag	CM	0	0	0	0	0	0	0	1	present	CP000488
*C. magnifica**	clam	Rmag	UCDCM	0	0	0	0	0	0	0	1	present	JARW00000000.1
*C. okutanii**	clam	Voku	HA	0	0	0	0	0	0	0	1	present	AP009247
Ca. T. autotrophicus*	Free-living	Tauto	EF1	0	0	0	0	0	0	0	1	present	NZ_CP010552.1
Ca. T. singularis*	Free-living	Tsing	GG2	0	0	0	0	0	0	0	1	present	CP008725.1
Ca. T. singularis*	Free-living	Tsing	PS1	0	0	0	0	0	0	0	1	present	CP006911.1
Ca. T. sp. strain*	Free-living	Tsp	MED612	1	1	1	1	0	1	1	1	present	2026721.3
Ca. T. sp.*	Free-living	Tsp	MED218	0	0	0	0	0	0	0	1	present	1986874.3
SUP05*	Free-living	Tsp	SUP05	0	0	0	0	0	0	0	1	present	655186.3
SUP05	Free-living	Tperd	C2484	0	0	0	0	0	0	0	1	present	C2484
SUP05	Free-living	Tperd	C2488	0	0	0	0	0	0	0	1	present	C2488

* Publications describing these genomes are (from top to bottom): Ikuta et al., 2015, Newton et al., 2007, Lee et al., 2014, Kuwahara et al., 2007, Shah and Morris, 2015, Marshall and Morris, 2015, Callbeck et al., 2018

The variability we observed in the Pho regulon is likely linked to variability in intergenic regions as well. Genes that are regulated by the Pho regulon are characterized by the presence of specific DNA sequences in the non-coding promoter region, called the 'PHO box'. If the response regulator PhoB is phosphorylated, which is the case for example during phosphate limitation, it binds to the PHO box and activates or represses the transcription of a particular gene (Santos-Beneit, 2015). The DNA sequences of these PHO boxes thus determine whether PhoB can bind. Therefore, we would expect nucleotide variation in the PHO box to have a fundamental effect on the gene expression in the SOX symbionts. Preliminary data indicate that the number of polymorphisms differs between inter-genic regions in the SOX symbiont (data not shown). However, to what extent promoter binding sites are affected, needs to be investigated. The PHO box often consists of two 11-nucleotide direct repeat units with seven well conserved nucleotides within bacterial species (Santos-Beneit, 2015). Between species, these sequences are less conserved which poses a challenge to the *de novo* detection of PHO boxes in bacterial genomes. One approach could be to create a database with all known PHO box sequences and blast these against the genomes of interest. In addition, promoter regions of genes that are likely regulated by PhoR-PhoB, such as phosphate-acquisition mechanisms, could be screened for direct repeat sequences. If this is successful, these potential PHO boxes could be blasted against the rest of the genome to detect other genes that may be regulated by this mechanism.

Considering the importance of inter-genic regions for cellular processes such as gene regulation, the question arises how natural selection affects these. Especially, for an endosymbiont with horizontal transmission mode such as the SOX symbiont, the regulation of gene expression may be essential to survive in very distinct

environments inside and outside of the host. Fluctuation in the availability of resources, as can be observed at hydrothermal vents (Zielinski et al., 2011) may impose an additional requirement for precise regulation of the symbiont metabolism. In fact, despite often considered neutral (Wang and Chen, 2013; Hu et al., 2006), bacterial inter-genic regions have recently been described to experience strong purifying selection and potentially even positive selection (Thorpe et al., 2017). The investigation of allele signatures in promotor binding sites, such as the PHO Box, could therefore help to identify evolutionary signatures in the regulatory machinery of endosymbionts. Population genetic concepts for the detection for selection could be one approach to identify inter-genic loci that have been subject to adaptive or purifying selection (see chapter IV, Hohenlohe et al., 2010).

5.3 Mediators of genomic plasticity

In Chapters II, III and IV we analyzed the striking plasticity in the genomes of SUP05 symbiont strains. We hypothesized that the variability among closely related strains can be explained by frequent horizontal acquisition and loss of genes and gene clusters, as well as genome rearrangements. Horizontal acquisition of genes needed for using additional energy sources as well as carbon fixation has been suggested before for the SOX symbionts of *Bathymodiolius* (Kleiner et al., 2012). Sayavedra et al. (*in prep.,* contributed works) showed that the *Bathymodiolus* symbionts are enriched in mobile genetic elements (MGEs) such as transposases and restriction-modification (RM) systems, compared to their free-living relatives and the related vertically transmitted clam symbionts. RM systems have been described to be have a variety of functions, many of which increase genomic variation (Vasu and Nagaraja, 2013). These functions include genome rearrangements, horizontal gene transfer and

phage defense. Therefore, the enrichment of RM systems, as often observed in naturally competent bacteria, could indicate an increased uptake of foreign DNA that needs to be controlled to avoid harmful effects on the cells (Vasu and Nagaraja, 2013).

5.3.1 Potential functions of CRISPR-Cas in the SOX symbionts

I discovered that the SOX strains symbiotic to *Bathymodiolus* mussels encoded Cas genes that are associated with clustered regularly interspaced short palindromic repeats (CRISPR). CRISPR-Cas systems have been described as an ‘adaptive immune system’ that protects prokaryotes from viruses and other sources of mobile genetic elements, such as plasmids (van der Oost et al., 2014; Makarova et al., 2011; Barrangou et al., 2007). The palindromic repeats are separated by short sequences called ‘spacers’ which represent a memory of foreign DNA encountered during past infections. Typically, the CRISPR-Cas interference happens in three stages. In the first stage, termed adaption, the CRISPR array expands by incorporation of another spacer sequence along with another repeat sequence. This stage is followed by the expression stage, where the Cas genes are transcribed to form the CRISPR RNAs (crRNAs) that each include a spacer sequence as target recognition. Finally, during the interference stage, crRNAs bind to the complementary target sequences and thus guide the Cas proteins towards the correct target sequence which then gets degraded (Westra et al., 2014; Yosef et al., 2012). My studies revealed that CRISPR-Cas systems and/or arrays were encoded in all *Bathymodiolus* and sponge SOX species, except for species Bh_SUP05. Despite this, the distribution of CRISPR-Cas systems was highly variable at the strain level and between sampling sites. For example, at site Menez Gwen not a single SOX symbiont strain encoded any CRISPR-

related feature. In addition, the representation of Cas-genes varied considerably among co-existing strains within single host individuals (Ansorge et al., 2019) (chapter II). Intriguingly, I found that the free-living relatives and vertically transmitted clam symbionts did not encode any CRISPR-Cas gene or array (chapter III, **Tab. 2**). This raises the intriguing possibility of a connection between the CRISPR-Cas system and the host-associated lifestyle with horizontal transmission mode. This was especially surprising as the free-living relatives were shown to be heavily infected by phages (Anantharaman et al., 2014; Roux et al., 2014) and nevertheless lack any CRISPR-Cas defense system. Apart from protection against foreign DNA, some types of CRISPR-Cas systems were shown to be involved in other processes. To investigate this, I identified the types of CRISPR-Cas systems and discovered that all three classified types I-III were present in the *Bathymodiolus* symbionts (**Tab. 2**). Surprisingly, CRISPR-Cas system type II was present in 45% of the 94 *Bathymodiolus* SOX draft genomes analyzed, and in the three sponge SOX as well and was therefore the most common type. Type II CRISPR-Cas is the rarest of all CRISPR-Cas systems and only present in 5% of all bacteria (Chylinski et al., 2014; Makarova et al., 2011). Interestingly, type II CRISPR-Cas are over-represented in pathogens and commensal bacteria (Chylinski et al., 2014), and the presence of this system has been described to be linked to virulence in *Campylobacter jejuni* (Louwen et al., 2013), *Legionella pneumophila* (Gunderson and Cianciotto, 2013), *Neisseria meningitidis* and *Francisella novicida* (Sampson et al., 2013). The latter, *F. novicida* has been shown to target its own mRNA to downregulate the expression of surface lipoproteins which allows the pathogen to invade, remain undetected in human epithelial cells and thus evade immune response. Like other mechanisms that were first detected in bacterial pathogens, this system may not be limited to pathogenic bacteria but also allow mutualistic symbionts to invade and persist in eukaryotic

cells. Considering that *Bathymodiolus* symbionts occur in epithelial gill tissue, one can imagine a similar mechanism as the one described for *F. novicida*. This would also explain the absence of these systems in the free-living relatives as well as the clam symbionts. The latter do not need to invade host tissue from the outside but instead are transmitted through the germline (Endow and Ohta, 1990).

Table 2 | Distribution of CRISPR-spacers and Cas genes and the identified types of CRISPR-Cas systems among SOX symbionts and relatives from the *Thioglobaceae* family. Grey markup indicates when a feature was present.

SUP05 species	site	CRISPR-positive fraction	# Cas genes	# spacers	Type II	Type I-F	Type III-B	# of genomes	# of positive genomes
Bbro	MC853	1.0	3-12	27-261	12	18	0	22	22
Bbro	CH	1.0	3-4	0-57	5	0	0	7	7
Bbro	MK	1.0	2-6	32-114	2	0	0	5	5
Bbro	TY	1.0	2-3	32-56	1	1	0	5	5
SOX2	CH	1.0	0	27-261	0	0	0	3	3
SOX2	DC673	1.0	1	33	0	0	0	1	1
Bcry_SUP05	DC673	1.0	5	34	1	0	0	1	1
Bh_SUP05	CH	0.0	0	0	0	0	0	6	0
Bsep	MyK	1.0	3	0	1	0	0	1	1
Bthe	CS	1.0	5-14	15-195	1	3	2	3	3
NMAR	SEM	0.8	0-5	0-17	5	1	0	6	5
NMAR	LG	0.7	0-2	0	2	0	0	3	2
NMAR	LSET	1.0	2-8	0-53	4	8	0	8	8
NMAR	LSMS	1.0	2-5	0-42	1	3	0	3	3
NMAR	MG	0.0	0	0	0	0	0	4	0
NMAR	RB	0.2	0	0-2	0	0	0	5	1
SMAR	CL	1.0	1-6	0-152	3	0	0	5	5
SMAR	LI	1.0	0-5	0-2	3	2	0	5	5
SMAR	WA	1.0	11	506	1	1	0	1	1
Sponge sox 1	CH	1.0	3-9	65-144	2	0	0	2	2
Sponge sox 2	MK	1.0	2	54	1	0	0	1	1
Ca. T. singularis 1		0.0	0	0	0	0	0	2	0
Ca. T. singularis 2		0.0	0	0	0	0	0	1	0
Ca. T. autotrophicus		0.0	0	0	0	0	0	1	0
Ca. T. perditus	Peru OMZ	0.0	0	0	0	0	0	2	0

The other common CRISPR-Cas system in *Bathymodiolus* was type I, subtype I-F. In addition to defense against foreign DNA, this system has been shown to affect gene regulation of group behavior, such as biofilm formation in *Pseudomonas ariguinosa* (Westra et al., 2014; Cady and O'Toole, 2011). Possibly, similar mechanisms could contribute to a differential regulation of genes among *Bathymodiolus* symbiont strains.

Without experimental evidence, the role of CRISPR-Cas in the *Bathymodiolus* symbionts remains unclear. However, with a more thorough mining of the available metagenomes we may be able to identify additional annotations of e.g. small CRISPR-associated RNAs (scaRNAs) which would help to narrow down the functional role of the CRISPR-Cas system. In addition, a comparison of all identified spacer sequences against i) virus sequence databases (e.g. IMG/VR), ii) general sequence databases (e.g. NCBI) or iii) against the symbiont genomes could aid the identification of target sequences and thus whether CRISPR-Cas systems in the SUP05 symbionts target self- or foreign DNA or both.

5.4 Why be diverse?

The central thread that runs through my doctoral studies always comes back to the extensive strain-level diversity in the SOX symbiont of *Bathymodiolus* mussels. In chapter II we have suggested that symbiotic communities, fed by environmental resources and not directly by the host are more permissive to symbiont strain-diversity. Yet, there are few systems that have been observed to have strain diversity as extensive as the *Bathymodiolus* SOX symbionts. In other chemosynthetic symbioses, e.g. in *Riftia* tubeworms and *Solemya* clams, high-resolution metagenomics has revealed strain diversity. But the level of heterogeneity was much lower than in *Bathymodiolus* SOX symbionts (Polzin et al., 2019; Perez and Juniper, 2017; Russell and Cavanaugh, 2017). In addition, these symbiont communities were usually dominated by a single strain, which does not seem to be the case in *Bathymodiolus*. Similar observations were made in honey bee gut and human gut, using high-resolution sequencing methods. These studies reported that within individuals single strains of the same species are either dominant or rare, which can

possibly be explained by competitive exclusion (Ellegaard and Engel, 2019; Truong et al., 2017; Schloissnig et al., 2013). This raises the question: is the high strain diversity between *Bathymodiolus* mussels the exception? And if so, how can this be explained? In symbiotic associations between bacteria and animal hosts there are two main factors that influence the symbiont strain diversity and composition: i) colonization and ii) persistence.

Strain diversity can only occur, if multiple symbiont strains are able to colonize a host individual. Here, the symbiont transmission mode has a strong impact. A strict vertical transmission mode usually imposes a strong physical bottleneck on the symbiont with each host generation. Often, only a few cells are passed on, which consequently leads to a drastic reduction in the number of strains that are passed on to the next host generation. This was for example described in *Buchnera* symbionts of aphids (Mira and Moran, 2002). Therefore, symbioses that undergo strict vertical transmission are expected to have restricted strain diversity. In the horizontal transmission mode there are different possibilities. On the one hand there can be physical bottlenecks similar to those observed during vertical transmission. This is the case if the symbiont transmission happens during a short time window in juvenile hosts. One example can be found in *Riftia* tubeworms. Here, only a few cells can colonize juvenile hosts, which implies that only a few strains can colonize the host tissue (Nussbaumer et al., 2006). In addition, the host can be involved in the reduction of strains during the colonization process. Such selection mechanisms were described in symbioses such in squids colonized by *Vibrio fischeri* (Bright and Bulgheresi, 2010; Nyholm and McFall-Ngai, 2003; Visick and McFall-Ngai, 2000). In addition, the founder effect plays a role which implies whoever is the first to occupy the tissue will persist, whereas other potential competitors are not able to invade this

habitat or niche. On the other hand, horizontal transmission can lead to the uptake of many different strains. This is the case if the colonization period is extended or even continuous throughout the lifetime of the host. In chapter II we have suggested such a colonization process for the SOX symbionts in *Bathymodiolus*.

The second factor that influences strain diversity is the stable co-existence of multiple strains. Here, a fundamental theoretical concept is important to consider: if two bacteria share the same resource, thus have overlapping ecological niches, they will compete. Competition ultimately leads to one strain out-competing the other, physical partitioning or niche differentiation (Russel et al., 2017; Ghoul and Mitri, 2016). Generally, the more genetically similar two organisms are, the less likely they can co-exist, because closely-related organisms are more likely to have the same resource requirements. This implies that two divergent species are more likely to co-exist than two highly similar strains of the same species. In symbioses this is one possible explanation why often many different species are found in the same host (e.g. the human microbiome). Instead, within a bacterial species these microbiomes are often dominated by one or few strains (Truong et al., 2017; Schloissnig et al., 2013). Competitive exclusion among closely-related strains were also mentioned as possible explanations leading to the individualized profiles of strains in the gut of individual honey bees (Ellegaard and Engel, 2019).

So how can two closely related strains of the same species stably co-exist? The options that could allow multiple strains to co-exist in a host are i) reduced competition, ii) niche-partitioning, iii) physical separation, or a combination of these. In chapter II we proposed that the non-limiting availability of the shared energy resource can reduce competition among the co-existing strains. This is for example

the case for reduced sulfur compounds. For limiting energy substrates, niche-partitioning could explain the emergence and co-existence of strains. One example is the capability to oxidize hydrogen. While all strains in *Bathymodiolus* hosts encode the machinery for hydrogen oxidation at one site (e.g. Lilliput), this capability is only encoded in a minor part of the strain population at another site (e.g. Clueless). This was also reflected in a higher hydrogen concentration at Lilliput in contrast to extremely low concentrations at Clueless. Consequently, when hydrogen is limiting only some strains can occupy the niche of hydrogen oxidation. And finally, symbiont strains could be physically separated in distinct bacteriocytes. This would prevent direct competition among strains within the same compartment and thus could allow multiple strains to co-exist within one individual (see **5.5.1**).

All these factors that allow persistence of multiple strains would likely also apply to other chemosynthetic symbioses such as in *Riftia* tubeworms, that acquire their symbionts horizontally; and *Solemya* clams that acquire their symbionts in a mixed mode of vertical and horizontal transmission. For *Riftia,* we know that different individuals harbor different dominant symbiont strains. This implies that the entire strain diversity at a single site is larger than the strain diversity within single individuals. And yet within single hosts there is usually only one dominant strain (Polzin et al., 2019; Perez and Juniper, 2017). However, the scarcity of high-resolution studies beyond a few marker genes makes it challenging to compare (chemosynthetic) symbioses and to pinpoint the factors that allow or prohibit strain co-existence in these. Below I speculate about the combination of factors in *Bathymodiolus* that I consider necessary to explain the patterns we observed.

First, *Bathymodiolus* acquire their symbionts from the environment. We have proposed that the evenness we observe between host individuals, with the exception of *B. brooksi* (see **5.1.2**) can only be explained if the host individuals exchange symbionts throughout their lifetime or at least over an extended period (chapter II). There might be processes within the host that could lead to the loss of symbiont strains. This could happen through competition among the symbionts, preferential digestion of specific strains, phage predation against a specific strain or by chance. However, if symbiont exchange is efficient, symbiont strains that get lost could be continuously replenished. This could be different in *Riftia* and *Solemya* symbionts, which are assumed not to be permissive to take up symbionts from the environment during their lifetime.

Second, there might be something inherent to the symbiont itself that makes the symbiosis special and high strain heterogeneity stable. In the *Bathymodiolus* symbiosis the mussel is often colonized by two primary symbiont types – the SOX and MOX symbiont. While we see this extensive strain diversity in the SOX, our preliminary data showed much lower diversity in the MOX symbiont. In chapter III and IV we suggested that evolvability could be a trait inherent to SUP05 bacteria, including the *Bathymodiolus* SOX symbionts. We suggested that evolvability could be increased by frequent events of horizontal gene transfer (HGT) and gene loss, genome rearrangements and high nucleotide variability in the SOX symbiont. This might allow a flexibility to fill the available niche space, possibly consisting of many different micro-niches. Jigsaw pieces of strains thus end up to build a functioning meta-organism with subdivided tasks. The lower heterogeneity in the MOX symbiont suggests that it is more static in its genomic features than the SOX. I suggest a targeted comparison of the degree of rearrangements, gene synteny and signatures

of HGT including phages between both symbionts. Based on my preliminary findings of low heterogeneity and gene content variation in the MOX symbiont, I would predict that we identify less rearrangements, HGT signatures, etc. in MOX symbionts, resulting in a more static genome composition than the SOX.

In summary, the strain diversity and evenness across *Bathymodiolus* symbiont populations appear to be an exception among symbioses that have been looked at in such high-resolution, also compared to human and bee gut microbiomes. Our theory that symbioses with environmentally-derived symbiont resources are more permissive towards higher strain diversity should hold true if also the other factors described above are given. The amount of high-resolution data that are now generated every year will probably soon provide additional examples of symbiont strain diversity in nature. These will help to refine these theories even further and shed more light on how symbiont strains can co-exist in animal-microbe symbioses.

5.4.1 The evolution of 'being ready'

I have been intrigued by the fact that SOX and MOX symbiont appear to have taken different evolutionary roads in this intimate association with *Bathymodiolus* hosts. I asked myself: how does evolution come up with a similar result – the intracellular association with deep-sea mussels – but with two different solutions? In chapters III and IV I suggested that in the SOX symbiont, and close relatives from the SUP05 clade, the trait evolvability might have been selected for. Evolvability refers to the cryptic genetic capacity of a population that can lead to adaptive phenotypes when the environment changes (Payne and Wagner, 2019). This means that an increased evolvability leads to the potential of a species to adapt to many different conditions

that might arise. For example, the extensive environmental abundance of antibiotics has been suggested to select for increased evolvability in bacteria which consequently leads to the even faster emergence of pathogenic strains with antibiotic resistence (Gillings and Stokes, 2012). The foundation for evolvability is diversity, which can be generated by higher mutation rates, recombination, and gene duplication, acquisition and loss. But how is the generation of variation, or the state of 'being ready' for any change, selected for? In a stable environment, generation of variability can be costly. Therefore the persistence of heterogeneity in a population needs a heterogeneous environment. Rainey and Cooper (2004) suggested that only with ecological opportunity, e.g. an additional energy substrate in the environment, the acquisition of novel genes or traits can be an advantage. In a homogeneous environment such variation would most likely be too costly and therefore purged by purifying selection. Thus, an increased evolvability needs both: ecological opportunity and genomic innovation. A support for this theory is that e.g. hyper-mutating strains occasionally develop in populations under stress whereas this rarely happens under stable conditions (Gillings and Stokes, 2012). Also rates of horizontal gene acquisition and recombinations can increase under stressful conditions (Gillings and Stokes, 2012). Therefore a heterogeneous environment and a decrease in the cost of acquiring foreign DNA can select for bacterial evolvability.

Interestingly, Rainey and Cooper (2004) suggested that the selection for duplication of regulatory features and their modularity can decrease the cost of acquiring foreign DNA. This is because the foreign DNA likely requires different regulatory cues and cascades. This is less deleterious to an organism when there is a duplicated version that can adapt whereas its twin can continue to regulate the inherent functions. Our observations of strain variability in regulatory elements of the SOX symbiont that

seemed surprising at first could thus make sense in the light of selection for modularity in regulatory components.

I could imagine that early in their evolutionary history, the SOX and MOX symbionts encountered very different circumstances. Whereas the MOX clade may have been evolving under relatively stable conditions with little ecological opportunity, the SOX might have been evolving under selective pressure towards evolvability. A decrease of costs in the acquisition of foreign DNA and heterogeneous conditions may have set the basis for the SOX clade to be highly evolvable and 'ready' for unknown environments. This remains speculative and provides hypothesis for future directions with a focus on regulatory components in the symbionts of *Bathymodiolus*.

5.5 Future directions

5.5.1 Sharing the compartment

One limitation of metagenomic analyses is the loss of spatial resolution. My studies could not reveal whether two functionally different strains can co-exist within a single bacteriocyte. However, this is a key question when we want to understand how such high strain diversity can be stable in single hosts. If two different strains share the same compartment, e.g. bacteriocyte, these will be competing for resources such as space within the host cell, and nutrients required by all strains such as reduced sulfur compounds. On the other hand, there might be strains that benefit from a direct interaction by the exchange of metabolic intermediates in pathways that are variable such as denitrification (Ansorge et al., 2019; Lilja and Johnson, 2016) (chapter II). Ongoing studies have visualized a gene encoding the enzyme methanol

dehydrogenase (MDH) in gill tissue of *Bathymodiolus* spp. This gene was strain specific in the SOX population and the images so far are inconclusive regarding the question whether all strains in a bacteriocyte encode this gene Sayavedra et al. (*in prep.*). We are currently working on creating a second set of probes for fluorescence in situ hybridization of the hydrogenase gene. This gene, encoding the capability of using hydrogen as an additional energy source, was also detected to be strain-specific in the SOX symbiont at multiple vent sites (Ansorge et al., 2019; Ikuta et al., 2016) (chapter II). The distribution of this gene in the gill tissue of *Bathymodiolus* mussels may potentially reveal that strains with different energy metabolisms can co-exist in single bacteriocytes. In addition, such image-based approaches also can help to understand fine-scale processes, such as the distribution of symbionts within host tissue which is still not fully understood. An additional approach to understand the diversity within single bacteriocytes is to perform single-cell sequencing on these cells. While our efforts of performing single-cell sequencing on the symbiont cells have been unsuccessful so far, sequencing single bacteriocytes may be more feasible. This has the potential to reveal which symbiont strains occur in direct contact to each other and can help to explain fine-scale population structure and its importance for co-existence. Finally, mass spectrometry imaging has the potential to reveal metabolic differences among co-existing strains in the gill tissue by the detection of metabolites or peptides (Watrous and Dorrestein, 2011). In addition to the spatial information this has great advantage of detecting phenotypic variation (see next paragraph).

5.5.2 From genotype to phenotype

In the present work most of the results shed light on the genotypic variation that is present in symbiont populations. Although some of our analyses also included transcriptomic data revealing that gene content variation among co-existing strains affect genes that are expressed and thus should be 'visible' also in the phenotype of these strains. Ultimately, selection can only act if a genotypic variation leads to a phenotypic change. Some of the variation we observe might lead to the same phenotype and may only be important when environmental change occurs. An identical phenotype outcome from different genotypes has been described as phenotypic robustness (Wagner, 2008). Given a robust phenotype, the underlying genotypic variation can therefore support evolvability in the population (see **5.4**). For example nonsynonymous polymorphisms that lead to amino acid substitutions in the protein have the potential to cause massive to no changes in protein function (chapter IV). We have shown that up to 15% of the nonynonymous substitutions affect the STOP codon and hence the length of the protein. Likely this has a huge impact on protein function, and shotgun proteomics would be one approach to investigate the phenotypic effects. However, if this protein is encoded in multiple copies the phenotype may be only slightly affected. Phenotypes also strongly depend on gene regulation that allows bacteria to react to different environmental cues. Teasing apart the regulatory components, their variability and potential signatures of directional selection may therefore help to understand some of the impact on phenotypic variation (see **5.2.**). This complexity makes it extremely difficult to predict the phenotypic effects of genotypic variation. Adding to this complexity, also identical genotypes can show different phenotypes leading to cell individualities (Nikolic et al., 2017; Ackermann, 2013). It is an intriguing thought that phenotypically, SOX and

MOX symbionts may have similar levels of population heterogeneity, which is invisible at the genotype level. The investigation of phenotypic variation in the symbiont population with additional methods is important to shed light on the above-mentioned questions and possibly to test some of the predictions that arise from the observed genotypic variations. Imaging-based methods such as mRNA-FISH (Coleman et al., 2007) or mass spectrometry imaging, detecting metabolites and peptides (Phelan et al., 2011) are therefore promising methods that can help to understand phenotypic variation in symbiont populations in their spatial context (Geier et al., 2019).

5.5.3 Catch me if you can

In order to fully understand the transmission and colonization of *Bathymodiolus* symbionts it is essential to compare the host-associated population with the free-living stage. Marker sequences of *Bathymodiolus* symbionts have been found in the environment, however it is still unclear whether these can proliferate or originate from mussel release waiting passively for their chance to colonize the next host (Fontanez and Cavanaugh, 2014; Crépeau et al., 2011). Metagenomics, metatranscriptomics and metaproteomics of the free-living symbiont population are needed to understand their heterogeneity and activity outside of the host. This will help to understand whether the entire free-living pool of symbionts can colonize single host individuals or only a subset of these. The latter could be due to selection mechanisms by the host, symbiont competition or founder effects. Therefore, the capture and comparison of free-living to host-associated populations, surely is an essential next step. This is especially important to fully understand symbiont transmission and to explain the differences we observed in *B. brooksi* (**5.1.2**).

5.6 Concluding remarks

The data presented in my study for the first time provide an overview of the intra-specific diversity in endosymbionts of multiple *Bathymodiolus* species. My observations in these natural associations have challenged theoretical concepts underlining the value of studying symbiotic communities in their natural context for the extension of evolutionary models. Theory, hypotheses and observations in nature go hand in hand if we want to understand the evolution and stability of intimate symbioses. Therefore it is essential to continue down this path of high-resolution meta-omic analyses and to test concrete hypotheses in laboratory experiments. Most analyses in this thesis were based on genotype variation of the SOX symbiont. In future studies these should be extended to phenotype variation in not only the SOX but also the MOX symbiont. This will help us to understand how the interaction between species connects to the interaction between single strains. Therefore, the integration of imaging technologies offering spatial resolution and a closer integration with environmental data will enable us to understand the *Bathymodiolus* endosymbiosis and possibly explain observations in other systems as well.

References for chapter 5

Ackermann, M. (2013). Microbial individuality in the natural environment. ISME J. *7*, 465–467.

Anantharaman, K., Duhaime, M.B., Breier, J.A., Wendt, K.A., Toner, B.M., and Dick, G.J. (2014). Sulfur oxidation genes in diverse deep-sea viruses. Science *344*, 757–760.

Ansorge, R., Romano, S., Sayavedra, L., Kupczok, A., Tegetmeyer, H.E., Dubilier, N., and Petersen, J. (2019). Diversity matters: Deep-sea mussels harbor multiple symbiont strains. BioRxiv 531459.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science *315*, 1709–1712.

Bright, M., and Bulgheresi, S. (2010). A complex journey: transmission of microbial symbionts. Nat. Rev. Microbiol. *8*, 218–230.

Cady, K.C., and O'Toole, G.A. (2011). Non-identity-mediated CRISPR-bacteriophage interaction mediated via the Csy and Cas3 proteins. J. Bacteriol. *193*, 3433–3445.

Chylinski, K., Makarova, K.S., Charpentier, E., and Koonin, E.V. (2014). Classification and evolution of type II CRISPR-Cas systems. Nucleic Acids Res. *42*, 6091–6105.

Claesson, M.J., Cusack, S., O'Sullivan, O., Greene-Diniz, R., de Weerd, H., Flannery, E., Marchesi, J.R., Falush, D., Dinan, T., Fitzgerald, G., et al. (2011). Composition, variability, and temporal stability of the intestinal microbiota of the elderly. Proc. Natl. Acad. Sci. U. S. A. *1*, 4586–4591.

Coleman, J.R., Culley, D.E., Chrisler, W.B., and Brockman, F.J. (2007). mRNA-targeted fluorescent in situ hybridization (FISH) of Gram-negative bacteria without template amplification or tyramide signal amplification. J. Microbiol. Methods *71*, 246–255.

Crépeau, V., Cambon Bonavita, M.-A., Lesongeur, F., Randrianalivelo, H., Sarradin, P.-M., Sarrazin, J., and Godfroy, A. (2011). Diversity and function in microbial mats from the Lucky Strike hydrothermal vent field. FEMS Microbiol. Ecol. *76*, 524–540.

DeChaine, E.G., and Cavanaugh, C.M. (2005). Symbioses of methanotrophs and deep-sea mussels (Mytilidae: Bathymodiolinae). In *Molecular basis of symbiosis*, P.D.J. Overmann, ed. (Springer Berlin Heidelberg), pp. 227–249.

Distel, D.L., Lee, H.K., and Cavanaugh, C.M. (1995). Intracellular coexistence of methano- and thioautotrophic bacteria in a hydrothermal vent mussel. Proc. Natl. Acad. Sci. U. S. A. *92*, 9598–9602.

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

Duperron, S. (2010). The diversity of deep-sea mussels and their bacterial symbioses. In *The vent and seep biota*, S. Kiel, ed. (Springer Netherlands), pp. 137–167.

Duperron, S., Bergin, C., Zielinski, F., Blazejak, A., Pernthaler, A., McKiness, Z.P., DeChaine, E., Cavanaugh, C.M., and 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. Environ. Microbiol. *8*, 1441–1447.

Duperron, S., Halary, S., Lorion, J., Sibuet, M., and 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.

Ellegaard, K.M., and Engel, P. (2019). Genomic diversity landscape of the honey bee gut microbiota. Nat. Commun. *10*, 446.

Endow, K., and Ohta, S. (1990). Occurrence of bacteria in the primary oocytes of vesicomyid clam *Calyptogena soyoae*. Mar. Ecol. Prog. Ser. *64*, 309–311.

Fontanez, K.M., and Cavanaugh, C.M. (2014). Evidence for horizontal transmission from multilocus phylogeny of deep-sea mussel (Mytilidae) symbionts. Environ. Microbiol. 16, 3608–21

Frank, S.A. (1996). Host–symbiont conflict over the mixing of symbiotic lineages. Proc R Soc Lond B *263*, 339–344.

Geier, B.K., Sogin, E., Michellod, D., Janda, M., Kompauer, M., Spengler, B., Dubilier, N., and Liebeke, M. (2019). Spatial metabolomics of in situ, host-microbe interactions. BioRxiv 555045.

Ghoul, M., and Mitri, S. (2016). The ecology and evolution of microbial competition. Trends Microbiol. *24*, 833–845.

Gillings, M.R., and Stokes, H.W. (2012). Are humans increasing bacterial evolvability? Trends Ecol. Evol. *27*, 346–352.

Goffredi, S.K., Johnson, S.B., and Vrijenhoek, R.C. (2007). Genetic diversity and potential function of microbial symbionts associated with newly discovered species of *Osedax* polychaete worms. Appl Env. Microbiol *73*, 2314–2323.

Gunderson, F.F., and Cianciotto, N.P. (2013). The CRISPR-associated gene cas2 of *Legionella pneumophila* is required for intracellular infection of amoebae. MBio *4*, e00074-13.

Hohenlohe, P.A., Phillips, P.C., and Cresko, W.A. (2010). Using population genomics to detect selection in natural populations: key concepts and methodological considerations. Int. J. Plant Sci. *171*, 1059–1071.

Hu, H., Lan, R., and Reeves, P.R. (2006). Adaptation of multilocus sequencing for studying variation within a major clone: evolutionary relationships of *Salmonella enterica* serovar Typhimurium. Genetics *172*, 743–750.

Ikuta, T., Takaki, Y., Nagai, Y., Shimamura, S., Tsuda, M., Kawagucci, S., Aoki, Y., Inoue, K., Teruya, M., Satou, K., et al. (2016). Heterogeneous composition of key metabolic gene clusters in a vent mussel symbiont population. ISME J.

Jannasch H. W. (1985). Review Lecture - The chemosynthetic support of life and the microbial diversity at deep-sea hydrothermal vents. Proc. R. Soc. Lond. B Biol. Sci. *225*, 277–297.

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.

Lamarche, M.G., Wanner, B.L., Crépin, S., and Harel, J. (2008). The phosphate regulon and bacterial virulence: a regulatory network connecting phosphate homeostasis and pathogenesis. FEMS Microbiol. Rev. *32*, 461–473.

Lilja, E.E., and Johnson, D.R. (2016). Segregating metabolic processes into different microbial cells accelerates the consumption of inhibitory substrates. ISME J. *10*, 1568–1578.

Lorion, J., Kiel Steffen, Faure Baptiste, Kawato Masaru, Ho Simon Y. W., Marshall Bruce, Tsuchida Shinji, Miyazaki Jun-Ichi, and Fujiwara Yoshihiro (2013). Adaptive radiation of chemosymbiotic deep-sea mussels. Proc. R. Soc. B Biol. Sci. *280*, 20131243.

Louwen, R., Horst-Kreft, D., Boer, A.G., Graaf, L., Knegt, G., Hamersma, M., Heikema, A.P., Timms, A.R., Jacobs, B.C., Wagenaar, J.A., et al. (2013). A novel link between *Campylobacter jejuni* bacteriophage defence, virulence and Guillain–Barré syndrome. Eur. J. Clin. Microbiol. Infect. Dis. *32*, 207–226.

Luyten, Y.A., Thompson, J.R., Morrill, W., Polz, M.F., and Distel, D.L. (2006). Extensive variation in intracellular symbiont community composition among members of a single population of the wood-boring bivalve *Lyrodus pedicellatus* (Bivalvia: Teredinidae). Appl Env. Microbiol *72*, 412–417.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J., Wolf, Y.I., Yakunin, A.F., et al. (2011). Evolution and classification of the CRISPR–Cas systems. Nat. Rev. Microbiol. *9*, 467–477.

Mira, A., and Moran, N.A. (2002). Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria. Microb. Ecol. *44*, 137–143.

Nagpal, R., Mainali, R., Ahmadi, S., Wang, S., Singh, R., Kavanagh, K., Kitzman, D.W., Kushugulova, A., Marotta, F., and Yadav, H. (2018). Gut microbiome and aging: Physiological and mechanistic insights. Nutr. Healthy Aging *4*, 267–285.

Nedoncelle, K., Lartaud, F., de Rafelis, M., Boulila, S., and Le Bris, N. (2013). A new method for high-resolution bivalve growth rate studies in hydrothermal environments. Mar. Biol. *160*, 1427–1439.

Nikolic, N., Schreiber, F., Co, A.D., Kiviet, D.J., Bergmiller, T., Littmann, S., Kuypers, M.M.M., and Ackermann, M. (2017). Cell-to-cell variation and specialization in sugar metabolism in clonal bacterial populations. PLOS Genet. *13*, e1007122.

Nussbaumer, A.D., Fisher, C.R., and Bright, M. (2006). Horizontal endosymbiont transmission in hydrothermal vent tubeworms. Nature *441*, 345–348.

Nyholm, S.V., and McFall-Ngai, M.J. (2003). Dominance of *Vibrio fischeri* in secreted mucus outside the light organ of *Euprymna scolopes*: the first site of symbiont specificity. Appl Env. Microbiol *69*, 3932–3937.

van der Oost, J., Westra, E.R., Jackson, R.N., and Wiedenheft, B. (2014). Unravelling the structural and mechanistic basis of CRISPR-Cas systems. Nat. Rev. Microbiol. *12*, 479–492.

Payne, J.L., and Wagner, A. (2019). The causes of evolvability and their evolution. Nat. Rev. Genet. *20*, 24.

Perez, M., and Juniper, S.K. (2017). Is the trophosome of *Ridgeia piscesae* monoclonal? Symbiosis 1–11.

Phelan, V.V., Liu, W.-T., Pogliano, K., and Dorrestein, P.C. (2011). Microbial metabolic exchange--the chemotype-to-phenotype link. Nat. Chem. Biol. *8*, 26–35.

Picazo, D.R., Dagan, T., Ansorge, R., Petersen, J.M., Dubilier, N., and Kupczok, A. (2019). Horizontally transmitted symbiont populations in deep-sea mussels are genetically isolated. BioRxiv 536854.

Polzin, J., Arevalo P., Nussbaumer T., Polz M.F., and Bright M. (2019). Polyclonal symbiont populations in hydrothermal vent tubeworms and the environment. Proc. R. Soc. B Biol. Sci. *286*, 20181281.

Ponnudurai, R., Kleiner, M., Sayavedra, L., Petersen, J.M., Moche, M., Otto, A., Becher, D., Takeuchi, T., Satoh, N., Dubilier, N., et al. (2017). Metabolic and physiological interdependencies in the *Bathymodiolus azoricus* symbiosis. ISME J.

Rainey, P.B., and Cooper, T.F. (2004). Evolution of bacterial diversity and the origins of modularity. Res. Microbiol. *155*, 370–375.

Rhoads, D.C., Lutz, R.A., Revelas, E.C., and Cerrato, R.M. (1981). Growth of bivalves at deep-sea hydrothermal vents along the Galápagos Rift. Science *214*, 911–913.

Robidart, J.C., Bench, S.R., Feldman, R.A., Novoradovsky, A., Podell, S.B., Gaasterland, T., Allen, E.E., and Felbeck, H. (2008). Metabolic versatility of the *Riftia pachyptila* endosymbiont revealed through metagenomics. Environ. Microbiol. *10*, 727–737.

Romano, S., Schulz-Vogt, H.N., González, J.M., and Bondarev, V. (2015). Phosphate limitation induces drastic physiological changes, virulence-related gene expression, and secondary metabolite production in *Pseudovibrio* sp. Strain FO-BEG1. Appl. Environ. Microbiol. *81*, 3518–3528.

Roux, S., Hawley, A.K., Beltran, M.T., Scofield, M., Schwientek, P., Stepanauskas, R., Woyke, T., Hallam, S.J., and Sullivan, M.B. (2014). Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. eLife *3*, e03125.

Russel, J., Røder, H.L., Madsen, J.S., Burmølle, M., and Sørensen, S.J. (2017). Antagonism correlates with metabolic similarity in diverse bacteria. Proc. Natl. Acad. Sci. *114*, 10684–10688.

Russell, S.L., and Cavanaugh, C.M. (2017). Intrahost genetic diversity of bacterial symbionts exhibits evidence of mixed infections and recombinant haplotypes. Mol. Biol. Evol. *34*, 2747–2761.

Sachs, J.L., Russell, J.E., Lii, Y.E., Black, K.C., Lopez, G., and Patil, A.S. (2010a). Host control over infection and proliferation of a cheater symbiont. J. Evol. Biol. *23*, 1919–1927.

Sachs, J.L., Ehinger, M.O., and Simms, E.L. (2010b). Origins of cheating and loss of symbiosis in wild Bradyrhizobium. J. Evol. Biol. *23*, 1075–1089.

Salazar, N., Valdés-Varela, L., González, S., Gueimonde, M., and Reyes-Gavilán, C.G. de los (2017). Nutrition and the gut microbiome in the elderly. Gut Microbes *8*, 82.

Sampson, T.R., Saroj, S.D., Llewellyn, A.C., Tzeng, Y.-L., and Weiss, D.S. (2013). A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. Nature *497*, 254–257.

Santos-Beneit, F. (2015). The Pho regulon: a huge regulatory network in bacteria. Front. Microbiol. *6*.

Sayavedra, L. (2016). Host-symbiont interactions and metabolism of chemosynthetic symbiosis in deep-sea *Bathymodiolus* mussels. PhD Thesis. University of Bremen.

Sayavedra, L., Ansorge, R., Rubin-Blum, M., Leisch, N., Dubilier, N., and Petersen, J. Horizontal acquisition followed by expansion and diversification of toxin-related genes in deep-sea bivalve symbionts. *in prep*.

Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., Waller, A., Mende, D.R., Kultima, J.R., Martin, J., et al. (2013). Genomic variation landscape of the human gut microbiome. Nature *493*, 45–50.

Schöne, B.R., and Giere, O. (2005). Growth increments and stable isotope variation in shells of the deep-sea hydrothermal vent bivalve mollusk *Bathymodiolus brevior* from the North Fiji Basin, Pacific Ocean. Deep Sea Res. Part Oceanogr. Res. Pap. *52*, 1896–1910.

Thorpe, H.A., Bayliss, S.C., Hurst, L.D., and Feil, E.J. (2017). Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. Genetics *206*, 363–376.

Truong, D.T., Tett, A., Pasolli, E., Huttenhower, C., and Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. *27*, 626–638.

Vasu, K., and Nagaraja, V. (2013). Diverse functions of Restriction-Modification systems in addition to cellular defense. Microbiol. Mol. Biol. Rev. MMBR *77*, 53–72.

Verna, C., Ramette, A., Wiklund, H., Dahlgren, T.G., Glover, A.G., Gaill, F., and Dubilier, N. (2010). High symbiont diversity in the bone-eating worm *Osedax* mucofloris from shallow whale-falls in the North Atlantic. Environ. Microbiol. *12*, 2355–2370.

Visick, K.L., and McFall-Ngai, M.J. (2000). An exclusive contract: specificity in the *Vibrio fischeri-Euprymna scolopes* partnership. J. Bacteriol. *182*, 1779–1787.

Wagner A. (2008). Robustness and evolvability: a paradox resolved. Proc. R. Soc. B Biol. Sci. *275*, 91–100.

Wang, T.-C., and Chen, F.-C. (2013). The evolutionary landscape of the *Mycobacterium tuberculosis* genome. Gene *518*, 187–193.

Watrous, J.D., and Dorrestein, P.C. (2011). Imaging mass spectrometry in microbiology. Nat. Rev. Microbiol. *9*, 683–694.

Wentrup, C., Wendeberg, A., Schimak, M., Borowski, C., and Dubilier, N. (2014). Forever competent: deep-sea bivalves are colonized by their chemosynthetic symbionts throughout their lifetime. Environ. Microbiol. *16*, 3699–3713.

Westra, E.R., Buckling, A., and Fineran, P.C. (2014). CRISPR–Cas systems: beyond adaptive immunity. Nat. Rev. Microbiol. *12*, 317–326.

Won, Y.-J., Hallam, S.J., O'Mullan, G.D., Pan, I.L., Buck, K.R., and Vrijenhoek, R.C. (2003). Environmental acquisition of thiotrophic endosymbionts by deep-sea mussels of the genus *Bathymodiolus*. Appl. Environ. Microbiol. *69*, 6785–6792.

Yosef, I., Goren, M.G., and Qimron, U. (2012). Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. Nucleic Acids Res. *40*, 5569–5576.

Zielinski, F.U., Gennerich, H.-H., Borowski, C., Wenzhöfer, F., and 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*, Q0AE04.

Acknowledgements

This thesis would not have been possible without the support, love and laughter from all the people that have been with me over the past few years.

Thank you, **Prof. Dr. Nicole Dubilier** for the possibility of conducting the studies, for your support along the way and for reviewing this thesis. Many thanks also to **Prof. Dr. Matthias Horn** for reviewing this thesis and for making time to join the examination committee. I also would like to thank **Dr. Bernhard Fuchs** and **Prof. Dr. Marko Rohlfs** for joining the examination committee.

Dr. Jillian Petersen, Jill, thank you for the supervision throughout my doctoral studies and for allowing me to spend a significant time of my studies in the Division of Microbial Ecology (DOME) at the University of Vienna. Also thank you for the insights on your magical writing skills - I learned a lot from you and I still have a lot to learn.

Thank you **Miguel Ángel González Porras** and **Caroline Zeidler** for being part of the examination committee.

In the **Symbiosis Department** I have always felt motivated by all my creative, smart, motivated, fun, crazy, shallow (water) and deep (sea) colleagues. Thanks for making science exciting, fun and cool! A special thanks goes to **Martina**, my mutual symbiotic office-partner who welcomes me every morning in the office and always has a friendly word. And also thank you for your technical support. Also thanks to **DOME** for kindly hosting me during 1.5 years of my PhD time. A special thanks goes to the **wolf pack office**!

Thank you, **MarMic Garlic Class**, I always had a lot of fun with you guys – I am curious to see were everyone will be 10 years from now.

A big thanks to **Oliver** for being a great friend and colleague – in the same boat – and for always having an open ear and a solution for almost everything. Your motto: *keep-calm-and-carry-on* (you don't have a choice) saved my day not only once. Also thank you for proof- reading part of this thesis.

Målin, thank you for your support especially throughout the last part of my thesis and for being a partner-in-crime when it comes to game evenings! And OF COURSE for your legendary cakes! Thank you for proof-reading part of this thesis.

Merle, your honest, organized and positive attitude was always helpful to realize that everything is just half as bad – it is great working with you. Thanks for proof-reading part of this thesis.

Thank you **Burak** for always reminding me of a life 'Ohne-Sorge'.

Miguel – the king of geneFISH – thank you for your efforts and expertise on making the geneFISH work. And also muchas gracias for bringing a bit of Spanish Sun to work every day.

Dear **Lizbeth**, you played a big part in forming my thesis project and I always enjoyed working and discussing with you. It was great in the good-old-office-times were we 'worked those mussels'. Thank you also for your contributions to the manuscripts in this thesis and and for valuable input on the second manuscript.

I would like to take the opportunity here to also thank **Maxim** for always having great ideas (and sharing them!) and for his contribution and input on one of the manuscripts in this thesis.

Sometimes one has to take a step back, even from science. **Jennifer**, you helped me to do this and come back to myself in stressy times. I never thought I can get so excited about walking bare-feet in the garden. Thank you for this and for always having a smile to give.

Josephine… there is so much I could write about you here. I think one of the pillars this thesis is build on is you. I know I can count on you all the time with anything. Science, work, private life – I know I get your honest advice and support. How often did we burst into laughter tears during MarMic and after. Thank you for all that and also for proof-reading my summary and especially untying the denglish I wrote in the german version – seriously one of the hardest parts of this thesis. Thank you for your friendship Dr. J. Z. Rapp!

Dear **Stefano**, I don't know were to start with thanking you. Thank you very much for your input on many parts on this thesis and for your contributions to the manuscripts (boring stuff first I thought). The last years have been a roller coaster – and you know I love roller coasters! You are the smartest person I know and I feel very privileged to have your support in science and every other part of my life. I cannot thank you enough for your contribution to the finishing of this thesis, for always being there in every moment along the way. I don't know HOW you manage to make me laugh and happy every day, but you do – Grazie che la vita con te è bella ogni giorno!

Mama, **Papa** und **Miri** – vielen Dank, dass ihr die beste, lustigste, verrückteste und liebste Familie seid, die ich mir wünschen könnte. Ohne eure Untertützung gäbe es diese Arbeit bestimmt nicht. Ich weiß, dass ich immer auf euch zählen kann und, dass ihr immer auf meiner Seite seid – egal was passiert. Danke für all die schönen Momente, Urlaube, Fahrradtouren und langen Sommerabende an denen wir schon oft die Welt neu erkärt haben. Ihr seid die Besten! Ich hab euch lieb.

Grazie **Angela** e **Mario** che siete diventati una seconda familia per me. Mi sento molto fortunata di avere due famiglie. Grazie per la vostra ospitalità ogni tempo che vengo a casa vostra. Un abbraccio! Anche voglio ringraziare **Gianluca** e **Laura** per il sostegno e per sempre fare ogni volta che vengo a Roma una bella esperienza!

Contribution to each manuscript included in this thesis

Manuscript 1 (chapter II)

Conceptual design: 50%

Data acquisition and experiments: 40%

Analysis and interpretation of results: 80%

Preparation of figures and tables: 95%

Writing the manuscript: 70%

Manuscript 2 (chapter III)

Conceptual design: 80%

Data acquisition and experiments: 40%

Analysis and interpretation of results: 90%

Preparation of figures and tables: 95%

Writing the manuscript: 90%

Manuscript 3 (chapter IV)

Conceptual design: 80%

Data acquisition and experiments: 40%

Analysis and interpretation of results: 90%

Preparation of figures and tables: 100%

Writing the manuscript: 95%

Universität Bremen

Bremen, 25. Februar 2019

Versicherung an Eides Statt

Ich, **Rebecca Ansorge**
Neuer Weg 22, 27798 Hude, Matrikelnummer: 2344485

.

versichere an Eides Statt durch meine Unterschrift, dass

1. ich die vorstehende Arbeit mit dem Titel „**Strain diversity and evolution in endosymbionts of Bathymodiolus mussels**“ selbständig und ohne fremde Hilfe angefertigt habe,

2. ich alle Stellen, die ich wörtlich dem Sinne nach aus Veröffentlichungen entnommen habe, als solche kenntlich gemacht habe, mich auch keiner anderen als der angegebenen Literatur oder sonstiger Hilfsmittel bedient habe.

3. die elektronische Version der Dissertation identisch mit der abgegebenen gedruckten Version ist

Ich versichere an Eides Statt, dass ich die vorgenannten Angaben nach bestem Wissen und Gewissen gemacht habe und dass die Angaben der Wahrheit entsprechen und ich nichts verschwiegen habe.

_____

Bremen, 25.02.2019, Rebecca Ansorge