

A drift-barrier model drives the genomic landscape of a structured bacterial population

Hélène Gardon, Corinne Biderre-petit, Isabelle Jouan-dufournel, Gisèle Bronner

▶ **To cite this version:**

Hélène Gardon, Corinne Biderre-petit, Isabelle Jouan-dufournel, Gisèle Bronner. A drift-barrier model drives the genomic landscape of a structured bacterial population. Molecular Ecology, 2020, 29 (21), pp.4143-4156. 10.1111/mec.15628 . hal-03428622

HAL Id: hal-03428622
https://hal.science/hal-03428622

Submitted on 15 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

MOLECULAR ECOLOGY

A drift-barrier model drives the genomic landscape of a structured bacterial population

Journal:	*Molecular Ecology*
Manuscript ID	MEC-20-0245.R3
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Gardon, Helene; Université Clermont Auvergne, CNRS, Laboratoire Microorganismes: Génome et Environnement Biderre-Petit, Corinne; Université Clermont Auvergne, CNRS, Laboratoire Microorganismes: Génome et Environnement Jouan-Dufournel, Isabelle; Université Clermont Auvergne, CNRS, Laboratoire Microorganismes: Génome et Environnement Bronner, Gisele; Universite Clermont Auvergne, CNRS, Laboratoire Microorganismes: Génome et Environnement
Keywords:	Bacterial genome diversity, Pangenome, Evolutionary mechanisms, Single-cell analyses, Prochlorococcus

SCHOLARONE™
Manuscripts

Title Page

A drift-barrier model drives the genomic landscape of a structured bacterial population.

Running title: Genome evolution of bacterial populations.

Authors/Affiliations

Hélène Gardon[a]*, Corinne Biderre-Petit[a], Isabelle Jouan-Dufournel[a], Gisèle Bronner[a]

[a] Université Clermont Auvergne, CNRS, Laboratoire Microorganismes: Génome et Environnement, F-63000 Clermont-Ferrand, France

*Corresponding author: Hélène Gardon

LMGE, UMR CNRS 6023

UNIVERSITE CLERMONT AUVERGNE, Campus Universitaire des Cézeaux

1 impasse Amélie Murat

63178 AUBIERE Cedex, France

Email: helene.gardon@uca.fr

For Review Only

Abstract

Bacterial populations differentiate over time and space to form distinct genetic units. The mechanisms governing this diversification are presumed to result from the ecological context of living units to adapt to specific niches. Recently, a model assuming the acquisition of advantageous genes among populations rather than whole genome sweeps has emerged to explain population differentiation. However, the characteristics of these exchanged, or flexible, genes and whether their evolution is driven by adaptive or neutral processes remain controversial. By analysing the flexible genome of single-amplified genomes of co-occurring populations of the marine *Prochlorococcus* HLII ecotype, we highlight that genomic compartments – rather than population units – are characterized by different evolutionary trajectories. The dynamics of gene fluxes vary across genomic compartments and therefore the effectiveness of selection depends on the fluctuation of the effective population size along the genome. Taken together, these results support the drift-barrier model of bacterial evolution.

Keywords

Bacterial genome diversity, pangenome, evolutionary mechanisms, single-cell analyses, *Prochlorococcus*

Introduction

The diversification of free bacterial species in the environment is assumed to result from their adaptation to specific ecological niches. However, the full understanding of the forces driving these differentiations also relies on evaluating their genome dynamics, in light of populational mechanisms such as selection, genetic drift and recombination. Based on Mayr's general species definition (Mayr, 1942), populations result from gene flow discontinuities within a species, leading to genetically cohesive units that can be distinguished according to their genome characteristics. However, this

For Review Only

definition of species hardly applies to bacteria as population boundaries remain elusive due to gene fluxes occurring even among distant relatives. Likewise, gene content variations of conspecific organisms, which gave rise to the concept of pangenome (Medini, Donati, Tettelin, Masignani, & Rappuoli, 2005; Tettelin et al., 2005), blur the genetic cohesion of the microbial population. Yet analyses based on comparative genomics have also suggested that high recombination rates lead to the exchange of advantageous genes within a bacterial population (Cadillo-Quiroz et al., 2012; Shapiro et al., 2012). These genes, rather than genomes, would sweep through the evolving population, leading to both its genetic cohesion and ecological differentiation at the species level. Genes acquired by horizontal transfer are indeed frequently reported as adaptive (McInerney, McNally, & O'Connell, 2017; Sela, Wolf, & Koonin, 2016). However, it has also been suggested that the distribution of flexible genes could be neutral (Baumdicker, Hess, & Pfaffelhuber, 2012). Furthermore, species with larger effective population size (*Ne*) have greater genetic diversity, and by extension a highly diverse pangenome (Andreani, Hesse, & Vos, 2017). As *Ne* affects the effectiveness of selection, it may impact the number of flexible genes that would be retained through selection (Bobay & Ochman, 2018).

In recent years, the tremendous progress of single-cell genomics (SCG) has greatly improved the sampling of coexisting subpopulations. This progress allowed the investigation of the factors that govern the diversification of the microbial genome structure and organization at a finer scale, such as for *Prochlorococcus marinus*. This cyanobacterium is one of the most abundant photosynthetic species in the ocean euphotic zone, responsible for up to 10% of the marine primary productivity (Flombaum et al., 2013; Partensky, Hess, & Vaulot, 1999). Its genetic diversity spans at least 12 distinct ecotypes (Biller, Berube, Berta-Thompson, Kelly, Roggensack, Awad, Roache-Johnson, Chisholm, et al., 2014; Kashtan et al., 2014; Malmstrom et al., 2010; Moore, Rocap, & Chisholm, 1998; Rocap et al., 2003), broadly separated into high-light (HL) and low-light (LL) ecotypes. All these ecotypes were shown to

For Review Only

contain different sets of functional genes and to adjust differently to environmental changes, suggesting a stable niche partitioning of ecologically distinct groups (Kent, Dupont, Yooseph, & Martiny, 2016; Larkin et al., 2016). From a large-scale SCG approach it was recently proposed that *Prochlorococcus* populations in the Atlantic Ocean are composed of hundreds of subpopulations resulting from an ancient niche partitioning (Kashtan et al., 2014) and that population differentiation was occurring among *Prochlorococcus* (Stolyar & Marx, 2019). Coexisting subpopulations showed a fine-scale sequence diversity, *i.e.*, a "genomic backbone" comprised primarily of within subpopulations core genes with distinct fixed alleles and several genomic islands (ISLs) mostly composed of flexible genes in combination with specific core alleles or shared among different backbones (Kashtan et al., 2014). This, in addition to the large population size and open pangenome of the *Prochlorococcus* genus, makes it a valuable taxon to study pan-genome evolution.

On the basis of the work of Kashtan et al. (2014), who suggest that variations in co-occurring subpopulations within the *Prochlorococcus* HLII ecotype are targeted on specific genome regions, we investigate the evolutionary underpinnings of bacterial genome differentiation among these subpopulations, with a focus on the flexible genome. By analysing synonymous *versus* nonsynonymous substitution rates (*dN*/*dS*) of single-amplified genomes (SAGs), we assess the nature and strength of selection on genomic compartments (core *versus* flexible, backbone *versus* ISLs). Overall, despite clear delineation of these subpopulations according to genome phylogeny, average nucleotide identity (ANI) analysis and content in flexible genes, we do not find significant differences in average *dN*/*dS* among clades. However, by analysing the evolutionary rates of clusters of orthologous genes (COGs), we demonstrate that ISLs are characterized by differences in selective pressures that shed light on different evolutionary trajectories. This variation in the efficacy of selection – associated with distinct sets of genes in specialized genomic compartments – could result from the fluctuating *Ne* along the genome.

For Review Only

Material and Methods

SAG datasets

In total, 87 SAGs of the marine cyanobacterium *Prochlorococcus* belonging to the HLII ecotype (Table S1) were examined to investigate the evolutionary dynamics of free-living bacteria. These SAGs were a subset of 96 SAGs collected at the Bermuda-Atlantic Times-series Study (BATS) site, during three samplings between November 2008 and April 2009 (Kashtan et al., 2014). The original set was reduced only to SAGs assigned to the seven phylogenetically delineated subpopulations distributed among three clusters defined at 98% identity of the ITS (Kashtan et al., 2014), *i.e.*, C1 to C5 within the cluster cN2, C8 within the cluster c9301 and C9 within the cluster cN1, and also excluded the contaminated SAG 518D8. The SAG sequences were downloaded from the National Center for Biotechnology Information (NCBI) (Table S2). Their assembly size ranged from 0.37 to 1.62 Mb, with an average GC content of 31.3%. We used CheckM (Parks, Imelfort, Skennerton, Hugenholtz, & Tyson, 2015) to estimate their completeness and contamination (Figure S1). Their completeness approximated 8.6 to 97.4%, with ~14% of SAGs being classified as partial (<50% of completeness; with an over-representation of SAGs from C1 clade (7 over 12)), ~ 45% as substantial (≥50 to 70% of completeness), ~19% as moderate (≥70 to 90% of completeness) and ~ 21% as near-complete (≥90% of completeness; fairly distributed over all clades, including those with a small number of SAGs). They all had less than 2.3% contamination (Figure S1B). Because of synteny of the HLII *Prochlorococcus* ecotype genomes (Yan et al., 2018) and to limit the complexity of the information, we used a reference genome in all analyses performed. We needed a reference genome that i) did not branch with any clades studied (that excludes two cultured strains, *i.e.* MIT9301, AS9601), ii) with a close relatedness with all clades iii) but not too much either (excluding the three most distant, *i.e.* MIT9107, MIT9123 and MIT9116, and those closest to MIT9301 and AS9601; (Kent

For Review Only

et al., 2019). Among the three remaining (*i.e.* MIT9302, MIT9311 and MIT9312), *P. marinus str.* MIT9312 (accession number ABB49062.1) was chosen because it historically denotes the eMIT9312 / HLII ecotype (Biller, Berube, Berta-Thompson, Kelly, Roggensack, Awad, Roache-Johnson, Ding, et al., 2014). Its genome, 1.71 Mb in size and with an average GC content of 31%, contained 1,962 CDS and showed the presence of six ISLs scattered all along its genomic backbone (Avrani, Wurtzel, Sharon, Sorek, & Lindell, 2011) (Table S3).

Genome scale comparisons

The ANI allowed for the delineation of operational units at the genome level (Varghese et al., 2015). ANI was calculated both within and among subpopulations using the pyani package (Pritchard, Glover, Humphris, Elphinstone, & Toth, 2016). It was estimated by aligning fragments of 1,020 nt (Klappenbach et al., 2007) with BLASTN+ (Altschul, Gish, Miller, Myers, & Lipman, 1990; Camacho et al., 2009) and averaging the sequence identity between pairs of genomes.

Genome synteny analysis was performed on the MIT9312 reference genome and representative SAGs were selected for each subpopulation (the largest ones). The whole genome alignment and the detection of LCBs were generated using Mauve 2.4 with the progressiveMauve algorithm (Darling, Mau, & Perna, 2010) and default settings.

COG assignments

Overall, 7,125 COGs previously defined (Kashtan et al., 2015) were analysed. These COGs were determined by inferring pairwise homologous relationships using the method described by Kelly and colleagues (Kelly, Huang, Ding, & Chisholm, 2012). Briefly, they first assigned orthology relationships between genes using reciprocal best BLASTP hits (*e*-value ≤ 1e-5; sequence identity

>35%; alignment length >75% of the length of the shorter protein of the two compared) followed by transitively clustering orthologs together. They then built Hidden Markov Model (HMM) profiles (Eddy, 2009) of each cluster to integrate most divergent homologous genes missed by the BLAST approach. It is assumed that homologous relationships are transitive within a COG; thus, all genes from a cluster are homologous to any other gene in the cluster.

First of all, we looked at for the balance between completeness and number of SAGs, some of them having less than 50% completeness, which could affect the result of the analysis. As we showed that many SAGs, even incomplete, were more informative than a reduced set of complete SAGs (Figure S2), they were all considered in the subsequent analysis as well as all COGs. Of all the 7,125 COGs, 1,410 were identified as core (Table S4; Figures S3 and S4A), namely, common to the available genomes for cultured strains of the HLII ecotype (*i.e.*, MIT9311, MIT9314, MIT9401, MIT9301, MIT9312, MIT9107, MIT9201, MIT9321, MIT9202, MIT9215, SB, GP2, and AS9601), among which 1,397 were composed of single-copy genes. The remaining COGs were considered as flexible (5,715 COGs in total) and were either shared by some but not all cultured strains (11.35%) or specific to SAGs. Flexible COGs detected in at least one SAG but absent from the MIT9312 reference genome were assigned to a genomic compartment (*i.e.*, ISLs or genomic backbone) according to the location of the closest pair of genes referenced in MIT9312 that bounds these flexible COGs (Figure S3). We assumed that if two contiguous genes found in MIT9312 belonged to a unique compartment, the flexible genes between them also belonged to this compartment, otherwise they were classified as *ambiguous*. The compartment assignment was subsequently inferred at the COG level on a majority rule basis. However, since flexible COGs might contain genes located in different compartments, the Shannon entropy was computed i) to evaluate the variability of compartment assignments at the gene level within a COG (a higher Shannon entropy reflected a higher variability of the gene distribution in the different compartments for the concerned COGs) and ii) to assess the accuracy of compartment

For Review Only

assignment at the COG level compared with its genes (a lower Shannon entropy reflected a more representative location at the COG level). For each COG, entropy was calculated as follows:

$$H(X) = -\sum_{i=1}^{n} P_i \log_2 P_i \quad (1)$$

where *n* is the number of genomic compartments (ISL1; ISL2; ISL2.1; ISL3; ISL4; ISL5; backbone; *ambiguous*) and *Pi* is the proportion of genes arising from genomic compartment *i* within the COG.

Taxonomic affiliation and functional enrichments

For each core and flexible COG, functional and taxonomic annotations of the genes they contained were performed using BLASTP against the EggNOG v4.5 database (Huerta-Cepas et al., 2016). Only genes with a minimum length of 60 amino acids and hits with an *e*-value lower than 1e-5, a minimum alignment coverage of 50% and an identity of 30% were kept. Since all genes within a COG may not have a unique taxonomic affiliation, we defined a category called *uncertain,* which stands for COGs encompassing genes at least affiliated with *Prochlorococcus* and/or *Synechococcus* and with other bacterial taxa. For each COG, a preliminary allocation of their genes to the different EggNOG functional categories was also performed. Genes from the "Poorly characterized" category were discarded from the functional analyses. The gene functional enrichment was subsequently assessed by computing observed/expected (O/E) ratios of functional categories according to the genomic location or taxonomic affiliation of corresponding genes. The expected values were obtained by multiplying the number of genes (core or flexible genes as a function of their genomic location or taxonomic affiliation) by the percent of total genes in each functional category. The enrichments were tested through chi-squared tests.

Multiple sequence alignments and phylogenetic analysis

For each core and flexible COG, gene sequence alignments were performed at the amino acid level using MAFFT v7.271 (Katoh & Standley, 2013) (linsi option), and DNA sequences were imposed on the protein alignments (tranalign, EMBOSS v6.6.0.0) (Rice, Longden, & Bleasby, 2000). Gaps were deleted with Gblocks (Castresana, 2000) (allowing smaller final blocks with gap positions).

The trimmed alignments of single-copy core COGs found in both the MIT9312 reference genome and at least one SAG of each subpopulation were concatenated with missing sequences treated as gaps. A maximum likelihood tree was inferred with PhyML v3.0 (Guindon et al., 2010), using a GTR+I+G model of evolution, as determined by jModelTest v2.1.10 (Darriba, Taboada, Doallo, & Posada, 2012), and a bootstrap threshold of 100.

Substitution rates estimation

The nonsynonymous substitutions per nonsynonymous site (*dN*), the synonymous substitutions per synonymous site (*dS*) and their ratio (*dN/dS*) were estimated for all single-copy COGs common to MIT9312 (either core or flexible) and for flexible single-copy COGs not found in MIT9312 but common to at least two subpopulations. Nonsynonymous and synonymous substitution rates were calculated using the maximum likelihood method as implemented in codeml from PAML v4.8a (Yang, 2007; Yang & Nielsen, 2000). Maximum likelihood phylogenetic trees were computed for each COG with the GTR+G model as implemented in PhyML v3.0 (Guindon et al., 2010).

The COGs dataset was clustered according to *dN*, *dS* and *dN/dS* values using *k*-means clustering (Hartigan & Wong, 1979). The optimal number of clusters defined by the elbow method was five.

Results

Prochlorococcus co-occurring subpopulation phylogenetics

For Review Only

We analysed 87 SAGs of the HLII ecotype (Table S1) spread over three clusters, *i.e.,* cN2, c9301 and cN1 as defined by phylogenetic analysis of their ITS by Kashtan and colleagues (2014) (Kashtan et al., 2014). Using a whole-genome sequence phylogeny, they showed that these SAGs were distributed over seven major subpopulations (C1 to C5, C8 and C9), also referred to as clades throughout the paper. Despite their congruency with both phylogenies supporting the same subpopulation delineation, the tree based on whole-genome sequences did not follow the monophyly of the three clusters defined with ITS sequences. To reinforce these data, we inferred a maximum likelihood phylogeny from the concatenated alignment of 1,202 core genes, using the MIT9312 strain as the outgroup (Figure 1A). Our results confirmed the robust delimitation of clades (bootstrap values >80%) and the paraphyly of the cluster cN2 because of the C8 and C3 clustering (100% bootstrap support). Our genome-wide ANI analysis was in accordance with the phylogeny, depicting the same subpopulation demarcation (Figure 1B), with the inter-clade ANI being close to 94% on average when analysing all pairwise comparisons from C1 to C8. The highest identity was observed for the closest relatives C1 and C2 (97% ANI on average) whereas C9 was the most divergent with 90% ANI on average with other subpopulations (Figure 1B). This finding is consistent with its emergence as the most basal branch of our tree (Figure 1A). By comparison, intra-clade ANI was higher (>98%), except for C8 (97%) and C9 (96%) (Table S5). This is in accordance with a low intra-clade polymorphism and allele differentiation between clades (Kashtan et al., 2014).

Genome organization among subpopulations

The subpopulations were also investigated for their gene content and shared genomic regions, by aligning the genomic sequences of one representative SAG for each clade (the longest near complete SAG; Figure S1; Table S1), with the MIT9312 strain being used as the reference genome because of its equidistance to all SAGs investigated. Conserved segments of locally collinear blocks

For Review Only

(LCBs) detected in at least seven genomes represented approximately 76% (1.33 Mb) of the MIT9312 genome length. SAG alignments covered between 79 and 86% of the MIT9312 genome length (the alignment fraction for each SAG against MIT9312 was as follow: C1 - 495K23: 1.42 Mb; C2 - 498C16: 1.35 Mb; C3 - 518A17: 1.47 Mb; C4 - 528N17: 1.38 Mb; C5 - 498I20: 1.41 Mb; C8 - 527L22: 1.37 Mb; C9 - 258J8: 1.38 Mb) (Figure S5). Therefore, genomes showed relatively high synteny, consistently to what was reported within *Prochlorococcus* ecotypes (Yan et al., 2018), however with slight shifts in the locations of the six ISLs previously characterized in MIT9312 (Avrani et al., 2011) (Table S3; Figure S5). In light of this collinearity, 5,290 single-copy COGs absent from MIT9312 were assigned to the chromosomal compartments (backbone or ISLs). A compartment was inferred for each gene within COGs, and the majority compartment was assigned at the COG level. These assignments were robust, as less than 8.5% of the COGs had a Shannon entropy equal to or higher than one (*i.e.*, for which at least two compartments had substantial occurrence). However, as the compartment boundaries can be fuzzy, more specifically those of ISLs, the assignment of a few of these COGs should be taken with caution. Overall, 63.1% of the COGs were assigned to the backbone, 8.2% and 14.5% were allocated to ISL3 and ISL4, respectively, 8.9% were spread over ISL1, ISL2, ISL2.1 and ISL5, and the remaining 5.3% were tagged as *ambiguous*. The relative density of the assigned COGs was 2.5-fold higher in ISLs (5.8 COGs per Mb) than in the backbone (2.3 per Mb) on average, except for ISL4 (11.2 per Mb) and ISL2.1 (2.3 per Mb). COGs shared by several subpopulations (≥5) were enriched in all ISLs except ISL4, whereas those found in a single subpopulation were enriched in ISL3 and ISL4.

Taxonomic affiliation of COGs

Taxonomic analyses were performed to assess the phylogenetic origin of COGs as well as their integrity (*i.e.*, homogeneity of their gene affiliations). Regarding the core COGs (1,309 in total; Figure

For Review Only

2A; Table S4; Figure S4B), 97.4% were clearly affiliated with the HLII *Prochlorococcus* ecotype, with 2.9% of them containing genes affiliated with other ecotypes. The last 2.6% of COGs were tagged as *uncertain*, as they clustered genes with varying taxonomy, including *Prochlorococcus* and/or *Synechococcus* (Figure 2A).

In contrast, the phylogenetic origin of the 5,715 flexible COGs was less obvious since almost half of them contained genes with no counterpart in the EggNOG database. Regarding those with taxonomic affiliation (*i.e.*, 2,424 in total; Figure 2A; Table S4; Figure S4C), 76.5% were related to *Prochlorococcus*, among which 94.4% consisted of genes affiliated with the HLII ecotype, 3.3% with the HLI ecotype and 2.3% with the LL ecotypes. Among the remaining COGs, 6% were related to *Synechococcus*, 13% to bacterial taxa other than *Prochlorococcus* and *Synechococcus*, and 4.5% were *uncertain* (Figure 2A). When affiliated with other bacterial taxa, COGs belonging to *Proteobacteria* overdominated (51.6%), followed by *Cyanobacteria* (14.4%), the *Bacteroidetes/Chlorobi* group (10.9%), *Firmicutes* (5.3%), *Actinobacteria* (2.3%) and *Spirochaetes* (2.3%) (Figure 2B). COGs tagged *uncertain* contained genes affiliated with *Prochlorococcus* (39.7%), *Synechococcus* (16.4%) or both, either associated with other taxa (16.4%) or not (27.6%) (Figure 2C). Taxa other than *Prochlorococcus* and *Synechococcus* were mostly *Cyanobacteria*, followed by *Proteobacteria, Bacteroidetes/Chlorobi* and *Firmicutes* (Figure 2C).

Functional characterization of flexible COGs

The functional potential of COGs was investigated at the genomic compartment level. Only genes with known function were considered, which represented 47% of all core and flexible genes. Excepting the over-representation of the functional category “Cell motility” in clades C1 and C9, our results showed no difference in the distribution of functional categories at the subpopulation scale

For Review Only

(*p=0.39*, chi-squared test) (Figure S6A). Therefore, subpopulations were considered as a whole in subsequent analyses.

The distribution of the functional categories assigned to genes from flexible COGs was compared to those from core COGs (Figure 3A) and was also analysed depending on the genomic compartments (Figure 3B) and taxonomic affiliations (Figure 3C). Our results highlighted an overall under-representation of flexible genes in the hierarchical categories "Information storage and processing" and "Metabolism", primarily impacting the functional categories involved in the mechanisms of transcription and translation as well as those in the energy production and the transport and metabolism of nucleotides, amino acids, coenzymes and lipids. This is in accordance with the fact that these categories mainly group housekeeping genes. When found in flexible COGs, these categories were preferentially located in the backbone (Figure 3B). Conversely, two functional categories were over-represented as a result of their enrichment in the ISLs (Figure 3B), namely the categories "Secondary metabolites biosynthesis, transport and catabolism", enriched in ISL3, ISL5 and in the *ambiguous* compartment (with most genes annotated as methyltransferase, thus possibly involved in DNA repair), and "Inorganic ion transport and metabolism", enriched in ISL2 and ISL3 (mainly transporters of inorganic and organic phosphate). The flexible genes encoding these functional categories mostly belonged to *Proteobacteria* and *Archaea* (Figure 3C). Interestingly, although exhibiting an O/E ratio close to 1, the category "Carbohydrate transport and metabolism" was under-represented in ISLs and over-represented in the *ambiguous* compartment (with genes encoding protein such as transketolase and transaldolase, thus possibly linked to the Calvin cycle) (Figure 3B).

Regarding the hierarchical category "Cellular processes and signaling", genes associated with the flexible COGs were over-represented in most functional categories, except for "Cell cycle control, cell division and chromosome partitioning", where they were under-represented (Figure 3A). These over-representations were especially marked in the ISL3 (four out of seven functional categories) and

For Review Only

the backbone, particularly the functional category "Signal transduction mechanisms" (with genes encoding proteins such as histidine kinases involved in response to nutrient stress), the associated genes being mainly affiliated with *Prochlorococcus* (Figure 3C). Finally, two functional categories over-represented in flexible COGs compared to core COGs, *i.e.*, "Cell wall biogenesis" (mostly genes involved in the biosynthesis of outer membrane lipopolysaccharide protein such as glycosyltransferases and GDP-mannose 4,6-dehydratase) and "Defense mechanisms" (such as endonuclease and transporters), were found in COGs specific to a single subpopulation (Figure S6B). Most genes from these COGs (95.33%) belonged to ISL3, ISL4 and *ambiguous* compartments and were affiliated with a large variety of taxa (Figure 3C). We can notice that, despite the high proportion of SAGs in C1 compared to the one in other clades, which increased its weight in the pool of COGs specific to one clade, the same enrichments were observed in all clades (Figure S6A).

Heterogeneity of nonsynonymous (dN) to synonymous (dS) substitution rate ratios between genomic compartments

Since the F_{ST} analysis by Kashtan et al. (2014) suggested the fixation of different alleles among subpopulations that diverged at least few million years ago, it is therefore permissible to use the *dN/dS* approach to have deeper insights into evolution processes. In this context, to evaluate the selective pressure on COGs, we computed the ratios of *dN/dS* for genes in COGs recovered in at least two clades for which *dS* values could guarantee reliable estimates (Figure S7A). Thus, the C1 and C2 clade comparison was not considered because of their low mean *dS* (±sd 0.05 ±0.07) (Kryazhimskiy & Plotkin, 2008). Moreover, to rule out possible bias in *dN/dS* ratios computed for inter-clade comparisons (dos Reis & Yang, 2013; Rocha et al., 2006; Wolf, Künstner, Nam, Jakobsson, & Ellegren, 2009), we first compared the distribution of *dN/dS* ratios of core COGs from the backbone (1,139 in total) estimated from either the comparisons to MIT9312 or inter-clade analyses. Overall, the

absence of significant difference of mean *dN/dS* between the two analyses (mean *dN/dS* ±sd = 0.21 ±0.14 and 0.22 ±0.19, respectively; *p=0.35*, Wilcoxon test) associated with high similarity between the density plots (Figure 4A) and a strong positive correlation between per COGs mean *dN/dS* values (*ρ=0.83, p<0.001*, Spearman rank correlation) (Figure 4B) support no bias in the evaluation of selective pressure. Similar results were obtained for the flexible COGs shared with MIT9312 (not shown). As these two approaches provided comparable results, we used the inter-clade analysis because it allowed taking into account the flexible genes not shared with MIT9312 and was not impacted by the over-representation of C1 members and the great number of COGs specific to this clade (Figure S7C). Then we investigated the evolutionary patterns of genomic compartments.

Overall, all pairwise comparisons between clades showed similar distributions of *dN/dS* values (Figure S7) suggesting that i) these estimations are independent of clade abundances and ii) selective constraints are homogeneous among clades. Most *dN/dS* ratios were below 1, whatever the nature of the COGs (*i.e.*, core or flexible), suggesting that the selective pressure was essentially negative as also found by Kashtan et al. (2014). However, the mean *dN/dS* ratios were significantly different between COGs assigned as core (1,202 COGs), flexible shared (310 COGs) and not shared with MIT9312 (1,033 COGs) (*p<0.001*, Kruskal-Wallis test) (Table S4; Figures S4C and D), the former experiencing the stronger negative selection (Figure 4C). At the compartment level, significant differences between the *dN/dS* ratios were also revealed (*p<0.001*, Kruskal-Wallis test) (Figure 4D). Thus, core and flexible genes shared with MIT9312 showed strong homogeneous selective constraints in the backbone (mean *dN/dS* ±sd ranging from 0.19 ±0.22 to 0.23 ±0.20 and from 0.21 ±0.19 to 0.29 ±0.26, respectively), while they were more variable in ISLs. Conversely, flexible genes not shared with MIT9312 exhibited variable selective constraints, unevenly scattered along the backbone. The lowest mean *dN/dS* ±sd ratio (0.29 ±0.17) was observed in the region between ISL2 and ISL2.1, while the highest (0.73 ±0.82) was between ISL4 and ISL5 (Figure 4E).

For Review Only

The analysis of ISLs also revealed contrasting patterns. ISL1, ISL2 and ISL2.1 experienced negative selective pressures similar to what was observed for core genes in the backbone except for ISL2 flexible COGs shared with MIT9312 (Figures 4D and 4E). Nonetheless, the profile of the ISL2.1 flexible genes not shared with MIT9312 (mean *dN/dS* ±sd = 0.17 ±0.13) suggested a negative selective pressure stronger than the one observed in the backbone (Figure 4D). Conversely, ISL1 core genes and ISL2.1 flexible genes shared with MIT9312 showed weaker selective constraints (Figure 4D). However, these higher *dN/dS* values (0.35 ±0.37 and 0.38 ±0.62 on average ±sd, respectively) might be the result of a sampling bias because there was only one COG in ISL1 and three in ISL2.1. Furthermore, we observed reduced selective constraints in ISL3, ISL4 and ISL5 (ranging from (mean *dN/dS* ±sd) 0.36 ±0.35 for ISL3 to 0.52 ±0.62 for ISL5), except for core genes in ISL3 (mean *dN/dS* ±sd = 0.26 ±0.25) (Figures 4D and 4E). ISL4 and genes assigned as *ambiguous* exhibited by far the least constrained selective pressures (mean *dN/dS* ±sd from 0.44 ±0.13 to 0.52 ±0.23, respectively). Apparent reduced constraints on core and flexible COGs shared with MIT9312 in ISL4 might be hazardous to interpret as these values were sustained by few COGs (three core and 14 flexible).

Substitution rate signatures of genomic compartments

To investigate substitution rate signatures depending on genomic compartments, we assessed the links between *dN*, *dS* and *dN/dS* estimated among genes within COGs. For the backbone, both *dN* and *dS* varied widely among COGs, whether core or flexible. Although the *dS* rates varied up to >1.5, more than 95% of estimated values were less than 0.35 (mean *dS* ±sd = 0.100 ±0.091). By comparison, *dN* rates displayed lower values and variations (mean *dN* ±sd = 0.026 ±0.003). Additionally, the relationship of *dN/dS* ratios versus *dN* or *dS* was similar when comparing core and flexible COGs (Figures S8 and S9). Five clusters of genes were distinguished based on *k*-means clustering (Figure S10). Among them, three were characterized by low *dS* values (ranging from 0.001 to 0.268), low *dN*

values (<0.340) and *dN/dS* ratios varying from 0 to >1.5 (clusters yellow: <0.314; orange: 0.296 – 1.146; red: 1.133 – >1.5) (Figures S8 and S9). Here, high *dN/dS* ratios were associated with low *dS* rather than high *dN* rates, suggesting a general trend of negative selective pressure and background selection. The fourth cluster (green) was characterized by intermediary *dS* values (from 0.173 to 1.076) and low *dN* values (<0.326), reflecting more divergent sequences but still negative selective pressure. The last cluster (dark blue) showed *dS* values ranging from 0.252 to >1.5 and *dN/dS* ratios <1. Here, the *dN/dS* ratios seemed governed by *dN* values (from 0.172 to 1.277), as illustrated by the dots linearly spread around the major diagonal of *dN/dS* versus *dN* plot in Figures S8 and S9.

Regarding ISLs, the patterns of *dN* and *dS* variations among genes were contrasted, whether for core or flexible COGs. Genes distributed over the ISL1, ISL2 and ISL2.1 mostly had low and homogeneous *dN* and *dS* values (yellow, orange, red and green clusters), suggesting substantial negative selective pressure. Though their *dN/dS* ratios were essentially low, a few genes had values close to or greater than 1 (Figure S9). By comparison, genes in ISL4 or tagged as *ambiguous* were characterized by higher *dN* and *dS* values, with *dN/dS* ratios linearly linked with *dN,* and *dS* close to or at saturation (dark blue cluster). Genes located in ISL3 and ISL5, in contrast, displayed a mixed profile (all clusters).

Evolutionary signature of COGs depending on their distribution in co-occurring subpopulations

Because flexible COGs are not recovered in all clades, we investigated the behaviours of *dN*, *dS* and *dN/dS* ratios according to the COGs distribution among clades. Overall, COGs were characterized by genes of low to high *dS* values, except those shared by all clades, which were depleted in genes with saturated *dS* (Figure S11). *dN/dS* estimates differed significantly depending on the number of clades where COGs were found ($p<0.001$, Kruskal-Wallis test), *i.e.*, those shared by a substantial number of clades tended to display lower *dN/dS* values.

Up to 65.5% of flexible COGs analysed for *dN/dS* ratios were taxonomically assigned, with a higher proportion for those found in ISL4 (78.7% affiliated) compared to other ISLs (53.8% affiliated on average). COGs with low *dS* values were essentially affiliated with *Prochlorococcus* (Figure 5A) and were mostly found in ISL3 and ISL5. COGs with saturated *dS* were assigned as *uncertain* (*i.e.*, with multiple affiliations including *Prochlorococcus* or *Synechococcus)* (Figure 5B) and were located in ISL4 and the *ambiguous* compartment. COGs affiliated with "other bacterial phyla" (neither *Prochlorococcus* nor *Synechococcus*) were characterized by both low *dN* and *dS* suggesting the close origin of their shared genes and were enriched in ISL3 and ISL4 ($p<0.005$, chi-squared test).

Discussion

Bacterial species diversification to adapt to specific niches is documented not only for micro-organisms in the context of experimental conditions (Wiser, Ribeck, & Lenski, 2013) but also for some environmental bacteria (Kent et al., 2016; Larkin et al., 2016; Shapiro et al., 2012). In this study, we consider co-occurring SAGs of the *P. marinus* HLII ecotype, for which phylogenetic analyses of ITS (Kashtan et al., 2014) or concatenated single-copy core COGs (this study) depicted a structuring into clades that might reflect ancient niche partitioning. This finding is also supported by the predominance of different alleles in core genes that are fixed within subpopulations and distinct sets of flexible genes among them (Kashtan et al., 2014). Moreover, these subpopulations were characterized by variations in their relative abundance with seasonality (Kashtan et al., 2014) suggesting an adaptation; however, the drivers of this differentiation remain elusive (Larkin et al., 2016). Here, *dN/dS* values supported general negative selection among clades, in the same way as what was observed at the intra-clade level (Kashtan et al., 2014), while no positive selection was found. This finding is congruent with the analysis of prevalent species in the human gut, which suggested that positive selection, if present, may not overpower the signal of negative selection (Garud, Good, Hallatschek, & Pollard, 2019).

For Review Only

Furthermore, Marttinen and colleagues (Marttinen, Croucher, Gutmann, Corander, & Hanage, 2015) showed that a parsimonious model without niche or diversifying selection, but including recombination, induced structured populations within a stable range of genetic diversity. Thus, adaptation might not be necessary to explain the structuring of this HLII population into clades. To decipher between masked or absence of positive selection, similar analysis of *dN/dS* ratio could be performed on HLI ecotype subpopulations, for which a stronger correlation between seasonality and environmental factors was observed (Larkin et al., 2016).

Uneven negative selective constraints were detected on flexible genes along the backbone and among ISLs (Figures 4D and 4E), supporting allelic variations along the genome (Kashtan et al., 2014). Overall, ISL1, ISL2 and ISL2.1 showed COGs under strongest purifying selection, whereas ISL3, ISL4 and ISL5 tended to concentrate COGs with relaxed selection. ISL1, ISL2 and ISL2.1 were predominantly characterized by COGs shared by all clades and affiliated with *Prochlorococcus* (Figures S8 and S9), suggesting their ongoing "fixation". Moreover, the high F_{ST} values observed for COGs in ISL2 and ISL2.1 (Figure S12) suggest their potential role in stable niche partitioning. For instance, in ISL2, two COGs shared by all subpopulations and subject to strong selective constraints (means $dN/dS < 0.19$) were related to phosphonate utilization, which could suggest adaptation to low-phosphate environments (Feingersch et al., 2012). Recently, Schmutzer and Barraclough (Schmutzer & Barraclough, 2019) suggested that, in the presence of gene fluxes among diverging populations, the concentration of locally adapted genes in a reduced number of loci could be favoured, as it would i) reduce the negative impact of insertions along the genome of horizontally transferred genes and ii) increase the relative efficacy of selection on a few "mega" loci compared to many dispersed loci of reduced effect. Thus, ISL1, ISL2 and ISL2.1 might concentrate "flexible" genes that are essential for all these clades. They might, therefore, be considered as "core" in these clades, as proposed for ISL2.1 regarding the HLII ecotype (Avrani et al., 2011).

For Review Only

Selective signatures in the flexible genome also revealed two sets of COGs. A first set (consisting of the yellow, orange and red clusters), found in all genomic compartments (*i.e.*, backbone and ISLs), was characterized by low *dN* values and *dS* below the mean *dS* of core genes (Figures S8 and S9). Although this is consistent with general background selection (Price & Arkin, 2015), the low *dS*, driving an unusually high *dN/dS* (red cluster), could also reflect strong negative selection on synonymous substitutions (Parmley & Hurst, 2007) or homogenization of sequence diversity among clades through homologous recombination (HR) (Hanage, 2016). HR could increase selection efficacy by reducing the Hill-Robertson effect (Hill & Robertson, 1966). In a context of structured populations, inter-clade HR could also increase *Ne* for genes whose circulation would spread beyond sub-populations, while constraining clade differentiation within a divergence mode (Marttinen et al., 2015). This is in line with the combinatorial nature of backbone and flexible genes as proposed by Kashtan et al. (2014) for the high-light or phosphonate related genes. COGs in the second set (dark blue cluster), primarily found in ISL3, ISL4 and ISL5, displayed relaxation of selective constraints associated with high *dS,* suggesting horizontal gene transfer (HGT) (Castillo-Ramírez et al., 2011). Moreover, they were not affiliated with *Cyanobacteria* (Figure 5). HGT is recognized as a driver of evolution, contributing to the adaptation to changing environments through the expansion and conversion of gene families (Gogarten, Doolittle, & Lawrence, 2002; Ochman, Lawrence, & Groisman, 2000; Wiedenbeck & Cohan, 2011). Therefore, the over-representation of genes involved in defense mechanisms or cell wall biogenesis in ISL3 and ISL4 (Figure 3) could reflect the acquisition of genes that may be transiently adaptive during phage infection periods (Avrani et al., 2011; Coleman et al., 2006; Kettler et al., 2007). However, adaptive HGTs are primarily documented for genes with large selective advantage, which might not be true for most of them. In case of near neutral HGT, selection depends on *Ne* (Kuo & Ochman, 2009), with potentially different outcomes on the flexible genome. For

McInerney and colleagues (McInerney et al., 2017), as large *Ne* enhances selection efficacy, the flexible genome is essentially adaptive, with slightly deleterious acquired genes being eliminated. However, neutral evolution of the pangenome could not be rejected (Andreani et al., 2017; Baumdicker et al., 2012; Vos & Eyre-Walker, 2017). Additionally, in the drift-barrier model (Bobay & Ochman, 2018; Lynch, 2010), the loss of flexible genes is random for small *Ne*, while larger *Ne* would increase i) the proportion of genes with selection coefficient $s < 0$ that would be perceived as deleterious, ii) the fixation probability of slightly advantageous genes, and iii) the fixation time or loss of quasi neutral genes, and thus the size and diversity of the flexible genome. It was proposed that the large pangenome of *Prochlorococcus* species might originate from their large *Ne* (estimated between 10^6 (Price & Arkin, 2015) and 10^{13} (Kashtan et al., 2014)). However, in the case of structured populations, as observed here, the evolution of a near neutral gene acquired from distant lineage might be restricted to the clade in which it was introduced. Thus, distant HGT might evolve in a context of lower *Ne,* a pattern that might explain the ISL4 characteristics (*i.e,* enriched in COGs specific to one clade, involved in defense mechanisms and not affiliated to *Cyanobacteria*) (Figure 5B). However, the *Ne* of HGT could also be enlarged through, *e.g.*, local HR, favouring a selective footprint and the persistence of acquired genes with marginal effect. The occurrence of both HGT (high *dS* with *uncertain* affiliation) and selected genes (low *dN/dS* with *Prochlorococcus* affiliation) in ISL3 and ISL5 (Figure 5A) could be the result of such processes. Overall, our results highlighting two sets of genes with distinct evolutionary trajectories (*i.e.*, strong negative selection *versus* selection relaxation) in a structured population are in accordance with the drift-barrier model. Furthermore, the structuring of the genetic information along the genome might depend on the dynamics of gene fluxes among clades within a structured population, especially for flexible genes. Rather than a non-random acquisition of genes with regard to their genomic location, we may consider the differential retention probability of transferred genes as a consequence of fluctuating *Ne* along the genome.

Acknowledgements

The work of H.G. was supported by a PhD fellowship funded by the Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. We are grateful to Cécile Lepère with the improvement of the manuscript.

For Review Only

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. doi: 10.1016/S0022-2836(05)80360-2

Andreani, N. A., Hesse, E., & Vos, M. (2017). Prokaryote genome fluidity is dependent on effective population size. *The ISME Journal*, *11*(7), 1719–1721. doi: 10.1038/ismej.2017.36

Avrani, S., Wurtzel, O., Sharon, I., Sorek, R., & Lindell, D. (2011). Genomic island variability facilitates Prochlorococcus–virus coexistence. *Nature*, *474*(7353), 604–608. doi: 10.1038/nature10172

Baumdicker, F., Hess, W. R., & Pfaffelhuber, P. (2012). The Infinitely Many Genes Model for the Distributed Genome of Bacteria. *Genome Biology and Evolution*, *4*(4), 443–456. doi: 10.1093/gbe/evs016

Biller, S. J., Berube, P. M., Berta-Thompson, J. W., Kelly, L., Roggensack, S. E., Awad, L., … Chisholm, S. W. (2014). Genomes of diverse isolates of the marine cyanobacterium Prochlorococcus. *Scientific Data*, *1*, 140034. doi: 10.1038/sdata.2014.34

Biller, S. J., Berube, P. M., Berta-Thompson, J. W., Kelly, L., Roggensack, S. E., Awad, L., … Chisholm, S. W. (2014). Genomes of diverse isolates of the marine cyanobacterium Prochlorococcus. *Scientific Data*, *1*, 140034. doi: 10.1038/sdata.2014.34

Bobay, L.-M., & Ochman, H. (2018). Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evolutionary Biology*, *18*(1), 153. doi: 10.1186/s12862-018-1272-4

Cadillo-Quiroz, H., Didelot, X., Held, N. L., Herrera, A., Darling, A., Reno, M. L., … Whitaker, R. J. (2012). Patterns of Gene Flow Define Species of Thermophilic Archaea. *PLoS Biology*, *10*(2), e1001265. doi: 10.1371/journal.pbio.1001265

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*(1), 421. doi: 10.1186/1471-2105-10-421

Castillo-Ramírez, S., Harris, S. R., Holden, M. T. G., He, M., Parkhill, J., Bentley, S. D., & Feil, E. J. (2011). The impact of recombination on dN/dS within recently emerged bacterial clones. *PLoS Pathogens*, *7*(7), e1002129. doi: 10.1371/journal.ppat.1002129

Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution*, *17*(4), 540–552. doi: 10.1093/oxfordjournals.molbev.a026334

Coleman, M. L., Sullivan, M., Martiny, A., Steglich, C., Barry, K., DeLong, E., & Chisholm, S. (2006).

For Review Only

Genomic Islands and the Ecology and Evolution of Prochlorococcus. *Science*, *311*(5768), 1768–1770. doi: 10.1126/science.1122050

Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS ONE*, *5*(6), e11147. doi: 10.1371/journal.pone.0011147

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012, August 1). JModelTest 2: More models, new heuristics and parallel computing. *Nature Methods*, Vol. 9, p. 772. doi: 10.1038/nmeth.2109

dos Reis, M., & Yang, Z. (2013). The unbearable uncertainty of Bayesian divergence time estimation. *Journal of Systematics and Evolution*, *51*(1), 30–43. doi: 10.1111/j.1759-6831.2012.00236.x

Eddy, S. R. (2009). A new generation of homology search tools based on probalistic inference. *Genome Informatics 2009*, 205–211. doi: 10.1142/9781848165632_0019

Feingersch, R., Philosof, A., Mejuch, T., Glaser, F., Alalouf, O., Shoham, Y., & Béjà, O. (2012). Potential for phosphite and phosphonate utilization by Prochlorococcus. *ISME Journal*, *6*(4), 827–834. doi: 10.1038/ismej.2011.149

Flombaum, P., Gallegos, J. L., Gordillo, R. A., Rincón, J., Zabala, L. L., Jiao, N., … Martiny, A. C. (2013). Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(24), 9824–9829. doi: 10.1073/pnas.1307701110

Garud, N. R., Good, B. H., Hallatschek, O., & Pollard, K. S. (2019). Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS Biology*, *17*(1), e3000102. doi: 10.1371/journal.pbio.3000102

Gogarten, J. P., Doolittle, W. F., & Lawrence, J. G. (2002). Prokaryotic Evolution in Light of Gene Transfer. *Molecular Biology and Evolution*, *19*(12), 2226–2238. doi: 10.1093/oxfordjournals.molbev.a004046

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. doi: 10.1093/sysbio/syq010

Hanage, W. P. (2016). Not So Simple After All: Bacteria, Their Population Genetics, and Recombination. *Cold Spring Harbor Perspectives in Biology*, *8*(7), a018069. doi: 10.1101/cshperspect.a018069

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, *28*(1), 100. doi: 10.2307/2346830

Hill, W. G., & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research*, *8*(3), 269–294. doi: 10.1017/S0016672300010156

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., … Bork, P. (2016).

For review only

eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, *44*(D1), D286–D293. doi: 10.1093/nar/gkv1248

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., … Chisholm, S. W. (2014). Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus. *Science*, *344*(6182), 416–420. doi: 10.1126/science.1248575

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., … Chisholm, S. W. (2015). Data from: Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus. *Dryad Digital Repository*. doi: 10.5061/dryad.9r0p6

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. doi: 10.1093/molbev/mst010

Kelly, L., Huang, K. H., Ding, H., & Chisholm, S. W. (2012). ProPortal: A resource for integrated systems biology of Prochlorococcus and its phage. *Nucleic Acids Research*, *40*(D1), D632–D640. doi: 10.1093/nar/gkr1022

Kent, A. G., Baer, S. E., Mouginot, C., Huang, J. S., Larkin, A. A., Lomas, M. W., & Martiny, A. C. (2019). Parallel phylogeography of Prochlorococcus and Synechococcus. *ISME Journal*, *13*(2), 430–441. doi: 10.1038/s41396-018-0287-6

Kent, A. G., Dupont, C. L., Yooseph, S., & Martiny, A. C. (2016). Global biogeography of Prochlorococcus genome diversity in the surface ocean. *The ISME Journal*, *10*(8), 1856–1865. doi: 10.1038/ismej.2015.265

Kettler, G. C., Martiny, A. C., Huang, K., Zucker, J., Coleman, M. L., Rodrigue, S., … D'Haeseleer, P. (2007). Patterns and Implications of Gene Gain and Loss in the Evolution of Prochlorococcus. *PLoS Genetics*, *3*(12), e231. doi: 10.1371/journal.pgen.0030231

Klappenbach, J. A., Goris, J., Vandamme, P., Coenye, T., Konstantinidis, K. T., & Tiedje, J. M. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, *57*(1), 81–91. doi: 10.1099/ijs.0.64483-0

Kryazhimskiy, S., & Plotkin, J. B. (2008). The Population Genetics of dN/dS. *PLoS Genetics*, *4*(12), e1000304. doi: 10.1371/journal.pgen.1000304

Kuo, C.-H., & Ochman, H. (2009). The fate of new bacterial genes. *FEMS Microbiology Reviews*, *33*(1), 38–43. doi: 10.1111/j.1574-6976.2008.00140.x

Larkin, A. A., Blinebry, S. K., Howes, C., Lin, Y., Loftus, S. E., Schmaus, C. A., … Johnson, Z. I. (2016). Niche partitioning and biogeography of high light adapted Prochlorococcus across taxonomic ranks in the North Pacific. *The ISME Journal*, *10*(7), 1555–1567. doi: 10.1038/ismej.2015.244

For Review Only

Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics*, *26*(8), 345–352. doi: 10.1016/j.tig.2010.05.003

Malmstrom, R. R., Coe, A., Kettler, G. C., Martiny, A. C., Frias-Lopez, J., Zinser, E. R., & Chisholm, S. W. (2010). Temporal dynamics of Prochlorococcus ecotypes in the Atlantic and Pacific oceans. *The ISME Journal*, *4*(10), 1252–1264. doi: 10.1038/ismej.2010.60

Marttinen, P., Croucher, N. J., Gutmann, M. U., Corander, J., & Hanage, W. P. (2015). Recombination produces coherent bacterial species clusters in both core and accessory genomes. *Microbial Genomics*, *1*(5). doi: 10.1099/mgen.0.000038

Mayr, E. (1942). Systematics and the Origin of Species. *Annals of the Entomological Society of America*, *36*(1), 138–139. doi: 10.1093/aesa/36.1.138a

McInerney, J. O., McNally, A., & O'Connell, M. J. (2017). Why prokaryotes have pangenomes. *Nature Microbiology*, *2*(4), 17040. doi: 10.1038/nmicrobiol.2017.40

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

Moore, L. R., Rocap, G., & Chisholm, S. W. (1998). Physiology and molecular phylogeny of coexisting Prochlorococcus ecotypes. *Nature*, *393*(6684), 464–467. doi: 10.1038/30965

Ochman, H., Lawrence, J. G., & Groisman, E. a. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, *405*(6784), 299–304. doi: 10.1038/35012500

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055. doi: 10.1101/gr.186072.114

Parmley, J. L., & Hurst, L. D. (2007). How Common Are Intragene Windows with K A > K S Owing to Purifying Selection on Synonymous Mutations? *Journal of Molecular Evolution*, *64*(6), 646–655. doi: 10.1007/s00239-006-0207-7

Partensky, F., Hess, W. R., & Vaulot, D. (1999). Prochlorococcus, a marine photosynthetic prokaryote of global significance. *Microbiology and Molecular Biology Reviews : MMBR*, *63*(1), 106–127.

Price, M. N., & Arkin, A. P. (2015). Weakly Deleterious Mutations and Low Rates of Recombination Limit the Impact of Natural Selection on Bacterial Genomes. *MBio*, *6*(6), e01302-15. doi: 10.1128/mBio.01302-15

Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G., & Toth, I. K. (2016). Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Analytical Methods*, *8*(1), 12–24. doi: 10.1039/C5AY02550H

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics : TIG*, *16*(6), 276–277. doi: 10.1016/S0168-9525(00)02024-2

For Review Only

Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., … Chisholm, S. W. (2003). Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature*, *424*(6952), 1042–1047. doi: 10.1038/nature01947

Rocha, E. P. C., Smith, J. M., Hurst, L. D., Holden, M. T. G., Cooper, J. E., Smith, N. H., & Feil, E. J. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of Theoretical Biology*, *239*(2), 226–235. doi: 10.1016/J.JTBI.2005.08.037

Schmutzer, M., & Barraclough, T. G. (2019). The role of recombination, niche-specific gene pools and flexible genomes in the ecological speciation of bacteria. *Ecology and Evolution*, *9*(8), 4544–4556. doi: 10.1002/ece3.5052

Sela, I., Wolf, Y. I., & Koonin, E. V. (2016). Theory of prokaryotic genome evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(41), 11399–11407. doi: 10.1073/pnas.1614083113

Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., Szabó, G., … Alm, E. J. (2012). Population genomics of early events in the ecological differentiation of bacteria. *Science (New York, N.Y.)*, *336*(6077), 48–51. doi: 10.1126/science.1218198

Stolyar, S., & Marx, C. J. (2019). Align to Define: Ecologically Meaningful Populations from Genomes. *Cell*, *178*(4), 767–768. doi: 10.1016/J.CELL.2019.07.026

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., … Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America*, *102*(39), 13950–13955. doi: 10.1073/pnas.0506758102

Varghese, N. J., Mukherjee, S., Ivanova, N., Konstantinidis, K. T., Mavrommatis, K., Kyrpides, N. C., & Pati, A. (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Research*, *43*(14), 6761–6771. doi: 10.1093/nar/gkv657

Vos, M., & Eyre-Walker, A. (2017). Are pangenomes adaptive or not? *Nature Microbiology*, *2*(12), 1576–1576. doi: 10.1038/s41564-017-0067-5

Wiedenbeck, J., & Cohan, F. M. (2011). Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiology Reviews*, *35*(5), 957–976. doi: 10.1111/j.1574-6976.2011.00292.x

Wiser, M. J., Ribeck, N., & Lenski, R. E. (2013). Long-term dynamics of adaptation in asexual populations. *Science*, *342*(6164), 1364–1367. doi: 10.1126/science.1243357

Wolf, J. B. W., Künstner, A., Nam, K., Jakobsson, M., & Ellegren, H. (2009). Nonlinear Dynamics of Nonsynonymous (dN) and Synonymous (dS) Substitution Rates Affects Inference of Selection. *Genome Biology and Evolution*, *1*, 308–319. doi: 10.1093/gbe/evp030

Yan, W., Wei, S., Wang, Q., Xiao, X., Zeng, Q., Jiao, N., & Zhang, R. (2018). Genome Rearrangement

For Review Only

Shapes Prochlorococcus Ecological Adaptation. *Appl. Environ. Microbiol.*, *84*(17), e01178-18. doi: 10.1128/AEM.01178-18

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. doi: 10.1093/molbev/msm088

Yang, Z., & Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution*, *17*(1), 32–43. doi: 10.1093/oxfordjournals.molbev.a026236

Data Accessibility Statement

All data used for this manuscript are available and open access. Genomic sequences are available on the NCBI genome assembly database, BioProject PRJNA239833, PRJNA239872 and PRJNA239873. FTP addresses of these genomic sequences dowloaded from the NCBI are provided in Table S2. Sequences of genes classified into clusters of orthologous genes (COGs) that were used for the analyses in this manuscript are available on DRYAD https://datadryad.org/stash/dataset/doi:10.5061/dryad.9r0p6 .

Author contributions

G.B. and C.B.-P. conceived and coordinated this study. G.B., H.G. and C.B.-P. designed the study. H.G., G.B., C.B.-P and I.J.-D. performed the analysis and analyzed the data. H.G., G.B., C.B.-P. and I.J.-D. contributed to the writing and the reviewing of the manuscript.

Competing interests

The authors declare no competing interests.

For Review Only

Figures

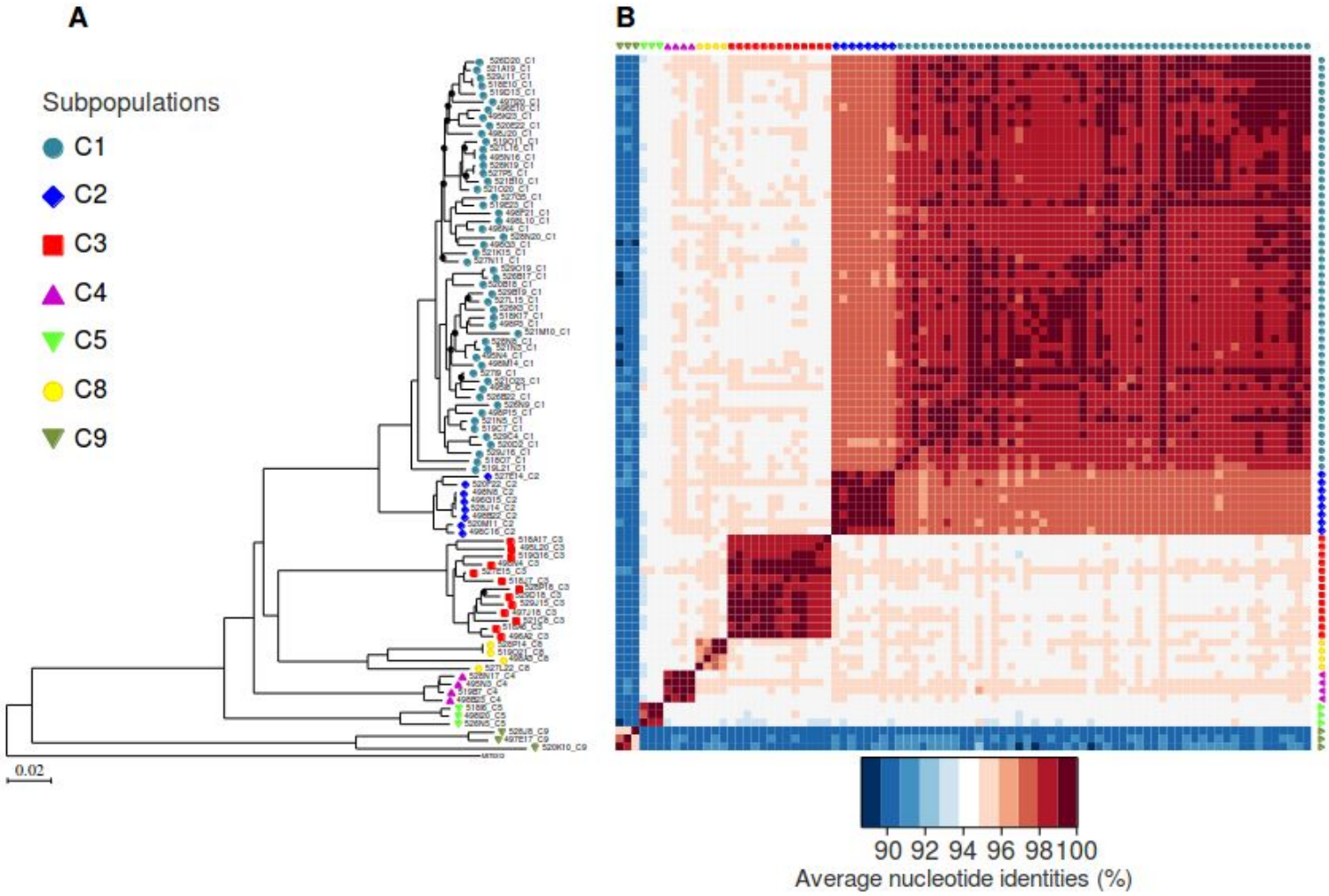


Figure 1: Phylogenetic relationships of 87 *Prochlorococcus* HLII ecotype single-amplified genomes (SAGs) distributed over seven major subpopulations (C1 to C5, C8 and C9). (A) The maximum likelihood phylogenetic tree inferred from the concatenated alignment of 1,202 single-copy core genes for all selected SAGs (within cN2 cluster: 52 C1, eight C2, 13 C3, four C4, three C5; within c9301 cluster: four C8 and within cN1 cluster: three C9). The reference genome MIT9312 was used to root the tree. Bootstrap supports < 80% are marked by black dots on the internal nodes. (B) Heatmap describing the pairwise average genome-wide nucleotide identity (ANI) (%) between SAG. Rows and columns are arranged according to the phylogenetic tree. Coloured dots used for both figures represent the different clades.

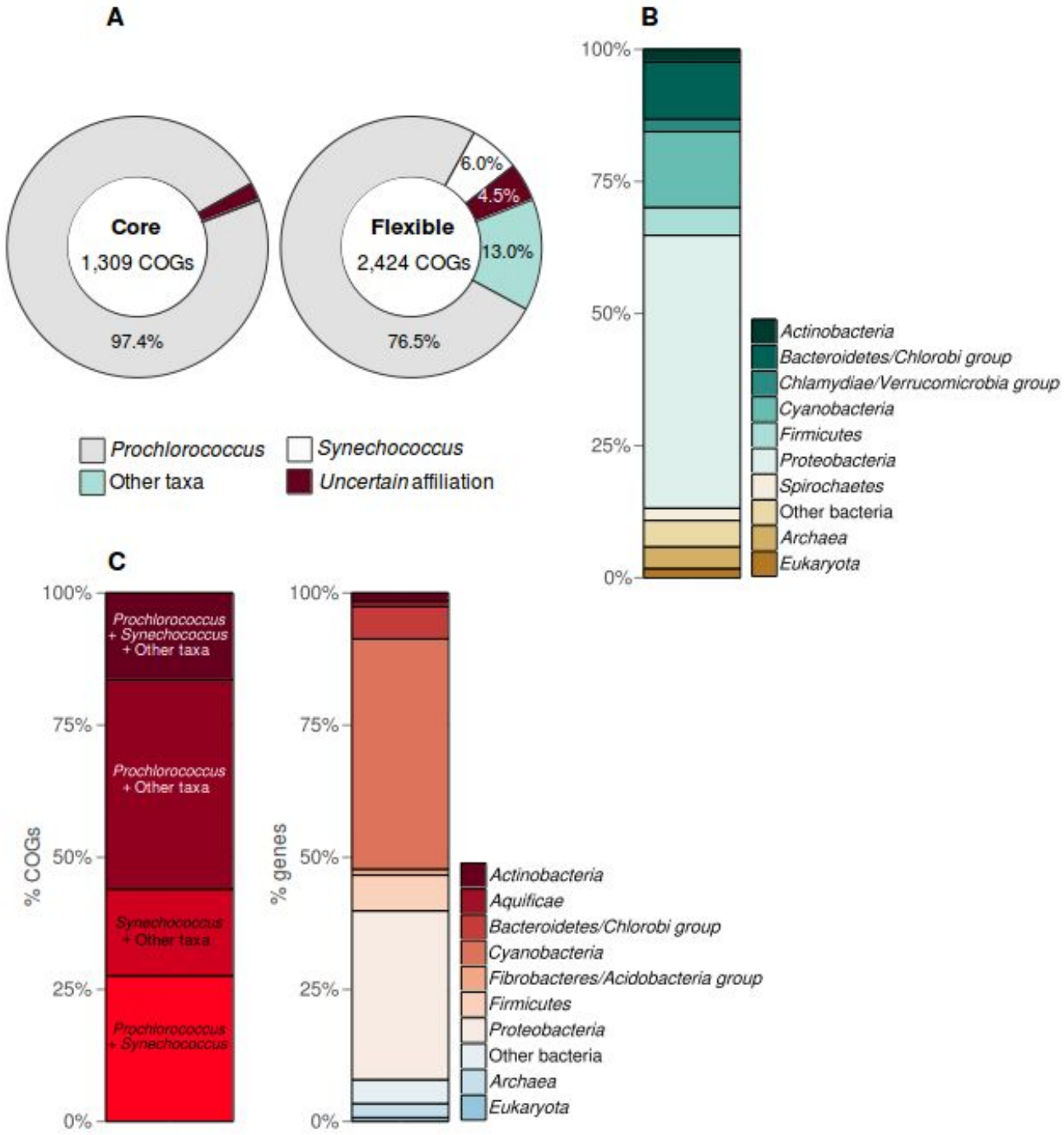
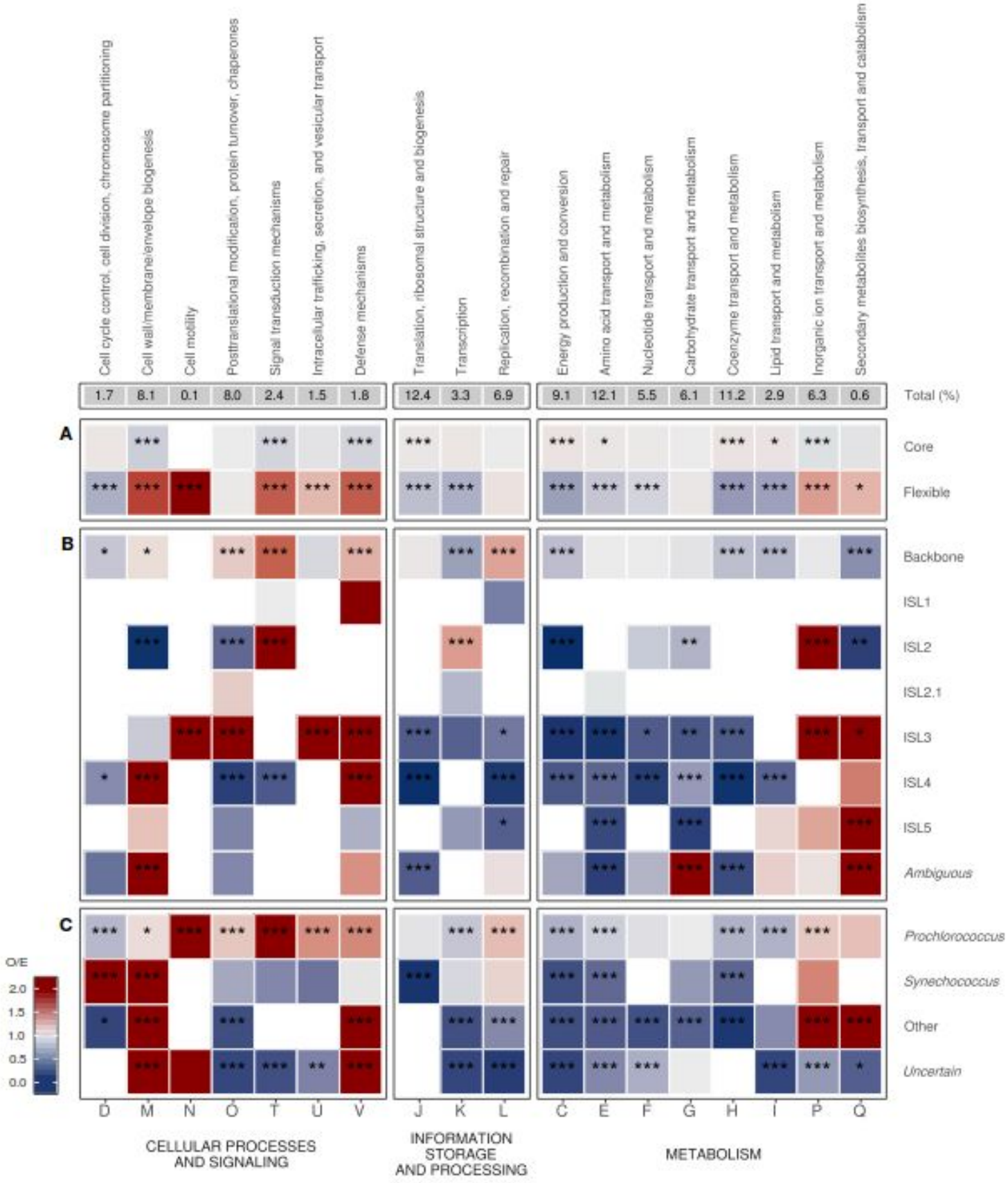


Figure 2: Taxonomic distributions of core and flexible clusters of orthologous genes (COGs) identified in SAGs. (A) Taxonomic affiliations of core and flexible COGs highlighting the proportion of those affiliated with the genera *Prochlorococcus* or *Synechococcus* or other taxonomic groups or having an *uncertain* affiliation, *i.e.*, containing genes assigned to various taxonomic groups, including *Prochlorococcus* and/or *Synechococcus*. (B) Taxonomic distributions of flexible COGs assigned to other taxa. (C) Taxonomic distributions of COGs tagged as *uncertain* with the proportion of those containing genes affiliated with *Prochlorococcus* and/or *Synechococcus* with or without other taxa (left) and the composition and abundance of genes within the category of other taxa (right). Bacterial taxa with less than 1% of abundance are grouped in the category "other bacteria" (B and C).


Cell cycle control, cell division, chromosome partitioning
Cell wall/membrane/envelope biogenesis
Cell motility
Posttranslational modification, protein turnover, chaperones
Signal transduction mechanisms
Intracellular trafficking, secretion, and vesicular transport
Defense mechanisms
Translation, ribosomal structure and biogenesis
Transcription
Replication, recombination and repair
Energy production and conversion
Amino acid transport and metabolism
Nucleotide transport and metabolism
Carbohydrate transport and metabolism
Coenzyme transport and metabolism
Lipid transport and metabolism
Inorganic ion transport and metabolism
Secondary metabolites biosynthesis, transport and catabolism
1.7 8.1 0.1 8.0 2.4 1.5 1.8 12.4 3.3 6.9 9.1 12.1 5.5 6.1 11.2 2.9 6.3 0.6
Total (%)
A
Core
Flexible
B
Backbone
ISL1
ISL2
ISL2.1
ISL3
ISL4
ISL5
Ambiguous
C
Prochlorococcus
Synechococcus
Other
Uncertain
O/E
2.0
1.5
1.0
0.5
0.0
D M N O T U V J K L C E F G H I P Q
CELLULAR PROCESSES AND SIGNALING
INFORMATION STORAGE AND PROCESSING
METABOLISM

1

Figure 3: COG functional annotations and distributions according to their category (A), location (B) and affiliation (C). The total percentages of genes (%) assigned to each EggNOG functional category (symbolized by a capital letter) are indicated. Observed/expected (O/E) ratios of core (A) and flexible (A-C) genes, according to their genomic location (backbone, genomic islands or *ambiguous*) in (B) and their taxonomic affiliation (*Prochlorococcus*, *Synechococcus*, other or *uncertain* taxonomic groups) in (C). The observed values (O) correspond to the number of genes assigned to each functional category. The expected values (E) were obtained by multiplying the number of genes (core, flexible, or flexible genes as a function of their genomic location or taxonomic affiliation) by the total percentage of genes in each functional category. The white boxes indicate the lack of genes involved in the functional category considered. Differences in the distribution of functional categories between core and flexible genes (A), over all genomic locations (B) and over all taxonomic affiliations (C) were tested using chi-squared tests ($p<0.005$ for all A, B and C groupings). Chi-squared tests were also performed for each line in the figure (each category against all others at once) to test the significance of enrichment for a given location or a given taxonomy. Chi-squared test: *, *p-value* < 0.05; **, *p-value* < 0.01; ***, *p-value* < 0.005. ISL: genomic island.

For Review Only

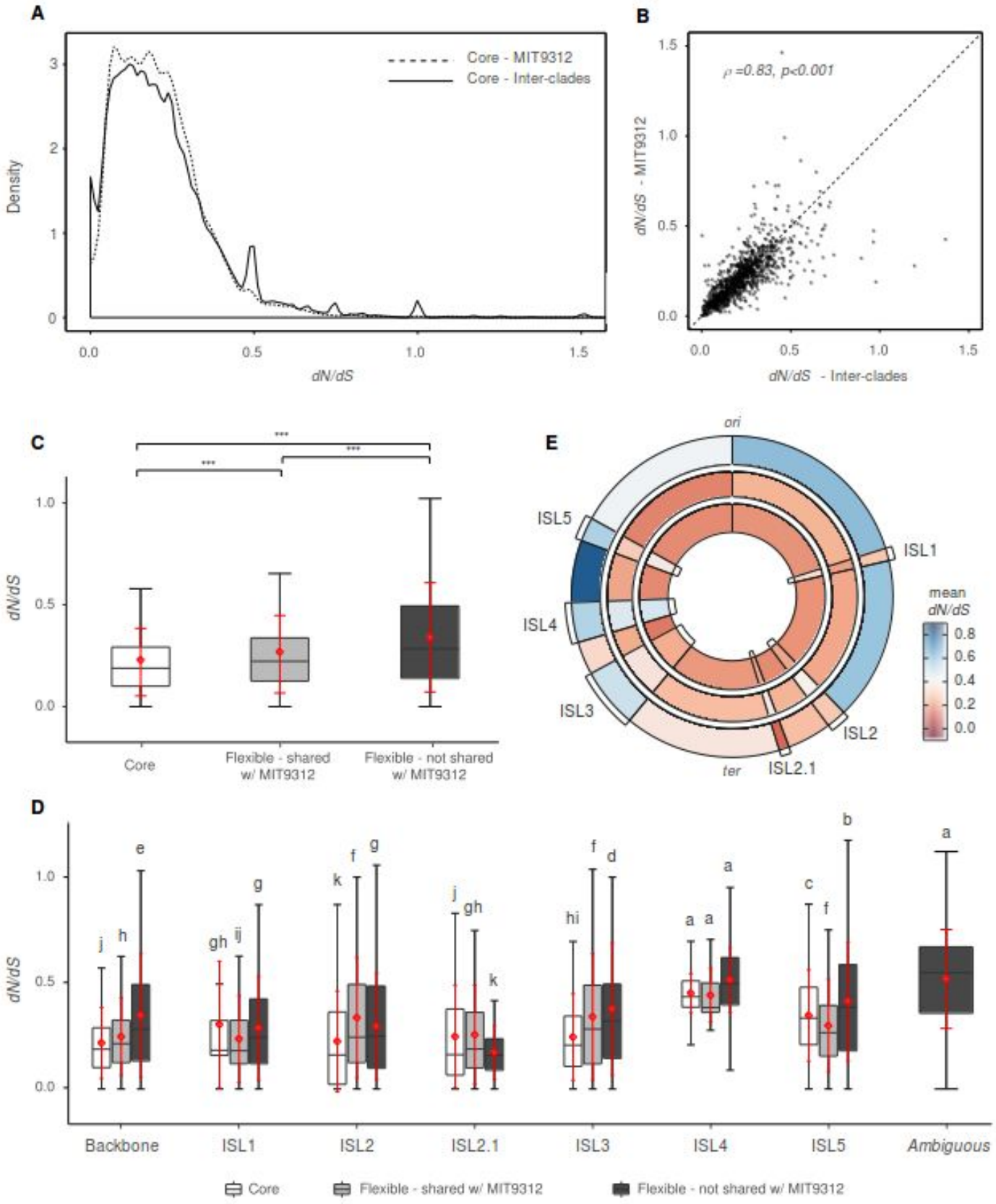


Figure 4: Selective pressure according to the genomic compartments. (A) Density of *dN/dS* ratios

estimated for core COGs found in the backbone. Dashed line: values for MIT9312-clade pairwise comparisons; solid line: values for inter-clade pairwise comparisons. (B) Correlation between mean *dN/dS* of MIT9312-clade pairwise comparisons and mean *dN/dS* of inter-clade pairwise comparisons. The points represent mean *dN/dS* values and were averaged per COG. Spearman's rank correlation test (ρ) and the associated *p-value* are indicated. The dashed line symbolizes the diagonal. (C) Boxplot of *dN/dS* value distributions in the core genes and flexible genes shared or not shared with MIT9312 (Kruskal-Wallis test and post hoc Wilcoxon signed-rank test with Bonferroni correction; ***, *p-value* < 0.001). (D) Boxplot showing *dN/dS* value distributions over the genomic compartments (backbone, genomic islands and *ambiguous*). Significant differences [thus categorizing compartments with similar values] are indicated by lowercase letters, a > b > c > d > e > f > g > h > i > j > k (Kruskal-Wallis test and post hoc Wilcoxon signed-rank test with Bonferroni correction, *p-value* < 0.05). (C) and (D) For a better understanding, only *dN/dS* <1.5 are shown; in red: mean ± the standard deviation (sd); white: core genes, light grey: flexible genes shared with MIT9312 and dark grey: flexible genes not shared with MIT9312. (E) Representation of mean *dN/dS* values along the chromosome (as organized in the MIT9312 reference genome) computed for genes within core COGs (inner circle) and flexible shared (middle circle) and not shared (outer circle) with MIT9312. The origin (*ori*) and terminus (*ter*) of replication are shown. ISL: genomic island.

For Review Only

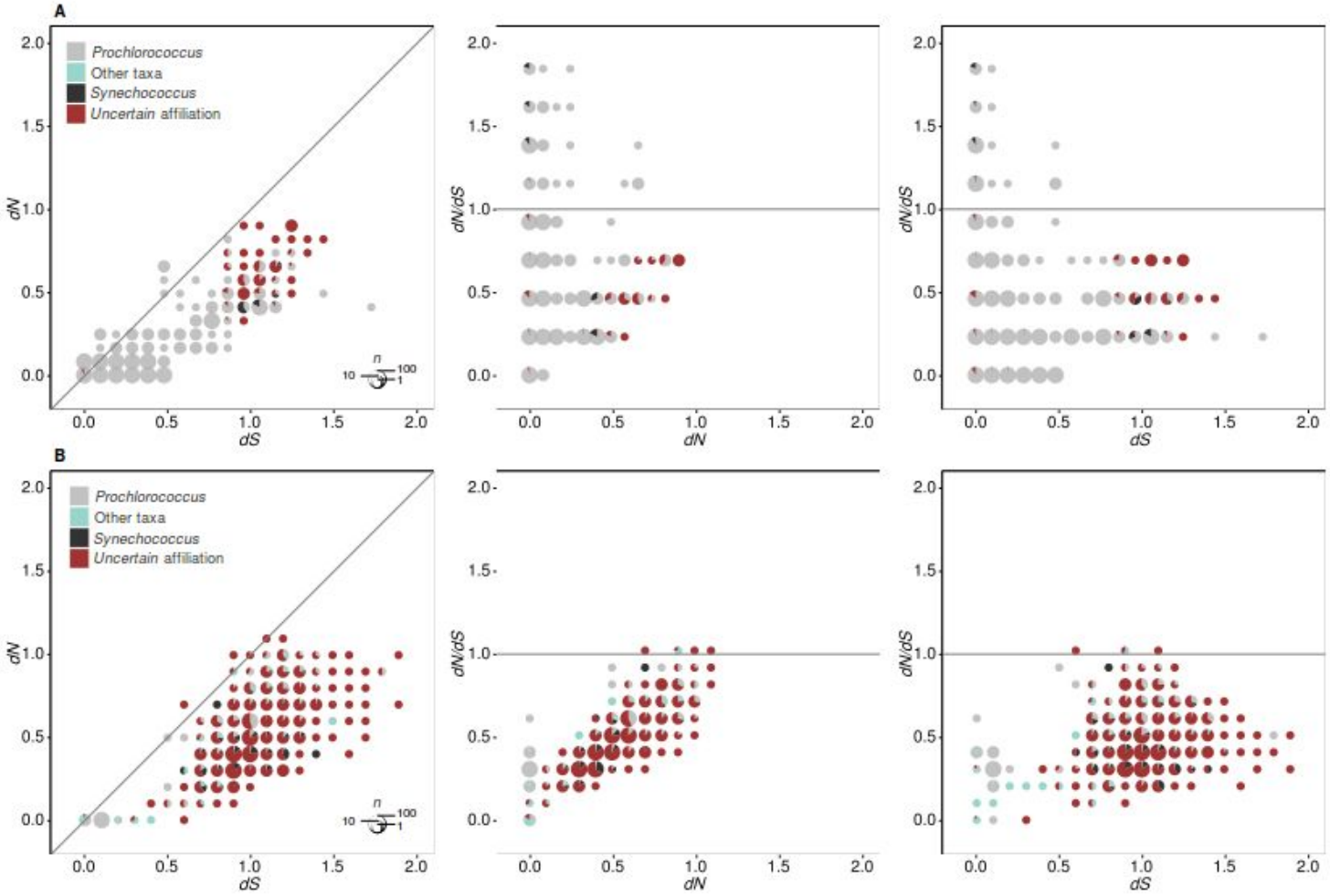


Figure 5: Relationships between substitution rates in flexible COGs and their taxonomic affiliations. Taxonomic affiliation – *i.e.*, *Prochlorococcus*, *Synechococcus*, other taxonomic groups or *uncertain* – of flexible COGs in ISL3 and ISL5 (A) and ISL4 and tagged as *ambiguous* (B) plotted against *dN*, *dS* and *dN/dS* estimates. The pie charts depict the distribution of taxonomic affiliation in a 2D region of the graph corresponding to *dN*, *dS* and *dN/dS* values within a range of 0.1. The size of each pie chart is proportional to the number of observations *n* for the 2D region considered (at least one observation, from 10 to 100 observations, more than 100 observations). Horizontal and diagonal lines in each panel represent a *dN/dS* ratio equal to 1. Left: *dN versus dS*, Middle: *dN/dS versus dN*, Right: *dN/dS versus dS*.