The environment drives microbial trait variability in aquatic habitats

Sara Beier, Anders F Andersson, Pierre Galand, Corentin Hochart, Jürg B Logue, Katherine Mcmahon, Stefan Bertilsson

Received: 6 January 2020 | Revised: 1 September 2020 | Accepted: 22 September 2020

DOI: 10.1111/mec.15656

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

The environment drives microbial trait variability in aquatic habitats

Sara Beier[1,2] | Anders F. Andersson[3] | Pierre E. Galand[4] | Corentin Hochart[4] | Jürg B. Logue[5] | Katherine McMahon[6] | Stefan Bertilsson[7]

[1]Biological Oceanography, Leibniz Institute for Baltic Sea Research Warnemünde (IOW), Rostock, Germany

[2]Laboratoire d'Océanographie Microbienne (LOMIC), Sorbonne Universités, CNRS, Observatoire Océanologique de Banyuls, Banyuls-sur-Mer, France

[3]Department of Gene Technology, School of Engineering Sciences in Chemistry, Biotechnology and Health, Science for Life Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden

[4]Laboratoire d'Ecogéochimie des Environnements Benthiques (LECOB), Sorbonne Universités, CNRS, Observatoire Océanologique de Banyuls, Banyuls-sur-Mer, France

[5]Department of Ecology and Genetics, Limnology, Uppsala University, Uppsala, Sweden

[6]Departments of Civil and Environmental Engineering, and Bacteriology, University of Wisconsin Madison, Madison, WI, USA

[7]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden

Correspondence

Sara Beier, Department of Biological Oceanography, Leibnitz Institute for Baltic Sea Research, Seestraße 15, 18119 Rostock, Germany.

Email: sara.beier@io-warnemuende.de

Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Number: BE5937/2-1; United States National Science Foundation, Grant/Award Number: MCB-0702395; Joint Genome Institute, Grant/Award Number: CSP1977; Agence Nationale de la Recherche, Grant/Award Number: ANR-14-CE02-0004-01; Long-Term Ecological Research Program, Grant/Award Number: NTL–LTER DEB-1440297; INSPIRE Award, Grant/Award Number: DEB-1344254; Swedish Research Council (VR)

[Correction added on 23-October-2020, after first online publication: Projekt Deal funding statement has been added.]

Abstract

A prerequisite to improve the predictability of microbial community dynamics is to understand the mechanisms of microbial assembly. To study factors that contribute to microbial community assembly, we examined the temporal dynamics of genes in five aquatic metagenome time-series, originating from marine offshore or coastal sites and one lake. With this trait-based approach we expected to find gene-specific patterns of temporal allele variability that depended on the seasonal metacommunity size of carrier-taxa and the variability of the milieu and the substrates to which the resulting proteins were exposed. In more detail, we hypothesized that a larger seasonal metacommunity size would result in increased temporal variability of functional units (i.e., gene alleles), as shown previously for taxonomic units. We further hypothesized that multicopy genes would feature higher temporal variability than single-copy genes, as gene multiplication can result from high variability in substrate quality and quantity. Finally, we hypothesized that direct exposure of proteins to the extracellular environment would result in increased temporal variability of the respective gene compared to intracellular proteins that are less exposed to environmental fluctuations. The first two hypotheses were confirmed in all data sets, while significant effects of the subcellular location of gene products was only seen in three of the five time-series. The gene with the highest allele variability throughout all data sets was an iron transporter, also representing a target for phage infection. Previous work has emphasized the role of phage–prokaryote interactions as a major driver of microbial diversity. Our finding therefore points to a potentially important role of iron transporter-mediated phage infections for the assembly and maintenance of diversity in aquatic prokaryotes.

KEYWORDS

community ecology, ecological genetics, marine environments, metagenomics, microbial ecology, time-series

1 | INTRODUCTION

For untangling the different factors that contribute to the temporal dynamics of communities, we first need to understand the mechanisms that underlie biological system variability. Identifying the factors that determine the assembly and maintenance of diversity in microbial communities over time is a central tenet in microbial ecology, and is a prerequisite to robustly predict the responses of communities to a dynamic environment (Nemergut et al., 2013).

Previous studies have demonstrated that temporal dynamics in the diversity of microorganisms is similar to that of macroorganisms (Oliver et al., 2012; White et al., 2006). Due to their huge population sizes, short generation times and relative ease of sampling, microorganisms can serve as valuable models to understand general mechanisms that drive temporal variability and changes in community diversity. For instance, it has been shown that microbial communities that switch from deterministic and niche-based, to more stochastic assembly mechanisms experience gradually higher rates of species turnover (Ayarza & Erijman, 2011; van der Gast et al., 2008). Stochastic-driven community assembly events have therefore been argued to be responsible for high species turnover rates in communities with large metacommunity sizes (Ayarza & Erijman, 2011). On the other hand, increased species turnover rates and temporal variability in species diversity and composition have also been linked to high levels of disturbance or environmental heterogeneity (Ager et al., 2010; Guo et al., 2019; Shade et al., 2013). These and other studies (Hatosy et al., 2013; Liang et al., 2015; Yan et al., 2012) have explored the temporal turnover or variability of taxonomic microbial units. However, while species are the natural evolutionary units that carry traits, they are selected for and assembled in a community based on their position in the niche space.

The importance of trait-based ecology for an improved understanding of the patterns and mechanisms that broadly determine the performance of complex communities has previously been highlighted (Escalas et al., 2019; McGill et al., 2006). Prokaryotes are furthermore characterized by large subspecies-level genomic plasticity due to genetic rearrangements that are often mediated by mobile genomic islands such as integrons, transposons, integrative and conjugative elements, and prophages (Bertelli et al., 2019). Such genetic rearrangements include horizontal gene transfer (HGT) via transformation, conjugation or transduction, gene loss or gene duplication, and blur the link between taxonomic identity and functional capabilities (Boon et al., 2014; Cordero & Polz, 2014). The phenomenon of prokaryotic subspecies-level genomic plasticity is addressed in the pan-genome concept, according to which at the species level 20%–35% of genes in a genome can be specific for a certain strain (Medini et al., 2005). One pertinent example are members of the aquatic SAR11 bacterial clade, which is among the most abundant organisms on Earth (Morris et al., 2002) and exhibits a large subspecies-level diversity (Ward et al., 2017). However, this subspecies-level diversity is not well resolved by the 16S rRNA gene commonly used as a phylogenetic marker: for example, genomes from the *Pelagibacter* strains HTCC1062 and HTCC1002 that were isolated from the same sample have been shown to differ by 31 genes inserted in the HVR3 region of HTCC1002, but only a single base in their 16S rRNA gene (Rappe et al., 2002; Wilhelm et al., 2007). Accordingly, the assessment of temporal dynamics in microbial communities based on phylogenetic marker genes may underestimate the true turnover of functional attributes in a community.

The importance of allelic variation of the same gene has been considered to be of inferior importance compared to the importance of differences in the gene content for determining physiological differences of prokaryotic cells among closely related lineages (Medini et al., 2005, 2008; Whitaker & Banfield, 2006). However, at the community level, allelic variations of the same gene orthologue arise often from (possibly distantly related) organisms that share the trait encoded by the shared gene orthologue, while otherwise occupying different niches in the environment. Allele diversity of gene orthologues in complex microbial communities can therefore be interpreted as a measure of their functional redundancy (Louca et al., 2018). Functional redundancy is considered to be a key parameter in community ecology, because it can buffer the functional performance of communities in a fluctuating environment (Jurburg & Salles, 2015).

In this study we address the temporal variability (betadiversity) of individual traits by assessing the variability of alleles representing individual gene orthologues in aquatic microbial communities. To avoid manual and possibly ambiguous classification of gene orthologue groups into complex traits, we used individual gene orthologues as a proxy for similar traits, as has been done previously, for example, to estimate functional diversity (Louca et al., 2018; Raes et al., 2011). We propose that this is biologically meaningful as the presence or absence of a gene in a particular genome should influence the performance of a measurable trait in the respective organism.

Increasing metacommunity size has previously been associated with higher taxon turnover because of an increasing importance of stochastic effects in the assembly of temporally linked communities

(Ayarza & Erijman, 2011). We therefore also hypothesized (i) that functional units would be more variable over time when they originate from a large seasonal metacommunity.

Beyond the impact of stochastic events on community assembly, previous work suggests that species sorting governed by prevailing environmental conditions is the most important factor for community assembly in aquatic bacterial communities (Langenheder & Lindström, 2019). We therefore propose that with increasing heterogeneity of the environmental parameters relevant for the functionality of a certain gene orthologue, the temporal allele variability of this specific gene will increase. For this, we consider environmental heterogeneity in its broadest sense, including the variability of exogenous parameters such as salinity or temperature, but also the variability created by biotic activities: these include for instance changing chlorophyll *a* concentration but also more elusive parameters such as the concentration of vitamins, certain metabolites or the presence of phages. To address the relationship between environmental heterogeneity and the variability of functional units, we hypothesized (ii) that genes that are present in multiple copies within a single genome will feature higher allele variability than single-copy genes. The rationale for this is that small differences between such copies within a single genome will enable the organism to perform well in a fluctuating environment as alleles/copies with different performance optima can be engaged. One example is the genes coding for chitinases, which can appear in up to 10 copies within individual genomes (Karlsson & Stenlid, 2009). The alleles representing these copies can differ in pH optima (Miyashita et al., 1991) or in substrate versatility with regard to the crystalline form of chitin that is processed, a feature that differs in chitin-containing organisms (Svitil et al., 1997). In agreement with the idea of better performance of multicopy genes in a changing environment, the presence of multiple copies should be particularly beneficial for proteins encountering fluctuations of parameters that are relevant for their functionality. We finally hypothesized (iii) that genes encoding products that are exposed to the extracellular milieu would feature higher temporal variability than gene products present in the cytoplasm, which are buffered against changing environmental conditions. Also here, chitinase genes can serve as an example, and it has indeed been suggested in a time-series study that compositional changes in chitinases, as typically multicopy extracellular genes, are highly dynamic over the annual cycle (Beier et al., 2012).

We have tested the above-mentioned hypotheses with data from five surface aquatic metagenomic time-series, assuming that allele variability increases with (a) larger seasonal metacommunity size, (b) higher gene copy number and (c) exposure to the extracellular environment. We estimated the temporal mean allele variability of genes that are not essential in prokaryotes and refer to them as auxiliary genes in the remainder of this paper. These genes reflect traits that are potentially, but not necessarily, present across prokaryotes in general. We additionally estimated the mean allele variability of a set of obligatory single-copy genes, considered as phylogenetic markers, in order to estimate and compare overall trait and species variability in each of the five data sets. The analysed data comprised surface ocean time-series from oligotrophic offshore sites from the Hawaii Ocean Time-series in the Pacific (HOT) and Bermuda Atlantic Time-series Study in the Atlantic (BATS) (Biller et al., 2018). We also included time-series from two marine sites closer to the coast, including the mesotrophic Linnaeus Microbial Observatory (LMO) in the Baltic Sea Proper (Alneberg et al., 2018; Hugerth et al., 2015) and the oligotrophic SOMLIT Observatory Laboratoire Arago (SOLA) station within the Mediterranean Sea (Galand et al., 2018), as well as a freshwater data set from the eutrophic Lake Mendota (MEN) (Linz et al., 2018).

2 | MATERIAL AND METHODS

2.1 | Data acquisition and bioinformatic processing

We downloaded and processed publicly available time-series metagenome data and the corresponding environmental data from two surface aquatic environments, including the Pacific (ALOHA) (Biller et al., 2018) and the Sargasso Sea (Biller et al., 2018). Three partly processed aquatic surface metagenome time-series and environmental data from Lake Mendota and the Northwest Mediterranean Sea and the Baltic Sea Proper (Hugerth et al., 2015) were provided from collaborators (Table S1).

Downstream analyses required large data sets and thus we included only metagenomes with a sequencing depth close to or exceeding 10 million reads. We analysed metagenomes from six to 10 sampling dates for each metagenome time-series, spanning over at least 6 months (Table S2). More details of the sampling procedure and the software used for bioinformatic processing of the metagenome data are summarized in Table S1. The code used in this study for data processing is available via https://www.protocols.io/view/allele-variability-bkhqkt5w.

In short, all raw sequence data were quality trimmed and co-assemblies were created for each time-series with the exception of the Baltic Sea metagenomes, where the publicly available BARM assembly (Alneberg et al., 2018) was used. CDS (coding DNA sequence) regions were identified by gene calling and subsequently clustered by 100% amino acid sequence similarity using the CD-HIT software (Li & Godzik, 2006). Functional annotation of the CDS regions was performed using the KEGG database (Kanehisa et al., 2007). We defined all gene variants coding for the same gene orthologue but differing in their amino acid sequence as alleles of this gene. We ignored DNA sequence differences causing synonymous replacements in the amino acid sequence. This is because we aimed in this study to explore the direct interaction between environmental dynamics and the assembly of alleles based on protein properties and therefore their amino acid but not their DNA sequences. We are aware of the possible inaccuracies introduced by the co-assembly of multiple time-series metagenome data sets: for instance, small-scale variations, such as single nucleotide polymorphisms (SNPs) causing nonsynonymous replacement that can discriminate between closely related alleles may be masked in the consensus sequences

(Dick, 2018). A reduced resolution of our assembly-based approach for the identification of closely related alleles will probably result in a general underestimation of the true temporal allele variability estimated as described below. However, we believe that such inaccuracy will be similar among the considered gene categories and should therefore not bias our downstream analyses in a certain direction.

The quality-trimmed data were mapped on the respective co-assemblies using BOWTIE2 set to very-sensitive-local (Langmead & Salzberg, 2012) and summarized using the FEATURECOUNTS software (Liao et al., 2014).

For downstream analyses, we focused on KEGG orthologues that were present in fewer than two-thirds of all prokaryotic genomes in the KEGG database (available at: https://www.kegg.jp/kegg/download/; release 78.1; downloaded: May 2016) containing data from 3,865 prokaryotic genomes), which we refer to as auxiliary genes. Our decision to exclude gene orthologues that are present in most of the prokaryotic genomes was motivated by the fact that a high proportion of these genes encode proteins involved in essential functions, such as replication, transcription or translation (Table S3). The evolution of these genes often correlates with the evolution of neutral markers (Medini et al., 2008), and they therefore represent taxonomic rather than functional markers. We are aware that among the genes here defined as auxiliary genes, some are still obligatory for certain phylogenetic lineages and their sequence diversification can correlate with the phylogenetic evolution of members of this specific lineage.

2.2 | Gene copy number and seasonal metacommunity size

For each gene orthologue identified via the KEGG ontology, the average per-cell gene copy number among carrier genomes was determined based on the average gene copy number in prokaryote organisms of all entries coding for the respective gene orthologue (KEGG; release 78.1). In the remainder of this paper we refer to the average copy number of a gene across all carrier genomes in the KEGG database simply as "gene copy number." While the presence of different communities at each of the examined sites will result in site-specific gene copy numbers, this information cannot be determined without access to the full sequence information for each genome present at this site. However, metagenome data do not allow us to reconstruct full genome data from all community members and we therefore use the full genome data of all prokaryotic genomes stored in the KEGG database to obtain a value for the gene copy number of individual genes that was then used to approximate the true values in the communities from all sample sites. We thereby assume that the ranking from single-copy genes to high-copy-number genes will be similar across different ecosystems.

The gene-specific seasonal metacommunity size was assessed by dividing the total number of alleles for the respective gene orthologue detected throughout each metagenome time-series divided by the gene copy number of this gene orthologue. To match data from the downstream allele variability analyses, we estimated the metacommunity size only after subsampling allele count-data for the respective genes to 500 counts per individual metagenome in each time-series. The resulting parameter estimates the seasonal metacommunity richness of carrier organisms for each gene orthologue and represents a parameter for temporal gamma diversity. The term metacommunity has been defined as local communities that are connected by dispersal of multiple potentially interacting species (Wilson, 1992), while for this study we use the term seasonal metacommunity to describe several communities from the same location but across different seasons. However, spatial and temporal community dynamics are linked because a recolonization of populations being temporally extinct from one site can only occur via the dispersal of these populations remaining at spatial refuges. Therefore, we expect some overlap between the theories concerning local metacommunities and seasonal metacommunities as defined here, as for instance exemplified in the similar properties of species–area relationships and species–time relationships (White et al., 2006).

It has been recently argued that metagenome-delineated measures of the richness of functional units, such as the seasonal metacommunity parameter introduced here, are ecologically meaningful measures to address functional redundancy (Louca et al., 2018).

2.3 | Subcellular location

The subcellular location of all entries in the KEGG database (release 78.1) was determined based on the presence of signal peptides in the amino acid sequences via the PSORTB (version 3.0) software (Yu et al., 2010). We considered all sequences affiliating according to the KEGG taxonomy with the phyla Firmicutes or Actinobacteria as originating from gram-positive bacteria (Gontang et al., 2007) and the remaining bacterial sequences as originating from gram-negative bacteria. We applied the settings for gram-negative bacteria, gram-positive bacteria or archaea for the KEGG entries affiliating with either of these groups, respectively. To test Hypothesis 3, gene orthologues were defined as cytoplasmatic if >80% of all entries were identified as coding for cytoplasmatic proteins. Gene orthologues were defined as noncytoplasmatic if >80% of all entries were identified as coding for either extracellular, cell-wall-bound (gram-positives, archaea) outer membrane-bound (gram-negatives) or periplasmatic proteins (gram-negatives). Gene-encoding proteins with a subcellular location predicted in the cytoplasmic membrane of gram negatives were ignored because their gene products can be in contact with both the cytoplasm and the outer environmental matrix. Of 8,127 auxiliary prokaryotic KEGG gene orthologues, 3,313 were identified as cytoplasmatic and 391 as noncytoplasmatic (Table S3).

2.4 | Temporal variability of gene-specific alleles

The temporal variability of the alleles for each individual auxiliary gene was estimated using the multivariate dispersion metric

(Anderson et al., 2006), which determines the average distance of individual samples within each time-series to the group centroid. We used this method analogously to earlier applications, with the only difference that instead of compositional data for taxonomic units (e.g., operational taxonomic units [OTUs], Grubisic et al., 2017), the input data consisted of compositional data of the alleles representing the respective auxiliary gene. We applied the Bray–Curtis index for estimating pairwise dissimilarity between individual samples and the sample centroid after subsampling for each auxiliary gene to 501 counts per sample and thus only considering genes exceeding 500 counts in each individual sample. Earlier estimates of sequencing depths necessary to estimate beta-diversity found that Bray–Curtis distance-based beta-diversity patterns are particularly insensitive to low sequencing depths as long as all compared samples are normalized to the same sequencing depths (i.e., 500 counts per sample were sufficient to capture roughly 90% of the total beta-diversity of microbial communities) (Lundin et al., 2012).

We additionally estimated the temporal variability of 27 obligatory single-copy genes (Raes et al., 2007) in each of the five time-series, as described above for auxiliary genes. These were used as a proxy for the taxonomic variability in community composition.

2.5 | Statistical analyses

To compare the overall temporal variability of both functional and taxonomic units across all five time-series, mean values and the standard variation of temporal variability for all auxiliary and taxonomic single marker genes (Table S4) within each time-series was estimated (Table 1). In this case the values from the different time-series were directly compared against each other. To improve compatibility across data sets, we subsampled all functionally annotated alleles from all individual samples to the minimal number of mapped and functionally annotated genes (1,740,778 in SRR2053279, Table S2) prior to the second subsampling step to 501 reads per gene per sample for computing allele variability data.

We applied analyses of covariance (ANCOVAs) to test the dependence of the temporal variability of auxiliary genes on their seasonal metacommunity size, gene copy number and the subcellular location of these genes defined as detailed above. In this case, analyses were performed for each time-series separately and in order to keep the maximal predictive power, no subsampling was carried out at the level of the total metagenome data across the time-series data. All genes exceeding 500 reads in individual samples of the time-series data, annotated as either cytoplasmatic or noncytoplasmatic (Table S3), were incorporated into the analyses. Residuals were inspected manually, and to improve their fit to a normal distribution the two continuous variables (seasonal metacommunity size, gene copy number) were log-transformed and subsequently mean-centred. Due to the highly unbalanced occurrence of cytoplasmatic versus noncytoplasmatic genes, significance was tested considering type III errors. To avoid biases due to a singular subsampling event to a reduced data set with only 501 counts per gene per sample, we estimated the *p*-values for our statistics using a bootstrapping approach with 1,000 permutations. An R script with the code to estimate gene allele variability and perform the described statistical tests is available via https://github.com/sarabeier/allele.variability. Gene orthologues that exceeded 500 counts in the individual samples for all time-series, and that were therefore included into the analyses, were ranked in each time-series by their allele variability. The variance of the gene variability rank was estimated in order to identify genes with similar or very distinct variability across all five time-series.

2.6 | Overall environmental heterogeneity

An index for environmental heterogeneity for each of the time-series was estimated by summing the coefficient of variance for temperature, salinity and chlorophyll *a* concentration (Table S2) as published elsewhere (Maurice et al., 2013). Environmental data for MEN were retrieved from the North Temperate Lakes Long Term Ecological Research site, which also documents detailed experimental methods (https://lter.limnology.wisc.edu). Salinity data for Lake MEN were not available, but we assumed the coefficient of salinity variance to be zero in this freshwater system without connection to marine waters. Environmental data for BATS and HOT were available via the homepage of the Bermuda Atlantic Time-Series Study (http://bats.bios.edu/) and via the Hawaii Ocean Time-Series data provided

TABLE 1 Overview of gene variability in five metagenome time series and environmental heterogeneity

Data set	Auxiliary genes					Core genes					Environment
	Min.	Max.	Mean	*SD*	*n*	Min.	Max.	Mean	*SD*	*n*	Heterogeneity index
SOLA	0.40	0.63	0.52	0.04	235	0.49	0.57	0.53	0.02	14	0.95
BATS	0.23	0.60	0.47	0.04	232	0.45	0.55	0.49	0.04	15	0.85
HOT	0.12	0.60	0.40	0.07	272	0.32	0.53	0.40	0.07	18	0.42
LMO	0.39	0.57	0.48	0.03	156	0.45	0.50	0.48	0.02	9	0.99
MEN	0.40	0.60	0.49	0.03	188	0.41	0.57	0.47	0.05	17	1.27

Note: Values are based on functionally annotated metagenome count data after subsampling to the minimal number of counts (*n*: number of considered genes).

online (http://hahana.soest.hawaii.edu/hot/hot-dogs/interface.html), respectively. Environmental data for SOLA and LMO were used as published earlier (Galand et al., 2018; Hugerth et al., 2015).

3 | RESULTS

Mean values for auxiliary gene variability ranged from 0.40 in the HOT time-series to 0.52 in the SOLA time-series. Mean allele variability of phylogenetic marker genes used as a proxy for the variability in taxonomic community composition exhibited similar values, ranging from 0.40 to 0.53 (Table 1). The ranking of mean allele variability in the five time-series was nearly identical for auxiliary genes and phylogenetic marker genes (Table 1). An index of environmental heterogeneity based on the temporal variability of temperature, salinity and chlorophyll *a* concentration in the five time-series (Table 1) indicated a higher positive correlation to the mean allele variability of auxiliary compared to core genes (Pearson's $r = .80$ and $r = .63$, respectively). However, due to the low number of data points ($n = 5$), none of these correlations was significant, apart from high correlation coefficients ($p = .10$ and $p = .26$, respectively).

In all five time-series, we identified a gene orthologue encoding an outer membrane iron receptor protein (KEGG ID: K02014) as the temporally most variable auxiliary gene (Figure 1; Table S5). The allele variability of other genes was less consistent across the five data sets. For instance, the alleles of restrictios enzymes encoded by the genes *hsdR* and *hasdM* (K01153 and K03427) were highly variable in SOLA, LMO and MEN, but belonged to genes with low variability in the offshore sites BATS and HOT (Table S5).

Gene variability assessed via the multivariate dispersion metric (Anderson et al., 2006) based on the Bray–Curtis dissimilarity index will, analogously to a simple pair-wise comparison, be zero if the composition of subunits (in our case the alleles of a gene) in all compared samples is identical. This is possible also for indefinitely large numbers of alleles contributing to the gene-specific allele composition in each individual sample within the time-series. In contrast to the pair-wise comparison, Bray–Curtis-based dispersion cannot reach 1, but will still approach unity if none of the included samples shares any alleles. Due to the fact that minimal temporal allele variability is possible with large allele richness and vice versa, there is no mathematical constraint that could cause a positive relationship between large allele richness and temporal variability.

The temporal variability of auxiliary genes exhibited a strong and highly significant positive correlation to seasonal metacommunity size ($p < .001$, $F > 125$; Table 2; Figure 1), which supports Hypothesis 1. The correlation with gene copy number was weaker, but still highly significant and positive in all five time-series ($p < .001$, $F > 61$; Table 2; Figure 1). Accordingly, our analyses also support Hypothesis 2.

Our results demonstrated that in the time-series HOT, LMO and MEN, the variability of genes coding for cytoplasmatic enzymes was lower than that of noncytoplasmatic ones ($p < .05$, $F > 5$; Table 2; Figure 1). However, the impact of the subcellular location of the gene product on their temporal variability could not be verified at a *p*-value cutoff $p < .05$ in the oligotrophic SOLA and BATS time-series, although in both cases a trend for higher allele variability of noncytoplasmatic genes was detected (Table 2; Figure 1). Hypothesis 3 was accordingly only partially supported by our analyses.

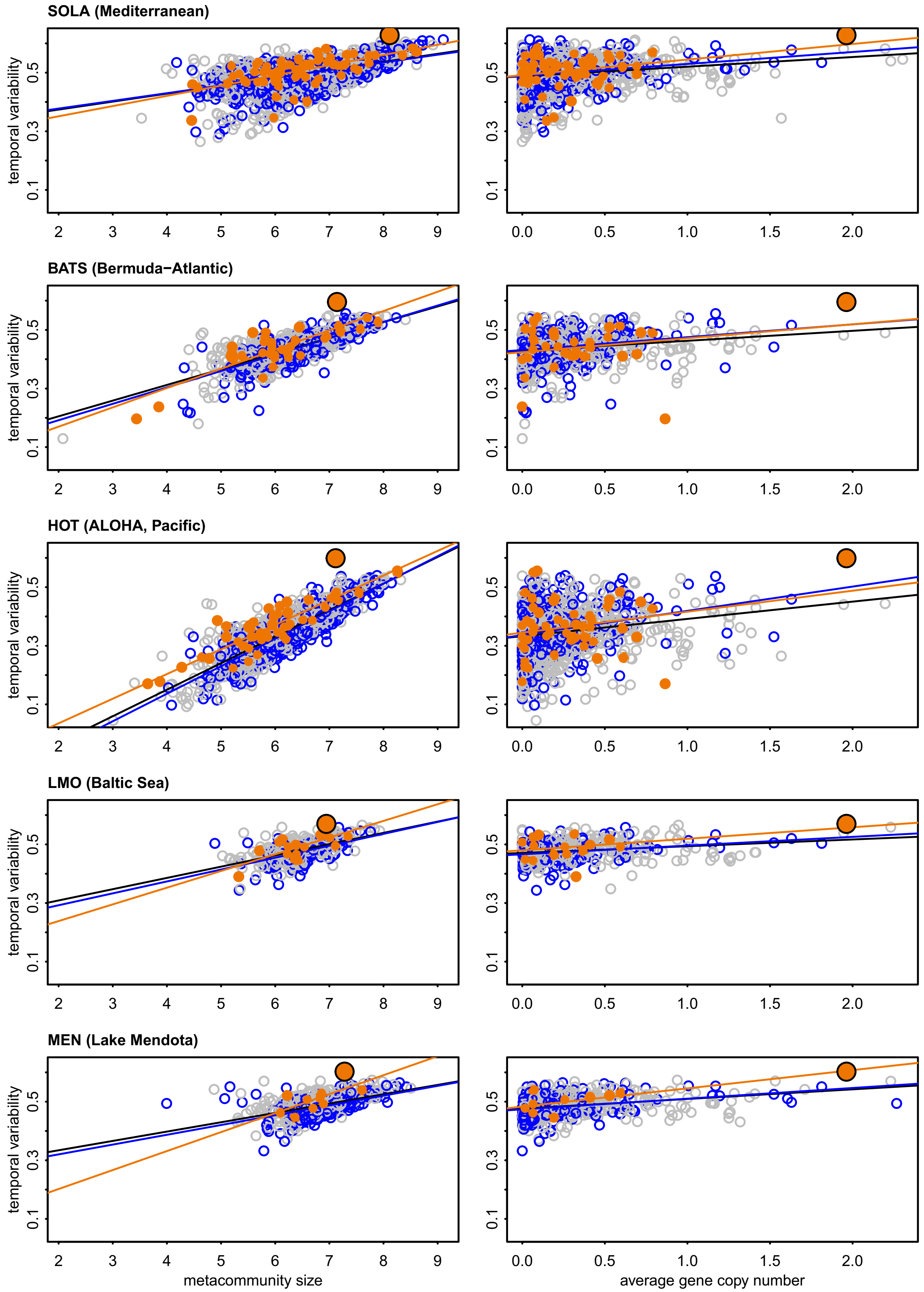
The interaction term between the parameters was not part of our hypothesis and we therefore focus on reporting the main effects of the tested variables. However, significant interaction terms among continuous variables can in extreme cases indicate that the direction of the main effect is only valid within a certain data range (Frost, 2019). Because we detected significant interaction terms in some cases (Table S6), we created interaction plots for both continuous variables with the other continuous variable set to the extremes of the data range (Frost, 2019). We thereby verified that the significant correlations reported in the main effect for each of the continuous variables were positive under all conditions for the respective other continuous variable.

4 | DISCUSSION

To our knowledge this is the first study that systematically compares multiple microbial metagenome time-series by assessing allele variability as a beta diversity measure to address the dynamics in the assembly of traits. Our data thus demonstrate that the assembly of functional units such as alleles follows similar patterns to that of taxonomic units by demonstrating that seasonal metacommunity size and environmental heterogeneity influence the temporal variability of functional units. This is in agreement with earlier taxon-based observations where species–time relationship analyses (STRs) have shown that these parameters influence variability in species composition (Ayarza & Erijman, 2011; Oliver et al., 2012; Shade et al., 2013). Although we did not analyse temporal variability via STRs as done in the above-mentioned studies, previous work indicated that steep STR slopes correlated with other measures of temporal variability in community composition (Shade et al., 2013).

Because we used metagenome time-series data from different laboratories for our analyses, there were some differences in the way samples were collected and sequenced (e.g., size fraction, time

FIGURE 1 Correlations of auxiliary gene variability with seasonal metacommunity size (log-transformed) or gene copy number (log-transformed) in the five metagenome time-series. Blue open data points and the blue trend line indicate cytoplasmatic genes, orange filled data points and the orange trend line indicate noncytoplasmatic genes, and light grey open data points indicate genes for which subcellular location could not be determined following our working definition. The black trend line represents all data points independent of their subcellular location. The highlighted enlarged data point indicates a gene orthologue encoding an outer membrane iron receptor protein (K02014), which in all time-series exhibited maximal temporal variability. All plotted data points present mean values obtained after 1,000 permutations repeating the subsampling processes


SOLA (Mediterranean)
BATS (Bermuda−Atlantic)
HOT (ALOHA, Pacific)
LMO (Baltic Sea)
MEN (Lake Mendota)
temporal variability
0.1
0.3
0.5
2
3
4
5
6
7
8
9
0.0
0.5
1.0
1.5
2.0
metacommunity size
average gene copy number

range covered, sequencing depth and data processing; Tables S1 and S2). This was of lesser importance as our primary focus was on testing our hypothesis by comparing variability patterns for genes only within each of the five time-series. Also, differences in the time range covered by the individual time-series did not seem to influence the observed patterns to any major extent. This is possibly because the largest temporal variability in microbial community composition in marine samples was shown to occur across intra- and interseasonal time frames, which were covered in all five time-series, while less variability was related to interannual timescales (Hatosy et al., 2013).

Among the variables that were tested for their impact on the temporal variability of functional units, seasonal metacommunity size clearly had the strongest effect (Table 2; Figure 1). It has been argued that a positive relationship between metacommunity size and variability in community composition is due to an increased importance of stochastic effects in the assembly of temporally consecutive communities (Ayarza & Erijman, 2011). However, in that study the authors compared the temporal variability in community composition for communities grown in bioreactors after manipulating the richness of the starting community and refer to this as metacommunity size. This is a different experimental design from our study, which is instead based on time-series data from natural aquatic environments where dispersal can bring in new species. Consequently, new alleles and their carrier species can emerge in samples that were not present at the start of the time-series. Hence in our study, seasonal metacommunity size was defined as the richness of carrier species of the investigated genes measured across all sample dates within each time-series. Possibly, the increasing importance of stochastic effects with increasing seasonal metacommunity size could have contributed to the positive relationship between seasonal metacommunity size and temporal variability observed across our five metagenome time-series. However, another mechanism may in our case further strengthen this relationship, because our analyses are based on data from natural environments where the functional attributes should reflect environmental heterogeneity. A larger seasonal metacommunity size for a specific gene could essentially be caused by higher diversification rates for this gene as a consequence of naturally occurring high temporal heterogeneity of parameters that impact functioning of this particular gene (McArthur et al., 1988). This can in turn increase the temporal variability of alleles representing this gene (Oliver et al., 2012).

TABLE 2 ANCOVA results

Data set	Tested variable	*F* value	*p* value
SOLA	Subcellular location	2	.154
	Metacommunity size***	453	<.001
	Gene copy number***	80	<.001
BATS	Subcellular location	3	.061
	Metacommunity size***	1,041	<.001
	Gene copy number***	139	<.001
HOT	Subcellular location***	55	<.001
	Metacommunity size***	3,152	<.001
	Gene copy number***	335	<.001
LMO	Subcellular location*	6	.026
	Metacommunity size***	119	<.001
	Gene copy number***	45	<.001
MEN	Subcellular location*	6	.017
	Metacommunity size***	338	<.001
	Gene copy number***	242	<.001

Note: Significance levels: ****p* < .001, ***p* < .01, **p* < .05.

We tested more specifically the impact of environmental heterogeneity on environmental variability of functional units by focusing on gene copy number and the subcellular location of gene products. As outlined in the Introduction, this is because we assumed that these variables are linked to the environmental heterogeneity to which a given gene responds in a selective process.

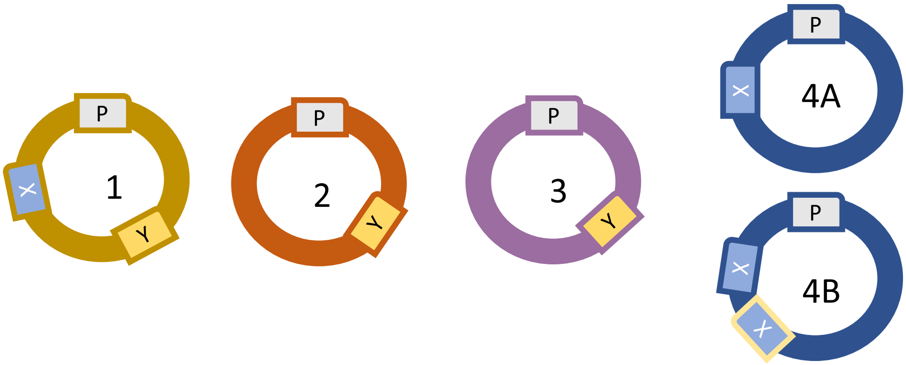


FIGURE 2 Schematic figure illustrating mechanisms that can lead to elevated allele variability of genes. Genomes 1–4 contain the phylogenetic marker gene P and two accessory genes X and Y. Genome 4 splits into the subgenomes 4A and 4B that share large parts of their genome but differ in gene copy number of gene X, for example due to the horizontal acquirement of the second gene copy of X in 4B. Temporal allele variability of accessory genes can reach high values if the population of carrier species for this specific gene is characterized by large temporal fluctuations in its composition or if genome rearrangement events lead to temporally alternating subspecies as exemplified in genome 4A/B

In agreement with our assumptions, data from all time-series indicated that temporal variability of alleles increased with increasing average copy number of the respective gene (Figure 1; Table 2). There are, broadly speaking, two different molecular mechanisms that may lead to increased temporal variability of alleles from multicopy genes: first as compared to single-copy genes they may be exposed to higher frequencies of genomic rearrangements, such as gene duplication, gene loss or gene gain via HGT. In such scenarios, temporally alternating subspecies whose genomes differ regarding some horizontally acquired genes but not in the remaining genome would lead to a higher allele variability of the horizontally acquired genes compared to the vertically inherited genes (Figure 2). The temporally alternating occurrence of close relatives indeed seems to be a common scenario in aquatic habitats (Cordero & Polz, 2014; Ward et al., 2017). Even though single-copy genes in a genome may also have been acquired from HGT, the appearance of multicopy genes must have been the consequence of a preceding HGT or gene duplication event. Furthermore, consecutive genomic rearrangements of genes that are already present in multiple copies within a genome, for example gene loss, are probably less deleterious than for single-copy genes. In agreement with these considerations, rearrangement events have indeed been reported to be more frequent for multicopy than for single-copy genes (Lerat et al., 2003) and this may be one molecular mechanism explaining the positive correlation of allele variability and gene copy number. A second possible mechanism that could underpin the observed positive relationships between gene copy number and temporal allele variability is that multicopy genes, independently from genome rearrangement events, may appear disproportionally often in taxa whose abundance typically fluctuates strongly over time. Multicopy genes are not equally distributed among different microbial lineages in marine environments, where taxa with oligotroph life strategies tend to feature genome streamlining (Lauro et al., 2009) and consequently a reduced occurrence of nonessential genes, such as multicopy genes as compared to copiotrophic community representatives. Indeed, in marine systems, populations adhering to the latter ecological strategy seem often to appear episodically with strongly fluctuating abundances (Giovannoni et al., 2014; Vergin et al., 2013).

As with multicopy genes, the literature proposes that cell surface proteins are disproportionally often exposed to gene rearrangement events (Nakamura et al., 2004). This has particularly been highlighted for members of the abundant SAR11 bacterial clade where the hypervariable HVR2 region mainly encodes genes that determine cell surface properties (Wilhelm et al., 2007). Therefore, elevated genetic rearrangement events at the subspecies level of noncytoplasmatic genes, which allow species featuring multiple ecotypes to deal with different or fluctuating environmental conditions, probably contributed to the increased temporal variability of these genes detected in three of the five time-series investigated. On the other hand, we believe that a pronounced accumulation of noncytoplasmatic proteins specifically in episodically appearing taxa to be less likely as this was the case for multicopy genes: all microorganisms need proteins that are in contact with the extracellular matrix in order to interact with their environment and this should be largely independent of their life history strategy.

Remarkably, while the subcellular location of gene products significantly influenced the temporal variability of alleles in only one of the three time-series from oligotrophic marine sites (HOT), a significant correlation was seen for both the LMO and MEN time-series, both of which are classified as productive. High productivity is generally coupled with more pronounced algal blooms and consequently larger fluctuations in an array of environmental variables likely to influence microbial populations. This includes the availability of labile dissolved organic matter (Bertilsson & Jones, 2003) or pH, particularly in less buffered freshwater systems (Verduin, 1956). It has furthermore been shown that heterotrophic bacterial community composition in aquatic habitats is strongly structured by the appearance of algal blooms (Eiler & Bertilsson, 2004; Kent et al., 2007). Possibly, these environmental conditions are of particular relevance for the allele variability of proteins that are in contact with the outer milieu. Yet, in contrast to the BATS and SOLA, the oligotroph HOT time-series also featured significantly higher allele variability for noncytoplasmatic as compared to cytoplasmatic genes. Strikingly, the variability of temperature was very low in the BATS time-series (*SD*: 1.14; Table S2) compared to the HOT and SOLA time-series (*SD*: 3.15 and 3.66, respectively; Table S2). We speculate that in contrast to HOT, higher temperature variability in SOLA and BATS may have masked the effect of low chlorophyll variability (Biller et al., 2018) and all linked parameters in these three oligotrophic sites. Temperature variations should, in contrast to variation in other environmental parameters, not affect allele variability in contrasting subcellular locations because prokaryotes do not regulate their intracellular temperature. Consequently, temperature changes would affect inner and outer cellular proteins equally. The crucial role of temperature in structuring marine bacterioplankton has been highlighted in earlier metagenome-based studies (Raes et al., 2011; Sunagawa et al., 2015) and by noting that temperature is an important parameter driving the abundance of different SAR11 ecotypes (Brown et al., 2012).

The indicated stronger correlation of average auxiliary gene allele variability with that of the average allele variability of phylogenetic marker genes (Table 1) confirms a better resolution of functional compared to taxonomic information for matching the metagenome with the corresponding environmental data, as suggested recently (Caputi et al., 2019). Note, however, that the heterogeneity index is only a rough estimator of true environmental heterogeneity: on the one hand, chlorophyll *a* concentration of Lake Mendota was only available with several days of mismatch (Table S2). Furthermore, only temperature, salinity and chlorophyll *a* were considered in this index, while marine microbial communities are exposed to a plethora of other environmental variables. For instance, parameters such as vitamin availability that often arise from biotic activity and indicate biological interactions may be essential for community assembly (Herren & McMahon, 2017) and accordingly also for temporal allele variability. The indicated correlation between the heterogeneity index and average allele variability therefore suggests that in our study temperature, salinity and chlorophyll *a* concentrations may have influenced the variability of other parameters not considered that have an impact on allele variability.

Interestingly, within all time-series, a gene encoding multicopy outer membrane iron receptor proteins (K02014, including FhuA) was detected as the gene with highest temporal allele variability (Figure 1; Table S5). Iron is a cofactor in many central reactions of organisms, such as respiration and photosynthesis, and is therefore essential for the viability of all cells (Holm et al., 1996). Proteins summarized as the KEGG orthologue K02014 interact with siderophore-bound iron, and the variability of siderophore types excreted by different siderophore producers may partly explain the high allele variability of K02014. However, besides their function as iron transporters, these proteins also serve as targets for phage infection. This was highlighted by the "Ferrojan Horse Hypothesis," according to which phages incorporate iron atoms into their tail to be recognized and enter host cells via K02014 transporters (Bonnain et al., 2016). A recent metagenome-enabled study demonstrated that iron-binding motives were present in 87% of unigenes encoding marine viral tail proteins. Furthermore, the corresponding viral contigs were distributed ubiquitously and with high abundance in marine sites worldwide sampled during the Tara Ocean Expedition (Caputi et al., 2019). These findings indicate that "ferrojan horse" phages may play an important role in the ecology of marine systems. It has been proposed previously that the interaction of prokaryotes with phages is a major mechanism for maintaining prokaryote diversity, particularly due to the variability of genes that interact with phage infections, such as K02014 (Cordero & Polz, 2014; Rodriguez-Valera et al., 2009). The large allele variability of K02014 across all five time-series is thus a further indication for a possibly high relevance of predation pressure from "ferrojan horse" phage infections in aquatic environments for the community assembly. Similar to K02014, *hsdR* and *hsdM* (K01153 and K03427) that encode restriction enzymes interacting with phages, exhibited very high allele variability in LMO, SOLA and MEN. However, these genes did not display such a pattern in the offshore sites BATS and HOT (Table S5). Possibly, infections by phages that are targeted by those specific restriction enzymes are of less relevance for community assembly in aquatic offshore sites compared to the sites closer to terrestrial habitats.

In summary, our study has revealed that the environment differentially impacts the temporal variability of functional marker gene diversity and that the response depends on seasonal metacommunity size, gene copy number and enzyme exposure to the extracellular environment. In particular, the two latter parameters appear to be linked with the temporal heterogeneity of environmental factors relevant for the functioning of the respective gene. As such, our trait-based analyses support findings from previous taxonomic-based studies concerning the impact of metacommunity size (Ayarza & Erijman, 2011; van der Gast et al., 2008) and environmental heterogeneity on the temporal variability within microbial communities (Ager et al., 2010; Guo et al., 2019; Shade et al., 2013). However, in contrast to earlier 16S rRNA gene-based taxonomic-centred studies comparing time-series data across different environments, we could confirm our findings repeatedly within individual time-series by focusing on individual traits with different characteristics. Our results further emphasize the importance of genomic rearrangements for functional adaptions to changing environments, specifically for genes encoding proteins with contact to the outer milieu. Overall, our findings indicate auxiliary multicopy genes encoding proteins with contact to the outer milieu as being highly sensitive marker genes to monitor and track environmental change in aquatic environments. Modelling of ecosystem characteristics using metagenome input data, for instance via machine learning techniques, provides promising approaches, but can suffer from overfitting due to the extensive number of potential predictor variables, particularly if only a few metagenomes are available (Angermueller et al., 2016). Focusing on a subset of relevant genes, such as those identified here empirically as highly sensitive marker genes, may improve the predictive power of such modelling approaches. A remarkable consistency of temporal patterns, such as the ranking of the tested variables with regard to their impact on allele variability, or the detection of an iron receptor and phage target protein as a maximally variable gene across all five time-series, points to relevance of trait-based analyses in predicting the assembly of complex microbial communities in a broad range of environments. We further want to corroborate previous suggestions (Boon et al., 2014; Escalas et al., 2019; Raes et al., 2011) that incorporating allele alpha and gamma diversity (e.g., allele metacommunity size) or also beta diversity patterns (e.g., allele variability) additional to gene abundance patterns into metagenome-based analyses might be a promising procedure to improve the fit of metagenome information to environmental data. Diversity measures seem to be particularly relevant for addressing the link between an ecosystem's heterogeneity and functional redundancy as well as assembly dynamics of inherent microbial communities.

ACKNOWLEDGMENTS

We thank Sarah Stevens for help using the Mendota time-series metagenome data and Christina Herren and Pavla Debeljak for commenting on an earlier version of the manuscript and giving advice concerning the statistical setup of our analyses or the ecology of prokaryote iron uptake. We also thank Moritz Buck for bioinformatic support with the PSORTB software. The PSORTB computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project snic2020-5-19 and functional annotations via diamond blast were supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, 031A532B). The work was supported by a grant from the German Science Foundation (DFG) awarded to S. Beier (BE 5937/2-1) and grants from the Swedish Research Council (VR) to S. Bertilsson. This research was also supported by the U.S. Department of Energy Joint Genome Institute through the Community Sequencing Program

(CSP 1977), which generated the Lake Mendota time-series. K.D.M. received funding from the United States National Science Foundation Microbial Observatories Program (MCB-0702395), the Long-Term Ecological Research Program (NTL–LTER DEB-1440297) and an INSPIRE Award (DEB-1344254). P.G. was supported by the Agence Nationale de la Recherche (ANR) through the project EUREKA (ANR-14-CE02-0004-01). Open access funding enabled and organized by Projekt DEAL.

AUTHOR CONTRIBUTIONS

S. Beier designed the study with support from S. Bertilsson and J.B.L. Data evaluation and statistical analyses were mainly performed by S. Beier partly supported by K.Mc.M., P.G., C.H. and A.F.A. All authors commented on and approved the final manuscript.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the European Nucleotide Archive (https://www.ebi.ac.uk/ena), NCBI (https://www.ncbi.nlm.nih.gov/bioproject/) and JGI/IMG (https://img.jgi.doe.gov/) with the following reference numbers (ENA Study PRJEB26919, NCBI BioProject PRJNA385855, NCBI BioProject PRJNA385855, NCBI BioProject PRJNA273799, JGI IMG/M study ID: Gs0063444).

ORCID

Sara Beier https://orcid.org/0000-0003-3707-4487
Anders F. Andersson https://orcid.org/0000-0002-3627-6899
Pierre E. Galand https://orcid.org/0000-0002-2238-3247
Corentin Hochart https://orcid.org/0000-0002-8508-7912
Jürg B. Logue https://orcid.org/0000-0001-8838-0914
Katherine McMahon https://orcid.org/0000-0002-7038-026X
Stefan Bertilsson https://orcid.org/0000-0002-4265-1835

REFERENCES

Ager, D., Evans, S., Li, H., Lilley, A. K., & van der Gast, C. J. (2010). Anthropogenic disturbance affects the structure of bacterial communities. *Environmental Microbiology*, *12*(3), 670–678. https://doi.org/10.1111/j.1462-2920.2009.02107.x

Alneberg, J., Sundh, J., Bennke, C., Beier, S., Lundin, D., Hugerth, L. W., Pinhassi, J., Kisand, V., Riemann, L., Jürgens, K., Labrenz, M., & Andersson, A. F. (2018). BARM and BalticMicrobeDB, a reference metagenome and interface to meta-omic data for the Baltic Sea. *Scientific Data*, *5*, 180146. https://doi.org/10.1038/sdata.2018.146

Anderson, M. J., Ellingsen, K. E., & McArdle, B. H. (2006). Multivariate dispersion as a measure of beta diversity. *Ecology Letters*, *9*(6), 683–693. https://doi.org/10.1111/j.1461-0248.2006.00926.x

Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, *12*(7), 878. https://doi.org/10.15252/msb.20156651

Ayarza, J. M., & Erijman, L. (2011). Balance of neutral and deterministic components in the dynamics of activated sludge floc assembly. *Microbial Ecology*, *61*(3), 486–495. https://doi.org/10.1007/s00248-010-9762-y

Beier, S., Mohit, V., Ettema, T. J. G., Ostman, O., Tranvik, L. J., & Bertilsson, S. (2012). Pronounced seasonal dynamics of freshwater chitinase genes and chitin processing. *Environmental Microbiology*, *14*(9), 2467–2479. https://doi.org/10.1111/j.1462-2920.2012.02764.x

Bertelli, C., Tilley, K. E., & Brinkman, F. S. L. (2019). Microbial genomic island discovery, visualization and analysis. *Briefings in Bioinformatics*, *20*(5), 1685–1698. https://doi.org/10.1093/bib/bby042

Bertilsson, S., & Jones, J. B. (2003). Supply of dissolved organic matter to aquatic ecosystems: Autochthonous sources. *Aquatic Ecosystems*, 3–24. https://doi.org/10.1016/B978-012256371-3/50002-0

Biller, S. J., Berube, P. M., Dooley, K., Williams, M., Satinsky, B. M., Hackl, T., Hogle, S. L., Coe, A., Bergauer, K., Bouman, H. A., Browning, T. J., De Corte, D., Hassler, C., Hulston, D., Jacquot, J. E., Maas, E. W., Reinthaler, T., Sintes, E., Yokokawa, T., & Chisholm, S. W. (2018). Marine microbial metagenomes sampled across space and time. *Scientific Data*, *5*, 180176. https://doi.org/10.1038/sdata.2018.176

Bonnain, C., Breitbart, M., & Buck, K. N. (2016). The Ferrojan Horse Hypothesis: Iron-Virus Interactions in the Ocean. *Frontiers in Marine Science*, *3*, UNSP 82. https://doi.org/10.3389/fmars.2016.00082

Boon, E., Meehan, C. J., Whidden, C., Wong, D.-H.-J., Langille, M. G. I., & Beiko, R. G. (2014). Interactions in the microbiome: Communities of organisms and communities of genes. *FEMS Microbiology Reviews*, *38*(1), 90–118. https://doi.org/10.1111/1574-6976.12035

Brown, M. V., Lauro, F. M., DeMaere, M. Z., Muir, L., Wilkins, D., Thomas, T., Riddle, M. J., Fuhrman, J. A., Andrews-Pfannkoch, C., Hoffman, J. M., McQuaid, J. B., Allen, A., Rintoul, S. R., & Cavicchioli, R. (2012). Global biogeography of SAR11 marine bacteria. *Molecular Systems Biology*, *8*, 595. https://doi.org/10.1038/msb.2012.28

Caputi, L., Carradec, Q., Eveillard, D., Kirilovsky, A., Pelletier, E., Pierella Karlusich, J. J., Rocha Jimenez Vieira, F., Villar, E., Chaffron, S., Malviya, S., Scalco, E., Acinas, S. G., Alberti, A., Aury, J.-M., Benoiston, A.-S., Bertrand, A., Biard, T., Bittner, L., Boccara, M., … Wincker, P. (2019). Community-level responses to iron availability in open ocean plankton ecosystems. *Global Biogeochemical Cycles*, *33*(3), 391–419. https://doi.org/10.1029/2018GB006022

Cordero, O. X., & Polz, M. F. (2014). Explaining microbial genomic diversity in light of evolutionary ecology. *Nature Reviews Microbiology*, *12*(4), 263–273. https://doi.org/10.1038/nrmicro3218

Dick, G. (2018). *Genomic approaches in earth and environmental sciences*. John Wiley & Sons.

Eiler, A., & Bertilsson, S. (2004). Composition of freshwater bacterial communities associated with cyanobacterial blooms in four Swedish lakes. *Environmental Microbiology*, *6*(12), 1228–1243. https://doi.org/10.1111/j.1462-2920.2004.00657.x

Escalas, A., Hale, L., Voordeckers, J. W., Yang, Y., Firestone, M. K., Alvarez-Cohen, L., & Zhou, J. (2019). Microbial functional diversity: From concepts to applications. *Ecology and Evolution*, *9*(20), 12000–12016. https://doi.org/10.1002/ece3.5670

Frost, J. (2019). Regression Analysis: An Intuitive Guide for Using and Interpreting Linear Models. Retrieved August 6, 2019, from Statistics By Jim website https://statisticsbyjim.com/regression/regression-analysis-intuitive-guide/

Galand, P. E., Pereira, O., Hochart, C., Auguet, J. C., & Debroas, D. (2018). A strong link between marine microbial community composition and function challenges the idea of functional redundancy. *The ISME Journal*, *12*(10), 2470–2478. https://doi.org/10.1038/s41396-018-0158-1

Giovannoni, S. J., Thrash, J. C., & Temperton, B. (2014). Implications of streamlining theory for microbial ecology. *The ISME Journal*, *8*(8), 1553–1565. https://doi.org/10.1038/ismej.2014.60

Gontang, E. A., Fenical, W., & Jensen, P. R. (2007). Phylogenetic diversity of gram-positive bacteria cultured from marine sediments. *Applied and Environmental Microbiology*, *73*(10), 3272–3282. https://doi.org/10.1128/AEM.02811-06

Grubisic, L. M., Bertilsson, S., Eiler, A., Heinrich, F., Brutemark, A., Alonso-Sáez, L., Andersson, A. F., Gantner, S., Riemann, L., & Beier, S. (2017). Lake bacterioplankton dynamics over diurnal timescales. *Freshwater Biology*, *62*(1), 191–204. https://doi.org/10.1111/fwb.12861

Guo, X., Zhou, X., Hale, L., Yuan, M., Ning, D., Feng, J., Shi, Z., Li, Z., Feng, B., Gao, Q., Wu, L., Shi, W., Zhou, A., Fu, Y., Wu, L., He, Z., Van Nostrand, J. D., Qiu, G., Liu, X., … Zhou, J. (2019). Climate warming accelerates temporal scaling of grassland soil microbial biodiversity. *Nature Ecology & Evolution*, *3*(4), 612–619. https://doi.org/10.1038/s41559-019-0848-8

Hatosy, S. M., Martiny, J. B. H., Sachdeva, R., Steele, J., Fuhrman, J. A., & Martiny, A. C. (2013). Beta diversity of marine bacteria depends on temporal scale. *Ecology*, *94*(9), 1898–1904. https://doi.org/10.1890/12-2125.1

Herren, C. M., & McMahon, K. D. (2017). Cohesion: A method for quantifying the connectivity of microbial communities. *The ISME Journal*, *11*(11), 2426–2438. https://doi.org/10.1038/ismej.2017.91

Holm, R. H., Kennepohl, P., & Solomon, E. I. (1996). Structural and functional aspects of metal sites in biology. *Chemical Reviews*, *96*(7), 2239–2314. https://doi.org/10.1021/cr9500390

Hugerth, L. W., Larsson, J., Alneberg, J., Lindh, M. V., Legrand, C., Pinhassi, J., & Andersson, A. F. (2015). Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biology*, *16*, 279. https://doi.org/10.1186/s13059-015-0834-7

Jurburg, S. D., & Salles, J. F. (2015). Functional Redundancy and Ecosystem Function — The Soil Microbiota as a Case Study. In Y. H. Lo, J. A. Blanco, & S. Roy (Eds.), *Biodiversity in Ecosystems - Linking Structure and Function*. Intech Europe.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., & Yamanishi, Y. (2007). KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, *36*(Database), D480–D484. https://doi.org/10.1093/nar/gkm882

Karlsson, M., & Stenlid, J. (2009). Evolution of family 18 glycoside hydrolases: Diversity, domain structures and phylogenetic relationships. *Journal of Molecular Microbiology and Biotechnology*, *16*(3–4), 208–223. https://doi.org/10.1159/000151220

Kent, A. D., Yannarell, A. C., Rusak, J. A., Triplett, E. W., & McMahon, K. D. (2007). Synchrony in aquatic microbial community dynamics. *The ISME Journal*, *1*(1), 38–47. https://doi.org/10.1038/ismej.2007.6

Langenheder, S., & Lindström, E. S. (2019). Factors influencing aquatic and terrestrial bacterial community assembly. *Environmental Microbiology Reports*, *11*(3), 306–315. https://doi.org/10.1111/1758-2229.12731

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357-U54. https://doi.org/10.1038/NMETH.1923

Lauro, F. M., McDougald, D., Thomas, T., Williams, T. J., Egan, S., Rice, S., DeMaere, M. Z., Ting, L., Ertan, H., Johnson, J., Ferriera, S., Lapidus, A., Anderson, I., Kyrpides, N., Munk, A. C., Detter, C., Han, C. S., Brown, M. V., Robb, F. T., … Cavicchioli, R. (2009). The genomic basis of trophic strategy in marine bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(37), 15527–15533. https://doi.org/10.1073/pnas.0903507106

Lerat, E., Daubin, V., & Moran, N. A. (2003). From gene trees to organismal phylogeny in prokaryotes: The case of the gamma-proteobacteria. *Plos Biology*, *1*(1), 101–109.

Li, W. Z., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Liang, Y., Jiang, Y., Wang, F., Wen, C., Deng, Y. E., Xue, K., Qin, Y., Yang, Y., Wu, L., Zhou, J., & Sun, B. O. (2015). Long-term soil transplant simulating climate change with latitude significantly alters microbial temporal turnover. *The ISME Journal*, *9*(12), 2561–2572. https://doi.org/10.1038/ismej.2015.78

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Linz, A. M., He, S., Stevens, S. L. R., Anantharaman, K., Rohwer, R. R., Malmstrom, R. R., Bertilsson, S., & McMahon, K. D. (2018). Freshwater carbon and nutrient cycles revealed through reconstructed population genomes. *Peerj*, *6*, e6075. https://doi.org/10.7717/peerj.6075

Louca, S., Polz, M. F., Mazel, F., Albright, M. B. N., Huber, J. A., O'Connor, M. I., Ackermann, M., Hahn, A. S., Srivastava, D. S., Crowe, S. A., Doebeli, M., & Parfrey, L. W. (2018). Function and functional redundancy in microbial systems. *Nature Ecology & Evolution*, *2*(6), 936–943. https://doi.org/10.1038/s41559-018-0519-1

Lundin, D., Severin, I., Logue, J. B., Ostman, O., Andersson, A. F., & Lindstrom, E. S. (2012). Which sequencing depth is sufficient to describe patterns in bacterial alpha- and beta-diversity? *Environmental Microbiology Reports*, *4*(3), 367–372. https://doi.org/10.1111/j.1758-2229.2012.00345.x

Maurice, C. F., Bouvier, C., de Wit, R., & Bouvier, T. (2013). Linking the lytic and lysogenic bacteriophage cycles to environmental conditions, host physiology and their variability in coastal lagoons. *Environmental Microbiology*, *15*(9), 2463–2475. https://doi.org/10.1111/1462-2920.12120

McArthur, J., Kovacic, D., & Smith, M. (1988). Genetic diversity in natural-populations of a soil bacterium across a landscape gradient. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(24), 9621–9624. https://doi.org/10.1073/pnas.85.24.9621

McGill, B. J., Enquist, B. J., Weiher, E., & Westoby, M. (2006). Rebuilding community ecology from functional traits. *Trends in Ecology & Evolution*, *21*(4), 178–185. https://doi.org/10.1016/j.tree.2006.02.002

Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development*, *15*(6), 589–594. https://doi.org/10.1016/j.gde.2005.09.006

Medini, D., Serruto, D., Parkhill, J., Relman, D. A., Donati, C., Moxon, R., Falkow, S., & Rappuoli, R. (2008). Microbiology in the post-genomic era. *Nature Reviews Microbiology*, *6*(6), 419–430. https://doi.org/10.1038/nrmicro1901

Miyashita, K., Fujii, T., & Sawada, Y. (1991). Molecular cloning and characterization of chitinase genes from Streptomyces lividans 66. *Journal of General Microbiology*, *137*, 2065–2072. https://doi.org/10.1099/00221287-137-9-2065

Morris, R. M., Rappe, M. S., Connon, S. A., Vergin, K. L., Siebold, W. A., Carlson, C. A., & Giovannoni, S. J. (2002). SAR11 clade dominates ocean surface bacterioplankton communities. *Nature*, *420*(6917), 806–810. https://doi.org/10.1038/nature01240

Nakamura, Y., Itoh, T., Matsuda, H., & Gojobori, T. (2004). Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature Genetics*, *36*(7), 760–766. https://doi.org/10.1038/ng1381

Nemergut, D. R., Schmidt, S. K., Fukami, T., O'Neill, S. P., Bilinski, T. M., Stanish, L. F., Knelman, J. E., Darcy, J. L., Lynch, R. C., Wickey, P., & Ferrenberg, S. (2013). Patterns and processes of microbial community assembly. *Microbiology and Molecular Biology Reviews*, *77*(3), 342–356. https://doi.org/10.1128/MMBR.00051-12

Oliver, A., Lilley, A. K., & van der Gast, C. J. (2012). Species-time Relationships for Bacteria. In L. A. Ogilvie, & P. R. Hirsch (Eds.), *Microbial ecological theory: Current perspectives* (pp. 71–85). Caister Academic Press.

P. White, E., B. Adler, P., K. Lauenroth, W., A. Gill, R., Greenberg, D., M. Kaufman, D., Rassweiler, A., A. Rusak, J., D. Smith, M., R. Steinbeck, J., B. Waide, R., & Yao, J. (2006). A comparison of the species-time relationship across ecosystems and taxonomic groups. *Oikos*, *112*(1), 185–195. https://doi.org/10.1111/j.0030-1299.2006.14223.x

Raes, J., Korbel, J. O., Lercher, M. J., von Mering, C., & Bork, P. (2007). Prediction of effective genome size in metagenomic samples. *Genome Biology*, *8*(1), R10. https://doi.org/10.1186/gb-2007-8-1-r10

Raes, J., Letunic, I., Yamada, T., Jensen, L. J., & Bork, P. (2011). Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Molecular Systems Biology*, *7*(1), 534. https://doi.org/10.1038/msb.2011.6

Rappe, M. S., Connon, S. A., Vergin, K. L., & Giovannoni, S. J. (2002). Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature*, *418*(6898), 630–633. https://doi.org/10.1038/nature00917

Rodriguez-Valera, F., Martin-Cuadrado, A.-B., Rodriguez-Brito, B., Pasic, L., Thingstad, T. F., Rohwer, F., & Mira, A. (2009). OPINION explaining microbial population genomics through phage predation. *Nature Reviews Microbiology*, *7*(11), 828–836. https://doi.org/10.1038/nrmicro2235

Shade, A., Caporaso, J. G., Handelsman, J., Knight, R., & Fierer, N. (2013). A meta-analysis of changes in bacterial and archaeal communities with time. *The ISME Journal*, *7*(8), 1493–1506. https://doi.org/10.1038/ismej.2013.54

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., d'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., … Velayoudon, D. (2015). Structure and function of the global ocean microbiome. *Science*, *348*(6237), 1261359. https://doi.org/10.1126/science.1261359

Svitil, A. L., Chadhain, S. M. N., Moore, J. A., & Kirchman, D. L. (1997). Chitin degradation proteins produced by the marine bacterium Vibrio harveyi growing on different forms of chitin. *Applied and Environmental Microbiology*, *63*(2), 408–413. https://doi.org/10.1128/AEM.63.2.408-413.1997

van der Gast, C. J., Ager, D., & Lilley, A. K. (2008). Temporal scaling of bacterial taxa is influenced by both stochastic and deterministic ecological factors. *Environmental Microbiology*, *10*(6), 1411–1418. https://doi.org/10.1111/j.1462-2920.2007.01550.x

Verduin, J. (1956). Energy fixation and utilization by natural communities in western Lake Erie. *Ecology*, *37*(1), 40–50. https://doi.org/10.2307/1929667

Vergin, K. L., Done, B., Carlson, C. A., & Giovannoni, S. J. (2013). Spatiotemporal distributions of rare bacterioplankton populations indicate adaptive strategies in the oligotrophic ocean. *Aquatic Microbial Ecology*, *71*(1), 1-U129. https://doi.org/10.3354/ame01661

Ward, C. S., Yung, C.-M., Davis, K. M., Blinebry, S. K., Williams, T. C., Johnson, Z. I., & Hunt, D. E. (2017). Annual community patterns are driven by seasonal switching between closely related marine bacteria. *The ISME Journal*, *11*(6), 1412–1422. https://doi.org/10.1038/ismej.2017.4

Whitaker, R. J., & Banfield, J. F. (2006). Population genomics in natural microbial communities. *Trends in Ecology & Evolution*, *21*(9), 508–516. https://doi.org/10.1016/j.tree.2006.07.001

Wilhelm, L. J., Tripp, H. J., Givan, S. A., Smith, D. P., & Giovannoni, S. J. (2007). Natural variation in SARII marine bacterioplankton genomes inferred from metagenomic data. *Biology Direct*, *2*, 27. https://doi.org/10.1186/1745-6150-2-27

Wilson, D. S. (1992). Complex interactions in metacommunities, with implications for biodiversity and higher levels of selection. *Ecology*, *73*(6), 1984–2000. https://doi.org/10.2307/1941449

Yan, Q., van der Gast, C. J., & Yu, Y. (2012). Bacterial community assembly and turnover within the intestines of developing zebrafish. *PLoS One*, *7*(1), e30603. https://doi.org/10.1371/journal.pone.0030603

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S. C., Ester, M., Foster, L. J., & Brinkman, F. S. L. (2010). PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, *26*(13), 1608–1615. https://doi.org/10.1093/bioinformatics/btq249

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Beier S, Andersson AF, Galand PE, et al. The environment drives microbial trait variability in aquatic habitats. *Mol Ecol*. 2020;00:1–13. https://doi.org/10.1111/mec.15656