Application of high-resolution metagenomics to study symbiont population structure across individual mussels

Submitted in fulfilment of the requirements for the degree
Doktor der Naturwissenschaften (Dr. rer. Nat.)
in the Faculty of Mathematics and Natural Sciences
of the Christian-Albrechts University Kiel

Submitted by
Devani Romero Picazo

Kiel, February 2020

First examiner: Prof. Dr. Tal Dagan

Second examiner: Prof. Dr. Nicole Dubilier

Date of the oral examination: 11.06.2020

Declaration

I hereby declare that thesis entitled “Application of high-resolution metagenomics to study symbiont population structure across individual mussels” has been carried out in the Institute of General Microbiology at the Christian-Albrechts University of Kiel, Germany, under the guidance of Dr. Anne Kupczok and Prof. Dr. Tal Dagan. The work is original and has not been submitted in part or full by me or any degree at other University. I further declare that the material obtained from other sources has been duly acknowledged in the thesis. My work has been produced in compliance to the principles of good scientific practice in accordance with the guidelines of the German science foundation.

Kiel, 11.06.2020

Devani Romero Picazo

Table of Contents

1 Abstract 6

2 Zusammenfassung (abstract in German) 7

3 Introduction 8

4 Methodology. 13

4.1 Collection and sequencing 14

4.2 Construction of the non-redundant gene catalog 15

4.3 Taxonomic annotation of gene catalog 15

4.4 Population pan-genomes reconstruction 16

4.5 Orthologous genes identification 17

4.6 Functional annotation 17

4.7 Estimation of the gene catalog coverages 17

4.8 Genome binning and symbiont core genome identification 18

4.9 SNV discovery on the core genomes 18

4.10 Measures of population structure 19

4.11 Strain deconvolution 20

4.12 Measures of community composition 21

4.13 Allele frequency spectra estimation 21

4.14 pN/pS and Neutrality Index estimation 21

5 Chapter I: on the population structure of deep-sea mussel symbionts. 23

5.1 Results 23

5.1.1 SOX is the dominant community member in the Bathymodolus microbiota. 23

5.1.2 Gene-based metagenomics binning recovers SOX and MOX core genomes 24

5.1.3 Bathymodiolus microbiota is composed of SOX and MOX strains from several clades 30

5.1.4 SOX strains evolve under purifying selection while MOX evolution is characterized by neutral processes 35

5.1.5 Intra-sample diversity is higher for SOX than for MOX. 39
5.1.6 Geographic isolation of bacterial communities associated with individual mussels. 40
5.1.7 Individual-specific microbiota genetic variants cannot be explained by host genetics. 45
5.2 Discussion 46
6 Chapter II: reconstructing the population pan-genomes of symbionts. 50
6.1 Results. 50
6.1.1 SOX and MOX symbiont populations are characterized by differences in their pan-genome sizes. 50
6.1.2 Strain clades are characterized by differential gene content. 54
6.1.3 Core and accessory genes show differential GC content and pN/pS. 63
6.1.4 Functionality of the accessory genomes. 71
6.1.5 Orthologous genes among SOX and MOX pan-genomes. 74
6.2 Discussion. 76
7 References 80
8 Acknowledgements 91
9 Supplementary Information 92
9.1 Impact of sequencing depth on diversity analysis 92
9.2 Strain symbiont composition based on ribosomal proteins 92
10 Supplementary tables. 93
11 Supplementary figures 94

1 Abstract

Eukaryotes are habitats for bacterial organisms where the host colonization and dispersal among individual hosts have consequences for bacterial ecology and evolution. Vertical symbiont transmission leads to geographic isolation of the microbial population and consequently to genetic isolation of microbiotas from individual hosts. In contrast, the extent of geographic and genetic isolation of horizontally transmitted microbiota and its consequences in shaping population pan-genomes is poorly characterized. Here we show that chemosynthetic symbionts (Sulfur-oxidizing or SOX and Methane-oxidizing or MOX) of individual *Bathymodiolus brooksi* mussels constitute genetically isolated subpopulations. The reconstruction of core genome-wide strain sequences from high-resolution metagenomes revealed distinct phylogenetic clades. Nucleotide diversity and strain composition vary along the mussel lifespan, and individual hosts show a high degree of genetic isolation. By additionally reconstructing population pan-genomes, we reveal that gene content differences between mussel symbiont communities reflect the differences in strain composition; thus, strains belonging to the same monophyletic group share most of their genes. Furthermore, for both symbionts, the accessory gene content is over-represented in functions related to genome integrity. Compared to SOX, the MOX pan-genome is larger and has a smaller fraction of accessory genes. We find that MOX contains more genes related to cell motility and mobile genetic elements. Altogether, our results suggest that the uptake of environmental bacteria is a restricted process in *B. brooksi*, where self-infection of the gill tissue results in serial founder effects during symbiont evolution. We suggest that this geographic isolation among symbiont populations from individual mussels limits the exposure of symbionts to mobile genetic elements. In addition, the differences between both species suggest that the two symbionts have different ecological traits, where the association of MOX with the host occurred more recently and has a more facultative character that may involve an active free-living phase. We conclude that bacterial colonization dynamics over the host life cycle are an important determinant of population structure and genome evolution of horizontally transmitted symbionts.

2 Zusammenfassung (abstract in German)

Eukaryoten sind Wirte von bakteriellen Organismen. Dabei hat die Besiedlung von Wirten und die Ausbreitung zwischen einzelnen Wirten Auswirkungen auf die Ökologie und Evolution der Bakterien. Die vertikale Übertragung von Symbionten führt zur geografischen Isolation der mikrobiellen Population und folglich zur genetischen Isolation der Mikrobiota einzelner Wirte. Im Gegensatz dazu ist das Ausmaß der geografischen und genetischen Isolation bei horizontal übertragenen Symbionten und die Konsequenzen für Pangenome von Populationen unzureichend charakterisiert. Die vorliegende Arbeit zeigt, dass chemosynthetische Symbionten (Schwefeloxidierend oder SOX und Methanoxidierend oder MOX) von einzelnen Muscheln der Art *Bathymodiolus brooksi* genetisch isolierte Subpopulationen bilden. Aus hochauflösenden Metagenomen wurden Strains, die das gesamte Kerngenom spannen, rekonstruiert. Diese Strains weisen deutlich getrennte phylogenetische Gruppen auf. Nukleotiddiversität und Strain-zusammensetzung variieren über die Muschellebensdauer und einzelne Wirtsmuscheln weisen einen hohen Grad an genetischer Isolation auf. Mithilfe einer Rekonstruktion von Pangenomen kann gezeigt werden, dass Unterschiede im Gengehalt zwischen Symbiontenpopulationen die Unterschiede in der Strain-zusammensetzung widerspiegeln. Demzufolge teilen Strains, die zur gleichen monophyletischen Gruppe gehören, die meisten Gene. Des Weiteren weist der variable Gengehalt beider Symbionten einen Überrepräsentierung von Funktionen, die im Zusammenhang mit Genomintegrität stehen, auf. Im Vergleich zu SOX ist das Pangenom von MOX größer und weist einen geringeren Anteil variabler Genen auf. MOX enthält mehr Gene, die mit Zellmotilität und mobilen genetischen Elementen zusammenhängen. Die Ergebnisse legen nahe, dass die Aufnahme von Bakterien aus der Umwelt ein restriktiver Prozess in *B. brooksi* ist, bei dem eine Selbstinfektion des Kiemengewebes zu fortlaufenden Gründereffekten während der Symbiontenevolution führt. Es wird vermutet, dass wegen der geografische Isolation zwischen den Symbiontenpopulationen einzelner Wirtsmuscheln die Symbionten mobilen genetischen Elementen wenig ausgesetzt sind. Darüber hinaus deuten die Unterschiede zwischen beiden Arten darauf hin, dass die beiden Symbionten unterschiedliche ökologische Merkmale aufweisen, wobei die Assoziation von MOX mit dem Wirt in jüngerer Zeit aufgetreten ist und einen fakultativeren Charakter aufweist, der auch eine aktive freilebende Phase der Symbionten beinhalten kann. Schlußfolgernd ist die Dynamik der bakteriellen Kolonisierung innerhalb des Lebenszyklus des Wirts eine wichtiger Faktor für die Populationsstruktur und die Genomevolution horizontal übertragener Symbionten.

3 Introduction

Bacteria inhabit most eukaryotes where their presence has consequences for key aspects of the host biology (McFall-Ngai et al. 2013), such as host development (McFall-Ngai 2014), nutrition (Shabat et al. 2016), or behavior (Schretter et al. 2018). From the bacterial perspective, animals constitute an ecological niche where microbial communities utilize the resources of their host habitat (Costello et al. 2012). The microbiota biodiversity over the host life cycle is determined by bacteria colonization dynamics and by host properties, including biotic and abiotic factors. For example, the microbiota can be affected by the host diet (David et al. 2014) or the host physiological state -e.g., hibernation (Sommer et al. 2016) or pregnancy (Koren et al. 2012). In addition, changes in the host environmental conditions such as temperature (Jones et al. 1998) or the availability of reduced compounds (Riou et al. 2008) can influence the microbiota community composition.

Microbiota dispersal over the host life cycle depends on the level of fidelity between the host and its microbiota. In faithful interactions, vertically transmitted bacteria are transferred from adults to their progeny during early host developmental stages, while in less faithful interactions, horizontally transmitted bacteria are acquired from the environment throughout the host life cycle (Bright and Bulgheresi 2010). Strictly vertically transmitted bacteria are specialized in their host niche and their association with the host imposes an extreme geographic isolation. Bacterial inheritance over host generations imposes a strong bottleneck on the microbiota population and leads to reduced intra-host genetic diversity (Wernegreen 2015). Examples are monoclonal or biclonal populations observed in symbiotic bacteria inhabiting grass sharpshooter (Woyke et al. 2010) and pea aphids (Guyomar et al. 2018). Furthermore, the geographic isolation of vertically transmitted bacteria leads to genetic isolation and to symbiont genome reduction over time as a consequence of genetic drift (Boscaro et al. 2017). In contrast, dispersal is expected to be higher for horizontally transmitted bacteria, where host-associated subpopulations are connected to one another through the environmental pool (Klose et al. 2015). Nonetheless, the genetic diversity of horizontally transmitted microbial populations may also be reduced due to bottlenecks during symbiont transmission and host colonization. Stochastic effects in the colonization of horizontally transmitted bacteria may manifest themselves in differences in microbiota strain composition among hosts (Hagen and Hamrick 1996; Vega and Gore 2017). This would lead to structured symbiont populations where the geographic isolation of the microbiota depends on the degree of symbiont dispersal among individual hosts. Geographic isolation between individual hosts over

the host life span would then lead to genetic isolation of the symbiont populations and to symbiont population structure. Genomic variation and genetic isolation have been observed for horizontally transmitted symbionts of the human gut microbiome (Schloissnig et al. 2013) and of the honey bee gut microbiome (Ellegaard and Engel 2019). Moreover, structured symbiont populations can also emerge within an individual host, as observed for *Vibrio fischeri* colonizing the squid light organ, where different light organ crypts are infected by a specific strain (Wollenberg and Ruby 2009). The degree of dispersal of horizontally transmitted symbionts remains understudied; hence, whether populations from different microbiomes are intermixing or are genetically isolated is generally unknown.

Here we study the microbiota strain composition of horizontally transmitted endosymbionts across individual *Bathymodiolus brooksi* deep-sea mussels. *Bathymodiolus* mussels live in a nutritional symbiosis with chemosynthetic sulfur-oxidizing (SOX) and methane-oxidizing (MOX) bacteria. The symbionts are acquired horizontally from the seawater and are harbored in bacteriocytes within the gill epithelium (Dubilier et al. 1998; Won et al. 2003). Most Bathymodiolus species harbor only a single 16S rRNA phylotype for each symbiont, including *B. brooksi* (Duperron et al. 2007). A recent metagenomic analysis of *Bathymodiolus* species from hydrothermal vents in the mid-Atlantic ridge showed the presence of different SOX strains with differing metabolic capacity (Ansorge et al. 2019). Mussel gills constantly develop new filaments that are continuously infected (Wentrup et al. 2014). However, whether the new gill filaments in *Bathymodiolus brooksi* are colonized predominantly by environmental bacteria or by symbionts from older filaments of the same host remains unknown. These two alternative scenarios are expected to impose different degrees of geographic isolation on the symbiont population: in continuous environmental acquisition, the level of inter-host dispersal is high while self-infection limits the symbiont dispersal. Here we studied the impact of tissue colonization dynamics of horizontally transmitted intracellular symbionts on the degree of symbiont diversity. Furthermore, we quantified the level of genetic isolation among communities across individual mussels and its impact on symbiont genome evolution.

We implemented a high-resolution metagenomics approach that captures genome-wide diversity for both symbionts in multiple *Bathymodiolus brooksi* individuals from a single site. The field of metagenomics underwent a tremendous expansion in the last decade. Metagenomes are the product of untargeted sequencing of the genomic content present in an environmental sample (Quince, et al. 2017). It has been crucial for the discovery of unculturable species by sequencing partly their genomes -i.e., 16S rRNA sequencing- or by inferring the genomic content of the whole

species (commonly known as metagenome assembled genomes or MAGs) (Parks et al. 2017). Although sequencing became more affordable with time, metagenomics analyses are computationally challenging, mainly because of the need for assembling short reads into whole genomes.

Nonetheless, new methods are continuously being introduced that enable to explore the genetic diversity present in the environment at unprecedented levels. Metagenomics is a very promising field, where not only we can study the whole variation in populations (including polymorphisms and strain specific genes), but also their extrachromosomal elements (Nielsen et al. 2014; Sibbesen et al. 2018). Here we used a deconvolution method that enabled us to reconstruct strain sequences from mussel metagenomes, and therefore study the strain diversity across individual mussel hosts (Quince, et al. 2017).

Geographic isolation promotes the fixation of mutations; therefore, it can importantly determine the evolution of bacterial genomes. While the impact of population structure on the fixation of single-point mutations has been extensively studied -e.g., in human gut microbial populations (Schloissnig et al. 2013) - the effect that it has on the gain and loss of genes in the population is often overlooked. The term pan-genome refers to all DNA sequences that are part of a population or species and encompasses the core genome, which includes genes present in all the individuals, and the dispensable or accessory genome, that contains genes present only in specific strains (Medini et al. 2005; Tettelin et al. 2005). Differences in the accessory genomes may involve functional differences among strains. These functional differences are very often associated to niche adaptation and the formation of ecotypes -e.g., in *Procclorococcus,* the adaptation of different ecotypes to high and low light exposure derives from the accessory region of their genomes (Kettler et al. 2007).

Although the role of accessory genes in adaptation is well known, the main evolutionary driver shaping gene content variation in species pangenomes is very controversial. Two recent articles were published nearly simultaneously and presented opposite conclusions on this matter; one of the studies argues, within a population genetics framework, that pan-genomes are the product of adaptive selection. They concluded that large pan-genomes will be found in populations with large effective population size (N_e) because selection is here more efficient in fixing adaptive newly acquired alleles (McInerney et al. 2017). On the other hand, the positive correlation between N_e and pan-genome size was suggested to result from genetic drift. The

authors of this work argue that populations with high N_e have a higher nucleotide diversity, and this might be possible to extrapolate also to gene content variation (Andreani et al. 2017).

There are two underlying molecular mechanisms that shape gene content variation: gene gain and gene loss. Gene loss occurs frequently in host-associated bacteria, (McCutcheon and Moran 2012; Bolotin and Hershberg 2015). Especially in vertically transmitted symbionts, populations are strongly affected by geographic isolation, what translates into a pervasive gene loss. Symbiont populations undergo bottlenecks during their transmission to new hosts, what leads to the fixation of non-sense mutations in non-essential genes by drift, and subsequent pseudogenization (Moran 2002). A very interesting example is the symbiosis between the bacterium *Buchnera* and its aphid host. Here, genes for the synthesis of essential amino acids became non-essential for the symbiont, as the metabolic pathway could be complemented by functions from the insect, and got lost (Hansen and Moran 2011).

On the other hand, genes can be acquired *de novo* via horizontal gene transfer (HGT) from other bacteria, which belong to either same or, less frequently, to different species (Popa et al. 2011). Function gain via HGT is commonly described among host-associated microbes. For instance, HGT among bacteria from the intestinal tract in humans has been shown to be involved in the acquisition of antibiotic-resistance genes (Huddleston 2014) and genes involved in the synthesis of insecticidal molecules have been shown to be horizontally transferred among strains of the plant symbiont *Burkholderia* (Pinto-Carbó et al. 2016). HGT is carried out by the action of mobile genetic elements (MGEs), which include any form of DNA that can “jump” within or between genomes. MGEs include insertion sequences, and their derivate transposons (IS elements), bacteriophages, plasmids and integrative and conjugative elements (ICEs) or sequences that are embedded into outer-membrane vesicles (OMVs). HGT rate is positively correlated with sequence similarity and for that reason it is more frequent among individuals of the same species than among individuals from different species. Nevertheless, HGT is also limited by ecological barriers, such as, for example, geographic distance (Popa and Dagan 2011). One example is the pan-genome of the species *Sulfolobus islandicus*, where clusters of accessory genes could be associated to the biogeography of this bacterium (Reno et al. 2009). MGEs have been suggested to play an important role in mediating niche adaptation of endosymbiotic bacteria. In this regard, the “intracellular arena” hypothesis posits that the host cell serves as an arena for the endosymbionts, where these horizontally acquire genes that are part of an intracellular-specific gene pool (Newton and Bordenstein 2011; Brockhurst et al. 2019).

Bacterial genomes from species with different ecology differ in their averaged GC content, therefore, GC-content is typically used as a proxy to identify genes that have been acquired via HGT (Lawrence and Ochman 1997). GC-content has also been shown to vary between core and accessory regions of the genome, independently of HGT. Core genomes maintain lower AT-content than accessory genomes by purifying selection. Nevertheless, it is debated whether selection or genetic drift is the main mechanism involved in generating low AT accessory genomes (Bohlin et al. 2017). A recent study shows that introducing lower AT-content plasmids to *E.coli* cells decreases the host's fitness and thus suggests that accessory genomes may also be selected to contain higher AT-content than core genomes (Dietel et al. 2018). AT-rich genome is furthermore a characteristic from endosymbiotic bacteria. These have been shown to be higher in AT-content compared to free-living microbes, and mutational AT-bias has been suggested to be the main driver behind this variation (Moran 1996). Most mutations in the genome occur by oxygen radicals that lead to the mismatching of DNA. Repair mechanisms will then bias the composition towards AT-richer sequences, which are energetically more expensive (Wang et al. 1998). For symbiotic bacteria, the small population sizes make random drift more effective, leading to the fixation of these mutations and therefore giving rise to AT-rich genomes.

The presence of multiple symbiont strains inhabiting the same mussel host brings us a unique opportunity to study how geographic isolation affects the exposure of the population to MGEs, and how this geographic isolation impacts the flexibility of the symbiont population pan-genomes. We ask how the strains are functionally different in order to better understand how symbiont diversity is generated. Additionally, we seek genes that may have been transferred between the two co-occurrent symbionts of deep-sea mussels and helped these bacteria to adapt to a symbiotic lifestyle.

4 Methodology.

For the development of this thesis, we have implemented a gene-based bioinformatics pipeline with the aim to analyze deeply sequenced metagenomes from deep-sea mussels at an unprecedented strain-level resolution. This pipeline can be divided into three modules, the first module consists of all the steps between the sequencing and the generation of a non-redundant gene catalog (NRGC). The second module focuses on identifying single-copy core genes (SCCG) and estimation of the frequencies of all single-nucleotide variants (SNVs) in the population. The third module of the pipeline aims to infer the population pangenomes. The last two modules converge in the identification of strain-specific genomes and the study of population structure (**Fig. 1**).

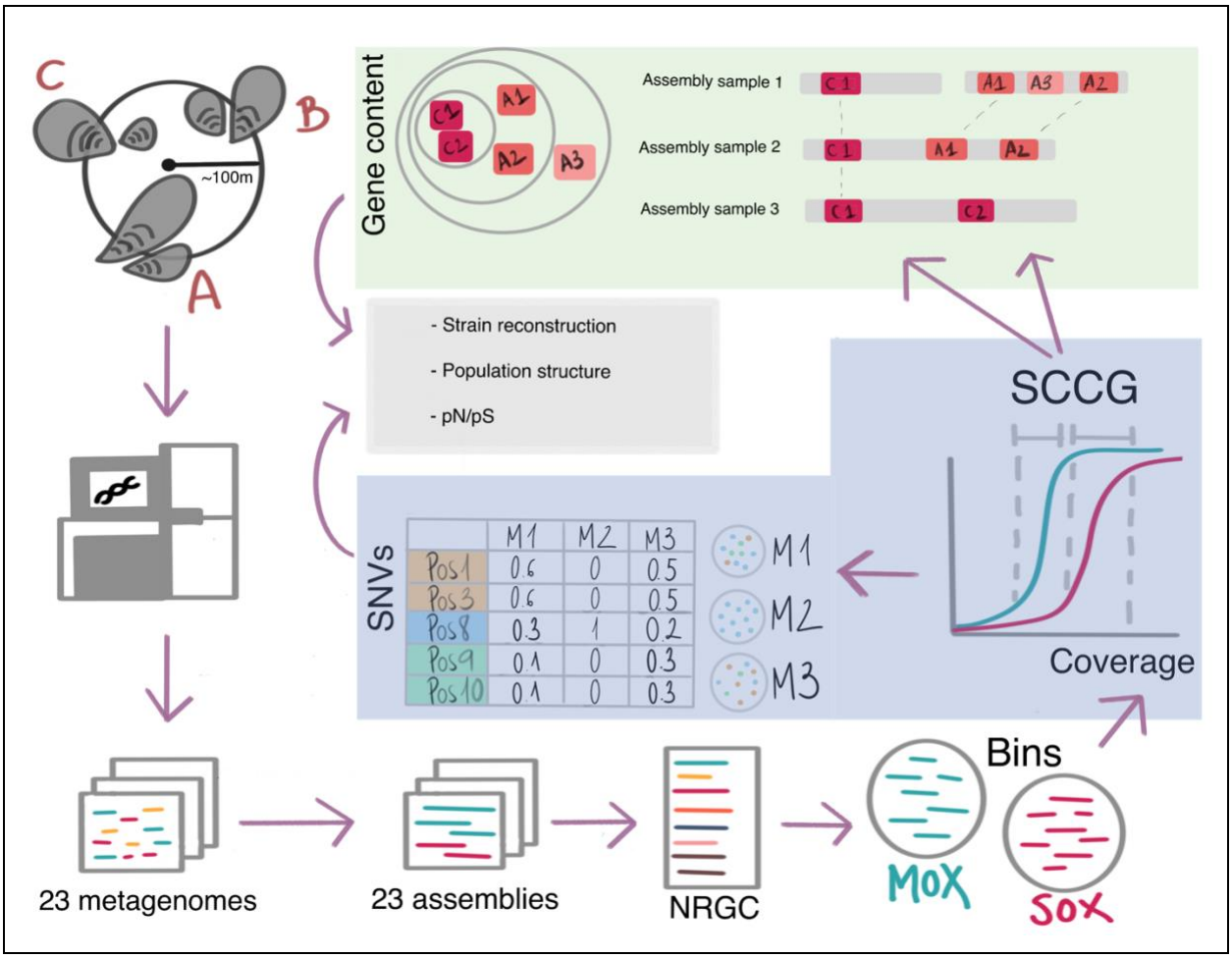


Figure 1: Metagenomics pipeline used in this thesis to analyze pan-genomes of two symbiotic species from the deep mussel *Bathymodiolus brooksi*. The pipeline consists of three main modules; the first module is meant to reconstruct a non-redundant gene catalog (NRGC) that contains all genes in the community. The second module infers single-copy core genes (SCCG) for the two metagenomic species and the frequency of the variants found among them. The third module identifies the accessory genomes. Posterior analyses involve the study of strain composition and population structure. 23 mussels have been sampled from a single cold seep in the northern Gulf of Mexico. These mussels are distributed into three distinct clumps (A, B and C). Symbiont-containing gills are extracted, homogenized and their metagenomes are sequenced. Per-sample assemblies and ORF predictions are performed. NRGC is reconstructed across samples. We used a canopy-based algorithm to bin genes into metagenomic species. SCCG are identified as those within a certain coverage range across all samples. SNVs are identified on these SCCG. We linked contigs across samples by using genes as connectors to reconstruct population pangenomes. We use SNV frequency covariation and gene content variation across samples to infer strain genomes. Additionally, we use this information to study population structure and degree of selection.

4.1 Collection and sequencing

Collection of samples and sequencing was performed thanks to our collaborators Rebecca Ansorge, Jillian M. Petersen and Nicole Dubilier from the symbiosis department at the Max Planck Institute for Marine Microbiology in Bremen, Germany.

Here, twenty-three individuals of *Bathymodiolus brooksi* mussels were collected during a research cruise with the E/V *Nautilus* from the cold seep location GC853 at the northern Gulf of Mexico in May 2015. The mussel distribution at the cold seep was patchy and mussel individuals were collected from three distinct clumps within a radius of 131 meters (coordinates clump a: 28.1237, -89.1404 depth: -1073m, clump b: 28.1241, -89.1401 + depth: -1073m, clump c: 28.1237, -89.1404 + depth: -1073 to 1078m) (**Fig. 4**). The gills from each mussel individual were dissected immediately after retrieval and homogenized with sterilized stainless steel beads, 3.2 mm in diameter (biostep, Germany). A subsample of the homogenate for sequencing analyses was preserved in RNA later (Sigma, Germany) and stored at -80°C. DNA was extracted from these subsamples as described by (Zhou et al. 1996). TruSeq library preparation and sequencing using Illumina HiSeq2500 was performed by the Max Planck Genome Centre in Cologne,

Germany, resulting in 250 bp paired-end reads with a median insert size of 400 bp. The raw reads were deposited in NCBI under BioProject PRJNA508280.

4.2 Construction of the non-redundant gene catalog

Illumina paired-end raw reads from the samples were trimmed for adapters and filtered by quality using BBMap tools (Bushnell 2014). Only reads with more than 30bp and quality above 10 were kept. This results in 37.7 million paired-end reads per sample on average (**Supplementary Table 1**).

We assembled each of the metagenomic samples individually using metaSPAdes (Nurk et al. 2017). Genes were predicted *ab initio* on contigs with metaProdigal (Hyatt et al. 2012). These predicted genes were clustered by single-linkage according to sequence similarity using BLAT (Kent 2002) (at least 95% of sequence identity in at least 90% of the length of the shortest protein and e-value < 10^{-6}). To reduce the potential inflation caused by the single-linkage clustering, we applied two additional filters to discard hits: the maximum ratio allowed between the two compared sequence lengths must be 4 and hits between partial and non-partial genes are discarded. These filters are meant to remove spurious links between sequences due to the presence of commonly spread protein domains. This clustering was performed in two successive steps; first, we obtained sample-specific gene catalogs by performing intra-sample clustering. This is meant to reduce sequence redundancy, resulting in an average of ~676 000 non-redundant genes per sample (**Supplementary Table 1**). Second, one-sided similarity search was performed across all pairs of sample catalogs. This resulted in 1 156 207 clusters (26.5%) and 3 207 869 (73.5%) singletons, which make up a catalog of 4 364 076 million non-redundant genes. For each of the clusters, we reconstructed a consensus sequence as a cluster representative. To this end, we took the majority nucleotide at each position (ties were resolved randomly).

4.3 Taxonomic annotation of gene catalog

Taxonomic annotation of the gene catalog was performed by aligning the translated genes to the non-redundant protein NCBI database (date: 24/05/18) using diamond (Buchfink et al. 2015) (e-value<10^{-3}, sequence identity ≥ 30%) and obtaining the best hit. Genes were annotated as MOX-related if their best hit is *Bathymodiolus platifrons* methanotrophic gill symbiont (NCBI Taxonomy

ID 113268) or *Methyloprofundus sedimenti* (NCBI Taxonomy ID 1420851). For SOX, the genomes of thioautotrophic symbionts belonging to four different Bathymodiolus species were used for annotation (NCBI Taxonomy IDs: 2360, 174145, 113267 and 235205). In addition, the gene catalog was screened for mitochondrial genes using best blastp hits against the *Bathymodiolus platifrons* mitochondrial protein sequences (NC_035421.1) (Sun et al. 2017) (all e-values $<10^{-40}$). The gene catalog was also screened for symbiont marker genes by best blastp hits to a published protein database for *Bathymodiolus azoricus* symbionts (Ponnudurai et al. 2017) (80% of protein identity and 100% of query coverage). This allowed the identification of 86 SOX and 39 MOX marker genes. The marker gene coverages are generally uniform across a sample, however, a high variance in coverage is present in two of the samples (**Fig. 5**). Since the binning method relies on the covariation of coverage across samples, the presence of a high variance in coverage can interfere with the proper clustering of genes, thus, two samples were discarded from further analysis (Dsc1, Dsc2).

4.4 Population pan-genomes reconstruction

Differences in strain composition creates different assembly fragmentation pattern across samples. Additionally, regions of the genome that are present only in certain strains tend to lay on independent contigs. Here, we restore the linkage between contigs that are part of the original same genome and identify, from all the genes present in the catalog, the accessory and multi-copy genes belonging to SOX and MOX symbiont species. To this end, we use a network traverse approach, where the genes identified as Single-Copy Core (SCCG) are initial seeds. The first layer of the pan-genome will be filled with all the genes that are found in the contigs where the SCCG are located. Note that for the pan-genomes reconstruction only genes of the non-discarded samples are considered. Thus, the size of SCCG sets from the second chapter differs from those from the first chapter, as singletons that originate from discarded samples were not further considered. Then, in an iterative manner, the genes from the recently added layer are seeds to expand the network. The search will conclude when no additional genes can be added to the pan-genome. This approach relies on clusters of homologous genes that are identified as one gene present across samples. For this reason, the presence of genes that can be found multiple times in the genomes -as it is the case for transposases- or the misclassification of different genes within the same cluster, may be responsible for the spurious linkage of genome fragments that do not belong to the same genome. In order to avoid such artifacts, we will only consider genes as seeds

if they originate from clusters with a global sequence identity no lower than 0.95. Additionally, the clusters should not be composed by more than one gene per sample.

4.5 Orthologous genes identification

We identified orthologous sequences between SOX and MOX genes. For that purpose, we extracted reciprocal best blast (Altschul et al. 1990) hits between the translated sequences of SOX and MOX pan-genomes. Then, full-length protein sequence alignments were conducted with the Needleman-Wunsch algorithm implemented in EMBOSS (Needleman and Wunsch 1970; Rice et al. 2000), and pairs of genes with pair-wise identities of at least 30% were assigned as orthologous.

4.6 Functional annotation

Genes that have been identified as belonging to one of the symbiont species were functionally annotated. To this end, protein homology search was conducted by blasting translated gene sequences to the symbionts-related protein database previously published (Ponnudurai et al. 2017). Hits with an e-value < 10^{-10} and at least 40% of protein sequence identity were kept. For genes with no hit to the database, functional annotation was performed by using eggNOG-mapper v2 (Huerta-Cepas et al. 2019).

4.7 Estimation of the gene catalog coverages

To estimate the gene abundances, we mapped the reads of each metagenomic sample to the gene catalog using bwa mem (Li and Durbin 2009). Reads below 95% of sequence identity or mapping quality of 20, as well as not primary alignments were discarded. Coverage per position for each gene in the catalog across samples was calculated using samtools depth (Li et al. 2009) and the gene coverage is given by the mean coverage across positions. We first downsampled the reads in each sample to the minimum number of reads found (33M, **Supplementary Table 1**) and calculated mean coverage per gene to perform the binning and the analyses of coverage variance across symbiont marker genes (see above).

4.8 Genome binning and symbiont core genome identification

Next, we performed co-abundance gene segregation by using a canopy clustering algorithm (Nielsen et al. 2014), which clusters genes into bins that covary in their abundances across the different samples. This approach allows to recover from chimeric associations obtained in the assembly process and to automatically separate core from accessory genes. Gene coverages across samples were used as the abundance profiles for binning. First, genes with a Pearson correlation coefficient (PCC) > 0.9 to the cluster abundance profile were clustered. Then, clusters with PCC > 0.97 between their median abundance profiles were merged and outlier clusters for which the coverage signal originates from less than three samples were removed. In addition, we removed a gene from a cluster if Spearman correlation coefficient to the median canopy coverage profile is lower than 0.7. Finally, overlaps among the clusters were removed by keeping a gene in the largest of the clusters in which it has been found.

This enabled us to cluster 900 310 genes into 98 944 co-abundant gene groups (3 to 699 genes) and three MetaGenomic Species (MGSs, ≥700 genes). An additional filter was applied to the MGSs to obtain final bins by removing outlier genes based on their coverage (**Fig. 2**). To this end, we used the Median Absolute Deviations (MAD) statistic as a cutoff to discard highly or lowly covered genes. We removed genes that are at least 24 times MAD far from the median in at least one sample. The bins after outlier gene removal constitute the core genomes of the MGSs. We checked for the completeness of the symbiont bins with CheckM, by screening for Gammaproteobacteria universal single copy marker genes (Parks et al. 2015).

4.9 SNV discovery on the core genomes

To perform single nucleotide variant (SNV) discovery, we mapped the downsampled reads individually for each sample to the gene catalog. Because sample size has been shown to influence variant detection (Subramanian 2016), we normalized the data across samples. To this end, we normalized each sample to the smallest median coverage found in a sample (482x coverage for SOX, 36x coverage for MOX and 568x for mitochondrial genes). LoFreq was used for probabilistic realignment and variant calling of each sample independently (Wilm et al. 2012). SNVs detected with LoFreq have been hard filtered using the parameters suggested by GATK best practices (Broad Institute). Briefly, SNVs with quality by depth below 2, Fisher's exact test Phred-scaled probability for strand bias above 60, root mean square of mapping quality below 40,

root mean square of base quality above 30, mapping quality rank sum test below -12.5 and read position rank sum test below -8 are kept for further analyses.

The resulting SNVs can be fixed or polymorphic in a sample. Polymorphic SNVs are characterized by the allele frequency of the alternative allele whereas fixed SNVs have an allele frequency of 1. Here, we define SNVs as polymorphic in a metagenomic sample if their frequency is between 0.05 and 0.95 in the sample.

4.10 Measures of population structure

SNV data is used for calculating intra-sample and inter-sample nucleotide diversity (π) as applied before to human gut microbiome species (Schloissnig et al. 2013). Intra-sample nucleotide diversity (π) is given as:

$$\pi(H,G) = \frac{1}{|G|}\sum_{i=1}^{|G|} \sum_{B_1 \in \{ACTG\}} \sum_{B_2 \in \{ACTG\} \setminus B_1} \frac{X_{i,B_1}}{C_i}\frac{X_{i,B_2}}{C_i - 1}$$

where *H* corresponds to the sample, *G* to the bacterial genome, |*G*| is the length of the analyzed genome and $X_{i,Bj}$ is the count of a specific nucleotide B_j at a specific locus *i* with coverage C_i. Inter-sample nucleotide diversity (π) is then given as follows, where H_1 and H_2 correspond to the two samples compared:

$$\pi(H_1, H_2, G) = \frac{1}{|G|}\sum_{i=1}^{|G|} \sum_{B_1 \in \{ACTG\}} \sum_{B_2 \in \{ACTG\} \setminus B_1} \frac{X_{i,B_1,S_1}}{C_{i,S_1}}\frac{X_{i,B_2,S_2}}{C_{i,S_2}}$$

To estimate the degree of genetic isolation based on gene content, we have derived the intra and inter-sample gene diversity measures, based on the previous formulas. Here, we estimate the probability of finding presence and absence of certain gene when sampling two genomes from a sample:

$$Gene\ diversity(S,P) = \frac{1}{|G|}\sum_{i=1}^{|G|} \left(\frac{\Delta_i(C_s - \Delta_i)}{C_s(C_s - 1)}\right)$$

Where G is the number of genes in the pan-genome, Δ_i is the coverage of gene i in sample S and C_s is the median coverage of core genes in sample S. Note that the difference between accessory gene coverage and median core coverage in a sample is always positive -i.e., C_s-Δ_i = min(0, C_s). The inter-sample gene diversity was estimated as follows:

$$Gene\ diversity(S1, S2, P) = \frac{1}{|G|}\sum_{i=1}^{|G|}\left(\frac{\Delta_{i,S1}(C_{S2}-\Delta_{i,S2})}{C_{i,S1}C_{i,S2}}\right)\left(\frac{\Delta_{i,S2}(C_{S1}-\Delta_{i,S1})}{C_{i,S2}C_{i,S1}}\right)$$

Where S1 and S2 correspond to the two samples compared.

Finally, these diversity measures are used to estimate the fixation index (F_{ST}), which measures genetic differentiation based on the nucleotide diversity present within and between populations.

$$F_{ST}(S_1, S_2, G) = 1 - \frac{\pi_{within}}{\pi_{between}} = 1 - \frac{\pi(S_1,G)+\pi(S_2,G)}{2\,\pi(S_1,S_2,G)}$$

The scripts to calculate genome-wide inter and intra-sample nucleotide diversity (π) and fixation index (F_{ST}) across all inter-sample comparisons from pooled SNV data have been deposited at https://github.com/deropi/BathyBrooksiSymbionts.

4.11 Strain deconvolution

We reconstructed the strains for the core genomes with DESMAN (Quince, et al. 2017). The SNVs with two states and their frequencies in each sample are used by DESMAN to identify strains in the core genomes that are present over multiple samples. Thereby, the program uses the SNV frequency covariation across samples to assign the SNV states to a specific genotype. For SOX, we ran the strain deconvolution five times using different seed numbers and 500 iterations. Due to computational limitations, a subset of 5 000 SNVs was used and the haplotypes considering the whole SNV dataset were inferred *a posteriori*. The five replicates were run for an increasing number of strains from seven to twelve. The program uses posterior mean deviance as a proxy for model fit. A posterior mean deviance lower than 5% was reached in the transition from eleven to twelve strains, therefore the number of inferred SOX strains is eleven. We did not run fewer numbers of strains because of the presence of large posterior mean deviances between runs with a small strain number. Additionally, we ran DESMAN for the SOX dataset that was subsampled to the MOX coverage with no replicates and eleven strains were found using posterior mean

deviance. For MOX, we ran four replicates using the whole SNV dataset and 500 iterations. The runs were performed by using an increasing number of strains from two to seven, reaching the optimal number of six strains. The consensus gene sequences of each strain were concatenated to generate the strain core genomes, which were used for further analyses. Splits network of the strain genome sequences were reconstructed using SplitsTree (Huson 1998) and uncorrected distances. The position of the root in the splits network was estimated by the minimum ancestral deviation (MAD) method (Tria et al. 2017), which uses maximum likelihood phylogenetic trees inferred with IQ-TREE (Nguyen et al. 2015).

4.12 Measures of community composition

To study the microbial community composition, we estimated α- and β-diversity accounting for strain relatedness in addition to species richness and evenness. α-diversity was estimated using phylogeny species evenness (PSE) (Helmus et al. 2007) implemented in the R package 'Picante' (Kembel et al. 2010). β-diversity was estimated using the weighted Unifrac distance, which is implemented in the R package 'GUniFrac' (Chen 2018). This measure quantifies differences in strain community composition between two samples and accounts for phylogenetic relationships.

4.13 Allele frequency spectra estimation

The unfolded allele frequency spectra were calculated from biallelic SNVs for each of the bacterial species within individual samples. The unfolded allele frequency spectrum estimation relies on the presence of ancestral states in the population. Because we have no information about the ancestry relationship among the strains present in the samples, we made one main assumption in this regard: the ancestral SNV state in the population corresponds to the one which is present in the higher number of strains. Ties are resolved by arbitrarily assigning one tip of the tree as ancestral state: M2.2 for MOX and S4 for SOX.

4.14 pN/pS and Neutrality Index estimation

We estimated pN/pS for both bacterial populations, which is a variant of dN/dS that can be used based on intra-species SNVs. To this end, we first calculated the expected ratio of

nonsynonymous and synonymous mutations for each gene by accounting for each possible mutation occurring in each of the codons. Then, we estimated the observed nonsynonymous to synonymous ratio by using the biallelic SNVs. These two measures are later compared, resulting in the pN/pS ratio. pN/pS was estimated genome-wide as well as individually for each of the genes in the two symbiont species. The per-gene pN/pS calculation results into undefined estimates for genes with no synonymous mutations. To circumvent this limitation, we added 1 to the number of observed synonymous mutations in each gene, which is a standard correction for dN/dS ratios (Stoletzki and Eyre-Walker 2011).

The neutrality index (NI) accounts for differences in the ratio of nonsynonymous to synonymous variants between divergent and polymorphic SNVs in order to quantify the departure of a population from neutral evolution (Rand and Kann 1996). $NI = \frac{pN/pS}{dN/dS}$, where *pN*, and *pS* are the number of polymorphic synonymous and nonsynonymous sites, respectively, and *dN* and *dS* are the number of divergent synonymous and nonsynonymous sites, respectively. For a coalescent population that evolves neutrally, the nature of fixed mutations that are involved in the divergence of the strains should not be different from that of the polymorphic mutations. An excess of divergent nonsynonymous mutations (NI<1) indicates that the population underwent positive selection or a large demographic change in the past (Rand and Kann 1996).

Here we used the NI to analyze if differences in selection have been involved in the evolution of SOX and MOX strains. Different strains are typically found in more than one sample, and this supports the notion that SNVs that characterize the strains constitute substitutions. We estimated NI by considering two different levels of divergence and polymorphism. First, we defined as divergent all those SNVs that have two possible states among the strains and as polymorphic all the invariant SNVs. Second, we used a more restrictive level of divergence. We excluded putative recently acquired SNVs from the set of divergent SNVs, by discarding those that have multiple states among strains from the same group. Polymorphic SNVs are all the remaining. The scripts to calculate the allele frequency spectra, pN/pS and NI have been deposited at https://github.com/deropi/BathyBrooksiSymbionts. Statistics and plotting were done in R.

5 Chapter I: on the population structure of deep-sea mussel symbionts.

Bathymodiolus brooksi is known to horizontally acquire their symbionts from the seawater. Understanding whether this filtering of bacteria occurs in a continuous or restricted manner is interesting because different degrees of bacterial dispersal impacts population structure. Consequently, colonization dynamics affects gene flow and shapes the genetic diversity of symbiont populations. Here, we inferred the degree of dispersal across bacterial populations that inhabit individual mussels by studying the extent of population structure revealed by the core genomes.

The contents of this chapter are published in ISME Journal:

Romero Picazo, D., Dagan, T., Ansorge, R. *et al.* Horizontally transmitted symbiont populations in deep-sea mussels are genetically isolated. *ISME J* **13,** 2954–2968 (2019). https://doi.org/10.1038/s41396-019-0475-z

5.1 Results

5.1.1 SOX is the dominant community member in the Bathymodolus microbiota.

In order to investigate the community composition, we quantified the relative abundance of the microbiota species based on their genomic coverage in the metagenomes. To compare the relative abundances of the two symbionts across the different mussels, we quantified the ratio of SOX to MOX abundance, as well as the symbiont abundance relative to the mitochondrion gene abundance. For this purpose, we annotated twelve *B. brooksi* mitochondrion genes by sequence similarity to the *B. platifrons* mitochondrial genome (Sun et al. 2017) (**Fig. 2f**). Our results show that SOX has a similar abundance as the mitochondrion, with a median of SOX to mitochondrion ratio of 1.06 (n=19). In contrast, MOX is much less abundant compared to the mitochondrion, with a median ratio of 0.0618 (n=19). The two ratios have a low variation across mussel shell sizes (**Fig. 3a,b**). The comparison of SOX and MOX abundance shows that the SOX coverage is on average 19-fold higher than that of the MOX. This shows that SOX is more abundant in the mussel microbiota (**Fig. 3c**). The comparison among sampling clumps shows a higher SOX to MOX ratio in sampling clump c compared to the other two (**Fig. 3c**). This observation is most likely explained by differences in the availability of H_2S and CH_4 among clumps, which is a known determinant of SOX and MOX abundance in *Bathymodiolus* (Riou et al. 2008). Thus, SOX is the dominant

member in metagenomes of the mussel microbiota analysed here, where differences in the SOX to MOX ratio among the mussel metagenomes are likely determined by environmental factors.

5.1.2 Gene-based metagenomics binning recovers SOX and MOX core genomes

To study the evolution of the SOX and MOX genomes in *Bathymodiolus* mussels we used a high-resolution metagenomics approach. Twenty-three adult *B. brooksi* individuals of shell sizes ranging between 4.8 cm and 24.3 cm were sampled from a single location at a cold seep site in the northern Gulf of Mexico. Shell size correlates with mussel age (Schöne and Giere 2005); thus, analyzing mussels within a wide shell size range allowed us to study the symbiont population structure across adult hosts of different ages.

The mussels were sampled from three separate mussel 'clumps' (small mussel patches residing on the sediment) that were at most 131m apart (**Fig. 4**). Such a 'patchy' distribution has often been observed in deep-sea mussels (Van Dover 2002). To obtain a comprehensive representation of the bacterial diversity in individual mussels and to accurately infer strain-specific genomes, homogenized gill tissue of each mussel was deeply sequenced (on average, 37.8 million paired-end reads of 250bp per sample, **Supplementary Table 1**). The resulting metagenomic sequencing data was analyzed by a gene-based binning approach (Nielsen et al. 2014).

The prediction of protein-coding genes from the assembled metagenomes yielded a non-redundant gene catalog of 4.4 million genes that potentially contained every gene present in the samples. This includes genes from the microbial community and from the mussel host. In the metagenomics binning step, genes that covary in their abundance across the different samples were clustered into metagenomic species (MGSs). Our analysis revealed two MGSs that comprise the SOX and MOX core genomes (**Fig. 2**). The distribution of gene coverage in individual samples shows that genes in each core genome have a similar abundance within each mussel. This confirms the classification of the SOX and MOX MGSs as core genomes. The MOX core genome is the largest MGS and it contains 2 518 genes with a total length of 1.97 Mbp. A comparison to Gammaproteobacteria marker genes shows that it is 96.2% complete. Furthermore, it contains 1 568 genes (62.3%) that have homologs in MOX-related genomes. The SOX core genome contains 1 439 genes, has a total length of 1.27 Mbp and is considered as 80.2% complete. It contains 1 188 genes (82.6%) with homologs in SOX-related genomes.

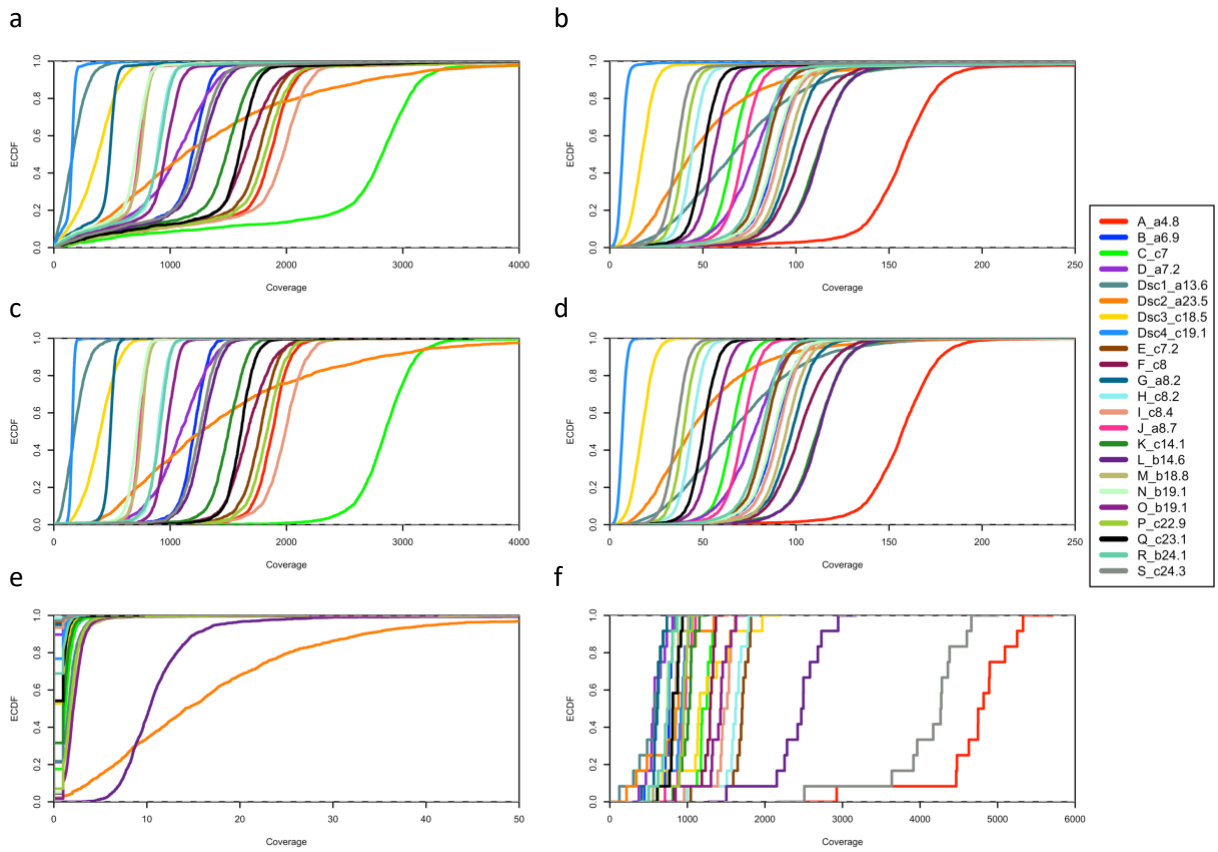


Figure 2: Coverage of symbiont bins. Empirical cumulative distribution function (ECDF) of MAGs and mitochondrion. Samples are additionally labeled with the location where they have been sampled and the shell size. 19 samples (A-R) are used for the analysis of population structure and the remaining 4 samples (Dsc1-Dsc4) were discarded during the analysis; Dsc1 and Dsc2 were discarded before the binning due to high variance in symbiont marker gene coverages (Dsc1_a13.6: median coverage=217.41 and SD=98.14 for SOX marker genes, median coverage=70.47 and SD= 41.3 for MOX marker genes; Dsc2_a23.5: median coverage=1368.28 and SD=785.73 for SOX marker genes, median coverage=44.60 and SD=185.12 for MOX marker genes). Dsc3 and Dsc4 were discarded after binning due to low coverage (median coverage < 15x for MOX core genes and median coverage <350x for SOX core genes). Median and standard deviation (SD) across genomes were estimated using only non-discarded samples. **a,** SOX incl. outlier genes (1 910 genes, median coverage range: 482-2 822, SD range: 157-891), **b,** MOX incl. outlier genes (2 618 genes, median coverage range: 36-157, SD range: 97-298), **c,** SOX (1 439 genes, median coverage range: 486-2 849, SD range: 52-294), **d,** MOX (2 518 genes, median coverage range: 36-157, SD range: 6-21), **e,**

MGS3 (1 449 genes, median coverage range: 0-10, SD range: 0-11), no outlier genes detected. Total genome length of 1.31 Mbp. The third MGS could not be assigned to a taxonomic level; the top three genera found are Oceanicella (Taxonomy ID:1233054) with 61 genes (4.21%), Neomegalonema (Taxonomy ID 356797, 26 genes, 1.79%), and Micavibrio (Taxonomy ID 213485, 26 genes, 1.79%). This unknown species is present in very low abundance and was mainly found in sample O. **f,** Mitochondrion (median coverage range: 568-4 784; SD range: 78-618).

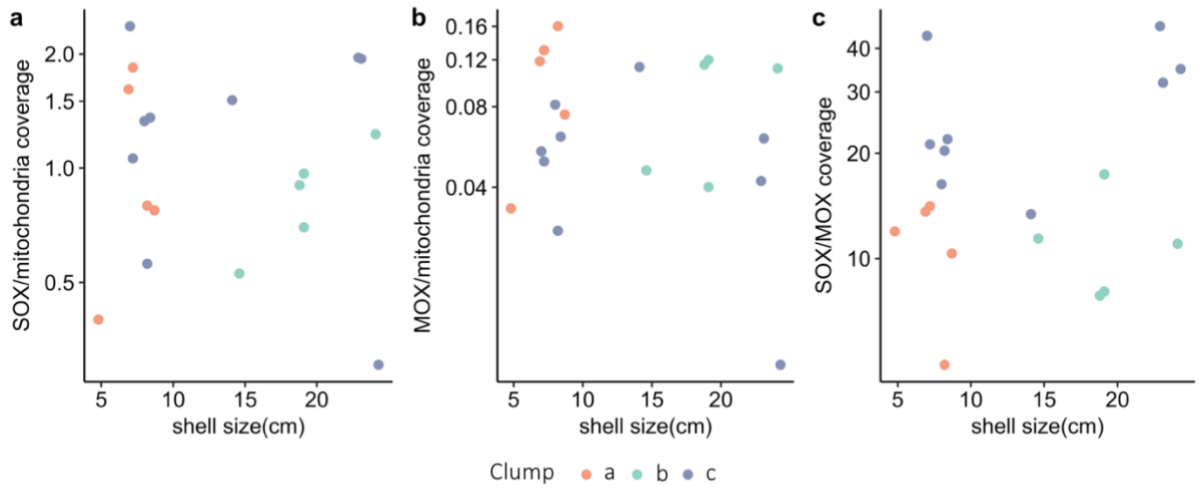


Figure 3: Relative symbiont abundances across mussels. Note that the y-axis (ratio) is displayed in log scale. Coverages are estimated by calculating the mean coverage across all genes for SOX, MOX, and mitochondria, respectively.

In addition to the SOX and MOX core genomes, our analysis revealed a third MGS of 1,449 genes (**Fig. 2e**) that was found in low abundance in a single mussel and, in addition, 98 944 co-abundant gene groups (CAGs, 3-699 genes). Of the 23 metagenomes, four samples were discarded during the metagenomics binning. Two samples were discarded prior to the binning due to high variance in symbiont marker gene coverages and two samples were discarded after binning due to low coverage for both symbionts (**Figs. 2,5**). To gain insight into the SOX and MOX population structure between hosts, we compared the characteristics of the core genomes across the remaining 19 samples. The analysis of the core genome coverages shows that SOX is the dominant member of the mussel microbiota. The differences in the SOX to MOX ratio among the mussel metagenomes are likely explained by differences in the availability of H_2S and CH_4 among clumps, which is a known determinant of SOX and MOX abundance in *Bathymodiolus* (Riou et al. 2008) (**Supplementary Information, Fig. 3**).

To study symbiont diversity below the species level, we analyzed single-nucleotide variants (SNVs) that were detected in the core genomes of the two symbionts. In this analysis, we considered SNVs that are fixed in a metagenome as well as polymorphic SNVs, i.e., SNVs, where both the reference and the alternative allele are observed in a single metagenome. We found 18 070 SNVs in SOX (SNV density of 14 SNVs/kbp, 49 multi-state, 0.27%) and 4 652 SNVs in MOX (SNV density of 2.4 SNVs/kbp, 5 multi-state, 0.11%). The number of polymorphic SNVs per sample ranges from 162 (0.9%) to 11 064 (61%) for SOX and from 27 (0.58%) to 3,026 (65%) for MOX (**Supplementary Table 1**), thus, most SNVs are polymorphic in at least one sample. It is important to note that the observed difference in strain-level diversity between SOX and MOX cannot be explained by the difference in sequencing depth (**Supplementary Information**). These results are in agreement with previous reports of SOX genetic diversity in other *Bathymodiolus* species (Ansorge et al. 2019). We further revealed that there is genetic diversity in the MOX symbiont.

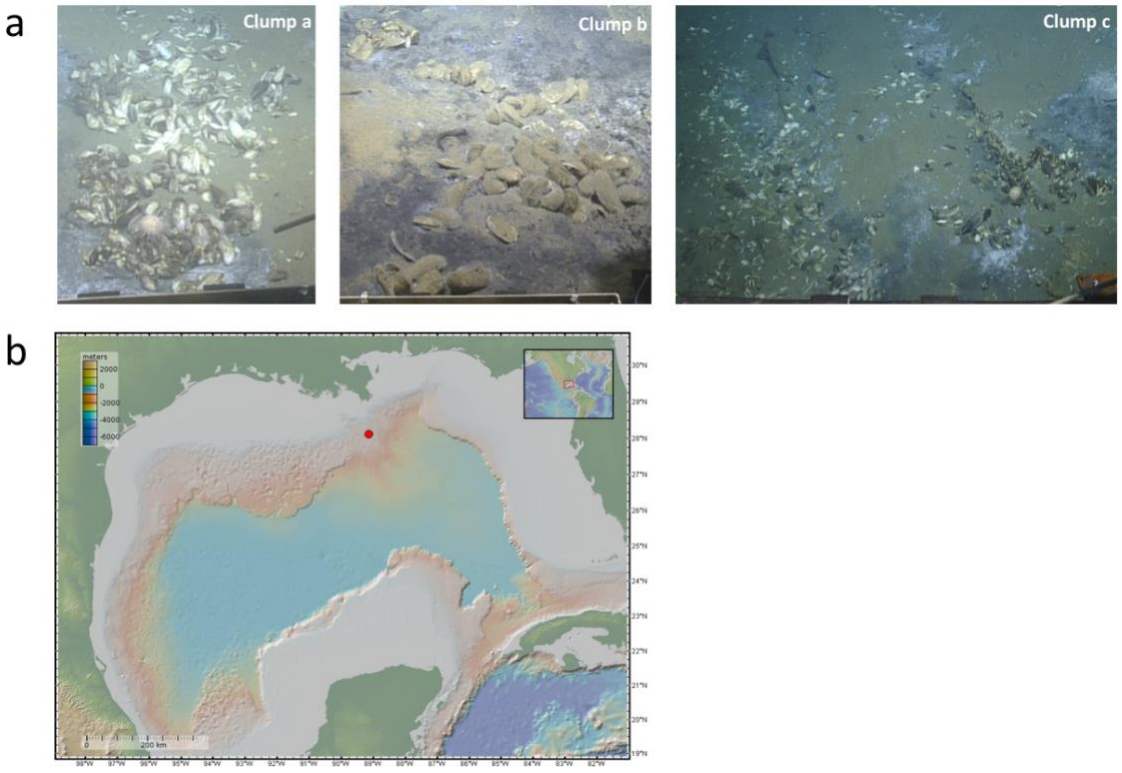


Figure 4: Sampling information. **a,** sampling clumps. **b,** sampling location at the Gulf of Mexico, different color shades represent sampling depth in meters.

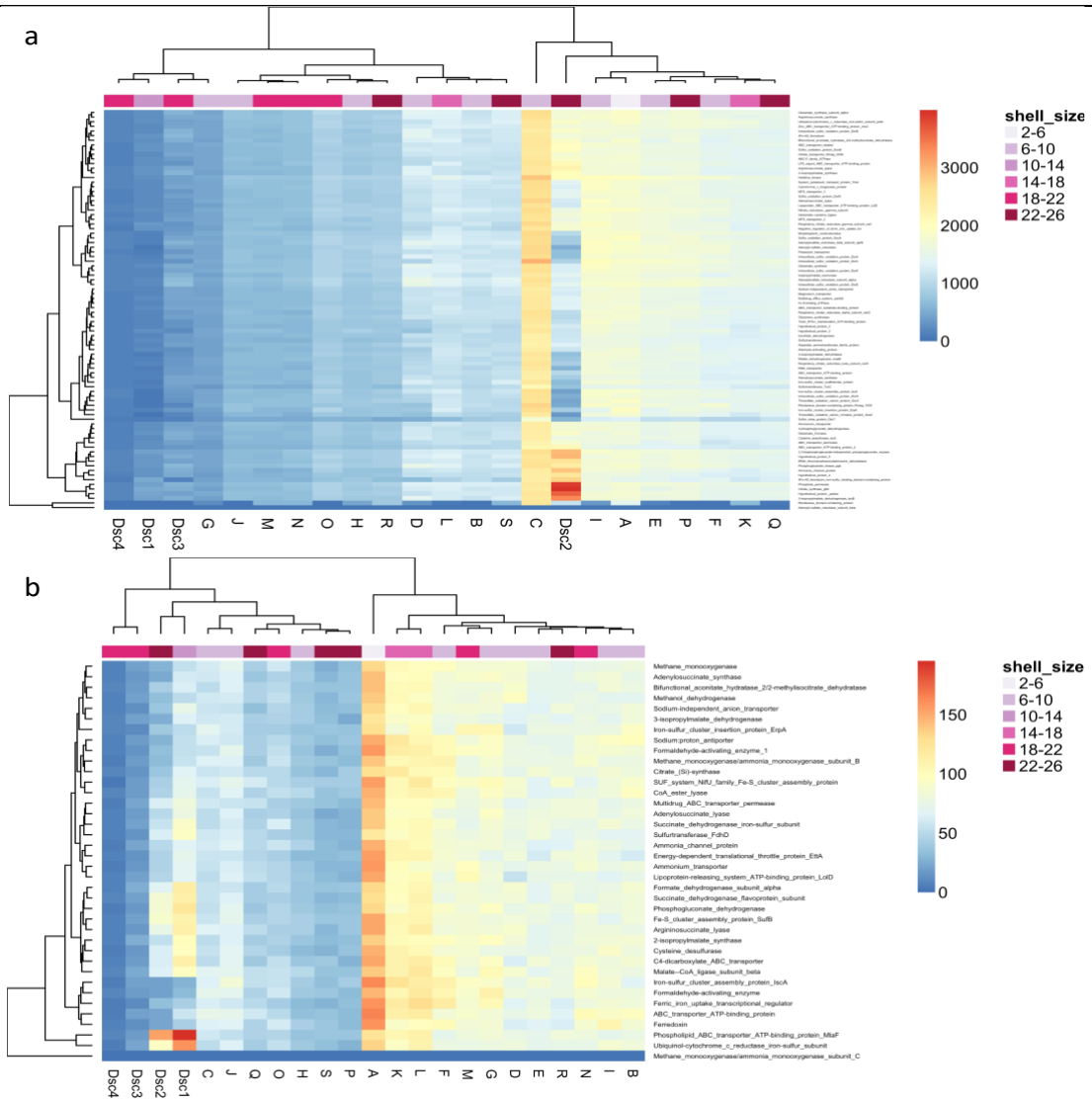


Figure 5: Gene coverages for **a,** 86 SOX and **b,** 39 MOX marker genes. The two core genomes were validated by screening for the presence of known SOX and MOX marker genes (Ponnudurai et al. 2017). These genes encode for protein functions in the methane and sulfur/thiosulfate metabolism pathways, such as methane monooxygensase (PmoB), methanol dehydrogenase (XoxF), intracellular sulfur-oxidation proteins (Dsr), sulfur-oxidation proteins (Sox), and adenylsulfate reductase (Apr), as well as for proteins associated to the carbon metabolism, such as malate, succinate and formate dehydrogenases (Mdh, Sdh, and Fdh). A total of 39 MOX and 86 SOX marker genes are present in the gene catalog.

5.1.3 Bathymodiolus microbiota is composed of SOX and MOX strains from several clades.

Diversity in natural populations of bacteria is characterized by cohesive associations among genetic loci that contribute to lineage formation and generate distinguishable genetic clusters beyond the species level (Shapiro and Polz 2014). The formation of niche-specific genotypes (i.e., ecotypes) has been mainly studied in populations of free-living organisms such as the cyanobacterium *Prochlorococcus* spp. (Kashtan et al. 2014). Here we consider a strain to be a genetic entity present in multiple hosts and characterized by a set of variants that are linked in the core genome. To study lineage formation in symbiont populations associated with *Bathymodiolus* mussels, we reconstructed the strain core genomes from strain-specific variants that show similar frequencies in a metagenomic sample.

The SNVs found in multiple samples and their covariation across samples were used for strain deconvolution of the core genomes using DESMAN (Quince, et al. 2017). This revealed that SOX is composed of eleven different strains with a mean strain core genome sequence identity of 99.52%. Phylogenetic reconstruction shows that the eleven strains cluster into four clades, which are separated by relatively long internal branches (**Fig. 6b**). Notably, 849 out of the total SNVs found on the SOX core genome (4.7%) could not be assigned to any particular strain. Thus, the resulting strain alignment is invariant for each of these positions and they are termed invariant SNVs from here on. For MOX, six strains with a mean core genome sequence identity of 99.88% were reconstructed. The phylogenetic network shows that the six strains cluster into two clades comprising three strains each (**Fig. 6e**). Of the total SNVs, 1 138 (24.4%) are invariant in the strain alignment. The overall MOX branch lengths are shorter than those of SOX. Furthermore, a pair of SOX strains differ by at most ~8 200 SNVs (0.44% different positions genome-wide) while two MOX differ by at most ~2 700 SNVs (0.19% different positions genome-wide) (**Fig. 6**), thus, SOX has higher genome diversity compared to MOX. We note that the SOX and MOX strain diversity is lower in comparison to coexisting strains of the free-living marine cyanobacterium *Prochlorococcus* - e.g., *Prochlorococcus* clades C1 and C3 differ in 3.2% of the positions (Kashtan et al. 2014). The absence of phylogenetic informative positions in the SOX and MOX ribosomal protein-coding genes serves as another evidence for the low SOX and MOX diversity (**Supplementary Information, Fig. 7**). Importantly, the observed difference in strain diversity between MOX and SOX cannot be explained by the difference in sequencing coverage (**Supplementary Information, Fig. 8**).

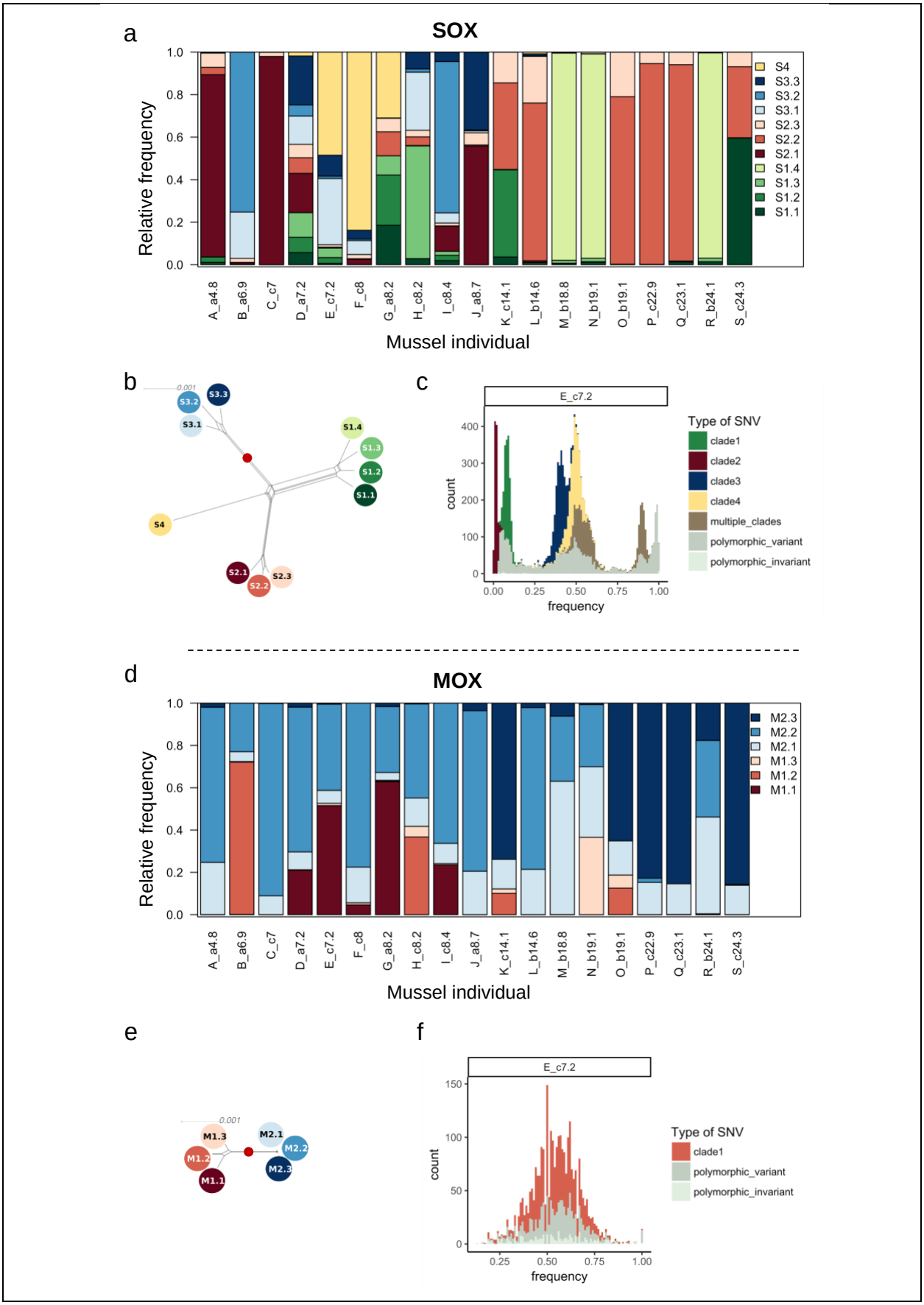

a
SOX
Relative frequency
1.0
0.8
0.6
0.4
0.2
0.0
S4
S3.3
S3.2
S3.1
S2.3
S2.2
S2.1
S1.4
S1.3
S1.2
S1.1
A_a4.8
B_a6.9
C_c7
D_a7.2
E_c7.2
F_c8
G_a8.2
H_c8.2
I_c8.4
J_a8.7
K_c14.1
L_b14.6
M_b18.8
N_b19.1
O_b19.1
P_c22.9
Q_c23.1
R_b24.1
S_c24.3
Mussel individual
b
0.001
S3.3
S3.2
S3.1
S1.4
S1.3
S1.2
S1.1
S4
S2.1
S2.2
S2.3
c
E_c7.2
count
400
300
200
100
0
0.00
0.25
0.50
0.75
1.00
frequency
Type of SNV
clade1
clade2
clade3
clade4
multiple_clades
polymorphic_variant
polymorphic_invariant
d
MOX
Relative frequency
1.0
0.8
0.6
0.4
0.2
0.0
M2.3
M2.2
M2.1
M1.3
M1.2
M1.1
A_a4.8
B_a6.9
C_c7
D_a7.2
E_c7.2
F_c8
G_a8.2
H_c8.2
I_c8.4
J_a8.7
K_c14.1
L_b14.6
M_b18.8
N_b19.1
O_b19.1
P_c22.9
Q_c23.1
R_b24.1
S_c24.3
Mussel individual
e
0.001
M1.3
M2.1
M2.2
M1.2
M2.3
M1.1
f
E_c7.2
count
150
100
50
0
0.25
0.50
0.75
1.00
frequency
Type of SNV
clade1
polymorphic_variant
polymorphic_invariant

Figure 6: Symbiont strain abundances (**a, d**), symbiont strain relationships (**b, e**), and example allele frequency spectra (**c, f**). **a, b, c,** 11 strains reconstructed for SOX. These cluster into four clades, with four, two times three and one strain per clade, labelled by shades of green, red, blue, and yellow. The strains differ by between 669 SNVs (strains S2.2 and S2.3, sequence identity 99.95%) and 8 171 SNVs (strains S3.2 and S4 sequence identity 99.36%). Minimum number of SNVs between strains of different clades is 6 451 (strains S1.1 and S2.1, sequence identify 99.49%). **d, e, f,** 6 strains reconstructed for MOX. These cluster into two clades, labelled by shades of red and blue. Strains differ by between 105 (strain M2.2 and M2.3, sequence identity 99.99%) and 2 677 SNVs (strain M1.1 and M2.1, sequence identity 99.81). The minimum number of SNVs differentiating strains from different clades is 2 224 (strains M2.2 and M1.3, sequence identity 99.85%). **a, d,** Stacked barplot of relative strain abundances for each individual mussel. Mussel individuals are labeled with an assigned letter (A-S), followed by the sampling clump (a, b or c) and the shell size (cm). **b, e,** Splits network of the strain genome sequences. Scale bar shows the number of differences per site. The red dots indicate the position of the root. **c, f,** Example of derived allele frequency spectra (sample E). Different colors represent different strain clades (see also **Supplementary Fig. 2).**

To study the community assembly at the strain level, we examined the strain distribution across individual mussels. Each SOX strain could be identified in between three and eight samples (frequency ≥5%; **Fig. 6a**). Only one or two strains were detected with a frequency of at least 5% in small mussels (≤7 cm), two to nine strains in medium-sized mussels (7.2 cm – 14.1 cm) and one to two strains in large mussels (14.6 cm – 24.1 cm). Notably, only strains from clades S1 and S2 are present in large mussels (≥14.6 cm). One of the large mussels (S) is an exception as it hosts three SOX strains and contains strains from both clades S1 and S2. Six mussels have one dominant SOX strain (frequency ≥90%). Five of these are large mussels (M, N, P, Q, R) and only one is a small mussel (C). The dominant strain is either S1.4, S2.1, or S2.2 (**Fig. 6a; Supplementary Table 1**). The MOX strain composition across mussels shows that each MOX strain occurs (frequency ≥5%; **Fig. 6d**) in four to 17 mussels and each mussel contains two to four MOX strains. Additionally, strains of clade M2 are dominant in ten of the mussels.

To investigate the degree of genetic cohesion within strain clades in the population, we studied the allele frequency spectrum (AFS) of each mussel. A visual inspection of the derived allele frequency spectra revealed multimodal distributions for both symbiont populations. The

modes reach high allele frequencies and are associated with the main phylogenetic clades; this suggests that the clades constitute cohesive genetic units (**Fig. 6c,f; Supplementary Fig. 1**). The presence of high-frequency modes is especially apparent for SOX in medium-sized mussels that contain multiple strains. To identify sample-specific strain sequences, we reconstructed dominant haplotypes (major allele frequency ≥90%) for the samples that contain a dominant strain (strain frequency ≥90%). By comparing dominant haplotypes among samples containing the same dominant strain, we found that these can contain between 42 and 74 differential SNVs (**Supplementary Table 1**). This suggests that the fixation of variants within individual mussels contributes to the observed population structure. Noteworthy, the downsampling of SOX to MOX coverage levels has an impact on the AFSs, which do not reveal as strong modes as the full coverage dataset (**Supplementary Fig. 3**).

Overall, our results revealed that the symbiont populations are composed of strains that cluster into a few clades, which appear to be maintained by strong cohesive forces. In addition, the strains are shared among multiple mussels, and multiple strains are capable of dominating different hosts. This suggests that stochastic processes are governing the symbiont community assembly, as previously proposed for other *Bathymodiolus* species (Ho et al. 2017).

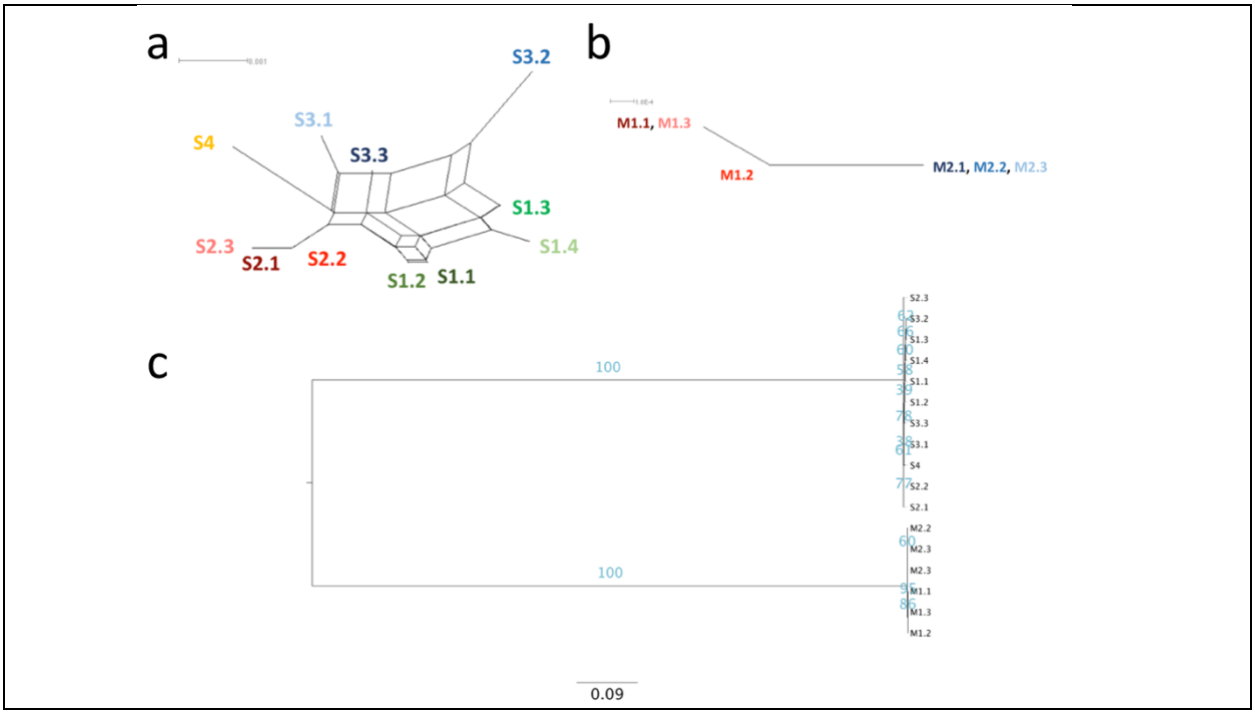


Figure 7: Splits networks and phylogenetic tree showing strain relationships based on a sequence alignment of ten ribosomal protein-coding genes. **a,** SOX splits network. The number of variant sites in the alignment is 26 out of 3 489 (0.7%), where nine sites are parsimony informative. Six variants are located at the first codon position and 20 variants are located at the third codon position. **b,** MOX splits network. The number of variant sites in the alignment is 3 out of 3 147 (0.1%), where none is parsimony informative. One variant is located at each codon position. **c,** Maximum likelihood phylogenetic tree reconstructed from the concatenated SOX and MOX merged alignment. The branch labels represent bootstrap values. The scale indicates the number of substitutions per site. The inferred length of the branch splitting SOX and MOX species ancestors is 1.75 substitutions per site.

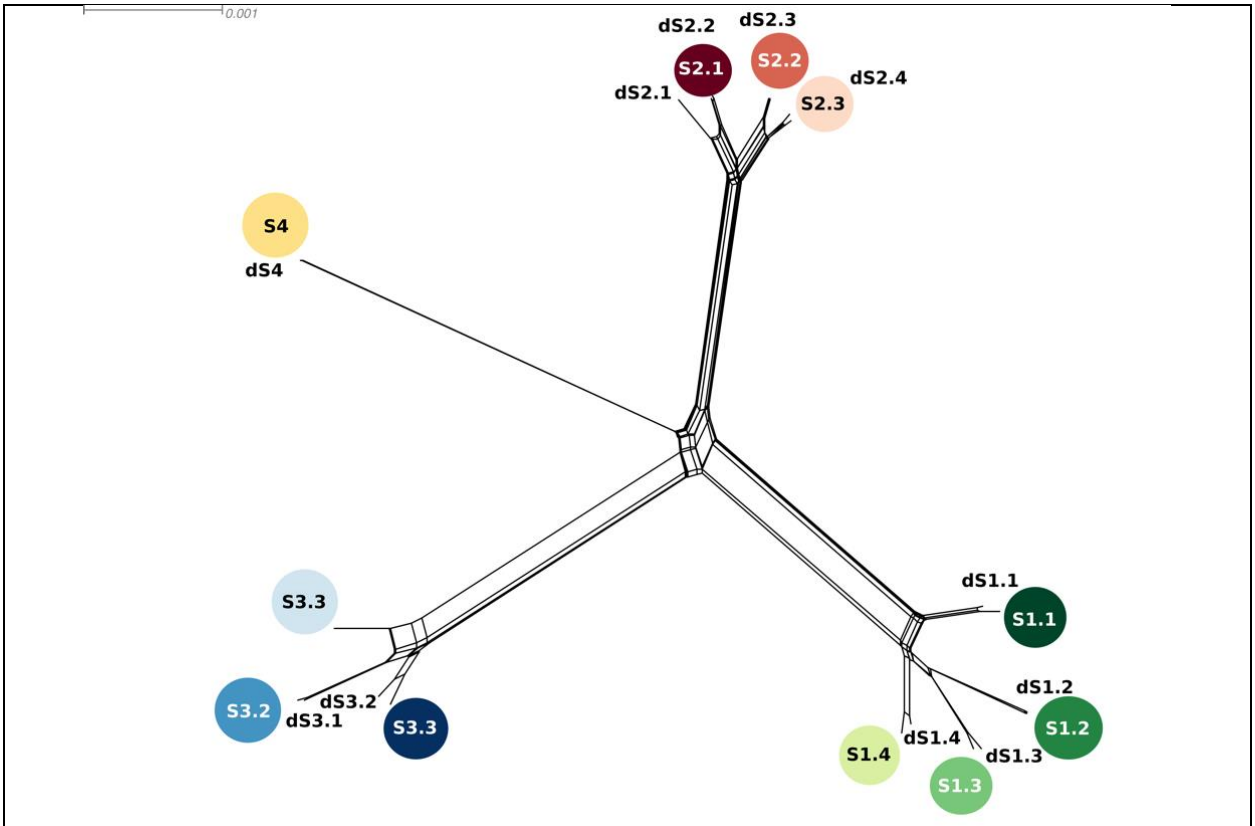


Figure 8: Splits network for SOX strains reconstructed using the full data set (colored) and SOX strains reconstructed after normalizing SOX to the median MOX coverage (black).

5.1.4 SOX strains evolve under purifying selection while MOX evolution is characterized by neutral processes

To study the evolution of SOX and MOX strains in *Bathymodiolus*, we examined the selection regimes that have been involved in the formation of cohesive genetic SOX and MOX units. The core genome-wide ratio of pN/pS is higher in MOX (pN/pS of 0.425) in comparison to SOX (pN/pS of 0.137), which indicates that the strength of purifying selection is higher for SOX. In addition, we estimated pN/pS for each of the symbiont core genes. This revealed that MOX genes are characterized by large pN/pS and small pS values, while SOX genes have small pN/pS and large pS values (**Fig. 9**). The relative rate of nonsynonymous to synonymous substitutions has been shown to depend on the divergence of the analyzed species (Rocha et al. 2006; Kryazhimskiy and Plotkin 2008). For populations of low divergence, SNVs comprise substitutions that have been fixed in the population and mutations that arose recently. The latter include slightly deleterious mutations that were not yet purged by selection, resulting in an elevated ratio of nonsynonymous to synonymous replacements. Thus, this ratio is not suitable for analyzing closely related genomes, which is usually the case when studying variation within bacterial species.

To circumvent the bias in pN/pS, we tested for differences in selection regimes in the evolution of SOX and MOX strains using the neutrality index (NI). NI is used to distinguish between divergent and polymorphic SNVs and to quantify the departure of a population from the neutral expectation. An excess of divergent nonsynonymous mutations (NI<1) indicates that the population underwent positive selection or an important demographic change in the past (Rand and Kann 1996). We estimated NI by considering two different levels of divergence and polymorphism. In the first level, all identified strains are considered as diverged taxonomic units; in the second level, we disregard the small-scale strain classification and consider only the clades as diverged taxonomic units (**Table 1**). Considering all strains as divergent, we observed a low NI^{MOX} (<1), which suggests that MOX evolved under a neutral (NI~1) or positive selection regime. NI^{MOX} increased when considering the clades as diverged, which suggests that the low NI^{MOX} observed at the strain level is the result of an excess of nonsynonymous SNVs within the strain clades that may constitute transient polymorphisms. Thus, the excess of nonsynonymous mutations observed for MOX is biased by the low level of divergence; hence, similar to the pN/pS ratio, it cannot serve as an indication for positive selection. On the other hand, we found that purifying selection is in action for SOX (NI^{SOX}>1), i.e., the divergent SNVs are enriched for synonymous SNVs in comparison to the polymorphic SNVs. Similar to MOX, when using the clades as divergent, NI^{SOX} slightly increases. This increase indicates that the SNVs that differ

between clades are more likely to be substitutions in comparison to those that differ among within-clade strains.

Altogether, these results suggest differences in the selection regimes during the evolution of the SOX and MOX strains. While the SOX core genome is shaped by purifying selection, we cannot detect deviation from the neutral expectation in the MOX core genome. These differences likely stem from the different divergence levels among the strains of both symbiont species populations. The association of SOX with *Bathymodiolus* mussels is considered to be ancient in chemosynthetic deep-sea mussels whereas the MOX association is thought to have evolved secondarily during *Bathymodiolus* diversification (Lorion et al. 2013). This agrees with the larger degree of divergence observed here for SOX. Since we observed no evidence for positive selection on the symbiont core genomes, we suggest that the strains constitute cohesive genetic units within one ecotype (Achtman and Wagner 2008), where all strains are functionally equivalent at the core genome level. Notwithstanding, the strains might be linked to differences in the accessory gene content, as observed, for example, in the free-living cyanobacterium *Prochlorococcus* spp. (Kashtan et al. 2014) and in SOX symbionts of other *Bathymodiolus* species (Ansorge et al. 2019).

Table 1. Neutrality index (NI) for the symbiont core genomes.

a, divergent SNVs are all those SNVs that differ between at least two strains, i.e., all identified strains are considered as diverged taxonomic units, and polymorphic SNVs are all the invariant SNVs. **b,** Divergent SNVs have the same state inside a strain clade and are not invariant and polymorphic SNVs are all the remaining, i.e., only the clades are considered as diverged taxonomic units.

a

	SOX		**MOX**	
	divergent	**polymorphic**	**divergent**	**polymorphic**
nonsynonymous SNVs	5004	990	2115	704
synonymous SNVs	10577	1450	1313	515
nonsynonymous SNVs/synonymous SNVs	0.47	0.68	1.61	1.37
NI	1.44		0.85	

b

	SOX		**MOX**	
	divergent	**polymorphic**	**divergent**	**polymorphic**
nonsynonymous SNVs	2549	3455	1041	1778
synonymous SNVs	6370	5657	649	1179
nonsynonymous SNVs/synonymous SNVs	0.40	0.61	1.60	1.51
NI	1.52		0.94	

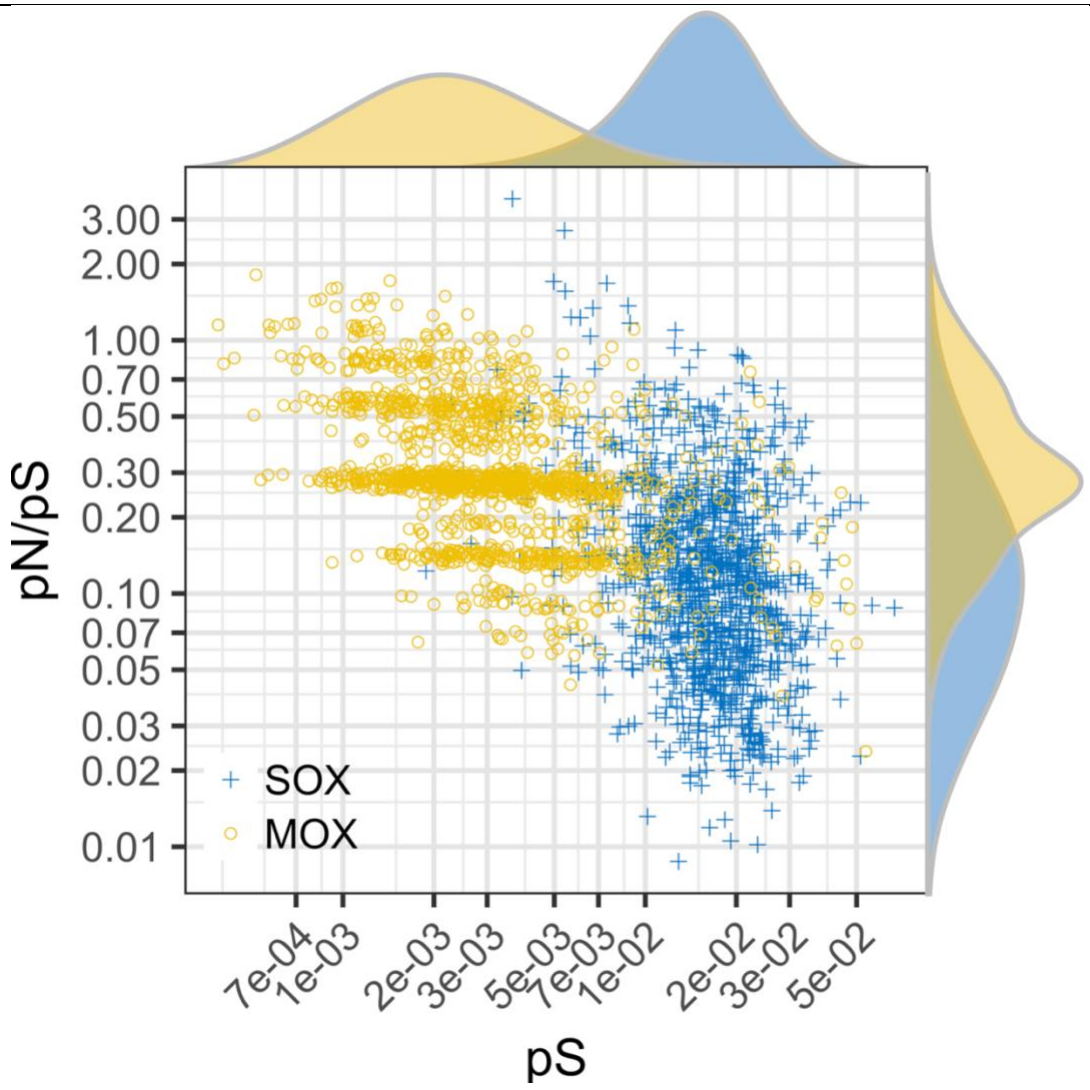


Figure 9: pS and pN/pS across genes for both symbionts. Density plots of the distributions are given in the margins. Note the log scale of both axes. Median MOX pS (0.0029) is smaller than median SOX pS (0.015) (Wilcoxon rank sum test, p-value < 10^{-6}). Median MOX pN/pS (0.28) is larger than median SOX pN/pS (0.12) (Wilcoxon rank sum test, p-value < 10^{-6}.). Only genes with SNVs are considered (1 117 genes for SOX and 1 359 genes for MOX).

5.1.5 Intra-sample diversity is higher for SOX than for MOX.

The association with the host limits the dispersal of bacterial populations where the association across generations is likely maintained by symbiont dispersal between host individuals. If symbionts are not continuously taken up from the environment, each individual host constitutes an isolated habitat over its lifetime (Costello et al. 2012). Geographic isolation between habitats results in genetic isolation and contributes to the formation of cohesive associations of genetic loci (Shapiro and Polz 2014). Previous studies showed that geographic isolation during vertical transmission can lead to the reduction of intra-host genetic diversity in the bacterial populations (Wernegreen 2015). Nonetheless, the degree of isolation remains understudied for horizontally transmitted microbes. To characterize the contribution of geographic isolation to strain formation in the *Bathymodiolous* symbiosis, we next studied the degree of genetic isolation. Our sample collection of mussels covering a range of sizes (and thus ages) enabled us to compare symbiont genome diversity among individual hosts of different age within a single sampling site, thus minimizing the putative effect of biogeography on population structure. The host species *B. brooksi* is ideal for such an analysis as it grows to unusally large sizes and possibly lives longer than many other *Bathymodiolus* species. To study differences in genome diversity of the two symbionts across individual mussels, we estimated the intra-sample nucleotide diversity (π) and the ecological measure α-diversity at the resolution of the SOX and MOX strains.

We found a high variability of π^{SOX} among different mussels (intra-sample π^{SOX} between $5.2x10^{-5}$ and $3.6x10^{-3}$, **Table 2, Fig. 10**). Furthermore, π^{SOX} and the SOX α-diversity are significantly positively correlated (ρ^2=0.98, $p<10^{-6}$, Spearman correlation, **Fig. 10a**); hence, the intra-sample strain diversity is well explained by the nucleotide diversity. The variability in π^{SOX} agrees with the three age-related groups observed before for the number of SOX strains across mussel size. Small mussels (≤7cm) and large mussels (14.6cm – 24.1cm) have a low π^{SOX} and harbor one to two strains. Medium-sized mussels (7.2cm – 14.1cm) have a high π^{SOX} and harbor two to nine strains. The community in the largest mussel is an exception, as it has a high π^{SOX}, similar to medium-sized mussels, which can be explained by the presence of three strains from two clades.

The MOX nucleotide diversity is significantly lower in comparison to SOX (intra-sample π^{MOX} between $5.6x10^{-6}$ and $7.0x10^{-4}$, **Table 2**, Wilcoxon signed rank test, p=0.015, **Fig. 10**). Similar to SOX, the MOX α-diversity is significantly positively correlated with π^{MOX} (ρ^2=0.89, $p<10^{-6}$, Spearman correlation) (**Fig. 10b**). One group of mussels harbors only MOX strains from clade 2

and is characterized by low MOX nucleotide diversity (A, C, J, L, M, P, Q, R, S, π^{MOX} between $5.6x10^{-6}$ and $2.1x10^{-5}$), while the other group habors MOX strains from both clades and is characterized by high MOX nucleotide diversity (B, D, E, F, G, H, I, K, N, O, π^{MOX} between $1.4x10^{-4}$ and $7.0x10^{-4}$). These groups are not associated with mussel size. Taken together, we observed a strong correlation between the nucleotide diversity π and α-diversity for both symbionts. Notably, π is based on all the detected SNVs whereas the α-diversity is based only on the strain composition and relatedness. Thus, the strong correlation demonstrates that the strain diversity captures most of the core genome-wide nucleotide diversity.

A comparison of the π values estimated here to other microbiome studies shows that higher π^{SOX} have been observed in other *Bathymodiolus* species (mean between 2.2 $x10^{-3}$ and $3.9x10^{-3}$) (Ansorge et al. 2019). The average SOX and MOX nucleotide diversity estimated here is within the range of values observed in the clam *Solemya velum* microbiome where the symbiont transmission mode is thought to be a mixture of vertical and horizontal transmission (Russell et al. 2017). Furthermore, our estimates are lower than those observed for most bacterial species in the human gut microbiome that are considered to be horizontally transmitted (Schloissnig et al. 2013).

5.1.6 Geographic isolation of bacterial communities associated with individual mussels.

Symbiont transmission mode is an important determinant of the community assembly dynamics (Bright and Bulgheresi 2010). For horizontally transmitted microbiota, similar community composition among hosts may develop depending on factors that affect the community assembly such as the environmental bacterial biodiversity or the order of colonization (Sprockett et al. 2018). To study the degree of geographic isolation between mussel hosts, we calculated genome-wide fixation index F_{ST} and the ecological measure β-diversity at the strain resolution across the metagenomic samples for the two symbionts. Small F_{ST} indicates that the samples stem from the same population whereas large F_{ST} indicates that the samples constitute subpopulations.

Our results revealed generally high pairwise F_{ST} values, indicating a strong genetic isolation between individual mussels (mean pairwise F_{ST}^{SOX} of 0.618, mean pairwise F_{ST}^{MOX} of 0.495, **Fig. 10**); hence, most mussels in our sample harbor an isolated symbiont subpopulation of SOX and MOX. The SOX β-diversity is significantly positively correlated with F_{ST}^{SOX}(ρ^2=0.7, $p < 10^{-6}$,

Spearman correlation). We observed groups of mussels that are characterized by a low pairwise F_{ST}^{SOX} within the group and a high pairwise F_{ST}^{SOX} with symbiont subpopulations from other mussels. This population structure is also represented in the distribution of β-diversity (**Fig. 10**). Thus, mussels from the same group harbor genetically similar SOX subpopulations and a similar strain composition. Examples are one group of mussels including L, O, P, and Q that contains only strains of clade S2 and another group including the mussels M, N, and R that contains only strains of clade S1 (**Fig. 10a**). Notably, both groups are composed of large mussels only that are characterized by a low π^{SOX}.

The distribution of pairwise F_{ST}^{MOX} revealed two main groups: one mussel group is characterized by high pairwise F_{ST}^{MOX} and low π^{MOX} while the other group is characterized by lower F_{ST}^{MOX} and high π^{MOX} (**Fig. 10b**). These correspond to the previously described groups, where one contains mussels with a low π^{MOX} and strains from clade M2 and the other group contains mussels with a high π^{MOX} and strains from both clades. We did not observe an association between MOX β-diversity and F_{ST}^{MOX} ($p>0.05$, Spearman correlation), which can be explained by the high proportion of invariant SNVs in MOX. Unlike SOX, the analysis of F_{ST}^{MOX} did not reveal groups of mussels with a low F_{ST}^{MOX} within the group and a high F_{ST}^{MOX} with other mussels. However, the pattern of β-diversity uncovered groups have a low β-diversity and a low nucleotide diversity. One group comprising large mussels (P, Q, S) is characterized by the presence of strain M2.3 and the absence of clade M1. Another group (A, C, J, L, M, R) containing mussels of different sizes is characterized by the dominance of strains M2.1 and M2.2 and the absence of clade M1. Thus, the comparison of strain composition across mussels revealed that the MOX population is structured similarly to SOX. However, unlike SOX, the MOX groups are not associated with specific mussel shell sizes.

The high F_{ST} values and the population structure we observed here reveal population stratification, that is especially pronounced for SOX. One possible factor that influences symbiont population structure is host genetics, whose impact on the composition of horizontally transmitted microbiota has been debated in the literature. Studies of the mammal gut microbiome showed that the host genotype had a contribution to the microbiome composition in mice (Benson et al. 2010), whereas the association with host genetics was reported to be weak in humans (Rothschild et al. 2018). Analyzing 175 SNVs in 12 mitochondrial genes, we detected no association between mussel F_{ST} and symbiont F_{ST} for any of the two symbionts (**Supplementary Information, Fig.**

11). Consequently, we conclude that the strong population structure observed for SOX and MOX cannot be explained by mussel relatedness (i.e., host genetics) or clump distribution.

Our results provide evidence for a strong genetic isolation between the symbiont subpopulations associated with individual mussels. This finding is consistent with the observed individual-specific symbiont strain composition. In contrast, much lower F_{ST} values were found for SOX populations in other *Bathymodiolus* species sampled from hydrothermal vents (mean F_{ST} per site between 0.05 and 0.17), which implies a weaker genetic isolation in these vents (Ansorge et al. 2019). Our analysis of cold seep *B. brooksi* data revealed SOX subpopulations with low genetic isolation that are observed using both F_{ST}, which takes all SNVs into account, and β-diversity at the level of strains. In contrast, only β-diversity disclosed subpopulations for MOX. Thus, strain-resolved metagenomics resolves similarities between individual mussel microbiomes below the species level.

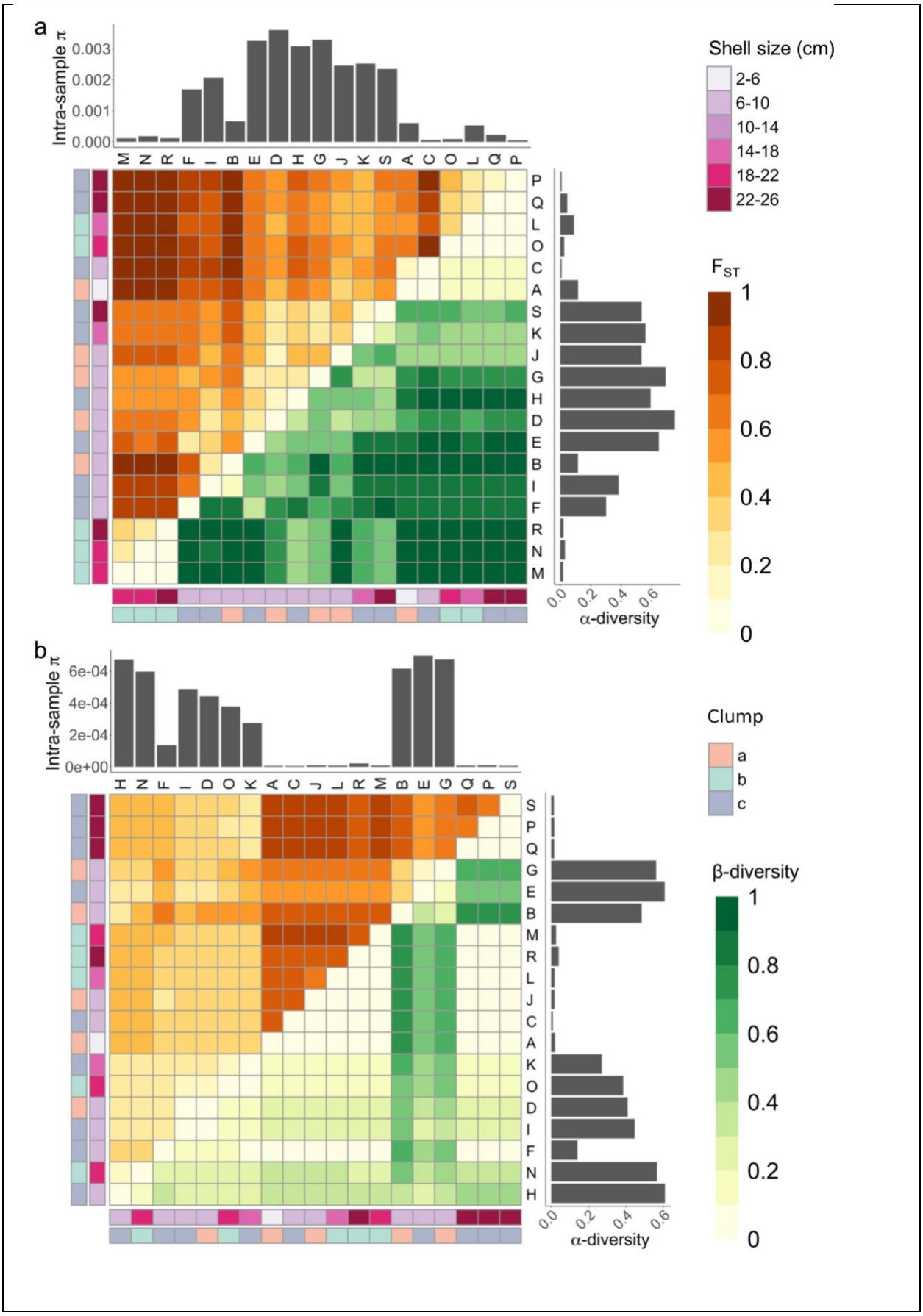

a
Intra-sample π
0.003
0.002
0.001
0.000
M N R F I B E D H G J K S A C O L Q P
P Q L O C A S K J G H D E B I F R N M
0.0 0.2 0.4 0.6
α-diversity
Shell size (cm)
2-6
6-10
10-14
14-18
18-22
22-26
F_ST
1
0.8
0.6
0.4
0.2
0
b
Intra-sample π
6e-04
4e-04
2e-04
0e+00
H N F I D O K A C J L R M B E G Q P S
S P Q G E B M R L J C A K O D I F N H
0.0 0.2 0.4 0.6
α-diversity
Clump
a
b
c
β-diversity
1
0.8
0.6
0.4
0.2
0

Figure 10: Symbiont population structure for a, SOX and b, MOX. Top left triangle: Intra-sample π and symbiont fixation index (F_{ST}) based on SNVs. Lower right triangle: α- and β-diversity based on reconstructed strains. Rows and columns are labelled by sample name, sample location, and shell size. Heatmap hierarchical clustering is based on Euclidean distance of F_{ST}. **a,** SOX: mean pairwise F_{ST} is 0.618. Two groups show an extreme degree of isolation: mean pairwise F_{ST} of group composed of M, N, R, is 0.313; mean pairwise F_{ST} of group composed of L, O, P, Q is 0.308; mean F_{ST} of sample pairs where one sample is M, N, or R and the other sample is L, O, P, or Q is 0.969. **b,** MOX: mean pairwise F_{ST} is 0.495. The clustering displays two groups: mean pairwise β-diversity of group composed of A, B, C, D, E, F, G, H, I, J, L is 0.099; mean pairwise β-diversity of group composed of K, M, N, O, P, Q, R, S is 0.383.

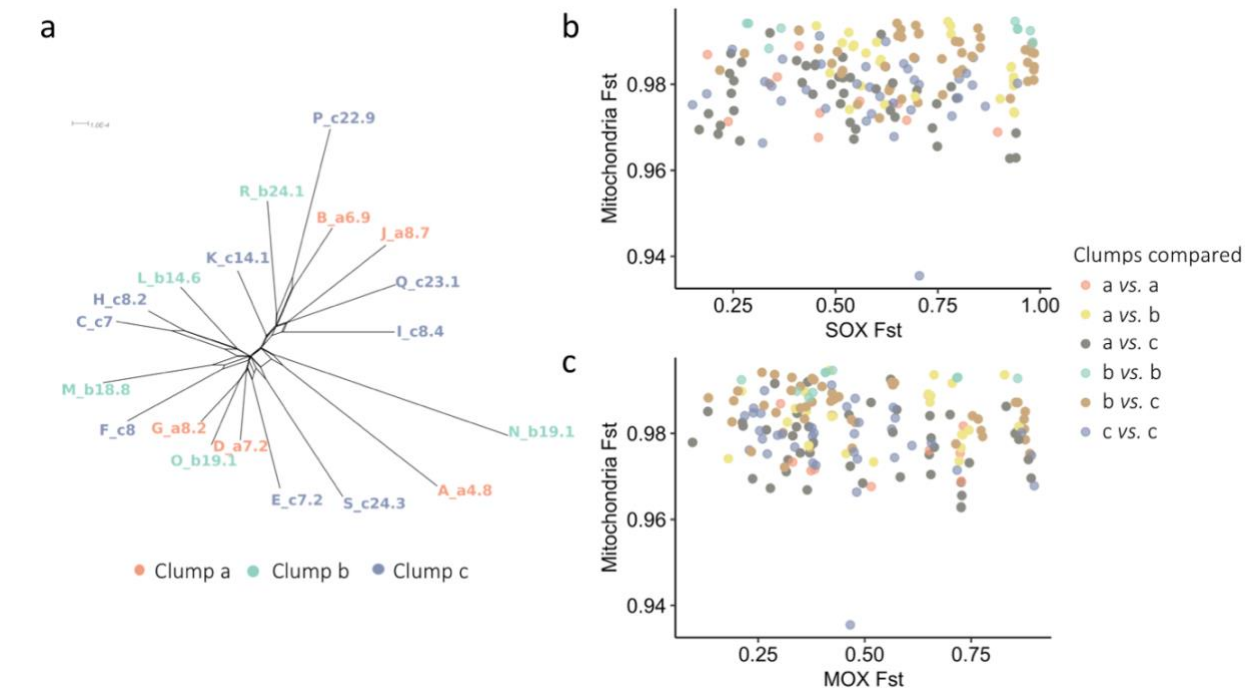


Figure 11: Mitochondria diversity. **a,** splits network (Huson 1998) (using Uncorrected P distances) based on the dominant haplotypes. We retrieved dominant mitochondrial haplotypes for each of the mussels by identifying SNVs with alternative frequency >0.5 and assigning the alternative nucleotide to these positions. **b,c,** relationship between **b,** SOX and c, MOX F_{ST} and mitochondria F_{ST}.

Table 2. Nucleotide diversity (π), Fixation Index (F_{ST}), and pN/pS calculations for both symbiont populations.

	SOX	**MOX**
Intra-sample π range	$5.2x10^{-5}$-$3.6x10^{-3}$	$5.6x10^{-6}$-$7.0x10^{-4}$
intra-sample π mean	$1.4x10^{-3} \pm 1.3x10^{-3}$ (s.d)	$2.7x10^{-4} \pm 2.8x10^{-4}$ (s.d)
intra-sample π median	$6.7x10^{-4}$	$1.4x10^{-4}$
Pairwise F_{ST} range	0.151-0.986	0.096-0.898
Mean pairwise F_{ST}	0.618	0.495
pN/pS	0.137	0.425

5.1.7 Individual-specific microbiota genetic variants cannot be explained by host genetics.

Mitochondrial genes have been previously used to identify lineages in *Bathymodiolus* mussels (Breusing et al. 2017). Here, we used genetic variants in the mitochondrion genome as a marker to investigate the contribution of mussel relatedness to the observed genetic isolation of the symbiont communities. Analysing the twelve mitochondrial genes, we detected 175 SNVs with a density of 15.6 SNVs/kbp (intra-sample $\pi^{Mitochondria}$ between $9.5x10^{-6}$ and $7.7x10^{-5}$, mean $3.3x10^{-5} \pm 2.2x10^{-5}$, s.d.). We found that most of the mitochondrial SNVs are fixed (frequency≥0.95) with an average of 99.8% fixed SNVs per sample. The phylogeny of the dominant mitochondrial haplotype for each sample shows no clustering of individuals according to sample clump or mussel size (**Fig. 11**), which indicates that the mussel individuals analyzed here belong to the same population. The high proportion of fixed SNVs results in high F_{ST} values (**Fig. 11b,c**), which is expected for vertically transmitted mitochondrial genomes. We further used the F_{ST} values to detect associations between mussel genetics and symbiont diversity. Our results reveal no association between mussel F_{ST} and symbiont F_{ST} for any of the two symbionts (**Fig. 11b,c**). Consequently, we conclude that the population structure observed for SOX and MOX cannot be explained by mussel relatedness or location.

5.2 Discussion

Our analysis revealed strong genetic isolation among subpopulations of symbiotic bacteria found in individual mussel hosts, indicating geographic isolation between mussels. The genetic isolation is independent of host genetics and of the mussel location in clumps. We hypothesize that the geographic isolation occurs through a restricted uptake of SOX and MOX symbionts from the environment over time. The lack of evidence for strong adaptive selection in SOX and MOX strains suggests that the inter-host population structure results from neutral processes rather than host discrimination against different strains. Here, we propose a neutral model for symbiont community assembly that explains how restricted symbiont uptake and colonization impose barriers to the symbiont dispersal, which can, over time, lead to inter-host population structure and contribute to the formation of cohesive genetic units within the symbiont population (**Fig. 12**). In our model, bacteria are acquired from the environmental symbiont pool in post-larvae mussels. The symbionts colonize every tissue in the beginning, to later restrict their localization to the gill tissue, as suggested by previous studies (Wentrup et al. 2013). Environmental symbiont acquisition in later developmental stages may occur due to symbiont loss and replacement driven by environmental changes or increased gill growth rate, which might explain the observed increase in symbiont diversity for middle-size mussels. The absence of clump-specific effects indicates the existence of a joint environmental pool across all sampled locations. The presence of a symbiont environmental pool was suggested before based on the detection of symbiont genes in adjacent seawater (Fontanez and Cavanaugh 2014; Ikuta et al. 2016). Nevertheless, the loss of central metabolic enzymes suggests that bacteria disperse in a dormant state (Ponnudurai et al. 2017). We hypothesize that the dormancy of free-living symbionts and the preservation of few symbiont cells inside bacteriocytes (Dubilier et al. 1998) contribute to the isolation of bacterial subpopulations inside the host cells from the overall population, which can lead to recombination barriers. Our results support the self-infection hypothesis (Wentrup et al. 2014), according to which, once the gill is first colonized, bacteria present in ontogenically older tissue infect newly formed gill filaments; thus, the uptake of symbionts from the environment is limited. In addition, decreased growth rate in older mussels may also lead to decreased symbiont uptake. This model plausibly explains the observed pattern of strong symbiont genetic isolation between mussels and of reduced SOX strain diversity in large mussels.

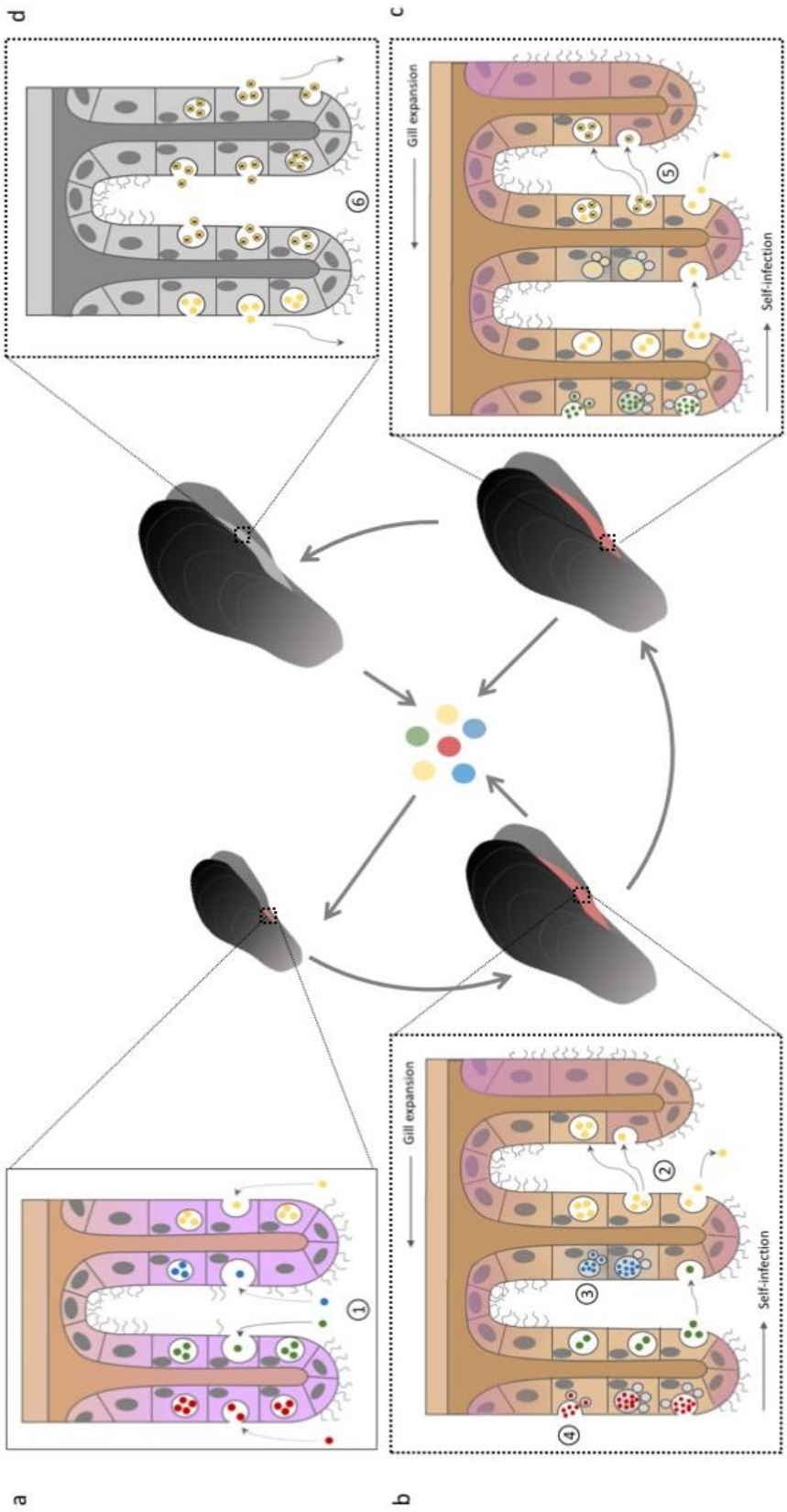

a
b
c
d
Gill expansion
Self-infection
Gill expansion
Self-infection
1
2
3
4
5
6

Figure 12: Symbiont colonization dynamics. **a,** The post larvae mussel gill does not take up endosymbionts until the gill presents several filaments and the gill epithelial cells reach a determined developmental stage (Wentrup et al. 2014). At this time point, the filaments are simultaneously infected by different strains via endocytosis (1). This imposes the first bottleneck in the symbiont population, since most likely, not all the strains from the environmental pool can infect the tissue**. b,** Bacteria are released from the host tissue to the environmental pool. As the mussel grows, new filaments are continuously formed in the gill throughout the mussel life span (growing cells shaded in purple). The new tissue is colonized by a self-infection process (Wentrup et al. 2014), which involves infection of the newly formed filaments via endocytosis with bacteria that are released from old tissue via exocytosis (2). The spatial distribution of strains within the gill tissue also supports self-infection (Ikuta et al. 2016). The continuous self-infection process imposes serial founder effects that lead to a reduction in strain diversity, which is mostly driven by drift. Additional sources of diversity loss are: tissue replacement (3) and regulated lysosomal digestion of symbionts (Sun et al. 2017) (4). **c,** In older mussels, a unique strain dominates the gill. In addition, *de novo* mutations occur in symbiont genomes (marked by x). Due to serial founder effects within the same mussel, those variants can be quickly fixed inside the mussel (5). **d,** As the mussel dies, bacteria are released from the gill, going back to the environmental pool (6), as reported for the environmentally acquired symbionts from the tube-worm belonging to the genus *Riftia* (Klose et al. 2015).

Notably, our results are in contrast to a recent study on other *Bathymodiolus* species from hydrothermal vents, that concluded that SOX populations from individual mussels of the same site intermix (Ansorge et al. 2019). This contrast may be explained by differences in the symbiont abundance in the seawater, which is expected to play a role in the colonization process. Our samples originate from a cold seep site with low mussel density (**Fig. 4**); thus, the concentration of symbionts in the surrounding seawater may be correspondingly low. The low symbiont abundance would cause a low probability of later infections and a prevalence of self-infection. In contrast, the symbiont abundance in the seawater at large and densely populated mussel beds at hydrothermal vents is expected to be higher, resulting in a higher probability of later infections.

The colonization of new filaments over the mussel lifespan via self-infection entails serial founder events on the bacterial population. Throughout this process, new mutations arising in the symbiont population during the lifetime of the mussel can reach fixation due to genetic drift following population bottlenecks. This process is expected to lead to a reduction of symbiont genetic diversity over the mussel lifetime. Thus, individual mussels develop into independent habitats that harbor individual symbiont subpopulations, which are genetically isolated from other mussel-associated subpopulations and from the environmental pool. The evolution of vertically transmitted endosymbiont populations is similarly affected by serial founder effects (Reuter et al. 2005), as we suggest here for horizontally transmitted bacteria. However, migration between host-associated subpopulations and the environmental pool results in an increased effective population size for horizontally transmitted bacteria; thus, the population is not subject to the fate of genome degradation as commonly observed in vertically transmitted symbionts (Boscaro et al. 2017). Serial founder effects and recombination barriers due to geographic isolation are important drivers of lineage formation in bacteria (Achtman and Wagner 2008). Reduction of genetic diversity due to transmission bottlenecks is considered a hallmark of pathogen genome evolution (Didelot et al. 2016); examples are *Yersinia pestis* (Gonzalez et al. 2015) and *Listeria monocytogenes* (Zhang et al. 2017). Our model demonstrates that, similar to pathogenic bacteria, genome evolution of bacteria with a symbiotic lifestyle can be affected by serial founder effects due to self-infection.

6 Chapter II: reconstructing the population pan-genomes of symbionts.

Diversity in bacterial populations is comprised of single-nucleotide variants (SNVs) and gene content variation. In chapter I, we analyzed SNVs in core genes and showed that geographic isolation impacts importantly the diversity of deep-sea mussel symbiont populations. Gene content variation can be vertically inherited or acquired horizontally from mobile genetic elements (MGEs). Therefore, gene acquisition via horizontal gene transfer (HGT) is theoretically less restricted by geographic isolation. Here, we study how symbiont population structure impacts gene gain and loss dynamics in the population pan-genomes. Moreover, we are interested in studying the implications of differences in accessory genomes among strains for microbial community composition across individual hosts. To that end, we inferred the population pan-genomes of the two co-occurring symbionts by combining the metagenomic assemblies of 19 mussel samples.

6.1 Results.

6.1.1 SOX and MOX symbiont populations are characterized by differences in their pan-genome sizes.

Using a network approach to reconstruct pan-genomes from metagenomic samples, we could initially recover 2,570 and 2,907 genes for SOX and MOX populations, respectively. The quality of these pan-genomes has been assessed by studying the distribution of two different gene cluster features across the pan-genome layers; these are the cluster size and the cluster sequence identity (**Table 3**). As expected for genes that are not present in every single strain, the gene cluster size decreases as the layer of the network increases. This means that the genes could not be recovered from every sample, therefore supporting their status as accessory. Additionally, the median sequence identity of the clusters is close to one (>0.99), and this indicates that the gene clusters added to the pan-genomes are not very likely affected by contamination – i.e., presence of homologous genes from a different metagenomic species. This further supports the robustness of the reconstructed pan-genome network. Note that the gene cluster size standard deviation (SD) is relatively high in some instances, indicating the presence of potential paralogous genes. Some paralogous genes are multicopy genes that can be transferred at a relatively high rate within and between genomes -e.g., transposons. These might

create spurious associations between sequences that do not belong to the same genome. To prevent erroneous linkages, we do not consider non-redundant genes that have multiple copies per sample as seed to expand the next layers of the network. Nevertheless, as they belong to the bacterial genomes, we included them into the pan-genomes.

We found that the pan-genome of SOX consists of five layers, where different numbers of genes are added (**Table 3**). From the total number of accessory genes recovered in the network, 83 genes were discarded as they showed a coverage of 0 across samples. Additional three genes were rejected, as they could also be added to the MOX pan-genome, and therefore their origin was unclear. In total, the reconstructed pan-genome for SOX population comprises 2,484 genes; 1,408 of them were previously identified as Single-Copy Core Genes (SCCG) and 1,076 are additional genes, which are potentially accessory. In total, the length of the genes in the SOX pan-genome is 2.27Mb.

On the other hand, the MOX population pan-genome is larger than that of SOX, with 2,866 genes, where 2,443 are SCCG and 423 are additional genes. Similarly to SOX, the coding fraction of the MOX pan-genome consists of 2.24Mb. Additionally, 38 genes were discarded, as no coverage was found across samples. Moreover, 3 genes were rejected as they were also linked to the SOX pan-genome. Note that the core genomes reported here contain fewer genes than the core genomes described in chapter I. This is because samples Dsc1-4 were used to construct the NRGC, but not for later analyses, including the pan-genome reconstruction.

To have an estimation of the fraction of the population pan-genome that we could recover from our total sampling set, we calculated the number of new genes (the presence of a gene is positive when its coverage is at least 5% of the median coverage of the core genes) that are added to the pan-genome for each newly sampled mussel in the two species (**Fig. 13**). We observed that, in the two species, the number of newly added genes per new sample reaches a plateau by sample size of 10 in SOX and 14 in MOX. This suggests that, most genes present in the population pan-genomes could be recovered with this sample size, although, this does not necessarily reflect the pan-genome size of the entire species.

Table 3: SOX and MOX population pan-genomes quality statistics shown per pan-genome layer. Cluster size indicates the number of genes that are comprised in each homology cluster, for which a representative is selected and included in the non-redundant gene catalog, and later in the pan-genome. Seqid: sequence identity of the multiple alignment of gene clusters. SD: standard deviation.

layer	#genes	Median cluster size	SD cluster size	Median seqid	SD seqid
SOX					
0	1408	19	5.04	0.99	0.01
1	907	13	6.70	1.00	0.02
2	202	3	39.06	1.00	0.06
3	42	4.5	6.09	1.00	0.01
4	6	3	14.81	1.00	0.09
5	5	1	1.20	1.00	0.00
MOX					
0	2443	19	3.87	1.00	0.01
1	427	9	7.62	1.00	0.02
2	26	5	5.61	1.00	0.00
3	5	4	3.50	1.00	0.00
4	6	3.5	1.11	1.00	0.00

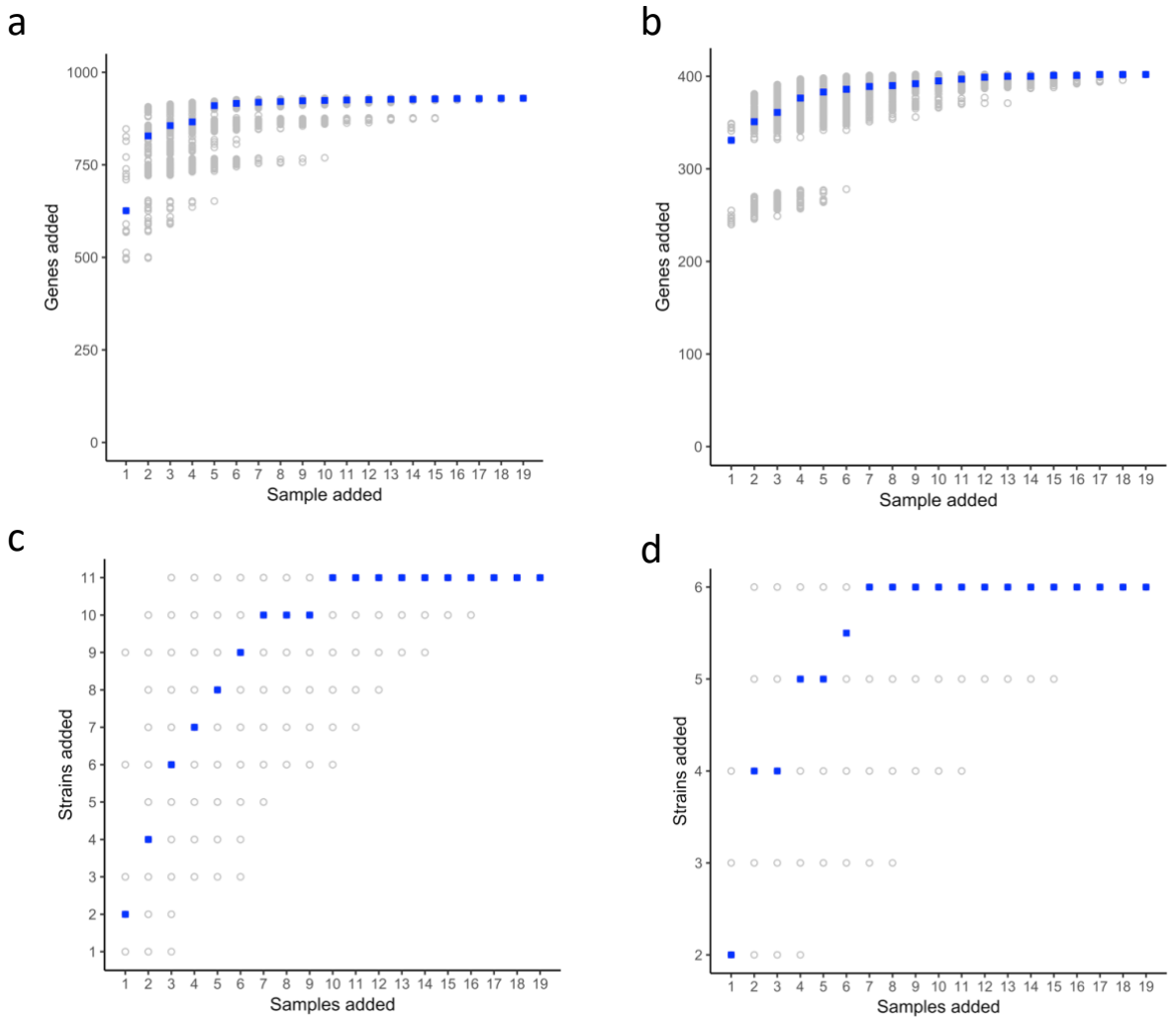


Figure 13: SOX and MOX population pangenomes size and strain composition across mussel samples. Number of new genes added to the **a,** SOX and **b,** MOX pan-genomes per each newly sampled mussel. Number of **c,** SOX and **d,** MOX strains added for each sampled mussel. Each gray dot corresponds to an individual data-point for a specific permutation (analysis performed by permutating 1,000 times the order of the added mussels). Blue squares correspond to the median number of genes added across permutations.

6.1.2 Strain clades are characterized by differential gene content.

In chapter I, we show that the deep-sea mussel symbiont populations are composed of multiple strains, which we characterized by a set of polymorphisms in their core genomes. Here, we aim to additionally identify the gene content of the strains. We previously proposed that the nucleotide diversity observed in the symbiont populations is mainly the product of neutral processes -namely, strong genetic isolation and genetic drift- rather than positive selection. Nevertheless, this conclusion was drawn regarding core genes. To study whether the presence of strain-specific functions could raise the frequency of determined strains in the population, we study the selective pressure on the accessory genome as well. Because the symbionts have a free-living phase, mobile gene elements mightcan potentially shape the symbiont genomes regardless of the strong population structure observed, by introducing genetic diversity in the form of newly acquired genes. Therefore, here we aim to study the contribution of gene gain and loss to the genetic diversity of symbiont populations relative to that of single point mutations.

We first studied the degree of isolation (F_{ST}) between samples by using gene diversity, which is based on the measured frequency of each gene within and between samples (see methods). We found that the overall degree of isolation based on the presence of genes across samples is lower than that of the previously reported (mean gene content F_{ST}^{SOX}: 0.347, mean gene content F_{ST}^{MOX}: 0.124). The lower F_{ST} found in accessory genes compared to core genes indicates that strains differ more in their core-genes polymorphisms than in their gene content. The gene content pairwise F_{ST} shows a very similar clustering pattern to the one previously observed, where samples cluster according to strain composition (**Fig. 14**). Such a strong signal further supports the successful recovery of accessory genes from metagenomes and the lack of contamination. Just as in the comparison between SNVs-based F_{ST} and β-diversity, larger clusters are formed when looking at gene content-based F_{ST} instead of SNVs-based F_{ST} (e.g., the annexing of samples A and C to the cluster formed by O, L, P, Q in SOX; **Fig. 14a**). This observation is explained by the significant positive correlation between gene diversity F_{ST} and β-diversity in the two symbiont populations (**Fig. 14c,d**). Such a strong positive correlation suggests that gene content differences are mainly found among strain clades, since the degree of genetic isolation among subpopulations containing strains from the same clade is small.

We studied how samples and accessory genes cluster according to their coverage across mussels. We found in both bacterial species, but especially in SOX, a very clear presence/absence pattern, where samples group according to the presence of genes and clades

(**Fig. 15, 16**). This observation further supports the hypothesis that gene content is mainly a characteristic of strain clades, rather than of the strains themselves. In addition, we identified, for SOX and MOX, two groups of 114 and 9 genes, respectively, that are found in all samples, with coverage similar to the SCCG median coverage (**Fig. 15, 16**). Those genes were previously discarded from the core pool, as they were coverage outliers, but will be added to the core set in the following analyses.

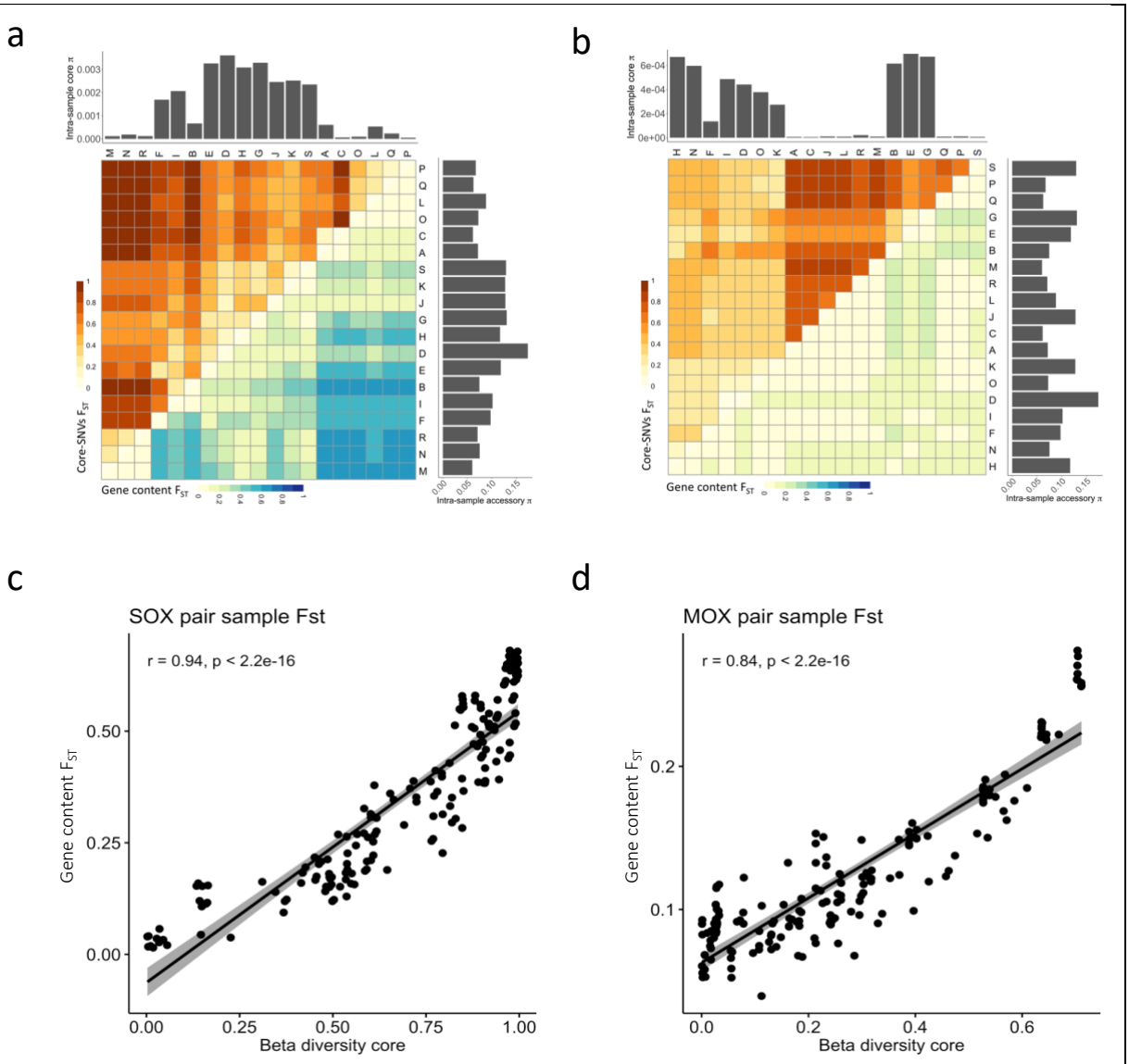


Figure 14: Symbiont population structure for **a,** SOX and **b,** MOX. Top left triangle: intra-sample π and symbiont fixation index (F_{ST}) based on SNVs. Lower right triangle: F_{ST} and gene diversity based on gene content. Rows and columns are labelled by sample name. Heatmap hierarchical clustering is based on Euclidean distance of core-SNVs F_{ST} Correlation between gene content-based F_{ST} and β-diversity for **c,** SOX and **d,** MOX symbionts rho and p-values shown for Spearman's correlation test. Linear regression lines and 95% confidence intervals are shown.

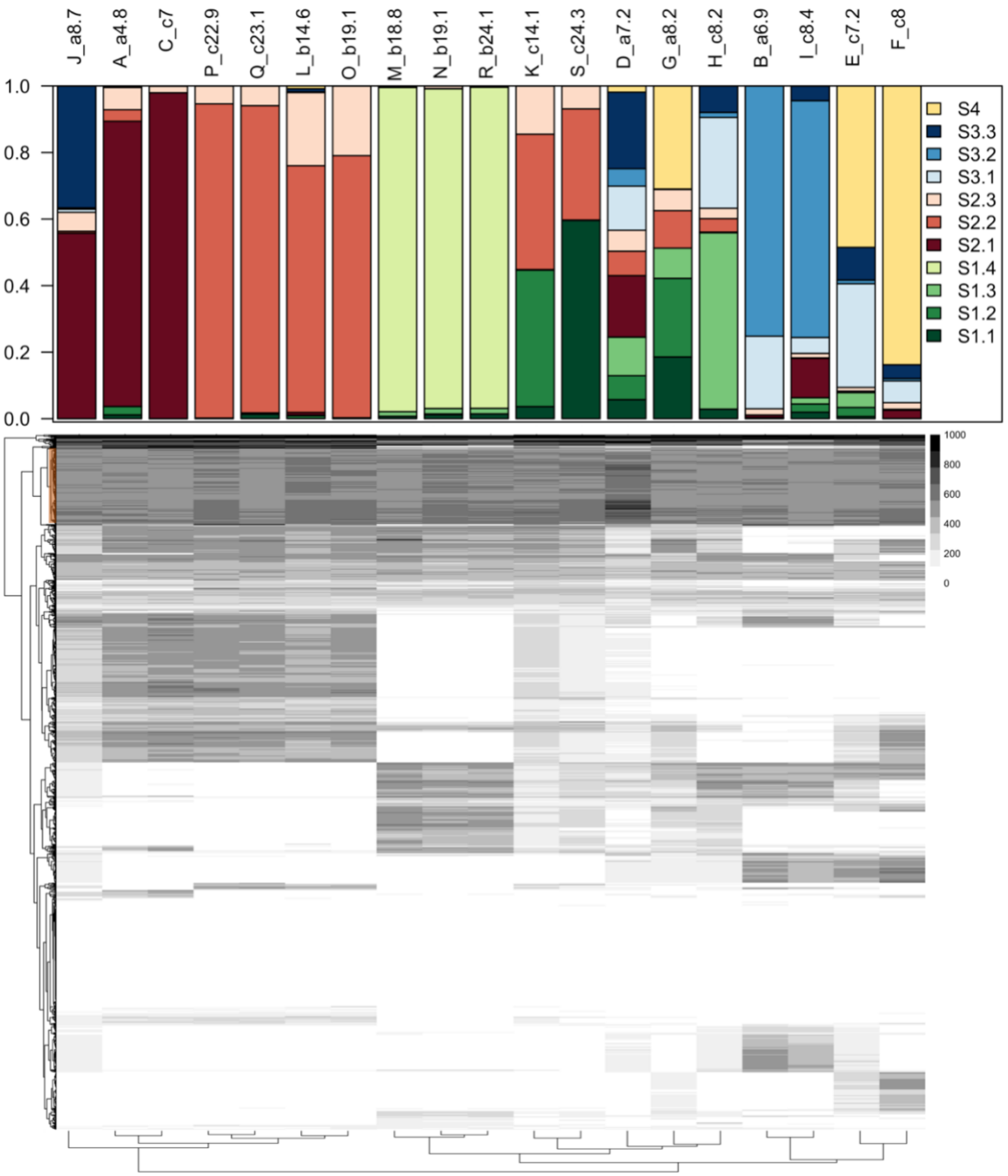


Figure 15: Heatmap of SOX accessory genes coverage across samples. For representative purposes, coverage is limited to 1000x. Genes and samples are hierarchically clustered. The coverage shown is normalized to the median SCCG coverage among samples and 18 genes are not shown as they show no coverage across samples after normalization. The orange square covers a cluster of 114 genes which are covered across all samples and will be added to the core gene set.

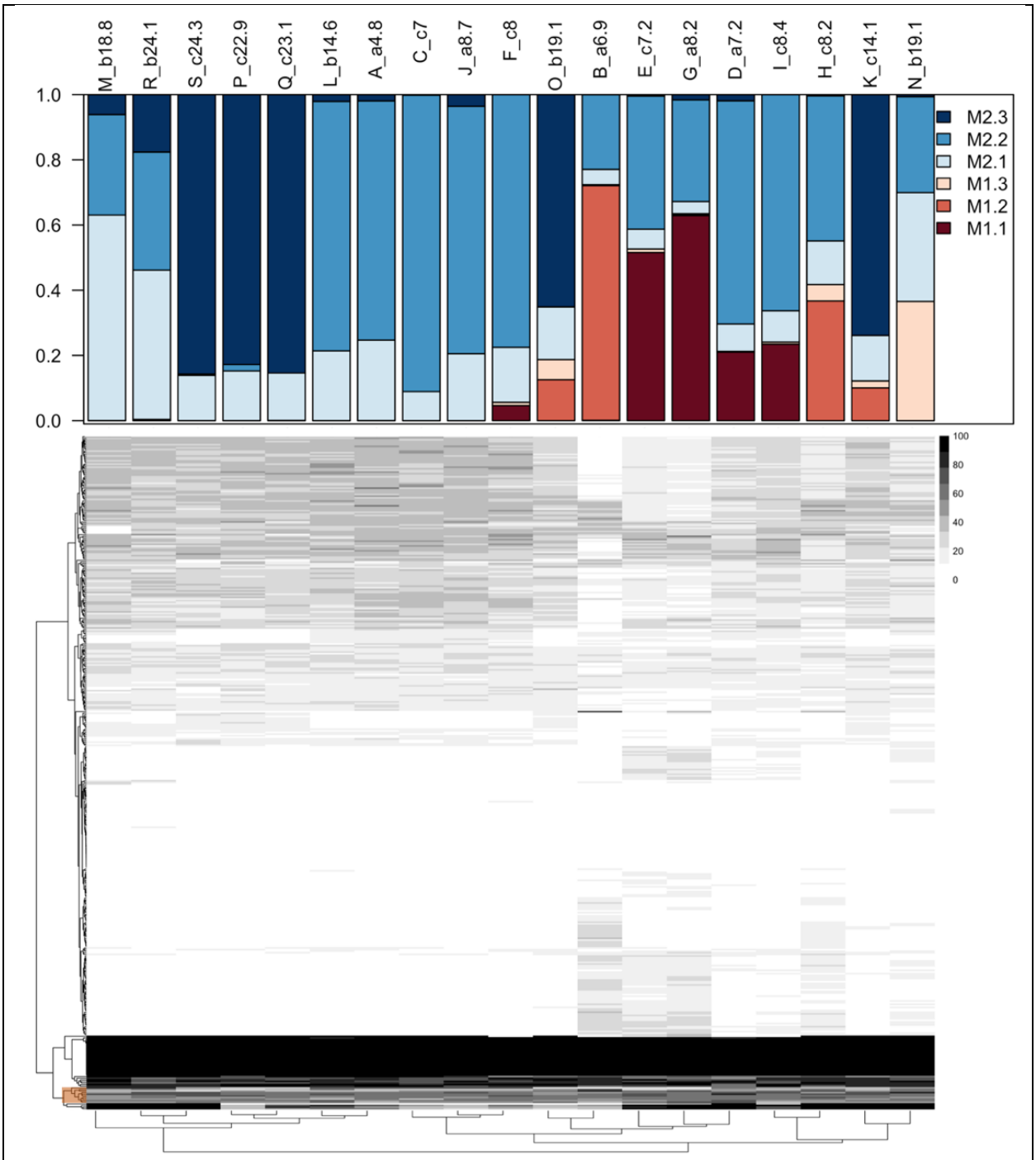


Figure 16: Heatmap of MOX accessory genes coverage across samples. For representative purposes, coverage is limited to 100x. Genes and samples are hierarchically clustered. The coverage shown is normalized to the median SCCG coverage among samples and 4 genes are not shown as they show no coverage across samples after normalization. The orange square covers a cluster of 9 genes which are covered across all samples and will be added to the core gene set.

To link the accessory genes to the core genomes of the different symbiont strains, we identified samples with a dominant strain – i.e., samples where the frequency of the dominant strain is at least 0.7- and determined, in these samples, the presence of accessory genes whenever their frequency is of at least 0.5 (considering that the maximum frequency is the sample-specific SCCG median coverage).

We linked 731 genes (76% of accessory genes) to 5 SOX strains dominant in 12 mussel samples (**Fig. 17 Supplementary Fig. 2**). Two of these strains belong to clade S2 (S2.1 and S2.2). We could identify the accessory genes of one strain for each of the remaining three clades (S1.4, S3.2 and S4). Strain assignment of genes was performed by merging the genes found across all samples where the particular strain was dominant. Similarly, clade-specific genes result from merging genes that are found across strains belonging to the same clade.

First of all, we found that the number of sample-specific genes is always smaller than 11, indicating that strain gene content assignment is highly consistent across samples (**Supplementary Fig. 2**). The clades differ in the number of accessory genes; S2 is the clade with a larger number of accessory genes, with 418 genes. It is followed by S4, with 322 genes. S3 and S1 contain almost the same number of accessory genes, with 295 and 293 genes, respectively. As expected, we observed that the two strains reconstructed for clade S2 share most of their accessory genome, with 356 genes common among strains, 20 genes present in S2.1 not found in S2.2 and 27 genes found in S2.2 not found in S2.1. Moreover, S2.1 and S2.2 contain 8 and 17 exclusive genes, respectively. Additionally, S2 contains the most genes exclusive for the clade, where the strains share 133 genes that can be found nowhere else. This is followed by S3, which contains 71 genes that are not found elsewhere. S1 contains 65 clade-specific genes. And finally, S4 contains 59 genes that could not be found in any other clade. We also identified 68 genes present across all samples. 18 of these genes were found to have a coverage that is larger than the maximum coverage found among core genes, and therefore they were classified as potential paralogous genes. We found that mutations exceed gene content variation. For instance, the most distant strains (based on core SNVs), S3.2 and S4, differ by 8,171 SNVs, and 315 genes (171 only present in S4 and 144 only present in S3.2). This is also observed among strains of the same clade -e.g., S2.1 and S2.2 differ by over 669 SNVs, while they have less than 20 strain-specific genes.

For MOX, we could assign 276 accessory genes (66.67% of the accessory genes) to 3 different strains covering the two phylogenetic clades -M2.2, M2.3 and M1.2- across 10 different

samples (**Fig. 18**). Similarly to SOX, the number of sample-specific genes never exceeds 10 (**Supplementary Fig. 1**) and the strains within the same clade share most of their genes. M2.2 contains 8 genes not present in M2.3 while M2.3 contains 13 genes not present in M2.2. M1 and M2 contain 144 and 211 accessory genes, where 65 are M1-specific and 132 are M2-specific. In addition, 79 genes were shared among all three strains. 34 out of these 79 genes were found to be potential paralogs, since their coverage is larger than the maximum coverage found among core genes.

After re-defining both core gene sets, we found that the pan-genome of MOX population contains a larger number of total genes than that of SOX population with 2,866 and 2,484 genes, respectively. In addition, we observed that a larger fraction of the SOX pan-genome is accessory (38.73% in SOX versus 14.45% in MOX). Together, this may suggest that SOX underwent genome erosion to a bigger extent. Moreover, MOX contains a larger number of paralogs than SOX, what may reflect the presence of a larger number of transposons, as expected for a species that is closer to a more facultative stage of the symbiosis (Newton and Bordenstein 2011). We also found that in the two bacterial species, the gene content is mostly different among strain clades. The absence of many strain-specific genes within the clades suggests that gene acquisition might not be a very pervasive phenomenon in the population, as differences in gene content are concentrated in deeper branches of the symbionts phylogenies.

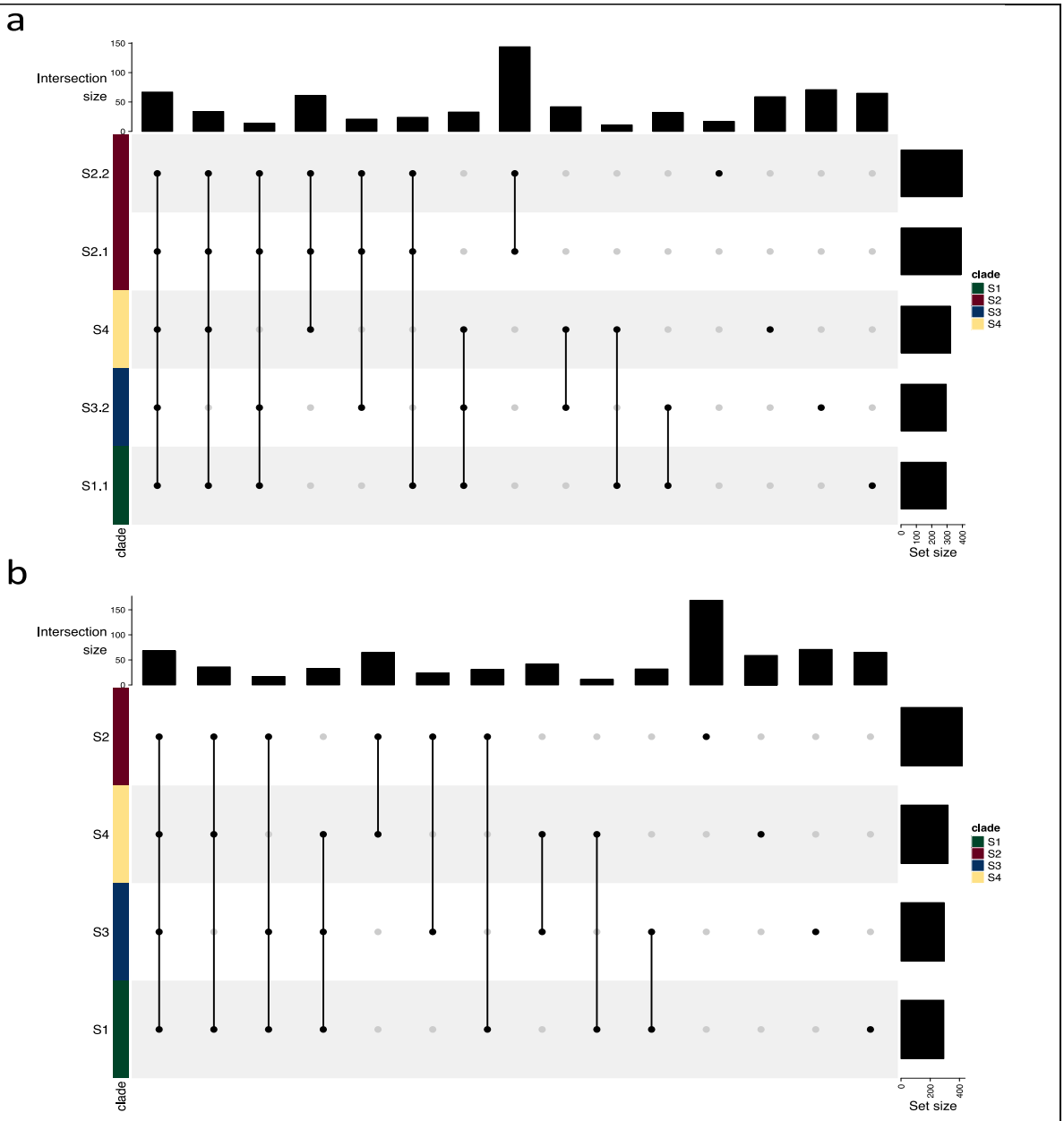


Figure 17: Interaction plots showing intersects for the gene content of 5 SOX strains. A total of 731 genes could be assigned to 5 strains across 12 samples. Note that only intersections larger than 10 are displayed. **a,** Intersects among the 5 recovered strains. These strains have been reconstructed by merging the genes found across all samples dominated by the same strain (**Supplementary Fig. 2**). **b,** Intersects among the four different clades. Clade-specific gene sets have been reconstructed by merging the gene sets from strains belonging to the same clade.

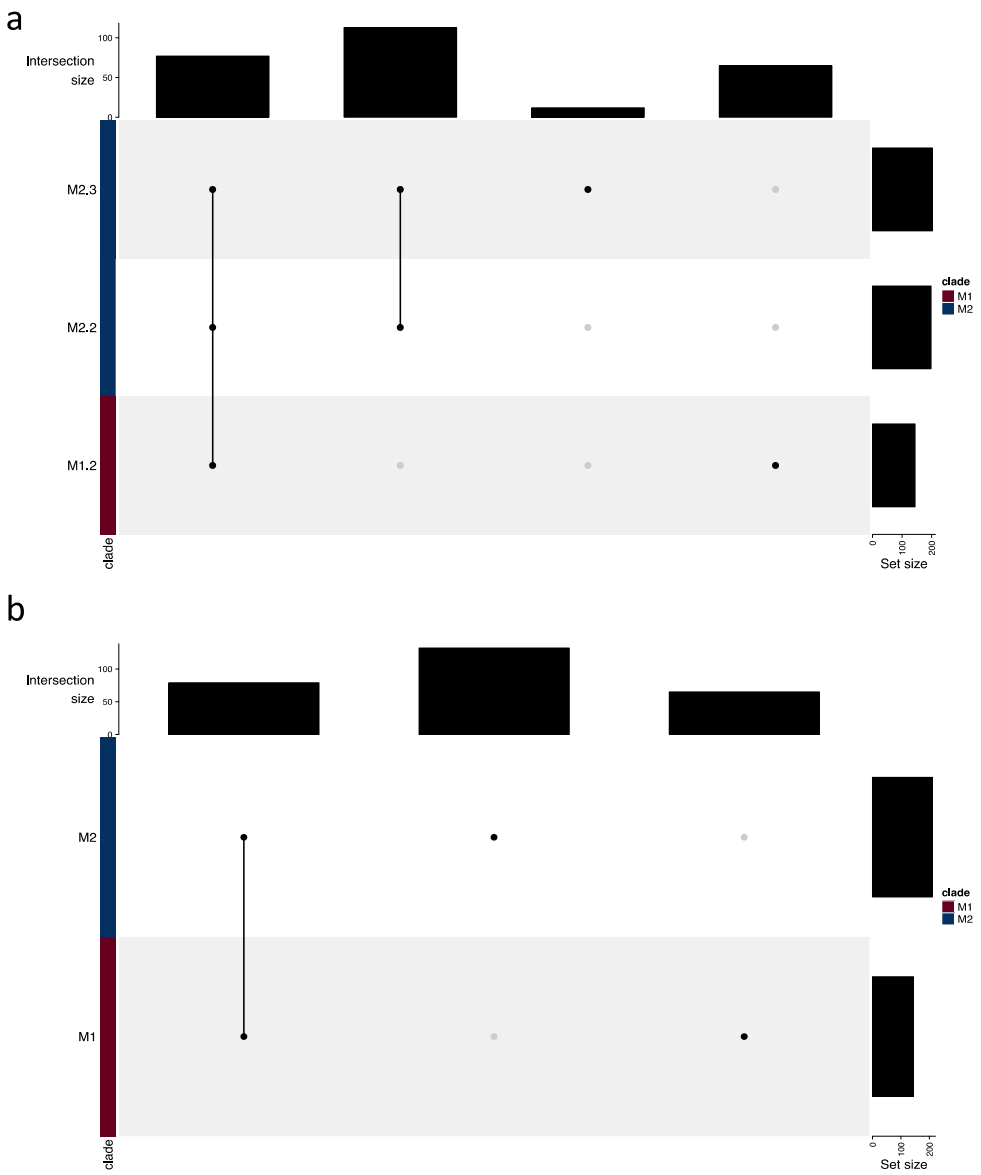


Figure 18: Interaction plots showing intersects for the gene content of 3 MOX strains. A total of 276 genes could be assigned to 3 strains across 10 samples. Note that only intersections larger than 10 are displayed. **a,** Intersects among the 3 recovered strains. These strains have been reconstructed by merging the genes found across all samples dominated by the same strain (**Supplementary Fig. 2**). **b,** Intersects among the two different clades. Clade-specific gene sets have been reconstructed by merging the gene sets from strains belonging to the same clade.

6.1.3 Core and accessory genes show differential GC content and pN/pS.

Core and accessory genomes have been shown to differ in their GC content. The primary force of evolution shaping the pangenomes is highly debated (Andreani et al. 2017; Bohlin et al. 2017; McInerney et al. 2017). Because GC-content is linked to bacterial ecology, it is characteristic for the species and can be used as a proxy to identify horizontally acquired genes (Lawrence and Ochman 1997). To explore the different hypotheses that link sequence composition and pan-genome element type (core or accessory), and to better understand the origin of the accessory genomes in symbiont populations of *B.brooksi*, we studied the GC content and selection regimes acting on accessory genes and core genomes.

Both SOX and MOX core genomes have a relatively low GC content (mean GC_{SOX}: 0.374, mean GC_{MOX}: 0.378), as expected for symbiotic bacteria (McCutcheon and Moran 2012). The unidentified MGS3 genome has a higher GC-content (mean GC: 0.423), suggesting that this species has a weaker association with the mussel host than the association previously described for MOX and SOX bacteria. Surprisingly, we found that although their medians are very similar, the core genes GC content distributions are significantly different for MOX and SOX bacteria (Wilcoxon rank sum test p-value = 0.03541). (**Fig. 19**). We found that the GC content of the accessory genomes belonging to both bacterial symbionts is significantly lower than that of their core genomes (Wilcoxon rank sum test p-value < $2.2x10^{-6}$).

To further understand whether these differences in GC-content can be explained by HGT, we looked at the taxonomic annotation of the genes in the pan-genome (**Table 5**). For SOX symbiont, the top three annotated species matched between the core and the accessory genome. The closest species to *B. brooksi* SOX symbiont is *Bathymodiolus thermophilus thioautotrophic gill symbiont*, followed by *Bathymodiolus azoricus thioautotrophic gill symbiont,* and *Bathymodiolus septemdierum thioautotrophic gill symbiont.* We found that the closest species to the *B. brooksi* MOX symbiont is *Bathymodiolus platifrons methanotrophic gill symbiont*. It is closely followed by *Methyloprofundus sedimenti*. The core genome shows similarities to *Methylomarinum vadi* (2.254%) and finally, 2.174% of the accessory genome matches with *Methylobacter tundripaludum*. We found that the top annotated species are mostly shared between core and accessory genomes, and that the percentages of each top annotated species are relatively similar between core and accessory genes. This indicates that the reconstructed pan-genome network is not much affected by contamination. The observed reduction in the

fraction of annotated genes among the accessory genome is expected because genes that are rare in the population are usually underrepresented in public databases.

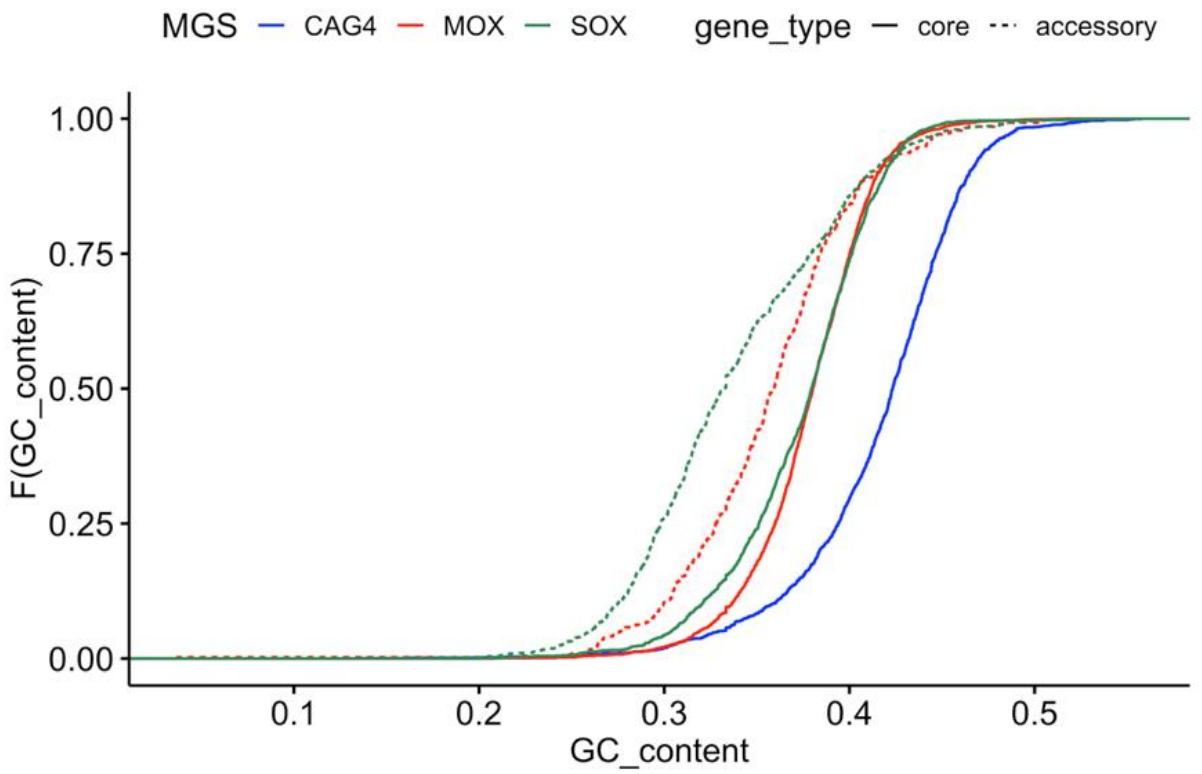


Figure 19: Empirical cumulative distribution of GC content of core and accessory genomes in SOX, MOX and the unidentified MGS3 (CAG4). Median core genome GC_{SOX}: 0.380, SD: 0.0401. Median accessory genome GC_{SOX}: 0.331, SD: 0.0549. Median core genome GC_{MOX}: 0.380, SD: 0.0342. Median accessory genome GC_{MOX}: 0.359, SD: 0.0487. Median genome GC_{CAG4}: 0.423, SD: 0.0465.

Additionally, we studied the mutations occurring in the accessory genomes. We first looked at the number of single nucleotide variants (SNVs) per base pair in the accessory genes and compared it to that of the core genomes. We found that MOX genes have, overall, a lower SNVs/bp ratio than SOX genes. For both species, the SNVs/bp ratio between the core and accessory genomes is significantly different (**Fig. 21**). The pS observed in SOX accessory genes

is lower than that of SOX core genes, what suggests that the core genome is more diverged (**Table 4**). On the other hand, we observed that the median pS for MOX accessory genes is higher than that of MOX core genes. We speculate that differences in pS between MOX core and accessory genomes most likely are due to the overestimation of divergence in paralogous genes (**Table 4**). Indeed, we observed that the median pS estimated for genes that have been identified as potential paralogs is higher than for non-paralogous genes in both bacterial species (**Fig. 22**).

We also estimated the degree of selection acting on the accessory genes and compared it to that of the core genomes. We observed a higher pN/pS value for SOX accessory genome compared to the SOX core genome. This suggests that the accessory genome of SOX is under lower selective constraints. The selective regime for the MOX accessory genome, measured as pN/pS, is highly similar to the one previously reported for the MOX core genome. This may be explained by a population undergoing stronger bottlenecks and/or having short divergence times (**Table 4, Fig. 20**).

We found that the accessory genomes of both symbionts have a lower GC-content than their respective core genomes. In SOX, pN/pS estimates suggest that these differences may be because of a lower selective constraint acting on accessory genomes. The estimated pS is commonly used as a proxy to date populations. We found that SOX core genes show lower divergence compared to SOX accessory genes. This suggests that the accessory genomes have been present in the population less time than the core, probably, because they were acquired at some point by HGT. Surprisingly, we found that the accessory genome of MOX is more diverged than its core genome. It is very unlikely that accessory genes, which are not essential (i.e., genes that are not required for the organism survival), are older than essential, core genes. The recent transfer of genes from MOX to SOX population might give rise to such results, if recent transfer occurred, the variation that we observe might belong additionally to the SOX population and not exclusively to the MOX population. An alternative explanation might be that these genes are under a higher evolutionary rate than the core genomes. Intra-genome evolutionary rate differences haven't been widely studied in bacteria. was shown that methylated GpC dinucleotides have higher evolutionary rates in eukaryotes (Cargill et al. 1999), and this could also explain the differences observed in MOX. Nonetheless, we found that the accessory genes with higher pS correspond to potential paralogous genes. This suggests that there might be an overestimation of the number of mutations that accumulate in the group of paralogous genes, which are prevalent in the MOX pan-genomes. This overestimation is due to the potential mapping of metagenomic reads that belong to several paralogous genes onto the same representative sequence.

Therefore, the artifact comes from the joined estimation of mutations that actually occur at different positions of the genome. On top of that, paralogous genes are thought to accumulate more variants since redundancy on the function will relax the selective constraints in one of the duplicated copies (Kondrashov et al. 2002). It is therefore not possible to distinguish the actual variation due to divergence of paralogs from the overestimation of variants in duplicated genes. To further understand the origin of the accessory genes, we next looked at the function of the genes in the symbionts pan-genomes and identified orthologous pairs between the two co-occurrent species.

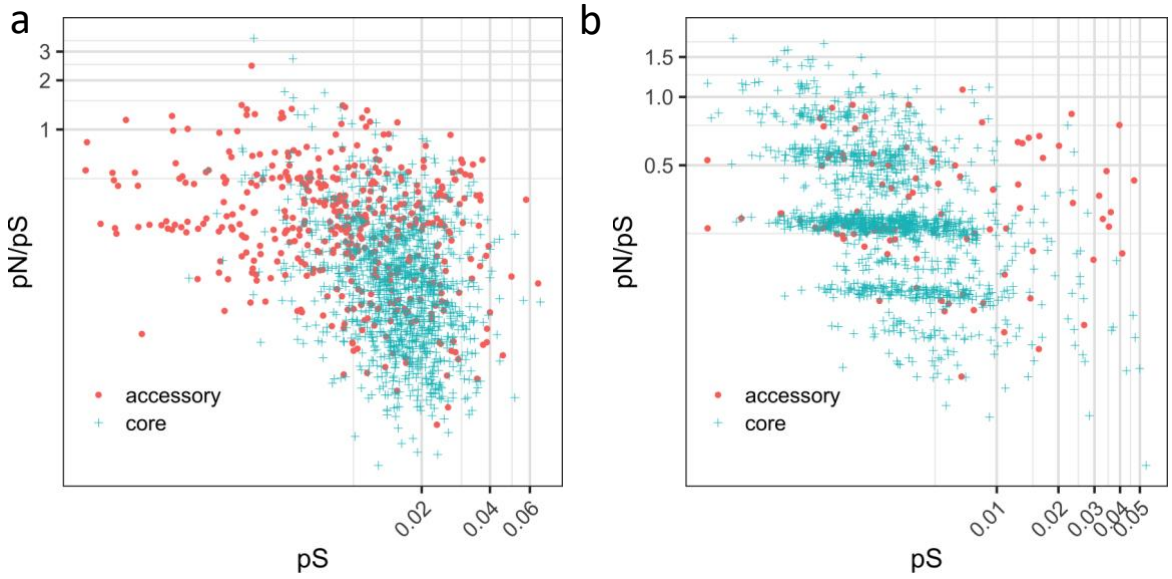


Figure 20. Distributions of pN/pS against pS shown for **a,** SOX and **b,** MOX populations. Dot color and shape indicates whether the gene is core or accessory. The distribution of estimated pN/pS for accessory SOX genes is significantly different than that of SOX core genes (Wilcoxon rank sum test p-value < $2x10^{-16}$), while there is no significant difference among MOX core and accessory genes (Wilcoxon rank sum test p-value = 0.3269).

Table 4: Distribution of variants and pN/pS estimates for SOX and MOX core and accessory genomes.

		#genes	#SNVs	#genes with SNVs	pN/pS	median per-gene pN/pS	median per-gene pS
MOX	Core	2,452	4,665	1,641	0.421	0.279	0.00294
	Accessory	414	735	112	0.437	0.288	0.00380
SOX	Core	1,522	19,349	1,445*	0.135	0.114	0.0161
	Accessory	962	7,580	486	0.332	0.263	0.00976

Table 5: Taxonomic affiliation of each SOX and MOX pan-genome element type to the top-three matched species. NA: Taxonomy not assigned.

Pan-genome element	Species	#genes (percentage)
SOX Core	*Bathymodiolus thermophilus thioautotrophic gill symbiont*	1,172 (77%)
	Bathymodiolus azoricus thioautotrophic gill symbiont	63 (4.139%)
	Bathymodiolus septemdierum thioautotrophic gill symbiont	40 (2.628%)
	NA	141 (9.264%)
SOX Accessory	*Bathymodiolus thermophilus thioautotrophic_gill_symbiont*	306 (31.81%)
	Bathymodiolus septemdierum thioautotrophic gill symbiont	63 (3.534%)
	Bathymodiolus azoricus thioautotrophic gill symbiont	34 (6.549%)
	NA	297 (3.087%)
MOX Core	*Bathymodiolus platifrons methanotrophic gill symbiont*	854 (26.67%)
	Methyloprofundus sedimenti	708 (28.87%)
	Methylomarinum vadi	56 (2.254%)
	NA	227 (9.258%)
MOX Accessory	*Bathymodiolus platifrons methanotrophic gill symbiont*	63 (15.22%)
	Methyloprofundus sedimenti	39 (9.42%)
	Methylobacter tundripaludum	9 (2.174%)
	NA	93 (22.46%)

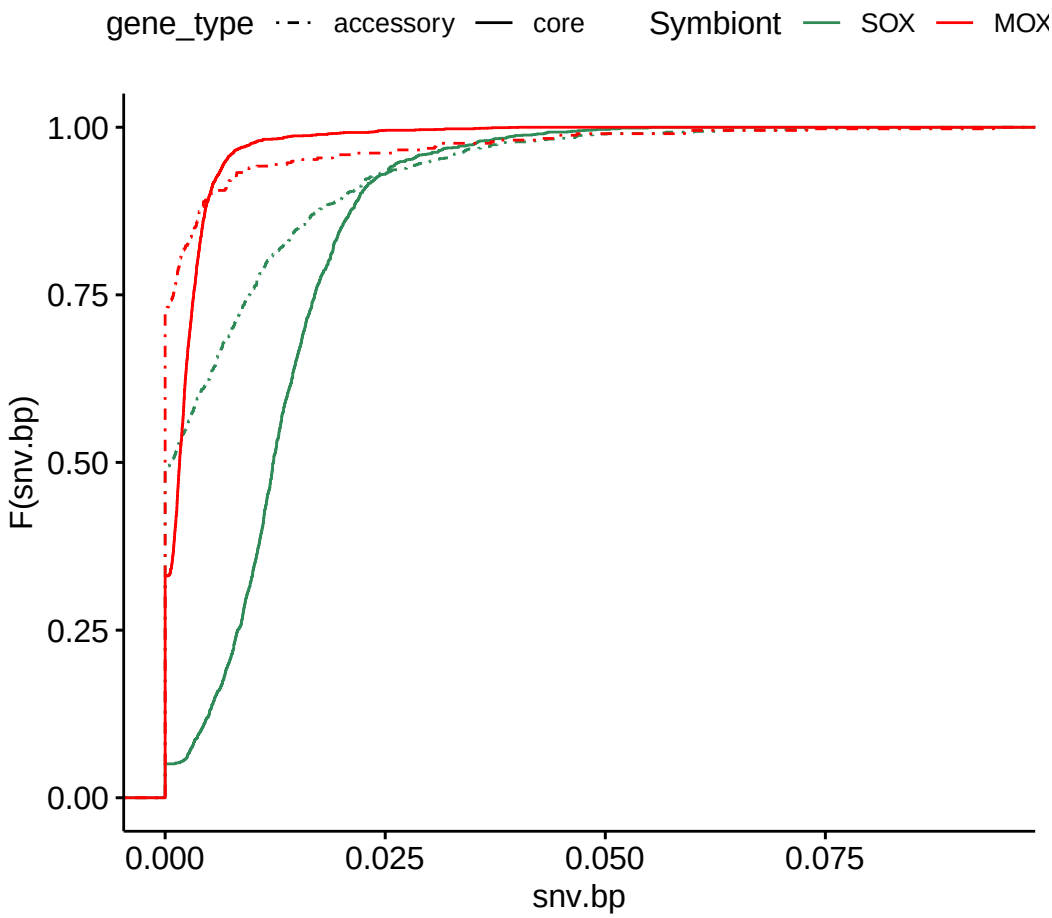


Figure 21: Empirical cumulative distribution of the number of identified SNVs per base pair for MOX and SOX core and accessory genomes. Median SNVs/bp for SOX core genes: 1.23×10^{-2}. Median SNVs/bp for SOX accessory genes: 7.74×10^{-4}. The distributions of SNVs/bp for SOX core and SOX accessory genes are significantly different (Wilcoxon rank-sum test p-value < 2.2×10^{-16}). Median SNVs/ bp for MOX core genes: 1.64×10^{-3}. Median SNVs/ bp for MOX accessory genes: 0. The distributions of SNVs/bp for MOX core and MOX accessory genes are significantly different (Wilcoxon rank-sum test p-value < 2.2×10^{-16}).

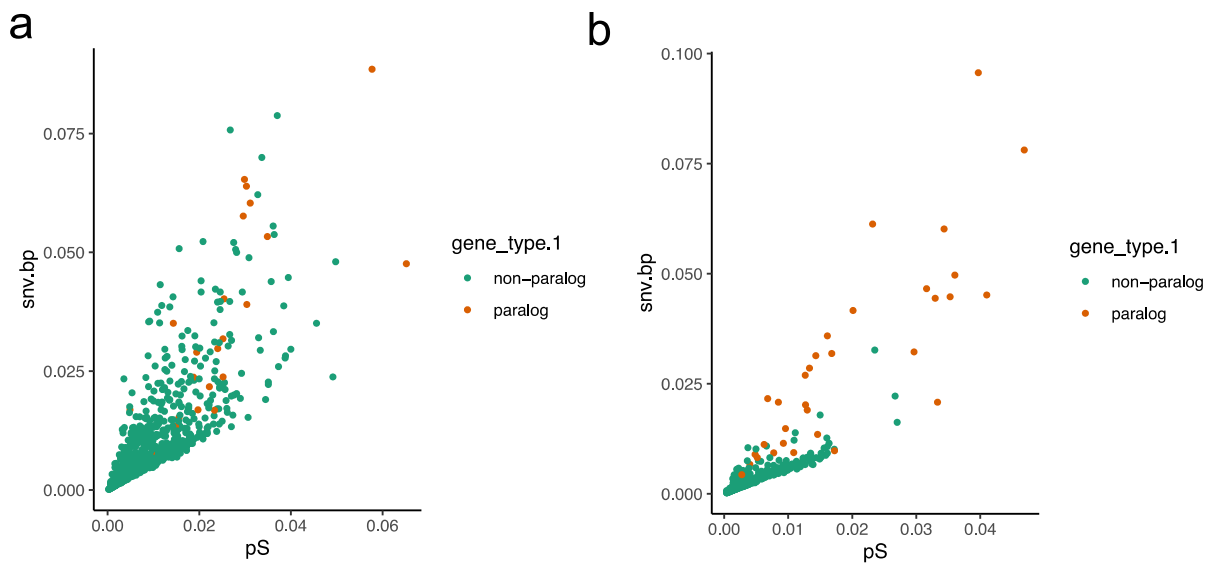


Figure 22: Distribution of pS estimates and SNVs/bp for **a,** SOX and **b,** MOX accessory genomes. Note that only genes with at least 1 SNV are represented. Orange dots indicate genes that have been identified as potential paralogs while green dots are single-copy accessory genes. **a.** Median pS for SOX paralog genes: 2.4×10^{-2}. Median pS for SOX non-paralog genes: 7.1×10^{-3} (Wilcoxon rank-sum test p-value: 1.2×10^{-10}). Median SNVs/bp for SOX paralog genes: 3.0×10^{-2}. Median SNVs/kb for SOX non-paralog genes: 5.7×10^{-3} (Wilcoxon rank-sum test p-value: 1.8×10^{-11}). **b.** Median pS for MOX paralog genes: 1.3×10^{-2}. Median pS for MOX non-paralog genes: 3.6×10^{-3} (Wilcoxon rank-sum test p-value: 1.8×10^{-13}). Median SNVs/bp for MOX paralog genes: 2.1×10^{-2}. Median SNVs/kb for MOX non-paralog genes: 2.7×10^{-3} (Wilcoxon rank-sum test p-value < 2.2×10^{-16}).

6.1.4 Functionality of the accessory genomes.

Keeping a flexible pan-genome has been shown to be important for niche adaptation. Here, we studied how the different functional categories distribute across the core and accessory genomes to identify common functions shared between the two symbionts that may be related to the adaptation to an endosymbiotic lifestyle.

As expected, functions associated to central metabolism -i.e., translation, ribosomal structure and biogenesis, post-translational modification, protein turnover and chaperones, nucleotide, lipid, inorganic ion, coenzyme carbohydrate and amino acid transport and metabolism and energy production and conversion- are mostly present in the core genomes of both symbionts (**Fig. 23**). Interestingly, the accessory genomes of both bacterial species are over-represented in functions related to DNA integrity: namely MGEs, defense mechanisms and DNA replication, recombination and repair functions. In MOX, 26.23% of the genes annotated as mobilome-related are accessory, while this percentage equals 73.08% in SOX. As for genes categorized as defense-related, 41.51% of MOX defense genes are accessory, while 68.06% are accessory for SOX. Moreover, 21.54% of genes annotated as related to replication, recombination and repair are accessory in MOX, and 45.86% are accessory in SOX (**Fig. 23**).

We observed that SOX contains a larger repertoire of genes related to defense -with 72 annotated genes- in comparison to MOX, that contains 54 genes. Additionally, a large fraction of them are related to restriction-modification systems, with 38 and 22 genes, respectively, annotated as such. A total of 5 and 4 additional genes annotated as restriction-modification systems were found, respectively for SOX and MOX pan-genomes, outside the COG category of defense. Interestingly, SOX genome contains 13 genes annotated as CRISPR-related, and only 1 gene is present in the core genome. Notably, the mobilome of MOX is larger than that of SOX. It contains 61 genes annotated as mobilome, versus the 26 that have been annotated in SOX. The number of genes that has been annotated as transposases is of 39 versus the 18 genes from SOX. Of these, 38.46% (15) are identified as paralogs in MOX, versus 16.67% (3) identified as paralogs in SOX. The larger percentage of transposases found among the paralogous set of genes in MOX might indicate that its mobilome is more active than that of SOX. Also, MOX contains a larger number of integrases (14 MOX genes versus 5 SOX genes). It is also interesting that MOX was found to contain many more functions related to cell motility and signal transduction/intracellular trafficking. MOX contains 50 genes for motility, 144 for signal

transduction mechanisms and 66 for intracellular trafficking, while SOX has 3 genes for motility, 31 for signal transduction and 39 for intracellular trafficking.

Altogether, these results suggest, first of all, that genes related to defense mechanisms, DNA recombination and repair, and mobilome make up most part of the flexible region of the genomes in the analyzed populations of symbionts. Differences on the pan-genomes of the two symbionts could be observed. The larger repertoire of mobilome-related genes, and the presence of more genes for cell motility may reflect differences in the ecological traits of the two species. In one hand, the amount of mobile elements in bacterial populations has been shown to correlate to their ecological trait, where facultative bacteria have the larger amount of mobile elements, followed by the free-living and finally the obligate symbionts (Newton and Bordenstein 2011). Additionally, the presence of genes related to cell motility is also an indicator that the species has a free-living phase, as previously suggested for the facultartive bioluminescent symbiont from the flashlight fish (Hendry et al. 2014). MOX might still be adapted to a more facultative lifestyle, likely having a free-living active state, while SOX might have a more obligate character.

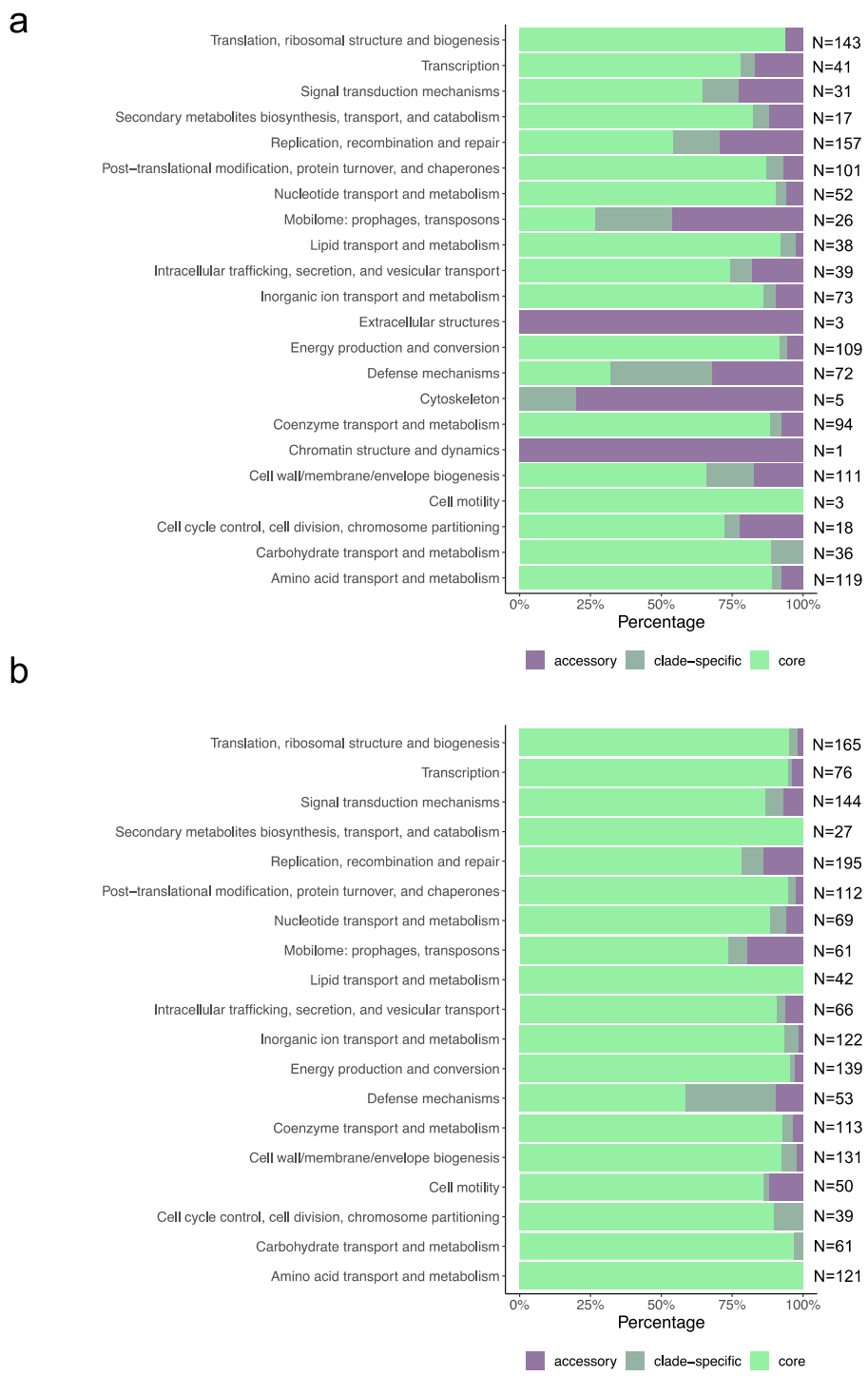


Figure 23: Assigned COG categories for core, accessory and clade-specific genes in **a,** SOX and **b,** MOX bacteria. Not assigned categories are not included in the analysis. N shows the total number of genes annotated for each category.

6.1.5 Orthologous genes among SOX and MOX pan-genomes.

Bacterial populations sharing a specific habitat are thought to share horizontally acquired genes from the same habitat-specific gene pool as well (Polz et al. 2013). This hypothesis has additionally been extended to endosymbiotic bacteria; the “intracellular arena” posits that the host cell serves as an arena for bacteria, where symbionts can acquire genes from a niche-specific gene-pool that allow them to succeed in the intracellular environment. The presence of similar functions among the accessory genes of both symbionts suggests that the symbionts may share genes. To test this hypothesis, we identified the orthologous pairs by comparing the translated sequences of the genes of MOX and SOX bacteria.

We found 761 orthologous sequences that share at least 30% of full-length protein sequence identity. Most of the pairs (93.29%) involve only core genes from both bacterial genomes. The most represented functional category among these orthologous pairs corresponds to very central functions such as translation, ribosomal and biogenesis category (125 genes). It is followed by the category associated with amino acid transport and metabolism (77 genes) and coenzyme transport and metabolism (61 genes). This result suggests that most of the genes functionally homologous between SOX and MOX very likely are related by ancestry (**Fig. 24**). On the other hand, only 51 pairs involve at least one accessory gene, and from those, just 10 are accessory in both genomes, where four pairs are mobilome-related. One of these four genes was identified as a transposase, another as an integrase and a third one was annotated as a transposase for the ortholog from SOX, and as integrase for the ortholog belonging to MOX. We find four orthologous gene pairs involving one accessory gene that are annotated as restriction-modification system-related. Our findings suggest that few events of transfer between MOX and SOX had probably occurred in the past. Nonetheless, HGT is not a pervasive phenomenon within the deep-sea mussels symbionts community.

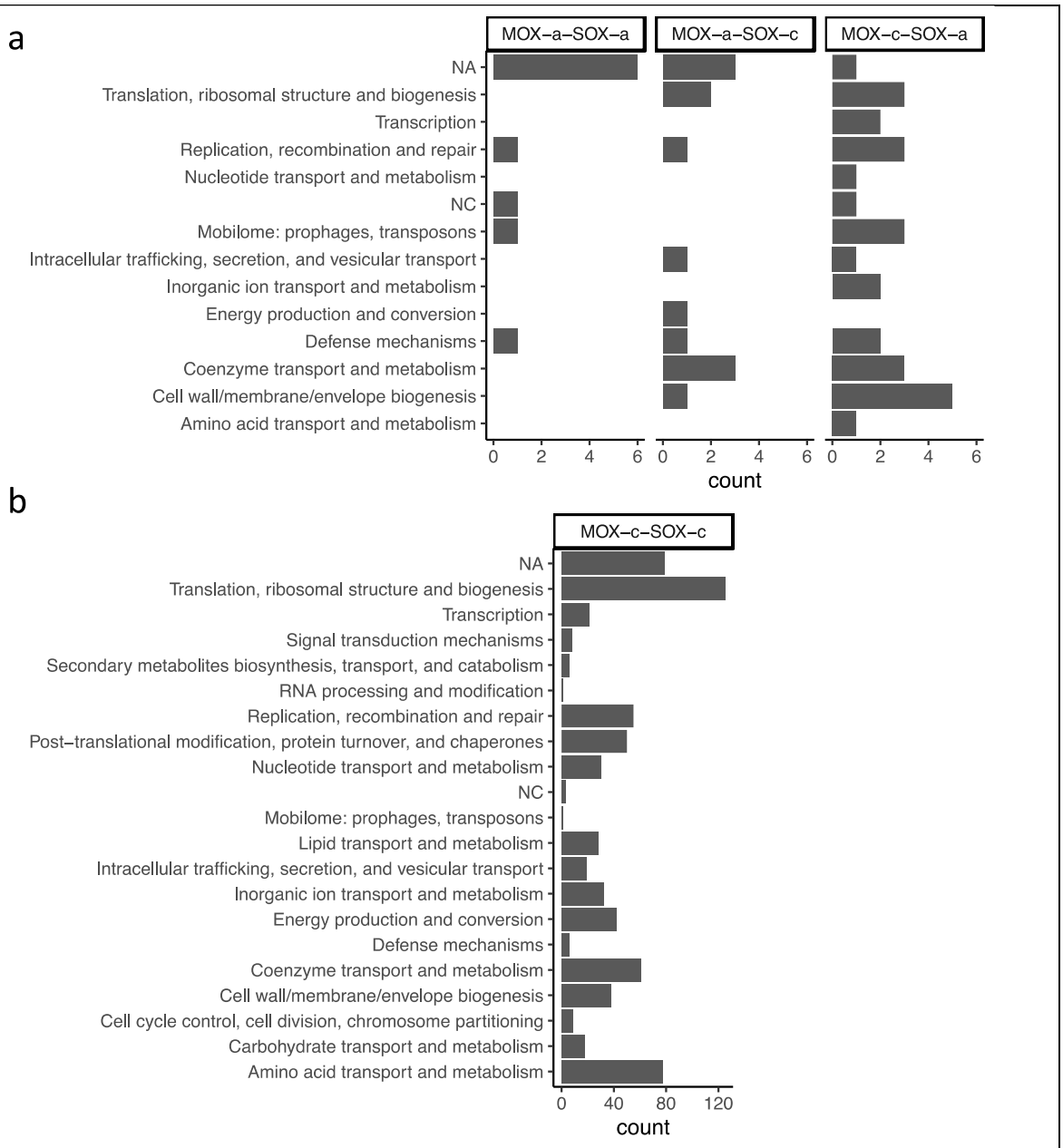


Figure 24: COG categories associated to orthologous gene pairs from SOX and MOX pan-genomes. **a,** Distribution of genes into COG categories for orthologous group pairs where at least one member of the pair is an accessory gene (MOX-a-SOX-a: pair of MOX accessory and SOX accessory genes. MOX-a-SOX-c: pair of MOX accessory and SOX core genes. MOX-c-SOX-a: pair of MOX core and SOX accessory genes). **b,** Distribution of genes into COG categories for MOX core and SOX core pairs. NA: Not Assigned, NC: Not Conclusive. Not conclusive pairs are those where each member of the pair was assigned to a different category.

6.2 Discussion.

Our previous results showed that the bacterial populations harbored across individual mussels are geographically isolated. The restricted uptake of bacteria explains the observed genetic isolation among symbiont subpopulations. Consequently, the bottlenecks associated with this isolation during the colonization process very likely result in the observed diversification of strains. This genetic isolation affects the fraction of the bacterial genome that is tied to a vertical inheritance -i.e., mostly the core genome. Nevertheless, MGEs can theoretically disperse among mussels and therefore act as genetic vectors that horizontally introduce accessory genes into symbiont populations.

We observed that the changes introduced by single point mutations exceed those introduced by gene gain and loss. Moreover, the strong correlation between F_{ST} and ß-diversity indicates that differences in gene content are mostly present among clades. In fact, we found that strains belonging to the same clade have very similar gene content, while they can differ by a larger number of SNVs. HGT is generally an important mechanism contributing to gene gain and loss rate. We observed that the GC-content of accessory genomes is lower than that of the core genomes. Whether this difference is driven by HGT or by a differential selective regime governing the two groups of genes is unclear. Our orthologous analysis enabled us to identify few potential HGT events between the two symbionts, thus HGT does not appear to be a main contributor to gene gain and loss in this population. Gene gain and loss importantly contributes to the diversity of bacterial populations, exceeding, in many cases, single-point mutation rates in the tip of phylogenies. Examples are populations of *Acinetobacter baumannii* or bacteria from the genus *Bacillus* (Hao and Golding 2006; Graña-Miraglia et al. 2017). We propose that the low gene gain and loss found in symbiont populations of deep-sea mussels is most likely explained by the intracellular condition of the symbionts, what translates into a strong geographic isolation. Such conditions reduce the exposure of the population to novel DNA elements that can be newly acquired via horizontal gene transfer, therefore decreasing the gene gain and loss rate. The uptake of genetic material might potentially still be possible when the symbionts are in their free-living phase. This uptake might be restricted if bacteria disperse in a dormant state, as already suggested for the SOX symbiont (Ponnudurai et al. 2017). One can speculate that differences in gene content within the *B.brooksi* symbiont populations are therefore importantly determined by gene loss events, which are very frequent among symbiotic species undergoing reductive evolution (Moran et al. 2008). The lower purifying selection acting on accessory genomes for SOX population supports this hypothesis; because accessory genes are under lower selective

constraints than core genes, the former are more prone to be lost. On the other hand, similar selective regimes have been found to govern the core and accessory genomes for MOX population. This might reflect the young character of this population, where selection did not yet have enough time to purge deleterious mutations in either core or accessory genomes.

One commonality among the pan-genomes of both symbionts is the functional annotation of the accessory genes. Most accessory genes in both species are involved in the maintenance of genome integrity (mobilome, DNA repair and defence mechanisms), with an over-representation of genes annotated as restriction-modification systems. Genes that encode for defence mechanisms often co-localize with MGEs (e.g., lysogenic phages and ICEs) in “defence islands”. Two different hypotheses have been postulated to explain the existence of defence islands as hotspots for DNA acquisition: the “garbage pile effect” refers to the widespread gain and loss of non-essential defence genes with limited fitness effect in these islands. Alternatively, the presence of defence islands may be adaptive, since the co-localization of functionally interacting defence genes may be beneficial (Koonin et al. 2017). Another explanation to the existence of defence islands is a rapid evolution of defence gene repertoire because of a continuous arms-race with the phage community, which leads to a high diversity of defence islands in the population. Along the same lines, the “pan-immune system” of bacteria suggests that selection on the maintenance of defence mechanisms acts at the group level, where the presence of strains encoding different defence systems is advantageous for the total community (Bernheim and Sorek 2020) (e.g., inhabiting a single mussel individual).

Considering that SOX and MOX are intracellular, it seems unlikely that the deep-sea mussel symbiotic communities are exposed to a high concentration of phages. Hence it is possible, that the defence islands we observe are better explained by other treats related to foreign DNA invasion. For example, RM systems have been described as part of self-mobilizable genetic elements (Furuta et al. 2010). They have been suggested to have a defensive role when bacteria colonize a new habitat but become genetic elements capable of spreading in the genomes when the population has adapted to it (Rocha 2001). The spread of RM systems occurs by “post-segregational cell killing” (also termed ‘addition mechanism); strains that have lost their RM system, and therefore can’t methylate their genomes, are eliminated from the population since they become sensitive to the restriction enzyme (Kobayashi 2001).The bacterial transition into a symbiotic life style is linked to insertion sequence (IS) proliferation, before the genomes undergo streamlining. Some suggest that the decrease in population size associated to host-restriction makes purifying selection less effective, and ISes get more frequently fixed in the population. The

more ISes, the more frequent transposition occurs. Transposition is self-limiting because the higher its frequency the more likely is to introduce deleterious mutations (Moran and Plague 2004). Since RM systems are tightly associated to MGEs, it is therefore possible, that the former spread into the population genomes as a consequence of this MGEs proliferation. For example, in *Helicobacter pylori,* genome rearrangements have been shown to contribute to the birth-and-death of genes, with especial influence on RM related genes (Furuta et al. 2011). Moreover, the presence of RM systems has been previously described in intracellular symbionts from termite gut flagellates, where they were shown to play an important role in genome rearrangements (Zheng et al. 2016). Our results show that SOX and MOX strains differ more in their polymorphisms than their genetic content and suggest that mobilization of genetic elements in the current symbiont populations is not prevalent. It is thus tenable to hypothesize that the distribution of gene function in the current pan-genomes reflects past events in the evolution of SOX and MOX populations, either during a free-living stage (when the symbionts where exposed to a high concentration of phages) or the transition to host restriction (when MGEs proliferation occurs). It is likely that by that time, RM systems, together with other MGEs, such as transposons, played an important role in shaping the accessory content across strains.

Differences in the pan-genomes of the two species may reflect the evolutionary histories of two symbionts at a different timepoint in their way to become obligate. The origin of the symbiosis between Bathymodiolus mussels and the sulfur-oxidizing symbiont was previously described to be older than that with the methane-oxidizing symbiont (Lorion et al. 2013). As we previously exposed in chapter I, this explains why MOX population was found to have a lower diversity than SOX population, and also, why it is not so strongly affected by purifying selection. Also, the observed differences between the pan-genomes of both species are in accordance with this hypothesis; MOX pan-genome is larger than the SOX pan-genome and contains a lower fraction of accessory genes. This indicates that the genomes of SOX population had more time to undergo erosion, thus decreasing in size. Differences in the age of the association between the host and its symbionts might also be related to the ecological trait of the later, where MOX and SOX species lie differently in the facultative to obligate spectrum. We found that MOX pan-genome contains a larger number of transposons than the SOX pan-genome. Facultative bacteria are usually found to have a larger number of mobile genetic elements compared to obligate and free-living bacteria (Newton and Bordenstein 2011). We found that the fraction of the total genes involved in replication, recombination, and repair that are accessory is larger in SOX in comparison to MOX. This is a common feature of obligate symbionts (Klasson 2004). The presence of cell motility

functions in MOX suggests that this symbiont may facultatively live as a free-living organism. Altogether, these results suggest that SOX might have a more obligate character than MOX.

We have found that the geographic isolation affecting symbiont populations of the deep-sea mussel *Bathymodiolus brooksi* restricts their access to novel DNA via HGT. This is reflected in pan-genomes with low flexibility, where gene content variation is mainly present on the internal branches of the strains phylogeny. Signatures of the different age of the symbiosis between the two symbionts and the mussel are imprinted in the population pan-genomes, where size and differential functionality indicate that MOX could keep functions related to a free-living phase, while SOX is fundamentally an obligate symbiont.

7 References

Achtman M, Wagner M. 2008. Microbial diversity and the genetic nature of microbial species. Nature Reviews Microbiology 6:431–440.

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

Andreani NA, Hesse E, Vos M. 2017. Prokaryote genome fluidity is dependent on effective population size. The ISME Journal:1–3.

Ansorge R, Romano S, Sayavedra L, Porras MÁG, Kupczok A, Tegetmeyer HE, Dubilier N, Petersen J. 2019. Functional diversity enables multiple symbiont strains to coexist in deep-sea mussels. Nature Microbiology 4:2487–2497.

Benson AK, Kelly SA, Legge R, Ma F, Low SJ, Kim J, Zhang M, Oh PL, Nehrenberg D, Hua K, et al. 2010. Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. Proceedings of the National Academy of Sciences 107:18933–18938.

Bernheim A, Sorek R. 2020. The pan-immune system of bacteria: antiviral defence as a community resource. Nature Reviews Microbiology 18:113–119.

Bohlin J, Eldholm V, Pettersson JHO, Brynildsrud O, Snipen L. 2017. The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. BMC Genomics 18:151.

Bolotin E, Hershberg R. 2015. Gene Loss Dominates As a Source of Genetic Variation within Clonal Pathogenic Bacterial Species. Genome Biol Evol 7:2173–2187.

Boscaro V, Kolisko M, Felletti M, Vannini C, Lynn DH, Keeling PJ. 2017. Parallel genome reduction in symbionts descended from closely related free-living bacteria. Nature Ecology & Evolution 1:1160.

Breusing C, Vrijenhoek RC, Reusch TBH. 2017. Widespread introgression in deep-sea hydrothermal vent mussels. BMC Evolutionary Biology 17:13.

Bright M, Bulgheresi S. 2010. A complex journey: transmission of microbial symbionts. Nature Reviews Microbiology 8:218–30.

Broad Institute. Best practices for variant calling with the GATK, https://www.broadinstitute.org/partnerships/education/broade/best-practices-variant-calling-gatk-1. Available from: https://www.broadinstitute.org/partnerships/education/broade/best-practices-variant-calling-gatk-1

Brockhurst MA, Harrison E, Hall JPJ, Richards T, McNally A, MacLean C. 2019. The Ecology and Evolution of Pangenomes. Current Biology 29:R1094–R1103.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nature methods 12:59–60.

Bushnell B. 2014. BBMap. Available from: sourceforge.net/projects/bbmap/

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, et al. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nature Genetics 22:231–238.

Chen J. 2018. GUniFrac: Generalized UniFrac Distances. Available from: https://CRAN.R-project.org/package=GUniFrac

Costello EK, Stagaman K, Dethlefsen L, Bohannan BJM, Relman DA. 2012. The Application of Ecological Theory Toward an Understanding of the Human Microbiome. Science 336:1255–1262.

David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, et al. 2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature 505:559–563.

Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. 2016. Within-host evolution of bacterial pathogens. Nature Reviews Microbiology 14:150–162.

Dietel A-K, Merker H, Kaltenpoth M, Kost C. 2018. Selective advantages favour high genomic AT-contents in intracellular elements. Available from: http://biorxiv.org/lookup/doi/10.1101/448324

Dubilier N, Windoffer R, Giere O. 1998. Ultrastructure and stable carbon isotope composition of the hydrothermal vent mussels Bathymodiolus brevior and B. sp. affinis brevior from the North Fiji Basin, western Pacific. Marine Ecology Progress Series 165:187–193.

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

Ellegaard KM, Engel P. 2019. Genomic diversity landscape of the honey bee gut microbiota. Nature Communications 10:446.

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

Furuta Y, Abe K, Kobayashi I. 2010. Genome comparison and context analysis reveals putative mobile forms of restriction–modification systems and related rearrangements. Nucleic Acids Res 38:2428–2443.

Furuta Y, Kawai M, Yahara K, Takahashi N, Handa N, Tsuru T, Oshima K, Yoshida M, Azuma T, Hattori M, et al. 2011. Birth and death of genes linked to chromosomal inversion. PNAS 108:1501–1506.

Gonzalez RJ, Lane MC, Wagner NJ, Weening EH, Miller VL. 2015. Dissemination of a Highly Virulent Pathogen: Tracking The Early Events That Define Infection. PLOS Pathogens 11:e1004587.

Graña-Miraglia L, Lozano LF, Velázquez C, Volkow-Fernández P, Pérez-Oseguera Á, Cevallos MA, Castillo-Ramírez S. 2017. Rapid Gene Turnover as a Significant Source of Genetic Variation in a Recently Seeded Population of a Healthcare-Associated Pathogen. Front. Microbiol. 8:1817.

Guyomar C, Legeai F, Jousselin E, Mougel C, Lemaitre C, Simon J-C. 2018. Multi-scale characterization of symbiont diversity in the pea aphid complex through metagenomic approaches. Microbiome 6:181.

Hagen MJ, Hamrick JL. 1996. Population level processes in Rhizobium leguminosarum bv. trifolii: the role of founder effects. Molecular Ecology 5:707–714.

Hansen AK, Moran NA. 2011. Aphid genome expression reveals host-symbiont cooperation in the production of amino acids. Proceedings of the National Academy of Sciences 108:2849–2854.

Hao W, Golding GB. 2006. The fate of laterally transferred genes: Life in the fast lane to adaptation or death. Genome Res 16:636–643.

Helmus MR, Bland TJ, Williams CK, Ives AR. 2007. Phylogenetic Measures of Biodiversity. The American Naturalist 169:E68–E83.

Ho P-T, Park E, Hong SG, Kim E-H, Kim K, Jang S-J, Vrijenhoek RC, Won Y-J. 2017. Geographical structure of endosymbiotic bacteria hosted by Bathymodiolus mussels at eastern Pacific hydrothermal vents. BMC Evolutionary Biology 17:121.

Huddleston JR. 2014. Horizontal gene transfer in the human gastrointestinal tract: potential spread of antibiotic resistance genes. Infect Drug Resist 7:167–176.

Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, et al. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research 47:D309–D314.

Huson DH. 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics 14:68–73.

Hyatt D, Locascio PF, Hauser LJ, Uberbacher EC. 2012. Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics 28:2223–2230.

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. The ISME Journal 10:990–1001.

Jones RJ, Hoegh-Guldberg O, Larkum AWD, Schreiber U. 1998. Temperature-induced bleaching of corals begins with impairment of the CO2 fixation mechanism in zooxanthellae. Plant, Cell and Environment 21:1219–1230.

Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, et al. 2014. Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. Science (New York, N.Y.) 344:416–20.

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26:1463–1464.

Kent WJ. 2002. BLAT — The BLAST -Like Alignment Tool. Genome Research 12:656–664.

Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J, et al. 2007. Patterns and Implications of Gene Gain and Loss in the Evolution of Prochlorococcus. PLOS Genetics 3:e231.

Klasson L. 2004. Evolution of minimal-gene-sets in host-dependent bacteria. Trends in Microbiology 12:37–43.

Klose J, Polz MF, Wagner M, Schimak MP, Gollner S, Bright M. 2015. Endosymbionts escape dead hydrothermal vent tubeworms to enrich the free-living population. Proceedings of the National Academy of Sciences of the United States of America 112:11300–11305.

Kobayashi I. 2001. Behavior of restriction–modification systems as selfish mobile elements and their impact on genome evolution. Nucleic Acids Res 29:3742–3756.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. Genome Biology 3:research0008.1.

Koonin EV, Makarova KS, Wolf YI. 2017. Evolutionary Genomics of Defense Systems in Archaea and Bacteria. Annu. Rev. Microbiol. 71:233–261.

Koren O, Goodrich JK, Cullender TC, Spor A, Laitinen K, Kling Bäckhed H, Gonzalez A, Werner JJ, Angenent LT, Knight R, et al. 2012. Host Remodeling of the Gut Microbiome and Metabolic Changes during Pregnancy. Cell 150:470–480.

Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. PLoS Genetics 4:e1000304.

Lawrence JG, Ochman H. 1997. Amelioration of Bacterial Genomes: Rates of Change and Exchange. J Mol Evol 44:383–397.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Lorion J, Kiel S, Faure B, Kawato M, Ho SYW, Marshall B, Tsuchida S, Miyazaki J-I, Fujiwara Y. 2013. Adaptive radiation of chemosymbiotic deep-sea mussels. Proc. R. Soc. B 280:20131243.

McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol 10:13–26.

McFall-Ngai M, Hadfield MG, Bosch TCG, Carey HV, Domazet-Lošo T, Douglas AE, Dubilier N, Eberl G, Fukami T, Gilbert SF, et al. 2013. Animals in a bacterial world, a new imperative for the life sciences. Proceedings of the National Academy of Sciences 110:3229–3236.

McFall-Ngai MJ. 2014. The Importance of Microbes in Animal Development: Lessons from the Squid-Vibrio Symbiosis. Annu. Rev. Microbiol. 68:177–194.

McInerney JO, McNally A, O'Connell MJ. 2017. Why prokaryotes have pangenomes. Nat Microbiol 2:1–5.

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. 2005. The microbial pan-genome. Current Opinion in Genetics & Development 15:589–594.

Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. PNAS 93:2873–2878.

Moran NA. 2002. Microbial Minimalism: Genome Reduction in Bacterial Pathogens. Cell 108:583–586.

Moran NA, McCutcheon JP, Nakabachi A. 2008. Genomics and Evolution of Heritable Bacterial Symbionts. Annual Review of Genetics 42:165–190.

Moran NA, Plague GR. 2004. Genomic changes following host restriction in bacteria. Current Opinion in Genetics & Development 14:627–633.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48:443–453.

Newton ILG, Bordenstein SR. 2011. Correlations Between Bacterial Ecology and Mobile DNA. Curr Microbiol 62:198–208.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution 32:268–274.

Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, et al. 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nature Biotechnology 32:822–828.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Research 27:824–834.

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

Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature Microbiology 2:1533–1542.

Pinto-Carbó M, Sieber S, Dessein S, Wicker T, Verstraete B, Gademann K, Eberl L, Carlier A. 2016. Evidence of horizontal gene transfer between obligate leaf nodule symbionts. The ISME Journal 10:2092–2105.

Polz MF, Alm EJ, Hanage WP. 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends in Genetics 29:170–175.

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

Popa O, Dagan T. 2011. Trends and barriers to lateral gene transfer in prokaryotes. Current Opinion in Microbiology 14:615–623.

Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T. 2011. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Research 21:599–609.

Quince C, Delmont TO, Raguideau S, Alneberg J, Darling AE, Collins G, Eren AM. 2017. DESMAN: a new tool for de novo extraction of strains from metagenomes. Genome Biology 18:181.

Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. 2017. review Shotgun metagenomics , from sampling to analysis. 35.

Rand DM, Kann LM. 1996. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. Molecular Biology and Evolution 13:735–748.

Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ. 2009. Biogeography of the Sulfolobus islandicus pan-genome. PNAS 106:8605–8610.

Reuter M, Pedersen JS, Keller L. 2005. Loss of Wolbachia infection during colonisation in the invasive Argentine ant Linepithema humile. Heredity 94:364–369.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16:276–277.

Riou V, Halary S, Duperron S, Bouillon S, Elskens M, Bettencourt R, Santos RS, Dehairs F, 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.

Rocha EPC. 2001. Evolutionary Role of Restriction/Modification Systems as Revealed by Comparative Genome Analysis. Genome Research 11:946–958.

Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, Feil EJ. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. Journal of Theoretical Biology 239:226–235.

Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, Zeevi D, Costea PI, Godneva A, Kalka IN, Bar N, et al. 2018. Environment dominates over host genetics in shaping human gut microbiota. Nature 555:210–215.

Russell SL, Corbett-Detig RB, Cavanaugh CM. 2017. Mixed transmission modes and dynamic genome evolution in an obligate animal–bacterial symbiosis. The ISME Journal 11:1359–1371.

Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, Waller A, Mende DR, Kultima JR, Martin J, et al. 2013. Genomic variation landscape of the human gut microbiome. Nature 493:45–50.

Schöne BR, 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 Research Part I: Oceanographic Research Papers 52:1896–1910.

Schretter CE, Vielmetter J, Bartos I, Marka Z, Marka S, Argade S, Mazmanian SK. 2018. A gut microbial factor modulates locomotor behaviour in Drosophila. Nature 563:402–406.

Shabat SKB, Sasson G, Doron-Faigenboim A, Durman T, Yaacoby S, Berg Miller ME, White BA, Shterzer N, Mizrahi I. 2016. Specific microbiome-dependent mechanisms underlie the energy harvest efficiency of ruminants. The Isme Journal 10:2958.

Shapiro BJ, Polz MF. 2014. Ordering microbial diversity into ecologically and genetically cohesive units. Trends in Microbiology 22:235–247.

Sibbesen JA, Maretty L, Krogh A. 2018. Accurate genotyping across variant classes and lengths using variant graphs. Nature Genetics 50:1054–1059.

Sommer F, Ståhlman M, Ilkayeva O, Arnemo JM, Kindberg J, Josefsson J, Newgard CB, Fröbert O, Bäckhed F. 2016. The Gut Microbiota Modulates Energy Metabolism in the Hibernating Brown Bear Ursus arctos. Cell Reports 14:1655–1661.

Sprockett D, Fukami T, Relman DA. 2018. Role of priority effects in the early-life assembly of the gut microbiota. Nature Reviews Gastroenterology & Hepatology 15:197–205.

Stoletzki N, Eyre-Walker A. 2011. The Positive Correlation between dN/dS and dS in Mammals Is Due to Runs of Adjacent Substitutions. Molecular Biology and Evolution 28:1371–1380.

Subramanian S. 2016. The effects of sample size on population genomic analyses – implications for the tests of neutrality. BMC Genomics 17:123.

Sun J, Zhang Yu, Xu T, Zhang Yang, Mu H, Zhang Yanjie, Lan Y, Fields CJ, Hui JHL, Zhang W, et al. 2017. Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. Nature Ecology & Evolution 1:0121.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, et al. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome." PNAS 102:13950–13955.

Tria FDK, Landan G, Dagan T. 2017. Phylogenetic rooting using minimal ancestor deviation. Nature Ecology & Evolution 1:0193.

Van Dover C. 2002. Community structure of mussel beds at deep-sea hydrothermal vents. Marine Ecology Progress Series 230:137–158.

Vega NM, Gore J. 2017. Stochastic assembly produces heterogeneous communities in the Caenorhabditis elegans intestine. PLOS Biology 15:e2000633.

Wang D, Kreutzer DA, Essigmann JM. 1998. Mutagenicity and repair of oxidative DNA damage: insights from studies using defined lesions. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 400:99–115.

Wentrup C, Wendeberg A, Huang JY, Borowski C, Dubilier N. 2013. Shift from widespread symbiont infection of host tissues to specific colonization of gills in juvenile deep-sea mussels. The ISME Journal 7:1244–1247.

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

Wernegreen JJ. 2015. Endosymbiont evolution: predictions from theory and surprises from genomes: Endosymbiont genome evolution. Annals of the New York Academy of Sciences 1360:16–35.

Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. 2012. LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Research 40:11189–11201.

Wollenberg MS, Ruby EG. 2009. Population Structure of Vibrio fischeri within the Light Organs of Euprymna scolopes Squid from Two Oahu (Hawaii) Populations. Applied and Environmental Microbiology 75:193–202.

Won Y-J, Hallam SJ, O'Mullan GD, Pan IL, Buck KR, Vrijenhoek RC. 2003. Environmental Acquisition of Thiotrophic Endosymbionts by Deep-Sea Mussels of the Genus Bathymodiolus. Applied and Environmental Microbiology 69:6785–6792.

Woyke T, Tighe D, Mavromatis K, Clum A, Copeland A, Schackwitz W, Lapidus A, Wu D, McCutcheon JP, McDonald BR, et al. 2010. One Bacterial Cell, One Complete Genome.Ahmed N, Ahmed N, editors. PLoS ONE 5:e10314.

Zhang T, Abel S, Wiesch zur PA, Sasabe J, Davis BM, Higgins DE, Waldor MK. 2017. Deciphering the landscape of host barriers to Listeria monocytogenes infection. Proc Natl Acad Sci U S A 114:6334–6339.

Zheng H, Dietrich C, Hongoh Y, Brune A. 2016. Restriction-Modification Systems as Mobile Genetic Elements in the Evolution of an Intracellular Symbiont. Mol Biol Evol 33:721–725.

Zhou J, Bruns MA, Tiedje JM. 1996. DNA Recovery from Soils of Diverse Composition. Applied and Environmental Microbiology 62:316–322.

8 Acknowledgements

Thank you for inviting me to explore the world no matter how frightening it was for you. You have enabled me to love what I do, and to do what I love. (Gracias por invitarme a explorar el mundo sin importar cuán aterrador fuese para vosotros. Me habéis permitido amar lo que hago y hacer lo que amo).

To my friends at home (Clara, Cristina, Laura, Marta, Menchu, Paula) and to Oliver; thank you for all the courage and support you give me.

I would like to express my special gratitude to Dr. Anne Kupczok for her guidance and support in science and in life. Many thanks for the long discussions that were very enlightening to me. Also, especial thanks go to Prof. Dr. Tal Dagan for her guidance and for giving me the opportunity to share niche with many great scientists and human beings. To all my colleagues and friends in the Genomics Microbiology Group; thanks for all your support in science but also for bringing some warmth to the grey Kiel. Especial thanks go to Ahmad Samer Kadib Alban (did I write it correctly this time?), for his honesty towards this thesis and so much emotional support. Also, many thanks to Dr. Robin Koch, for the very extended scientific discussions on this thesis and outside the scope of it. I would like to thank Almut Werner for her great job analyzing the functional categories in the metagenomes and Ryszard Soluch for critical comments on this thesis.

I reserve especial thanks to both members of my thesis advisor committee: Prof. Dr. Thorsten B. Reusch and Prof. Dr. John Baines for their many useful comments on this thesis.

I additionally would like to acknowledge the IMPRS graduate school and finally, the CRC1182 (Origin and Function of Metaorganisms) for funding this project.

9 Supplementary Information

9.1 Impact of sequencing depth on diversity analysis

We tested if the higher SNV density in SOX compared to MOX could be explained by the higher sequencing depth for SOX. To this end, we repeated part of the analyses by analyzing a subset of the orginal SOX reads that were subsampled to achieve MOX coverage levels. Normalizing SOX to the median MOX coverage of 36x resulted in a SNV density of 12.1 SNVs/kpb. This density is similar to the original estimate of 14 SNVs/kbp and still highly elevated compared to MOX (2.4 SNVs/kbp), which indicates that the difference in diversity is not driven by bias due to sequencing depth. Nucleotide diversity of SOX normalized to the MOX coverage (intra-sample π between $1.4x10^{-5}$ and $1.4x10^{-3}$, mean $4.9x10^{-4}$, $\pm$ $4.9x10^{-4}$, s.d.) results in lower estimates than for the full coverage, however, these estimates are still higher than the MOX nucleotide diversity (Table 2).

To test the effect of sequencing coverage on strain inference, we repeated the strain deconvolution for samples where the SOX coverage is decreased to the median MOX coverage. This yielded eleven SOX strains, where some are not identical to the SOX strains reconstructed from the full dataset, yet, the four SOX clades remain well supported (**Fig. 8**). Hence, we consider the strain relationships as more reliable indicators of strain diversity than the number of strains alone. Nevertheless, the resulting AFS does not reveal modes according to the strain clades as observed previously in the full coverage analyses (**Supplementary Fig. 3**).

9.2 Strain symbiont composition based on ribosomal proteins

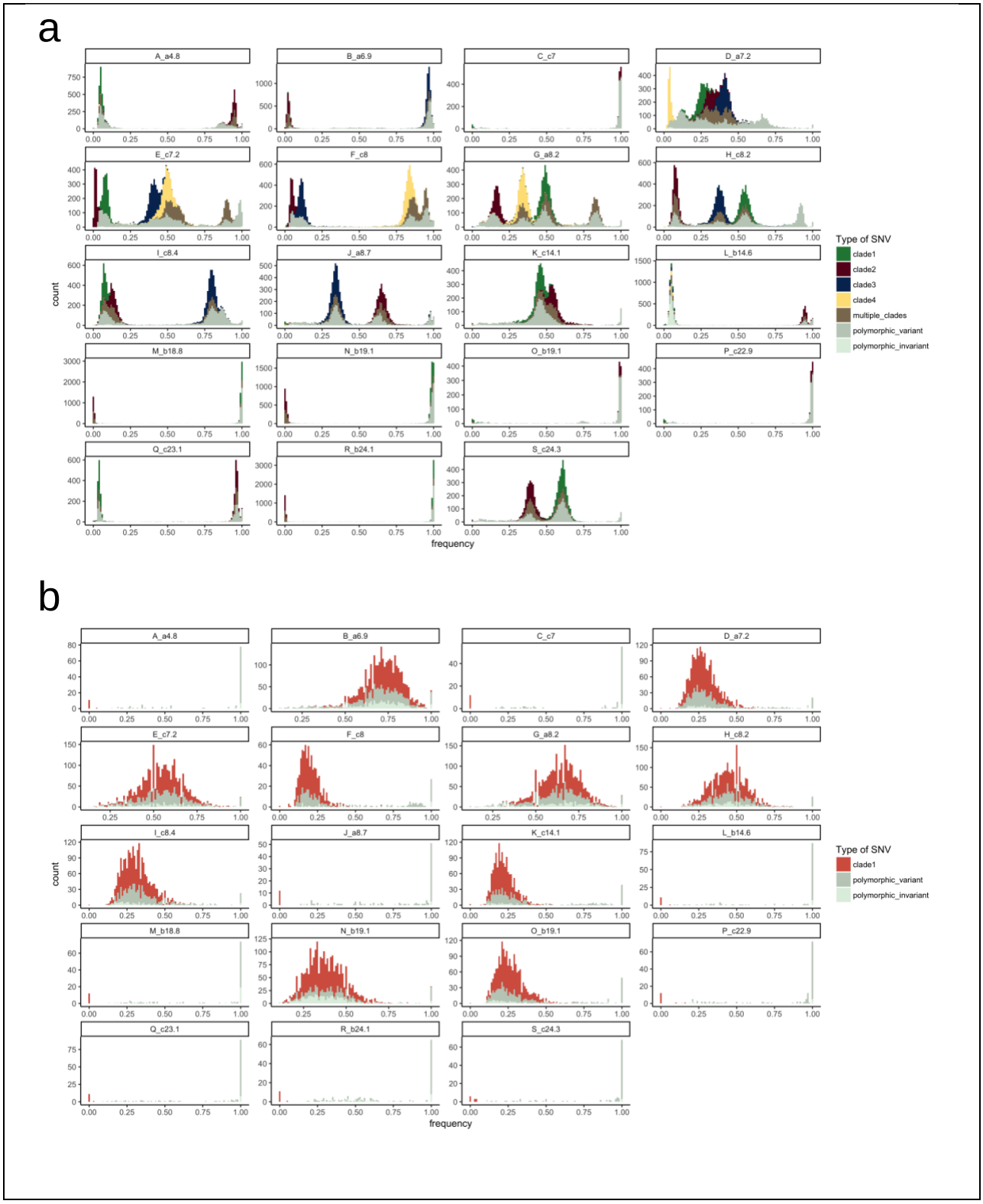
Ribosomal proteins are frequently used for analyzing bacterial diversity. Here we studied the phylogenetic relationships among the reconstructed strains by analyzing ten different ribosomal protein-coding genes present in the core genomes of the two symbiotic species. The phylogenetic tree reconstruction shows that both species clearly split in the tree (**Fig. 7**). However, the low intra-species bootstrap values indicate that relationships among strains can not be confidentially inferred. This can be traced back to the low number of variants within SOX and MOX, respectively. This results in poorly resolved phylogenetic networks for the SOX and MOX populations. In conclusion, because ribosomal proteins are highly conserved, using them for the analysis of within-species diversity leads to the underestimation of the existing diversity.

10 Supplementary tables.

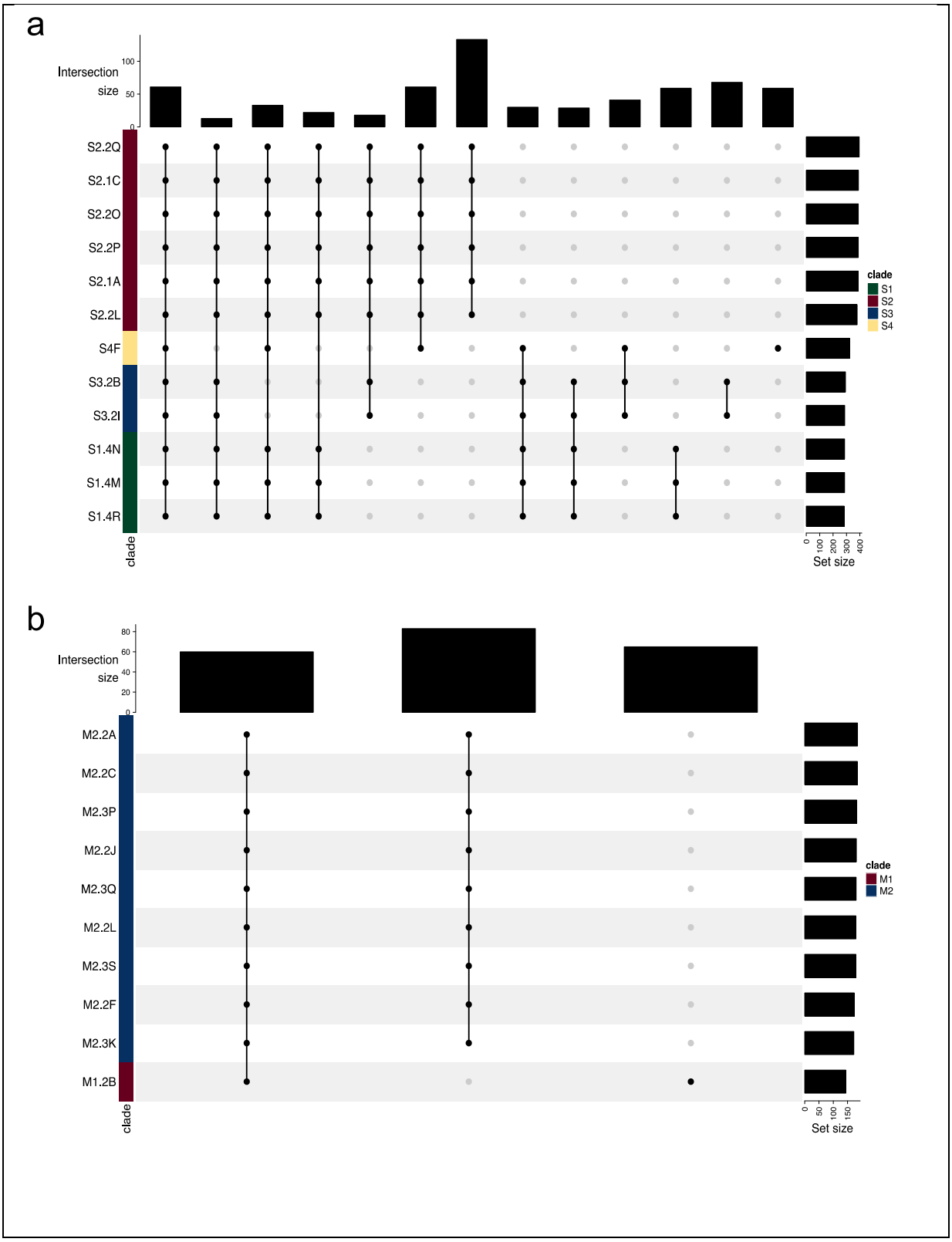
Sample	Size (cm)	Clump	# Reads	# Trimmed reads	# Contigs	# Genes	# Non-redundant genes	# Polymorphic SNVs SOX	# Polymorphic SNVs MOX	# Polymorhic SNVs mitochondria	# Sample-specific SNVs SOX	# Sample-specific SNVs MOX	Dominant strain (≥90%)	#Differential dominant strain SNVs
A_a4.8	4.8	a	37,061,240	37,028,936	1,394,249	1,044,714	697,354	3740	30	3	10	6	-	-
B_a6.9	6.9	a	37,133,252	37,106,306	1,486,937	1,116,948	688,707	1852	3026	1	140	169	-	-
C_c7	7	c	38,285,976	38,252,301	1,418,404	1,006,084	693,453	207	27	0	17	2	S2.1	-
D_a7.2	7.2	a	36,755,366	36,731,976	1,560,929	1,129,326	683,199	11064	2138	0	20	7	-	-
E_c7.2	7.2	c	37,933,850	37,904,916	1,498,588	1,091,911	706,758	10505	2831	1	13	43	-	-
F_c8	8	c	38,060,325	38,033,705	1,432,041	1,074,170	697,580	8923	826	0	242	2	-	-
G_a8.2	8.2	a	37,230,553	37,203,081	1,453,270	1,133,968	679,642	10404	2934	0	22	73	-	-
H_c8.2	8.2	c	38,070,443	38,034,179	1,548,988	1,150,117	699,488	10812	2717	0	28	51	-	-
I_c8.4	8.4	c	38,134,036	38,103,876	1,470,797	1,067,621	706,089	10148	2232	0	52	7	-	-
J_a8.7	8.7	a	37,807,348	37,782,857	1,463,875	1,138,457	678,274	7147	48	0	49	2	-	-
K_c14.1	14.1	c	36,693,086	36,661,232	1,460,859	1,084,916	698,127	6742	1542	0	24	6	-	-
L_b14.6	14.6	b	37,412,053	37,386,973	1,459,725	1,106,129	687,149	3338	39	0	1903	3	-	-
M_b18.8	18.8	b	39,118,344	39,086,779	1,781,668	1,033,575	693,534	314	40	0	14	12	S1.4	42
N_b19.1	19.1	b	37,259,078	37,231,365	1,479,982	1,141,813	691,418	397	2536	0	18	398	S1.4	74
O_b19.1	19.1	b	38,172,030	38,137,180	1,505,470	1,132,309	707,064	315	1934	0	16	26	-	-
P_c22.9	22.9	c	39,623,187	39,589,960	1,473,410	1,101,708	693,062	162	48	0	10	3	S2.2	44
Q_c23.1	23.1	c	38,172,522	38,140,344	1,457,910	1,097,379	683,899	615	43	2	12	3	S2.2	44
R_b24.1	24.1	b	37,435,085	37,410,384	1,492,285	1,134,884	689,875	341	89	0	37	44	S1.4	45
S_c24.3	24.3	c	38,814,311	38,775,353	1,549,186	1,151,312	705,170	6423	32	1	31	5	-	-
Dsc1_a13.6	13.6	a	33,794,815	33,711,379	1,599,186	1,058,588	585,047	-	-	-	-	-	-	-
Dsc2_a23.5	23.5	a	39,994,621	39,945,569	1,285,996	862,788	402,105	-	-	-	-	-	-	-
Dsc3_c18.5	18.5	c	38,976,397	38,944,379	1,734,610	1,211,676	670,826	-	-	-	-	-	-	-
Dsc4_c19.1	19.1	c	36,555,894	36,518,986	1,604,292	1,176,485	703,808	-	-	-	-	-	-	-

Supplementary Table 1: Metagenomes sequencing ans assembly report.

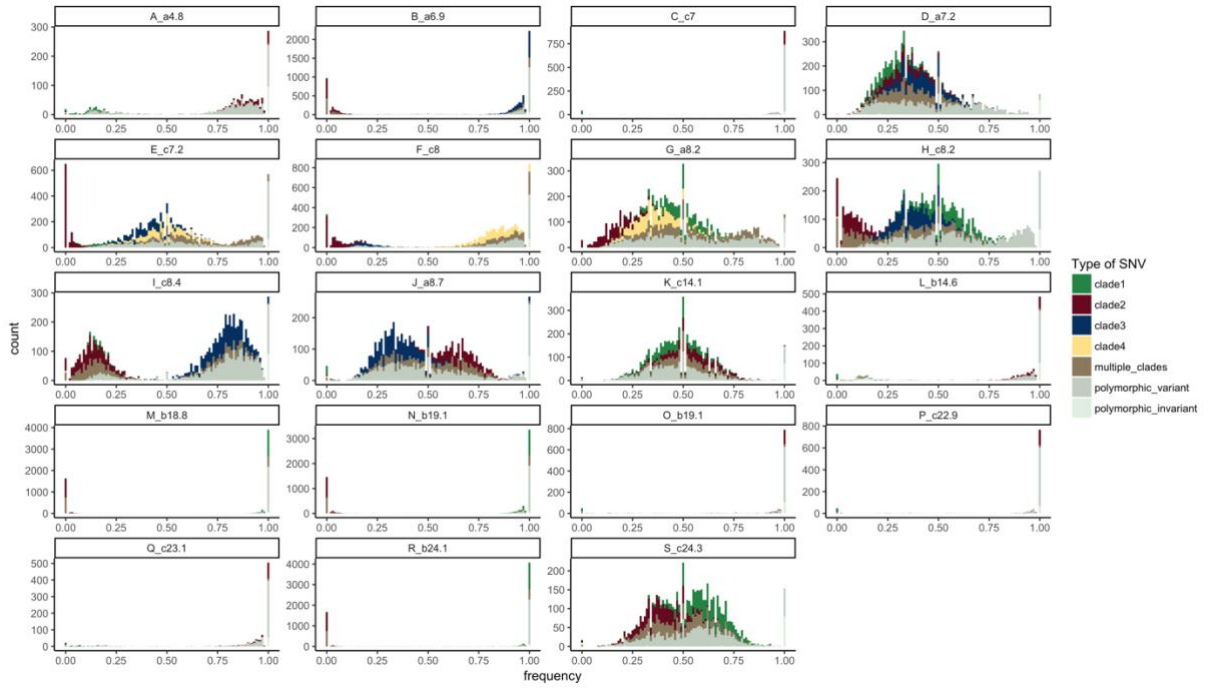
11 Supplementary figures.



Supplementary Figure 1: Intra-sample unfolded allele frequency spectra for **a,** SOX and **b,** MOX. Stacked histogram of the frequencies colored by strain clades (see colors in **Fig. 6**). SNVs are defined as clade-specific when they occur at the same state in every strain from a clade, as variant, if they differ among strains from one clade and invariant if they are not recovered by the strain deconvolution.


a
Intersection size
100
50
0
S2.2Q
S2.1C
S2.2O
S2.2P
S2.1A
S2.2L
S4F
S3.2B
S3.2I
S1.4N
S1.4M
S1.4R
clade
clade
S1
S2
S3
S4
0
100
200
300
400
Set size
b
Intersection size
80
60
40
20
0
M2.2A
M2.2C
M2.3P
M2.2J
M2.3Q
M2.2L
M2.3S
M2.2F
M2.3K
M1.2B
clade
clade
M1
M2
0
50
100
150
Set size

Supplementary Figure 2. **a,** Interaction plots showing exclusive intersects for 731 genes that could be assigned to 5 strains across 12 samples. Note that only intersections larger than 10 are displayed. **b,** Interaction plots showing exclusive intersects for 276 genes that could be assigned to 3 strains across 10 samples.



Supplementary Figure 3. Intra-sample unfolded allele frequency spectra for downsampled SOX to MOX coverage (~36x).