The ISME Journal (2019) 13:2954–2968
https://doi.org/10.1038/s41396-019-0475-z

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

Devani Romero Picazo[1] · Tal Dagan[1] · Rebecca Ansorge[2] · Jillian M. Petersen[3] · Nicole Dubilier[2] · Anne Kupczok[1]

Received: 20 February 2019 / Revised: 22 May 2019 / Accepted: 16 June 2019 / Published online: 8 August 2019

## Abstract

Eukaryotes are habitats for bacterial organisms where the host colonization and dispersal among individual hosts have consequences for the 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 is poorly characterized. Here we show that chemosynthetic symbionts of individual *Bathymodiolus brooksi* mussels constitute genetically isolated subpopulations. The reconstruction of core genome-wide strains from high-resolution metagenomes revealed distinct phylogenetic clades. Nucleotide diversity and strain composition vary along the mussel life span and individual hosts show a high degree of genetic isolation. 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 conclude that bacterial colonization dynamics over the host life cycle is thus an important determinant of population structure and genome evolution of horizontally transmitted symbionts.

## Introduction

Bacteria inhabit most eukaryotes where their presence has consequences for key aspects of the host biology [1], such as host development [2], nutrition [3], or behavior [4]. From the bacterial perspective, animals constitute an ecological niche, where microbial communities utilize the resources of their host habitat [5]. 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 [6] or the host physiological state (e.g., hibernation [7] or pregnancy [8]). In addition, changes in the host environmental conditions such as temperature [9] or the availability of reduced compounds [10] can have an effect on 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 [11]. 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 [12]. Examples are monoclonal or biclonal populations observed in symbiotic bacteria inhabiting grass sharpshooter [13] and pea aphids [14]. Furthermore, the geographic isolation of vertically transmitted bacteria leads to genetic isolation and to symbiont genome reduction

**Supplementary information** The online version of this article (https://doi.org/10.1038/s41396-019-0475-z) contains supplementary material, which is available to authorized users.

✉ Devani Romero Picazo
dpicazo@ifam.uni-kiel.de

✉ Anne Kupczok
akupczok@ifam.uni-kiel.de

[1] Genomic Microbiology Group, Institute of General Microbiology, Christian-Albrechts University, Kiel, Germany

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

[3] Division of Microbiology and Ecosystem Science, University of Vienna, Wien, Austria

over time as a consequence of genetic drift [15]. 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 [16]. Nonetheless, the genetic diversity of horizontally transmitted microbial populations may also be reduced due to bottlenecks during symbiont transmission and host colonization. Stochastic effects on the colonization of horizontally transmitted bacteria may manifest themselves in differences in microbiota strain composition among hosts [17, 18]. 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 [19] and of the honey bee gut microbiome [20]. 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 [21]. 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 the 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 [22, 23]. Most Bathymodiolus species harbor only a single 16S rRNA phylotype for each symbiont, including *B. brooksi* [24]. Nevertheless, 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 [25]. Mussel gills constantly develop new filaments that are continuously infected [26]. However, whether the new gill filaments in *B. 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 genetic diversity. Furthermore, we quantified the level of genetic isolation among communities across individual mussels and its impact on symbiont genome evolution. For that, we implemented a high-resolution metagenomics approach that captures genome-wide diversity for both symbionts in multiple *B. brooksi* individuals from a single site.

## Results

### 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 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 [27]; 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 131 m apart (Supplementary Fig. 1). Such a ‘patchy’ distribution has often been observed in deep-sea mussels [28]. 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 250 bp per sample, Supplementary Table 1). The resulting metagenomic sequencing data were analyzed by a gene-based binning approach [29].

The prediction of protein-coding genes from the assembled metagenomes yielded a nonredundant gene catalog of 4.4 million genes that potentially contains 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 (Supplementary 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 2518 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 1568 genes (62.3%) that have homologs in MOX-related genomes. The SOX core genome contains 1439 genes, has a total length of 1.27 Mbp and is considered as 80.2%

complete. It contains 1188 genes (82.6%) with homologs in SOX-related genomes. In addition to the SOX and MOX core genomes, our analysis revealed a third MGS of 1449 genes (Supplementary Fig. 2) 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 (Supplementary Figs. 2 and 3). 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* [10] (Supplementary Fig. 4).

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 4652 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 3026 (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 [25]. We further revealed that there is genetic diversity in the MOX symbiont.

## 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 [30]. 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. [31]. Here we consider a strain to be a genetic entity that is present in multiple hosts and is 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 [32]. This revealed that SOX is composed of 11 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. 1b). Notably, 849 out of the total SNVs found in 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. 1e). Of the total SNVs, 1138 (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 ~8200 SNVs (0.44% different positions genome-wide) while two MOX differ by at most ~2700 SNVs (0.19% different positions genome-wide) (Fig. 1), thus, SOX has higher genome diversity in comparison 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 [31]). 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 Fig. 5). Importantly, the observed difference in strain diversity between MOX and SOX cannot be explained by the difference in sequencing coverage (Supplementary Fig. 6).

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. 1a). Only one or two strains were detected with a frequency of at least 5% in small mussels (≤7 cm), 2–9 strains in medium-sized mussels (7.2–14.1 cm) and 1–2 strains in large mussels (14.6–24.1 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, and R) and only one is

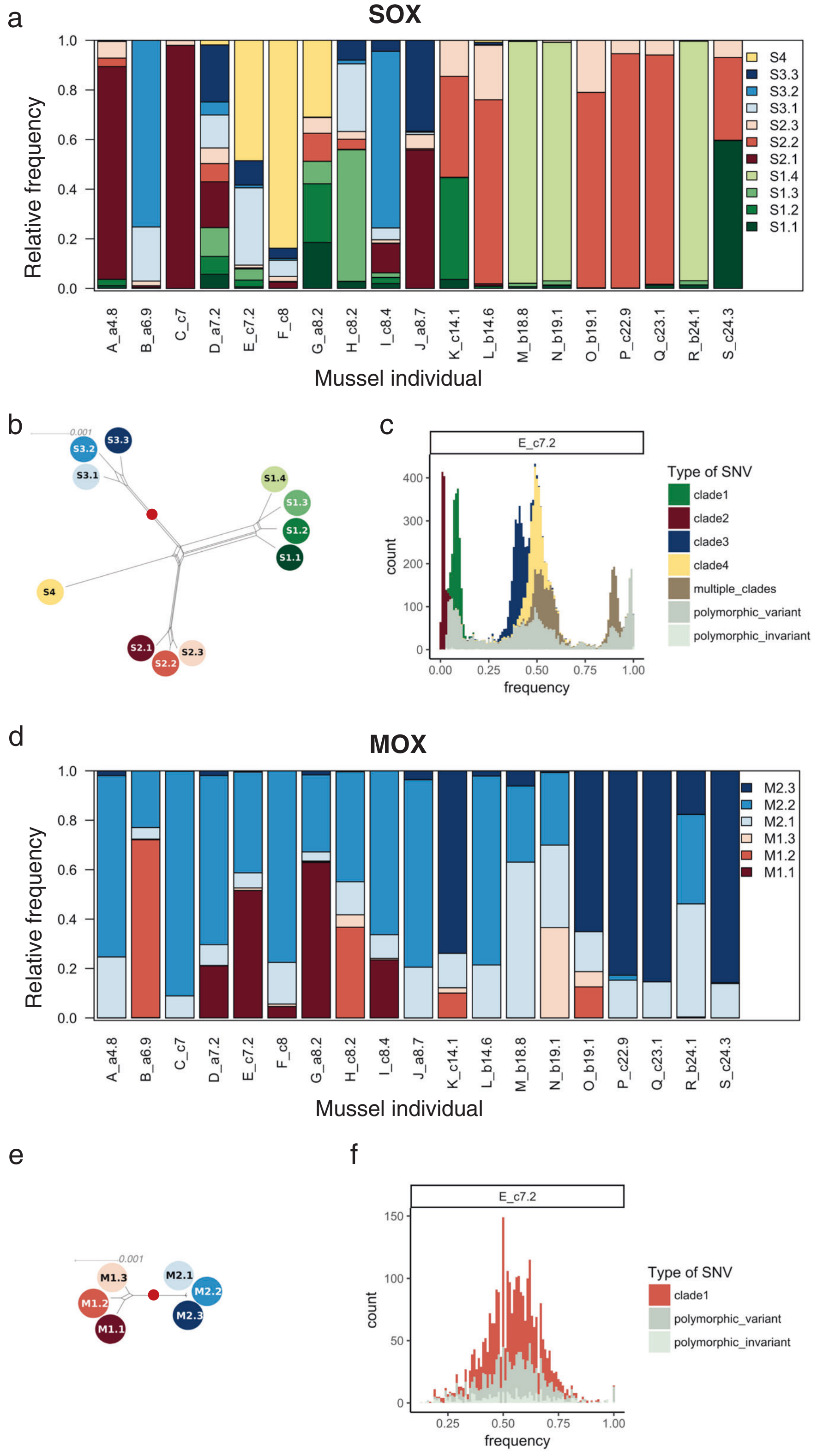

**Fig. 1** Symbiont strain abundances (**a**, **d**), symbiont strain relationships (**b**, **e**), and example allele frequency spectra (**c**, **f**). **a**, **b**, **c** A total of 11 strains reconstructed for SOX. These cluster into four clades, with four, two times three and one strain per clade, labeled 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 8171 SNVs (strains S3.2 and S4 sequence identity 99.36%). Minimum number of SNVs between strains of different clades is 6451 (strains S1.1 and S2.1, sequence identify 99.49%). **d**, **e**, **f**, Six strains reconstructed for MOX. These cluster into two clades, labeled by shades of red and blue. Strains differ by between 105 (strain M2.2 and M2.3, sequence identity 99.99%) and 2677 SNVs (strain M1.1 and M2.1, sequence identity 99.81). The minimum number of SNVs differentiating strains from different clades is 2224 (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. 7)

a small mussel (C). The dominant strain is either S1.4, S2.1, or S2.2 (Fig. 1a; Supplementary Table 1). The MOX strain composition across mussels shows that each MOX strain occurs (frequency ≥5%; Fig. 1d) in 4–17 mussels and each mussel contains 2–4 MOX strains. In addition, 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. 1c, f; Supplementary Fig. 7). 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. 8).

Overall, our results revealed that the symbiont populations are composed of strains that cluster into 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 [33].

## 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 with 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 (Supplementary Fig. 9). The relative rate of nonsynonymous to synonymous substitutions has been shown to depend on the divergence of the analyzed species [34, 35]. 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). The 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 [36]. 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 ($NI \sim 1$) 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 with 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 with 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 [37]. This agrees with the larger degree of divergence observed here for SOX. Since we observed no evidence for positive selection in the symbiont core genomes, we suggest that the strains constitute cohesive genetic units within one ecotype [38], 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

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

a

| | SOX | | MOX | |
|---|---|---|---|---|
| | Divergent | Polymorphic | Divergent | Polymorphic |
| Nonsynonymous SNVs | 5004 | 990 | 2115 | 704 |
| Synonymous SNVs | 10,577 | 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 | |

**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

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

| | SOX | MOX |
|---|---|---|
| Intra-sample π range | $5.2 \times 10^{-5} - 3.6 \times 10^{-3}$ | $5.6 \times 10^{-6} - 7.0 \times 10^{-4}$ |
| Intra-sample π mean | $1.4 \times 10^{-3} \pm 1.3 \times 10^{-3}$ (s.d) | $2.7 \times 10^{-4} \pm 2.8 \times 10^{-4}$ (s.d) |
| Intra-sample π median | $6.7 \times 10^{-4}$ | $1.4 \times 10^{-4}$ |
| 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 |

*Prochlorococcus* spp. [31] and in SOX symbionts of other *Bathymodiolus* species [25].

## 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 [5]. Geographic isolation between habitats results in genetic isolation and contributes to the formation of cohesive associations of genetic loci [30]. Previous studies showed that geographic isolation during vertical transmission can lead to the reduction of intra-host genetic diversity in the bacterial populations [12], nonetheless, the degree of isolation remains understudied for horizontally transmitted microbes. To characterize the contribution of geographic isolation to strain formation in the *Bathymodiolus* 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 unusually 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 $\pi^{SOX}$ among different mussels (intra-sample $\pi^{SOX}$ between $5.2 \times 10^{-5}$ and

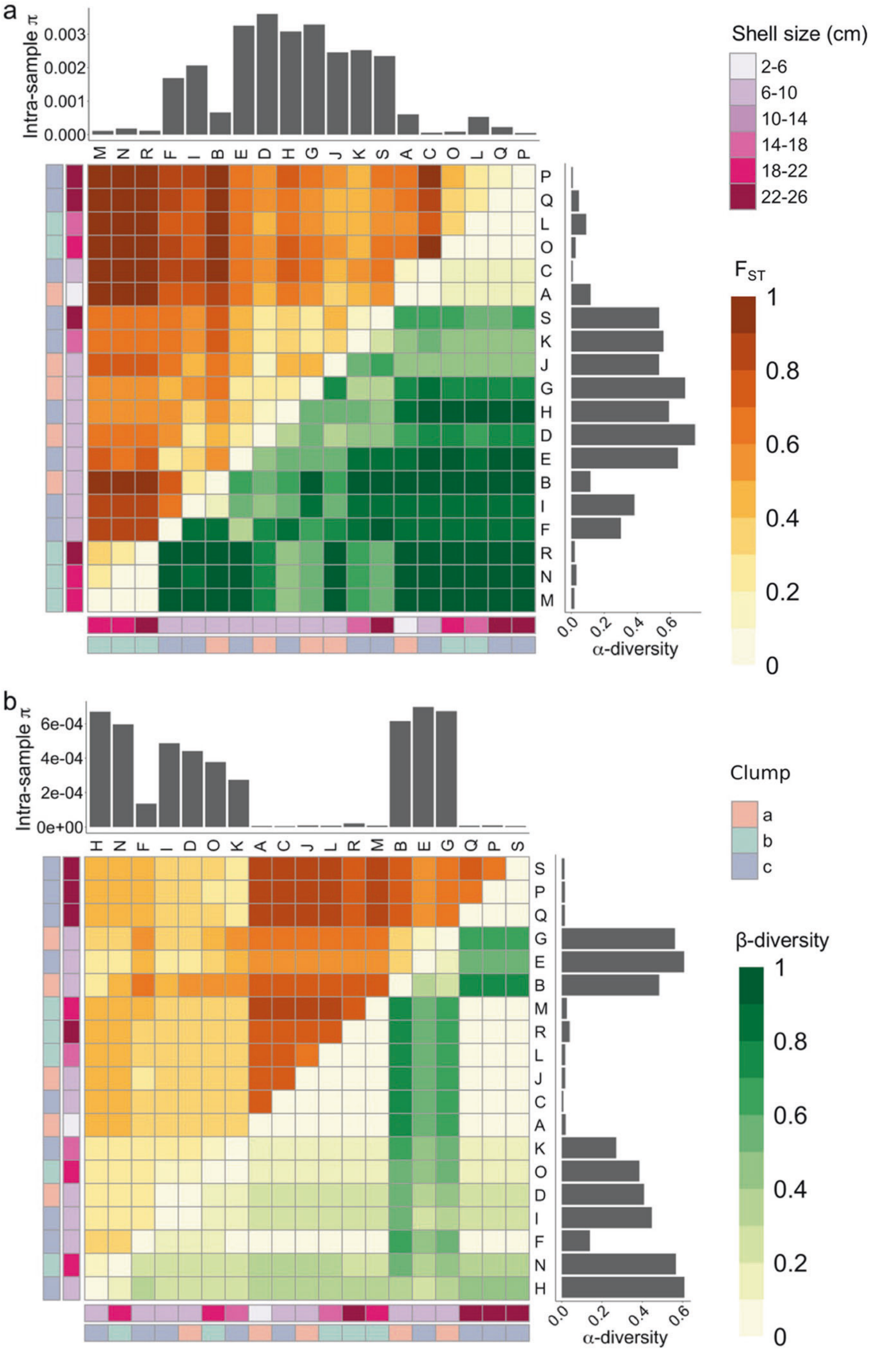

**Fig. 2** Symbiont population structure for **a** SOX and **b** MOX. Top left triangle: intra-sample π and symbiont fixation index ($F_{ST}$) based on SNVs. Lower right triangle: α- and β-diversity based on reconstructed strains. Rows and columns are labeled by sample name, sample location, and shell 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 P, Q, S is 0.0033; mean pairwise β-diversity of group composed of A, C, J, L, M, R is 0.012

$3.6 \times 10^{-3}$, Table 2, Fig. 2). Furthermore, $\pi^{SOX}$ and the SOX α-diversity are significantly positively correlated ($\rho^2 = 0.98$, $p < 10^{-6}$, Spearman correlation, Fig. 2a); hence, the intra-sample strain diversity is well explained by the nucleotide diversity. The variability in $\pi^{SOX}$ agrees with the three age-related groups observed before for the number of SOX strains across mussel size. Small mussels (≤7 cm) and large mussels (14.6–24.1 cm) have a low $\pi^{SOX}$ and harbor one to two strains. Medium-sized mussels (7.2–14.1 cm) have a high $\pi^{SOX}$ and harbor two to nine strains. The community in the largest mussel is an exception, as it has a high $\pi^{SOX}$, 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 with SOX (intra-sample $\pi^{MOX}$ between $5.6 \times 10^{-6}$ and $7.0 \times 10^{-4}$, Table 2, Wilcoxon signed rank test, $p = 0.015$, Fig. 2). Similar to SOX, the MOX α-diversity is significantly positively correlated with $\pi^{MOX}$ ($\rho^2 = 0.89$, $p < 10^{-6}$, Spearman correlation) (Fig. 2b). 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, and $\pi^{MOX}$ between $5.6 \times 10^{-6}$ and

$2.1 \times 10^{-5}$), 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, and $\pi^{MOX}$ between $1.4 \times 10^{-4}$ and $7.0 \times 10^{-4}$). These groups are not associated with mussel size. Taken together, we observed a strong correlation between the nucleotide diversity $\pi$ and α-diversity for both symbionts. Notably, $\pi$ 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 $\pi$ values estimated here to other microbiome studies shows that higher $\pi^{SOX}$ have been observed in other *Bathymodiolus* species (mean between $2.2 \times 10^{-3}$ and $3.9 \times 10^{-3}$) [25]. The average SOX and MOX nucleotide diversity estimated here is within the range of $\pi$ values observed in the clam *Solemya velum* microbiome where the symbiont transmission mode is thought to be a mixture of vertical and horizontal transmission [39]. Furthermore, our $\pi$ estimates are lower than those observed for most bacterial species in the human gut microbiome that are considered to be horizontally transmitted [19].

## Geographic isolation of bacterial communities associated with individual mussels

Symbiont transmission mode is an important determinant of the community assembly dynamics [11]. 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 [40]. 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. 2); hence, most mussels in our sample harbor an isolated symbiont subpopulation of SOX and MOX. The SOX β-diversity is significantly positively correlated with $F_{ST}^{SOX}$ ($\rho^2 = 0.7$, $p < 10^{-6}$, 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. 2). 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. 2a). Notably, both groups are composed of large mussels only that are characterized by a low $\pi^{SOX}$.

The distribution of pairwise $F_{ST}^{MOX}$ revealed two main groups: one mussel group is characterized by high pairwise $F_{ST}^{MOX}$ and low $\pi^{MOX}$ while the other group is characterized by lower $F_{ST}^{MOX}$ and high $\pi^{MOX}$ (Fig. 2b). These correspond to the previously described groups, where one contains mussels with a low $\pi^{MOX}$ and strains from clade M2 and the other group contains mussels with a high $\pi^{MOX}$ 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 consisting of large mussels (P, Q, and S) is characterized by the presence of strain M2.3 and the absence of clade M1. Another group (A, C, J, L, M, and 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 [41], whereas the association with host genetics was reported to be weak in humans [42]. 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 Fig. 10). 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

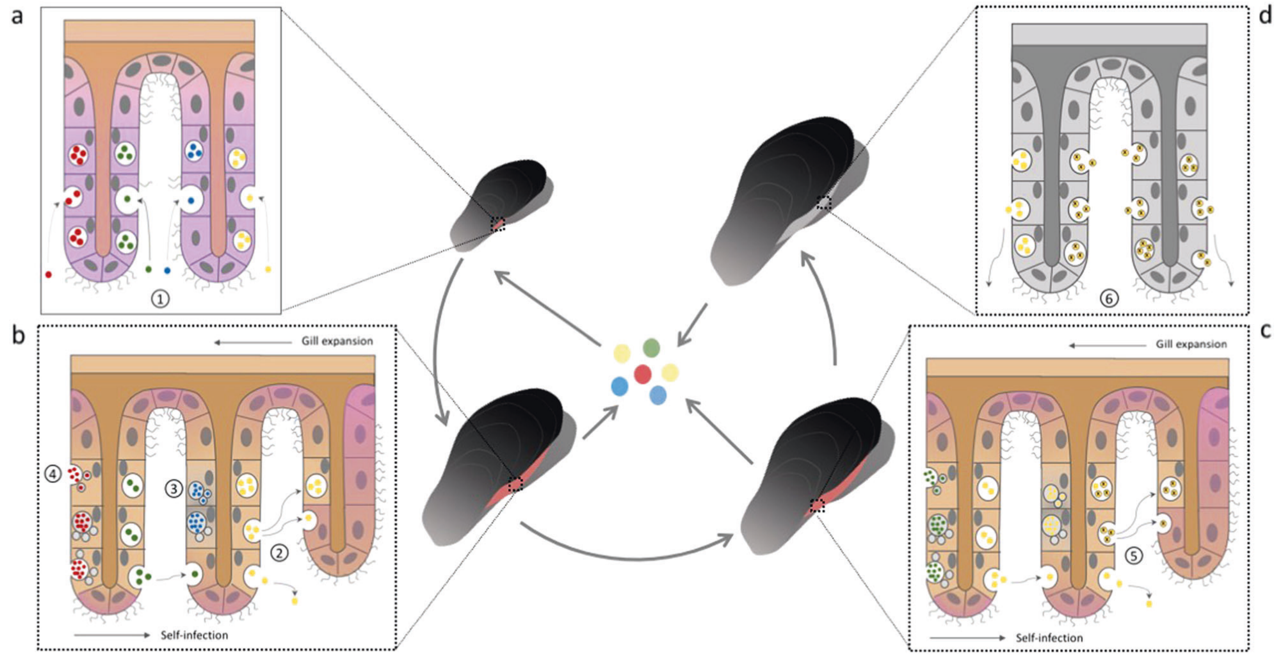

**Fig. 3** Symbiont colonization dynamics. **a** The postlarvae mussel gill does not take up endosymbionts until the gill presents several filaments and the gill epithelial cells reach a determined developmental stage [26]. At this time point, the filaments are simultaneously infected by different strains 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 [26], 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 [44]. 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 [57] (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* [16]

vents [25]. 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.

## 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 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 is the result of 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. 3). In our model, bacteria are acquired from the environmental symbiont pool in postlarvae mussels. The symbionts colonize every tissue in the beginning, to later restrict their localization to the gill tissue, as suggested by previous studies [43]. 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 [44, 45]. Nevertheless, the loss of central metabolic enzymes suggests that bacteria disperse in a dormant state [46]. We hypothesize that the dormancy of free-living symbionts and the preservation of few symbiont cells inside bacteriocytes [23] 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 [26], 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 provides a plausible explanation for the observed pattern of strong symbiont genetic isolation between mussels and of reduced SOX strain diversity in large mussels.

Notably, our results are in contrast to a recent study on other *Bathymodiolus* species from hydrothermal vents, which concluded that SOX populations from individual mussels of the same site intermix [25]. 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 (Supplementary Fig. 1); thus, the concentration of symbionts in the surrounding seawater may be correspondingly low. The low symbiont abundance would result in 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 life span 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 life time. 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 [47], 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 [15]. Serial founder effects and recombination barriers due to geographic isolation are important drivers of lineage formation in bacteria [38]. Reduction of genetic diversity due to transmission bottlenecks is considered a hallmark of pathogen genome evolution [48]; examples are *Yersinia pestis* [49] and *Listeria monocytogenes* [50]. 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.

## Methods

### Collection and sequencing

Twenty-three individuals of *B. 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: −1073 m, clump b: 28.1241, −89.1401 depth: −1073 m, clump c: 28.1237, −89.1404 depth: −1073 to 1078 m). The gills from each mussel individual were dissected immediately after retrieval and homogenized with sterilized stainless steal 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. [51]. TruSeq library preparation and sequencing using Illumina HiSeq2500 were 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 have been deposited in NCBI under BioProject PRJNA508280.

### Construction of the nonredundant gene catalog

Illumina paired-end raw reads from the samples were trimmed for adapters and filtered by quality using BBMap tools [52]. Only reads with more than 30 bp 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 [53]. Genes were predicted ab initio on contigs with metaProdigal [54]. These predicted genes were clustered by single-linkage according to sequence similarity using BLAT [55] (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 nonpartial 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 nonredundant 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 nonredundant genes. For each of the clusters, we reconstructed a consensus sequence as cluster representative. To this end, we took the majority nucleotide at each position (ties were resolved randomly).

## Taxonomic annotation of gene catalog

Taxonomic annotation of the gene catalog was performed by aligning the translated genes to the nonredundant protein NCBI database (date: 24/05/18) using diamond [56] (*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 *B. platifrons* mitochondrial protein sequences (NC_035421.1) [57] (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 [46] (80% of protein identity and 100% of query coverage). This allowed to identify 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 (Supplementary Fig. 3). 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).

## 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 [58]. 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 [59] 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).

## Genome binning and symbiont core genome identification

Next, we performed co-abundance gene segregation by using a canopy clustering algorithm [29], 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 allowed us to cluster 900,310 genes into 98,944 co-abundant gene groups (3 to 699 genes) and three MGSs (≥700 genes). An additional filter was applied to the MGSs to obtain final bins by removing outlier genes based on their coverage (Supplementary 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 of the samples. The bins after outlier gene removal constitute the core genomes of the MGSs. From the marker gene catalog, 37 (of 39) MOX marker genes and 77 (of 86) SOX marker genes were included in the core genomes. In addition, we checked for the completeness of the symbiont bins with CheckM, by screening for the presence of Gammaproteobacteria universal single copy marker genes [60].

## SNV discovery on the core genomes

To perform SNV discovery, we mapped the downsampled reads individually for each sample to the gene catalog. Because sample size has been shown to have an effect on variant detection [61], we normalized the data across samples. To this end, we normalized each sample to the smallest median coverage found in a sample (482× coverage for SOX, 36× coverage for MOX and 568× for mitochondrial genes). LoFreq was used for probabilistic realignment and variant calling of each sample independently [62]. SNVs detected with LoFreq have been hard filtered using the parameters suggested by GATK best practices [63]. 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.

## Population structure analyses

SNV data are used for calculating intra-sample and inter-sample nucleotide diversity (π) as applied before to human gut microbiome species [19]. Intra-sample nucleotide diversity (π) is given as:

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

where $S$ 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 $S_1$ and $S_2$ correspond to the two samples compared:

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

Finally, these diversity measures are used to estimate the fixation index ($F_{ST}$), which measures genetic differentiation based on the intra and inter-sample nucleotide diversity.

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

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.

## Strain deconvolution

We reconstructed the strains for the core genomes with DESMAN [32]. 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 5000 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 11 to 12 strains, therefore the number of inferred SOX strains is 11. We did not run fewer numbers of strains due to the presence of large posterior mean deviances between runs with a small strain number. In addition, we ran DESMAN for the SOX dataset that was subsampled to the MOX coverage with no replicates and 11 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 was reconstructed using SplitsTree [64] and uncorrected distances. The position of the root in the splits network was estimated by the minimum ancestral deviation method [65], which uses maximum likelihood phylogenetic trees inferred with IQ-TREE [73].

## α- and β-diversity

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 eveness [66] implemented in the R package 'Picante' [67]. β-diversity was estimated using the weighted Unifrac distance, which is implemented in the R package 'GUniFrac' [68]. This measure quantifies differences in strain community composition between two samples and accounts for phylogenetic relationships.

## 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 AFS 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.

## pN/pS and Neutrality Index estimation

We estimated *pN/pS* for the two symbiont populations, which is a variant of *dN/dS* that can be used based on intraspecies 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 undefinded 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 [69].

The 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 [36]. 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 [36].

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, i.e., the SNVs that do not differentiate among strains. Second, we used a more restrictive level of divergence. To this end, we excluded putative recently acquired SNVs from the set of divergent SNVs by discarding those that have multiple states among strains from the same clade. 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 [70].

## Strain genetic diversity based on ribosomal proteins

The core genomes from the two bacterial species were screened for the presence of ribosomal proteins. To this end, the amino acidic translations of the largest coding sequence found in a gene cluster have been functionally annotated with eggNOG [71]. A ribosomal protein was detected in the core genomes of both bacterial species when the same KEGG orthologous group could be found in the two genomes. This led to a set of ten ribosomal proteins which were used for further analyses (50S ribosomal proteins L10, L27, L28, L29, L30, l32 and l35, and 30S ribosomal proteins S2, S9, and S21). Concatenated alignments for the 17 SOX and MOX strains were calculated with mafft [72] and the maximum likelihood phylogeny for this alignment was reconstructed using IQ-TREE [73] with 1000 bootstrap replicates.

**Acknowledgements** We thank the captain and crew of the E/V *Nautilus* 2015 Expedition, the team of the ROV *Hercules*, the chief scientist on this expedition leg, Eric Cordes, and the expedition leader, Mike Brennan. In addition, we thank Elie Jami, Robin Koch, Tanita Wein, and Christian Wöhle for critical comments on the manuscript. This work was supported by the CRC1182 *Origin and Function of Metaorganisms*, the Bioinformatics Network at Kiel University, and the North-German Supercomputing Alliance (HLRN, project shb00002). Additional funding support for ND, JMP, and RA came from the Max Planck Society, the Gordon and Betty Moore Foundation Marine Microbiology Initiative Investigator Award through grant GBMF3811 to ND and a European Research Council Advanced Grant (BathyBiome, Grant 340535 to ND).

**Author contributions** AK, TD, JMP, and ND designed the study; RA, JMP, and ND collected the data; DRP analyzed the data; DRP, AK, and TD interpreted the results with contribution from RA, DRP, AK, and TD wrote the manuscript with contributions from all authors.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## References

1. McFall-Ngai M, Hadfield MG, Bosch TCG, Carey HV, Domazet-Lošo T, Douglas AE, et al. Animals in a bacterial world, a new imperative for the life sciences. Proc Natl Acad Sci. 2013;110:3229–36.
2. McFall-Ngai MJ. The importance of microbes in animal development: lessons from the squid-vibrio symbiosis. Annu Rev Microbiol. 2014;68:177–94.
3. Shabat SKB, Sasson G, Doron-Faigenboim A, Durman T, Yaacoby S, Berg Miller ME, et al. Specific microbiome-dependent mechanisms underlie the energy harvest efficiency of ruminants. ISME J. 2016;10:2958.
4. Schretter CE, Vielmetter J, Bartos I, Marka Z, Marka S, Argade S, et al. A gut microbial factor modulates locomotor behaviour in *Drosophila*. Nature. 2018;563:402–6.
5. Costello EK, Stagaman K, Dethlefsen L, Bohannan BJM, Relman DA. The application of ecological theory toward an understanding of the human microbiome. Science. 2012;336:1255–62.
6. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 2014;505:559–63.
7. Sommer F, Ståhlman M, Ilkayeva O, Arnemo JM, Kindberg J, Josefsson J, et al. The gut microbiota modulates energy metabolism in the hibernating brown bear *Ursus arctos*. Cell Rep. 2016;14:1655–61.
8. Koren O, Goodrich JK, Cullender TC, Spor A, Laitinen K, Kling Bäckhed H, et al. Host remodeling of the gut microbiome and metabolic changes during pregnancy. Cell. 2012;150:470–80.
9. Jones RJ, Hoegh-Guldberg O, Larkum AWD, Schreiber U. Temperature-induced bleaching of corals begins with impairment of the CO2 fixation mechanism in zooxanthellae. Plant Cell Environ. 1998;21:1219–30.
10. Riou V, Halary S, Duperron S, Bouillon S, Elskens M, Bettencourt R, et al. Influence of $CH_4$ and $H_2S$ availability on symbiont distribution, carbon assimilation and transfer in the dual symbiotic vent mussel *Bathymodiolus azoricus*. Biogeosciences. 2008;5:1681–91.
11. Bright M, Bulgheresi S. A complex journey: transmission of microbial symbionts. Nat Rev Microbiol. 2010;8:218–30.
12. Wernegreen JJ. Endosymbiont evolution: predictions from theory and surprises from genomes: endosymbiont genome evolution. Ann N Y Acad Sci. 2015;1360:16–35.
13. Woyke T, Tighe D, Mavromatis K, Clum A, Copeland A, Schackwitz W, et al. One bacterial cell, one complete genome. PLoS ONE. 2010;5:e10314.
14. Guyomar C, Legeai F, Jousselin E, Mougel C, Lemaitre C, Simon J-C. Multi-scale characterization of symbiont diversity in the pea aphid complex through metagenomic approaches. Microbiome. 2018;6:181.
15. Boscaro V, Kolisko M, Felletti M, Vannini C, Lynn DH, Keeling PJ. Parallel genome reduction in symbionts descended from closely related free-living bacteria. Nat Ecol Evol. 2017;1:1160.
16. Klose J, Polz MF, Wagner M, Schimak MP, Gollner S, Bright M. Endosymbionts escape dead hydrothermal vent tubeworms to enrich the free-living population. Proc Natl Acad Sci USA. 2015;112:11300–5.
17. Hagen MJ, Hamrick JL. Population level processes in *Rhizobium leguminosarum* bv. *trifolii*: the role of founder effects. Mol Ecol. 1996;5:707–14.
18. Vega NM, Gore J. Stochastic assembly produces heterogeneous communities in the *Caenorhabditis elegans* intestine. PLOS Biol. 2017;15:e2000633.
19. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, et al. Genomic variation landscape of the human gut microbiome. Nature. 2013;493:45–50.
20. Ellegaard KM, Engel P. Genomic diversity landscape of the honey bee gut microbiota. Nat Commun. 2019;10:446.
21. Wollenberg MS, Ruby EG. Population Structure of *Vibrio fischeri* within the light organs of *Euprymna scolopes* Squid from Two Oahu (Hawaii) Populations. Appl Environ Microbiol. 2009;75:193–202.
22. Won Y-J, Hallam SJ, O'Mullan GD, Pan IL, Buck KR, Vrijenhoek RC. Environmental acquisition of thiotrophic endosymbionts by deep-sea mussels of the genus *Bathymodiolus*. Appl Environ Microbiol. 2003;69:6785–92.
23. Dubilier N, Windoffer R, Giere O. 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. Mar Ecol Prog Ser. 1998;165:187–93.
24. Duperron S, Sibuet M, MacGregor BJ, Kuypers MMM, Fisher CR, Dubilier N. Diversity, relative abundance and metabolic potential of bacterial endosymbionts in three B*athymodiolus* mussel species from cold seeps in the Gulf of Mexico. Environ Microbiol. 2007;9:1423–38.
25. Ansorge R, Romano S, Sayavedra L, Kupczok A, Tegetmeyer HE, Dubilier N, et al. Diversity matters: deep-sea mussels harbor multiple symbiont strains. https://www.biorxiv.org/content/10.1101/531459v1 2019.
26. Wentrup C, Wendeberg A, Schimak M, Borowski C, Dubilier N. Forever competent: Deep-sea bivalves are colonized by their chemosynthetic symbionts throughout their lifetime. Environ Microbiol. 2014;16:3699–713.
27. Schöne BR, Giere O. Growth increments and stable isotope variation in shells of the deep-sea hydrothermal vent bivalve mollusk *Bathymodiolus brevior* from the North Fiji Basin, Pacific Ocean. Deep Sea Res Part Oceano Res Pap. 2005;52:1896–910.
28. Van Dover C. Community structure of mussel beds at deep-sea hydrothermal vents. Mar Ecol Prog Ser. 2002;230:137–58.
29. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol. 2014;32:822–8.
30. Shapiro BJ, Polz MF. Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol. 2014;22:235–47.
31. Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. Science. 2014;344:416–20.
32. Quince C, Delmont TO, Raguideau S, Alneberg J, Darling AE, Collins G, et al. DESMAN: a new tool for de novo extraction of strains from metagenomes. Genome Biol. 2017;18:181.
33. Ho P-T, Park E, Hong SG, Kim E-H, Kim K, Jang S-J, et al. Geographical structure of endosymbiotic bacteria hosted by *Bathymodiolus* mussels at eastern Pacific hydrothermal vents. BMC Evol Biol. 2017;17:121.
34. Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, et al. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol. 2006;239:226–35.
35. Kryazhimskiy S, Plotkin JB. The population genetics of dN/dS. PLoS Genet. 2008;4:e1000304.
36. Rand DM, Kann LM. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. Mol Biol Evol. 1996;13:735–48.
37. Lorion J, Kiel S, Faure B, Kawato M, Ho SYW, Marshall B, et al. Adaptive radiation of chemosymbiotic deep-sea mussels. Proc R Soc B. 2013;280:20131243.
38. Achtman M, Wagner M. Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol. 2008;6:431–40.
39. Russell SL, Corbett-Detig RB, Cavanaugh CM. Mixed transmission modes and dynamic genome evolution in an obligate animal–bacterial symbiosis. ISME J. 2017;11:1359–71.

40. Sprockett D, Fukami T, Relman DA. Role of priority effects in the early-life assembly of the gut microbiota. Nat Rev Gastroenterol Hepatol. 2018;15:197–205.
41. Benson AK, Kelly SA, Legge R, Ma F, Low SJ, Kim J, et al. Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. Proc Natl Acad Sci. 2010;107:18933–8.
42. Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, Zeevi D, et al. Environment dominates over host genetics in shaping human gut microbiota. Nature. 2018;555:210–5.
43. Wentrup C, Wendeberg A, Huang JY, Borowski C, Dubilier N. Shift from widespread symbiont infection of host tissues to specific colonization of gills in juvenile deep-sea mussels. ISME J. 2013;7:1244–7.
44. Ikuta T, Takaki Y, Nagai Y, Shimamura S, Tsuda M, Kawagucci S, et al. Heterogeneous composition of key metabolic gene clusters in a vent mussel symbiont population. ISME J. 2016;10:990–1001.
45. Fontanez KM, Cavanaugh CM. Evidence for horizontal transmission from multilocus phylogeny of deep-sea mussel (Mytilidae) symbionts. Environ Microbiol. 2014;16:3608–21.
46. Ponnudurai R, Kleiner M, Sayavedra L, Petersen JM, Moche M, Otto A, et al. Metabolic and physiological interdependencies in the *Bathymodiolus azoricus* symbiosis. ISME J. 2017;11:463–77.
47. Reuter M, Pedersen JS, Keller L. Loss of Wolbachia infection during colonisation in the invasive Argentine ant Linepithema humile. Heredity. 2005;94:364–9.
48. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host evolution of bacterial pathogens. Nat Rev Microbiol. 2016;14:150–62.
49. Gonzalez RJ, Lane MC, Wagner NJ, Weening EH, Miller VL. Dissemination of a highly virulent pathogen: tracking the early events that define infection. PLOS Pathog. 2015;11:e1004587.
50. Zhang T, Abel S, Wiesch PAzur, Sasabe J, Davis BM, Higgins DE, et al. Deciphering the landscape of host barriers to *Listeria monocytogenes* infection. Proc Natl Acad Sci USA. 2017;114:6334–9.
51. Zhou J, Bruns MA, Tiedje JM. DNA recovery from soils of diverse composition. Appl Environ Microbiol. 1996;62:316–22.
52. Bushnell, Brian. BBMap. sourceforge.net/projects/bbmap/.
53. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017; 27:824–34.
54. Hyatt D, Locascio PF, Hauser LJ, Uberbacher EC. Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics. 2012;28:2223–30.
55. Kent WJ. BLAT — The BLAST —Like Alignment Tool. Genome Res. 2002;12:656–64.
56. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.
57. Sun J, Zhang Y, Xu T, Zhang Y, Mu H, Zhang Y, et al. Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. Nat Ecol Evol. 2017;1:0121.
58. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60.
59. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
60. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.
61. Subramanian S. The effects of sample size on population genomic analyses — implications for the tests of neutrality. BMC Genom. 2016;17:123.
62. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, et al. LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012;40:11189–201.
63. Broad Institute. Best practices for variant calling with the GATK, https://www.broadinstitute.org/partnerships/education/broade/best-practices-variant-calling-gatk-1.
64. Huson DH. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics. 1998;14:68–73.
65. Tria FDK, Landan G, Dagan T. Phylogenetic rooting using minimal ancestor deviation. Nat Ecol Evol. 2017;1:0193.
66. Helmus MR, Bland TJ, Williams CK, Ives AR. Phylogenetic measures of biodiversity. Am Nat. 2007;169:E68–E83.
67. Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, et al. Picante: R tools for integrating phylogenies and ecology. Bioinformatics. 2010;26:1463–4.
68. Chen J GUniFrac: Generalized UniFrac distances. https://CRAN.R-project.org/package=GUniFrac.
69. Stoletzki N, Eyre-Walker A. The positive correlation between dN/dS and dS in mammals is due to runs of adjacent substitutions. Mol Biol Evol. 2011;28:1371–80.
70. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2017.
71. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47:D309–D314.
72. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
73. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74.