Molecular Phylogenetics and Evolution 107 (2017) 324–337

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Are ocean currents to slow to counteract SAR11 evolution? A next-generation sequencing, phylogeographic analysis

Julieta M. Manrique, Leandro R. Jones *

Consejo Nacional de Investigaciones Científicas y Técnicas, Av. Rivadavia 1917, (C1083ACA) Buenos Aires, Argentina
Laboratorio de Virología y Genética Molecular, Facultad de Ciencias Naturales sede Trelew, Universidad Nacional de la Patagonia San Juan Bosco, Argentina

ARTICLE INFO

Article history:
Received 1 June 2016
Revised 18 November 2016
Accepted 24 November 2016
Available online 25 November 2016

Keywords:
Bacteria
SAR11
Phylogeography
Biogeography
Phylogenetic
16S

ABSTRACT

This work set out to shed light on the phylogeography of the SAR11 clade of Alphaproteobacteria, which is probably the most abundant group of heterotrophic bacteria on Earth. In particular, we assessed the degree to which empirical evidence (environmental DNA sequences) supports the concept that SAR11 lineages evolve faster than they are dispersed thus generating vicariant distributions, as predicted by recent simulation efforts. We generated 16S rRNA gene sequences from surface seawater collected at the South West Atlantic Ocean and combined these data with previously published sequences from similar environments from elsewhere. Altogether, these data consisted in about 1e6 reads, from which we generated 355,306 high quality sequences of which 95,318 corresponded to SAR11. Quantitative phylogeographic analyses supported the existence of a spatially explicit distribution of SAR11 species and provided evidence in favor of the idea that dispersal limitations significantly contribute to SAR11 radiation throughout the world's oceans. Likewise, pairwise phylogenetic distances between the communities studied here were significantly correlated with the genetic divergences predicted by a previously proposed neutral model. As discussed in the paper, these findings are compatible with the concept that the ocean surface constitutes a homogeneous environment for SAR11, in agreement with previous experimental data. We discuss the implications of this hypothesis in a global change scenario. This is the first study combining high throughput sequencing and phylogenic analysis to study bacterial phylogeography and reporting a distance decay pattern of phylogenetic distances for bacteria.

1. Introduction

A relatively reduced group of cosmopolitan microorganisms, such as the SAR11 clade of Alphaproteobacteria, consistently dominates marine ecosystems (Giovannoni et al., 2005; Morris et al., 2002; Zhao et al., 2016). Besides profusion and ubiquity, the noteworthy features of SAR11 include proteorhodopsin phototrophy (Steindler et al., 2011), the capacity of maintaining an extracellular phosphate buffer, as well as probably other metabolite buffers (Zubkov et al., 2015), and the ability to grow on highly abundant osmolytes produced by other marine microorganisms (Lidbury et al., 2014). In addition, transcriptional network analyses have shown that SAR11 and other microorganisms' metabolisms are highly coordinated, with SAR11 transcript nodes displaying the highest degree centrality (Aylward et al., 2015). Remarkably, this coordination seems to be conserved among communities from very disparate marine environments. Furthermore, recent experimental studies showed that transcriptional levels in *Pelagibacter ubique*, a cultivated representative of the SAR11 clade, are highly recalcitrant to external stimuli, suggesting a high resilience to environmental variation (Cottrell and Kirchman, 2016). A current challenge is to understand the diversity patterns and diversification mechanisms of these and other ubiquitous and profuse marine bacteria.

Selection has been long recognized as a major force in microbial evolution. A previous work have suggested that SAR11 phylotypes defined based on internal transcribed spacer (*ITS*) sequences are adapted to temperature and latitude (Brown et al., 2012). Brown et al. (2012), however, neither evaluated the influence of dispersal limitations (*DL*) nor corrected for the phylogeny in their models. By the other side, evidence is accumulating that neutral evolution coupled to *DL* can generate and maintain significant biodiversity patterns (Martiny et al., 2011; Whitaker et al., 2003; Zinger et al., 2014). For the SAR11 case, a recent simulation effort showed that ocean currents may be slow enough to allow for evolutionary drift of marine populations (Hellweger et al., 2014), though empirical

* Corresponding author at: Laboratorio de Virología y Genética Molecular, Facultad de Ciencias Naturales sede Trelew, Universidad Nacional de la Patagonia San Juan Bosco, 9 de Julio y Belgrano s/n (9100), Trelew, Chubut, Argentina.
E-mail addresses: ljones@conicet.gov.ar, lrj000@gmail.com (L.R. Jones).

http://dx.doi.org/10.1016/j.ympev.2016.11.015
1055-7903/

corroboration is lacking. Thus, the present study set out to evaluate the degree to which empirical evidence (environmental DNA sequences) supports the concept that SAR11 lineages evolve faster than they are dispersed.

The possibility of combining next generation sequencing technologies with modern phylogenetic algorithms capable of inferring reliable phylogenies from tens of thousands of sequences (Goloboff et al., 2009; Price et al., 2010), represents an outstanding opportunity to shed new insights on the historical factors underlying the current distributions of marine microorganisms. Here, we used high throughput sequencing (*HTS*) data and explicit, quantitative phylogeographic models to check if SAR11 populations are subjected to *DL*. We also evaluated the degree to which pairwise phylogenetic distances between SAR11 communities matched the divergences predicted by Hellweger et al.'s (2014) neutral agent-based model. We generated *HTS* data for the South West Atlantic Ocean (*SWAO*), for which no previous SAR11 data were available, and compiled previously published data that (i) corresponded to the same target gene as our own data, (ii) were generated from surface seawater, (iii) for which geographic and environmental data were published and (iv) for which data in standard flowgram format (aka SFF) were available allowing to apply adequate quality controls. From a total of about 1e6 16S rDNA reads, we obtained 95,318 SAR11 high-quality 16S sequences, encompassing data from 11 worldwide distributed locations. Analyses of these data combined with geographic and ecological covariates, supported the existence of a spatially explicit SAR11 biogeography and revealed a significant covariation between the group's phylogeny and distribution, compatible with *DL*. Furthermore, pairwise phylogenetic distances between SAR11 populations were congruent with the divergences in the surface ocean predicted by Hellweger et al.'s (2014) neutral model.

2. Materials and methods

2.1. Sampling

Samples from the *SWAO* were collected in January 2014 (austral summer) at 39.95° S 55.68° W (*MDQ* sample), 45.93° S 57.7° W (*BH* sample) and 46° S 59.39 °W (*TAL* sample) during R/V "Coriolis II" expedition. Each of these samples consisted of approximately 3 L of water that were collected at ~1 m depth using niskin bottles attached to a rosette and processed immediately on board of "Coriolis II". Oceanographic data were monitored with a Sea-Bird CTD, attached to the rosette. The samples from the seawater section of the Chubut River (*ChR*) estuary (43.44° S 65.11° W) were taken in January 2013 (*S13*), July 2013 (*W13*), January 2014 (*S14*) and July 2014 (*W14*) during the high tide. These samples (~3 L) were collected in acid-cleaned carboy-tanks using a peristaltic pump with the intake submerged at a depth of ~1 m, and immediately transported to the laboratory in a portable refrigerated cabinet (~4 °C). All the samples were prefiltered with a 100 μm pore size Nitex mesh to remove large particulate material and zooplankton. After that, picoplankton DNA was isolated as described in the following section.

2.2. Generation of 16S gene libraries

The prefiltered samples were successively filtered through 47 mm diameter polycarbonate membrane filters (MSI Westboro) of decreasing pore sizes (20, 10, 5 and 0.22 μm) to separate the different cell size fractions, by applying a pressure of 20 mmHg. The picoplankton was therefore concentrated and immobilized onto the 0.22 μm membranes. The filters were replaced when clogged, and were stored in 1.5 ml eppendorf tubes at −30 °C until processed. Each filter was cut in small pieces (~1/8 of the filter) with a 70% ethanol-cleaned, flame-sterilized scissor and the picoplankton DNA was recovered using a CTAB extraction method described before (Manrique et al., 2012). The DNAs obtained for each location/sampling time were pooled, precipitated with 0.1 vol of 3 M Sodium Acetate (pH 5.3) and 2.5 volumes of ethanol for 1 h at 4 °C, centrifuged at 21,000*g* for 30 min at 4 °C (Sorval Legend Micro 17 R, Thermo Scientific), and finally resuspended in 50 μl of ultrapure, DNase-free water. DNA quality was evaluated by gel electrophoresis and spectrophotometry using a Nanovue Plus (GE healthcare) spectrophotometer, as outlined elsewhere (Jones and Manrique, 2015). DNA yields were determined by densitometry analysis against standards of 10, 20, 30, 40, 60 and 100 ng of DNA (High DNA Mass Ladder, Invitrogen) using the software ImageJ. DNA samples were kept at −30 °C until used. DNA samples were PCR amplified (done in triplicate) for amplicon pyrosequencing using primers targeting bacterial 16S rRNA genes as described previously (Manrique et al., 2012). Pyrosequencing reactions were performed at MrDNA sequencing service (http://mrdnalab.com/) with the Roche 454 FLX titanium and reagents, following standard manufacturer's instructions. The obtained data were deposited in GenBank under accession numbers SRR3176906, SRR3180667, SRR3180668, SRR3180669, SRR3180670, SRR3180671 and SRR3180683.

2.3. Datasets

In addition to the *SWAO* samples, we compiled pyrosequencing data from Delaware Bay (*DEL*), Pearl River Estuary (*PEA*), Columbia River Estuary (*CoR*) and a coastal region close to *CoR* (*NP*), and the China Sea (*CS1*, *CS2* and *CS3*) (Table 1; Fig. 1). The *HTS* data from these last locations were downloaded from GenBank using the Sequence Read Archive Toolkit (http://www.ncbi.nlm.nih.gov/sra). These samples were selected based on environmental similarity with our samples (surface seawater), targeted gene region and availability of geographic and environmental information and data in SFF format. Some of these studies generated data for several environments. In that cases, only the data corresponding to surface seawater were gathered to minimize among sample environmental variation.

For sequence comparisons and quality controls (described in the following sections) we used a reference dataset containing sequences from the different groups previously defined inside SAR11 (Clade Ia: X52280, L10935, U13159, U75253, AF245616, AJ400350, AF327027, AF268220, AY033306 and AY033320; Clade Ib: X52172, U75649 and AY033312; Clade II: U75254, U75256, U75257, AJ400351, AY033299, AY033303 and AY033322; Clade III: Z99997, AF418965 and AY145598; Clade IV: U70686, AF353226 and AF353229).

2.4. Quality control and data selection

In order to minimize potential PCR and sequencing errors, the *HTS* data from all the studied sites (Table 1; Fig. 1) were preprocessed using Mothur following the general guidelines described in reference (Schloss et al., 2011). As a first step, flowgrams displaying differences of 3 and 2 nucleotides in the primer and barcode sequences, respectively, homopolymer stretches longer than 9 positions and less than 360 or more than 720 flows were dismissed. The remaining data were denoised using the PyroNoise algorithm (Quince et al., 2011). After that, the primer and barcode sequences were eliminated and the sequences were aligned against the SAR11 reference sequences (please see Section 2.3, Datasets) in order to detect and eliminate sequences that didn't overlap with the target gene region. Before aligning the *HTS* data, the reference sequences were pre-aligned as detailed in

Table 1
Origins of the samples analyzed, codes used along this work, geographic coordinates, and number of high-quality SAR11 sequences.

Site	Code	Coordinates	# reads	Reference
Chubut River Estuary[a]	*ChR*	43.35° S, 65.01° W	34351	This work
Columbia River Estuary	*CoR*	45° N, 124° W	3555	Fortunato et al. (2013)
Delaware Bay	*DEL*	38.85° N, 75.1° W	30476	Campbell et al. (2011)
Newport Coast	*NP*	44.65° N, 125.36° W	11,569	Fortunato et al. (2013)
Pearl Estuary	*PEA*	22° N, 114.07° E	643	Liu et al. (2015)
China Sea A	*CSA*	16.6° N, 113.6° E	1172	Sun et al. (2014)
China Sea B	*CSB*	19.1° N, 113.6° E	1111	Sun et al. (2014)
China Sea C	*CSC*	21.5° N, 115.5° E	954	Sun et al. (2014)
South West Atlantic A	*MDQ*	39.95° S, 55.68° W	6824	This work
South West Atlantic B	*BH*	45.93° S, 57.7° W	3873	This work
South West Atlantic C	*TAL*	46° S, 59.39° W	790	This work

[a] Four samples (*S13*, *W13*, *S14* and *W14*) were collected along two years in the seawater section of the Chubut River estuary. The numbers of SAR11 sequences per sample were 8069, 5547, 8421 and 12,314, respectively.

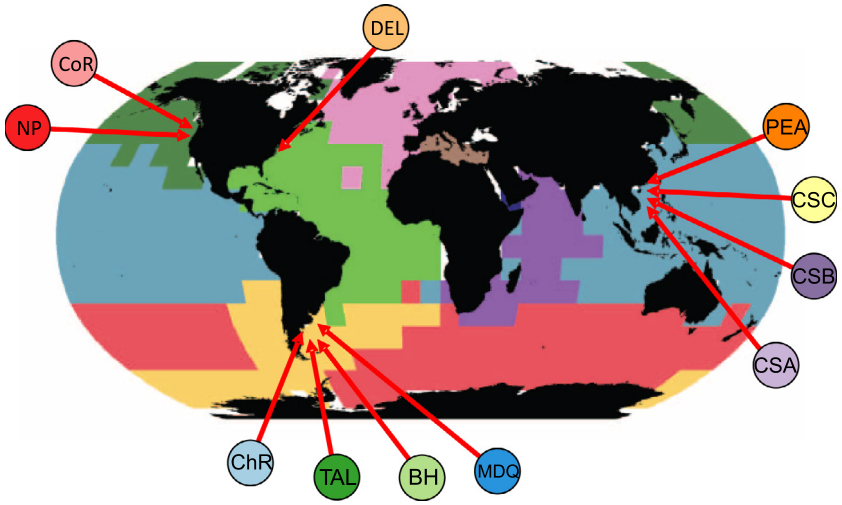


Fig. 1. Origins of the SAR11 sequences analyzed here. *ChR* Chubut River estuary; *MDQ*, *TAL*, *BH* oceanic samples from the South West Atlantic Ocean; *CoR* Columbia River Estuary; *NP* ocean coast close to *CoR*; *DEL* Delaware bay; *PEA* Pearl estuary; *CSA*, *CSB* and *CSC*, China Sea. The map displays biogeographic patterns in global surface ocean microbes predicted by a neutral agent-based model. From Hellweger, F.L., van Sebille, E., Fredrick, N.D., 2014; Biogeographic patterns in ocean microbes emerge in a neutral agent-based model; Science 345, 1346–1349. Reprinted with permission from AAAS.

Section 2.5.2, after which the reads from each dataset were aligned against the obtained reference alignment using the *align.seqs* function of Mothur. Then, we merged the sequences that were within 2 bp of a more abundant sequence to further reduce potential PCR and sequencing errors. Finally, chimeras were identified using the UCHIME algorithm implemented in Mothur and removed from the datasets.

The high quality 16S sequences were classified against the SILVA database (http://www.arb-silva.de) using the *classify.seqs* function of Mothur. After that, the putative SAR11 reads identified were submitted to phylogenetic comparisons against the SAR11 reference sequences (Section 2.3) and an outgroup sequence (*Rickettsia bellii*, accession number U11014.1) in order to confirm the classification according to the topological placement in the trees (please see the following section for details on phylogenetic methods).

2.5. Data analyses

2.5.1. OTU-based analyses

For the OTU-based studies, the high-quality SAR11 sequences obtained as described in Section 2.4 were aligned using Mothur (*align.seqs* command) and a prealigned SAR11 reference dataset (Section 2.3) as template. After that, uncorrected pairwise distances were obtained with Mothur counting internal gaps of any extension as a single mutation and ignoring terminal gaps. The obtained distance matrix was used to generate 100% (aka *unique*), 99% and 97% clusters with the average neighbor linkage method, which were processed with the *make.shared* command with default settings. To identify the reads corresponding to 100% similarity OTUs that were present just in a single sample (hereafter *endemic OTUs*), we created a Mothur database and then parsed it with an R script that returned, for each location, a list containing the endemic OTUs and their frequencies. Please notice that the need of using a computer assisted approach (*i.e.* an R script) does not obey to the use of any particular criterion to define endemism but to the large number of OTUs analyzed, which precludes doing these analyses by hand. When needed, and as indicated along the paper, the data were normalized using rarification to avoid potential biases due to sequencing depth differences between the samples. To this aim, random samples of 600 reads were taken from each community using the *sub.sample* command of Mothur.

2.5.2. Phylogenetic analyses

For phylogenetic analyses, sequence alignments were obtained with MAFFT (Katoh and Standley, 2014) and visually inspected with Jalview (Waterhouse et al., 2009). Given the large number of sequences studied, direct alignment (that is, aligning all the sequences from scratch) using the iterative refinement methods implemented in MAFFT (Katoh and Standley, 2014) was unfeasible.

By the other side, the progressive alignment approaches produced poor results, as revealed by a profusion of obviously misaligned positions. Thus, we generated a reliable alignment of the SAR11 reference sequences (please see the section *Datasets* above) by the MAFFT's *linsi* iterative refinement method, and then the *HTS* reads were incorporated into this alignment using the *add fragments* option of MAFFT. Maximum Likelihood phylogenies were obtained with *FastTree*, which can infer reliable trees from thousands of sequences (Price et al., 2010). The evolutionary model (*JC69*) was inferred using *Mr.AIC* (Nylander, 2004) using the reference alignment as input. *FastTree* was run with default parameters, excepting that a single rate category was implemented in the evolutionary model. Under these conditions, the program performs up to $4 \times \log 2(N)$ rounds of minimum-evolution Nearest Neighbor Interchange (*NNI*), 2 rounds of Subtree Pruning Regrafting (*SPR*) moves and up to $2 \times \log(N)$ rounds of maximum-likelihood *NNIs*, where N is the number of unique sequences in the dataset. Parsimony trees were obtained with the program TNT (Goloboff and Catalano, 2016). We set the program to perform 100 random addition sequences followed by *SPR*, holding one tree while swapping. Each of the obtained trees were subjected to a Sectorial Search (*SS*) round and to a final SPR round. For *SS*, the trees were divided in 40–60 sectors covering all the tree. Sectors greater than 500 were analyzed by a combined analyses [*RAS* + Tree Drift (*TD*) + Tree Fusing (*TF*)], which implemented 5 cycles of *TD* and 3 rounds of *TF*. Sectors greater than 1000 were analyzed by *TD*. All the trees were saved and used in the quantitative phylogeographic analyses as described below.

2.5.3. Quantitative phylogeographic analyses

Phylogenetic distances were estimated by the Generalized Unique Fraction Metric (GUniFrac) (Chen et al., 2012). The metric works as follows: Let A and B be two communities and let $[S_A]$ and $[S_B]$ be the species sets found in communities A and B, respectively. In addition, let T be a rooted tree containing, at least, all the species present in $[S_A]$ and $[S_B]$. Now consider the set of tree branches $[B_A]$, whose descendants include the sequences from $[S_A]$ and the set of branches $[B_B]$, whose descendants include the sequences present in $[S_B]$. The amount of evolution underwent by community A can be estimated by L_A, the sum of the lengths of the branches included in $[B_A]$ and, likewise, community B evolution can be estimated by the sum of the lengths of the branches included in $[B_B]$, L_B. Notice that some of the branches present in $[B_A]$ can be also present in $[B_B]$. The *A-B* GUniFrac distance is proportional to the fraction of the branch length of the tree that leads to descendants from either one community or the other, but not both, that is, the sum of L_A and L_B minus the sum of the branch lengths in $[B_A] \cap [B_B]$. Conveniently, the GUniFrac metric have a parameter, α, that controls the contribution of high-abundance branches. Here, we used an α of 0.5 for achieving the overall best power (Chen et al., 2012).

The relationships between phylogenetic distances and water temperature, latitude, geographic distance and expected divergences were assessed by Mantel and Partial Mantel tests and distance-based linear models in the *ecodist* R package (Goslee and Urban, 2007). To reduce de effect of different measurement scales, variables were normalized by subtracting the corresponding means and dividing by the standard deviations. We used 10,000 permutations and $\rho = 0$ as the null hypothesis in Mantel tests.

3. Theory

Figs. 2–4 describe different possible phylogeographic scenarios of ubiquitous and profuse marine bacteria. With ubiquitous, or cosmopolitan, we mean that the lineage is present everywhere. However, some genetic variants (hereafter *ribotypes*) may have circumscribed or discontinuous distributions, as detailed later. We define Richness (R) as the number of ribotypes in a given sample (square boxes labeled 1., 2. and 3. in Figs. 2–4). Geographic distances between sampling sites are represented by the letter D. Pairwise phylogenetic distances (*PD*) between samples are proportional to the sum of the lengths of the tree branches span by one or the other sample but not both (represented by red lines in Figs. 2–4).

In the case of a homogeneous distribution, that is in the absence of a biogeography, R would be the same for all the samples, regardless the sampled area. Likewise, pairwise *PD*s expectance is zero for all D (Fig. 2).

In the case that habitat filtering (*HF*) determines microbial distributions, everything is everywhere but the environment selects.

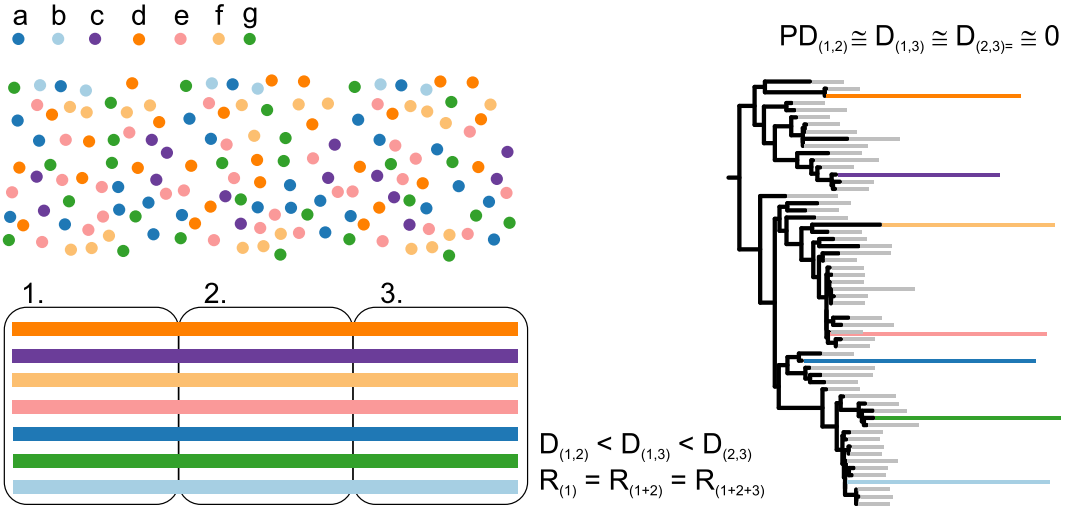


Fig. 2. Phylogeography of abundant marine bacteria with a homogeneous distribution. The colored points labeled *a* to *g* in the upper left corner of the figure represent different ribotypes inside a worldwide distributed lineage (represented by the cloud of points). The boxes labeled 1–3 represent three samples among which the geographic distances between samples 1 and 3 are greater than the distances between samples 1 and 2 and 2 and 3; that is, $D_{(1,2)} < D_{(1,3)} > D_{(2,3)}$. The colored rectangles drawn along these boxes represent the distribution of ribotypes *a–g* among the samples. Species richness (measured as number of ribotypes) is represented by the letter *R*. To the right, a phylogenetic tree is shown with ribotypes *a–g* indicated by colors. The bar heights represent ribotypes' abundances. The gray bars represent unsampled or extinct ribotypes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

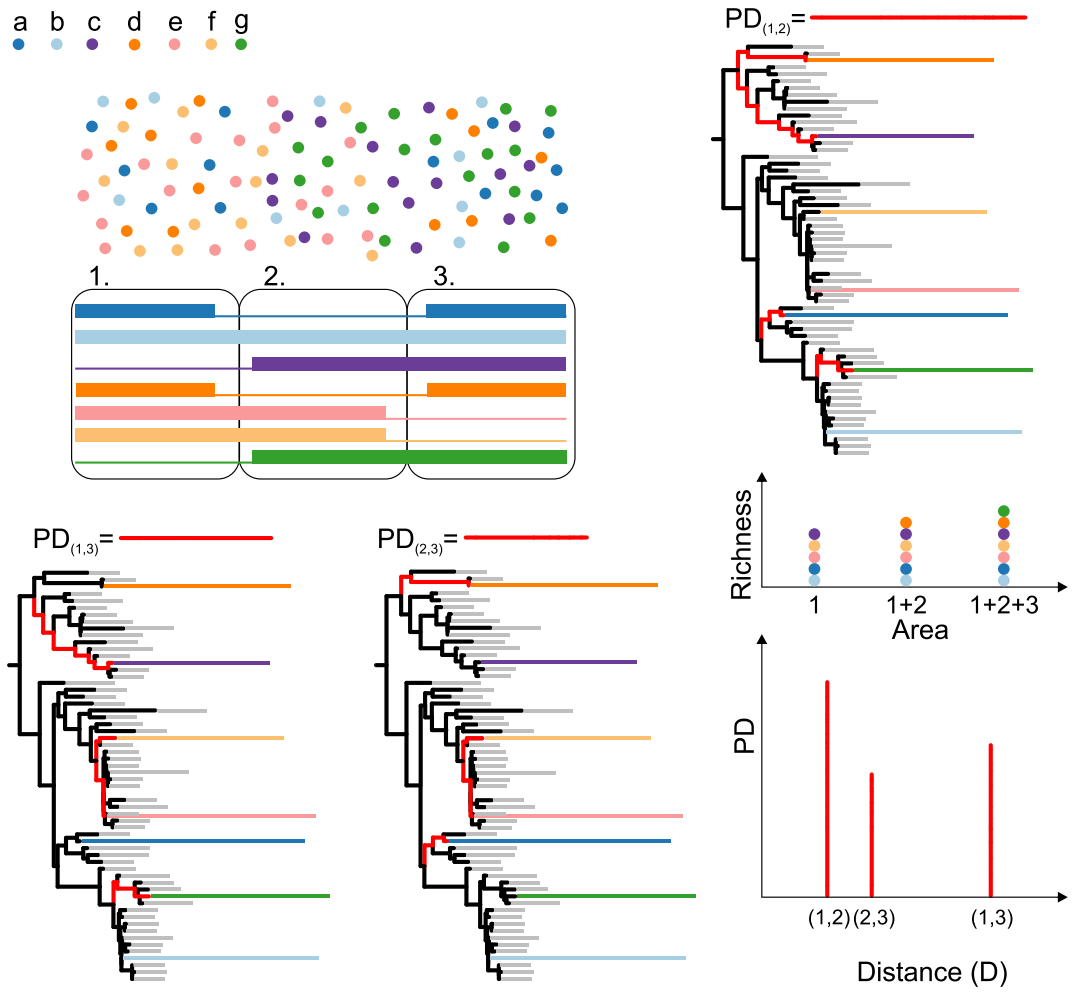


Fig. 3. Phylogeography of abundant marine bacteria under *HF*. Ribotypes, samples and the ribotypes distribution among the samples are represented as in Fig. 1. The red colored branches of the trees, which are also drawn as in Fig. 1, represent pairwise phylogenetic distances (*PD*) between the samples. The distances are also drawn over the trees using red lines equivalent to concatenating the colored branches of the trees. The figure also depicts the expected taxa-area relationship and the relationship between pairwise phylogenetic and geographic (*D*) distances. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Dispersal is very efficient, ensuring that new genetic variants rapidly become cosmopolitan. Thus, historical factors are generally irrelevant regarding the group's distribution. Richness is expected to increase as the sampling area does because increasing the sampled area results in an increase of the niches sampled (Fig. 3). However, one might expect the taxa-area relationship to reach a plateau if the number of sampled ribotypes approaches the number of ecotypes present in the lineage. It must be expected for geographically close samples to share several ribotypes, because many environmental variables are spatially autocorrelated (ribotype *e* and *f* in samples 1 and 2, and ribotypes *c* and *g* in samples 2 and 3; Fig. 3). However, when two samples are far enough apart from one another, the corresponding locations can be environmentally the same, in terms of ecotype niche. Thus, samples that are sufficiently far apart from one another also will share a certain amount of ribotypes due to habitat selection (ribotypes *a* and *d* in samples 1 and 3; Fig. 3). This implies that covariation between the ribotypes distribution among particular communities and the phylogeny should be limited, even if phylogenetic inertia (that is phylogenetically related variants sharing similar environmental preferences) were widespread in the group (Table 2). Endemic ribotypes can of course exist. But endemicity is expected to be ephemeral, and thus rare among samples, due to rapid dispersal (Table 2). Under this scenario everything is everywhere but some ribotypes remain undetected due to their very low frequencies (thin lines in boxes 1, 2 and 3 of Fig. 3). This can obey to low growth rates due to environmental constraints and/or the presence of inactive cells or resistance forms dispersed from other regions. Nearby samples can share phylogenetically related sequences due to phylogenetic inertia (ribotypes *e* and *f* in samples 1 and 2, Fig. 3). However, convergent evolution should be common too, due to the group's profusion and ubiquity (ribotypes *c* and *g* in samples 2 and 3; Fig. 3). Finally, some ribotypes may present widespread distributions due to adaptation to commonplace conditions (ribotype *b*, Fig. 3).

In a vicariance model in which ribotype distributions can be circumscribed due to *DL*, global diversity is governed by allopatric diversification and (limited) dispersal. Under dispersal limiting conditions, dispersal chance is usually negatively correlated with geographical distance, which therefore can be used as a proxy for the strength of *DL* between pairs of communities. Richness increases as the sampled area does, as in the *HF* model (Fig. 4). However, in this case the function shouldn't reach a plateau because diversification is happening everywhere, in an independent fashion, thus generating endemic ribotypes everywhere (ribotypes *a*, *c* and *e* being present just in samples 2, 3 and 1, respectively; Fig. 4). Endemic ribotypes should therefore be common (Table 2). Under *DL*, close locations are expected to share some ribotypes as in the HF scenario. However, in this case the phenomenon should be interpreted mainly as dispersal among nearby places rather than due to constraints imposed by spatially

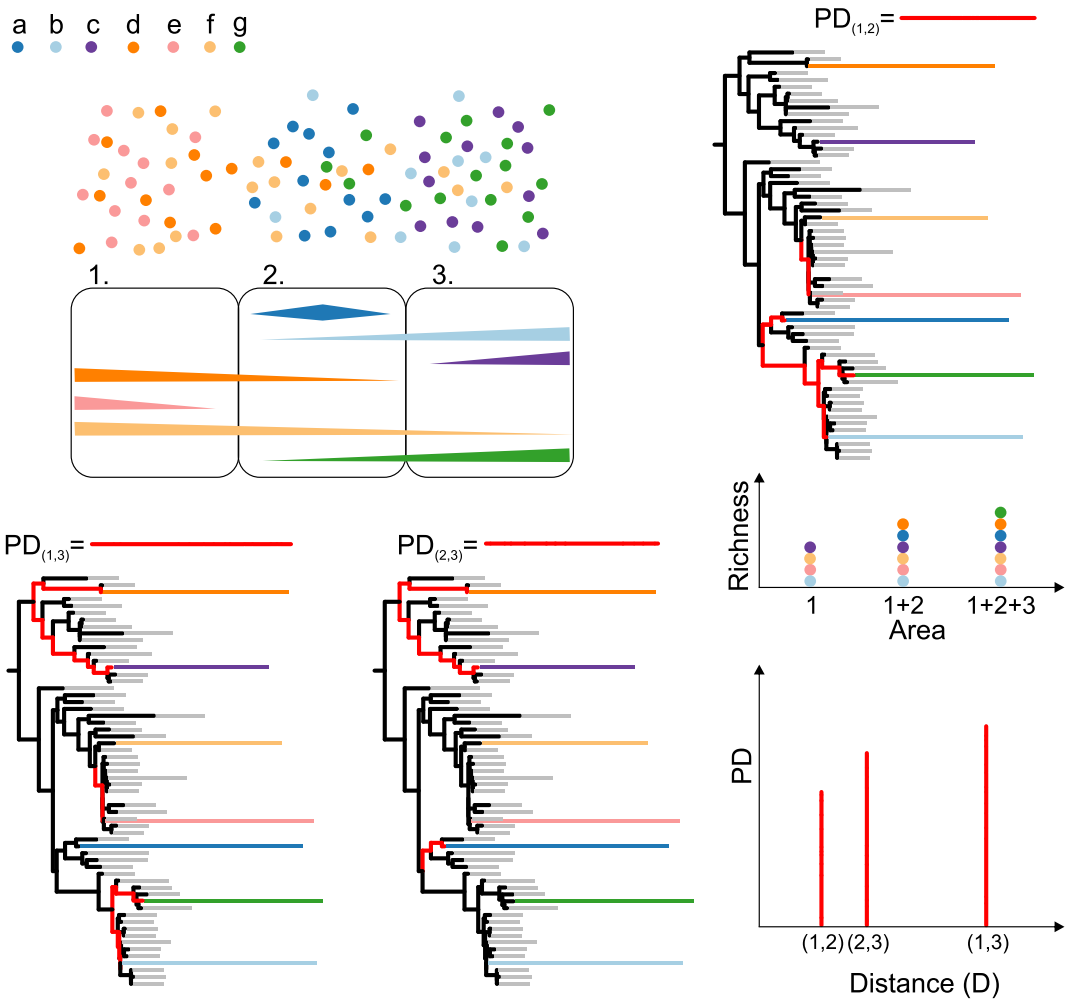


Fig. 4. Phylogeography of abundant marine bacteria under *DL*. Ribotypes, samples and the ribotypes distribution among the samples are represented as in Fig. 1. Phylogenetic distances, the expected taxa-area relationship and the expected relationship between phylogenetic and geographic distances are represented as in Fig. 2.

Table 2
Key features of phylogeographic models of ubiquitous and profuse marine bacteria.

Model	Endemic variants	Taxa/area relationship	Covariation[a]
HF[b]	Rare	Lineal/plateau	Limited
DL[c]	Common	Lineal	Yes

[a] Distribution of ribotypes between communities exhibiting covariation with the phylogeny.
[b] Habitat filtering.
[c] Dispersal limitations.

autocorrelated environmental covariates. As a consequence, many of the lineages common to nearby samples will be identical due to dispersal and/or phylogenetically related due to diversification coupled to dispersal (ribotypes *b* and *g* in samples 2 and 3 and ribotype *d* in samples 1 and 2, Fig. 4). This allows deriving the property that pairwise phylogenetic distances are expected to be positively correlated with geographic distances (Fig. 4), that is that the distribution of ribotypes between different communities covary with the phylogeny (Table 2). In a tree, such covariation is reflected in the "clumping" of sequences of individual and nearby samples. The phenomenon can be quantitatively tackled by distance decay analyses, which under *DL* have to give negative correlations between geographic and phylogenetic distances. Conveniently, the correlation between distance matrices can be statistically assessed by the Mantel test (Legendre and Legendre, 2003; Mantel, 1967). Finally, there can be cosmopolitan ribotypes in response to long term dispersal (ribotype *f*; Fig. 4). It is important to remark that the *DL* model requires new genetic variants to be capable of thriving in a variety of environments such that dispersion to nearby locations is generally possible. As discussed in Section 5, there are several characteristics of SAR11 that might be responsible for such capability.

4. Results

In an effort to minimize environmental differences, we limited our study to surface seawater. Samples were collected at three offshore locations (*MDQ, BH* and *TAL*) and from the seawater section of an estuary (*S13*, *W13*, *S14* and *W14*) from the *SWAO*, for which SAR11 data were lacking so far (Table 1). Picoplanktonic DNA was isolated by membrane immobilization, and 16S sequence libraries were obtained by *HTS* of the V1–V3 region of the gene. Additionally, we compiled previously published *HTS* data encompassing or significantly overlapping the same 16S gene region, and for which environmental records and standard flowgram format files were available (Table 1; Fig. 1). The geographically closest samples corresponded to the *ChR* ones, whereas the farthest sites were located 19,720 km apart. The average geographic distance between the samples was 8611 km, the first and third quartiles were 4668 km and 11,540 km, respectively, and the standard

deviation was 4915. The raw data from the 11 locations studied here consisted in 957,561 reads. After quality controls aimed to minimize PCR and sequencing errors, we obtained 355,306 high-quality 16S sequences, which were classified by comparison to the SILVA database in order to identify the sequences belonging to SAR11. The vast majority of such sequences corresponded to SAR11 clades I and II. Thus, all the subsequent analyses were focused on these two groups. The putative SAR11 sequences were submitted to further phylogenetic analyses in order to visually confirm the classification. After that, we finally obtained 95,318 high-quality SAR11 16S sequences. These sequences were accrued into Operational Taxonomic Units (OTUs) and analyzed using standard procedures. In addition, the data were submitted to phylogenetic analyses to characterize the spatial distribution of species across the phylogeny.

Spatial changes in biodiversity were quantitatively evaluated by analyses of the taxa-area relationships after grouping the sequences in identical, 99% and 97% OTUs. To this aim, we implemented a resampling scheme consisting of series of nested spherical caps. Starting at each of the 11 locations studied, the rest of the samples were successively accrued, in order of proximity. Each time a new location was visited, the hemi-arch defined by the line joining the new location and the current center was used to calculate a spherical cap area, using the Earth radius. Concurrently, normalized samples were taken from each of the sites present in that area, and the corresponding OTU richness were determined. These analyses showed that taxa richness significantly increased as the sampled area did (Fig. 5).

Phylogenetic analyses of the *SWAO* data alone showed an heterogeneous geographic distribution of ribotypes, reflected by the clumping of the sequences from the different samples in the tree (Fig. 6). These analyses also showed that the *ChR* SAR11 community was much variable over time, but that this variation was overwhelmed by spatial biodiversity. In agreement with this, substantial degrees of endemism and temporal variation were evident after the *SWAO* sequences were grouped into 97% and 99% similarity OTUs (Fig. 7). Despite quite a few OTUs were shared between the samples that were taken along two years at *ChR* (Fig. 7B), only 276 out of 1639 99% OTUs and 105 out of 319 97% OTUs detected along this period at *ChR* were also detected at other *SWAO* locations (Fig 7A).

To characterize the distribution of ribotypes across the phylogeny at a global scale, we first identified all the sequences corresponding to 100% similarity OTUs that were present just in a single location (that is, *endemic* OTUs), and mapped these data on a Maximum Likelihood phylogeny (Fig. 8). These analyses showed that endemic sequences (red dots in Fig. 8) tended to form clumps along the tree, that is that the ribotypes distribution and phylogeny covary, as can be expected under *DL* (Table 2). Using the scale drawn in the lower right corner of Fig. 8 (indicating the span of 1000 tree terminals), it can be appreciated that extensive parts of the phylogeny were unrepresented even in the more densely sampled locations. As observed among the *SWAO* samples, endemic OTUs were abundant (heights of the red-colored bars of the barplots in Fig. 8), as could be expected under *DL* (Table 2). Similar patterns were observed using Parsimony analyses (please see quantitative analyses below).

In order to quantitatively assess the covariation between the ribotypes' distribution and phylogeny, we performed distance decay analyses of phylogenetic distances obtained from both Maximum Likelihood and Parsimony trees. To cope with phylogenetic uncertainty, the phylogenetic distances were averaged from both 100 Maximum Likelihood bootstrap trees and 100 Parsimony trees obtained as described in Section 2.5.2. A previous study have suggested that SAR11 diversity is related to latitude and temperature (Brown et al., 2012). Thus, we used remoteness, temperature and latitude as covariates in graphical (Fig. 9), Mantel tests and dbLM analyses. When considered in isolation, the three covariates were significantly associated with pairwise phylogenetic distances, though the correlation and the amount of variation explained were greater for remoteness than for latitude and temperature (not shown). However, when the three explanatory variables were combined in partial Mantel tests, which allow to assess the relationship between two variables while controlling for the effects of one or more extra variables, only remoteness was significantly correlated with phylogenetic distance (Table 3). Likewise, when the three explanatory variables were combined in a single dbLM, the model still explained a substantial fraction of the response variable variation, but remoteness was the only significant predictor (Table 3). We obtained similar results using the optimal Maximum Likelihood and Parsimony trees (Fig. S1; Table S1).

Hellweger et al. (2014) used a neutral agent-based model to create an atlas of neutral biogeography of the ocean surface. Thus, we finally evaluated the correspondence between the phylogenetic distances obtained here and the divergences predicted by Hellweger et al.'s neutral agent-based model. These analyses showed that both the mean and maximum nucleotide divergences predicted by the neutral model were significantly correlated with the observed phylogenetic distances (Fig. 10; Table 4). In agreement with this, dbLM analyses showed that both the mean and maximum expected nucleotide divergences significantly explained the phylogenetic distances obtained here (Table 4). Equivalent results were obtained using optimal Maximum Likelihood and Parsimony trees (Fig. S2; Table S2).

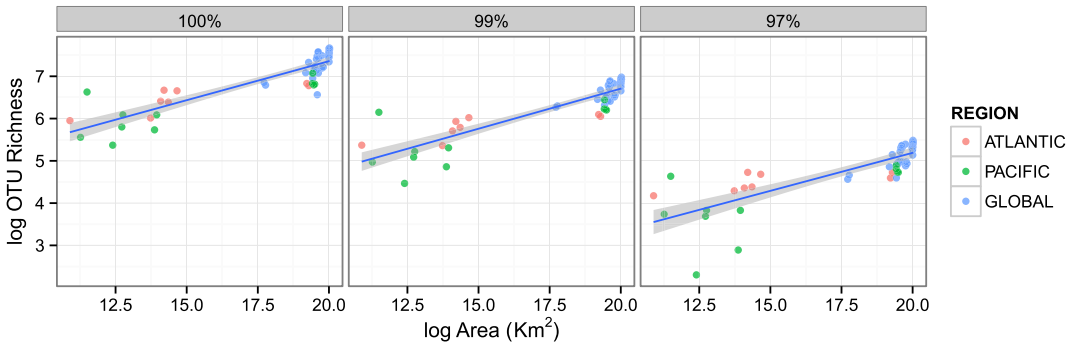
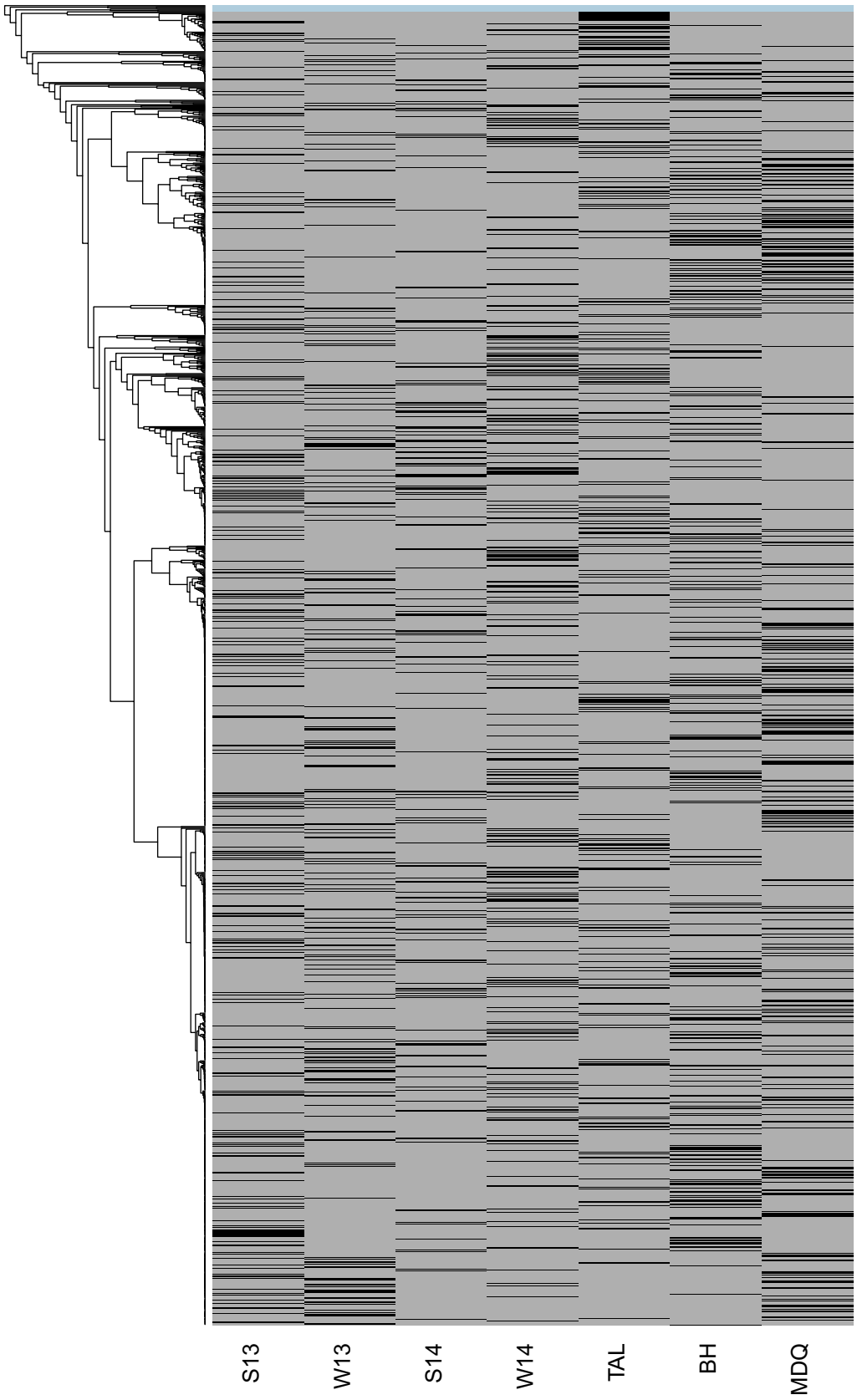


Fig. 5. Taxa-area relationships of SAR11 communities. Worldwide taxa-area relationships for 100, 99 and 97% similarity OTUs are displayed. The lines and shaded confidence regions denote standard regression analyses (100%: b = 0.18, $p < 2e{-}16$, $R^2 = 0.77$; 99%: b = 0.18, $p < 2e{-}16$, $R^2 = 0.77$; 97%: b = 0.17, $p = 1.86e{-}13$, $R^2 = 0.64$). The data were rarified previous to the analyses to balance the sequencing depths (please see the text for details).

5. Discussion

The ubiquitous SAR 11 clade of marine Alphaproteobacteria contributes about a quarter of the surface seawater bacterioplankton. To better understand the group's evolutionary dynamics, we performed a quantitative phylogeographic analysis of 95,318 high-quality 16S sequences from surface seawater collected at 11 worldwide distributed sites. These analyses revealed



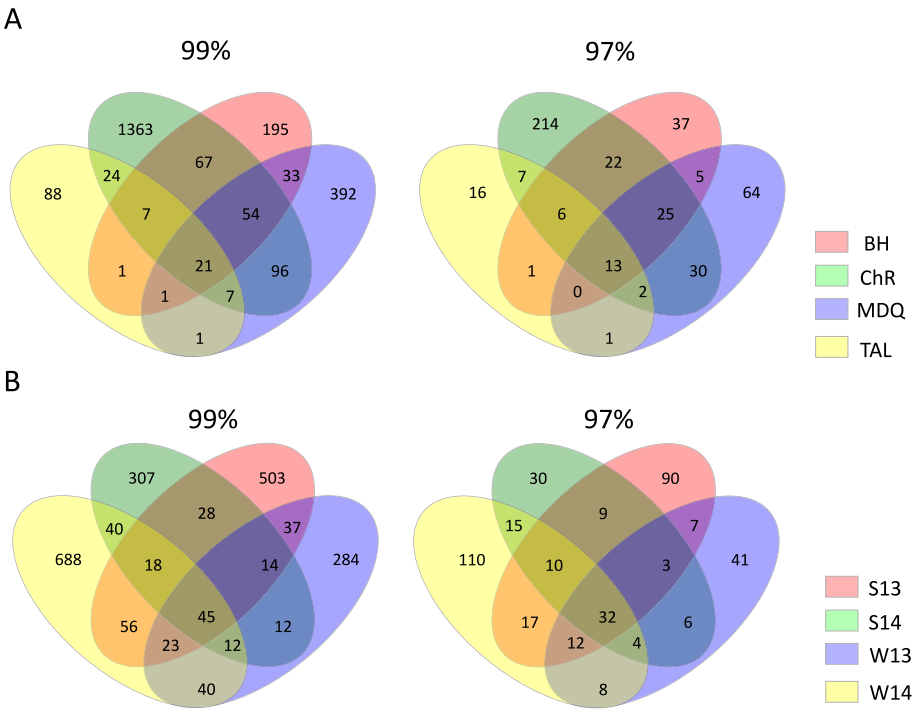


Fig. 7. Distribution of SAR11 97 and 99% similarity OTUs among the South West Atlantic Ocean samples. Panel *A* displays the distribution of OTUs among individual sampling sites, whereas panel *B* shows data corresponding only to the Chubut River (*ChR*) estuary (*S13*, *W13*, *S14*, *W14*). Samples *S13* and *S14* were taken during the austral summer on years 2013 and 2014, respectively. Samples *W13* and *W14* were taken during the 2013 and 2014 austral winter. The *ChR* OTUs (panel A) were obtained from the pooled *S13*, *S14*, *W13* and *W14* reads. High rates of endemism, revealed by the presence of many unique OTUs in all the samples, can be appreciated from panel *A*. The data in panel *B* shows that temporal variations are large regardless the similarity cutoff used to group the sequences. Notwithstanding that, comparison to panel *A* indicates that this variability is overwhelmed by spatial biodiversity, as many 99% and 97% OTUs from *ChR* were not detected among the *MDQ*, *TAL* and *BH* samples (1363 and 214, respectively).

substantial levels of endemicity and the existence of significant taxa-area relationships after grouping the sequences at either identical, 99% and 97% similarity OTUs. Distance decay analyses of phylogenetic distances between the different SAR11 communities revealed a significant covariation between the group's phylogeny and distribution, compatible with *DL*. In addition, the phylogenetic distances between the samples studied here were correlated with the divergences predicted by a neutral agent-based model. To the best of our knowledge, this is the first study using *HTS* combined with phylogenic analysis to untangle global bacterial phylogeography. Furthermore, there are no previous reports of distance decay patterns of phylogenetic distances among marine bacteria. Our results support the concept that *DL* are strong enough to allow for evolutionary drift of SAR11 populations, in agreement with previous simulation studies (Hellweger et al., 2014).

As we were interested in incorporating the SAR11 phylogeny in our models, we used sequences from a region of the 16S gene that was previously determined to be the optimal region of the gene to use for phylogenetic analysis from pyrosequencing reads (Liu et al., 2007). However, a potential problem of our approach is that the sequences of some of the locations investigated were amplified and sequenced differently, which could give raise to two difficulties. First, the sequences studied don't span exactly the same region of the gene but present slightly different degrees of overlapping. Second, although it wasn't studied specifically for the case of SAR11 bacteria, the use of different primers can result in amplification biases. Regarding the first issue, we treated gaps as missing data. In this way, gapped regions have a minimal influence on positioning the sequences in the phylogeny. Previous studies have shown that the decrease in resolution and accuracy due to inclusion of incomplete taxa is mostly associated with sequences displaying too few complete characters (Wiens, 2003). By the other side, it has been demonstrated that adding incomplete data generally results in an increase and, in extreme cases, only a slight decrease of phylogenetic accuracy (Jiang et al., 2014). This is to say that excluding incomplete taxa isn't necessarily a conservative approach, as it could actually result in a lack of phylogenetic resolution, and no impact on accuracy. By the other side, it is very unlikely for missing data to induce significant statistical associations such as those observed here between phylogenetic distances and remoteness (Fig. 9; Table 3) and the correlation between the phylogenetic distances measured here and the divergence values predicted independently by Hellweger et al.'s (2014) neutral model (Fig. 10; Table 4). In fact, analyses of the *SWAO* sequences alone, which were generated using the same primers and in a single 454 run, revealed the same patters as the whole dataset did (Figs. 6 and 7). The problem that could be associated with potential primer biases is that samples amplified with the same primer pair could

Fig. 6. SAR11 ribotypes distribution among the South West Atlantic sites. Each tree terminal corresponds to a row of the presence/absence matrix drawn to the right (presence: black; absence: gray). Each of the matrix columns correspond to a location, or sampling time in the case of the Chubut River estuary, as indicated in the bottom of the figure: *S13*, *W13*, *S14* and *W14* Chubut River (*ChR*) estuary; *MDQ*, *TAL*, *BH* oceanic samples. The *ChR* samples were collected along two years during the austral summer (*S13*, *S14*) and winter (*W13*, *W14*) of 2013 and 2014, respectively. The cells shaded in light blue on top of the matrix correspond to clades III and IV reference sequences, used to root the tree. Sequencing depths were normalized previous to phylogenetic analysis. Branch lengths were set arbitrarily to facilitate data display. The clumping of sequences in the tree according to the corresponding origins, that is the covariation between sequences' distribution and phylogeny, is supportive of the efficacy of dispersal limitations (Table 2; Section 3). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

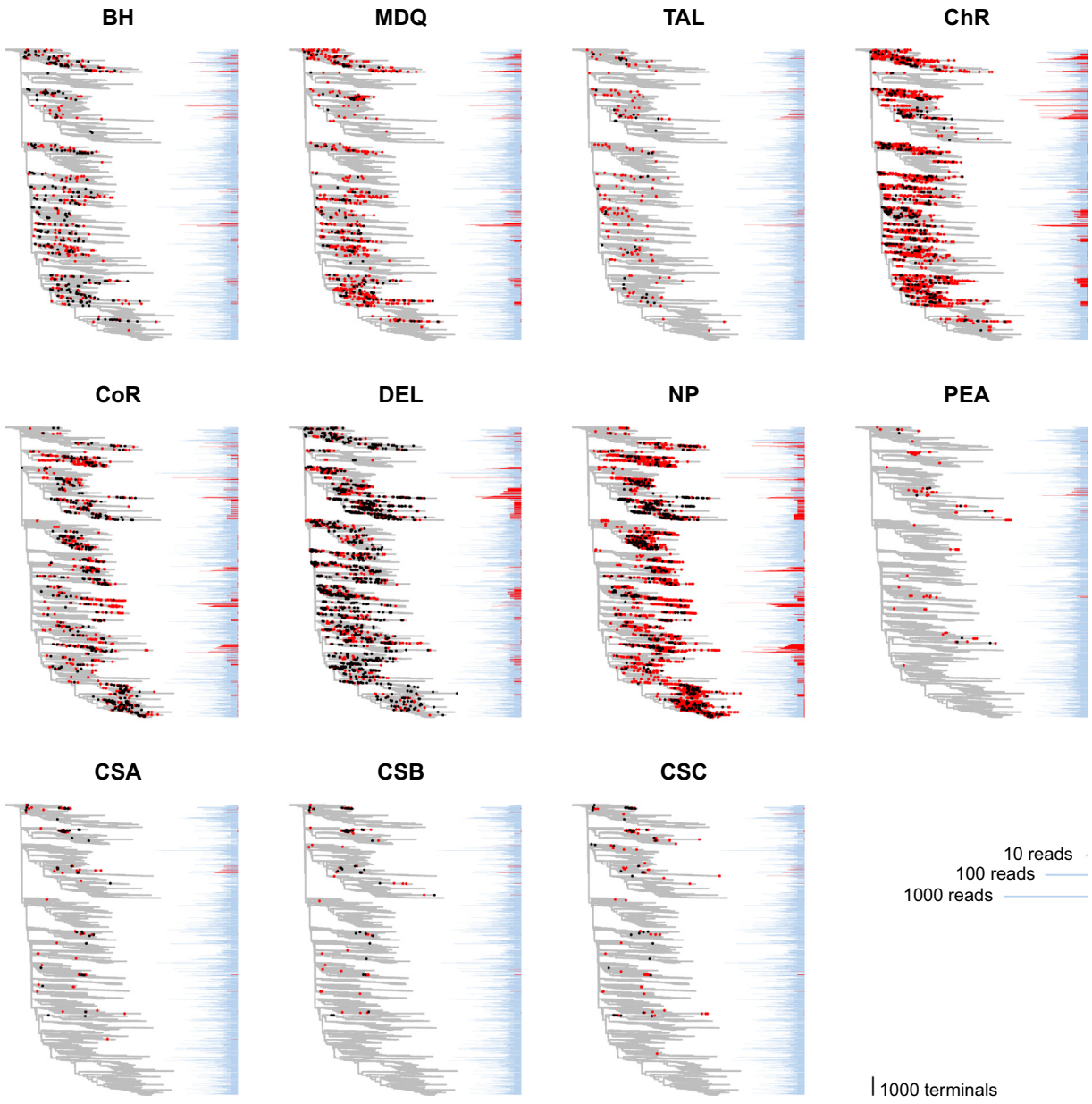


Fig. 8. Maximum likelihood phylogenetic analysis of 95,318 SAR11 sequences. Each panel depicts the distribution of the sequences from each location across the phylogeny. *ChR* Chubut River estuary; *MDQ*, *TAL*, *BH* oceanic samples from the South West Atlantic Ocean; *CoR* Columbia River Estuary; *NP* ocean coast close to CoR, *DEL* Delaware bay; *PEA* Pearl estuary; *CSA*, *CSB* and *CSC*, China Sea. The sequences present at each site are indicated by dots. Red dots correspond to endemic sequences, whereas black dots correspond to sequences that were observed in two or more locations. Sequence abundances, in logarithmic scale, are given by the barplots located to the right of the trees. The abundances of endemic OTUs are depicted in red. The horizontal lines at the bottom right corner depict bar heights equivalent to frequencies of 10, 100 and 1000 reads. Using these scale bars as references, it can be appreciated that endemic OTUs are very abundant. The upright bar at the bottom right corner indicates the vertical span of 1000 tree terminals. This scale bar is aimed to help appreciating that large parts of the tree are underrepresented in all the samples, reflecting the covariation of SAR11 distribution and phylogeny (Table 2; Section 3). Please see Fig. 9 and Table 3 for quantitative analyses.

be biased towards the same SAR11 lineages. The data studied here were produced using four primer pairs; however, when all de phylogenetic distances obtained from samples sequenced with different primer pairs were removed from the analyses, we still observed a significant association between remoteness and relatedness (Fig. S3).

The existence of functional relationships between the number of species of plants and animals in an area and the size of that area has been known for long (Drakare et al., 2006). Recently, this ecological law was demonstrated to hold for bacteria (Horner-Devine et al., 2004). Taxa-area relationships for marine bacterial communities, similar to those observed in this work, have been reported recently (Zinger et al., 2014). The slope coefficients (or z parameter of the taxa-area relationship) observed by us (0.17–0.18; Fig. 5) were small compared to the ones reported by Zinger et al. (2014), who obtained values in the range of about 0.2–0.4. A difference between this study and Zinger et al.'s is that Singer et al. used data from the V6 region of the 16S gene, which show a considerable sequence variability and can differ, regarding taxonomic information, from the gene region targeted by us (Liu et al., 2008; Wang et al., 2007). Additionally, the taxonomic resolutions used in both studies are very different. While Zinger et al. focused on whole bacterial communities, we focused on a single bacterial clade. Thought both approaches can capture habitat heterogeneity, one might

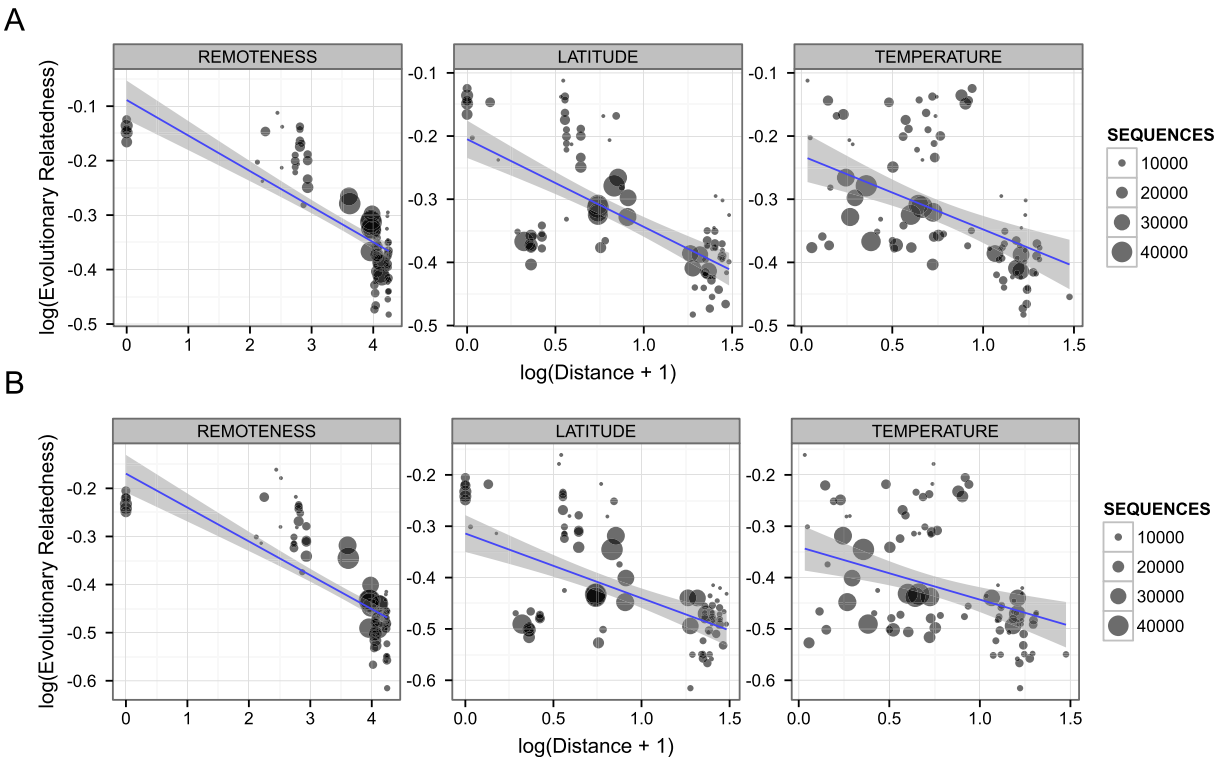


Fig. 9. Distance-decay analyses of phylogenetic distances calculated from Maximum Likelihood (A) and Parsimony (B) trees. Pairwise community relatedness (dots) were measured as 1 minus the corresponding *GUniFrac* distance. *GUniFrac* distances were averaged from 100 bootstrapped Maximum Likelihood trees (A) or a collection of 100 Parsimony trees (B). The dots' diameters are proportional to the number of sequences used to estimate the distances. The lines and shaded confidence regions denote standard linear regression analyses. Distance-based regression and Mantel test analyses revealed significant associations only for remoteness (Table 3). As detailed in Section 3 and summarized in Table 2, these analyses corroborate that SAR11 distribution and phylogeny covary, thus supporting the concept that *DL* play a role in SAR11 diversification.

Table 3
Distance decay of phylogenetic distances obtained from Maximum Likelihood and Parsimony trees.

Test	Covariate	Likelihood		Parsimony	
		ρ/R^2	*p*	ρ/R^2	*p*
Partial Mantel	Remoteness	0.80	1e−4	0.80	1e−4
	Temperature	−0.06	0.58	−0.04	0.66
	Latitude	0.03	0.76	−0.11	0.27
dbLM		0.75	–	0.77	–
	Remoteness	–	1e−3	–	1e−3
	Temperature	–	0.60	–	0.65
	Latitude	–	0.75	–	0.28

expect the whole community to be more sensitive to habitat-niche convergence of taxonomic entities across the whole bacterial phylogeny. By the other side, the low slope values observed here could obey to a geographical range effect (Drakare et al., 2006). It is worth mentioning that the existence of differences between the slope of taxa-area relationship of distinct taxonomic groups is indicative of differences in sensitivity to niche or potential range loss (Drakare et al., 2006). Thus, that the *z* parameters observed for whole bacterial communities (Zinger et al., 2014) are greater than the SAR11 ones suggests that SAR11 lineages might be relatively resilient. Finally, also noteworthy is that the *z* parameters reported here were almost identical regardless the similarity threshold used to define SAR11 OTUs (Fig. 5). Thus, the differences between populations that are expected due to drift along a relatively short period of time, *i.e.* those observed at the greater similarity cutoffs, seem to have a very similar effect on diversity patterns than the processes that have been shaping the group's diversity on the long term; that is, the patterns of very broad taxa at 97% sequence similarity.

As mentioned in the Introduction section, evidence is accumulating that *DL* can generate and maintain significant biogeographic patterns in bacteria, although empirical corroboration was missing so far for ubiquitous and profuse groups such as SAR11. The taxa-area patterns and distance decay relationships observed here (Figs. 5 and 9) are compatible with the concept that *DL* do have a role in SAR11 radiation in surface ocean. Thus, this study provides the first empirical evidence that SAR11 lineages can be subjected to *DL*, in agreement with recent simulations efforts (Hellweger et al., 2014). It could be argued that if the environmental variables relevant to SAR11 thriving would change constantly and monotonically with distance, and assuming that phylogenetic relatedness were proportional to similitude in environmental preferences, these same patterns would also be expected if SAR11 populations were tracking their environment. In addition, the clumping of endemic OTUs along the phylogeny (Fig. 8), could be compatible also with processes of habitat-niche convergence of species, provided that ecological traits were phylogenetically clustered (Gerhold et al., 2015; Webb et al., 2002). However, we consider this

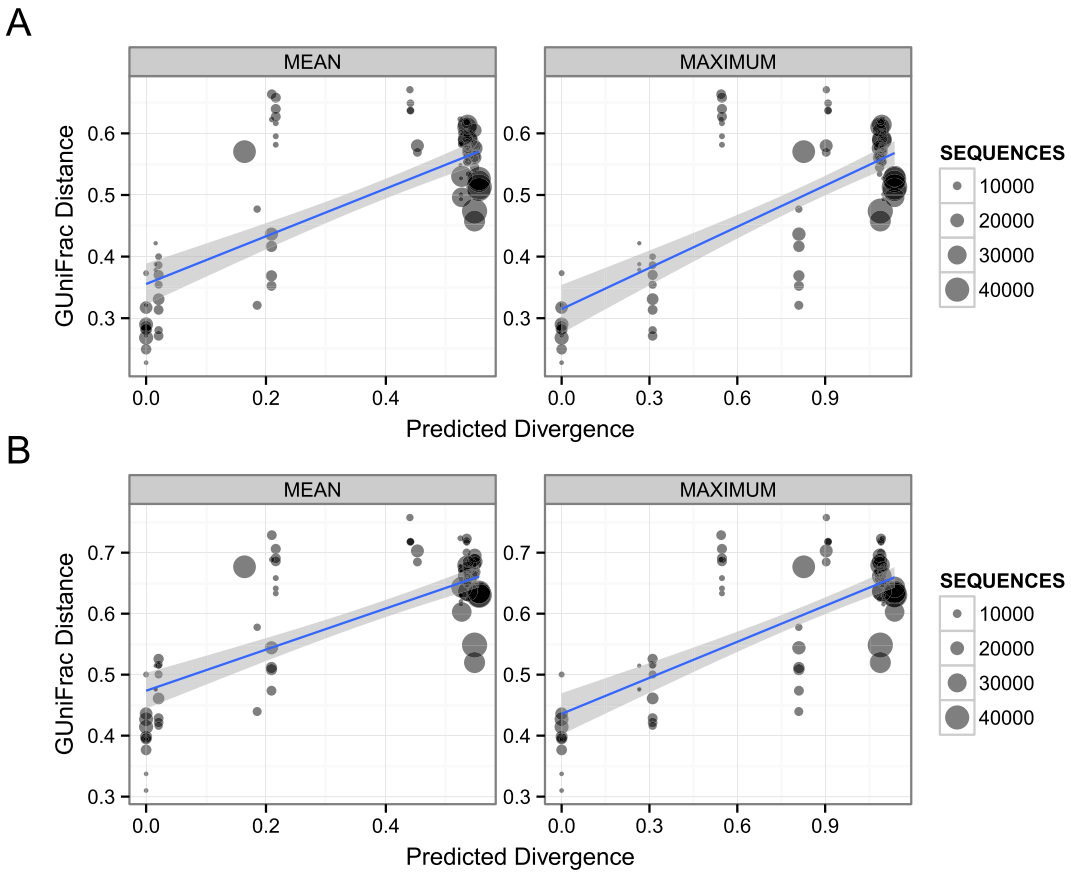


Fig. 10. Correspondence between predicted and observed divergences between SAR11 communities. Pairwise phylogenetic distances (*y*-axis) were estimated by the *GUniFrac* metric, averaged from 100 bootstrapped Maximum Likelihood trees (A) or a collection of 100 Parsimony trees (B). The expected divergences (*x*-axis) correspond to the mean and maximum divergences predicted by Hellweger et al.'s (2014) neutral agent-based model in the global surface ocean. Mantel tests and distance-based linear model analyses revealed significant correspondences between both *ML* and Parsimony distances and the expected divergences (Table 4). These analyses show that SAR11 diversity can be predicted by neutral models, thus offering empirical support to the concept that ocean currents are too slow to counteract evolutionary drift.

Table 4
Correspondence between predicted divergences and phylogenetic distances obtained from Maximum Likelihood and Parsimony trees.

		Likelihood		Parsimony	
		ρ/R^2	*p*	ρ/R^2	*p*
Mantel	Mean[a]	0.77	1e−4	0.79	1e−4
	Maximum[b]	0.76	1e−4	0.79	1e−4
dbLM	Mean	0.59	1e−3	0.63	1e−3
	Maximum	0.57	1e−3	0.63	1e−3

[a] Mean divergence predicted by a neutral model (Hellweger et al., 2014).
[b] Maximum divergence predicted by a neutral model (Hellweger et al., 2014).

possibility as very unlikely. Assuming that environmental variables change constantly and monotonically is equivalent to assume that no places along the ocean surface will offer similar niches and also that any two opposite points on Earth would present the greatest possible environmental differences, which is obviously incorrect. In addition, it is implausible that only phylogenetic inertia determines niche preferences in SAR11, due to the abundance and profusion of the group. By the other side, Brown et al. (2012) have observed a disparate distribution of SAR11 phylotypes defined based on *ITS* sequences. They attributed this distribution to niche preferences, and showed that temperature and latitude were the covariates that better explained the relative abundance of each phylotype. However, Brown et al. (2012) neither corrected for the phylogeny in their models nor evaluated the effect of *DL*. When we combined latitude, temperature and remoteness in a dbLM and Partial Mantel analyses aimed to assess the relationship of these covariates and phylogenetic distances between SAR11 communities, only remoteness resulted to be a significant explanatory variable (Table 3). We attribute these differences to the different experimental designs and model implementations used in both studies. Ignoring the phylogeny and the potential role of *DL* can result in misleading conclusions. For example, that a monophyletic phylotype be present exclusively in both Antarctic and Arctic waters can be interpreted as the phylotype having a distribution conditioned by ecological niche differentiation. However, the phylotype's distribution and phylogeny can still covary due to *DL*, from a mild case in which some variants inside the phylotype's clade are under- or overrepresented in either the Arctic or Arctic regions, to an extreme in which the Arctic and Antarctic sequences are clustered in two separate tree branches inside the phylotype's clade.

This can be tackled only implementing models that include phylogenetic evidence and explicit quantitative phylogeographic approaches. Thus, despite Brown et al. (2012) data suggests a relationship between ITS phylotypes and the environment, further research is required to weigh the importance of evolutionary drift in the phylotypes' radiation.

As mentioned above, we interpret our results as supporting the concept that *DL* contribute to SAR11 diversification in the ocean surface. A potential implication of the phenomenon is that the spatial scale could be relevant for SAR11 studies since everything isn't everywhere and thus results obtained at a given location may require further research to be extrapolated to distant places. Also worth of noticing is that, as explained in Section 3, this scenario requires that the different SAR11 genetic variants were capable of thriving under many ecological conditions. The ability to thrive in disparate marine environments could rely on properties such as proteorhodopsin phototrophy (Steindler et al., 2011) and the capabilities of maintaining extracellular buffers (Zubkov et al., 2015) and growing on osmolytes produced by other organisms (Lidbury et al., 2014). In addition, recent studies have shown that transcription in *Pelagibacter ubique*, a cultivated representative of the SAR11 clade, appears to be controlled by factors other than the environment (Cottrell and Kirchman, 2016), suggesting that SAR11 metabolism is resilient to environmental variation. Together, these and our data suggest that the ocean surface may constitute a homogeneous niche for these microorganisms, which could consist for example in certain ranges of solar radiation conditions, CO_2 pressures and temperatures, as well as particular interactions with other organisms. If the hypothesis is correct, and considering the importance of oceanic bacteria in processes of biogeochemical magnitude, global-change induced impacts on the SAR11 niche might jeopardize the thriving of the lineage resulting in unpredictable ecological consequences of a global reach.

Acknowledgements

The authors are members of the National Council of Scientific and Technological Research (CONICET). The data reported here was deposited in GenBank under accession numbers SRR3176906, SRR3180667, SRR3180668, SRR3180669, SRR3180670, SRR3180671 and SRR3180683. This work was partially supported by Asociación Civil Argentina Genetics (ArGen) and grant PRH 120 from the National Agency for the Promotion of Science and Technology of the Ministry of Science, Technology and Productive Innovation (MINCyT). The R/V "Coriolis II" expedition was a joint initiative of MINCyT, the University of Quebec and CONICET.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2016.11.015.

References

Aylward, F.O., Eppley, J.M., Smith, J.M., Chavez, F.P., Scholin, C.A., DeLong, E.F., 2015. Microbial community transcriptional networks are conserved in three domains at ocean basin scales. Proc. Natl. Acad. Sci. USA 112, 5443–5448.

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. Mol. Syst. Biol. 8, 595.

Campbell, B.J., Yu, L., Heidelberg, J.F., Kirchman, D.L., 2011. Activity of abundant and rare bacteria in a coastal ocean. Proc. Natl. Acad. Sci. USA 108, 12776–12781.

Cottrell, M.T., Kirchman, D.L., 2016. Transcriptional control in marine copiotrophic and oligotrophic bacteria with streamlined genomes. Appl. Environ. Microbiol. 82, 6010–6018.

Chen, J., Bittinger, K., Charlson, E.S., Hoffmann, C., Lewis, J., Wu, G.D., Collman, R.G., Bushman, F.D., Li, H., 2012. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics 28, 2106–2113.

Drakare, S., Lennon, J.J., Hillebrand, H., 2006. The imprint of the geographical, evolutionary and ecological context on species-area relationships. Ecol. Lett. 9, 215–227.

Fortunato, C.S., Eiler, A., Herfort, L., Needoba, J.A., Peterson, T.D., Crump, B.C., 2013. Determining indicator taxa across spatial and seasonal gradients in the Columbia River coastal margin. ISME J. 7, 1899–1911.

Gerhold, P., Cahill, J.F., Winter, M., Bartish, I.V., Prinzing, A., 2015. Phylogenetic patterns are not proxies of community assembly mechanisms (they are far better). Funct. Ecol. 29, 600–614.

Giovannoni, S., Tripp, H., Givan, S., Podar, M., Vergin, K., Baptista, D., Bibbs, L., Eads, J., Richardson, T., Noordewier, M., MS, R., Short, J., Carrington, J., Mathur, E., 2005. Genome Streamlining in a Cosmopolitan Oceanic Bacterium. Science 309, 1242–1245.

Goloboff, P., Catalano, S., 2016. TNT version 1.5, including a full implementation of phylogenetic morphometrics. Cladistics 32, 221–238.

Goloboff, P., Catalano, S., Mirande, J., Szumik, C., Arias, J., Källersjö, M., Farris, J., 2009. Phylogenetic analysis of 73060 taxa corroborates major eukaryotic groups. Cladistics 25, 211–230.

Goslee, S., Urban, D., 2007. The ecodist package for dissimilarity-based analysis of ecological data. J. Stat. Softw. 22, 1–19.

Hellweger, F.L., van Sebille, E., Fredrick, N.D., 2014. Biogeographic patterns in ocean microbes emerge in a neutral agent-based model. Science 345, 1346–1349.

Horner-Devine, M.C., Lage, M., Hughes, J.B., Bohannan, B.J., 2004. A taxa-area relationship for bacteria. Nature 432, 750–753.

Jiang, W., Chen, S.Y., Wang, H., Li, D.Z., Wiens, J.J., 2014. Should genes with missing data be excluded from phylogenetic analyses? Mol. Phylogenet. Evol. 80, 308–318.

Jones, L.R., Manrique, J.M., 2015. The peril of PCR inhibitors in environmental samples: the case of *Didymosphenia geminata*. Biodivers. Conserv. 24, 1541–1548.

Katoh, K., Standley, D.M., 2014. MAFFT: iterative refinement and additional methods. Methods Mol. Biol. 1079, 131–146.

Legendre, P., Legendre, L., 2003. Numerical Ecology. Elsevier Science B.V, Amsterdam.

Lidbury, I., Murrell, J.C., Chen, Y., 2014. Trimethylamine N-oxide metabolism by abundant marine heterotrophic bacteria. Proc. Natl. Acad. Sci. USA 111, 2710–2715.

Liu, Z., Lozupone, C., Hamady, M., Bushman, F.D., Knight, R., 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res. 35, e120.

Liu, Z., DeSantis, T.Z., Andersen, G.L., Knight, R., 2008. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Res. 36, e120.

Liu, J., Fu, B., Yang, H., Zhao, M., He, B., Zhang, X.H., 2015. Phylogenetic shifts of bacterioplankton community composition along the Pearl Estuary: the potential impact of hypoxia and nutrients. Front. Microbiol. 6, 64.

Manrique, J.M., Calvo, A.Y., Halac, S.R., Villafane, V.E., Jones, L.R., Walter Helbling, E., 2012. Effects of UV radiation on the taxonomic composition of natural bacterioplankton communities from Bahia Engano (Patagonia, Argentina). J. Photochem. Photobiol., B 117, 171–178.

Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27, 209–220.

Martiny, J.B., Eisen, J.A., Penn, K., Allison, S.D., Horner-Devine, M.C., 2011. Drivers of bacterial beta-diversity depend on spatial scale. Proc. Natl. Acad. Sci. USA 108, 7850–7854.

Morris, R., Rappé, M., Connon, S., Vergin, K., Siebold, W., Carlson, C., Giovannoni, S., 2002. SAR11 clade dominates ocean surface bacterioplankton communities. Nature 420, 806–810.

Nylander, J.A., 2004. Program Distributed by the Author. Evolutionary Biology Centre, Uppsala University.

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS ONE 5, e9490.

Quince, C., Lanzen, A., Davenport, R.J., Turnbaugh, P.J., 2011. Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12, 38.

Schloss, P.D., Gevers, D., Westcott, S.L., 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS ONE 6, e27310.

Steindler, L., Schwalbach, M.S., Smith, D.P., Chan, F., Giovannoni, S.J., 2011. Energy starved Candidatus Pelagibacter ubique substitutes light-mediated ATP production for endogenous carbon respiration. PLoS ONE 6, e19725.

Sun, F.L., Wang, Y.S., Wu, M.L., Jiang, Z.Y., Sun, C.C., Cheng, H., 2014. Genetic diversity of bacterial communities and gene transfer agents in northern South China Sea. PLoS One 9, e111892.

Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73, 5261–5267.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., Barton, G.J., 2009. Jalview Version 2 – a multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189–1191.

Webb, C.O., Ackerly, D.D., McPeek, M.A., Donogue, M.J., 2002. Phylogenies and community ecology. Annu. Rev. Ecol. Syst. 33, 475–505.
Whitaker, R.J., Grogan, D.W., Taylor, J.W., 2003. Geographic barriers isolate endemic populations of hyperthermophilic archaea. Science 301, 976–978.
Wiens, J.J., 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52, 528–538.
Zhao, X., Wan, X., He, R.L., Yau, S.S., 2016. A new method for studying the evolutionary origin of the SAR11 clade marine bacteria. Mol. Phylogenet. Evol. 98, 271–279.
Zinger, L., Boetius, A., Ramette, A., 2014. Bacterial taxa-area and distance-decay relationships in marine environments. Mol. Ecol. 23, 954–964.
Zubkov, M.V., Martin, A.P., Hartmann, M., Grob, C., Scanlan, D.J., 2015. Dominant oceanic bacteria secure phosphate using a large extracellular buffer. Nat. Commun. 6, 7878.